ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

WIP: Automatic DNN Deployment on Heterogeneous Platforms: the GAP9 Case Study

Energy-Efficient Embedded Systems Laboratory (UNIBO)

Luka Macan <luka.macan@unibo.it>
Alessio Burrello <alessio.burrello@polito.it>
Luca Benini <luca.benini@unibo.it>
Francesco Conti <f.conti@unibo.it>

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

PULP Platform

Open Source Hardware, the way it should be!

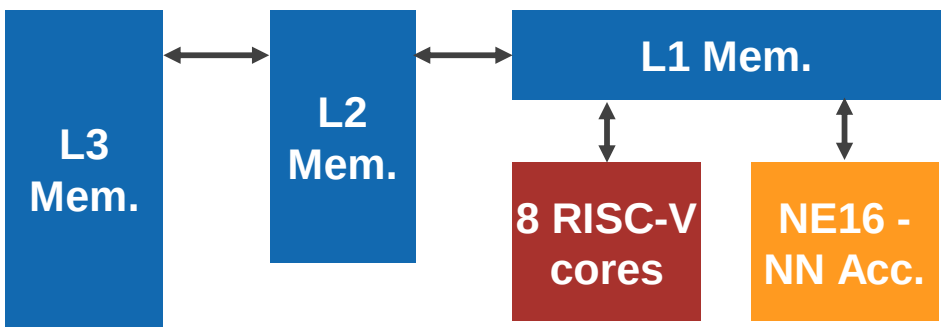
Goal: NN Deployment on the Edge

Our target: NN accelerated embedded edge devices

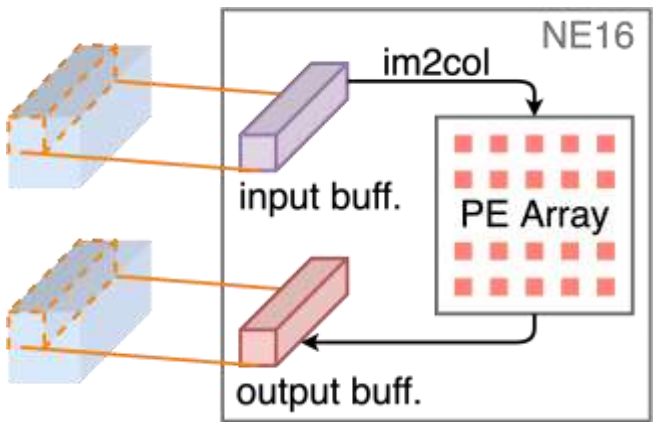
- Heterogenous
- Ultra-low-power
- Software managed data ”caches”

Case study: **GAP9 SoC**

- Compute: Cluster of 8 RISC-V cores + NN Accelerator
- 0.33 mW/GOP[1]
- 3 levels of memory + DMAs



Simplified system view of the GAP9 SoC

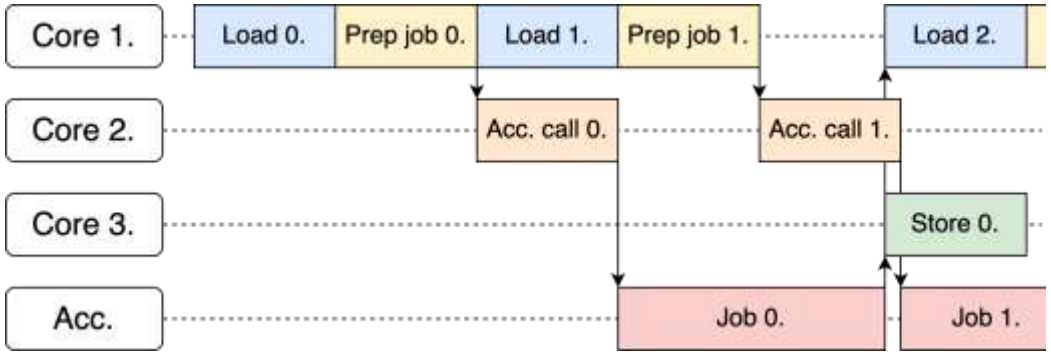


NE16 Accelerator high-level view

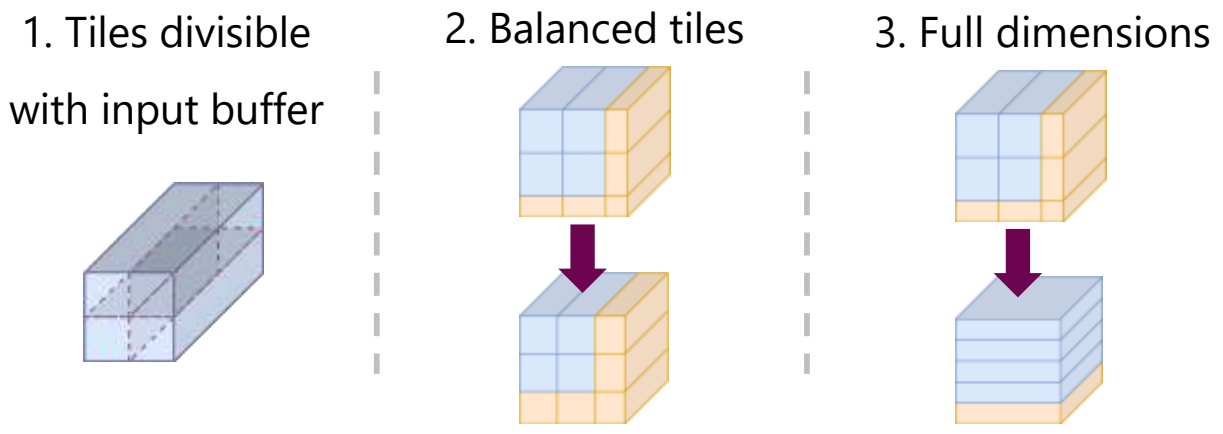
[1]https://greenwaves-technologies.com/gap9_processor/

Inter-layer optimizations

- **Software pipelining & Task parallelization**

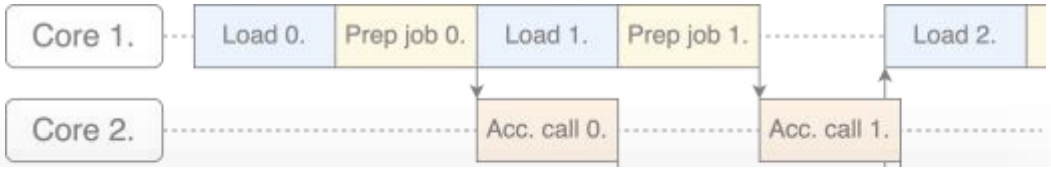


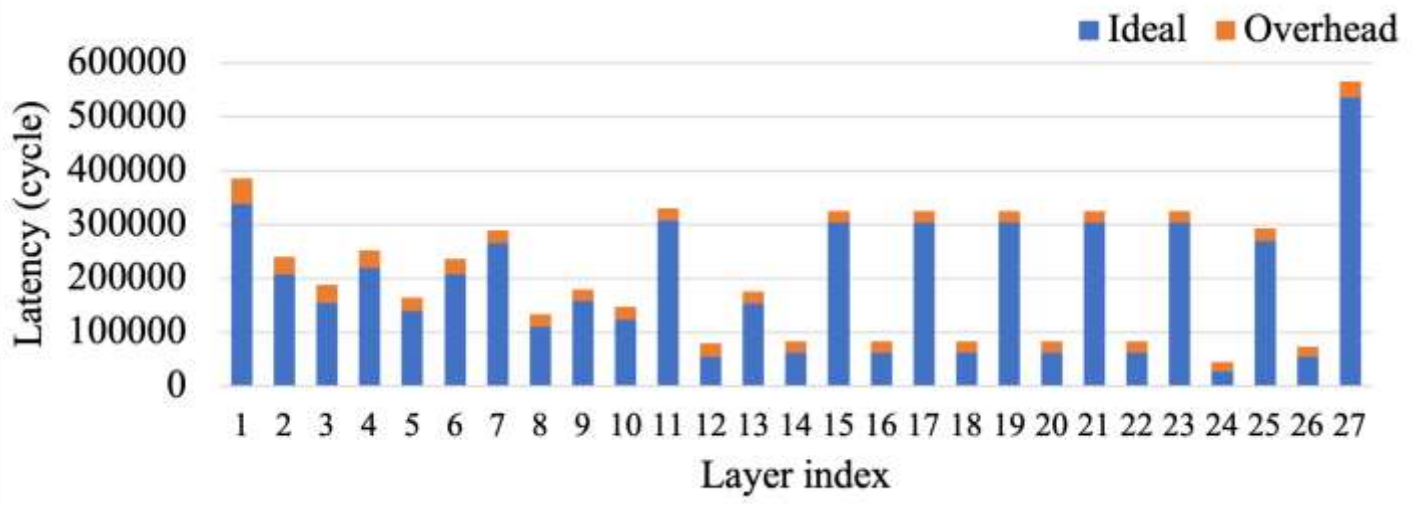
- **Tiling heuristics**



Inter-layer optimizations

- **Software pipelining & Task parallel**
- **Tiling heuris**

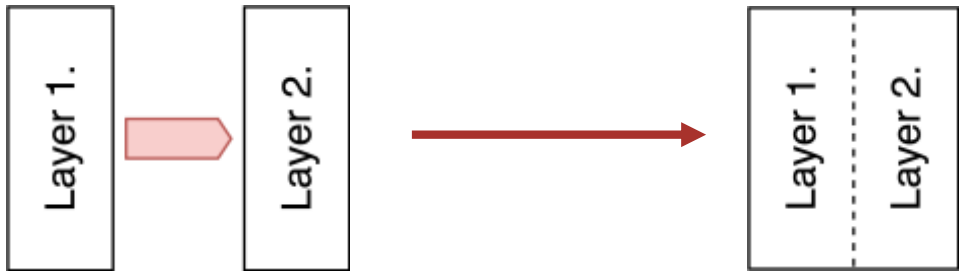




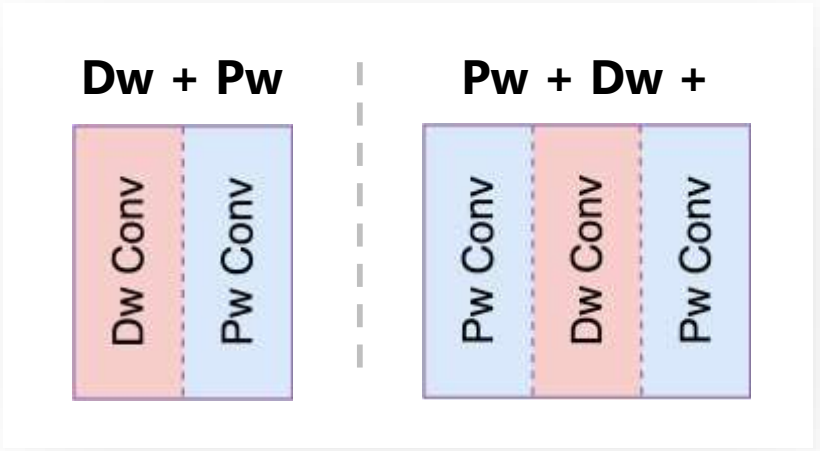
12% overall latency overhead over ideal

Intra-layer optimizations

- **Layer fusion**

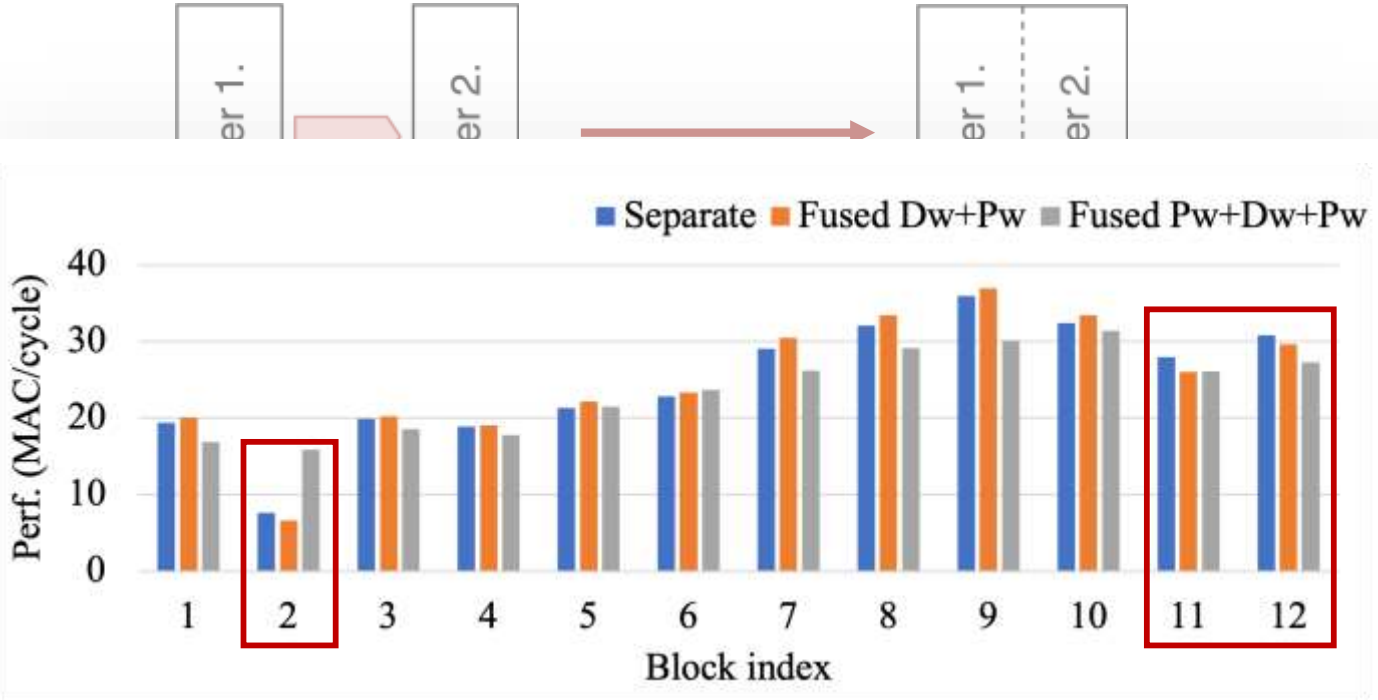


- **Choice of layers to fuse**



Intra-layer optimizations

- Layer fusion



MN-V2 unique Inverted Bottleneck blocks

- Choice of laye

Conclusion

- **Achieved high accelerator utilization (~90%) and 3.44x performance improvement over the predecessor, GAP8**
- **Accelerators need several techniques to be applied for them to come close to ideal utilization**
- **Automatic deployment tools will play a more and more critical role as the hardware gets more complex**