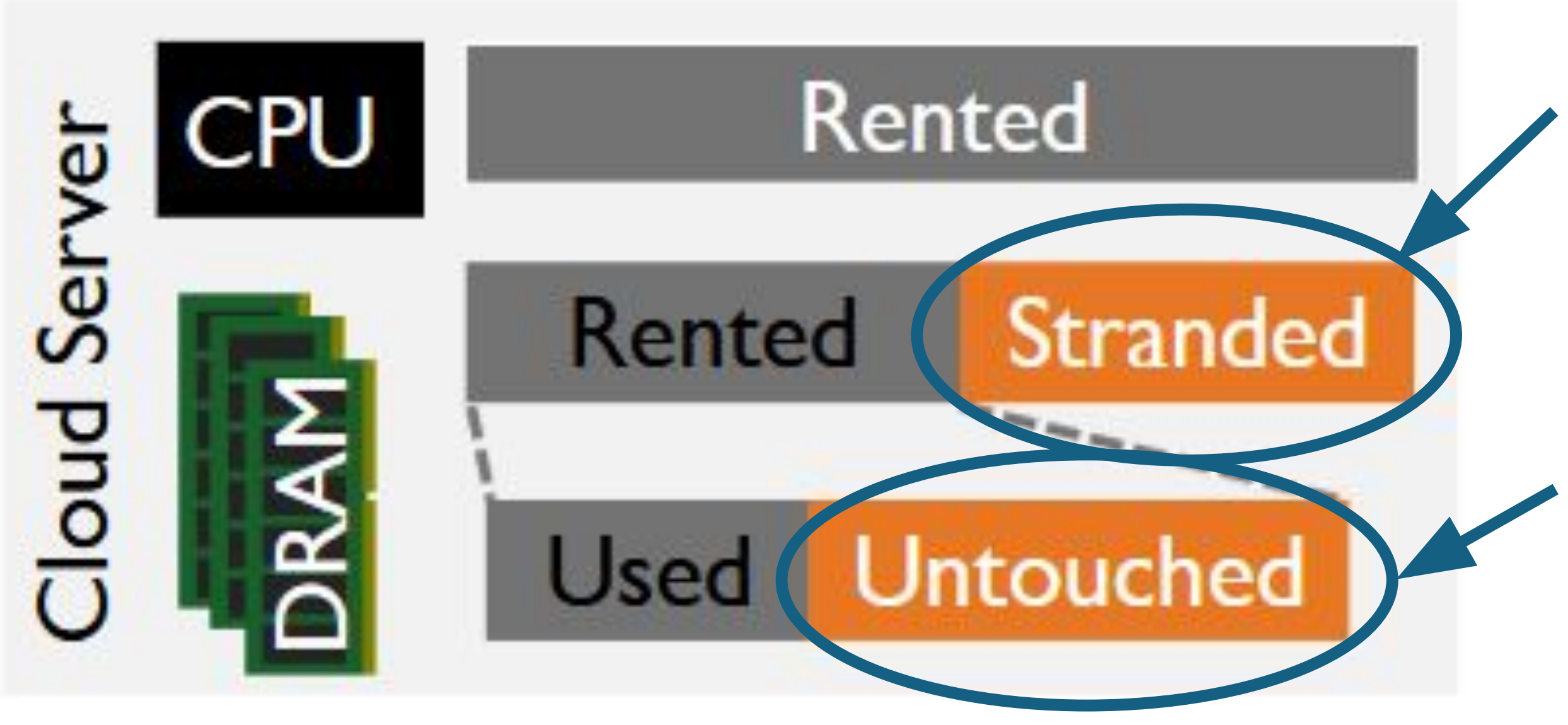
Rack-Scale Servers for the Post-Moore Era

Pooria Poorsarvi Tehrani, Mohammad Arman Soleimani

Motivation

•Hardware resources are underutilized and wasted in data centers



Up to 25% under high load

50% of VMs use <50% of provisioned memory

[Li '23]

•**Hardware Resource Disaggregation** is a promising solution

- ✔ Lower maintenance costs
- ✔ Easier resource pooling
- ✔ Isolated points of failure

- Requires rearchitecting hardware and software
- Needs new tools and simulators for research

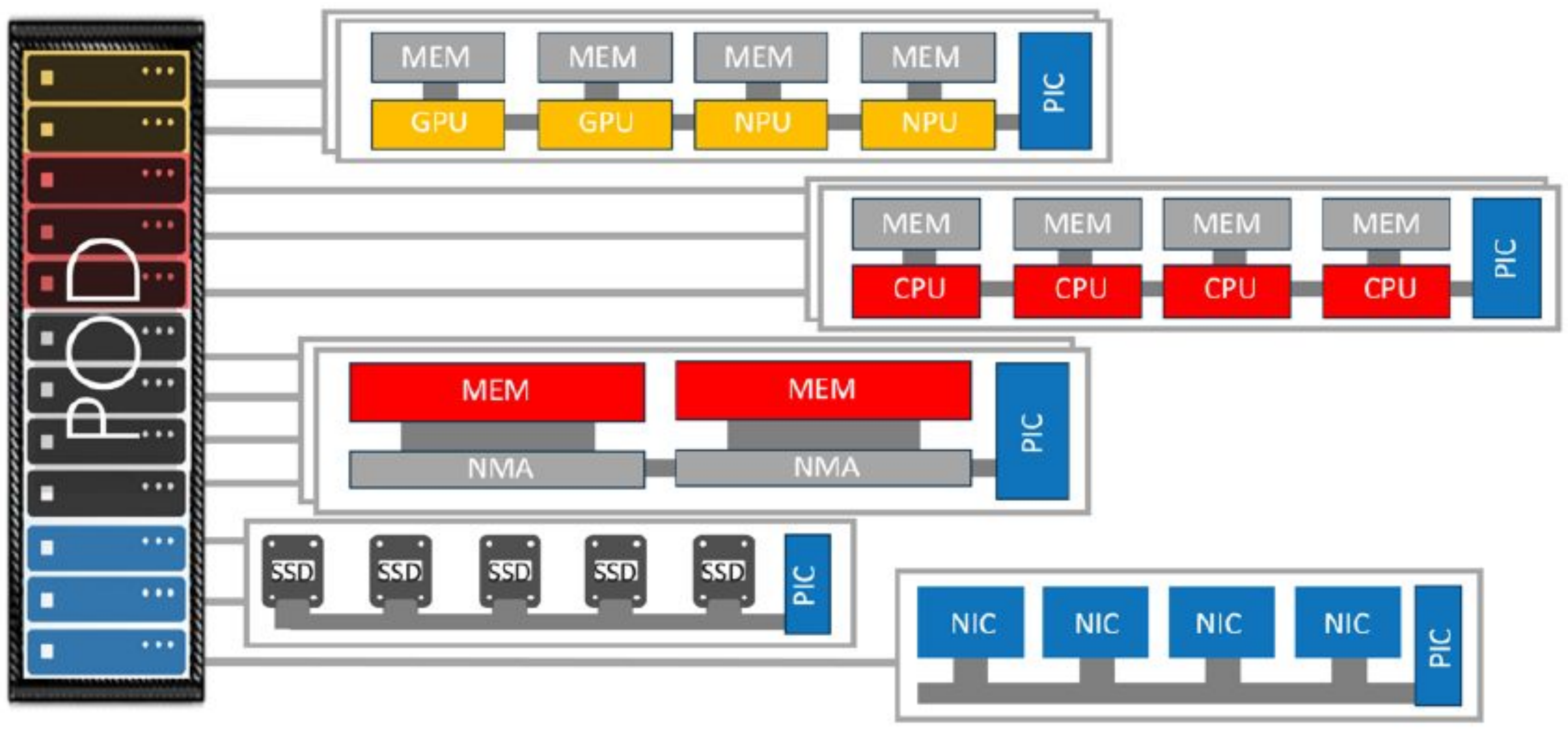
Current Solutions

Compute Express Link (CXL)
Interconnect between processors and devices such as memory and accelerators

- ✔ Coherent access to system and device memory
- ✔ Enables memory pooling and sharing devices
  Example: Pond [Li '23] saves 7% DRAM using a shared memory pool

- Cost of CXL hardware diminishes cost saved from pooling [Levis '23]
- High latency hurts performance
- Coherency needs might be different from what CXL implements

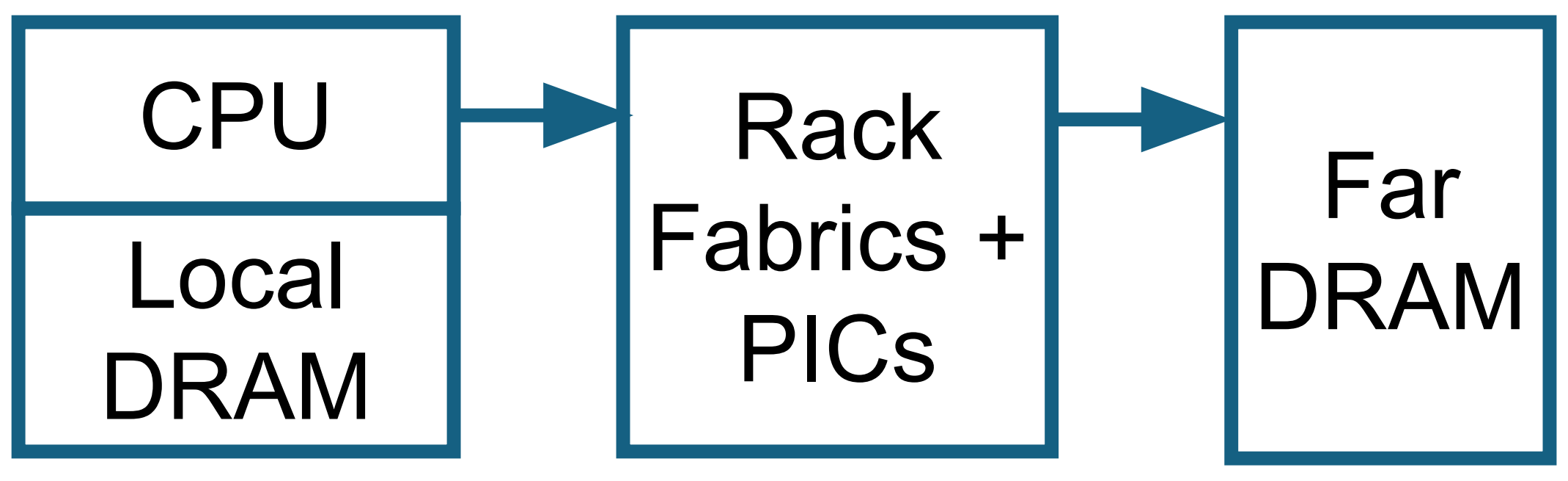
Rack-Scale Computing



•Disaggregate hardware resources across a rack
•Use intra-rack fabrics for rapid data movement
•Communicate among servers via Pod Interface Chip (PIC)
•Many unclear hardware and software design elements

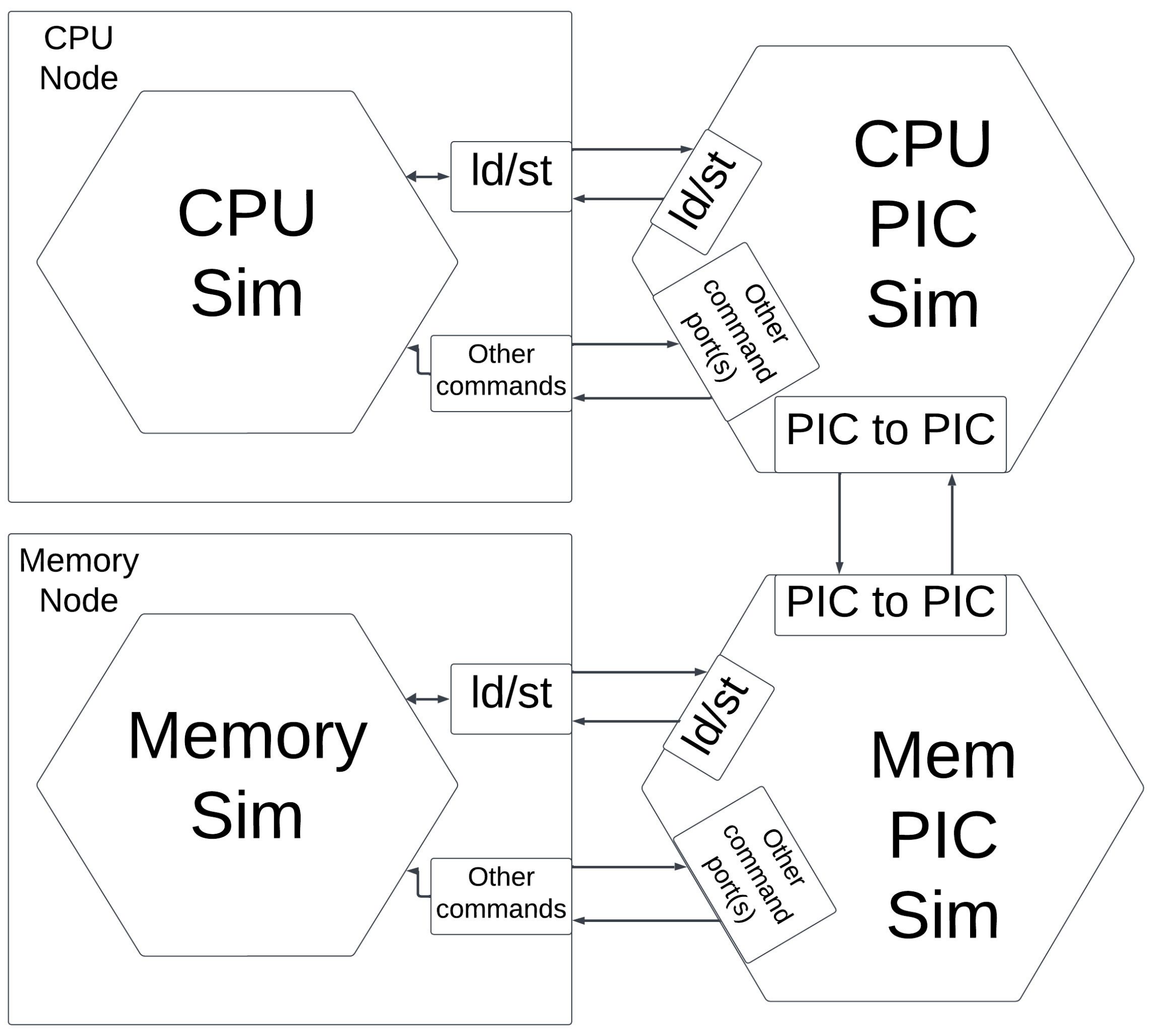
Proof of Concept

- Goal: Flexible Multi-Node Design
  - Slot in different simulators
  - Easily extend to add new nodes
  - Easily modify existing nodes
- Support different levels of simulation fidelity
- PoC scope: emulate just the CPU and disaggregated memory with the PICs on their path

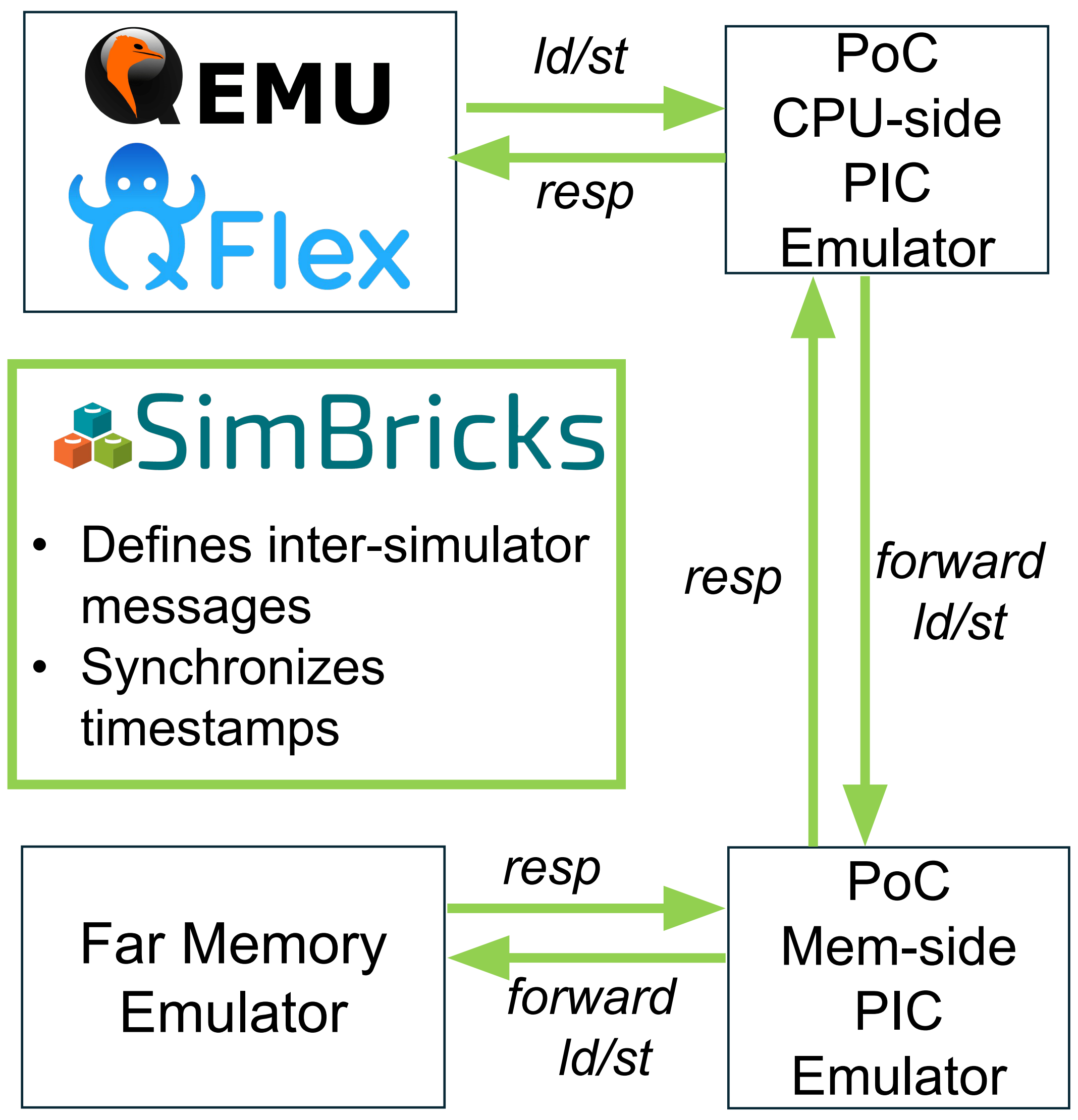


How do we design a simulation infrastructure to use for deeper studies into rack-scale design?

Design Interface



Baseline Architecture



- Defines inter-simulator messages
- Synchronizes timestamps

Current + Future Work

- Current areas of work
  - Simbricks communication cost
  - Simbricks sync cost
  - Memory characteristics of data centers
  - Qflex: timing simulator, supporting statistical sampling on 128 cores
- Next steps for rack-scale research
  - Add more PIC functionality
  - Investigate higher bandwidth serial fabrics to interface with remote memory
  - Study global address translation and coherence in memory pooling
  - Expand simulator functionality

PARSA PARALLEL SYSTEMS ARCHITECTURE LAB
EPFL
EcoCloud