Imperial College London

fpgaConvNet: A Framework for Mapping Convolutional Neural Networks on FPGAs

Stylianos I. Venieris, Christos-Savvas Bouganis
stylianos.venieris10@imperial.ac.uk

FCCM 2016, Washington DC
2 May 2016

Imperial College London

Deep Learning and AI

Industry Voice SMB Spotlight

the INQUIRER

MIT Technology Review

10 Breakthrough Technologies

2014

Neuromorphic Chips

Microprocessors configured more like brains than traditional chips could soon make computers far more astute about what's going on around them.

Imperial College London

Deep Learning Success Stories - ConvNets

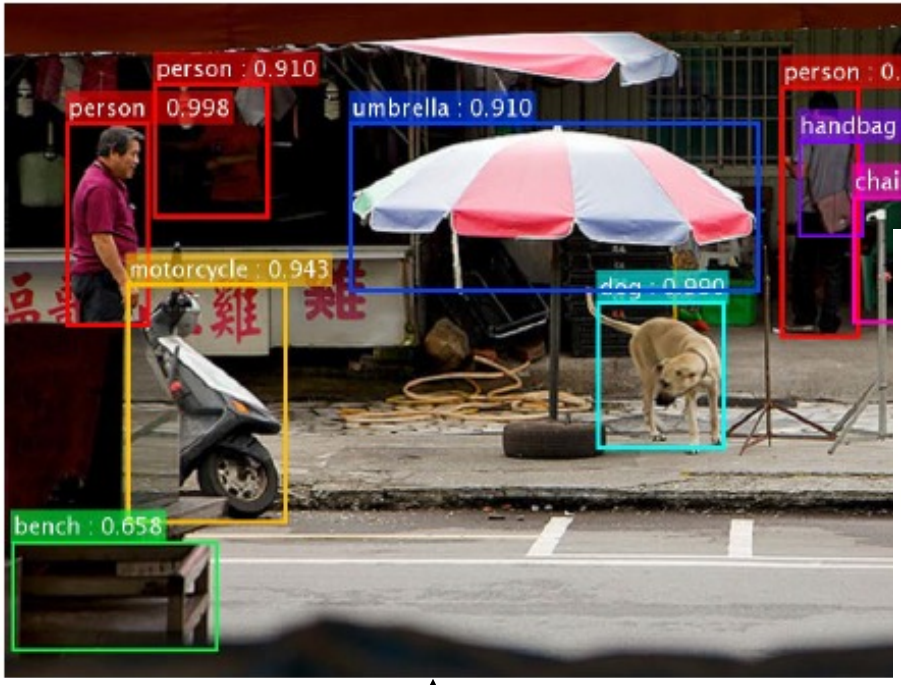
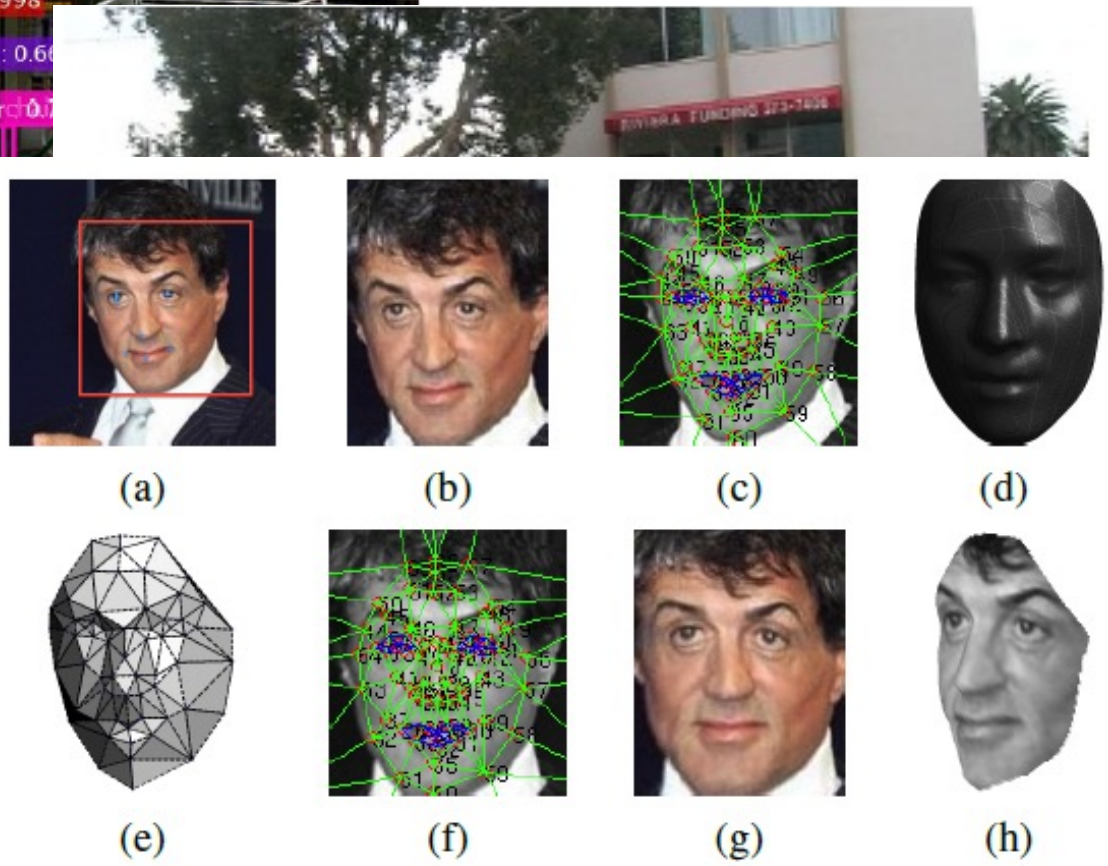


Image Recognition (Microsoft, 2015



"Deep Face" (Facebook, 2014)

Image Captioning (Microsoft, 2015)

Imperial College London

Deep Learning on FPGAs

- Hand-tuned implementations

Hardware Accelerated Convolutional Neural Networks for Synthetic Vision Systems

Clément Farabet[1,2], Berin Martini[2], Polina Akselrod[2], Selçuk Talay[2], Yann LeCun[1] and Eugenio Culurciello[2]
[1] The Courant Institute of Mathematical Sciences and Center for Neural Science, New York University, USA
[2] Electrical Engineering Department, Yale University, New Haven, USA

Going Deeper with Embedded FPGA Platform for Convolutional Neural Network

Jiantao Qiu[1,2], Jie Wang[1], Song Yao[1,2], Kaiyuan Guo[1,2], Boxun Li[1,2],Erjin Zhou[1], Jincheng Yu[1,2], Tianqi Tang[1,2], Ningyi Xu[3], Sen Song[2,4], Yu Wang[1,2], and Huazhong Yang[1,2]

- Memory I/O Optimisation

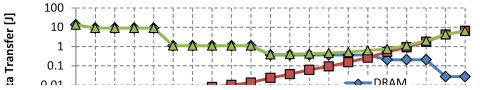
Memory Access Optimized Routing Scheme for Deep Networks on a Mobile Coprocessor

Aysegul Dundar*, Jonghoon Jin†, Vinayak Gokhale†, Berin Martini* and Eugenio Culurciello*†
*Weldon School of Biomedical Engineering, Purdue University

Memory-Centric Accelerator Design for Convolutional Neural Networks

Maurice Peemen, Arnaud A. A. Setio, Bart Mesman and Henk Corporaal
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
Email: m.c.j.peemen@tue.nl, arnaud.arindra.adiyoso@student.tue.nl, b.mesman@tue.nl, h.corporaal@tue.nl

***Abstract*—In the near future, cameras will be used everywhere as flexible sensors for numerous applications. For mobility and privacy reasons, the required image processing should be local on embedded computer platforms with performance requirements**



- Design Space Exploration

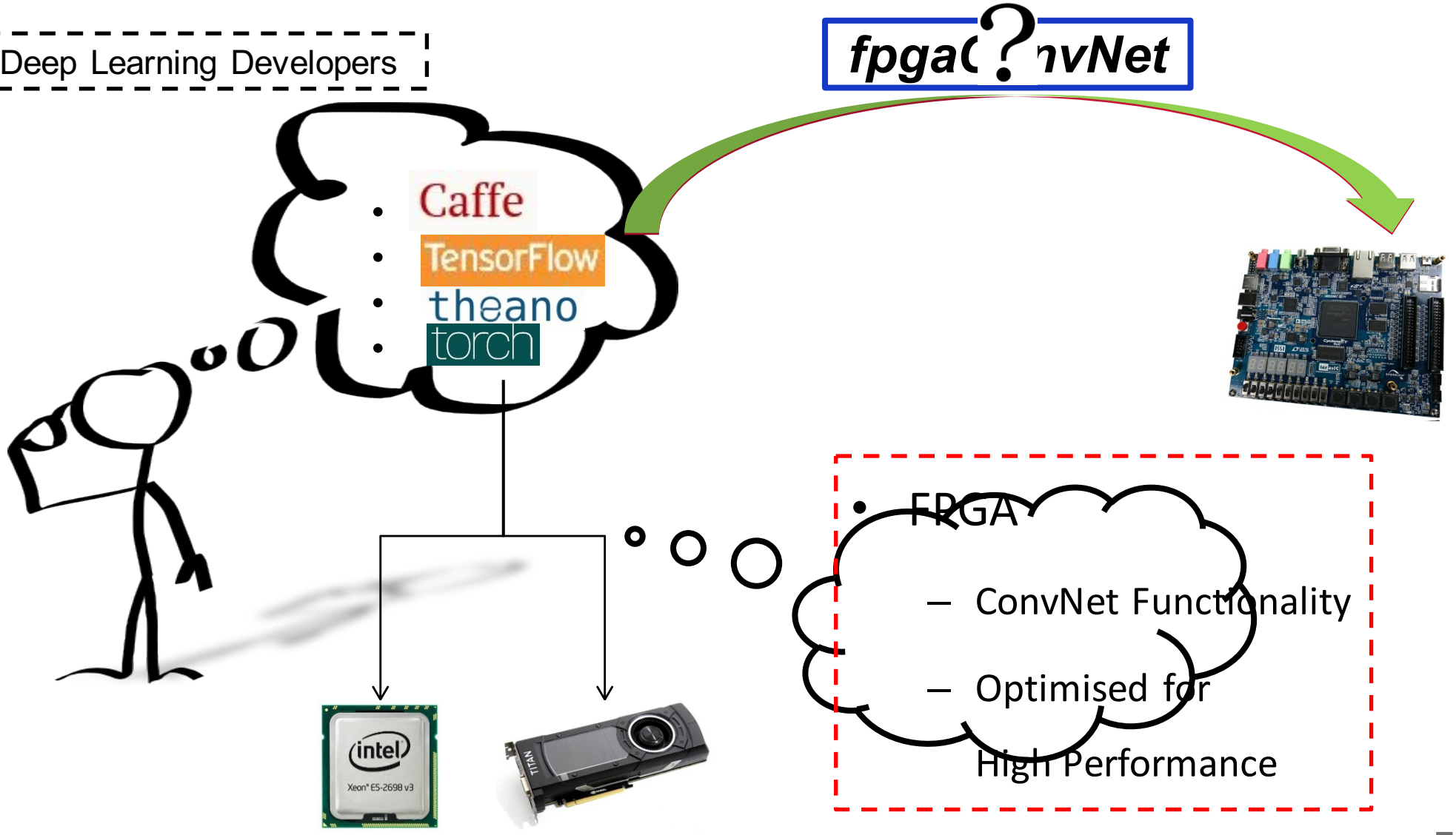
Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks

Chen Zhang[1] chen.ceca@pku.edu.cn
Peng Li[2] pengli@cs.ucla.edu
Guangyu Sun[1,3] gsun@pku.edu.cn
Yijin Guan[1] guanyijin@pku.edu.cn
Bingjun Xiao[2] xiao@cs.ucla.edu
Jason Cong [2,3,1,*] cong@cs.ucla.edu

Throughput-Optimized OpenCL-based FPGA Accelerator for Large-Scale Convolutional Neural Networks

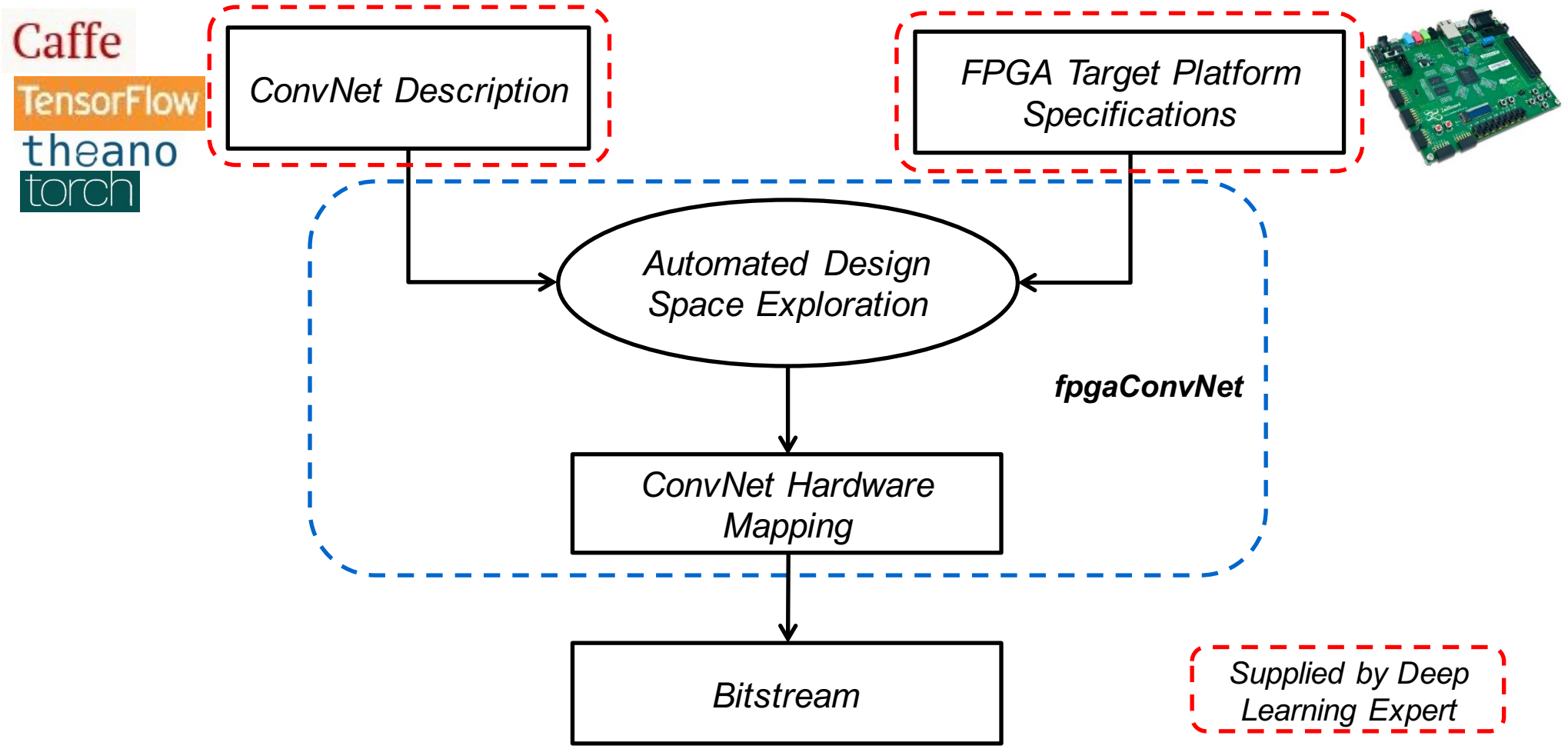
Naveen Suda, Vikas Chandra‡, Ganesh Dasika*, Abinash Mohanty, Yufei Ma, Sarma Vrudhula†, Jae-sun Seo, Yu Cao

What is missing?



Imperial College London

Our approach - fpgaConvNet



Imperial College London

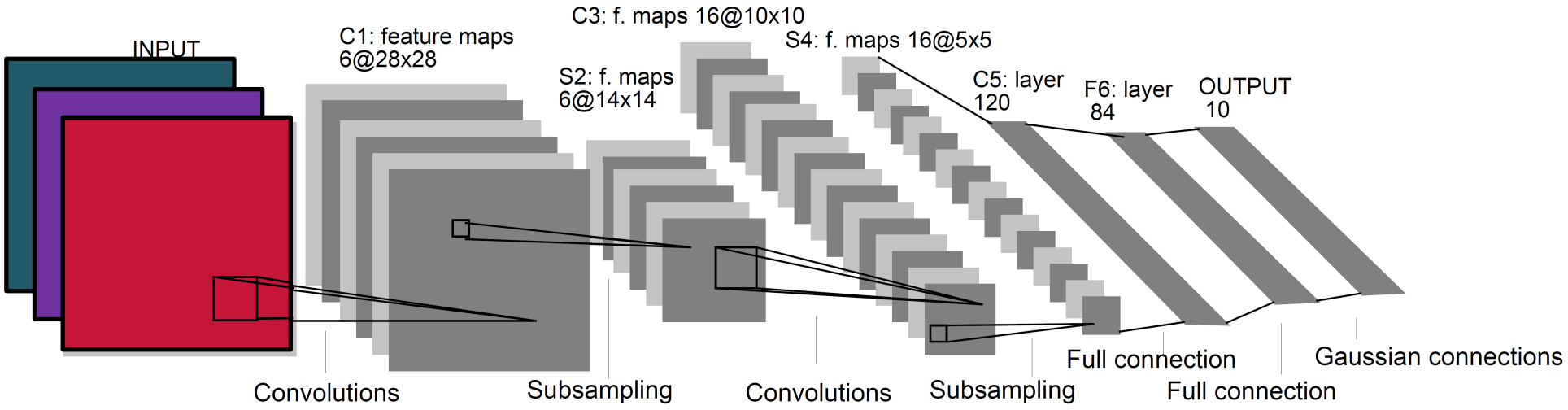
Convolutional Neural Networks (ConvNets)

convolutional + nonlinearity

pooling

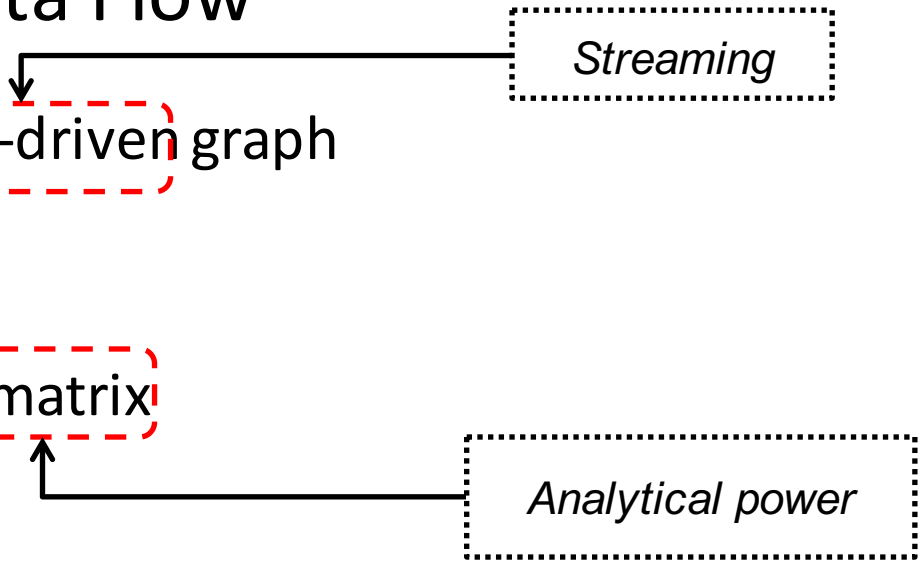
convolutional + nonlinearity

pooling



fpgaConvNet – ConvNet Modelling Framework

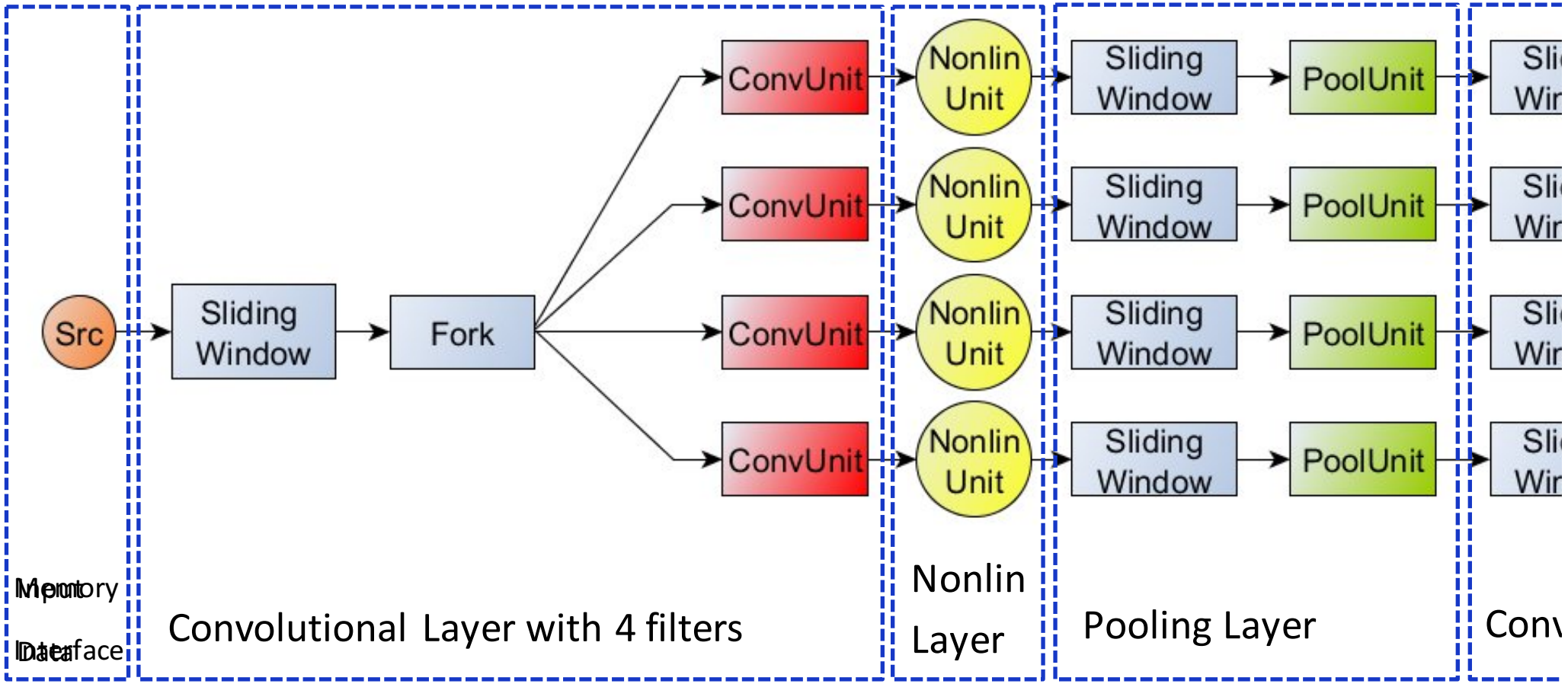
- Synchronous Data Flow
  - ConvNet as a data-driven graph
  - Represented as a matrix
  - Each layer mapped to a tunable set of *hardware building blocks*



Imperial College London

fpgaConvNet – Modelling ConvNets with SDF

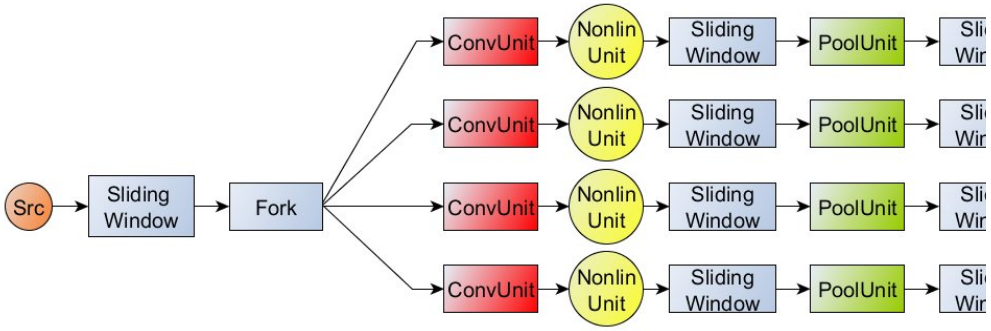
ConvNet Hardware SDF Graph

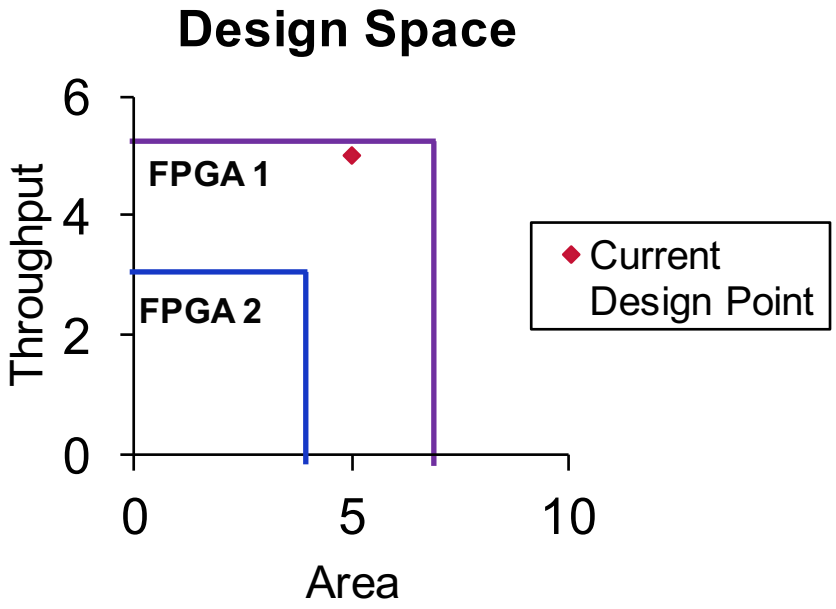


Imperial College London

fpgaConvNet – Design Space Perspective

ConvNet Hardware SDF Graph



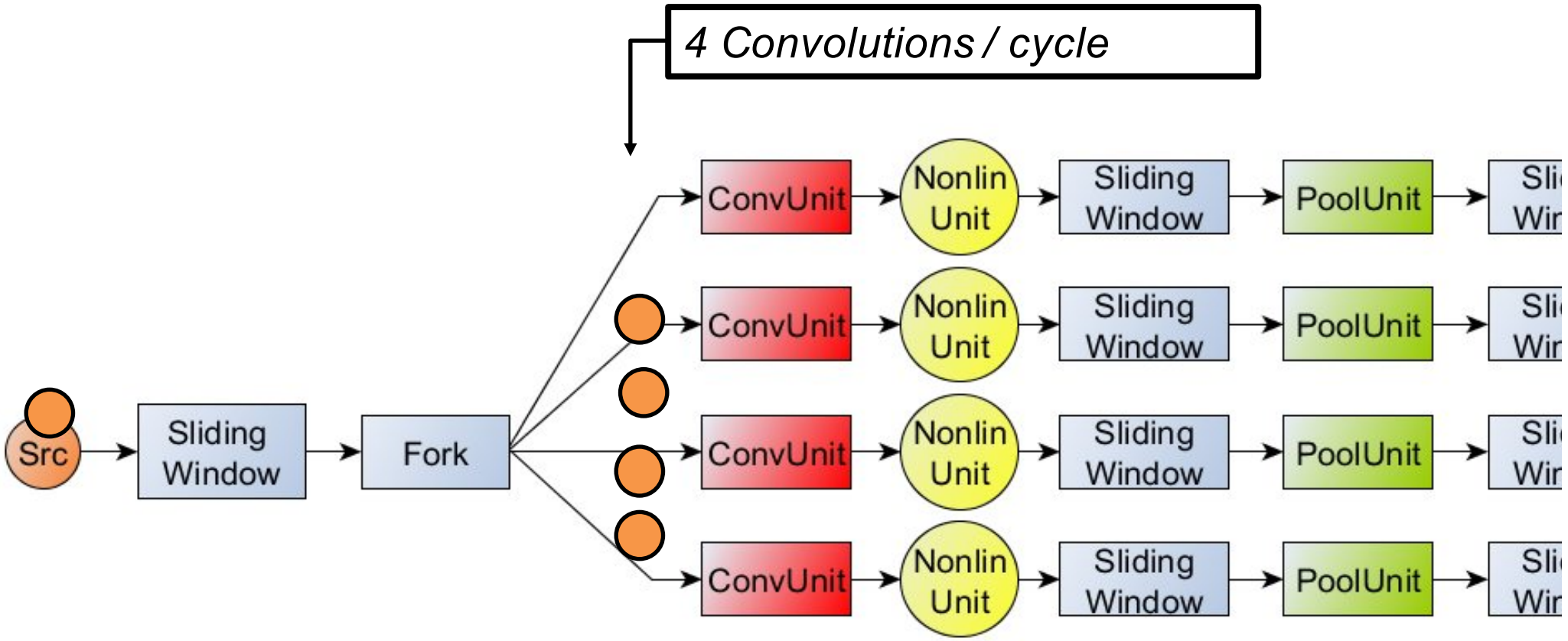


Bottlenecks:

- Limited *compute resources*
- Limited *off-chip memory bandwidth*
- Limited *on-chip memory* for model parameters

Define a set of *actions* to move around the design space

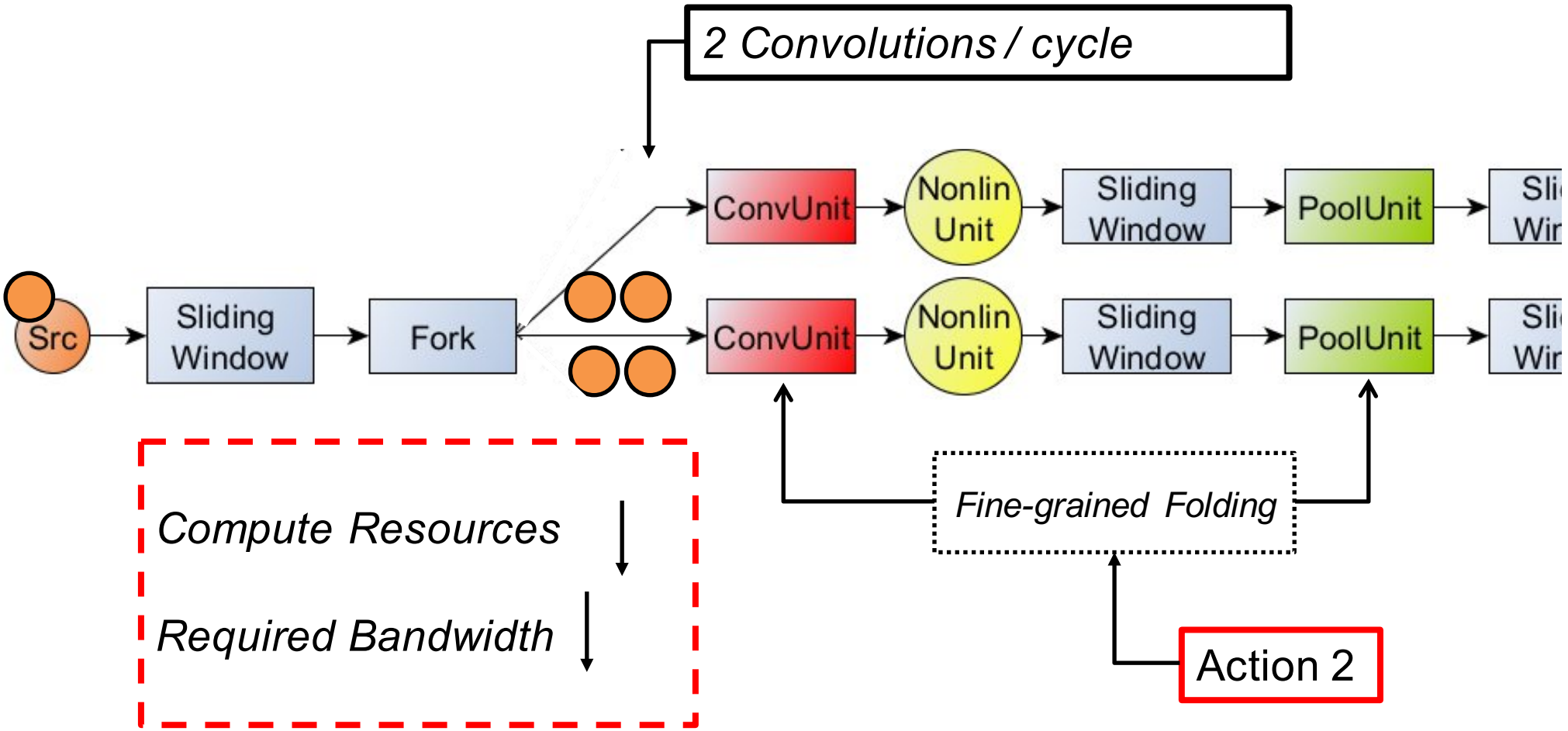
Action 1: *Coarse-grained Folding*



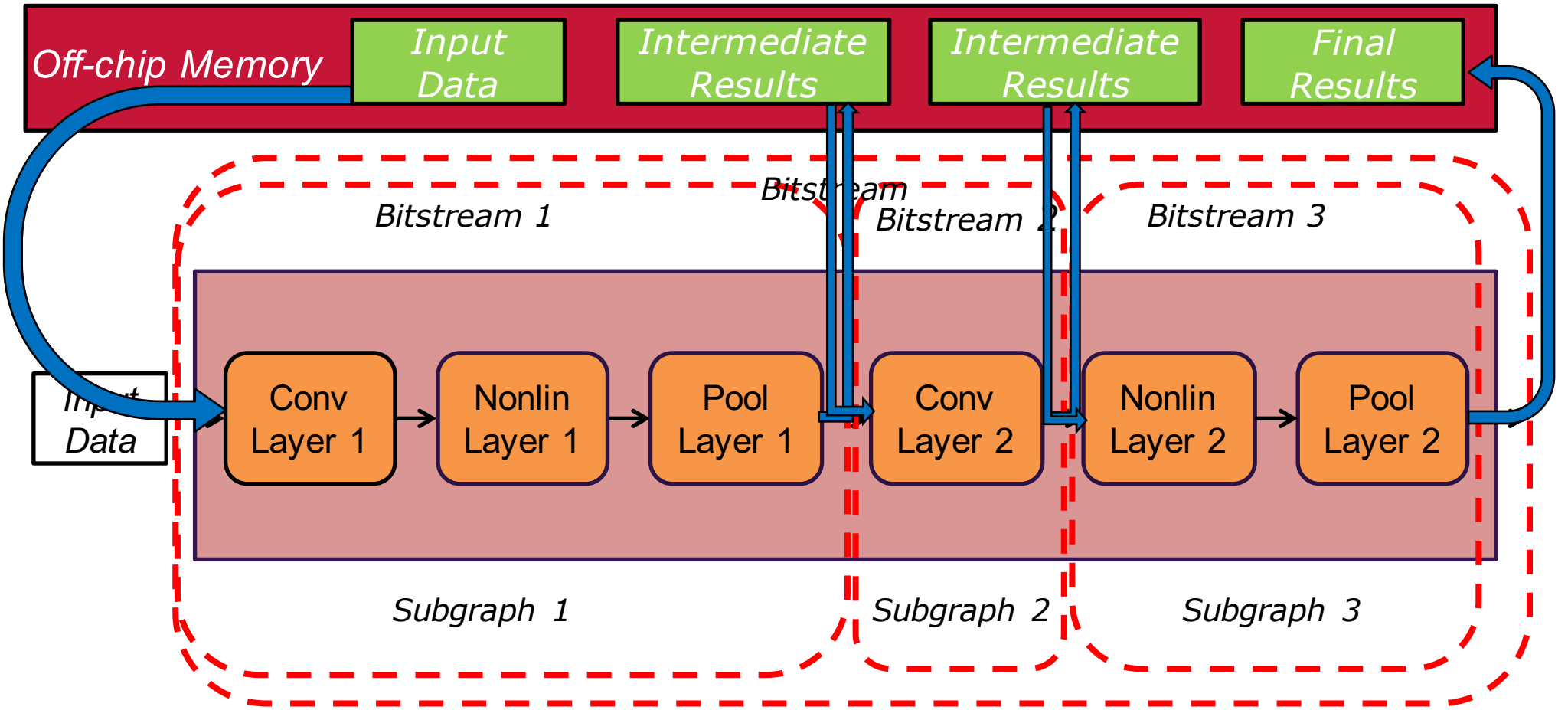
1) Exceeding the available compute resources

2) Not enough off-chip memory bandwidth

Action 1: *Coarse-grained Folding*



Action 3: *Partitioning through Reconfiguration*



Exceeding the available compute resources

Not enough on-chip memory

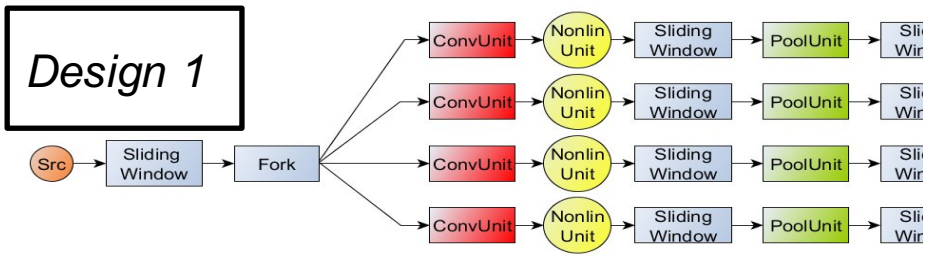
→ *FPGA Reconfiguration*

Imperial College London

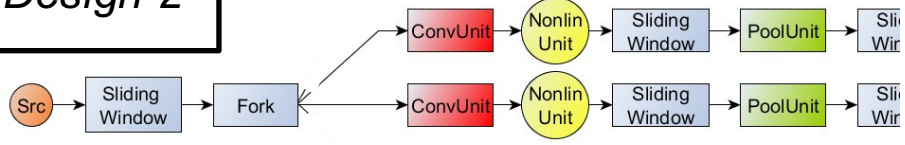
fpgaConvNet – SDF Analytical Power

Window Size = K
Pool Size = P

Design 1



Design 2



Hardware Stages →

Interconnections ↓

$$\Gamma_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & K^2 & -K^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4K^2 & -4K^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4P^2 & -4P^2 \end{bmatrix}$$

$$\Gamma_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & K^2 & -K^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2K^2 & -2K^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2P^2 & -2P^2 \end{bmatrix}$$

- Synchronous Data Flow
  - Actions as *algebraic operations*
  - Any local action *propagates* through the network
  - *Static scheduling*
  - Analytical *Performance Model*
  - Cast DSE to formal *resource-constrained optimisation*

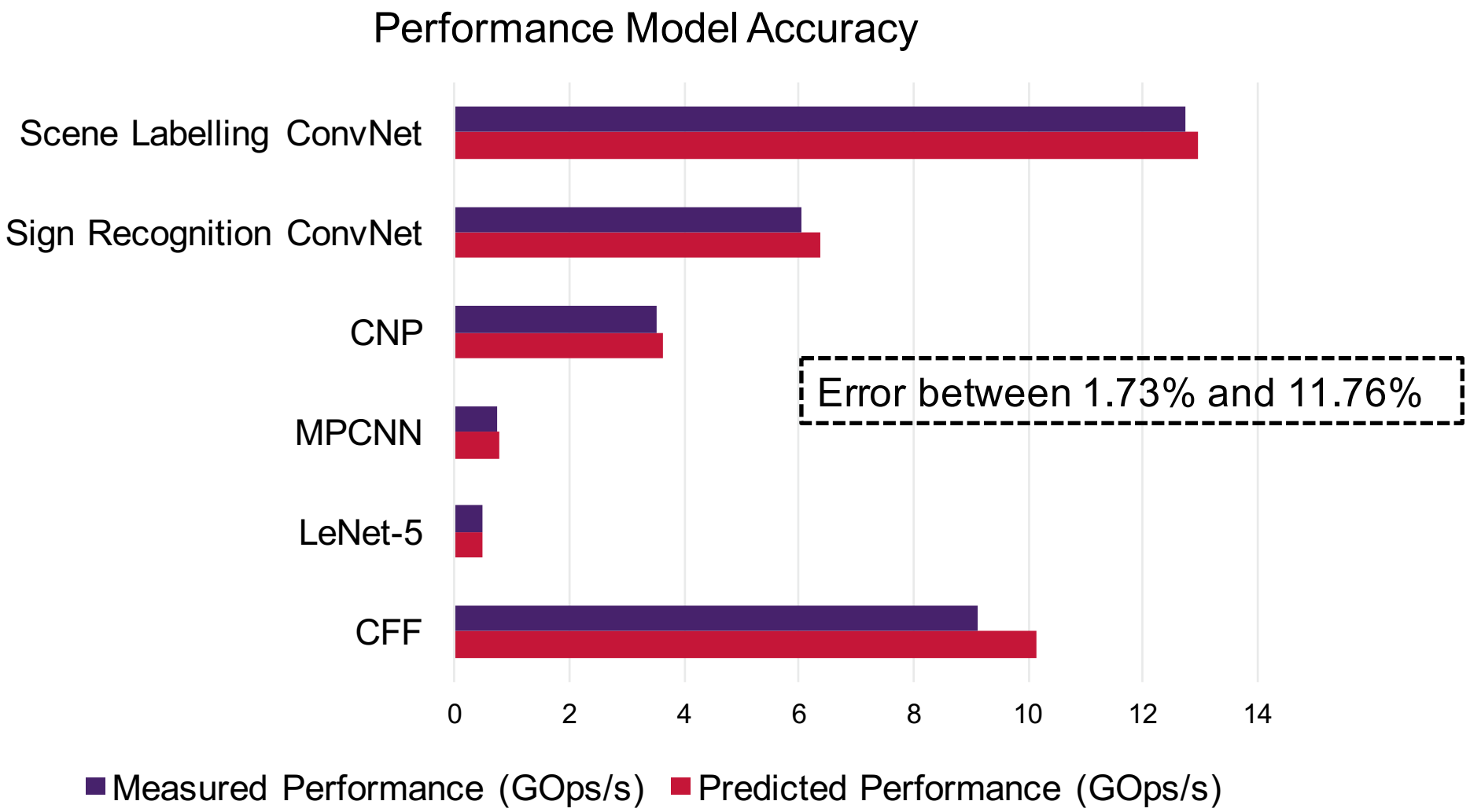
Evaluation - Experimental Setup

- *fpgaConvNet*
  - Xilinx Zynq-7000 XC7Z020 SoC with 220 DSPs at 100 MHz
  - Q8.8 fixed-point precision to match existing work

    (also supports floating-point)
  - Current toolflow supports the Vivado HLS toolchain

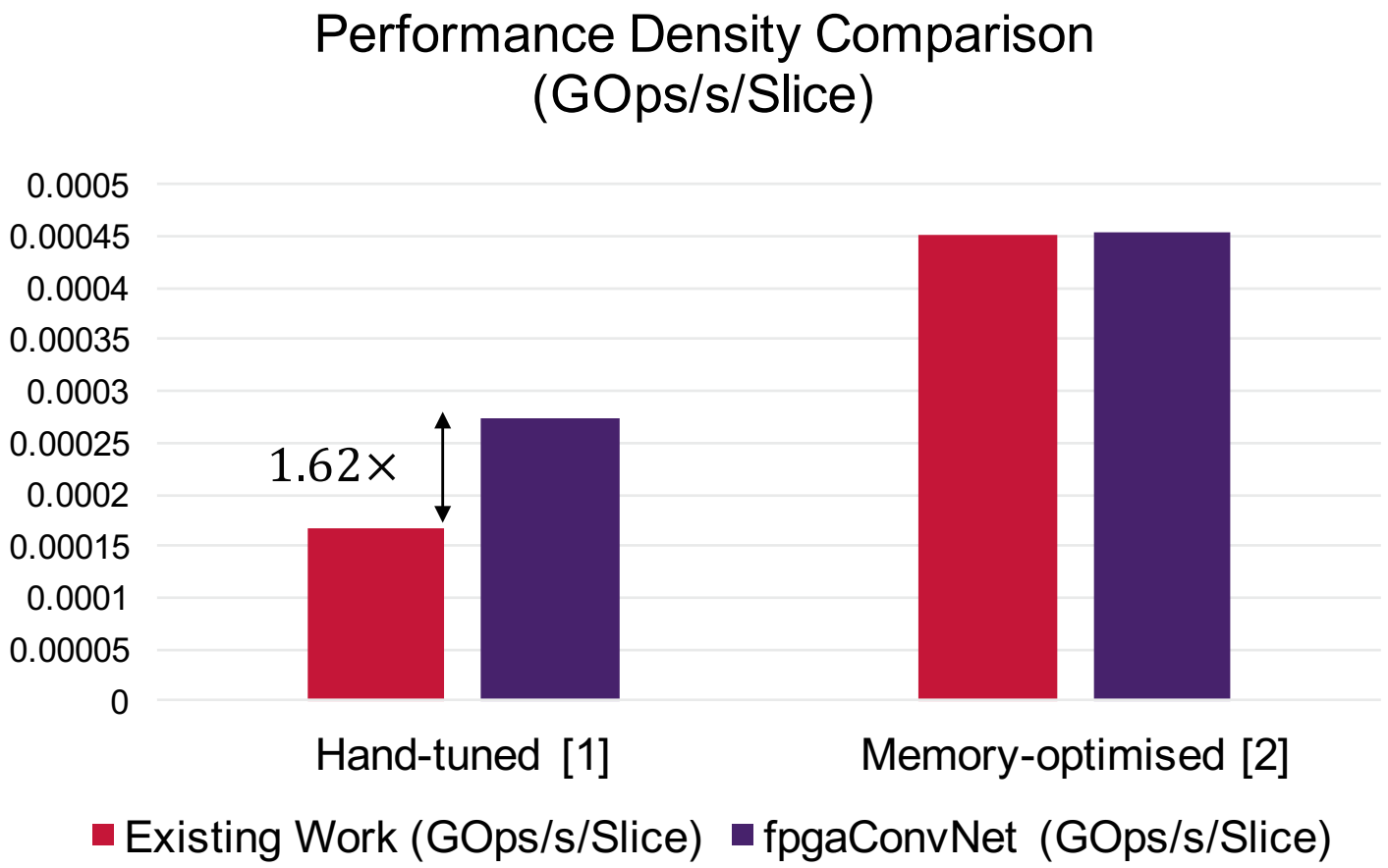
Imperial College London

Performance Model Accuracy



Imperial College London

fpgaConvNet vs. Existing FPGA Work

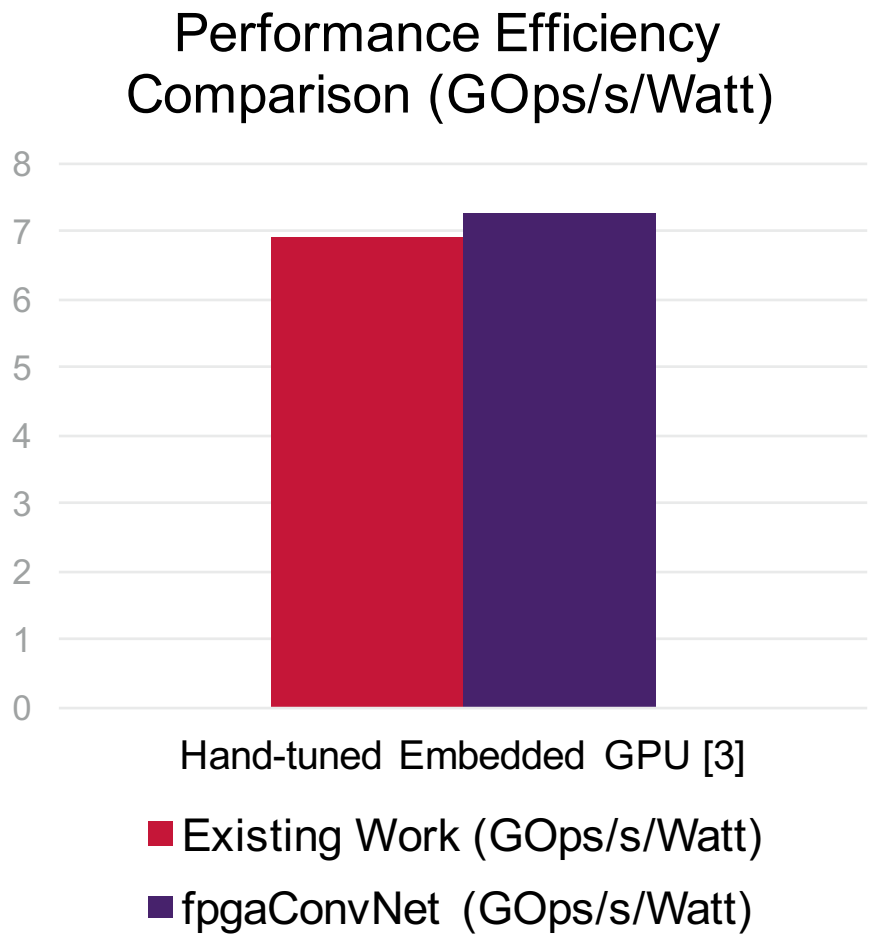


[1] C. Farabet et al., "CNP: An FPGA-Based Processor for Convolutional Networks", in FPL, IEEE, 2009.

[2] M. Peemen et al., "Memory-centric accelerator design for Convolutional Neural Networks", in ICCD, IEEE, 2013.

Imperial College London

fpgaConvNet vs. Existing Embedded GPU Work



Hand-tuned Embedded GPU

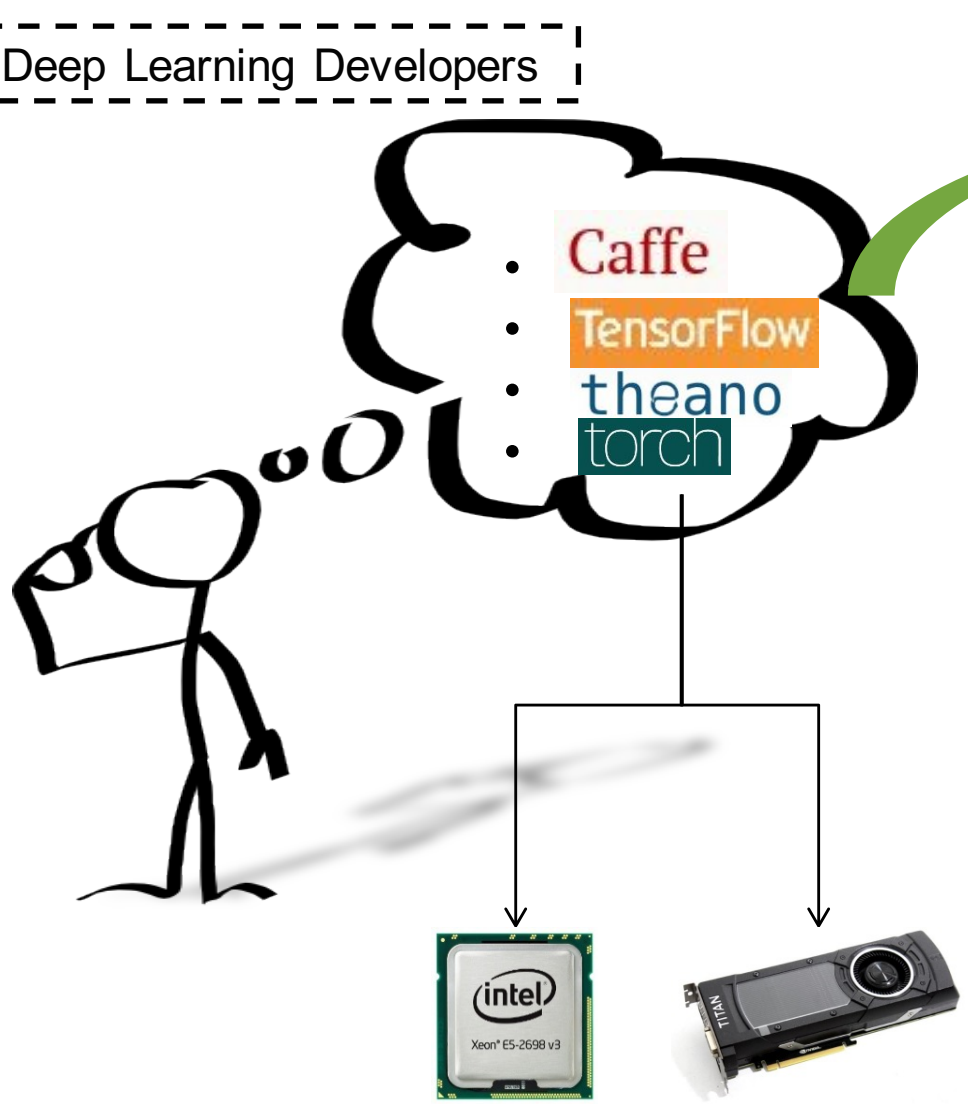
- Tegra K1 at 800 MHz
- Memory Bandwidth: 12 GB/s

fpgaConvNet

- Zynq-7000 XC7Z020 at 100 MHz
- Memory Bandwidth: 4.26 GB/s

[3] L. Cavigelli et al., "Accelerating real-time embedded scene labeling with convolutional networks", in DAC, ACM/EDAC/IEEE, 2015.

Conclusions



fpgaConvNet