LLNL Deploys Storage-class Memory to Build Highest Performance Storage Array with Two Rack Footprint

Neil Carson

Chief Technology Officer

Fusion-io

Agenda

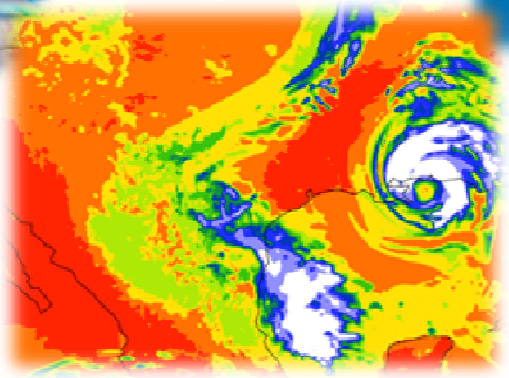
1: About LLNL and Hyperion

2: Applying the architecture in enterprise

Computed and I/O Intensive Applications

Physical Science

Energy



Climate

Unbalanced Systems

- Balanced systems–gigaflop of computation vs gigabytes of memory
- RAM density not keeping up
- Flash memory:
  - 100x capacity density of DRAM

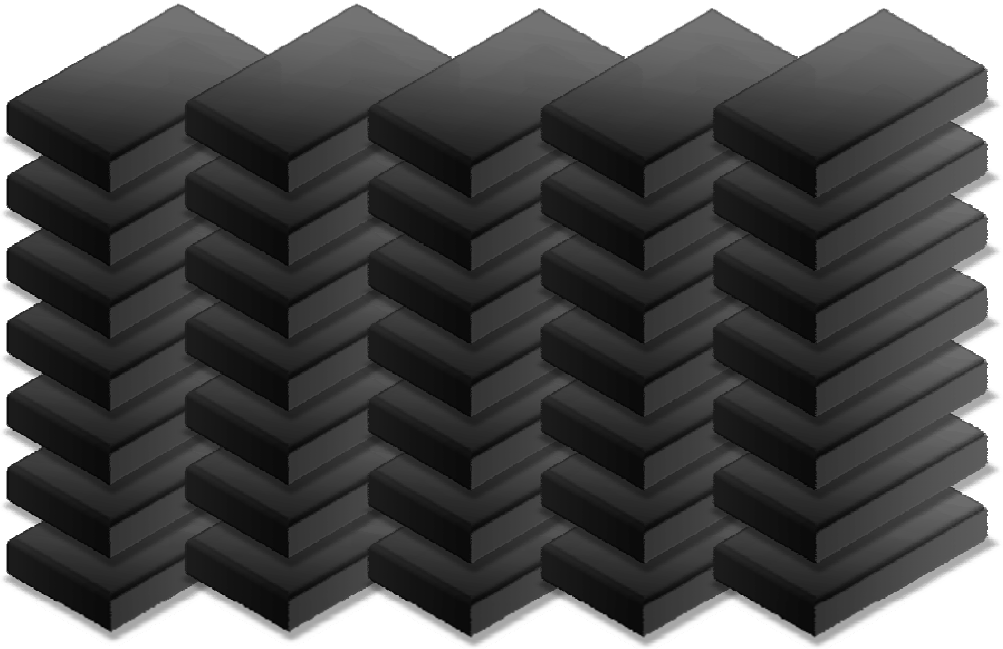
Utilization

- Quarter to a third of system's time is spent idle:
  - Waiting for data sets to load
  - Saving checkpoints
  - Application file writes
- To solve checkpoint problem, need only as much flash as RAM

Work with Lawrence Livermore

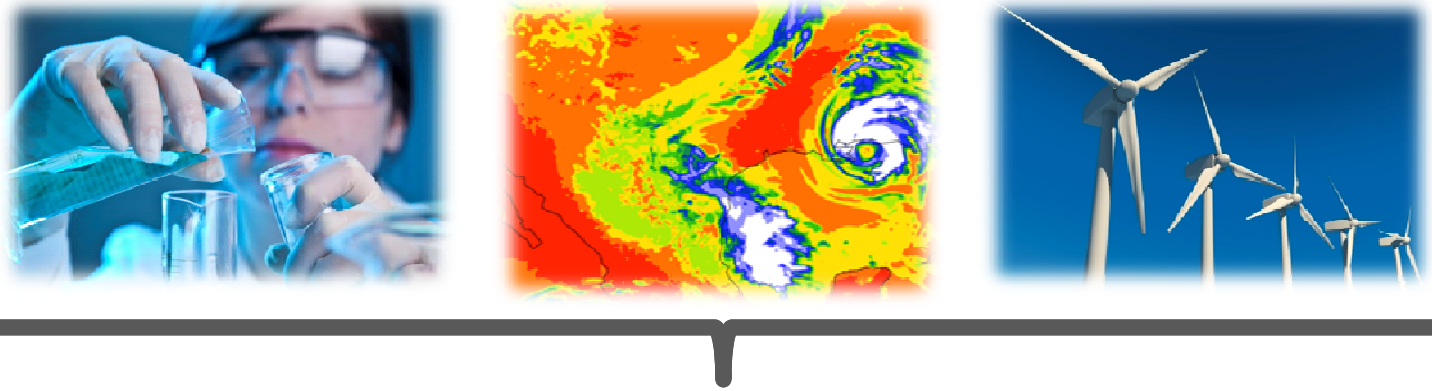
- Hyperion test-bed
- Testing HPC technologies critical to maintaining US nuclear weapons stockpile without underground testing
- >1100 nodes with ~100 teraflop compute capacity
- Over 9TB memory
- Over 100TB flash memory

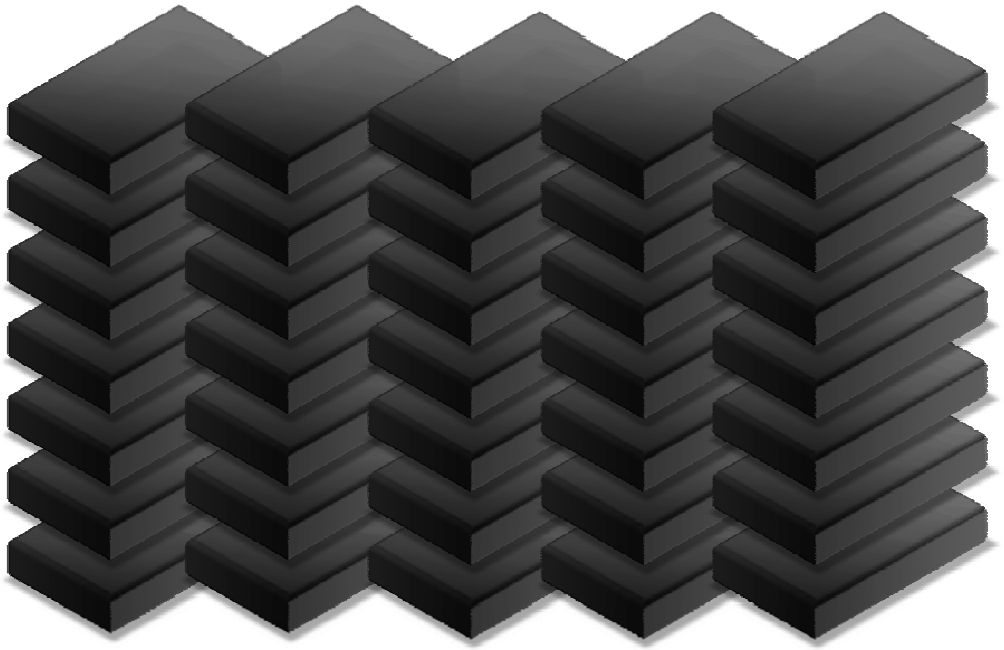
How to Allocate Enough I/O



Need Massive I/O Throughput

Single Storage Cluster Supports Multiple Applications



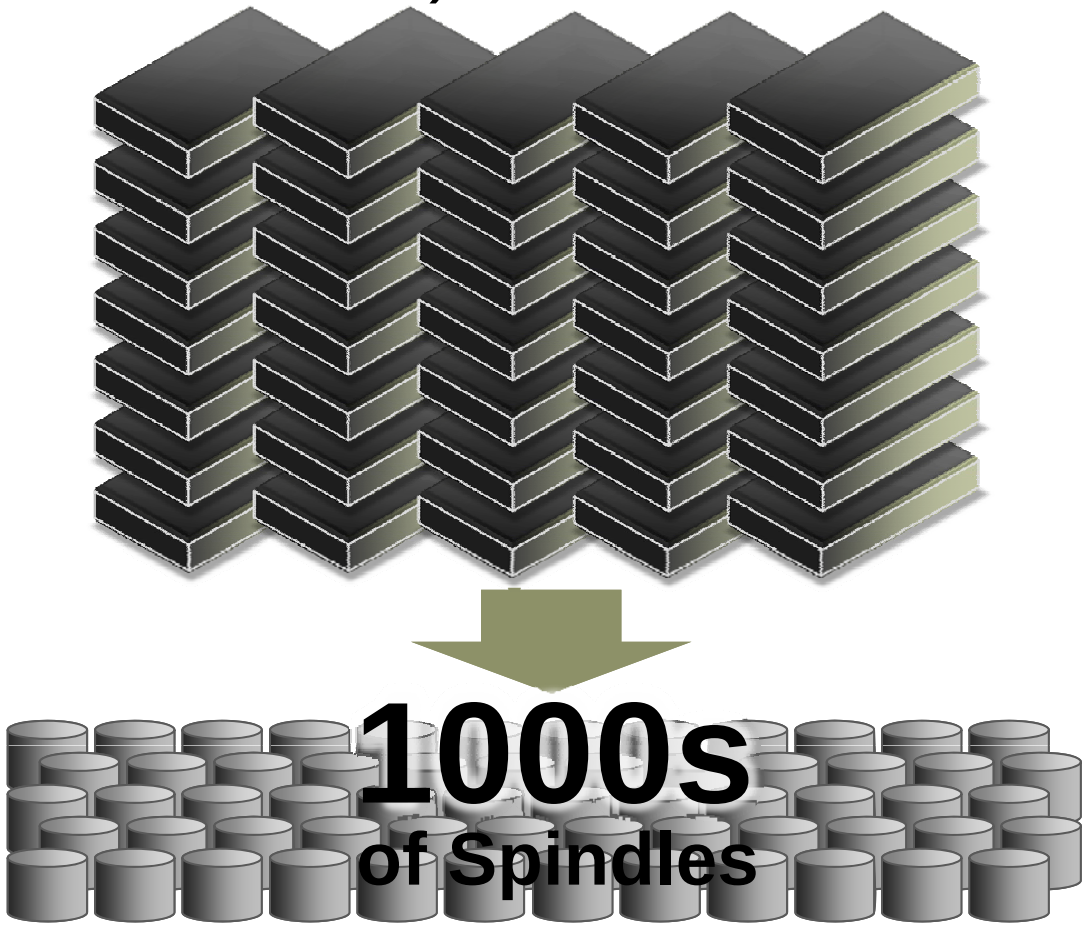


Lustre Intro

- Lustre is a massively parallel distributed file system generally used for cluster computing, providing tens of thousands of nodes with petabytes of storage capacity.
- Servers: MDS, MDT, OSS, OST
- RDMA support

More on Lustre

- Intended to aggregate **LOTS** of spindles (or block device)



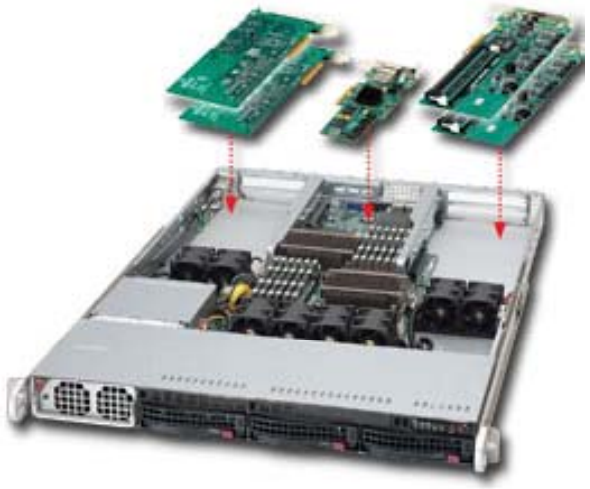
Managing Disk Spindles is Painful

- Disk failures
- Need to swap disks
- Unpredictable
- Power and space consuming
- Wasteful

LLNL Deployment

Supermicro server

- 80 servers
  - 6016XT-TF

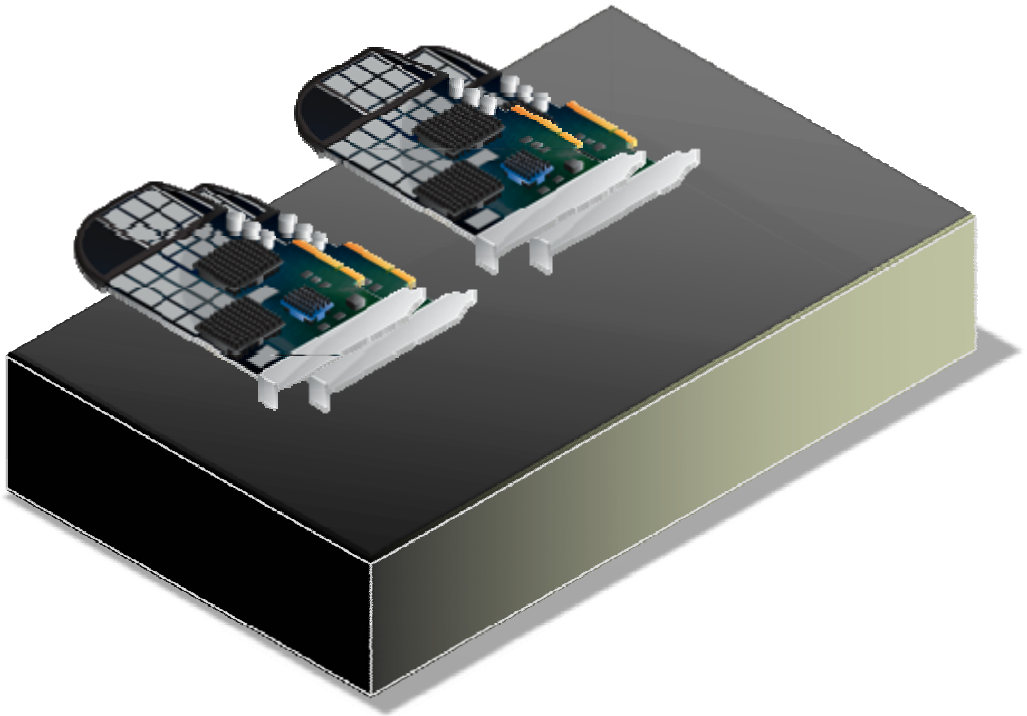


Specs

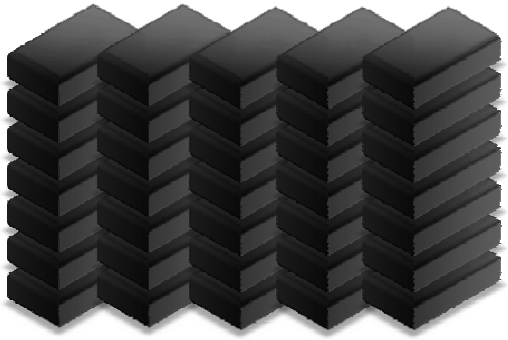
- Processor 5600/5500
- Up to 192GB DDR3
- 4x PCI-E 2.0 (x8) slots
  - (in x16 slots, 2 on each side)
- Integrated IPMI 2.0 with KVM and Dedicated LAN
- 1400W Gold-level High-efficiency power supply

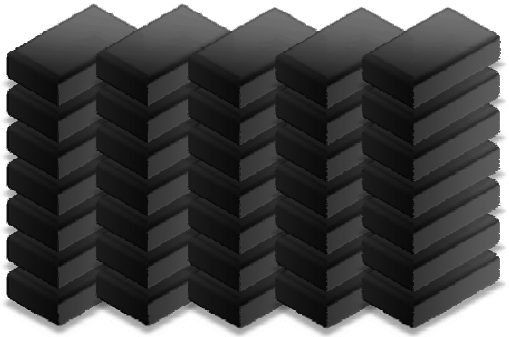
Flash and Network per-U

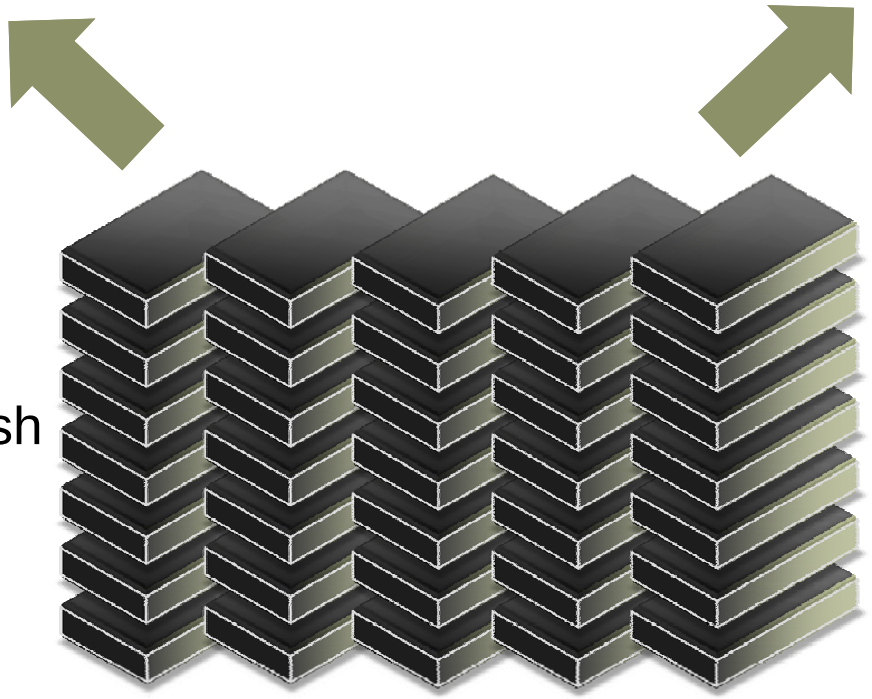
- 8 Flash modules on four cards: 1.3TB
- Two 40 Gigabit Infiniband ports
- Two 10 Gigabit Ethernet ports



Lustre Needs to be Presented Datacenter Wide







Each server has four Flash cards

Massive IOPS—Tiny Footprint

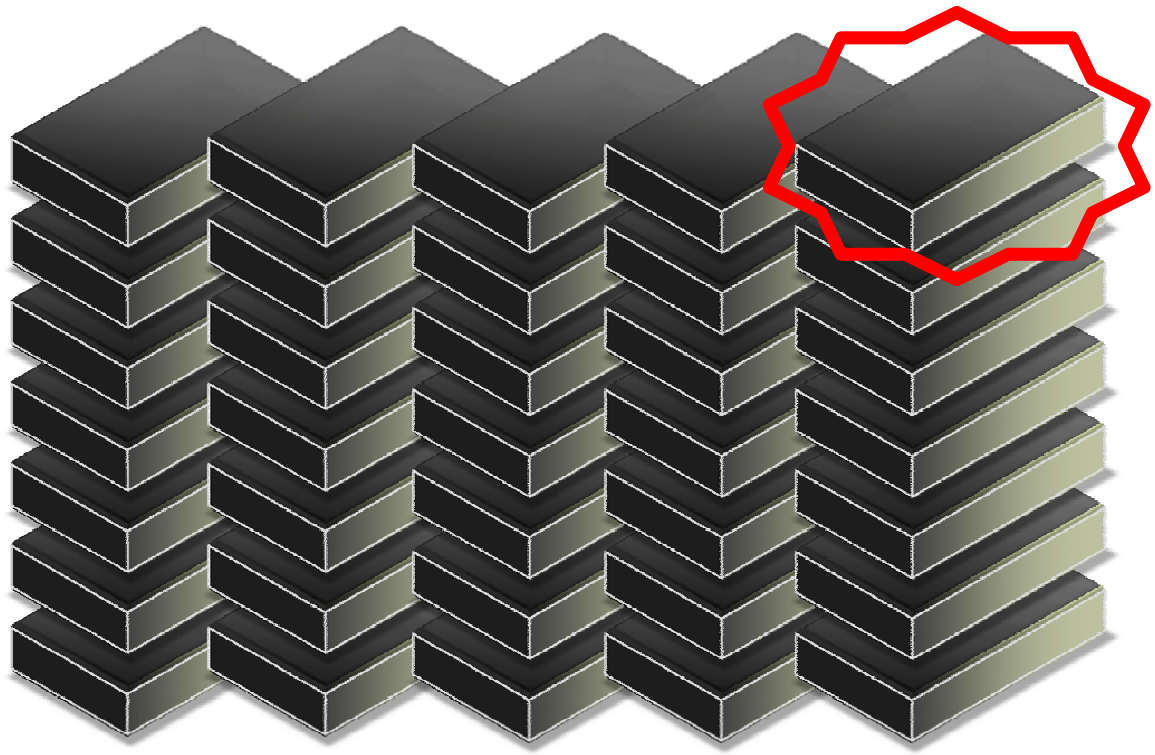
80 servers in only two 42U racks

Results for the Cluster

- With 160 initiators:
  - 40,800,000 sustained IOPS
  - 400GB/sec sustained bandwidth
- Bandwidth expectations exceeded by 80GB/sec

Lustre Metadata

- Lustre metadata still tops out at 40K IOPS
- With improvements, results can be higher

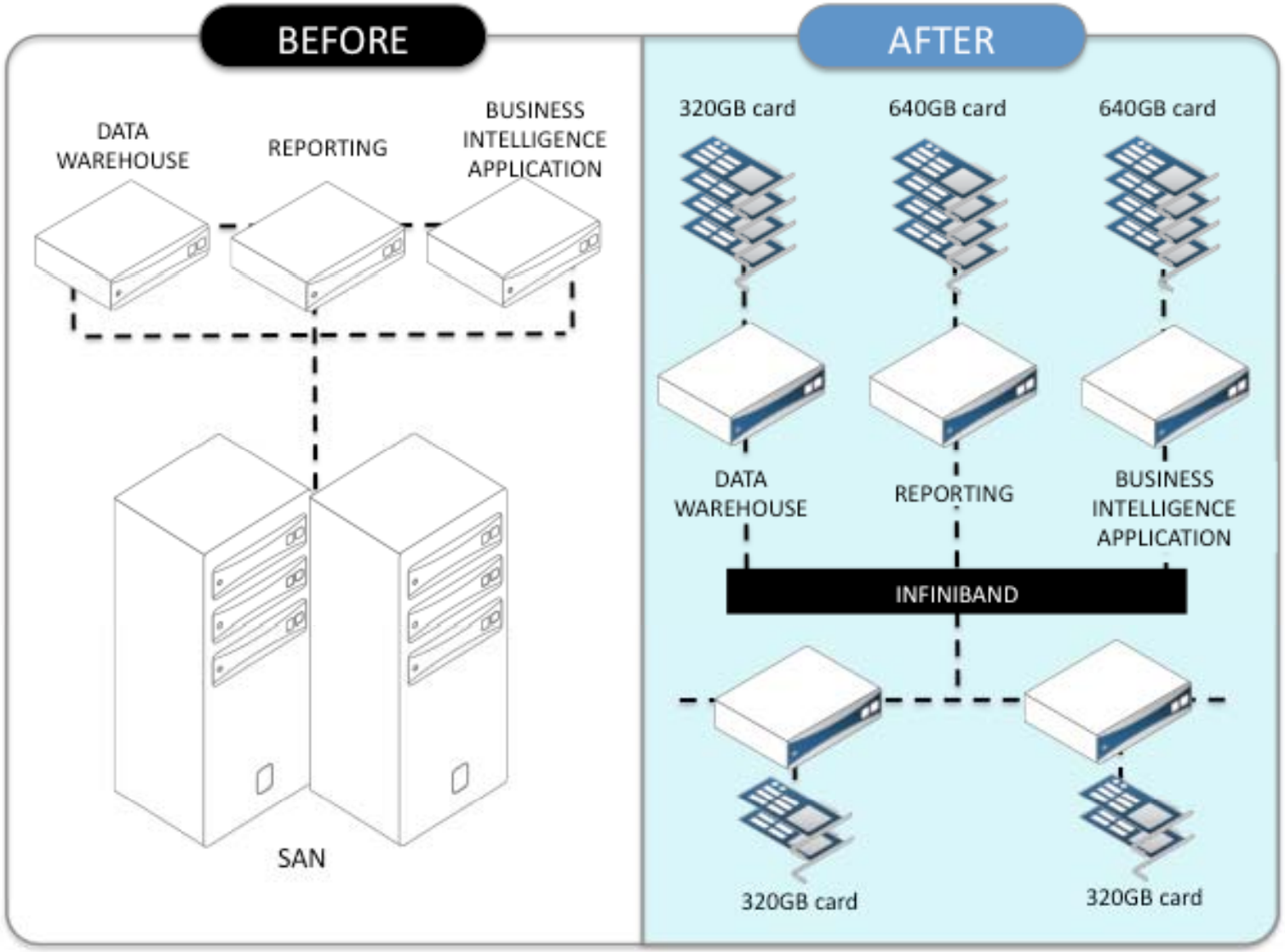


Building Smaller Storage Targets

Aggregating Storage

- Parallel file systems
- Block targets

Smaller-scale Shared Storage

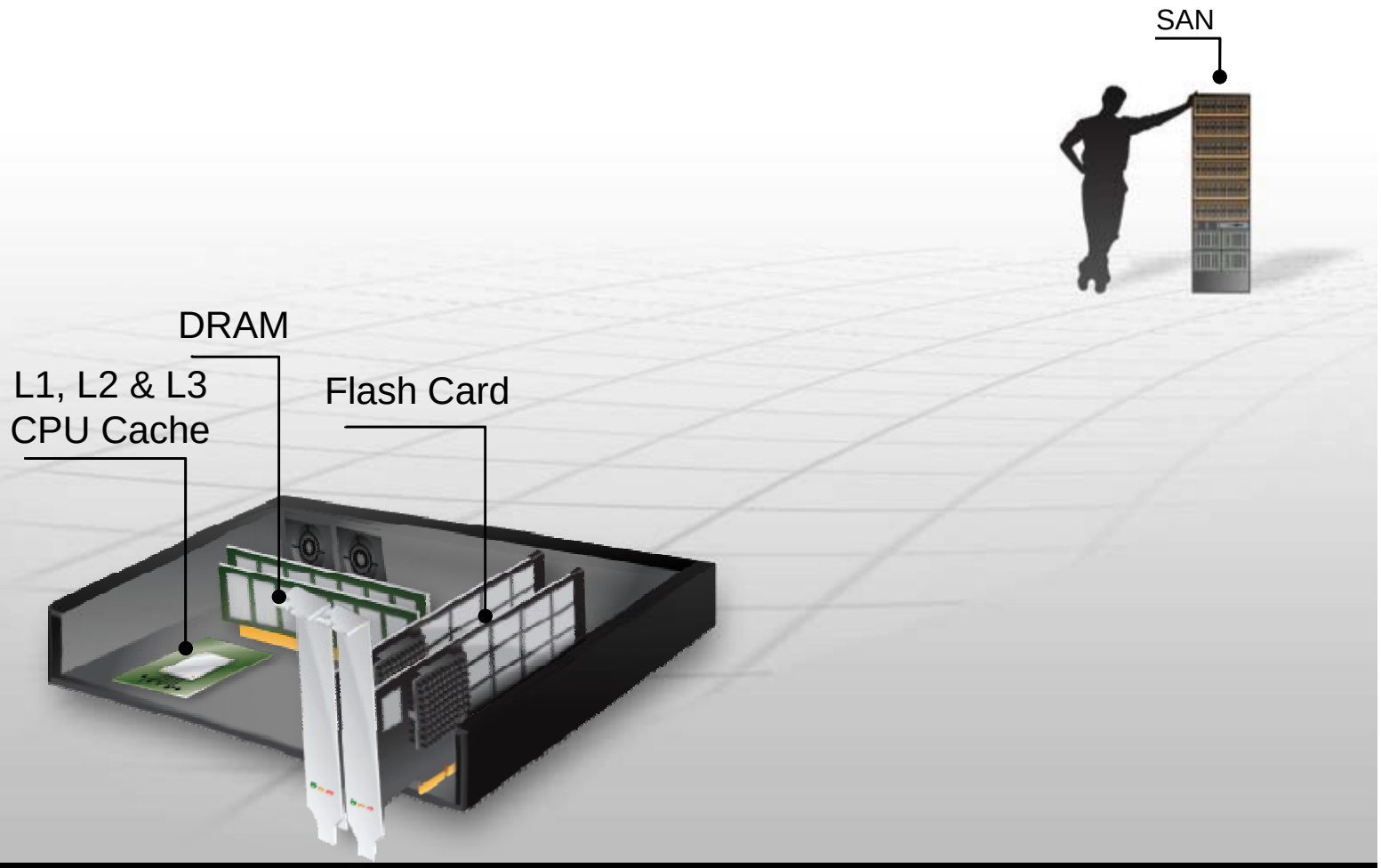


Benefits to the Financial Institution

- “Process more data, faster”
- 75% faster reports, ensuring completion before business hours
- Simultaneous reports
- Nightly back-ups
- ROI under 6 months:
  - Maintenance, power, cooling savings
  - SAN fabric
  - Rack space

It's About Performance

Flash Succeeds Near the CPU



Critical Reliability Elements

- Validation in a variety of server platforms
- Simplified data path
- Sophisticated error correction
- Ability to handle flash errors or flash failures seamlessly
- Field upgradable products
- Monitoring and management

Flash Inside the Server

- A recent customer story: TPC-E-style transactional workload, business-critical data

  1. Two T5440 servers, 2 NetApp 6080s, FlashCache module, 12 disc shelves: Insufficient performance

  2. Swapped 4 disc shelves to SSDs: Better but insufficient performance

  3. Tried a ½ rack of Exadata: Better still, but insufficient, performance

  4. Two DL580 G7 servers, 10 cards per server:
     - ~13TB Flash, millions of IOPS per server
     - 4x performance at a fraction of the cost


THANK YOU

COMPUTERWORLD SNIA

SNW

Driving Innovation Through the Information Infrastructure

SPRING 2011