WIP: AUTOMATIC DNN DEPLOYMENT ON HETEROGENEOUS PLATFORMS: THE GAP9 CASE STUDY

Luka Macan*, Alessio Burrello, Luca Benini, Francesco Conti

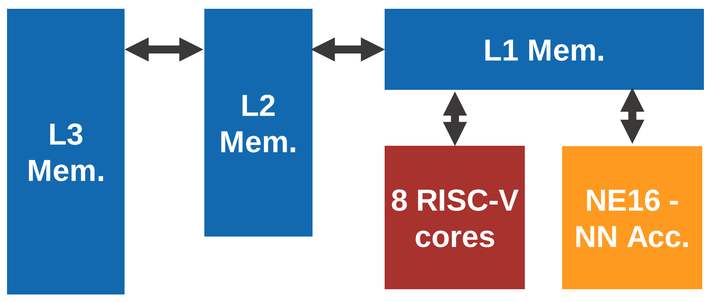
*luka.macan@unibo.it

Background

Target systems:

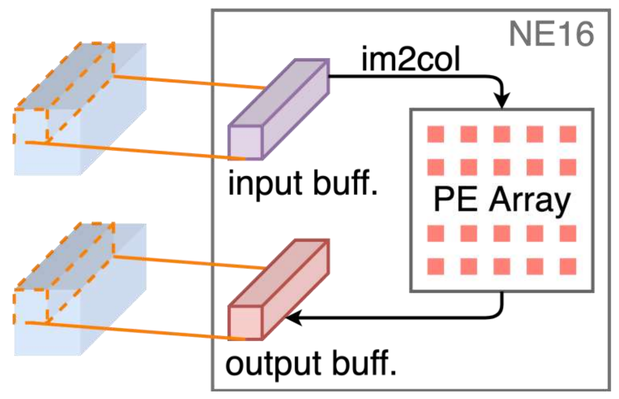
- Heterogeneous
- Embedded
- Low-power

Case study: **GAP9**



Simplified system view

NE16 overview



DORY

- *DNN deployment*
- *Parallel ultra-low-power platforms*
- *Tiling ILP*
- *Open-source*

Intra-Layer Optimizations

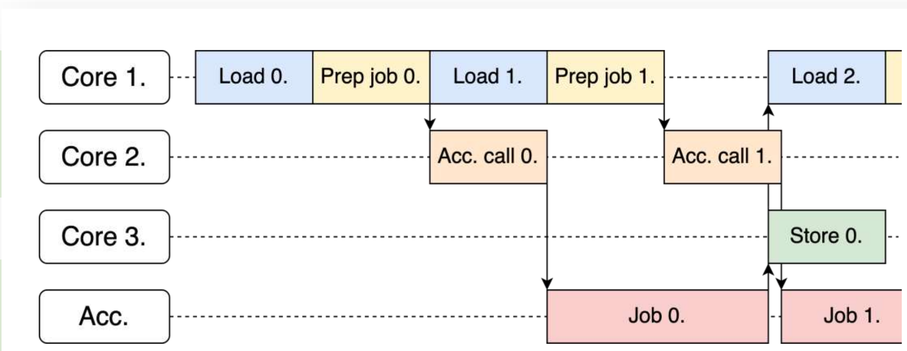
Problem

Data movement overhead

Layer size preferences

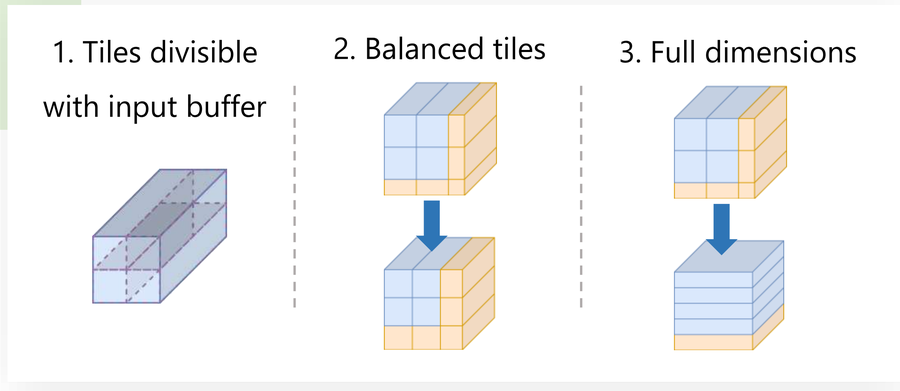
Applied solution

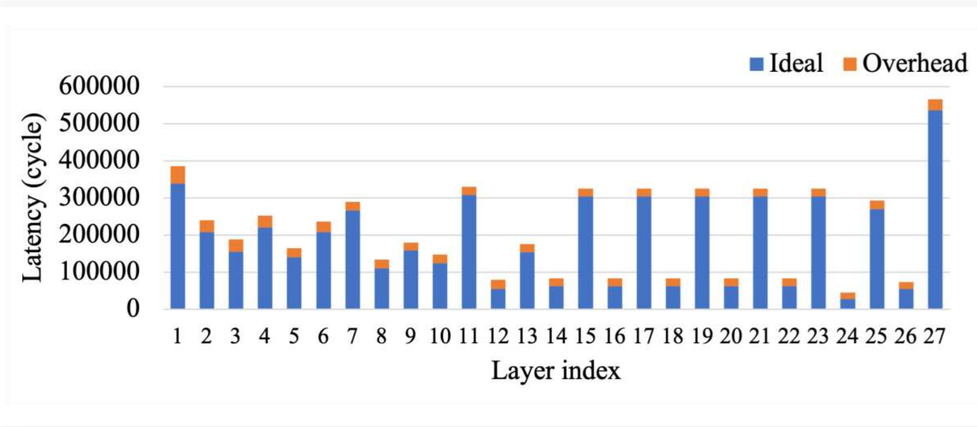
Software Pipelining



Task-Level Parallelism

Tiler Heuristics

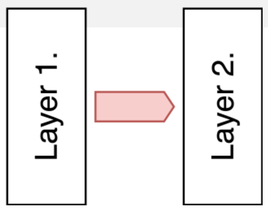




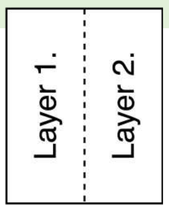
Result 1: With above mentioned techniques, achieved 12% overall latency overhead over ideal execution on MobileNetV1

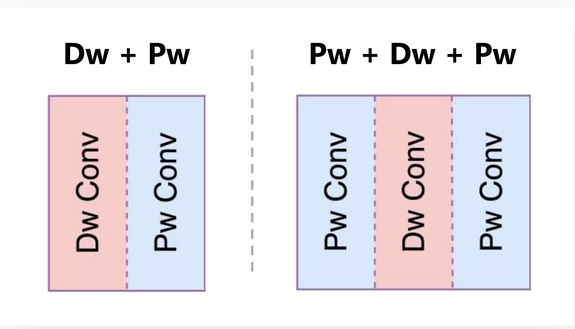
Inter-Layer

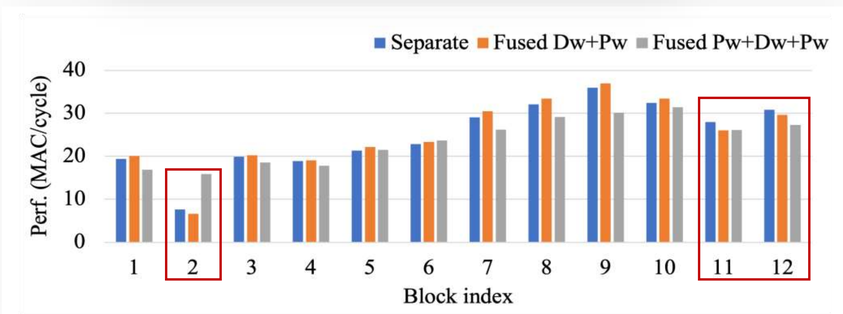
Data Movement Overhead



Layer Fusion







Result 2: Layer Fusion is not a one-shoe-fits-all solution, and the choice of layers is important

pulp-platform/dory.git

	Perf. (MAC/cycle)	Peak-Perf. Perc.
GAP8 MN-V1	8.09	N/A
GAP9 MN-V1	26.04	91%
GAP9 MN-V2	24.23	89%

Result 3: Achieved 91% and 89% accelerator utilization on MobileNet-V1 and MobileNet-V2 respectively. Speedup of 3.44x over DORY on GAP8 (no acc.).

Acknowledgements: Funded by *ROADSTER* and *NeuroSoC* (g.a. 101070634) projects

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

EMBEDDED SYSTEMS WEEK