Lunch Seminar

Memory Study and Optimisation within a Neural Network Hardware Accelerator

Fabio Maschi

23 November 2018

Presentation

I. Education

- 2015-2018 **Master of Engineering**
  Electrical and Electronic Engineering – Paris-Saclay University, France
  2017-18 Double degree, **M.Sc** Embedded Systems and Information Processing
  2016-17 **Erasmus** at University College London, England

- 2014-2015 **Bachelor of Sciences**
  Electrical Engineering – Paris-Sud University, France

- 2011-2014 **Bachelor of Engineering** (not completed)
  Computer Engineering – Pontifical Catholic University RS, Brazil

Presentation

II. Research

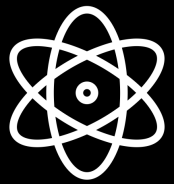
- Jun - Sep 2017 - Internship at **INESC-ID**, University of Lisbon, Portugal
  Exploitation of Data-Flow Parallelism and Stream Processing Paradigms within Reconfigurable Processor Architecture

- Feb - Sep 2018 - Master Thesis, **CEA**, France
  Memory Study and Optimisation within a Neural Network Hardware Accelerator

Introduction

I. Working environment

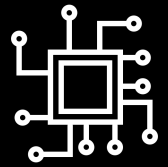
- French Commission for Atomic Energy
- Created at the end of World War II by General Charles de Gaulle
- “From research to industry”
- Research, development, education and deployment
- Four operational instances:



Nuclear Energy Division

Military Applications Division

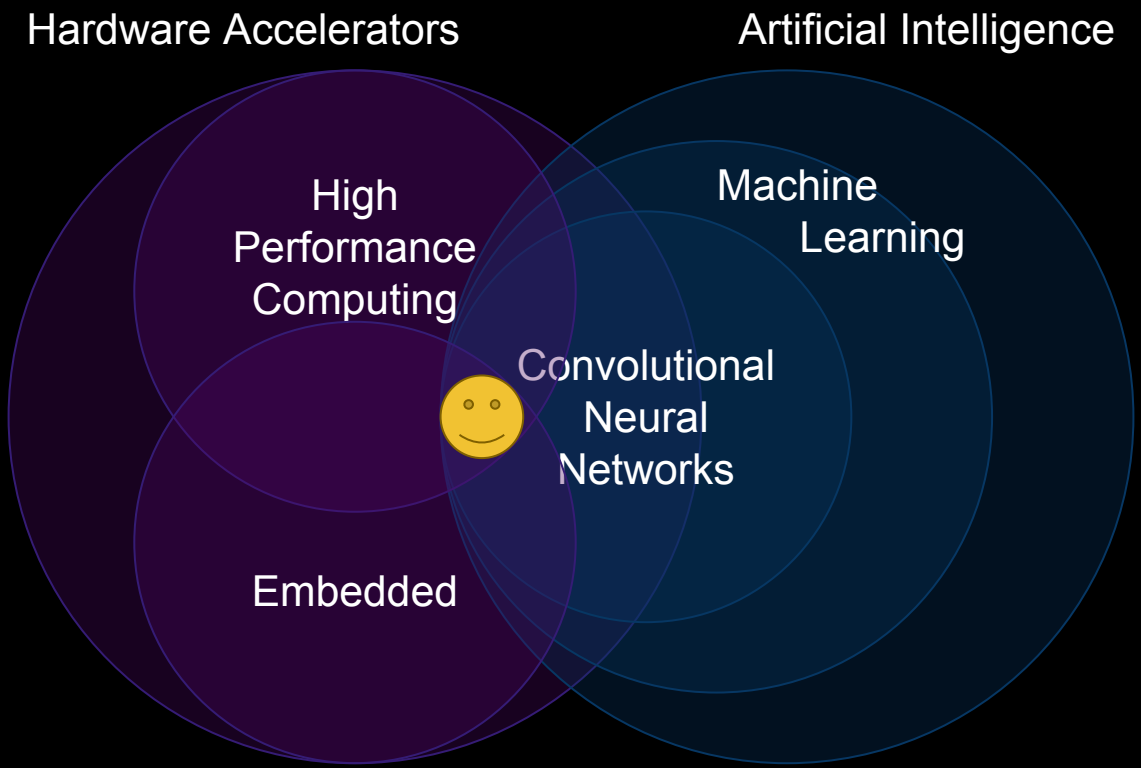
Fundamental Research Division



Technology Research Division

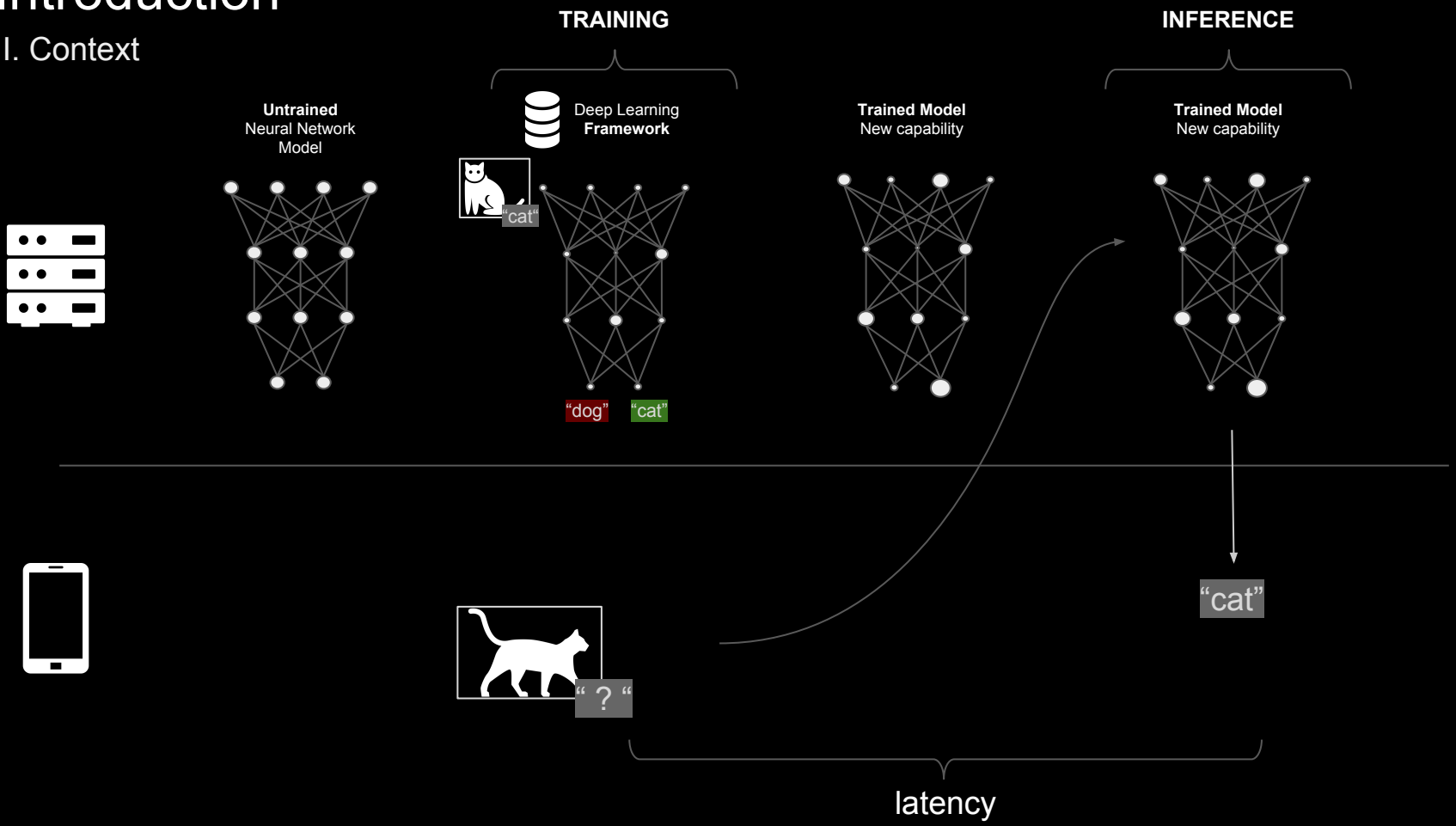
Introduction

II. Context



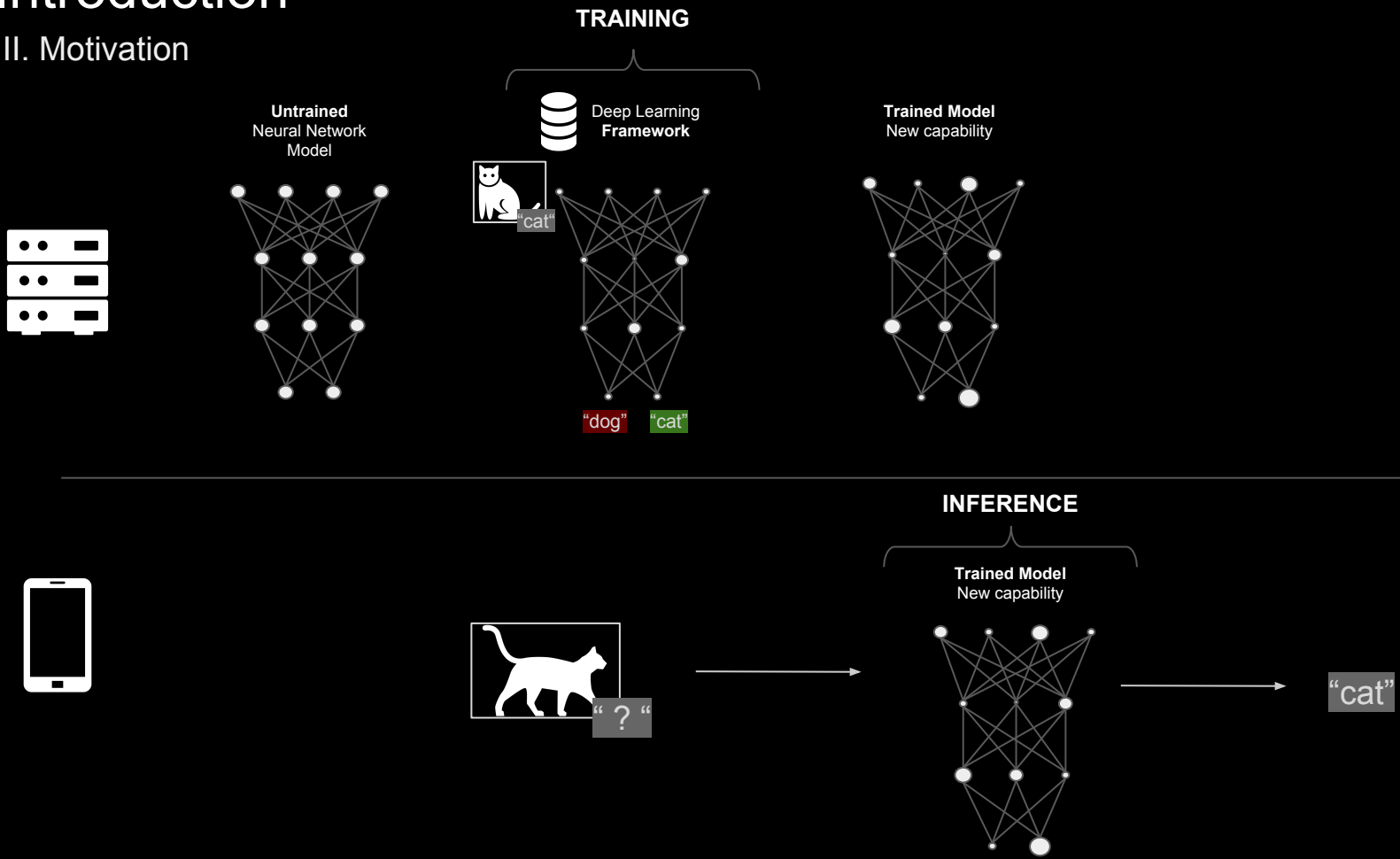
Introduction

II. Context



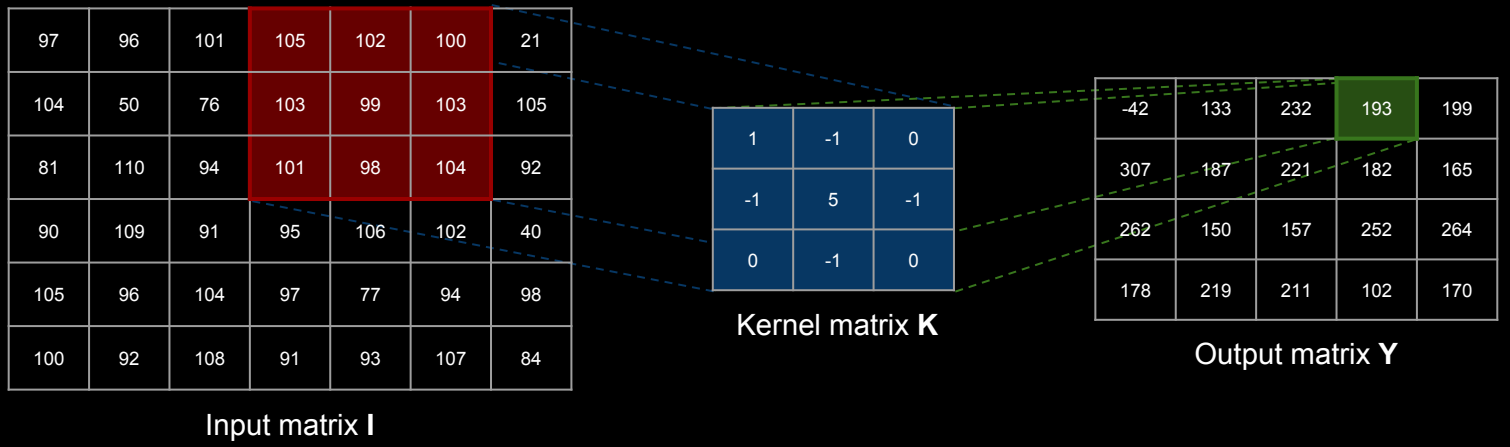
Introduction

III. Motivation



Introduction

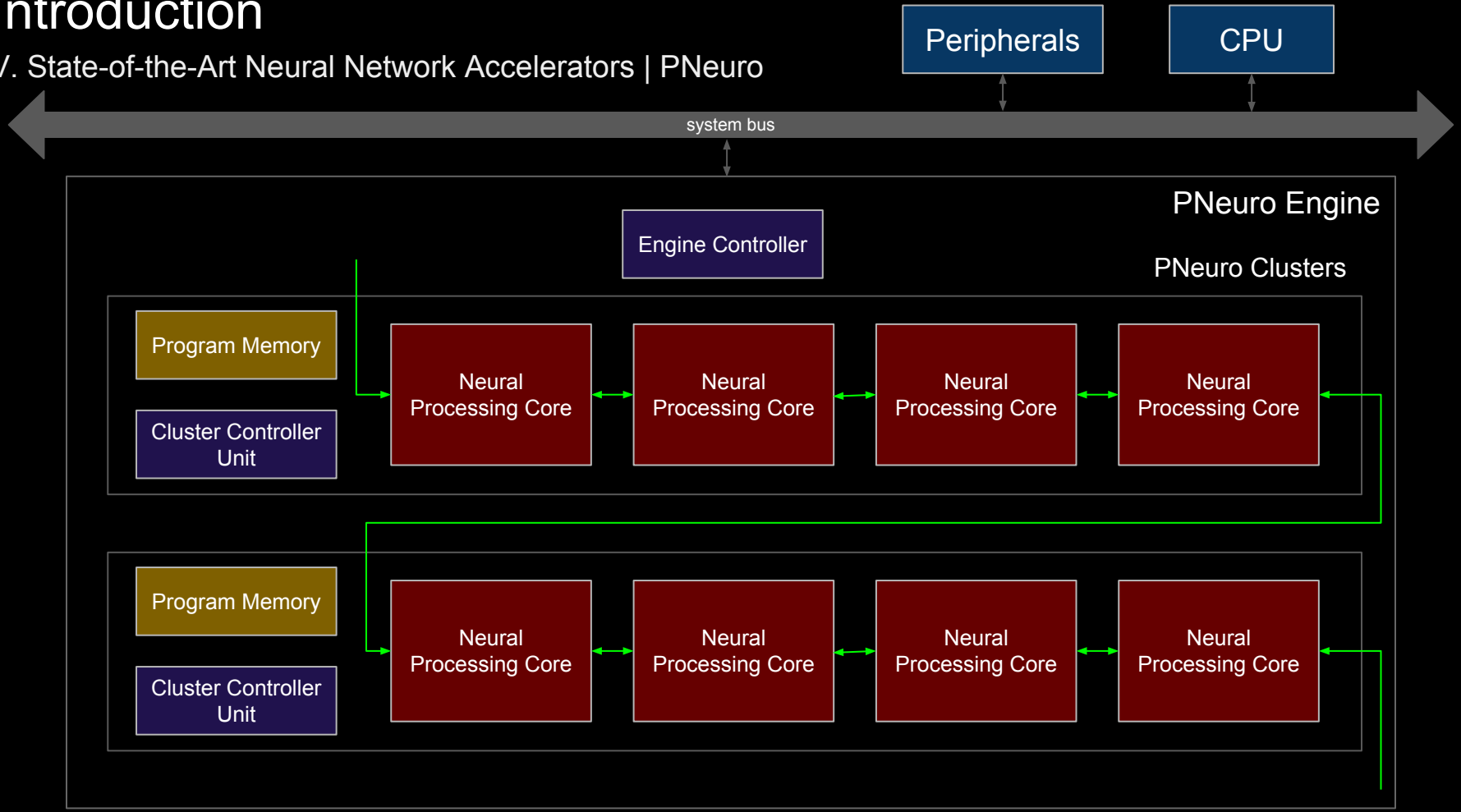
IV. Background



$$Y(j,k) = I * K = \sum_{p} \sum_{q} I(p,q) \times K(j-p, k-q)$$

Introduction

V. State-of-the-Art Neural Network Accelerators | PNeuro

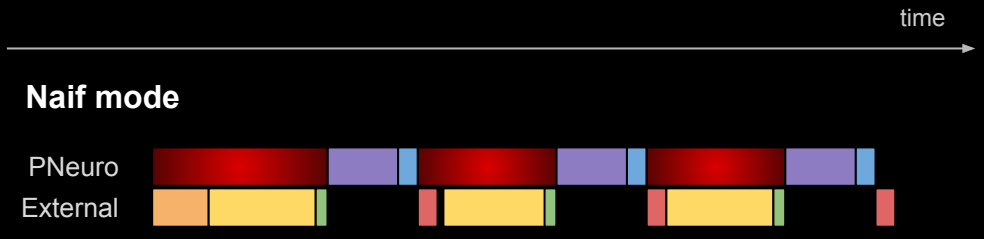


Task I: Communication Protocol

I. Design

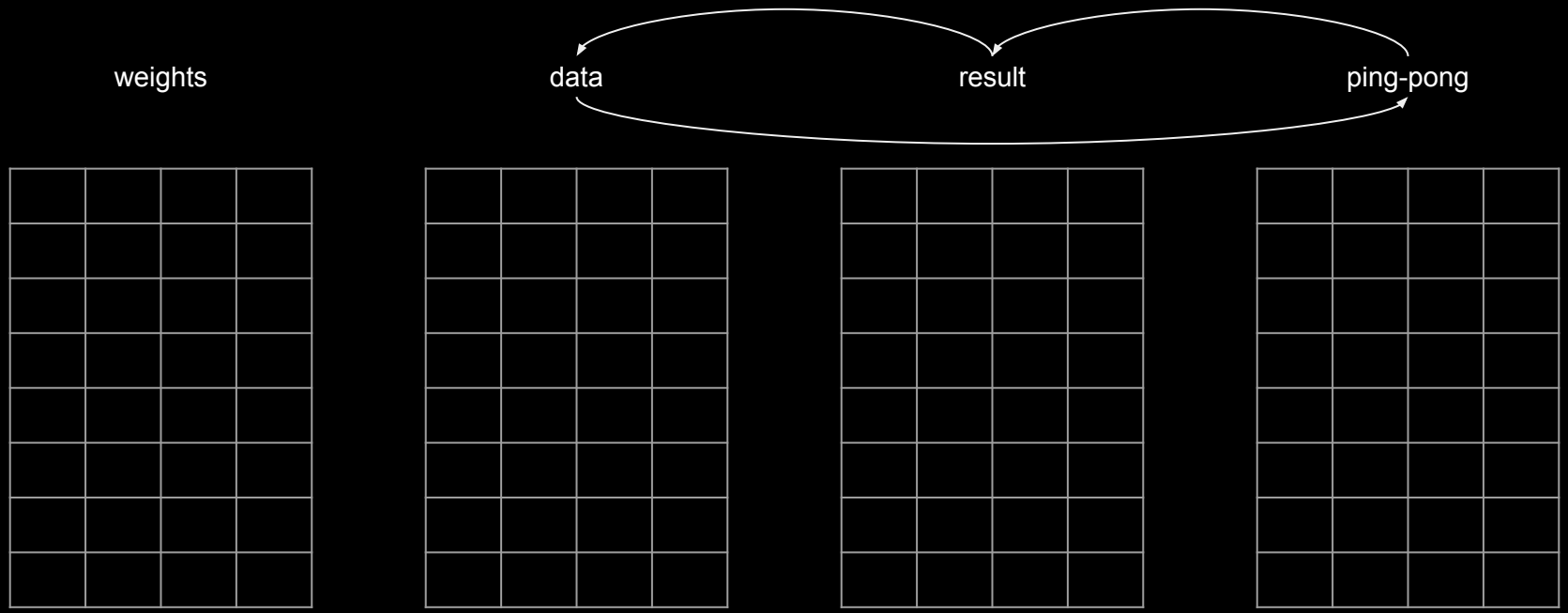
Task II: Memory Investigation

I. Problem Statement



Task II: Memory Investigation

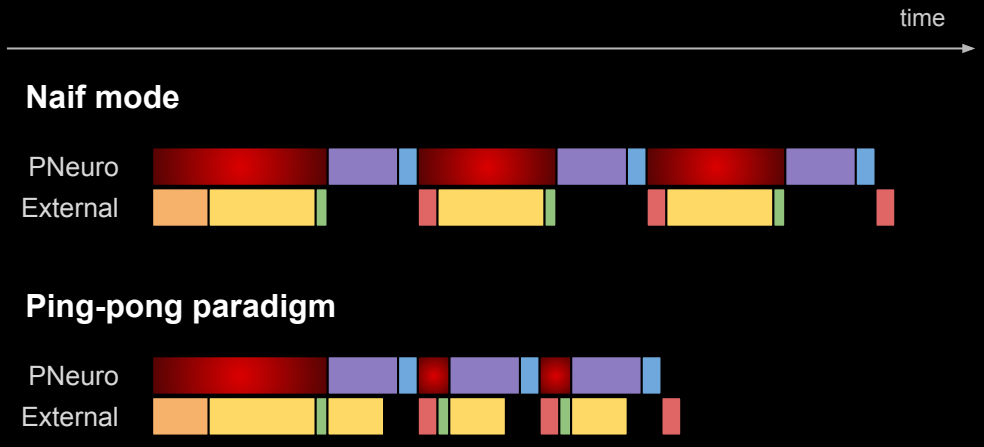
I. Problem Statement



- Single-port cuts

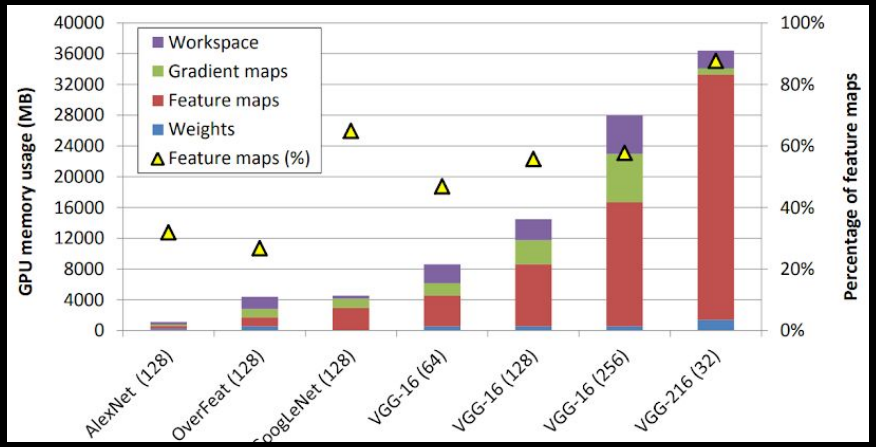
Task II: Memory Investigation

I. Problem Statement



Task II: Memory Investigation

I. Problem Statement

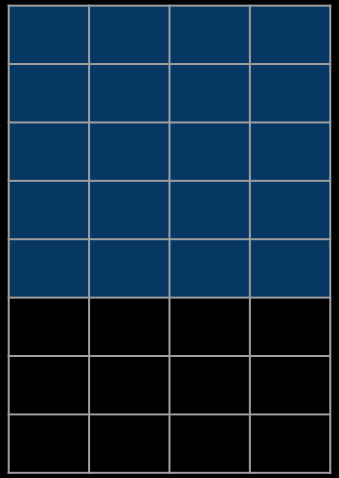


M. Rhu, N. Gimelshein, J. Clemons, A. Zulfiqar, and S. Keckler. "**vDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design**". In *The 49th Annual IEEE/ACM International Symposium on Microarchitecture* (p. 18). IEEE Press, Oct. 2016.

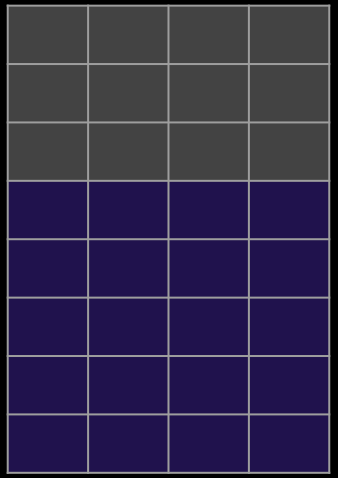
Task II: Memory Investigation

II. Design

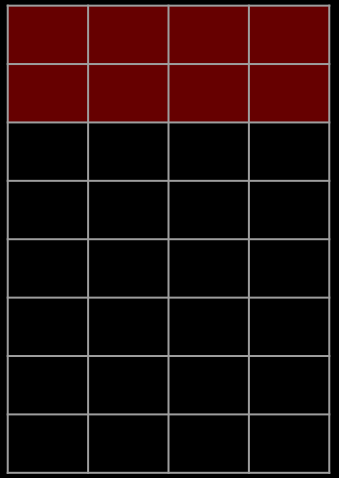
weights



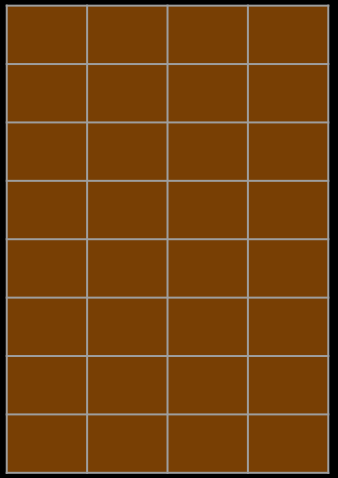
data



result



ping-pong



Task II: Memory Investigation

II. Design

Application criteria:

- Data production rate
- Data consumption rate
- CNN Model

Architecture influence:

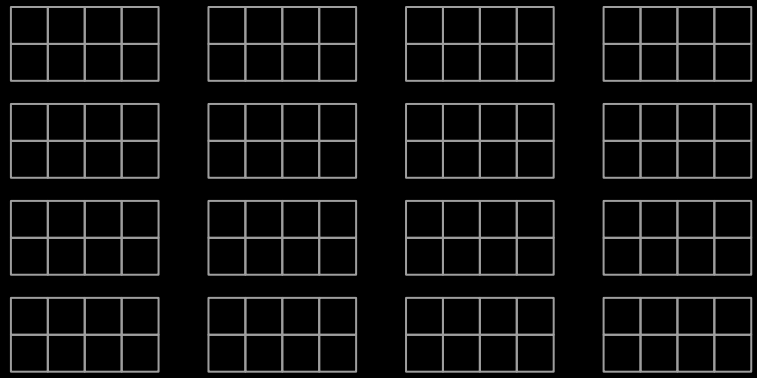
- Burst size (optimal transfer)
- Synchronisation latency
- Bus width

Hardware (memory cuts) criteria:

- Surface density
- Timing
- Power
- Leakage

Compromise between accelerator flexibility, target application(s) and performance.

Geometry	Timing		Powers		Leakage	
Density	Cycle time	Data setup	Read	Write	Active	Standy
0.83	1.54	0.62	1.07	1.11	0.20	0.21



Task II: Memory Investigation

III. Discussion

Contributions

- Granted a fine-grain memory management
- Synchronisation mechanism

Results

- Speed-up of 3.91 compared to the original platform (same non-optimised CNN implementation)

Perspectives

- Optimised and more complex implementations of neural network models on the accelerator

References and credits

1. A. Carbon, J.-M. Philippe, O. Bichler, R. Schmit, B. Tain, D. Briand, N. Ventroux, M. Paindavoine, and O. Brousse, “**PNeuro: A scalable energy-efficient programmable hardware accelerator for neural networks**”, in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 1039–1044, Mar. 2018.

2. A. Oleksiak, M. Kierzynka, W. Piatek, G. Agosta, A. Barenghi, C. Brandolese, and J. Činkelj. “**M2DC–Modular Microserver Data Centre with heterogeneous hardware”**. *Microprocessors and Microsystems*, 52, 117-130, 2017.

3. M. Rhu, N. Gimelshein, J. Clemons, A. Zulfiqar, and S. Keckler. “**vDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design”**. In *The 49th Annual IEEE/ACM International Symposium on Microarchitecture* (p. 18). IEEE Press, Oct. 2016.

4. M. Copeland, “**What’s the difference between deep learning training and inference?**” Online, Aug. 2018. https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/

5. Images from the Noun Project: **nuclear** by Ralf Schmitzer; **military** by myiconfinder; **research** by Minnie Pigeon; **technology** by Aiden Icons; **server** by Chunk Icons; **phone** by REVA; **database** by Noura Mbarki; **cat** by Hea Poh Lin; **cat** by Gan Khoon Lay.