SUSHI: SubGraph Stationary HW-SW Co-design for ML Inference

Payman Behnam*[1], Jianming Tong*[2], Alind Khare [1], Tushar Krishna [2], Alexey Tumanov [1]

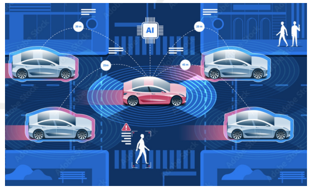
[1]SAIL Lab, School of Computer Science, [2]Synergy Lab, School of ECE

* Equal Contribution

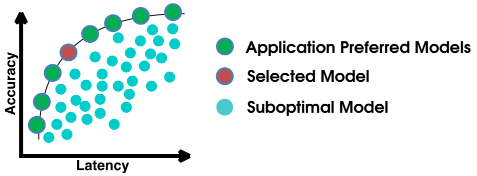
Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

Motivation

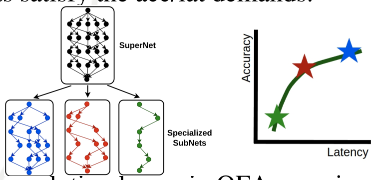
- ML applications come with dynamic accuracy/latency demand and are sensitive to SLO attainment.



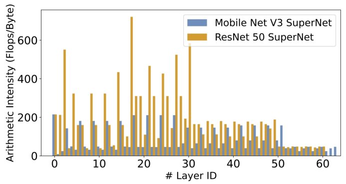
- Current works improve latency & accuracy trade-off for a single point of a specific DNN model, which is suboptimal under dynamic deployment conditions.
- Arbitrary single model from the latency/accuracy trade-off space may be suboptimal for wide-range acc-lat demand.



- Weight-shared DNNs offer a trade-off between accuracy and latency (i.e., $\mathrm{OFA}_{\mathrm{ICLR'20}}$) -> Dynamically changing of SubNets satisfy the acc/lat demands.

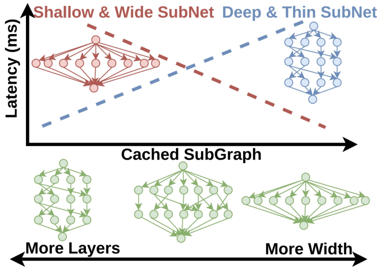


- Most convolution layers in OFAs running on an edge accelerator are memory-bound (i.e., less reuse).

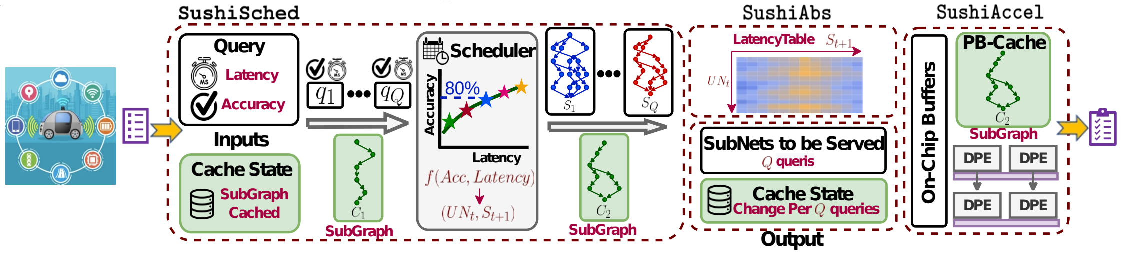


Challenges

- **Caching:** Due to the resource-restricted nature of many deployment devices, finding the best size of cache to keep a SubNet (SubGraph) is challenging.
- **Hardware:** hardware should support rapid switching among different SubNets and reuse as many weights as possible to reduce off-chip DRAM accesses.
- **Scheduler:** The latency of served SubNets depends on the SubGraph cached in the on-chip buffer.
- **Abstraction:** Scheduler decisions should be generalizable to any hardware that supports WS-DNN inference.

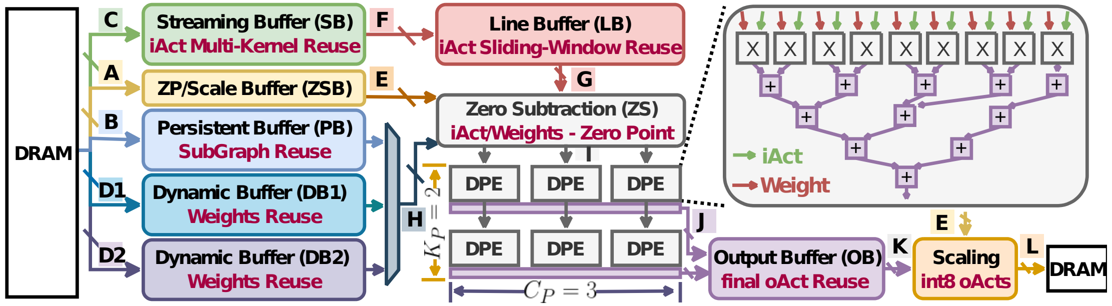


Proposed Solution: SUSHI



- SUSHI has three major novel components: scheduler, abstraction, and accelerator.

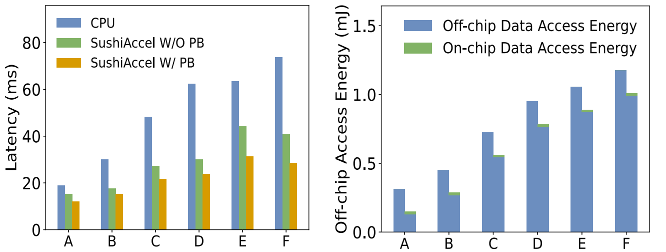
Architecture of SUSHI Accelerator



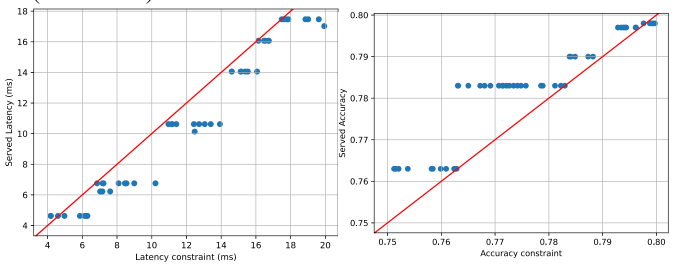
- SUSHI utilizes ***SubGraph Reuse*** in addition to input, weight, output reuse.

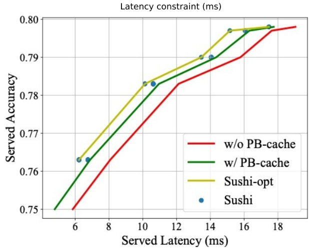
Evaluation

- Results from Alevo U50 board shows SUSHI achieves [1.16X, 1.43X] speedup than Sushi W/O cache. SUSHI saves energy up to 52.6%.



- SUSHI can serve strictly better accuracy & lesser latency (ResNet50).





- SUSHI increases the accuracy up to 1% given the same latency.
- SUSHI improves the latency by 41% given the same accuracy.

Georgia Tech