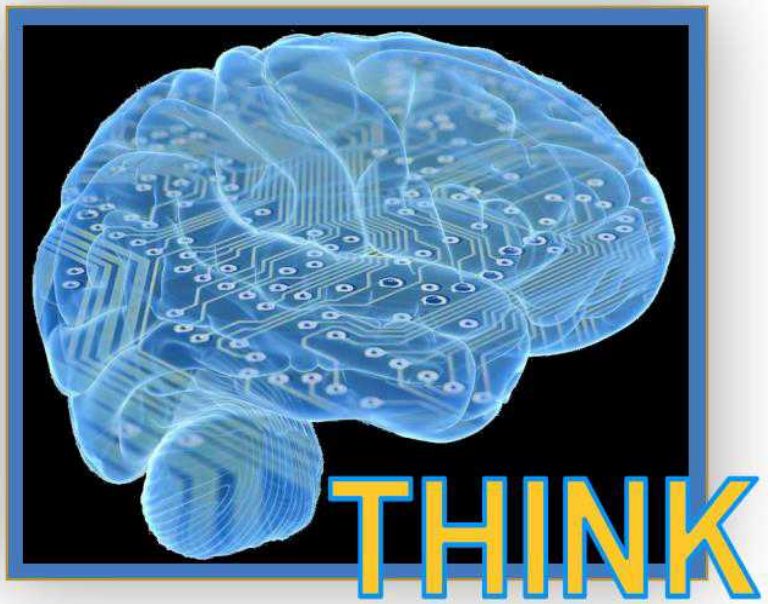
Projet transverse THINK

Testing Hardware Instantiations of Neural Kernels



J.-P. Cachemiche
CPPM

Motivation

Systèmes neuronaux et deep learning

- Montrent leur efficacité dans de nombreux domaines, principalement pour toute forme de reconnaissance
- Technique ancienne mais qui monte en puissance sous deux effets :
  - Disponibilité de larges bases de données
  - Augmentation de puissance des processeurs

Application dans le domaine de la physique

- Augmentation des luminosités et du bruit de fond rends les détections par algorithme classique plus difficiles
  - ➡ Compensable par plus d'intelligence dans la sélection

    2 approches :
    - Architectures triggerless (filtrage logiciel), mais infléchissement de la loi de Moore→ techniques d'accélération impératives
    - Architectures trigger classiques mais rendues plus intelligentes par l'apport d'algorithmes neuronaux
- Est-il possible d'introduire les techniques neuronales très tôt dans la chaîne ? Avec quelle performance ?
  - L'objet de cette étude
  - ➡ Premiers résultats très intéressants dans CMS pour l'extraction de Jets

Situation et roadmap

Constat

- Techniques et supports matériels peu connues dans l'IN2P3
- Besoins de formation exprimés lors de l'ANR DAQ Emergeants sur techniques sous-jacentes
  - Calcul GPU, Architectures MPPA type Kalray, Langages de haut niveau pour FPGAs, Calcul neuronal

But du projet

- Augmenter le niveau de connaissance des ingénieurs et techniciens sur les **techniques d'apprentissage et les techniques neuronales**
  - Organisation de webinaires avec les meilleurs spécialistes
- Identifier quelques applications typiques
- Définir une architecture neuronale appropriée
    - Voir présentation Frédéric Magniette
- Effectuer une phase d'apprentissage
- Effectuer un portage sur différents candidats matériels
  - FPGA, MPPA Kalray, GPU, Processeur neuromorphique (Movidius Intel, BrainChip d'Akira, ...)
- Comparer les performances
- Élaborer des méthodes, des blocs réutilisables
- Diffuser les résultat à la communauté sous forme de workshops

Applications potentielles

A discuter

- Le projet **Amidex OWEN** (Optimal Waveform recognition Electronic Node) qui consiste à développer un nouvel instrument pour traiter le signal venant d'un détecteur innovant, une TPC sphérique à haute pression. Son but est la recherche d'un phénomène rare tel que la détection directe de matière noire et l'observation de la décroissance double béta sans neutrino. Dans ce contexte, il s'agit de développer un système d'acquisition intégrant un algorithme de problème inverse basé sur les réseaux de neurones pour l'identification des formes d'ondes

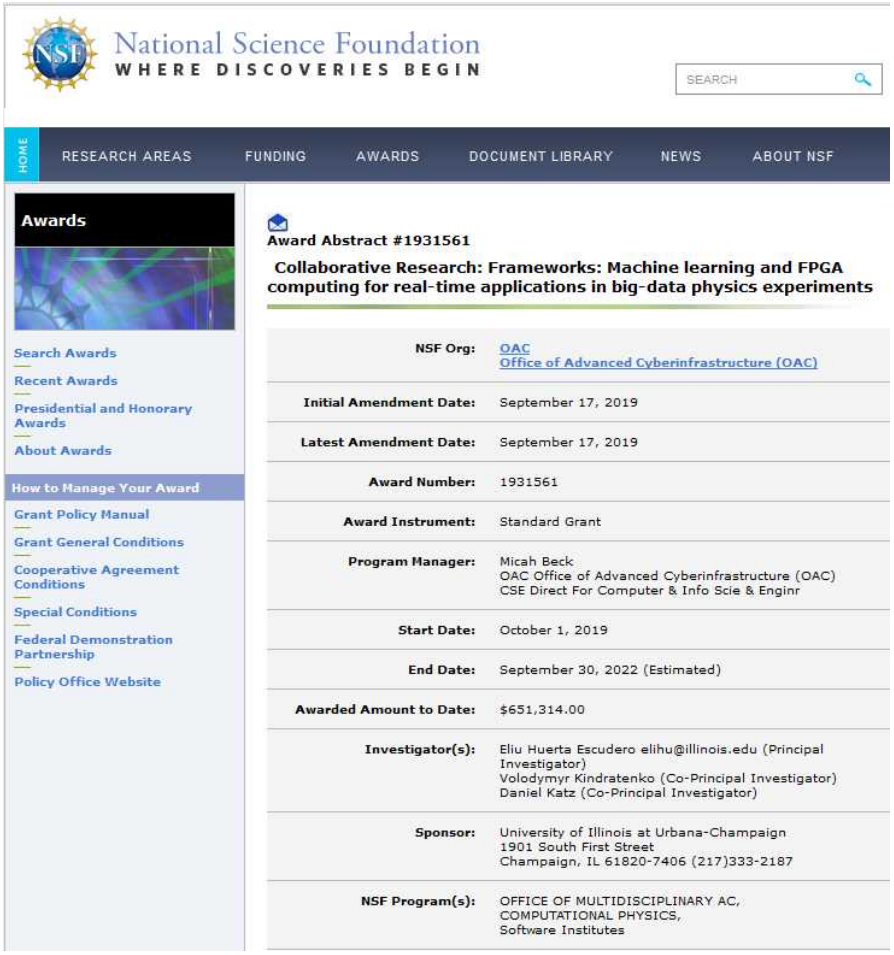
- Le projet **RTA** (Real-Time Analysis) dans l'expérience LHCb qui consiste à traiter 40 Tb des donnees par seconde pour n'en garder que 80Gb/s pour une analyse plus profonde offline. Pour ce faire RTA doit à la fois utiliser efficacement les architectures modernes de calcul, et mettre en place des algorithmes avancés tels que les réseaux neurones.

- Le projet **Amidex AIDAQ** qui consiste à implémenter des algorithmes de reconnaissance neuronale sur FPGA dans le calorimètre à argon liquide d'ATLAS pour réaliser les fonctions de trigger de premier niveau en environnement fortement bruité et avec des niveaux de pile-up variables.

- Le projet **HGCNN** qui consiste à développer des outils d'analyse pour les données des calorimètres à haute granularité (comme le futur calorimètre HGCal de CMS). Ces outils doivent être intégrés dans des FPGA et fournir des primitives de déclenchement avec des latences de l'ordre de la microseconde.

- Les projets d'**imagerie médicale** et en particulier ceux articulés autour des **tomographes** où les problèmes de reconnaissance sont cruciaux.

Projets similaires et premiers travaux



National Science Foundation
WHERE DISCOVERIES BEGIN

SEARCH

HOME | RESEARCH AREAS | FUNDING | AWARDS | DOCUMENT LIBRARY | NEWS | ABOUT NSF

Awards

Search Awards
Recent Awards
Presidential and Honorary Awards
About Awards

How to Manage Your Award
Grant Policy Manual
Grant General Conditions
Cooperative Agreement Conditions
Special Conditions
Federal Demonstration Partnership
Policy Office Website

Award Abstract #1931561

Collaborative Research: Frameworks: Machine learning and FPGA computing for real-time applications in big-data physics experiments

NSF Org:	OAC Office of Advanced Cyberinfrastructure (OAC)
Initial Amendment Date:	September 17, 2019
Latest Amendment Date:	September 17, 2019
Award Number:	1931561
Award Instrument:	Standard Grant
Program Manager:	Micah Beck OAC Office of Advanced Cyberinfrastructure (OAC) CSE Direct For Computer & Info Scie & Enginr
Start Date:	October 1, 2019
End Date:	September 30, 2022 (Estimated)
Awarded Amount to Date:	$651,314.00
Investigator(s):	Eliu Huerta Escudero elihu@illinois.edu (Principal Investigator) Volodymyr Kindratenko (Co-Principal Investigator) Daniel Katz (Co-Principal Investigator)
Sponsor:	University of Illinois at Urbana-Champaign 1901 South First Street Champaign, IL 61820-7406 (217)333-2187
NSF Program(s):	OFFICE OF MULTIDISCIPLINARY AC, COMPUTATIONAL PHYSICS, Software Institutes

Fast inference of deep neural networks in FPGAs for particle physics

Javier Duarte, Song Han, Philip Harris, Sergo Jindariani, Edward Kreinar, Benjamin Kreis, Jennifer Ngadiuba, Maurizio Pierini, Ryan Rivera, Nhan Tran, Zhenbin Wu

(Submitted on 16 Apr 2018 (v1), last revised 28 Jun 2018 (this version, v3))

Recent results at the Large Hadron Collider (LHC) have pointed to enhanced physics capabilities through the improvement of the real-time event processing techniques. Machine learning methods are ubiquitous and have proven to be very powerful in LHC physics, and particle physics as a whole. However, exploration of the use of such techniques in low-latency, low-power FPGA hardware has only just begun. FPGA-based trigger and data acquisition (DAQ) systems have extremely low, sub-microsecond latency requirements that are unique to particle physics. We present a case study for neural network inference in FPGAs focusing on a classifier for jet substructure which would enable, among many other physics scenarios, searches for new dark sector particles and novel measurements of the Higgs boson. While we focus on a specific example, the lessons are far-reaching. We develop a package based on High-Level Synthesis (HLS) called hls4ml to build machine learning models in FPGAs. The use of HLS increases accessibility across a broad user community and allows for a drastic decrease in firmware development time. We map out FPGA resource usage and latency versus neural network hyperparameters to identify the problems in particle physics that would benefit from performing neural network inference with FPGAs. For our example jet substructure model, we fit well within the available resources of modern FPGAs with a latency on the scale of 100 ns.

Planning

Projet sur 3 ans

	2020				2021				2022			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Trainings												
Choix d'applications pertinentes												
Apprentissage												
Implémentation matérielle												
Evaluation												
Mise à disposition outils et blocs réutilisables												

Participants

Physiciens

Nom	Prénom	Laboratoire	Statut	% ETP
Gligorov	Vladimir	LPNHE	DR	5%
Monnier	Emmanuel	CPPM	DR	5%
Aad	George	CPPM	CR	10%
Calvet	Thomas	CPPM	CR	10%
Boursier	Yannick	CPPM	MDC	5%

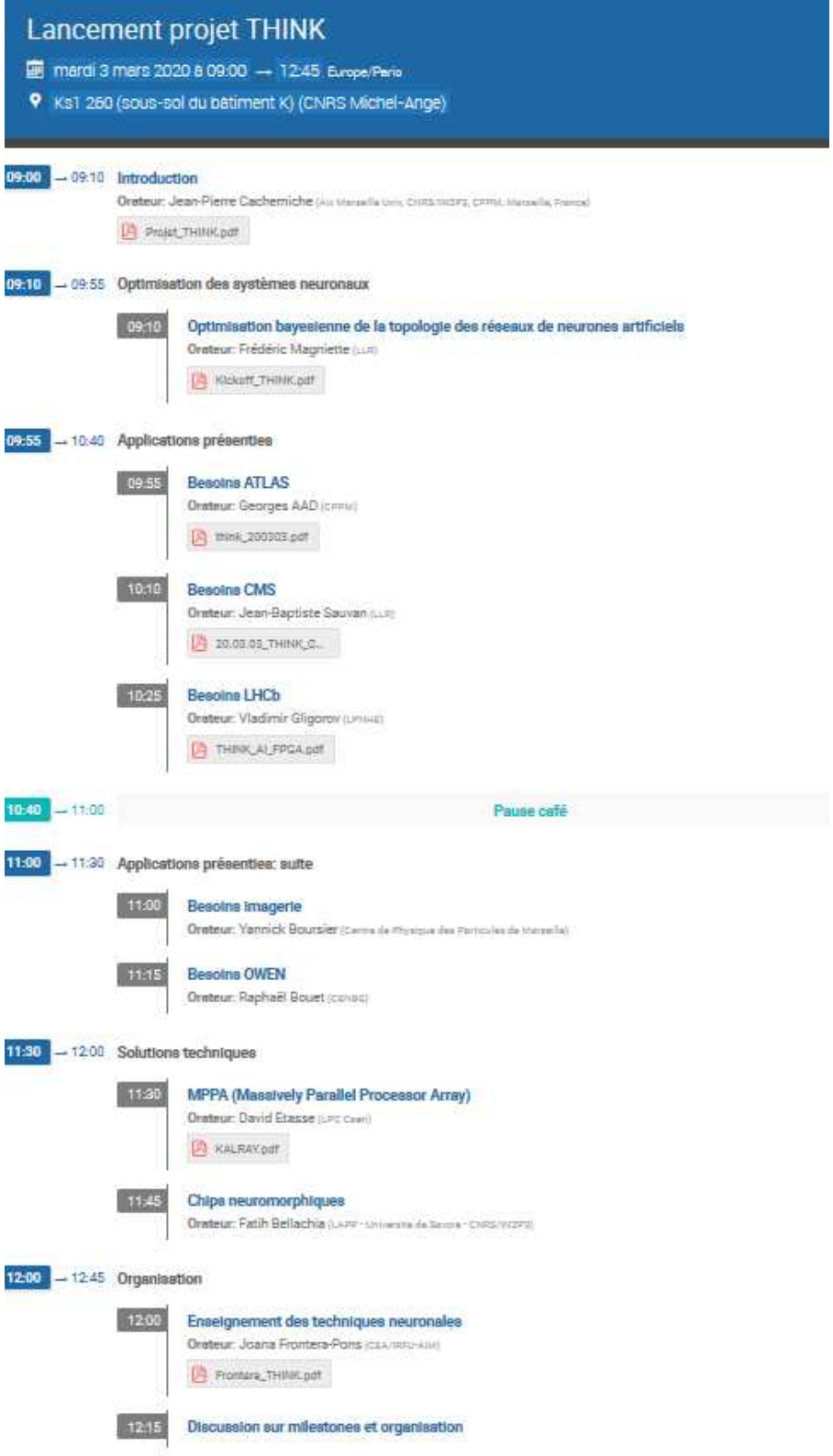
Ingénieurs

Nom	Prénom	Laboratoire	Statut	% ETP
Cachemiche	Jean-Pierre	CPPM	IR	10%
Le Dortz	Olivier	LPNHE	IR	10%
Druillole	Frédéric	CENBG	IR	15%
Bouet	Raphaël	CENBG	CDD IE	20%
Rebii	Abdel	CENBG	IR	15%
Etasse	David	LPC Caen	IR	10%
Hommet	Jean	LPC Caen	IR	10%
Bellachia	Fatih	LAPP	IE	10%
Lafrasse	Sylvain	LAPP	IE	10%
Magniette	Frédéric	LLR	IR	10%
Frontera-Pons	Joana	IRFU/AIM	IR	5%

Responsabilités

- **IRFU/AIM :** aspects théoriques et formation
- **Tous** : choix d'applications représentatives
- **LLR :** sélection des structures neuronales
- **LPC Caen** : portage sur MPPA, éventuellement sur carte développée par le laboratoire
- **LAPP** : portage sur processeur neuromorphique
- **LPNHE** : portage sur FPGA et GPU
- **CENGB** : portage sur FPGA Xilinx
- **CPPM** : coordination du projet, portage sur FPGA Intel et sur GPU

Agenda



Lancement projet THINK

mardi 3 mars 2020 à 09:00 → 12:45 Europe/Paris

Ks1 260 (sous-sol du bâtiment K) (CNRS Michel-Ange)

09:00 → 09:10 **Introduction**
Orateur: Jean-Pierre Cachemiche
Projet_THINK.pdf

09:10 → 09:55 **Optimisation des systèmes neuronaux**

09:10 **Optimisation bayesienne de la topologie des réseaux de neurones artificiels**
Orateur: Frédéric Magniette
Kickoff_THINK.pdf

09:55 → 10:40 **Applications présentées**

09:55 **Besoins ATLAS**
Orateur: Georges AAD
think_200303.pdf

10:10 **Besoins CMS**
Orateur: Jean-Baptiste Sauvan
20.03.03_THINK_C...

10:25 **Besoins LHCb**
Orateur: Vladimir Gligorov
THINK_AI_FPGA.pdf

10:40 → 11:00 **Pause café**

11:00 → 11:30 **Applications présentées: suite**

11:00 **Besoins imagerie**
Orateur: Yannick Boursier

11:15 **Besoins OWEN**
Orateur: Raphaël Bouet

11:30 → 12:00 **Solutions techniques**

11:30 **MPPA (Massively Parallel Processor Array)**
Orateur: David Etasse
KALRAY.pdf

11:45 **Chips neuromorphiques**
Orateur: Fatih Bellachia

12:00 → 12:45 **Organisation**

12:00 **Enseignement des techniques neuronales**
Orateur: Joana Frontera-Pons
Frontera_THINK.pdf

12:15 **Discussion sur milestones et organisation**