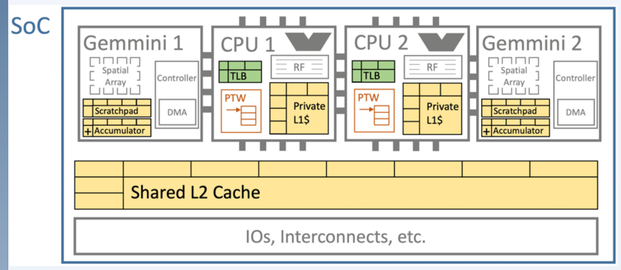
Enabling Systematic Deep-Learning Architecture Evaluation via Full-Stack Integration

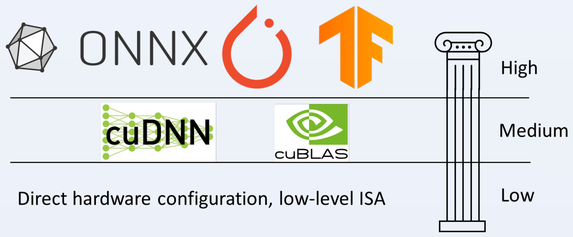
Yakun Sophia Shao (ysshao@berkeley.edu)
University of California, Berkeley

Motivation

- DNN accelerators are often developed in isolation, without considering the cross-stack, system-level effects in real workloads.
- DNN accelerators must cope with
  - SoC resource contention
  - Data movement across cores/accelerators
  - OS overheads
  - Programming stack inefficiencies
- At the SoC level:
  - Memory hierarchy: resource contention, cache coherence
  - Virtual Address Translation: page faults, TLB latency
  - Host CPUs: unaccelerated kernels



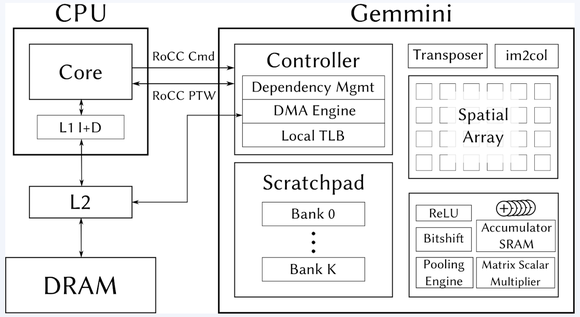
- Across the programming stack:



- Enable full-system evaluation and design-space exploration

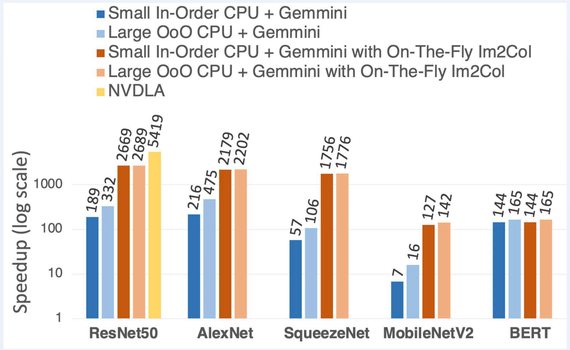
The Gemmini Infrastructure

- Flexible hardware template
  - Spatial array: dataflow, dimensions, pipelining
  - Non-GEMM functionality: transpose, im2col, ReLU…
  - Scratchpad: capacity, banks, single- or dual-port
  - Virtual address translation
  - Host CPU
  - Memory Hierarchy



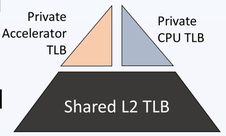
Measured Performance

- Using cloud-based FPGA
  - ResNet50: 40.3 FPS
  - AlexNet: 79.3 FPS
  - MobileNet: 18.7 FPS
  - BERT: 167x speedup

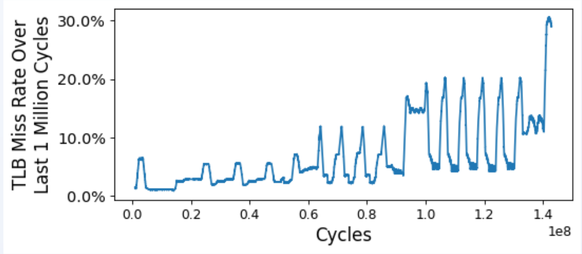


Case Study: Virtual Memory

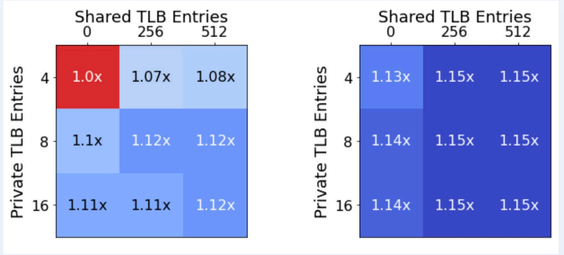
- Gemmini enables researchers to investigate virtual memory in accelerators.
- We can configure Gemmini to include a two-level TLB hierarchy, with one private TLB for the accelerator and one larger shared L2 TLB.



- Accelerator's TLB miss rate can be orders-of-magnitude higher than the one in CPUs for non-DNN benchmarks.

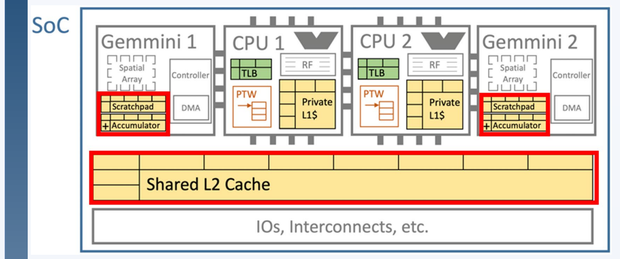


- Small, private TLB is much more impactful.
- Low-cost optimization:
  - Single-entry L0 TLB filters out consecutive TLB requests to the same page



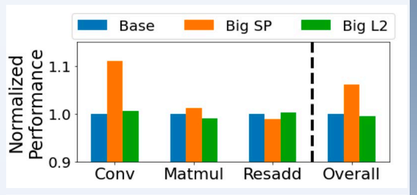
Case Study: Memory Partition

- Gemmini enables application-system co-design for real-world deployment
- Memory partition strategies in SoCs

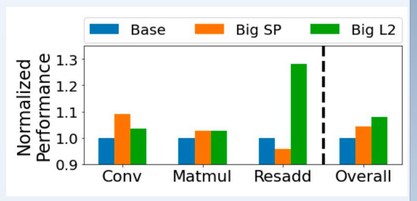


Config Name	Scratchpad (per core)	Accumulator (per core)	L2 Cache
Base	256 KB	256 KB	1 MB
BigSP	512 KB	512 KB	1 MB
BigL2	256 KB	256 KB	2 MB

- Single core:
  - Private spad more helpful
  - Better for convolutions



- Dual core:
  - Shared L2 more helpful
  - Better for residual additions



To appear at DAC 2021

- **Best Paper Candidate**
- https://github.com/ucb-bar/gemmini