IPDPS 2020 New Orleans

RAW Reconfigurable Architectures Workshop

POLITECNICO MILANO 1863 NECST laboratory

POLITECNICO MILANO 1863

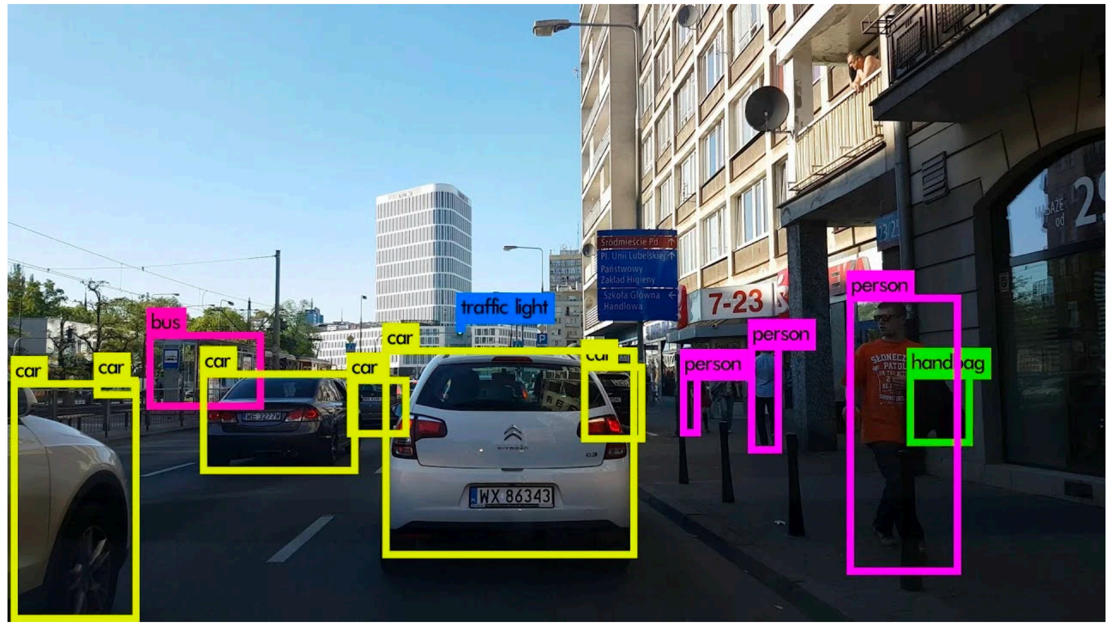
Hardware resources analysis of BNNs splitting for FARD-based multi-FPGAs Distributed Systems

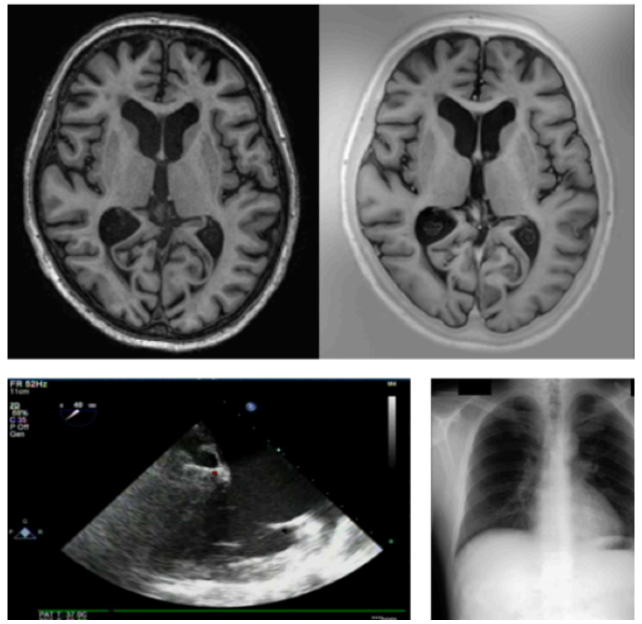
Giorgia Fiscaletti, Marco Speziali, Luca Stornaiuolo,
Marco D. Santambrogio, Donatella Sciuto

Dipartimento di Elettronica Informazione e Bioingegneria (DEIB)
{ giorgia.fiscaletti, marco.speziali }@mail.polimi.it
{ luca.stornaiuolo, marco.santambrogio, donatella.sciuto }@polimi.it

Context Definition

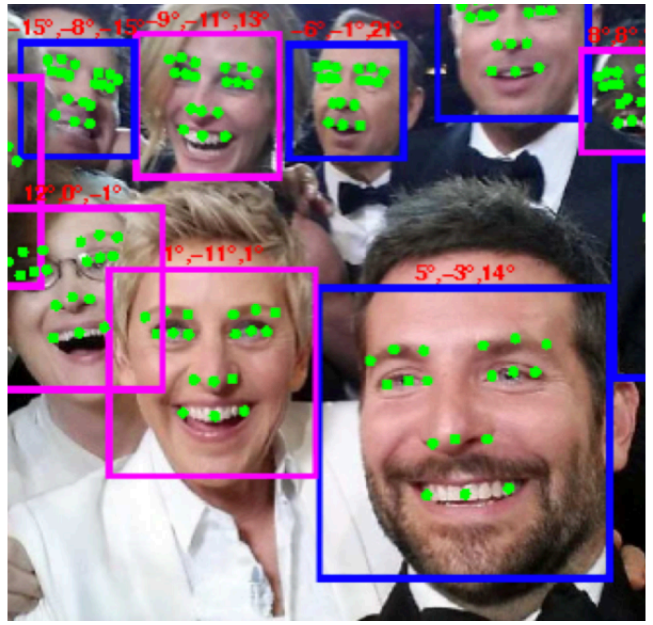
Autonomous Driving

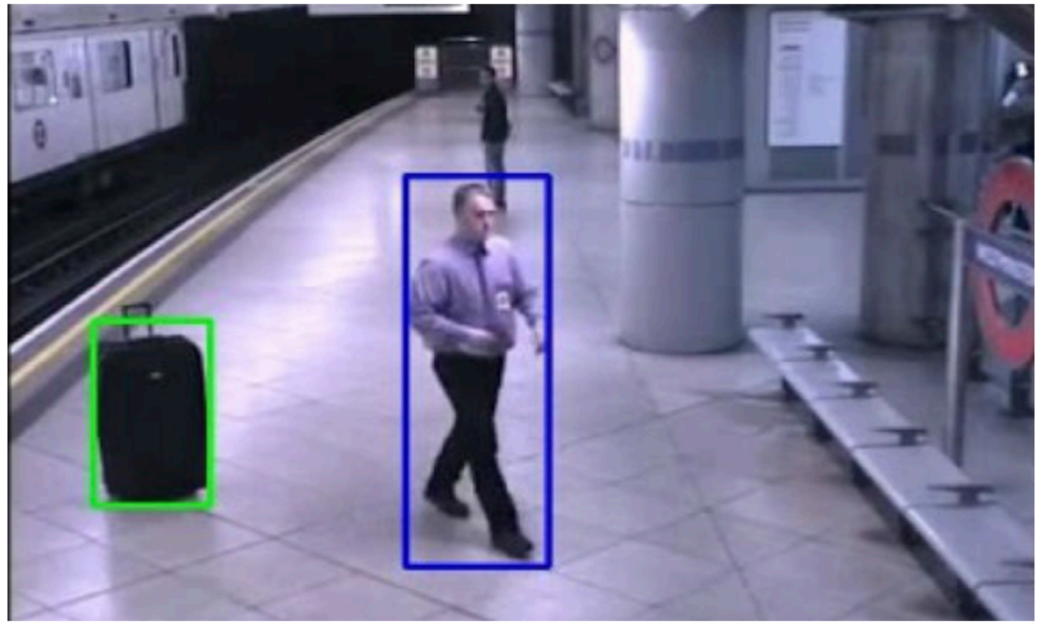




Medical Application

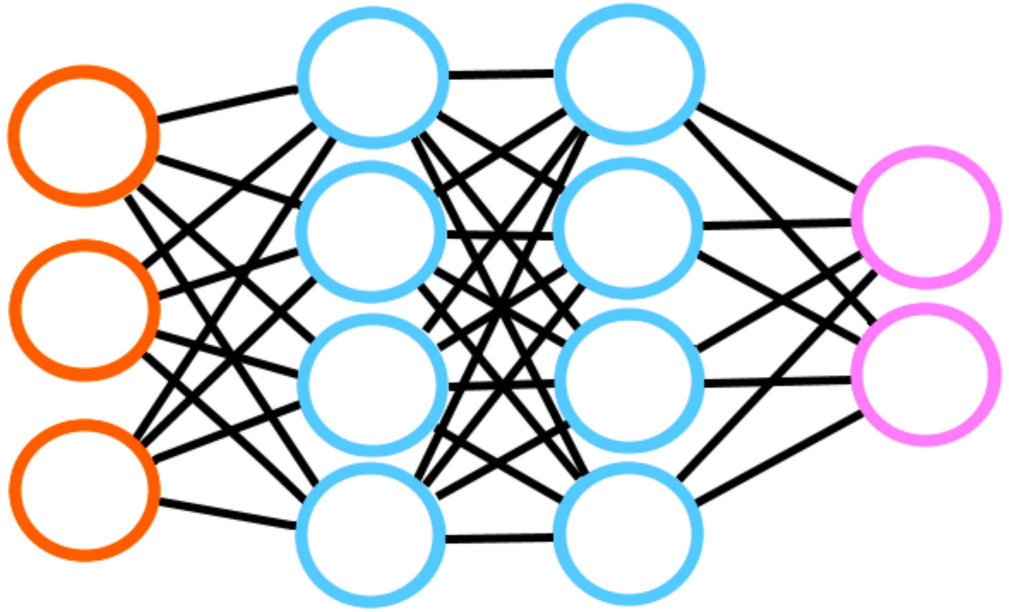
Face Detection



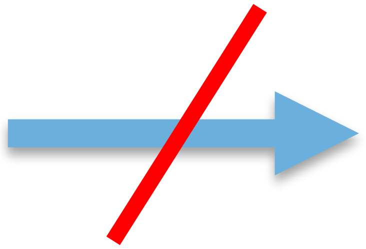


Security Systems

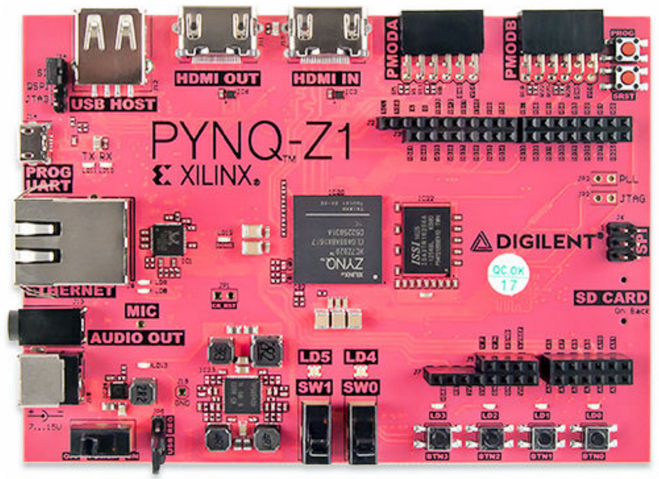
Problem Description



Binarized Neural Network

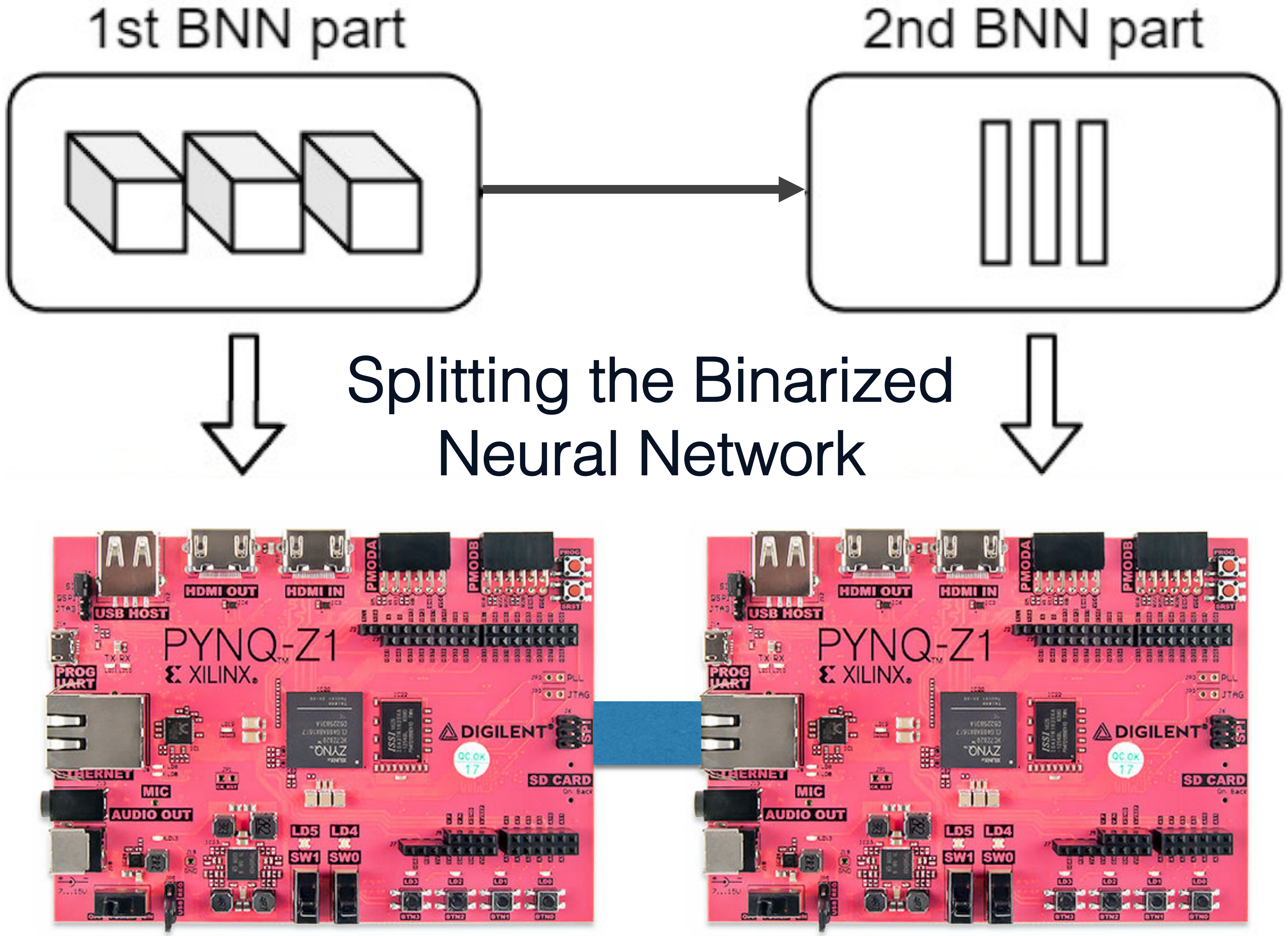


Limited resources



FPGA-based Embedded System

Proposed Solution



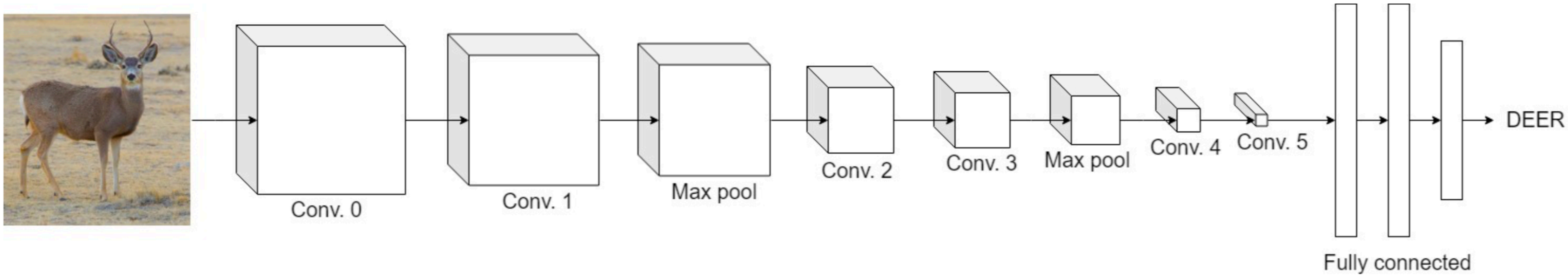
1st BNN part

2nd BNN part

Splitting the Binarized Neural Network

- Balancing the use of hardware resources
- Jobs and pipelines management for the distributed system

The BNN selected



CnvW2A2

- 6 Convolutional Layers
- 3 Fully-Connected Layers

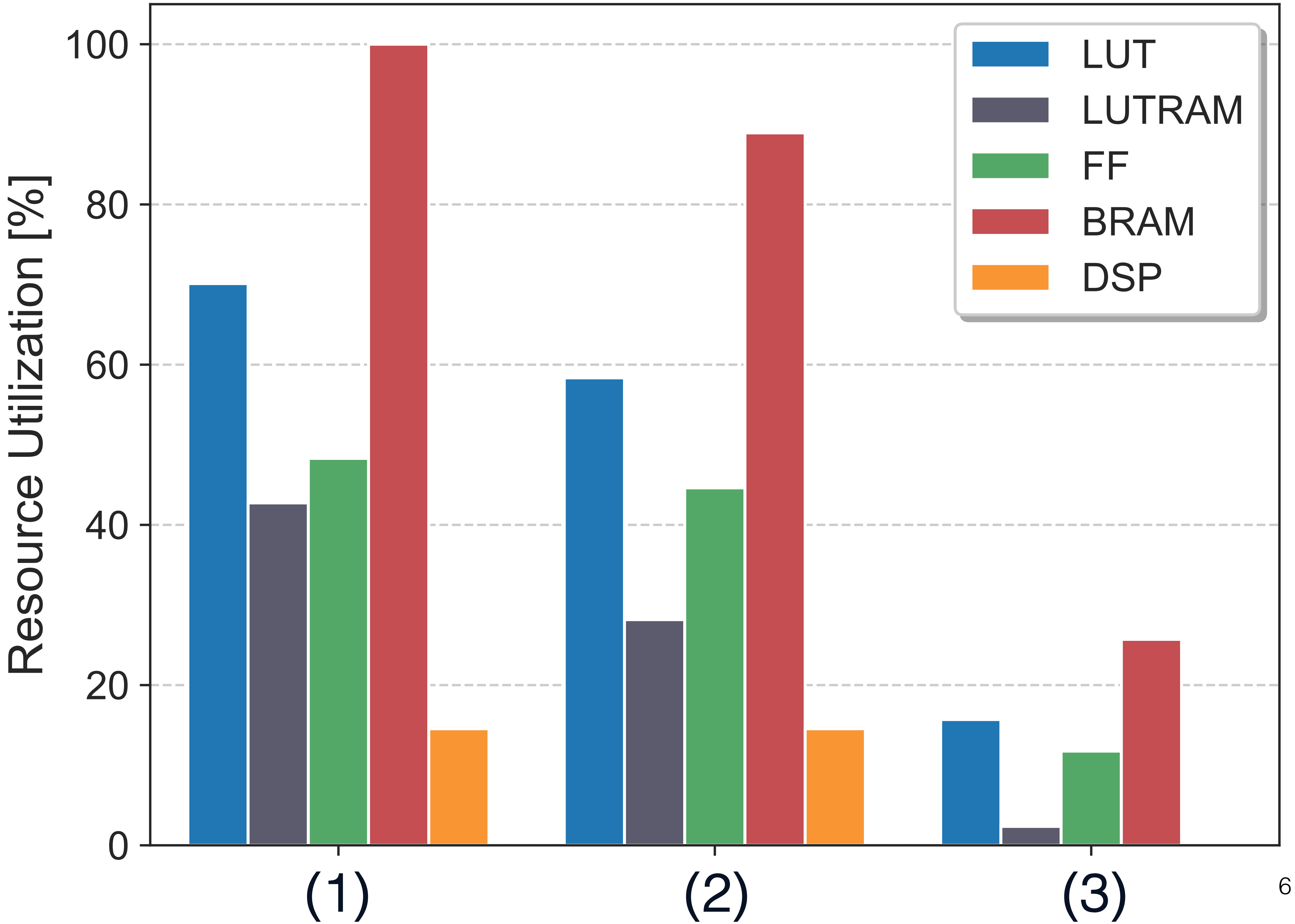
Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, "*Finn: A framework for fast, scalable binarized neural network inference*," in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, ser. FPGA '17.

Resources Comparison

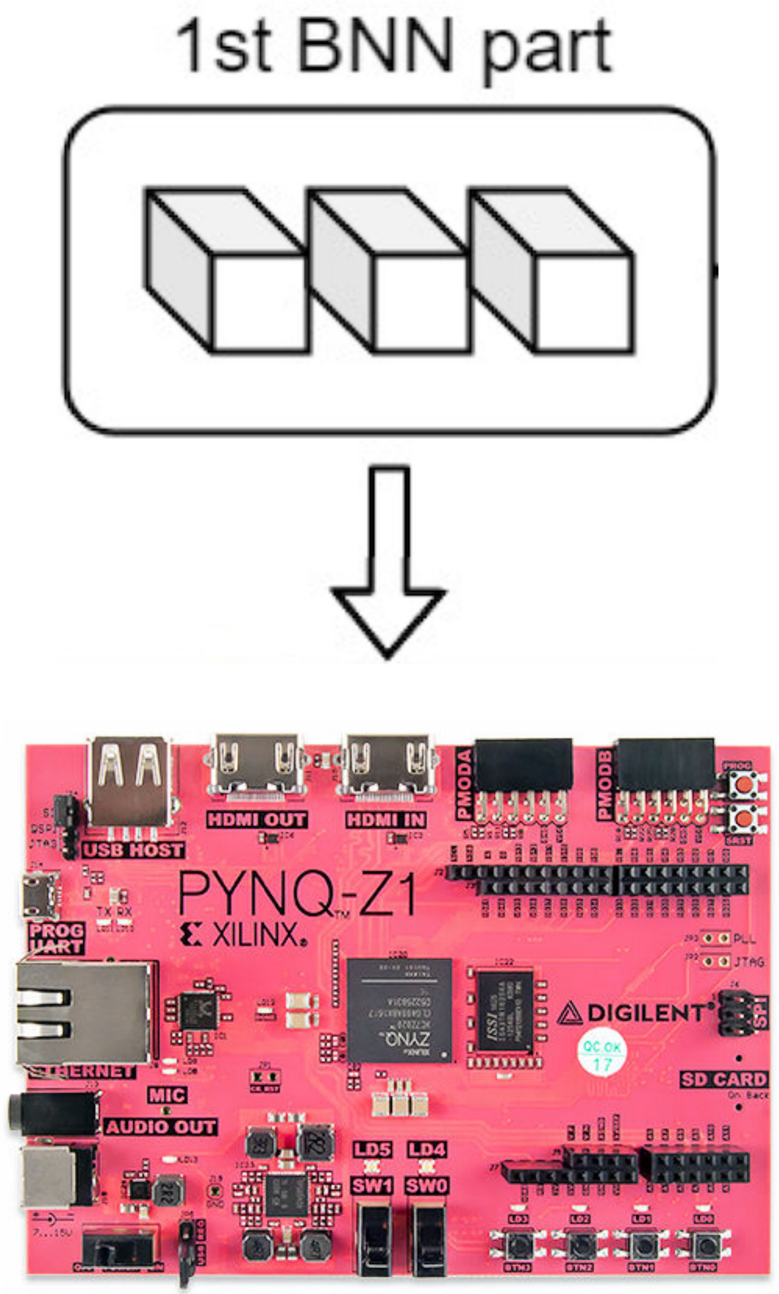
(1) CnvW2A2

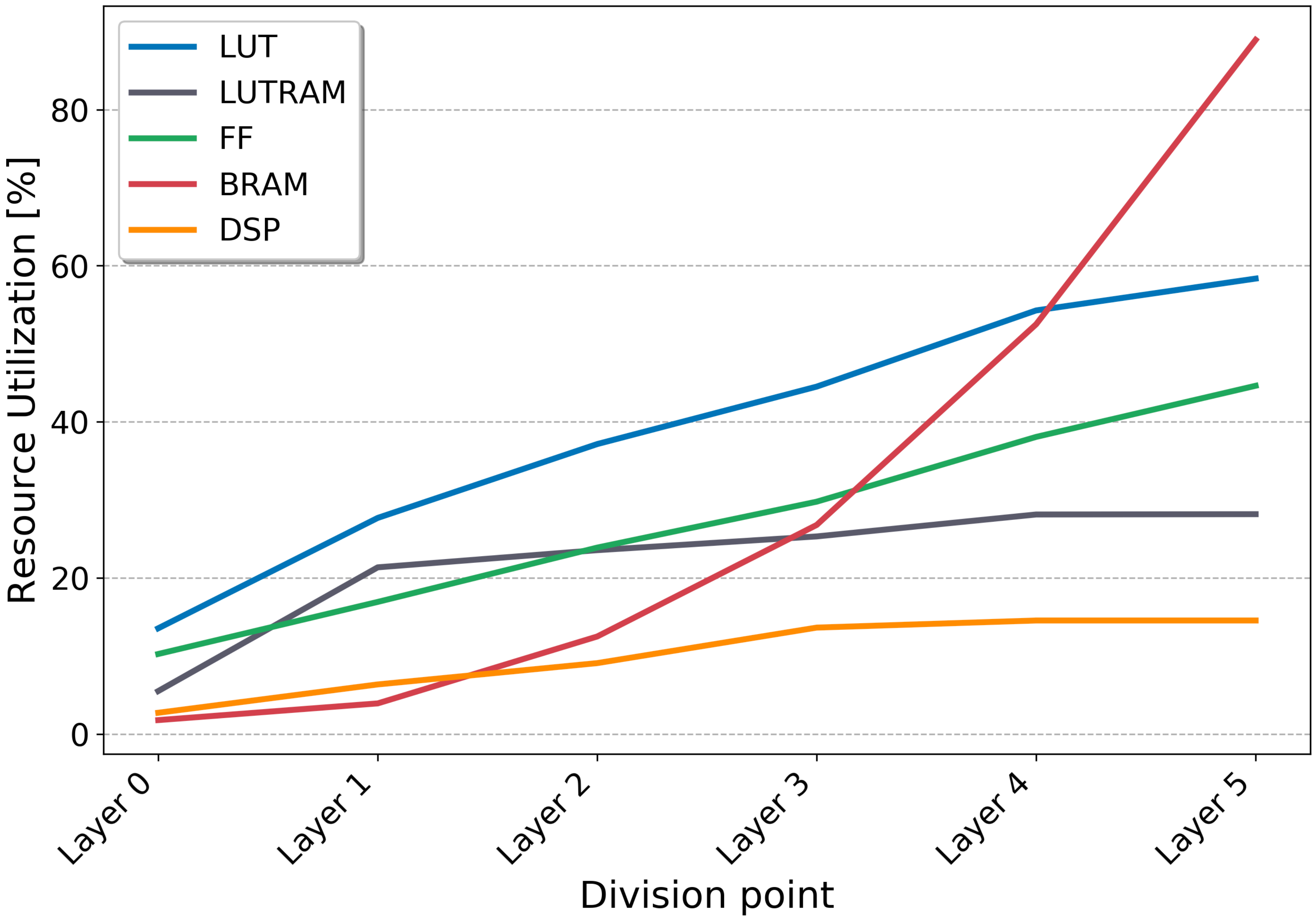
(2) Convolutional

(3) Fully-Connected



Resources Evaluation

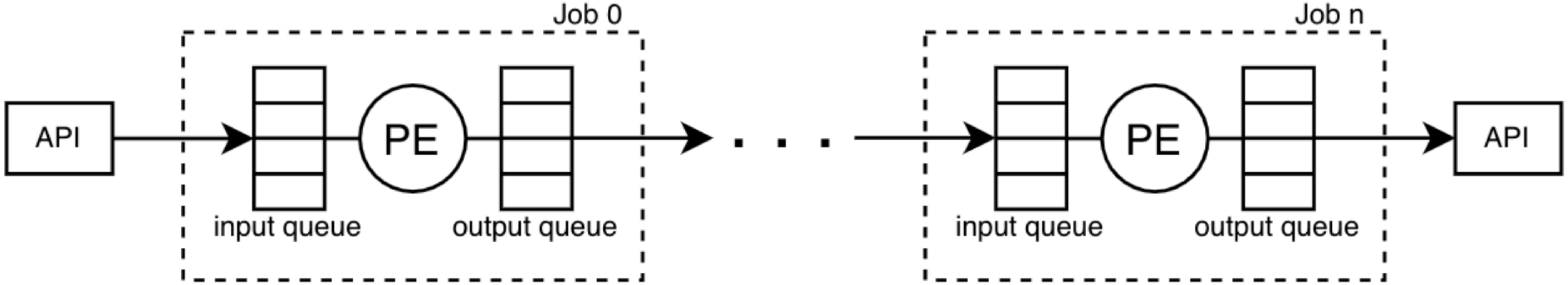




POLITECNICO MILANO 1863
NECST laboratory

POLITECNICO
MILANO 1863

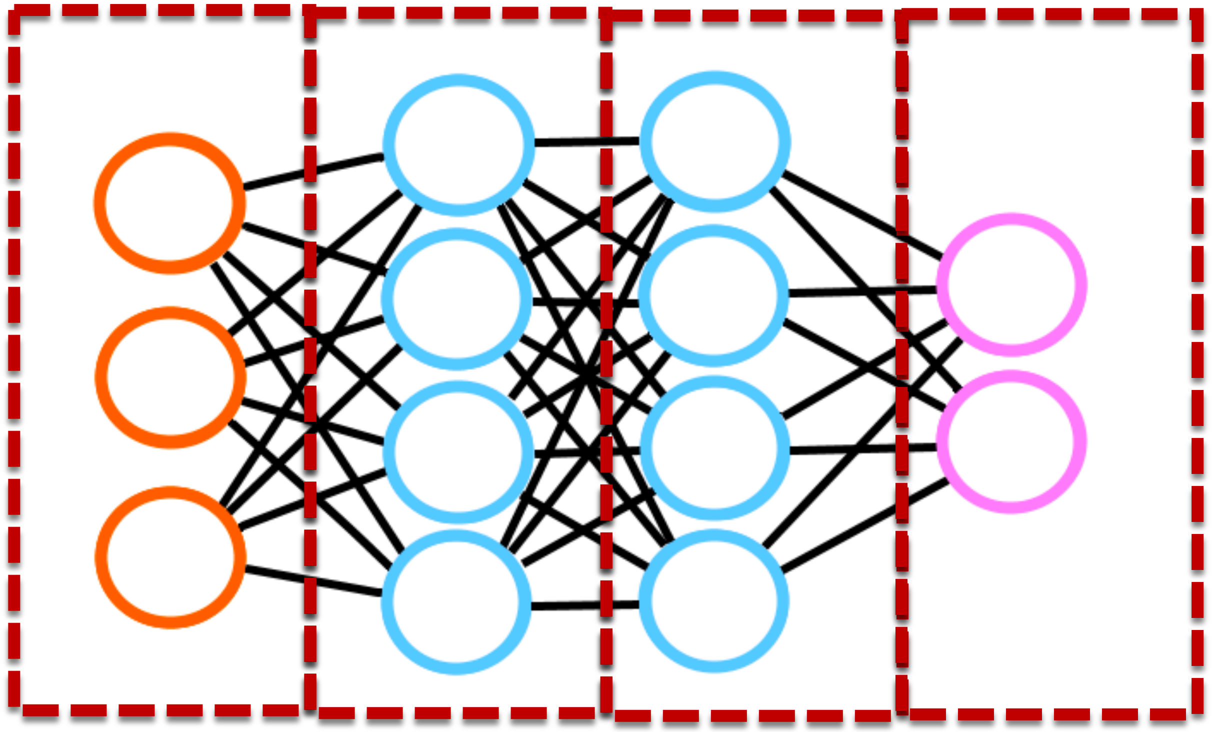
FARD improvements



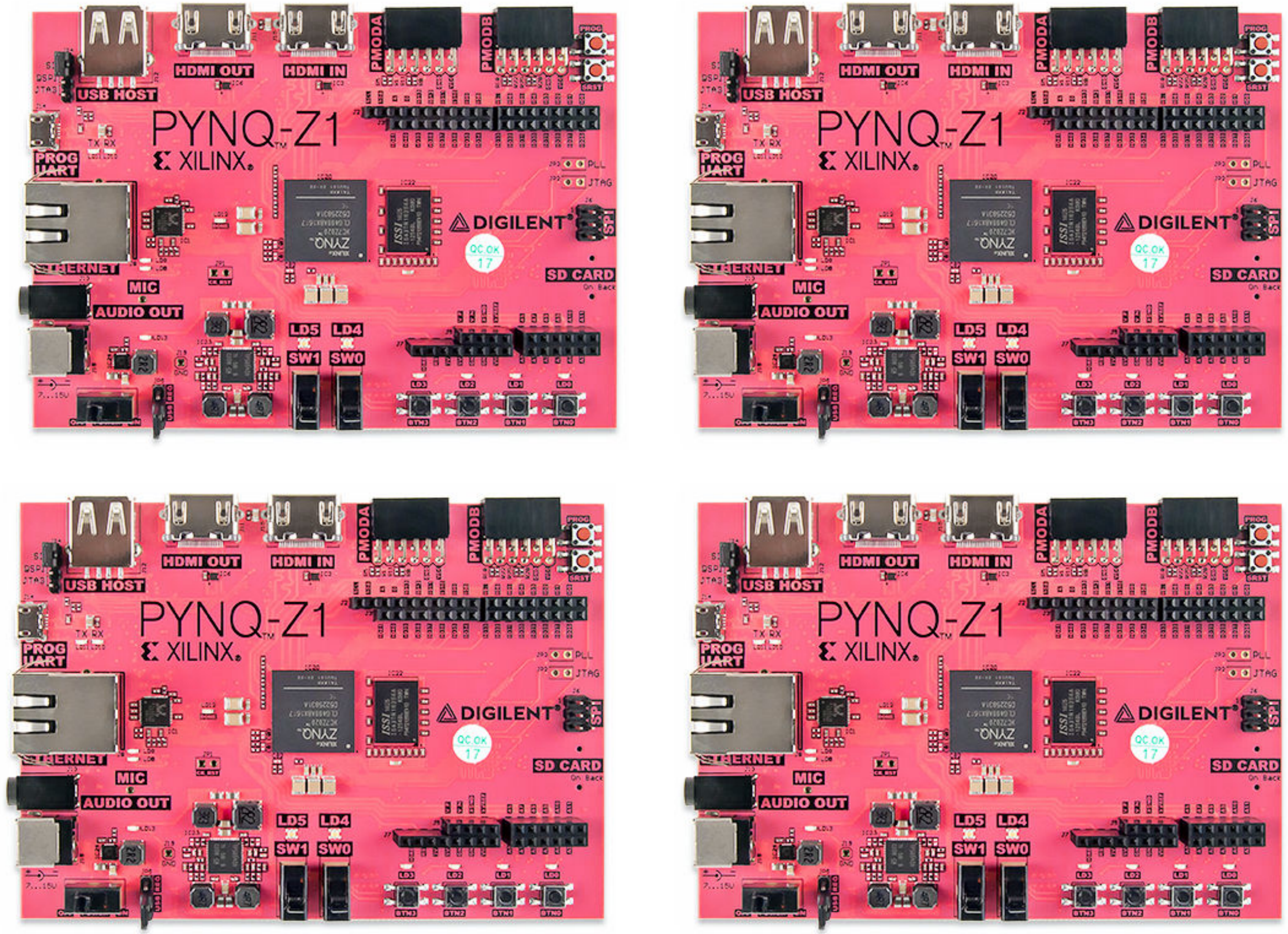
FARD extension with pipelines. A **pipeline** is a set of *n* **jobs**, each one with a specific position in the flow of data.

Barbieri, Samuele and Casasopra, Fabiola and Brondolin, Rolando and Santambrogio, Marco D, "*Fog Acceleration through Reconfigurable Devices*," in Proceedings of the 2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI) 2019.

Future Directions



Binarized Neural Network



Multiple Splitting Strategy

IPDPS 2020 New Orleans

RAW Reconfigurable Architectures Workshop

POLITECNICO MILANO 1863 NECST laboratory

POLITECNICO MILANO 1863

Hardware resources analysis of BNNs splitting for FARD-based multi-FPGAs Distributed Systems

https://necst.it/
https://www.slideshare.net/necstlab

Giorgia Fiscaletti, Marco Speziali, Luca Stornaiuolo,
Marco D. Santambrogio, Donatella Sciuto

Dipartimento di Elettronica Informazione e Bioingegneria (DEIB)
{ giorgia.fiscaletti, marco.speziali }@mail.polimi.it
{ luca.stornaiuolo, marco.santambrogio, donatella.sciuto }@polimi.it