Traiter et transmettre l'information à hautdébit, à énergie réduite, avec flexibilité : le grand écart dans trois directions

Matthieu Arzel

UNIVERSITÉ DE BRETAGNE SUD
ÉCOLE DOCTORALE MATHSTIC

Traiter et transmettre l'information à haut débit, à énergie réduite, avec flexibilité : le grand écart dans trois directions.

Mémoire présenté le 7 janvier 2021 par

Matthieu ARZEL

en vue de l'obtention de l'

HABILITATION À DIRIGER DES RECHERCHES

Mention : "Sciences et Technologies de l'Information et de la Communication"

devant le jury composé de :

Virginie FRESSE Maître de conférence HDR, Université Jean Monnet Saint Etienne	Rapporteur
Christophe JEGO Professeur, Bordeaux INP ENSEIRB MATMECA	Rapporteur
Olivier ROMAIN Professeur, CERGY-PARIS Université	Rapporteur
Frédéric ROUSSEAU Professeur, Université Grenoble Alpes	Président du jury
Emmanuel BOUTILLON Professeur, Université Bretagne Sud	Examinateur
Michel JEZEQUEL Professeur, IMT Atlantique	Examinateur
Fan YANG Professeur, Université de Bourgogne	Examinateur

Équipe d'acceuil : Laboratoire CNRS Lab-STICC UMR 6285
Composante universitaire : UFR Sciences et Sciences de l'ingénieur - Université Bretagne Sud

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Université Bretagne Sud

Lab-STICC

Matthieu ARZEL

Traiter et transmettre l'information à haut débit, à énergie réduite, avec flexibilité : le grand écart dans trois directions.

Résumé

Enseignant-chercheur à IMT Atlantique au sein du département Électronique, et membre du Lab-STICC (CNRS, UMR 6285), pôle Communications, Architecture et Circuits (CACS), je présente dans ce document mon activité depuis mon doctorat. Je détaille notamment mon implication en enseignement tout au long des différentes formations d'ingénieurs d'IMT Atlantique et les grands axes de mes travaux de recherche. Ceux-ci ont concerné l'adéquation algorithme-architecture en allant jusqu´à l'adéquation au niveau du signal physique et du mode de représentation électrique adapté à l'information traitée. Ainsi, j'ai investigué des solutions électroniques mixtes analogique-numérique et stochastiques pour traiter les signaux dans un récepteur de télécommunications mais aussi pour intégrer des réseaux de neurones. L'association information-signal-transistor a été un objet de réflexion depuis ma thèse et m'a permis d'explorer le triptyque débit-consommation de ressources (matérielles et énergétiques)-flexibilité. J'ai ainsi montré qu'une association bien pensée information-signal-transistor peut permettre d'améliorer conjointement au moins deux axes, bien qu'ils paraissent souvent orthogonaux. Tenter le grand écart simultanément dans trois directions d'amélioration de l'état de l'art est ainsi intellectuellement faisable.

Les télécommunications sont certes mon domaine de prédilection mais ne sont pas l'unique objet de mes recherches. Je les ai progressivement étendues à des applications biomédicales et plus récemment à l'intelligence artificielle pour les systèmes autonomes. Ces travaux de recherche ont fait l'objet de nombreuses collaborations, académiques et industrielles, nationales et internationales, et ont permis de former douze docteurs diplômés à cette date, dont plusieurs ont fait le choix de poursuivre d'ailleurs dans l'enseignement-recherche.

Table des matières

Table des figures

Acronymes

ACAP Adaptive Compute Acceleration Platform. 126

ADC Analogue to Digital Converter. 83

ALU Arithmetic Logic Unit.

APP *A Posteriori* Probability. 61

APSK Amplitude and Phase Shift Keying. 75

ASIC Application Specific Integrated Circuit. vii, xii, 52, 66, 74, 79, 85, 91, 94, 112–115, 117, 122, 125

ASIP Application Specific Instruction-set Processor. 74

AWGN Additive White Gaussian Noise. 55, 56

BAN Body Area Network. 49

BCH Bose Ray-Chaudhuri Hocquenghem. 68

BCJR Bahl Cocke Jelinek Raviv. 69, 70

BER Bit Error Rate. 70

BiCMOS Bipolar and Complementary MOS. 61, 65

BJT Bipolar Junction Transistor. 55, 56

BP Belief-Propagation. 75

BPSK Binary Phase-Shift Keying. 55

CCSDS Consultative Committee for Space Data Systems. 81

CMOS Complementary MOS. xii, 65, 74, 109, 110, 113–119

COP Center Of Pressure. 108–110

CPU Central Processing Unit. 47, 48, 84, 91, 94, 101, 126

DAC Digital to Analogue Converter. 83

DDoS Distributed Denial of Service. 90

DSP Digital Signal Processor (-ing). 79, 85

DVB Digital Video Broadcasting.

DVB-RCS Digital Video Broadcasting - Return Channel by Satellite. vii, 51, 61, 65, 82

DVB-S Digital Video Broadcasting - Satellite. 82

DVB-S2 Digital Video Broadcasting - Satellite - Second generation. 82

DVB-S2X Digital Video Broadcasting - Satellite - Second generation eXtension. 82

EEG Électro-EncéphaloGraphie. xii, 49, 110–113

EM Edge Memory. 69, 70

ESA European Space Agency. 66, 82

EXIT EXtrinsic Information Transfer. 75, 76

FER Frame Error Rate.

FPGA Field-Programmable Gate Array. vii, viii, 52, 68, 74, 77–80, 84, 85, 89, 91–96, 98, 99, 101–104, 112, 113, 121, 124–126, 128–130

Gbps Giga bits per second.

GPGPU General Purpose computing on Graphics Processing Units. 93

GPP General Purpose Processor. 79, 93, 111

GPU Graphics Processing Unit. 79, 112, 113, 122, 126

Gsps Giga samples per second.

HDL Hardware Description Language. 103

HLS High Level Synthesis. 103

ICM Interface Cerveau Machine. 110, 111

IFoIP Intermediate Frequency over Internet Protocol. 82

IoT Internet of Things. 90

IP Internet Protocol. 82, 84, 86, 87

LDPC Low-Density Parity-Check. xi, 54, 59, 60, 65, 66, 68, 69, 73–76, 79

LLR Log-Likelihood Ratio. 58

LTE Long Term Evolution. 74, 75

MAP Maximum A Posteriori. 74

Mbps Mega bits per second.

MIMO Multiple-Input Multiple-Output. vii, xi, 51, 60, 75, 76

ML Maximum Likelihood. 76

MO Medial Offset. xii, 108

MOS Metal Oxyde Semiconductor. 56

MOSFET Metal Oxyde Semiconductor Field Effect Transistor. 55

MPLS Multiprotocol Label Switching. 87

Msps Mega samples per second.

NB-LDPC Non Binary Low-Density Parity-Check. 75, 76

NDS Noise Dependent Scaling. 69

NIC Network Interface Card.

NMOS N-type MOS. 115

NoC Network on Chip.

NPU Network Processing Unit.

NRZ Non-Return to Zero.

OFDM Orthogonal Frequency Division Multiplexing.

OOK On-Off Keying.

OSI Open System Interconnection.

P-FEC Packet Forward Error Coding. 87, 88

PCB Printed Circuit Board.

PCI Peripherical Component Interconnect.

PCIe Peripherical Component Interconnect express.

PE Polyéthylène.

PMOS P-type MOS. 115

PON Passive Optical Network.

PSK Phase Shift Keying. 75

QAM Quadrature Amplitude Modulation. 75

QC-LDPC Quasi-Cyclic Low-Density Parity-Check. 79, 80

QoS Quality of Service.

RAM Random Access Memory.

RAN Radio Acces Network. xi, 81, 82, 85–88

RFoIP Radio-Frequency over Internet Protocol. xi, 82–85, 87

RFSoC Radio Frequency System on Chip. 83

RNG Random Number Generator.

RS Reed-Solomon. 87, 88

RTP Real-Time Protocol. 84, 87

S/H Sample and Hold.

SCCC Serially Concatenated Convolutional Code.

SD Sphere Decoding.

SDN Software-Defined Network.

SDR Software-Defined Radio.

SIMD Single Instruction Multiple Data.

SISO Soft-Input Soft-output (decoder) *ou* Single-Input Single-Output (par opposition à MIMO).

SLL Super Long Line.

SLR Super Logic Region.

SM Spatial Multiplexing.

SNR Signal-to-Noise Ratio. 56

SoC System on Chip.

SOVA Soft-Output Viterbi Algorithm.

SSI Stacked Silicon Interconnect.

STBC Space-Time Block Code.

SVM Support Vector Machine.

T/H Track and Hold.

Tbps Tera bits per second.

TCP Transmission Control Protocol. 87

TEB Taux d'Erreur Binaire (BER en anglais). xi, 63

TEP Taux d'Erreur Paquet (FER en anglais). xi, 63

TFM Tracking Forecast Memory.

TPU Tensor Processing Unit.

Tsps Tera samples per second.

UDP User Datagram Protocol. 84, 87

UMTS Universal Mobile Telecommunications System.

VHDL VHSIC Hardware Description Language. 103

VHSIC Very High Speed Integrated Circuits.

VLSI Very Large Scale Integration.

VPLS Virtual Private LAN Service. 87

VPN Virtual Private Network. 87

WAN Wide Area Network. 86

XDP eXpress Data Path.

Première partie

Présentation du parcours professionnel

1

Curriculum vitæ

1.1 État civil

Matthieu ARZEL
Né à Brest, le 5 mai 1978
Marié, 3 enfants

Maître de conférences
IMT Atlantique Bretagne – Pays de la Loire
Département Électronique
Technopôle Brest-Iroise – CS 83818
29 238 Brest Cedex 3 – France

Téléphone : 02 29 00 13 11
Courriel : matthieu.arzel@imt-atlantique.fr
Page web : https://cv.archives-ouvertes.fr/matthieu-arzel

1.2 Résumé

Enseignant-chercheur à IMT Atlantique et membre du Lab-STICC (CNRS, UMR 6285), j'équilibre mes activités entre la formation, la recherche académique et la recherche contractuelle avec des entreprises partenaires. Cela m'a permis de développer une expertise en

- conception de circuits dédiés mixtes analogiques-numériques (4 puces) pour des systèmes flexibles et basse-consommation dans les domaines des communications numériques, des réseaux de neurones artificiels et du génie biomédical,
- conception d'architectures de circuits reconfigurables pour les communications numériques haut-débit, la sécurité des réseaux informatiques et l'apprentissage-machine,
- techniques de traitement pour les communications numériques, notamment en codes correcteurs d'erreur et solutions de chaînes d'émission-réception mobiles, optiques et satellitaires.

Cette recherche me permet d'illustrer et d'orienter la formation en Électronique, et plus largement en Télécommunications, selon les besoins avérés de l'industrie et ceux que je pressens.
J'utilise aussi ce lien industriel pour orienter et placer les élèves d'IMT Atlantique en stages, thèses CIFRE et premier emploi, principalement en France, Norvège, Allemagne et Canada. Pour cela, il faut connaître ses élèves, ce que je fais au travers de mes nombreux enseignements, favorisant les activités de mini-projets et projets pour garder un lien étroit. Curieux d'innovations pédagogiques, je diversifie mes modes d'enseignement et me suis impliqué dans plusieurs réformes de la formation dans mon établissement. J'ai par ailleurs co-publié et présenté des travaux conjoints en conférences sur la pédagogie dans l'enseignement supérieur. J'ai pris des responsabilités de domaine, puis celles d'adjoint à l'enseignement au chef du département Électronique. Je suis actuellement membre du Comité de Pilotage de la Formation à IMT Atlantique.

1.3 Formation

Doctorat en Sciences de l'Ingénieur, mention "Très honorable avec les félicitations du jury" — **22 juin 2006**, Brest, France

Ecole Nationale Supérieure des Télécommunications (ENST) de Bretagne en habilitation conjointe avec l'Université de Bretagne Sud

Semi-iterative analogue turbo decoding —an application to DVB-RCS-like codes

Rapporteurs :	Hans-Andrea LOELIGER	Professeur à l'ETH Zürich
	Georges ALQUIÉ	Professeur à l'Université Pierre et Marie Curie
Examinateurs :	Emmanuel BOUTILLON	Professeur à l'Université de Bretagne Sud
	Michel JÉZÉQUEL	Directeur d'Études à l'ENST Bretagne, directeur de thèse
	Cyril LAHUEC	Maître de conférences à l'ENST Bretagne, encadrant
	Fabrice SEGUIN	Maître de conférences à l'ENST Bretagne, encadrant
Invité :	Patrice GAMAND	PHILIPS Semiconductors

Diplôme d'Études Approfondies (DEA) en Électronique, mention "Bien" — **2002**, Lorient, France

Université de Bretagne Sud

Diplôme d'ingénieur — **2002**

Ecole Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France

Etudes couplées à de nombreux stages, dont une année de ***césure****, pour un total de 24 mois en entreprises (Thalès SA, Coframi/Alcatel, TurboConcept) et laboratoire (TUM, Münich, équipe de Joachim Hagenauer).*

1.4 Expérience professionnelle

Enseignant-chercheur, Maître de conférences, département Électronique
ENST Bretagne → Télécom Bretagne → IMT Atlantique Bretagne –Pays de la Loire

Nov. 2006-Présent
Brest, France

- Co-responsable des enseignements en Électronique et Physique (ELP) depuis 2015, puis adjoint à l'enseignement du département Électronique
- Membre du Comité de Pilotage de la Formation et du pôle Compétences de l'IMT Atlantique
- Membre du Lab-STICC (CNRS UMR 6285) / pôle Communications, Architectures, Circuits et Systèmes (CACS)/ équipe Interaction Algorithme Silicium (IAS) puis du pôle Traitement et Transmission de l'Information, algorIthme et Intégration (T2I3) / équipe Algorithm/Architecture Interactions (2AI)
- Encadrant de 12 docteurs diplômés et de 2 doctorants actuellement
- Membre du comité de programme technique du IEEE International Symposium on Turbo Codes & Iterative Information Processing '10, '16, '18
- Relecteur pour IEEE Transactions on Signal Processing, EURASIP Journal on Advances in Signal Processing,IEEE Transactions on Circuits and Systems, IEEE Transactions on VLSI Systems, IET Electronics Letters, IEEE Transactions on Communications, IEEE Communications Letters, IEEE ISCAS, IEEE NEWCAS, IEEE ICECS
- Contributeur/Responsable scientifique de contrats de recherche institutionnels pour l'ESA (European Space Agency), l'ANR (Agence Nationale de la Recherche), du LABEX Cominlabs et de contrats de recherche industriels (Orange Labs, EMC Norway, Widenorth, OVH, Huawei)

Ingénieur R&D
TurboConcept
Concepteur d'architectures et de circuits de synchronisation et de correction d'erreur pour émetteurs-récepteurs numériques sur FPGA.

Déc. 2005 - Nov. 2006
Plouzané, France

1.5 Fonctions d´élu au sein de l'Institut Mines-Télécom

Représentant du collège Maîtres de conférences et ingénieurs d'études — **Déc. 2012 - Juin 2017**
Conseil d'École de Télécom Bretagne

Représentant du collège Maîtres de conférences et ingénieurs d'études — **Déc. 2013 - Juin 2017**
Comité de l'Enseignement de Télécom Bretagne

Représentant du collège Maîtres de conférences et ingénieurs d'études — **Déc. 2012 - Déc. 2015**
Comité d'évaluation des appellations de Télécom Bretagne

2 Enseignement

2.1 Résumé de mes activités

Depuis mon recrutement à l'ENST Bretagne, je me suis impliqué fortement dans l'enseignement sur l'**intégralité des cursus Ingénieur**, aussi bien sur la formation d'ingénieur en partenariat (FIP) que sur la formation d'ingénieur généraliste (FIG), devenue formation d'ingénieur sous statut étudiant (FISE) depuis la création d'IMT Atlantique (1er janvier 2017).

J'ai enseigné dans un **large panel de disciplines**, en électronique analogique (fondamentaux d'adaptation, amplification, filtrage et théorie des lignes) et numérique (des bases à des techniques avancées, à l´état de l'art de l'ingénierie), en informatique (bases de programmation, environnement Linux) sur différentes applications de traitement du signal. J'ai par ailleurs encadré des **projets** sur tout le cursus (de la première à la troisième année) sur ces thématiques ainsi que sur des applications réseau et de traitement d'antenne. Cette **pluri-disciplinarité** a été permise par mes études antérieures en Télécommunications et ma curiosité naturelle.

Rapidement en charge de modules, j'ai participé de manière active à différentes **réformes de l'enseignement**. La première en 2010 concernait la refonte du tronc commun de la FIG, avec une recherche de trans-disciplinarité au sein des unités d'enseignement (UE). J´ai fait partie du comité à l'origine de cette réforme et ai pris par la suite la responsabilité d'une UE

trans-disciplinaire “Outils de l’ingénieur” (60h), organisée en 4 modules (Codage, logique et langage C ; Méthodes et outils de calcul numérique ;Maîtrise de l’environnement numérique de travail ;Bases de la programmation objet). Mes responsablités croissant, j’ai accepté la **co-responsablité de l’un des 5 domaines disiciplinaires, Électronique et Physique (ELP)**, en 2015, dont j’ai piloté les différentes évolutions jusqu’à la réforme suivante, issue de la création d’IMT Atlantique.

Je suis aussi **membre du groupe Compétences et Pédagogie** depuis 2013, devenu **Pôle Compétences** en 2018. Je contribue ainsi à la mise en place de l’évaluation par compétences et la nécessaire re-considération des enseignements. Auparavant, une note pour toute une unité d’enseignement était un indicateur global, certes pondéré, mais discriminant des connaissances et peu les compétences, c’est-à-dire des savoir-faire permettant de résoudre des problèmes complexes, requérant une réflexion élaborée, croisant différents savoirs, en fait ce que l’ on attend d’un ingénieur. IMT Atlantique a élaboré tout un corpus de compétences auquel j’ai contribué. Ce corpus décrit des compétences spécifiques à des domaines d’ingénierie (comme une d’électronique ANALYSER, DIMMENSIONNER ET METTRE EN ŒUVRE LES MONTAGES FONDAMENTAUX DE L’ÉLECTRONIQUE ET LES CHAÎNES D’INSTRUMENTATION), des compétences spécifiques transverses (comme POSER, ANALYSER, REFORMULER, STRUCTURER ou TRAVAILLER ET APPRENDRE ENSEMBLE) et des compétences génériques (comme COMPRENDRE ET ANALYSER, SYNTHÉTISER UN PROBLÈME ET/OU UNE SITUATION COMPLEXES ou CONCEVOIR ET RÉALISER DES SYSTÈMES ET DES ORGANISATIONS) qui intègrent les deux précédentes catégories.

Dès lors, évaluer (correctement) par compétences requiert de revoir l’alignement pédagogique et les situations d’apprentissage pour permettre ces évaluations. En définissant et listant les compétences que nous souhaitons évaluer, il faut aussi s’assurer que les activités pédagogiques et la stratégie d’évaluation soient cohérentes. En outre, il faut que cet alignement soit visible tant par les enseignants que par les élèves, ce qui me semble le point le plus difficile à concrétiser. J’ai adopté ce mode d’évaluation pour les UE dont je suis responsable et tente, commes les autres membres du Pôle, d’y former mes collègues avec le soutien de nos Ingénieurs Pédagogiques.

En outre, depuis 2018, je suis **adjoint à l’enseignement du chef du département Électronique**. A ce titre, je suis **membre du Comité de Pilotage de la formation**. Je m’assure de l’engagement de mon département dans les différentes formations, et de son équilibre en son sein, gérant ainsi les ressources de vacation par les doctorants et la charge de mes collègues, coordonnant l’évolution des enseignements. Je suis aussi l’interface avec les autres départements et la Direction de la Formation et de la Vie Scolaire (DFVS), portant la parole des uns et des autres et contribuant à l’évolution permanente de la formation. Au sein de

mon département, je suis consulté sur les besoins présents et futurs pour les recrutements des enseignants-chercheurs.

Pour finir, j'ai contribué à des activités d'enseignement innovantes que nous avons présentées dans différentes conférences nationales et internationales sur la pédagogie dans l'enseignement supérieur.

2.2 Publications en innovation pédagogique

- Myriam Le Goff-Pronost, Matthieu Arzel, Antoine BEUGNARD, Jean-Philippe COUPEZ, François GALLÉE, Claire LASSUDRIE, Michel Morvan, Bruno Vinouze, Richard NAËL et Didier BAUX. "Introducing complexity into project management through multi-stakeholder interactions". In : *SEFI 2014 : 42th annual conference*. Birmingham, United Kingdom : SEFI, sept. 2014, p. 135-135. URL : https://hal.archives-ouvertes.fr/hal-01170285
- Claire LASSUDRIE, Marie-Pierre Adam, Matthieu Arzel, Antoine BEUGNARD, Jean-Philippe COUPEZ, François GALLÉE, Sylvie KEROUEDAN, Myriam Le Goff-Pronost, Michel Morvan, Bruno Vinouze et Didier BAUX. "Score distribution as a tool to reveal group dynamics in student projects ?" In : *SEFI 2015 : Annual Conference of the European Society for Engineering Education*. T. SEFI 2015. Orléans, France : SEFI, 39 rue des Deux Eglises, 1000 Brussels, BELGIUM, juin 2015, p. 111-111. URL : https://hal.archives-ouvertes.fr/hal-01185711
- Marie-Pierre Adam, Matthieu Arzel, Antoine BEUGNARD, Jean-Philippe COUPEZ, François GALLÉE, Claire LASSUDRIE, Myriam Le Goff-Pronost, Michel Morvan, Bruno Vinouze et Didier BAUX. "Analyse d'une formation à la conduite de projets selon une grille de maturité de processus". In : *QPES 2015 : Colloque Questions de pédagogies dans l'enseignement supérieur : Innover : pourquoi et comment ?* Brest, France, juin 2015, p. 125-130. URL : https://hal.archives-ouvertes.fr/hal-01174273
- Marie-Pierre Adam, Matthieu Arzel, Antoine BEUGNARD, Jean-Philippe Coupez, Myriam Le Goff-Pronost, Michel Morvan, Pierre Tremenbert, Bruno Vinouze et Didier BAUX. "Boosting advanced skills in project management thanks to complex human and technical situations". In : *SEFI 2016 : European Society for Engineering Education annual conference*. Tampere, Finland, 2016, p. 1-11. URL : https://hal.archives-ouvertes.fr/hal-01494159, que j'ai présenté en conférence.
- Michel Morvan, Bruno Vinouze, Marie-Pierre Adam, Priscillia Creach, Matthieu Arzel, Didier BAUX, Antoine BEUGNARD, Jean-Philippe Coupez, Myriam Le Goff-Pronost et Camilla Kärnfelt. "How to apprehend leadership related skills in a project management experiment ?" In : *SEFI 2017 : 45th Conference on Education Excellence For*

Sustainable Development. Azores, Portugal, 2017, p. 536-543. URL : https://hal.archives-ouvertes.fr/hal-01661642

2.3 Activités de formation

En plus de mon rôle d'**adjoint à l'enseignement du chef du département Électronique**, j'interviens dans la formation sur un total de 205 h d'enseignements qui auraient dû être dispensés en 2019/20, sans la crise sanitaire. Il s'agit d'un volume acceptable, loin de celui tenu certaines années (près de 400 heures en 2010-11), et atteint après plusieurs recrutements au département Électronique qui ont comblé des départs successifs et ainsi allégé ma charge. Dans la pratique, sur 2019-20, le temps accordé à l'enseignement fut bien supérieur, en raison de l'enseignement à distance dans un contexte difficile de confinement généralisé. Il m'occupa quasiment à temps plein du 16 mars au 5 mai 2020 pour les enseignements de tronc commun d'Électronique dont je suis en outre co-responsable.

2.3.1 Un aperçu des formations IMT Atlantique

La formation Ingénieur de trois ans à IMT Atlantique se répartit entre la FISE, la FIP et la FIL (Formation d'Ingénieur par apprentissage en Ingénierie Logicielle, sur le campus de Nantes). La FISE, qui fournit le plus gros volume de diplômés, concerne les élèves admis à l'issue du concours commun Mines-Ponts et ceux admis sur titre après une formation universitaire, ainsi que les étudiants en séjour d'échange (sur les deuxième et troisième années). La FISE est organisée autour d'un parcours commun de formation et de Thématiques d'ApprofFondissement (TAF), structures de formations thématiques déroulées chacune sur presqu'une année universitaire. La première année de FISE est entièrement sous le sceau du parcours commun. Les deuxième et troisième années intègrent encore quelques enseignements de ce parcours et surtout les TAF.

La FIP propose une formation dédiée sur les deux premières années puis intègre les TAF dans son cursus.

Les Master of Science (MSc) ont aussi une partie qui leur est propre avant d'intégrer les TAF en dernière année. Il reste les Mastères spécialisés dans lesquels je n'interviens pas.

En parallèle du parcours commun de première année et des TAF, les élèves mènent des projets :

- projet de 1ère année "CoDEV" de conception, en équipe de $\sim$4 étudiants,
- projet de 2ème année "Commande entreprise" d'étude et/ou conception sur un sujet proposé par une entreprise partenaire, en équipe de $\sim$4-6 étudiants,

- projet de 3ème année “Ingénieur” (1 à 3 étudiants), plutôt orientés sur des thématiques R&D de la TAF.

La formation doctorale complémente ces parcours via 5 Ecoles Doctorales interrégionales. La formation continue existe aussi au sein d'IMT Atlantique. N'y contribuant pas, je ne la détaillerai pas ici.

2.3.2 Responsabilités d'unités d'enseignement

2.3.2.1 UE Électronique (60 h) de 1ère année

Cette UE apporte des connaissances et compétences fondamentales en Électronique analogique et numérique qui permettent de comprendre le fonctionnement et les limites des systèmes d'acquisition et de traitement de information, fondamentales à la culture d'un ingénieur généraliste. Cette UE met donc en avant les concepts de base, à savoir l'utilisation et la mise en commun de composants simples pour réaliser des fonctions plus complexes, et les contraintes physiques qui y sont rattachées.

A l'issue de l'UE, les élèves ingénieurs sont capables de,

- en électronique numérique,
  - effectuer des opérations logiques à partir d'opérateurs élémentaires par l'établissement de fonctions logiques simplifiées,
  - réaliser des opérateurs arithmétiques à partir d'opérateurs logiques,
  - différencier la logique combinatoire de la logique séquentielle,
  - connaître l'usage des bascules synchrones sur front et leur montage en registres,
  - modéliser un système séquentiel à base de registres et logique combinatoire,
  - connaître les règles de conception séquentielle CMOS,
  - connaître les limites et avantages d'une technologie CMOS d'intégration microélectronique,
  - analyser le fonctionnement d'une machine à états finis,
- en électronique analogique de
  - connaître les différents régimes de fonctionnement d'un transistor MOSFET,
  - analyser un circuit amplificateur à transistor MOSFET,
  - analyser un circuit à transistor MOSFET en régime de commutation (bloqué/passant),
  - connaître et reconnaître les différents régimes de fonctionnement d'un amplificateur opérationnel,
  - analyser un circuit à amplificateur opérationnel en régime linéaire,
  - analyser un circuit à amplificateur opérationnel en régime non linéaire,
  - déterminer les gains et impédances d'entrée et de sortie d'un quadripôle.

Ces différents résulats d'apprentissages visés contribuent aux compétences suivantes, formalisées au sein d'IMT Atlantique :

- ANALYSER, DIMENSIONNER ET METTRE EN ŒUVRE LES MONTAGES FONDAMENTAUX DE L'ÉLECTRONIQUE ANALOGIQUE ET LES CHAÎNES D'INSTRUMENTATION (compétence spécifique d'un domaine d'ingénierie),
- EXPLIQUER LES DÉPENDANCES ET LES LIENS ENTRE MATÉRIEL ET LOGICIEL (compétence spécifique d'un domaine d'ingénierie),
- CONCEVOIR, MODÉLISER ET SIMULER (compétence spécifique transverse).

2.3.2.2 UE Conception haut-niveau de circuits (80 h) de 2ème/3ème année

Cette UE est fait partie des enseignements d'approfondissement en Systèmes Embarqués et Hétérogènes. Elle propose aux étudiants une approche exploitant les outils et méthodes les plus récents de la conception de circuits électroniques, basée sur des représentations de haut-niveau et sur une conception conjointe matérielle/logicielle (codesign). Ces compétences permettent d'atteindre un niveau très avancé de compréhension et de mise en pratique dans la conception en électronique numérique. Elles permettent également de renforcer le lien entre logiciel et matériel, en travaillant sur les deux domaines simultanément. Finalement, elle amplifie les capacités d'intégration des étudiants et leur garantit une formation à l'état de l'art en conception de systèmes.

A l'issue de l'UE, les élèves ingénieurs sont capables de

- modéliser, simuler et concevoir des systèmes numériques complexes avec le langage SystemC,
- concevoir un circuit numérique à l'aide d'un outil de synthèse d'architecture ,
- exploiter un flot de conception système dédié aux SoCs ,
- concevoir un circuit numérique avec méthode et fiabilité,
- utiliser les principes architecturaux utilisés dans les microprocesseurs et processeurs de traitement de signal pour concevoir des circuits performants,
- utiliser les flots de conception de processeurs à jeu d'instructions dédié à l'application de type ASIP,
- justifier un choix de circuit en fonction de l'application,
- concevoir un système sur puce à l'aide d'un flot de conception dédié,
- appliquer des méthodes permettant d'approcher rapidement et sûrement le résultat espéré.

Ces différents résulats d'apprentissages visés contribuent aux compétences suivantes :

- INNOVER ET ENTREPRENDRE DANS UN CONTEXTE COMPLEXE ET INCERTAIN

(compétence générique),

- RÉSOUDRE UN PROBLÈME COMPLEXE EN ALLIANT THÉORIE ET PRATIQUE (compétence générique),
- COMPRENDRE ET ANALYSER, SYNTHÉTISER UN PROBLÈME ET/OU UNE SITUATION COMPLEXES (compétence générique).

2.3.2.3 UE Intégration électronique – de l'algorithme au prototype (80 h) de 2ème/3ème année

L'actuelle révolution industrielle s'appuie sur les systèmes électroniques qui ont développé nos moyens de production, de communication, et de traitement de l'information. Un ingénieur doit donc connaître les atouts et limites de l'intégration électronique afin d'en faire un usage pertinent dans son domaine applicatif. Dans le cadre de la TAF suivie, l'électronique est le support du traitement du signal et/ou de l'information et fait bien souvent appel à des compétences avancées. Néanmoins, une connaissance des techniques fondamentales de conversion du signal d'analogique vers numérique puis du traitement en électronique numérique permet à un ingénieur de développer des modèles réalistes et/ou des prototypes et ainsi d'expérimenter des innovations.

A l'issue de l'UE, les élèves ingénieurs sont capables de

- justifier les performances d'une implantation matérielle (fréquence maximale de fonctionnement, adaptation entre un algorithme et une architecture),
- définir une solution d'intégration matérielle adaptée à un cahier des charges,
- définir une architecture de circuit en cohérence avec les contraintes de l'application (précision, débit, latence, consommation d'énergie, consommation de ressources, matérielles)
- concevoir une architecture de circuit numérique,
- concevoir un processeur générique ou dédié,
- décrire un circuit en langage VHDL,
- prototyper un circuit électronique sur FPGA,
- intégrer une chaîne d'acquisition et de traitement du signal.

Ces différents résulats d'apprentissages visés contribuent aux compétences suivantes :

- COMPRENDRE ET ANALYSER, SYNTHÉTISER UN PROBLÈME ET/OU UNE SITUATION COMPLEXES (compétence générique),
- RÉSOUDRE UN PROBLÈME COMPLEXE EN ALLIANT THÉORIE ET PRATIQUE (compétence générique),
- CONCEVOIR ET RÉALISER DES SYSTÈMES ET DES ORGANISATIONS (compé-

tence générique).

2.3.3 Interventions

2.3.3.1 Cours, travaux dirigés et micro-projets

J'interviens principalement dans le parcours commun et la TAF Systèmes Embarqués Hétérogènes (SEH), avec des enseignements supplémentaires dans les TAF Oberservation et Perception de l'Environnement (OPE), Ingénierie des Systèmes de Communication (ISC) et Health. Le cœur de ces interventions est **l'enseignement des bases de l'électronique numérique jusqu'aux techniques avancées en conception de circuits numériques (techniques de parallélisme, adéquation algorithme-architecture, synthèse de haut niveau, langage SystemC), en passant par l'enseignement du langage VHDL et des architectures classiques de processeurs numériques, dédiés et généralistes**. Plutôt que des travaux pratiques classiques, je favorise des **micro-projets** qui me permettent d'avoir un fil rouge pour le reste de mes enseignements et d'apporter aux élèves, en plus de l'acquisition des compétences par la pratique, le **plaisir de réaliser et concrétiser des "objets" complexes**.

Ainsi, les micro-projets que je propose permettent tant de réaliser des jeux rétro sur circuit FPGA que de réaliser des chaînes de traitement du son ou de signaux d'électrocardiogramme, directement perceptibles par les élèves et le plus souvent sources de **motivation** intense. Cette appropriation du savoir par les élèves me tient à cœur. En outre, par ces nombreux micro-projets, je suis en mesure d'**évaluer les élèves par compétences**, profitant de situations complexes, encourageant la réflexion et l'application adéquate des connaissances et méthodes acquises dans les enseignements magistraux. Néanmoins, le suivi et l'accompagnement des micro-projets requièrent une forte disponibilité de ma part : j'organise les enseignements de manière à ce que les étudiants bénéficient de créneaux supplémentaires pour accéder aux salles de projet, et je leur fournis alors un soutien. Ceci amplifie grandement ma charge horaire mais nous permet d'atteindre **un niveau de compétences en accord avec nos exigences**.

2.3.3.2 Encadrement de projets sur les différentes formations d'IMT Atlantique.

En plus des micro-projets au sein des TAF, j'encadre des projets qui couvrent les 3 années de formation, tant FISE que FIP. En général, j'encadre **1 projet CoDEV (première année), entre 1 et 2 projets Commande Entreprise (2ème année) et entre 1 et 2 projets Ingénieur (3ème année) par an**. Les projets, au delà de l'aspect purement pédagogique, permettent de mieux connaître les étudiants et d'**identifier les élèves ayant un potentiel élevé** pour

intégrer en stage les entreprises et laboratoires les plus exigeants (et donc offrant des stages de grande richesse) et/ou poursuivre en doctorat. Bien souvent, c'est ainsi que je prépare mon recrutement de doctorants. Les projets permettent aussi de **détecter les élèves en difficulté** et de leur apporter un soutien, au-delà des enseignements. Ainsi, j'ai aidé des élèves à organiser et orienter leur études, à effectuer leurs recherches de stage, la rédaction de CV et leurs candidatures à des bourses.

J'ai par ailleurs été membre du comité de pilotage "Projet complexe" de troisième année de FIP de 2014 à 2019, avant qu'il ne disparaisse, la troisième année de FIP intégrant ensuite les TAF, ne permettant plus toute l'organisation et les cours associés de gestion de projet. Ce projet était particulièrement riche pour les étudiants et les enseignants. Nous divisions la promotion FIP en deux équipes conséquentes (plus d'une dizaine d'élèves par équipe) qui devaient s'organiser comme au sein d'une entreprise, avec un chef de projet, des responsables de communication externe/interne et des responsables de lots : il leur fallait gérer la **complexité** de l'organisation d'une équipe nombreuse. Ils devaient répondre à un appel d'offre pour un client, qui pouvait être la direction de l'École mais aussi des extérieurs, avec toute l'exigence que l'on attend de **professionnels**. Sur la base de leurs réponses et après négociations avec le client, chaque équipe réalisait une étude. Le sujet était systématiquement choisi en dehors du champ de compétences techniques des étudiants. En plus d'être une difficulté, un élément de complexité, cela permettait de focaliser l'attention des étudiants sur leur **organisation** plus que sur les solutions techniques pour traiter efficacement le problème. Le rôle du comité de pilotage était lui aussi complexe : en plus de proposer le sujet, nous suivions les élèves par des **points méthodologiques réguliers** avec un échantillon variable de chaque équipe. Nous pointions leurs forces et faiblesses, prodiguions des conseils et aiguillonnions quand nécessaire. Nous mettions en évidence la nécessité d'outils et méthodes pour s'organiser, l'importance du leadership et de la synergie d'équipe. Pour augmenter la complexité et évaluer l'adaptabilité et la robustesse de leur organisation, nous introduisions des aléas, comme des changements de dates dans les rendus, des modifications de cahier des charges par le client et des transferts d'individus entre équipes.

Tous ces aspects de la complexité de la profession d'ingénieur sont généralement occultés aux élèves, même en alternance. Nous avons en effet pris conscience par ce projet à quel point les maîtres d'apprentissage protègent leurs apprentis et, ce faisant, les privent d'un vécu utile à l'ingénieur. Exposer les élèves à la complexité, aux risques, à leur gestion, avant le début de leur carrière nous a semblé extrêmement bénéfique. Nous avons été surpris par l'implication si totale de nombreux élèves qui ont ainsi largement **anticipé les points durs du métier d'ingénieur**.

De cette **richesse pédagogique**, nous avons publié plusieurs aspects en conférences de

pédagogie dans l'enseignement supérieur. Je suis extrêmement reconnaissant aux membres de ce comité de pilotage pour cette collaboration et la joie d'enseigner différemment que cela m'a procurée.

Je regrette par ailleurs que ce type de projet n'existe plus dans cette forme. Certes ce projet complexe mobilise beaucoup de ressources d'encadrement pour assurer un passage à l'échelle de toute la formation ingénieur, mais il offre une telle richesse pédagogique qu'il doit **inspirer** les actuels projets en maturation au sein de la toute nouvelle formation.

2.3.3.3 Tutorat de stages de fin d'études et de stages de césure (25 étudiants de 2008 à 2020), pour beaucoup en coopération avec les partenaires de recherche.

Plaçant les élèves bien souvent en stage chez mes partenaires de recherche, je m'investis dans leur suivi et m'assure de la bonne **adéquation entre les compétences des stagiaires, leur souhaits de formation, leurs aspirations professionnelles et ce qu'offrent et recherchent les entreprises ou laboratoires d'accueil**. Parmi les destinations récurrentes, l'actuelle entreprise WideNorth est un partenaire privilégié, qui accueille chaque année plusieurs de nos élèves tant en stage court, de fin d'études et/ou long (une année de césure). Spécialisée dans les communications satellitaires et située en périphérie d'Oslo, dynamique, employant d'anciens élèves, elle attire de nombreux étudiants. Par ce biais, j'attire et sélectionne par ailleurs mes futurs doctorants.

2.3.3.4 Tuteur de 13 élèves en Formation d'Ingénieur en Partenariat

Par la FIP, je contribue aux bonnes relations avec nos entreprises partenaires et au suivi de nos élèves à travers le tutorat. Ainsi, j'ai encadré 13 de nos élèves depuis 2008 :

Ludovic Boué / SIGMA Informatique (2008-2011), Qinlin Zha / Orange Labs (2009-2012), Quentin Feraboli / Orange Labs(2011-2014), Mohammed Al Sadan / DCI NAVFCO (2012-16), Mohammed Al Sheddi / DCI NAVFCO (2012-16), Salman Al Osaimi / DCI NAVFCO (2013-16), Bassam Al Yahyan / DCI NAVFCO (2013-16), Hamoud Al Sharif / DCI NAVFCO (2014-19), Swilem Al Swilemy / DCI NAVFCO (2014-19), Faisal Al Qahtani / DCI NAVFCO (2015-19), Abdulrahman Al Abdullatif / DCI NAVFCO (2015-19), Stéphane Hervé / Naval Group (2017-2020), Alexis Aragnouet / Dassault Systems (2020-2023).

2.3.3.5 Co-coordinateur des contrats de formation pour la Marine Royale Saoudienne avec DCI/NAVFCO.

Finalement, je contribue aux contrats de formation pour la Marine Royale Saoudienne avec DCI/NAVFCO depuis 2012. Je l'avais fait par ailleurs en tant que "précepteur" lors de mon doctorat sur le premier contrat. J'encadre ces étudiants (élèves officiers) et organise leur soutien académique, facilitant le dialogue entre les différentes parties prenantes, dont les entreprises d'accueil en stage. Cela demande de part et d'autre une ouverture à d'autres cultures et contribue au **rayonnement de notre savoir-faire tant en termes de formation que d'industrie**.

3

Recherche

3.1 Thématiques de recherche

Mon activité de recherche est menée au sein du département Électronique d'IMT Atlantique et de l'actuel pôle Communications, Architectures et Circuits (CACS) du laboratoire CNRS des Sciences et Techniques de l'Information, Communication et Connaissance (Lab-STICC), UMR 6285, qui positionne son expertise "des capteurs à la connaissance". Ce pôle est reconnu pour son activité sur un large spectre, de la théorie de l'information à la conception de circuits et systèmes numériques. Au sein de ce pôle, je suis membre de l'équipe Interaction Algorithme-Silicium (IAS). Le laboratoire se réorganisant, il est probable que j'intègre le futur pôle Traitement et Transmission de l'Information, algorIthme et Intégration (T2I3) et la future équipe Algorithm/Architecture Interactions (2AI). L'adéquation algorithme-architecture est au cœur de mes thématiques de recherche que sont le traitement haut débit pour les systèmes de communications numériques, l'analyse de trafic réseau à haut débit et flexible et enfin les systèmes autonomes « intelligents » à faible consommation et/ou haute-capacité de traitement.

3.1.1 Architectures à haut débit pour les systèmes de communication numérique

Les systèmes de communication numériques font l'objet de mes centres d'intérêts depuis mes premières années de chercheur débutant en tant que stagiaire au sein du laboratoire du Professeur Joachim Hagenauer à la *Technische Universität München* (TUM) jusqu'à mes tout derniers travaux en tant qu'enseignant-chercheur à IMT Atlantique, en passant par la case ingénieur de recherche à TurboConcept. J'ai investigué majoritairement les systèmes de réception pour les domaines des communications mobiles, satellitaires et optiques où les débits visés exigent des solutions matérielles hautement parallèles. J'y ai développé mes compétences en adéquation algorithme-architecture et ai ainsi acquis une expertise plus large aussi bien en conception de circuits analogiques, numériques et mixtes ainsi qu'en communications numériques. Cette thématique de recherche est le fruit de nombreuses collaborations académiques et industrielles, locales, nationales et internationales. Les principaux résulats obtenus sont 3 brevets dont un en cours de dépôt sur des techniques de correction d'erreur, la démonstration du décodage stochastique des turbocodes convolutifs (en échec dans les autres laboratoires avant les travaux que j'ai menés avec C. Jégo, Q.T. Dong et trois stagiaires), la réalisation de 2 puces mixtes de décodage validées par le test dans ce domaine. Voici la liste de ces activités avec les thèses encadrées, les projets collaboratifs et contrats associés.

- **Traitement par circuits analogiques de l'information numérique**
  - 2 thèses (la mienne et celle de Daniel GOMEZ TORO, 2011-14).
  - Participation à la conception de 2 puces, validées par tests.
  - 1 contrat industriel d'expertise pour l'ESA (2013), 1 contrat Orange Labs (2007-10).
  - 2 brevets.

- **Traitement stochastique de codes convolutifs, turbo-codes, codes Cortex et Reed-Solomon**
  - 1 thèse (Quang Trung DONG, 2008-11) et 3 stages de master recherche encadrés.
  - Collaboration avec J.C. Carlach, Orange Labs.
  - Collaboration avec Warren Gross, McGill University, Canada (stages et articles).

- **Traitement numérique pour les communications mobiles, satellitaires et optiques, au-delà du Gbit/s**
  - 2 thèses CIFRE Orange Labs (Jean DION, 2010-13 et Tuan Anh TRUONG, 2011-14).
  - 1 thèse (Ali HAROUN, 2010-14) conjointe avec Christophe Jégo, IMS, Bordeaux.
  - 1 thèse CIFRE SAFRAN (Aomar BOURENANE, 2019-).
  - Projets collaboratifs ANR AFANA (2007-10), ANR UDEC (2008-11) et FUI 100GFlex

(2010-13).
- Contrats[1] industriels de conseil pour STM Norway puis Widenorth (Norvège, 2008-09 et 2014-15).
- Contrat[2] ARTES Advanced Technnology de l'Agence Spatiale Européenne (ESA) "Advanced Modem Prototype for Interactive Satellite Terminals", en consortium avec STM Norway, Alcatel Alenia Space España, DLR, TurboConcept, VeriSat(2007-10).
- Contrat[2] ARTES Advanced Technnology ESA "User Terminal Wideband Modem for Very High Throughput Satellites", en consortium avec WideNorth, University of Luxembourg, NTNU et Space Norway (2018-20).
- Contrat[1] de réalisation d'un démonstrateur FPGA de décodeur LDPC quasi-cyclique convolutif pour liaisons optiques à 100Gbit/s pour Huawei Paris Research Center (2018-19).
- Contrat[2] ARTES Advanced Technnology ESA "Wideband RF over IP demonstrator" en consortium avec WideNorth et Simula UiB (Norvège, 2020-22).
- 1 brevet en cours de dépôt.

3.1.2 Architectures à haut débit et flexibles pour l'analyse de trafic réseau

Contribuant à la problématique du haut débit pour la couche physique des réseaux, j'ai étendu mon activité aux couches supérieures. En effet, profitant des progrès technologiques de la couche physique, les débits ont explosé alors que la tâche des opérateurs pour gérer les données s'est complexifiée. En effet, ils doivent assurer d'une part la sécurité de leur réseau et d'autre part la qualité de service (*Quality of Service*, QoS) pour des applications exigeantes comme la vidéo à la demande. La sécurité des réseaux est par ailleurs critique tant d'un point de vue économique que politique. Ainsi, pour offrir des solutions robustes face aux attaques, les opérateurs disposent de solutions logicelles qu'ils maîtrisent bien mais qui peinent à suivre la montée en débit et la diversification des attaques, notamment avec la croissance fulgurante de l'Internet des Objets, *Internet of Things* (IoT). En outre, la QoS est la clé du succès commercial et ne doit pas être garantie par des investissements exagérés dans un réseau physique surdimensionné. Elle doit être garantie par un usage "intelligent" du réseau qui s'adapte à la nature de son trafic, en étant capable de l'analyser le plus finement et le plus rapidement possible, permettant l'usage optimal des ressources à tout instant. J'ai proposé des solutions à base de circuits reconfigurables FPGA pour ces différents problèmes

1. Contrat dont je fus responsable scientifique pour mon établissement et principal investigateur.
2. Contrat dont je fus responsable scientifique pour mon établissement.

et j'ai défendu une approche de collaboration efficace logiciel-matériel exploitant de manière originale les différentes capacités de reconfiguration des FPGA. Les résultats marquants sont la proposition d'implantation sur FPGA de l'algorithme *Support-Vector Machine* (SVM) qui dépassa l'état de l'art et diffusa jusqu'à la communauté de l'apprentissage automatique, puis les travaux d'intrication logiciel-matériel pour les sondes réseau qui furent récompensés par le prix du meilleur poster à la conférence IEEE CNS 2017 à Las Vegas.

Voici la liste de ces activités avec les thèses encadrées, les projets collaboratifs et contrats associés.

- **Analyse et classification de trafic sur FPGA**
  1 thèse (Tristan GROLÉAT, 2011-14)
- **Architectures hybrides matériel-logiciel**
  - 2 thèses (André LALEVÉE, 2014-17 et Franck CORNEVAUX-JUIGNET, 2014-18).
  - 1 contrat[1] industriel avec OVH Brest (2017-18).
  - 1 plateforme de démonstration financée par Brest Métropole et le Conseil général du Finistère.

3.1.3 Architectures faible consommation et/ou haute-capacité de traitement pour les systèmes autonomes « intelligents »

Les concepts et un savoir-faire acquis dans le domaine des télécommunications peuvent être appliqués à des domaines connexes où les traitements sont aussi fortement contraints et où l'autonomie est fortement recherchée. L'autonomie est à considérer tant en termes d'usage d'une source limitée d'énergie que l'indépendance vis à vis d'autres ressources à disposition dans les réseaux, comme des ressources de calcul ou des bases de données volumineuses. L'ingénierie biomédicale fut le premier domaine d'application des compétences acquises en traitement microélectronique analogique et mixte, notament au travers des travaux d'intégration de traitements au sein de prothèses, en collaboration avec le LaTIM, INSERM UMR 1101. Par la suite, dans des systèmes complexes, comme celui d'un réseau de capteurs corporels, j'ai considéré des traitements avancés comme la classification des signaux captés. Nous avons donc cherché à offrir un compromis d'efficacité de traitement et d'intégration pour minimiser la consommation d'énergie des capteurs, notamment par l'implantation en circuit de réseaux de neurones artificiels à cliques, inventés au département. Enfin, nous avons investigué le traitement au plus proche des capteurs pour de l'électroencéphalographie, afin d'en développer l'usage. Dans la continuité des réseaux de neurones artificiels intégrés avec le moins de ressources possibles, j'ai co-encadré une thèse sur l'accélération matérielle de l'apprentissage profond. Les résultats les plus marquants sont, outre

les différentes démonstrations de traitements de très faible consommation énergétique, la réalisation de 2 puces de réseaux de neurones artificiels ultra-basse consommation, dont une flexible, toutes deux validées par test et des solutions originales pour l'accélération de l'apprentissage profond qui ont valu à Ghouthi Boukli Hacène, doctorant que j'ai co-encadré, le prix de la meilleure thèse du programme Futur et Fuptures 2020 de la Fondation Mines-Télécom et premier prix de la thèse 2020 de l'Agence Francaise pour l'Intelligence Artificielle.

Voici la liste de ces activités avec les thèses encadrées, les projets collaboratifs et contrats associés.

- **Traitement biomédical enfoui**
  - Collaboration avec le LaTIM, INSERM UMR 1101
  - Participation à un groupe inter-GDR SOC-SIP, ISIS et Stic et Santé sur le thème des dispositifs biomédicaux.
- **Accélération matérielle pour les interfaces Cerveau-Machine**
  - 1 thèse (Erwan LIBESSART, 2015-18), 1 post-doc (Fabio TONI BRAZ, 2015-16).
  - Conception d'un ASIC CMOS 65nm et d'un démonstrateur FPGA.
  - Projet LABEX Cominlabs SABRE, en collaboration avec Francesco ANDRIULLI (Politecnico di Torino) et Anatole LÉCUYER (INRIA Rennes).
- **Réseaux de neurones parcimonieux pour une intégration matérielle à faible coût**
  - 2 thèses (Benoît LARRAS, 2012-15 et Paul CHOLLET, 2014-17).
  - 2 puces mixtes en technologie CMOS 65nm, testées avec succès.
- **Accélération de l'apprentissage profond pour une Intelligence Artificielle accessible à tous**
  - 1 thèse (Ghouthi BOUKLI HACENE, 2016-19), puis CDD CNRS, prix de la meilleure thèse du programme Futur et Fuptures 2020 de la Fondation Mines-Télécom et premier prix de la thèse 2020 de l'Agence Francaise pour l'Intelligence Artificielle.
  - 1 thèse CIFRE PSA (Hugo TESSIER, 2019-)
  - 1 stage de Master Recherche (Hugo LE BLÉVEC, 2020)
  - Collaboration avec Yoshua BENGIO, laboratoire MILA (Montréal, Canada).
  - Collaboration de recherche avec Interface Concept au sein de Pracom.

3.2 Activité contractuelle

Depuis 2007, j'ai financé ma recherche en propre par des contrats d'accompagnement et des contrats bilatéraux dont j'ai été responsable à hauteur de 650 k€, dont 428 k€ depuis 2014, et des contrats collaboratifs dont j'ai assumé la responsabilité pour un montant de 505 660 euros entre 2011 et 2019. J'ai en outre été contributeur d'autres contrats que je n'inclus

pas dans ces sommes mais que je liste par la suite. Je ne décompte pas ici les contrats de formation avec DCI/Navfco, gérés par la direction de la formation de mon établissement et auxquels j'ai énergiquement contribué, ni les différentes thèses CIFRE qui faisaient partie d'accords plus larges au sein de l'École avec nos partenaires, comme via Pracom, ni les financements Pracom des thèses de Tristan GROLÉAT et d'André LALEVÉE.

100GFlex Réseaux optiques Flexibles à base de modulation OFDM Multi-bandes à 100Gbps

Type Collaboratif FUI

Dates 08/03/2010 - 31/05/2013

Implication Contributeur

Budget 390 036 € (établissement)

Partenaires MitsubishiElectricR&D Centre Europe, Ekinops, France Telecom/Orange Labs, Yenista Optics, Institut Télécom et l'Université de Rennes I

Description Le projet 100G-FLEX s'est intéressé au développement de systèmes de transmission optique multiplexés en longueur d'onde (WDM) ayant un débit de 100 Gbit/s par canal. Les applications visées pour ces systèmes sont les réseaux métropolitains pour lesquelles les enjeux en terme de déploiement de nouveaux systèmes semblaient cruciaux au vu de la capacité croissante des réseaux d'accès, alimenteurs des réseaux métropolitains, et les contraintes de coût, de consommation et de flexibilité sont très fortes. Face à ces exigences, 100G-FLEX a proposé un ensemble de techniques matérielles et algorithmiques permettant un multiplexage OFDM optique.

AFANA Application-Field-aware Adaptive Network on chip Architecture

Type ANR

Dates 01/01/2007-30/09/2010

Implication Contributeur

Budget 803 681 € (IMT-UBS)

Partenaires Université de Bretagne Sud, TurboConcept et Technicolor

Description Ce projet visait une nouvelle approche orientée application pour la conception de NoC (Network on Chip) en vue de solutionner le goulot d'étranglement des communications au sein des futurs SoC (System on Chip) pour les télécommunications et le codage vidéo.

AMICEM Accélération matérielle pour une interface cerveau-machine temps réel

Type ARED

Dates 01/10/2015-30/09/2018

Implication Co-responsable

Budget 45 000 €

Description Cette bourse a financé partiellement la thèse d'Erwan Libessart dans le cadre du projet Cominlabs SABRE, décrit plus loin.

AMPIST Advanced Modem Prototype for Interactive Satellite Terminals

Type Projet ESA ARTES 5

Dates 11/05/2007-03/01/2010

Implication Co-responsable

Budget 41 933 €

Partenaires STM Norway, Alcatel Alenia Space España, DLR, TurboConcept, VeriSat

Description Ce projet répondait à un appel à projet ESA ARTES 5. Notre proposition consistait en une phase d'étude et la mise en œuvre d'un démonstrateur pour un système DVB-RCS en acccord avec les dernières évolutions du standard. Plusieurs problèmes concernant la réalisation du démonstrateur ont été analysés au cours de la phase d'étude. Nous avons notamment été en charge de l'annulation d'interférence, due aux canaux adjacents, et avons proposé des techniques de faible complexité. Nous avons par ailleurs montré les enjeux de ces traitements préalables à la synchronisation.

BP-MIMO Belief-Propagation for low-complexity MIMO receivers

Type ARED

Dates 01/10/2010-30/09/2013

Implication Co-responsable

Budget 86 430 €

Description Il s'agit du financement de la thèse d'Ali HAROUN dont Christophe JÉGO était directeur et Charbel ABDEL-NOUR co-encadrant. Dans le cadre de cette thèse, nous avons investigué la conception conjointe de détecteurs MIMO et de décodeurs de canal en exploitant une représentation unifiée par graphe.

CIFRE TRUONG Contrat d'encadrement de thèse de Tuan Anh TRUONG

Type CIFRE

Dates 04/10/2011-05/10/2014

Implication Responsable

Budget 30 000 €

Partenaires Orange Labs

Description Dans cette thèse, nous avons cherché à optimiser l'efficacité spectrale des transmissions pour les réseaux d'accès optique passif (PON) au moyen d'algorithmes de traitement numérique du signal. Nous avons notamment montré tout l'intérêt des techniques OFDM associées à des adaptations propres aux ressources matérielles du PON.

CIFRE TESSIER Contrat d'encadrement de thèse d'Hugo TESSIER

Type CIFRE

Dates 06/01/2020-06/01/2023

Implication Co-responsable

Budget 75 000 €

Partenaires PSA

Description Cette thèse vise à permettre l'embarquement efficace de solutions d'apprentissage profond pour la conduite autonome. Nous abordons ainsi aussi bien l'optimisation de la structure des réseaux que celle des représentations numériques au sein de ces réseaux. Les solutions apportées devront être exploitables par les ressources matérielles intégrables au sein d'un véhicule.

CYBER-THD Plate-forme très haut débit pour l'accélération de l'analyse de trafic réseau

Type Subventions de Brest Métropole et Conseil Genéral 29

Dates 09/10/2015-31/12/2017

Implication Responsable

Budget 13 127 €

Description Ces subventions ont permis l'acquisition de serveurs et cartes FPGA à l´état de l'art pour monter une plate-forme d'expérimentation pour nos activités d'accélération de l'analyse de trafic réseau et contribuer ainsi au succès de 3 thèses.

DICTOD Contrat de formation à la conception de turbo décodeurs pour STM Norway

Type Contrat industriel

Dates 15/12/2008-14/01/2009

Implication Responsable

Budget 4 300 €

Description Nous avons formé les ingénieurs à l'état de l'art des architectures et des techniques de décodage des turbocodes convolutifs.

EMCNorway Contrat de formation/conseil pour EMC Norway

Type Contrat industriel

Dates 01/02/2014-31/01/2015

Implication Responsable

Budget 30 000 €

Description Nous avons encadré et formé aux codes fontaines les équipes d'EMC Norway pour intégration dans un de leur produits.

HIRP OPEN Contrat d'étude pour Huawei Paris Research Center

Type Contrat industriel

Dates 01/09/2018-31/08/2019

Implication Responsable

Budget 79 564 €

Description Dans ce contrat, nous avons proposé des solutions architecturales et des algoritmes permettant d'effectuer la correction d'erreur sur cible reconfigurable pour les systèmes optiques à plus de 100Gbps. Ces travaux ont été menés par 3 ingénieurs de recherche, dont un CDD, Benoît PORTEBŒUF, et un post-doc, André LALEVÉE.

OVHFPGA Contrat d'études pour OVH

Type Contrat industriel

Dates 07/03/2017-06/03/2018

Implication Responsable

Budget 60 000 €

Description Ce contrat a porté sur l'analyse des opportunités apportées par les solutions à base de FPGA pour les infrastuctures d'OVH, tant pour les aspects sécurité que performance accrue de traitements critiques. Pierre-Henri Horrein en a été le principal contributeur.

RENPRAC Contrat d'études sur le réseaux de neurones parcimonieux et échantillonnage compressé (CS) pour radio cognitive

Type Contrat industriel

Dates 01/11/2013-31/10/2014

Implication Contributeur

Budget 40 000 €

Description Ce contrat a porté sur les apports possibles des réseaux de neurones parcimonieux en combinaison avec de l'échantillonnage compressé (CS) pour la radio cognitive.

RLP-AD Contrat d'expertise pour SITAEL

Type Contrat industriel

Dates 03/01/2013-02/05/2013

Implication Contributeur

Budget 10 000 €

Description Sur recommendation de l'ESA, nous avons effectué un audit du circuit de décodage analogique proposé par l'entreprise italienne SITAEL dans le cadre d'un démonstrateur de récepteur basse consommation. Nous avons souligné les défauts et suggéré des améliorations à leur solution initiale.

SABRE Stratégies computationnelles pour des interfaces cerveaux-machine basées sur des formulations hybrides volume / surface

Type Projet LABEX Cominlabs

Dates 01/09/2014-30/04/2019

Implication Co-responsable

Budget 417 533 €

Partenaires Inria Rennes/IRISA

Description Ce projet a cherché à fournir des solutions d'imagerie EEG haute résolution pour perfectionner les interfaces cerveau-machine. D'une part, SABRE à étudié des méthodes innovantes de solutions EEG qui fonctionnent dans une complexité linéaire au lieu de cubique en ce qui concerne les degrés de liberté physiques. Cela a permis déjà de réaliser d'énormes économies en termes de temps de calcul et de complexité. D'autre part, ces méthodes d'EEG ont été accélérées par la mise en œuvre d'accélérateurs dédiés sur FPGA. Des solutions ASIC ont aussi été proposées pour permettre une efficacité de traitement sans égal.

TURBODECODER IP Contrat d'expertise pour STM Norway

Type Contrat industriel

Dates 01/08/2010-31/08/2011

Implication Responsable

Budget 18 905 €

Description Dans ce contrat, nous avons formé et conseillé des ingénieurs dans la conception de turbodécodeurs.

UDEC Universal channel DECoder

Type ANR

Dates 01/01/2008-30/09/2011

Implication Contributeur

Budget 1 506 006 € (budget intégral)

Partenaires CEA LETI, Thalès Communications

Description Ce projet visait à unifier l'approche orientée sur la flexibilité et celle orientée sur l'optimalité pour la conception d'un décodeur canal multi-standards.

UTWN User Terminal WideBand Modem for Very High Throughput Satellites

Type Contrat industriel ESA ARTES Advanced Technology

Dates 11/10/2018-10/10/2020

Implication Responsable pour mon établissement

Budget 75 000 €

Partenaires Université de Bretagne Sud, Widenorth, University of Luxembourg, NTNU, Space Norway

Description L'objectif de cette activité est de mettre en œuvre un modem DVB-S2X entièrement reconfigurable, basé sur la radio logicielle, pour des bandes passantes ultra larges allant jusqu'à 1,5 GHz, répondant à des besoins concrets comme le téléchargement de données à partir de capteurs LEO et de satellites d'imagerie utilisant la bande Ka étendue. Nous avons fourni conjointement avec l'UBS une solution de décodeur FEC DVB-S2X (LDPC+BCH) capable de traiter les symboles au rythme de 1.4 Gsps sur un SoC Xilinx du marché.

WRFoIP Wideband RF over IP

Type Contrat industriel ESA ARTES Advanced Technology

Dates 16/04/2020-15/11/2022

Implication Responsable pour mon établissement

Budget 78 999 €

Partenaires Widenorth, Simula UiB

Description Le but de ce projet est de concevoir et de mettre en œuvre un module bidirectionnel, à large bande, de transport sur réseaux de données de type IP de signaux RF numérisés (cloudRAN) avec pour objectif de prendre en charge des bandes passantes allant jusqu'à 5 GHz. Les scénarios d'utilisation visés sont la communication bidirectionnelle par VSAT et la liaison descendante depuis les satellites d'observation de la Terre. Ce projet finance Ali MOHYDEEN, post-doc sur 10 mois.

3.3 Diffusion et prix

- Membre des GDR-ISIS et SOC2.
- Évaluations pour les revues suivantes : IEEE Transactions on Signal Processing, EURASIP Journal on Advances in Signal Processing, IEEE Transactions on Circuits and Systems, IEEE Transactions on VLSI Systems, IET Electronics Letters, IEEE Transactions on Communications, IEEE Communications Letters.
- Évaluation pour les conférences suivantes : IEEE ISCAS, IEEE NEWCAS, IEEE ICECS.
- Membre du comité de programme technique du IEEE International Symposium on Turbo Codes & Iterative Information Processing '10, '16, '18
- Prix du meilleur poster à la conférence IEEE CNS 2017 à Las Vegas.
- Travaux de thèse de Ghouthi Boubli Hacène récompensés par le prix de la meilleure thèse du programme Futur et Fuptures 2020 de la Fondation Mines-Télécom et premier prix de la thèse 2020 de l'Agence Francaise pour l'Intelligence Artificielle

3.4 Participation à des comités de suivi individuel de thèse et jurys de recrutement

Je participe au Comité de Suivi Invidividuel de la thèse de Jean Bruant, sur le sujet *Abstraction du flot de développement FPGA pour l'intégrer dans un flot de développement logiciel moderne*, au sein d'une thèse CIFRE OVH, sous la co-direction d'Olivier Muller, maître de conférences à Grenoble INP, et Frédéric Pétrot, professeur à Grenoble INP, et l'encadrement de Pierre-Henri Horrein et Tristan Groléat au sein d'OVH.

En plus de jurys de recrutement en interne de mon établissement, j'ai participé au jury suivant :

2017, CentraleSupélec Recrutement d'un.e nseignant.e-chercheur.se en Communications numériques et électronique numérique, Campus de Rennes de CentraleSupélec, Laboratoire IETR (Institut d'Électronique et de Télécommunications de Rennes, UMR CNRS 6164), équipe SCEE. CDI de droit public niveau maître de conférences. Section CNU : 61/63. Candidate recrutée : Haïfa Fares.

3.5 Encadrements

3.5.1 Doctorants

Je liste ici l'ensemble des thèses que j'ai encadrées.

12 ont été soutenues pour une durée moyenne des thèses d'environ 1196 jours (soit 3 ans 3 mois et 9 jours).

2 thèses sont en cours pour des soutenances prévues en 2022 et 2023.

Les doctorants en contrat CIFRE sont identifiés par le symbole *.

1. **Quang Trung DONG**

   Titre : Le principe du calcul stochastique appliqué au décodage des turbocodes : conception, implémentation et prototypage sur circuit FPGA

   Dates : 01/10/2008-20/12/2011

   Encadrement : 50%, direction de Christophe Jégo

   Jury :

Jean-Philippe DIGUET,	Directeur de Recherche CNRS, Lab-STICC/UBS Lorient (Président)
Emmanuel CASSEAU,	Professeur des Universités, ENSSSAT Lannion (Rapporteur)
Dominique DALLET,	Professeur des Universités, IPB Bordeaux (Rapporteur)
Christophe JÉGO,	Professeur des Universités, IPB Bordeaux (Directeur de thèse)
Matthieu ARZEL,	Maître de conférences, Telecom Bretagne Brest (Encadrant)

   Situation actuelle : Manager R&D, VHT, Vietnam

2. **Jean DION***

Titre : Etude et implémentation d'une architecture de décodage générique et flexible pour codes correcteurs d'erreurs avancés

Dates : 04/10/2010-05/11/2013

Encadrement : 50%, direction de Michel Jézéquel

Jury :

Christophe JÉGO,	Professeur des Universités - IPB/ENSEIRB-MATMECA (Rapporteur)
Jean-Pierre CANCES,	Professeur des Universités, ENSIL (Rapporteur)
Maryline HÉLARD,	Professeur des Universités, INSA (Examinateur)
David GNAEDIG,	Ingénieur de recherche, TurboConcept (Examinateur)
Marie-Hélène HAMON,	Ingénieur de recherche, Orange Labs (Encadrante de thèse)
Pierre PÉNARD,	Ingénieur de recherche, Orange Labs (Encadrant de thèse)
Matthieu ARZEL,	Maître de conférences, Télécom Bretagne (Encadrant de thèse)
Michel JÉZÉQUEL,	Professeur Institut télécom, Télécom Bretagne (Directeur de thèse)

Situation actuelle : Ingénieur R&D, B-COM, France

3. **Ali HAROUN**

Titre : Récepteur itératif pour système à multi-antennes basé sur l'algorithme de propagation de croyance

Dates : 04/10/2010-21/11/2014

Encadrement : 33%, Charbel ABDEL-NOUR, 33%, direction de Christophe Jégo

Jury :

Emmanuel BOUTILLON,	Professeur des Universités, Université de Bretagne Sud (Président)
Charly POULLIAT,	Professeur des Universités, INP - ENSEEIHT (rapporteur)
Jean-Pierre CANCES,	Professeur des Universités, ENSIL (Rapporteur)
Matthieu ARZEL,	Maître de conférences, Télécom Bretagne (Encadrant de thèse)
Charbel ABDELNOUR,	Maître de conférences, Télécom Bretagne (Encadrant de thèse)
Christophe JÉGO,	Professeur des Universités, INP/ENSEIRB-MATMECA (Directeur de thèse)

Situation actuelle : Enseignant-chercheur à l'Université Internationale du Liban

4. **Tuan Anh TRUONG***

Titre : Digital signal processing for next-generation passive optical networks

Dates : 03/10/2011-28/11/2014

Encadrement : 50%, direction de Michel Jézéquel

Jury :

Jean-François HÉLARD,	Professeur, INSA Rennes (Président)
Michel JOINDOT,	Professeur, ENSSAT (Rapporteur)
Didier ÉRASME,	Professeur, Télécom ParisTech (Rapporteur)
Naveena GENAY,	Ingénieur R&D, Orange Labs Lannion (Examinatrice)
Michel JÉZÉQUEL,	Professeur, Télécom Bretagne (Directeur de thèse)
Matthieu ARZEL,	Maître de Conférences, Télécom Bretagne (Encadrant de thèse)
Hao LIN,	Ingénieur R&D, Orange Labs Rennes (Examinateur)
Bruno JAHAN,	Ingénieur R&D, Orange Labs Rennes (Examinateur)

Situation actuelle : Ingénieur R&D, SPHEREA, France

5. **Tristan GROLEAT**

Titre : High performance traffic monitoring for network security and management

Dates : 03/01/2011-18/03/2014

Encadrement : 50%, direction de Sandrine Vaton

Jury :

Guy GOGNIAT,	Professeur, Université de Bretagne Sud (Président)
Philippe OWEZARSKI,	Chargé de Recherches, LAAS/CNRS (Rapporteur)
Dario ROSSI,	Professeur, Télécom ParisTech (Rapporteur)
Sandrine VATON,	Professeure, Télécom Bretagne (Directrice de thèse)
Matthieu ARZEL,	Maître de Conférences, Télécom Bretagne (Encadrant)
Isabelle CHRISMENT,	Professeure, Télécom Nancy (Examinatrice)
Stefano GIORDANO,	Professeur, University of Pisa (Examinateur)
Ludovic NOIRIE,	Chercheur Senior, Alcatel Lucent (Invité)

Situation actuelle : Ingénieur R&D OVH, France

6. **Daniel GOMEZ TORO**

Titre : Temporal Filtering with Soft Error Detection and Correction Technique for Radiation Hardening Based on a C-element and BICS

Dates : 03/10/2011-12/12/2014

Encadrement : 33%, Fabrice Seguin, 33 %, direction de Michel Jézéquel

Jury :

Patrick LOUMEAU,	Professeur, Télécom ParisTech (Rapporteur)
Vahid MEGHDADI,	Professeur, Ecole Nationale Supérieure d'Ingénieurs de Limoges (Rapporteur)
Camille LEROUX,	Maître de conférences, Institut Polytechnique de Bordeaux (Examinateur)
Michel JÉZÉQUEL,	Professeur, Télécom Bretagne (Directeur de thèse)
Matthieu ARZEL,	Maître de conférences, Télécom Bretagne (Encadrant)
Fabrice SEGUIN,	Maître de conférences, Télécom Bretagne (Encadrant)

Situation actuelle : Ingénieur R&D Plexus, Allemagne

7. **Benoît LARRAS**

Titre : Intégration CMOS analogique de réseaux de neurones à cliques

Dates : 01/10/2012-03/12/2015

Encadrement : 33%, Fabrice Seguin, 33 %, direction de Cyril Lahuec

Jury :

Olivier ROMAIN,	Professeur des Universités, Université de Cergy-Pontoise (Rapporteur)
Sylvain SAÏGHI,	Maître de Conférences HDR, Université de Bordeaux (Rapporteur)
Patrick LOUMEAU,	Professeur, Télécom ParisTech (Examinateur)
Simon THORPE,	Directeur de recherche CNRS, Université de Toulouse III (Examinateur)
Claude BERROU,	Professeur, Télécom Bretagne (Examinateur)
Cyril LAHUEC,	Maître de Conférences HDR, Télécom Bretagne (Directeur de thèse)
Matthieu ARZEL,	Maître de Conférences, Télécom Bretagne (Encadrant)
Fabrice SEGUIN,	Maître de Conférences, Télécom Bretagne (Encadrant)

Situation actuelle : Enseignant-chercheur, ISEN Lille, France

8. **Paul CHOLLET**

Titre : Traitement parcimonieux de signaux biologiques

Dates : 01/10/2014-24/11/2017

Encadrement : 33%, Fabrice Seguin, 33 %, direction de Cyril Lahuec

Jury :

Mustapha NADI,	Professeur des Universités , Université de Lorraine - Institut Jean Lamour - Vandoeuvre-lès-Nancy (Président)
Noëlle LEWIS,	Professeur des Universités, IMS - Bordeaux (Rapporteur)
Hervé BARTHELEMY,	Professeur des Universités, Université Sud Toulon Var (Rapporteur)
Patricia DESGREYS,	Professeur, Télécom ParisTech (Examinatrice)
Fabrice SEGUIN,	Maître de conférences, IMT Atlantique (Encadrant)
Matthieu ARZEL,	Maître de conférences, IMT Atlantique (Encadrant)
Cyril LAHUEC,	Maître de conférences (HDR), IMT Atlantique (Directeur de thèse)

Situation actuelle : Enseignant-chercheur, Telecom ParisTech, France

9. **André LALEVÉE**

Titre : Towards highly flexible hardware architectures for high-speed data processing : a 100 Gbps network case study

Dates : 01/10/2014-28/11/2017

Encadrement : 33%, Pierre-Henri Horrein, 33 %, direction de Michel Jézéquel

Jury :

Christophe JÉGO,	Professeur, Bordeaux INP / ENSEIRB-MATMECA (Président)
Daniel CHILLET,	Professeur, Université de Rennes 1 / ENSSAT (Rapporteur)
Virginie FRESSE,	Maître de Conférences / HDR, Université de St-Etienne (Rapporteur)
Michel JÉZÉQUEL,	Professeur, IMT Atlantique (Directeur de thèse)
Matthieu ARZEL,	Maître de Conférences, IMT Atlantique (Encadrant)
Pierre-Henri HORREIN,	Maître de Conférences, IMT Atlantique (Encadrant)
Michael HÜBNER,	Prof. Dr.-Ing, Ruhr-Universität-Bochum (examinateur)
Olivier MULLER,	Maître de Conférences, Ensimag / Institut polytechnique de Grenoble (examinateur)

Situation actuelle : Enseignant-chercheur, ISEN Brest, France

10. **Franck CORNEVAUX-JUIGNET**

Titre : Hardware and software co-design toward flexible terabitsper second traffic processing

Dates : 01/10/2014-04/07/2018

Encadrement : 33%, Pierre-Henri Horrein, 33 %, direction de Christian Person

Jury :

Frédéric ROUSSEAU,	Professeur, Université Grenoble Alpes (Président)
Yvon SAVARIA,	Professeur, Polytechnique Montréal (Rapporteur)
Christophe JÉGO,	Professeur, Bordeaux INP (Rapporteur)
Christine HENNEBERT,	Docteure Ingénieure de recherche, CEA (Examinatrice)
Matthieu ARZEL,	Maître de Conférences, IMT Atlantique (Encadrant)
Pierre-Henri HORREIN,	Docteur Ingénieur de recherche, OVH (Encadrant)
Christian PERSON,	Professeur, IMT Atlantique (Directeur)
Tristan GROLÉAT,	Docteur Ingénieur de recherche, OVH (Invité)

Situation actuelle : Ingénieur R&D, Widenorth, Norvège

11. **Erwan LIBESSART**

Titre : Interface cerveau-machine : de nouvelles perspectives grâce à l'accélération matérielle

Dates : 05/10/2015-30/10/2018

Encadrement : 33% co-direction de Cyril Lahuec et Francesco Andriulli

Jury :

Fan YANG,	Professeur, Université de Bourgogne (Présidente)
Bertrand GRANADO,	Professeur, Sorbonne Université (Rapporteur)
Olivier ROMAIN,	Professeur, Université de Cergy Pontoise (Rapporteur)
Matthieu ARZEL,	Maître de Conférences, IMT Atlantique (Encadrant)
Cyril LAHUEC,	Maître de Conférences (HDR), IMT Atlantique (Encadrant)
Francesco ANDRIULLI,	Professeur, Politecnico di Torino (Directeur)

Situation actuelle : Enseignant-chercheur, CentraleSupélec, France

12. **Ghouthi BOUKLI HACENE**

Titre : Processing and learning deep neural networks on chip

Dates : 03/10/2016-03/10/2019

Encadrement : 25%, Nicolas Farrugia, 25%, Vincent Gripon, 25%, direction de Michel Jézéquel

Jury :

Julie GROLLIER,	Directrice de recherche, CNRS/Thales (Rapporteur)
Warren GROSS,	Professeur, McGill University (Rapporteur)
Hervé JEGOU,	Chercheur, Facebook AI Research (Examinateur)
Yoshua BENGIO,	Professeur, Université de Montréal (Examinateur)
Vincent GRIPON,	Chercheur permanent, IMT Atlantique (Encadrant)
Nicolas FARRUGIA,	Maître de conférence, IMT Atlantique (Encadrant)
Matthieu ARZEL,	Maître de conférence, IMT Atlantique (Encadrant)
Michel JÉZÉQUEL,	Professeur, IMT Atlantique (Directeur)

Situation actuelle : Chercheur CNRS, France

13. **Aomar BOURENANE***

Titre : Architectures de décodeurs de canal à très haut débit pour voie descendante satellitaire

Dates : 01/02/2019-

Encadrement : co-direction avec Frédéric Guilloud

Situation actuelle : Thèse CIFRE Safran en cours

14. **Hugo TESSIER***

Titre : Implémentation embarquée de réseaux de neurones pour véhicule autonome

Dates : 06/01/2020-

Encadrement : 25%, Mathieu Léonardon, 25%, Vincent Gripon, 25%, direction de Michel Jézéquel

Situation actuelle : Thèse CIFRE PSA en cours

3.5.2 Ingénieurs de recherche et post-doctorants

1. **Arun KUMAR**

Titre : Intégration et prototypage de la partie réception numérique d'une transmission optique OFDM multi-bandes

Dates : 01/09/2010-31/10/2011

Financement : FUI 100GFLEX

Encadrement : 50%, avec Gérald Le Mestre

2. **Fabio TONI BRAZ**

Titre : Intégration ASIC d'accélérateurs mixtes pour une inteface cerveau-machine à base d'EEG

Dates : 23/10/2014-30/10/2015

Financement : Projet Cominlabs SABRE

Encadrement : 100%

3. **Benoît PORTEBŒUF**

   Titre : Etude et conception de décodeurs de codes LDPC quasi-cycliques convolutifs pour les communications optiques 100G sur cible reconfigurable

   Dates : 01/10/2018-27/09/2019

   Financement : Contrat HIRP OPEN

   Encadrement : 100%

4. **André LALEVÉE**

   Titre : Automatisation de flot de conception multi-FPGA pour une conception de décodeurs flexibles

   Dates : 01/10/2018-31/08/2019

   Financement : Ecole

   Encadrement : 100%

3.5.3 Stagiaires

1. **Colas GÉRANTON**

   Titre : Étude du décodage stochastique appliqué aux codes et turbocodes convolutifs.

   Dates : Mars-juillet 2008

   Niveau : Master recherche Electronique à l'IUP de Lorient

   Encadrement : 100%

2. **Yvain BRUNED**

   Titre : Décodage stochastique de turbocodes convolutifs

   Dates : Mai-juillet 2010

   Niveau : Stage obligatoire 1A ENS CACHAN

   Encadrement : 100%

3. **Arun KUMAR**

Titre : FFT parallèle pipelinée pour l'OFDM en communications optiques 100G

Dates : Mai-Août 2010

Niveau : Master IMARS

Encadrement : 50%, avec Gérald Le Mestre

4. **Amine SEMMA**

   Titre : Etude, conception et intégration de codes fontaines et raptor

   Dates : Février 2014 - janvier 2015

   Niveau : Stage de césure en 2ème année à Télécom Bretagne sur un contrat de recherche

   Encadrement : 50%, avec Romain Héloir de STM Norway, 50%

5. **Paul CHOLLET**

   Titre : Reconnaissance de spectre pour la radio logicielle

   Dates : Mars-septembre 2014

   Niveau : Master recherche IMARS

   Encadrement : 50%, avec Cyril Lahuec

6. **Hugo LE BLÉVEC**

   Titre : Intelligence artificielle autonome : le défi de la performance à moindre coût matériel et énergétique

   Dates : Mai-octobre 2020

   Niveau : Master recherche IMARS

   Encadrement : 50%, avec Mathieu Léonardon

3.6 Bibliométrie (04/11/2020)

h-index 13, source Google Scholar (https://scholar.google.fr/citations?user=2YX46XEAAAA

Citations 684, source Google Scholar

Publications 18 articles en revues (majoritairement IEEE), 63 publications en conférences, 2 chapitres de livres.

Liste disponible https://cv.archives-ouvertes.fr/matthieu-arzel et en Section 10.1

Brevets 2 + 1 déposé en août 2020

3.7 Rayonnement international

Allemagne Michael HÜBNER, Ruhr-Universität Bochum, collaboration au cours de la thèse d'André Lalevée avec publication commune

Canada Warren GROSS, McGill University, partenaire de recherche (publications communes, stagiaires)
Yoshua BENGIO, MILA - Université de Montréal, partenaire de recherche (publications communes, stagiaires)
Vincent GAUDET, Waterloo University, accueil de stagiaires

Italie Francesco ANDRIULLI, Politecnico di Torino, partenaire de recherche (publications communes et direction de thèse)

Luxembourg Steven KISSELEFF, SNT Université du Luxembourg, collaboration sur un projet ESA avec publications communes

Norvège Helge FANEBUST, CEO Widenorth, partenaire de recherche (contrats bilatéraux et 3 contrats ESA en collaboration) et accueil de stagiaires, recrutement d'élèves
Eirik ROSNES, Simula UiB, partenaire de recherche sur projet ESA

Deuxième partie

Présentation des travaux de recherche

4

Introduction

Concevoir un système de traitement et de transmission de l'information exige le respect d'un cahier des charges complexe, mariant des contraintes en termes de précision de l'information, de débit de traitement et de transmission, de consommation énergétique, de flexibilité et de coût financier. Trop souvent ces contraintes paraissent antagonistes et trouver une solution qui les satisfasse toutes relève de la gageure. Prenons le cas d'usage quotidien d'un terminal de communication mobile de type smartphone. L'utilisateur souhaite une solution mobile, réactive, offrant de multiples services et jouissant d'une longue autonomie, le tout pour un coût qu'il juge raisonnable. Les solutions actuelles sont le fruit de plusieurs décennies de recherche intensive et d'une ingénierie efficace qui en limite le coût. Néanmoins, l'utilisateur lui-même prend vite conscience que certaines contraintes sont difficilement conciliables et requièrent des compromis. Si en plus du débit, l'utilisateur cherche à maximiser l'autonomie de son smartphone, il a généralement deux réflexes. Il réduit la luminosité et le temps d'usage de l'écran d'une part, et limite les temps de communication d'autre part. Il a donc l'intuition qu'il s'agit là de deux gouffres à énergie. Mais qu'en est-il plus précisément ? [CH10] a montré qu'un smartphone consommait typiquement entre 31 et 51% de son énergie dans le module GSM de communication radio, la gestion de l'écran étant le poste de consommation suivant, talonné de près par le processeur. La mobilité a donc un prix énergétique conséquent et impose une contrainte forte sur le débit accessible par l'utilisateur. Lorsqu'aucune

mobilité n'est requise, le triptyque du débit, de l'énergie et de la flexibilité reste un défi dans bien des domaines, avec un accent plus ou moins fort sur l'un de ces trois points. L'innovation attendue par les concepteurs et utilisateurs d'un système est celle qui permettra de pousser plus loin dans une direction sans que cela ne se fasse au détriment des deux autres. J'ai ainsi investigué dans deux domaines qui privilégient des directions différentes à l'heure actuelle : les systèmes en réseaux, mobiles ou non, et les systèmes embarqués autonomes, notamment pour la santé et plus généralement intégrant une "intelligence artificielle".

D'une part, le monde des systèmes en réseau explose et appelle des besoins en débit de transmission et traitement sans cesse croissants. Prenons encore l'exemple des réseaux mobiles. Le réseau GSM (*Global System for Mobile Communication*) en norme numérique 2G et 2.5G permettait, dès les années 1990, des débits au niveau du terminal de l'ordre de quelques kbit/s à quelques dizaines de kbit/s. Ensuite, l'UMTS (*Universal Mobile Telecommunications System*) et toutes ses évolutions, de la 3G initiale commercialisée en France en 2004 à la version *DC-HSPA+ (Dual-Carrier High Speed Packet Access +)*, ont offert des débits de quelques centaines de kbit/s à quelques dizaines de Mbit/s. La norme LTE (*Long-Term Evolution*) peut offrir en France depuis 2014 plusieurs dizaines de Mbit/s voire plusieurs centaines de Mbit/s en conditions optimales en version LTE-*Advanced*, dite 4G+. Évidemment, l'avenir des réseaux mobiles suit cette course puisque la 5G, au-delà d'ouvrir la porte à de nouveaux usages et une interconnection plus larges d'objets autres que des smartphones, promet des débits de plusieurs Gbit/s. Cette croissance des débits n'impacte pas que le terminal mais tout le réseau sous-jacent qui doit adapter son infrastructure sans cesse, les normes, contraintes et besoins évoluant en permanence. Nous avons donc considéré que la montée en débit devait se faire en garantissant une flexibilité croissante des solutions matérielles sous-jacentes. Sans cela, nos solutions ne pouvaient être applicables raisonnablement par nos partenaires industriels, qui recherchent bien souvent une infrastructure matérielle durable pour limiter les coûts à moyen et long termes. Au-delà de ces deux enjeux majeurs de débit de données et de flexibilité, les concepteurs doivent aussi s'inquiéter de la consommation, qui s'avère bien souvent un enjeu incontournable, même sans mobilité. Par exemple, un datacentre ne peut consommer sans fin, pour des raisons certes de coût de consommation énergétique mais aussi de dissipation de la chaleur produite par les systèmes électroniques. Un exploitant de datacentre ne peut qu'apprécier une alternative aux architectures classiques reposant sur des Central Processing Unit (CPU) multi-coeurs afin de traiter plus de données à plus faible consommation d'énergie, tout en restant flexible ! J'ai ainsi investigué la problématique de la montée en débit du réseau à coût raisonnable en m'attaquant aux cas mobiles et satellitaires sous l'angle des décodeurs correcteurs d'erreurs et à l'apport du traitement numérique pour les réseaux optiques, comme cela est détaillé au

Chapitre 5. Ensuite, selon la même logique de soutenir la montée en débit des réseaux, j'ai contribué à la conception de solutions permettant leur sécurité et leur usage optimisé via la surveillance de trafic à très haut débit, et tout de même flexible. Ceci est détaillé au Chapitre 6.

D'autre part, l'industrie bio-médicale est également en croissance forte, répondant aux besoins d'une médecine de pointe mais aussi d'un système médical qui cherche à toujours mieux veiller sur ses patients. Si l'on considère très probable la multiplication des capteurs corporels au sein d'un réseau corporel, Body Area Network (BAN), dans les prochaines décennies, on peut facilement imaginer des systèmes complexes aggrégeant de grandes quantités de données à traiter et donc faisant face à un problème de débit de données. Cependant, nous restons pour l'instant à une échelle bien plus modeste où le débit de données reste encore acceptable et bien en deçà de ce qu'un terminal tel qu'un smartphone accepte. L'enjeu pour l'instant est ailleurs. Au-delà des aspects éthiques, l'acceptabilité par le patient requiert une discrétion et une simplicité d'usage qui vont de paire avec une forte autonomie des implants ou des capteurs **et** un encombrement réduit : il s'agit d'un verrou technologique majeur. En effet, nous ne pouvons pas envisager des capteurs généralisés au grand public exigeant des batteries imposantes ou des recharges trop répétées. Ainsi, les systèmes sans-fil, consommant peu et tout de même riches d'informations font l'objet d'un intérêt croissant aussi bien pour le corps médical que les usagers immédiats, comme le montre l'engouement pour les capteurs à usage des sportifs. Les concepteurs de circuit doivent alors imaginer les moyens d'embarquer efficacement des traitements plus ou moins élaborés qui apportent une valeur ajoutée au capteur, communicant plus ou moins avec d'autres, mais surtout consommant très peu pour maximiser leur acceptabilité. Ainsi, j'ai collaboré à la conception de prothèses instrumentées du genou avec le LaTIM puis au sein du projet SABRE au développement d'interfaces cerveau-machine à base d'EEG haute résolution. Dans les deux cas, nous avons proposé des solutions matérielles qui permettent d'absorber la charge de calcul de traitements plus ou moins conséquents dans un faible volume et avec une consommation énergétique compatibles avec une intégration embarquée.

En outre, par mes travaux de recherche sur l'intégration de solutions de classification embarquées au plus proche du capteur, j'ai contribué à l'exploration de solutions pour ne transmettre des données que lorsque cela est utile. Le département Électronique d'IMT Atlantique bénéficiant d'une expertise dans le domaine des réseaux de neurones, nous avons notamment cherché à tirer profit d'une solution originale prometteuse car parcimonieuse et donc *a priori* favorable à une intégration embarquée, les réseaux de neurones à cliques. Deux thèses ont permis d'explorer ce nouveau champ de recherche et m'ont par la suite encouragé à poursuivre mes travaux autour de l'intégration embarquée de réseaux de neurones

profonds, et plus seulement pour la santé. Tous ces aspects concernant la conception de systèmes autonomes “intelligents” à faible consommation et/ou haute capacité de traitement sont abordés dans le Chapitre 7.

Chacun des Chapitres 5, 6 et 7 inclut les perspectives qui lui sont associées. Un dernier Chapitre 8 dresse le bilan de mon activité de recherche passée, présente et future.

5

Architectures à haut débit pour les systèmes de communication numérique

Sommaire de ce chapitre

5.1 Problématique investiguée

Les systèmes de communications numériques font l'objet de mes centres d'intérêts depuis mes premières années de chercheur débutant en tant que stagiaire au sein du Laboratoire du Professeur Joachim Hagenauer à la *Technische Universität München* (TUM) jusqu'à mes tout derniers travaux en tant qu'enseignant-chercheur à IMT Atlantique, en passant par la case ingénieur de recherche à Turbo Concept.

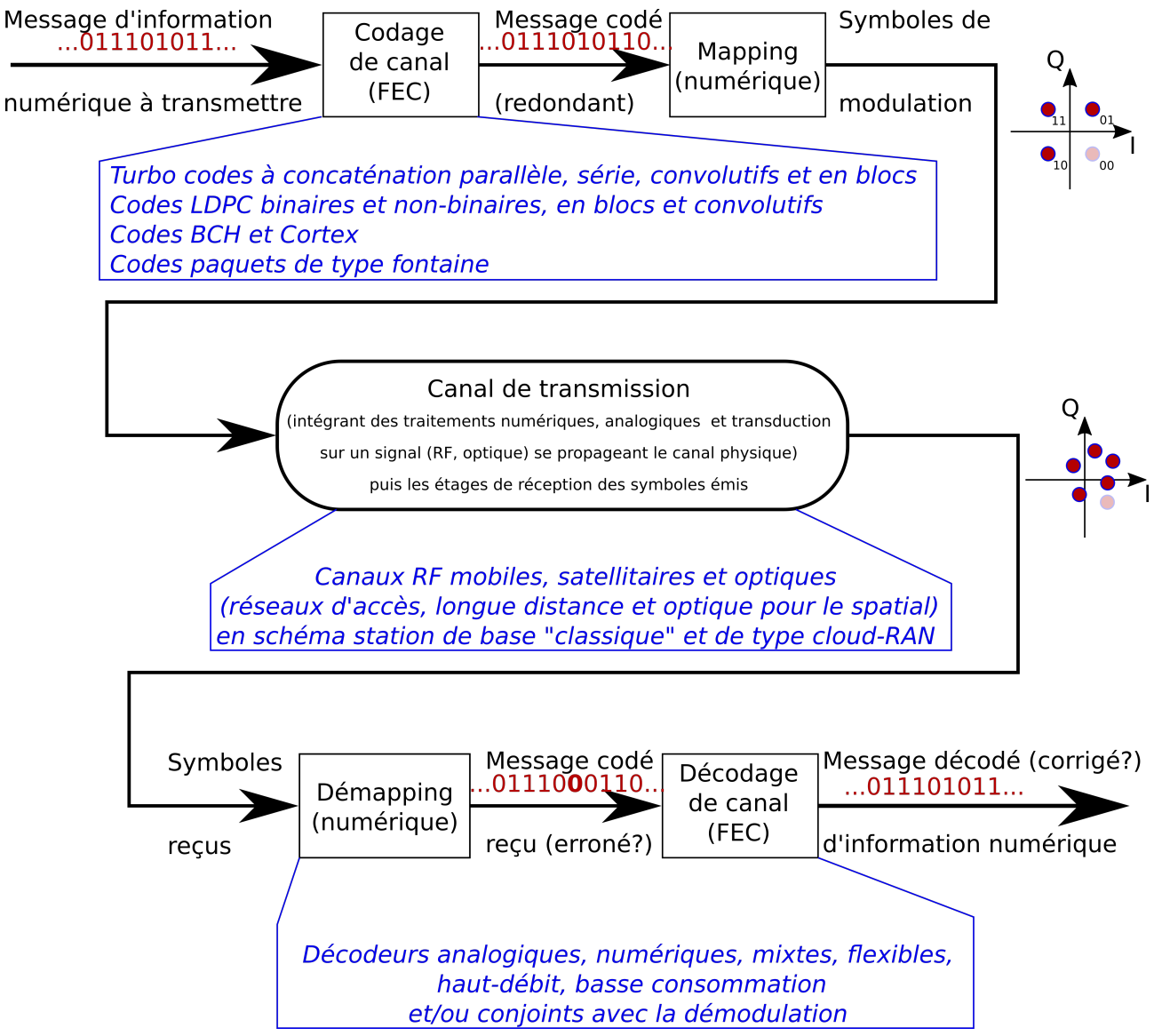


Figure 5.1.1 – Principe d'une chaîne de communication numérique (annotée des variations que j'ai étudiées).

Comme illustré en Fig.5.1.1, j'ai investigué majoritairement les systèmes de réception pour les domaines des communications mobiles, satellitaires et optiques (notamment le réseau d'accès et les futures communications optiques en espace libre) où le débit visé exige des solutions matérielles hautement parallèles. Pour cela, il faut extraire le parallélisme des algorithmes étudiés, au risque de dégrader la performance de la réception (donnée par une métrique comme le taux d'erreur en communications), et adopter une représentation de l'information adaptée. Nous allons voir que le choix de l'un comme de l'autre n'est pas trivial. Pour expliciter mes propos, je m'appuierai ici le plus souvent sur mes travaux sur les décodeurs (analogiques, mixtes et numériques) de canal, de type turbo, LDPC et Cortex, ainsi que sur les modules connexes en réception (synchronisation et démodulation).

Parmi les nombreux défis à relever pour un récepteur numérique, la synchronisation et le décodage correcteur d'erreur ont particulièrement attiré mon attention. Ce dernier a été mon objet premier et ensuite privilégié de recherche. Sur lui repose beaucoup d'enjeux : un codage correcteur d'erreur efficace fiabilise un lien de communication, ce qui permet pour un débit d'information donné de réduire la puissance d'émission, donc la consommation d'un émetteur. Mais le décodage correcteur d'erreur est aussi un goulot d'étranglement pour le débit, et potentiellement un gouffre énergétique pour le récepteur : il repose pour les codes les plus récents (turbo, LDPC, polaires et autres) sur un traitement itératif qui requiert du temps et des ressources, tant de calcul que de mémorisation. Le codage correcteur d'erreur au sein d'une chaîne requiert donc un compromis complexe entre puissance à l'émission et puissance de traitement pour corriger les erreurs en réception.

5.2 L'analogique au sein du récepteur numérique

5.2.1 Signal et information, analogique et numérique, quels mariages ?

Dans la chaîne modélisée en Fig.5.1.1, on considère un canal de transmission comme ceux utilisés pour les communications mobiles, satellitaires ou optiques, qui introduit un bruit globalement continu en temps et en valeur sur le signal transmis. Le signal reçu à l'entrée du décodeur est alors équivalent à celui émis mais altéré par une grandeur analogique. Par conséquent, la nature physique de l'entrée du décodeur de canal est intrinsèquement analogique. Comme proposé par H.-A. Loeliger et J. Hagenauer en 1998 [HW98 ; Loe+98], au lieu de numériser les signaux analogiques et de les traiter avec les circuits numériques, pourquoi ne pas traiter directement les signaux analogiques ? Est-il même possible de décoder des informations numériques à partir des signaux analogiques uniquement avec des circuits analogiques ? Cela présenterait-il un intérêt ?

J'ai tenté de répondre à ces questions lors de mon doctorat et de mes premières années en tant qu´enseignant-chercheur, en proposant des architectures mixtes de décodage, pour allier la simplicité et la basse consommation des nœuds de calcul analogique à base de Bipolar Junction Transistor (BJT) ou Metal Oxyde Semiconductor Field Effect Transistor (MOSFET) et la flexibilité du traitement et de la mémorisation numériques.

5.2.1.1 La base : des courants et probabilités aux opérateurs fondamentaux des algorithmes de passage de message

Les sorties d'un canal de propagation

Les algorithmes de propagation de croyances et de turbodécodage ne nécessitent que peu d´opérateurs : somme et produit sur des probabilités, normalisation et comparaison/sélection. Mais les probabilités utilisées pour alimenter l'algorithme doivent d'abord être extraites de la sortie du canal démodulé. Soit c^i un symbole binaire émis et $\widehat{c}^{\,i}$ la sortie de canal démodulée correspondante. Dans de nombreux modèles, le canal est considéré comme un canal à bruit blanc gaussien additif (Additive White Gaussian Noise (AWGN)) caractérisé par son écart-type σ. Supposons en outre que l'on utilise la modulation par déplacement de phase binaire (Binary Phase-Shift Keying (BPSK)), c'est-à-dire que chaque bit b est converti en une valeur antipodale b^a pour être ensuite modulé :

$$b \longrightarrow b^a \tag{5.1}$$

$$0 \longrightarrow -1 \tag{5.2}$$

$$1 \longrightarrow +1 \tag{5.3}$$

Ainsi, le log-rapport de vraisemblance de c^i conditionné à $\widehat{c}^{\,i}$ est

$$L(c^i|\widehat{c}^{\,i}) = \ln\left(\frac{Pr(c^i=1|\widehat{c}^{\,i})}{Pr(c^i=0|\widehat{c}^{\,i})}\right) \tag{5.4}$$

Si nous appliquons la loi de Bayes,

$$\begin{aligned} L(c^i|\widehat{c}^{\,i}) &= \ln\left(\frac{Pr(c^i=1,\widehat{c}^{\,i})}{Pr(c^i=0,\widehat{c}^{\,i})}\right) \\ &= \ln\left(\frac{Pr(\widehat{c}^{\,i}|c^i=1)\,Pr(c^i=1)}{Pr(\widehat{c}^{\,i}|c^i=0)\,Pr(c^i=0)}\right) \end{aligned} \tag{5.5}$$

Si l'on suppose une source binaire uniforme, $Pr(c^i = 0) = Pr(c^i = 1) = 1/2$. De plus, dans le cas d'une modulation BPSK sur un canal à bruit de type AWGN dont l'écart-type est σ, pour tout $b \in \{0, 1\}$ dont la valeur antipodale associée est b^a :

$$Pr(\widehat{c}^{\,i} | c^i = b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2} \|\widehat{c}^{\,i} - b^a\|^2\right)} \tag{5.6}$$

Combiner Eq. 5.5 et Eq. 5.6 permet de conclure que

$$L(c^i | \widehat{c}^{\,i}) \quad = \quad 4 \times \text{SNR} \times \widehat{c}^{\,i} \tag{5.7}$$

où Signal-to-Noise Ratio (SNR) est le rapport signal/bruit par bit égal ici à $\frac{1}{2\sigma^2}$.
En d'autres termes, une sortie de canal démodulée est proportionnelle au log-rapport de vraisemblance du symbole émis correspondant. Par conséquent, il existe deux solutions pour mettre en œuvre des algorithmes de propagation de probabilité. En effet, la sortie de canal démodulée peut être directement utilisée pour alimenter un décodeur fonctionnant uniquement dans le domaine logarithmique avec des opérateurs non linéaires [Hag97 ; HW98]. Autrement, l'exponentiation peut être implémentée pour convertir la sortie de canal démodulée en probabilités. Ensuite, l'algorithme est appliqué, ne nécessitant que des sommes, des produits et des normalisations sur ces probabilités [Loe+98].

Extraire de manière analogique des probabilités sur l'information transmise grâce à quelques transistors

Chaque fois que le décodage par propagation message probabiliste est effectué dans le do-

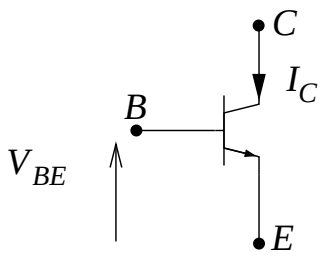


Figure 5.2.1 – Transistor bipolaire NPN.

maine logarithmique ou non, l'idée de base est d'utiliser la caractéristique exponentielle des transistors bipolaires dans la région active ou des transistors Metal Oxyde Semiconductor (MOS) en faible inversion. Par exemple, si un transistor bipolaire, BJT en anglais, illustré sur la figure 5.2.1, est dans la région active, son courant de collecteur I_C dépend de la tension

base-émetteur V_{BE} et est approximé par :

$$I_C \approx I_S \, e^{\frac{V_{BE}}{V_T}} \tag{5.8}$$

où $V_{BE} \gg V_T$. I_S est égal à

$$I_S = \beta I_{SE} \tag{5.9}$$

où I_{SE} est le courant de saturation de la jonction base-émetteur et β est le *gain en courant* du transistor. V_T est la *tension thermique*. Elle dépend de la constante de Boltzmann k, de la température T et de la charge de l'électron q :

$$V_T = \frac{kT}{q} \tag{5.10}$$

À 300K, V_T=26mV. Une paire de BJT en émetteur commun est schématisée en figure 5.2.2. Cette paire est polarisée avec un courant fixe I_{bias}, fourni par un autre BJT par exemple.

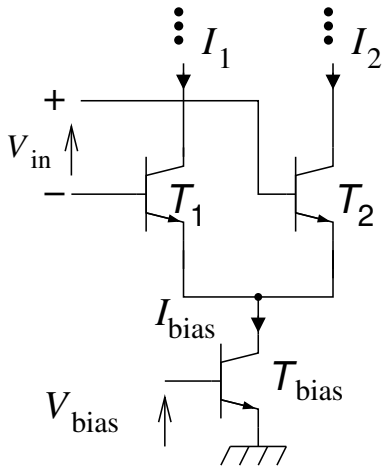


Figure 5.2.2 – Paire bipolaire à émetteur commun.

L'entrée est une tension différentielle V_{in}. On suppose que les deux transistors T_1 et T_2 sont polarisés dans la région active. Par conséquent, les courants de collecteur sont :

$$I_1 = I_S \, e^{\frac{V_{BE_1}}{V_T}} \tag{5.11}$$

$$I_2 = I_S \, e^{\frac{V_{BE_2}}{V_T}} \tag{5.12}$$

Le rapport des courants de sortie est alors :

$$\frac{I_2}{I_1} = e^{\frac{V_{BE_2} - V_{BE_1}}{V_T}} \tag{5.13}$$

Puisque l'émetteur est commun, $V_{E_1} = V_{E_2}$ et

$$\begin{aligned} V_{BE_2} - V_{BE_1} &= V_{B_2} - V_{B_1} \\ &= V_{in} \end{aligned} \tag{5.14}$$

Par conséquent, Eq. 5.14 et Eq. 5.13 impliquent :

$$\frac{I_2}{I_1} = e^{\frac{V_{in}}{V_T}} \tag{5.15}$$

Ainsi, **si V_{in} est associée à un log-rapport de vraisemblance d'un symbole binaire X tel que :**

$$V_{in} = V_T \ln\left(\frac{Pr(X=1)}{Pr(X=0)}\right) \tag{5.16}$$

alors

$$\frac{I_2}{I_1} = \frac{Pr(X=1)}{Pr(X=0)} \tag{5.17}$$

En outre, pour tout BJT en région active, $I_C = \beta I_B$ avec β communément dans un intervalle de 50 à 200, dépendant de la technologie employée, $I_{C_1} \approx I_{E_1}$, $I_{C_2} \approx I_{E_2}$ et

$$I_{\text{bias}} = I_1 + I_2 \tag{5.18}$$

Finalement, puisque $Pr(X=1) + Pr(X=0) = 1$:

$$I_1 = Pr(X=0)\ I_{\text{bias}} \tag{5.19}$$

$$I_2 = Pr(X=1)\ I_{\text{bias}} \tag{5.20}$$

Ceci montre que si la tension d'entrée d'une paire à émetteur commun est proportionnelle à $\ln\left(\frac{Pr(X=1)}{Pr(X=0)}\right)$, alors les deux courants de sortie sont proportionnels aux probabilités $Pr(X=1)$ et $Pr(X=0)$ par un facteur égal au courant de polarisation. Par conséquent, **si une tension est associée à un log-rapport de vraisemblance, Log-Likelihood Ratio (LLR), et un courant à une probabilité, une paire de transistors à émetteur commun agit comme un convertisseur LLR-à-probabilités**. L'exponentiation analogique n'est plus un problème. Ensuite, **la somme, le produit et la normalisation des courants sont des opérations de**

base pour les concepteurs analogiques. Selon le type des entrées et des sorties — courant ou tension, il est très simple de construire et assembler soit des modules de somme-produit comme proposés dans [Loe+98] soit des contreparties non linéaires comme définies dans [Hag97 ; Moe+00]. Tout cela est la base qui a permis de réaliser des décodeurs de type turbo [Ama+03 ; GG03 ; WGS04 ; Vog+05], LDPC [Hal+03 ; GC11a ; LCK13 ; Miy+13 ; ASC13 ; Bac+16 ; Zha+17] et autres familles plus ou moins exotiques comme les codes Cortex [Per+09a ; Per+09b].

Bien-sûr, ce modèle de circuit très simple ne considère pas les non-idéalités qui dégradent les performances de ces cellules. Mais par mes travaux, ceux de l'équipe IAS et ceux de la communauté, des solutions pour les contrer furent proposées avec succès comme nous le verrons par la suite.

5.2.1.2 Contexte de ma thèse de doctorat

Les premiers prototypes de décodeurs analogiques élémentaires [Lus+99 ; Moe+00] ont montré que les décodeurs analogiques surpassaient leurs homologues numériques en termes de taille, de consommation d'énergie et/ou de vitesse. Par exemple, la puce de Moerz et al [Moe+00] décodait 3,3 fois plus vite qu'un homologue numérique, consommait 8 fois moins d'énergie et était 5,2 fois plus petite. Néanmoins, le concept du décodage analogique passe-t-il à l'échelle des normes industrielles avec des codes complexes et des longueurs de trames de plusieurs centaines à milliers de bits ? Pour répondre à ces questions, j'ai participé à l'effort de recherche de la communauté à travers mon doctorat, puis la collaboration aux travaux du laboratoire après mon intégration en tant qu'enseignant-chercheur.

En tant que doctorant à partir de 2002, je suis parti des bases énoncées dans la section précédente et du constat suivant. Le turbo-décodage bénéficie de l'échange itératif de données probabilistes des messages entre deux décodeurs pour approcher la limite de Shannon. Chaque décodeur implémente un algorithme basé sur la propagation de probabilités dans un réseau de nœuds de calcul. Ces probabilités sont, en théorie, des quantités réelles. Elles sont donc mieux représentées par des signaux analogiques plutôt que par des signaux quantifiés qui les approchent. De plus, la propagation de ces signaux de type analogique n'est limitée que par la loi naturelle de cause à effet et les délais physiques de propagation. Par conséquent, l'algorithme de décodage n'est par définition ni discret en valeur ni discret en temps, c'est-à-dire non pas numérique, mais analogique. De plus, l'idée originale derrière l'invention de turbo-codes était l'utilisation d'une rétroaction appropriée, un concept purement analogique. Ainsi, si les données codées sont numériques et sont donc évidemment codées par des circuits numériques, les signaux reçus sont analogiques après avoir

été transmis sur des canaux réels et devraient être traités de la manière la plus appropriée par des algorithmes de type analogique. Par conséquent, le contrôle des erreurs est évidemment numérique à partir du point de vue de l'émetteur, mais le contrôle des erreurs est plus analogique que numérique du point de vue du récepteur. Ainsi, la mise en œuvre d'un turbo-décodeur analogique fut ma mission.

Ces deux premières puces ont été suivies par quelques autres puces développées par d'autres laboratoires universitaires [Win+04], implémentant des décodeurs itératifs traitant parfois plusieurs dizaines de bits. Par exemple, un turbo-décodeur pour la plus petite trame de la norme Universal Mobile Telecommunications System (UMTS) - taux 1/3, 40 bits d'information - a été implémenté et testé avec succès [Vog+05]. Un autre turbo-décodeur analogique a été développé pour la norme IEEE802.16a [WGS04]. Le turbo-code utilisé était un turbocode produit $(16,11)^2$. De nombreuses pistes ont ensuite été explorées pour interpeller l'industrie numérique. Ainsi, la consommation des circuits [Gau+04] et leur surface [MSN03] n'ont cessé de diminuer. Les décodeurs analogiques ont été rendus reconfigurables [GGG02]. Une méthodologie de conception et une conception automatique ont été proposées [Dai02]. Même une solution BIST (Built-In Self Test) [Yiu+05] et un codec LDPC simple [Hal+03] ont été présentés. Nous avons même étendu les concepts du décodage analogique aux étages supérieurs d'un récepteur en proposant un circuit analogique simple pour le *demapping* des modulations numériques telles que les modulations d'amplitude en quadrature (MAQ) et les modulations par déplacement de phase (MDP) [Seg+04]. De même, [Sol+06] proposa un circuit analogique permettant la détection MIMO en amont d'un décodeur de canal analogique. Cependant, alors que les applications industrielles traitent généralement des longueurs de trame allant jusqu'à quelques milliers de bits, la conception de décodeurs analogiques pour des trames aussi grandes est irréalisable avec l'architecture entièrement parallèle habituelle. Bien que les cellules analogiques de base soient beaucoup plus petites que leurs homologues numériques, la relation un à un entre le symbole à décoder et l'élément matériel de décodage requis par le traitement entièrement parallèle donne une surface sur puce prohibitive dans le cas de longueurs de trames de quelques milliers de bits. Moerz a proposé une solution innovante à ce problème : un décodeur à signaux mixtes [Moe04]. Il s'agit essentiellement d'un noyau de décodeur mixte réutilisant un morceau de réseau de décodage analogique connecté à des mémoires numériques par l'intermédiaire de convertisseurs analogique vers numérique et numérique vers analogique. Cette proposition a mis en évidence non seulement le fait que la réutilisation du matériel analogique est possible mais aussi le fait qu'elle ne peut être évitée pour une application industrielle requérent des trames de plusieurs centaines à milliers de bits.

5.2.2 Décodage mixte de turbocodes de type DVB-RCS

5.2.2.1 Décodage analogique de codes convolutifs

J'ai conçu dans le cadre de ma thèse un décodeur analogique en technologie Bipolar and Complementary MOS (BiCMOS) 0.25μm de code convolutif récursif systématique double binaire à 8 états tel que l'on rencontrait alors dans les standards en tant que code composant de turbocodes. Je l'ai dimensionné pour pour un taille de bloc de 24 symboles double-binaires et pour un taux de codage égal à 2/3. L'algorithme *A Posteriori* Probability (APP) présenté dans [AH98] a été appliqué et associé directement à une architecture parallèle.

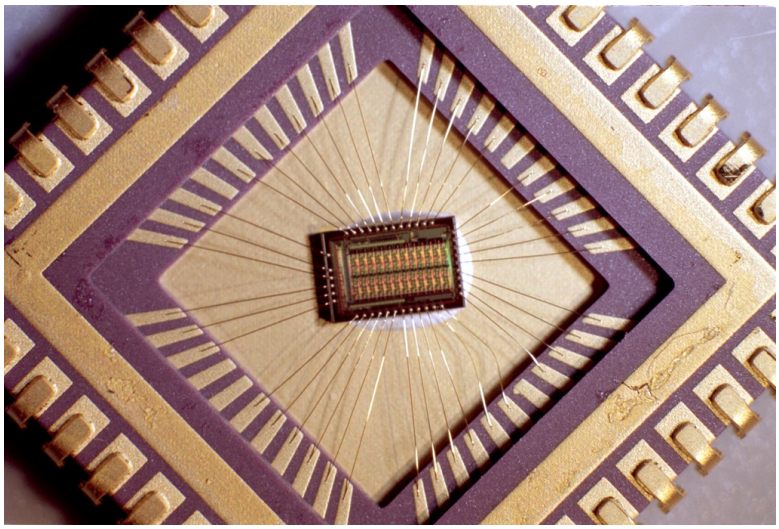


Figure 5.2.3 – Puce ANAMAP dont le boîtier a été ouvert : la structure répétitive du décodeur en treillis à 24 sections est reconnaissable.

La figure 5.2.3 montre la puce dans son boîtier. La surface complète du circuit, y compris les plots, est égale à 6.80mm². La surface du noyau, y compris l'anneau de décodage et les interfaces d'entrée/sortie, est égale à 4.31mm². Les blocs principaux du noyau peuvent être clairement identifiés dans la figure 5.2.4 qui est une microphotographie du circuit. Ces blocs sont :

- les mémoires en entrée,
- l'anneau de décodage BCJR, constitué de 24 sections identiques de décodage et occupant la majeure partie de la surface du circuit, soit 3.03mm²,
- l'interface de sortie — deux registres à décalage de 24 bits —.

Enfin, la consommation électrique du circuit a été mesurée pour une tension d'alimentation analogique AVDD=2,8V et une tension d'alimentation numérique VDD=1,5V. Elle est égale à

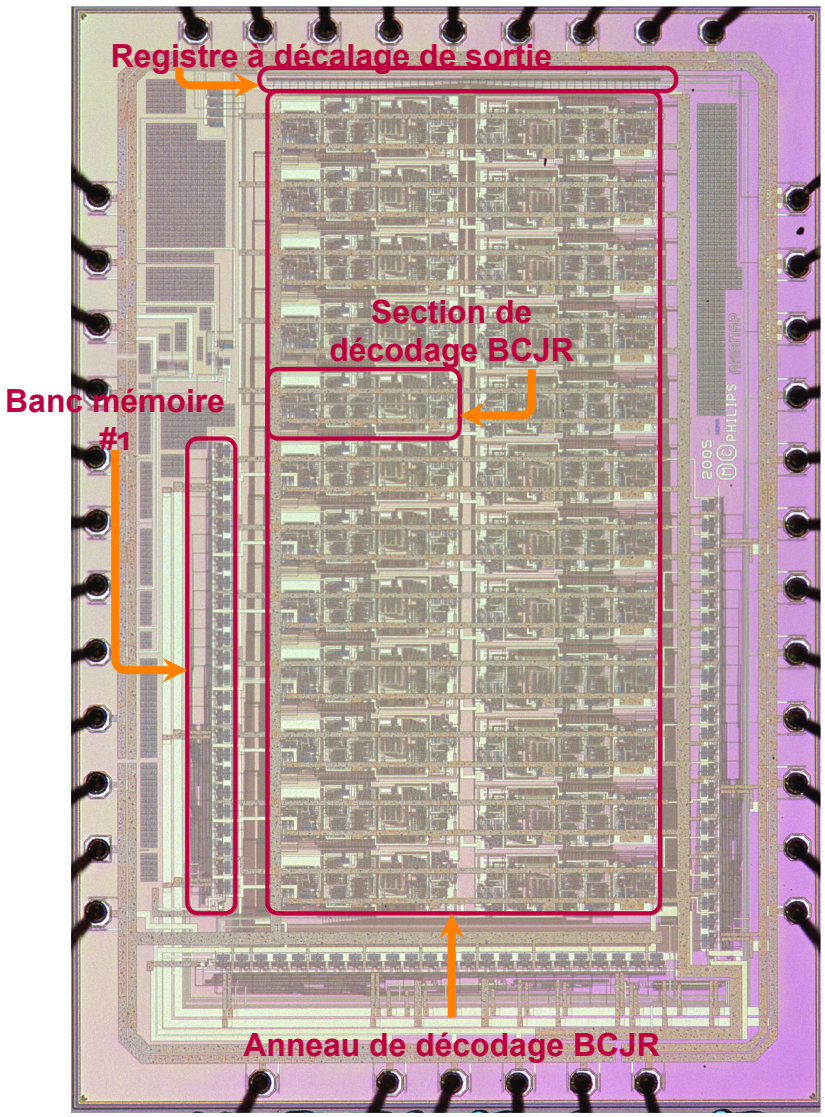


Figure 5.2.4 – Microphotographie de la puce ANAMAP.

414mW et est presque *indépendante du débit de données*, puisque la consommation électrique est principalement statique. Ceci est conforme à la consommation simulée de 440mW. Par conséquent, l'énergie mesurée par bit décodé est égale à 4nJ à un débit de données de 100Mbit/s.

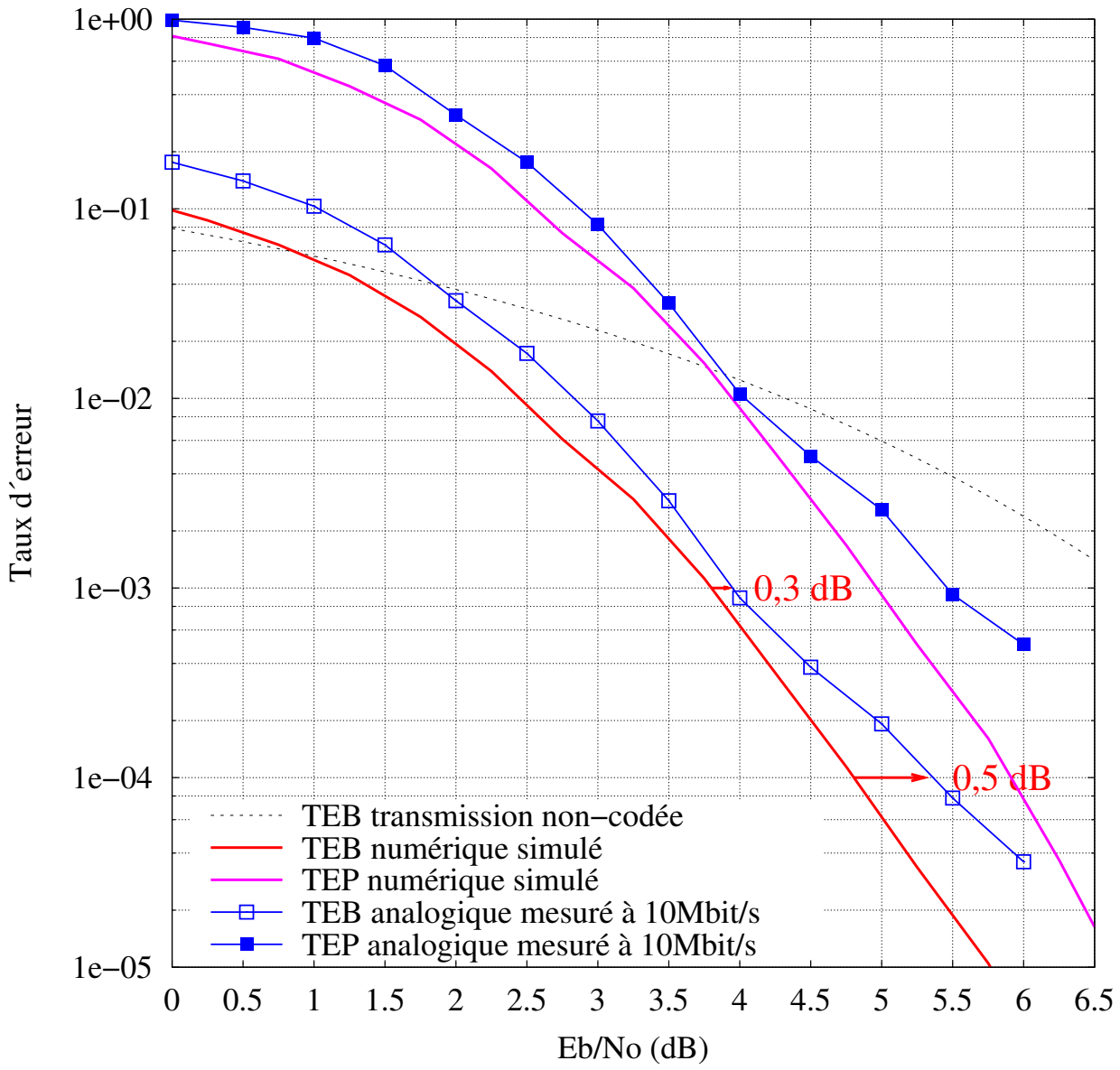


Figure 5.2.5 – Comparaison des TEB et TEP mesurés et simulés.

Les résultats des mesures sont présentés dans la figure 5.2.5. Ils sont comparés aux résultats de la simulation de haut niveau. Une perte de 0,3dB se produit pour un taux d'erreur binaire (TEB) de 10^{-3} et augmente ensuite à mesure que le taux d'erreur diminue — de plus de

0,5dB pour un TEB supérieur à 10^{-4}. Cette perte peut être en partie attribuée à un couplage de bruit déterministe entre l'horloge maître qui cadence les quelques circuits numériqes et les signaux analogiques, comme la tension d'alimentation analogique AVDD et l'entrée analogique VREF qui définit le niveau d'équiprobabilité pour les tensions d'entrée unipolaires correspondant aux log-rapports de vraisemblance. Lorsque le rapport signal/bruit appliqué à l'entrée du décodeur augmente, ce bruit supplémentaire reste constant et le rapport signal/bruit réel dans le circuit converge vers une limite. Ceci explique l'allure des courbes de la figure 5.2.5. Evidemment, pour un premier prototype, il était illusoire d'espérer anticiper et contrer toutes les imperfections possibles.

Ce décodeur analogique double binaire à 8 états peut être utilisé seul mais peut également être utilisé comme un décodeur de composants dans un schéma turbo, comme nous l'avons proposé par la suite.

5.2.2.2 Concept du décodage semi-itératif

Réaliser un décodeur analogique complètement parallèle traitant des trames de plusieurs milliers de bits comme dans les normes actuelles requiert des surfaces de silicium inacceptables comme montré par Moerz dans [Moe04] qui propose une architecture mixte. La limite à cette architecture est l'usage massif de convertisseurs numérique/analogique, qui cassent le modèle tout analogique des messages échangés. Ils corrompent l'idéal analogique initial. Nous avons tenté de le conserver en exploitant des mémoires analogiques et non numériques et de les associer à des turbocodes astucieusement construits de manière conjointe au décodeur analogique. Ce fut la deuxième contribution majeure de ma thèse de doctorat, associée à un brevet [ASL06] et qui fut résumée dans [Arz+07], qui est intégré en annexe de ce document en Chapitre 9 et section 9.1 Cet article présente un algorithme semi-itératif novateur et son architecture correspondante basée sur la réutilisation du matériel et l'utilisation de turbo-codes à "roulettes" (*slice*) [Gna+05]. L'algorithme proposé rompt la relation un à un entre la longueur du bloc de code et la taille du décodeur par rapport à la solution entièrement parallèle. Une seule puce de décodeur analogique capable de supporter des longueurs de trame jusqu'à quelques milliers de bits n'occupe ainsi que quelques dizaines de mm^2 . Cela est fait sans aucune perte de performance de correction, puisque l'architecture conserve une caractéristique essentielle du turbo-décodage analogique totalement parallèle : l'échange continu d'information extrinsèque. Il est utilisé à chaque étape du processus de décodage pour améliorer la vitesse et la qualité de la convergence. La réduction du parallélisme avec un décodeur semi-itératif permet de diviser par 10 la surface de silicium, et avec une perte de débit plus faible qu'avec un décodeur analogique qui aurait été totalement itératif. Ainsi,

un décodeur semi-itératif offre un compromis entre la réduction de la surface sur la puce et le débit de données. De plus, une architecture semi-itérative peut facilement être reconfigurée grâce aux propriétés de l'entrelaceur d'un turbocode à roulettes. J'ai montré qu'une puce de turbo-décodeur semi-itératif de 37 mm^2 en technologie BiCMOS 0.25μm peut traiter n'importe quelle longueur de mot de code allant de 40 à 2432 bits, c'est-à-dire les longueurs de trames de la norme DVB-RCS.

5.2.2.3 Evolution et perspectives pour le décodage analogique

Ce travail fournit des réponses à certaines des principales questions de conception d'un décodeur analogique signalées dans [Vog+05] concernant la réutilisation du matériel, la reconfigurabilité et les grandes tailles de bloc de code. Néanmoins, d'autres problèmes restaient à résoudre. Par exemple, de grands mots de code de quelques milliers de bits signifie que le temps entre le premier et le dernier symbole échantillonné devient grand. Cela entraînerait nécessairement une dégradation des valeurs de maintien et donc affecterait le décodage avec les capacités basiques utilisées dans les circuits d'échantillonnage proposés dans la thèse. Un travail sur les mémoires analogiques et les interfaces étaient donc une suite attendue. En outre, le manucrit de thèse a soulevé le problème de la sensibilité du décodeur proposé aux effets parasites dus à la technologie BiCMOS sous-jacente. L'usage d'une technologie CMOS n'aurait d'ailleurs pas évité cet écueil. Mes travaux de thèse furent donc repris par Nicolas Duchaux qui travailla à rendre robuste les décodeurs analogiques BiCMOS et intégra un turbo-décodeur proche de ce que je proposais. Nous publiâmes notamment quelques articles [DUC+10 ; Duc+09b ; Duc+09a ; Duc+08] montrant que l'on pouvait concevoir des turbo décodeurs analogiques fiables. Malheureusement, Nicolas Duchaux n'a pas conclu son doctorat. En parallèle, notre communauté du décodage analogique initialement centrée sur l'Allemagne, la Suisse, la France, le Canada et les Etats-Unis s'est étoffée et les tentatives se multiplièrent pour réaliser des décodeurs analogiques dignes d'une industrialisation. Ainsi, [HY10] montra que les décodeurs analogiques offrent des propriétés de convergence inattaignables par les circuits numériques en temps discret, expliquant les résultats de simulation et encourageant à poursuivre l'effort de recherche dans le domaine. Progressivement, la communauté a pris conscience que battre les circuits numériques sur le plan du débit était irréalisable, le coût du parallélisme étant trop important et l'efficacité des mémoires numériques étant insurpassable en analogique. Par contre, pour toute application requérant une faible consommation d'énergie, le décodage analogique reste un concurrent sérieux ! En effet, [GC11a] a proposé un circuit CMOS pour de telles applications en décodant à 100 pJ/bit un code LDPC (32,8). De même, [LCK13], toujours avec un décodeur

analogique LDPC (32,8), décrit un circuit capable d'atteindre 216Mb/s en consommant 4.98 mW soit 23pJ/bit grâce à un critère d'arrêt. De même, [Miy+13] a atteint une consommation de 10.4pJ/bit pour un décodeur LDPC (32, 8) mixte (mélangeant analogique et numérique). Ensuite, un décodeur LDPC (120, 75) plus puissant, fut testé [ASC13] avec un débit mesuré à 750Mbit/s et une consommation de 17pJ/bit. Montré comme étant plus performant que l'état de l'art numérique basse consommation en ASIC, les conclusions furent un peu abusives. En effet, cette consommation de 17pJ/bit ne prenait en compte que la consommation du cœur analogique et en rien celle des mémoires qui stockent les messages à décoder et surtout les bancs de convertisseurs parallèles numérique vers analogique, que l'on ne retrouve pas dans une intégration numérique en telle quantité.

Nous notons que les décodeurs implantés sur circuit ont atteint leur plus haute complexité avec un LDPC (480,240) [Zha+17] en 2017. Les auteurs conclurent alors que leur solution de décodage analogique n'était concurrentielle avec le numérique que dans le cas d'applications à budget d'énergie limité, débit "modéré" (quelques dizaines de Mbit/s) et des gains de codage relativement faibles (quelques dB par rapport à une transmission non codée). Même si cela ne semble pas réjouissant, c'est une conclusion intéressante à l'ère de l'IoT et des réseaux de capteurs. **Les capteurs répondent aux contraintes de (très, voire ultra) basse consommation d'énergie, sans exiger des débits importants ni des codes puissants. Il leur faut transmettre des données et contrôler avec le moins d'énergie possible. Et dans ce domaine, les décodeurs analogiques sont compétitifs !** Nous avons ainsi présenté nos travaux à différents entreprises (Orange Lab, Renesas), malheureusement sans jamais convaincre totalement. En effet, l'écueil majeur reste l'expertise requise en électronique analogique qui est de moins en moins répandue alors que les outils de conception numérique se démocratisent à un niveau de fiabilité et répétabilité imbattable.

Néanmoins, nous avons travaillé en tant que consultants de l'entreprise SITAEL dans le cadre d'un projet de l'European Space Agency (ESA) dont une partie des résultats furent rendus publiques [Bac+16]. Nous fûmes sollicités pour fiabiliser la conception et éviter des chausse-trapes propres à la conception d'un décodeur analogique. SITAEL parvint à réaliser et tester une puce intégrant la démodulation cohérente IF, la poursuite de porteuse, la récupération d'horloge, la conversion SP-L vers NRZ, la détection d'arrivée de trame de données et une unité de contrôle ainsi que les mémoires analogiques et un décodeur LDPC (128,64). L'application considérée est la télécommande d'engins spatiaux, donc pas du tout des communications à hauts débits mais plutôt du contrôle à basse consommation, moins d'un demi-watt pour toutes ces fonctions. Il aurait été objectivement intéressant de comparer cela à la consommation d'un équivalent numérique, mais cela reste un ordre de grandeur intéressant pour dimensionner des solutions implantées dans des réseaux de capteurs. Ces

travaux ont montré que le décodage analogique est une solution d'intérêt, maintenant à portée de toute entreprise disposant des compétences en analogique au vu de l'important travail de recherche mené par la communauté académique depuis 1998.

J'ai gardé de mes travaux en décodage analogique l'intérêt pour des traitements simples qui exploitent une représentation originale de l'information. C'est donc avec enthousiasme que je me suis impliqué dans la recherche autour du décodage stochastique qui allie le concept analogique de flux d'information convergeant rapidement par un échange quasi instantané entre des noeuds de calculs simples et la fiabilité du numérique.

Par ailleurs, ces travaux m'ont bien fait comprendre l'intérêt majeur de l'électronique analogique pour le traitement basse consommation et faible encombrement, notamment dans le cas d'applications profondément embarquées telles que nous le verrons par la suite.

5.3 Décodage stochastique de codes convolutifs, turbocodes, codes Cortex et Reed-Solomon

5.3.1 Concepts fondamentaux du décodage stochastique

En 2003, V. Gaudet, membre de la communauté du décodage analogique, passé en séjour de recherche au département Électronique, et A. Rapley, ont proposé une nouvelle approche [GR03] basée sur le calcul stochastique. Les principes du calcul stochastique ont été décrits dans les années 1960 par Gaines [Gai69] et Poppelbaum et al. [PAE67] comme une méthode permettant d'effectuer des opérations complexes avec peu de ressources matérielles. La principale caractéristique de cette méthode est que les probabilités sont converties en flux de bits stochastiques en utilisant des séquences de Bernoulli, dans lesquelles l'information est portée par les statistiques des flux de bits comme illustré en Fig. 5.3.1 où un nouveau bit est produit à chaque cycle d'horloge pour chaque flux, augmentant progressivement la précision de la conversion. Bien-sûr, par nature, une même probabilité peut être représen-

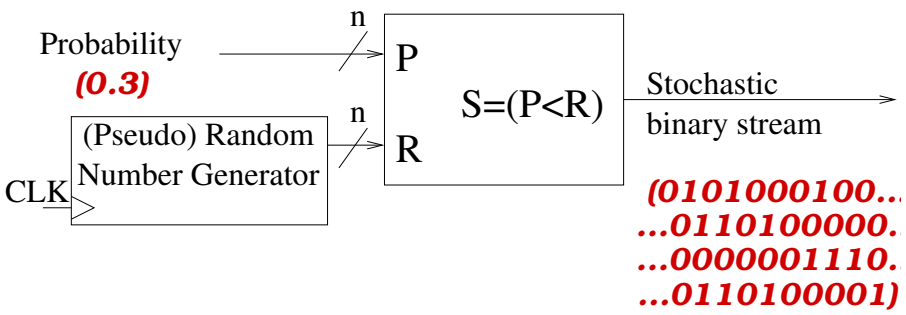


Figure 5.3.1 – Principe de la conversion de probabilité en flux binaire stochastique.

tée par plusieurs flux stochastique distincts. Ensuite, les opérations arithmétiques sur les probabilités peuvent être réalisées par l'intermédiaire de simples portes logiques, comme la multiplication portée par une porte ET (Fig. 5.3.2), la normalisation avec une bascule JK et l'addition normalisée avec des multiplexeurs.

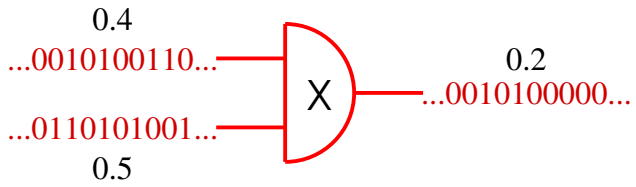


Figure 5.3.2 – Principe de la conversion de probabilité en flux binaire stochastique.

Le décodage stochastique a été appliqué initialement à certains codes de correction d'erreurs courts tels que le code de Hamming (7,4) [GR03] et un turbo-code en bloc (256,121) basé sur deux codes de Hamming (16,11) [Win+05]. La première implémentation d'un décodeur stochastique avec un code LDPC (16,8) a été décrite dans [GGM05]. Une approche améliorée de décodage stochastique a ensuite été proposée pour décoder des codes LDPC de taille plus importante [SGM06 ; TMG07]. Cette approche a également été étendue à des codes en blocs linéaires bien connus avec des matrices de contrôle de parité à haute densité, à savoir les codes Bose Ray-Chaudhuri Hocquenghem (BCH), les codes Reed Solomon et les codes produit [Sha+08]. Le potentiel du traitement stochastique en terme de faible complexité et de haut débit a été démontré par l'implantation sur circuit reconfigurable de type FPGA d'un décodeur LDPC (1056, 528) [SMG08] qui a atteint un débit de 1,66 Gb/s, clairement digne d'un exploit à l'état de l'art lors de sa publication.

5.3.2 Innovations apportées inspirées de mon expérience analogique

Dans ce contexte, j'ai proposé des implantations stochastiques de turbocodes à concaténation parallèle de codes convolutifs. Cela avait été tenté par d´autres laboratoires, notamment celui de W. Gross à l'Université McGill à Montréal, mais sans succès. Christophe Jégo, alors en séjour d'études à McGill avant de revenir à Télécom Bretagne, et moi-même espérions que notre expertise en décodage analogique serait un atout pour s'attaquer à ce défi du haut débit à faible complexité par le traitement stochastique, et ce d'autant plus que dépasser le Gbit/s était un enjeu majeur alors pour les turbo décodeurs.

Deux stagiaires en fin d'études, Colas Géranton en Master recherche à l'IUP Lorient en 2007 et Yvain Bruned de l'ENS Cachan en 2010 et un doctorant, Quand Trung Dong de 2008 à 2011 s'engagèrent avec moi et Christophe Jégo dans le projet de recherche de turbo décodage convolutif stochastique.

Parmi les problèmes rapidement soulevés par la communauté du décodage stochastique, l'activité de commutation aléatoire des flux fut le plus important. En effet, à haut rapport signal à bruit, les messages sont rapidement très fiables et convertis en probabilités quasiment constantes et égales à 0 ou 1, c'est à dire des flux stochastiques variant extrêmement peu, rendant alors le décodage stochastique très peu performant, voire inopérant. [SMG08] a introduit la mise à l'échelle en fonction du bruit, Noise Dependent Scaling (NDS), des symboles reçus afin de réduire la fiabilité et de faire en sorte que les flux changent régulièrement. Un autre problème était la corrélation et le verrouillage des flux, qui restaient bloqués à '0' ou '1'. Les super-nœuds [Win+05], les Edge Memory (EM) [SMG08] et les *(Majority-Based) Tracking Forecast Memories* (TFM) [Sha+10 ; Sha+10] ont permis de re-générer continuellement des flux stochastiques afin de rompre la corrélation.

Nous partîmes alors du constat qu'un turbo décodeur convolutif exploite un échange itératif de messages probabilistes entre les noeuds d'un réseau en treillis, ce qui est similaire à ce qui se faisait au sein des décodeurs LDPC, avec des additions, multiplications et normalisations de probabilités, avec quelques différences tout de même. En effet, le décodage selon l'algorithme Bahl Cocke Jelinek Raviv (BCJR) requiert de nombreuses sommes de probabilités et plusieurs boucles d'échanges, en anneau fermé aussi bien pour les informations extrinsèques du procédé turbo que pour les transferts de métriques aller et retour dans le cas de codes circulaires. Ces derniers posaient des problèmes d'initialisations entre mots de codes et constituaient des processus convergeant vers des états où l'information était sans cesse atténuée par les nombreuses additions normalisées. Ces phénomènes existent aussi en décodage analogique et je sus les détecter puis les contrer.

De l'addition à la multiplication

L'addition stochastique fut vite mise en avant comme un enjeu pour le décodage BCJR des turbocodes. En effet, sa précision est requise pour une correction d'erreur efficace mais elle impacte grandement l'efficacité d'un turbo décodeur stochastique. Précisément, le problème est le suivant. Pour additionner n probabilités, il faut multiplexer n flux de manière aléatoire. Chaque flux ne contribue alors que $1/n^{\text{ième}}$ du temps. Dans un décodage en treillis à plusieurs états où l'on somme plusieurs dizaines de métriques, le temps requis pour une précison acceptable permettant la correction d'erreurs devient rapidement incompatible avec un débit plusieurs Gbit/s. Nous avons proposé d'appliquer au turbo décodage une solution proposée par Janer et al. [Jan+96] qui convertit l'addition en multiplication dans le domaine exponentiel (Fig. 5.3.3) , tout comme nous le faisions en jonglant entre courants et tensions via des transistors en électronique analogique pour traiter les probabilités au sein d'un décodeur analogique. Néanmoins, l'application de cette seule technique n'est pas suffisante pour atteindre le Gbit/s à un Bit Error Rate (BER) acceptable et nous avons proposé des techniques complémentaires.

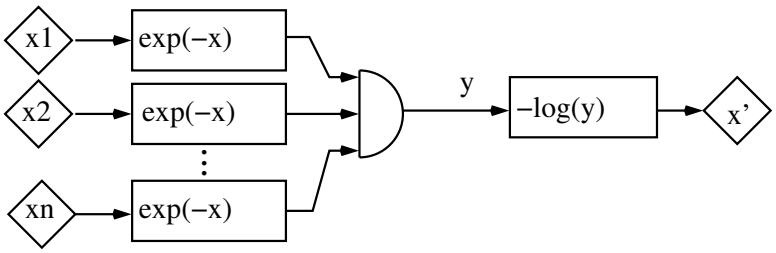


Figure 5.3.3 – Addition stochastique par transformation exponentielle.

Décodage à flux multiples

Une architecture basée sur EM est illustrée dans la Fig.5.3.4. Chaque probabilité P_i est portée par un flux stochastique s_i et est traitée avec les autres probabilités par une unité logique – A, B ou C – ou une EM pour éviter le verrouillage.

Cette EM capte un bit régénérateur d'une réserve (*pool*) lorsque la corrélation se produit. Mais, il serait également possible de prendre un bit régénératif d'un autre flux stochastique indépendant représentant la même probabilité. Pour réduire la corrélation entre les flux concurrents, il faut alors en avoir plus de deux. De plus, le bit régénératif doit être choisi au hasard parmi eux. Contrairement aux TFM et aux EM, une valeur de bit courante serait utilisée au lieu d'une valeur mémorisée. Comme il n'y aurait pratiquement pas d'effet mémoire, les nœuds de variables cachées ne pourraient pas rester systématiquement bloqués

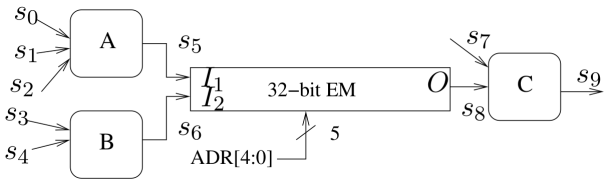


Figure 5.3.4 – Architecture basée sur une *Edge Memory* où chaque probabiblité est représentée par un unique flux stochastique.

à l'équiprobabilité.

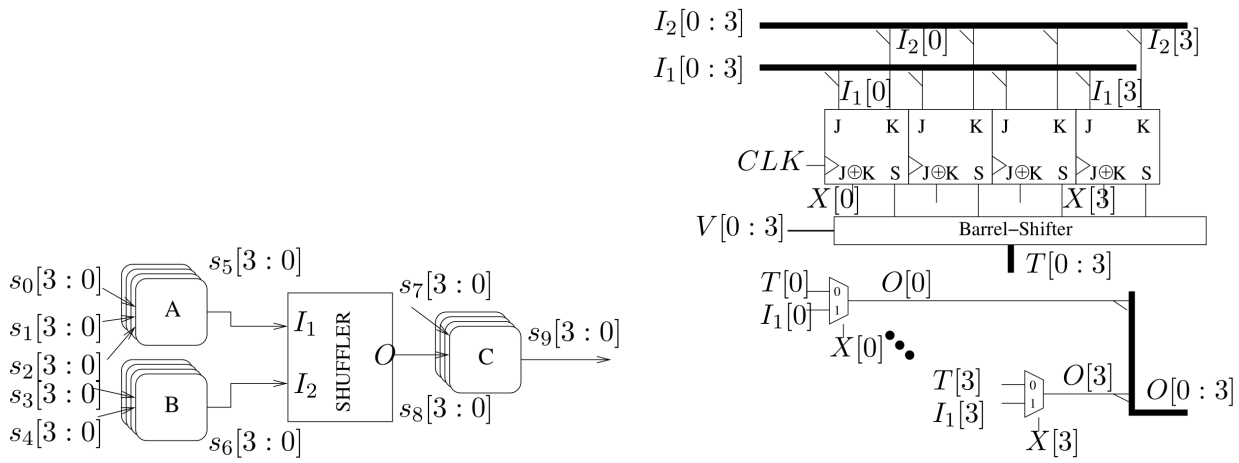
J'ai ainsi proposé de remplacer l'architecture conventionnelle par l'architecture à flux multiples [Arz+11] pour offrir une efficacité d'architecture, c'est-à-dire de rapport entre le débit et la complexité du matériel, supérieure à celle l'état de l'art. L'architecture à flux multiples est illustrée dans la Fig. 5.3.5a où tous les flux et les portes logiques sont dupliqués p fois (p>2) et la sélection aléatoire des bits est faite par un simple brasseur ou *shuffler* illustré dans la Fig. 5.3.5b. Le *shuffler* est constitué de p bascules JK – fournissant la normalisation stochastique –, d'un *barrel-shifter* de p bits et de p multiplexeurs. Le *barrel-shifter* est purement combinatoire. Sa valeur de décalage V est mise à jour à chaque cycle de décodage (DC) – qui correspond à la sortie d'un nouveau bit de tout module de conversion stochastique illustré en Fig.5.3.1– et est la même pour tout *barrel-shifter*.

Pour simplifier davantage l'architecture, la règle de brassage peut être déterministe pour éviter des RNG supplémentaires, par exemple en décalant circulairement les bits d'une position vers la gauche à chaque cycle d'horloge. De plus, la représentation d'une probabilité par p flux stochastiques au lieu d'un seul divise le nombre de DCs par p pour assurer la même précision. Ainsi, cette nouvelle architecture a un débit similaire à toute autre architecture stochastique parallélisée au degré p comme suggéré dans [GR03]. Par conséquent, l'architecture à flux multiples offre une solution performante en terme d'efficacité d'architecture, c'est-à-dire de rapport entre le débit et la complexité du matériel.

Nos différents travaux sur décodage stochastique ont donné lieu à une publication majeure sur le décodage stochastique de turbocodes convolutifs [Don+10] que nous intégrons dans le document en Chapitre 9 section 9.2. Elle résume les travaux que j'ai conjointement menés avec Christophe Jégo et Quang Trung Dong dans le cadre de son doctorat et qui ont fait l'objet de nombreux échanges avec Warren J. Gross de l'Université McGill.

5.3.2.1 Bilan et perspectives pour le décodage stochastique

En parallèle de ces travaux sur les turbocodes convolutifs, j'ai collaboré aver Camille Leroux lors de son séjour à l'Université de McGill dans le cadre du stage de fin d'études de



(a) Architecture à flux multiples.

(b) Brasseur (*Shuffler*) mélangeant $p = 4$ flux représentant la même probabilité.

Figure 5.3.5 – Architecture proposée où chaque probabilité est représentée par plusieurs flux stochastiques (4 ici) traités par des unités logiques parallèles et interconnectées par des *shufflers*.

Romain Héloir. Celui-ci s'est intéressé au décodage stochastique de codes Reed-Solomon selon l'algorithme de Chase [HEL+12] et a atteint un débit de 800Mb/s sur Virtex 5 pour un code RS(255,239). J'ai aussi collaboré au décodage des codes Cortex proposés par J.C. Carlach [CV99]. Un décodage analogique performant fut proposé [Per+09c] et j'ai donc proposé par la suite un décodage stochastique [Arz+11] d'une classe particulière, les codes C4 qui furent brevetés [Per+10].

Pour résumer mes travaux, **j'ai collaboré à démontrer que le traitement stochastique est possible et performant pour décoder les turbocodes et les codes Cortex. Pour cela j'ai innové en proposant des schémas de circuits proches de ce que l'analogique avait démontré comme efficaces, transformant les additions en multiplications dans le domaine exponentiel et en inventant le concept de traitement stochastique à flux multiples.**

Néanmoins, les décodeurs industriels restent, à notre connaissance, conventionnels, sans aucun traitement stochastique. Warren J. Gross a largement œuvré pour la promotion du traitement stochastique auprès de l'industrie, mais sans succès. Le frein principal reste l'usage de flux aléatoires ou pseudo-aléatoires, compliquant la prédictibilité des comportements, et imposant donc probablement des certifications très coûteuses voire impossibles. Cela n'empêche cependant pas la communauté scientifique de toujours innover en traitement stochastique. En outre, il a été montré que le décodage stochastique peut être avantageux dans les systèmes contraints en énergie ou à haut débit [HEG17]. Il a été montré que grâce à la robustesse aux fautes du traitement stochastique on peut grandement réduire la consommation

énergétique ou le débit en abaissant la tension d'alimentation ou en augmentant fortement la fréquence d'horloge du système tout en maintenant de bonnes performances [HEG17].

François Leduc-Primeau a grandement contribué et ouvert plusieurs voies d'intérêt. La première est de tirer le meilleur de chaque monde en mélangeant les représentations quantifée et stochastique au sein d'un décodeur LDPC [Led+13], offrant ainsi des performances de décodage remarquables. La seconde est de ré-exploiter le traitement stochastique dans un domaine de traitement de l'information au sein de résaux de calculs fortement connectés comme les réseaux de neurones profonds [Ard+17]. Cette solution stochastique entière (ne se limitant pas à des flux binaires mais à des flux entiers de faible radix, sur 2 ou 3 bits par exemple) a permis de réduire la consommation énergétique de 21% par rapport à l'état de l'art, sans perte de précision de classification. Il nous semble que ces derniers travaux méritent de plus amples investigations, **notamment en concevant de manière conjointe des réseaux profonds et des classifieurs**. Si l'apprentissage se fait en connaissant la nature stochastique de l'inférence, il est probable que la classification soit bien plus efficace que celle de l'état de l'art.

5.4 Adéquation algorithme-architecture pour des récepteurs flexibles à complexité réduite

5.4.1 Le Graal des décodeurs universels

En 2010, au début de la thèse de Jean Dion en contrat CIFRE avec Orange Labs, que j'ai encadré, turbocodes et LDPC étaient en concurrence aussi bien dans le monde académique que dans les comités de normalisation, en raison de leurs performances en termes de correction d'erreurs, de complexité, de latence de décodage, et en raison de leur adaptabilité et de leur flexibilité en termes de longueurs et de débits de mots de code. Ainsi, plusieurs normes IEEE, DVB et 3GPP les ont intégrés.

Puisque les terminaux de communication et multimédia doivent être conformes à plusieurs de ces normes de télécommunication simultanément, ils intègrent plusieurs récepteurs, ce qui augmente linéairement leurs coûts. Pour réduire ces derniers, la radio cognitive a été proposée comme une solution élégante, en essayant de maximiser la polyvalence des récepteurs au moindre coût. Le dernier étage d'un récepteur polyvalent est le décodeur canal qui n'est pas polyvalent par nature, les décodeurs turbo et LDPC classiques n'ayant que peu de propriétés en commun. La thèse de Jean Dion a eu pour objet une architecture de décodeur unique, capable de prendre en charge des mots de code de type LDPC et turbo, même s'ils sont multiplexés dans le temps, tout en offrant les performances d'un système

conventionnel multi-décodeurs mais à un coût matériel réduit. Cela était perçu par Orange Labs comme un verrou technologique majeur.

Le département Électronique traitait déjà en partie cette problématique par les travaux d'Amer Baghdadi qui explorait les solutions à base de processeurs à jeux d'instruction dédiés, *Application Specific Instruction-set Processor* (Application Specific Instruction-set Processor (ASIP)). A travers les projet ANR AFANA (*Application-Field-Aware Adaptive Network on chip Architecture* et UDEC (*Universal channel DECoder*) auxquels je pris part à mon arrivée au département, et les thèses de Purushotham Murugappa et Rachid Al-Khayat, il fut proposé une architecture multi-ASIP, interconnectée par un *Network on Chip* (NoC) flexible et haut-débit capable de décoder des codes LDPC et turbo [Mur+11]. Stimulés par une activité forte dans la communauté scientifique (avec notamment les résultats prometteurs de [CMM12 ; CMM13] et les travaux de l'équipe de N. Wehn [AVW08 ; BIW11]) ces travaux ont perduré et ont abouti sur une solution exploitant astucieusement les capacités de reconfiguration dynamique des FPGA [Lap+16]. L'état de l'art s'est enricihi et propose même une solution à quatre modes : codes convolutifs, turbo, LDPC et polaires [QLL18].

En alternative à cette approche ASIP, j'ai proposé que nous nous intéressions à une approche algorithmique en cherchant une technique de décodage conjointe aux codes LDPC et turbo convolutifs pour laquelle nous pourrions proposer une architecture dédiée pour une intégration ASIC la plus efficace possible en termes de consommation de surface de silicium, d'énergie et de débit, telle que cela semblait possible en étudiant la littérature. Partant des idées originales de représentations de codes LDPC sous formes de treillis similaires à ceux des codes convolutifs exploités par les turbocodes [MS02] et les idées alternatives à bases de codes à répétition généralisés [NBC06] et des architectures communes turbo convolutif et LDPC [SC08 ; Chi09 ; GTJ10 ; Nae+10 ; GRF10], nous nous sommes lancés dans cette voie.

L'algorithme et l'architecture proposés ont visé les codes LDPC quasi-cycliques des normes telles que 802.11n et 802.16e [DIO+12a] et les turbocodes Long Term Evolution (LTE) avec un décodage selon une représentation commune en treillis [DIO+12b]. Les mots de code de chaque norme ont été décodés avec une dégradation maximale de 0,20 dB par rapport aux meilleurs algorithmes de décodage, les algorithmes Log-Maximum A Posteriori (MAP) ou Layered Belief-Propagation Sum-Product, dans leur version en virgule flottante, et avec un nombre équivalent d'itérations. La structure multistandard que nous avons proposée ne requiert que 5% de ressources matérielles supplémentaires par rapport à une structure compatible uniquement avec le standard 3GPP LTE. Nous avons aussi montré qu'une réalisation ASIC en CMOS 65nm peut fonctionner à une fréquence de 500 MHz pour une surface de 1,7 mm^2 décodant les mots de code 3GPP LTE et IEEE 802.11n, et acceptant une reconfiguration dynamique entre deux mots de code consécutifs. Dans ce cas, la thèse a montré que

les débits atteignent 475 Mbps en mode 3GPP LTE et 482 Mbps en mode IEEE 802.11n.

5.4.2 Détecteurs-décodeurs conjoints au secours du MIMO

Mes travaux de recherche m'ont amené aussi à revoir l'architecture du récepteur pour que celui-ci profite plus pleinement du **potentiel du traitement sur circuit**. En effet, la conception des systèmes de réception est généralement confiée à des experts en traitement du signal plus qu'en conception de circuits, ce qui bride parfois la performance des systèmes. Dans le cadre de la thèse d'Ali Haroun, co-encadrée avec Charbel Abdel-Nour et dirigée par Christophe Jégo, nous avons cherché à associer la détection *Multiple-Input Multiple-Output* (MIMO) et le décodage correcteur d'erreur dans un cadre commun d'application de l'algorithme de propagation de croyance. Pour cela, nous avons proposé d'exploiter des codes LDPC non-binaires pour les associer le plus directement à des symboles de modulation et permettre un échange itératif efficace entre détection et décodage.

Les systèmes multi-antennaires (MIMO) sont devenus un élément essentiel des normes de communication sans fil, notamment IEEE 802.11n (Wi-Fi), IEEE 802.16 (WiMAX) et LTE. Malgré les nombreux avantages que présente l'application d'un traitement MIMO itératif au niveau du récepteur, la complexité de calcul et la latence de cette approche représentent toujours un défi. En effet, une détection conventionnelle par calcul de distances euclidiennes du symbole reçu vis-à-vis des différents symboles candidats peut vite devenir un cauchemar quand les modulations sont d'ordre élevé (8-Phase Shift Keying (PSK), 16-Quadrature Amplitude Modulation (QAM),32-Amplitude and Phase Shift Keying (APSK) et même au-delà) sur plusieurs antennes.

Le traitement conjoint à passage de message profite d'une alternative pour la détection MIMO qui est basée sur le principe de la propagation des croyances (Belief-Propagation (BP)) [KDK05], largement étudiée pour le décodage LDPC. D'autres études sur le détecteur MIMO basées sur l'algorithme BP ont été proposées dans [Kao1 ; YXW07]. Nous avons contribué à cette voie [Har+14c ; Har+14a ; Har+14b] en associant itérativement la détection MIMO et le décodage de codes LDPC non-binaires, Non Binary Low-Density Parity-Check (NB-LDPC), tous les deux basés sur l'algorithme BP de telle sorte que le récepteur fonctionne sur un grand graphe facteur conjoint [Har+14b] illustré en Fig.5.4.1.

L'échange de messages sur un tel graphe n'est pas sans rappeler ce qui fut fait au sein des décodeurs massivement parallèles analogiques et stochastiques de mes travaux précédents, à une différence près et non des moindres : l'étage de détection a un sous-graphe très local et prend une décision forte qui peut impacter violemment les couches de décodage. La méthode des diagrammes EXtrinsic Information Transfer (EXIT) [Brio1] a donc été

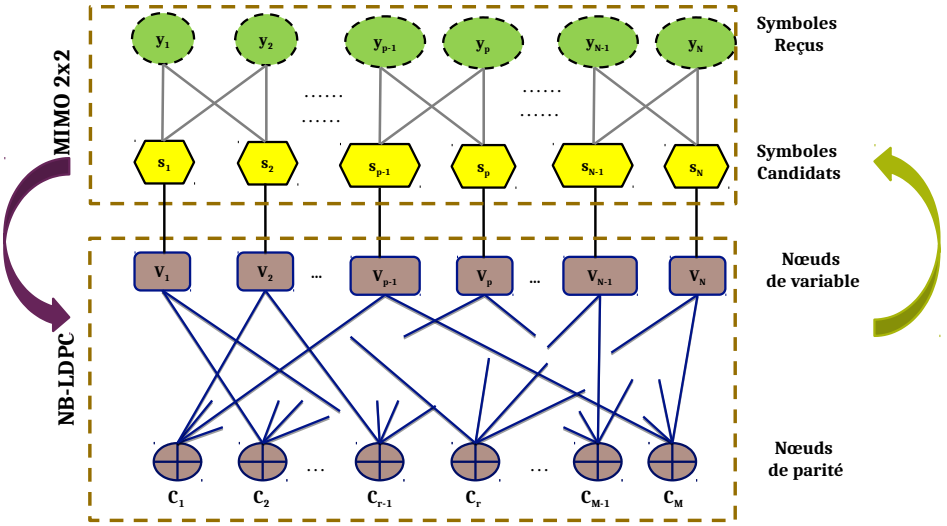


Figure 5.4.1 – Graphe facteur conjoint à la détection MIMO et au décodage LDPC.

appliquée au cas non-binaire que nous avons considéré afin d'analyser la convergence d'un récepteur itératif favorable à une intégration matérielle. En effet, une détection MIMO-BP et un décodage NB-LDPC peuvent être combinés sur différents ordonnancements, dont l'ordonnancement turbo, très générique et très simple à mettre en œuvre et l'ordonnancement *shuffle* ayant une convergence plus rapide. Les départager demanda un usage éclairé des diagrammes EXIT. Ensuite, nous avons visé une réduction de la complexité calculatoire du récepteur en vue d'une réalisation matérielle. Nous avons ainsi proposé une solution algorithmique qui réduit grandement la complexité de calcul de l'estimation de la distance euclidienne [Har+16] tout en offrant une détection efficace. Cette approche évite de calculer toutes les distances euclidiennes des différents symboles candidats afin d'identifier la solution à maximum de vraisemblance, Maximum Likelihood (ML). En effet, une fois la solution ML trouvée, nous limitons le calcul des distances aux points de la constellation les plus proches de cette solution. Nous avons montré que l'algorithme résultant offre un très bon compromis entre performance et complexité de calcul. Une réduction de 95% du nombre de multiplications nécessaires pour estimer les distances euclidiennes est obtenue sans aucune dégradation des performances par rapport à la détection BP conventionnelle. De plus, la technique proposée nécessite un nombre plus faible de multiplications réelles par rapport à d'autres algorithmes comme le décodage sphère [AA09].

5.5 Le défi de l'optique numérique

5.5.1 Mon initiation dans le projet FUI 2009 100GFLEX

Dans cadre du projet FUI 2009 100GFLEX, nous avons contribué à étudier l'intérêt des techniques Orthogonal Frequency Division Multiplexing (OFDM) multi-bandes pour la montée en débit des transmissions optiques. Les nombreux partenaires étaient Mitsubishi-Electric R&D Centre Europe, Ekinops, France Telecom/Orange Labs, Yenista Optics, Institut Télécom et l'Université de Rennes I. Bien avant de traiter le problème de la correction d'erreur dans un système, il faut être capable en réception de détecter correctement les signaux émis. Cela requiert en communications numériques une synchronisation efficace en amont de la détection pour retrouver avec une précision suffisante la porteuse, le rythme d'échantillonnage et sa phase. Dans le cadre d'un récepteur optique cohérent OFDM, la synchronisation est d'autant plus un défi que les performances de la modulation OFDM sont très sensibles aux défauts de synchronisation. Certains algorithmes de synchronisation en temps et en fréquence mis au point pour les radiocommunications peuvent alors présenter un intérêt pour la mise en œuvre d'un récepteur OFDM optique cohérent comme l'ont montré Raphaël Le Bidan et Thierry Le Gall dans ce projet. Ils ont notamment passé en revue les algorithmes de synchronisation temps et fréquence de l'état de l'art et ont comparé leurs complexités et performances par des simulations sur canal à dispersion chromatique Pour ma part, j'ai participé avec Gérald Le Mestre (ingénieur de recherche) et Arun Kumar (ingénieur sous contrat le temps du projet) sur la validation de ces estimations de complexité et leur parallélisation pour permettre un traitement au débit de plusieurs Gbit/s sur circuit. Ensuite, nous avons réalisé l'intégration de la meilleure solution dans un prototype basé sur un FPGA. Cette solution exploitait des transformations de Fourier rapide (FFT) pour permettre à des algorithmes itératifs de travailler dans le domaine fréquentiel et temporel.

Ces premiers résultats m'ont convaincu de l'intérêt de m'impliquer dans la recherche autour des récepteurs numériques pour les communications optiques. Les besoins étaient clairs, les enjeux technologiques, pour l'industrie et la société numérique, majeurs et les possibilités d'innover me sont apparues comme nombreuses si je tirais partie de mes connaissances en adéquation algorithme-architecture pour aider dans les choix algorithmiques et proposer des solutions d'intégration matérielle à haut débit.

5.5.2 Traitement numérique du signal pour les futures générations de réseau d'accès optique passif

Par la suite, j'ai donc poursuivi ces travaux dans le cadre non plus des réseaux optiques longue distance mais plutôt dans celui des réseaux d'accès optique passifs, *Passive Optical Network* (PON), qui ont la contrainte de devoir rester extrêment peu coûteux car distribuées à très grande échelle en desservant les utilisateurs terminaux. J'ai notamment encadré la thèse CIFRE de Tuan Anh Truong, dirigé par Michel Jézéquel, et supervisé par Bruno Jahan et Hao Lin à Orange Labs.

La thèse évalua notamment l'intérêt de la modulation OFDM pour les PON notamment associée à des techniques de répartition de charge de type Levin-Campello. Sur la base de la technique de précodage par transformée de Fourier discrète (DFT), un système OFDM précodé fut également proposé [TRU+14a]. Ayant le même rendement de débit de données que le système Levin-Campello de référence, le système proposé réduit le facteur de crête, *Peak-to-Average Power Ratio* (PAPR), d'un signal transmis de 2dB (pour un *clipping rate* raisonnable de de 10^{-3}), ce qui entraîne une réduction de la consommation d'énergie de l'amplificateur de puissance du côté émetteur. Cette contribution fut marquante puisqu'elle démontrait très clairement que **le traitement du signal permet de conserver le matériel en place dans les PON et d'en augmenter les performances avec un budget énergétique contenu**.

De plus, par rapport à la modulation OOK, la modulation OFDM s'est révélée plus robuste contre la dispersion chromatique et les pics de laser. Enfin, des techniques innovantes furent proposées pour compenser les inconvénients de l'OFDM, notamment l'important *Peak-to-Average-Power-Ratio* (PAPR) et la sensibilité à la synchronisation. Il a été démontré que les techniques de synchronisation proposées sont plus performantes que les techniques conventionnelles [TRU+14b]. L'attrait de cette approche fut notamment d'allier des traitements simples et un ordonnancement en deux temps pour réduire la complexité et augmenter les performances des PON comme je l'avais appris lors de mes travaux antérieurs sur le projet 100GFLEX.

Enfin, une technique de *Tone Reservation* fut proposée, réduisant le PAPR jusqu'à 4 dB dans un système OFDM précodé [TRU+13]. Ces travaux furent toujours ramenés à la considération d'une intégration matérielle rendue possible par une complexité algorithmique contenue. Je me suis continuellement efforcé d'y veiller et de l'inculquer au doctorant que nous avons formé. L'intégration sur circuit serait une poursuite intéressante dans ce domaine du PON où le coût est un enjeu majeur et le champ de recherche est ouvert. **Proposer des solutions portables sur peu de ressources matérielles au sein de FPGA de faible coût**

est une perspective d'intérêt pour promouvoir des architectures de PON évolutives et pérennes.

5.5.3 FPGA en remplacement d'ASIC pour l'optique 100Gbps flexible

En communications optiques, même si les FPGA sont souvent considérés comme un moyen de démontrer les performances de décodage et d'assurer que le taux d'erreur post-correction ciblé est atteint, comme dans [ZD15], les ASIC restent des solutions privilégiées pour les implémentations à haut débit [TLW17 ; Kum+15]. Ces puces bénéficient des performances des circuits numériques dédiés mais au prix d'une absence de flexibilité. Si des modifications sur le schéma de correction d'erreur sont nécessaires, un nouveau circuit doit être conçu, avec un coût considérable en termes de temps (plusieurs mois à années) et de ressources humaines (possédant un large spectre de compétences). Pour répondre à ce même problème mais à des débits moindres, les systèmes sans-fil exploitent des solutions flexibles depuis de nombreuses années. Ces solutions sont basées sur des dispositifs matériels reconfigurables tels que les FPGA ou sur des dispositifs matériels programmables tels que les Digital Signal Processor (-ing) (DSP), les General Purpose Processor (GPP) et même les Graphics Processing Unit (GPU), certes pour des débits plus faibles d'un ordre ou deux de grandeur.

Toutes ces solutions programmables/reconfigurables offrent différents compromis entre la flexibilité de traitement et le débit de données. Grâce aux dernières avancées technologiques, de telles solutions peuvent être intéressantes pour concevoir des liaisons optiques 100G flexibles, et surtout pour mettre en œuvre le codec de correction d'erreurs qui est à la fois le goulot d'étranglement du débit et la technologie clé vers la performance. Grandement intéressé par ces perspectives, j'ai répondu à un appel à projet lancé par Huawei en 2018 afin d'étudier la mise en œuvre de codes de correction d'erreurs avancés tels que les codes LDPC convolutifs sur du matériel programmable/reconfigurable afin de permettre de retirer les puces spécialisées des systèmes optiques et d'offrir ensuite plus de flexibilité. Des codes convolutifs de type Quasi-Cyclic Low-Density Parity-Check (QC-LDPC) [FZ99 ; Tan+04 ; Pus+11] ont été sélectionnés pour leur performance en correction et leurs décodeurs à coût matériel et à latence maîtrisés.

Atteindre des débits de correction d'erreurs de plusieurs dizaines de Gbps sur du matériel programmable/reconfigurable est un défi car il nécessite un niveau élevé de parallélisme de traitement. Non seulement chaque unité matérielle doit exploiter pleinement son parallélisme interne, mais de nombreuses unités matérielles doivent également être mises en parallèle. Cela suppose une conception fine du système et un schéma de correction d'erreur adapté.

Nous avons étudié les solutions possibles pour différentes plateformes matérielles programmables/reconfigurables et sélectionné celle qui est la plus adaptée aux spécifications des communications optiques 100G, en fonction des entrées de Huawei (codes QC-LDPC ciblés, contraintes telles que la latence et la consommation d'énergie). Ensuite, nous avons conçu une architecture de décodeur QC-LDPC capable de supporter le débit ciblé. Nous avons fourni une implantation logicielle optimisée sur cœur Intel I9 et une autre matérielle sur FPGA Xilinx Ultrascale+ VU13P pour un décodeur QC-LDPC convolutif flexible et à haut débit. Plus spécifiquement, nous avons automatisé par des scripts en Python l'analyse du code QC-LDPC convolutif et de ses propriétés pour optimiser l'ordonnancement et le contrôle du décodeur et conçu un outil de placement/routage complètement automatisé pour la conception sur FPGA multiples. Nous avons testé et caractérisé plusieurs configurations (nombre d'itérations, algorithme de décodage, protomatrice...) démontrant ainsi la flexibilité offerte par notre solution, mais aussi validant les principes et équivalences proposés.

Ce domaine du décodage haut-débit flexible sur FPGA est d'un intérêt notable pour mes activités de recherche et requiert une expertise forte difficilement disponible. Cette voie de recherche fait partie de mes perspectives décrites ci-après.

5.6 Perspectives en traitement numérique haut-débit pour les communications

5.6.1 Décodeurs haut-débit et flexibles pour les communications satellitaires en bande Ka et optiques

Dans le cadre de la thèse CIFRE d'Aomar Bourenane avec Safran, j'investigue la conception d'un récepteur à haut débit et consommation restreinte pour les futures générations de communication satellitaire en orbite basse. La direction est assurée par Frédéric Guilloud (IMT Atlantique) et la supervision chez Safran par Alain Thomas.

L'observation de la Terre est au cœur de nombreux enjeux civils et militaires qui justifient le lancement et le déploiement croissant de satellites en orbite basse. Leur nombre croît massivement, amenant une saturation de la bande de fréquence actuelle (bande Ku entre 12 et 18 GHz), ce qui impose que la prochaine génération de satellites exploite de nouvelles ressources de transmission. Une première solution exploiterait la bande Ka entre 27 et 40 GHz tandis qu'une seconde se tournerait vers des transmissions optiques en espace libre. En effet, la transmission optique présente un meilleur rapport masse / consommation au Mbps que la transmission par radiofréquence pour les satellites d'observation. De nombreux démonstrateurs sont donc en cours de développement, voire en vol comme OSIRIS2 du

DLR [SF17] ou Optel-mu de Ruag [Dre+12].

Pour la première solution de montée en bande Ka, le codage de canal doit évoluer. L'ESA a notamment fait la promotion d'un turbocode convolutif à concaténation série, *Serially Concatenated Convolutional Code* (SCCC), que l'on retrouve dans les propositions du Consultative Committee for Space Data Systems (CCSDS) [CCS12 ; CCS19]. Or, à notre connaissance, il n'existe pas de solution très haut débit dans l'état de l'art pour les SCCC, ce qui représente donc un sujet d'investigation à privilégier. La seconde solution basée sur une transmission optique du satellite vers la Terre semble particulièrement adaptée pour les micro/nanosatellites dans une plage allant jusque 10Gbps, ou pour les plateformes où des débits de plus de 100Gbps sont envisageables. Nos travaux antérieurs sur les communications optiques terrestres y trouveront une continuité en s'adaptant aux spécificités de l'espace libre et aux contraintes de systèmes satellitaires à ressources de calcul et d'énergie limitées.

D'un point de vue scientifique les défis sont conséquents et nécessitent la mise en œuvre de techniques avancées en traitement du signal et en conception conjointe d'architectures de récepteurs. La recherche portera ici sur des innovations tant en algorithmie qu'en architecture matérielle sous des contraintes de consommation d'énergie et d'optimisation conjointe.

Les premiers travaux du doctorant ont d'ailleurs fait l'objet d'un dépôt de brevet dans le sens d'adéquation algorithme de décodage - architecture basse consommation.

Ces travaux ont en outre une portée sur les communications terrestres pour lesquelles les communications optiques en espace libre sont de plus en plus considérées pour pallier le manque de ressources dans le spectre radio et éviter le coût d'infrastructures optiques fibrées. Il s'agit clairement d'un domaine de recherche ouvert à l'innovation, et appelé par l'industrie des télécommunications.

5.6.2 Revoir le segment sol des communications satellitaires avec *RF over IP*

5.6.2.1 Evolution du réseau d'accès satellitaire

Conventionnellement, les antennes au sol des réseaux satellitaires sont adossées à des stations de base qui effectuent le traitement du signal requis pour recevoir/émettre l'information portée par le signal électromagnétique capté/à émettre par l'antenne. Une alternative déjà exploitée par les réseaux mobiles les plus récents (5G notamment) consiste en un déport de ce traitement du signal dans des datacentres (ou au minimum des sites de traitement distants, d'accès aisé et centralisant les ressources de calcul) qui collectent les signaux radio-fréquence (ou en fréquence intermédiaire) numérisés à haute fréquence et qui y appliquent les traitements auparavant effectués directement dans les stations de base : on peut dénommer ce concept (au sens large) le cloud-RAN. Il recouvre une problématique

complexe en architecture des réseaux, gestion des réseaux et en traitement du signal et de l'information.

Dans le cadre d'un contrat pour l'ESA, en collaboration avec Abdeldjalil Aïssa El Bey (Département Signal et Communications), Widenorth et Simula UiB, j'explore cette problématique du *cloud-RAN* **satellitaire** et notamment la question de la transmission des signaux radiofréquence (RF)[1], numérisés et formatés en paquets, entre l'antenne et les unités de traitement du signal qui sont déportées dans un datacentre. Cette technique est souvent appelée RFoIP, *RF over IP* ou Intermediate Frequency over Internet Protocol (IFoIP) puisque les réseaux de transport de paquets sur Internet Protocol (IP) sont souvent retenus dans l'actuel segment sol des réseaux satellitaires.

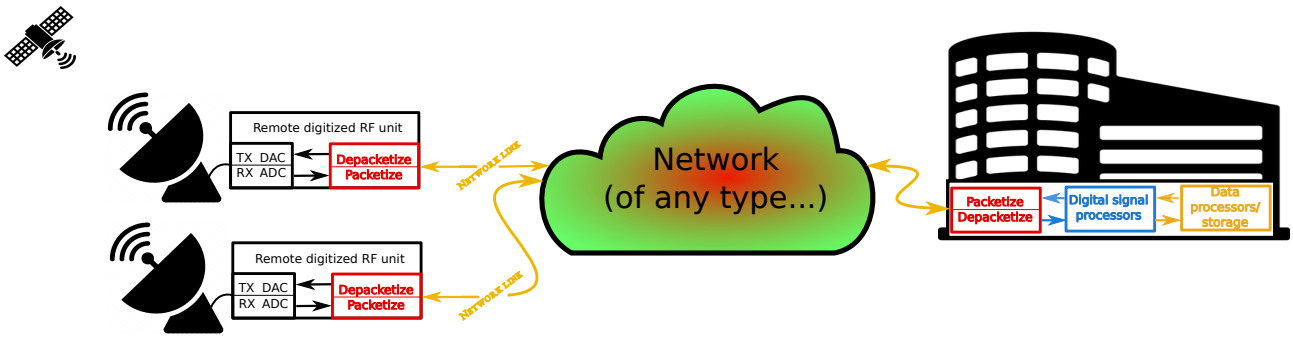


Figure 5.6.1 – Principe du RFoIP, première étape vers le cloud-RAN satellitaire.

Ce concept de RFoIP illustré en Fig.5.6.1 offre des opportunités d'une grande valeur pour les opérateurs. Cela permet de concentrer et mutualiser la puissance de calcul requise par ces traitements de l'interface air dans des datacentres (qui peuvent être au sein du *cloud* ou directement chez l'opérateur/exploitant satellitaire), rendant les stations de base plus simples et plus génériques, car limitées aux étages RF et de conversion analogique-numérique couplés à des interfaces réseaux par paquets, donc plus facilement implantables dans différents environnements. Les opérateurs et exploitants peuvent ainsi multiplier les antennes satellites au sol à moindre coût et augmenter les performances de leurs réseaux d'accès.

Actuellement, deux géants du numérique proposent des services de cloud-RAN satellitaire : Amazon Web Services (AWS) Ground Station[2] et Microsoft Azure Orbital[3], tous les deux en collaboration notamment avec Kratos, actuel leader du marché RFoIP satellitaire. Ces deux géants maîtrisent leurs réseaux et cependant, le RFoIP large bande semble être

1. de type DVB-RCS, Digital Video Broadcasting - Satellite (DVB-S), Digital Video Broadcasting - Satellite - Second generation (DVB-S2) ou Digital Video Broadcasting - Satellite - Second generation eXtension (DVB-S2X)
2. https://aws.amazon.com/fr/ground-station/
3. https://azure.microsoft.com/fr-fr/services/orbital/

un réel défi, même pour eux. Comme expliqué durant l'*AWS Public sector summit*, en juin 2019 (https://www.youtube.com/watch?v=YhZe6eOPoRQ), AWS peut transporter les signaux des bandes S et X. En bande S, avec des bandes passantes de 54 MHz au maximum, les signaux radio sont bien numérisés et transportés par paquets en voies montante et descendante, mais en bande X en voie descendante (bande passante maximale de 500 MHz), AWS ne peut pas faire mieux que de transmettre une bande passante de 325 MHz mais sous forme de **signaux démodulés** et absolument pas les signaux radio originaux à la sortie de l'antenne.

Passer de la bande S à la bande X semble imposer à l'actuel réseau d'AWS un surcoût qui rende impossible RFoIP au delà de la bande S. Passer à la bande Ka et au-delà avec des bandes passantes de plusieurs GHz semble donc totalement déraisonnable avec les seules techniques actuelles.

5.6.2.2 Nécessité et défis d'une compression de signal large bande

Il serait tentant de conclure hâtivement en prétendant que les solutions du cloud-RAN mobile sont disponibles et peuvent être réutilisées pour le RFoIP satellitaire large bande. Cependant, l'interface air des réseaux d'accès mobile est très différente de celle du satellitaire : elle repose sur des techniques comme l'OFDM, MIMO et *beamforming*, pour augmenter la capacité de l'interface, tandis qu'un lien satellite pointé tire profit d'une large bande passante allouée sans aucun besoin des techniques précédemment citées. Ainsi, les normes LTE et LTE *Advanced* exploitent des bandes radio d'au plus 20 MHZ (100 MHz en LTE *Advanced*) de 450 MHz à 3,8 GHz qui peuvent être agrégées pour fournir un débit utilisateur au maximum de 300 Mbps (452 Mbps en LTE *Advanced*). Les produits réellement RFoIP ainsi actuellement disponibles sont le plus souvent limités à une bande passante de 125 MHz.

Pour permettre la montée en bande radio, nous visons **une bande passante numérisée de manière instantanée de 5GHz, typiquement requise pour les bandes satellitaires comme la Ka et la Q/V**, en tirant profit des toutes dernières puces de conversion mises sur le marché par Texas Instruments et Analog Devices. Il s'agit d'une bande passante instantanée numérisée surpassant grandement celle des actuels systèmes. Nous pouvons en outre voir un lien fort avec la problématique de la précédente section. Je suis aussi particulièrement inspiré par les dernières avancées technologiques apportées par les Radio Frequency System on Chip (RFSoC) de Xilinx avec des convertisseurs analogique-numérique haute vitesse intégrés. En concurrence directe avec les puces produites par TI et Analog Devices en termes de bande passante (4 à 6 GHz) pour la RF et de fréquence d'échantillonnage (5 Géch/s sur les Analogue to Digital Converter (ADC) et 10 Géch/s sur les Digital to Analogue

Converter (DAC)), ils ont l'avantage d'une intégration étroite avec le cœur de traitement FPGA, ouvrant des perspectives d'architectures très performantes pour intégrer une chaîne d'**acquisition-numérisation-traitement-transport de signal**. En effet, ils permettent d'intégrer dans la même puce, si les ressources sont suffisantes, la pile protocolaire pour le transport de paquets données RF sur les réseaux de données numériques. Cet accès direct au réseau pour le signal RF sans passer par un CPU est un avantage majeur pour **réduire la latence** du transport.

Néanmoins, prenons le temps de bien évaluer le problème qui ne se limite pas à la disponibilité de matériel compatible. Afin de numériser une bande passante analogique, nous devons l'échantillonner avec un suréchantillonnage d'au moins 2 fois selon le théorème de Nyquist. Le rapport signal à bruit et la dynamique du signal des échantillons dépendent de la qualité et du nombre de bits permis par la conversion analogique - numérique. Si nous utilisons un CAN 16 bits à la fréquence de Nyquist, nous générons un débit de 16 x 2 x 5 Gbps = 160 Gbps lors de la numérisation d'une bande de 5 GHz. Ensuite, le surdébit supplémentaire des protocoles de transport (par exemple Real-Time Protocol (RTP)/User Datagram Protocol (UDP)/IP/Ethernet), ainsi que le surdébit de signalisation sont ajoutés, ce qui donne généralement un débit de 200 Gbps pour transférer une bande de 5 GHz sans faire de compression de données.

Le réseau qui supporterait ce RFoIP "brut" échangeant des signaux RF large-bande numérisés devrait supporter une charge de données considérablement supérieure à ce que l'on connait aujourd'hui. Prenons deux cas classiques suivants en norme DVB-S2X avec une unique porteuse exploitant 5GHz de bande. Dans des conditions de canal favorables, le système peut utiliser une modulation efficace 256APSK (efficacité spectrale maximale de 8 bits/s/Hz), avec un rendement de codage relativement élevé 3/4 et un *roll-off* égal à 1,05. Cela correspond à débit d'environ 28,6 Gbit/s pour des conditions de transmission de bonne qualité (canal favorable). Dans des conditions bien moins favorables (météo capricieuse), il est raisonnable d'augmenter la protection de l'information transmise et donc de réduire l'efficacité spectrale avec une modulation QPSK (2 bits par échantillon IQ) et un rendement de codage de 9/20. Le débit d'information utile est alors d'environ 4,3 Gbit/s. Si l'on exploite une voie RFoIP à 200 Gbit/s pour transférer cette information source, cela implique une charge du réseau multipliée par un facteur compris entre 7 et 46 selon les conditions de transmission !

Un tel système nécessiterait forcément l'augmentation physique des capacités des réseaux des opérateurs satellitaires et de leurs clients. De tels investissements peuvent être réduits par des **techniques de compression de données appliquées aux signaux RF**. Les algorithmes appliqués peuvent être aussi bien sans qu'avec pertes d'information tant que le système complet offre les mêmes performances de transmission de l'information que

son prédécesseur. Nous souhaitons ainsi proposer des solutions pour réduire la charge de données sur le réseau d'accès afin de le maintenir à un coût acceptable par les opérateurs. Il faut investiguer tant des techniques classiques de compression (comme un encodage adapté [Gol66 ; Kim+06]), avec ou sans pertes, combinées ou pas [VWC12] que des techniques sub-Nyquist telles que le Xampling [ME10].

Cependant, nous voyons plusieurs verrous à cette évolution technologique.

Le premier est la capacité du matériel actuel à traiter de tels débits. Quelles sont les technologies (conversion, traitement sur puce ASIC, FPGA, DSP) compatibles ?

Le second verrou est d'ordre algorithmique au premier abord mais profondément matériel au final : quelles sont les techniques les plus adaptées pour compresser le signal RF ou IF suffisament pour rendre l'exploitation du cloud-RAN rentable ? Cette question de codage/décodage de source (voire au niveau système avec le Xampling en remplacement du modèle de Nyquist) a un double matériel : quelles sont les architectures matérielles et les cibles d'implantation qui sauront exploiter l'éventuel parallélisme des algorithmes disponibles pour tenir le débit associé à des bandes passantes instantanées supérieures au GHz ?

Par ailleurs, cette adéquation algorithmes-architectures en codage/décodage de source offrira-t-elle suffisamment de performance pour garantir une fiabilité de transmission identique à celle offerte par les solutions actuelles ? En effet, nous craignons que les solutions actuelles de synchronisation et détection soient sensibles aux conditions de compression et nous supposons que le codage de canal ne pourra suffir à contenir la dégradation de performance du système si la compression-décompression dégrade trop fortement le signal original.

5.6.2.3 Les codes correcteurs d'erreur à la rescousse du cloud-RAN satellitaire

La problématique du cloud-RAN satellitaire est riche de défis, car même si l'on arrive à compresser le signal large bande efficacement sur un lien RFoIP, le débit requis par flux de données restera problématique pour le réseau terrestre de transport entre antenne et datacentre. Si l'on reconsidère l'objectif d'une bande passante instantanée de 5GHz à transporter par paquets sur un réseau terrestre, que l'on fait l'hypothèse d'une compression à 50% (comme dans le standard mobile Open Radio equipment Interface (ORI) [ETS14]), celui-ci doit être capable de supporter un flux de 100Gbps, soit l'équivalent d'un aggrégat de 2000 flux vidéo 4k60p (50 Mbps sur YouTube), avec fiabilité. Très clairement, tout réseau n'en est pas capable et cela suppose une architecture adaptée.

De mon point de vue, il est possible de classer les réseaux d'accès selon 4 grands modèles illustrés en Fig. 5.6.2 :

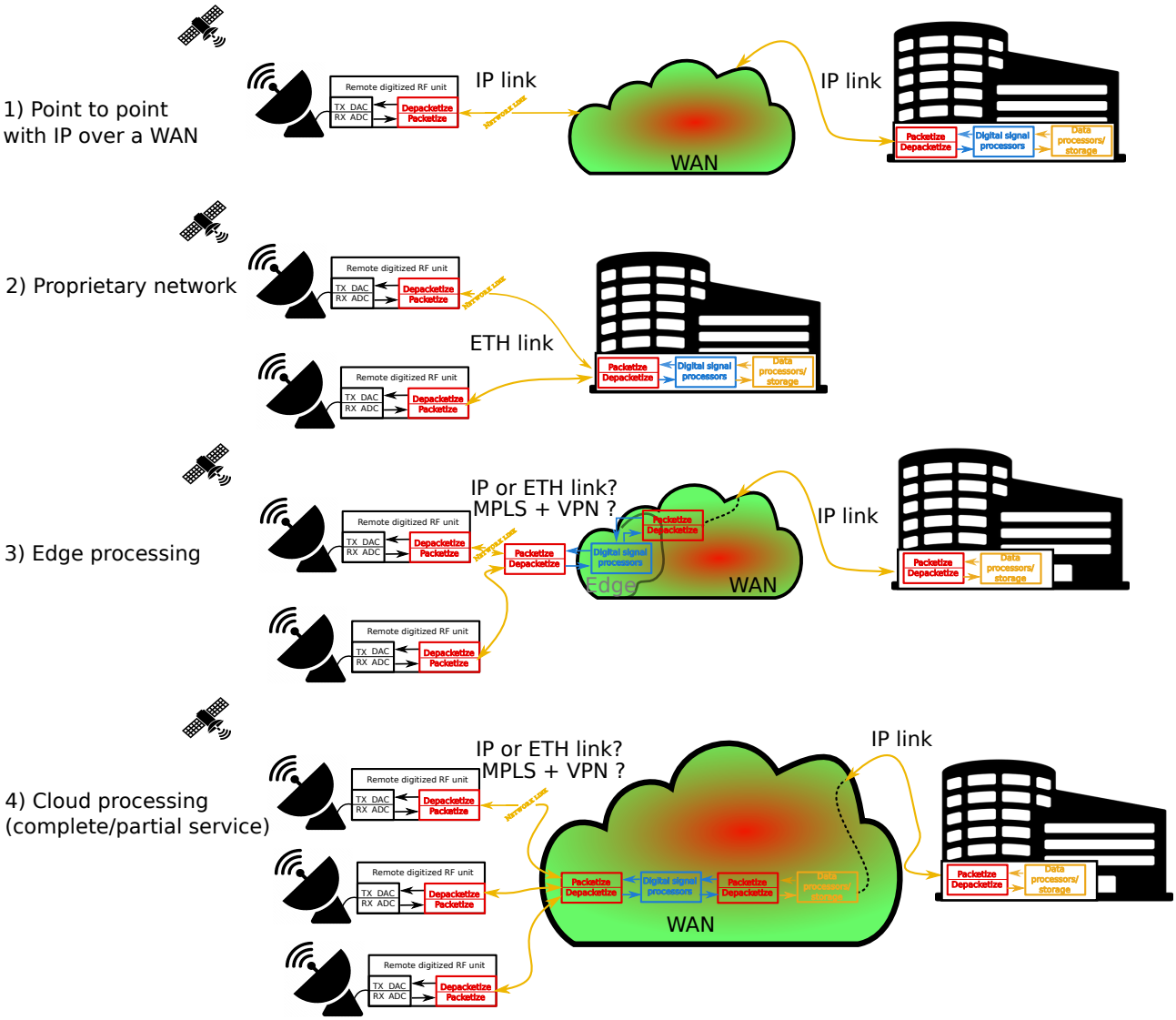


Figure 5.6.2 – Quatre modèles de réseaux vers le cloud-RAN satellitaire.

1. une connection point à point entre l'antenne distante et le datacentre est réalisée sur un réseau IP de type Wide Area Network (WAN) (ou une combinaison de tels réseaux), avec de nombreux goulots d'étranglement et une fiabilité de connexion très incertaine ; c'est un cas typique pour les actuelles stations sol pour l'observation de la Terre ;
2. un réseau propriétaire utilisant des ressources efficaces, comme des liens Ethernet dédiés, mais (trop) coûteuses supprime (ou du moins réduit) le risque de goulot d'étranglement et permet de garantir la fiabilité du transport de paquets ;
3. une architecture de réseau avec traitement en périphérie est une solution intermédiaire et se rapproche du choix fait pour la 5G (avec des sous-réseaux Ethernet spécialisés et spécifiques pour le RAN qui permet de réduire la charge du réseau à une quantité

réduite de ressources au plus près des sites d'antennes distants ; ces réseaux peuvent utiliser des solutions de type Multiprotocol Label Switching (MPLS) avec Virtual Private Network (VPN) ou Virtual Private LAN Service (VPLS) pour assurer la qualité du service ;

4. un service de traitement totalement dans le *cloud* peut être envisagé dans un avenir proche, grâce à l'augmentation de la bande passante du réseau et à la réduction des coûts, avec de fortes exigences en matière de qualité de service qui devraient être satisfaites à un coût raisonnable (également envisagé pour la 5G et au-delà).

Dans les cas 1 (actuelles stations sol d'observation de la Terre) et 4, les signaux de la chaîne RFoIP sont transportés par des paquets IP. Des normes industrielles telles que VITA-49 [09 ; CNC12] (pour le domaine satellitaire et la défense) ou eCPRI (pour les communications 5G) existent pour interconnecter un système RFoIP afin d'améliorer la maintenabilité et l'interopérabilité. Mais elles ne normalisent pas les solutions permettant d'utiliser le réseau pour transporter les paquets de données RF. Ces normes exigent des couches de protocole supplémentaires pour garantir une fiabilité de transport. VITA-49 est un standard L6, c'est-à-dire un protocole à la couche présentation, au-dessus de L5 (couche session) et L4 (couche de transport). Un protocole L5 couramment utilisé pour la diffusion de données en temps réel pour la vidéo ou l'audio sur IP est le protocole de transport en temps réel, RTP, sur un protocole de transport L4, comme UDP. C'est la solution retenue pour la voix sur IP (VoIP), la vidéoconférence et les services de *streaming* puiqu'elle est très simple et sans délai de retransmission. Malheureusement, elle est d'une fiabilité limitée : les paquets peuvent être perdus ou arriver dans le désordre. Transmission Control Protocol (TCP) est fiable et permet d'éviter la perte de données et de gérer leur ordonnancement, mais à un coût trop important sur le débit de transmission. Aucune combinaison de ces protocoles conventionnels ne peut être utilisée seule pour permettre un transport fiable des paquets RF sur un réseau de transport IP qui souffre d'une gigue et de pertes telles que la reconstruction du signal pour traitement de réception ou émission par l'antenne soit impossible avec les exigences de qualité requises par les services exploitant le lien satellite.

Une solution à ce problème peut être apportée par du codage paquet communément appelé Packet Forward Error Coding (P-FEC). J'ai déjà investigué cette technique pour un contrat de recherche avec EMC Norway que j'ai mené avec Amine Semma (stagiaire très talentueux) et Frédéric Guilloud (département Signal et Communications) en 2014. J'en retiens que pour notre problème du *cloud*-RAN satellitaire il serait pertinent de considérer tant des codes fontaines [Mac05 ; Lub02 ; Sho04]) que des codes anti-effacement plus anciens tels que les codes Reed-Solomon (RS) largement utilisés dans les systèmes de stockage (des

CD aux systèmes RAID6) et les systèmes de transmission comme code externe (comme dans de nombreuses normes DVB). Ainsi, une norme dédiée aux systèmes SATCOM, ANSI/ TIA-5041 [16] répond à ce problème de fiabilité avec une variante du schéma RS, les codes Cauchy RS [Blö+95] comme P-FEC sur la couche transport avec un protocole pour gérer la connexion et la transmission des paquets FEC. Les codes RS de Cauchy sont utilisés dans de nombreux systèmes de stockage et ont donc prouvé leur efficacité. Cependant, le code Cauchy RS de la norme ANSI/ TIA-5041 a un taux fixe de 192/255$\sim$75% qui pourrait trop affecter le débit du système si les paquets sont perdus très rarement, ce qui est difficile à prévoir et varierait selon l'état du réseau. Les codes de type Fontaine pourraient permettre un taux flexible mais certainement avec une quantité de paquets à coder trop importante pour la latence maximale requise par un système SATCOM. Quelle adéquation matériel-code-réseau est possible pour en plus accepter les débits visés de plusieurs dizaines à centaines de Gpbs ? Existe-t-il des solutions flexibles ? génériques ?

Les travaux de recherche sont donc nombreux dans ce domaine du *cloud*-RAN, allant du codage de source au codage paquet réseau, et requièrent une attirance forte pour l'adéquation algorithme-architecture en regard des débits visés. Il s'agit clairement d'une de mes perspectives majeures de recherche à court et moyen termes, et étroitement connexe aux travaux que je présente au chapitre suivant.

6

Architectures à haut débit et flexibles pour l'analyse de trafic réseau à plusieurs Tbps

Sommaire de ce chapitre

6.1 Problématique investiguée

Par les réseaux de données transite l'intégralité des échanges d'information. Ils offrent des services de plus en plus nombreux et variés aussi bien aux particuliers qu'aux entreprises. Ainsi, profitant des progrès technologiques de la couche physique, les débits ont explosé alors que la tâche des opérateurs pour gérer les données s'est complexifiée. En effet, ils doivent assurer d'une part la qualité de services exigeants comme la vidéo à la demande et d'autre part la sécurité de leur réseau. Ce dernier point est critique d'un point de vue économique comme politique. Pour offrir des solutions robustes face aux attaques, les opérateurs disposent de solutions logicielles qu'ils maîtrisent bien mais qui peinent à suivre la montée en débit et la diversification des attaques, notamment avec la croissance fulgurante de l'Internet des Objets, Internet of Things (IoT). Par exemple, en septembre 2016, plusieurs entreprises (KrebsonSecurity, Dyn, OVH, etc.) ont fait l'objet d'attaques par déni de service distribué, Distributed Denial of Service (DDoS). Ces attaques ont été perpétrées par un *botnet*, nommé "Mirai", d'équipements IoT (plus de 300 000 entités, contributrices malgré elles), des caméras de surveillance notamment. Cette attaque d'ordre planétaire a impacté des sites et services web très populaires comme Twitter, Netflix, Paypal et Spotify, certains étant alors devenus inaccessibles. Le record en septembre 2016 revient à OVH, hébergeur mondial, qui a subi une attaque historique de plus d'1Tbps ! Depuis ce franchissement du Tbps, deux attaques majeures par DDoS ont établi des records successifs : 1.3Tbps contre GitHub en mars 2018 suivie dans le même mois d'une attaque aux États-Unis à 1.7 Tbps.

La sécurité à très haut débit n'est pas le seul enjeu. Les réseaux supportent une diversité croissante de services avec des exigences plus ou moins fortes en termes de qualité de services. Fournir des services de navigation sécurisés, effectuer des transferts massifs de données pour de l'analyse ou de la sauvegarde sans exigence de réactivité ou fournir un service de visio-conférence multi-sites (voire mobile) haute définition (HD), ne requièrent pas les mêmes protocoles, ressources et priorités par exemple, pour garantir la qualité de service, *Quality of Service* (QoS), attendue. Et pourtant cette QoS, qui est la clé du succès commercial, ne doit pas être garantie par des investissements exagérés dans un réseau physique sur-dimensionné. Elle doit être garantie par un usage "intelligent" du réseau qui adapte son traitement et son transfert à la nature de son trafic, en étant capable de l'analyser le plus finement et le plus rapidement possible, permettant l'usage optimal des ressources à tout instant.

Mais est-il même possible de continuer à analyser et classifier le trafic avec justesse avec de telles contraintes de latence et débit ? En effet, comme l'ont analysé Noa Zilberman, Andrew W. Moore et Jon A. Crowcroft en 2016 [ZMC16], un fossé se creuse entre les capacités

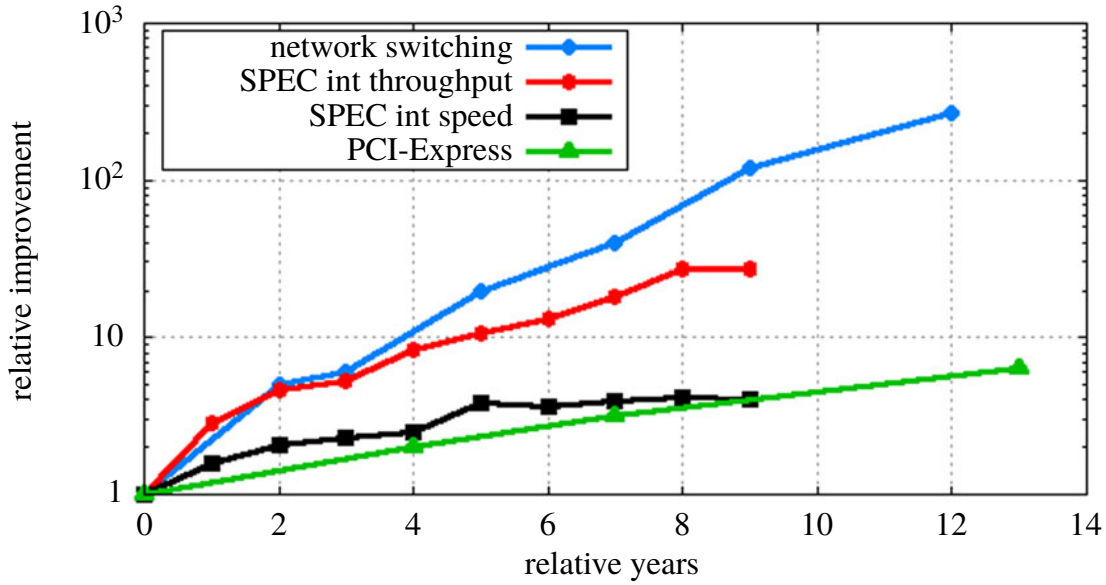


Figure 6.1.1 – Evolution des débits de traitement et transfert à partir de 2001 [ZMC16].

de transfert des réseaux et les capacités de traitement disponibles sur les unités de type CPU. La figure 6.1.1 souligne cette évolution en comparant l'évolution de la performance des CPU au banc de test *Standard Performance Evaluation Corporation* (SPEC) CPU2006 et l'évolution des capacités de transfert des liens réseau et bus PCIe. On y lit clairement la saturation des CPU pour l'analyse des données en transit sur un réseau. Même le PCIe ne peut pas suivre la croissance du débit sur les réseaux pour remonter les données aux CPU qui sont encore bien plus submergés par le volume de données alors produites. Les auteurs concluent à la nécessité de nouvelles architectures réseau combinant logiciel et matériel.

Pour aller dans ce sens de la collaboration logiciel-matériel, les FPGA ont été proposés comme solutions au plus proche des liens physiques de données dès les années 2006-2007 [Bia+06 ; Loc+07 ; SR07]. Devant permettre une plus grande agilité et des expérimentations aisées, à faible coût (comparées à des solutions ASIC interfacées avce des logiciels), ces solutions FPGA se sont vite révélées comme incontournables, de telle sorte que les grands fournisseurs de services exploitent de plus en plus les FPGA. Ainsi, Microsoft a déployé massivement ses solutions de SmartNIC au sein de son infrastructure Azure [Fir16], exploitant les FPGA pour décharger les CPU de tâches de traitement de données, arguant de la reconfigurabilité des FPGA comme avantage majeur vis à vis de solutions ASIC, trop coûteuses, non-agiles, requérant trop de temps pour être disponibles dans en environnement réseau qui change en quelques mois, voire semaines ou jours et dont le modèle actuel est le *Software Defined Network*. [Zil+15] détaille d'ailleurs les enjeux du SDN et les solutions matérielles disponibles pour permettre la reconfiguration dynamique des réseaux, et les FPGA

y tiennent un rôle majeur. Cela se confirme par la prise de position de grands acteurs des réseaux autres que Microsoft. Par exemple, *Amazon Web Services* (AWS) offre un accès à ses FPGA installés au sein de ses datacentres et OVH a intégré les FPGA dans son système de mitigation des attaques. En bref, les FPGA émergent au cœur des réseaux de données et deviennent donc disponibles à grande échelle : comment pouvons nous en tirer profit pour améliorer les performances des composants des réseaux ?

6.2 Analyse des forces et faiblesses des solutions matérielles et logicielles

Dans le cadre d'un projet de dernière année de cycle ingénieur à Télécom Bretagne dont je fus le co-superviseur avec Sandrine Vaton (département Informatique d'IMT Atlantique), j'ai découvert ce problème de la montée en débit pour les techniques usuelles d'analyse de trafic qui permettent aussi bien de garantir la qualité de service que la sécurité des réseaux et de leurs usagers. Parmi les étudiants impliqués dans ce projet, Tristan Groléat accepta de s'engager dans une thèse (2011-2014), que j'ai co-supervisée de nouveau avec Sandrine Vaton, financée par Pracom et le projet européen FP7 DEMONS (auquel je n'ai pas contribué). Cette thèse avait pour objectif de comprendre les besoins actuels et futurs de ces techniques d'analyse de trafic et de proposer des solutions matérielles et/ou logicielles capables de passer à l'échelle du Tbps.

6.2.1 Les besoins de l'analyse de trafic

Outre les débits croissants de données, nous avons identifié dans cette thèse plusieurs contraintes et critères de sélection et conception des plateformes d'analyse de trafic. En premier lieu, la résilience à une attaque ou un "stress" (comme un pic de charge non-malveillant, mais violent, dû à un comportement des usagers) est un problème majeur. Sur certaines plateformes (logicielles notamment), il est très difficile d'éviter de perdre des paquets dans des conditions de stress qui apportent une surcharge de traitement, rendant le traitement courant des paquets impossible par manque de ressources restantes. Ce problème est donc connexe à celui de la puissance de calcul permettant d'effectuer l'analyse de trafic en temps réel, avec la latence la plus faible. Si l'analyse *offline* ou *post-mortem* a souvent été utile pour comprendre des attaques ou des événements autres sur un réseau, elle ne permet pas de réagir immédiatement face à une anomalie pour éviter une coupure de service ou contrer des tentatives de piratage. Pour permettre cette capacité d'analyse au débit des données, certaines plateformes offrent un niveau de parallélisme élevé, d'autres offrent un nombre élevé

d'opérations par seconde ou la possibilité de réaliser des opérations plus complexes. Le meilleur choix dépend bien-sûr des spécificités de l'algorithme. Exige-t-il des calculs lourds et/ou une forte dépendance des données par exemple ? En outre, les plateformes d'analyse se doivent d'être flexibles pour faire face à un réseau en perpétuelle évolution, avec des trafics de données qui évoluent à l'échelle d'une journée, des services qui apparaissent en permanence, des protocoles qui évoluent, des réseaux qui s'interconnectent de plus en plus (filaire/non-filaire, public/privé/professionnel). Certaines plateformes sont plus adaptées aux applications flexibles, comme les plateformes logicielles, notamment de type GPP ou General Purpose computing on Graphics Processing Units (GPGPU), bien que la flexibilité puisse être obtenue sur toutes les plateformes avec différents degrés d'effort (plus marqué en matériel). Ensuite, la fiabilité et la sécurité sont deux exigences majeures des opérateurs de réseau, même si certaines applications peuvent être moins critiques que d'autres. Garantir conjointement ces deux critères est bien plus simple sur des plateformes simples, dont on maîtrise tous les composants et qui limite l'usage de plusieurs couches logicielles.

A ces contraintes d'action/exécution d'une plateforme, s'ajoutent des contraintes de conception. La souveraineté est une dimension qui est apparue à la suite de discussions avec des entreprises comme OVH et Orange Labs et des institutions comme DGA-MI. Il n'est plus concevable de conserver des composants dans un réseau avec des zones d'ombre, et encore moins lorsque ces composants ont la responsabilité de la sécurité et fiabilité de ce réseau. Cela exige des solutions logicielles comme matérielles totalement maîtrisables par l'exploitant. Une solution est donc de favoriser l'*open-source*. Elle n'est peut-être pas importante ou recommandée pour toutes les applications (notamment sur des aspects de confidentialité ou pour bloquer certains actes malveillants), mais l'existence d'une communauté ouverte pour une plateforme peut faciliter le développement et la maintenance. Le temps de développement est un aussi un critère essentiel puisqu'il détermine une part importante du coût d'une application et la réactivité de mise à jour pour faire face aux évolutions des systèmes et des techniques d'attaques. A cela s'ajoute, la simplicité de mise à jour. Les applications déployées sur un réseau doivent être maintenues, et certaines plateformes rendent cette tâche plus facile que d'autres. Enfin, le coût du matériel est un point important dans la considération du passage à l'échelle à tout un réseau et peut encourager à développer un système fortement distribué ou centralisé.

6.2.2 Les FPGA au sein des réseaux

Un état de l'art des solutions aussi bien académiques qu'industrielles a révélé les nombreux avantages de solutions à base de FPGA. En effet, les FPGA fournissent un parallé-

lisme massif et un développement de très faible niveau, au niveau du transfert de registre. Cela permet d'optimiser le transfert de paquets du lien de communication vers les unités de traitement au sein de la même puce et d'éviter différentes interfaces physiques et logicielles entre l'interface réseau et le cœur de calcul, réduisant latence et risque de congestion. En outre, les capacités de traitement de données en temps réel des FPGA sont bien supérieures à ce que qu'offre l'approche logicielle sur CPU et sont certes plus faibles que celles de l'approche ASIC mais à coût aussi bien plus faible. Par ailleurs, il nous a semblé que les capacités de reconfiguration des FPGA étaient encore sous-exploitées dans les réseaux et que nous pouvions proposer des solutions performantes **et** flexibles. Cependant, il faut bien reconnaître que le développement sur FPGA est difficile et long en comparaison d'alternatives logicielles. Comme les FPGA sont configurés au lieu d'être programmés, ils sont également moins flexibles que les plateformes logicielles. Nous verrons que ces deux faiblesses des FPGA, au regard de leurs forces et des opportunités de la reconfiguration, ont motivé deux thèses après celle de Tristan Groléat qui a intégré le groupe OVH où il a construit l'équipe FPGA qu'il dirige.

6.3 Classification de trafic sur FPGA

Durant la thèse de Tristan Groléat, nous avons considéré un cas d'application typique en analyse de traffic, l'analyse de paquet avec un algorithme puissant et complexe, bien connu en apprentissage machine : la machine à vecteurs de support, *Support Vector Machine* (SVM). Nous avons repensé l'algorithme et les fonctions connexes, notamment pour le stockage des flux, initialement très gourmand en mémoire, et donc inadapté aux faibles ressources d'un FPGA. **La version proposée de SVM a été montrée comme adaptée à la classification de trafic réseau sur FPGA, ne causant aucune perte en termes de précision de classification tout en traitant l'intégralité des paquets en transit, ce dont les alternatives logicielles sont incapables (elles sous-échantillonnent pour pouvoir simultanément transférer et analyser les paquets)** [GAV12 ; GVA14 ; GAV14]. Les performances mesurées à l'aide d'un générateur de trafic ont montré que le classificateur supporte 10 Gb/s sans problème sur une carte FPGA modeste. La limite se situe plutôt dans le nombre de flux que le classificateur peut traiter par seconde. Mais ce nombre est encore bien supérieur en utilisant un FPGA qu'en logiciel pur. Notre démonstration a mis en évidence que les FPGA permettent de traiter facilement des paquets à des débits élevés. Mais cette mise en œuvre a aussi montré un inconvénient majeur des FPGA : la modification des paramètres de la SVM peut prendre beaucoup de temps, notamment par la génération d'une configuration propre à chaque jeu de paramètres de SVM, ce qui est long et peut demander des

optimisations pour tenir les fréquences de fonctionnement requises sur chaque FPGA cible. Les contributions sur la mise en œuvre matérielle du classificateur de trafic ont conduit à des publications en revue [GVA14 ; GAV14] et plusieurs en conférence dont [GAV12] qui eut un impact dans la communauté de l'analyse de trafic et au-delà, notamment dans le monde du *machine learning* et de l'intelligence artificielle, pour l'originalité de notre approche matérielle d'un problème important de cette communauté. Cet impact m'a interpelé et encouragé à considérer plus généralement l'intérêt des FPGA pour l'intelligence artificielle comme nous le verrons plus tard. A ce titre, cet article est fourni en annexe de ce document (chapitre 9 / section 9.4).

6.4 La nécessité de l'union des forces du logiciel et du matériel

6.4.1 Vers plus de puissance et de flexibilité

Pour tester les solutions d'analyse de trafic sur FPGA, Tristan Groléat développa un générateur de trafic *open-source* [Gro+13] disponible en téléchargement (https://github.com/tristan-TB/hardware-traffic-generator). Un générateur de trafic doit pouvoir saturer les composants réseaux que l'on teste et pouvoir aisément faire varier la composition des flux émis. Nous nous trouvons dans un cas typique de besoin simultané de haut-débit et de flexibilité. La solution alors proposée correspondait à notre besoin, relativement limité.

Nous avons remarqué que nous pourrions augmenter grandement la flexibilité et l'évolutivité de notre solution en exploitant des techniques avancées associées aux FPGA comme la reconfiguration dynamique partielle et l'usage de réseaux sur puce, *Network on Chip* (NoC). Ces techniques associées à des pré-traitements logiciels semblaient prometteuses. En outre, nous nous sommes mis à envisager d'améliorer la programmabilité des FPGA de différentes façons. Nous pouvions développer une surcouche de framework sur FPGA ou une surcouche de type *coarse-grain array* (CGA) et utiliser des outils de conception de haut-niveau. Nous avons estimé qu'une sonde d'analyse de trafic pourrait même être une hybride matérielle-logicielle pour exploiter les forces des deux mondes : la capacité d'analyse avancée de tous les paquets, sans exception, permise par les FPGA couplée à la flexibilité du traitement logiciel qui ne travaillerait que sur des métriques extraites par le FPGA, mais à la demande précise du logiciel.

Ces pistes d'union du logiciel et du matériel ont ensuite été explorées à travers deux thèses connexes menées simultanément, financées par la région Bretagne, l'Institut Mines-Telecom et les fonds propres aux département Électronique. Du matériel additionnel à ce dont disposaient les départements Informatique et Électronique était requis pour mener ces deux

thèses. Il fut financé conjointement par les fonds de recherche du Conseil Général du Finistère et de Brest Métropole Océane ainsi que complété par les investissements du Contrat Plan Etat-Région 2014-20. Nous avons cherché à dépasser l'obsolescence des Virtex 7 disponibles sur les NetFPGA SUME utilisées par la communauté en privilégiant des cartes exploitées par les datacentres et à base de Virtex Ultrascale+ (comme les cartes XUP-P3R ou XUP-VV4 de Bittware), similaires à ce qu'exploite OVH à Brest. Nous avons ainsi collaboré sur cette thématique avec OVH, notamment par un contrat d'expertise pour évaluer ce que pouvaient apporter les dernières solutions reconfigurables à la sécurité des réseaux.

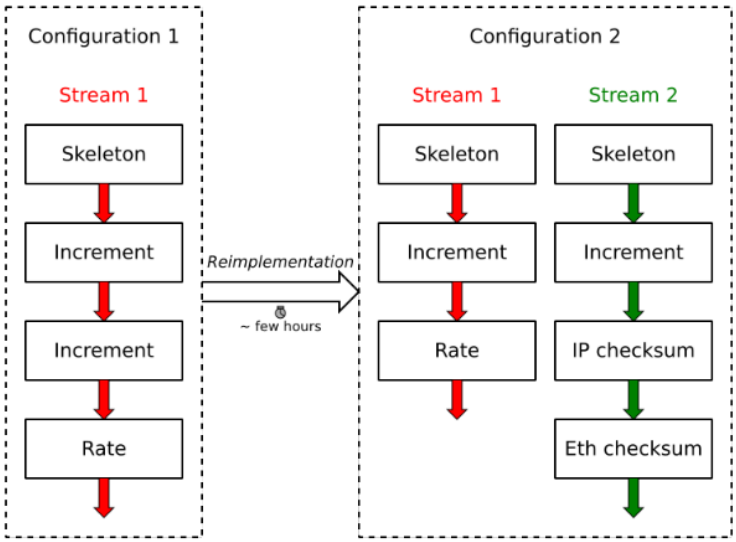
6.4.2 Architectures hybrides matérielles/logicielles

6.4.2.1 Reconfiguration et relocalisation pour plus de flexibilité

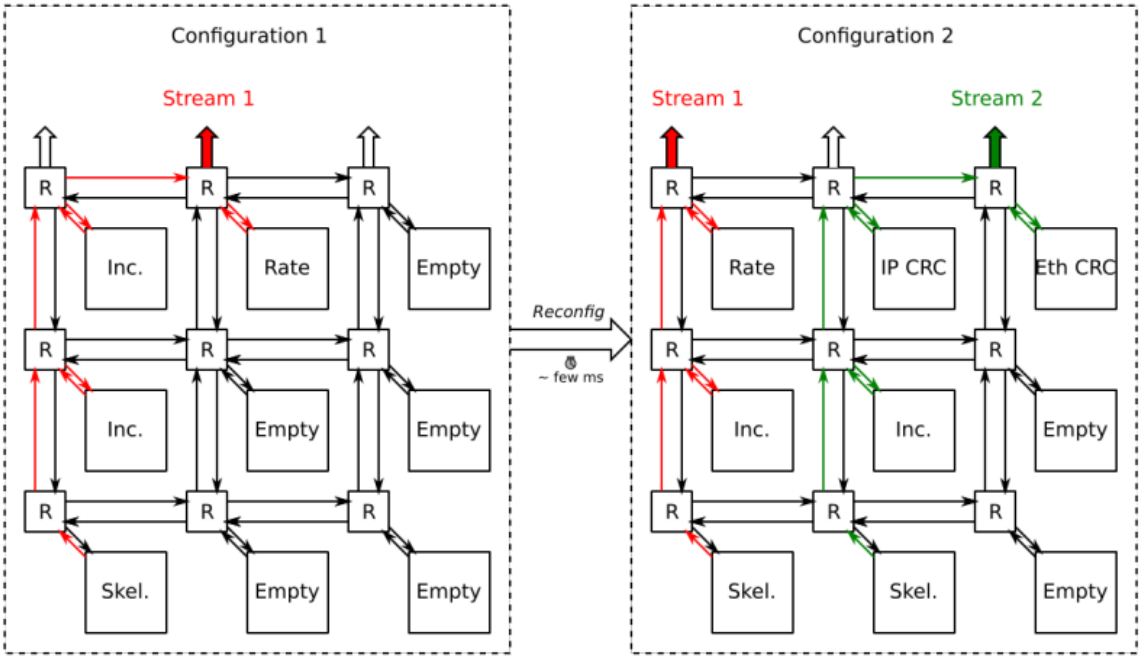
L'objectif était de proposer des solutions d'architectures conjointes logicielles-matérielles pour rendre possible des traitements jusqu'alors inenvisageables en temps-réel à plusieurs dizaines, voire centaines de gigabits par seconde tout en restant flexibles. Les applications visées concernaient tant la sécurité des réseaux que la cryptanalyse. Pierre-Henri Horrein a rejoint le département en tant qu'enseignant-chercheur dans cette phase de la recherche. Grâce à son expertise en systèmes hétérogènes et flexibles, il été un moteur essentiel pour cette activité de recherche.

Parmi les solutions disponibles, nous avons voulu tirer profit des solutions de reconfiguration disponibles pour les FPGA, qui nous semblaient totalement ignorées par les concepteurs de solutions de test et d'analyse de trafic. Nous avons décidé d'investiguer un composant que Tristan Groléat avait déjà identifié comme essentiel, notamment pour le test de sécurité et fiabilité des composants et systèmes en réseau : la génération de flux réseaux [Gro+13].

André Lalevée a alors rejoint l'équipe en tant que doctorant. Notre objectif était de développer et d'appliquer une architecture combinant la relocalisation de *bitstream* de configuration [Lal+16] et un réseau sur puce pour améliorer à la fois la flexibilité et le débit de la solution initiale de Tristan Groléat pour la génération de trafic (Fig.6.4.1b). Ainsi, si un nouveau trafic doit être généré, il suffit alors de reconfigurer partiellement le générateur en fonction du nouveau modèle au lieu d'exiger une réimplémentation complète, ce qui réduit le temps entre chaque test de quelques heures à quelques millisecondes. En outre, aucune réimplémentation n'est alors nécessaire si la connectivité entre les unités du générateur doit être changée, car cela est assuré par une architecture NoC. Ce concept de NoC entre unités relogeables peut aussi être appliqué à une sonde d'analyse de trafic, la rendant flexible et néanmoins performante en terme de débit de traitement. Après un séjour d´échange dans



(a) Modification de la configuration du générateur de flux pipeline proposé par Tristan Groléat.



(b) Modification de la configuration du générateur de flux proposé lors de la thèse d'André Lalevéee en exploitant un NoC pour interconnecter des régions relogeables.

Figure 6.4.1 – Augmenter la flexibilité d'un composant réseau par la relocalisation de *bitstream* et l'usage d'un NoC.

l'équipe de Michael Huebner, de la Ruhr Universität Bochum, spécialiste des technologies reconfigurables, André Lalevée a donc développé un outil d'automatisation de relocalisation de *bitstream* de configuration [Lal+16], *AutoReloc*, totalement scripté et validé sur plateforme

Xilinx. Cet outil n'est pas spécifique au cas d'application et vaut pour tout circuit constitué d'unités relogeables. Il a appliqué cet outil à une architecure exploitant un NoC de type *mesh-2D* qu'il a conçu pour du transfert de paquets en respectant plusieurs contraintes majeures. La flexibilité de l'architecture complète doit être permise par relocalisation de *bitstream* : il a donc fallu fallu adapter le protocole du NoC à cette spécificité ainsi que ses interfaces d'interconnection aux unités relogeables. Ensuite, la contrainte de débit (de plusieurs dizaines de Gbps pour saturer des composants réseaux) a imposé des choix forts dans le dimensionnement du NoC et la simplification de son protocole. Nous avons ainsi démontré qu'une solution FPGA exploitant l'état de l'art de la technologie fournit une flexibilité à l'échelle attendue. En quelques millisecondes, un circuit peut être adpaté au besoin, qu'il soit dicté par une opérateur humain ou bien par une supervision machine, typiquement logicielle. On commence alors à envisager une **intrication entre logiciel et matériel pour un système agile et performant à des débits de plusieurs dizaines à centaines de Gbit/s**.

6.4.2.2 Intriquer logiciel et matériel

En parallèle de la thèse d'André, nous avons souhaité pousser le concept de sonde d'analyse conjointe entre logiciel et matériel au plus loin de la flexibilité et de la performance. Franck Cornevaux-Juignet fut le doctorant qui s'y attela.

Une carte SmartNIC, qui peut être rendue configurable par l'emploi d'un FPGA comme au sein des datacentres Microsoft (Azure), Amazon ou OVH, analyse les paquets à la volée pour remonter des métriques, et éventuellement des paquets sélectionnés, selon les règles imposées par son *firmware* et sa configuration, à une entité maître, comme une application d'analyse exécutée sur un serveur. Cette application peut renvoyer elle aussi des métriques issues de traitements sur les données remontées à un service de supervision et déclencher potentiellement une mise à jour de la configuration de la Smart NIC, voire du firmware.

La figure 6.4.2 montre la boucle de rétroaction matériel-logiciel lorsqu'une modification est nécessaire pour une carte réseau de type SmartNIC, c'est-à-dire reconfigurable. Les processus de surveillance du trafic reposent largement sur la génération de nouvelles configurations pour gérer la volatilité du trafic, et non pas les paramètres d'une configuration. En plus du temps de génération, le processus de reconfiguration rend la puce FPGA incapable de traiter les paquets entrant, de sorte que la sonde est aveugle pendant ce temps. Ce problème est facilement atténué par la redondance des systèmes critiques en réseau, la reconfiguration d'un système à la fois évitant toute interruption de traitement. Cependant, cette procédure est lourde à mettre en place et nécessite de nombreuses opérations. Par conséquent, il est souvent réservé à la maintenance planifiée ou à des urgences pour ré-

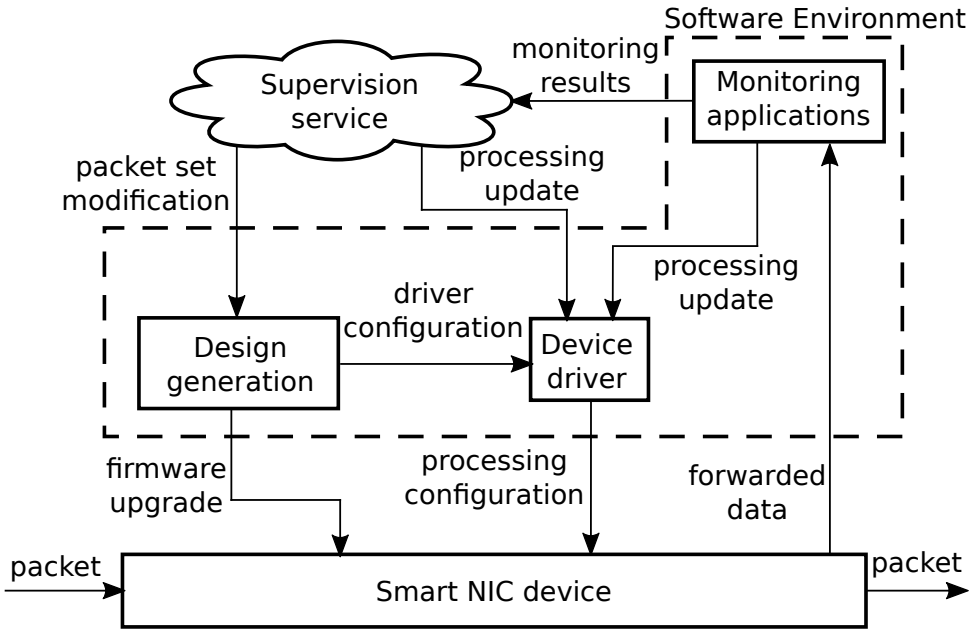


Figure 6.4.2 – *Framework* conventionnel de l'analyse de trafic réseau en smartNIC : .

soudre des failles de sécurité. Ainsi, un flux de conception FPGA standard limite la réactivité des solutions SmartNIC classiques aux rares mises à jour de firmware. Bien qu'il soit suffisant pour la flexibilité du réseau actuel, il ne suffit pas pour l'adaptation en direct au trafic entrant. Il est par exemple impossible de réadapter finement l'analyse par la SmartNIC en fonction de ce que la supervision observe et déduit des retours de cette SmartNIC.

Afin de résoudre ce problème, le travail de thèse de Franck Cornevaux-Juignet se concentra sur des architectures basées sur un paradigme différent. Il est possible de gagner en flexibilité des SmartNIC grâce à la conception et à la réutilisation d'éléments de traitement génériques, spécialisables pour différents types de traitements de paquets sans reconfiguration du FPGA mais par application de paramètres. Ainsi, avec des paramètres de configuration fournis dynamiquement par le logiciel, l'architecture générique n'est configurée qu'une seule fois sur le FPGA, puis le comportement peut être modifié en fonction des besoins. Les paramètres sont utilisés pour l'adaptation de l'exécution, tandis que la reconfiguration est conservée comme moyen de fournir des mises à jour moins fréquentes. Si une architecture matérielle statique n'est pas idéale pour la flexibilité, l'utilisation d'un traitement basé sur les paramètres de mise à jour réduit cette limitation. Une API logicielle est un moyen efficace d'intégrer la configuration dans un flux logiciel classique pour les utilisateurs finaux. De plus, le traitement de tous les paramètres dans le logiciel permet de suivre la configuration pour une application de surveillance rapide et consciente du contenu. Combiner ainsi le matériel et le logiciel est la clé d'un développement aisé pour le développeur logiciel, garant d'un haut

débit et d'applications de surveillance adaptatives.

Pour valider et tester ces idées, une sonde conjointe entre matériel et logiciel fut proposée et développée selon l'architecture illustrée en figure 6.4.3. Nous avons intégré d'emblée dans notre solution la possibilité de gérer plusieurs sondes matérielles distribuées dans le réseaux qui collaborent avec une même entité logicielle, qui peut donc profiter d'une réelle synergie avec de multiples accélérateurs de traitement.

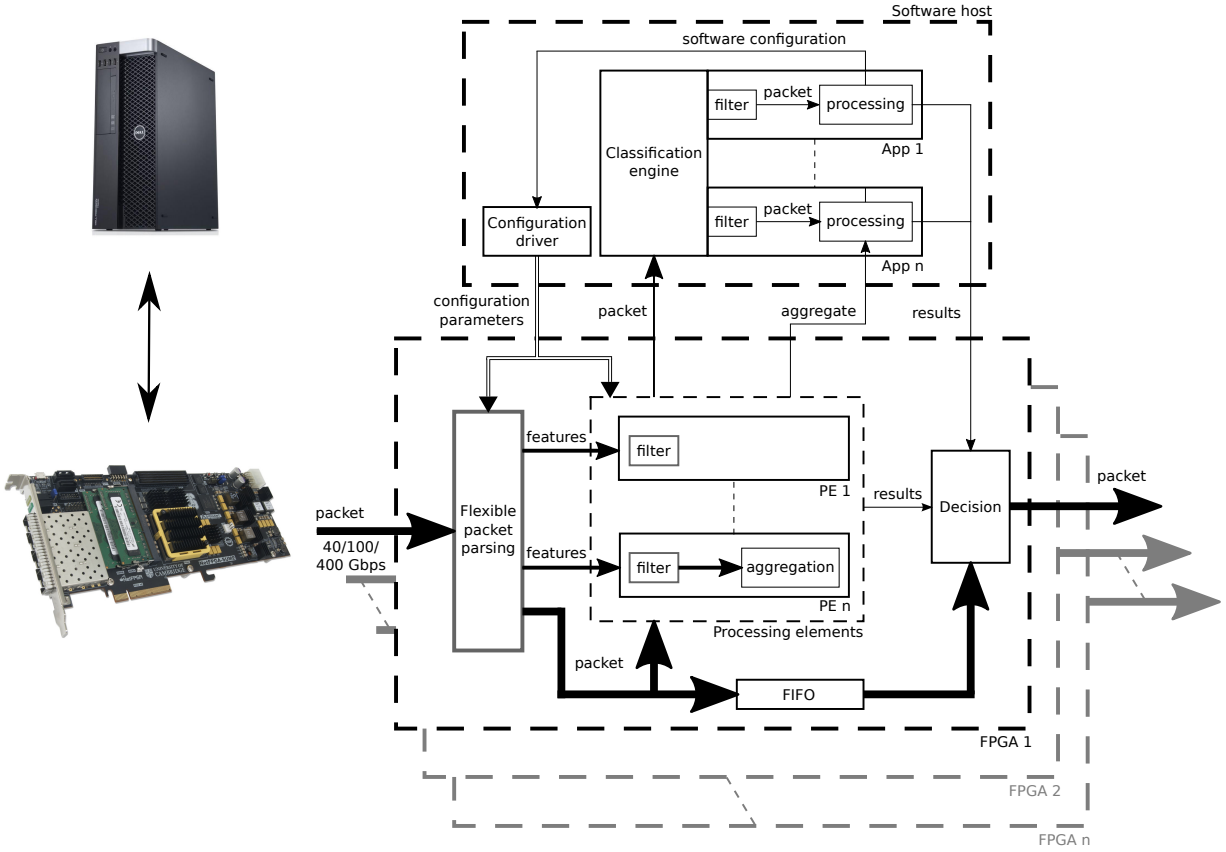


Figure 6.4.3 – Sonde conjointe entre matériel et logiciel proposée, développée et testée lors de la thèse de Franck Cornevaux-Juignet.

Capable d'être intégralement re-paramétrée à la volée, notre solution conserve une capacité d'analyse permanente et totale des paquets en transit (en differents points du réseau), grâce notamment à un analyseur de paquets performant et conçu pour être adaptable, au rythme des paquets, à une analyse ou une autre, en fonction des besoins exprimés par la supervision. Deux applications furent expérimentées sur cette plateforme d'analyse. La première utilise uniquement l'analyseur de paquets proposé [Cor+17b] pour créer une application de surveillance simple. Grâce à la flexibilité de l'analyseur de paquets, il est possible d'affiner les fonctions de surveillance afin d'avoir une vue précise de la composition du tra-

fic lorsqu'une anomalie est détectée. **La sonde est alors capable de détecter différentes attaques volumétriques de différents ordres sur différents types de paquets sur une liaison saturée à 40 Gbps. Nous avons testé avec succès la capacité de notre sonde à supporter et détecter une attaque massive tout en analysant paquet par paquet les flux, évitant que des actions malveillantes passent inaperçues. Cette capacité ne peut être apportée par aucun autre analyseur de paquets de la littérature qui laisse passer les flux les plus petits que des attaques massives peuvent masquer en forçant le système à sous-échantilloner les paquets pour tenir la charge.**

La deuxième application combine l'analyseur de paquets, un réseau d'interconnexion et des processeurs de règles pour créer une solution avancée de classification des paquets hybride entre matériel et logiciel. Ce classificateur unique permet de mettre à jour l'ensemble des paramètres et l'ensemble des règles au moment de l'exécution sans perte de paquets. L'utilisation orientée logiciel grâce à l'API permet de définir toute règle comme n'importe quel classificateur de logiciel conventionnel le fait. Alors que le test est effectué sur une saturation de liens à 40 Gbps, **la classification hybride est capable de traiter 120 Gbps de paquets de 64 octets, soit le stress maximal apporté par de petits paquets**. Le filtrage est raffiné progressivement en matériel sous le contrôle du logiciel. Cela permet de contourner le problème de la mémoire réduite disponible sur FPGA : elle ne permet de stocker qu'un nombre limité de règles, alors que ces règles sont très nombreuses dans une application de surveillance de trafic !

En conclusion, l'approche originale combinant logiciel et matériel a permis de fournir des solutions de surveillance de trafic capables de tenir une **analyse exhaustive des paquets au rythme des liens, avec une flexibilité totale dans l'analyse**, en controlant la charge imposée au CPU exécutant la partie logicielle (garantissant ainsi sa **robustesse aux attaques**), et en offrant une **réactivité permettant une mesure toujours adaptée aux observations**. Ainsi, notre sonde innovante est capable de détecter toute attaque volumétrique comme toute tentative malveillante insidieuse, et d'assurer une analyse de trafic en temps réel, bien utile pour le *Software-Designed Network* qui cherche une adaptation la plus juste au trafic. **Ces travaux furent récompensés par le prix du meilleur poster à la conférence CNS 2017 à Las Vegas** [Cor+17a]. Cette solution pourrait encore évoluer et intégrer les travaux d'André Lalevée sur l'association d'un NoC à des unités relogeables, ou bien des solutions comme des architectures reconfigurables gros-grain comme celles proposées par [KSD17], où logiciel et matériel collaborent au sein d'un FPGA pour assurer flexibilité, performance et facilité de conception.

6.5 Perspectives de l'accélération sur FPGA des traitements pour le réseau

Nous avons montré au cours de ces trois thèses et d'un contrat de recherche avec OVH que les FPGA offrent une solution inégalable pour combiner flexibilité et traitement des paquets à plusieurs dizaines et centaines de Gbps. Les FPGA permettent de résoudre le problème de l'analyse du trafic face à la montée en débit des réseaux tout en conservant une flexibilité similaire aux plateformes actuelles. Ce constat a d'ailleurs conduit de nombreux acteurs majeurs du réseau, comme Microsoft, Amazon ou OVH, à utiliser les FPGA dans leurs infrastructures. L'émergence des SmartNIC a changé le statut du FPGA, qui est passé de plate-forme de prototypage à support majeur de l'accélération des traitements pour le réseau, stimulant ainsi le marché et la recherche.

La technologie FPGA étant en pleine croissance, avec des acteurs historiques actifs comme Intel ou Xilinx et de nouveaux arrivants percutants comme Achronix (dont les produits sont déjà disponibles sur des cartes Bittware), elle développe sans cesse les opportunités pour les composants réseaux. Par exemple, nous avons montré qu'avec la reconfiguration partielle et la relocalisation de *bitstream*, il serait facile de combiner la mise à jour des architectures établies et l'intégration de nouvelles applications accélérées, répondant à de nouveaux besoins, et ceci sans exiger un lourd travail de conception. Nous avons considéré les cas d'application de la génération de trafic pour le test de composant et celui d'une sonde flexible pour l'analyse de trafic, mais nos travaux peuvent être étendus à d'autres composants. Par exemple, assurer les services vidéos interactifs supportés par les réseaux 5G requiert des routeurs aux latences de traitement les plus faibles possibles. Les réseaux doivent aussi suivre des besoins nouveaux liés à de nouveaux schémas dans les systèmes de télécommunication. Le développement du cloud-RAN comme décrit au chapitre précédent crée de nouveaux besoins avec des **composants réseaux embarquant des capacités de compresssion/décompression adaptative, de récupération d'information après perte de paquet, le tout à des débits qui visent le Tbps avec des latences de traitement au-delà des exigences actuelles**. **Cet environnement est donc très riche tant en cas d'applications qu'en opportunités d'innovation où l'association des compétences en architecture numérique, FPGA, traitement du signal et de l'information et réseaux est un avantage majeur pour se démarquer.**

Les acteurs de cette industrie du réseau misent sur les FPGA mais avec le problème des ressources humaines limitées dans ce domaine d'expertise. Sur 100 ingénieurs en génie logiciel, combien sont capables de produire du VHDL ou du Verilog et proposer des architectures performantes sur FPGA ? Trop peu pour les besoins actuels et futurs. Au-delà de

ce que nous avons proposé comme hybridation logiciel-matériel qui permet d'augmenter les contributeurs avec un nombre réduit d'experts FPGA, il peut être intéressant d'investiguer sur l'apport des solutions de synthèse de haut niveau, High Level Synthesis (HLS), ou de langages plus évolués que les actuels Hardware Description Language (HDL) comme VHSIC Hardware Description Language (VHDL), Verilog ou SystemVerilog. Ainsi, je participe au comité de suivi individuel de Jean Bruant, doctorant CIFRE à OVH, sous la supervision de Pierre-Henri Horrein (OVH), Tristan Groléat (OVH) et l'encadrement d'Olivier Muller (TIMA/SLS) et Frédéric Pétrot (TIMA/SLS). Cette thèse porte sur l'"abstraction du flot de développement FPGA pour l'intégrer dans un flot de développement logiciel moderne" et ouvre des pistes prometteuses pour améliorer la productivité, la maintenance et l'évolutivité des composants réseaux à base de FPGA.

Dans les perspectives d'intérêt, il me semble que les nouveaux modèles de FPGA proposées par Achronix, tels les *Speedster7t* en technologie TSMC 7nm FinFET, méritent une attention particulière. Outre l'avance en terme de technologie (à l'heure de la rédaction, Xilinx n'emploie au mieux "que" la techonologie 16Fin FET+), ils tirent profit d'une architecture de reconfiguration différente avec un *Network-on-Chip* (NoC) 2D pour l'interconnection de blocs logiques reconfigurables, d'interfaces GDDR6, de *transceivers* à des débits jusqu'à 112Gbps, des *Hard IP* (modules câblés dédiés et optimisés) MACs 400G Ethernet, PCIe Gen5 et, originalité, jusqu'à 1 760 unités de type *machine learning processors* (MLP) bien utiles pour les opérations mathématiques requérant des formats de représentation à précision variable. Pour de nombreux traitements complexes tels que SVM, ces unités semblent une belle opportunité pour augmenter la précision et le débit de traitement. En outre, l'architecture NoC avec des accès privilégiés (jusqu'à 4Tbps cummulés) à des mémoires GDDR6 permet de lever en partie le verrou du stockage et de l'accès à de nombreux compteurs, comme traité par Tristan Groléat.

Le verrou classique de la congestion de routage a été un frein majeur à nos développements sur cible Xilinx. Les travaux d'André Lalevée et Franck Cornevaux-Juignet ont requis de nombreuses semaines pour simplement trouver des solutions à la congestion, notamment du fait que nos bus étaient relativement larges (typiquement 128 bits appairés pour être affectés au mieux aux bus AXI de 256 bits pour le NoC d'André et 512 bits pour l'analyseur de paquets proposé par Franck). Or, le NoC des *Speedster7t* repose sur des lignes et colonnes, chacune sous la forme de 2 bus AXI 256 bits unidirectionels capables d'un débit de 512 Gbps dans les 2 directions, qui gèrent intégralement le routage. Aucune unité de logique reconfigurable n'est requise pour la gestion du transfert de paquets. Aucun transfert de type *Stacked Silicon Interconnect* (SSI) n'est à gérer. Or ces derniers sont problématiques. Incontournables dans les derniers modèles *Ultrascale+* de Xilinx, ils permettent l'interconnexion

entre les différentes *dies*, associées à des *Super Logic Regions* (SLR) qui constituent un FPGA. Malheureusement, lorsqu'un signal passe d'une SLR à une autre, il quitte une *die* SLR par des vias pour emprunter une *Super Long Line* (SLL) dans une *die* d'interconnexion puis finalement va remonter par des vias dans une autre *die* SLR. Même si cette technologie SSI est le résultat d'un travail d'une très grande qualité, elle limite l'usage de FPGA massifs par les problèmes de routage et de congestion associés. Le NoC des *Speedster7t* semble dénué de tout problème de ce type. Evidemment, les données constructeurs sont à prendre avec précaution et doivent être vérifiées par des applications à des cas réels. Néanmoins, offrir un NoC pour gérer les nombreux transferts de paquets internes **sans exploiter la logique reconfigurable** pour router les signaux est un réel avantage pour tenir les débits sans des semaines à chercher une solution au problème de congestion de chaque cas traité. En outre, il serait intéressant d'analyser en quoi la solution *Speedster7t* contraint les architectures pour tirer profit au mieux de ce NoC.

7

Architectures faible consommation et/ou haute-capacité de traitement pour les systèmes autonomes « intelligents »

Sommaire de ce chapitre

7.1 Problématique investiguée

Les précédentes sections ont traité de mes activités de recherche dans le domaine des télécommunications, aussi bien en communications numériques qu'en analyse de trafic des réseaux. J'y ai développé des concepts et un savoir-faire qui cependant peuvent être appliqués à des domaines connexes où les traitements sont aussi fortement contraints et où l'autonomie est fortement recherchée. Par autonomie, j'entends autant l'usage d'une source limitée d'énergie que l'indépendance vis à vis d'autres ressources à disposition dans les réseaux, comme des ressources de calcul ou des bases de données de grande ampleur.

En collaboration avec Cyril Lahuec, l'ingénierie biomédicale fut le premier domaine d'application des compétences acquises en traitement micro-électronique analogique et mixte, notament au travers des travaux d'intégration de traitements au sein de prothèses, au plus proche des capteurs et avec une totale autonomie énergétique (aucun transfert d'énergie venant de l'extérieur du système). Les traitements considérés dans ces travaux étaient extrêmement simples. Par la suite, dans des systèmes complexes, comme celui d'un réseau de capteurs corporels, j'ai considéré des traitements avancés comme la classification des signaux captés. Nous avons donc cherché à offrir un compromis d'efficacité de traitement et d'intégration pour minimiser la consommation d'énergie des capteurs, notamment par l'implantation en circuit de réseaux de neurones artificiels à cliques, inventés au département. Enfin, nous avons investigué le traitement au plus proche des capteurs pour de l'électroencéphalographie, afin d'en développer l'usage. Dans la continuité des réseaux de neurones artificiels intégrés avec le moins de ressources possibles, j'ai co-encadré une thèse sur l'accélération matérielle de l'apprentissage profond et ainsi amorcé une thématique de recherche que je développe à la fin de cette section.

7.2 Traitement profondément embarqué, voire enfoui

En collaboration avec le Laboratoire de traitement de l'information médicale (LaTIM, INSERM UMR 1101), Cyril Lahuec et moi-même avons investigué le défi de l'autonomie énergétique pour des prothèses instrumentées pour le genou. Une prothèse totale de genou est composée de deux parties métalliques, le plateau tibial et la partie fémorale, et d'un insert en polyéthylène (PE) imitant le ménisque, tels que montrés en Fig. 7.2.1.

Supportant de nombreuses contraintes mécaniques, la partie en polyéthylène est sujette à l'usure, ce qui entraîne une défaillance prématurée d'environ 10% des implants. Quand on considère qu'environ 100 000 prothèses de ce type sont posées chaque année en France (60

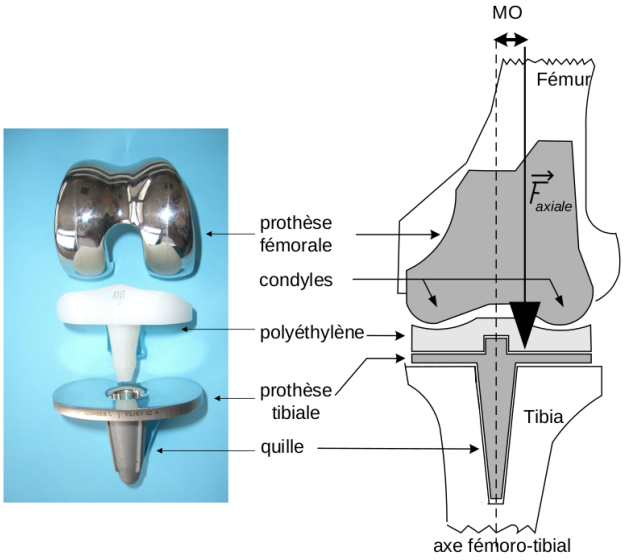


Figure 7.2.1 – Prothèse totale de genou et définition du décalage d'application de la force axiale, MO, simulant un déséquilibre ligamentaire.

000 en 2008, 80 000 en 2013), cela fait une quantité significative d'opérations chirurgicales coûteuses qui pourraient être évitées si l'on pouvait déceler le déséquilibre et le compenser pour réduire l'usure du PE. En outre, seuls des patients de plus de 65 ans sont généralement équipés, puisque la durée de vie des solutions actuelles est de 15 à 20 ans. Un meilleur suivi de l'état de la prothèse serait un atout pour prolonger cette durée de vie et donc permettre d'équiper des patients plus jeunes.

Connaître le taux d'usure du polyéthylène est donc crucial pour estimer la durée de vie de l'implant.

Comme l'a proposé le LaTIM, le centre de pression, Center Of Pressure (COP), peut être utilisé comme une mesure du déséquilibre des ligaments collatéraux [Alm+10] associable au décalage d'application de la force axiale et à l'usure du PE. La mesure du COP est donc fortement intéressante mais implique d'équiper la prothèse d'un système de télémétrie, si possible autoalimenté. En effet, transmettre l'énergie par induction nécessite un équipement externe contraignant pour un patient et surtout une bobine qui doit être placée dans la quille de la partie inférieure de la prothèse, comme proposé dès 1999 dans [TGW99] et comme développé dans [Gra+07]. L'ajout du système de télémétrie et d'alimentation par induction impose une quille tibiale plus importante, ce que les chirurgiens souhaitent éviter, comme nous l'avons appris auprès des collègues du LaTIM. Une quille plus importante requiert un volume d'os conséquent, alors que les patients concernés sont généralement agés, avec des os fragiles et une morphologie rarement compatible avec des quilles profondes.

L'équipe du LaTIM a eu l'idée alternative d'exploiter les capteurs de pression servant à

calculer le COP comme sources d'énergie pour tout le système électronique de mesure et transmission, grâce à des composants piézo-électriques [Gou+09 ; Alm+11] judicieusement implantés sur le plateau tibial comme montré en Fig. 7.2.2.

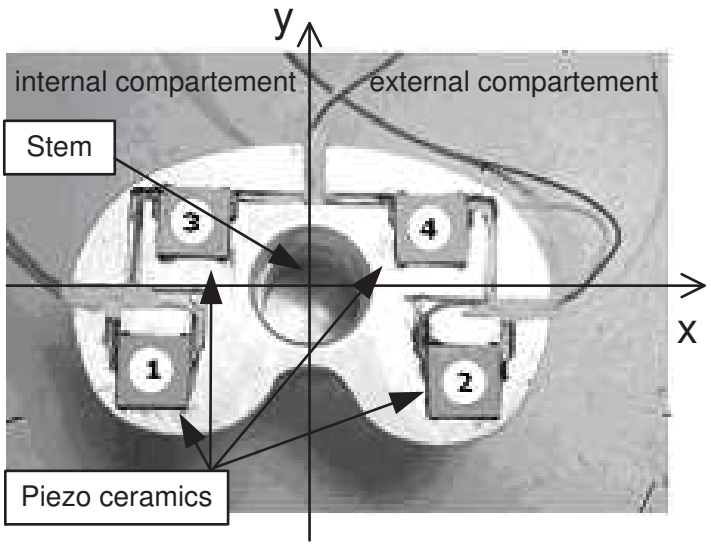


Figure 7.2.2 – Proposition du LaTIM d'implanter des composants piézo-électriques pour mesurer le centre de pression et alimenter le système de télémétrie.

Il fut montré, par des mesures menées par le LaTIM et Cyril Lahuec, qu'un cycle de marche produit 1.8mW grâce à ces composants piézo-électriques pour une charge de 35 kΩ. Ce faible budget était-il suffisant pour permettre une télémétrie du centre de pression ? Cette question a fait l'objet de travaux de recherche que Cyril Lahuec et moi avons menés en collaboration avec le LaTIM et deux étudiants, Colas Géranton puis Deepak Gupta. Nous avons alors participé à un groupe inter-GDR (STIC-Santé, ISIS, SoC2) sur les systèmes embarqués pour la santé.

Le centre de pression est obtenu par un calcul de barycentre des forces mesurées par chaque composant piézoélectrique [Alm+10]. Chacun de ces composants produit ainsi une tension proportionnelle à la pression qu'il subit, fournissant un signal qui peut ensuite être traité par un circuit électronique. Nous avons ainsi proposé une première architecture de mesure de ce COP à base d'unités numériques pour les calculs [Lah+09 ; Lah+11], en supposant une solution de récupération d'énergie adaptée et avec l'émission par une antenne dans la bande de fréquence 402-405 MHz (norme MICS [ETS02]) utilisant la prothèse elle-même.

Comme détaillé dans [Lah+09], en considérant une technologie CMOS AMS 0,35μm alimentée en 1,3V, et sur la base de l'architecture proposée, nous avons établi la consommation des différentes unités et conclu sur la puissance disponible pour l'émission radio des me-

sures. Ainsi, avec un cœur numérique de traitement, nous disposons au mieux de 1,6mW pour émettre le COP moyen calculé. Néanmoins, nous avons considéré que nous pouvions faire une usage bien plus pertinent de la faible énergie à disposition en réduisant la part du calcul numérique et en limitant ainsi le poids de la conversion analogique-numérique dans le bilan énergétique. Tirant profit de notre expérience acquise par la conception de décodeurs analogiques de canal, nous avons ensuite cherché à exploiter un cœur de calcul analogique. Ces travaux, détaillés dans [LA11], ont montré que l'on pouvait traiter les tensions analogiques directement issues des composants piézoélectriques, sans conversion analogique-numérique immédiate. Une conversion analogique-numérique du résultat final, et non des 4 tensions issues des composants piézoélectriques du premier schéma, permet de fournir une donnée numérique à l'étage d'émission. Cette solution analogique CMOS sous le seuil est bien moins complexe que la première traitant le signal en numérique. Nous avons montré que la surface de silicium dédiée au calcul est en effet divisée par 10 en passant du numérique à l'analogique, et la consommation d'énergie du système complet est divisée par 2 grâce à la suppression d'un oscillateur.

Ainsi, nous avons démontré qu'il est tout à fait possible de réaliser **un système de télémétrie pour prothèse instrumentée qui soit totalement autonome en énergie**. Bien-sûr, les traitements considérés étaient extrêmement simples et reversaient l'essentiel de la contrainte énergétique sur la transmission radio que nous avons identifiée comme bien plus gourmande en énergie (trois ordres de grandeur dans notre cas). Dans quelle mesure est-ce généralisable ? Quels sont les traitements d'intérêt pour un système embarqué qui peuvent concurrencer la transmission radio en terme de consommation des ressources énergétiques et qui justifient toujours un intérêt d'une implantation au plus près de la source pour limiter le transfert de données ? Autrement dit, quels traitements sont puissants en extraction d'information mais peuvent être parcimonieux en ressources matérielles et énergétiques ? Nous avons identifié différentes fonctions dans la suite de nos travaux qui répondent favorablement à ces questions.

7.3 Accélération matérielle pour un système autonome d'interface cerveau-machine à forte exigence calculatoire

Au sein du projet SABRE (*Seizing Advances in Bci from high Resolution EEG imaging in runtime*) financé par le LabEx Cominlabs et la région Bretagne, j'ai collaboré avec Cyril Lahuec, Francesco Andriulli (professeur au Politecnico di Torino) et l'équipe Hybrid d'Anatole Lécuyer à l'Inria afin de montrer tout le potentiel d'une analyse par électro-encéphalographie (EEG) raffinée pour une Interface Cerveau Machine (ICM). Une telle interface consiste en un

système de communication entre le cerveau et le monde extérieur qui permet la traduction d'une activité cérébrale en l'activation d'une machine, comme la commande d'une prothèse, l'écriture sur un écran, le contrôle d'un fauteuil roulant [VWD96 ; Wol+00]. Parmi les techniques non-invasives d'acquisition du signal cérébral, l'EEG est une solution de bien plus faible coût que les alternatives (magnéto-encéphalographie, imagerie à résonance magnétique, spectroscopie proche infra-rouge). Néanmoins, l'EEG requiert un traitement mathématique lourd pour obtenir une haute résolution exploitable pour l'application médicale attendue qui doit être effective en temps réel. Une ICM sera donc d'autant plus performante que la résolution et la complexité des modèles cérébraux seront élevées, ce qui s'oppose très vite à une application en temps réel [GF11 ; LLA09 ; CLL06]. En conclusion, **une solution d'ICM reposant sur une acquisition par EEG peut être efficace et bon marché, donc disponible à un grand public, si l'on résout le problème du traitement à haute résolution en temps réel**.

Pour résoudre ce problème d'imagerie EEG à haute résolution et temps réel mais faible coût, plusieurs voies existent.

D'une part, il faut chercher à définir les modèles les plus efficaces pour extraire l'information utile. Une des pistes les plus prometteuses étudiées au sein du laboratoire CERL (*Computational Electromagnetics Research Laboratory*) d'IMT Atlantique repose sur l'ICM exploitant l'analyse haute résolution des courants électriques dans le cortex sensori-moteur [NG12]. Le CERL travaille notamment à améliorer l'extraction d'information sur ces courants par l'intermédiaire des potentiels mesurés à la surface du crâne et qui résultent de ces courants, comme illustré en Fig. 7.3.1. C'est ce que l'on appelle la résolution du problème inverse de l'EEG. Cette résolution du problème inverse nécessite plusieurs solutions du problème direct, à savoir modéliser les potentiels mesurables en surface par des électrodes en connaissant les dipôles sources et la structure du cerveau [Gre+08]. En effet, chaque problème direct, donc toutes les combinaisons source électrique/électrode possibles, doit être préalablement résolu. Cela signifie que la moindre avancée dans les méthodes de résolution du problème direct a une grande importance pour les méthodes d'imagerie utilisant le problème inverse.

D'autre part, une fois que les modèles les plus adaptés sont sélectionnés, il faut leur fournir une intégration temps réel qui rende l'ICM opérante. Les techniques mises en avant par le CERL, même si elles améliorent l'état de l'art restent d'une complexité élevée puisqu'elles requièrent des milliers d'intégrations surfaciques, extrêmement gourmandes en termes de ressources de calcul, dépassant très vite les capacités de calcul d'un logiciel sur processeur de type GPP que l'on trouve dans les ordinateurs grand-public, même s'ils bénéficient d'une

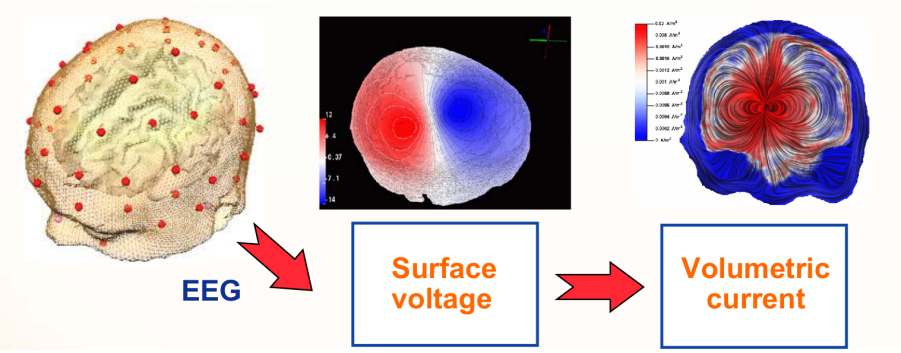


Figure 7.3.1 – Principe d'EEG volumique à partir de mesures en surface.

accélération partielle par GPU. En outre, il est difficile d'envisager une solution faible coût peu encombrante et temps réel associée à la puissance de calcul d'un parc d'ordinateurs. Et pourtant, l'accès à la haute résolution passe par une puissance de calcul indéniablement accrue.

A travers la thèse d'Erwan Libessart, nous avons étudié différentes alternatives d'intégration d'un accélérateur pour la résolution du problème inverse de l'EEG sur ASIC et choisi de considérer en premier lieu des architectures numériques performantes en terme de latence, débit, surface de silicium, consommation énergétique et fiabilité. La flexibilité n'a pas été requise.

Deux opérations non-linéaires grandement utilisées par les algorithmes proposés par le CERL, dont l'opérateur de Calderon [And+08 ; Guz+16], ont été rapidement identifiées comme problématiques : l'inverse et la racine carrée inverse. Cet opérateur de racine carrée inverse est par ailleurs essentiel en électromagnétisme car il est utilisé pour calculer la valeur d'un potentiel électrostatique, qui est inversement proportionnel à une distance. Nous avons investigué ce problème en considérant ce qu´offrait l'état de l'art. Ce dernier est riche de solutions pour l'inversion en virgule flottante mais bien plus pauvre pour la virgule fixe, ce qui n'est pas surprenant. Les développeurs de circuit en virgule fixe sont culturellement conditionnés pour éviter l'usage de la division, jugée trop complexe pour être acceptable un intégration de circuit. Comme nous l'avons détaillé dans deux articles [Lib+17b ; Lib+17a] qui traitent de l'inverse et de l'inverse de la racine carrée, cet *a priori* est abusif. Par un usage astucieux des ressources FPGA et une reprise de l'algorithme de Newton-Raphson [Lib+17b ; Lib+17a], nous avons proposé des architectures d'opérateurs inverse et inverse de racine carrée en virgule fixe offrant un rapport entre débit de traitement et complexité plus élevé que ce qui est proposé par l'état de l'art. De plus, la méthode proposée ne nécessite aucun bloc mémoire pour stocker les coefficients et peut être entièrement pipelinée, ce qui permet un calcul à

haute fréquence de fonctionnement de fort intérêt aussi bien en intégration FPGA qu'ASIC.

En technologie CMOS ST 65nm, nous avons cherché à maximiser la fréquence de fonctionnement ET la surface de silicium, pour finalement maximiser la métrique de débit de traitement par une unité de surface. Nous avons ainsi comparé deux solutions concurrentes à précision identique et montré que l'implantation de notre architecture originale offre un débit de 41Gop/s/mm² en CMOS 65nm [Lib+18]. Comparé à l'état de l'art précédent des opérateurs de multiplication sur 16 bits en technologie CMOS 65nm [SSW14], notre solution est 4 fois plus performante en termes de débit par unité de surface silicium.

Une plateforme FPGA d'accélération de la fonction d'intégration surfacique a été réalisée et testée [Lib18]. La solution proposée permet de réaliser en 1 minute ce que réalise un cœur de processeur Xeon en 1 heure avec une consommation du FPGA estimée à 4,369 W. A titre de comparaison, une fonction d'intégration similaire a été accélérée sur un GPU Nvidia Tesla K40 [AGD17] et a bénéficié ainsi d'un facteur d'accélération par rapport à un cœur de processeur de 80. Il faut noter qu'une Tesla K40 consomme 235 W en fonctionnement, soit 54 fois la consommation de notre prototype FPGA, avec une interface PCIe $\times$16 gen3, offrant donc un débit double à ce que permet la plateforme FPGA utilisée. Le ratio accélération sur consommation est donc largement à l'avantage de la solution FPGA qui pourrait en outre aisément dépasser le GPU en accélération brute si l'interface PCIe était équivalente.

Nous avons aussi porté notre architecture sur ASIC en technonologie CMOS 65nm de ST-Microelectronics, selon les recommandations précédemment énoncées et obtenu le circuit illustré en Fig. 7.3.2.

En disposant d'une interface au débit suffisant, le circuit proposé a une capacité d'accélération de 600 par rapport au cœur de processeur Xeon de même génération en occupant 16,2 mm². Autrement dit, **une heure de traitement sur un cœur Xeon passe à une minute sur FPGA Virtex7 690T en PCIe gen2 et 6 secondes sur la puce ASIC conçue avec une interface adaptée**. Clairement, une telle accélération permet de disposer d'une unité d'intégration surfacique qui autoriserait une modélisation à haute résolution de l'interaction des mesures de potentiel EEG en surface du cerveau et de l'activité de différentes zones au sein du cerveau. Le champ d'exploitation est vaste alors, notamment pour des interfaces cerveau-machine fiables et réactives.

En conclusion, nous avons montré par cette étude au sein du projet SABRE que l'accélération matérielle est la clé technologique d'une montée en résolution des modèles qui permettront alors un usage des interfaces cerveau-machine par le grand public, tant à des fins médicales, que professionnelles ou récréatives.

L'activité du cerveau n'a pas été qu'un objet de mesure pour mes activités de recherche

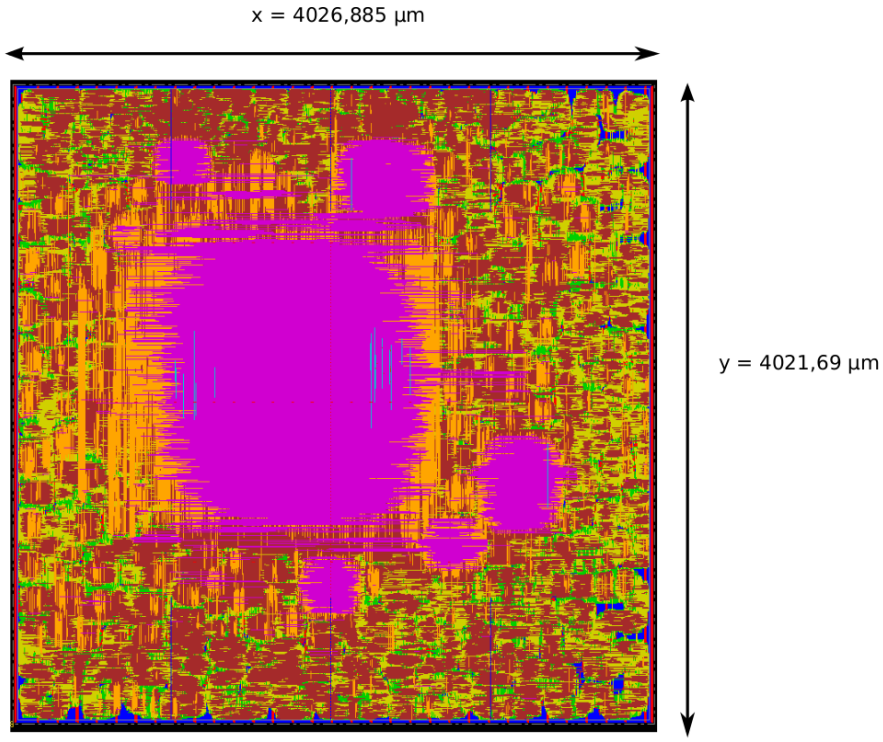


Figure 7.3.2 – Masques ASIC CMOS 65nm généré pour l'accélérateur de la fonction d'intégration.

mais aussi une source d'inspiration pour tout un ensemble de travaux du département Électronique de l'IMT Atlantique et plus largement du Lab-STICC, auxquels j'ai contribué. Ceci est explicité dans la prochaine section.

7.4 L'intelligence embarquée pour tous

7.4.1 Problématique considérée

Le cerveau humain est un réseau asynchrone hautement parallèle composé de centaines de milliards de nœuds, les neurones, capables de traiter une quantité impressionnante de données avec peu d'énergie (l'ordre de grandeur est de 20W). Ces caractéristiques ont suscité l'intérêt de la communauté scientifique pour la conception de circuits à bases de réseaux de neurones artificiels. Ceux-ci sont conçus pour imiter le fonctionnement du cerveau humain afin d'effectuer des tâches complexes tout en consommant aussi peu d'énergie que possible.

Dans de tels circuits, de multiples calculs peuvent être effectués en parallèle en utilisant un grand nombre de neurones entièrement connectés comme dans le cas de TrueNorth. Ces "processeurs neuronaux" sont efficaces sur des tâches complexes en vision artificielle par

exemple [Mer+14] mais sont surdimensionnés pour des tâches telles que les mémoires associatives intégrées. Ils sont alors avantageusement remplacés par des réseaux plus simples basés sur des cliques, soit des réseaux Willshaw-Palm [WBL69], soit des réseaux en cluster comme ceux conçus pour être **parcimonieux** par Claude Berrou et Vincent Gripon [GB11]. Ces derniers sont exploitables pour de la classification et comme mémoires associatives, pour lesquels nous avons cherché à apporter une intégration matérielle qui soit capable de passer à l'échelle lorsque la taille des réseaux augmente et qui soit fiable (ce qui est un enjeu majeur pour les solutions analogiques).

L'apprentissage profond étant devenu incontournable, nous avons en outre tenté d'y contribuer sur les aspects de parcimonie et d'adéquation réseau-architecture de circuit pour proposer des solutions performantes à faible coût calculatoire et énergétique.

7.4.2 Des réseaux de neurones parcimonieux pour une intégration matérielle à faible coût

Nous avons étudié les réseaux à cliques, à clusters ou non, comme ceux de Willshaw-Palm [WBL69 ; Pal13] et conçu **deux puces ASIC** [Lar+16 ; Lar+18a]. Deux doctorants ont alors collaboré, Benoît Larras et Paul Chollet, encadrés par moi-même, Cyril Lahuec et Fabrice Seguin. Nous avons proposé des architectures mixtes innovantes réduisant considérablement la consommation des circuits par rapport à l'état de l'art et nous les avons validées sur nos plateformes de test. Tout au long de notre étude, nous avons cherché à toujours associer l'information au signal le plus adéquat, numérique ou analogique, porté par une tension ou un courant. Nous avons montré que des tâches de classification spécifiques peuvent être effectuées par de tels réseaux de quelques milliers de neurones comme la gestion des modes de fonctionnement d'un processeur [LAR+14] ou de la détection d'anomalies cardiaques [Cho+17a].

Nous avons notamment cherché à réaliser un circuit qui consomme aussi peu d'énergie que possible, en instanciant les neurones à l'aide de circuits CMOS analogiques et mixtes [LAR+13]. Un cluster de 4 neurones est ainsi décrit en Fig. 7.4.1. Les synapses sont des miroirs de courant N-type MOS (NMOS) (transistor M_7 associé aux différents transistors M_{S1-1}) dont les courants sont ajoutés ou non au point A_1 en fonction des tensions $w_{i,j}$ représentant les informations $E_{i,j}$ en entrée du réseau. Le courant alors produit en A_1 représente le nombre de contributions actives. Un miroir de courant P-type MOS (PMOS) $M_4 - M_3$ transfère le courant au point B_1 d'une branche de l'unité gérant la règle d'activation *Winner-Takes-All*. Cette unité fut à l'origine proposée pour un tout autre objectif. Je l'ai découverte lors de la lecture de [GC11b] qui offre une solution pour la conception de décodeurs analogiques de codes

LDPC [Gu+09] et l'ai donc proposée ici. Sans mon passage par le décodage analogique, je n'aurais pas eu connaissance de cette technique simple et efficace, mais confidentielle.

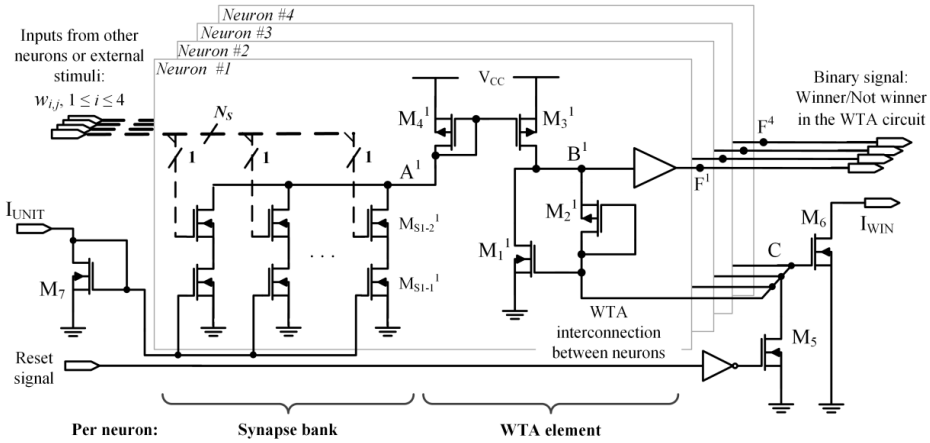


Figure 7.4.1 – Réalisation d'un cluster de 4 neurones à base de transistors CMOS.

Il faut noter que toutes les branches des neurones d'un cluster sont ensuite connectées au même point C. Cette unité permet de transférer tout le courant I_{WIN} au point $B_{j\in[1,4]}$ du seul neurone vainqueur. Ce courant est finalement converti en une tension grâce au buffer de sortie au point $F_{j\in[1,4]}$. La tension produite est un signal binaire donc numérique.

Ce dernier point est essentiel. **Le traitement est intégralement analogique mais les informations en entrée et sortie d'un cluster sont uniquement numériques, et sans implanter aucun convertisseur analogique-numérique coûteux ! Traiter l'information par un circuit analogique permet d'espérer une consommation plus faible (ce qui fut confirmé par nos travaux comme expliqué par la suite) et propager l'information en numérique simplifie son stockage et la fiabilise.** Cela permet en outre d'accéder à toute la flexibilité du numérique et à la densité de ses mémoires, ce qui est d'un grand intérêt car propager un signal analogique est compliqué : l'information portée par un signal analogique est corrompue par tous les éléments parasites et le bruit sur la puce sans que cela soit aisément compensable contrairement à l'usage d'un signal numérique.

De **nombreuses architectures mixtes** sont donc possibles, du tout parallèle intégralement câblé à une version itérative extrême où seule une unité de traitement analogique pour un cluster est intégrée et réutilisée grâce à un ordonnancement itératif et l'usage de contrôle et mémoires numériques.

Ces deux modèles ont été considérés et intégrés sur deux puces distinctes pour lesquelles ont été conçues une unité de compensation des variations de paramètres environnementaux (processus technologique, tensions d'alimentation et température) altérant la précision du

buffer de sortie de chaque neurone et donc des traitements analogiques [Lar+16].

La première puce que nous avons réalisée a permis de tester l'intégration de 3 réseaux à cliques (Fig. 7.4.2), avec une consommation de puissance par neurone (appelé aussi fanal) de l'ordre de grandeur du microwatt, avec une consommation d'énergie par événement synaptique extrêmement faible de 7fJ [Lar+17]. Nous avons ainsi démontré la robustesse du circuit, dont les performances ne changent pas dans différentes conditions environnementales de fonctionnement des transistors. Enfin, nous avons montré que **le circuit produit est dix fois plus efficace qu'un équivalent numérique en termes de surface de silicium occupée et de latence**.

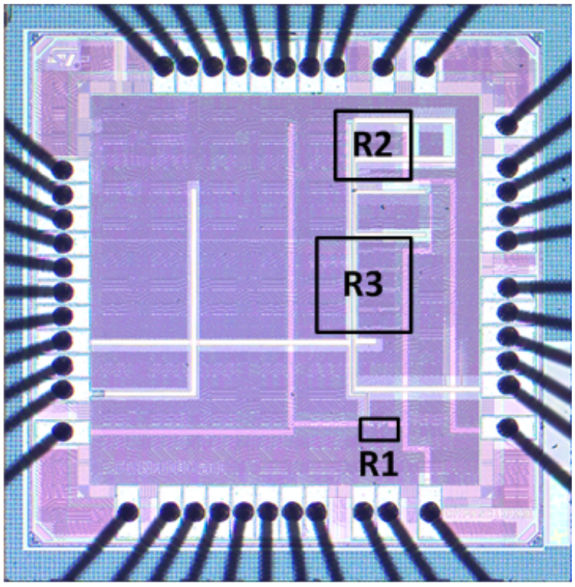


Figure 7.4.2 – Première puce mixte de 16 470 μm² intégrant 3 réseaux à cliques en technologie CMOS 65nm.

Ensuite, nous avons proposé une architecture de calcul pour **réseau de neurone flexible, configurable et itératif** capable de mettre en œuvre de multiples types de réseaux neuronaux sur clique et comprenant jusqu'à 3968 neurones. Le circuit, dont une photographie est présentée en Fig. 7.4.3, a été intégré dans un ASIC CMOS ST 65-nm. Le cœur du réseau réagit en 83ns à une stimulation et occupe une surface de silicium de 0,21mm² [Lar+18b].

Cette puce fut testée pour une application de classification d'électro-cardiogramme (ECG) dont nous détaillons le contexte et les résultats dans la section suivante.

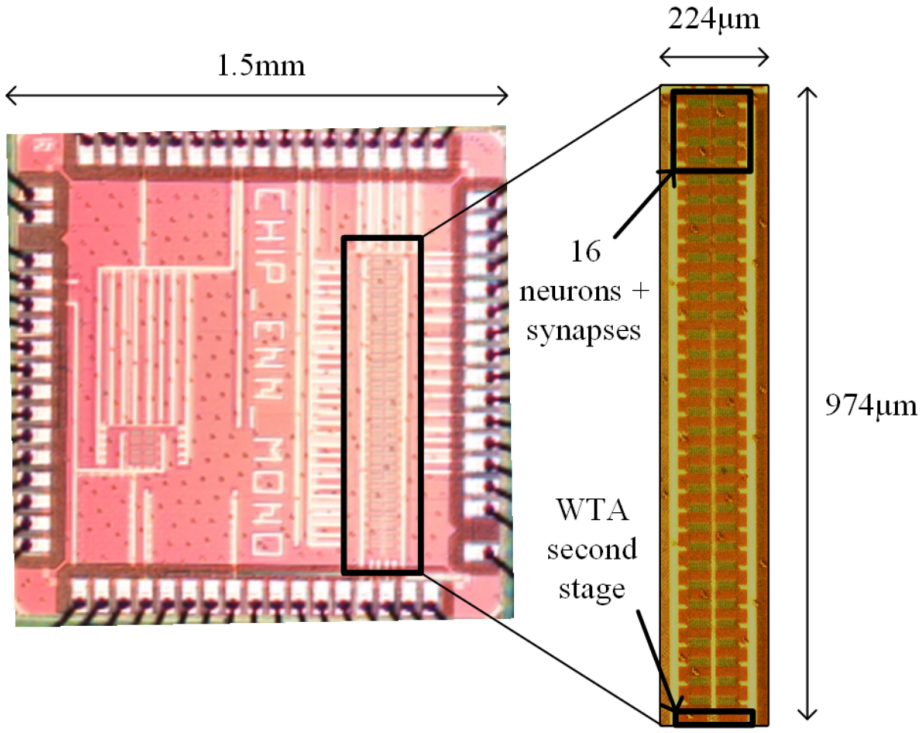


Figure 7.4.3 – Seconde puce mixte flexible pour traitement neuronal itératif en technologie CMOS 65nm.

7.4.3 Application en génie biomédical : l'intelligence au plus près du capteur pour une autonomie augmentée

Les réseaux de capteurs corporels présentent des opportunités fortes pour un système de soins médicaux de meilleure qualité et moins coûteux. Ils permettent en effet un suivi continu des maladies chroniques depuis le domicile grâce à des capteurs qui acquièrent des paramètres physiologiques en continu. Ces paramètres sont utilisés pour détecter les anomalies et les traiter dès qu'elles surviennent. Les capteurs sont soumis à de lourdes contraintes telles que des critères éthiques, d'acceptabilité mais aussi techniques comme la fiabilité, la robustesse, la taille et la consommation d'énergie. La contrainte énergétique est amplifiée par le fait que les capteurs peuvent être portés sous la peau. Changer de source d'énergie est donc un grand défi dans de nombreux cas. Les travaux que nous avons menés autour de la thèse de Paul Chollet ont permis une analyse des besoins d'un réseau de capteurs corporels et de solutions pour y répondre. Notamment, les différents besoins en énergie ont été évalués afin de choisir l'axe de recherche pour améliorer la durée de vie des sources d'énergie des capteurs. L'analyse montre que l'intégration du traitement dans le capteur semble être la meilleure solution. Cela permet une utilisation parcimonieuse des éléments de transmission qui consomment la plus grande partie de l'énergie. Nous avons choisi de traiter un

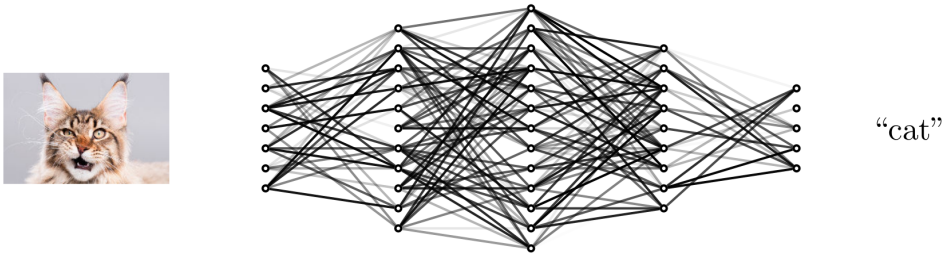
problème bien connu et offrant de grandes bases de données publiques, l'analyse d'électrocardiogrammes pour la détection d'arythmie. Un capteur a été proposé la permettant. Il comprend un cœur de traitement du signal basé sur un réseau neuronal sur cliques [Cho+17b ; Cho+17a]. La simulation du système permet une classification entre trois types d'arythmie avec une précision de 95 %. Le prototype, basé sur un circuit mixte analogique/numérique CMOS 65 nm ne nécessite que 1,4 μJ pour réaliser l'acquisition du signal, sa classification et sa transmission radio. C'est **105 fois moins qu'un capteur transmettant les données directement acquises sans classification**. Pour réduire encore la consommation d'énergie, une nouvelle méthode de détection fut proposée. L'idée est de concevoir un convertisseur analogique vers caractéristique pour simplifier l'architecture du capteur. Une architecture de convertisseur analogique vers caractéristique fut élaborée pour l'acquisition des battements du cœur et leur classification. Des simulations ont montré un besoin en énergie de 1,18 nJ pour l'acquisition des paramètres tout en offrant une précision de classification de 98 %. **En ajoutant cette nouvelle forme d'acquisition au système déjà proposé, nous avons atteint une diminution d'énergie dans un rapport de 2500 par rapport au système qui émetterait les données brutes vers une plate-forme de traitement**. Ce travail a ouvert la voie au développement d'un **capteur à faible énergie pouvant ainsi s'accomoder d'une durée de vie limitée d'une pile et donc être plus qu'embarqué : enfoui.**

7.5 Accélération de l'apprentissage profond

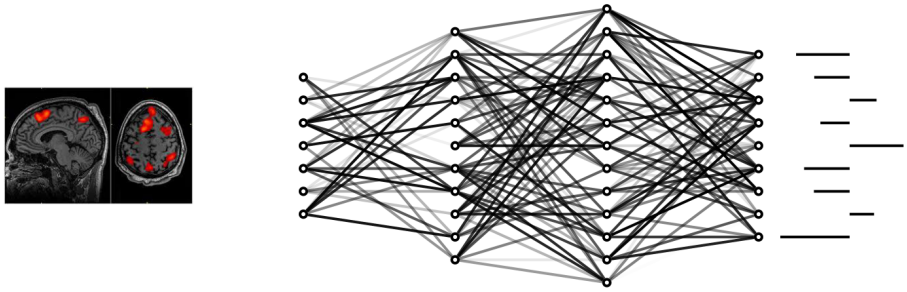
Par la suite, en collaboration avec Vincent Gripon et Nicolas Farrugia, j'ai abordé le problème de la mise en œuvre de solutions d'apprentissage profond dans le contexte de dispositifs aux ressources limitées, comme les systèmes embarqués que sont les smartphones ou bien les systèmes informatiques "grand public", c'est-à-dire ne disposant pas de parcs de serveurs similaires à ceux de Google. Nous avons notamment recruté Ghouthi Boukli Hacène pour une thèse sur le sujet auquel il a brillamment contribué.

Nous avons investigé plusieurs voies pour réduire à la fois la mémoire requise pour l'apprentissage profond et sa complexité calculatoire sans compromettre sa précision. Pour ce faire, nous avons notamment considéré les techniques de l'état de l'art comme l'élagage, la quantification ou la factorisation. Nous avons également introduit de nouvelles méthodes pour traiter le cas de l'apprentissage incrémental, c'est-à-dire capable d'apprendre de nouvelles informations au fil du temps sur un modèle existant sans avoir à conserver l'intégralité de la base de données antérieure et sans détruire les connaissances précédemment acquises [Gri+17 ; Bou+17a ; Bou+17b]. L'apprentissage par transfert illustré en Fig.7.5.1 est

1. Entraînement de l'apprentissage d'un réseau profond en utilisant des datasets **génériques** sur de grandes bases de données.



2. Après suppression du dernier étage du réseau précédent, ce qui offre une représentation intermédiaire de l'image en entrée (donnant des informations génériques sur l'image, comme la présence de certains motifs par exemple), calcul de vecteurs caractéristiques.



3. Classification par une technique classique appliquée sur les vecteurs caractéristiques.

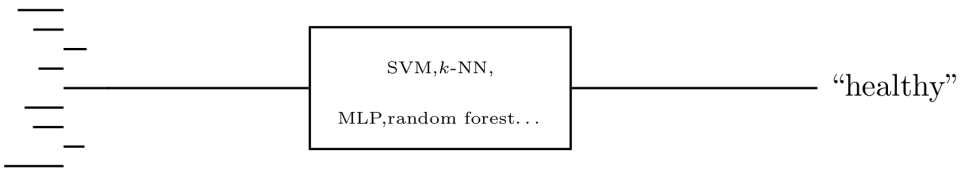


Figure 7.5.1 – Principe de l'apprentissage par transfert, extrait de la thèse de Ghouthi Boukli Hacène.

une technique de l'état de l'art que nous avons exploitée et améliorée. En effet, nous avons proposé une technique d'apprentissage incrémental à budget limité (BRIL) [Bou+17a] et l'apprentissage par transfert avec augmentation de données (TILDA) [Bou+18] pour permettre une intégration aisée sur cible à ressources limitées, typiquement un FPGA ou un processeur embarqué avec peu de mémoire par rapport à un serveur de datacentre.

TILDA s'appuie sur un réseau issu d'un apprentissage antérieur sur une base de données importante, et dont on a supprimé les dernières couches qui effectuent la classification. Le réseau ainsi tronqué est performant pour extraire le vecteur caractéristique d'une entrée donnée. Ce vecteur caractéristique est ensuite divisé en sous-vecteurs pour améliorer les performances, et quantifié pour réduire la complexité du calcul et la mémoire nécessaire et on utilise finalement un classificateur inspiré des techniques de plus proche voisin pour apprendre de manière incrémentale un exemple à la fois. TILDA offre ainsi une solution d'apprentissage incrémental où l'algorithme est adapté à la volée en utilisant de nouvelles données tout en conservant les connaissances acquises précédemment. Une architecture matérielle pour intégration sur FPGA fut également proposée, démontrant la faisabilité d'un apprentissage incrémental sur puce, **tout en conservant la précision de classification égale à celle de l´état de l'art.**

En outre, une nouvelle méthode d'élagage basée sur l'attention, *Shift Attention Layer* (SAL), a été introduite. Elle remplace une couche convolutionnelle conventionnelle par la concaténation d'une opération de décalage et d'une simple convolution 1x1. L'idée est d'utiliser l'élagage non seulement pour réduire la taille des réseaux de neurones mais aussi pour réduire considérablement le nombre d'opérations. À cette fin, nous avons équipé chaque noyau convolutif d'un mécanisme d'attention visant à apprendre quel poids doit être conservé dans la couche de décalage résultante. Nous avons démontré que SAL réduit à la fois la mémoire et les calculs requis par une solution d'inférence basée sur l'apprentissage profond. Par ailleurs, nous avons proposé une architecture d'implantation sur FPGA pour SAL [Hac+20].

La qualité et la pertinence des travaux de Ghouthi Boukli Hacène lui ont valu, outre un nombre conséquent de publications, de nombreuses reconnaissances. La première fut d'être **accueilli au sein du laboratoire MILA de Yoshua Bengio à Montréal** en tant que stagiaire lors du doctorat puis d'être associé à ses travaux. Les dernières en date sont le **premier prix de la thèse du programme Futur et Ruptures 2020 de la Fondation Mines-Télécom ainsi que le premier prix de thèse IA 2020 par l'Association française pour l'Intelligence Artificielle**.

A la suite de ces travaux, en janvier 2020, Vincent Gripon, Mathieu Léonardon et moi-même avons débuté l'encadrement de la thèse de Hugo Tessier en contrat CIFRE avec

PSA pour justement étudier l'implémentation de réseaux de neurones sur système embarqué permettant d'augmenter l'autonomie de conduite des véhicules. Les premiers résultats sur des techniques d'élagage et de régularisation sélective sont prometteurs et ont débouché sur le dépôt d'un brevet par PSA.

7.6 Perspectives pour une intelligence artificielle autonome : le défi de la performance à moindre coût matériel et énergétique

L'intelligence artificielle investit de nombreux systèmes et environnements avec facilité tant qu'une connexion aux centres de calculs massifs est disponible. En effet, lorsque de l'apprentissage profond est considéré notamment, l'intelligence résultante est exceptionnellement puissante tant que les phases d'apprentissage et d'inférence disposent de grands volumes de données et d'une puissance calculatoire massive. Ceci requiert un déport des tâches de calcul et de mémorisation des terminaux vers des centres disposant de fortes capacités de stockage et d'unités de calcul distribuées à bases de processeurs généralistes, graphiques ou dédiés. Ainsi, le paradigme actuel est d'associer l'apprentissage profond à des datacentres, généralement propriétés de grandes firmes internationales. Les applications d'inférence ou de prédiction peuvent aussi être exécutées au sein de datacentres ou sur des serveurs de calculs indépendants, avec souvent des contraintes de faible latence et de fort débit de traitement pour permettre l'exécution d'applications utilisatrices. Ainsi, Nvidia a optimisé ses GPU de la gamme Volta pour inclure des unités dédiées aux calculs tensoriels, les *Tensor Cores*, que l'on retrouve sur les cartes V100 avec 640 unités pour une puissance nominale de 125Tflops/s [NVI17], même si en pratique elle n'a été mesurée au maximum qu'à 74% de sa capacité [WB+19]. De même, Google a investi dans la conception d'ASIC dédiés à la performance de ses services, les *Tensor Processing Units* (TPU) adaptés initialement aux traitements tensoriels avec une arithmétique sur 8 bits [Jou+17]. Les puces TPU de 2017 consommaient 40W en activité (28 W en attente) pour fournir 92 Tera opérations sur 8 bits par seconde (TOp8b/s). Si l'on veut exécuter un algorithme issu d'un apprentissage machine sur des appareils portables ou autonomes tels que des des robots ou des smartphones ne possédant qu'une quantité limitée de ressources (mémoire et calcul) et d'énergie, un circuit similaire à un TPU est clairement inapproprié. Il faut donc des circuits dédiés bien mieux adaptés à ces terminaux "légers". L'argument couramment employé contre cette nécessité de solutions d'intelligence embarquée, donc faible coût, est que ces terminaux ont un accès au *cloud* et donc à ses ressources de calculs. Certes, mais ce n'est pas satisfaisant et ce n'est pas systématique !

Considérons quelques cas d'application distincts pour étayer notre vision du problème : le

véhicule autonome, l'analyse d'environnement sonore pour de l'assistance à la personne et la station de base d'un réseau cellulaire.

Un véhicule autonome doit être capable de prendre une décision vitale pour ses passagers indépendamment de la disponibilité d'un accès au réseau ou de la charge calculatoire d'un datacentre qui ne répondrait pas (à temps). L'intelligence alors embarquée doit offrir une garantie de décision de très faible latence et exacte, que le réseau de données soit disponible ou pas.

De la même manière, si l'on considère l'assistance à la personne par analyse de son environnement sonore, qui souhaite voir l'ensemble de sa vie analysée par des entreprises aux pratiques non maîtrisées ? Dans le cas particulier d'une personne souffrant d'un handicap, la solution d'assistance doit elle contraindre cette personne à une plus faible protection de sa vie privé ou tout simplement de son intimité ? Non seulement l'inférence se doit alors d'être réalisée localement et cependant efficacement, sans échange avec un centre de calcul, mais aussi l'apprentissage doit être autonome et incrémental tout au long de la vie de la personne, ce qui est bien plus complexe que l'inférence. Cela doit se faire efficacement et à un coût acceptable par un individu.

Enfin une station de base dans un réseau cellulaire doit disposer de facultés d'adaptation et de décision extrêmement rapides face à un environnement très changeant. Si elle doit en permanence analyser son environnement radio et son lien au réseau pour ajuster au mieux l'usage des ressources et la qualité des communications, elle est soumise à des contraintes de réactivité qui peuvent être incompatibles avec l'usage d'un datacentre (plusieurs –dizaines de – millisecondes s'écoulent avant un premier retour). D'une manière plus générale, l'usage du *machine learning* pour les réseaux sans fil a été analysé [Sun+19] et soulève très clairement cette limite de réactivité. L'inférence de l'état de fonctionnement d'une station de base se doit d'être alors localement calculée, mais est-ce faisable ? Outre ces considérations de sécurité des véhicule autonomes, de préservation des données personnelles et de réactivité face à un environnement complexe et extrêmement variable, l'usage massif de bande passante sur le réseau pour échanger des données est une aberration énergétique quand les mêmes services pourraient être rendus en local sans le surcoût énergétique de transport de données.

De cette analyse, nous retenons des contraintes d'autonomie et d'indépendance des centres de calcul pour aussi bien l'apprentissage que l'inférence, ce qui requiert d'investiguer conjointement des méthodes d'apprentissage profond adaptées et des systèmes embarqués performants, pour sortir du schéma actuel d'apprentissage profond forcément adossé à des centres de calcul.

En outre, le domaine de l'apprentissage profond, très compétitif, se focalise essentielle-

ment sur des techniques validées sur des bases de données de vision artificielle (ImageNET, CIFAR-10/100, MNIST). Cet étalon de la vision artificielle biaise trop souvent la réflexion au sein de la communauté scientifique car il privilégie la précision de classification sur le cas d'usage associé et ne se soucie qu'exceptionnellement des conditions d'usage (énergie réduite, ressources de calcul limitées, environnement et conditions variables). Il faut donc concevoir aussi bien des techniques d'apprentissage profond autonome que des systèmes embarqués eux-aussi autonomes pour permettre un paradigme d'intelligence artificielle indépendante de datacentres (tant pour l'apprentissage que pour l'inférence) capables d'exploiter des ressources limitées efficacement. Nous estimons en outre que les solutions autonomes seront applicables aux cas plus généraux : favoriser des solutions simples, adaptées au cas d'usage et localisées permettront de réduire l'usage des datacentres et donc des transferts massifs d'information par le réseau, dont le coût environnemental est non négligeable. D'après les données de l'ADEME en novembre 2019 [1] , le secteur du numérique est responsable de 4 % des émissions mondiales de gaz à effet de serre (dont 25 % dus aux datacentres et 28 % dus aux infrastructures réseau) et ceci devrait doubler d'ici 2025.

Google a d'ailleurs anticipé cette évolution et a dévoilé dès 2018 deux gammes de puces dans ce sens. Les *Cloud TPUs* de troisième génération qui poursuivent la course à la performance massive et les *Edge TPUs* qui visent les terminaux à la frontière du *cloud*, autrement dit ceux du grand public, avec une capacité dédiée de 4TOp8b/s pour une consommation de 2W. Ainsi, toute une gamme de puces, dénommée *Coral*, est annoncée pour 2020, pour fournir des accélérateurs d'inférence comme d'apprentissage "allégé" avec des techniques de type *transfer learning* [Goo20]. Malheureusement, nous ne savons pas encore dans quelles mesures ces puces seront ouvertes et indépendantes des datacentres pour des transferts de modèles.

Les solutions actuelles basées sur le paradigme du datacentre nous enchaînent donc à des multinationales, ce qui n'est pas souhaitable pour différentes raisons développées précédemment. Notre ambition est de contribuer à une indépendance vis-à-vis de ces grands groupes.

Comme l'ont montré les travaux menés autour de la thèse Ghouthi Boukli Hacène au sein du département Électronique de l'IMT Atlantique et du MILA à Montréal, l´apprentissage et l'inférence sur système embarqué et autonome est faisable et offre des performances en adéquation avec l'état de l'art d'un point de vue de la précision de classification pour les datasets communs. Néanmoins, ces travaux ont aussi relevé de nombreux défis comme l'intégration sur circuits de type FPGA et la pleine exploitation de leurs spécificités, tant calculatoires que

1. Face cachée du numérique (La), Réduire les impacts du numérique sur l'environnement, ADEME, nov. 2019, https ://www.ademe.fr/face-cachee-numerique

de reconfiguration. Les processeurs classiques disponibles au sein des terminaux communs (Smartphones, tablettes, ordinateurs grand public) n'offrent en outre pas suffisamment de performance calculatoire et tout individu ne peut pas disposer d'une machine flexible de traitement coûtant plusieurs (dizaines de) milliers d'euros. En outre, le matériel reconfigurable est à privilégier sur les circuits dédiés pour leur facilité de mise à jour et donc la pérennité du matériel. Une puce dédiée de type ASIC (Application Specific Integrated Circuit) est certes très performante mais ne peut pas s'adapter aux évolutions des besoins et des techniques qui rendent ces solutions technologiques obsolètes trop rapidement. Un FPGA est certes plus cher à l'unité mais offre une durée de vie supérieure au matériel. En outre, il permet une configuration répondant exactement au besoin de tout cas d'usage qui peut évoluer. En résumé, une solution reconfigurable de type FPGA est bien plus **durable** pour un niveau donné de ressources matérielles.

Quelles sont les performances opérationnelles que l'on peut attendre d'un système intelligent embarqué et autonome ainsi implanté ? Peut-il prendre la bonne décision sous les contraintes fortes des applications précédemment citées ? Quelles sont les ressources requises ? Quelles combinaisons algorithmes-matériels sont-elles les plus adaptées pour répondre aux contraintes applicatives ?

Des trois cas d'usage que nous avons listés précédemment un point commun surgit : le signal traité (comme un signal radio ou sonore) est bien souvent un signal scalaire temporel (ou issu d'une fusion de tels signaux) et il serait donc pertinent de tester nos idées sur des datasets plus proches que des images de chat ou des chiffres manuscrits. Si nous cherchons des datasets sur des signaux de type scalaire temporel pour effectuer de l'apprentissage machine, il nous faut de grands volumes de données annotées et disponibles à toute la communauté scientifique, qui peut ainsi valider par la suite nos travaux. Sous ces contraintes, les datasets audio correspondent parfaitement à nos attentes. Ils permettraient de considérer immédiatement le cas d'application des assistants vocaux précédemment cités. Ensuite, nous pourrions extrapoler nos résultats, à un premier niveau d'approximation, à des signaux radio par exemple en passant à l'échelle fréquentielle.

Des travaux sont actullement menés en ce sens en collaboration avec Interface Concept, entreprise membre de Pracom et experte en conception de systèmes embarqués, notamment pour la défense et l'aéronautique. Ghouthi Boukli Hacène y contribue notamment en tant que chercheur CNRS au département, ainsi qu'un stagiaire de Master Recherche, Hugo Le Blevec de mai à octobre 2020 qui a proposé une intégration matérielle sur FPGA de réseaux particulièrement favorables à une telle intégration [CHG20]. Nous projetons de démarrer une thèse sur cette thématique dans le courant de l'année universitaire 2020-21.

Par ailleurs, je souhaite explorer des solutions combinant des cibles reconfigurables de

différentes complexités et des systèmes logiciels sur CPU/GPU. En effet, à ma connaissance, aucune tentative d'intrication logiciel-matériel aussi avancée que celle menée avec les thèses d'André Lalevée et Franck Cornevaux-Juignet n'a été tentée dans le domaine de l'intelligence artificielle. Les obstacles sont en effet nombreux : nécessité d'un lien haut-débit entre les différents cœurs de calcul, flots de conception distincts, représentation de l'information sur des formats traditionnellement différents (virgule flottante en logiciel *versus* virgule fixe en matériel). Et pourtant, j'y vois des forces complémentaires que nous pourrions combiner. Les solutions à base de CPU/GPU bénéficient de ressources de calcul flottant adaptées à la diversité des signaux traités, naturellement adaptables à des dynamiques et précisions extrêmement variables mais aussi extrêmement énergivores, tandis que des solutions FPGA classiques ne sont performantes que sur des représentations malheureusement limitées de l'information mais à des débits de traitement et des besoins énergétiques bien avantageux. En adoptant une démarche similaire à celle menée pour l'analyse de trafic réseau, nous pourrions exploiter les différentes capacités de reconfiguration des FPGA et leur performance ciblée en adéquation avec des CPU/GPU qui géreraient en priorité les points nécessitant une large dynamique et haute précision et adapteraient les opérateurs sur FPGA en conséquence. En outre, les réseaux profonds pourraient être conçus pour faciliter cela, voire en tirer profit. L'adéquation algorithme-architecture devrait ainsi être orientée vers une adéquation {réseau profond spécifique}-{traitement flottant et supervision logicielle sur CPU/GPU}-{accélérateur reconfigurable}. Dans ce contexte, nous sommes particulièrement intéressés pour explorer ce que peuvent offrir les solutions reconfigurables émergentes telles que les Adaptive Compute Acceleration Platform (ACAP) comme les VERSAL de Xilinx ou Speedster7t d'Achronix qui intègrent un NoC câblé, pour répondre au défi du transfert de données entre de multiples instances de calcul et des mémoires conséquentes. Cette technonologie est au cœur de nombreux enjeux pour mes travaux de recherche en Intelligence Artificielle mais aussi en traitement haut débit pour les télécommunications, comme expliqué aux chapitres précédents.

8

Conclusion sur mes travaux et mes perspectives de recherche

L'**adéquation entre mes enseignements et ma recherche** est multiple. Tout d'abord, mes travaux de recherche inspirent mes enseignements, tant pour les illustrations dans les enseignements fondamentaux que pour former les étudiants à l'état de l'art des techniques de conception de systèmes embarqués, et ainsi répondre aux attentes de leurs futurs employeurs. Je diversifie les domaines d'enseignement où j'interviens, à l'image de ce que je fais en recherche, soulignant dans mes cours les invariants de l'intégration de systèmes et les spécialités de chaque domaine. Ensuite, par mon implication en enseignement et dans l'**orientation indvidualisée des élèves**, je peux identifier des profils d'intérêt et effectuer un **recrutement de doctorants de grande qualité**, ayant bien souvent effectué des stages chez mes partenaires de recherche. La sélection exigeante des candidats à la thèse est un élément clé de mon activité.

Ma recherche est avant tout **collaborative**. Elle a été souvent adossée à des contrats bilatéraux avec des entreprises et à des collaborations au sein de mon département mais aussi en dehors. C'est de ces collaborations que vient la diversité des domaines d'application que j'ai considérés. Au travers des précédents chapitres, j'ai dépeint mes travaux de recherche autour du **traitement et de la transmission de l'information selon le triptyque**

débit — ressources matérielles et énergétiques — flexibilité. Ces travaux ont été menés en développant une forme personnelle d'**adéquation algorithme-architecture qui a intégré régulièrement l'adéquation à l'information, au signal et au transistor**. En effet, j'ai contribué en conception de circuits mixtes analogique-numérique aussi bien pour des récepteurs que pour des unités d'intelligence artificielle avec des choix originaux d'association. Par exemple, pour les décodeurs de canal, j'ai étudié l'intérêt de traitements analogiques de l'information numérique par des transistors opérant sur des tensions ou des courants et la force d'une approche analogique pour exploiter le traitement stochastique au sein de structures complexes, ce qui est détaillé au Chapitre 5. De même, j'ai contribué à la conception de puces de réseaux de neurones en associant toujours information, signal et transistor de manière adéquate afin de proposer une solution efficace (Chapitre 7), c'est-à-dire maximisant un compromis cible entre débit, ressources et flexibilité. Évidemment, d'autres critères sont à considérer dans la conception d'un système de traitement et/ou transmission de l'information, comme la latence, la robustesse aux changements de conditions de fonctionnement, la généricité ou la portabilité, mais ils ont globalement été secondaires pour les cas d'application que j'ai considérés. Tant pour l'analyse de trafic réseau (Chapitre 6) que pour la conception d'interfaces cerveau-machine ou l'usage d'une intelligence artificielle (Chapitre 7), **des choix peuvent permettre de satisfaire ce triptyque si l'on sort des outils employés par la communauté d'origine**. Ces domaines qui exploitent grandement des solutions logicielles sous-estiment, voire méconnaissent, les solutions reconfigurables que sont les FPGA et leurs dernières évolutions. Pourtant, elles offrent des performances surprenantes en termes de débit, d'économie d'énergie et restent suffisament flexibles pour être aisément adaptées à différents cas d'usage, mais toujours **à condition de savoir adapter le traitement à leurs spécificités**. C'est ce que nous avons montré en apprentissage machine, avec l'adaptation de SVM (Chapitre 6) ou de techniques d'apprentissage profond (Chapitre 7) aux spécificités des FPGA, et ceci sans perte de précision.

J'ai ainsi contribué à montrer que les trois éléments de ce triptyque ne sont pas forcément orthogonaux, à condition de proposer une adéquation algorithme-architecture poussée jusqu'à la représentation de l'information.

J'ai acquis notamment une expertise en intégration de solutions de correction d'erreurs en communications numériques et d'accélérateurs matériels pour les réseaux, ce qui a contribué à l'obtention de contrats de recherche industrielle, qui m'ont offert de nouveaux problèmes, de nouveaux points de vue et ainsi de nouvelles idées qui accroissent et diversifient mon expertise. Ce cercle vertueux fait l'objet de toute mon attention et je souhaite l'enrichir par plus de contrats collaboratifs, selon les perspectives détaillées dans chacun des Chapitres 5, 6 et 7. En bref, **trois voies sont initiées et seront développées dans les**

prochaines années. La première concerne les communications numériques haut-débit satellitaires et terrestres appellant à développer des systèmes autour de **liaisons optiques en espace libre**, pour lesquelles des récepteurs aux performances extrêmes doivent être conçus. Ces récepteurs devront intégrer notamment des décodeurs de canal offrant des taux d'erreur faibles et des débits attendus de plusieurs centaines de Gbps, tout en ayant une consommation permettant une intégration "thermiquement" possible. Sinon, l'énergie à dissiper sera telle que le système ne pourra pas être intégré à coût raisonnable. Tant le choix des codes correcteurs d'erreur que la conception de l'architecture et le choix de la cible matérielle sont essentiels. Initiés via un contrat industriel et connexe à d'autres travaux pour les communications optiques sur fibre, ces travaux nécessiteront des collaborations étendues, notamment avec des collègues compétents en conception de codes et de procédés de synchronisation pour ces canaux spécifiques.

La seconde voie est connexe à la première et liée au débit croissant des communications satellitaires et le besoin d'une intégration flexible, évolutive et compétitive pour le segment sol, ce que le *cloud-RAN* offre. Néanmoins, comme je le constate dans un actuel contrat de recherche pour l'ESA, le ***cloud-RAN* dans sa version satellitaire**, dans les bandes de fréquence actuellement visées, se trouve face au défi d'une bande passante RF instantanée à transporter sur un réseau de plusieurs GHz. Je considère actuellement deux pistes. La première est de doter les chaînes actuelles de communication de solutions de compression ET de transport réseau au plus proche du lien physique, pour éviter tout goulot d'étranglement aux interfaces des réseaux et fournir des capacités de traitement parallèle. Les FPGA et notamment leur évolution en RFSoC, intégrant les convertisseurs au sein de la puce reconfigurable, me semblent une priorité à investiguer. La seconde piste consiste à revoir la chaîne de communication et notamment la conversion analogique-numérique. Les convertisseurs analogique-numérique peinent à suivre la disponibilité de bande-passante et pourraient être mieux exploités en intégrant des techniques de *compressed sensing*. Celles-ci requièrent de concevoir des étages analogiques en amont de chaque convertisseur analogique-numérique qui ne doit plus échantillonner qu'une bande passante réduite. Cependant, ces étages analogiques doivent être conçus pour fonctionner à des fréquences élevées (bandes Ka et Q/V) et être suffisament flexibles pour être adaptés selon le besoin du système. Le système résultant serait globalement plus performant mais aussi plus complexe à concevoir, requérant simultanément plusieurs compétences, de l'analogique à la théorie de l'information.

Ces deux premières voies me semblent proches des défis considérés dans le prochain programme Horizon Europe, et qui pourrait être une source de financement de travaux collaboratifs avec mes actuels (et d'autres) partenaires industriels et académiques, nationaux et internationaux. En prolongement et complément de ces travaux, donc à moins court terme,

j'anticipe une nécessaire investigation de techniques de codage correcteur au niveau paquet, s'adaptant aux contraintes et spécificités du cloud-RAN naissant et à celles des applications émergentes.

J'envisage ainsi un programme de recherche à moyen terme prolongeant les activités fraîchement débutées et débouchant sur un transfert industriel.

La troisième voie que je considère est l'**apport des solutions reconfigurables aux systèmes autonomes à base d'intelligence artificielle**. Cela peut sembler distant du reste de mes activités, mais en réalité cela reste très proche. Les systèmes d'apprentissage-machine reposent sur des puissances de calcul et de mémoire actuellement gigantesques, limitant l'accès à l'innovation à un nombre réduit de contributeurs ayant accès à cette puissance et limitant le déploiement aux systèmes embarqués. Comme exprimé dans les perspectives du Chapitre 7, la communauté scientifique avance en ce sens mais il me semble que l'adéquation algorithme-architecture reste encore balbutiante. L'usage des FPGA et plus généralement des solutions reconfigurables (comme les ACAP) est encore réellement sous-estimé, les réseaux d'apprentissage profond n'étant que trop rarement élaborés pour une intégration sur ressources limitées et encore moins sur cible reconfigurable associée à un CPU autrement que pour une accélération locale. En continuant sur la voie ouverte par la thèse de Ghouthi Boukli Hacène, nous devrions considérer le potentiel de l'intrication logiciel-matériel telle que considérée dans les thèses d'André Lalevée et Franck Cornevaux-Juignet et ne pas voir le FPGA comme une seule solution d'accélération locale et statique, mais comme une ressource dynamique. En effet, les solutions à base de CPU/GPU bénéficient de ressources de calcul flottant adaptées à la diversité des signaux traités, naturellement adaptables à des dynamiques et précisions extrêmement variables mais aussi extrêmement énergivores, tandis que des solutions FPGA classiques ne sont performantes que sur des représentations malheureusement limitées de l'information mais à des débits de traitement et des besoins énergétiques bien avantageux. En associant astucieusement les capacités de reconfiguration des circuits à l'analyse large dynamique et large précision des solutions CPU/GPU, et grâce à l'augmentation des débits intra-cartes, des avancées majeures peuvent être obtenues pour une intelligence autonome. Au vu des enjeux de ce domaine, de nombreuses collaborations sont en cours et appelées à se développer, tant au niveau national qu'international, avec des perspectives d'activité de l'ordre d'au moins quelques thèses successives. Par ailleurs, je veillerai à poursuivre l'effort de transfert industriel déjà initié dans ce domaine.

L'ensemble de ces perspectives m'enthousiasme et le plus dur reste de faire le choix des priorités.

9 Publications sélectionnées

9.1 Article IEEE TCAS1 sur le décodage semi-itératif pour les turbocodes [Arz+07]

Semi-Iterative Analog Turbo Decoding

Matthieu Arzel, Cyril Lahuec, Fabrice Seguin, David Gnaedig, and Michel Jézéquel, *Member, IEEE*

***Abstract*—Based on multiple-slice turbo codes, a novel semi-iterative analog turbo decoding algorithm and its corresponding decoder architecture are presented. This work paves the way for integrating flexible analog decoders dealing with frame lengths over thousands of bits. The algorithm benefits from a partially continuous exchange of extrinsic information to improve decoding speed and correction performance. The proposed algorithm and architecture are applied to design an analog decoder for double-binary codes. Taking full advantage of multiple slice codes, the on-chip area is shown to be reduced by ten when compared to a conventional fully parallelized analog slice turbo decoder. The reconfigurable analog core area for frames of 40 bits up to 2432 bits is 37 mm^2 in a 0.25-μm BiCMOS process.**

***Index Terms*—Analog decoding, multiple-slice turbo codes, turbo codes.**

I. Introduction

ANALOG decoding has been a research topic for about a decade now. Treating frame lengths of only a few bits at early stages, the decoders have continued to grow in size and complexity to deal with longer frame lengths. Even though the latest implemented analog decoders can treat a few dozens information bits (121 bits for the largest decoder to date [1]), industrial applications commonly deal with far larger frame lengths: up to a few thousand bits. Designing analog decoders for such large frames is impracticable with the usual fully parallelized architecture. Despite the fact that the basic analog cells are far smaller than their digital counterparts (by a factor of five in a BiCMOS process [2]), the one-to-one relation between symbol to decode and decoding element required by the fully parallelized processing yields a prohibitive on-chip area in the case of frame lengths of a few thousand bits. If the parallelism of analog decoders is not reduced, analog decoding will go no further than being an academic research topic despite its clear advantages in terms of speed and power consumption [2], [3] over digital decoders.

An obvious solution to this area problem is to reuse hardware somehow. A first attempt to provide such a solution was introduced in [4]. It is basically a mixed decoder core reusing a piece of analog decoding network connected to digital memories through digital-to-analog converters (DACs) and analog-to-digital converters (ADCs). Thanks to analog hardware reuse and interleaving via digital memories, the complexity can be reduced by an order of one hundred when compared to a fully parallelized analog decoder. This mixed architecture reduces chip area but unfortunately it reduces the data throughput too. Indeed, access to digital memories and use of DACs and ADCs are a bottleneck in terms of data rate. Moreover, even though the analog core converges thanks to the exchange of probabilistic information related to the trellis states, it does not benefit from a continuous exchange of extrinsic information. Yet, such an exchange yields powerful information and improves decoding speed and correction performance as will be shown in this paper. Then, the question to address is: how to reduce the parallelism without losing the benefits of a continuous exchange of extrinsic information? A possible solution to this problem lies in the properties of multiple slice turbo codes which were introduced to improve decoder performance not only for digital implementations, but also for analog ones, as explained by Gaudet in his Ph.D thesis [5]. Nevertheless, to the authors' best knowledge, no actual digital implementation of multiple slice turbo codes has been reported to date.

Reference [6] proposed a joint code-decoder architecture for the analog implementation of digital video broadcasting (DVB)-RCS-like decoders using these multiple-slice turbo codes. They were shown in [7] to introduce no performance degradation compared to conventional turbo codes. However, although slice turbo codes enhance the design of the analog decoder in terms of simplicity, reliability and reusability, the implemented decoder still occupies a large on-chip area. This paper proposes a modified architecture of the fully parallelized analog slice decoder presented in [6] yielding a significant reduction of occupied area with a minimal loss of data throughput. This is achieved by reducing parallelism and keeping a partially continuous exchange of extrinsic information. Moreover, the proposed architecture is reconfigurable. A single analog turbo decoder chip can treat different frame lengths ranging from a few dozens up to a couple of thousands of bits. Along with detailing the architecture, the paper provides some insights on the area reduction achievable using this architecture compared to a traditional fully parallel one. Based on a designed and tested elementary decoder some figures are given.

Section II gives the principles of multiple slice turbo codes and a first solution to design a corresponding analog turbo decoder. Section III shows how to fully exploit the properties of these codes to obtain an innovative semi-iterative decoding algorithm. Section IV presents the proposed reconfigurable semi-iterative architecture which is shown to offer a good compromise between on-chip area and data rate. Section V details the area of a semi-iterative architecture. It shows that this area is reduced by ten when compared to a fully parallelized counterpart. Section VI concludes this article and highlights some issues which remain to be solved.

Manuscript received January 23, 2006; revised July 12, 2006 and November 9, 2006. This work was supported in part by Brest Métropole Océane under the "Fond de Concours au Développement de la Recherche." This paper was recommended by Associate Editor B. C. Levy.

M. Arzel, C. Lahuec, F. Seguin, and M. Jézéquel are with the École Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France, 29238 Brest, France.

D. Gnaedig is with TurboConcept, 29280 Plouzané, France.

Digital Object Identifier 10.1109/TCSI.2007.897770

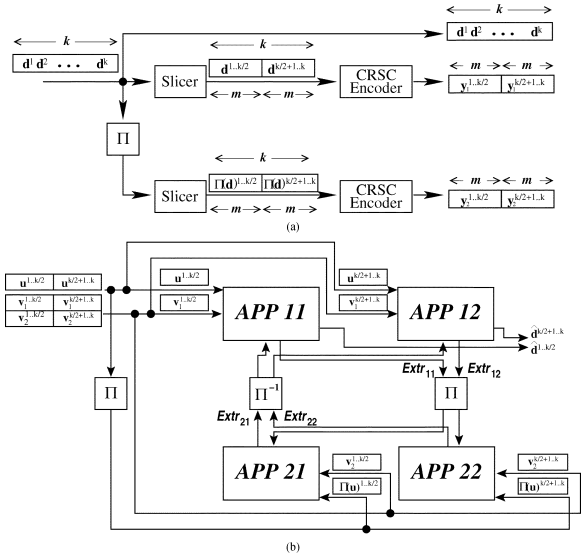


Fig. 1. (a) 2-slice turbo encoder. (b) Turbo decoder.

II. Principles of Multiple Slice Turbo Codes

A. Slice Turbo Encoding and Decoding

A frame $\mathbf{d}$ of length k is sliced into p parts of same length m. Each sub-frame, or slice, is independently encoded using a circular recursive systematic convolutional (CRSC) encoder. The interleaving is done over the full-length k of the initial frame. Finally, the interleaved frame is sliced again into p sub-frames which are again independently encoded. The principle of slice turbo codes is illustrated in Fig. 1(a) with $p = 2$.

The corresponding decoding requires $2 \times p$ (four in the present example) elementary *a posteriori* probability (APP) decoding processes—one per slice to decode and p slices per dimension, two interleavers, and one deinterleaver, as shown in Fig. 1(b). Each decoder treats one slice in the natural order (APP 11 and APP 12 in the present case) or in the interleaved order (APP 21 and APP 22) with the corresponding redundancy bits. As an example, APP 11 decodes the received slice $\mathbf{u}^{1\ldots k/2}$ with the redundancy bits $v_1^{1\ldots k/2}$—corresponding to emitted sequences $\mathbf{d}^{1\ldots k/2}$ and $y_1^{1\ldots k/2}$, respectively—and the first half of the interleaved extrinsic data produced by APP 21 and APP 22. The elementary decoder APP 11 finally provides the decoded slice $\widehat{\mathbf{d}}^{1\ldots k/2}$ corresponding to the emitted slice $\mathbf{d}^{1\ldots k/2}$.

B. Multiple Slice Interleaver

A critical part in the slice architecture is the design of the interleaver [8]. A hierarchical interleaver with two levels of permutations has been designed for a frame split into p slices of m symbols. A first level called the spatial permutation exchanges the symbols between the slices, and a second level called the temporal permutation shuffles the symbols within one slice as presented in Fig. 2. Hence, the m symbols in one slice in the natural order are distributed over the p slices in the interleaved order.

1) Design of the Interleaver: The interleaver Π is defined by its function $\Pi(k)$ given by (1), which associates each symbol

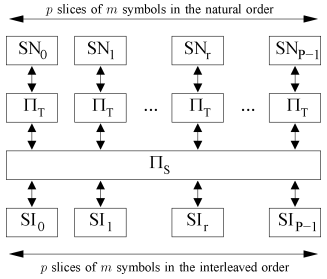


Fig. 2. Interleaving with slices.

with index k in the interleaved order with a symbol with index $\Pi(k)$ in the natural order. Let l and k denote the indexes of the symbols in the natural and interleaved order, respectively. The coding process is performed in the natural order on independent consecutive blocks of m symbols. The symbol with index l is used in slice $\lfloor l/m \rfloor$ at temporal index time $l \bmod m$, where $\lfloor . \rfloor$ denotes the integral part function. Likewise, in the interleaved order, the symbol with index k is used in slice $r = \lfloor k/m \rfloor$ at temporal index $t = k \bmod m$. Note that $k = m \cdot r + t$, where $r \in \{0, \ldots, p-1\}$ and $t \in \{0, \ldots, m-1\}$. The interleaving function is then given by

$$\Pi(k) = \Pi(t, r) = \Pi_S(t, r) \cdot m + \Pi_T(t). \tag{1}$$

2) Design of the Temporal and Spatial Permutations: The temporal and spatial permutations are designed to optimize the performance of the turbo code. The temporal permutation is then given by

$$\Pi_T(t) = \alpha_\Pi \cdot t + \beta_\Pi(t \bmod 4) \bmod m \tag{2}$$

where α_Π is relatively prime with m, and $(\beta_\Pi(i))_{0 \le i < 4}$ are four coefficients smaller or equal to m, which verify that their values modulo four are all different. The spatial permutation is defined as a circular rotation

$$\Pi_S(t, r) = (A(t \bmod p) + r) \bmod p \tag{3}$$

where A is a bijection of variable $t \in \{0, \ldots, m-1\}$ to $\{0, \ldots, m-1\}$. This optimization is described in [7] in detail and recalled here. The choice of the α_Π parameter of the temporal permutation maximizes the spread of the interleaver. Then, the bijection A is chosen to be irregular in order to introduce a high dispersion in the interleaver. Finally, the β_Π parameters are chosen in order to maximize the minimum distance of the code. It has been shown in [7] that this optimized interleaver leads to very good performance and therefore slice turbo codes introduce no performance degradation compared to conventional turbo codes.

C. Analog Decoder Design

Reference [6] explained how a complex analog turbo decoder can be designed by reusing a reduced set of small elementary decoders. Each one decodes a circular convolutional code with an APP algorithm [9]. Such an algorithm can be summarized as being a sequence of sums and products on probabilities. [10]

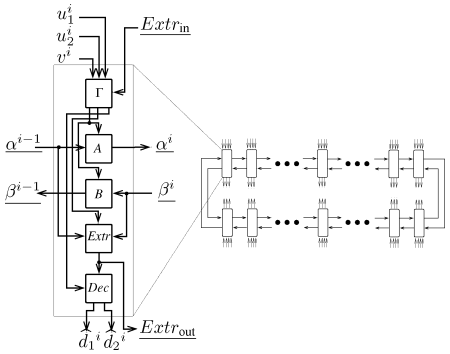


Fig. 3. Analog circular APP decoder. Each section is made up of modules defined by the APP algorithm used. There are as many sections as symbols to decode.

and [11] showed that it is possible to associate a voltage with a log-likelihood ratio (LLR) and a current with a probability using transistors and diodes. This is a clear advantage of analog decoders over digital decoders since currents can be easily added and multiplied. Thus, the APP algorithm can be implemented using either a bipolar junction transistor (BJT)-based analog network [2], [6] or a subthreshold CMOS one [3], [12]–[14]. The frame is decoded by letting this network converge to a stable state. The on-chip slice decoding network is the direct mapping of the APP algorithm normally used to decode CRSC codes. It is divided into as many sections as symbols to decode. Fig. 3 gives the example of such a network decoding a slice of double-binary symbols. Each section is made up of four modules: a Γ module to compute the branch metrics, an A module for the forward metrics, a B module for the backward metrics and a Dec module to decide on the value of the double-binary symbol. There are two sets of inputs to the section. The first set of section $\#i$ is made up of the demodulated outputs of the channel u_1^i, u_2^i and v^i, which are associated with the ith emitted symbol $d_1^i d_2^i$ and its parity bit y^i. The second set of inputs is composed of the forward and backward metric vectors $\underline{\alpha^{i-1}}$ and $\underline{\beta^i}$ produced by the adjacent trellis sections. The outputs are the metric vectors $\underline{\alpha^i}$ and $\underline{\beta^i - 1}$, fed to the adjacent sections, and the decisions $\widehat{d_1}^i \widehat{d_2}^i$ for the transmitted symbol $d_1^i d_2^i$. A fifth module is required if the APP decoder is part of a turbo decoder: the Extr module. This module computes the output extrinsic information vector $\underline{\text{Extr}_{\text{out}}}$ which is then used by the Γ module of the second APP decoder as the input $\underline{\text{Extr}_{\text{in}}}$. All the modules are connected in voltage mode. Each module, except the Dec one, is divided into a computing BJT-based core and two MOS interface circuits. [2] showed that the BJT-based core is a simple analog multiplier: the well-known Gilbert cell [15]. Designed for two differential inputs, this cell can be extended to treat m-ary symbols and also to implement any type of APP decoder [16].

III. Semi-Iterative Analog Algorithm

A. Principle

Multiple slice turbo codes were designed to enhance the parallelism of digital architectures without losing coding gains [7]. [6] showed that these codes are, in addition, well-suited

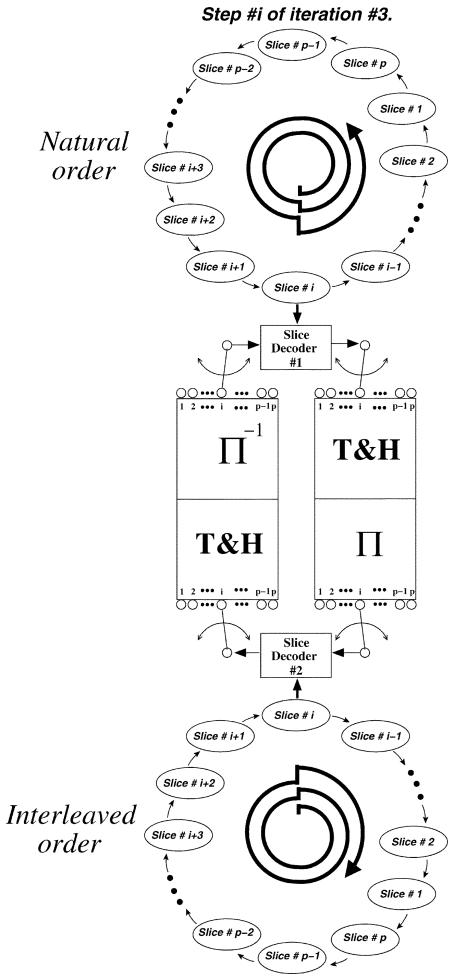


Fig. 4. Semi-iterative algorithm for multiple slice analog decoding.

to analog designs since they enhance the reliability and the reusability of the decoders. This section explains how they can also be used to reduce the parallelism of an analog decoder. Providing that a frame is divided into p slices of m symbols, the proposed new turbo decoding algorithm consisting of a number *nbIt* of turbo iterations, each divided into p steps, begins once the frame is completely stored in memories. A step corresponds to the simultaneous decoding of one slice in the natural order and of another one in the interleaved order. As shown in Fig. 4, the algorithm uses two circular APP slice decoders, two interleavers and two track-and-hold (T/H) blocks. Each slice decoder has two inputs—the slice to be decoded at the current step and the corresponding extrinsic information produced by the other decoder—and one output—the extrinsic information corresponding to the current slice. The interleavers simply shuffle and transmit the extrinsic symbols between the two decoders. Each T/H block tracks the extrinsic information currently produced and holds it only at the end of the current step. In tracking mode, this block connects continuously—via

an interleaver—one of the extrinsic outputs of a decoder to one of the extrinsic inputs of the other decoder. The algorithm is described below.

Initialize all the T/Hs to a value representing a probability equal to 0.5.

For turbo iteration it=1 **to** ***nbIt*** **do:**

For step i=1 **to** p **do:**

Map the ith slice in the natural order to the first analog circular APP slice decoder, and the ith slice in the interleaved order to the second analog circular APP slice decoder.

Let the turbo network converge and track the extrinsic information being produced. Each decoder uses and produces extrinsic symbols.

Some of the extrinsic symbols used are being produced by the decoder in the other dimension *and* some are read from the T/H memory. The former are continuously transmitted by T/Hs in tracking mode and an interleaver. The latter were computed at another step by the other decoder and are held by the T/Hs.

The extrinsic symbols being produced are tracked: some are directly used to decode the current slice in the other dimension *and* some will be used to decode another slice in the other dimension at another step.

At the end of this step, hold the extrinsic information produced in memory.

Only a pth of the extrinsic information produced by the first decoder is continuously exchanged with the second decoder and vice versa. The remaining part of the extrinsic information produced is simply tracked in order to be used at the next steps to feed the decoders. Thus, this new algorithm is partially iterative and allows a partially continuous exchange of extrinsic information between the two dimensions of a turbo decoding process. Therefore, it is neither a conventional iterative turbo algorithm—treating iteratively the two dimensions one after the other, and each slice one after the other—nor a completely continuous analog turbo process. It can be referred to as a "semi-iterative" analog algorithm performing turbo iterations.

B. Advantages

One may ask why is it so important to keep such a continuous exchange of extrinsic information? To answer this question, transistor-level simulations of an iterative 2-slice analog decoder, a semi-iterative 2-slice analog decoder and a fully parallelized two-slice analog decoder are run. Before going any further, a few words to describe the operation of the iterative two-slice analog decoder are necessary. This decoder is made up of a single APP decoder, thus it uses the minimum hardware possible. It decodes the sub-frame in the natural order and stores the produced extrinsic information in memory when done. Then, the decoder processes the corresponding interleaved sub-frame using, as a starting point, the extrinsic information previously computed and stored. The data are hence processed sequentially. This is the kind of technique used in [17]. All the decoders are fed with the same noisy frame whose signal to noise ratio is equal to 4 dB. The simulation results are then compared when the three decoders attempt to correct the same error. At the first step of the first iteration, the iterative analog tailbiting decoder and of the semi-iterative analog tailbiting decoder converge, as shown in Fig. 5(a) and (b), respectively. Note that for the iterative decoder this is simply equivalent to simulating a stand-alone circular APP decoder. Fig. 5(c) corresponds to the fully parallelized 2-slice turbo decoder. In the present case, the three decoders converge to the same solution and finally distinguish the probabilities associated with the four possible values of a received double-binary symbol. The lower the voltage, the higher the probability of the corresponding value. The hard decision at the channel output (not illustrated) is 10, which is an error, whereas all three analog decoders converge to the correct value 00.

Significant differences exist between the three convergence speeds. At the first step of the first iteration, the iterative analog circular APP decoder corrects the error within 50 ns whereas the fully parallelized and semi-iterative decoders correct within 20 ns. In this particular case, the two decoders continuously exchange half the total extrinsic information. This clearly shows that continuous exchange, even partial, enhances the decoding rate. Moreover, the decision is more reliable with such an exchange. Comparing Fig. 5(a) and (b), there is a larger gap between the decided value and its closest challenger in the case of the semi-iterative decoder than in the case of the iterative decoder.

Therefore, if the parallelism has to be reduced, implementing only one slice decoder and decoding one dimension after the other in $2 \times p$ steps—which is called iterative in this paper—is not an optimal solution. At the first step of the first iteration, the iterative decoder converges at a lower speed and with a lower reliability than the semi-iterative decoder. At the second step, the iterative decoder also benefits from less reliable extrinsic information than the semi-iterative decoder. The iterative decoder thus converges with an even lower reliability and at an even lower speed than the semi-iterative circuit.

C. BER/FER Performance

Based on simulation results, [16] and [18] both concluded that analog turbo decoding achieves better error control performance than digital turbo decoding. In [16], the model implemented an ideal fully parallelized analog decoder with RC lines connecting computing nodes. An analog slice decoder was simulated in the same manner in order to compare its theoretical error control performance with that of a digital floating point counterpart. Fig. 6 shows the simulation results. The two decoders offer the same error control, since here the analog network is finally a continuous time version of the digital one. No additional information or other handling of the existing information is provided by the analog decoder when compared to the digital decoder. If an analog slice decoder is used in an iterative way, i.e., slices are decoded one after the other and dimensions one after the other, the extrinsic information has to be stored between the iterations. Therefore, the iterative analog

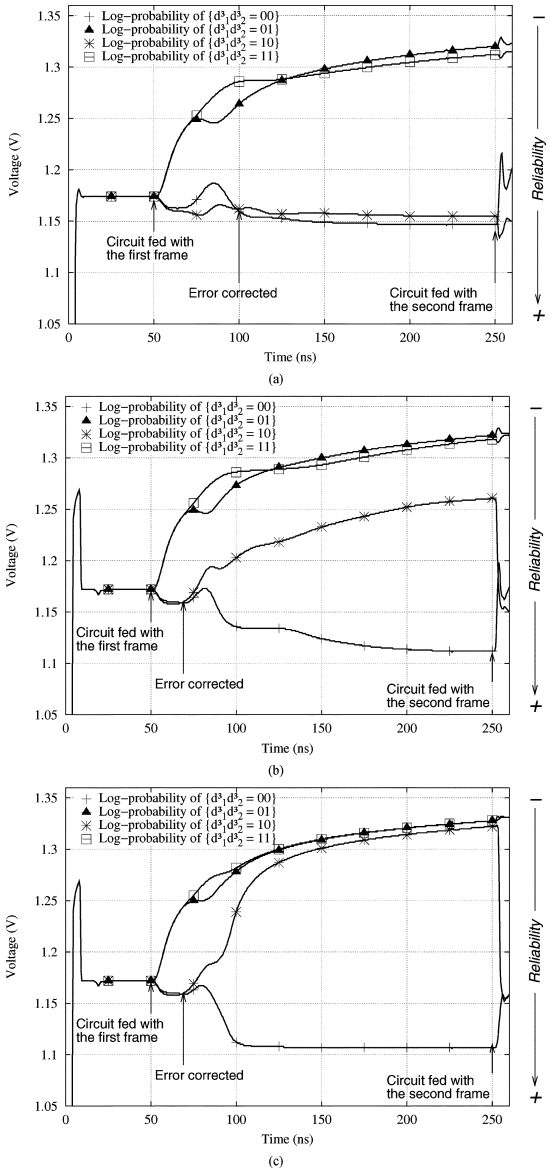


Fig. 5. (a) Correction of an error by three turbo analog decoders fed with the same noisy frame: iterative decoder, at the first step of the first iteration. (b) Semi-iterative decoder at the first step of the first iteration (c) Fully parallelized decoder (transistor level simulations).

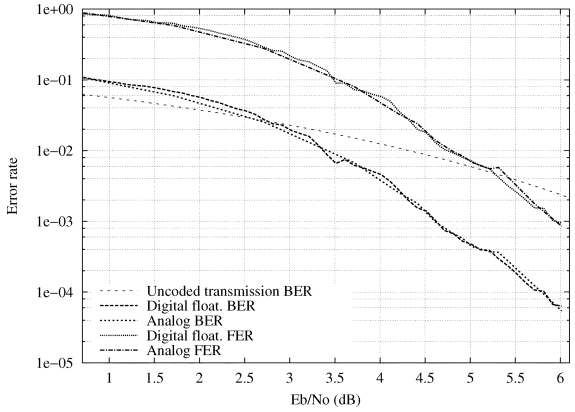


Fig. 6. Comparison of the error rates achieved by a digital floating point circular APP decoder and by an analog counterpart (8 states, 24 double-binary symbols, $R = 2/3$).

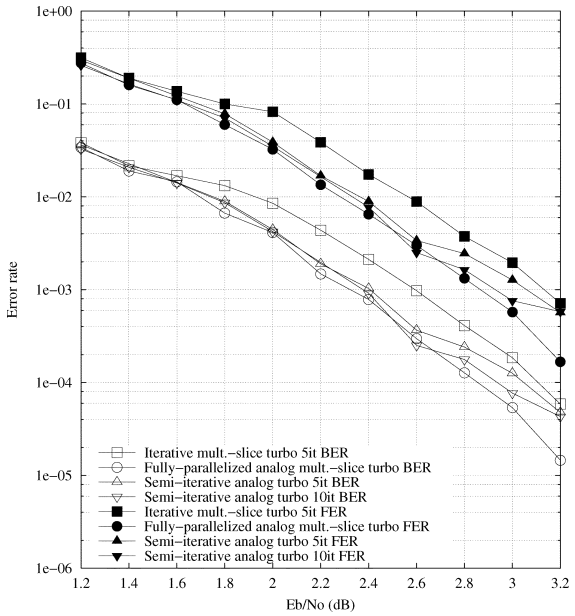


Fig. 7. Performance comparison of three analog turbo algorithms with two slices (48 symbols, $R = 1/2$): fully parallelized, semi-iterative and iterative.

turbo decoder offers the same error control as its digital floating point counterpart, for the same reason that the analog and digital slice decoders provide the same error control. Similarly, Moerz proposed an analog sliding window decoder which provides the same error correction as its digital counterpart, both being fully iterative [4]. Iterative analog turbo decoders do not benefit from all the possibilities offered by a non-iterative analog network.

The same analog modeling with RC lines is used in the present paper to implement three different *analog* turbo decoding schemes with multiple slice codes: fully parallelized, semi-iterative and iterative. Figs. 7–9 show the simulated behavioral performance of these three analog decoding schemes with two and four slices. The following remarks can be made from these curves. First, comparing the performance of the iterative analog decoding with the semi-iterative analog one shows that the latter solution outperforms the former when using the same number of turbo iterations, five in the present case. Therefore, to yield the same error control, the iterative analog decoder requires more time per step and more iterations than the semi-iterative one. As an example, Fig. 8 shows that ten turbo iterations of the iterative decoder achieve almost the same performance as five turbo iterations of the semi-iterative

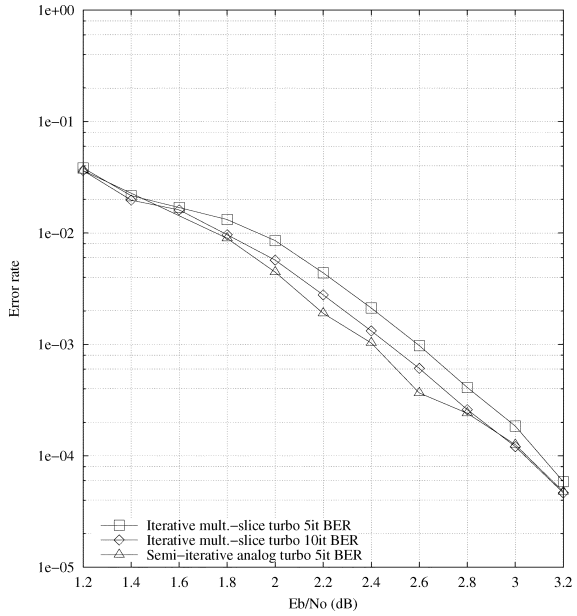


Fig. 8. BER comparison of three analog turbo decoders with two slices (48 symbols, $R = 1/2$): iterative with five iterations, iterative with ten iterations and semi-iterative with five iterations.

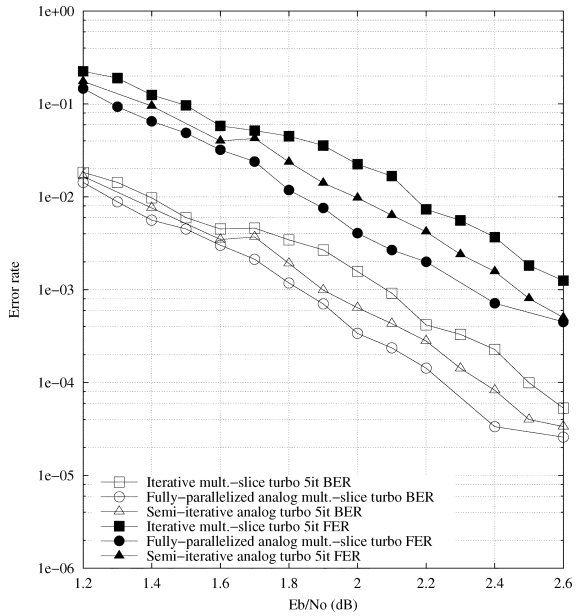


Fig. 9. Performance comparison of three analog turbo algorithms with four slices (96 symbols, $R = 1/2$): fully parallelized, semi-iterative and iterative.

decoder. Second, and as intuitively expected, the decoding performance of the semi-iterative decoder lies in between the performance of the iterative slice decoder and that of the fully parallelized decoder. These two remarks highlight again the efficiency of continuous extrinsic information exchange.

Another question arising is: does increasing the number of turbo iterations improve the performance of the semi-iterative decoder? Simulations run for ten turbo iterations show a slight improvement compared to five turbo iterations, but not a really significant one: 0.07 dB at most.

D. Data Throughput Performance and Power Consumption

The maximum data throughput of the proposed semi-iterative architecture is the fully parallel architecture throughput divided by the number of slices used and the number of turbo iteration performed. Nevertheless, it is instructive to give some figures. Considering a block of 48 double-binary symbols, i.e., 96 data bits, organized into two slices, the fully parallel, the iterative and the semi-iterative decoders are made using the same 24-double-binary symbol analog decoder. The time required by a decoder to converge gives an indication on the achievable data throughput. Using the transistor level simulations shown in Fig. 5(a) and (b), the iterative and semi-iterative analog decoders provide distinct log probabilities within 100 ns. Moreover, the rank of the log probabilities after 100 ns is the same as the rank after 200 ns in the illustrated example. Therefore, stopping the convergence process after 100 ns simply reduces the range of the log probabilities, which is similar to applying feedback coefficients to the extrinsic information between consecutive turbo iterations. Fig. 5(c) shows that the fully parallel analog decoder corrects an error within 20 ns. It is assumed that the hard decision circuit *Dec* used is the same for the three decoders and that it has a 30-ns latency [6]. It means that the fully parallel architecture, processing the 96 bits at once, can achieve a 1.92 Gb/s data throughput. The semi-iterative decoder with five turbo iterations, each one performed in 2×100ns, can achieve a 93 Mb/s data throughput. An iterative decoder of a similar complexity is made up of two elementary decoders, processing two slices of the same dimension in parallel. It must perform ten turbo iterations to achieve the decoding performance of the 5-iteration semi-iterative decoder as shown in Fig. 8. Each turbo iteration is performed in 2×100ns. Thus, with the same complexity as the semi-iterative decoder, this iterative decoder only achieves 47 Mb/s. Hence, with a complexity similar to the one of the iterative decoder, the semi-iterative architecture has a data throughput two times higher in the present case. Since the fully parallel, semi-iterative and iterative decoders use the same component decoder duplicated many times, their power consumptions depend on the power consumption of the component decoder. Using the component decoder presented in [19], the semi-iterative and iterative decoders both consume 2×300 mW $= 600$ mW for the analog decoding core at any data rate, and the fully parallel decoder consume $4 \times 300\,\text{mW} = 1200\,\text{mW}$. These figures were obtained with an analog supply voltage of 2.5 V with a 0.25-μm BiCMOS technology. The decoding core power consumption is almost static. The total power consumption is mainly due to the component decoder cores and the power consumption of the T/Hs can be neglected. Thus, the power consumption would be equal to 18 mW per decoded bit for the semi-iterative and iterative decoders and to 36 mW per decoded bit for the fully parallel decoder.

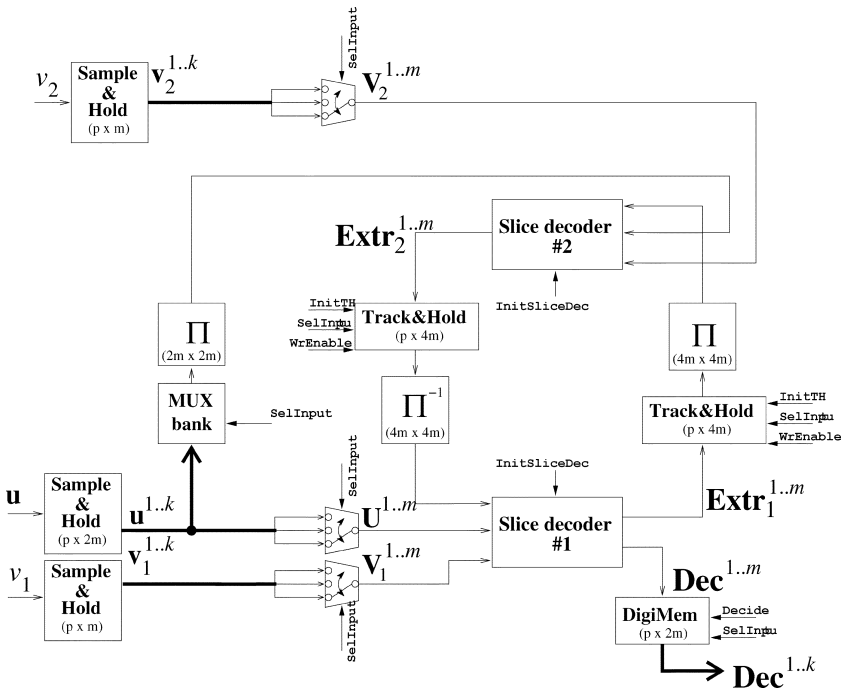


Fig. 10. Analog semi-iterative turbo decoder architecture.

IV. Semi-Iterative Analog Decoder Design

A. Architecture

Semi-iterative turbo decoding of multiple slice frames provides two axes of on-chip area reduction. First, as only one slice decoder is needed per dimension at each step of the decoding process, the same APP decoding hardware can be reused at each step. Second, thanks to multiple slice properties, the interleaving network can be re-arranged to occupy less on-chip area. Fig. 10 presents a novel architecture which exploits these two properties. The decoding circuit is fed with serial channel outputs $\mathbf{u}$, y_1 and y_2 which are sampled and held in analog memory banks. The systematic sequence $\mathbf{u}$ of k symbols and the redundant sequences $\mathbf{y}_1$ and $\mathbf{y}_2$ of k symbols—relative to the first dimension of the turbo scheme and to the second one, respectively—are then available in parallel for the decoding core.

The following example considers DVB-RCS component codes using double-binary symbols with a rate of 2/3. Therefore, each input symbol $\mathbf{u}$ corresponds to two bits, each redundancy symbol—y_1 or y_2—correspond to one bit and each extrinsic symbol is defined by four probabilities. The circuit is also made up of input sample-and-holds, multiplexers—to select the data needed at each step of the decoding process —, two slice decoders, analog T/H banks, two interleavers and one digital memory *DigiMem* according to these conditions. Of course, the proposed architecture can be applied to binary codes of different rates.

The previous semi-iterative algorithm is applied to this circuit once a frame, made up of p slices, coming from the channel is completely sampled. Therefore, the scheduling of the circuit is as follows.

Sample-and-hold (S/H) the channel outputs until a whole frame is stored.

Switch on InitTH and wait for the circuit state to become stable.

All the T/H memories are initialized with equiprobable values. Thus, the decoding process will begin with no extrinsic information.

Switch off InitTH.

For iteration it= 1 to *nbIt*:

For step i= 1 to p:

Switch on InitSliceDec: the two slice decoders are initialized with equiprobable states in order not to bias their convergences.

Put SelInput at step number i, which is the number of the current slice. Input multiplexers also select m systematic symbols $\mathbf{U}^1 \dots \mathbf{U}^m$ from the received sequence $\mathbf{u}^{1\cdots k}$ and m redundant symbols $\mathbf{Y}_1^1 \dots \mathbf{Y}_1^m$ from $\mathbf{y}_1^{1\cdots k}$ to feed the first slice decoder with the current slice. In the same manner, m redundant symbols $\mathbf{Y}_2^1 \dots \mathbf{Y}_2^m$ are selected from $\mathbf{y}_2^{1\cdots k}$ to feed the second slice decoder. The input multiplexer bank selects m systematic symbols from $\mathbf{u}^{1\cdots k}$ which are then interleaved to provide the current slice in the interleaved order to the second slice decoder.

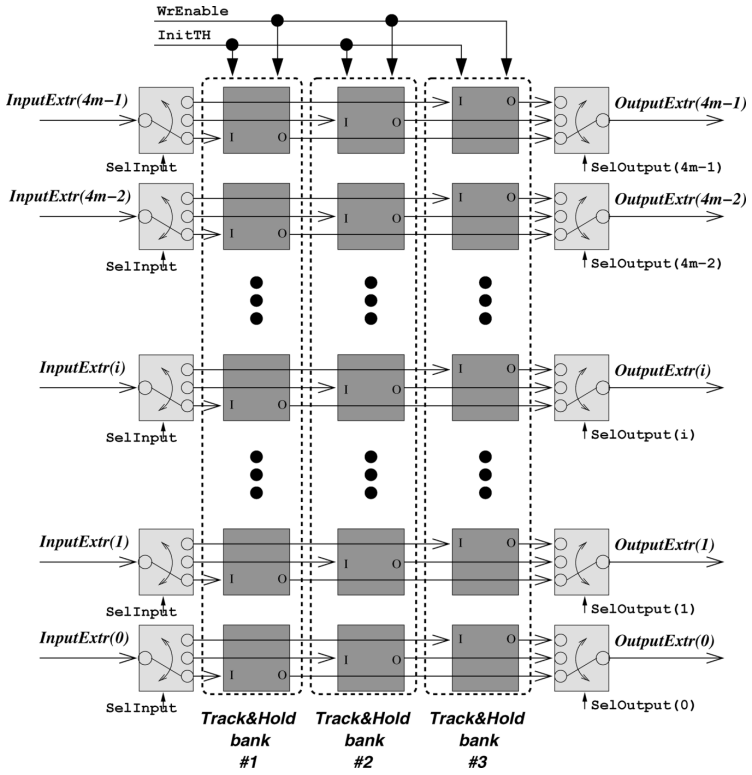


Fig. 11. Multiplexed T/H banks for storing extrinsic symbols with $p = 3$.

Switch off InitSliceDec.

If it= 1, then switch on WrEnable, and let the network converge.

The first slice decoder continuously provides m extrinsic symbols $\mathbf{Extr}_1^{1\cdots m}$ to an analog T/H bank which tracks them. At the same time this bank provides m extrinsic symbols to an interleaver which shuffles and transmits them as inputs to the second slice decoder. One part of these inputs comes directly from the first decoder, whereas another part remains constant, since it does not correspond to the slice being decoded in the first dimension but to another one decoded and held at another step.

In parallel, the second slice decoder produces m other extrinsic symbols $\mathbf{Extr}_2^{1\cdots m}$ which feed a T/H bank in tracking mode. The T/H banks provide m extrinsic symbols, which are interleaved and transmitted to the first slice decoder. Some of these symbols come directly from the second slice decoder *and* some are held values computed at another step.

Thus, the two decoders continuously exchange a pth of their extrinsic information.

Else:

switch off WrEnable. If the current iteration is not the first, all the T/Hs contain extrinsic information which was computed at the previous iterations. If the slice decoders can write in extrinsic memories as soon as a new slice is loaded, the continuously exchanged extrinsic symbols are immediately written by a decoder while the other one reads them. Thus, the result of the previous iteration is lost. To prevent this happening, the write access is not enabled until each decoder has read the extrinsic information and used it to bias the convergence.

switch on WrEnable, and let the network converge.

At the end of this step,

hold the produced extrinsic information in analog memories. This information will be used at another step of this iteration or another step of the next iteration.

switch off WrEnable,

if this iteration is the last one, switch on

Decide: the decoded slice is stored in the digital memory DigiMem.

At the end of the last iteration, the whole frame of p slices is decoded and stored in DigiMem.

The two slice decoders exchange extrinsic information through analog T/Hs and interleavers which were optimized in terms of on-chip area thanks to the properties of multiple slice turbo codes. Instead of interleaving all the extrinsic symbols and then selecting only the ones needed by the current slice decoding, only the required extrinsic symbols are selected and then interleaved. Fig. 11 illustrates the design of such a

multiplexed extrinsic T/H block with $p = 3$. It is made up of three blocks: an input multiplexer bank, $p = 3$ T/H banks, each one tracking and holding m extrinsic symbols, and an output multiplexer bank. Three external signals control this block: InitTH , WrEnable and SelInput . The latter feeds a logic block—not illustrated—which computes at each step the intern bus signal SelOuput . This signal controls the output multiplexer bank. The logic block is simple to design since the spatial permutation, as described in Section II-B2), is a circular rotation.

To illustrate how it operates, the following example is taken. For instance, assume that signal SelInput is equal to 3. Consequently, the third slice is being decoded and the corresponding extrinsic symbols are computed. They are being tracked by the third T/H bank. Then, each multiplexer of the output bank selects an extrinsic symbol in one of the T/H banks. For example, the first symbol is selected in the first bank: it is a held value since computed at step 1. The second symbol is selected in the third bank: this symbol is being tracked and also continuously exchanged since the current slice is the third. The third symbol comes from the second bank: it is a held symbol, and so on. Thus, the output multiplexers of the extrinsic T/H block interleave spatially and the following wired shuffling network interleaves temporally. The order of these two permutations is inverted when compared to the original scheme presented in Section II-B. The original scheme was designed for a digital implementation using p processing units, all of them treating, during the same clock cycle, the same symbol position but from different slices. Therefore, permuting temporally and then spatially was more suited to this digital scheme than the inverse.

Using T/Hs to store extrinsic information allows the data rate to be independent of any data storage between decoding steps. Since the data to be stored are continuously tracked, storing them after tracking requires no additional time than the time needed to switch the pass-gates. Therefore, storage of extrinsic information is transparent and introduces *no bottleneck in terms of data throughput*, unlike digital memories coupled with ADCs and DACs, as proposed in [4].

B. Reconfigurable Architecture

Standards such as DVB-RCS [20] often define several frame lengths that a single-chip decoder must accept. To cope with a large number of them, an analog decoder must also be reconfigurable. This is not a feature easy to implement with a conventional design mainly because it implies the design of large circuits. However, the semi-iterative decoder architecture is well suited for reconfigurability as shown next. To be able to cope with various frame lengths implies that the decoder must be able to treat multiple slice frames with different values of p and m. For instance, in the case of the DVB-RCS-like frames, p ranges from 2 up to 32 and m ranges from 20 up to 38. Each elementary slice decoder is implemented as a ring as shown in Fig. 3. The number of decoding sections in the decoder must hence vary with m. The ring structure of the decoder makes it easy to add or remove sections by means of simple multiplexers connecting successive sections. The size of the decoding ring is then selected by means of an N-bit word as shown in Fig. 12. As an example, the elementary slice circular APP decoder implemented

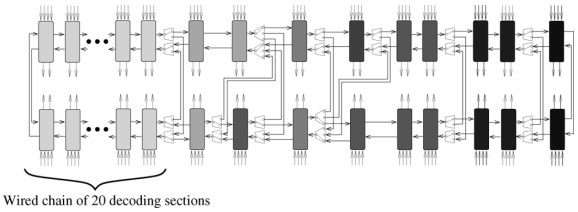


Fig. 12. Reconfigurable analog circular APP slice decoder.

in this paper has a minimum size of 20 decoding sections to which 3, 4, 6, 7, 12, 16, or 18 decoding sections can be added by means of a 3-bit control signal. It can thus decode all the slices made up of 20, 23, 24, 26, 27, 32, 36, and 38 double-binary symbols.

All the memories and multiplexers must be sized up according to the largest couple (p, m) which is (32,38). As an example, if p is equal to 12 and m is equal to 36, the multiplexers will only select the 12 first memory banks and only the first 36 positions will be used by the slice decoders.

Finally , as the interleavers shuffle m symbols, they must be reconfigurable too. [21] proposes a solution to this problem based on (P, Q) switching networks, where P and Q are particular integers. With the same notation defined in this paper, if $P = 19$ and $Q = 8$, all the shuffling schemes of 20 to 38 extrinsic symbols are possible.

V. Area Study and Design Issues

Although the presented architecture has not yet been laid out, it is nevertheless possible to give some meaningful figures about the on-chip area. A fully functional double-binary circular APP decoder designed for a 0.25-μm BiCMOS process from Philips was presented in [19]. This chip included input S/Hs as memory elements and a 24-double-binary decoding core. The next subsections detail each block of the semi-iterative architecture in terms of occupied area using data from [19]. An example is taken and actual figures given for 32 slices of 38 double-binary symbols, i.e., the total frame length is 2432 systematic bits.

A. Decoding Section and SISO Decoder

The size of the decoding section, shown in Fig. 3, is 0.124 mm^2. Thus, the elementary decoder has a total area of $0.124 \times m$, where m is the number of symbols per slice. In the example taken, $m = 38$ and the elementary tailbiting decoder decoder area is 4.7 mm^2.

B. S/Hs Modules

Since the semi-iterative scheme requires analog values to be stored in extrinsic T/Hs and input S/Hs, it is necessary to design low-leakage memory elements such as the one presented in [19]. An elementary input memory is made up of two sample-and-hold cells put in parallel as shown in Fig. 13. This architecture allows dealing with a continuous flow of data from the channel. The data pertaining to the current frame, saved in one cell, are processed while the next frame is sampled in the other one. The unity gain buffers in Fig. 13 are operational transconductance amplifiers (OTAs). The switches are implemented as

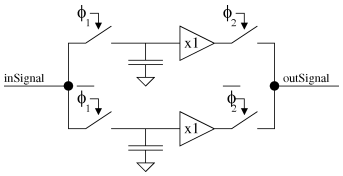


Fig. 13. Parallel S/Hs used as analog memory.

CMOS pass-gates for better reduction of charge injection. The parallel S/Hs have an area of 2565 μm^2. Hence, the total input memory, area in the case of 2432-bit long frames is 12.47 mm^2.

C. T/Hs and Interleavers

The extrinsic interleavers in the semi-iterative architecture are a little complex to estimate since they use multiplexers and T/Hs. $p:1$ and $1:p$ multiplexers are designed with CMOS pass-gates and the extrinsic T/Hs are simply implemented using the same cell composing the parallel S/Hs. Nevertheless, the sampling frequency of the extrinsic information is larger than the sampling frequency of the input data. Therefore, the T/Hs require smaller capacitors than the input S/Hs, and also less area. Since only m extrinsic symbols are selected and then interleaved to finally feed the other slice decoder, the extrinsic interleaver also has $4 \times m$ signals to shuffle instead of $4 \times m \times p$, which requires less on-chip area. This is a key feature of the semi-iterative architecture. Finally, interleaving the input systematic symbols after they have been selected helps reduce the input interleaver area. The most basic interleaver for n signals is a simple hard-wired shuffling network whose area is equal to $(n \times ((\mathrm{minWidth}) + (\mathrm{minSpace})))^2$ where minWidth is the minimal width of a wire and minSpace is the minimal space between two wires. For the process chosen, $(\mathrm{minWidth}) + (\mathrm{minSpace})$ is equal to 1 μm, thus shuffling n signals requires an area simply equal to n^2 μm^2. In the chosen example, the extrinsic interleaving, including the T/Hs, requires an area of 11.1 mm^2 while the input interleaver is 0.63 mm^2.

D. Area Comparison Between Fully Parallel and Semi-Iterative Architectures

From the above descriptions, the chip area is computed as a sum of wiring area and block area. Therefore, the areas of the fully parallelized and semi-iterative architectures can be accurately estimated. Fig. 14(a) and (b) show the areas of a fully parallelized decoder and its semi-iterative counterpart versus the number of slices to decode. These two examples use slices made up of 38 symbols (this choice is explained in Section IV-B). First, one has to note that there is a scale factor greater than ten between the two graphs. On the one hand, Fig. 14(a) shows that the most area-consuming parts of a fully parallelized decoder are the single-input single-output (SISO) decoding part, made up of $2 \times p$ circular APP slice decoders. The second largest parts are the extrinsic interleaving parts. Finally, Fig. 15 compares the on-chip area of the reconfigurable semi-iterative decoder with

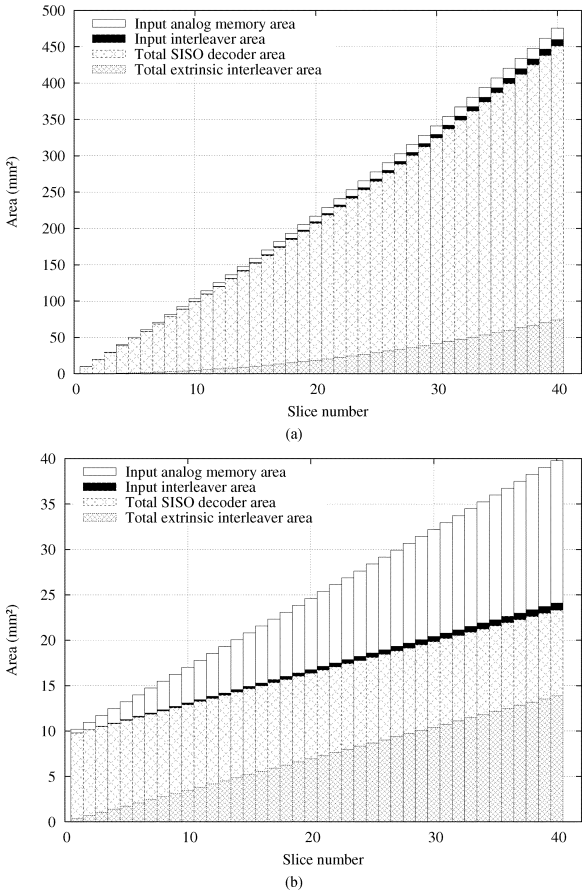


Fig. 14. Areas of (a) *fully parallelized* and of (b) *semi-iterative* architecture in Philip's 0.25-μm BiCMOS process versus the number of slices decoded, each slice being made up of 38 symbols.

the area required by a fully parallelized decoder. Both decoders are able to decode 32 slices of 38 symbols. The total area required by the fully parallel analog decoder is equal to 367 mm^2 while the corresponding semi-iterative decoder's area is only equal to 37 mm^2. Detailing the area of each block, one can make the following remarks. The input memory areas are the same, since each frame has to be stored before being decoded and since the semi-iterative memory is sized up to the largest frame length. Even though the semi-iterative input interleaver is reconfigurable, its area is so reduced that it does not appear on the graph. The reconfigurable temporal extrinsic interleaver has an estimated area of 1.2 mm^2 in the chosen 0.25-μm BiCMOS process. The total extrinsic interleaver area—including memories—is larger by 2.4 mm^2 when compared to the non-reconfigurable circuit but is divided by more than three when compared to a fully parallelized circuit. This is due to multiple slice code properties, as explained is Section V-C. Finally, the total SISO decoding area is reduced by 32 thanks to hardware reuse. The most area-consuming parts are now the input memory and the extrinsic interleaving part. The latter includes the multiplexed

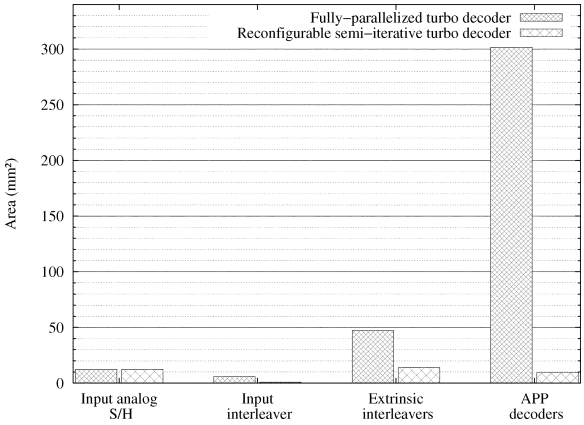


Fig. 15. On-chip area comparison of a fully parallelized analog turbo decoder and a reconfigurable semi-iterative decoder: the first is sized up to decode 1216 symbols whereas the second is able to cope with frame lengths ranging from 20 symbols up to 1216 symbols.

TABLE I
COMPARISON OF THE FULLY-PARALLEL, ITERATIVE AND SEMI-ITERATIVE ARCHITECTURES FOR 32 SLICES OF 38 SYMBOLS. ALL THE VALUES ARE NORMALIZED WITH RESPECT TO THE FULLY-PARALLEL FIGURES

Architecture	Fully-parallel	Iterative	Semi-iterative
Power consumption	1	31.2×10^{-3}	31.2×10^{-3}
Data rate	1	1.6×10^{-3}	3.1×10^{-3}
Energy / decoded bit	1	20	10
Area / decoded bit	1	0.08	0.1

T/Hs. Therefore, the design of the memories is now one of the keys to reducing the area even further.

VI. CONCLUSION

This paper reports an innovative semi-iterative algorithm and its corresponding architecture based on hardware reuse and the use of multiple-slice turbo codes [8]. The proposed algorithm breaks the one-to-one relation between code block length and size of the decoder when compared to the fully parallel solution. A single analog decoder chip able to cope with frame lengths up to a couple of thousand bits occupies only a few tens of mm^2. As shown in the paper, this is done with no correction performance degradation since the architecture retains a key feature of parallelized analog turbo decoding: continuous exchange of extrinsic information. It is used at each step of the decoding process to improve the convergence speed and quality. The proposed semi-iterative architecture is compared with the well-known iterative and fully parallel architectures. Table I sums up the characteristics of the three architectures extrapolated from [19]. The figures are normalized with respect to the fully parallel characteristics and are technology independent. Reducing the parallelism with a semi-iterative decoder yields a dramatic reduction in occupied on-chip area but with a smaller loss of data rate than with an iterative decoder. Thus, a semi-iterative decoder offers a compromise between on-chip area reduction and data rate. Moreover, a semi-iterative architecture can easily be made reconfigurable thanks to the multiple-slice interleaver properties. The design example of this paper shows that a single 37 mm^2 analog turbo decoder chip can treat any frame lengths ranging from 40 bits up to 2432 bits. This range includes the range of the frame lengths of the DVB-RCS standard. Indeed, this work provides answers to some of the main issues in designing an analog decoder pointed out in [22] pertaining to hardware reuse, re-configurability and large code block size.

Nevertheless, other issues remain to be solved. For instance, holding large frames of a few thousand bits means that the time between the first and the last sampled symbols becomes large. This will necessarily lead to a degradation in the hold values and hence affect the decoding. An increase in the hold capacitor of the sampling circuit is not a good solution since it will slow down the circuit. Therefore, after spending time on designing decoders one must spend time on designing the decoder's interfaces.

REFERENCES

[1] C. Winstead, "Analog Iterative Error Control Decoders," Ph.D. dissertation, University of Alberta, Edmonton, AB, Canada, 2005.
[2] M. Moerz, T. Gabara, R. Yan, and J. Hagenauer, "An analog 0.25-μm BiCMOS tailbiting MAP decoder," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 2000, pp. 356–357.
[3] F. Lustenberger, M. Helfenstein, H.-A. Loeliger, F. Tarköy, and G. S. Moschytz, "An analog VLSI decoding technique for digital codes," in *Proc. IEEE Int. Symp. Circuits Syst.*, 1999, vol. 2, pp. 424–427.
[4] M. Moerz, "Analog sliding window decoder core for mixed signal turbo decoder," in *Proc. Int. ITG Conf. Source and Channel Coding*, Erlangen, Germany, Jan. 2004, pp. 63–70.
[5] V. Gaudet, "Architecture and Implementation of Analog Iterative Decoders," Ph.D. dissertation, University of Toronto, Toronto, ON, Canada, 2003.
[6] M. Arzel, C. Lahuec, F. Seguin, D. Gnaedig, and M. Jézéquel, "Analog slice turbo decoding," in *Proc. IEEE Int. Symp. Circuits Syst.*, Kobe, Japan, May 2005.
[7] D. Gnaedig, E. Boutillon, M. Jézéquel, V. Gaudet, and P. Gulak, "On multiple slice turbo codes," *Annales des Télécommun.*, vol. 60, no. 1–2, Jan. 2005.
[8] D. Gnaedig, E. Boutillon, M. Jézéquel, V. Gaudet, and P. Gulak, "On multiple slice turbo codes," in *Proc. 3rd Int. Symp. Turbo Codes and Related Topics*, Brest, France, Sep. 2003, pp. 343–346.
[9] J. B. Anderson and S. M. Hladik, "Tailbiting MAP decoders," *IEEE J. Select. Areas Commun.*, vol. 16, no. 2, pp. 297–302, Feb. 1998.
[10] H.-A. Loeliger, F. Lustenberger, M. Helfenstein, and F. Tarköy, "Probability propagation and decoding in analog VLSI," in *Proc. 1998 IEEE Int. Symp. Inf. Theory*, Aug. 16–21, 1998, p. 146.
[11] J. Hagenauer and M. Winklhofer, "The analog decoder," in *Proc. 1998 IEEE Int. Symp. Inf. Theory*, Aug. 16–21, 1998, p. 145.
[12] C. Winstead, J. Dai, S. Yu, C. Meyers, R. Harrison, and C. Schlegel, *CMOS Analog MAP Decoder for (8,4) Hamming Code*, vol. 39, no. 1, pp. 122–131, Jan. 2004.
[13] A. G. i. Amati, S. Benedetto, G. Montorsi, D. Vogrig, A. Neviani, and A. Gerosa, "An analog turbo decoder for the UMTS standard," in *Proc. IEEE Int. Symp. Inf. Theory*, Chicago, IL, Jun. 2004, p. 296.
[14] A. Mondragon-Torres, E. Sanchez-Sinencio, and K. Narayanan, "Floating-gate analog implementation of the additive soft-input soft-output decoding algorithm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 50, no. 10, pp. 1256–1269, Oct. 2003.
[15] B. Gilbert, "A precise four-quadrant multiplier with subnanosecond response," *IEEE, J. Solid State Circuits*, vol. SC-3, no. 6, pp. 353–372, Dec. 1968.
[16] M. Arzel, C. Lahuec, M. Jézéquel, and F. Seguin, "Analogue decoding of duo-binary codes," in *Proc. Int. Symp. Inf. Theory and its Applications*, Parma, Italy, Oct. 2004.
[17] M. Moerz, "Quantization of soft-information in analogue mixed-signal turbo decoding," in *Proc. 4th Analog Decoding Workshop*, Zurich, Switzerland, Jun. 2005.
[18] S. Hemati and A. Banihashemi, "Comparison between continuous-time asynchronous and discrete-time synchronous iterative decoding," in *Proc. IEEE Global Telecommun. Conf.*, Nov.-Dec. 29–3, 2004, vol. 1, pp. 356–360.
[19] C. Lahuec, G. L. Mestre, M. Arzel, F. Seguin, and M. Jézéquel, "Design and test of a 0.25-μm BiCMOS double-binary analogue APP decoder," in *Proc. 5th Analog Decoding Workshop*, Jun. 2006.

[20] "Digital video broadcasting: Interaction channel for satellite distribution systems," ETSI, Cedex, France, EN 301 790, 2002.
[21] V. C. Gaudet, R. J. Gaudet, and P. G. Gulak, "Programmable interleaver design for analog iterative decoders," *IEEE Trans. Circuits Syst II, Analog Digit. Signal Process.*, vol. 49, no. 7, pp. 457–464, Jul. 2002.
[22] D. Vogrig, A. Gerosa, A. Neviani, A. G. i. Amat, G. Montorsi, and S. Benedetto, *A 0.35-μm CMOS Analog Turbo Decoder for the 40-Bit Rate 1/3 UMTS Channel Code*, vol. 40, no. 3, pp. 753–762, Mar. 2005.

Matthieu Arzel was born in Brest, France, in 1978. He received the Engineering Diploma and the Ph.D degree from the École Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France, in 2002 and 2006, respectively.

In 2006, he was with Turboconcept as a Research Engineer. He joined the Electronic Engineering Department of ENST de Bretagne as a Full-Time Lecturer in 2006. His research interests are in iterative decoding techniques, analog/mixed integrated circuit achitectures and design.

Cyril Lahuec was born in Orléans, France, in 1972. He received the B.Sc (Hon.) degree from the University of Central Lancashire, Lancashire, U.K., in 1993, the M.Eng and Ph.D. degrees from Cork Institute of Technologies, Cork, Ireland, in 1999 and 2002, respectively.

He was with Parthus Technologies (now Ceva), Cork, for his Ph.D work and then as a Consultant. He joined the Electronic Engineering Department of École Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France, as a Full-Time Lecturer in 2002. His research interests are in frequency synthesis, analogue integrated circuit design, and channel decoding.

Fabrice Seguin was born in Talence, France, in 1973. He received the Ph.D. degree from the Universite Bordeaux 1, Bordeaux, France, in 2001.

His doctoral research concerned the current-mode design of high-speed current conveyors and applications in RF circuits. In 2002, he joined the Electronic Engineering Department of École Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France, as a Full-Time Lecturer. He is currently involved with design issues of analog channel decoders and related topics.

David Gnaedig was born in Altkirch, France, in 1978. He received the Engineering Diploma from the École Nationale Supérieure des Télécommunications (ENST) de Paris, Paris, France, in 2001, and the Ph.D. degree from the Université de Bretagne Sud, Lorient, France, and the École Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France, in 2005.

In 2002, he joined TurboConcept, Plouzané, France, while working toward his doctoral degree. His research interests are in high throughput iterative decoding techniques for turbo and low-density parity check (LDPC) decoders and especially in joint code/architecture design. He is currently working with TurboConcept as a Research Engineer.

Michel Jézéquel (M'02) was born in Saint Renan, France, on February 26, 1960. He received the degree of "Ingénieur" in electronics from the École Nationale Supérieure de l'Électronique et de ses Applications, Paris, France in 1982.

In the period 1983–1986 he was a Design Engineer at CIT ALCATEL, Lannion, France. Then, after an experience in a small company, he followed a one year course about software design. In 1988, he joined the École Nationale Supérieure des Télécommunications de Bretagne, where he is currently Professor, head of the Electronics Department. His main research interest is circuit design for digital communications. He focuses his activities in the fields of Turbo codes, adaptation of the turbo principle to iterative correction of intersymbol interference, the design of interleavers and the interaction between modulation and error correcting codes.

9.2 Article IEEE TSC sur le décodage stochastique de turbocodes [Don+10]

Stochastic Decoding of Turbo Codes

Quang Trung Dong, Matthieu Arzel, Christophe Jego, and Warren J. Gross

***Abstract*—Stochastic computation is a technique in which operations on probabilities are performed on random bit streams. Stochastic decoding of forward error-correction (FEC) codes is inspired by this technique. This paper extends the application of the stochastic decoding approach to the families of convolutional codes and turbo codes. It demonstrates that stochastic computation is a promising solution to improve the data throughput of turbo decoders with very simple implementations. Stochastic fully-parallel turbo decoders are shown to achieve the error correction performance of conventional *a posteriori* probability (APP) decoders. To our knowledge, this is the first stochastic turbo decoder which decodes a state-of-the-art turbo code. Additionally, an innovative systematic technique is proposed to cope with stochastic additions, responsible for the throughput bottleneck.**

***Index Terms*—Iterative decoding, stochastic decoding, turbo codes.**

I. Introduction

Iterative soft-input soft-output (SISO) decoding was first presented by Berrou *et al.* in 1993 [1] for the turbo decoding of two parallel concatenated convolutional codes, widely known as turbo codes. Since their invention, turbo codes have received considerable attention due to their performance close to the theoretical limits. They are especially attractive for mobile communication systems and have been adopted as part of several channel coding standards for high data rates such as UMTS and CDMA2000 (third-generation) or 3GPP-LTE (the last step toward the fourth generation). The general concept of iterative SISO decoding has been extended to other families of error-correcting codes such as product codes. It also prompted the rediscovery of low-density parity-check (LDPC) codes. After many years of research, many decoding algorithms, decoder architectures and circuits were proposed. Although the industrial products were digitally designed, Hagenauer [2] and Loeliger [3] simultaneously proposed to apply the SISO concept to a continuous-time continuous-value decoding scheme with analog circuits to provide high decoding speeds and/or low power consumptions with extremely simple computation units working in parallel. In 2003, Gaudet, a member of the analog decoding community, and Rapley proposed a novel approach [4] based on stochastic computation.

Principles of stochastic computation were described during the 1960s by Gaines [5] and Poppelbaum *et al.* [6] as a method to carry out complex operations with a low hardware complexity. The main feature of this method is that the probabilities are converted into streams of stochastic bits using Bernoulli sequences, in which the information is given by the statistics of the bit streams. As a result, complex arithmetic operations on probabilities such as multiplication and division are transformed into operations on bits using elementary logic gates. This advantage allows architectures to be designed with low computational complexity and enables high data rates to be achieved.

Stochastic computations have been recently considered to decode FEC codes. Early stochastic decoding has been applied to some short error correcting codes such as the (7,4) Hamming code [4] and a (256,121) block turbo code based on two (16,11) Hamming codes [7]. The first implementation of a stochastic decoder with a (16,8) LDPC code was described in [8]. An improved stochastic decoding approach was then proposed to decode practical LDPC codes [9], [10]. This approach was also extended to well-known linear block codes with high-density parity-check matrices, namely BCH codes, Reed Solomon codes and product codes [11]. When compared with conventional Sum-Product implementations, stochastic decoding could provide near-optimal performance for practical LDPC codes. The potential of the stochastic technique for low complexity and high throughput was recently demonstrated by the FPGA implementation of a (1056,528) LDPC decoder [12] which achieved a throughput of 1.66 Gb/s. Thus, state-of-the-art decoders combine high throughput and low complexity thanks to the stochastic approach.

This paper proposes to extend stochastic computation to the design of turbo decoders. A major challenge in the implementation of turbo decoders is to achieve high-throughput decoding. Indeed, the next generations of mobile communication systems will require data rates of 1 Gb/s and beyond. Thanks to stochastic decoding, a fully-parallel architecture is a promising response to this challenge. In order to provide a typical study case, the investigation is limited to a single-binary turbo code similar to the ones adopted for the next generation of mobile systems (3GPP-LTE).

This paper is organized as follows. Section II provides a brief overview of the turbo codes, the APP algorithm and the principles of stochastic computation. Section III describes the APP-based stochastic processing applied to the iterative decoding of practical turbo codes. Section IV introduces a method to increase the stochastic decoding throughput. Some simulation results are given in Section V to compare the stochastic processing with a conventional decoding using the *a posteriori* probability algorithm.

II. Background

A. Turbo Codes

Fig. 1(a) shows the structure of a turbo encoder made up of two tail-biting recursive systematic convolutional (RSC) encoders concatenated in parallel thanks to an interleaver. Each RSC code has a coding rate $R = 1/2$, a codeword length $n = 2k$ and a constraint length $\nu = 4$. It can be represented by means of a trellis diagram as shown in Fig. 1(b). The overall code rate of the turbo code is $R = 1/3$. At each time i, the information bit (or systematic bit) d^i and two redundancies (or parity bits) y_1^i and y_2^i corresponding to the contributions of each RSC code are provided by the encoder.

The architecture of the turbo decoder illustrated in Fig. 1(c) is composed of two SISO decoders that exchange some probabilities thanks to an interleaver (Π) and a de-interleaver (Π^{-1}). Each SISO decoder is fed with three different inputs: the channel output corresponding to the systematic bit (u^i), the parity bit produced by the corresponding component encoder (v_1^i or v_2^i), and the extrinsic probabilities computed by the other component decoder. The iterative exchange of extrinsic probabilities between the SISO decoders greatly improves the error correction performance.

Manuscript received March 05, 2010; accepted August 09, 2010. Date of publication September 02, 2010; date of current version November 17, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tong Zhang.

Q. T. Dong and M. Arzel are with the Institut Telecom/Telecom Bretagne, CNRS Lab-STICC UMR 3192, Technopôle Brest-Iroise F-29238 Brest Cedex 3, France (e-mail: qt.dong@telecom-bretagne.eu; matthieu.arzel@telecom-bretagne.eu).

C. Jego is with the CNRS IMS, UMR 5218 351, Cours de la Libération, F-33405 Talence Cedex (e-mail: christophe.jego@ims-bordeaux.fr).

W. J. Gross is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada (e-mail: wjgross@ece.mcgill.ca).

Digital Object Identifier 10.1109/TSP.2010.2072924

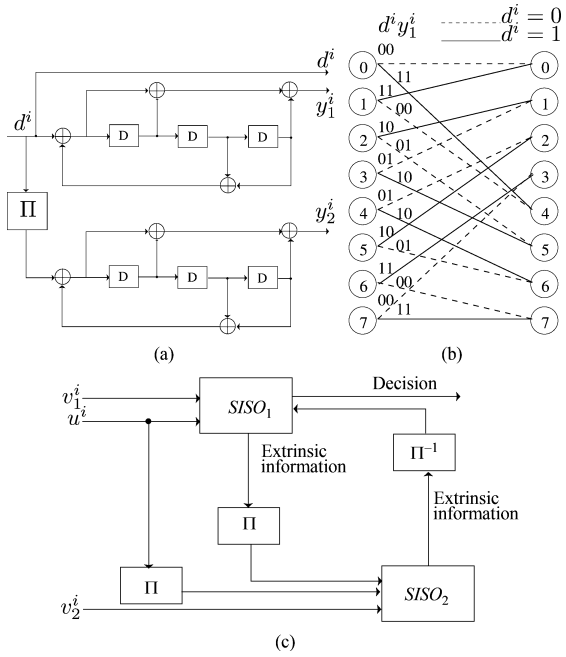


Fig. 1. (a) Turbo encoder; (b) trellis diagram; and (c) turbo decoder architecture.

B. SISO Decoding Algorithm

In order to decode convolutional codes, an algorithm known as BCJR was introduced by Bahl *et al.* [13]. It was adapted by Anderson and Hladik to deal with tail-biting codes [14]. The APP decoding process performed by each SISO component decoder can be summarized by the following steps.

1) Branch Metric Computation: First, the branch metrics $\gamma^i(s', s)$ can be expressed as

$$\gamma^i\left(s', s\right) = \Pr{}^a\left(d^i = j\right) \Pr{}^{ex}_{\text{in}}\left(d^i = j|u, v_2\right) \Pr\left(u^i, v_1^i|d^i, y^i\right) \tag{1}$$

where d^i is the information bit for the transition from state s' to state s of the trellis at time i. $\Pr^a(d^i = j)$ is the *a priori* probability corresponding to the transition $d^i = j$. If a uniform source is considered, all the symbols have the same probability during the transmission, then $\Pr^a(d^i = j) = 1/2$. $\Pr^{ex}_{\text{in}}\left(d^i = j|u, v_2\right)$ is the incoming extrinsic probability computed by the other component decoder. It is calculated from the input sequences u and v_2. In the case of an Additive White Gaussian Noise (AWGN) channel, the third factor is given by

$$\Pr\left(u^i, v_1^i|d^i, y^i\right) = \exp\left(\frac{\left\langle u^i, d^i\right\rangle + \left\langle v_1^i, y_1^i\right\rangle}{\sigma^2}\right) \tag{2}$$

where σ^2 is the variance of the AWGN and $\langle a, b\rangle$ represents the scalar product of two symbols a and b.

2) State Metric Computation: Second, the forward and backward metrics are recursively calculated as follows:

$$\alpha^{i+1}(s) = \sum_{s'=0}^{2^{\nu-1}-1} \alpha^i\left(s'\right)\gamma^i\left(s', s\right) \tag{3}$$

$$\beta^i\left(s'\right) = \sum_{s=0}^{2^{\nu-1}-1} \beta^{i+1}(s)\,\gamma^i\left(s, s'\right). \tag{4}$$

The state metric values are initialized at the same probability, i.e., $\alpha^0(s) = \beta^k(s) = 1/2^{\nu-1}$ for any $s \in [0 \ldots 2^{\nu-1} - 1]$. During the decoding process, the state metrics have to be kept in a given range and therefore are normalized regularly.

3) Extrinsic Probability Computation: Third, in the context of an iterative process, the component decoders exchange extrinsic probabilities calculated as

$$\Pr{}^{ex}_{out}\left(d^i = j|u, v_1\right) = \frac{\sum_{(s',s)/d^i(s',s)=j} \phi_e^i(s', s)}{\sum_{(s',s)} \phi_e^i(s', s)} \tag{5}$$

where

$$\phi_e^i\left(s', s\right) = \alpha^i\left(s'\right)\beta^{i+1}(s)\,\gamma_e^i\left(s', s\right) \tag{6}$$

$$\gamma_e^i(s', s) = \exp\left(\frac{\left\langle v_1^i, y_1^i\right\rangle}{\sigma^2}\right). \tag{7}$$

4) A Posteriori Probability Computation: Finally, *a posteriori* probabilities are computed so that

$$\Pr\left(d_i = j|u, v_1\right) = \sum_{(s',s)/d^i(s',s)=j} \phi^i\left(s', s\right) \tag{8}$$

where

$$\phi^i\left(s', s\right) = \alpha^i\left(s'\right)\beta^{i+1}(s)\,\gamma^i\left(s', s\right). \tag{9}$$

The decoded symbol $\hat{d}^i$ at time i is equal to the value j that maximizes this *a posteriori* probability.

A suboptimal version in the logarithmic domain with an acceptable loss of performance referred to as Max-Log-MAP (or Sub-MAP) algorithm was introduced by Robertson *et al.* [15].

C. Stochastic Decoding Principles

1) Stochastic Computation: In a stochastic computing process, the probabilities are converted into Bernoulli sequences using random number generators and comparators [5]. The number of bits at "1" in a stream represents the corresponding probability. For instance, a 10-bit sequence with 4 bits equal to "1" represents a probability of 0.4. Therefore, different stochastic streams may represent the same probability. In order to obtain a good precision, the length of a sequence has to be large. The conventional arithmetic operations, such as multiplication or division, are thus processed by simple logic gates. For instance, the multiplication of a set of N probabilities $p_0, p_1, \ldots, p_{N-1}$ can be achieved by an N-input AND logic gate fed with N mutually independent stochastic streams. The output probability of the AND logic gate is exactly equal to $\prod_{i=0}^{N-1} p_i$. At each time, the bit of each input sequence contributes directly to the output bit. Similarly, the normalization of two Bernoulli sequences is carried out by means of JK flip-flops [4].

2) The Thorny Addition: From the equations of the APP algorithm, it can be noted that besides the multiplication and division operations, a huge number of additions is necessary. Since the addition of N values in the interval [0, 1] may take values bigger than 1, this operation cannot

TABLE I
COMPLEXITY OF ONE SECTION OF A STOCHASTIC SINGLE-BINARY 8-STATE TURBO DECODER WITH MULTIPLEXERS FOR ADDITIONS

Module	Elementary hardware resources								Random bits	
	NAND2	AND2	OR2	XOR2	Mux2:1	Mux8:1	3-bit counter	D Flip-flop	7 bits	1 bit
Γ		34	12						2	
A / B		32	8	8	24	8		256		96
Ext		32	2	2	4	2		64		16
Dec		37	2			2	2			6
Total		167	32	18	52	20	2	576	2	214

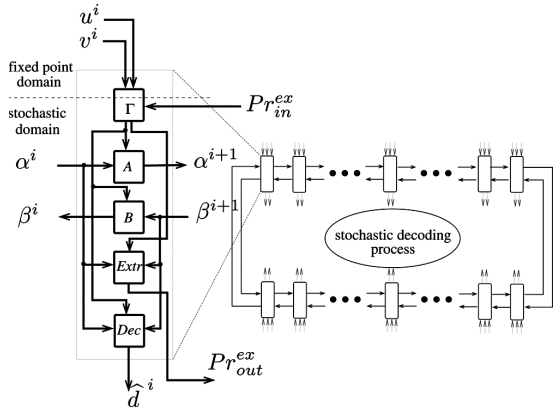


Fig. 2. Stochastic tail-biting APP decoder.

be done directly with stochastic streams. The addition operands can be scaled equally so that the sum always lays in the interval $[0, 1]$. In practice, a multiplexer that randomly selects one of the N inputs with probability $1/N$ will produce an output stream that is the scaled sum of the input probabilities $\sum_{i=0}^{N-1} 1/N p_i$. At each time, each input bit does not contribute directly to the output bit. Consequently, the output sequence length has to be about N times larger than the input sequence lengths to achieve the same precision. This constraint is particularly problematic for the APP-based decoding process. Indeed, many additions are necessary to normalize the state metrics and to compute the extrinsic probabilities. Thus, processing addition operations with multiplexers severely slows down the decoding convergence speed of a turbo decoder.

III. STOCHASTIC DECODING APPLIED TO TURBO CODES

A. SISO Component Decoder Architecture

The stochastic decoding of turbo codes requires the stochastic computation to be applied to a tail-biting APP algorithm, which relies on the trellis representation. Fig. 2 details the exchange of information between the various sections of a tail-biting APP decoder. There are as many sections as symbols to decode and each section is made up of four modules. A Γ module is fed by the channel outputs u^i and v^i, which are associated with the i^{th} transmitted symbol d^i and its parity bit y^i. This module converts u^i and v^i into *a priori* probabilities, represented by two stochastic streams to compute the branch metrics and then the forward metrics in an A module and the backward metrics in a B module. These modules are involved in a recursive process since they use the forward and backward metrics α^i and β^{i+1} from their neighbors and provide them α^{i+1} and β^i. A Dec module decides the final value of each binary symbol, $\hat{d}^i$ for the transmitted symbol d^i. A last module is also required if the APP decoder is part of a turbo decoder: the Ext module. This module computes the output extrinsic probability Pr_{out}^{ex} which is then used by a Γ module of the second APP decoder as the input Pr_{in}^{ex}. All the modules exchange stochastic streams over a logic gate network based on the code trellis representation. Each stochastic decoding step is referred to as a *decoding cycle* (DC) and corresponds to the output of one new bit for each stochastic unit. The decoding process terminates when a maximum number of DCs is reached.

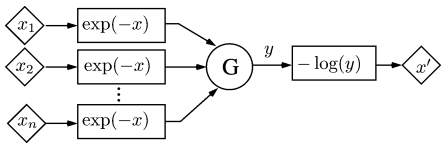


Fig. 3. Principle of exponential-domain computation.

B. Hardware Complexity

One major problem in stochastic decoding that deeply degrades the decoding performance is known as the *latching problem* [7]. It is related to the sensitivity to the level of random switching activity (bit transition) [16]. This problem can be easily observed at high signal-to-noise ratios (SNRs). Different solutions have been suggested to solve the latching problem, and thus, to improve the BER performance of stochastic decoding, such as using supernodes [7], scaling the received log-likelihood ratios (LLRs) up to a maximum value [16], edge memories (EMs) insertion and noise-dependent scaling (NDS) [12]. The APP decoders proposed in this paper take advantage of EMs and NDS. In particular, EMs are assigned to stochastic streams that represent forward and backward metric values α^i and β^i to break the correlation using re-randomization. Similarly, EMs are assigned to stochastic streams used for the output extrinsic computation in the module Ext. Overall, ten 32-bit EMs are necessary for each section of the stochastic SISO decoder to circumvent the *latching problem*. The complexity of one section of the stochastic decoder in terms of elementary hardware resources and random bits is detailed in Table I. Vectors of 7 random bits are used by the stochastic SISO decoder to convert the channel outputs into stochastic streams. As already mentioned in this paper, the main drawback of this architecture is the need of N-to-1 multiplexers to perform additions. For this reason, solutions have to be investigated to replace these large multiplexers.

IV. STOCHASTIC ADDITION IN THE EXPONENTIAL COMPUTATION DOMAIN

In order to remove N-to-1 multiplexers for stochastic addition operations, a novel approach is proposed. The main idea is to carry out a critical operation $\mathbf{F}$ in the exponential domain thanks to the $\exp(-x)$ function. The output values of $\exp(-x)$ modules are then processed by using a simple operation $\mathbf{G}$. Then, the result is converted back into a probability thanks to the function $-\log(x)$ as illustrated in Fig. 3. If $\mathbf{F}$ is the addition operation, then $\mathbf{G}$ is the multiplication operation, processed by an AND logic gate. Therefore, no large multiplexer is required to perform the stochastic addition operation.

TABLE II
COMPLEXITY OF ONE SECTION OF A STOCHASTIC SINGLE-BINARY 8-STATE TURBO DECODER WITH ADDITIONS IN THE EXPONENTIAL DOMAIN

Module	Elementary hardware resources								Random bits	
	NAND2	AND2	OR2	XOR2	Mux2:1	Mux8:1	3-bit counter	D Flip-flop	7 bits	1 bit
Γ		34	12						2	
A / B	48	120	8	8	32			288		88
Ext	34	52	2	2	6			82		16
Dec	32	46		1			1	16		16
Total	162	372	30	19	70		1	674	2	208

A. Exponential and Logarithmic Transformations

The idea of processing stochastic streams in the exponential domain was first introduced by *Janer et al.* [17]. The $\exp(-x)$ function is chosen instead of $\exp(x)$ so that the output value can be represented by stochastic streams. In practice, the exponential function can be easily approximated by the first terms of its Taylor's expansion. In [17], the authors described some circuits for the first-, second- and third-order approximations. They also demonstrated that the accuracy of this approximation does not depend on the number of input probabilities that are being added. Therefore, this stochastic exponential transformation opens an efficient way to carry out the conversion of stochastic additions into stochastic multiplications.

In [17], the result in the exponential domain was sufficient to end the data processing. Unfortunately in a turbo decoder architecture, the result of the addition operation has to be used by another module. Thus, the exponential stochastic stream has to be converted back into a conventional stochastic stream that corresponds to the addition of n terms. A logarithm function is necessary to perform this transformation. A Taylor's expansion is also considered in this case.

B. Hardware Complexity

Table II gives a summary of the complexity of one section of the stochastic single-binary 8-state turbo decoder with additions in the exponential domain. Expanding the Taylor series to the second order is sufficient for both exponential and logarithmic modules. The additional cost of addition operations in the exponential domain in terms of hardware resources is reasonable. Indeed, 162 NAND2 logic gates, 205 AND2 logic gates, 98 D Flip-flops and eighteen 2-to-1 multiplexers are necessary to replace the twenty 8-to-1 multiplexers used for addition operations in the probability domain. The hardware complexity of one section of a stochastic single-binary 8-state turbo decoder has to be compared with an fixed-point Sub-MAP counterpart. For such a SISO decoder, the received symbols are 5-bit quantized while the extrinsic information and state metrics are both 7-bit quantized to achieve almost ideal performance [18]. A conventional Sub-MAP decoder is composed of three main parts, namely processing, memory and control. The major problem of the turbo decoders is the memory bottleneck. In order to reduce the state metric memory size, the sliding window principle can be applied, where each received frame has to be divided into several sliding windows. Such a sub-block processing is constrained by the sliding window initialization. To solve this constraint, additional costs in terms of resources and/or latency have to be considered. For a stochastic decoder, a randomization engine is necessary for providing random bits. These random bits are used in 2-to-1 multiplexers and as the addresses of stochastic stream generators. Although this amount of random bits for one section might seem large, as shown in Table II, random bits can be significantly shared by different modules without having an impact in terms of BER performance [12]. Moreover, random number generators using unreliable device behavior have to be considered since they require less hardware resources than conventional linear feedback shift registers. It means that a direct comparison between the two decoding techniques can only be done for the processing unit of one section. The FPGA implementation cost of a fixed-point Sub-MAP decoder must be compared with the results given in Table II. One LUT is allocated for each elementary hardware resources of the Table II. In this case, 633 and 638 LUTs are necessary for the fixed-point Sub-MAP and the stochastic versions, respectively. In contrast, the stochastic decoding of one section is less costly in terms of flip-flops. Indeed, the flip-flop number can be decreased from 1398 down to 680 if a stochastic decoder is considered. It means that stochastic decoding is competitive in terms of hardware complexity for turbo codes.

V. SIMULATION RESULTS

In this section, the decoding performance is given for different versions of stochastic decoders for both convolutional and turbo codes. Fig. 4(a) shows the BER performance of the stochastic decoding of a tail-biting RSC code ($n = 400$ bits, code rate $R = 1/2$) with 30 K DCs and different optimizations. A decoder combining NDS and EMs provides a BER performance similar to the one of a conventional APP floating-point algorithm. Moreover, processing additions in the exponential domain enables a decrease of the number of DCs from 30 K to 4 K with an acceptable performance loss of 0.1 dB when compared with the floating-point APP algorithm. Similar conclusions are obtained with a 4000-bit RSC code as shown in Fig. 4(b). Thus, the extension of the stochastic decoding to convolutional codes is possible. The BER performance of the proposed stochastic decoding method is provided also for a ($n = 600$, $R = 1/3$) turbo code in Fig. 4(c). The turbo code is designed with an S-Random interleaver [19]. The EM and NDS techniques are required to achieve good decoding performance. Stochastic turbo decoding needs 250 K DCs to achieve the performance of the floating-point Sub-MAP decoding with six iterations. Fortunately, the exponential stochastic approach proposed in this paper enables the number of DCs to be reduced from 250 to 32 K without any performance degradation. Thus, the proposed summation is a necessary step toward the implementation of high-speed stochastic turbo decoders. To compete with state-of-the-art turbo decoders, a stochastic decoder requires a higher level of parallelism. Two ways have to be explored. First, parallel processing of larger frames of a few thousands of bits—as in wireless communications standards—would be of major interest. Second, representing any probability with p parallel independent streams could divide the number of DCs by p and multiply the throughput by p. Naturally, a higher parallelism will impact the decoder complexity, which is the price to pay for high throughput devices.

VI. CONCLUSION

This paper extends the application of the stochastic decoding to the families of convolutional codes and turbo codes. Simulation results show performance close to the floating-point Sub-MAP decoding algorithm for ($n = 600$, $R = 1/3$) turbo codes. One major problem of a conventional stochastic decoding of turbo codes is the large number of decoding cycles. To reduce the number of cycles, a novel technique for implementing the stochastic addition operation has been investigated. It consists of transforming the stochastic additions into stochastic multiplications in the exponential domain. The number of decoding cycles is thus considerably reduced with no performance degradation. The

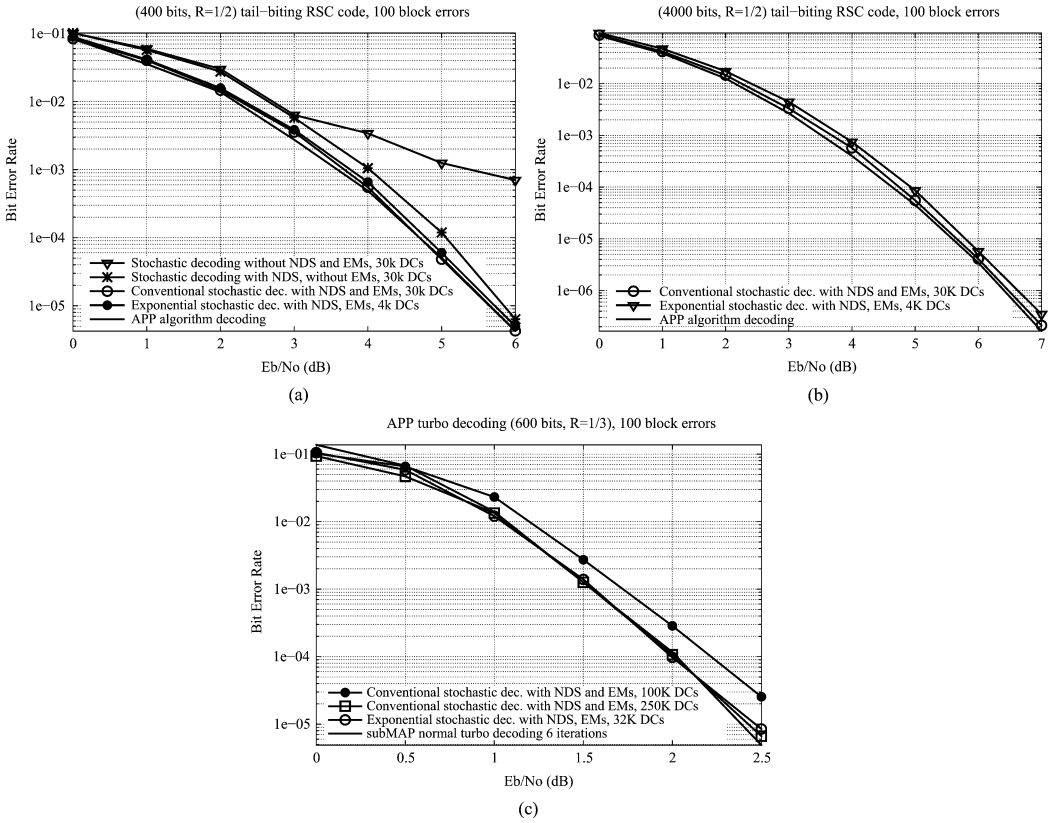


Fig. 4. Performance of the stochastic decoding of (a) a rate-1/2 convolutional code for codewords of 400 bits and (b) 4000 bits and of (c) a rate-1/3 turbo code for codewords of 600 bits .

results provided in this paper validate the potential of stochastic decoding as a practical approach for high-throughput turbo decoders and encourage to keep on investigating in this way.

REFERENCES

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *IEEE Int. Conf. Commun. (ICC) Conf. Rec.*, Geneva, Switzerland, May 1993, vol. 2, pp. 1064–1070.

[2] J. Hagenauer and M. Winklhofer, "The analog decoder," in *Proc. 1998 IEEE Int. Symp. Information Theory*, Aug. 16–21, 1998, p. 145.

[3] H.-A. Loeliger, F. Lustenberger, M. Helfenstein, and F. Tarköy, "Probability propagation and decoding in analog VLSI," in *Proc. 1998 IEEE Int. Symp. Information Theory*, Aug. 16–21, 1998, p. 146.

[4] V. Gaudet and A. Rapley, "Iterative decoding using stochastic computation," *Electron. Lett.*, vol. 39, no. 3, pp. 299–301, Feb. 2003.

[5] B. Gaines, "Stochastic computing," *AFIPS SJCC*, no. 30, pp. 149–156, 1967.

[6] W. Poppelbaum, C. Afuso, and J. Esch, "Stochastic computing elements and systems," *Amer. Fed. Inf. Process. Societies Fall Joint Comput. Conf. (AFIPS FJCC)*, no. 31, pp. 635–644, 1967.

[7] C. Winstead, V. Gaudet, A. Rapley, and C. Schlegel, "Stochastic iterative decoders," in *Proc. Int. Symp. Information Theory (ISIT)*, Sep. 2005, pp. 1116–1120.

[8] W. Gross, V. Gaudet, and A. Milner, "Stochastic implementation of LDPC decoders," in *39th Asilomar Conf. Signals, Systems, Computers Conf. Rec.*, Nov. 1, 2005, pp. 713–717.

[9] S. S. Tehrani, W. Gross, and S. Mannor, "Stochastic decoding of LDPC codes," *IEEE Commun. Lett.*, vol. 10, no. 10, pp. 716–718, Oct. 2006.

[10] S. S. Tehrani, S. Mannor, and W. Gross, "Survey of stochastic computation on factor graphs," in *Proc. 37th Int. Symp. Multiple-Valued Logic (ISMVL)*, May 2007, pp. 54–59.

[11] S. S. Tehrani, C. Jego, B. Zhu, and W. Gross, "Stochastic decoding of linear block codes with high-density parity-check matrices," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5733–5739, Nov. 2008.

[12] S. S. Tehrani, S. Mannor, and W. Gross, "Fully parallel stochastic LDPC decoders," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5692–5703, Nov. 2008.

[13] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

[14] J. B. Anderson and S. M. Hladik, "Tailbiting MAP decoders," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 297–302, Feb. 1998.

[15] P. Robertson, E. Villebrun, and P. Hoeher, "A comparison of optimal and sub-optimal map decoding algorithms operating in the log domain," in *IEEE Int. Conf. Communications (ICC)*, Seattle, WA, Jun. 1995, vol. 2, pp. 1009–1013.

[16] C. Winstead, "Error-control decoders and probabilistic computation," presented at the 3rd SOIM-COE Conf., Tohoku Univ., Sendai, Japan, Oct. 2005.

[17] C. Janer, J. Quero, J. Ortega, and L. Franquelo, "Fully parallel stochastic computation architecture," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2110–2117, Aug. 1996.

[18] G. Montorsi and S. Benedetto, "Design of fixed-point iterative decoders for concatenated codes with interleavers," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 5, pp. 871–882, May 2001.

[19] S. Dolinar, D. Divsalar, and F. Pollara, "Weight distributions for turbo codes using random and non-random permutations," TDA Progress Rep. 42-122, 1995.

9.3 Article IEEE/OSA Journal of Lightwave Technology sur la technique d'OFDM précodé par DFT pour les prochaines générations de PON [TRU+14a]

DFT Precoded OFDM—An Alternative Candidate for Next Generation PONs

Tuan-Anh Truong, Matthieu Arzel, Hao Lin, Bruno Jahan, and Michel Jézéquel, *Member, IEEE*

***Abstract*—Recently, the orthogonal frequency division multiplexing (OFDM) technique has been extensively studied for fiber-based optical transmissions in the context of access networks. The adaptive modulation optical (AMO) OFDM system has been proved to be one of the cost-effective solutions. Among adaptive modulation techniques, the Levin–Campello (LC) bit/power loading, which is widely implemented for xDSL systems, is shown to bring excellent performances in unamplified optical intensity modulated/direct detected (IMDD) transmissions. In this paper a novel adaptive discrete Fourier transform precoded OFDM (POFDM) is proposed and investigated. By means of numerical simulation, the proposed modulation scheme is shown to reduce the peak-to-average power ratio of the transmitted OFDM signal up to 2 dB at 10^{-3} clipping rate. As a consequence, a smaller input back-off can be applied to the laser-driving current before the power amplifier, resulting in a lower power consumption of the POFDM system when compared to the conventional LC AMOOFDM system. According to simulation results, the energy consumption of the power amplifier is reduced by a factor of two in unamplified optical IMDD transmissions. Moreover, in terms of capacity-versus-reach performance, the proposed system achieves the same performance as the conventional adaptive modulation scheme.**

***Index Terms*—Discrete Fourier transform (DFT), intensity modulation direct detection, optical fiber communication, orthogonal frequency division multiplexing (OFDM), power amplifiers (PAs).**

I. Introduction

DUE to the ever growing data traffic demand of applications for access networks, extensive research has been carried out to look for cost-effective, high-speed (>10 Gbps) optical transmission systems. In the context of access networks, the system costs including the installation, operation and maintenance costs are always the first concerns of a system designer. In selecting the optical modulator, direct modulation of laser's intensity is proved to be a great benefice since a directly-modulated laser (DML) is far cheaper than an external modulator. Moreover, single-mode-fiber (SMF) links are widely installed all around the world and the actual optical transmission standards are extensively based on intensity-modulated direct detection (IMDD) technique. Hence, transmission systems exploiting SMF links and IMDD technique are strong candidates for the future generation of access and metropolitan networks.

In order to exploit this cost-effective solution, one of the interesting technical approaches is to compensate all the detriments of an optical transmission by means of advanced modulation schemes and digital signal processing. In [1] adaptive modulation optical orthogonal frequency division multiplexing (AMOOFDM) system is shown to be an excellent cost-effective solution for short-reach and metropolitan optical transmissions since the OFDM modulation has a high spectral efficiency and the chromatic dispersion of the fiber link can be easily compensated with simple frequency-domain equalization and adaptive modulation. Moreover, the OFDM modulation makes the transmission system very flexible and robust against the chromatic dispersion thanks to the parallel transmission of narrowband subcarriers. The discrete Levin–Campello (LC) bit/power loading technique [2], which is shown to bring a quasi-optimal solution of water-filling technique and is widely implemented in xDSL systems, has also been shown to give excellent performance in optical IMDD transmissions [3], [4]. In [4] a 12.5 Gbps AMOOFDM transmission is proved to be possible over more than 20-km SMF link with low-cost DFB and VCSEL laser intensity modulators and a 2.5 GHz band-width PIN-based photodetector. These results show a great potential of such a modulation scheme for future generations of passive optical networks (PONs).

However, the OFDM modulation has its own drawbacks. And one of the important disadvantages of such a modulation technique is an important peak-to-average power ratio (PAPR). That is, the OFDM signal has large maximum power peak when compared to its average power. Due to the important PAPR, the efficiency of power amplifiers (PAs) in OFDM transmissions might be very low [5]. Moreover, large power peaks of the OFDM signal may also saturate the laser, resulting in in-band noise which degrades the signal to noise ratio (SNR) and out-of-band radiations which cause inter-channel interference. Recently, a multiband discrete Fourier transform (DFT)-spread OFDM system has been proposed for long-haul coherent optical (CO) transmissions [6]. The idea is to divide the whole OFDM band into several subbands then precode the data carried by the subcarriers on each subband with a DFT matrix. The technique is shown to bring good performance in terms of PAPR reduction and fiber nonlinearity mitigation. Sung *et al.* proposed later in [7] a new precoding technique which consists in coding separately the real and imaginary parts of the transmitted data symbols by a DFT matrix. The precoded complex symbols are then input to a conventional OFDM modulator.

Manuscript received June 13, 2013; revised November 4, 2013 and December 20, 2013; accepted January 12, 2014. Date of publication January 20, 2014; date of current version February 10, 2014. This work was supported by Orange Labs, France.

T.-A. Truong, H. Lin, and B. Jahan are with Orange Labs, 4 rue du Clos Courtel, 35510, Cesson Sevigne, France (e-mail: tuananh.truong@orange.com; hao.lin@orange.com; bruno.jahan@orange.com).

M. Arzel and M. Jézéquel are with the Electronic Engineering Department of the School of Telecommunication Telecom Bretagne, 29238 Brest, France (e-mail: matthieu.arzel@telecom-bretagne.eu; Michel.jezequel@telecom-bretagne.eu).

Digital Object Identifier 10.1109/JLT.2014.2301632

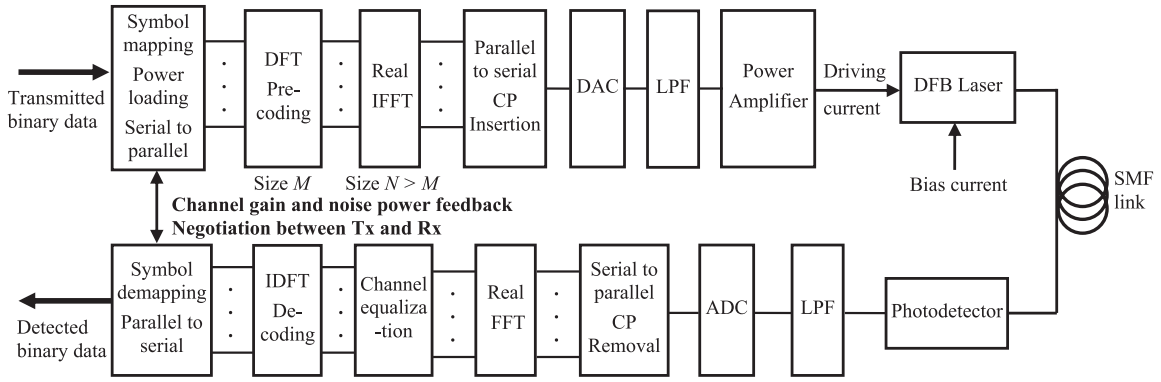


Fig. 1. Transmission link diagram of LC AMOOFDM and POFDM modems. The Precoding and the Decoding blocks are only used in POFDM modem. At the initialization stage, the estimated channel gain and noise power are sent back to the Tx to find the optimal bit/power distribution of the data subcarriers.

The new precoding technique is shown to bring better performance in terms of PAPR maintenance along the fiber and nonlinearity mitigation at the expense of an increased system complexity. DFT-Spread OFDM technique has also been used in wireless radio communications, and is already adopted by the next-generation 4G mobile standard for the uplink [8]. It is important to note that the precoding techniques proposed in [6] and [7] are only suitable for CO OFDM transmissions because in this context the chromatic dispersion results in only a phase rotation of the OFDM subcarriers. This means that at the receiver side a simple channel phase compensation can be used to equalize the received signal. The equalized OFDM symbols can then be decoded to detect the transmitted data. On the contrary, in the context of an optical IMDD transmission, the interplay between the laser chirps and the chromatic dispersion results in attenuation dips in the channel frequency response [9]. The noise power on deeply attenuated subcarriers might be increased after the channel equalizer. In DFT-Spread systems, the power-increased noise is also spread over different data subcarriers. This results in a degradation of the overall bit error rate (BER) performance. Hence, in the context of transmissions in PONs, conventional DFT-spread techniques cannot guarantee a good performance without an appropriate choice of null subcarriers. In order to solve this problem, we propose in this paper a rate adaptive DFT-precoded system which optimizes the transmission data rate under the constraint of the transmitted electrical power and the symbol error Rate (SER) in the context of optical IMDD transmissions. To the best of authors' knowledge, the performance of the DFT-spread technique has not yet been adequately analyzed in this context. And this is the first time a rate adaptive algorithm is proposed for optical IMDD DFT-spread OFDM transmissions, although the LC rate adaptive technique is well-known in conventional OFDM transmissions. In this paper the term POFDM, which stands for precoded OFDM, denotes the proposed system. According to numerical simulations, it is shown that the proposed POFDM system has the same performance in terms of data bit rate when compared to that of the conventional LC AMOOFDM system. However, the proposed transmission scheme gives a large power saving due to the PAPR reduction. This paper is organized as follows. In Section II, principal functions of a conventional LC AMOOFDM modem are summarized. The proposed POFDM modem is described in Section III. The theoretical basis of the proposed rate adaptive algorithm for POFDM system is also detailed in this section. Section IV presents the modeling of each component in the transmission link. Simulation parameters, which are extracted from commercially available component's datasheet or from experimental measurements, are also given. The optimum operating parameters of the DFB laser are also discussed in this section. The power saving due to the PAPR reduction of a POFDM system when compared to an LC AMOOFDM system is assessed in Section V. In Section VI, the performance in terms of capacity-versus-reach of the conventional LC AMOOFDM system and the proposed POFDM system in unamplified transmissions is given. Key capacity-limiting factors of an optical IMDD transmission are also identified and discussed. Finally, conclusion is drawn in Section VII.

II. Conventional LC AMOOFDM Modem

A DML-based SMF-channel IMDD transmission diagram is illustrated in Fig. 1. The precoding block at the transmitter and the decoding block at the receiver are only used in POFDM systems. By referring to Fig. 1, the general functions of an LC AMOOFDM modem can be described as follows. At the transmitter, the transmitted binary sequence is mapped into a complex quadrature amplitude modulation (QAM) sequence. A serial-to-parallel converter distributes the complex data to a large number of subcarriers to form frequency-domain OFDM symbols. According to the SNR on each subcarrier, the order of QAM constellation and the data power can differ from one subcarrier to another. An inverse fast Fourier transform (IFFT) is then applied to each frequency-domain OFDM symbol to generate real-valued time domain OFDM symbols. It is noted that the real-valued symbols are obtained by using the Hermitian symmetry, which means that the data carried by the negative frequency subcarriers are the complex conjugates of the data on the positive frequency subcarriers, and the DC subcarrier

carries no data. Here a cyclic prefix is inserted into each OFDM symbol to combat the chromatic dispersion of the fiber link. A parallel-to-serial converter serializes the IFFT-output symbols to form a digital LC AMOOFDM transmitted sequence. According to a pre-determined clipping rate, a digital to analog converter (DAC) performs a clipping of the digital signal to limit the power dynamic range. The digital clipped signal is then quantized on a given number of bits. A zero-order hold followed by a low pass filter is then implemented in order to generate an analog signal waveform, which is then biased with a DC component before driving a DFB laser. The peak-to-peak value of the analog laser-driving signal can be adjusted by means of a PA. It is shown later that the optimal driving current peak-to-peak value, which brings the best performance in terms of throughput, varies according to the transmission distance. Before that the optical signal is injected into the fiber link, its power is limited by an optical attenuator. At the receiver side, after a photodiode, the received signal is demodulated with a processing dual to the transmitter's one.

According to the principle of the LC AMOOFDM technique [4], the transmitter consists in finding the best bit and power combination for each subcarrier in order to maximize the overall transmission bit rate under a constraint of the laser-driving signal power and the SER. In order to calculate the bit and power combination, one of the input parameters of the LC algorithm is the SNR of each subcarrier [1]. At the initial stage of the transmission, this parameter can be obtained via noise estimation and feedbacks between the transmitter and the receiver. A certain number of OFDM training symbols, which are known by the receiver, can be sent so that the SNR on each subcarrier can be estimated.

III. The Proposed POFDM Modem

The transmission procedure of the proposed POFDM modem is similar to that of the LC AMOOFDM modem described previously except two important differences:

1) A same QAM constellation is used for all active subcarriers and the power distribution between different subcarriers is uniform. If a subcarrier is inactive, it carries null data. And whether a subcarrier carries null data or not is defined by a null-subcarrier pattern.
2) Before the IFFT, the frequency-domain data of the active subcarriers are coded by a DFT matrix.

It is noted that according to the channel gain there are some subcarriers that carry null data because of deep attenuation. The precoding of frequency-domain OFDM symbols is twofold. First, it results in a correlation between the input data of the IFFT. This decreases the appearance probability of large peaks of the transmitted signal. Hence, the PAPR of the laser-driving current is reduced. The interests of this PAPR reduction are discussed in detail in section V. Second, the precoding is also a good way to benefit from the frequency selectivity of an optical IMDD channel. The overall BER of a transmission is no longer dominated by subcarriers having the worst SNRs. It is shown later that by appropriately selecting null-subcarrier patterns and the corresponding signal constellations, the performance in terms of data rate of the POFDM technique can be equivalent to that of the LC adaptive modulation technique, which is known as an approximate solution of the optimal water-filling technique.

In order to find the optimal QAM constellation and null-subcarrier pattern in a POFDM system, a rate adaptive algorithm, which is based on the equivalent noise power of the system, is proposed. In the next paragraphs the analytical expression of the equivalent noise power and methods for estimating its value are discussed.

A. Equivalent Noise Power Estimation

A transmitted data symbol is referred to as $\boldsymbol{X} = [X_0 X_1 \dots X_{M-1}]^T$ and its frequency domain symbol after precoding is $\boldsymbol{Y} = [Y_0 Y_1 \dots Y_{M-1}]^T = \text{DFT}\{\boldsymbol{X}\}$ where $\{X_k | k = 0, \dots, M-1\}$ is taken from QAM constellation and carried by a set of M corresponding active subcarriers. In general, M is smaller than the total number of data subcarriers. The received sample Z_k corresponding to the kth active subcarrier after the FFT at the receiver side is expressed as

$$Z_k = Y_k H_k + \text{ICI}_k + N_k \tag{1}$$

where H_k denotes the channel gain and ICI_k denotes the nonlinear inter-carrier interference (ICI) including the intermodulation terms [10], the Four Wave Mixing products [11] that fall into the $k\text{th}$ sub-channel, the clipping noise and the quantization noise of the ADC and N_k denotes the photodetector noise. Supposing that the channel response is known at the receiver side and the zero-forcing channel equalization technique is used, the received data symbol before the IDFT decoding can be written as

$$\hat{\boldsymbol{Y}} = \left[\hat{Y}_0 \hat{Y}_1 \dots \hat{Y}_{M-1}\right]^T = \left[\frac{Z_0}{H_0} \frac{Z_1}{H_1} \dots \frac{Z_{M-1}}{H_{M-1}}\right]^T$$

and after the IDFT decoding

$$\hat{\boldsymbol{X}} = \left[\hat{X}_0 \hat{X}_1 \dots \hat{X}_{M-1}\right]^T = \text{IDFT}\left\{\hat{\boldsymbol{Y}}\right\}$$

where

$$\begin{aligned} \hat{Y}_n &= Y_n + \frac{\text{ICI}_n}{H_n} + \frac{N_n}{H_n} \\ &= Y_n + \text{ICI}'_n + N'_n \end{aligned} \tag{2}$$

and

$$\begin{aligned} \hat{X}_n &= \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} \left(Y_k + \frac{\text{ICI}_k}{H_k} + \frac{N_k}{H_k}\right) e^{j2\pi nk/M} \\ &= X_n + \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} \frac{\text{ICI}_k}{H_k} e^{j2\pi nk/M} \\ &\quad + \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} \frac{N_k}{H_k} e^{j2\pi nk/M} \\ &= X_n + \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} \text{ICI}'_k e^{j2\pi nk/M} \\ &\quad + \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} N'_k e^{j2\pi nk/M} \end{aligned} \tag{3}$$

with $\mathrm{ICI}'_k = \mathrm{ICI}_k/H_k$ and $N'_k = N_k/H_k, n = 0, \ldots, M-1$.

It can be seen in (3) that the detected data symbol consists of the transmitted data symbol and an equivalent noise which includes the nonlinear ICI term and the photodetector noise term. Because the transmitted data and the photodetector noise are independent, the nonlinear ICI noise and the photodetector noise are hence independent. Hence, the equivalent noise power can be expressed as

$$\begin{aligned}\sigma^2_{\mathrm{equiv},n} &= E\left\{N_{\mathrm{equiv},n}\left(N_{\mathrm{equiv},n}\right)^*\right\} \\ &= \frac{1}{M}E\left\{\sum_{k=0}^{M-1}\left|\mathrm{ICI}'_k\right|^2\right\} + \frac{1}{M}E\left\{\sum_{k=0}^{M-1}\left|N'_k\right|^2\right\} \\ &\quad + \frac{1}{M}E\left\{\sum_{i=0}^{M-1}\sum_{j=0,j\neq i}^{M-1}\mathrm{ICI}'_i\mathrm{ICI}'^*_j e^{\frac{j2\pi n(i-j)}{M}}\right\} \\ &= \frac{1}{M}\sum_{k=0}^{M-1}\left(\sigma^2_{\mathrm{ICI}',k} + \sigma^2_{N',k}\right) \\ &\quad + \frac{1}{M}E\left\{\sum_{i=0}^{M-1}\sum_{\mathrm{j}=0,\mathrm{j}\neq\mathrm{i}}^{M-1}\mathrm{ICI}'_i\mathrm{ICI}'^*_j e^{\frac{j2\pi n(i-j)}{M}}\right\} \quad (4)\end{aligned}$$

where the first term stands for the power of the nonlinear ICI noise and the photodetector noise, and the second term corresponds to the correlation between the nonlinear ICI terms at different sub-channels.

For the estimation of the equivalent noise power, two methods are proposed. The first one involves estimating the noise power at the receiver side after the IDFT decoding. The estimated noise power can be expressed as

$$\hat{\sigma}^2_{\mathrm{equiv},n} = \frac{1}{L}\sum_{i=1}^{L}\left|\hat{X}^i_n - X^i_n\right|^2, n = 0, \ldots, M-1 \quad (5)$$

where L denotes the number of training OFDM symbols. This estimation technique gives a good estimate of the equivalent noise power since it takes into account all possible noise sources in the system. However, by estimating the noise power after the decoding, the noise power on each individual sub-channel is not available. Hence, in order to find the best null-subcarrier pattern, it is necessary to send at least one training sequence for each pattern. Therefore, the initialization of the modem may take a long time to finish.

To reduce the modem initialization time, the second method of noise power estimation consists in approximating the equivalent noise power. By neglecting the second term in (4), from (2) and (4) the approximate noise power can be estimated at the receiver side before the IDFT decoding

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{equiv,appro},n} &= \frac{1}{M}\sum_{k=0}^{M-1}\left(\hat{\sigma}^2_{\mathrm{ICI}',k} + \hat{\sigma}^2_{N',k}\right) \\ &= \frac{1}{M}\sum_{k=0}^{M-1}\hat{\sigma}^2_k = \frac{1}{M}\sum_{k=0}^{M-1}\frac{1}{L}\sum_{i=1}^{L}\left|\hat{Y}^i_k - Y^i_k\right|^2. \quad (6)\end{aligned}$$

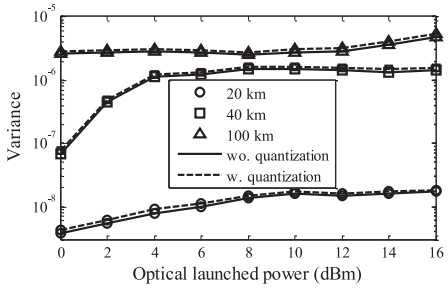


Fig. 2. Variance of the approximation error as a function of transmitted optical power and fiber length. 16-QAM constellation, 8-bit quantization. The modulation index $m = 0.9$. The received signal is equalized and its power is normalized to unity.

It is noted that by neglecting the second term in (4) the equivalent noise power is the same for all detected data symbols $\hat{X}_n, n = 0, \ldots, M-1$. And by estimating the noise power before decoding, the noise power for each individual sub-channel is available. With just one training sequence the noise power on the kth sub-channel can be estimated by

$$\hat{\sigma}^2_k = \hat{\sigma}^2_{\mathrm{ICI}',k} + \hat{\sigma}^2_{N',k} = \frac{1}{L}\sum_{i=1}^{L}\left|\hat{Y}^i_k - Y^i_k\right|^2, k = 0, \ldots, N_{\mathrm{SC}} - 1 \quad (7)$$

where N_{SC} denotes the number of sub-channels. Once the noise power on each sub-channel is available, the equivalent noise power in transmissions using different null-subcarrier patterns can be estimated with (6). In practice, this method allows a faster initialization as only one training sequence is required.

B. Approximation Error Analysis

In the second method of noise power estimation, the approximation is only valid when the power of the second term in (4) is negligible when compared to that of the first term. In an OFDM transmission where frequency subcarrier spacing is in the order of hundreds of MHz, the strength of FWM products may become important [11], [12]. Particularly, in the context of access networks where the power budget is relatively low, the transmitted optical power may need to be high to obtain a high bit-rate transmission. The effects of nonlinearity noise could become important. Hence, the power of the neglected term in (4) is expected to be important. Fig. 2 illustrates the power of the approximation error as a function of the transmitted optical power and the transmission distance. Simulation parameters are given in Table I and the transmission configuration is detailed in Section IV. The photodetector noise is not introduced in this simulation in order to distinguish the nonlinear ICI noise from the receiver noise. As expected, when transmitted optical power and/or fiber length increase, the power of the correlation between FWM products falling into different sub-channels increases because the strength of each FWM product increases. In addition, the noise power is also increased because of frequency response dips caused by chromatic dispersion. In Fig. 2 quantization noise is shown to have a negligible impact on the

TABLE I
PARAMETERS USED IN THE SIMULATIONS

Symbol	Physical meaning	Unit	Value
F	newly defined laser parameter *	A/W	5.32
I_{th}	laser threshold current	mA	20
I_S	laser characteristic current	μA	5.65
B	newly defined laser parameter *	Hz^2/A	1.3e23
τ_p	photon lifetime	ps	4.12
τ_n	carrier lifetime	ns	0.14
τ_c	newly defined laser parameter *	ns	1.65
α	laser linewidth enhancement factor	---	2.68
D	dispersion parameter	$ns/(nm.km)$	17
α_{att}	fiber attenuation coefficient	dB/km	0.2
λ	optical wavelength	μm	1.55
A_{eff}	fiber effective area	m^2	80e-12
n_2	nonlinear index	m^2W^{-1}	2.5e-20
Se	photodetector sensitivity	dBm	-20

The laser parameters are extracted from a commercialized DFB laser developed by the EM4 company. The laser's model is E0037751
* The analytical expression and the physical meaning of these parameters can be found in [17].

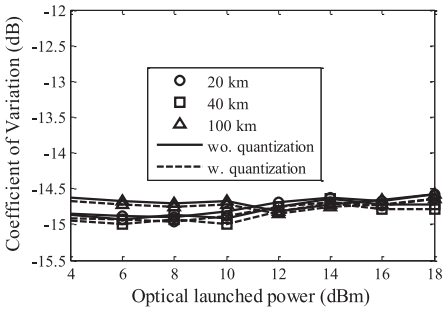


Fig. 3. Coefficient of variation as a function of transmitted optical power and fiber length. The approximation error is found to be negligible. The quantization noise also has insignificant impacts on the approximation.

approximation error. Indeed, as reported in [13] the quantization noise is found to be a quasi white noise. At the receiver side, after the FFT it can be also considered white. Therefore, the power of the neglected term in (4), which is the correlation between noise samples at different sub-channels, is expected to be negligible. The same conclusion is found for the clipping noise provoked by laser saturation. According to [14] the clipping noise is an uncorrelated noise. After the FFT at the receiver side, according to the central limit theorem it is found to be an uncorrelated Gaussian noise. Because the noise is uncorrelated, the correlation between noise samples at different sub-channels is also expected to be negligible.

In order to analyze the importance of the approximation error when compared to the overall noise power we consider the coefficient of variation of the root mean square deviation (RMSD)

$$CV_{\mathrm{RMSD}} = \frac{\sqrt{E\left\{\left(\hat{\sigma}^2_{\mathrm{equiv,appro},n} - \hat{\sigma}^2_{\mathrm{equiv},n}\right)^2\right\}}}{E\left(\hat{\sigma}^2_{\mathrm{equiv,appro},n}\right)} \quad (8)$$

which is the ratio between the RMSD and the mean value of the estimate of the second noise power estimation method. In dB the coefficient of variation is expressed as $CV_{\mathrm{RMSD},dB} = 10\log_{10}\left(CV_{\mathrm{RMSD}}\right)$. As shown in Fig. 3, even in a relatively long transmission (100 km) and with a strong transmitted power the coefficient of variation is always in the order of -15 dB. It is important to note that when the transmitted power and/or the fiber length increase, the power of the first term in (4) also increases. Hence, even though the approximation error power increases due to the fiber nonlinearity, the coefficient of variation is still very small. Obviously, the approximation error can be considered negligible in the context of access networks, which is the scope of this paper.

It must be noted that the nonlinear ICI noise depends itself on the null-subcarrier pattern. Hence, the nonlinear ICI noise power in transmissions with optimal null-subcarrier patterns may differ from that in transmissions of training sequences. Indeed, in a system using rate adaptive algorithm, it is important to always achieve the targeted SER while optimizing the capacity of the transmission. It is shown by simulation that even with only one training sequence the second method of noise power estimation still gives optimal null-subcarrier patterns that guarantee the desired SER of a transmission. For the sake of simplicity, the second method is used for all simulations in this paper.

C. The Proposed Rate Adaptive Algorithm

Inspired by the rate adaptive algorithm proposed by Levin and Campello [2], [15] which is used in LC AMOOFDM system, the proposed rate adaptive algorithm for POFDM transmissions involves optimizing the capacity of the transmission under a constraint of the electrical transmitted signal power and the desired SER of the transmission. The objective of the algorithm is to obtain the optimal active subcarrier pattern (or null-subcarrier pattern) and the corresponding QAM constellation based on the estimated equivalent noise power. In a system where the number of subcarriers is important, the total number of possible null-subcarrier patterns might be huge. Hence, an exhaustive searching for the optimal pattern may result in long initialization time of a POFDM modem. From (6), the equivalent noise power is shown to be the average of the noise power on each sub-channel. Hence, in order to maximize the equivalent SNR of the POFDM system, if a subcarrier must carry null data, it must be the subcarrier having the worst SNR. In a POFDM system, the transmission capacity depends on two factors: the number of active subcarriers and the corresponding constellation of each subcarrier, which is determined by the equivalent noise's power in the system. Giving a fixed number of active subcarriers, the optimal null-subcarrier pattern is the one in which the subcarriers having the worst SNRs carry null data. By using this remark the number of possible null-subcarrier patterns is reduced to N_{SC}, which is the total number of subcarriers. The algorithm needs to check only N_{SC} patterns in which the number of active subcarriers varies from 1 to N_{SC}.

Algorithm 1: Rate Adaptive POFDM

input : Vector nse_pwr of size 1 x N_{SC} with positive elements containing noise power on each sub-channel

output : Optimal active-subcarrier pattern $opt\_actv\_pattern$
Optimal number of bits per QAM symbol b_{opt}

$B_{max} \leftarrow 0$; /* Initialize the maximum bit rate */

for $i \leftarrow N_{SC}$ **to** 1 **do** /* loop to check N_{SC} possible patterns */

$M \leftarrow i$; /* M : the number of active subcarriers */

$N_{null} \leftarrow (N_{SC} - M)$; /* N_{null} : the number of null subcarriers */

/* Initialization of noise power on active subcarriers */

$nse\_pwr\_actv\_sc \leftarrow nse\_pwr$;

/* Discard the null subcarriers which have the worst SNRs. The corresponding noise power values are set to 0 so that they are not considered in the equivalent SNR. */

set N_{null} largest values in $nse\_pwr\_actv\_sc$ to zero ;

/* Calculate the equivalent SNR from (6) */

$$SNR_{equiv} \leftarrow \left(\frac{M}{\text{sum}(nse\_pwr\_actv\_sc)}\frac{N_{SC}}{M}\right) ; \quad (9)$$

/* N_{SC}/M is the energy normalizing factor. The number of active subcarriers can change but the electrical power budget is fixed. */

/* Calculate the number of bits carried by the corresponding QAM symbol according to the SNR gap Γ [15], [16] */

$$b \leftarrow floor\left(log_2\left(1 + \frac{SNR_{equiv}}{\Gamma}\right)\right) ; \quad (10)$$

/* Calculate the transmission bit rate */

$$B \leftarrow (b * M) ; \quad (11)$$

if $B \geq B_{max}$ **then** /* Update the optimal values */

$B_{max} \leftarrow B$; /* Update the maximum bit rate */

$b_{opt} \leftarrow b$; /* Optimal number of bits per QAM symbol */

/* As the values corresponding to null subcarriers in *nse_pwr_actv_sc* are set to 0, we deduce the active subcarrier pattern by finding the positions of positive values. */

$opt\_actv\_pattern \leftarrow (nse\_pwr\_actv\_sc > 0)$;

end if

end for

IV. Modeling and Simulation Parameters

A. Laser Modeling

In order to simulate the DFB laser, the large signal rate equation model described in [17] is adopted. The evolution of the laser's intensity and phase is modeled by a set of rate equations whose parameters can be totally extracted by experimental measurements. In this paper the same extraction procedure as described in [17] is applied on a commercialized DFB laser. The extracted parameters, which are coherent with the constructor datasheet and reasonable physical values, are given in Table I. The extracted parameters are also validated by comparing the experimental data with the simulated data.

1) Laser Parasitic: For the first-order approximation, the filtering effect of the laser parasitic can be approximated by an *RC* low pass filter [18]. In the simulation an *RC* low pass filter with 6 GHz cutoff frequency is used.

2) Laser Phase Noise to Intensity Noise Conversion: It is well known that the PM-to-AM conversion noise may become one of the principal limiting factors of gigabit-per-second optical IMDD transmissions [19]. According to [20], the laser phase noise can be modeled by Wiener process where the phase change between two consecutive samples can be modeled by a central Gaussian random variable whose variance can be written as $E\left\{\left(\Delta\phi\left(T_S\right)\right)^2\right\} = 2\pi\Delta\upsilon T_S$. T_S is the sampling time and $\Delta\upsilon$ is the laser linewidth which is inversely proportional to the optical power [19]. For a DFB laser, the typical value of $\Delta\upsilon$ is 5 MHz at 1 mW output optical power.

B. Fiber Modeling

The propagation of an optical signal through an SMF can be described by the nonlinear Schrödinger equation [21]

$$\frac{\partial A}{\partial z} = -\frac{j}{2}\beta_2\frac{\partial^2 A}{\partial t^2} - \frac{\alpha}{2}A + j\gamma\left|A\right|^2 A \quad (12)$$

where $\beta_2\left(s^2/m\right)$ is the dispersion parameter, $\alpha\left(dB/m\right)$ is the fiber attenuation coefficient and γ is the fiber nonlinearity coefficient. The numerical solution of (12) can be obtained by using the symmetrized split-step-Fourier method [21].

C. Photodetector Modeling

In general, the bandwidth of a photodetector is larger than that of other components in the transmission link. In the simulation where the sampling frequency is only 12 GS/s, the frequency response of the photodetector can be considered flat. Hence, at the receiver side only the noise is considered. According to the receiver sensitivity, all the possible noise sources can be modeled [1]. In this paper a PIN-based photodetector is used, and the typical value of sensitivity of such a receiver is -20 dBm (corresponding to a non-return to zero modulation at 10^{-9} BER).

D. Simulation Parameters

The parameters used in the simulation of different components in the transmission link are given in Table I. The numerical simulation is done with MATLAB software. For the OFDM modulation, the FFT size is 256. However, only 107 subcarriers can carry useful data. The DC subcarrier carries null data and 20 subcarriers at the edge of the spectrum are set to zero to reduce the aliasing impact. The 128 negative-frequency subcarriers are the complex conjugate of the 128 corresponding positive-frequency subcarriers. After the IFFT, 16 last samples of the time-domain OFDM symbol are copied at the beginning of the symbol to form the cyclic prefix. In the context of access networks this cyclic prefix length is sufficient to combat the chromatic dispersion. The DAC has 8 quantization bits and the clipping rate, which is the ratio between the maximal quantized power and the signal average power, is 13 dB. These values are adopted according to [1], which are shown to provoke negligible quantization noise and clipping noise. The optical power coupled into the fiber link is always 0 dBm. The transmission distance can vary from 0 to 100 km in the simulations. At the receiver side the received optical power can vary according to a loss profile, which stands for possible insertion losses in the transmission link. For the channel estimation, because the IMDD optical channel is relatively static, the simple least-square

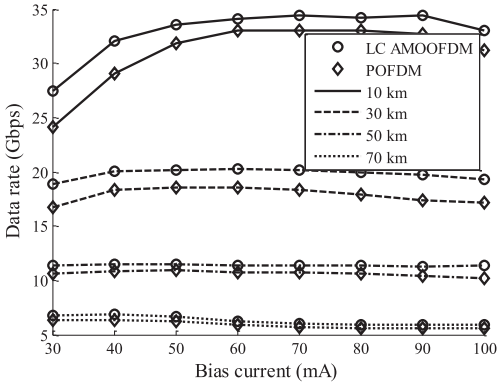


Fig. 4. Transmission data rate as a function of laser bias current and fiber length at optimal modulation indices. Loss profile 4 dB.

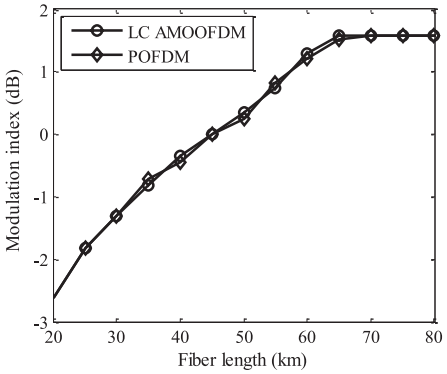


Fig. 5. Optimal modulation index as a function of fiber length. The modulation index increases with the fiber length to compensate for the fiber loss. Laser bias current 60 mA. Loss profile 4 dB.

method is used to estimate the channel response. In order to reduce the impact of noise on the channel estimation performance, the estimated channel response is then averaged over 30 training symbols. Further increase of the number of training symbols does not improve the estimation performance. For the noise estimation, during the modem initialization 1000 OFDM training symbols are transmitted. The approximation method described in Section II is adopted to estimate the noise power, which is the input argument of the LC rate adaptive algorithm and the proposed algorithm for POFDM transmissions.

E. Optimum Operating Parameters of the Laser

Fig. 4 illustrates the transmission data rate as a function of laser bias current and transmission distance in an unamplified system. It is noted that in both LC AMOOFDM and POFDM transmissions, the SNR gap Γ is chosen so that the SER at each subcarrier converges at 10^{-3}. Hence, the corresponding BER is always lower than 10^{-3}. This allows an error-free transmission with an appropriate forward error correction. The transmission data rate is then calculated as the sum of the bits transmitted with each subcarrier per second. In short-distance transmissions ($<$30 km), the data rate drops when the bias current decreases. This is because when the laser bias current decreases, the adiabatic chirp also decreases. The first attenuation zone in the channel transfer function might be very deep due to the small value of adiabatic chirp [22]. Contrarily, in transmissions longer than 30 km, the nonlinear ICI noise becomes important when the transmission distance increases [10]. It is shown in Fig. 5 that the driving current's peak-to-peak value might become large to adapt to fiber loss. Hence, an increase of adiabatic chirp (equivalent to an increase of bias current) might enhance the FM modulation of the laser. This results in an increased nonlinear ICI noise power which degrades the system performance.

The transmission data rate becomes relatively stable for bias currents larger than 60 mA. When the bias current continues to increase the data rate begins to drop due to important adiabatic chirp and the saturation of the laser. In practice, a larger bias current also results in a more power-consuming transmitter. For a compromise between the channel throughput and the power consumption, the bias current is chosen to be 60 mA from now on. The corresponding adiabatic frequency is 2 GHz. It is noted that in each transmission the modulation index, which is the ratio between the clipped amplitude of the driving current and the bias current, can vary from -10 to 1.6 dB in order to find the optimal value which is the compromise between different noise sources. In most cases the performance of the POFDM system is equivalent to that of the LC AMOOFDM system. In short-distance transmissions ($<$30 km), thanks to an adaptive modulation, the LC AMOOFDM system brings a little gain in terms of data rate. However, in all cases, no important gain is observed.

V. Power Saving Due to PAPR Reduction

A. PAPR Reduction

The PAPR of a discrete OFDM signal is defined as the ratio of the maximum peak power to the average power over a symbol interval

$$\text{PAPR}\{X_k\} = \max_{0\le k\le N-1} \frac{|X_k|^2}{E\left\{|X_k|^2\right\}} \tag{13}$$

where N is the number of samples in an OFDM symbol. Indeed, the fact that the data carried by different subcarriers are independent makes the appearance probability of large peaks high. It is well known that the important PAPR is one of the major disadvantages of an OFDM modulation. In practice, the power of an OFDM signal is amplified by a PA before modulating the laser. And the largest peaks of the OFDM signal are always clipped due to the amplifier saturation. Given a fixed input average power and a fixed saturation power of the amplifier, when the PAPR increases the clipping probability also increases. The signal clipping results in an in-band noise, which degrades the SNR, and out-of-band radiations which cause interchannel interference. For this reason, the average power of the OFDM signal must be adjusted so that it is rarely clipped due to the amplifier saturation. That is, before the amplifier one has

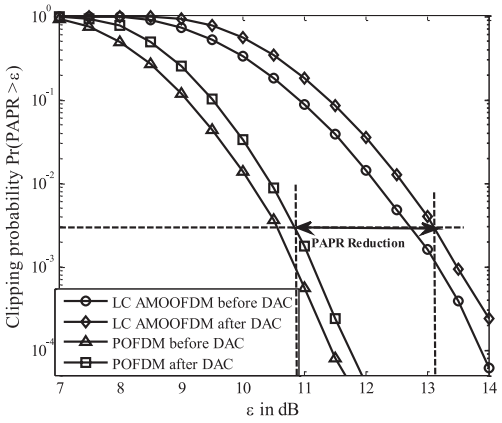


Fig. 6. CCDF comparison. For a same clipping probability, the amount of PAPR reduction is also the amount of IBO reduction, which improves the efficiency of the PA.

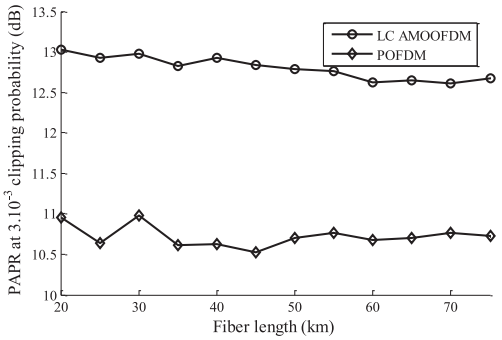


Fig. 7. PAPR performance as a function of transmission distance in unamplified transmissions. Loss profile 4 dB.

to apply to the OFDM signal an input backoff (IBO), which is the ratio between the minimum input power that saturates the amplifier and the average power of the input signal. In order to examine PAPR characteristics of an OFDM modulation it is common in literature for authors to use the complementary cumulative distribution function (CCDF) of the PAPR, which gives the clipping probability of the OFDM signal.

Fig. 6 shows the CCDF of OFDM and POFDM signals before and after the DAC. It is noted that in this simulation all the data subcarriers carry 16-QAM symbols. Due to the precoding by a DFT matrix, the PAPR of POFDM signal is strongly reduced when compared to that of OFDM signal. This is because of the correlation between different frequency-domain subcarriers data after the precoding. For example, at a same clipping probability of 10^{-3}, the PAPR is reduced by 2 dB. It is shown later that this PAPR reduction has a straightforward relation with the reduction of the PA power consumption.

B. PA Efficiency and Power Saving

The PA efficiency is defined as $\eta = P_{\mathrm{out,av}}/P_{\mathrm{DC}}$ where $P_{\mathrm{out,av}}$ is the average power output and P_{DC} is the delivered DC power which is also the power consumed by an amplifier. Under the assumption of ideal linear amplifier, according to [5] the PA efficiency can be written as

$$\eta = 0.5/\varepsilon \tag{14}$$

where ε is the inverse function of the CCDF at a given clipping probability p. Another speaking ε is the corresponding abscissa value of the CCDF in Fig. 6 given the ordinate value p. It is noted that ε is also the IBO one must apply to the OFDM signal given the clipping probability p. Hence, it is straightforward that, for the same clipping probability p, when the PAPR is reduced (ε is reduced) the amplifier efficiency is also improved, resulting in a reduction of amplifier power consumption. The power saving due to PAPR reduction can be expressed as

$$P_{\mathrm{saving}} = P_{\mathrm{DC},1} - P_{\mathrm{DC},2} = 2\left(P_{\mathrm{out,av},1}\varepsilon_1 - P_{\mathrm{out,av},2}\varepsilon_2\right) \tag{15}$$

where the subscript 1 denotes the LC AMOOFDM system and the subscript 2 denotes the POFDM system.

For a fixed clipping probability of 3.10^{-3}, the PAPR of POFDM and LC AMOOFDM signals in an unamplified transmission is illustrated in Fig. 7. The PAPR of LC AMOOFDM signal decreases at longer transmission distances because the number of active subcarriers also decreases with distance due to the fiber loss. The PAPR of POFDM signal depends on three factors: the number of active subcarriers, their position in the signal spectrum and the data constellation [8]. Because of the adaptive modulation, all these three factors can be adapted to the channel condition. As a consequence, in Fig. 7 the PAPR of POFDM signal varies as a function of the fiber length. Unlike the conventional OFDM signal's PAPRs, the PAPR of a POFDM signal is more sensitive to the data constellation [8]. The PAPR fluctuations when the fiber length increases from 20 to 25 km, 25 to 30 km, 30 to 35 km and 45 to 50 km correspond to the constellation changes from 64-QAM to 32-QAM, 32-QAM to 64-QAM, 64-QAM to 32-QAM and 32-QAM to 16-QAM, respectively. The PAPR increase or decrease when the constellation changes is also validated according to [8]. When the fiber length increases from 25 to 30 km, although the constellation size increases from 32-QAM to 64-QAM, the number of active subcarriers is strongly reduced from 92 to 68. Hence, the resulting bit rate decreases. When the data constellation is fixed, a small PAPR fluctuation is observed due to the variation of the number of active subcarriers and the change of their position in the signal spectrum. Supposing a 50-Ohm equivalent load resistance in the system, the power of laser-driving signals can be calculated. Due to an IBO applied to the driving signals, the DC powers which supply the PAs are much more important than the average power of the driving signals as shown in Fig. 8. The gross power saving of the POFDM system can be deduced by using (15).

In order to calculate the net power saving of the POFDM system when compared to the LC AMOOFDM system, the power cost due to the precoding at the transmitter and the decoding

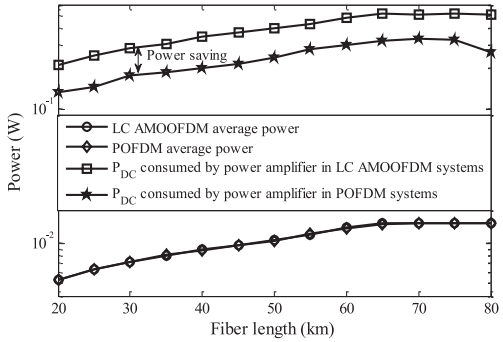


Fig. 8. Average power of driving signals and the corresponding DC power supply for PAs as a function of transmission distance in unamplified transmissions. Loss profile 4 dB.

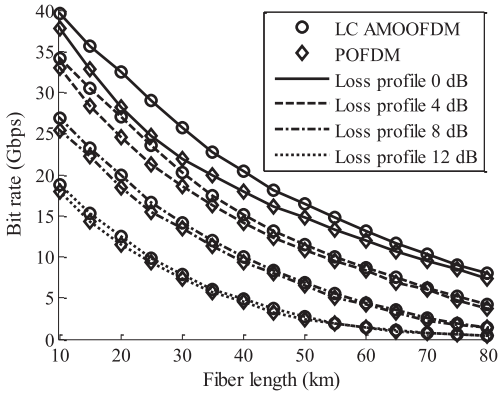


Fig. 9. Capacity-versus-reach performance in unamplified transmissions with different loss profiles.

at the receiver must be taken into account. Supposing that the precoding can be done by using an FFT and the DSP works for the same amount of time as the PA, it is shown in [5] that the power cost is on the order of μW, whereas the gross power saving is on the order of W. Obviously the POFDM technique provides a large net power saving when compared to the LC AMOOFDM technique. It is interesting to note that the amount of the power saving is comparable to the power consumed by the laser, which is one of the most power-consuming components of the transmitter.

VI. Optimal Capacity-Versus-Reach Performance

In the previous section the proposed POFDM system is shown to outperform the conventional LC AMOOFDM system in terms of PAPR and power consumption. It is also shown in [1], [4] that unamplified AMOOFDM systems employing IMDD technique are suitable for cost-effective optical access networks. Therefore, it is interesting to compare the performance in terms of data rate of the proposed POFDM and the LC AMOOFDM systems under a same transmission configuration. Fig. 9 illustrates the data line rate in function of fiber length in unamplified transmissions with different loss profiles. In most cases, the POFDM system has the same performance as the LC AMOOFDM system. When the loss profile is small, in short transmissions (<30 km) the LC AMOOFDM system outperforms POFDM system. However, when the loss profile becomes important the adaptive modulation becomes unnecessary and the performance of POFDM system is almost the same as that of LC AMOOFDM system. This is because when the channel gain and the SNR are low due to chromatic dispersion and fiber loss, the number of active subcarriers in LC AMOOFDM system becomes small. Hence, the adaptive modulation does not bring much gain when compared to the precoding technique where all the active subcarriers have the same modulation constellation.

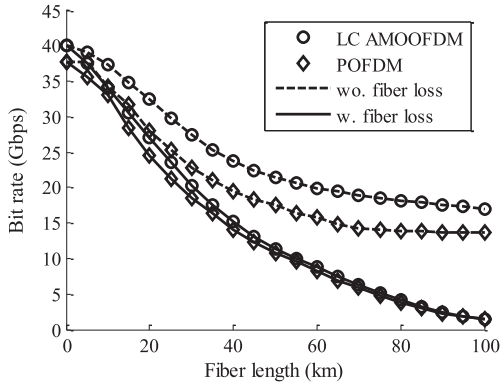


Fig. 10. Capacity-versus-reach performance comparison in unamplified transmissions with and without fiber loss. Loss profile 4 dB.

A. Impact of Fiber Loss

From Fig. 9 it is shown that in both systems the data rate drops sharply as the fiber length increases. In order to analyze the impact of fiber loss on the transmission data rate, the performance of the two modulation techniques in unamplified configuration with and without fiber loss is illustrated in Fig. 10. It is observed that in both cases with or without fiber loss, for any transmission distance shorter than 30 km the data rate drops sharply when the distance increases. This is because of the interplay between the laser frequency chirps and the chromatic dispersion. Indeed, when the distance increases, the first deep null of the channel frequency response shifts into the low-frequency zone and begins to exist in the signal frequency band. In case of no link loss, longer transmission distance does not decrease much the transmission capacity because within the transmission range and within the signal band of interest, the second deep null of the frequency response still does not exist. However, if the fiber loss is taken into account, the transmission throughput continues to drop at longer transmission distance due to poor SNRs at the receiver input.

B. Impact of Laser Frequency Chirps

Considering the "small signal" optical channel frequency response, the transient laser chirp amplifies the channel gain of all subcarriers by a factor of $\sqrt{1+\alpha^2}$ [9], and when the adiabatic

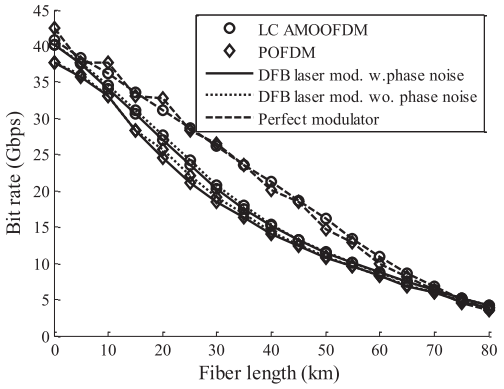


Fig. 11. Capacity-versus-reach performance of transmissions with different modulators: Perfect modulator and DFB laser modulator with and without laser phase noise. Loss profile 4 dB.

chirp is taken into account, the first dip of the channel frequency response is improved [23]. These results may lead to the conclusion that laser frequency chirps may be an advantage of a DFB-laser intensity modulator when compared to a no-chirp intensity modulator. However, as mentioned before, in a dispersive channel the laser frequency chirps cause intermodulation noise. Indeed, because of the laser chirps, an intensity modulation provokes itself a phase modulation. After the dispersive channel, at the receiver side the phase modulation results in an intensity distortion. In an OFDM system, the intermodulation distortion causes ICI that cannot be suppressed even with a sufficient cyclic prefix. It is also reported in [1] that the laser-induced frequency chirp is one of the key limiting factor of an optical IMDD OFDM transmission. In order to show the impact of the laser frequency chirps on the transmission capacity, the performance of LC AMOOFDM and POFDM systems with a DFB-laser intensity modulator and that with a perfect intensity modulator (no-chirp) are illustrated in Fig. 11. For transmission distances shorter than 60 km, the laser frequency chirps degrade the transmission capacity due to the intermodulation noise. In practice under a large signal modulation regime the laser chirps are considered as a detriment factor rather than a good factor even though the gain it presents in the small signal frequency response. For transmission distances longer than 60 km, the performance of a perfect modulator is almost the same as that of a DFB-laser modulator. This is because in long-distance transmissions the fiber loss becomes the principal capacity-limiting factor. The performance of a DFB-laser modulator with and without laser phase noise is also shown in Fig. 11. It is observed that the laser phase noise-induced PM-to-AM noise conversion has little impact on the transmission throughput.

It can be concluded that for transmission distances shorter than 60 km, the principal capacity-limiting factors are the chromatic dispersion and the laser frequency chirps. And for fiber lengths longer than 60 km, the fiber loss becomes the main capacity-limiting factor. Moreover, in unamplified transmissions, in most cases the performance in terms of useful data rate of the proposed POFDM system is almost the same as that of the conventional LC AMOOFDM system. In particular, when the noise margin is important (or important loss profile), the adaptive modulation brings negligible gain when compared to the proposed modulation scheme.

VII. Conclusion

Based on the proposed rate adaptive algorithm, the performance of a DFT Precoded OFDM modulation technique is analyzed in unamplified optical IMDD transmissions. It is shown that the power consumption of the proposed POFDM modem is largely reduced due to the strong PAPR reduction of the transmitted signal. In addition, it is also shown that the proposed transmission scheme is capable of achieving the same data rate as the conventional LC AMOOFDM system. The proposed POFDM system also has all the advantages of an AMOOFDM modulation, which are: flexibility and robustness, cost-effectiveness and high spectral efficiency. Hence, the POFDM system can be a cost-effective solution for optical transmissions in the context of access networks.

Acknowledgment

T. A. Truong thanks N. Genay, research engineer at Orange Labs, Lannion, France for her help during the extraction of lasers parameters for the simulation used in this paper.

References

[1] J. M. Tang and K. A. Shore, "30-gb/s signal transmission over 40-km directly modulated DFB-laser-based single-mode-fiber links without optical amplification and dispersion compensation," *J. Lightw. Technol.*, vol. 24, no. 6, pp. 2318–2327, Jun. 2006.

[2] J. Campello, "Optimal discrete bit loading for multicarrier modulation systems," presented at the IEEE Int. Symp. Inform. Theory, Cambridge, MA, USA, 1998.

[3] B. Charbonnier, P. Urvoas, M. Ouzzif, and J. Le Masson, "Capacity optimisation for optical links using DMT modulation, an application to POF," in *Proc. 34th Eur. Conf. Opt. Commun.*, 2008, pp. 1–2.

[4] T.-N. Duong, N. Genay, M. Ouzzif, J. Le Masson, B. Charbonnier, P. Chanclou, and J.-C. Simon, "Adaptive loading algorithm implemented in AMOOFDM for NG-PON system integrating cost-effective and low-bandwidth optical devices," *IEEE Photon. Technol. Lett.*, vol. 21, no. 12, pp. 790–792, Jun. 2009.

[5] R. J. Baxley and G. T. Zhou, "Power savings analysis of peak-to-average power ratio in OFDM," *IEEE Trans. Consumer Electron.*, vol. 50, no. 3, pp. 792–798, Aug. 2004.

[6] Y. Tang, W. Shieh, and B. S. Krongold, "DFT-spread OFDM for fiber nonlinearity mitigation," *IEEE Photon. Technol. Lett.*, vol. 22, no. 16, pp. 1250–1252, Aug. 2010.

[7] M. Sung, S. Kang, J. Shim, J. Lee, and J. Jeong, "DFT-precoded coherent optical OFDM with hermitian symmetry for fiber nonlinearity mitigation," *J. Lightw. Technol.*, vol. 30, no. 17, pp. 2757–2763, Sep. 2012.

[8] H. G. Myung, J. Lim, and D. Goodman, "Peak-to-average power ratio of single carrier FDMA signals with pulse shaping," in *Proc. IEEE 17th Int. Symp. Personal, Indoor Mobile Radio Commun.*, 2006, pp. 1–5.

[9] F. Devaux, Y. Sorel, and J. F. Kerdiles, "Simple measurement of fiber dispersion and of chirp parameter of intensity modulated light emitter," *J. Lightw. Technol.*, vol. 11, no. 12, pp. 1937–1940, Dec. 1993.

[10] E. Peral and A. Yariv, "Large-signal theory of the effect of dispersive propagation on the intensity modulation response of semiconductor lasers," *J. Lightw. Technol.*, vol. 18, no. 1, pp. 84–89, Jan. 2000.

[11] A. J. Lowery, S. Wang, and M. Premaratne, "Calculation of power limit due to fiber nonlinearity in optical OFDM systems," *Opt. Exp.*, vol. 15, no. 20, pp. 13282–13287, 2007.

[12] R. W. Tkach, A. R. Chraplyvy, F. Forghieri, A. H. Gnauck, and R. M. Derosier, "Four-photon mixing and high-speed WDM systems," *J. Lightw. Technol.*, vol. 13, no. 5, pp. 841–849, May 1995.

[13] D. Dardari, "Joint clip and quantization effects characterization in OFDM receivers," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 8, pp. 1741–1748, Aug. 2006.
[14] S. Dimitrov, S. Sinanovic, and H. Haas, "Clipping noise in OFDM-based optical wireless communication systems," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 1072–1081, Apr. 2012.
[15] J. M. Cioffi, *A Multicarrier Primer*, ANSI T1E1.4 Committee Contribution, Nov. 1991.
[16] A. Garcia-Armada, "SNR gap approximation for M-PSK-based bit loading," *IEEE Trans. Wireless Commun.*, vol. 5, no. 1, pp. 57–60, Jan. 2006.
[17] L. Bjerkan, A. Royset, L. Hafskjaer, and D. Myhre, "Measurement of laser parameters for simulation of high-speed fiberoptic systems," *J. Lightw. Technol.*, vol. 14, no. 5, pp. 839–850, May 1996.
[18] T. Ioannis, R. Ioannis, H. Robert, A. Neophytos, B. Aleksandra, and V. Richard, "Extraction of laser rate equations parameters for representative simulations of metropolitan-area transmission systems and networks," *Opt. Commun.*, vol. 194, pp. 109–129, 2001.
[19] S. Yamamoto, N. Edagawa, H. Taga, Y. Yoshida, and H. Wakabayashi, "Analysis of laser phase noise to intensity noise conversion by chromatic dispersion in intensity modulation and direct detection optical-fiber transmission," *J. Lightw. Technol.*, vol. 8, no. 11, pp. 1716–1722, Nov. 1990.
[20] W. Shieh and I. Djordjevic, *Orthogonal Frequency Division Multiplexing for Optical Communications*. San Diego, CA, USA: Elsevier, 2010.
[21] G. P. Agrawal, *Nonlinear Fiber Optics*, 4th ed. San Diego, CA, USA: Academic, 2007.
[22] B. Wedding, "Analysis of fibre transfer function and determination of receiver frequency response for dispersion supported transmission," *Electron. Lett.*, vol. 30, no. 1, pp. 58–59, 1994.
[23] J. Wang and K. Petermann, "Small signal analysis for dispersive optical fiber communication systems," *J. Lightw. Technol.*, vol. 10, no. 1, pp. 96–100, Jan. 1992.

Tuan-Anh Truong was born in Kien Giang, Vietnam, in 1987. He received the B.S. and M.S. degrees both in mathematics and signal processing from the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 2011. He is currently working toward the Ph.D. degree in , Orange Labs, (formerly France Telecom), Orange Labs, Rennes, France.

His research interests include digital signal processing for fiber-based optical communications in the context of access network.

Matthieu Arzel was born in Brest, France, in 1978. He received the Engineering Diploma and the Ph.D degree from the École Nationale Supérieure des Télécommunications (ENST) de Bretagne, Brest, France, in 2002 and 2006, respectively.

In 2006, he was with Turboconcept as a Research Engineer. He joined the Electronic Engineering Department of ENST de Bretagne as a Full-Time Lecturer in 2006. His research interests include iterative decoding techniques, analog/mixed integrated circuit architectures and design.

Hao Lin (M'06) received the Ph.D. degree in electric and communications from Ecole Nationale Supérieure des Télécommunications Paris/Eurecom, France, in 2009. He is now a Research Engineer at Orange Labs, (formerly France Telecom), Rennes, France.

His research interests include multicarrier-related signal processing toward various applications.

Bruno Jahan was born in Chinon, France, in 1966, He received the M.S. degree in optical and photonics in 1989 and the M.S. degree in electronic systems in 1990 from the University of Paris Sud, Orsay, France. In 1991, he was with Telediffusion de France as a Research Engineer. He joined the Orange Labs, (formerly France Telecom), Rennes, in 1998.

His research interests include digital signals processing for wire and wireless communications.

Michel Jézéquel (M'02) was born in Saint Renan, France, on February 26, 1960. He received the degree of "Ingénieur" in electronics from the École Nationale Supérieure de l'Électronique et de ses Applications, Paris, France, in 1982.

From 1983 to 1986, he was a Design Engineer at CIT ALCATEL, Lannion, France. Then, after an experience in a small company, he followed a one year course about software design. In 1988, he joined the École Nationale Supérieure des Télécommunications de Bretagne, where he is currently a Professor, Head of the Electronics Department. His major research interest includes circuit design for digital communications. He focuses his activities in the fields of Turbo codes, adaptation of the turbo principle to iterative correction of intersymbol interference, the design of interleavers, and the interaction between modulation and error correcting codes.

9.4 Article IEEE IWCMC'12 sur l'accélération matérielle de classification de trafic à base de SVM sur FPGA [GAV12]

Hardware Acceleration of SVM-Based Traffic Classification on FPGA

Tristan Groleat
Télécom Bretagne
Brest, France
tristan.groleat@telecom-bretagne.eu

Matthieu Arzel
Télécom Bretagne
Brest, France
matthieu.arzel@telecom-bretagne.eu

Sandrine Vaton
Télécom Bretagne
Brest, France
sandrine.vaton@telecom-bretagne.eu

***Abstract*—Understanding the composition of the Internet traffic has many applications nowadays, mainly tracking bandwidth consuming applications, QoS-based traffic engineering and lawful interception of illegal traffic. Although many classification methods such as Support Vector Machines (SVM) have demonstrated their accuracy, not enough attention has been paid to the practical implementation of lightweight classifiers. In this paper, we consider the design of a real-time SVM classifier at many Gbps to allow online detection of categories of applications. Our solution is based on the design of a hardware accelerated SVM classifier on a FPGA board.**

***Index Terms*—Traffic classification, SVM, FPGA, acceleration.**

I. Introduction

Traffic classification is the task of associating network traffic with the generating application or category of application. In some cases operators would appreciate knowing which application packets belong to in order to better engineer the traffic, charge their customers, etc. All operators are continuously tracking the composition of traffic per category of applications. They are analyzing trends and tracking the emergence of new bandwith consuming applications. Quality of Service (QoS) solutions which segregate traffic into classes and give them different priorities also require the ability to associate traffic to applications since the applications do not tag their traffic by themselves. Traffic classification is also useful for differentiated charging and for Service Level Agreements (SLA) verification. Lawful Interception of illegal traffic makes also mandatory for Internet Service Providers (ISP) to analyze their customers' traffic and recognize some illegal or critical traffic.

Traffic classification is a challenging task for several reasons. First, operators have to deal with a huge amount of traffic to analyze as reported by the annual CISCO report [1]. Second, traditional techniques for traffic classification have some limitations. Traditionally traffic could be classified either by the analysis of port numbers or by Deep Packet Inspection (DPI). Port-based classification is not always reliable since many applications change dynamically their port number or hide themselves behind well-known ports belonging to other applications (e.g. port 80 for HTTP). DPI techniques recognize specific character strings in the packet payload. These techniques can be evaded by applications which cipher their traffic. They are moreover under distress when a large number of signatures have to be recognized in a high bit rate traffic.

This claims for the design of lightweight techniques for traffic classification that do not rely on port numbers or DPI. There has been a large body of litterature on traffic classification [2]–[15] which is an evidence of the interest of the academics towards this topic. A few surveys are available, for example [16]–[18]. Our contribution to this already deeply investigated subject focuses on the architectural design of some ligthweight traffic classification techniques. The performance of different classification techniques has been deeply investigated in terms of the obtained classification rates (i.e. % of flows which are correctly associated to the generating application). But in spite of the plethora of litterature about traffic classification the question of how these methods must be implemented *in practice* to enable online traffic analysis has not received enough attention. There are a few studies that investigate the impact, for example, of packet or flow level subsampling [19] or feature selection [20] on classification accuracy. In [21], authors study the performance of a software version of the SVM algorithm and optimize the processing time in order to deal with high-speed links. But in general there is a lack of litterature about boosting lightweight traffic classification algorithms with hardware and/or software acceleration techniques.

We study the real implementation of a lightweight classification algorithm with a high-performance hardware accelerated solution. We take as a baseline the well-known Support Vector Machine (SVM) [22] which separates flows in a virtual space by hyperlanes. Flows are described by simple packet level features, in that case the size of the first three data packets in the flow [23] [24]. As any supervised classification method, the SVM algorithm consists of two main phases: *a training phase* and *a detection phase*. During the training phase, the algorithm starts from a learning trace labelled with categories of applications and computes the classification model namely the separating hyperplanes. Using this model, the detection phase decides of the category of application of new flows.

We consider a hardware accelerated implementation of the detection phase of the SVM algorithm. As the rate of traffic can be larger than tens of Gb/sec. in access and core networks, one needs to accelerate the detection algorithm to perform real time classification. We implement a SVM classifier on FPGA to boost the performance and adapt it to high rate capture

978-1-4577-1379-8/12/$26.00 ©2012 IEEE

points. To validate our contributions, we consider datasets generated in real networks and conduct extensive experiments.

II. Validation of SVM based traffic classification

A. Background on Support Vector Machine (SVM)

SVM [22] is a supervised classification algorithm. It transforms a non linear classification problem into a linear one, using a so called "kernel trick". Given a set of sample points in a multi-dimensional space one would like to separate them by hyperplanes, thus defining different classes. It is often impossible to separate sample points of different classes by hyperplanes and the separating surface is extremely difficult to compute. The idea of SVM is to map, by means of the kernel function, training points to a transformed space where it is possible to find separating hyperplanes. In the target space SVM must find the hyperplanes which separate points belonging to different classes and have a maximum distance between points of both classes and the separating hyperplane. The output of the training phase is made up of the parameters of the kernel and a set of support vectors x_i that define the separating hyperplane. During the detection phase SVM simply classifies new points according to the subspace they belong to.

SVM is often regarded as the best performing algorithm for traffic classification [16] [20] and has been adopted by several authors [15] [24] [21]. The accuracy depends on the selection of the kernel functions where Radial Basis Function (RBF) kernels usually give good results. We use in our implementations the LibSVM [25] library, which is an integrated software for support vector classification allowing multiclass classification, learning, cross-validation and different kernel functions. LibSVM implements different algorithms for applications of SVM to classification, to distribution estimation and to regression problems. Several algorithms exist for SVM-based classification. We have used the C-Support Vector Classification (C-SVC) algorithm [26].

Let us assume that we have a set of training points $x_i \in \mathbb{R}^n, i = 1, \ldots, l$ in two classes and a set of indicator values $y_i \in \{-1, +1\}$ such that $y_i = +1$ if x_i belongs to class 1 and $y_i = -1$ if x_i belongs to class 2. Let us also assume that we have selected a function ϕ such that $\phi(x_i)$ maps training point x_i into a higher dimensional space.

The training phase searches for an hyperplane that separates points $\phi(x_i)$ belonging to classes 1 and 2. The criterion is to maximize the distance of misclassified points to the separating hyperplane. The direction of the separating hyperplane is defined by a vector $w = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i)$ where only a few of coefficients α_i are non null. Non null coefficients define the so-called support vectors which characterize the separating hyperplane. The equation of the separating hyperplane is given by $w^T \phi(x) + b = 0$ that is $\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b = 0$ where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the so called "kernel" function.

In the detection phase, any new point x is classified according to the following decision function:

$$\text{sign}(w^T \phi(x) + b) = \text{sign}(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b) \quad (1)$$

x is classified into class 1 if $w^T \phi(x) + b$ is positive and into class 2 if $w^T \phi(x) + b$ is negative.

In this article we use the Radial Basis Function kernel as good results are often obtained with this kernel:

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \quad (2)$$

From this simple two-class SVM problem, one can easily deal with multi-class SVM classification problems. A usual approach is the so called "one versus one" (1 vs 1) approach. In this approach $\frac{n(n-1)}{2}$ two-class SVM problem are considered, one for each pair of classes. A training phase is performed for each two-class problem thus producing $\frac{n(n-1)}{2}$ separating hyperplanes. Each new point is then classified according to each of those two-class classification problems. The final decision is taken on the basis of a majority vote, that is to say that the new point is allocated to the class which has obtained the highest number of votes.

B. Accuracy of the SVM algorithm

In order to assess the accuracy of the SVM-based classifier we have performed validation over three different datasets. The learning and detection phases have been performed using the libSVM library [25]. The traffic descriptor that is used as input to the SVM classifier is made up of the size of the first three non empty packets of each flow, where a flow is defined as a set of packets with identical 5-tuples (IP Src adress, IP Dest adress, Src port, Dest port, protocol) [23] [24].

We have used for validation three datasets with groundtruth. The groundtruth identifies the application that has generated the traffic flow. It has been obtained either by Deep Packet Inspection (DPI) with for example Linux L7-filter [27] or by using a tool such as GT [11].

The characteristics of the three traffic traces used as benchmarks are listed in Table I. Those three traces correspond to three very different scenarios: campus network, laboratory environment and residential access network. As a consequence the composition of traffic is significantly different from one trace to the other.

1) The FT (France Telecom) dataset has been provided by France Telecom under the terms of a Non Disclosure Agreement. Traffic has been dumped on one geographical zone of an ADSL France Telecom access network and groundtruth has been established by DPI.
2) The Ericsson dataset corresponds to some trafic that has been generated in a laboratory environment of Ericsson research.
3) The Brescia dataset is a public dataset [11]. It corresponds to some traffic captured on a campus network. The groundtruth has been obtained with the GT tool.

The definition of classes is not universal. It mainly depends on the filters that have been defined for packet payload

Trace label	Network of capture	Bytes	Flows	Flows with a known classification	Capture mean rate (kb/s)
Ericsson	Local Area Network at an Ericsson Laboratory	1 755 816 952	39 056	12 858	315.18
Brescia	Campus trace generated at University of Brescia, Italy	746 850 665	153 237	76 182	1 042.9
FT	DSL Link of France Telecom	1 041 481 214	1 065 836	428 794	3 383.1

TABLE I
TRAFFIC TRACES AND THEIR PROPERTIES

inspection (DPI). In order to enable a comparison between traces we have merged applications into different categories that are listed in Table II.

Class label	Class name
1	Web
2	P2P download
3	Direct download
4	Streaming
5	Game
6	Mail
7	Instant messaging
8	Distant control

TABLE II
TRAFFIC CLASSES

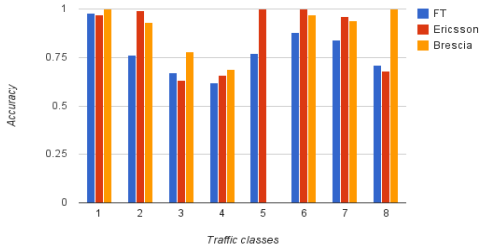


Fig. 1. Accuracy per traffic class

The traffic classification accuracy, that is to say the overall percentage of flows which are correctly classified is 94.43 % for the FT trace, 98.53 % for Ericsson and 97.41 % for Brescia.

A global accuracy figure is usually not considered as sufficient to demonstrate the performance of a classifier. Some classes could be frequently misclassified with not much impact on the global figure if only few flows correspond to those classes. A usual representation of results is given by the confusion matrix. Figure 1 provides the accuracy per category of applications, that is to say the percentage of flows of each category of applications that has been accurately classified.

As one can see from this figure, the accuracy of the SVM algorithm differs from one category of applications to another and from one trace to another. The proportion of a category of applications in a trace impacts the ability of the SVM algorithm to detect it. For example, as class 1 (Web) is present with a good proportion in all three traces, the accuracy of the detection is high. However, as class 4 (Streaming), is almost absent in the three traces it has the worst classification accuracy. Another reason for the low classification rate of Streaming traffic might be that the size of the first packets is not an accurate descriptor for this traffic.

III. TOWARDS ON-LINE TRAFFIC CLASSIFICATION

A. Requirements

In what follows the implementation of on-line SVM traffic classification is studied. In our scenario, probes are located on an operator access or core network to monitor traffic at packet level, reconstruct flows and classify them with a SVM. Only the detection phase of SVM is made on-line. The learning phase is made off-line periodically with a groundtruth generated by tools such as GT. We want to support data rates going up to tens of Gb/sec as equipments such as NetFPGA 10G [28] or COMBOv2 [29] are available to test algorithms at this speed.

The goal for on-line traffic classification is to handle all flows on a saturated 10 Gb/s link. Two main functions will be required to achieve this goal:

- The flow reconstruction reads each packet, identifies to which flow it belongs, and stores the packet lengths required for classification. The processing speed depends on the number of packets per second in the traffic.
- The SVM classification runs the SVM algorithm once for each received flow. The processing speed depends on the number of flows per second in the traffic.

The traces used to test the classification algorithm are described in Table I. The average sizes of packets and flows in bytes vary for these traces. Requirements in terms of packets/sec. and flows/sec. supported by the algorithm to reach a 10 Gb/s speed are described for each trace in Table III. To support each trace sent at 10 Gb/s, the flow reconstruction should support at least 55 861 124 packets/sec. and the SVM classifier should support at least 1 279 231 flows/sec.

We have first developed a software version of the classifier that is fed by a trace, to assess the possible performance in software. The classifier is made up of 3 main processes: (i) read the trace, (ii) rebuild flows from the stream of packets (iii) classify flows. For flow reconstruction, an algorithm proposed for a Netflow hardware implementation [30] is used.

Trace	Packets per second	Flows per second
Ericsson	6 809 840	27 805
Brescia	55 861 124	256 472
FT	48 310 718	1 279 231

TABLE III
REQUIREMENTS FOR A CLASSIFICATION AT 10 GB/S FOR EACH TRACE

It has the advantage of requiring a constant time per packet and a bounded memory, which fits well with a hardware implementation. For the SVM algorithm, the libSVM [25] library (written in C) was chosen. To use all the cores of the processor, openMP [31] for libSVM is enabled.

Table IV shows the performance of the software implementation on a 2.66 GHz 6-core Xeon X5650 with hyper-threading enabled and 12 GB of DDR3 RAM. It shows that the software implementation is not able to support 10 Gb/s. The best supported speed ranges from 2.32 Mb/s to 1597 Mb/s depending on the trace.

- The flow reconstruction speed does not depend on the trace as the flow reconstruction algorithm requires a constant time per packet. The only noticeable difference is for the biggest trace, FT, where the flow reconstruction probably suffers from the heavy CPU usage of the SVM classification.
- SVM classification is always more limiting than flow reconstruction (0.024 % of the requirements for 10 Gb/s in the worst case). Its speed depends on different factors including the number of support vectors in each SVM model: Brescia is the trace for which the learnt model has the most support vectors (24 758), then come FT (6 296) and Ericsson (4 341).

Trace	Packets per second (flow reconstruction)	Flows per second (classification)
Ericsson	5 189 293 76 % of 10Gb/s req.	4 655 17 % of 10Gb/s req.
Brescia	5 153 675 9.2 % of 10Gb/s req.	1 031 0.40 % of 10Gb/s req.
FT	4 336 677 9.0 % of 10Gb/s req.	311 0.024 % of 10Gb/s req.

TABLE IV
PERFORMANCE OF THE SOFTWARE IMPLEMENTATION COMPARED TO 10GB/S REQUIREMENTS

Even with a powerful computer, a software implementation is not able to reach a 10 Gb/sec. speed, mainly due to its limited ability to parallelize the computation. This justifies the use of hardware acceleration. Different platforms may be used to provide hardware acceleration for network monitoring:

- Network processors are programmable in software and provide hardware-accelerated tools for tasks commonly required in network monitoring.
- Programmable cards with an integrated Field-Programmable Gate Array (FPGA) are very flexible and provide hardware access to the network interfaces.

To be able to explore fully the parallelism possibilities in the SVM classification algorithm, we have chosen to use a card with an FPGA that is more flexible than network processors. Two main vendors provide such cards: NetFPGA with the NetFPGA 10G card, which has 4 interfaces at 10 Gb/s, and INVEA TECH with the Combov2 card, which has 2 to 4 interfaces at 10 Gb/s. Both cards integrate a Xilinx Virtex-5 XC5VTX240 FPGA. We are going to present an implementation of the flow reconstruction and SVM classification on this FPGA.

For flow reconstruction we use the same principles as [30] which considers an FPGA implementation of Netflow. SVM implementations on FPGA have also been proposed, but they are either focused on the learning phase [32] or not adapted to our classification algorithm [33] as they are restricted to two-class problems or using different kernels.

B. The SVM classification algorithm

The classification part of the SVM algorithm takes a vector as input and returns the class of that vector as an output. It works with few steps, repeated for each support vector. Algorithm 1 describes these steps. It is the multi-class implementation of the decision making procedure described in Section II-A. This pseudo-code has been written in order to enlight the possibilities to parallelize the algorithm.

Algorithm 1 SVM classification algorithm

$x \leftarrow$ the vector to classify
for all support vector x_i **do** {Main loop}
  $c_i \leftarrow$ the class of x_i
  $k_i \leftarrow K(x_i, x)$
  for all class $c_j \neq c_i$ **do** {Sum loop}
    $d \leftarrow$ index of the decision between c_i and c_j
    $S_d \leftarrow S_d + y_{d,i} \times \alpha_{d,i} \times k_i$
  end for
end for
for all decision d between c_i and c_j **do** {Comparison loop}
  if $S_d - b_d > 0$ **then**
    Votes $V_i \leftarrow V_i + 1$
  else
    Votes $V_j \leftarrow V_j + 1$
  end if
end for
Select class $c_n \leftarrow$ class with the highest votes V_n

The support vectors and the y, α and b values are part of the SVM model. Compared to the notations used in Section II-A, index d is added to identify the binary decision problem considered for the model values.

C. Operations

The important operations of the classification algorithm are in the main loop, for each iteration:

- the kernel computation (Equation 2) requires 3 integer additions (one per vector component), 3 integer multipli-

cations (to compute the squares), one multiplication by a floating-point constant, and one exponential computation.
- the sum computation requires one floating-point multiplication and one floating-point addition. It is run 7 times for the 8 classes defined in Table II.

D. Parallelism

The Main loop is where most of the computation time is spent. It iterates many times (from 4 341 to 24 758 support vectors for the different traces presented in Table I) and includes complicated operations (exponential, multiplications). But it can be easily parallelized as each iteration does not depend on others. The only shared data is in the additive S values. These values have to be duplicated so that iterations are computed in parallel. Then as S is additive, the duplicated values can be merged by summing them.

The Sum and Comparison loops have few iterations: one per class. In this article there are 8 classes defined in Table II, so the loops can be totally parallelized. The Sum loop iterations are totally independent, while the Comparison loop iterations share the votes counter, which is additive. So it can be duplicated and then merged.

All loops can be removed by using parallel processing except the main loop, that has too many iterations and would require more area than is available on the Virtex-5. But it is possible to implement more than once the Main loop, so that less iterations are required to process one vector. Section IV-B describes an architecture with an adjustable level of duplication of this loop.

IV. Hardware implementation

A. Fixed-point model

Floating-point operations are complex to realize in hardware and use too much area on the FPGA. The best solution is to transform the algorithm to use a fixed-point model instead of a floating-point model. Table V shows the bit widths of different variables used in the SVM fixed-point model.

Variable	Integer part	Decimal part
Vector component	11	0
α	7	11
γ	0	18
b	15	11
S	15	11

TABLE V
Quantization of the main SVM fixed-point model values

These quantization parameters have been chosen so that the mathematical operations are made on values as small as possible, without loosing too much precision for the classification. Some sizes are quite large because the classifier should work whatever the SVM model, so that a new synthesis is not required to change the model. The possible values of the variables have been determined by analyzing SVM models learnt in different conditions. For example the precision of the γ parameter is very important (decreasing it leads to a drop in classification accuracy), but its absolute value never reaches 1. The 11-bit width of a vector component has been chosen because we assume that the size of a packet will not be more than 1500 bytes.

Multiplications are complex to realize in hardware. They are required to compute the squares in the kernel, but squares are symmetric functions with one integer parameter varying from -1500 to 1500. A ROM with 1501 values is used to emulate squares. Similarly, a ROM is used to emulate the exponential function. Finally, to avoid the $y_{d,i} \times \alpha_{d,i} \times k_i$ multiplication, $ln\left(|y_{d,i} \times \alpha_{d,i}|\right)$ is precomputed, and the exponential used to compute k_i is computed only after the addition of this term. Delaying the exponential computation transforms the multiplication into an addition. This way only one multiplication by a constant remains in the kernel computation, which is much simpler than a multiplication of two variables.

To check that the loss in precision is not too important, a software implementation of the classification algorithm with the fixed-point model has been implemented. Figure 2 compares the accuracy of the fixed-point model to the results of the float model as described in Section II-B. It shows that the transition to fixed-point decreases the accuracy of the algorithm, but it remains around 90 %. Depending on the requirements, a higher accuracy can be achieved using wider fixed-point values, but it will require more space on the FPGA.

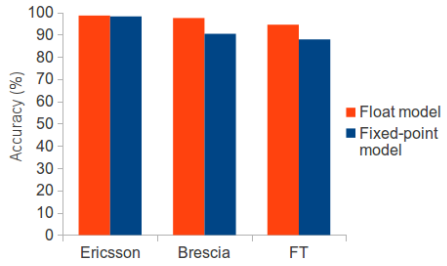


Fig. 2. Accuracy of the fixed-point model compared to the float model

B. Architecture

The architecture of a traffic-processing module on NetFPGA or Combov2 cards is very similar. It uses a block with an input bus for input traffic, and an output bus for output traffic. The classifier block is described in Figure 3.

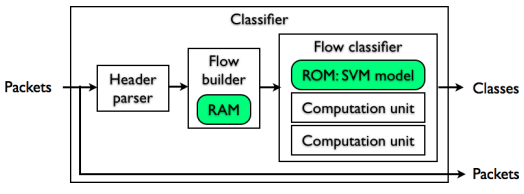


Fig. 3. Architecture of the classifier

Trace	Ericsson			Brescia		FT		
Computation units	2	4	8	2	8	2	4	8
Occupied slices	8 414	14 186	26 350	24 340	32 679	10 174	15 643	26 846
Occupied slice registers	9 221	21 864	45 967	9 272	40 658	8 966	21 287	44 356
FPGA usage (% of slices)	22.47	37.89	70.38	65.01	87.28	27.17	41.78	71.70
Maximum frequency (MHz)	174	156	165	51	62	157	164	139
Cycles per flow	2 193	1 110	569	12 401	3 121	3 170	1 598	813
Flows per second	79 733.6	140 766	290 107	4 122.42	20 164.3	49 741.0	102 840	171 861
% of 10Gb/s requirements	286.8	506.3	1043	1.61	7.862	3.888	8.039	13.43

TABLE VI
SYNTHESIS RESULTS OF SVM TRAFFIC CLASSIFICATION ON A VIRTEX-5 XC5VTX240 FPGA

The computation units represent the most important part of this architecture: they implement the computation of the main loop described in Algorithm 1. To get the best performance from the FPGA, operations of the algorithm must be parallelized. As seen in Section III-D, all loops can be totally unrolled by duplicating the hardware for each iteration except the main loop. The computation unit is duplicated as much as the Virtex-5 supports.

As the computation in the main loop is complicated, each iteration will take many clock cycles in hardware. To improve the throughput of the loop and reduce its computation time, the iterations can be pipelined: one new support vector is processed by the first operation of the computation unit at each clock cycle, and then forwarded to the next operation. This way all operations work in parallel and each computation unit accepts one support vector at each time step.

As Figure 3 shows, the SVM model is currently stored in ROMs. This forces to synthesize the design again to change the SVM model used. In future works, ROMs will be converted into RAMs, so that it is possible to change the SVM model faster. This will not require to change the implementation.

C. Results of the hardware accelerated traffic classifier

The proper behavior of the hardware implementation has been tested by checking that its results are exactly identical to the software implementation of the fixed-point model, first in simulation, and then implemented on a NetFPGA card. So the classification results of the hardware implementation are exactly the ones of figure 2. This section focuses on the performance in classified flows per second.

To assess the performance of the hardware implementation and compare it to the software implementation, it has been synthesized on a Virtex-5 XC5VTX240. Three different SVM models (one for each trace) have been tested. The number of processing units has been changed as well, to exploit the maximum parallelism on the FPGA. Table VI presents the results of these synthesis.

The number of occupied slices and slice registers as well as the maximum frequency are given by the synthesis tool. They are an indication of the hardware complexity of the implementation. The number of cycles required per flow has been determined by analyzing the code of the hardware implementation. It increases with the number of support vectors in the model, and decreases with the number of parallel computation units.

Thanks to massive parallelism, hardware implementations all have better performance in terms of flows per second than software implementations. The implementation for the Brescia trace gives poor results because of its low working frequency. The particularity of this trace is that the SVM model contains more support vectors than the others. They use too much space on the FPGA, which creates long and slow routes in the design an decrease its maximum frequency. The Ericsson and FT traces SVM models have less support vectors. Even with only 2 computation units, the implementation for the Ericsson trace gives results much higher than the requirements to support a 10 Gb/s speed (286 % of the requirements). The implementation for the FT trace brings roughly the same performance improvements, but the requirements in terms of flows per second are very high because the trace contains many very small flows. So with 8 computation units, it fulfills only 13 % of the requirements.

A performance result that is not visible in Table VI is the delay that the classifier adds to the packets if it is used directly on a network link to tag packets with their class number (and not just as a passive probe set up in derivation on the link). The current implementation sends the classification to the host computer instead of tagging the packets, but it could be modified without overhead. For now, 10 clock cycles are required between the arrival of the packet in the classifier and the time when the class of the packet is known. This delay is constant because a packet is considered unknown if the flow has not yet been classified (its class is not in RAM), so it does not depend on the classification time. At the frequencies obtained in Table VI, the delay ranges from 57.2 ns to 254 ns. These figures do not include the time required to handle the Ethernet protocol.

To improve the supported speed for all traces, many directions are possible. A better management of the memory used to store the SVM models, by reducing it or putting it in an external memory, would bring more room to parallelize the logic on the FPGA. Critical paths in the design may be improved to achieve higher frequencies. It is also possible to use more powerful FPGAs, or to use multiple FPGAs in parallel to reach the 10 Gb/s speed for each trace. Algorithmic changes to reduce the number of support vectors may help too.

V. Conclusion and Future Works

We have studied the practical implementation of SVM-based traffic classification. We have adapted the classifier to high rate traffic with hardware acceleration of SVM on a FPGA. Our hardware implementation outperforms tremendously the software implementation and allows real-time classification. In the future, we will enhance our FPGA implementation to support higher speeds and dynamic classification models and test it within a production network. It would also be interesting to compare our results to more optimized software implementations such as [21] [34]. Massively parallel implementations of SVM also exist using GPUs [35], [36], and recent works have shown the ability to process packets with a GPU [37]. Globally, this work is a step further to the real implementation of a classification architecture.

Acknowledgment

This work has been partly funded by the EU FP7 DEMONS (257315) project and by the VIPEER project of the french National Research Agency (ANR-09-VERS-014).

References

[1] CISCO Systems, "Cisco Visual Networking Index: Forecast and Methodology, 2010-2015."

[2] M. M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement (IMC'04)*. New York, NY, USA: ACM, 2004, pp. 135–148.

[3] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: Automated construction of application signatures," in *In SIGCOMM 2005 MineNet Workshop*, 2005.

[4] A. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, ser. SIGMETRICS '05. New York, NY, USA: ACM, 2005, pp. 50–60.

[5] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: multilevel traffic classification in the dark," *SIGCOMM Comput. Commun. Rev.*, vol. 35, pp. 229–240, 2005.

[6] L. Bernaille, R. R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," *SIGCOMM Comput. Commun. Rev.*, vol. 36, pp. 23–26, 2006.

[7] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, ser. MineNet '06. New York, NY, USA: ACM, 2006, pp. 281–286.

[8] H. Dahmouni, S. Vaton, and D. Rossé, "A markovian signature-based approach to IP traffic classification," in *MineNet 2007 : ACM Sigmetrics Workshop on Mining Network Data*, 2007, pp. 29 – 34.

[9] T. Auld, A. Moore, and S. Gull, "Bayesian neural networks for internet traffic classification," *Neural Networks, IEEE Transactions on*, vol. 18, no. 1, pp. 223 –239, 2007.

[10] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," *SIGCOMM Comput. Commun. Rev.*, vol. 37, pp. 5–16, 2007.

[11] F. Gringoli, L. Salgarelli, M. Dusi, N. Cascarano, F. Risso, and K. Claffy, "GT: picking up the truth from the ground for Internet traffic," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 5, pp. 13–18, 2009.

[12] A. Este, F. Gringoli, and L. Salgarelli, "On the stability of the information carried by traffic flow features at the packet level," *SIGCOMM Comput. Commun. Rev.*, vol. 39, June 2009.

[13] W. Li, M. Canini, A. Moore, and R. Bolla, "Efficient Application Identification and the Temporal and Spatial Stability of Classification Schema," *Computer Networks, Special Issue on Traffic classification and its applications to modern networks*, vol. 53, pp. 790–809, 2009.

[14] C. Rotsos, J. J. Van Gael, A. Moore, and Z. Ghahramani, "Probabilistic graphical models for semi-supervised traffic classification," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ser. IWCMC '10. New York, NY, USA: ACM, 2010, pp. 752–757.

[15] P. Bermolen, M. Mellia, M. Meo, D. Rossi, and S. Valenti, "Abacus: Accurate behavioral classification of P2P traffic," *Elsevier Computer Networks*, vol. 55, no. 6, pp. 1394–1411, 2011.

[16] H. Kim, D. Barman, M. Faloutsos, M. Fomenkov, and K. Lee, "Internet traffic classification demystified: The myths, caveats and best practices," in *In Proc. ACM CoNEXT*, 2008.

[17] T. Nguyen and G. Armitage, "A survey of techniques for Internet traffic classification," *IEEE Communications Surveys and Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[18] S. Valenti, *Dealing with P2P traffics in modern networks: measurement, identification and control*. Paris: TELECOM ParisTech, 2011.

[19] D. Rossi and S. Valenti, "Fine-grained traffic classification with netflow data," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ser. IWCMC '10, 2010, pp. 479–483.

[20] N. Williams, S. Zander, and G. Armitage., "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *ACM SIGCOMM Computer Communication Review*, 2006.

[21] A. Este and F. Gringoli, "On-line svm traffic classification," in *Proceedings of the 7th IWCMC Conference (IWCMC TRAC'2011)*, 2011.

[22] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[23] G.Gomez and P.Belzarena, "Early Traffic Classification using Support Vector Machines," in *Fifth International Latin American Networking Conference (LANC'09)*, 2009.

[24] A. Este, F. Gringoli, and L. Salgarelli, "Support vector machines for tcp traffic classification," *Computer Networks*, vol. 53, no. 14, pp. 2476 – 2490, 2009.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[26] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, 1992, pp. 144–152.

[27] Clear Foundation, "l7-filter: application layer packet classifier for Linux," http://l7-filter.clearfoundation.com/.

[28] NetFPGA, "NetFPGA: a line-rate, flexible, and open platform for research, and classroom experimentation," http://netfpga.org/.

[29] CESNET, "Our hardware," http://www.liberouter.org/hardware.php?flag=U, Nov. 2011.

[30] M. Žádník and L. Lhotka, "Hardware-accelerated netflow probe," Technical Report 32/2005, CESNET, Praha, Tech. Rep., 2005.

[31] L. Dagum and R. Menon, "Openmp: an industry standard api for shared-memory programming," *Computational Science Engineering, IEEE*, vol. 5, no. 1, pp. 46 –55, jan-mar 1998.

[32] D. Anguita, A. Boni, and S. Ridella, "A digital architecture for support vector machines: theory, algorithm, and fpga implementation," *Neural Networks, IEEE Transactions on*, vol. 14, no. 5, pp. 993 – 1009, sept. 2003.

[33] D. Anguita, S. Pischiutta, S. Ridella, and D. Sterpi, "Feed-forward support vector machine without multipliers," *Neural Networks, IEEE Transactions on*, vol. 17, no. 5, pp. 1328 –1331, sept. 2006.

[34] F. Gringoli, L. Nava, A. Este, and L. Salgarelli, "MTCLASS: enabling statistical traffic classification of multi-gigabit aggregates on inexpensive hardware," in *Proceedings of the 8th IWCMC Conference (IWCMC TRAC'2012)*, 2012.

[35] B. C. Catanzaro, N. Sundaram, and K. Keutzer, "Fast support vector machine training and classification on graphics processors," in *Technical Report No. UCB/EECS-2008-11*. EECS Department, University of California, Berkeley, 2008.

[36] A. Carpenter, "Cusvm: A cuda implementation of support vector classification and regression," http://patternsonascreen.net/cuSVM.html.

[37] S. Han, K. Jang, K. Park, and S. Moon, "PacketShader: a GPU accelerated software router," http://shader.kaist.edu/packetshader/.

9.5 Article IEEE SPL sur la conception d'un opérateur d'inverse en virgule fixe [Lib+17b]

A Scaling-Less Newton–Raphson Pipelined Implementation for a Fixed-Point Reciprocal Operator

Erwan Libessart, Matthieu Arzel, Cyril Lahuec, and Francesco Andriulli, *Senior Member, IEEE*

***Abstract*—The reciprocal is a widespread operation in digital signal processing architectures. A usual method consists in using the Newton–Raphson algorithm or its derivatives, either in floating or in fixed-point formats. With the former format, the standardized format of the mantissa makes the implementation easier, but for the fixed-point format there are many possibilities. This forces a design with scaling of the input in order to respect a predetermined work range. Having the input in a known range makes it possible to compute a first approximation with coefficients stored in memory blocks. With this method, it is hard to propose a "ready to use" IP for all the fixed-point formats. In this letter, a novel architecture, which does not require scaling, is proposed. This design is totally pipelined, ROM-less and can be directly used in any architecture. The implementation was optimized to reach a maximum clock frequency of 740 MHz on a Virtex-7 Field-Programmable Gate Array (FPGA).**

***Index Terms*—Fixed-point representation, FPGA, leading one detector, Newton–Raphson, reciprocal.**

I. Introduction

DIVISION is a key operation in many DSP algorithms, but with a dramatic cost in term of hardware resources. The common way to compute division is to calculate the reciprocal of the divisor and then multiply the result by the dividend. The implemented algorithms can use the floating-point format or the fixed-point format but the latter is generally more efficient in execution time and resources used. Different methods can be employed to compute the reciprocal. Look-up tables or digit recurrence methods such as Sweeney, Robertson, and Tocher (SRT) can be used but there are some issues, for example, necessary resources or latency, when the size of the input increases [1]. Another solution is the Newton–Raphson method, which allows to roughly double the number of bits of accuracy at each iteration, with an acceptable amount of necessary resources. However, it requires input scaling in a certain working range, for example, $[1, 2]$ or $[0.5, 1]$ [2], [3]. Of course, rescaling is then mandatory. The computation and the storage of coefficients are often required to compute the first iteration. This means that additional work is necessary to integrate this operator as an IP core in a whole signal processing architecture. Avoiding the scaling makes the integration easier but the classical methods for the mantissa's reciprocal cannot be used in such a case.

In this letter, a way of using the Newton–Raphson algorithm for fixed-point reciprocal computation without scaling the data is proposed. So, a generic IP core, which is adapted for any design, can be proposed. This is possible by computing a first approximation of the reciprocal, which allows meeting the condition of convergence, whatever the value of the algorithm's input is. Moreover, the proposed method does not require any memory block to store coefficients and can be fully pipelined, which permits high-frequency computing.

This letter is organized as follows. Section II presents the Newton–Raphson algorithm and its typical use. The technique that allows to avoid the scaling in described in Section III. Then, the implementation of the whole Newton–Raphson algorithm is described in Section IV. Section V presents the results of the scaling-less Newton–Raphson FPGA implementations. Finally, Section VI concludes the letter.

II. Related Work

The Newton–Raphson method is an iterative algorithm to compute the reciprocal of a number a. At each iteration, the number of bits of accuracy is roughly doubled. x_n, the final estimation of $\frac{1}{a}$, is obtained after n iterations of this equation:

$$x_{i+1} = x_i(2 - ax_i), \tag{1}$$

where x_0 is the first approximation of $\frac{1}{a}$, and is another input of the algorithm. This method is usable for both formats of representation: fixed-point and floating-point. The usual approach is to consider that a belongs to a predefined interval such as $[1, 2]$, which is the range of the floating-point's mantissa, $[0.5, 1]$ or $[0, 1]$ [2]–[5]. With a belonging to a predetermined range, it is possible to compute the first approximation with great accuracy by using a method based on a polynomial approximation. So, this first step requires the use of a memory block in order to store the different coefficients. With this method, the obtained accuracy for the first approximation decreases the number of necessary Newton–Raphson iterations. This approximation may be rather complex and this strategy requires scaling at the input and output of the algorithm. So, this method requires pre- and postprocessing in order to manage the scaling that brings

Manuscript received February 10, 2017; accepted April 1, 2017. Date of publication April 12, 2017; date of current version April 24, 2017. This work was supported by Labex CominLabs through SABRE project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Cavallaro. *(Corresponding author: Erwan Libessart.)*

The authors are with the IMT Atlantique, Brest 29238, France (e-mail: erwan.libessart@imt-atlantique.fr; matthieu.arzel@telecom-bretagne.eu; cyril.lahuec@telecom-bretagne.eu; francesco.andriulli@telecom-bretagne.eu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2694225

additional resources and constraints, and consequently additional integration time.

In this work, an alternative method for fixed-point representation without any scaling of the data, whatever the initial value, is proposed. So, the scaling-less Newton–Raphson architecture is "ready to use" and does not need any additional element.

III. A Technique to Avoid the Scaling

A. 1-Bit Approximation Using a Leading One Detector

The main idea is based on the condition that has to be respected in order to be sure that the algorithm converges:

$$0 < a \times x_0 < 2. \tag{2}$$

The strategy presented here consists in adapting the computation of x_0 to a, whereas in the literature a is adapted to the method of x_0's computation, by being in a predefined range. To the authors' knowledge, no other technique to avoid the scaling can be found in the literature.

It is important to note that the closer to 1 the product is, the faster the convergence is. The format of a is $uQm.p$, which means that it is an unsigned number with m integer bits and p fractional bits. It is then extended to the $uQn.n$ format with $n = \max(m, p)$. Thus, the operator has an absolute accuracy of 2^{-n} and the output $\frac{1}{a}$ can be represented in $uQn.n$ format too.

According to the condition presented above, a is represented in base 2 as

$$a = a_{n-1}a_{n-2}\ldots a_0 a_{-1}\ldots a_{-n}. \tag{3}$$

Then, let j be the index of the leading one of a. So, a respects the following inequality:

$$2^j \leq a < 2^{j+1}. \tag{4}$$

Then, a value for x_0 which respects the condition of convergence can be deduced:

$$x_0 = 2^{-(j+1)}. \tag{5}$$

From (4) and (5),

$$0.5 \leq a \times x_0 < 1. \tag{6}$$

Equation (6) implies that x_0 approximates $\frac{1}{a}$ with one bit of accuracy. This value of x_0 can be easily obtained by applying a leading one detector (LOD) on a and then taking the bit-reversal (BR) of the result.

This LOD can be easily computed by an architecture that cascades AND and NOT gates.

Such an architecture is well-adapted for application-specified integrated circuit (ASIC) implementation. For an FPGA target, it could be better to use another solution. The method presented here consists in the following steps:

1) Take the bit-reversal of a named BR(a).
2) Calculate the two's complement of BR(a).
3) Apply a bitwise AND between this two's complement and BR(a).

This method suits the FPGA target because it can take benefit from the carry propagation blocks for the two's complement operation to achieve higher clock frequency than with the conventional cascaded solution. The result of this bitwise operation gives exactly the value $2^{-(j+1)}$, which is the desired value for x_0. The proof is quite simple. From (4), a is represented like this:

$$a = 0\ldots 01 a_{j-1}\ldots a_{-n}. \tag{7}$$

So, BR(a) is equal to

$$\mathrm{BR}(a) = a_{-n}\ldots a_{j-1}10\ldots 0. \tag{8}$$

And, for its two's complement TC(BR(a)):

$$\mathrm{TC}(\mathrm{BR}(a)) = \overline{a_{-n}}\ldots\overline{a_{j-1}}10\ldots 0. \tag{9}$$

And, then the bitwise AND operation gives the expected value for x_0 as below:

$$a_{-n}\ldots a_{j-1}10\ldots 0 \quad \& \quad \overline{a_{-n}}\ldots\overline{a_{j-1}}10\ldots 0 = 2^{-(j+1)}. \tag{10}$$

This method is not dependent on the size of the input, so it is possible to implement it in a generic way with a hardware description language (HDL). Moreover, this architecture requires only one clock cycle to deliver the result and achieve high clock frequency (as it will be shown in the next section).

B. 2-Bit Approximation

In the literature, the usual strategy is to optimize the number of bits of accuracy for x_0 [2], [3]. It enables to reduce the number of Newton–Raphson iterations required to reach the desired accuracy. So, the method presented above has to be improved. In fact, it is more interesting to use a simple combinatorial circuit to gain this second bit of accuracy than to compute a whole Newton–Raphson iteration that contains two multiplications as shown in (1). This combinatorial circuit can be deduced easily by distinguishing the possible cases for a_{j-1}. Having $a_{j-1} = 1$ implies

$$1.5 \times 2^j \leq a < 2^{j+1}, \tag{11}$$

and then, if x_0 is still equal to $2^{-(j+1)}$

$$0.75 \leq a \times x_0 < 1, \tag{12}$$

which means that in this case the approximation already has 2 bits of accuracy.

Now, if $a_{j-1} = 0$,

$$2^j \leq a < 1.5 \times 2^j, \tag{13}$$

and then, if x_0 is still equal to $2^{-(j+1)}$

$$0.5 \leq a \times x_0 < 0.75. \tag{14}$$

So, x_0 can be multiplied by 1.5 in order to have the 2 expected bits of accuracy, so that

$$x_0 = 2^{-(j+1)} + 2^{-(j+2)}, \tag{15}$$

and then $0.75 \leq a \times x_0 < 1.125$. So, the condition on x_0 to have the required accuracy is that the bit at the index $-(j+2)$ must be the complement of a_{j-1}. The implementation of this condition is quite simple and requires only one more cycle of computation and one bit in the fractional part in order to be sure not to be confronted to rounding error. This modification of the generation of x_0 reduces the number of Newton–Raphson iterations by 1. As the result, a simple way to approximate the reciprocal of a with a generic combinatorial architecture can be

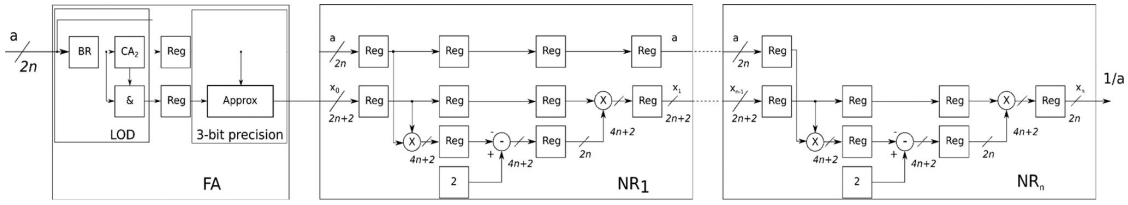


Fig. 1. Architecture for the reciprocal operator.

determined. The same kind of reasoning can be used to have a accuracy of 3 bits with

$$x_{0_{-(j+1)}} = a_j, x_{0_{-(j+2)}} = \overline{a_{j-1}}, x_{0_{-(j+3)}} = \overline{a_{j-2}}. \quad (16)$$

And. therefore, $0.875 \leq a \times x_0 < 1.125$.

Henceforth, the 3-bit-accuracy solution is considered. In fact, in the literature, the typical case for the reciprocal operator computes a 16-bit input. So, using the 3-bit-accuracy solution requires three Newton–Raphson iterations to be sure to reach the desired accuracy. Five and four iterations are required with the 1-bit and 2-bit-accuracy solutions, respectively.

IV. Reciprocal Architecture

Fig. 1 shows the scaling-less Newton–Raphson architecture for the reciprocal operator. The FA (first approximation) block is the 3-bit-accuracy FA block described in Section III, which transmits the value of a and delivers x_0 in $2n+2$ bits in two clock cycles. The LOD implementation for the FA block is adapted for any input sizes. The LOD architectures in [6], [7] use an elementary 4-bit or 8-bit LOD block, which is used to implement the others 2^n-LOD blocks. This means that more resources than necessary are used for input sizes between two powers of 2. Concerning this aspect, the solution presented in this letter is more efficient.

A Newton–Raphson iteration block (NR_i) in Fig. 1 computes the ith evaluation x_i (1). A block first computes $2 - ax_n$ and then multiplies by x_n. Each of these blocks contains two multipliers and is totally pipelined. Bus reductions are required during the process. In fact, the different arithmetic operations generate useless bits for the required accuracy so these can be ignored. All the blocks are identical, except for the last one, which has not to transmit the value of a.

V. Implementation Results

This section is composed of three parts. First, the chosen LOD method in Section III and the one proposed in [6] are implemented and compared on a Virtex-7 690T FPGA. Then, the implementation results for the reciprocal operator on the same target are presented. Finally, the scaling-less Newton–Raphson architecture is implemented on a Virtex-4 SX35-12 FPGA, the same as in [4], which also aims at increasing the computation frequency and so is a good reference for a comparison.

A. First Approximation

The 1-bit FA block consists of an LOD. The proposed method and the work of [6] were implemented into the FPGA Virtex-7 690T. The obtained comparison is shown in Fig. 2.

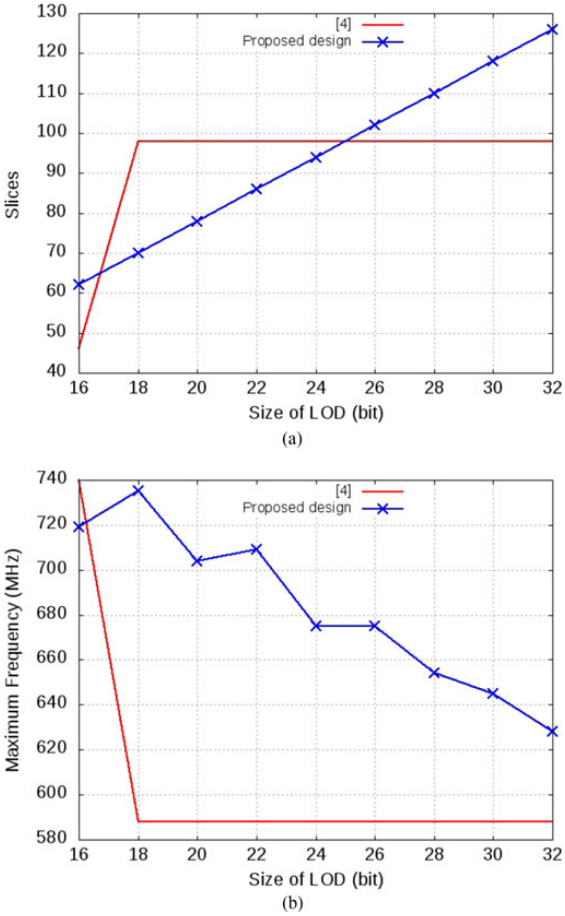


Fig. 2. Plot of resources (number of slices) (a) and maximum frequency (b) for various sizes of LOD.

The design proposed in [6], which is the result of a genetic algorithm, is better in term of resources and frequency for a 16-bit LOD. For larger inputs, the 32-bit architecture has to be used. This explains the visible step on the plots. The design proposed in this letter can be used for all input sizes. This allows us to be more efficient in resources and maximum frequency for 18- to 24-bit inputs. The proposed architecture has a better maximum frequency for each size larger than 16 bits. So, this architecture is used henceforth.

TABLE I
FPGA IMPLEMENTATION RESULTS FOR THE FIRST APPROXIMATION ON VIRTEX-7 690T

	1-bit FA	1-bit FA+NR	2-bit FA
LUT	31	31	47
Flip-Flop	31	64	64
DSP	0	2	0
Clock cycles	1	7	2
Max. frequency (MHz)	719	740	714

TABLE II
IMPLEMENTATION RESULTS: 16-BIT RECIPROCAL ON VIRTEX-7 690T

	Proposed design with 1-bit FA	Proposed design with 3-bit FA
LUT	159	111
Flip-flop	325	240
DSP	10	6
Clock cycles	31	20
Need of scaling	No	No
Pipelined	Yes	Yes
Coefficients storage	No	No
Max. frequency (MHz)	740	740

Table I presents the implementation results for 1-bit FA, 1-bit FA followed by an Newton-Raphson (NR) iteration and 2-bit FA blocks. The 2-bit FA requires a 50% increase in Look-Up Table (LUT), an additional clock cycle and twice more registers, compared to the 1-bit FA block. But it replaces a Newton–Raphson iteration, which requires two DSP cells and five additional clock cycles for the same result. The only benefit is the gain of a few LUTs and a slight increase of the maximum clock frequency.

B. Reciprocal Operator

The global architecture for the reciprocal operator was also implemented on FPGA target. The implementation is generic, so that the number of iterations and the size of the input can be chosen. The design is totally pipelined, in order to maximize the computation frequency. The reciprocal architecture was implemented on the Virtex-7 690T FPGA. The results are presented in Table II and highlight the difference between using the 1-bit FA and the 3-bit one for a 16-bit reciprocal operator. As the accuracy roughly doubles with each iteration, the 1-bit FA solution needs five iterations, whereas the 3-bit FA one only needs three iterations. The input size permits to use the DSP48E1 of the Virtex-7 FPGA for the multiplications required by the algorithm. DSP48E1 contains a 25×18 two's complement multiplier. So, for larger input sizes, LUT-based multipliers or DSP combinations have to be used. This DSP cell allows to reach great frequency clock but by adding some latency. A maximum frequency of 740 MHz can be reached. This is the maximum frequency that can be obtained when all the registers of DSP48E1 cells are used [8], [9]. This optimal use of DSP cells explains why stages of registers were added compared to the architecture presented in Fig. 1.

TABLE III
IMPLEMENTATION RESULTS: 16-BIT DIVISION ON VIRTEX-4 SX35

	Proposed design with 3-bit FA	[4]
Slices	347	1478
LUT	372	2091
Flip-flop	568	1820
DSP	7	7
Clock cycles	25	112
Need of scaling	No	Yes
Pipelined	Yes	Yes
Coefficients storage	No	Yes (34Kb)
Max. frequency (MHz)	294.1	294.1

This solution is totally pipelined and does not require any coefficients storage. Moreover, it can be used for each even sizes of the input. The architecture is easily portable to all the FPGA targets. Very few LUTs and DSPs are used (less than 0.1% of the Virtex-7 690T capacity).

C. Comparison With Another High-Frequency Design

The scaling-less Newton–Raphson architecture is compared to [4], which aims at high-frequency computation on a Virtex-4 SX35-12 FPGA using a minimax polynomial approximation and binary arithmetic and by using the same iteration method. The design was ported on the same target and modified to a complete division operator by adding a shift register for the 16-bit dividend and the final multiplication. On the basis of the architecture given in Fig. 1, some register stages have been added in the NR blocks in order to fit to the DSP cells architecture. This allows, as [4], to reach the frequency of 294.1 MHz. So, there is a fix point for the comparison. Moreover, both architecture use seven DSP cells. So, the resources and latency comparison can legitimately be made. The results are presented in Table III.

Finally, the scaling-less Newton–Raphson architecture requires 25 clock cycles to compute the division, whereas that in [4] requires 112 cycles. Both design use seven DSP cells but the proposition of this letter reduces the number of LUTs in 82% and the number of Flip-flop in 69%. No scaling or ROM block is necessary. Thus, the proposed design is more flexible and its integration as an IP core is easier.

VI. CONCLUSION

This letter presents a fixed-point implementation perspective for the Newton–Raphson algorithm. We proved that it is possible to have a design in which no scaling is obligatory by adapting the first approximation, which allows to implement a "ready to use" IP core. This adaptation is made with a LOD operator, which is size generic in order to be more flexible and save resources. This LOD operator can be made with a two's complement on FPGA to benefit from carry propagation blocks. The final reciprocal operator design is a low-resources and totally pipelined architecture. In addition to the absence of obligatory scaling, the proposition does not require any ROM block to store coefficients. Thus, this is an architecture that can be used directly and easily in every digital signal processing architecture regardless of the representation format of the input. Moreover, the presented work is available as an open-source project [10].

REFERENCES

[1] S. F. Obermann and M. J. Flynn, "Division algorithms and implementations," *IEEE Trans. Comput.*, vol. 46, no. 8, pp. 833–854, Aug. 1997.

[2] A. Rodriguez-Garcia, L. Pizano-Escalante, R. Parra-Michel, O. Longoria-Gandara, and J. Cortez, "Fast fixed-point divider based on Newton-Raphson method and piecewise polynomial approximation," in *Proc. Int. Reconfigurable FPGAs*, pp. 1–6, Dec. 2013.

[3] H. C. Neto and M. P. Vestias, "Very low resource table-based FPGA evaluation of elementary functions," in *Proc. Int. Reconfigurable FPGAs*, pp. 1–6, Dec. 2013.

[4] M. P. Vestias and H. C. Neto, "Revisiting the Newton-Raphson iterative method for decimal division," in *Proc. Int. Field Programmable Logic Appl.*, pp. 138–143, Sep. 2011.

[5] M. Ito, N. Takagi, and S. Yajima, "Efficient initial approximation for multiplicative division and square root by a multiplication with operand modification," *IEEE Trans. Comput.*, vol. 46, no. 4, pp. 495–498, Apr. 1997.

[6] K. Kunaraj and R. Seshasayanan, "Leading one detectors and leading one position detectors—An evolutionary design methodology," *Can. J. Electr. Comput. Eng.*, vol. 36, no. 3, pp. 103–110, 2013.

[7] K. H. Abed and R. E. Siferd, "VLSI implementations of low-power leading-one detector circuits," in *Proc. IEEE SoutheastCon, 2006*, pp. 279–284, Mar. 2006.

[8] 7 Series DSP48E1 Slice-User Guide. Xilinx, 2016. Accessed on: February 10, 2017.

[9] Virtex-7 T and XT FPGAS Data Sheet: DC and AC Switching Characteristics. Xilinx, 2016. Accessed on: February 10, 2017.

[10] (2017). [Online]. Available: https://redmine.telecom-bretagne.eu/projects/scaling-less-newton-raphson

9.6 Article IEEE TCAS1 sur l'implantation d'un circuit flexible pour l'inférence de réseaux de neurones à cliques en CMOS 65nm [Lar+18a]

A Fully Flexible Circuit Implementation of Clique-Based Neural Networks in 65-nm CMOS

Benoit Larras, *Member, IEEE*, Paul Chollet, Cyril Lahuec, Fabrice Seguin, and Matthieu Arzel

(Invited Paper)

***Abstract*— Clique-based neural networks implement low-complexity functions working with a reduced connectivity between neurons. Thus, they address very specific applications operating with a very low-energy budget. However, the implementation in the state of the art is not flexible and a fabricated circuit is only usable in a unique use case. Besides, the silicon area of hardwired circuits grows exponentially with the number of implemented neurons that is prohibitive for embedded applications. This paper proposes a flexible and iterative neural architecture capable of implementing multiple types of clique-based neural networks of up to 3968 neurons. The circuit has been integrated in an ST 65-nm CMOS ASIC and occupies a 0.21-mm^2 silicon surface area. The proper functioning of the circuit is illustrated using two application cases: a keyword recovery application and an electrocardiogram classification. The neurons outputs are updated 83 ns after a stimulation, and a neuron needs an energy of 115 fJ to propagate a change at the input to its output.**

***Index Terms*— Neural networks circuit, clique-based neural networks, analog/mixed-signal circuit, iterative circuit structure, classification circuit.**

I. Introduction

A. Different Types of Artificial Neural Networks and Their Applicative Contexts

NEURO-INSPIRED computing is an alternative to Von Neumann computing for solving multiple-variable problems in a limited amount of both operations and energy [1]. They consist of associations, additions and subtractions of weighted signals, and activations upon threshold. Applications relying on these operations are thus benefiting from using neuro-inspired computing implementations. The application domains are very diverse, such as financial data prediction [2], voice-activity detection (VAD) [3]–[5] and bio-medical signals classification [6]. Moreover, in embedded applications where energy is at stake, neuro-inspired computing is favored by using circuits implementing Artificial Neural Networks (ANNs), using different neural network models [7]–[9]. For this purpose, ANN circuit implementations have to combine low-energy designs and low-complexity algorithms. This is achieved when the circuit does not mimic the behavior of the neurons at the biological level, for example, as in [10]–[12], but instead at the information exchange level, as in [13]–[17].

Multiple computations can be performed in parallel using a vast number of fully-connected neurons. These are efficient, for instance, at decomposing image prior to searching for a specific pattern in a subset of pixels [18]; however, their complexity is further increased when implementing synaptic weights. Composed of about 1 million neurons, circuits implementing such large ANNs can be termed "neural processors" [13]–[16]. Oversized for tasks such as embedded associative memories, they are advantageously replaced by simpler clique-based ANNs, either clustered [19] or Willshaw-Palm networks [20]. Specific classification tasks can be performed by networks of few thousands neurons [21], [22]. These ANNs are not fully connected, have binary weighted synapses and use simple activation functions. Clustered networks are formed by grouping neurons per category of information, Figure 1(a). Neurons from different clusters are connected together to form a clique. A local "Winner-Takes-All" (WTA) rule activates or not the neurons. Neurons in Willshaw-Palm networks are not clustered, Figure 1(b). The cliques are formed by connecting k neurons together. The activation function is then a global "k-Winners-Take-All." Low complexity and, hence, low power consumption are better exploited by implementing the neurons using analog CMOS circuits [19]. Unfortunately, creating the neural connections *a priori* implies that they are not reconfigurable. This limits their uses for patterns recognition in a non-variable context once the system is calibrated. Proper calibration avoids the necessity of a self-configuration feature.

B. Contributions and Problems Addressed by Present Work

The connections between neurons in implemented clique-based ANNs are hardwired. Even though it offers the best performance in terms of energy consumption, the main drawback is the lack of flexibility. Once the circuit has been produced, it cannot be modified and is thus very task-specific. This is problematic in an application requiring a change in the classification parameters over time.

Manuscript received May 5, 2018; revised August 3, 2018 and September 17, 2018; accepted November 7, 2018. Date of publication December 14, 2018; date of current version April 15, 2019. This work was supported in part by the Research Institute Pracom, in part by the SENSE CominLabs Project, and in part by the Brittany Regional Council Grant CPER. This paper was recommended by Associate Editor A. James. *(Corresponding author: Benoit Larras.)*

B. Larras is with the Department of SMART, ISEN Lille, 59046 Lille, France. He is also with CNRS, Institute of Electronics, Microelectronics and Nanotechnologies, 59650 Villeneuve d'Ascq, France (e-mail: benoit.larras@yncrea.fr).

P. Chollet is with Télécom ParisTech, 75013 Paris, France (e-mail: pachollet@telecom-paristech.fr).

C. Lahuec, F. Seguin, and M. Arzel are with the Department of Electronics, IMT Atlantique, 29238 Brest, France, and also with Lab-STICC, CNRS, 29238 Brest, France (e-mail: cyril.lahuec@imt-atlantique.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2018.2881508

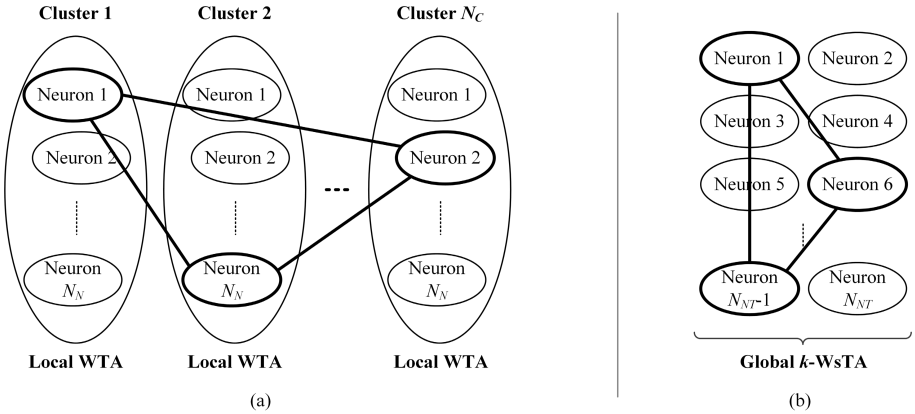


Fig. 1. Topologies of clique-based neural networks. (a) is a clustered clique-based network composed of N_C clusters of N_N neurons each. (b) is a Willshaw-Palm network composed of N_{NT} neurons in total. For both cases, a clique is highlighted in black.

Moreover, in certain applications, it is not efficient to separate the network in clusters. For example, in electrocardiograms (ECG) classification, several classification parameters have 2 or 3 possible values, which leads to 2 or 3 neurons per cluster, respectively [23]. Yet, clusters of small size reduce the number of cliques that can be stored in the network [24]. In that case, using the k-WsTA activation rule proves to be more efficient. In order to address a maximum of applicative use-cases with a single circuit, it is therefore interesting to implement both WTA and k-WsTA activation rules.

A flexible circuit addressing both aforementioned issues, by implementing Willshaw-Palm networks as well as clustered networks, is presented in [25]. A 2-stage WTA architecture is designed to perform the neural activation function. By iterating the WTA operation several times, it is possible to either reproduce the iterative process of message recovery in clique-based networks, or emulate different clusters by storing the temporary data in a dedicated memory. Designed for the ST 65-nm CMOS process, an ASIC prototype has been fabricated and tested. The ASIC implements a single 128-neuron cluster and can emulate a network of up to $N_C = 31$ clusters of $N_N = 128$ neurons each. This corresponds to 3968 neurons in a clustered network. The ASIC can also emulate a multiple of 128 neurons in a Willshaw-Palm network, each clique containing up to 31 neurons. Possible activation rules are WTA (in clustered networks) and k-WsTA (in Willshaw-Palm networks). The former is implemented in hardware while the latter requires iterations. The ASIC has shown the same recovery performance when tested in the same context and conditions as [19]. The time response of the cluster, defined as the time a neuron needs to be activated from a stimulation, is 83 ns in measurement. The amount of energy needed to change the state of a neuron from a stimulation, called energy consumption per synaptic event, is measured to be 115 fJ. The ASIC has also been tested and verified in the context of ECG classification described in [22], achieving a recovery accuracy of 93.6%. The present work, based on the work of [25], includes the following additional contributions:

- A thorough description of the original single-cluster architecture is provided. It includes new circuit-level simulations demonstrating the WTA behavior, as well as a complete characterization of the Digital Computing Unit (memories and logic gates).
- The fabricated chip is positioned in the state of the art based on unitary measurements. Both ASIC silicon area and energy consumption at the neuron level are considered.
- Proper functioning of the circuit implementing larger clustered ANNs is demonstrated in a network of 1024 neurons. The targeted application is 8-character keyword recovery, implementing the complete ASCII table. The context of ECG classification application is also extended to include state-of-the-art classifiers.

The paper is organized as follows. The existing building blocks for the WTA operation design are described in Section II. The implemented network architecture and recovery procedures are shown in Section III. Measurement results and applicative use-cases are given in Section IV and Section V, respectively. Section VI concludes the paper.

II. Neuron Model and Implementation

Clique-based networks implement binary synapses, addition of the contributions and an activation function. Clustered networks implement a local WTA operation. It means that in each cluster, the neuron having the most active synapses is activated. Willshaw-Palm networks implement a global k-WsTA operation, *i.e.*, in the whole network the k neurons having the most active synapses are activated. However, by iterating a WTA operation k times, and each time switching off the previous winners, the k neurons having the most contributions can be detected. Thus, a WTA operation coupled with external computation can be used for both types of networks. The adapted model to implement the neurons in different types

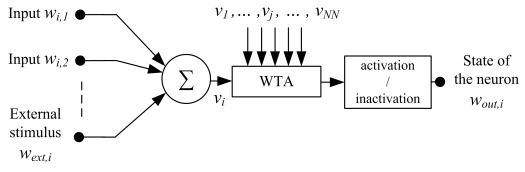


Fig. 2. Representation of a neuron implementing binary synapses and a WTA decision rule. In the schematic, $w_{i,j}$ represents the binary input #j of neuron #i, and v_i represents the number of active contributions for neuron #i.

of clique-based networks is shown in Figure 2. In this model, the binary contributions from other neurons and an external stimulus are summed:

$$\forall i, \quad 1 \leq i \leq N_{\mathrm{N}} : \; v_i = \sum_{j=1}^{n} w_{i,j} + w_{\mathrm{ext},i}, \tag{1}$$

where $w_{i,j}$ is the value of input #j of neuron #i in a cluster of length N_{N} and $w_{\mathrm{ext},i}$ is the contribution of the external stimulus. The external stimulus is a binary input that can be set or reset from outside the network. These stimuli can represent erroneous input information to be corrected or partial information to be recovered. Setting a certain number of external stimuli means stimulating a set of neurons, and thus starts the recovery. The activation of a neuron is decided by a WTA rule comparing the v_i of each neuron. The binary output of the winning neuron is thus set to '1' and the state of the neuron is said to be *active*:

$$\forall i, \quad 1 \leq i \leq N_{\mathrm{N}} : \; w_{\mathrm{out},i} = \begin{cases} 1 & \text{if } v_i = \max\limits_{1 \leq j \leq N_{\mathrm{N}}} (v_j), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The neuron circuit from [19], Figure 3, is used to implement the functions in Figure 2. It has been validated in measurements in terms of complexity, silicon occupation and energy consumption. It is less complex when voltages are used to exchange data between neurons and computations in the neurons are done with currents. A synapse bank per neuron converts the binary output voltages from other neurons into binary currents. A single synapse consists of a switched current source outputting either a unitary current I_{UNIT} or no current. The circuit is designed with transistors sized at 3 times the minimal length, operating in moderate inversion with currents in the hundreds of nanoamperes range. These currents are then summed in a single node, *e.g.* node A^1 for neuron #1. The resulting current is then input in the WTA element of the neuron by means of the current mirror formed by M_3^1 - M_4^1. The WTA circuit, adapted from [26], obtained by connecting the different WTA elements through node C, compares the currents input and activates the neuron corresponding to the highest current value. The current flowing in the WTA element #1 sets the drain-source voltage of M_1^1, depending on the other currents flowing in the other neurons. If the current flowing in neuron #1 is the highest in the cluster, the drain-source voltage of M_1^1 is set over its saturation voltage. Otherwise, M_1^1 is blocked and its drain-source voltage is set below its saturation voltage. The operation results in a binary voltage at node B^1, converted into a standard binary voltage at node F^1. The logic output of an activated neuron is '1', represented by a voltage V_{CC}, else it is set to '0', represented by a voltage of 0 V. The highest current is also copied at the I_{WIN} output. Finally, the WTA circuit is reset by tying node C to ground via the Reset signal.

As shown in [21], this circuit implementation has 2 kinds of physical drawbacks. At the synapse level, there is a leakage current of 70 pA flowing in the synapse even when the synapse is off. As a result, if too many synapses are connected to a single neuron, static current flows in that neuron, leading to its automatic activation. The maximal amount of leakage current acceptable to prevent self-activation of a neuron is 50 nA, which corresponds to a maximum of 700 synapses connected per neuron. Another option could be using non-volatile memories as synapses in order to reduce the leakage current, but it would also increase the synapse complexity. At the neuron level, there is also a leakage current of 3.6 nA in each losing element of the WTA connected to the same node. The sum of these leakage currents is subtracted from the input current in the winning neuron, biasing the WTA operation. If the sum is in the hundreds of nanoamperes range, it is comparable to the currents in the synapses. Therefore, these currents degrade the WTA behavior, and it is not possible to complete any message recovery. Keeping the total WTA leakage current below 100nA imposes a maximum of 28 WTA elements connected to C.

III. Proposed Iterative Architecture

A. Limitations of Parallel Hardwired Architecture

Implementing a fully parallel hardwired binary ANN, like in Figure 4(a), either Wilshaw *et al.* [20] or clustered [24], has 2 main limitations. The most obvious drawback is the total lack of flexibility. A hardwired network implies that the ANN is designed for a single application; any change implies the redesign of the full chip. Second, a parallel architecture yields a large integrated circuit. Clustering the neurons helps reducing the silicon area as compared to a fully-meshed network. However, the silicon area of the circuit increases as a squared function when scaling up the number of neurons. For instance, estimations in [19] show that a hardwired network of 31 clusters of 128 neurons each would occupy an area of 354 mm^2. Moreover, implementing the flexibility feature in a parallel design increases all the more the silicon area of the circuit. A different circuit organization is thus necessary when the number of neurons to implement increases.

B. Single Cluster Architecture

In order to overcome the limitation of the hardwired fully parallel architecture described in Section III-A, this work proposes to implement a single group of neurons performing a WTA operation, and to iterate the recovery process to emulate a bigger network, either Willshaw-Palm, clustered or mixed. Iterations also have the advantage to allow implementing more elaborate activation rules. Connections between potential clusters and data exchange operation are managed by a dedicated Digital Computing Unit (DCU). The proposed architecture is shown in Figure 4(b). This way, the network topology can

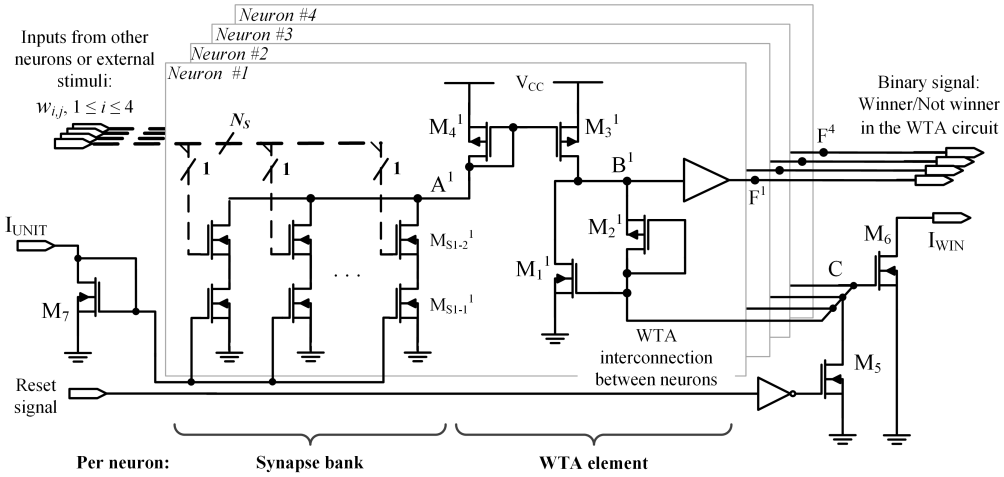


Fig. 3. Schematic of a neuron in a sub-cluster of 4, connected through node C.

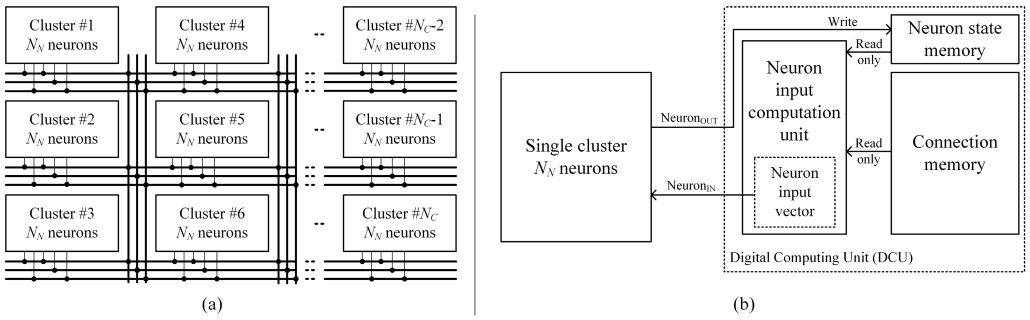


Fig. 4. Circuit structures allowing the implementation of clique-based networks. (a) shows a fully parallel circuit where neurons are hardwired. (b) depicts an iterative structure where a single cluster is implemented. A Digital Computing Unit (DCU) iterates the recovery process and stores the connections between the neurons, as well as the neurons states.

be easily modified by changing the content of the connection memory, the activation function, the number of emulated clusters N_C or the number of iterations N_{IT} of the recovery process.

It is not possible to design a single WTA circuit with a large number of elements connected to a single node C, because of leakage currents flowing in the WTA circuit, as explained in Section II. Thus, the WTA circuit has been designed like a 2-stage WTA, as depicted in Figure 5. The first stage is composed of groups of 16 neurons connected through a single node. By connecting only 16 neurons in a group, the sum of leakage currents in the circuit decreases to 57 nA and does not bias the WTA operation. After each group of 16 neurons, *e.g.*, neurons #1 to #16, the highest current $I_{\text{WIN\_1}}$ is fed into the second stage of the WTA. The corresponding first stage output, represented by node F in Figure 3, is also set to '1'.

Figure 6 depicts the behavior of the WTA circuit between neurons #1 to #16, simulated using *Spectre*®, including process variability (SS: slow corner with $V_{CC} = 0.9$ V and $T = 0$ °C, TT: typical corner with $V_{CC} = 1$ V and $T = 27$ °C, FF: fast corner with $V_{CC} = 1.1$ V and $T = 80$ °C). In the simulation, neuron #1 receives a current corresponding to 3 active contributions, and neuron #2 a current corresponding to 1 active contribution. Neuron #17 receives a current corresponding to 1 active contribution, but has no influence in the depicted WTA circuit. The remaining neurons do not receive any input current. After the stimulation, node B^1 is set over the decision threshold and thus the signal at node F^1 is set to logical '1'. The output current $I_{\text{WIN\_1}}$ corresponds to 3 active contributions. The signals at nodes F^2 to F^{16} stay at logical '0'. In parallel, the signal at node F^{17} is set to '1', and the output current $I_{\text{WIN\_2}}$ corresponds to 1 active contribution. The second stage is composed of $N_N/16$ WTA elements where the currents $I_{\text{WIN\_1}}$ to $I_{\text{WIN\_N}_N/16}$ are directly input. These WTA elements are connected together to perform a second WTA operation. Figure 7 describes the second stage WTA operation, simulated using *Spectre*®. Since $I_{\text{WIN\_1}}$ is superior to $I_{\text{WIN\_2}}$, the output of WTA element #1 is set

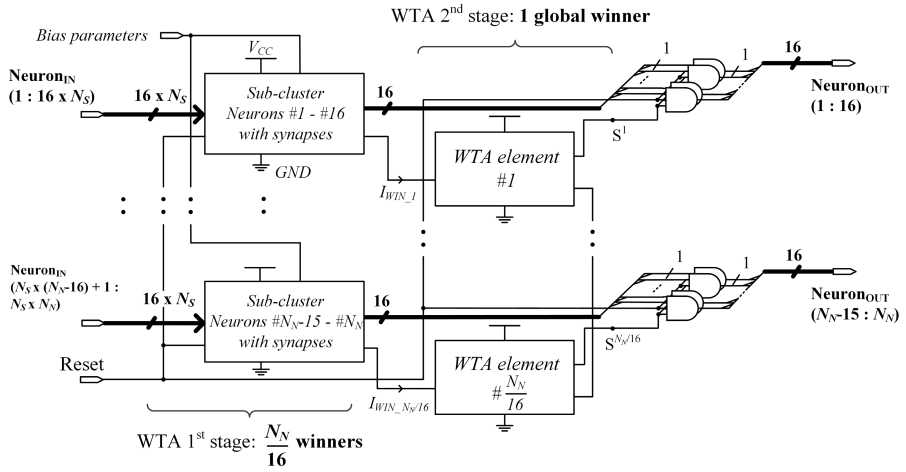


Fig. 5. Structure of the 128 implemented neurons grouped in a single cluster, with the 2-stage WTA circuit. N_S is the number of synapses per neuron.

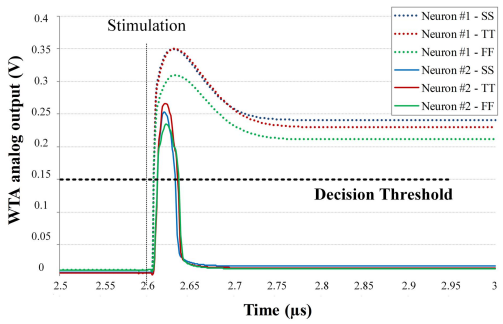


Fig. 6. *Spectre*® transient simulation of 2 neurons competing in the first WTA stage, including process variability (SS: slow corner with $V_{CC} = 0.9$ V and $T = 0$ °C, TT: typical corner with $V_{CC} = 1$ V and $T = 27$ °C, FF: fast corner with $V_{CC} = 1.1$ V and $T = 80$ °C). In the simulation, neuron #1 has 3 active contributions, while neuron #2 has 1 active contribution. The stimulation starts at 2.6 μs, and the output buffer, before node F, imposes the decision threshold.

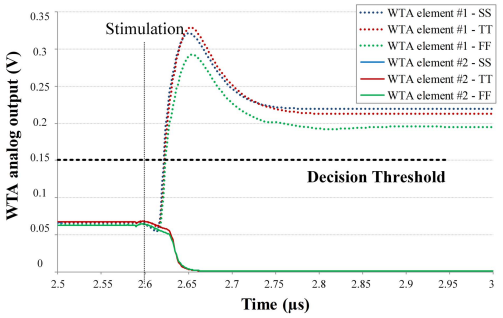


Fig. 7. *Spectre*® transient simulation of 2 WTA elements competing in the second WTA stage, including process variability (SS: slow corner with $V_{CC} = 0.9$ V and $T = 0$ °C, TT: typical corner with $V_{CC} = 1$ V and $T = 27$ °C, FF: fast corner with $V_{CC} = 1.1$ V and $T = 80$ °C). In the simulation, neuron #1 feeds a current equal to 3 active contributions in WTA element #1, while neuron #17 feeds a current equal to 1 active contribution in WTA element #2. The stimulation starts at 2.6 μs, and the output buffer, before node S, imposes the decision threshold.

over the decision threshold and the signal at node S^1 is set to logical '1'. The signals at nodes S^2 to $S^{N_N/16}$ stay at logical '0'. The global winner in the cluster is decided by an AND-logic operation between the signals at nodes F and S for each neuron, as shown in Figure 5. The process between the neurons stimulation and the availability of the results is asynchronous. Thus, there is no need for a clock signal in the network core.

This second stage is also limited to 16 elements, as a result N_N is capped at 256 for 2 stages. It is possible to increase further the cluster size by adding more stages in the structure. However, for energy consumption purpose, it is better to minimize the number of WTA stages. A lower amount of current is copied, and it also reduces the number of potential sources of transistor mismatch.

C. Description of the DCU

The DCU implements memories to store the states of the neurons and the connections between them, Figure 4(b). To evaluate the size of these memories, clustered clique-based networks are considered, since a larger amount of neurons and connections are emulated in this type of network.

Data concerning connections between neurons are stored in a *connection memory*, Figure 4(b). A straightforward architecture is to design it as a binary $N_C \times N_N$-by-$N_C \times N_N$ matrix of unidirectional connections. The rows and the columns represent all the neurons in the network, where the rows are the emitting side of a connection, and the columns are the receiving side. However, since the connections in cliques are bidirectional, the memory size can be halved.

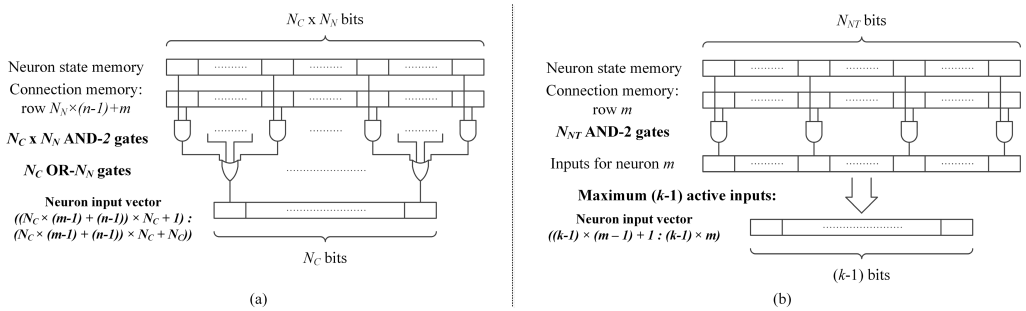


Fig. 8. Detailed operation for the next-state computation in the case of (a) clustered clique-based networks, and (b) Willshaw-Palm clique-based networks. The operation is shown for neuron #*m* of cluster #*n* and stores the result in the *neuron input vector*.

Finally, the *connection memory* needs $\frac{(N_C \times N_N)^2}{2}$ bits to store all the connections between neurons. The *neuron state memory* stores 1 bit per neuron signaling if it is active. The storage capacity of this memory is thus $N_C \times N_N$ bits.

The data output by the analog neurons $Neuron_{OUT}$ are received and stored in the DCU in the *neuron state memory*. These data are used by the *neuron input computation unit*, weighed by the data stored in the *connection memory*, to generate the input signal $Neuron_{IN}$ sent to the neurons. The details of this operation are shown in Figure 8(a) for clustered networks, and in Figure 8(b) for Willshaw-Palm networks.

For clustered networks, each input of neuron #*m* of cluster #*n* is the result of a logic AND operation between the actual states of all the neurons and the connection bits for the corresponding neuron, as shown in Figure 8(a). The latter are stored in row number $N_N \times (n-1) + m$ in the *connection memory*. Since in this kind of network the WTA rule is used, there is only one neuron activated per cluster. Logic OR operations between the N_N weighted neurons states of each cluster bring therefore the resulting data to input into the corresponding neuron, on N_C bits, 1 bit for each cluster. The number of synapses per neuron N_S is thus equal to the number of clusters N_C.

For Willshaw-Palm networks, the network is composed of N_{NT} neurons. For each neuron #*m*, the neurons states are weighted by row #*m* in the *connection memory* through a logic AND operation, as shown in Figure 8(b). As stated in Section III-B, since the size of a clique is at most k neurons, a neuron can receive at most $k-1$ active contributions. The relevant data input in the network for neuron #*m* are also a $(k-1)$-bit binary word. The number of synapses per neuron N_S is thus equal to $k-1$.

In both cases of networks, data regarding the inputs of the network are temporarily stored in the *neuron input vector*, until the next iteration. It needs at most $N_C \times N_N \times N_C$ bits of memory for clustered networks, and $N_{NT} \times (k-1)$ bits of memory for Willshaw-Palm networks. The *neuron input vector* is split in vectors of $N_S \times N_N$ bits fitting the $Neuron_{IN}$ signal width, and separately input in the analog cluster. The *neuron input computation unit* has "Read only" access to both the *neuron state memory* and the *connection memory*.

D. Recovery Procedures

The following algorithms, one for the WTA operation and one for the k-WsTA operation, describe the entire recovery procedure and detail which parts of the process are done in the analog cluster and in the DCU. One can note that the operations in the analog cluster are exactly the same in both types of networks. The circuit behavior can thus be changed by reconfiguring the DCU.

For the WTA operation, several clusters are emulated one after another, Figure 9(a). The data input into the network are updated after each iteration, and between two cluster computations until the number of clusters N_C is attained.

For the k-WsTA operation, only one cluster is emulated k times per iteration in order to determine the k winners, Figure 9(b). The data input into the network are updated after each iteration, and between cluster computations one winner is subtracted from the next data word to input.

IV. Prototype Characterization and Validation

A. Prototype Unitary Measurements

The circuit presented in the previous section has been implemented in an ASIC using the ST 65-nm CMOS process. The implemented cluster contains 128 neurons, which can represent 7-bit variables. Each neuron is connected to 31 synapses, *i.e.*, 3 times as many as in [19]. This number is low enough to prevent self-activation of a neuron due to leakage currents of 70 pA in inactive synapses. Except for the initial excitation, a neuron can receive a contribution from 30 other neurons, *i.e.*, the size of the cliques can go up to 31 neurons. In a clustered network, a neuron can receive a contribution from 30 other clusters, since one neuron per cluster is active. It is possible to emulate 31 clusters of 128 neurons, *i.e.*, 3968 neurons in total.

Considering the network size, the DCU has to implement 3968 AND-2 and 31 OR-128 logic gates. Besides, the *connection memory*, the *neuron state memory* and the *neuron input vector* contain 8 Mbits, 4 kbits and 123 kbits, respectively. The silicon area of the DCU implemented on an ASIC has been estimated for the 65-nm CMOS technology process using *Synopsys*® and SRAM memory density provided by the circuit manufacturer. The estimated silicon area is 8.5 mm^2.

Clustered clique-based networks

Initialize all the inputs of the network.
For iteration $i = 1$ to N_{IT} **do**
  For cluster $j = 1$ to N_C **do**
    Analog cluster:
    Serially input the stimuli and input signals to all the neurons with LR_{IN}.
    Load the stimuli and input signals with L_{IN}.
    Compute the neurons outputs in the cluster.
    Memorize all the neurons outputs in D-flip-flops for serialization with L_{OUT}.
    Serially output neurons output signals with LR_{OUT}.
    Digital computing unit:
    Store neurons activation data.
    If $j < N_C$ **do**
      Select input data for next cluster.
    End if
  End for
  Digital computing unit:
  If $i < N_{IT}$ **do**
    Compute and update network inputs for next iteration.
  Else
    Display final results.
  End if
End for

(a)

Willshaw-Palm networks

Initialize all the inputs of the network.
For iteration $i = 1$ to N_{IT} **do**
  While number of activated neurons $< k$ **do**
    Analog cluster:
    Serially input the stimuli and input signals to all the neurons with LR_{IN}.
    Load the stimuli and input signals with L_{IN}.
    Compute the neurons outputs in the cluster.
    Memorize all the neurons outputs in D-flip-flops for serialization with L_{OUT}.
    Serially output neurons output signals with LR_{OUT}.
    Digital computing unit:
    Store neurons activation data.
    Set to '0' activated neurons inputs.
    Compute updated input signal.
  End while
  Digital computing unit:
  If $i < N_{IT}$ **do**
    Compute and update network inputs for next iteration.
  Else
    Display final results.
  End if
End for

(b)

Fig. 9. Recovery algorithm detailing computation happening in the analog cluster or in the DCU in the case of (a) clustered clique-based networks, and (b) Willshaw-Palm clique-based networks.

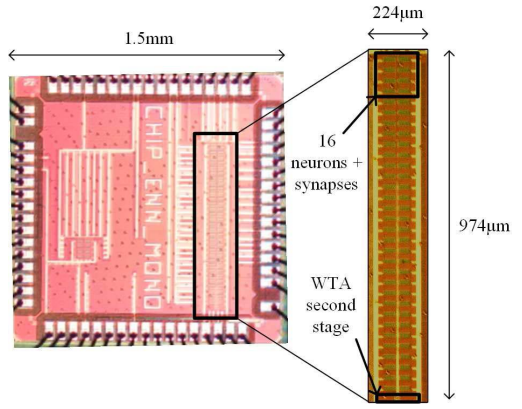


Fig. 10. Photograph of the fabricated die.

For simplicity reasons, the DCU is implemented externally on a *Xilinx*® *Spartan6*-based FPGA board. To reduce the number of I/O pins of the ASIC, the input and output messages are serialized. Therefore, Serial-In-Parallel-Out (SIPO) Parallel-In-Serial-Out (PISO) interfaces are inserted at the network's input and output. In order to decrease the latency of each sequence of message input, the SIPO interface is separated in 8 parallel independent interfaces dispatching their input data into different neurons. The fabricated ASIC is shown in Figure 10 and occupies 0.21 mm^2 of silicon area.

The ASIC operates at $V_{\mathrm{CC}} = 1$ V and with synaptic currents I_{UNIT} equal to 300 nA. The transistors composing the current mirrors operate thus in moderate inversion. In order to secure current measurements with a 100 pA resolution, the circuit core is supplied using a *Keysight*® *B2902A* Source/Measure Unit. The static current delivered to the ASIC is measured at 23.4 μA and corresponds to the biasing of each neuron with the synaptic current.

Figure 11 represents the test-bench used to characterize the ASIC. The *Reset* signal is used to clear all the registers in the circuit, as well as resetting the reference nodes for the WTA operation, *i.e.*, node C in Figure 3. The *Reset* signal is active-low. The input messages are sent serially by the FPGA to the test-chip through the $SerialIn(1:8)$ inputs, when the control signal LR_{IN} is high. After a message has gone through the SIPO registers, the L_{IN} signal is set to '1' and the network is let to execute the WTA operations (first and second stage) for a user-defined duration, arbitrarily set off-line. At the end of the WTA operation, the LR_{OUT} signal goes high and the outputs of the neurons are loaded in the PISO interface. The resulting message is serially output when the L_{OUT} signal is high. Depending on the position in the recovery process, defined by the corresponding algorithm in Section III-D, this procedure is repeated until the algorithm is completed, for a single recovery. After, the FPGA can either sort the results of multiple recoveries, or output the raw results for analysis by the user.

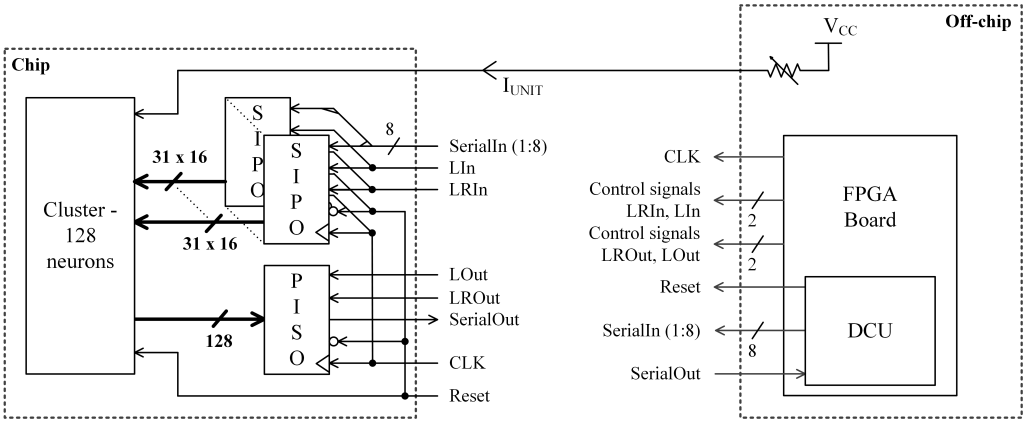


Fig. 11. Schematic of the test-bench used for the chip measurements. An FPGA board generates the control signals and implements the DCU. The biasing signals are generated separately outside the chip.

The FPGA implementing the DCU operates at 137 MHz and consumes a total power of 172 mW, out of which 24 mW correspond to the dynamic power consumption. Computing with the DCU lasts 4.56 μs each time. It corresponds to 625 clock cycles. To load the SIPO register in the ASIC, 496 cycles are needed, and 128 more cycles are needed to unload the PISO register. One more clock cycle is used for computing new data to input in the network. The longest response time of the analog cluster is measured to be 83 ns. Moreover, for each active input in the network, an additional dynamic current of $4 \times I_{\text{UNIT}}$ flows in the analog circuit. This current is due to the successive mirroring of I_{UNIT}, twice per WTA element, in the 2 stages of the cluster. It corresponds to a dynamic power consumption of $4 \times I_{\text{UNIT}} \times V_{\text{CC}}$ per active input, equal to 1.2 μW. The energy consumption per synaptic event per neuron, *i.e.*, the amount of energy needed to propagate a stimulation from an input to the output of a neuron, is measured to be 115 fJ. This value is calculated considering the static power divided by the number of neurons, plus the dynamic power generated by the activation of 1 synapse, during a response time of 83 ns. The chip performance is summarized in Table I.

Besides, the WTA resolution, defined as the ability for the WTA circuit to discriminate neurons having a different number of contributions, is tested. Two different neurons are stimulated in different synaptic activation cases, repeated for 120 different couples of neurons. It shows that the WTA is not able to discriminate a winning neuron when the neurons have more than 18 active synapses. This is caused by current mismatch in the successive current mirrors in the 2-stage WTA. As a result, it limits the number of neurons in a clique to 18, and impacts therefore the number of emulated clusters N_C. This is to be taken into account when storing the messages in the network. But, at the expense of larger silicon area and increased delay, mismatch can be reduced so that the number of neurons in a clique can be made larger than 18.

TABLE I
NEURAL NETWORK CHARACTERIZATION SUMMARY

ASIC

Process	ST 65-nm CMOS
Network silicon area	0.21 mm^2
Supply voltage	1 V
Synaptic current	300 nA
Static current	23.4 μA
Analog cluster response time	83 ns
Energy consumption per synaptic event per neuron	115 fJ

FPGA board

FPGA type	*Xilinx® Spartan6*
DCU maximum frequency	137 MHz
Latency per data exchange between DCU and analog cluster	4.56 μs
FPGA power consumption	172 mW
FPGA core dynamic power consumption	24 mW

B. Implemented Network for Characterization

In order to compare the recovery power of the flexible ASIC with theoretical and measured performance, the ASIC is first tested in the application of word recognition with a limited alphabet. This comparison also allows quantifying the amount of resources needed to bring flexibility in the network, in terms of power consumption and computation time. The considered clustered clique-based network is made of 5 clusters containing 6 neurons each. It can be used to store 5-letter words with the limited alphabet *{A, E, I, R, S, T}*. Ten messages are stored in the network and form its dictionary. It yields a network density of 26%, *i.e.*, the theoretical maximum amount of information stored without degrading the recovery. The stored messages are not random, but chosen to allow an equal connection distribution between all the neurons. Figure 12 represents the network with all 10 cliques stored. As an example, the clique storing the message STARS is highlighted in black.

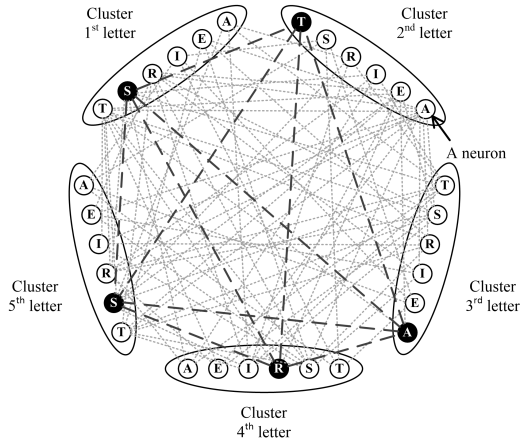


Fig. 12. A 5-cluster 6-neuron clique-based network with 10 cliques stored. Highlighted in black, the clique storing the message *STARS*.

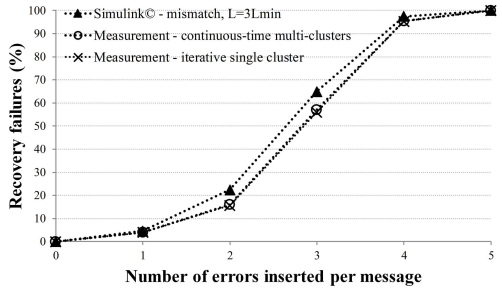


Fig. 13. Percentage of recovery failures depending on the number of inserted errors in the input message, for the considered set of stored messages. The density is 26%, equal to the theoretical maximum value. The results are displayed for behavioral simulation with *Simulink*®, for measurement results from [19], and for the measurement results of the flexible ASIC.

For each simulation or measure, a message modified by an error mask is input in the network. An erasure is counted as an error. The network is then put to convergence and the output of the neurons is recorded. If the output of the neurons does not correspond to any stored message, the decoding is considered as a failure. This procedure is repeated for every possible error mask and for every message, leading to a total of 168,070 simulations or measures.

Figure 13 displays the percentage of failed recoveries depending on the number of inserted errors, for behavioral simulation with *Simulink*® (including transistor mismatch), for measurement results from [19], and for the measurement results of the flexible ASIC. The performance in terms of recovery power of the flexible ASIC is on par with that of the previous ASIC. This validates the behavior of the flexible ASIC coupled with the DCU, and allows the use of this circuit in extended applications.

On the one hand, a neuron in the analog implementation of [19] has a response time of 14 ns. Since the analog clusters of neurons are implemented in parallel and hardwired in the ASIC, there is no need for iterating the recovery process. The network is let to converge to a stable state, and the output is available in a convergence time of 58 ns after a stimulation. The latency introduced by data exchanges with the ASIC, to input the stimulation and output the results, is 317 ns. On the other hand, the neurons in the flexible ASIC have a response time of 83 ns for an analog operation. As the architecture is iterative at the cluster level, the analog operation has to be repeated 5 times, 1 time per emulated cluster, to output the states of all the emulated neurons. The recovery process on the complete network has to be iterated 4 times to converge to stable neuron outputs. In total, the analog operation is repeated 20 times. Each time, data exchanges between the ASIC and the DCU introduce an additional latency of 4.56 μs. Overall, considering both the analog response times and the total digital latency, a message is retrieved in 92.9 μs.

Table II sums up the different features implemented by state-of-the-art ASICs and their corresponding performance in terms of silicon area and computation time. From data in Table II, it appears that the application fields of [13]–[16] are completely different from that of [19] and the proposed work. While [13], [14], and [16] have a higher computational power thanks to the higher number of neurons and connections, they are less adapted than [19] and the proposed work to small-scale embedded applications, where the energy budget is limited. As expected, the flexibility feature introduced in the current work has repercussions on both the silicon area of the ASIC and its computation time, compared to [19]. But, estimations of the silicon area of a hardwired network, as well as the silicon area occupied by [13]–[16], show that iterating the computations on a limited set of neurons is the best choice to increase the number of computing neurons, as the size of the circuit increases exponentially with the number of implemented neurons.

V. Prototype Applicative Use-Cases

Since the functionality of the fabricated ASIC has been validated in Section IV by referencing to a state-of-the-art circuit, this section aims at extending the applicative fields in order to show out the potential of the flexible ASIC. First, the context of [19] is extended in terms of cluster size and storage memory depth. Second, the performance of the ASIC is demonstrated in an ECG classification application, using a Willshaw-Palm network.

A. Keyword Recovery: Extended Clustered Network

The amount of neurons that can be emulated in the flexible ASIC allows increasing the storage capability of the considered network. In that sight, the applicative use-case of Section IV-B is extended. The objective is to retrieve 8-character keywords in a dedicated set, used in user ID research in a list, for instance. The number of neurons per cluster N_N is increased to 128 neurons, so that the possible letters represent the whole ASCII table. The number of clusters N_C is set to 8. In this network, a number of messages M equal

TABLE II
FEATURE SUMMARY OF DIFFERENT ANN CIRCUIT IMPLEMENTATIONS

	[13]	[14]	[15]	[16]	[19]	**This work**
Technology process	130-nm CMOS	28-nm CMOS	180nm CMOS	14-nm CMOS	65-nm CMOS	**65-nm CMOS**
Silicon area	102 mm^2	4,300 mm^2	43.8 mm^2	60 mm^2	0.016 mm^2	**0.21 mm^2 + 8.5 mm^2 (DCU)**
Number of implemented neurons	16 computing cores per chip	4096 computing cores per chip	1,024 per chip	128 neuro-morphic cores per chip	30	**128**
Number of computing neurons	16,000 per chip	1,000,000 per chip	1,024 per chip	131,072 per chip	30	**Up to 3968**
Flexibility	topology and activation function	topology and activation function	topology	topology and activation function	none	**topology and activation function**
Neuron response time	0.25 s	1 ms	12.5 ms	8.4 ns	14 ns	**83 ns**
Energy consumption per synaptic event	450 nJ	26 pJ	134 fJ	23.6 pJ	68 fJ	**115 fJ**
Latency for data exchanges per recovery	-	-	-	-	317 ns	**145.9 μs**

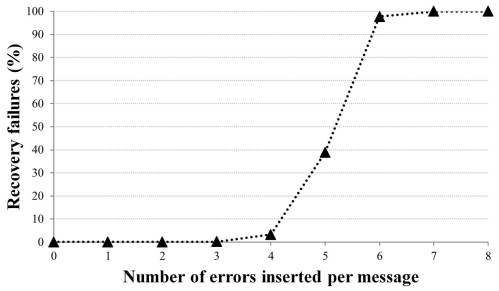


Fig. 14. Measured percentage of recovery failures depending on the number of inserted errors in the input message for the application of password recovery. The implemented network is composed of 8 clusters of 128 neurons each. The number of stored messages is 1024, yielding a network density of 6%.

to 1024 is stored. The network density, expressed as the ratio M/N_N^2 [24], is 6%.

Figure 14 depicts the measurement results of the flexible ASIC implementing the network of 8 clusters of 128 neurons each. For each number of inserted errors, 10000 random messages are generated from the original stored messages and input in the network. The recovery rate is indicated for each number of inserted errors in Figure 14. If the input message includes a number of errors lower than 5 (half the characters are wrong), the network is still able the find the stored original message without failure.

Moreover, using the full capabilities of the flexible ASIC does not change both the response time of the network and the computation time of the DCU per iteration per cluster, as all the computations are done in parallel in analog and digital units. The extended network thus takes 148.6 μs in total to process a single recovery. It corresponds to 4×8 iterations in the analog cluster of 83 ns each, and 4×8 computations of the DCU of 4.56 μs each. The energy consumption per neuron per synaptic event does not change and is 115 fJ. The total energy consumption of the DCU per recovery is 25.1 μJ. The scaling of the network is thus assured with a validated recovery rate, while keeping energy consumption per neuron per synaptic event independent of the number of implemented neurons.

B. ECG Classification: Willshaw-Palm Network

The fabricated chip has been also tested in an ECG classification application, in order to detect different types of cardiac beats [22]. The considered beat types are: Right Bundle Branch Block (RBBB), Paced Beat (PB) and Left Bundle Branch Block (LBBB). The signals are taken from the MIT-BIH arrhythmia database [27] and are preprocessed to extract features from the raw data. These features are the minima, maxima and standard deviations of 5 detail coefficients DC1 to DC5, which are produced by a 5-level Daubechies-2 wavelet decomposition of the signals from the database, *i.e.*, 15 features. The features that are not relevant in the different beat types classification are discarded. For the rest of the features, 1 to 3 ranges of values are represented by a neuron and form a 24-neuron Willshaw-Palm clique-based network, shown in Figure 15. For example, 3 neurons represent 3 ranges of values for the maximum of DC1. Three cliques, one for each beat type, are formed and connect 8 characteristic features, *i.e.*, 8 neurons, for each beat type with one another. They are represented in black, red and blue colors in Figure 15. The aim of the ANN is to output a beat type, *i.e.*, a clique of 8 neurons. Thus, the activation function for this ANN is a 8-WsTA operation. Furthermore, 4 iterations of the neuron activation process are done in order to ensure that the ANN converges to a stored clique.

The entire test set is made of 1572 beat samples: 523 RBBB, 486 LBBB and 563 PB. The retrieving process is repeated 50 times to verify the consistency of the ANN response.

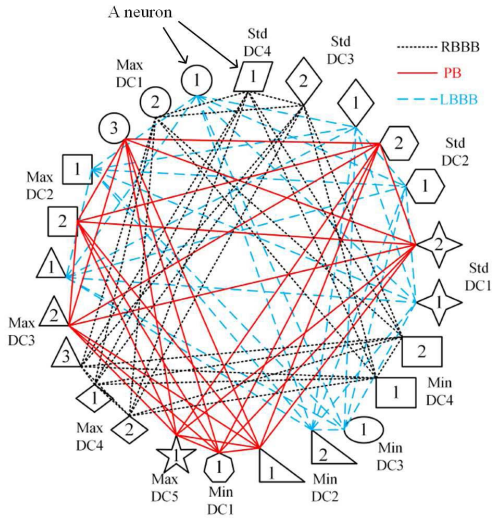


Fig. 15. Implemented network for ECG classification. A neuron corresponds to quantization value of a decomposition feature (the shape of a neuron indicates the corresponding feature). The network includes 3 cliques, indicated in blue, black and red colors. The activation function for this ANN is an 8-WsTA operation.

TABLE III
STATE OF THE ART FOR ECG CLASSIFICATION METHODS

	[28]	[29]	[30]	**This work**
Implementation type	ASIC 130nm-CMOS	FPGA-based	Software	**ASIC 65-nm CMOS**
Classification algorithm	SVM	Deep Neural Network	Deep Neural Network	**CBNN**
Number of classes	2	2	5	**3**
Classification accuracy	95%	97%	98.1%	**93.6%**
Energy per classification	273 μJ	n.a.	n.a.	**233.8 pJ**

First, using a behavioral simulation model designed with *Matlab*®, the simulated ANN achieves a classification accuracy of 93.5%. Second, using the flexible ANN chip, the implemented network achieves a measured classification accuracy of 93.6%. The measurement results are even slightly better since, in particular cases where neurons in 2 different cliques are ex-aequo, the ideal simulation model outputs both cliques at the same time, resulting in a wrong decision. Because of transistor mismatch in the circuit, the fabricated ANN always converges to a complete clique and does not activate neurons split in different cliques after 4 iterations. The measured classification results are also similar to the ones obtained with a hardware classifier using an SVM accelerator in [28], which consumes 273 μJ per classification, or with a hardware classifier implemented using an FPGA [29]. Moreover, it is only 4.5% lower than a software classifier using deep neural networks, for the same database [30].

The energy per classification is 233.8 pJ for the analog part, and 25.1 μJ for the DCU. Even if the prototype is not optimized for the targeted application and if the DCU is implemented off-chip on a FPGA, it shows that using an analog clique-based network in this context is relevant and reduces the energy per classification by an order of magnitude, compared to [28]. The fabricated ASIC has also an implementation complexity far lower than those of [29] and [30], in terms of silicon area and energy consumption. It also successfully demonstrates the behavior of the flexible ASIC for Willshaw-Palm networks applications.

VI. CONCLUSION

This paper presents a flexible ASIC in a 65-nm technology node able to implement clique-based ANNs using different topologies and neuron activation functions. The implemented ANNs can be configured off-chip to be used in different dedicated applications, using up to 3968 neurons. The iterative structure of the implemented network is the best way, in terms of silicon area, to implement clique-based ANNs with this number of neurons, as the silicon area of a hardwired network grows exponentially with the number of neurons. The ASIC has been fabricated and successfully tested in an extended context of keyword recognition, using the complete ASCII table, and in the context of ECG classification as a Willshaw-Palm classifier. Therefore, the presented ASIC is validated and can be used in both clustered networks and Willshaw-palm networks implementations. The analog circuit outputs the neurons states 83 ns after stimulation, while computing once with the DCU takes 4.56 μs. Each neuron bears an energy consumption of 115 fJ per synaptic event. This performance allows the use of this network in real-time classification in the context of kHz-range signal applications, such as bio-medical signals or audio signals. Further work on this topic will focus on realizing circuits integrating pre-processing accelerators on-chip in the context of "Near-Sensor computing."

ACKNOWLEDGMENT

The authors would like to thank R. Pallas for providing ASIC measurements.

REFERENCES

[1] C. Mead, *Pulsed Neural Networks* (VLSI Systems Series). Reading, MA, USA: Addison-Wesley, 1989.
[2] D. Reid, A. Hussain, and H. Tawfik, "Spiking neural networks for financial data prediction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–10.
[3] K. M. H. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6 μW power-proportional acoustic sensing frontend for voice activity detection," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, Jan. 2016.
[4] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. Tschanz, and V. De, "A 2.3 nJ/frame voice activity detector-based audio front-end for context-aware system-on-chip applications in 32-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, Aug. 2013.
[5] M. Yang, C. H. Chien, T. Delbruck, and S. C. Liu, "A 0.5 V 55 μW 64 $\times$ 2 channel binaural silicon cochlea for event-driven stereo-audio sensing," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2554–2569, Nov. 2016.
[6] T.-E. Chen *et al.*, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, Feb. 2017.

[7] F. Rosenblatt, "The perceptron—A perceiving and recognizing automaton project para," Cornell Aeronautical Lab., New York, NY, USA, Tech. Rep. 85-460-1, 1957.

[8] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554–2558, Apr. 1982.

[9] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1063–1070, Sep. 2004.

[10] H. Y. Hsieh and K. T. Tang, "VLSI implementation of a bio-inspired olfactory spiking neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1065–1073, Jul. 2012.

[11] J. Chen and T. Shibata, "A neuron-MOS-based VLSI implementation of pulse-coupled neural networks for image feature generation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 6, pp. 1143–1153, Jun. 2010.

[12] X. Wu, V. Saxena, K. Zhu, and S. Balagopal, "A CMOS spiking neuron for brain-inspired neural networks with resistive synapses and *in situ* learning," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 11, pp. 1088–1092, Nov. 2015.

[13] E. Painkras *et al.*, "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, Aug. 2013.

[14] F. Akopyan *et al.*, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 10, pp. 1537–1557, Oct. 2015.

[15] N. Qiao *et al.*, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128 k synapses," *Frontiers Neurosci.*, vol. 9, p. 141, Apr. 2015

[16] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018.

[17] J. Schemmel *et al.*, "Live demonstration: A scaled-down version of the BrainScaleS wafer-scale neuromorphic system," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2012, p. 702.

[18] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[19] B. Larras, C. Lahuec, F. Seguin, and M. Arzel, "Ultra-low-energy mixed-signal IC implementing encoded neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 11, pp. 1974–1985, Nov. 2016.

[20] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, "Non-holographic associative memory," *Nature*, vol. 222, pp. 960–962, Jun. 1969.

[21] B. Larras, B. Boguslawski, C. Lahuec, M. Arzel, F. Seguin, and F. Heitzmann, "Analog encoded neural network for power management in MPSoC" *Analog Integr. Circuits Signal Process.*, vol. 81, no. 3, pp. 595–605, Dec. 2014.

[22] P. Chollet, R. Pallas, C. Lahuec, M. Arzel, and F. Seguin, "A sub-nJ CMOS ECG classifier for wireless smart sensor," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 3840–3843.

[23] P. Chollet, K. Colombier, C. Lahuec, M. Arzel, and F. Seguin, "Toward sub-pJ per classification in body area sensor networks," in *Proc. 14th IEEE Int. New Circuits Syst. Conf. (NEWCAS)*, Jun. 2016, pp. 1–4.

[24] V. Gripon and C. Berrou, "Sparse neural networks with large learning diversity," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1087–1096, Jul. 2011.

[25] B. Larras, P. Chollet, C. Lahuec, F. Seguin, and M. Arzel, "A fully flexible circuit implementation of clique-based neural networks in 65-nm CMOS," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–4.

[26] M. Gu and S. Chakrabartty, "Synthesis of bias-scalable CMOS analog computational circuits using margin propagation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 2, pp. 243–254, Feb. 2012.

[27] *MIT-BIH Arrhythmia Database*. Accessed: Apr. 2016. [Online]. Available: http://physionet.org/physiobank/database/mitdb/

[28] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *IEEE J. Solid-State Circuits*, vol. 48, no. 7, pp. 1625–1637, Jul. 2013.

[29] J. P. Dominguez-Morales, A. F. Jimenez-Fernandez, M. J. Dominguez-Morales, and G. Jimenez-Moreno, "Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 24–34, Feb. 2018.

[30] M. M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Inf. Sci.*, vol. 345, pp. 340–354, Jun. 2016.

Benoit Larras was born in Nancy, France, in 1988. He received the Bachelor's and Master's degrees in telecommunications from IMT Atlantique, Brest, France, in 2012, and the Ph.D. degree in electrical engineering from IMT Atlantique in 2015. He is currently an Associate Professor with ISEN–IEMN, Lille, France. His research topics are analog/mixed-signal IC design and the circuit implementation of neural networks and associative memories, in the context of near-sensor computing.

Paul Chollet was born in Besançon, France, in 1990. He received the Bachelor's and Master's degrees in telecommunications from IMT Atlantique, Brest, France, in 2014, and the Ph.D. degree in electrical engineering from IMT Atlantique in 2017. He is currently an Associate Professor with Télécom ParisTech, Paris, France. His research topics are analog/mixed signal IC design and analog-to-information/features conversion. He received a Fulbright Fellowship from Cornell University for a five-month research project.

Cyril Lahuec was born in Orléans, France, in 1972. He received the B.Sc. degree (Hons.) from the University of Central Lancashire, U.K., in 1993, and the M.Eng. (mode A, by research) and Ph.D. degrees from the Cork Institute of Technology, Ireland, in 1999 and 2002, respectively, and the Habilitation degree from the University of South Brittany, Brittany, in 2012. The Habilitation degree is the highest French university degree passed after a few years of active research and student supervisions. He was with Parthus Technologies (now Ceva) Cork for the Ph.D. work, where he was also Consultant. He was a Visiting Scholar with the University of Edinburgh for 4 months in 2011. He joined the Department of Electronic Engineering, IMT Atlantique, as a full-time Lecturer in 2002. His research interests are in frequency synthesis, analog IC design, channel decoding, and biomedical applications.

Fabrice Seguin was born in Talence, France, in 1973. He received the Ph.D. degree from the Université Bordeaux 1, France, in 2001. His Ph.D. research concerned the current mode design of high-speed current-conveyors and applications in RF circuits. In 2002, he joined the Electronics Engineering Department, IMT Atlantique, Brest, France, as a full-time Lecturer. He is currently involved in the design issues of analog channel decoders and related topics, energy harvesting, neural coding, and reliability in nanoscale technologies.

Matthieu Arzel was born in Brest, France, in 1978. He received the Dipl.Ing. and the Ph.D. degrees from the Ecole Nationale Supérieure des Télécom munications de Bretagne, Brest, in 2002 and 2006, respectively. In 2006, he was with Turboconcept as a Research Engineer. He joined the Electronic Engineering Department, IMT Atlantique, as a full-time Lecturer in 2006. His research interests include iterative decoding techniques and analog/mixed integrated circuit architectures and design.

9.7 Article Springer JSPS sur l'apprentissage incrémental à budget limité avec des réseaux de neurones convolutionnels pré-entraînés et des mémoires associatives binaires [Bou+19]

Budget Restricted Incremental Learning with Pre-Trained Convolutional Neural Networks and Binary Associative Memories

Ghouthi Boukli Hacene, Vincent Gripon, Nicolas Farrugia, Matthieu Arzel and Michel Jezequel
IMT Atlantique, Brest, France
name.surname@imt-atlantique.fr

***Abstract*—Thanks to their ability to absorb large amounts of data, Convolutional Neural Networks (CNNs) have become state-of-the-art in numerous vision challenges, sometimes even on par with biological vision. They rely on optimisation routines that typically require intensive computational power, thus the question of embedded architectures is a very active field of research. Of particular interest is the problem of incremental learning, where the device adapts to new observations or classes. To tackle this challenging problem, we propose to combine pre-trained CNNs with binary associative memories, using product random sampling as an intermediate between the two methods. The obtained architecture requires significantly less computational power and memory usage than existing counterparts. Moreover, using various challenging vision datasets we show that the proposed architecture is able to perform one-shot learning – and even use only a small portion of the dataset – while keeping very good accuracy.**

***Index Terms*—Incremental Learning, Transfer Learning, Convolutional Neural Networks, Associative Memories**

I. Introduction

For the past few years, Deep Neural Networks (DNNs), and in particular Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance [1], [2], [3] in several domains of supervised learning, sometimes even being on par with the visual cortex [4]. DNNs rely on hundreds of millions of parameters that are trained to deal with large amounts of data. In this context a major drawback of the method is the need for intensive computation and memory usage during the learning phase. This limitation is critical for embedded systems such as smartphones or sensor networks.

A lot of effort has been driven towards optimized hardware implementations of DNNs [5]. For example in [6], the authors propose an architecture with state-of-the-art performance on the ImageNet challenge [7] using less than one megabyte, and in [8] a fixed point quantization of CNNs was introduced to reduce the network size. However, learning its parameters still require the processing of the whole dataset, involving many gradient computations. Moreover, the whole training dataset has to be stored in memory for learning.

Incremental methods provide solutions to process the learning data sequentially, using subsets of the training dataset. An incremental technique is defined as such [9], [10]: a) it is able to learn additional information from new data (example-incremental), b) it does not require access to the original data used to train the existing classifiers (in order to limit memory usage), c) it preserves previously acquired knowledge (avoid catastrophic forgetting) and d) it is able to accommodate new classes that may be introduced with new data (class-incremental). Although models have been proposed and studied extensively during the last decades, finding a good compromise between accuracy and required resources remains challenging. Indeed, most of existing works retrain the model when receiving new data [11], [12], and reuse some prior data for the retraining process [10], [13].

In this paper we propose an incremental learning model with the following claims:

- It is possible to adapt the model to new data without retraining it,
- It uses much less computational power than existing counterparts,
- It approaches state-of-art accuracy on challenging vision datasets (CIFAR10, ImageNet),
- It dramatically decreases the memory usage (by several orders of magnitude compared to nearest neighbour search),
- It only requires a few learning examples.

We point out that these claims are of particular interest when targeting embedded applications.

We rely on an increasingly popular method to benefit from the accuracy of DNNs without the need to train them on the targeted dataset, termed "transfer learning" [14], [2]. The idea is to use pre-trained CNNs on large datasets as feature extractors, and retrain the final layer of DNNs.

In this work, we propose to combine transfer learning with binary associative memories to achieve fully incremental learning. Binary associative memories are devices that are able to perform one-shot learning with very limited resources. The output of the DNN is quantized in a specific manner to combine it with the binary associative memories. This solution allows processing data sequentially one subset at a time, without forgetting initially processed data, and using only a sample of the database for learning. An overview of the proposed method is depicted in Figure 1. We evaluate the proposed method on challenging vision datasets (ImageNet and CIFAR10), and compare both accuracy and resources with alternative methods.

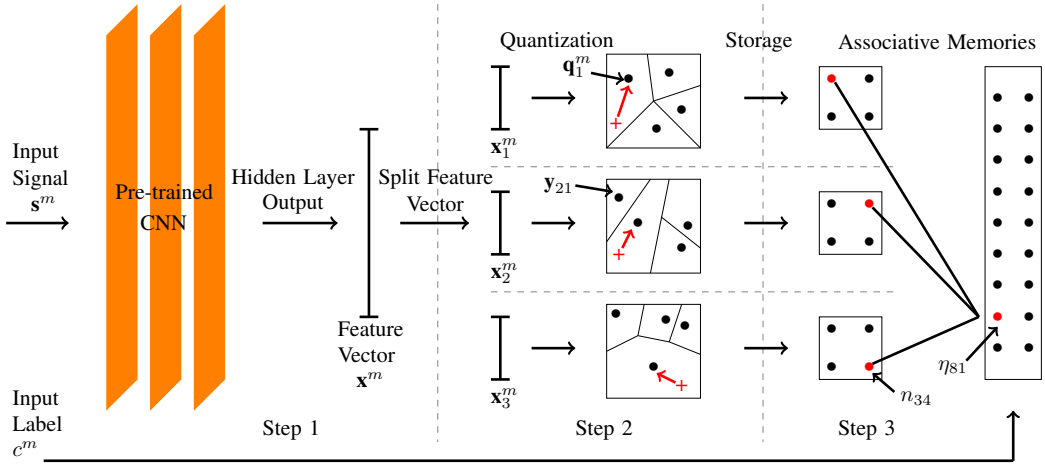


Figure 1. Overview of the proposed method, comprising three main steps. Given a set of samples, we first use a pre-trained CNN for feature extraction. Subsequently, we use a PQ technique to quantize the feature vectors. Finaly, we use a binary associative memory to store and classify the quantized data

The outline of the paper is as follows. In Section II we introduce related work. We present the proposed method in Section III. The experimental results are outlined in Section IV. Finally, Section V is a conclusion.

II. Related work

The term *incremental* usually refers to the ability of a learning process to learn sequentially, thus being able to handle new data and new classes without the need to retrain the whole system [10]. As an example, the "learn++" algorithm introduced in [10] accommodates new classes using weak one-vs-all classifiers. This approach conveniently manages the insertion, deletion and recurrence of classes over learning data [13]. However, this method requires to continuously train new classifiers in order to accommodate for new data, resulting in a potentially large computational intensiveness and memory usage.

Another approach was proposed to deal with a large amount of data [11], [12], [15]. The idea is to replace batches in classical learning methods with a process based on Support Vector Machines (SVM), in which learning is performed using only one subset at a time, independently of the others. As a consequence, it is possible to limit memory usage. However, training the SVMs can be computationally expensive.

In [16] incremental learning refers to three distinct problems: example-incremental learning [11], [12], [15], class-incremental learning [10], [13], and attribute-incremental learning. In [17], the authors propose a SVM inspired method to handle both the first two concepts defined above. However, it requires the training of novel SVMs using new examples and old SVMs. In addition, SVMs suffer from *catastrophic forgetting*, which is the loss of previously learned information [18], [19]. To address this problem a combination between SVMs and learn++ method called "SVMlearn++" [20] was proposed, showing a promising improvement on biological datasets [21]. However, this method still needs to retrain a new SVM each time new data is processed, and some knowledge is forgotten while new information is being learned.

The method we propose in this paper is quite different as it combines incremental aspects with one-shot learning using binary associative memories. Consequently, there is no need to retrain the system with old data, nor to perform computationally intensive processing with a new one. In addition, learning new data does not damage previously learned information, and only a few examples are required for learning, resulting in substantial savings in memory during the learning process.

III. The Proposed Method

The proposed method is built upon three main ideas: 1) using a pre-trained deep CNN to perform features extraction of signals, 2) using product quantizing techniques to embed data in a finite alphabet and 3) using binary associative memories to store and classify data as a proxy to a nearest neighbour search. In the next paragraphs, we detail these three steps.

Step 1: The first key idea is to use the internal layers of a pre-trained deep CNN [3] which acts as a generic feature extractor and associates an input signal $\mathbf{s}^m$ with a feature vector $\mathbf{x}^m$ (cf. Figure 1 step 1).

Step 2: Having a feature vector set $X = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^M\}$, the next step is to embed $\mathbf{x}^m$ in a finite alphabet. This step is crucial as it allows to map outputs of step 1) to the inputs of step 3). There is a lot of literature dedicated to this problem, including methods relying on Product Quantization (PQ) [22]. Because we aim at providing computationally light solutions, we rather use product random sampling in this work. Basically, we split each $\mathbf{x}^m$ into P subvectors of equal sizes

denoted $\left(\mathbf{x}_p^m\right)_{1\leq p\leq P}$, which are quantized independently from each other using random selection of K anchor points $Y_p = \mathbf{y}_{p1}, ..., \mathbf{y}_{pK}$, where each $\mathbf{y}_p^k$ is such that $\exists \mathbf{x}^m \in X, \mathbf{x}_p^m = \mathbf{y}_p^k$.

After step 2), each feature vector $\mathbf{x}^m$ is transformed into a word of fixed length $\left(\mathbf{q}_p^m\right)_{1\leq p\leq P}$ over a finite alphabet (the alphabet of the anchor points) (cf. Figure 1 step 2). This process is performed as follows:

$$\left\{ \begin{array}{lcl} k^\star(m,p) & = & \arg\min_k \|\mathbf{x}_p^m - \mathbf{y}_{pk}\|_2 \\ \mathbf{q}_p^m & = & y_{pk^\star(m,p)} \end{array} \right. . \quad (1)$$

Step 3: The outputs of product random sampling are words of fixed length P over a finite alphabet (containing K distinct symbols). The idea here is to use a neural network comprising two layers, the input one that is organized in P clusters containing K neurons each, and the output one containing RC neurons, where R is the number of neurons for each class and C is the number of classes in our dataset. Consider the neurons in the input layer to be indexed by two variables $p, 1 \leq p \leq P$ and $k, 1 \leq k \leq K$, where p denotes the index of the cluster and k the index of the neuron inside the cluster, and the neurons in the output layer to be indexed by two variables $c, 1 \leq c \leq C$ and $r, 1 \leq r \leq R$. We denote neurons in the input layer n_{pk} and neurons in the output layer η_{cr}.

When processing a training input signal $\mathbf{s}^m$, a number of neurons are activated in the network. Namely, we activate the neurons in the input layer whose indexes p, k are corresponding to the indexes of the activated anchor subvectors $\mathbf{q}_p^m$ obtained in Step 2. We activate in the output layer a neuron whose first index c is the index of the class c^m the training vector is part of, the second index r being drawn uniformly at random. Then we add connections (since the network is binary, there is no connection weight but only presence or absence of connections) between η_{cr} and all n_{pk}, printing a bipartite clique into the network (note that if a connection already existed, it is left unchanged) (c.f. Figure 1, Step 3).

We do the same process for every new data allowing incremental learning. Our method is a combination of a deep pre-trained CNN that does not change during the training process, and associative memories that are modified after each newly observed example or class. This combination allows to handle both example and class incremental approaches with no other prior about the learning dataset, using only few learning examples and without having to retrain the model or damage the previously obtained knowledge [23]. The overall process is depicted in Figure 1.

IV. Experiments

In this section we first describe the implementation details and strategies followed to quantize feature vectors $\mathbf{X}$. The accuracy of each strategy is then presented and discussed. We also investigate the behaviour of the accuracy, when changing some parameters, as the number of parts P to split feature vectors, as the number of neurons $\tilde{k}$ for each class in the input layer, and as the number of neurons R for each class in the output layer.

A. Implementation Details and Strategies Followed

For our method, we adopt the Inception V3 CNN model. It takes as input an image which is resized to $299 \cdot 299$ pixels and outputs a 2048 dimensional vector from the layer before the first fully-connected one [24]. The output vector represents the feature vector of the input image. To get the subvectors $\mathbf{y}_{pk}$ introduced in the previous section, we split the feature vectors $\mathbf{X}$ into P parts and we choose randomly $\tilde{k}$ subvectors from each class for each part (*i.e.* $K = C\tilde{k}$, C is the number of classes in our dataset and K is the total number of neurons in each cluster of the input layer).

We compare three strategies to emphasize the interest of the proposed method. They are described in the following paragraphs.

1) The "Independent Incremental" Approach (I-I): In this approach, training is not necessary: from each class we sample a portion of the example vectors and directly associate them to the corresponding output neurons using the associative memory. New data does not impact previously acquired knowledge, avoiding catastrophic forgetting.

In the case where $R = 1$, note that the associative memory is equivalent to counting how many quantized subvectors belong to each class and selecting the maximal one.

2) The "Non-Independent Incremental" Approach (N-I): In this approach, learning new elements can affect previously learned data. More precisely, each new input vector is quantized using all the already acquired anchor subvectors, independently of the class of the example that added them. The learning procedure is therefore computationally more complex than for the I-I method.

3) The "Non-Independent Offline" Approach (N-O): In this approach, the selection of anchor vectors is performed prior to any storing in the associative memory, so that the latter becomes independent of the order on which examples and classes are presented to the network.

B. Evaluation

We evaluate the proposed methods using three distinct datasets. The two first datasets (called in this paper ImageNet1 and ImageNet2) use 10 different classes of imageNet which were not used to train the CNN model. We use Cifar10 as the third dataset. Throughout the experimental part, the given accuracy is the average one over 10 realisations of each experiment.

1) Comparing the approches: Our first experiments consist of stressing the effect of the number of neurons per class R in the output layer of the associative memory on the accuracy of the three proposed approaches, for fixed values of the quantization parameters P and $\tilde{k}$. Figure 2 depicts the evolution of the accuracy of the methods as a function of the number of R. Expectedly, we observe that performance increases as a function of the number of neurons in the output layer. More interesting is the behaviour of the I-I method, which seems almost independent on R, while staying very close to the N-O method even when the latter is using a large value of R.

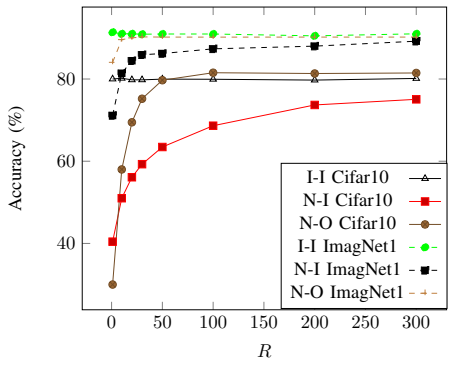


Figure 2. Evolution of the accuracy of the I-I approch ($P = 16$ and $\tilde{k} = 20$) as a function of the number of neurons R in the output layer for each class (ImageNet1 and Cifar10).

Additionally, the I-I approach shows better accuracy than the N-I one even when varying the parameters P and $\tilde{k}$, as shown in Table I. With this in mind, we focus on the I-I approach with $R = 1$ in the following experiments.

Table I
COMPARING THE ACCURACY OF I-I APPROACH WITH N-I APPROACH WITH $R = 300$ AND VARIOUS CLUSTER PARAMETERS (P AND $\tilde{k}$) (CIFAR10).

	I-I accuracy (%)	N-I accuracy (%)
$P = 16, \tilde{k} = 20$	**80.14**	75.06
$P = 16, \tilde{k} = 10$	**77.72**	65.52
$P = 64, \tilde{k} = 15$	**80.09**	75.61
$P = 32, \tilde{k} = 5$	**76.36**	59.38

Next, we consider two incremental protocols: class-incremental and example-incremental.

2) Class-Incremental Protocol: We first evaluate the effect of adding a new class on the accuracy. To do so, we start with an empty quantizer and an empty associative memory, and we add classes one by one. In order to avoid arbitrary decisions in the order in which classes are presented to the method, we perform experiments with random shuffles (200 times) and plot the average. We consider the following parameters: $P = 16$, $\tilde{k} = 20$ and $R = 1$. The accuracy a_c when introducing a novel class C_c is computed from scratch by adding new test examples to old one, according to Equation (2), where z_c represents the number of well classified test examples of all classes (for C_1 to C_c), m_c is the number of test examples of the class C_c and M_c is the total number of all test examples from classes C_1 to C_c.

$$\begin{cases} M_c = M_{c-1} + m_c \\ a_c = \frac{z_c}{M_c} \\ \text{with } M_0 = 0 \end{cases} \quad (2)$$

Each time a novel class is to be learned, we randomly sample $\tilde{k}$ subvectors for each of the P subspaces. We jointly add $\tilde{k}$ corresponding neurons in each cluster of the input layer. Finally, we add a new neuron corresponding to the newly added class in the output layer.

The obtained results are depicted in Figure 3. Of course the effect of adding new classes is more significant for a few number of classes, as it considerably strengthens the problem. The accuracy obtained after training all 10 classes approaches the one corresponding to the N-O approach for each dataset.

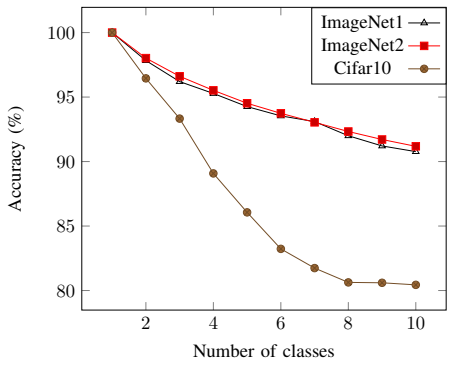


Figure 3. Evolution of the accuracy of the proposed method as a function of number of classes for $P = 16$, $\tilde{k} = 20$ and $R = 1$ (ImageNet1, ImageNet2 and Cifar10).

3) Example-Incremental Protocol: Next we evaluate the effect of adding new learning examples, without introducing new classes, on the accuracy of the I-I approach. To do so, we split the learning database into 5 parts and we learn one part at a time. The same testing dataset is used to measure accuracy at each step. For each part to be learned, we proportionally sample subvectors for each of the P subspaces and add the corresponding neurons in each cluster of the input layer. In Figure 4, for the three datasets, our method handles the incremental learning improving its accuracy and reaching the same final results as in the previous tests.

C. Complexity and memory usage

A key factor in proposing interesting solutions when targeting embedded architectures are complexity and memory usage. We refer to complexity as the number of arithmetical operations needed to learn the database (complexity-ℓ) or to classify an unlabelled input (complexity-p). The complexity-ℓ of the proposed method is negligible because we do not have to train the model with the whole dataset (c.f. subsection IV-D), while complexity-p is defined as $TK + PRC$ where T is the feature vector size ($T = 2048$ in our case due to the use of inception v3).

Figure 5 represents the accuracy as a function of the complexity-p when varying K and P (the other parameters are fixed to $T = 2048$, $R = 1$ and $C = 10$) and shows the best accuracy-complexity ratio (BACR). For a given K, BACR is obtained as the maximum in accuracy for similar values of complexity-p. We use the set of parameters reaching

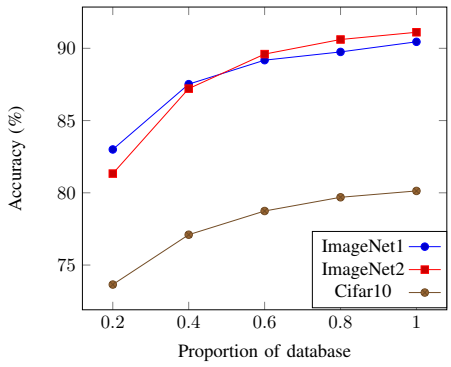


Figure 4. Evolution of the accuracy as a function of the number of learning examples ($P = 16$ and $\tilde{k} = 20$) (ImageNet1, ImageNet2 and Cifar10).

the BACR to compare our method with a nearest neighbour search.

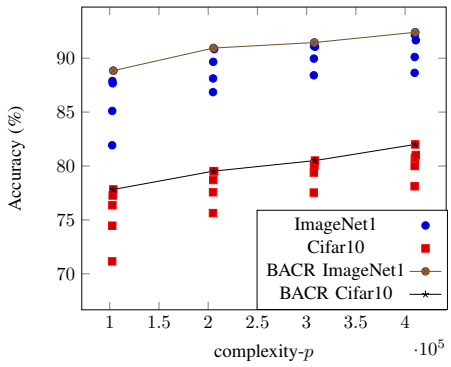


Figure 5. Evolution of the accuracy of I-I approach as a function of complexity-p (ImageNet1 and Cifar10) when varying K and P ($T = 2048$, $R = 1$ and $C = 10$).

Memory usage is defined by the size of clusters on input or output layers, and of the binary matrix which stores the connections between neurons. Memory usage sums up to $KTf + KPRC$ where f is the number of bits used per vector coordinates (we use $f = 32$ bits). Note that the size of the pre-trained CNN is not considered and the memory usage for learning and classifying are considered similar. We estimate accuracy, complexity and memory usage to compare the proposed method with a λ nearest neighbour (λ-NN) approach. The complexity-ℓ of the λ-NN search is also negligible, the complexity-p is defined by MT and memory usage is MTf. Results are shown in Table II.

We observe a loss in accuracy using our method, from 87% for nearest neighbour to 82%. On the other hand we obtain important gains in complexity and memory usage. One of the reason is that nearest neighbour search requires storing all

Table II
ACCURACY, COMPLEXITY AND MEMORY USAGE OF I-I APPROACH ($P = 64$, $K = 200$ AND $R = 1$) COMPARED TO λ-NN SEARCH FOR CIFAR10.

	Proposed Method	Other techniques	
		1-NN	5-NN
Accuracy(%)	82	85	**87**
complexity-ℓ	**negligible**	**negligible**	**negligible**
complexity-p	$\mathbf{4.1 \cdot 10^5}$	10^8	10^8
Memory usage-ℓ	$\mathbf{1.3 \cdot 10^7}$	$3.3 \cdot 10^9$	$3.3 \cdot 10^9$
Memory usage-p	$\mathbf{1.3 \cdot 10^7}$	$3.3 \cdot 10^9$	$3.3 \cdot 10^9$

training examples, which does not meet the criteria that define incremental learning algorithms [10].

Instead of using λ-NN search, we can accelerate it using PQ (Product Quantification). Namely, we split all feature vectors $\mathbf{x}^m$ into P of equal size denoted $\left(\mathbf{x}_p^m\right)_{1 \leq p \leq P}$, and for each subspace, we perform K-means on the feature vector set $X_p = \{\mathbf{x}_p^1, ..., \mathbf{x}_p^M\}$ to extract K centroids. When using K-means, we lowerbound the complexity-ℓ by taking into account only the MTK operations needed to quantize the learning dataset before storing it, with no consideration for the price of performing K-means. We motivate this choice as one could instead use product random sampling as described in our method. The complexity-p is $TK + MP$ and the memory usage is $TfK + MP\dot{l}og_2(K)$. Table III shows the obtained results.

Table III
ACCURACY, COMPLEXITY AND MEMORY USAGE RATIO OF I-I APPROACH ($P = 64$, $K = 200$ AND $R = 1$) COMPARED TO λ-NN SEARCH USING PQ ($K = 200$, $P = 64$) FOR CIFAR10. NUMBERS BETWEEN BRACKETS ACCOUNTS FOR PRODUCT RANDOM SAMPLING INSTEAD OF PQ.

	Proposed method	Other techniques	
		1-NN	5-NN
Accuracy(%)	82	82.6(82)	**86.07**(83)
complexity-ℓ	**negligible**	$\geq 2 \cdot 10^{10}$	$\geq 2 \cdot 10^{10}$
complexity-p	$\mathbf{4.1 \cdot 10^5}$	$3.2 \cdot 10^6$	$3.2 \cdot 10^6$
Memory usage-ℓ	$\mathbf{1.3 \cdot 10^7}$	$3.7 \cdot 10^7$	$3.7 \cdot 10^7$
Memory usage-p	$\mathbf{1.3 \cdot 10^7}$	$3.7 \cdot 10^7$	$3.7 \cdot 10^7$

NN search using PQ not only gives a good accuracy (86.07% compared with 82% of the I-I approach), but also reduces the complexity-p and memory usage by a factor of 100. However it requires a large computational power for learning process and stores a quantized version of the whole dataset, again not complying with the incremental learning algorithms criteria [10]. In addition both complexity and memory usage depends of number of learning examples M and could quickly become problematic.

Note that the proposed method uses product random sampling because it offers almost the same accuracy as using K-means instead. Moreover to use K-means it is required to store the whole database and perform expensive operations.

Finally, to assess the robustness of the proposed method

with regards to the chosen CNN feature extractor, we perform similar experiments using the SqueezeNet [6] architecture. This network makes even more sense with regards to embedded platforms given its very small memory usage. Table IV shows the obtained results that comfort the ones obtained using Inception V3.

Table IV
ACCURACY, COMPLEXITY AND MEMORY USAGE RATIO OF I-I APPROACH ($P = 64, K = 200$ AND $R = 1$) USING SQUEEZENET COMPARED TO λ-NN SEARCH FOR IMAGENET2.

	Proposed method	Other techniques	
		1-NN	5-NN
Accuracy(%)	84	88	**89**
complexity-ℓ	**negligible**	**negligible**	**negligible**
complexity-p	$\mathbf{2 \cdot 10^5}$	$8.4 \cdot 10^5$	$8.4 \cdot 10^5$
Memory usage-ℓ	$\mathbf{6.5 \cdot 10^6}$	$3.2 \cdot 10^8$	$3.2 \cdot 10^8$
Memory usage-p	$\mathbf{6.5 \cdot 10^6}$	$3.2 \cdot 10^8$	$3.2 \cdot 10^8$

D. Discussion

The proposed method achieves incremental learning (Figures 3 and 4) with substantial reduction of computational complexity and memory usage, compared to nearest neighbour search, without compromising classification accuracy (Tables II and III). Note that we preferred the I-I approach in our tests, since it obtained better accuracy than N-I. Since we also use product random sampling to feed the associative memories and require only one neuron per class in the output layer, we obtain that the neuron n_{pk} corresponding to y_{pk} is only connected to the neuron η_c, where $\exists \mathbf{x}^m \in X, \mathbf{x}_p^m = \mathbf{y}_{pk}$ and x^m belongs to the class C_c. As a consequence, knowing the connections of neurons obtained from random sampling with the neurons of the output layer, we need only few examples to train our model. Thus, the proposed method needs only a portion of the learning dataset to train, resulting in even lighter computational intensiveness and memory usage.

V. CONCLUSION

We introduced a novel incremental algorithm based on pre-trained CNNs and associative memories to classify images, the first ones using connection weights to process images, the second one using existence of connections to store them efficiently. This combination of methods allows to learn and process data using very few examples, memory usage and computational intensiveness. The obtained accuracy is close to other state-of-the-art methods based on transfer learning. As a consequence, we believe this method is promising for embedded devices and consider proposing thrifty hardware implementations of it as future work.

REFERENCES

[1] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," *CoRR*, vol. abs/1502.06796, 2015. [Online]. Available: http://arxiv.org/abs/1502.06796

[2] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS Comput Biol*, vol. 10, no. 12, p. e1003963, 2014.

[5] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song *et al.*, "Going deeper with embedded fpga platform for convolutional neural network," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2016, pp. 26–35.

[6] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[7] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and F.-F. Li, "Large scale visual recognition challenge," *www. image-net. org/challenges/LSVRC/2012*, vol. 1, 2012.

[8] D. D. Lin, S. S. Talathi, and V. S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *International Conference on Machine Learning (ICML)*, June 2016.

[9] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: an incremental learning algorithm for multilayer perceptron networks," in *Acoustics, Speech, and Signal Processing. ICASSP'00. Proceedings.IEEE International Conference on*, vol. 6. IEEE, 2000, pp. 3414–3417.

[10] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, vol. 31, no. 4, pp. 497–508, 2001.

[11] N. A. Syed, S. Huan, L. Kah, and K. Sung, "Incremental learning with support vector machines," 1999.

[12] T. Poggio and G. Cauwenberghs, "Incremental and decremental support vector machine learning," *Advances in neural information processing systems*, vol. 13, p. 409, 2001.

[13] Y. Sun, K. Tang, L. L. Minku, S. Wang, and X. Yao, "Online ensemble learning of data streams with gradually evolved classes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, pp. 1532–1545, 2016.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[15] V. Lomonaco and D. Maltoni, "Comparing incremental learning strategies for convolutional neural networks," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2016, pp. 175–184.

[16] Z.-H. Zhou and Z.-Q. Chen, "Hybrid decision tree," *Knowledge-based systems*, vol. 15, no. 8, pp. 515–528, 2002.

[17] J. Zheng, F. Shen, H. Fan, and J. Zhao, "An online incremental learning support vector machine for large-scale data," *Neural Computing and Applications*, vol. 22, no. 5, pp. 1023–1035, 2013.

[18] N. Kasabov, *Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines*. Springer Science & Business Media, 2013.

[19] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[20] Z. Erdem, R. Polikar, F. Gurgen, and N. Yumusak, "Ensemble of svms for incremental learning," in *International Workshop on Multiple Classifier Systems*. Springer, 2005, pp. 246–256.

[21] J. F. G. Molina, L. Zheng, M. Sertdemir, D. J. Dinter, S. Schönberg, and M. Rädle, "Incremental learning with svm for multimodal classification of prostatic adenocarcinoma," *PloS one*, vol. 9, no. 4, p. e93600, 2014.

[22] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2011.

[23] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv preprint arXiv:1312.6211*, 2013.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.

10 Références

10.1 Mes publications

Livres

[Ber+07] Claude Berrou, Karine Amis Cavalec, Matthieu Arzel, Catherine Douillard, Alain Glavieux, Frédéric Guilloud, Michel Jezequel, Sylvie Kerouedan, Charlotte Langlais, Christophe Laot, Raphaël Le Bidan, Emeric MAURY, Samir Saoudi, Yannick Saouter, Gérard Battail et Emmanuel Boutillon. *Codes et turbocodes (sous la direction de Claude Berrou)*. Iris. Springer, 2007, p. 397. URL : https://hal.archives-ouvertes.fr/hal-01801570.

[Ber+10] Claude Berrou, Karine Amis Cavalec, Matthieu Arzel, Catherine Douillard, Alexandre Graell I Amat, Frédéric Guilloud, Michel Jezequel, Sylvie KEROUEDAN, Charlotte Langlais, Christophe Laot, Raphaël Le Bidan, Samir Saoudi, Yannick Saouter, Youssouf OULD CHEIKH MOUHAMEDOU, Emeric MAURY, Alain Glavieux, Emmanuel Boutillon et Gérard Battail. *Codes and Turbo Codes*. IRIS international series. Springer-Verlag, 2010, p. 400. URL : https://hal.archives-ouvertes.fr/hal-01170419.

Articles de revues

[Arz+07] Matthieu Arzel, Cyril Lahuec, Fabrice Seguin, David Gnaedig et Michel Jezequel. “Semi-iterative analog turbo decoding”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* 54.6 (juin 2007), p. 1305-1316. DOI : 10.1109/TCSI.2007.897770. URL : https://hal.archives-ouvertes.fr/hal-01801701.

[Arz+11] Matthieu Arzel, Cyril Lahuec, Christophe Jego, Warren J. Gross et Yvain Bruned. “Stochastic multiple-stream decoding of Cortex codes”. In : *IEEE Transactions on Signal Processing* 59.7 (juil. 2011), p. 3486-3491. URL : https://hal.archives-ouvertes.fr/hal-00617865.

[Bou+18] Ghouthi Boukli Hacene, Vincent Gripon, Nicolas Farrugia, Matthieu Arzel et Michel Jezequel. “Transfer Incremental Learning Using Data Augmentation”. In : *Applied Sciences* 8.12 (déc. 2018), p. 2512. URL : https://hal.archives-ouvertes.fr/hal-01950211.

[Bou+19] Ghouthi Boukli Hacene, Vincent Gripon, Nicolas Farrugia, Matthieu Arzel et Michel Jezequel. “Budget Restricted Incremental Learning with Pre-Trained Convolutional Neural Networks and Binary Associative Memories”. In : *Journal of Signal Processing Systems* 91.9 (sept. 2019), p. 1063-1073. DOI : 10.1007/s11265-019-01450-z. URL : https://hal.archives-ouvertes.fr/hal-02273161.

[Don+10] Q. T. Dong, Matthieu Arzel, Christophe Jego et W. J. Gross. “Stochastic Decoding of Turbo Codes”. In : *IEEE Transactions on Signal Processing* 58.12 (2010), p. 6421-6425. DOI : 10.1109/TSP.2010.2072924. URL : https://hal.archives-ouvertes.fr/hal-00538602.

[Duc+09a] Nicolas Duchaux, Cyril Lahuec, Matthieu Arzel et Fabrice Seguin. “Analog decoder performance degradation due to BJT’s parasitic elements”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* 56.11 (nov. 2009), p. 2402-2410. URL : https://hal.archives-ouvertes.fr/hal-01801572.

[GAV14] Tristan GROLEAT, Matthieu Arzel et Sandrine Vaton. “Stretching the edges of SVM traffic classification with FPGA acceleration”. In : *IEEE Transactions on Network and Service Management* 11.3 (sept. 2014), p. 1-14. DOI : 10.1109/TNSM.2014.2346075. URL : https://hal.archives-ouvertes.fr/hal-01058332.

[GOM+14] Daniel GOMEZ TORO, Matthieu ARZEL, Fabrice SEGUIN et Michel JEZEQUEL. “Soft Error Detection and Correction Technique for Radiation Hardening Based on C-element and BICS”. In : *IEEE Transactions on Circuits and Systems II : Express Briefs* 61.12 (déc. 2014), p. 952-956. URL : https://hal.archives-ouvertes.fr/hal-01194933.

[GVA14] Tristan GROLEAT, Sandrine VATON et Matthieu ARZEL. “High-Speed Flow-Based Classification on FPGA”. In : *International Journal of Network Management* 24.4 (août 2014), p. 253-271. DOI : 10.1002/nem.1863. URL : https://hal.archives-ouvertes.fr/hal-01058333.

[Hac+18] Ghouthi Boukli HACENE, Vincent GRIPON, Matthieu ARZEL, Nicolas FARRUGIA et Yoshua BENGIO. “Quantized Guided Pruning for Efficient Hardware Implementations of Convolutional Neural Networks”. In : *CoRR* abs/1812.11337 (2018). arXiv : 1812.11337. URL : http://arxiv.org/abs/1812.11337.

[Har+16b] Ali HAROUN, Charbel ABDEL NOUR, Matthieu ARZEL et Christophe JEGO. “Low-Complexity Soft Detection of QAM demapper for a MIMO System”. In : *IEEE Communications Letters* 20.4 (avr. 2016), p. 732-735. DOI : 10.1109/LCOMM.2016.2525722. URL : https://hal.archives-ouvertes.fr/hal-01453396.

[Lah+11] Cyril LAHUEC, Shaban ALMOUAHED, Matthieu ARZEL, Deepak GUPTA, Chafiaa HAMITOUCHE-DJABOU, Michel JEZEQUEL, Christian ROUX et Eric STINDEL. “A self-powered telemetry system to estimate the postoperative instability of a knee implant”. In : *IEEE Transactions on Biomedical Engineering* 58.3 (mar. 2011), p. 822-825. URL : https://hal.archives-ouvertes.fr/hal-00609270.

[LAR+14] Benoît LARRAS, Bartosz BOGUSLAWSKI, Cyril LAHUEC, Matthieu ARZEL, Fabrice SEGUIN et Frédéric HEITZMANN. “Analog Encoded Neural Network for Power Management in MPSoC”. In : *Analog Integrated Circuits and Signal Processing* 81.3 (déc. 2014), p. 595-605. DOI : 10.1007/s10470-014-0420-z. URL : https://hal.archives-ouvertes.fr/hal-01170221.

[Lar+16] Benoît LARRAS, Cyril LAHUEC, Fabrice SEGUIN et Matthieu ARZEL. “Ultra-Low-Energy Mixed-Signal IC Implementing Encoded Neural Networks”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* 63.11 (2016), p. 1974-1985. DOI : 10.1109/TCSI.2016.2600663. URL : https://hal.archives-ouvertes.fr/hal-01394054.

[Lar+18a] Benoît Larras, Paul Chollet, Cyril Lahuec, Fabrice Seguin et Matthieu Arzel. “A Fully Flexible Circuit Implementation of Clique-Based Neural Networks in 65-nm CMOS”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* (déc. 2018), p. 1-12. URL : https://hal-imt-atlantique.archives-ouvertes.fr/hal-01983523.

[Lib+17b] Erwan Libessart, Matthieu Arzel, Cyril Lahuec et Francesco Andriulli. “A Scaling-Less Newton-Raphson Pipelined Implementation for a Fixed-Point Reciprocal Operator”. In : *IEEE Signal Processing Letters* 24.6 (juin 2017), p. 789-793. DOI : 10.1109/LSP.2017.2694225. URL : https://hal.archives-ouvertes.fr/hal-01617306.

[TRU+14a] Tuan-Anh TRUONG, Matthieu Arzel, Lin HAO, Bruno JAHAN et Michel Jezequel. “DFT Precoded OFDM - an Alternative Candidate for Next Generation PONs”. In : *Journal of Lightwave Technology* 32.6 (mar. 2014), p. 1228-1238. DOI : 10.1109/JLT.2014.2301632. URL : https://hal.archives-ouvertes.fr/hal-01061713.

[TRU+14b] Tuan-Anh TRUONG, Matthieu Arzel, Hao LIN, Bruno JAHAN et Michel Jezequel. “New low-complexity and robust time synchronization technique for optical IMDD OFDM transmissions”. In : *Optics Express* 22.12 (juin 2014), p. 14322-14340. URL : https://hal.archives-ouvertes.fr/hal-01067689.

Articles de conférences

[Ada+15] Marie-Pierre Adam, Matthieu Arzel, Antoine BEUGNARD, Jean-Philippe COUPEZ, François GALLÉE, Claire LASSUDRIE, Myriam Le Goff-Pronost, Michel Morvan, Bruno Vinouze et Didier BAUX. “Analyse d’une formation à la conduite de projets selon une grille de maturité de processus”. In : *QPES 2015 : Colloque Questions de pédagogies dans l’enseignement supérieur : Innover : pourquoi et comment ?* Brest, France, juin 2015, p. 125-130. URL : https://hal.archives-ouvertes.fr/hal-01174273.

[Ada+16] Marie-Pierre Adam, Matthieu Arzel, Antoine BEUGNARD, Jean-Philippe Coupez, Myriam Le Goff-Pronost, Michel Morvan, Pierre Tremenbert, Bruno Vinouze et Didier BAUX. “Boosting advanced skills in project management thanks to complex human and technical situations”. In : *SEFI 2016 : European Society for Engineering Education annual conference*. Tampere, Finland, 2016, p. 1-11. URL : https://hal.archives-ouvertes.fr/hal-01494159.

[Arz+04a] Matthieu Arzel, Cyril Lahuec, Michel Jezequel et Fabrice Seguin. “Analogue decoding of duo-binary codes”. In : *ISITA 2004 : International Symposium on Information Theory and its Applications*. Parma, Italy, 2004, p. . URL : https://hal.archives-ouvertes.fr/hal-01809308.

[Arz+04b] Matthieu Arzel, Cyril Lahuec, Michel Jezequel et Fabrice Seguin. “Décodage analogique de codes duo-binaires DVB-RCS”. In : *ISIVC'04 : 2nd International Symposium on Image/Video Communications over fixed and mobile networks, 7-9 juillet , Brest, France*. Brest, France, 2004, p. . URL : https://hal.archives-ouvertes.fr/hal-01809306.

[Arz+05a] Matthieu Arzel, Cyril Lahuec, Fabrice Seguin, David Gnaedig et Michel Jezequel. “Analog Slice Turbo Decoding”. In : *ISCAS 2005 : IEEE International Symposium on Circuits And Systems, Kobé, Japan, May 23-26*. Kobé, Japan, 2005, p. 332-335. DOI : 10.1109/ISCAS.2005.1464592. URL : https://hal.archives-ouvertes.fr/hal-01809349.

[Arz+05b] Matthieu Arzel, Cyril Lahuec, Fabrice Seguin et Michel Jezequel. “Semi-iterative analogue turbo decoding”. In : *ADW 2005 : 4th Analog Decoding Workshop, June 16-17, Rennes, France*. Rennes, France, 2005, p. . URL : https://hal.archives-ouvertes.fr/hal-01809320.

[Arz+06] Matthieu Arzel, Fabrice Seguin, Cyril Lahuec et Michel Jezequel. “Semi-Iterative Analog Turbo Decoding”. In : *ISCAS 2006 : International Symposium on Circuits and Systems, Kos, Greece, May 21-24*. Kos, Greece, 2006, p. 3562-3565. URL : https://hal.archives-ouvertes.fr/hal-01809339.

[Bou+17a] Ghouthi Boukli Hacene, Vincent Gripon, Nicolas Farrugia, Matthieu Arzel et Michel Jezequel. “Budget Restricted Incremental Learning with Pre-Trained Convolutional Neural Networks and Binary Associative Memories”. In : *SIPS 2017 : IEEE International Workshop on Signal Processing Systems*. Lorient, France, 2017, p. 1-4. DOI : 10.1109/SiPS.2017.8109978. URL : https://hal.archives-ouvertes.fr/hal-01656152.

[Bou+17b] Ghouthi Boukli Hacene, Vincent Gripon, Nicolas Farrugia, Matthieu Arzel et Michel Jezequel. “Finding All Matches in a Database using Binary Neural Networks”. In : *COGNITIVE 2017 : The Ninth International Conference on Advanced Cognitive Technologies and Applications*. Athènes, Greece, 2017, p. 59-64. URL : https://hal.archives-ouvertes.fr/hal-01522646.

[Bou+17c] Ghouthi Boukli Hacene, Vincent Gripon, Nicolas Farrugia, Matthieu Arzel et Michel Jezequel. “Incremental Learning on Chip”. In : *GlobalSIP 2017 : 5th IEEE Global Conference on Signal and Information Processing - Symposium on Signal Processing for Accelerating Deep Learning*. Montréal, Canada, 2017, p. . URL : https://hal.archives-ouvertes.fr/hal-01754847.

[Cho+16] Paul Chollet, Kevin Colombier, Cyril Lahuec, Matthieu Arzel et Fabrice Seguin. “Toward sub-pJ per classification in Body Area Sensor Networks”. In : *NEWCAS 2016 : 14th IEEE International on New Circuits and Systems*. Vancouver, Canada, 2016, p. 1-4. DOI : 10.1109/NEWCAS.2016.7604764. URL : https://hal.archives-ouvertes.fr/hal-01493883.

[Cho+17a] Paul Chollet, Cyril Lahuec, Matthieu Arzel et Fabrice Seguin. “A Sub-nJ CMOS ECG Classifier for Wireless Smart Sensor”. In : *EMBC 2017 : 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Jeju, South Korea : IEEE, 2017, p. 3840-3843. DOI : 10.1109/EMBC.2017.8037694. URL : https://hal.archives-ouvertes.fr/hal-01616520.

[Cho+17b] Paul Chollet, Benoît Larras, Cyril Lahuec, Fabrice Seguin et Matthieu Arzel. “An ultra-low power iterative clique-based neural network integrated in 65-nm CMOS”. In : *NEWCAS 2017 : 15th IEEE International New Circuits and Systems Conference*. Strasbourg, France, 2017, p. 5-8. DOI : 10.1109/NEWCAS.2017.8010091. URL : https://hal.archives-ouvertes.fr/hal-01596340.

[Cor+17a] Franck Cornevaux-Juignet, Matthieu Arzel, Pierre-Henri Horrein, Tristan Groleat et Christian Person. “Combining FPGAs and processors for high-throughput forensics”. In : *CNS 2017 : IEEE Conference on Communications and Network Security*. Las Vegas, United States : IEEE, 2017, p. . URL : https://hal.archives-ouvertes.fr/hal-01742964.

[Cor+17b] Franck Cornevaux-Juignet, Matthieu Arzel, Pierre-Henri Horrein, Tristan Groleat et Christian Person. “Open-source flexible packet parser for high data rate agile network probe”. In : *CNS 2017 : IEEE Conference on Communications and Network Security*. Las Vegas, United States, 2017, p. . URL : https://hal.archives-ouvertes.fr/hal-01740903.

[Cue+03] Javier Cuevas Ordaz, Patrick Adde, Sylvie Kerouedan, Matthieu Arzel et Jérôme Le Masson. “Turbo décodage de code produit haut débit utilisant un code BCH étendu”. In : *GRETSI’ 03 : 19ème colloque sur le traitement du signal*

et des images, 8-11 septembre, Paris. Paris, France, 2003, p. 349-352. URL : https://hal.archives-ouvertes.fr/hal-01809325.

[DAJ11] Quang Trung DONG, Matthieu ARZEL et Christophe JEGO. "Design and FPGA Implementation of Stochastic Turbo Decoder". In : *NEWCAS 2011*. Bordeaux, France, juin 2011, p. 21-24. URL : https://hal.archives-ouvertes.fr/hal-00617892.

[DIO+12a] Jean DION, Marie-Hélène HAMON, Pierre PENARD, Matthieu ARZEL et Michel JEZEQUEL. "Adapted scheduling of QC-LDPC decoding for multistandard receivers". In : *International Symposium on Turbo Codes & Iterative Information*. Gothenburg, Sweden, août 2012, p. 111-115. DOI : 10.1109/ISTC.2012.6325209. URL : https://hal.archives-ouvertes.fr/hal-00948457.

[DIO+12b] Jean DION, Marie-Hélène HAMON, Pierre PENARD, Matthieu ARZEL et Michel JEZEQUEL. "Multi-standard Trellis-based FEC Decoder". In : *DASIP 2012 : Conference on Design & Architectures for Signal & Image Processing*. Karlsruhe, Germany, oct. 2012. URL : https://hal.archives-ouvertes.fr/hal-00948464.

[Duc+08] Nicolas DUCHAUX, Cyril LAHUEC, Fabrice SEGUIN, Matthieu ARZEL et Michel JEZEQUEL. "Effect of BJT's parasitics on computing cells for analog decoders". In : *Proc. Joint 6th International IEEE Northeast Workshop on Circuits and Systems and TAISA Conference NEWCAS-TAISA, 22-25 juin 2008, Montréal, Québec, Canada*. Montréal, Canada, 2008, p. . URL : https://hal.archives-ouvertes.fr/hal-01809330.

[Duc+09b] Nicolas DUCHAUX, Cyril LAHUEC, Fabrice SEGUIN, Matthieu ARZEL et Michel JEZEQUEL. "Trade-off between surface, biasing current and performance of an analog turbo decoder". In : *IEEE joint conference NEWCAS-TAISA'09, June 28 - July 01, Toulouse, France*. Toulouse, France, 2009, p. 199-202. URL : https://hal.archives-ouvertes.fr/hal-01801573.

[DUC+10] Nicolas DUCHAUX, Cyril LAHUEC, Gurvan PORHIEL, Matthieu ARZEL, Fabrice SEGUIN et Michel JEZEQUEL. "Decreasing the effects of BJT's parasitics of computing cells for analog decoders". In : *6th Symposium of Turbo codes and related topics*. Brest, France, sept. 2010, p. 261-264. URL : https://hal.archives-ouvertes.fr/hal-00540872.

[GAV12] Tristan GROLEAT, Matthieu ARZEL et Sandrine VATON. “Hardware Acceleration of SVM-Based Traffic Classification on FPGA”. In : *IWCMC TRAC : International Wireless Communications and Mobile Computing Conference, International Workshop on TRaffic Analysis and Classification*. Limassol, Cyprus, août 2012. URL : https://hal.archives-ouvertes.fr/hal-00797503.

[GOM+13] Daniel GOMEZ TORO, Fabrice SEGUIN, Matthieu ARZEL et Michel JEZEQUEL. “Study of a Cosmic Ray Impact on Combinatorial Logic Circuits of an 8bit SAR ADC in 65nm CMOS Technology”. In : *MWSCAS 2013 : IEEE 56th International Midwest Symposium on Circuits and Systems*. Columbus, United States, juil. 2013. URL : https://hal.archives-ouvertes.fr/hal-00868594.

[Gri+17] Vincent GRIPON, Ghouthi Boukli HACENE, Nicolas FARRUGIA, Matthieu ARZEL et Michel JEZEQUEL. “Incremental learning on chip”. In : *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Montreal, France : IEEE, nov. 2017. DOI : 10.1109/GlobalSIP.2017.8309068. URL : https://hal.archives-ouvertes.fr/hal-01875912.

[Gro+13] Tristan GROLEAT, Matthieu ARZEL, Sandrine VATON, Alban BOURGE, Yannick LE BALCH et Hicham BOUGDAL. “Flexible, extensible, open-source and affordable FPGA-based traffic generator”. In : *HPDC 2013 : 22nd International ACM Symposium on High Performance Parallel and Distributed Computing*. New-York, United States, juin 2013. URL : https://hal.archives-ouvertes.fr/hal-00859291.

[Gup+10a] Deepak GUPTA, Shaban ALMOUAHED, Cyril LAHUEC, Matthieu ARZEL, Michel JEZEQUEL et Chafiaa HAMITOUCHE-DJABOU. “In-vivo polyethylene wear and knee prosthesis longevity estimation”. In : *ITAB 2010 : IEEE International Conference on Information Technology and Applications in Biomedicine*. Corfou, Greece, nov. 2010, p. 1-4. DOI : 10.1109/ITAB.2010.5687610. URL : https://hal.archives-ouvertes.fr/hal-00565619.

[Gup+10b] Deepak GUPTA, Cyril LAHUEC, Matthieu ARZEL, Chafiaa HAMITOUCHE-DJABOU et Michel JEZEQUEL. “Ligament imbalance metrics and an autonomous measurement system for post TKA”. In : *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Buenos Aires, Argentina, août 2010, p. 6417-6420. URL : https://hal.archives-ouvertes.fr/hal-00540867.

[GVA13] Tristan Groleat, Sandrine Vaton et Matthieu Arzel. “Accélération matérielle pour le traitement de trafic sur FPGA”. In : *15èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel)*. Pornic, France, mai 2013, p. 1-4. URL : https://hal.archives-ouvertes.fr/hal-00817038.

[Hac+20] Ghouthi Boukli Hacene, Vincent Gripon, Matthieu Arzel, Nicolas Farrugia et Yoshua Bengio. “Quantized Guided Pruning for Efficient Hardware Implementations of Deep Neural Networks”. In : *NEWCAS 2020 : 18th IEEE International New Circuits and Systems Conference*. Montréal, Canada : IEEE, juin 2020, p. 206-209. DOI : 10.1109/NEWCAS49341.2020.9159769. URL : https://hal-imt-atlantique.archives-ouvertes.fr/hal-02934543.

[HAR+13] Ali HAROUN, Charbel Abdel Nour, Matthieu Arzel et Christophe Jego. “Réception itérative MIMO basée sur la propagation de croyance et des codes LDPC non-binaires”. In : *GRETSI 2013 : 24ème colloque du Groupement de Recherche en Traitement du Signal et des Images*. Brest, France, sept. 2013. URL : https://hal.archives-ouvertes.fr/hal-00860658.

[HAR+14a] Ali HAROUN, Charbel Abdel Nour, Matthieu Arzel et Christophe Jego. *An efficient MIMO receiver based on BP algorithm with truncated message-passing*. GDR SoC-SiP 2014 : 9ème colloque national du GDR SoC-SiP du CNRS. Poster. Juin 2014. URL : https://hal.archives-ouvertes.fr/hal-01061840.

[HAR+14b] Ali HAROUN, Charbel Abdel Nour, Matthieu Arzel et Christophe Jego. “Low-Complexity LDPC-coded Iterative MIMO Receiver Based on Belief Propagation algorithm for Detection”. In : *ISTC 2014 : 8th International Symposium on Turbo Codes and Iterative Information Processing*. Bremen, Germany, août 2014, p. 213-217. DOI : 10.1109/ISTC.2014.6955116. URL : https://hal.archives-ouvertes.fr/hal-01170240.

[HAR+14c] Ali HAROUN, Charbel Abdel Nour, Matthieu Arzel et Christophe Jego. “Low-Complexity LDPC-coded Iterative MIMO Receiver Based on Belief Propagation algorithm for Detection”. In : *GDR SoCSiP-ISIS 2014 : journée thématique Architectures de Codes Correcteurs d’Erreurs*. Plouzané, France, nov. 2014. URL : https://hal.archives-ouvertes.fr/hal-01170374.

[HAR+14d] Ali HAROUN, Charbel Abdel Nour, Matthieu Arzel et Christophe Jégo. “Low-Complexity layered BP-based Detection and Decoding for a NB-LDPC Coded MIMO System”. In : *ICC 2014 : IEEE International Conference on Communi-*

cations. Sydney, Australia, juin 2014, p. 5107-5112. DOI : 10.1109/ICC.2014.6884131. URL : https://hal.archives-ouvertes.fr/hal-01174282.

[HAR+14e] Ali HAROUN, Charbel ABDEL NOUR, Matthieu ARZEL et Christophe JÉGO. "Symbol-based BP Detection for MIMO Systems associated with Non-Binary LDPC Codes". In : *WCNC 2014 : IEEE Wireless Communications and Networking Conference*. Istanbul, Turkey, avr. 2014, p. 212-217. DOI : 10.1109/WCNC.2014.6951949. URL : https://hal.archives-ouvertes.fr/hal-01174280.

[HAR+15] Ali HAROUN, Charbel ABDEL NOUR, Matthieu ARZEL et Christophe JEGO. "Algorithme de détection à très faible complexité pour des systèmes MIMO basés sur la propagation de croyance". In : *GRETSI 2015 : 25ème colloque du Groupement de Recherche en Traitement du Signal et des Images*. Lyon, France, sept. 2015. URL : https://hal.archives-ouvertes.fr/hal-01214329.

[Har+16a] Ali HAROUN, Charbel ABDEL NOUR, Matthieu ARZEL et Christophe JEGO. "Architecture d'un détecteur MIMO souple basé sur l'algorithme de propagation de croyance". In : *GDR SoC-SiP 2016 : colloque National du Groupe de Recherche System on Chip -System in Package*. Nantes, France, 2016, p. . URL : https://hal.archives-ouvertes.fr/hal-01451200.

[Har+16c] Ali HAROUN, Charbel ABDEL NOUR, Matthieu ARZEL et Christophe JEGO. "Soft detector architecture based on Belief Propagation for MIMO systems". In : *DASIP 2016 : Conference on Design and Architectures for Signal and Image Processing*. Rennes, France, 2016, p. 1-4. URL : https://hal.archives-ouvertes.fr/hal-01450885.

[Har+17] Ali HAROUN, Charbel ABDEL NOUR, Matthieu ARZEL et Christophe JEGO. "Architecture de détecteur MIMO-BP itératif associé à un décodeur LDPC non binaire". In : *GRETSI 2017 : 26ème colloque du Groupement de Recherche en Traitement du Signal et des Images*. Juan Les Pins, France, 2017, p. . URL : https://hal.archives-ouvertes.fr/hal-01617324.

[HEL+12] Romain HELOIR, Camille LEROUX, Saied HEMATI, Matthieu ARZEL et Warren J. GROSS. "Stochastic Chase Decoder for Reed-Solomon Codes". In : *NEWCAS 2012*. Montreal, Canada, juin 2012. URL : https://hal.archives-ouvertes.fr/hal-00725059.

[Hor+16] Pierre-Henri HORREIN, Philip-Dylan GLEONEC, Erwan LIBESSART, André LALEVEE et Matthieu ARZEL. *Ouessant : Flexible Integration of Dedicated Coprocessors in Systems On Chip*. DATE 2016 : Design, Automation & Test in Europe Confe-

rence & Exhibition. Poster. Mar. 2016. URL : https://hal.archives-ouvertes.fr/hal-01343408.

[Kis+19] Steven KISSELEFF, Nicola MATURO, Symeon CHATZINOTAS, Helge FANEBUST, Bjarne RISLOW, Kimmo KANSANEN, Matthieu ARZEL et Hans C HAUGLI. "User Terminal Wideband Modem for Very High Throughput Satellites". In : *37th International Communications Satellite Systems Conference (ICSSC)*. Okinawa, Japan, oct. 2019. URL : https://hal-imt-atlantique.archives-ouvertes.fr/hal-02433769.

[LA11] Cyril LAHUEC et Matthieu ARZEL. "An analog core computing the center of pressure in in a knee replacement prosthesis". In : *IEEE NEWCAS 2011 : IEEE 9th International conference on New Circuits and Systems Conference*. Bordeaux, France, juin 2011, p. 105-108. DOI : 10.1109/NEWCAS.2011.5981230. URL : https://hal.archives-ouvertes.fr/hal-00624136.

[Lah+06] Cyril LAHUEC, Gérald LE MESTRE, Fabrice SEGUIN, Matthieu ARZEL et Michel JEZEQUEL. "Design and test of a 0.25um-BICMOS double binary analogue APP Decoder". In : *ADW ´06 : 5th Analog Decoding Workshop, June 5-6, Torino, Italy*. Torino, Italy, 2006, p. 35-38. URL : https://hal.archives-ouvertes.fr/hal-01809302.

[Lah+09] Cyril LAHUEC, Matthieu ARZEL, Manuel GOURIOU et François GALLÉE. "A ligament laxity telemetry system architecture for a knee replacement prosthesis". In : *2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*. Bratislava, Slovakia, 2009, p. 1-6. URL : https://hal.archives-ouvertes.fr/hal-01801574.

[Lal+16] André LALEVEE, Pierre-Henri HORREIN, Matthieu ARZEL, Michael HÜBNER et Sandrine VATON. "Autoreloc : Automated Design Flow for Bitstream Relocation on Xilinx FPGAs". In : *DSD 2016 : Euromicro Conference on Digital System Design*. Limassol, Cyprus, 2016, p. 14-21. DOI : 10.1109/DSD.2016.92. URL : https://hal.archives-ouvertes.fr/hal-01393973.

[LAR+13a] Benoît LARRAS, Bartosz BOGUSLAWSKI, Cyril LAHUEC, Matthieu ARZEL, Fabrice SEGUIN et Frédéric HEITZMANN. "Analog Encoded Neural Network for Power Management in MPSoC". In : *NEWCAS 2013 : proceedings of the 11th IEEE international NEWCAS conference*. Paris, France, juin 2013, p. 1-4. URL : https://hal.archives-ouvertes.fr/hal-00843312.

[LAR+13b] Benoît LARRAS, Cyril LAHUEC, Matthieu ARZEL et Fabrice SEGUIN. “Analog implementation of encoded neural networks”. In : *ISCAS 2013 : IEEE International Symposium on Circuits and Systems*. Beijing, China, mai 2013, p. 1-4. URL : https://hal.archives-ouvertes.fr/hal-00836649.

[LAR+15] Benoît LARRAS, Cyril LAHUEC, Fabrice SEGUIN et Matthieu ARZEL. “Design of analog subthreshold encoded neural network circuit in sub-100nm CMOS”. In : *IJCNN 2015 : IEEE International Joint Conference on Neural Networks*. Killarney, Ireland : IEEE, juil. 2015, p. 1-7. DOI : 10.1109/IJCNN.2015.7280672. URL : https://hal.archives-ouvertes.fr/hal-01217841.

[Lar+17] Benoît LARRAS, Paul CHOLLET, Cyril LAHUEC, Fabrice SEGUIN et Matthieu ARZEL. “A 65-nm CMOS 7fJ per synaptic event clique-based neural network in scalable architecture”. In : *ISCAS 2017 : IEEE International Symposium on Circuits and Systems*. Baltimore, United States, 2017, p. 1-4. DOI : 10.1109/ISCAS.2017.8050658. URL : https://hal.archives-ouvertes.fr/hal-01656139.

[Lar+18b] Benoît LARRAS, Paul CHOLLET, Cyril LAHUEC, Fabrice SEGUIN et Matthieu ARZEL. “A fully flexible circuit implementation of clique-based neural networks in 65-nm CMOS”. In : *ISCAS 2018 : IEEE International Symposium on Circuits and Systems (ISCAS)*. Firenze, Italy, 2018, p. . DOI : 10.1109/ISCAS.2018.8350954. URL : https://hal.archives-ouvertes.fr/hal-01849349.

[LAS+15] Claire LASSUDRIE, Marie-Pierre ADAM, Matthieu ARZEL, Antoine BEUGNARD, Jean-Philippe COUPEZ, François GALLÉE, Sylvie KEROUEDAN, Myriam LE GOFF-PRONOST, Michel MORVAN, Bruno VINOUZE et Didier BAUX. “Score distribution as a tool to reveal group dynamics in student projects ?” In : *SEFI 2015 : Annual Conference of the European Society for Engineering Education*. T. SEFI 2015. Orléans, France : SEFI, 39 rue des Deux Eglises, 1000 Brussels, BELGIUM, juin 2015, p. 111-111. URL : https://hal.archives-ouvertes.fr/hal-01185711.

[Le +14] Myriam LE GOFF-PRONOST, Matthieu ARZEL, Antoine BEUGNARD, Jean-Philippe COUPEZ, François GALLÉE, Claire LASSUDRIE, Michel MORVAN, Bruno VINOUZE, Richard NAËL et Didier BAUX. “Introducing complexity into project management through multi-stakeholder interactions”. In : *SEFI 2014 : 42th annual conference*. Birmingham, United Kingdom : SEFI, sept. 2014, p. 135-135. URL : https://hal.archives-ouvertes.fr/hal-01170285.

[Lib+17a] Erwan Libessart, Matthieu Arzel, Cyril Lahuec et Francesco Andriulli. “A scaling-less Newton-Raphson pipelined implementation for a fixed-point inverse square root operator”. In : *NEWCAS 2017 : 15th IEEE International New Circuits and Systems Conference*. Strasbourg, France, 2017, p. . DOI : 10.1109/NEWCAS.2017.8010129. URL : https://hal.archives-ouvertes.fr/hal-01617301.

[Lib+17c] Erwan Libessart, Matthieu Arzel, Cyril Lahuec et Francesco Andriulli. “Implantation en virgule fixe d’un opérateur de calcul d’inverse à base de Newton-Raphson, sans normalisation et sans bloc mémoire”. In : *GRETSI 2017 : 26ème colloque du Groupement de Recherche en Traitement du Signal et des Images*. Juan-Les-Pins, France, 2017, p. . URL : https://hal.archives-ouvertes.fr/hal-01630144.

[Lib+17d] Erwan Libessart, Adrien Merlini, Matthieu Arzel, Cyril Lahuec et Francesco Andriulli. *Accélération matérielle pour l’imagerie cérébrale par EEG*. 13ème Colloque du GDR SoC/SiP. Poster. Juin 2017. URL : https://hal.archives-ouvertes.fr/hal-01754865.

[Lib+18] Erwan Libessart, Matthieu Arzel, Cyril Lahuec et Francesco Andriulli. “40 Gop/S/mm² Fixed-Point Operators for Brain Computer Interface in 65nm CMOS”. In : *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. Florence, Italy, 2018, p. . DOI : 10.1109/ISCAS.2018.8351028. URL : https://hal.archives-ouvertes.fr/hal-01813164.

[Mor+17] Michel Morvan, Bruno Vinouze, Marie-Pierre Adam, Priscillia Creach, Matthieu Arzel, Didier BAUX, Antoine BEUGNARD, Jean-Philippe Coupez, Myriam Le Goff-Pronost et Camilla Kärnfelt. “How to apprehend leadership related skills in a project management experiment ?” In : *SEFI 2017 : 45th Conference on Education Excellence For Sustainable Development*. Azores, Portugal, 2017, p. 536-543. URL : https://hal.archives-ouvertes.fr/hal-01661642.

[San+09] Oscar David Sanchez Gonzalez, Matthieu Arzel, Christophe Jego, Mauricio Guerrero et Antonio Garcia. “Design and implementation of a MIMO channel emulator onto FPGA device”. In : *IWS’09 : 1XV proyecto Iberchip*. Buenos Aires, Argentina, mar. 2009. URL : https://hal.archives-ouvertes.fr/hal-00424233.

[TRU+12] Tuan-Anh TRUONG, Hao LIN, Bruno JAHAN, Luiz ANET NETO, Matthieu Arzel et Michel Jezequel. “On the Performance of Timing Synchronization Tech-

niques for Optical OFDM IMDD Transmission". In : *IPC 2012 : IEEE Photonics conference*. Burlingame, Ca, United States, sept. 2012, p. 179-180. URL : https://hal.archives-ouvertes.fr/hal-00811742.

[TRU+13] Tuan-Anh TRUONG, Lin HAO, Bruno JAHAN, Matthieu ARZEL et Michel JEZEQUEL. "PAPR reduction using contiguous-tone Tone Reservation technique in optical OFDM IMDD transmissions". In : *OFC/NFOEC 2013 : Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC), 2013*. Anaheim, United States, mar. 2013. URL : https://hal.archives-ouvertes.fr/hal-00940412.

BREVETS

[ASL06] Matthieu ARZEL, Fabrice SEGUIN et Cyril LAHUEC. "Procédé et dispositif de décodage de codes à roulettes". FR2883121 (France). Sept. 2006. URL : https://hal.archives-ouvertes.fr/hal-01807308.

[Per+10] Jorge Ernesto PEREZ CHAMORRO, Fabrice SEGUIN, Cyril LAHUEC et Matthieu ARZEL. "Procédé de codage de données à au moins deux étapes d'encodage à et au moins une étape de permutation, dispositif de codage, programme d'ordinateur et signal correspondants". 2 941 829 (France). Août 2010. URL : https://hal.archives-ouvertes.fr/hal-01170863.

10.2 PUBLICATIONS CITÉES

[09] *VITA Radio Transport (VRT) Standard, ANSI/VITA 49.0-2009*. Vita Standards Organization. Mar. 2009.

[16] *TIA-5041 Future Advanced SATCOM Technologies (FAST) Open Standard Digital-If Interface (OSDI) for SATCOM Systems*. Telecommunications Industry Association (TIA). Mai 2016.

[AA09] L AZZAM et E. AYANOGLU. "Reduced Complexity Sphere Decoding via a Reodered Lattice Representation". In : *IEEE Trans. on Commun.* 57 (2009), p. 2564-2569.

[AGD17] Ross Adelman, Nail Gumerov et Ramani Duraiswami. “FMM/GPU-Accelerated Boundary Element Method for Computational Magnetics and Electrostatics”. In : *IEEE Transactions on Magnetics* PP (août 2017), p. 1-1. DOI : 10.1109/TMAG.2017.2725951.

[AH98] John B. Anderson et Stephen M. Hladik. “Tailbiting MAP Decoders”. In : *IEEE Journal on selected areas in communications* 16.2 (fév. 1998).

[Alm+10] Shaban Almouahed, Manuel Gouriou, Chafiaa Hamitouche, Eric Stindel et Christian Roux. “Design and evaluation of instrumented smart knee implant”. In : *IEEE Transactions on Biomedical Engineering* 58.4 (2010), p. 971-982.

[Alm+11] S. Almouahed, M. Gouriou, C. Hamitouche, E. Stindel et C. Roux. “Design and Evaluation of Instrumented Smart Knee Implant”. In : *IEEE Transactions on Biomedical Engineering* 58.4 (avr. 2011), p. 971-982. ISSN : 1558-2531. DOI : 10.1109/TBME.2010.2058806.

[Ama+03] A. Graell i Amat, G. Montorsi, S. Benedetto, D. Vogrig, A. Gerosa et A. Neviani. “A Full CMOS Analog Turbo Decoder for UMTS Coding Schemes”. In : *Proc. 2nd Analog Decoding Workshop*. Sept. 2003.

[And+08] Francesco P Andriulli, Kristof Cools, Hakan Bagci, Femke Olyslager, Annalisa Buffa, Snorre Christiansen et Eric Michielssen. “A multiplicative Calderon preconditioner for the electric field integral equation”. In : *IEEE Transactions on Antennas and Propagation* 56.8 (2008), p. 2398-2412.

[Ard+17] A. Ardakani, F. Leduc-Primeau, N. Onizawa, T. Hanyu et W. J. Gross. “VLSI Implementation of Deep Neural Network Using Integral Stochastic Computing”. In : *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 25.10 (oct. 2017), p. 2688-2699. ISSN : 1557-9999. DOI : 10.1109/TVLSI.2017.2654298.

[ASC13] A. R. Abolfazli, Y. R. Shayan et G. E. R. Cowan. “750Mb/s 17pJ/b 90nm CMOS (120,75) TS-LDPC Min-Sum based analog decoder”. In : *2013 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. Nov. 2013, p. 181-184.

[AVW08] M. Alles, T. Vogt et N. Wehn. “FlexiChaP : A reconfigurable ASIP for convolutional, turbo, and LDPC code decoding”. In : *2008 5th International Symposium on Turbo Codes and Related Topics*. Sept. 2008, p. 84-89.

[Bac+16] M. Bacci, F. Bigongiari, S. Chicca, A. Colonna, W. Errico, V. Nurra, G. Piscopiello, P. Tosi, G. Tuccio, N. Toptsidis, A. Busso, R. Jansen et N. Alagha. “Low-power analogue receiver ASIC for space telecommand applications”. In : *2016 International Workshop on Tracking, Telemetry and Command Systems for Space Applications (TTC)*. Sept. 2016, p. 1-5.

[Bia+06] A. Bianco, R. Birke, G. Botto, M. Chiaberge, J. M. Finochietto, G. Galante, M. Mellia, F. Neri et M. Petracca. “Boosting the performance of PC-based software routers with FPGA-enhanced network interface cards”. In : *2006 Workshop on High Performance Switching and Routing*. 2006, 6 pp.-. doi : 10.1109/HPSR.2006.1709693.

[BIW11] C. Brehm, T. Ilnseher et N. Wehn. “A scalable multi-ASIP architecture for standard compliant trellis decoding”. In : *2011 International SoC Design Conference*. Nov. 2011, p. 349-352. doi : 10.1109/ISOCC.2011.6138782.

[Blö+95] Johannes Blömer, Malik Kalfane, Richard Karp, Marek Karpinski, Michael Luby et David Zuckerman. *An XOR-Based Erasure-Resilient Coding Scheme*. Rapp. tech. TR-95-048. available at http ://www.icsi.berkeley.edu/ftp/global/pub/techrepo 95-048.pdf. International Computer Science Institute (ICSI), août 1995.

[Bri01] S. ten Brink. “Convergence behavior of iteratively decoded parallel concatenated codes”. In : *IEEE Trans. on Commun.* 49 (2001), p. 1727-1737.

[CCS12] CCSDS. *LEXIBLE ADVANCED CODING AND MODULATION SCHEME FOR HIGH RATE TELEMETRY APPLICATIONS, RECOMMENDED STANDARD CCSDS 131.2-B-1*. available at https ://public.ccsds.org/Pubs/131x2b1e1.pdf. Consultative Committee for Space Data Systems (CCSDS), 2012.

[CCS19] CCSDS. *SCCC—SUMMARY OF DEFINITION AND PERFORMANCE, INFORMATIONAL REPORT CCSDS 130.11-G-1*. available at https ://public.ccsds.org/Pubs/130x Consultative Committee for Space Data Systems (CCSDS), 2019.

[CH10] Aaron Carroll et Gernot Heiser. “An Analysis of Power Consumption in a Smartphone”. In : *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference*. USENIXATC’10. Boston, MA : USENIX Association, 2010, p. 21-21. url : http://dl.acm.org/citation.cfm?id=1855840.1855861.

[CHG20] Guillaume Coiffier, Ghouthi Boukli Hacene et Vincent Gripon. “ThriftyNets : Convolutional Neural Networks with Tiny Parameter Budget”. In : (2020). arXiv : 2007.10106 [cs.LG].

[Chi09] M. Chiu. “Low-Density Parity-Check Codes with 2-State Trellis Decoding”. In : *IEEE Commun. Lett.* 57 (jan. 2009), p. 12-16.

[CLL06] Marco Congedo, Fabien Lotte et Anatole Lécuyer. “Classification of movement intention by spatially filtered electromagnetic inverse solutions”. In : *Physics in Medicine & Biology* 51.8 (2006), p. 1971.

[CMM12] C. Condo, M. Martina et G. Masera. “A Network-on-Chip-based turbo/LDPC decoder architecture”. In : *2012 Design, Automation Test in Europe Conference Exhibition (DATE)*. Mar. 2012, p. 1525-1530. doi : 10.1109/DATE.2012.6176715.

[CMM13] C. Condo, M. Martina et G. Masera. “VLSI Implementation of a Multi-Mode Turbo/LDPC Decoder Architecture”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* 60.6 (juin 2013), p. 1441-1454. issn : 1558-0806. doi : 10.1109/TCSI.2012.2221216.

[CNC12] Todor Cooklev, Robert Normoyle et David Clendenen. “The VITA 49 analog RF-digital interface”. In : *IEEE Circuits and Systems Magazine* 12 (déc. 2012). doi : 10.1109/MCAS.2012.2221520.

[CV99] J.C. Carlach et C. Vervoux. “A new family of block turbo codes”. In : *Proceedings of 13th Applicable Algebra in Engineering Communication and Computing (AAECC 13)*. Hawaï, USA, nov. 1999, p. 15.

[Dai02] Jie Dai. “Design Methodology for Analog VLSI Implementations of Error Control Decoders”. Thèse de doct. University of Utah, déc. 2002.

[Dre+12] Thomas Dreischer, Björn Thieme, Michael Bacher, Klaus Buchheim et P Hyvnen. “OPTEL μ : a compact system for optical downlinks from LEO satellites”. In : *Proc. 12th SpaceOps, Stockholm, Sweeden* (2012).

[ETS02] ETSI-EN-301-839-1. “Electromagnetic compatibility and Radio spectrum Matters (ERM) ;Radio equipment in the frequency range 402 MHz to 405 MHz for Ultra Low Power Active Medical Implants and Accessories ;Part 1 : Technical characteristics, including electromagnetic compatibility requirements, and test methods.” In : 2002.

[ETS14] GSORI ETSI. *001 V4.1.1 (2014-10),”Open Radio equipment Interface (ORI) ; Requirements for Open Radio equipment Interface (ORI) (Release 4),”*. 2014.

[Fir16] Daniel Firestone. “SmartNIC : Accelerating Azure’s Network with FPGAs on OCS servers”. In : *OCP U.S. SUMMIT 2016*. San Jose, CA, mar. 2016.

[FZ99] A.J. Felström et K.S. Zigangirov. “Time-Varying Periodic Convolutional Codes with Density Parity-Check Matrix”. In : *IEEE Transactions on Information Theory* 45 (6 1999). URL : https://ieeexplore.ieee.org/document/782171 (visité le 10/10/2018).

[Gai69] B. Gaines. “Advances in information Systems Science”. In : *chapter 2, Plenum, New York*. 1969, p. 37-172.

[Gau+04] V. Gaudet, N. Nguyen, C. Winstead et C. Schlegel. “A 0.8V CMOS analog decoder for an (8,4,4) extended Hamming code.” In : *Proc. IEEE International Symposium on Circuits and Systems*. T. 1. Vancouver, Canada, mai 2004, p. 1116-1119.

[GB11] V. Gripon et C. Berrou. “Sparse Neural Networks With Large Learning Diversity”. In : *IEEE Transactions on Neural Networks* 22.7 (juil. 2011), p. 1087-1096. ISSN : 1941-0093. DOI : 10.1109/TNN.2011.2146789.

[GC11a] M. Gu et S. Chakrabartty. “A 100 pJ/bit, (32,8) CMOS Analog Low-Density Parity-Check Decoder Based on Margin Propagation”. In : *IEEE Journal of Solid-State Circuits* 46.6 (juin 2011), p. 1433-1442. ISSN : 1558-173X. DOI : 10.1109/JSSC.2011.2134550.

[GC11b] Ming Gu et Shantanu Chakrabartty. “Synthesis of bias-scalable CMOS analog computational circuits using margin propagation”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* 59.2 (2011), p. 243-254.

[GF11] Joseph T Gwin et Daniel Ferris. “High-density EEG and independent component analysis mixture models distinguish knee contractions from ankle contractions”. In : *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2011, p. 4195-4198.

[GG03] V. Gaudet et G. Gulak. “A 13.3Mbps 0.35μm CMOS Analog Turbo Decoder IC with a Configurable Interleaver”. In : *IEEE Journal of Solid-State Circuits* 38.11 (nov. 2003), p. 2010-2015.

[GGG02] V. C. Gaudet, R. J. Gaudet et P. Glenn Gulak. “Programmable Interleaver Design for Analog Iterative Decoders”. In : *IEEE Transactions on Circuits and Systems-II : Analog and Digital Signal Processing* 49.7 (juil. 2002), p. 457-464.

[GGM05] W.J. Gross, V.C. Gaudet et A. Milner. “Stochastic Implementation of LDPC Decoders”. In : *Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on*. 28 2005-Nov. 1 2005, p. 713-717. DOI : 10.1109/ACSSC.2005.1599845.

[Gna+05] D. Gnaedig, E. Boutillon, M. Jézéquel, V. Gaudet et P. Gulak. “On Multiple Slice Turbo Codes”. In : *Annales des Télécommunications* 60.1-2 (jan. 2005).

[Gol66] Solomon Golomb. “Run-Length Encodings”. In : *Information Theory, IEEE Transactions on* 12 (août 1966), p. 399-401. DOI : 10.1109/TIT.1966.1053907.

[Goo20] Coral Team Google Research. *Retrain a classification model on-device with weight imprinting*. Jan. Accessed in 2020.

[Gou+09] M. Gouriou, S. Almouahed, C. Hamitouche, C. Roux et E. Stindel. “Predictive autonomous orthopaedic device”. In : 9th *Workshop on Computer Assisted Orthopeadic Surgery*. Boston, USA, juin 2009, p. -.

[GR03] V.C. Gaudet et A.C. Rapley. “Iterative decoding using stochastic computation”. In : *Electronics Letters* 39.3 (fév. 2003), p. 299-301. ISSN : 0013-5194. DOI : 10.1049/el:20030217.

[Gra+07] F. Graichen, R. Arnold, A. Rohlmann et G. Bergmann. “Implantable 9-Channel Telemetry System for In Vivo Load Measurements With Orthopedic Implants”. In : *IEEE Transactions on Biomedical Engineering* 54.2 (fév. 2007), p. 253-261. ISSN : 1558-2531. DOI : 10.1109/TBME.2006.886857.

[Gre+08] Roberta Grech, Tracey Cassar, Joseph Muscat, Kenneth P Camilleri, Simon G Fabri, Michalis Zervakis, Petros Xanthopoulos, Vangelis Sakkalis et Bart Vanrumste. “Review on solving the inverse problem in EEG source analysis”. In : *Journal of neuroengineering and rehabilitation* 5.1 (2008), p. 25.

[GRF10] G. Gentile, M. Rovini et L. Fanucci. “A multi-standard flexible turbo/LDPC decoder via ASIC design”. In : *2010 6th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*. Sept. 2010, p. 294-298.

[GTJ10] T.S.V. Gautham, A. Thangaraj et D. Jalihal. “Common architecture for decoding turbo and LDPC codes”. In : (jan. 2010), p. 1-5.

[Gu+09] Ming Gu, Kiran Misra, Hayder Radha et Shantanu Chakrabartty. “Sparse decoding of low density parity check codes using margin propagation”. In : *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*. IEEE. 2009, p. 1-6.

[Guz+16] JE Ortiz Guzman, Axelle Pillain, Lyes Rahmouni et Francesco P Andriulli. “On the preconditioning of the symmetric formulation for the EEG forward problem by leveraging on calderon formulas”. In : *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, p. 755-758.

[Hag97] J. Hagenauer. *Der analoge Decoder*. German Pat. Appl. No 197 25 275.3. Juin 1997.

[Hal+03] D. Haley, C. Winstead, C. Schlegel et A. Grant. “An analog LDPC codec core.” In : *Proc. International Symposium on Turbo Codes*. Brest, France, 2003, p. 391-394.

[HEG17] A. Hussein, M. Elmasry et V. Gaudet. “On the Fault Tolerance of Stochastic Decoders”. In : *2017 IEEE 47th International Symposium on Multiple-Valued Logic (ISMVL)*. Mai 2017, p. 219-223. DOI : 10.1109/ISMVL.2017.50.

[HW98] J. Hagenauer et M. Winklhofer. “The Analog Decoder”. In : *Proc. 1998 IEEE Int. Symp. on Information Theory*. 16-21 Aug 1998, p. 145.

[HY10] S. Hemati et A. Yongacoglu. “Dynamics of analog decoders for different message representation domains”. In : *IEEE Transactions on Communications* 58.3 (mar. 2010), p. 721-723. ISSN : 1558-0857. DOI : 10.1109/TCOMM.2010.03.080011.

[Jan+96] C.L. Janer, J.M. Quero, J.G. Ortega et L.G. Franquelo. “Fully parallel stochastic computation architecture”. In : *Signal Processing, IEEE Transactions on* 44.8 (août 1996), p. 2110-2117. ISSN : 1053-587X. DOI : 10.1109/78.533736.

[Jou+17] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers et al. “In-datacenter performance analysis of a tensor processing unit”. In : *Proceedings of the 44th Annual International Symposium on Computer Architecture*. 2017, p. 1-12.

[Kao1] F. R. Kschischang et al. “Factor graphs and the sum-product algorithm”. In : *IEEE Trans. on Infor. Theory* 47 (2001), p. 498-519.

[KDK05] Mustafa Kaynak, Tolga Duman et Erozan Kurtas. “Belief Propagation over MIMO Frequency Selective Fading Channels”. In : *IEEE Inter. Conf. on Net. and Services* 34 (2005), p. 564-570. ISSN : 0090-6778.

[Kim+06] Hong-Sik Kim, Youngha Jung, Hyunjin Kim, Jin-Ho Ahn, Woo-Chan Park et Sungho Kang. “A high performance network-on-chip scheme using lossless data compression”. In : *IEICE Electronics Express IEEE Trans. Computer Aided Design Integr. Circuits Syst* 74 (jan. 2006), p. 791-796. DOI : 10.1587/elex.7.791.

[KSD17] H. Khanzadi, Y. Savaria et J. P. David. “A data driven CGRA Overlay Architecture with embedded processors”. In : *2017 15th IEEE International New Circuits and Systems Conference (NEWCAS)*. Juin 2017, p. 269-272. DOI : 10.1109/NEWCAS.2017.8010157.

[Kum+15] S. Kumawat, R. Shrestha, N. Daga et R. Paily. “High-Throughput LDPC-Decoder Architecture Using Efficient Comparison Techniques & Dynamic Multi-Frame Processing Schedule”. In : *IEEE Transactions on Circuits and Systems I : Regular Papers* 62.5 (mai 2015), p. 1421-1430. ISSN : 1558-0806. DOI : 10.1109/TCSI.2015.2403032.

[Lap+16] V. Lapotre, P. Murugappa, G. Gogniat, A. Baghdadi, M. Hübner et J. Diguet. “A Dynamically Reconfigurable Multi-ASIP Architecture for Multistandard and Multimode Turbo Decoding”. In : *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 24.1 (jan. 2016), p. 383-387. ISSN : 1557-9999. DOI : 10.1109/TVLSI.2015.2396941.

[LCK13] W. Lee, S. Chiu et Y. Ke. “IC Design of a Low-Power Analog LDPC Decoder Employing New Stopping Iteration Method”. In : *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*. Août 2013, p. 311-313. DOI : 10.1109/GreenCom-iThings-CPSCom.2013.69.

[Led+13] F. Leduc-Primeau, S. Hemati, S. Mannor et W. J. Gross. “Relaxed Half-Stochastic Belief Propagation”. In : *IEEE Transactions on Communications* 61.5 (mai 2013), p. 1648-1659. ISSN : 1558-0857. DOI : 10.1109/TCOMM.2013.021913.120149.

[Lib18] Erwan Libessart. “Interface cerveau-machine : de nouvelles perspectives grâce à l’accélération matérielle”. 2018IMTA0105. Thèse de doct. 2018. URL : http://www.theses.fr/2018IMTA0105/document.

[LLA09] Fabien Lotte, Anatole Lécuyer et Bruno Arnaldi. “FuRIA : an inverse solution based feature extraction algorithm using fuzzy set theory for brain–computer interfaces”. In : *IEEE transactions on signal processing* 57.8 (2009), p. 3253-3263.

[Loc+07] J. W. Lockwood, N. McKeown, G. Watson, G. Gibb, P. Hartke, J. Naous, R. Raghuraman et J. Luo. “NetFPGA–An Open Platform for Gigabit-Rate Network Switching and Routing”. In : *2007 IEEE International Conference on Mi-*

croelectronic Systems Education (MSE'07). Juin 2007, p. 160-161. DOI : 10.1109/MSE.2007.69.

[Loe+98] H.-A. LOELIGER, F. LUSTENBERGER, M. HELFENSTEIN et F. TARKÖY. "Probability Propagation and Decoding in Analog VLSI". In : *Proc. 1998 IEEE Int. Symp. on Information Theory*. 16-21 Aug 1998, p. 146.

[Lub02] M. LUBY. "LT codes". In : *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* Nov. 2002, p. 271-280. DOI : 10.1109/SFCS.2002.1181950.

[Lus+99] F. LUSTENBERGER, M. HELFENSTEIN, G. S. MOSCHYTZ, H. A. LOELIGER et F. TARKOY. "All analog decoder for a binary (18,9,5) tail-biting trellis code". In : *Proc. 25th European Solid-State Circuits Conference*. Sept. 1999, p. 362-365.

[Mac05] D. J. C. MACKAY. "Fountain codes". In : *IEE Proceedings - Communications* 152.6 (déc. 2005), p. 1062-1068. ISSN : 1350-2425. DOI : 10.1049/ip-com:20050237.

[ME10] M. MISHALI et Y. C. ELDAR. "Xampling : Analog Data Compression". In : *2010 Data Compression Conference*. Mar. 2010, p. 366-375. DOI : 10.1109/DCC.2010.39.

[Mer+14] Paul A MEROLLA, John V ARTHUR, Rodrigo ALVAREZ-ICAZA, Andrew S CASSIDY, Jun SAWADA, Filipp AKOPYAN, Bryan L JACKSON, Nabil IMAM, Chen GUO, Yutaka NAKAMURA et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface". In : *Science* 345.6197 (2014), p. 668-673.

[Miy+13] D. MIYASHITA, R. YAMAKI, K. HASHIYOSHI, H. KOBAYASHI, S. KOUSAI, Y. OOWAKI et Y. UNEKAWA. "A 10.4pJ/b (32, 8) LDPC decoder with time-domain analog and digital mixed-signal processing". In : *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*. Fév. 2013, p. 420-421. DOI : 10.1109/ISSCC.2013.6487796.

[Moe+00] M. MOERZ, T. GABARA, R. YAN et J. HAGENAUER. "An Analog 0.25μm BiCMOS Tailbiting MAP Decoder". In : *Proc. IEEE International Solid-State Circuits Conference*. Fév. 2000, p. 356-357.

[Moe04] Matthias MOERZ. "Analog Sliding Window Decoder Core for Mixed Signal Turbo Decoder". In : *Proc. Int. ITG Conference on Source and Channel Coding*. Erlangen, Germany, jan. 2004, p. 63-70.

[MS02] M.M. Mansour et N.R. Shanbhag. “Turbo decoder architectures for low-density parity-check codes”. In : *Global Telecommunications Conference, 2002. GLOBECOM ’02. IEEE*. T. 2. Nov. 2002, p. 1383-1388.

[MSN03] A.F. Mondragon-Torres, E. Sanchez-Sinencio et K.R. Narayanan. “Floating-Gate Analog Implementation of the Additive Soft-Input Soft-Output Decoding Algorithm”. In : *IEEE Transactions on Circuits and Systems I : Fundamental Theory and Applications* 50.10 (oct. 2003), p. 1256-1269.

[Mur+11] P. Murugappa, R. Al-Khayat, A. Baghdadi et M. Jezequel. “A flexible high throughput multi-ASIP architecture for LDPC and turbo decoding”. In : *Design, Automation Test in Europe Conference Exhibition (DATE), 2011*. Mar. 2011, p. 1-6.

[Nae+10] F. Naessens, V. Derudder, H. Cappelle, L. Hollevoet, P. Raghavan, M. Desmet, A.M. AbdelHamid, I. Vos, L. Folens, S. O’Loughlin, S. Singirikonda, S. Dupont, J.-W. Weijers, A. Dejonghe et L. Van der Perre. “A 10.37 mm2 675 mW reconfigurable LDPC and Turbo encoder and decoder for 802.11n, 802.16e and 3GPP-LTE”. In : *VLSI Circuits (VLSIC), 2010 IEEE Symposium on*. Juin 2010, p. 213-214.

[NBC06] A. Nimbalker, Y. Blankenship et B. Classon. “Turbo-like Decoding Algorithm for Structured LDPC codes”. In : *2006 IEEE International Symposium on Information Theory*. Juil. 2006, p. 1708-1712. DOI : 10.1109/ISIT.2006.261646.

[NG12] Luis Fernando Nicolas-Alonso et Jaime Gomez-Gil. “Brain computer interfaces, a review”. In : *Sensors* 12.2 (2012), p. 1211-1279.

[NVI17] Tesla NVIDIA. *V100 GPU architecture*. 2017.

[PAE67] W. Poppelbaum, C. Afuso et J. Esch. “Stochastic computing elements and systems”. In : *AFIPS FJCC*. 31. 1967, p. 635-644.

[Pal13] Günther Palm. “Neural associative memories and sparse coding”. In : *Neural Networks* 37 (2013), p. 165-171.

[Per+09a] Jorge Ernesto Perez Chamorro, Cyril Lahuec, Fabrice Seguin, Gérald Le Mestre et Michel Jezequel. “A subthreshold PMOS analog cortex decoder for the (8,4,4) hamming code”. In : *Journal of the electronics and telecommunications research institute* 31.5 (oct. 2009), p. 585-592. DOI : 10.4218/etrij.09.0109.0207. URL : https://hal.archives-ouvertes.fr/hal-01845148.

[Per+09b] Jorge Ernesto Perez Chamorro, Fabrice Seguin, Cyril Lahuec, Gérald Le Mestre et Michel Jezequel. “Decoding a family of dense codes using the sum-product algorithm and subthreshold PMOS”. In : *IEEE International Symposium on Circuits And Systems, Taiwan*. Taipei, Taiwan : IEEE ISCAS, 2009, p. 2685-2688. URL : https://hal.archives-ouvertes.fr/hal-01853651.

[Per+09c] J. Perez-Chamorro, F. Seguin, C. Lahuec, M. Jezequel et G. Le Mestre. “Decoding a family of dense codes using the Sum-Product Algorithm”. In : *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*. Mai 2009, p. 2685-2688. DOI : 10.1109/ISCAS.2009.5118355.

[Pus+11] A.E. Pusane, R. Smarandache, P.O. Vontobel et D.J. Costello Jr. “Deriving Good LDPC Convolutional Codes from LDPC Block Codes”. In : *IEEE Transactions on Information Theory* 57 (2 2011). URL : https://ieeexplore.ieee.org/document/5695133 (visité le 10/10/2018).

[QLL18] W. Qiao, D. Liu et S. Liu. “QFEC ASIP : A Flexible Quad-Mode FEC ASIP for Polar, LDPC, Turbo, and Convolutional Code Decoding”. In : *IEEE Access* 6 (2018), p. 72189-72200. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2018.2883292.

[SC08] Y. Sun et J.R. Cavallaro. “Unified decoder architecture for LDPC/turbo codes”. In : *IEEE Workshop on Signal Processing Systems, 2008. SiPS 2008*. T. 2. Oct. 2008, p. 13-18.

[Seg+04] F. Seguin, C. Lahuec, J. Lebert, M. Arzel et M. Jézéquel. “Analogue 16-QAM demodulator”. In : *IEE Electronics Letters* 40.18 (2004), p. 1138-1139. DOI : 10.1049/el:20046146. URL : https://hal.archives-ouvertes.fr/hal-01807307.

[SF17] C. Schmidt et C. Fuchs. “The OSIRIS program — First results and Outlook”. In : *2017 IEEE International Conference on Space Optical Systems and Applications (ICSOS)*. Nov. 2017, p. 19-22. DOI : 10.1109/ICSOS.2017.8357205.

[SGM06] S. Sharifi Tehrani, W.J. Gross et S. Mannor. “Stochastic decoding of LDPC codes”. In : *Communications Letters, IEEE* 10.10 (oct. 2006), p. 716-718. ISSN : 1089-7798. DOI : 10.1109/LCOMM.2006.060570.

[Sha+08] S. Sharifi Tehrani, C. Jego, Bo Zhu et W.J. Gross. “Stochastic Decoding of Linear Block Codes With High-Density Parity-Check Matrices”. In : *Signal Processing, IEEE Transactions on* 56.11 (nov. 2008), p. 5733-5739. ISSN : 1053-587X. DOI : 10.1109/TSP.2008.929337.

[Sha+10] S. Sharifi Tehrani, A. Naderi, G. Kamendje, S. Hemati, S. Mannor et W. J. Gross. “Majority-Based Tracking Forecast Memories for Stochastic LDPC Decoding”. In : *IEEE Transactions on Signal Processing* 58.9 (sept. 2010), p. 4883-4896. ISSN : 1941-0476. DOI : 10.1109/TSP.2010.2051434.

[Sho04] A. Shokrollahi. “Raptor codes”. In : *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.* Juin 2004, p. 36-. DOI : 10.1109/ISIT.2004.1365073.

[SMG08] S. Sharifi Tehrani, S. Mannor et W.J. Gross. “Fully Parallel Stochastic LDPC Decoders”. In : *Signal Processing, IEEE Transactions on* 56.11 (nov. 2008), p. 5692-5703. ISSN : 1053-587X. DOI : 10.1109/TSP.2008.929671.

[Sol+06] J. Soler-Garrido, R. J. Piechocki, K. Maharatna et D. McNamara. “MIMO detection in analog VLSI”. In : *2006 IEEE International Symposium on Circuits and Systems*. Mai 2006, 4 pp.-. DOI : 10.1109/ISCAS.2006.1693728.

[SR07] Jeffrey Shafer et Scott Rixner. “RiceNIC : A reconfigurable network interface for experimental research and education”. In : *Proceedings of the 2007 workshop on Experimental computer science*. ACM. 2007, p. 21.

[SSW14] M. Sadeghian, J. E. Stine et E. G. Walters. “Optimized cubic chebyshev interpolator for elementary function hardware implementations”. In : *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. Juin 2014, p. 1536-1539. DOI : 10.1109/ISCAS.2014.6865440.

[Sun+19] Yaohua Sun, Mugen Peng, Yangcheng Zhou, Yuzhe Huang et Shiwen Mao. “Application of machine learning in wireless networks : Key techniques and open issues”. In : *IEEE Communications Surveys & Tutorials* 21.4 (2019), p. 3072-3108.

[Tan+04] R.M. Tanner, D. Sridhara, A. Sridharan, T.E. Duja et D.J. Costello Jr. “LDPC Block and Convolutional Codes Based on Circulant Matrices”. In : *IEEE Transactions on Information Theory* 50 (12 2004). URL : https://ieeexplore.ieee.org/document/1362891 (visité le 10/10/2018).

[TGW99] S. J. G. Taylor, J. Gorjon et P. S. Walker. “An instrumented prosthesis for knee joint force measurement in vivo”. In : *IEE Colloquium on Innovative Pressure, Force and Flow Measurements (Ref. No. 1999/089)*. Oct. 1999, p. 6/1-6/4. DOI : 10.1049/ic:19990496.

[TLW17] Jing Tian, Jun Lin et Zhongfeng Wang. “A 21.66 Gbps Nonbinary LDPC Decoder for High-Speed Communications”. In : *IEEE Transactions on Circuits and Systems II : Express Briefs* PP (mai 2017), p. 1-1. DOI : 10.1109/TCSII.2017.2706273.

[TMG07] S.S. Tehrani, S. Mannor et W.J. Gross. “Survey of Stochastic Computation on Factor Graphs”. In : *Multiple-Valued Logic, 2007. ISMVL 2007. 37th International Symposium on*. Mai 2007, p. 54-54. DOI : 10.1109/ISMVL.2007.53.

[Vog+05] D. Vogrig, A. Gerosa, A. Neviani, A. Graell i Amat, G. Montorsi et S. Benedetto. “A 0.35-μm CMOS Analog Turbo Decoder for the 40-bit Rate 1/3 UMTS Channel Code”. In : *IEEE J. Solid-State Circuits* 40.3 (mar. 2005), p. 753-762.

[VWC12] A. Vosoughi, M. Wu et J. R. Cavallaro. “Baseband signal compression in wireless base stations”. In : *2012 IEEE Global Communications Conference (GLOBECOM)*. Déc. 2012, p. 4505-4511. DOI : 10.1109/GLOCOM.2012.6503828.

[VWD96] Theresa M Vaughan, Jonathan R Wolpaw et Emanuel Donchin. “EEG-based communication : prospects and problems”. In : *IEEE transactions on rehabilitation engineering* 4.4 (1996), p. 425-430.

[WB+19] Gu-Yeon Wei, David Brooks et al. “Benchmarking tpu, gpu, and cpu platforms for deep learning”. In : *arXiv preprint arXiv :1907.10701* (2019).

[WBL69] David J Willshaw, O Peter Buneman et Hugh Christopher Longuet-Higgins. “Non-holographic associative memory”. In : *Nature* 222.5197 (1969), p. 960-962.

[WGS04] C. Winstead, V. Gaudet et C. Schlegel. “A CMOS Analog $(16,11)^2$ Turbo Product Decoder”. In : *Proc. 3rd Analog Decoding Workshop,* Banff, Canada, juin 2004.

[Win+04] C. Winstead, J. Dai, S. Yu, C. Meyers, R.R. Harrison et C. Schlegel. “CMOS Analog MAP Decoder for (8,4) Hamming Code”. In : *IEEE J. Solid-State Circuits* 39.1 (jan. 2004), p. 122-131.

[Win+05] C. Winstead, V.C. Gaudet, A. Rapley et C. Schlegel. “Stochastic iterative decoders”. In : *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*. Sept. 2005, p. 1116-1120. DOI : 10.1109/ISIT.2005.1523513.

[Wol+00] Jonathan R Wolpaw, Niels Birbaumer, William J Heetderks, Dennis J McFarland, P Hunter Peckham, Gerwin Schalk, Emanuel Donchin, Louis A Quatrano, Charles J Robinson et Theresa M Vaughan. “Brain-computer interface technology : a review of the first international meeting”. In : *IEEE transactions on rehabilitation engineering* 8.2 (2000), p. 164-173.

[Yiu+05] M. Yiu, V. C. Gaudet, C. Schlegel et C. Winstead. “Digital built-in selt test of CMOS analog iterative decoders”. In : *Proc. IEEE International Symposium on Circuits and Systems*. Kobe, Japan, mai 2005, p. 2204-2207.

[YXW07] Xiumei Yang, Yong Xiong et Fan Wang. “An Adaptive MIMO System Based on Unified Belief Propagation Detection”. In : *IEEE Int. Conf. on Commun. (ICC)* (2007), p. 209-213. ISSN : 0090-6778.

[ZD15] Ding Zou et Ivan B. Djordjevic. “FPGA implementation of concatenated non-binary QC-LDPC codes for high-speed optical transport”. In : *Opt. Express* 23.11 (juin 2015), p. 14501-14509. DOI : 10.1364/OE.23.014501. URL : http://www.opticsexpress.org/abstract.cfm?URI=oe-23-11-14501.

[Zha+17] Z. Zhao, K. Yang, H. Zheng, F. Gao et X. Bu. “Design, Simulation, and Implementation of a CMOS Analog Decoder for (480,240) Low-Density Parity-Check Code”. In : *IEEE Access* 5 (2017), p. 17381-17391. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2017.2742531.

[Zil+15] N. Zilberman, P. M. Watts, C. Rotsos et A. W. Moore. “Reconfigurable Network Systems and Software-Defined Networking”. In : *Proceedings of the IEEE* 103.7 (juil. 2015), p. 1102-1124. ISSN : 0018-9219. DOI : 10.1109/JPROC.2015.2435732.

[ZMC16] Noa Zilberman, Andrew W. Moore et Jon A. Crowcroft. “From photons to big-data applications : terminating terabits”. In : *Philosophical Transactions of the Royal Society of London A : Mathematical, Physical and Engineering Sciences* 374.2062 (2016). ISSN : 1364-503X. DOI : 10.1098/rsta.2014.0445.