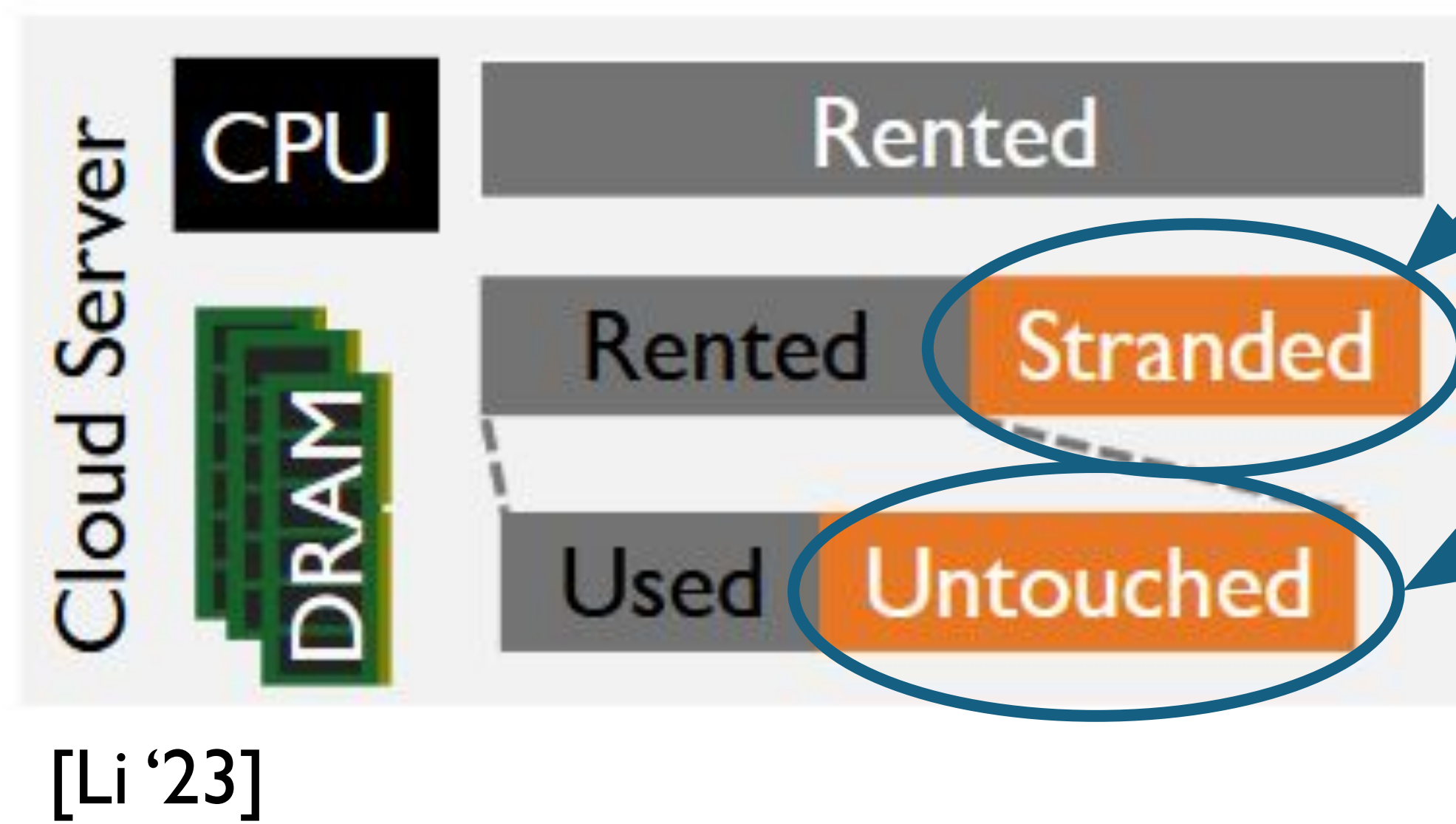


PolyRack: Servers for the Post-Moore Era

Pooria Poorsarvi Tehrani, Shanjing Lin, August Ning, Babak Falsafi, David Atienza, Thomas Bourgeat, Sanidhya Kashyap

Motivation

- Hardware resources are underutilized and wasted in data centers



- **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

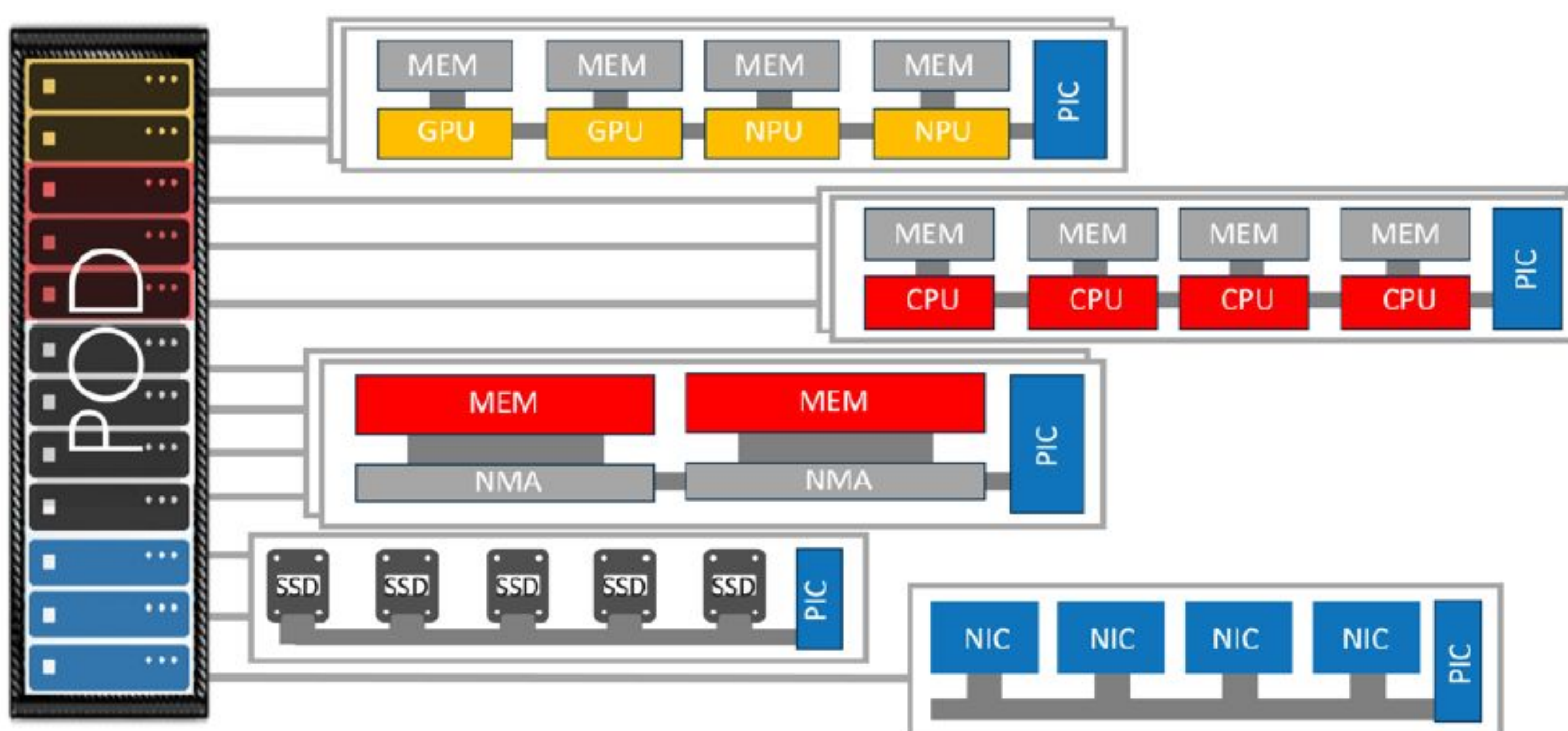
Current Solutions

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

PolyRack Computing



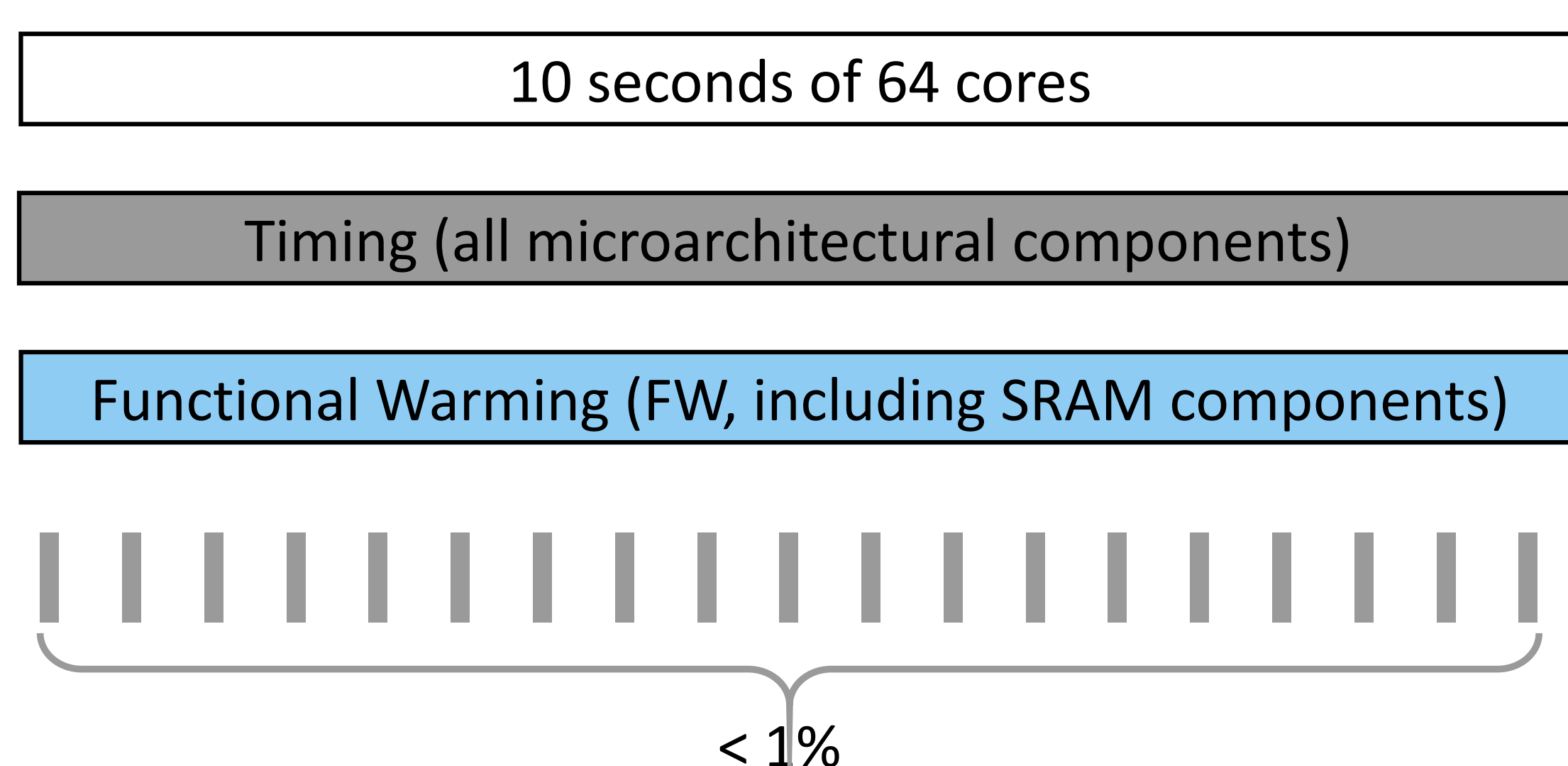
- Disaggregate hardware resources across a rack
- Use intra-rack fabrics for rapid data movement
- Communicate among servers via Pod Interface Chip (PIC)
- Many unclear hardware and software design elements

Current + Future Work

- Multi-node and multi-modal simulation
- LLM inference benchmarking:
 - LLM inference frameworks
 - Vector Databases
 - Agent frameworks
- LLM node : CPU and NPU design for inference
- Future areas of work
 - CPU node design: RPC, tool invocation for agents
 - Fabric design
 - Memory node design: data analytics, vector databases, KV caches

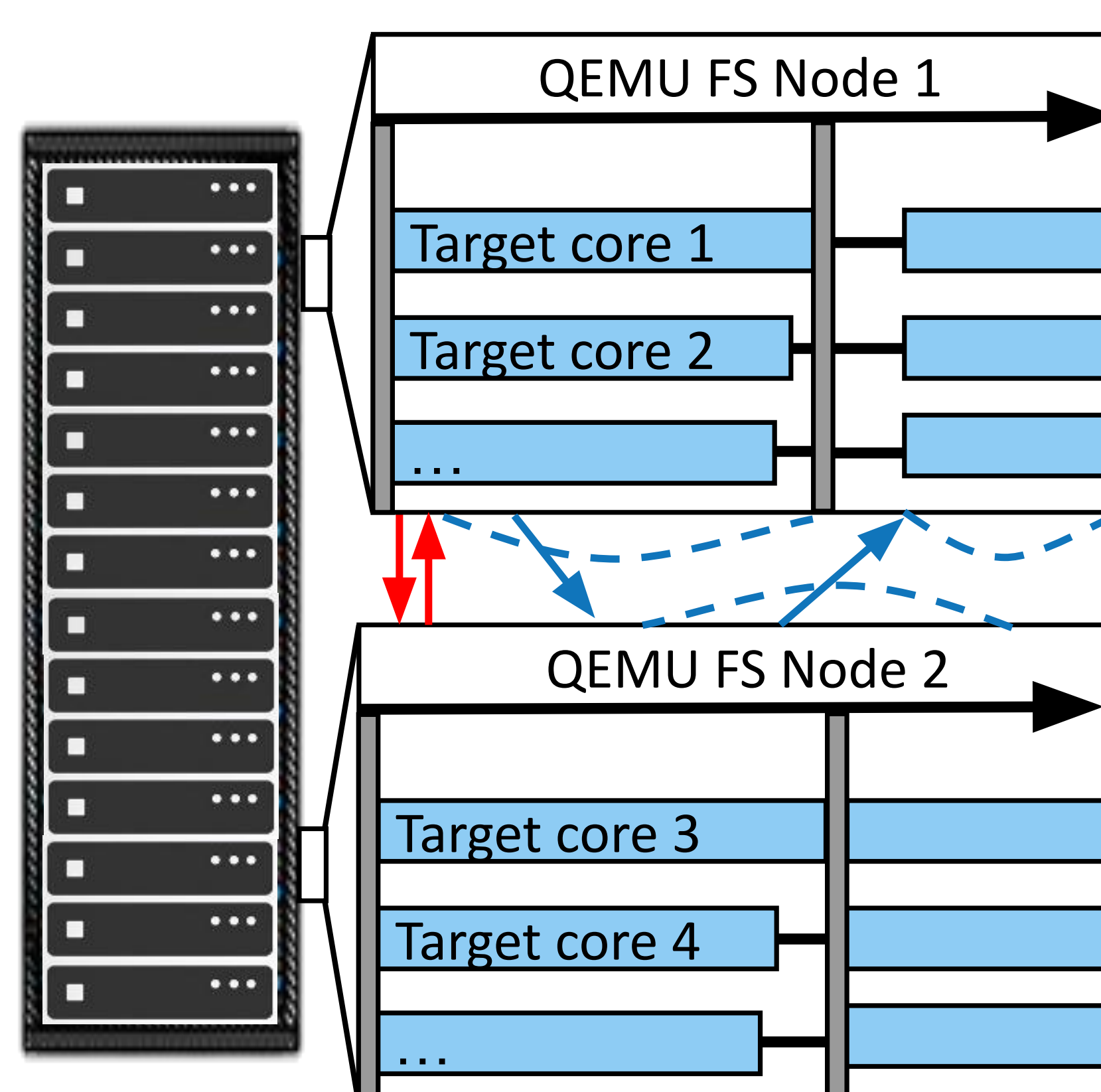
How do we design a simulation infrastructure to use for deeper studies into PolyRack design?

Near-GIPS Simulation



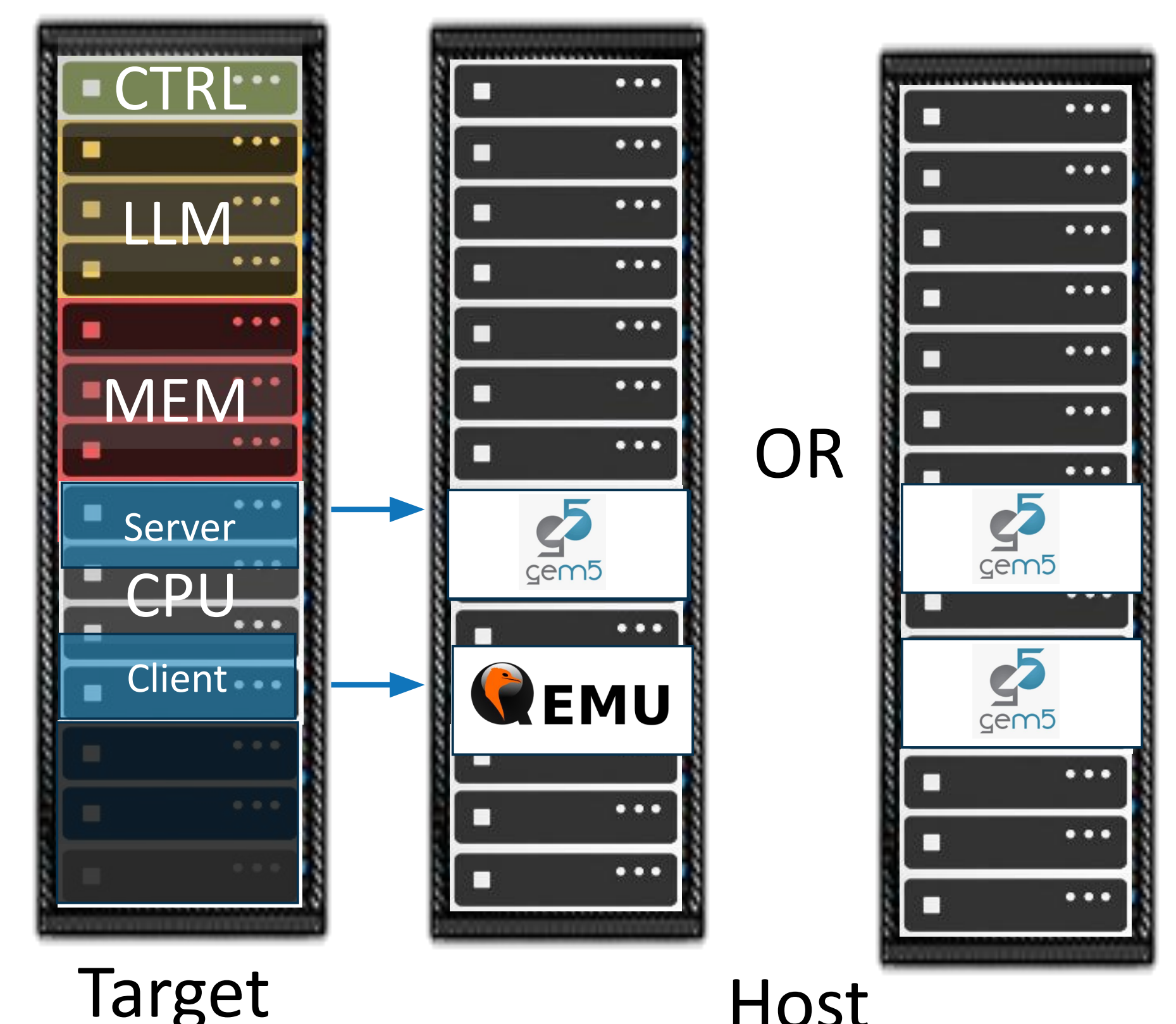
- Functional simulation
 - Only translates architecture
- Timing simulation: the oracle
 - 3-4 order of magnitudes slower than functional simulation
- Statistical sampling: functional warming and timing on sampling units [Wunderlich '03]

Multi-Node Simulation



- PolyRack needs simulation of racks
- Scale for many nodes with low latency

Multi-Modal Simulation



- Only model minimum needed state
- Universal checkpoints to be able to swap simulators