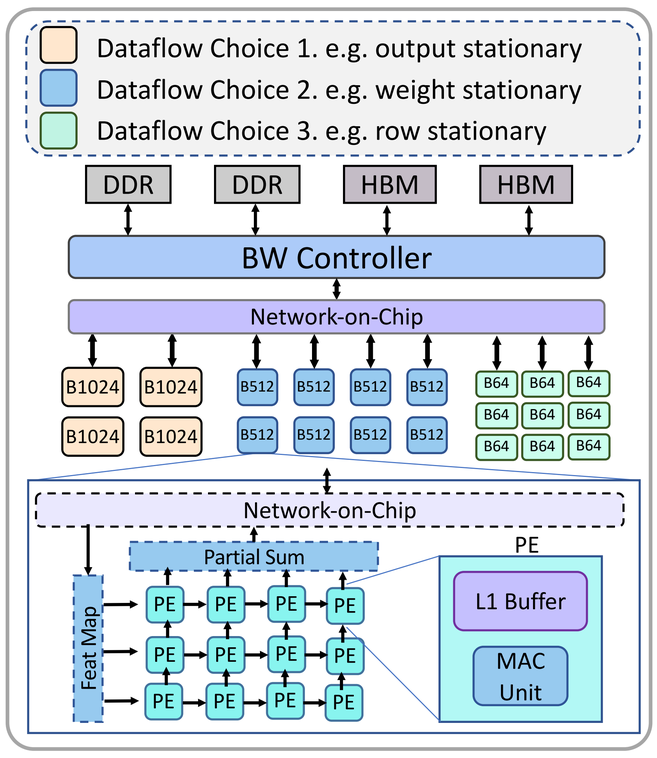
Serving Multi-DNN Workloads on FPGAs: a Coordinated Architecture, Scheduling, and Mapping Perspective

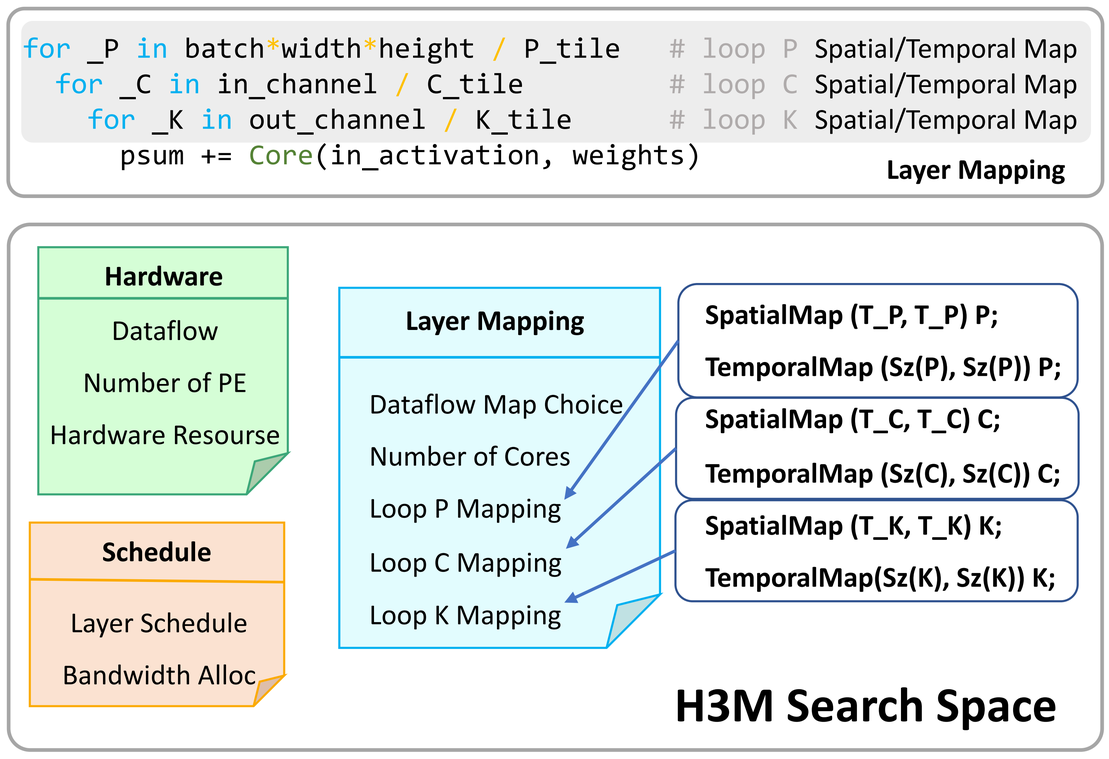
Shulin Zeng, Guohao Dai, Niansong Zhang*, Xinhao Yang, Haoyu Zhang, Zhenhua Zhu,
Huazhong Yang, Yu Wang
Tsinghua University

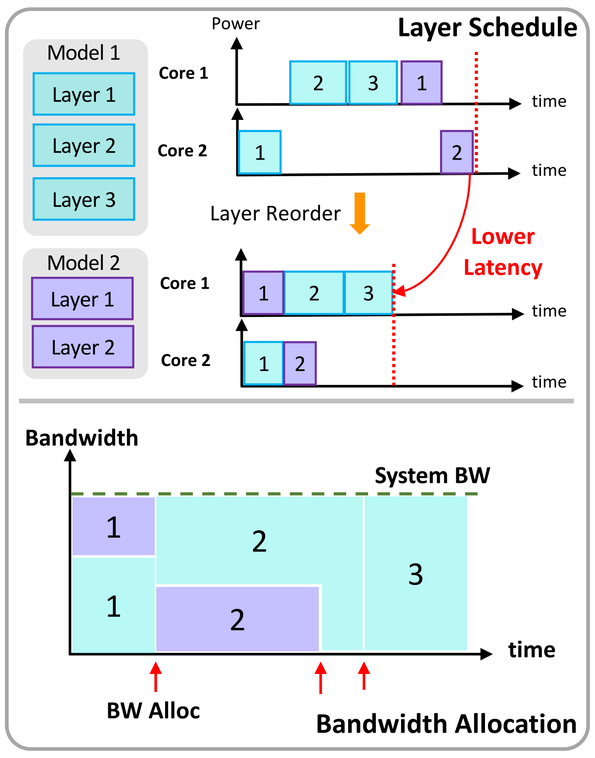
Published in IEEE Transactions on Computers

Key Takeaways

- **Heterogenous dataflow** and **homogenous multi-core** offer superior performance for DNN inference in the cloud.
- Optimizing **architecture**, **schedule**, and **compiler mapping** together finds better overall system designs.







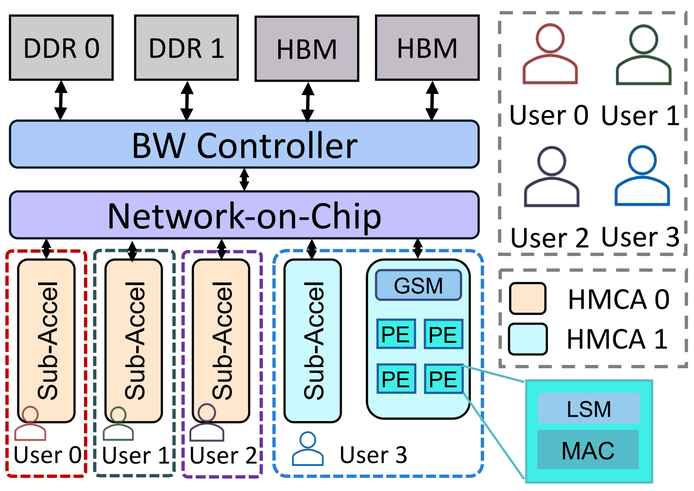
Motivation

DNN INFerence-as-a-Service (INFaaS)

- Serve mixed DNN workloads on virtualized multi-tenant cloud FPGA accelerators
- Heterogeneous **dataflow** for different workloads
- Homogeneous multi-core for **spatial multi-tenancy**

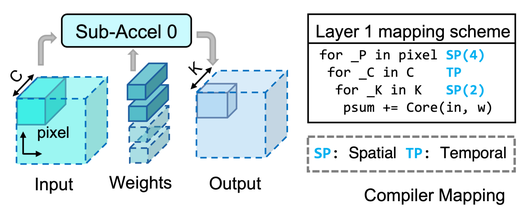
Trade-offs

- Small cores: flexible and secure
- Large core: resource and power efficient



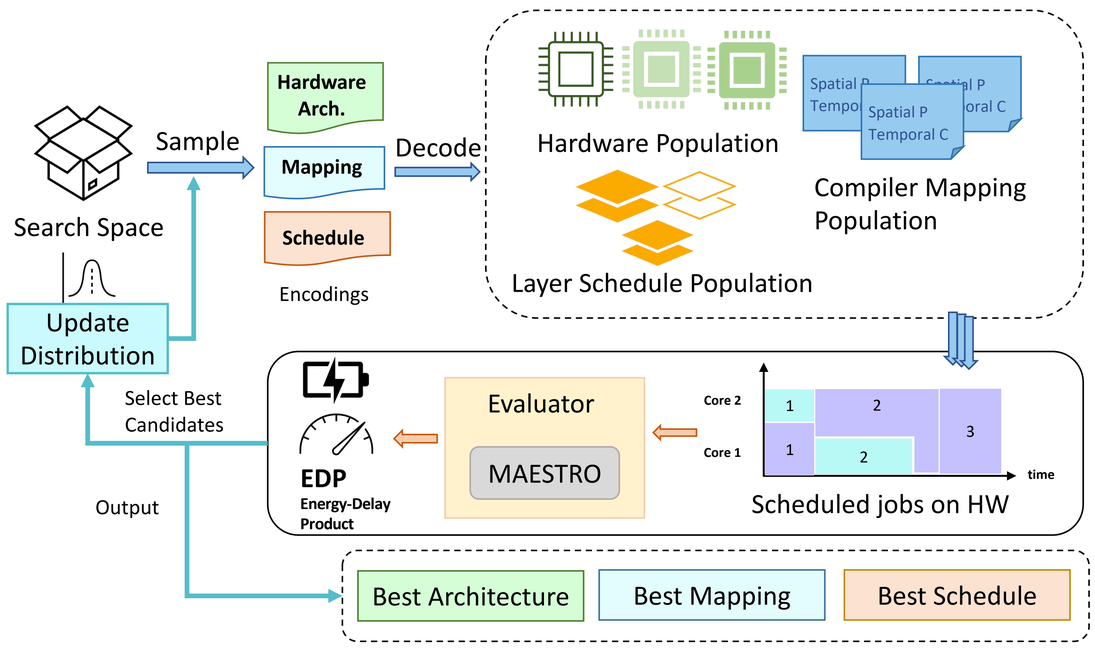
Search Space

- How to choose **dataflow**?
- What's the right **granularity**?
- How to allocate **resources**?
- How to **map** NN to multi-core architectures?



Method

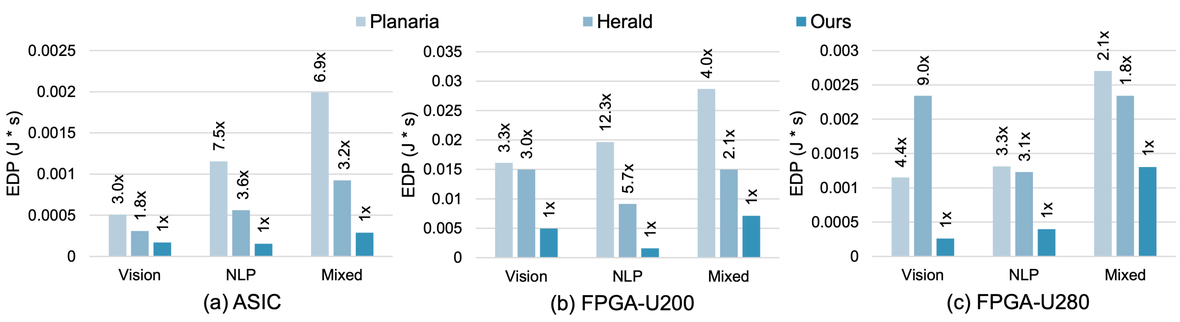
- Formulate a joint optimization problem
- **Energy-Delay-Product** (EDP) as the goal for given workloads
- Co-optimize to find best solution with CMA-ES



Results

- 3.0–7.5X better EDP than homogeneous multi-core accelerators
- 1.8–3.6X better EDP than heterogenous dataflow accelerator with fixed core gradularity

Workload	DNN Models
Vision	ResNet50 [29], MobileNetV2 [39], GoogleNet [40], VGG16 [17]
NLP	GNMT [16], ncf [41], Transformer (12 layers) [31]
Mixed	ResNet50 [29], MobileNetV2 [39], GNMT [16], Transformer (12 layers) [31]



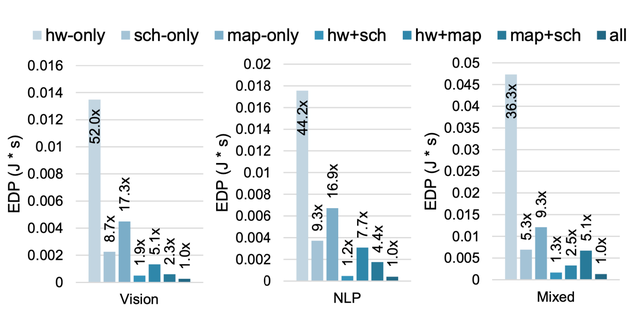


TABLE 6: Optimal Configuration and resource utilization of H3M over the three workloads on the Xilinx U280 FPGA.

Modules		Optimal Configs	LUTs	FFs	BRAMs	URAMs	DSPs
HyperConnect		-	15.6K	7.5K	0	0	0
Load/Save Module			24.8K	61.4K	0	0	0
NoC			50.1K	65.3K	0	0	0
Cores	Vision	ws:16*B1152 + rs:37*B800	1091.0K	2042.2K	1700	567	6110
	NLP	os:33*B512 + ws:10*B3136	1056.6K	1969.2K	1443	481	6118
	Mixed	os:12*B3136 + ws:8*B4096	760.9K	1831.9K	1479	493	8840
Total Utilization		(Vision/NLP/Mixed)	91%/88%/65%	84%/81%/75%	84%/72%/73%	59%/50%/51%	68%/68%/98%

Insights

- Vision: many small cores (B512, B800)
- NLP: large core + small cores (B3136, B800)
- Mixed: large cores (B4096, B3136)

Paper link

*This work is done during internship at Tsinghua University