OpenSDS mini summit COPENHAGEN

Storage Intelligence for the Data Center

Allen Samuels

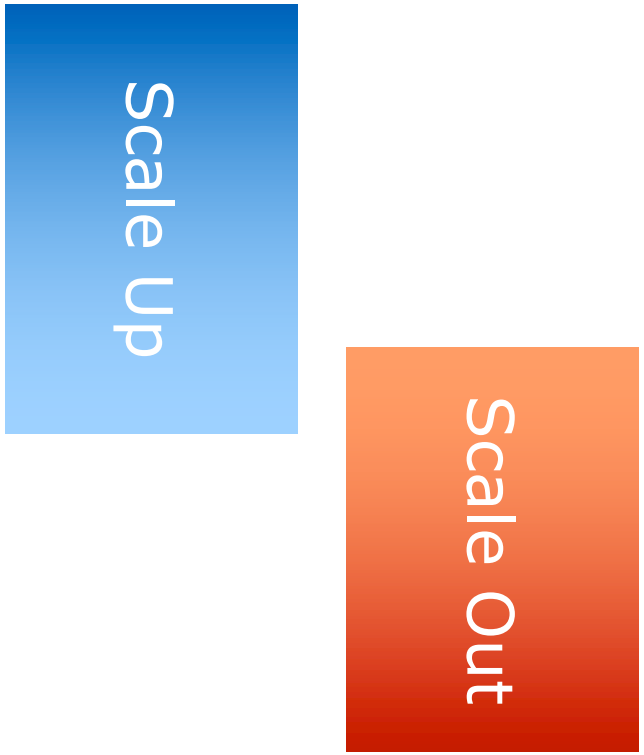
R&D Engineering Fellow, CTO Office

Western Digital

Computational System Architecture Evolution

- Moore’s law drives major upheavals about every 10 years

- 60’s Mainframes
- 70’s Minicomputers
- 80’s uProcessors
- 90’s Client/Server
- 00’s Web
- 10’s Virtualization & Cloud computing
- 20’s ?



Vectors of Innovation

What do you do with a quadrillion transistors?

- Enhancing the current architecture
  - More Cores
  - Bigger Caches
  - Faster pipes

Vectors of Innovation

What do you do with a quadrillion transistors?

- Continued Innovation within traditional architectural boundaries
  - Machine Learning processors
  - Siliconization of common software – SDN, DPDK, SPDK, RDMA, etc.
  - NVMe Over Fabrics
  - Rack-Scale architecture and disaggregation

Vectors of Innovation

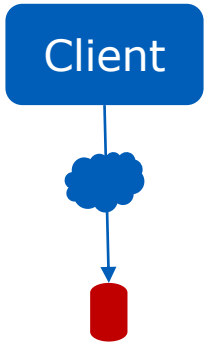
What do you do with a quadrillion transistors?

- Innovation transcending traditional boundaries
  - Persistent memory
  - Intelligence metastasizes
    - Into the network
    - Into the storage

Innovating in Networking and Storage

Current Best Practices

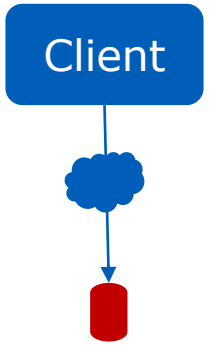
"Logical" view of networked storage



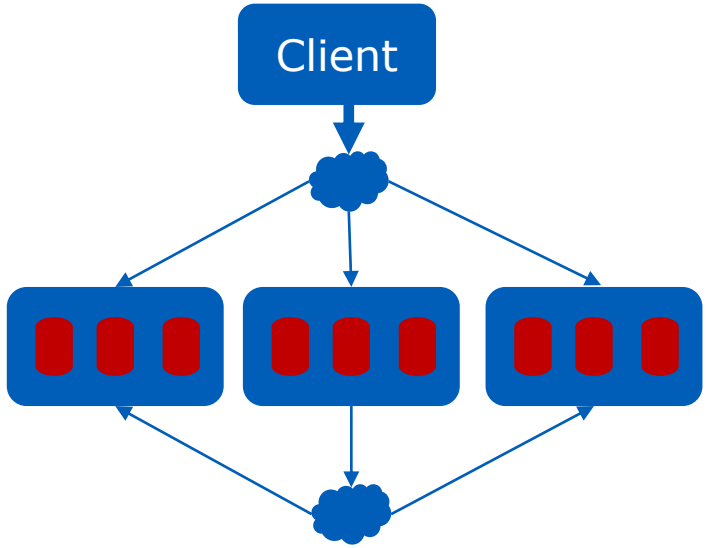
Innovating in Networking and Storage

Current Best Practices

"Logical" view of networked storage



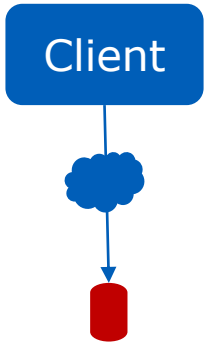
Is usually built like this



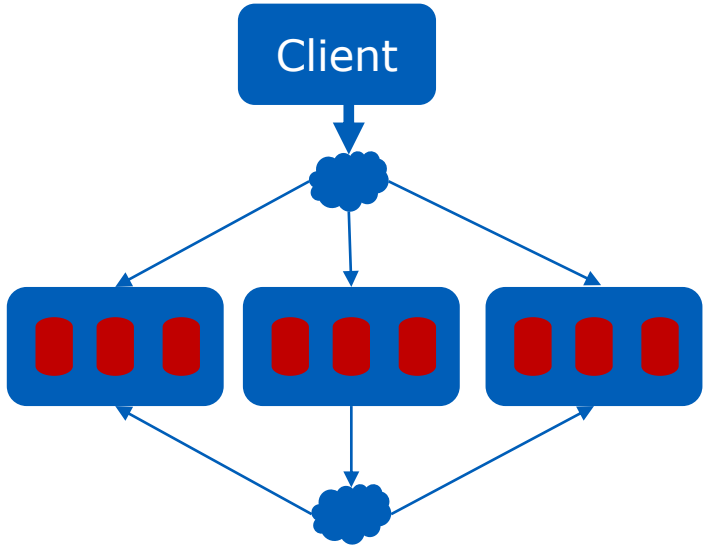
Innovating in Networking and Storage

Current Best Practices

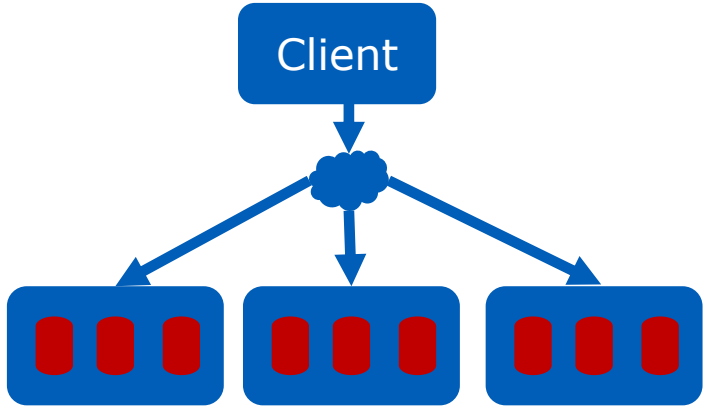
"Logical" view of networked storage



Is usually built like this



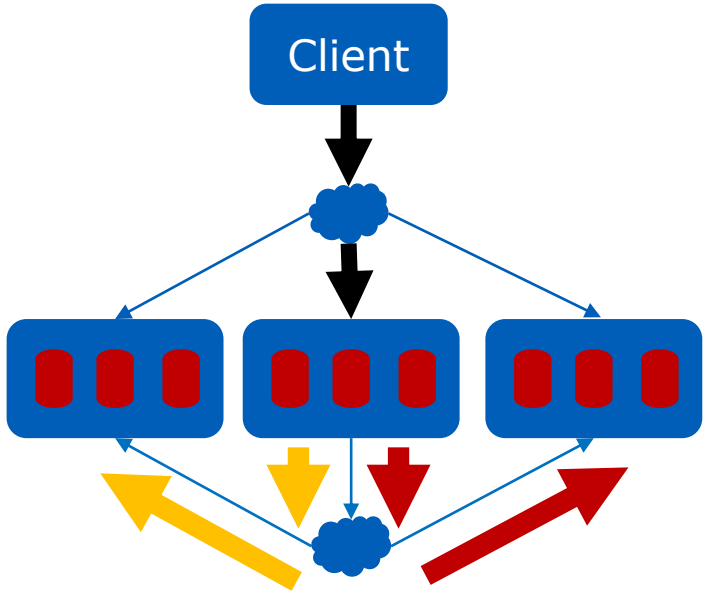
Or sometimes like this



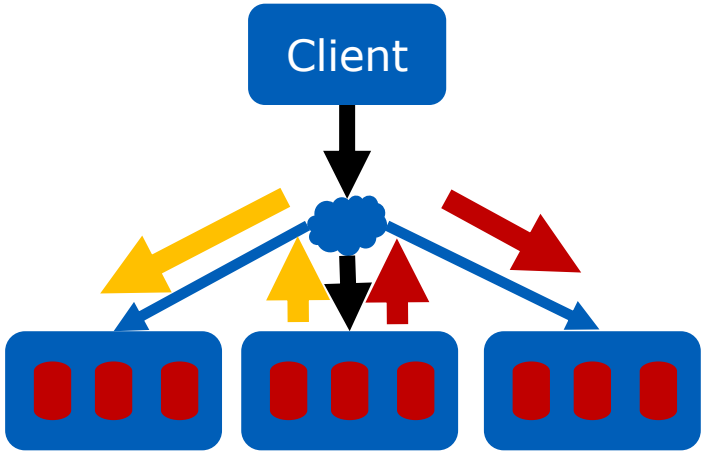
Innovating in Networking and Storage

Current Best Practices

A write operation must be conveyed to multiple systems in the cluster



Or like this

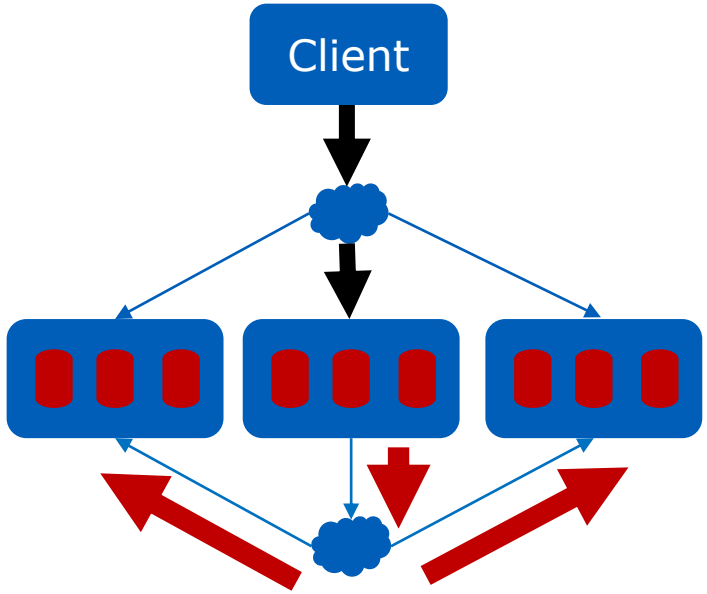


- Substantial network duplication in write operations

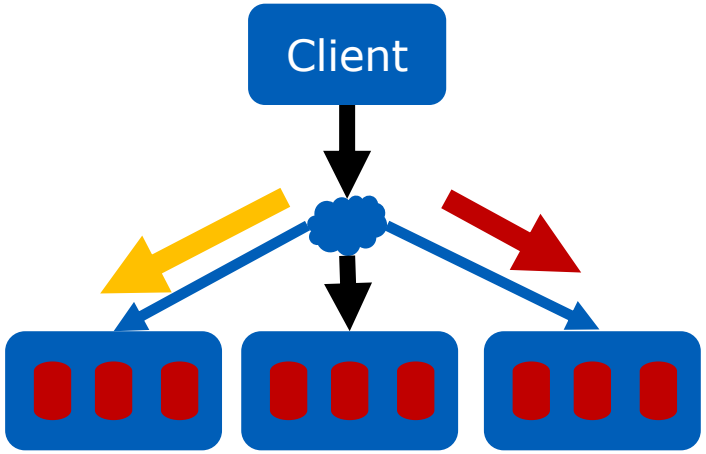
Innovating in Networking and Storage

Injecting Storage Intelligence into the Network

Intelligent switch handles replication/erasure coding



Or like this



- Intelligent switch handles replication/erasure coding
- Up to 33% (6 ➔ 4) reduction in bandwidth consumption for writes
- Similar reduction in bandwidth consumption for erasure coded reads

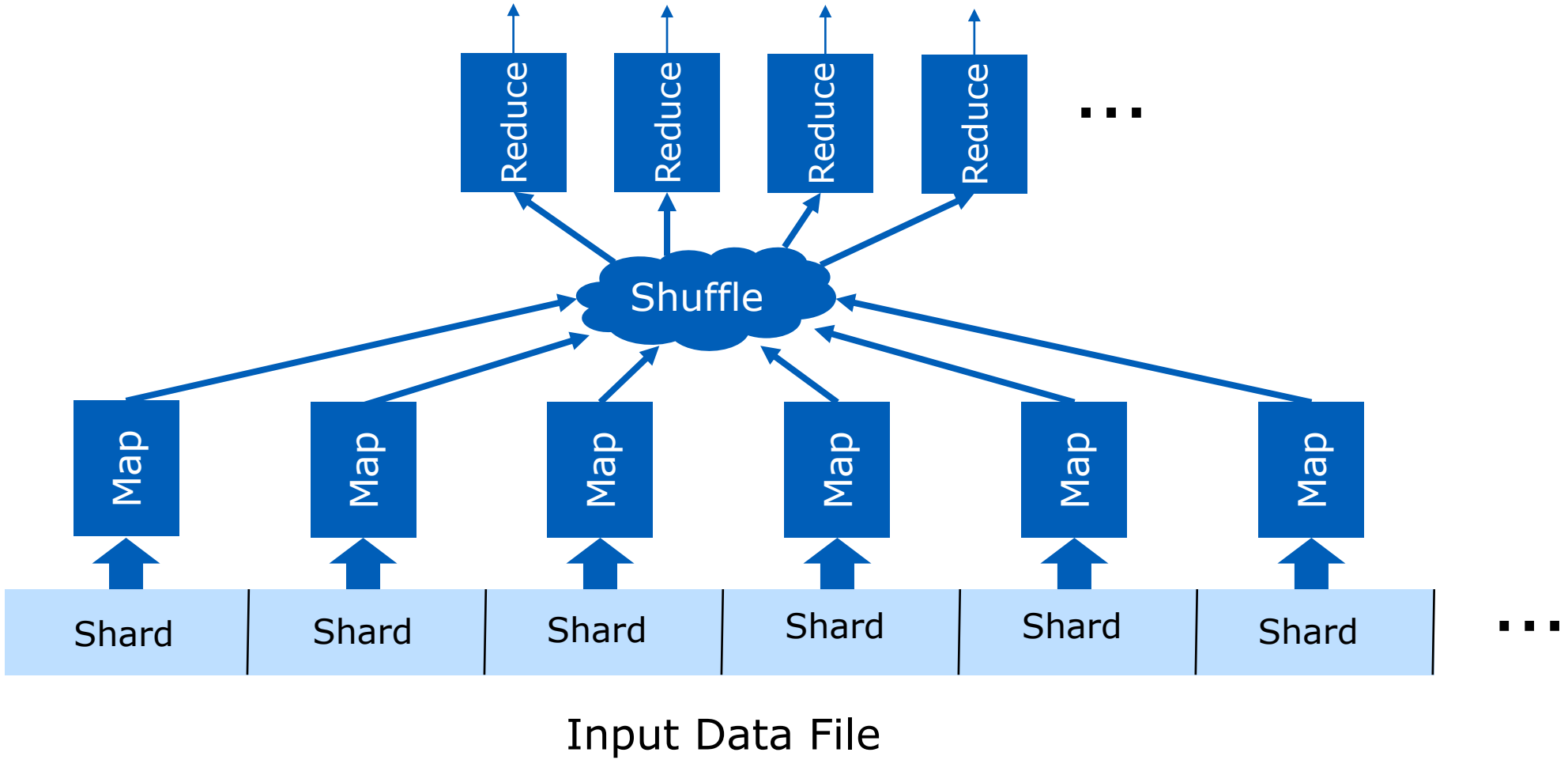
Storage Intelligence in the Network

Architectural Issues to resolve

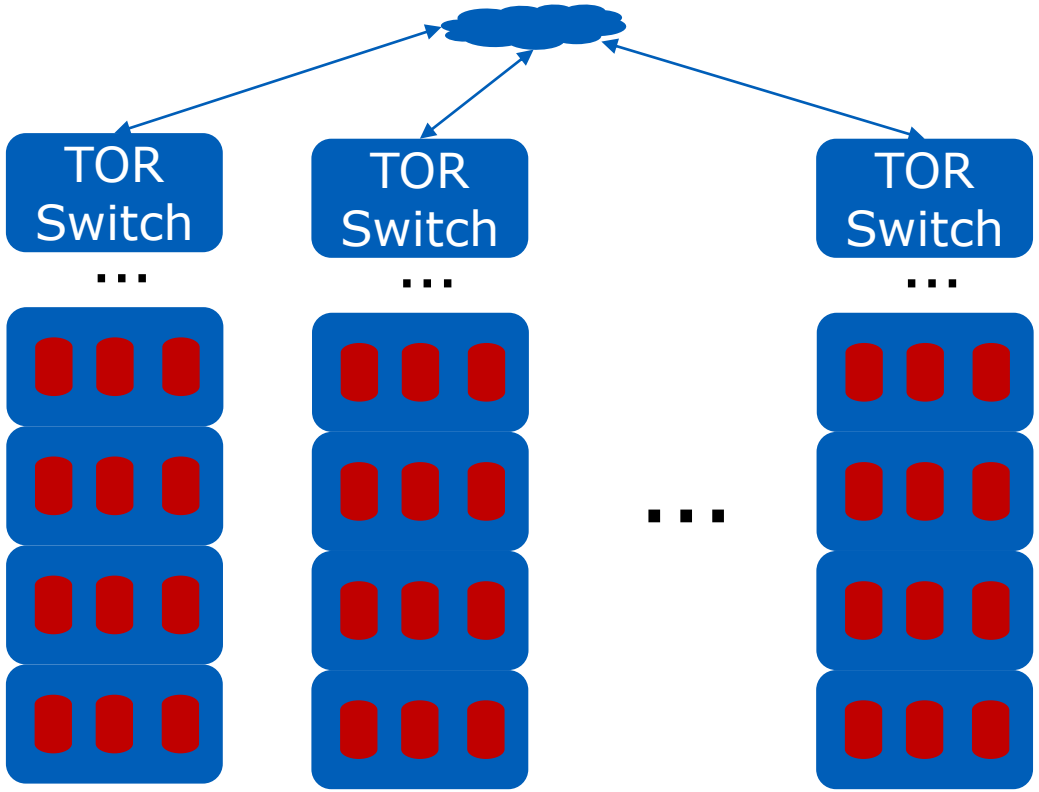
- Protocols
  - TCP, UDP, SCTP
- Error Handling
  - Failure reporting
  - Recovery responsibility
- Backward Compatibility
  - iSCSI Proxy
- *Watch this space!*

Innovating in Compute and Storage

MapReduce, the conceptual cartoon version



Typical MapReduce Cluster



- Hadoop file system distributes and replicates data for availability and durability
- Large Chunks (64MB+) are used to allow HDD to operate sequentially

Innovating with Compute and Storage

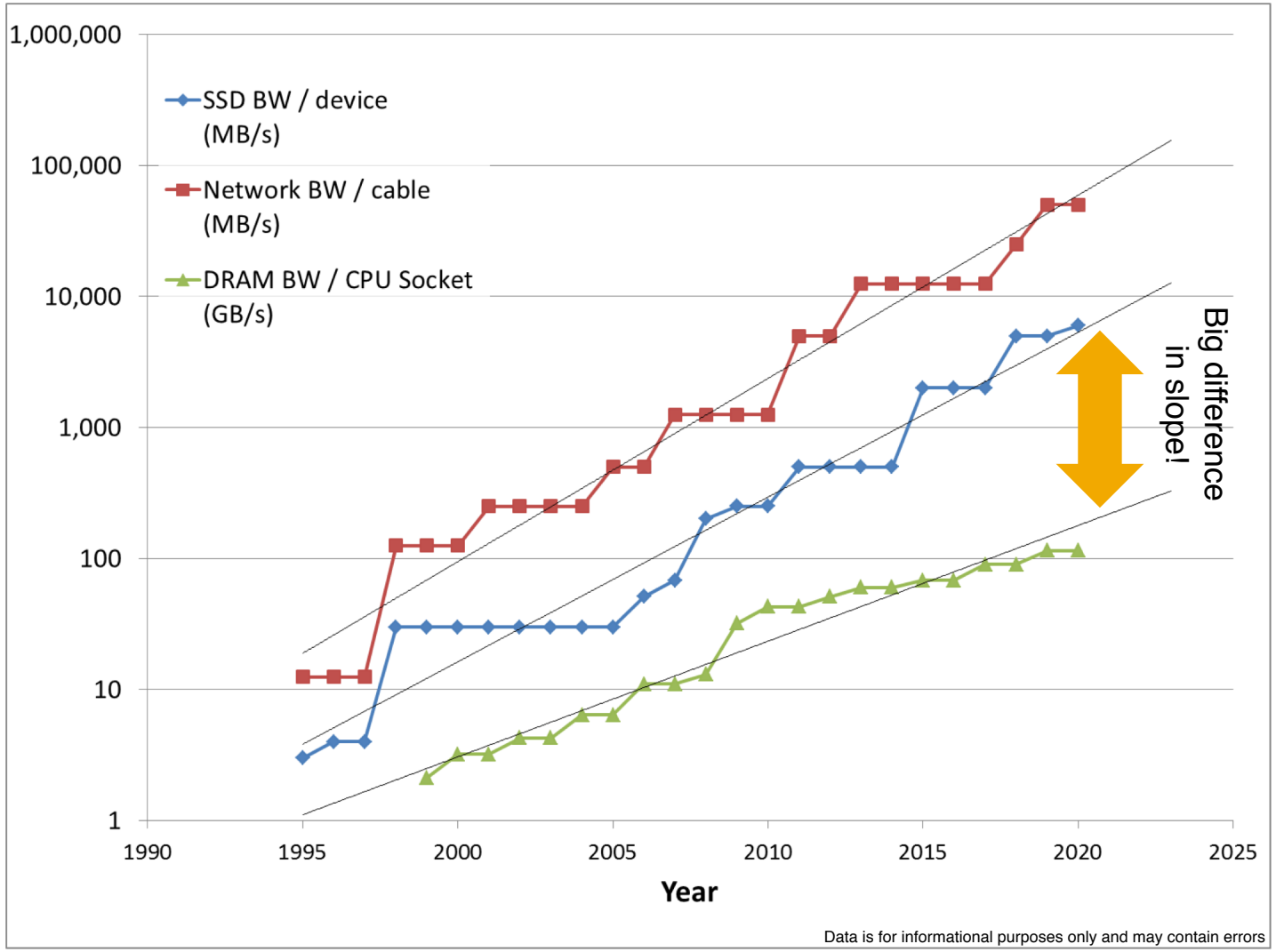
MapReduce

- A typical balanced MapReduce node in the HDD era (circa 2008) was
  - 8-32 CPU cores
  - 12-24 HDD
  - 4-16 GB DRAM
  - 1-2 Gbs NIC

Network, Storage and DRAM Trends

Log scale

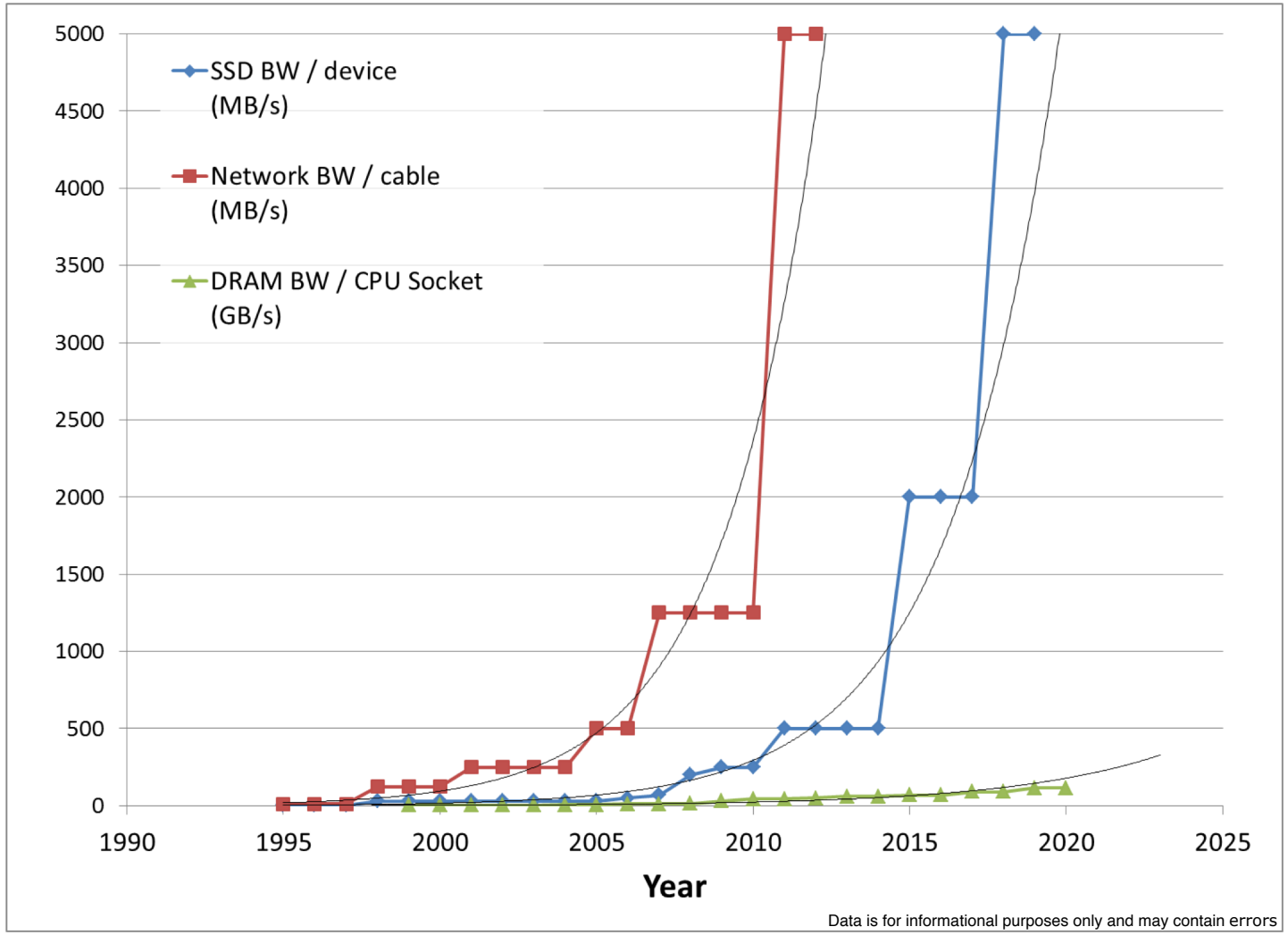
- Use DRAM Bandwidth as a proxy for CPU throughput



Network, Storage and DRAM Trends

Linear scale

- Same data as last slide, but for the Log-impaired
- The ***ratio*** of Network and Storage to CPU throughput is widening very quickly



Innovating with Compute and Storage

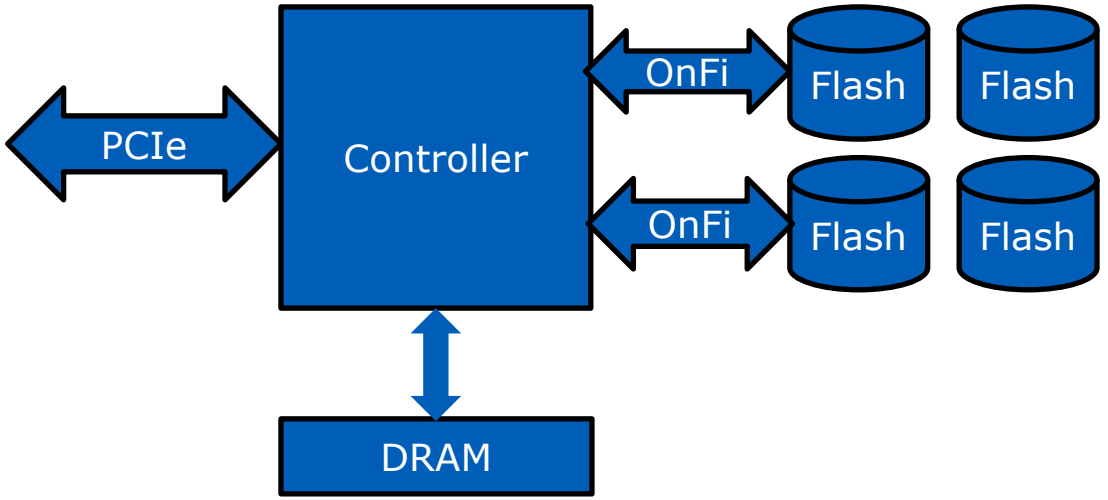
MapReduce

- A decade later the balance is disrupted
  - Networks are up to 25x faster
  - Flash provides 50-200x more BW for same storage or 5-20x more for same cost vs. HDD
  - CPUs and DRAM have lagged, growing only about 8-9x in the same period

- Server CPU performance limited by DRAM BW
- Need DRAM BW that scales with the problem

Injecting Compute into the Storage

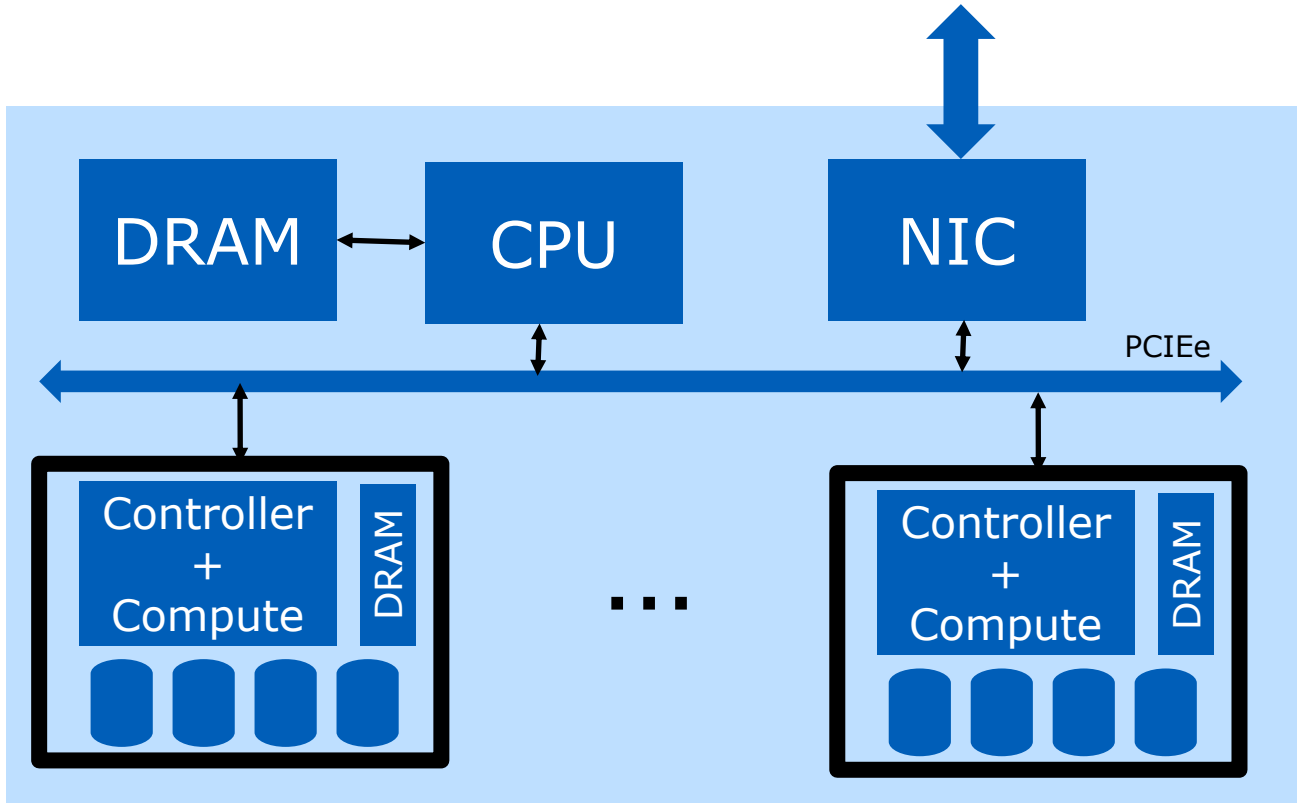
Enterprise SSD, the cartoon version

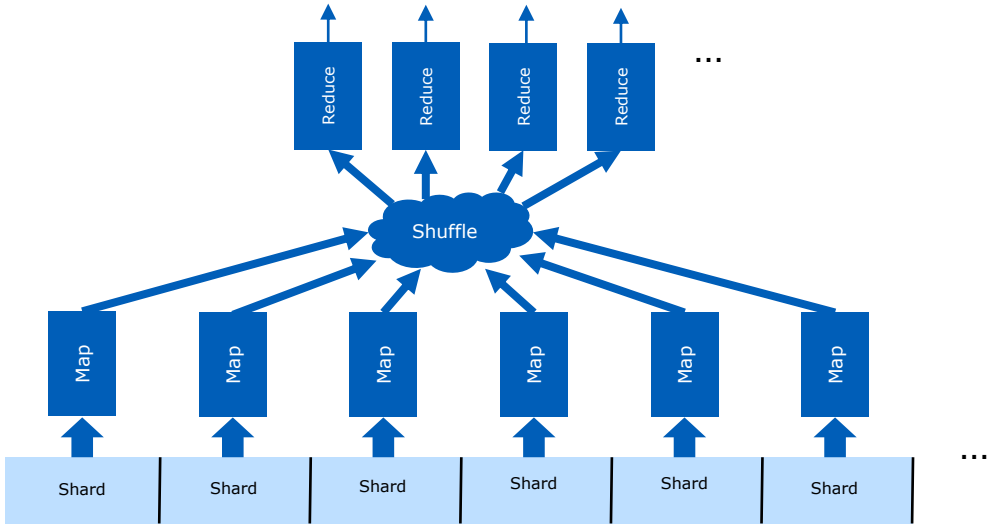


- DRAM bandwidth underutilized by controller
- DRAM capacity sized for 4KB mapping tables (FTL)
- *DRAM just waiting for some compute*

Injecting Compute into the Storage

Rebalancing MapReduce





- Move Map phase into the compute on the storage device
  - Mappers often I/O and scan dominated
  - Substantial reduction in vertical traffic in many applications

Putting Intelligence into Storage

Exposing the Compute

- What's the "right" way to expose the compute?
- Compute as Server
  - Run your favorite operating system
  - Bridge external network across interconnect bus (PCIe)
  - Use all your favorite storage and server management tools
- Compute as Coprocessor (Master/Slave)
  - Minimal overhead resource consumption
  - Avoid the server management headache (updating, etc.)

Putting Intelligence into Storage

Exposing the Storage

- What's the "right" way to expose the Storage?
- Host Managed
  - Run your favorite file system and storage management tools
  - Communicate at the LBA level
  - Complex consistency model
    - Same as a SAN block device
- Device Managed
  - Communicate at the object level (File?)
  - Free to disobey the Posix imperative
  - Potentially the simplest consistency model

Putting Intelligence into Storage

Sweet Spot Applications

- Bandwidth Intensive Applications
- Scan oriented
- Unstructured Data
- Relatively low compute per byte of data

- Real-time analytics on large-scale data
- Classic Big Data batch processing

Summary

- Substantial benefits to liberating compute from the centralized CPU prison
-

Thank You

@opensds_io

OpenSDS mini summit COPENHAGEN