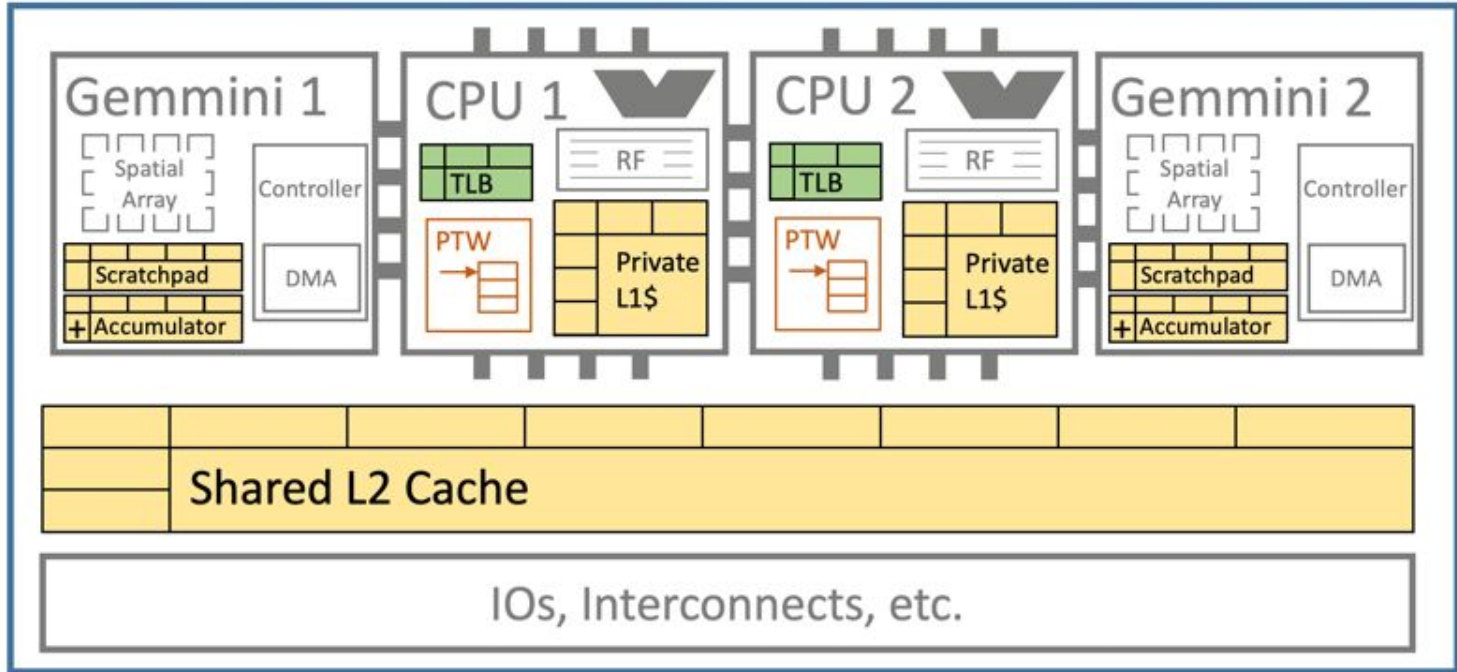# Enabling Systematic Deep-Learning Architecture Evaluation via Full-Stack Integration

Sophia Shao
ysshao@berkeley.edu
Electrical Engineering and Computer Sciences
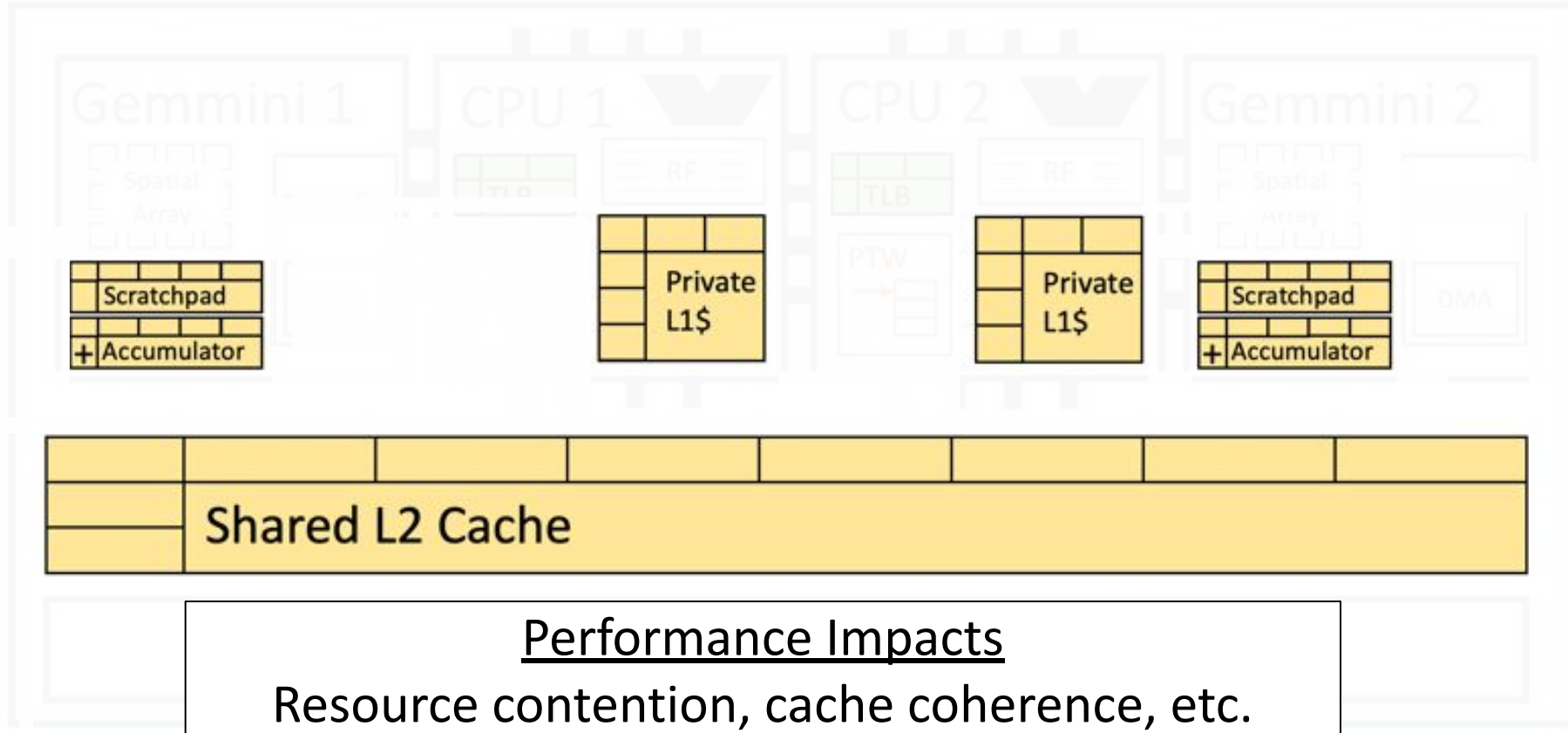
Berkeley
UNIVERSITY OF CALIFORNIA

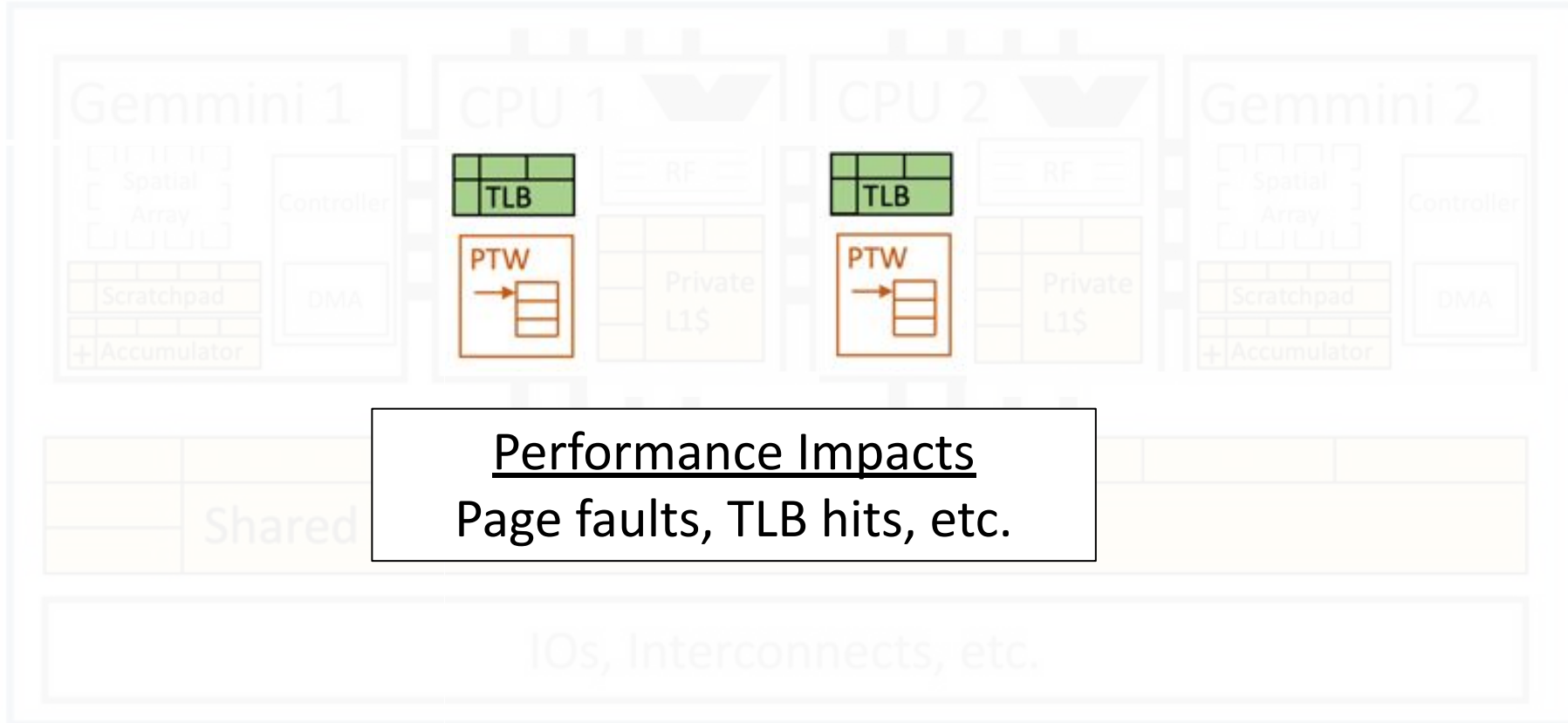# DL accelerators don't exist in isolation.

Full-System Visibility: Memory Hierarchy

Performance Impacts
Resource contention, cache coherence, etc.

Full-System Visibility: Virtual Addresses

Performance Impacts
Page faults, TLB hits, etc.

Full-System Visibility: Host CPUs

Performance Impacts
Unaccelerated kernels, etc.
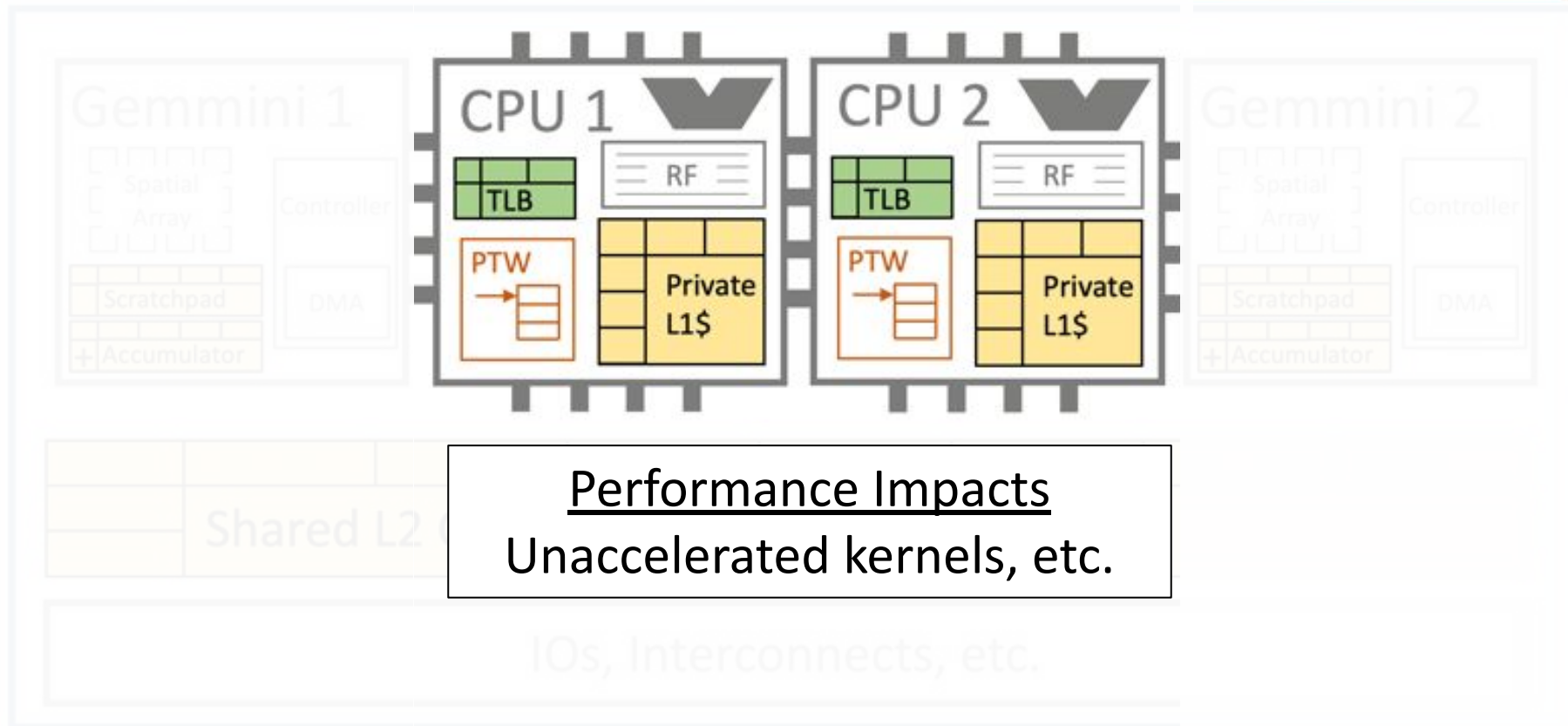
Gemmini

- DNN accelerator generator
- Flexible hardware template
- Full-system
- Open-source!
  - https://github.com/ucb-bar/gemmini