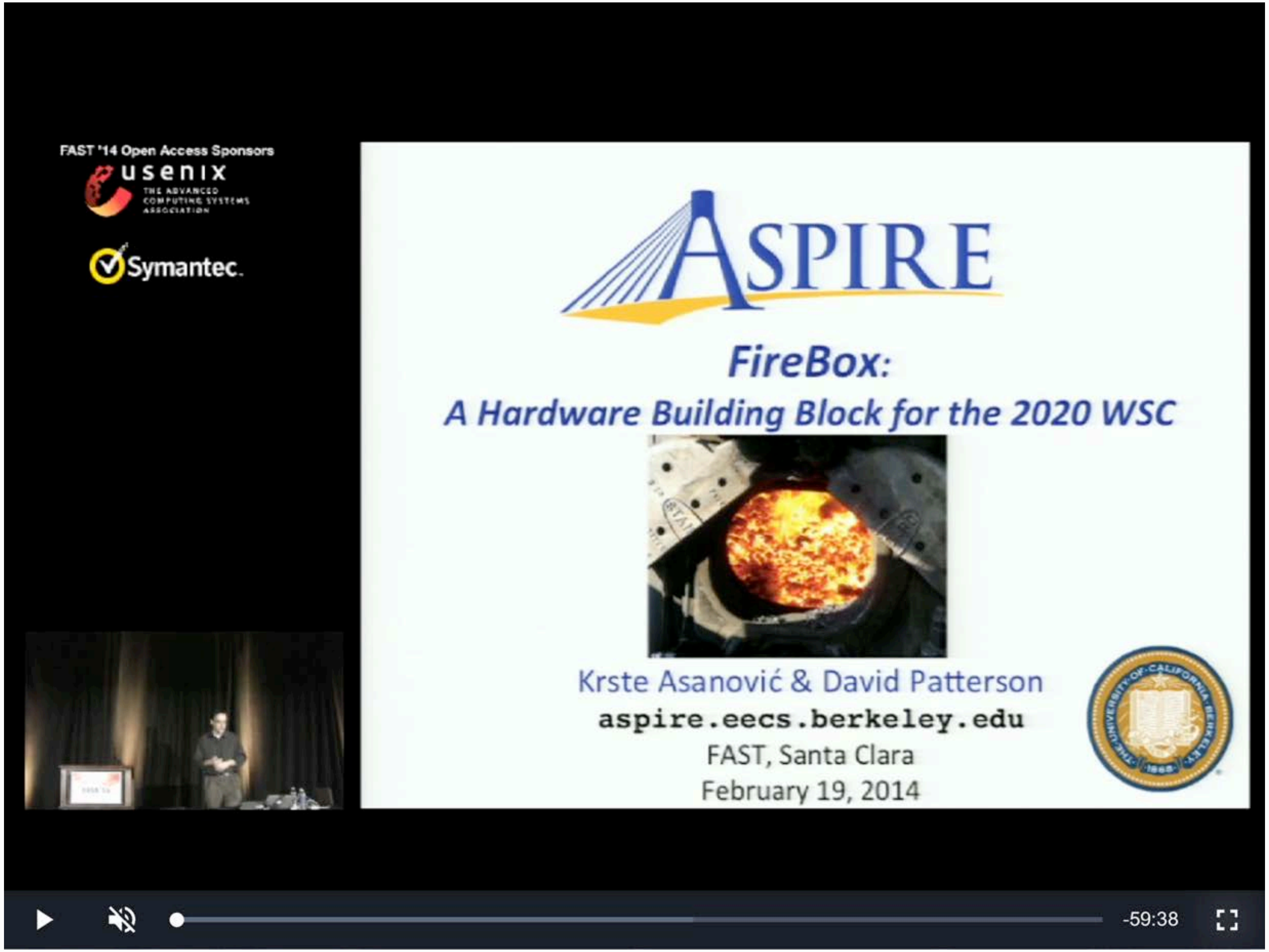
FireBox: Past, Present, and Future

NATHAN PEMBERTON

12/12/2017

In the year 2014…



FAST '14 Open Access Sponsors

usenix THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Symantec.

ASPIRE

FireBox:

A Hardware Building Block for the 2020 WSC

Krste Asanović & David Patterson

aspire.eecs.berkeley.edu

FAST, Santa Clara

February 19, 2014

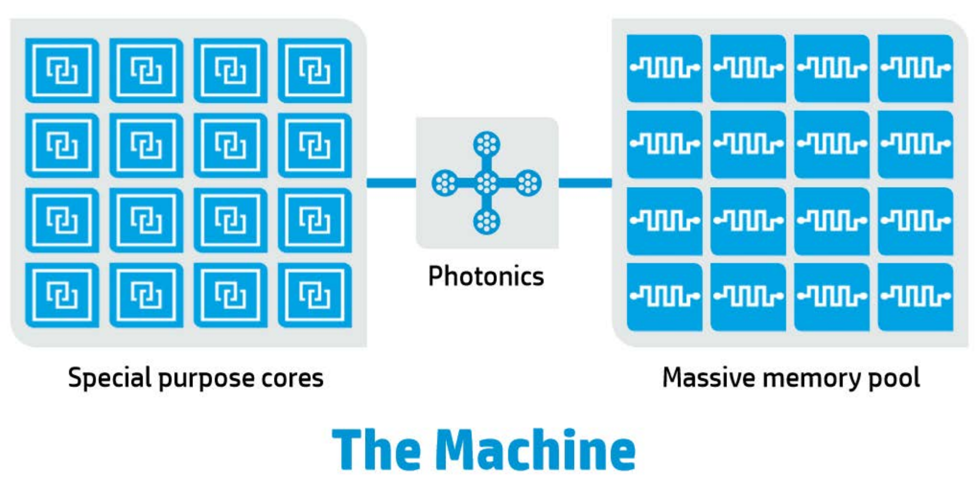
Motivating Trends

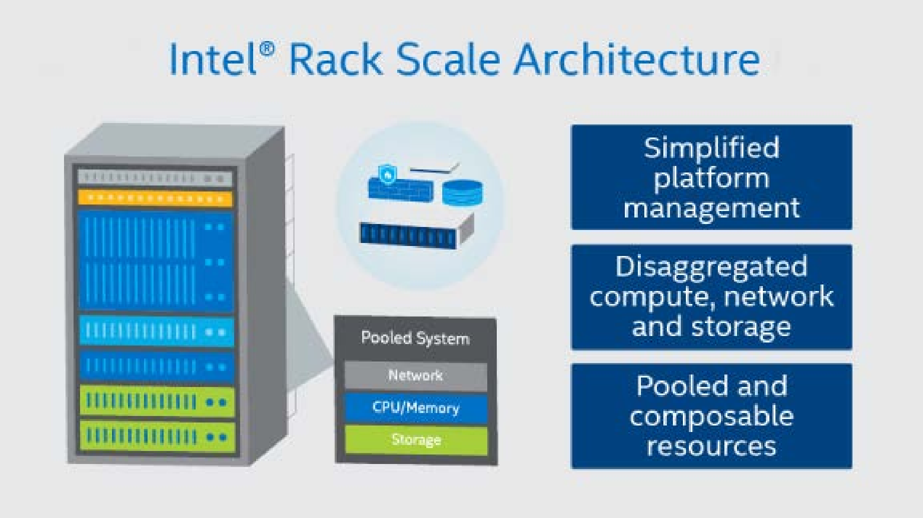
Compute moving to extremes: edge + cloud

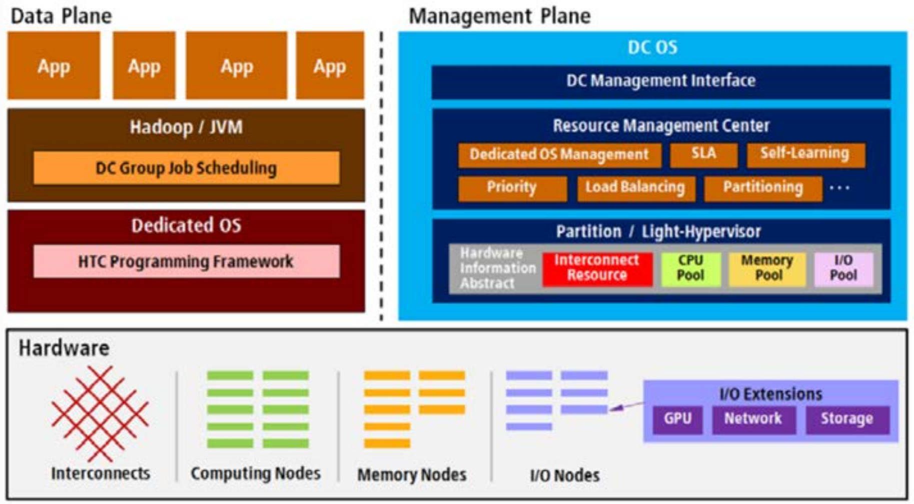
Specialization for the datacenter

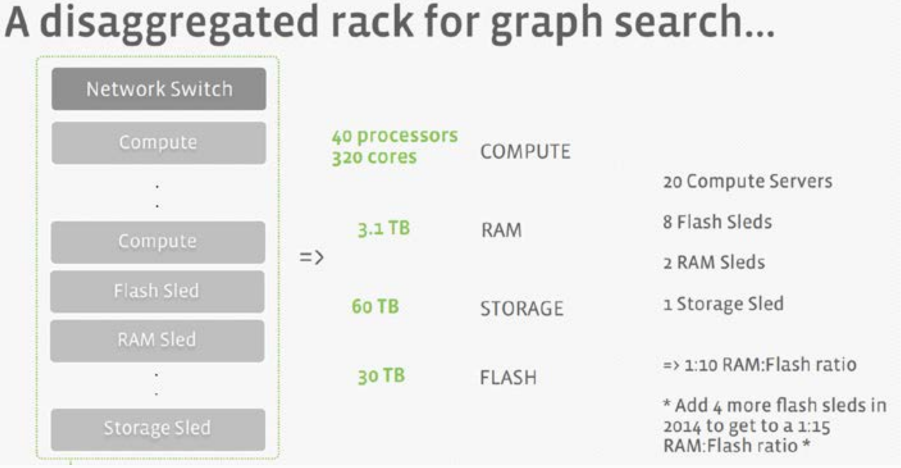
- ~2000
  - Commercial Off-The-Shelf (COTS) computers, switches, & racks
- ~2010
  - Custom computers, switches, & racks but build from COTS chips
- ~2020
  - Custom computers, switches, & racks using custom chips

WSCs in Industry




Intel® Rack Scale Architecture
Pooled System
Network
CPU/Memory
Storage
Simplified platform management
Disaggregated compute, network and storage
Pooled and composable resources




A disaggregated rack for graph search...

20 Compute Servers

8 Flash Sleds

2 RAM Sleds

1 Storage Sled

=> 1:10 RAM:Flash ratio

* Add 4 more flash sleds in 2014 to get to a 1:15 RAM:Flash ratio *

Hardware Specialization

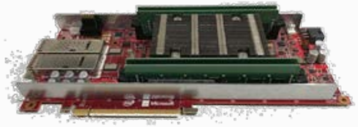
Compute

Moved faster than expected

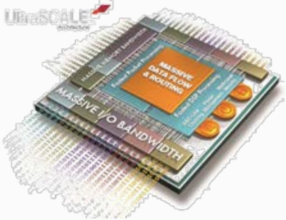
GPGPU

Google TPU

Microsoft Catapult

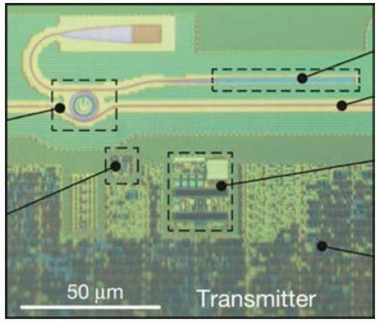


Amazon F1



Networks

Si Photonics not quite here yet



But NWs have improved rapidly

INFINIBAND™

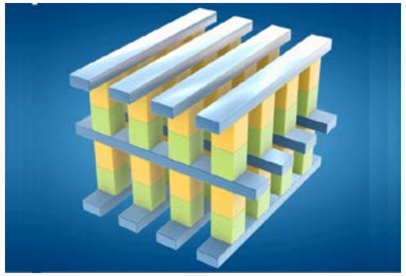
100G

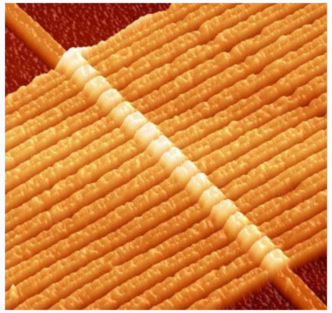
GENZ

RoCE™

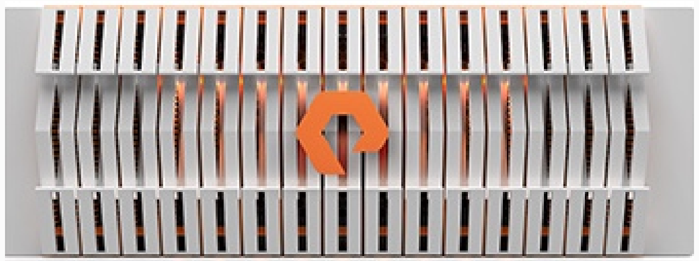
Storage and Memory

New physics moved slower than expected





Huge growth in flash



Applications and Operating Systems

New Software Models

- Service-Oriented Architecture (SOA) - All components are designed to be services
  - Communicate only via the network (cluster-first mentality)
  - Services reusable and interoperable
- Serverless – Event-driven stateless functions

Cluster-wide management

- "Datacenter OS" (Mesos, Kubernetes, etc…)
- Local OS less powerful (cluster-first mentality)

Challenges

Tail Latency

- At scale, slow == failed
- Need techniques to get good latency from unpredictable parts
- New Metric: 99% tail (average much less important)

Complex Memory Hierarchies

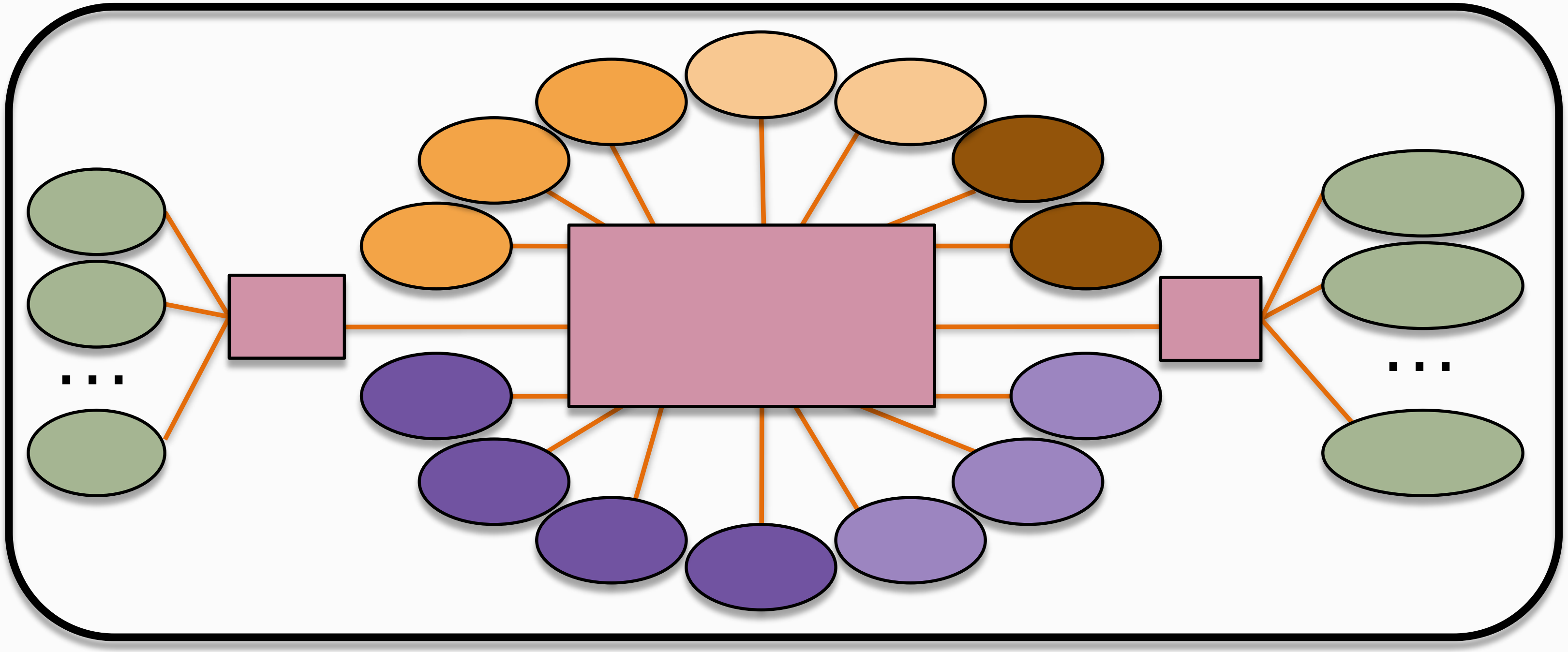
- Deeper: HBM, DRAM, NVM, Flash, Disk, Tape
- Wider: NUMA, RDMA, Disaggregation

Security

- No longer considered safe within WSC
- Encryption demanded at all times

Firebox at Berkeley

What’s Happened?



Systems and Runtimes

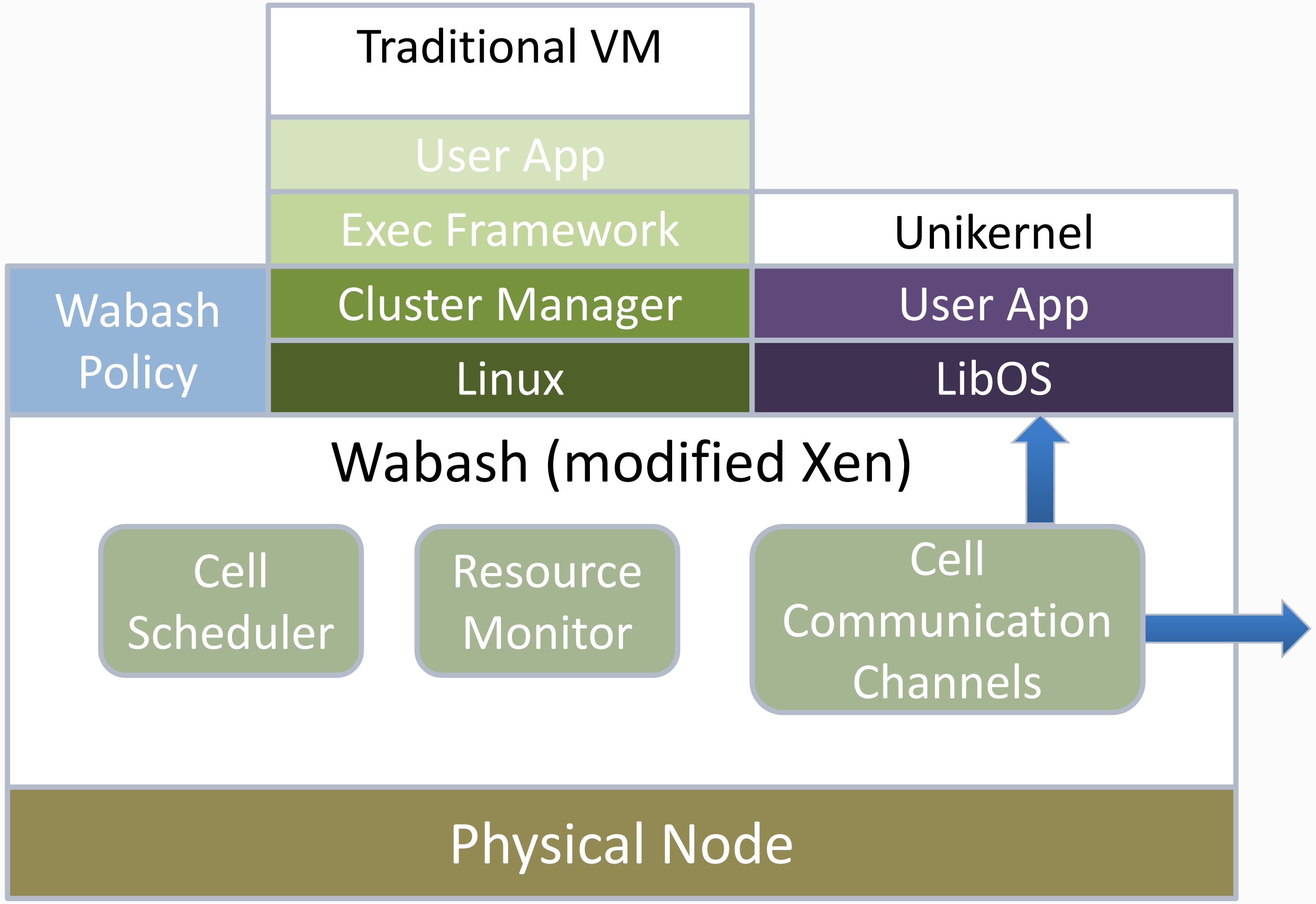
Wabash OS

Project Goals and Approach:

- Push “DCOS” concept further
- Based on concepts from Tessellation
- Explored OS noise (and other overheads)

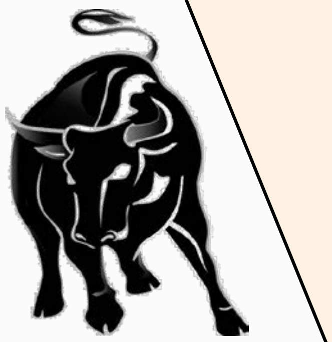
Conclusions

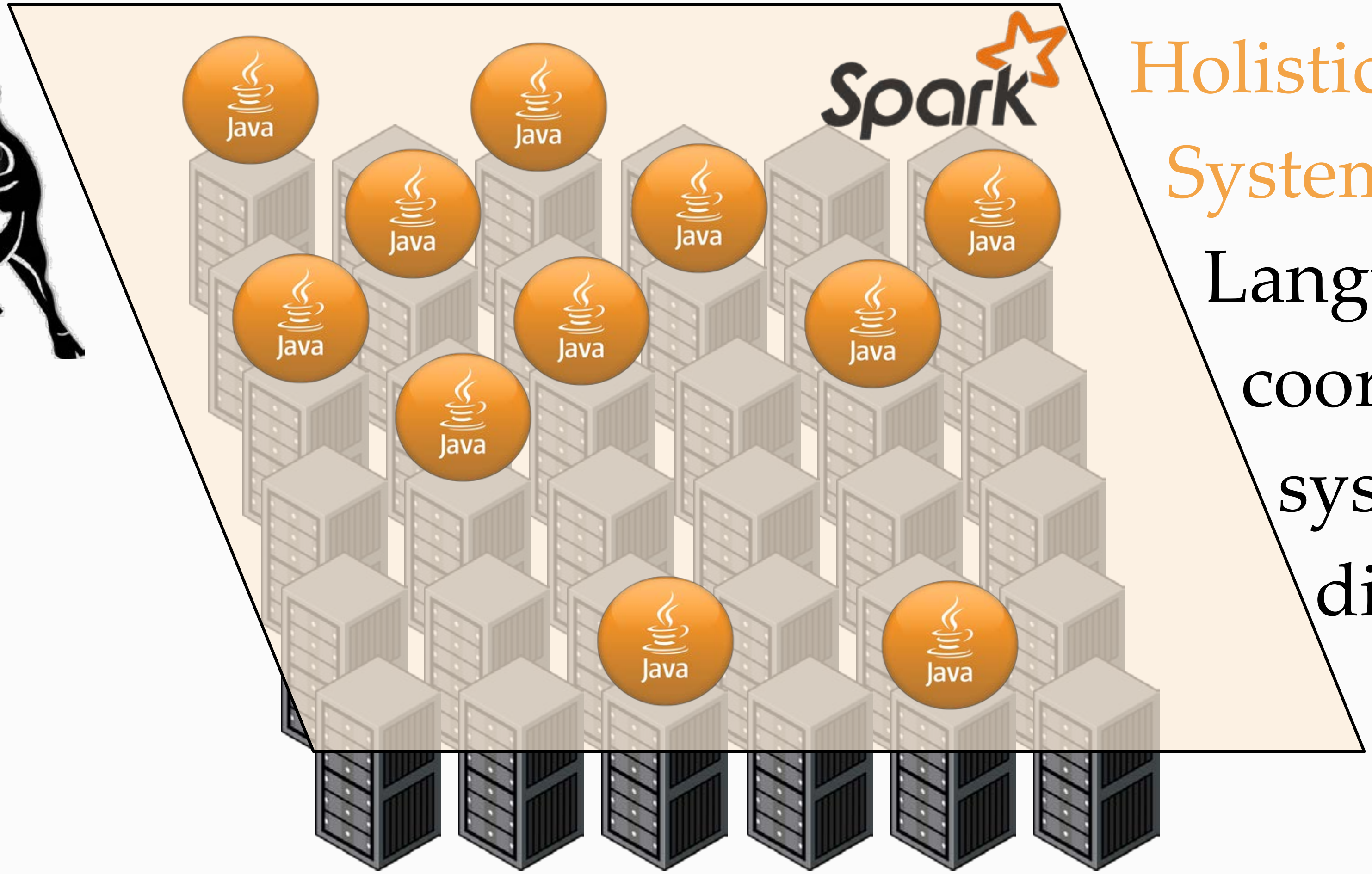
- “Embrace the Noise”
  - Better to be tail-tolerant than perfectly predictable
- Path to performance without new OS
  - New kernel-bypass techniques
  - Self-virtualizing hardware (SR-IOV)



Systems and Runtimes

Taurus and the Holistic Runtime

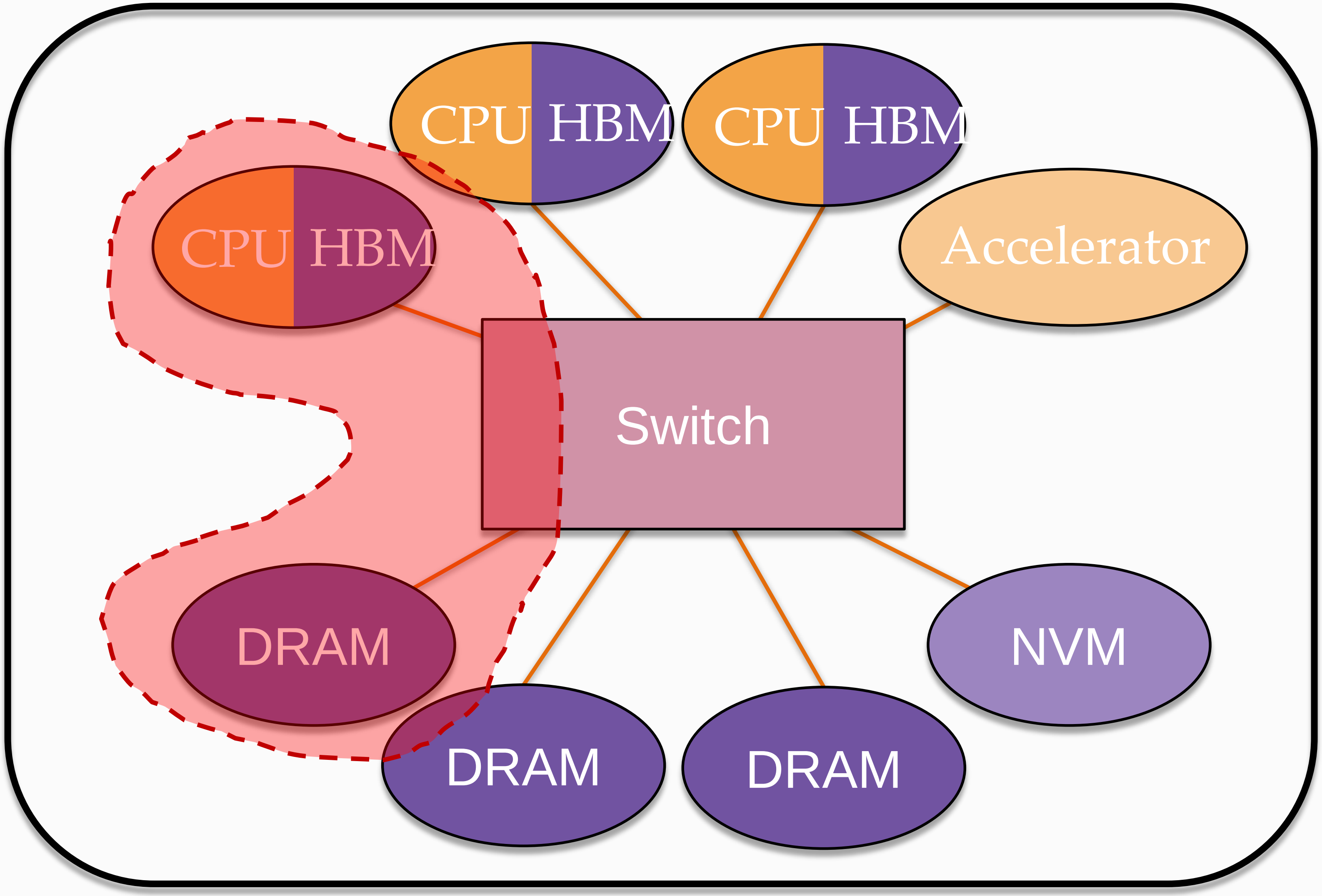




Holistic Language Runtime System: A Distributed Language Runtime to coordinate the runtime systems underlying a distributed application

Remote Memory

What is the right interface?

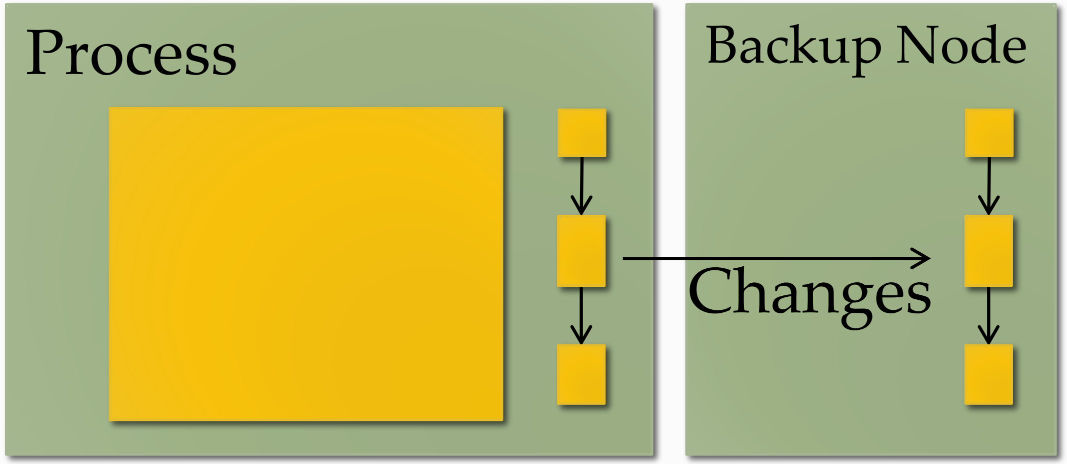


Remote Memory

Explicit Interfaces

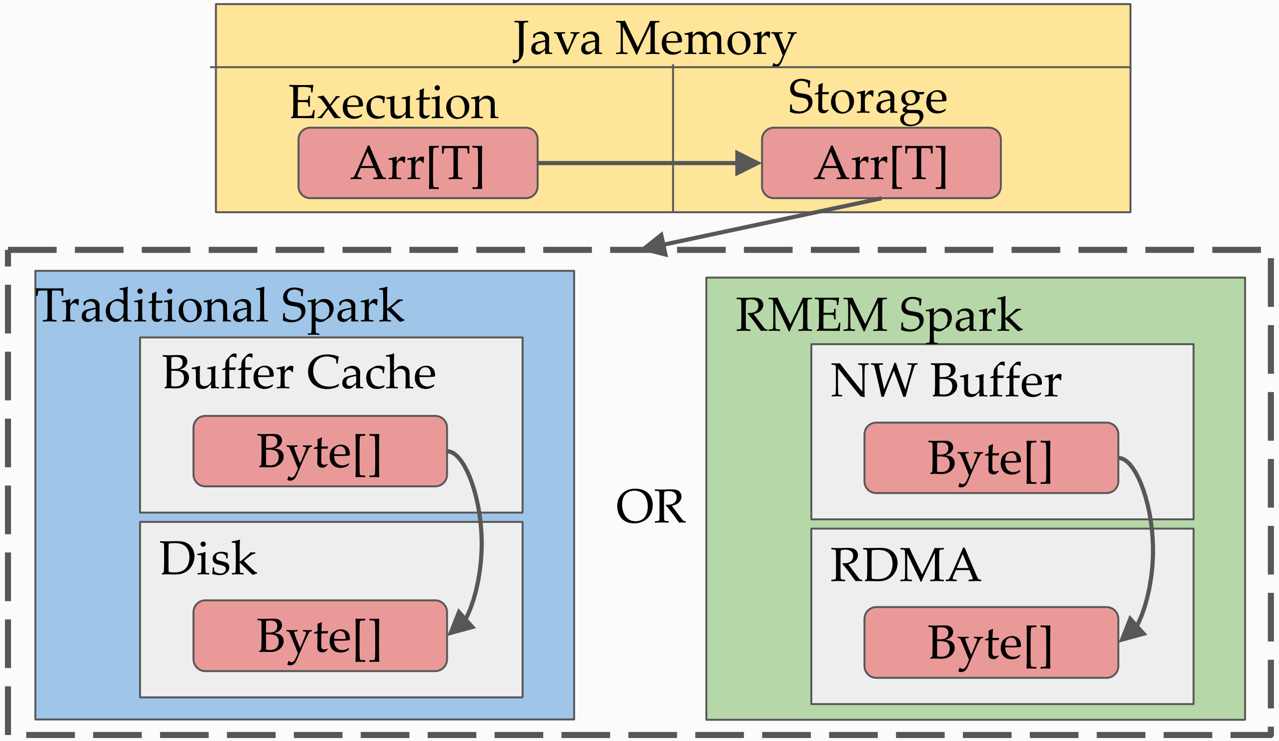
Nephele

→Transactional memory backup to NVM



Spark over RDMA

→Use remote memory to cache Spark objects

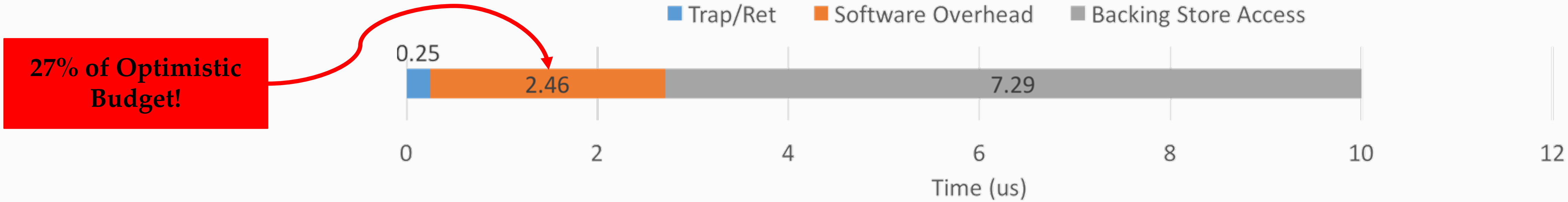


Remote Memory

Cache-Like Interfaces

Network Requirements for Disaggregation

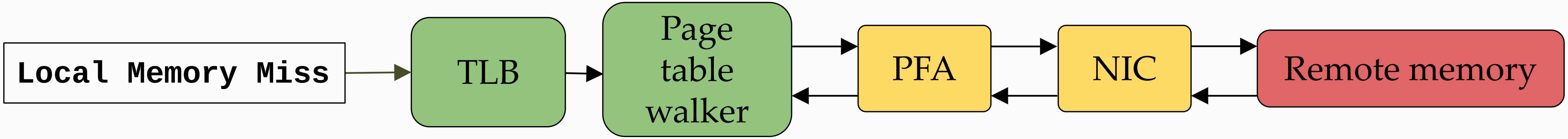
- ⟶Found 100Gbps@3us may be sufficient for memory disaggregation
- ⟶But! SW overheads are significant



27% of Optimistic Budget!

Page-Fault Accelerator

- ⟶Handle page-faults in HW
- ⟶Manage evictions asynchronously in OS

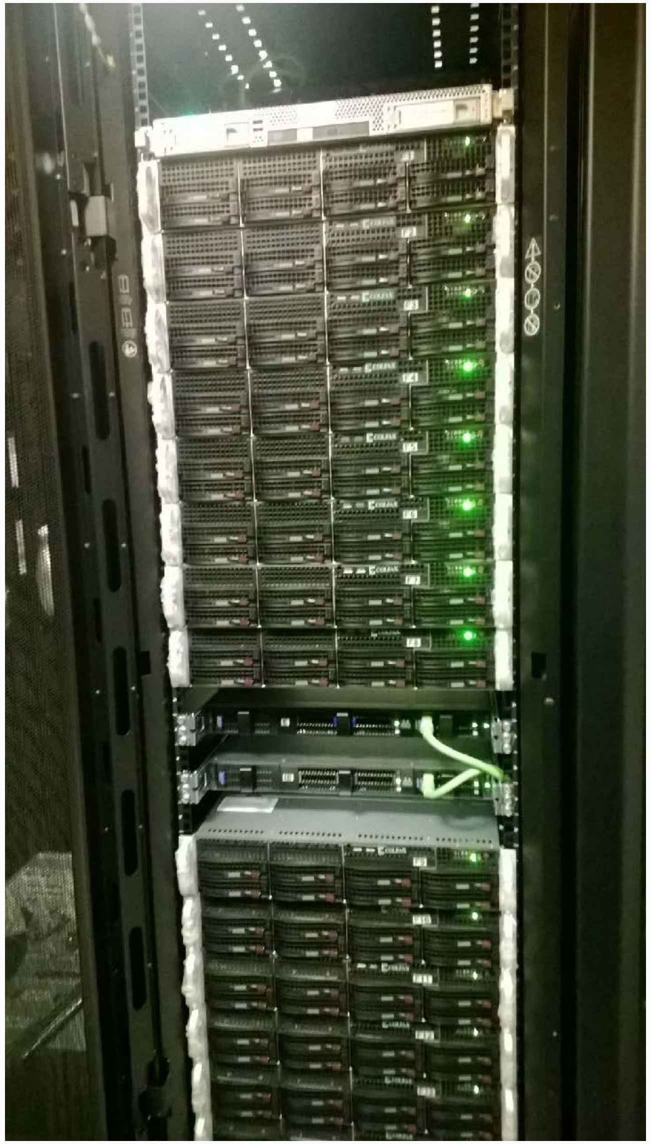


The Future of Firebox

Biggest Challenge so far: Lack of scalable and flexible experimentation platform

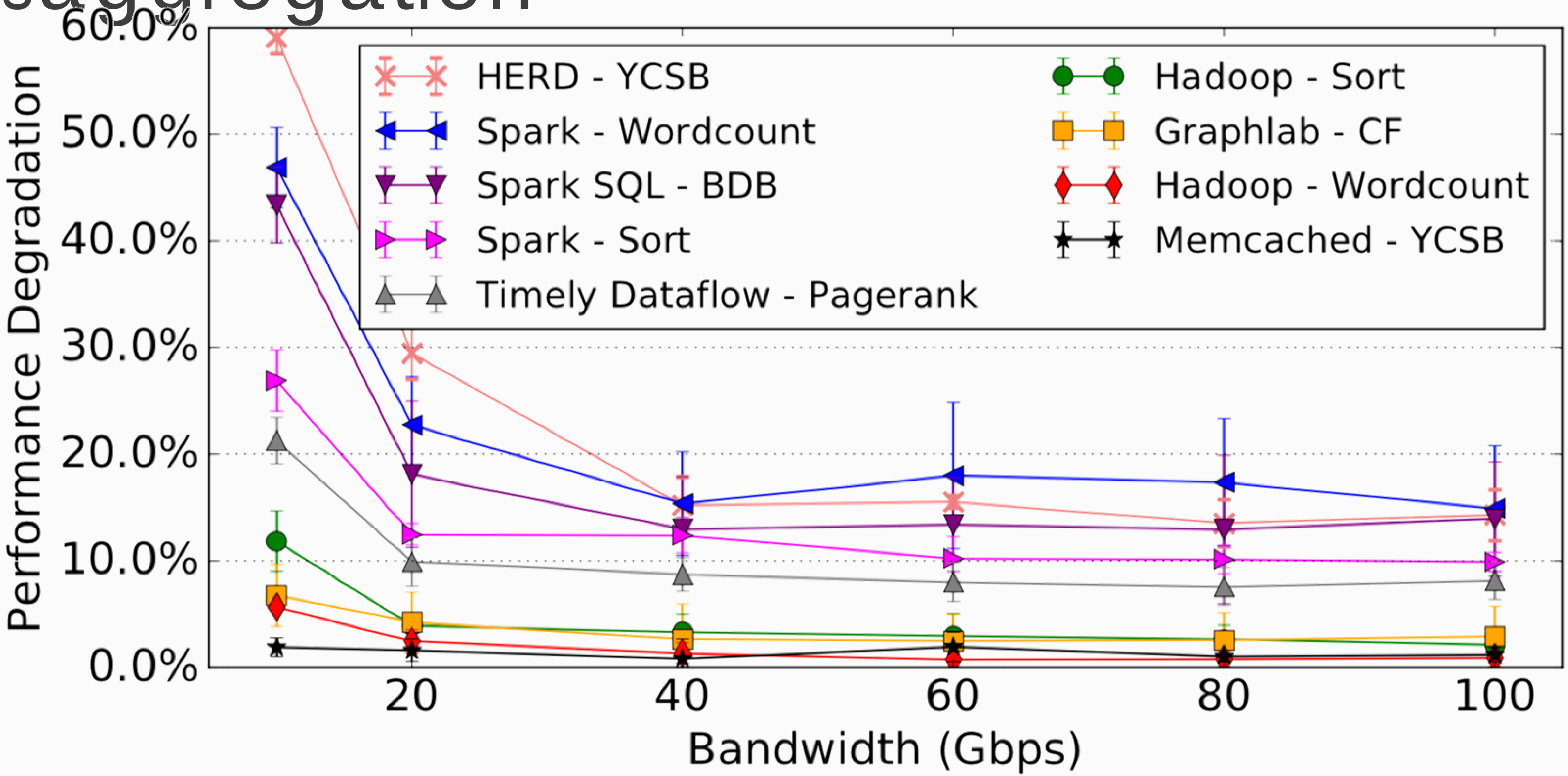
What if we could simulate thousands of nodes with arbitrary RTL?

- Evaluate impact of swapping on NW
- Place accelerators at network endpoints (”bump in the wire”)
- Experiment with memory blade design (compute-in-memory?)
- Extend hardware enclaves across NW and heterogeneous devices
- Gather detailed traces without impacting performance
- …

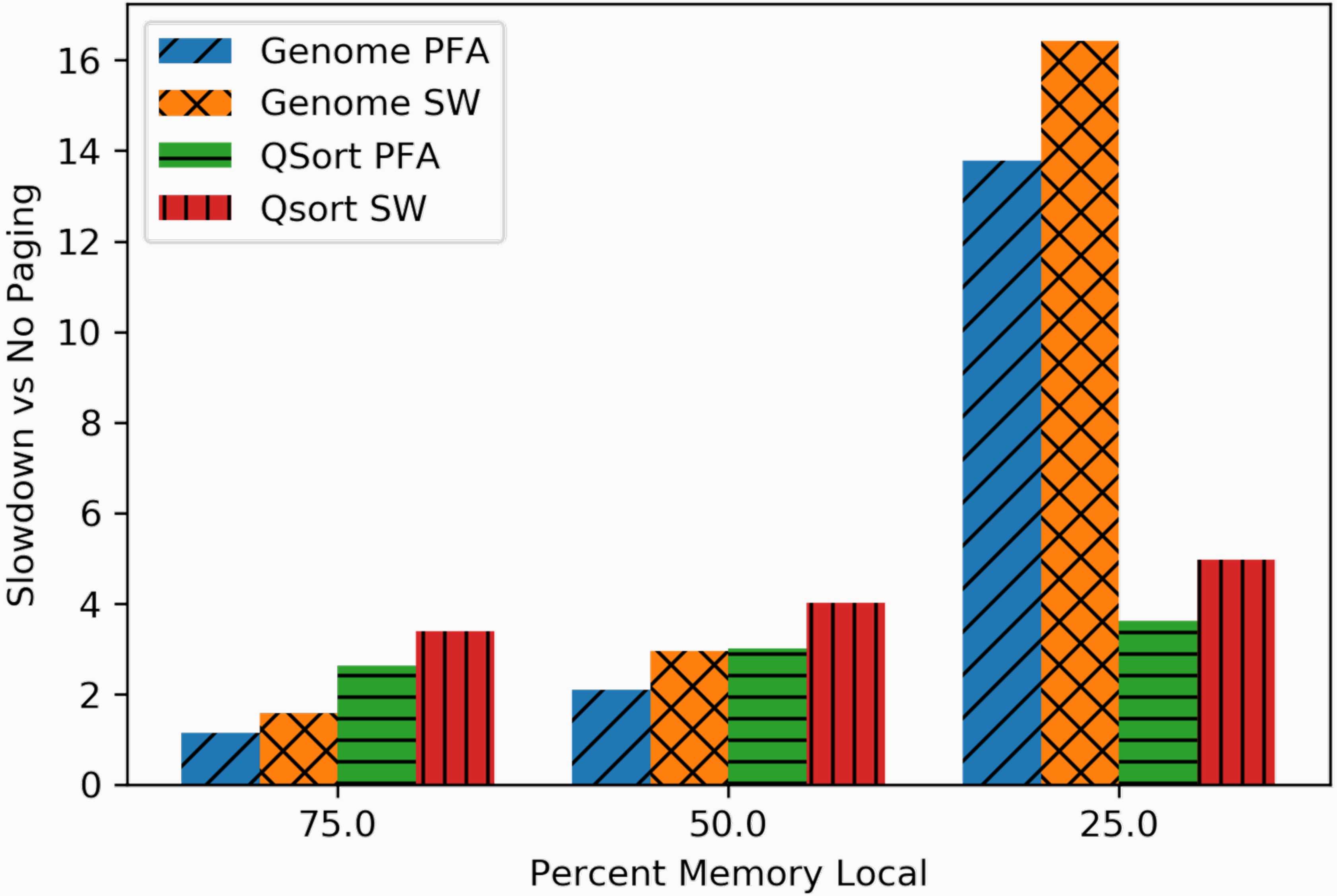


BACKUP AND REFERENCE SLIDES

Network Requirements for Disaggregation



Page-Fault Accelerator


Genome PFA
Genome SW
QSort PFA
Qsort SW
Slowdown vs No Paging
Percent Memory Local