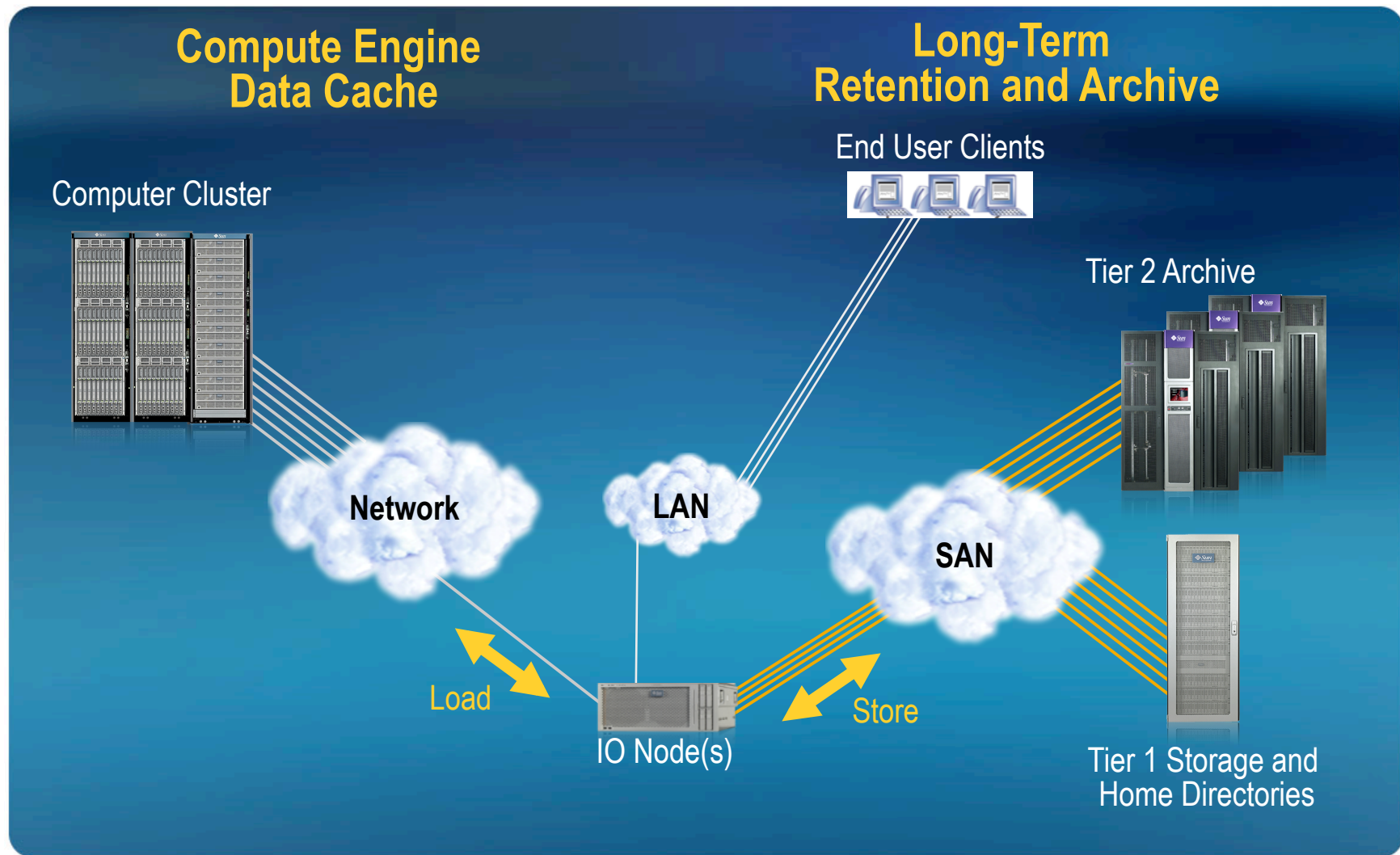# Managing Mountains of Data in Large Scale HPC Systems

Roger Goff

HPC Architect

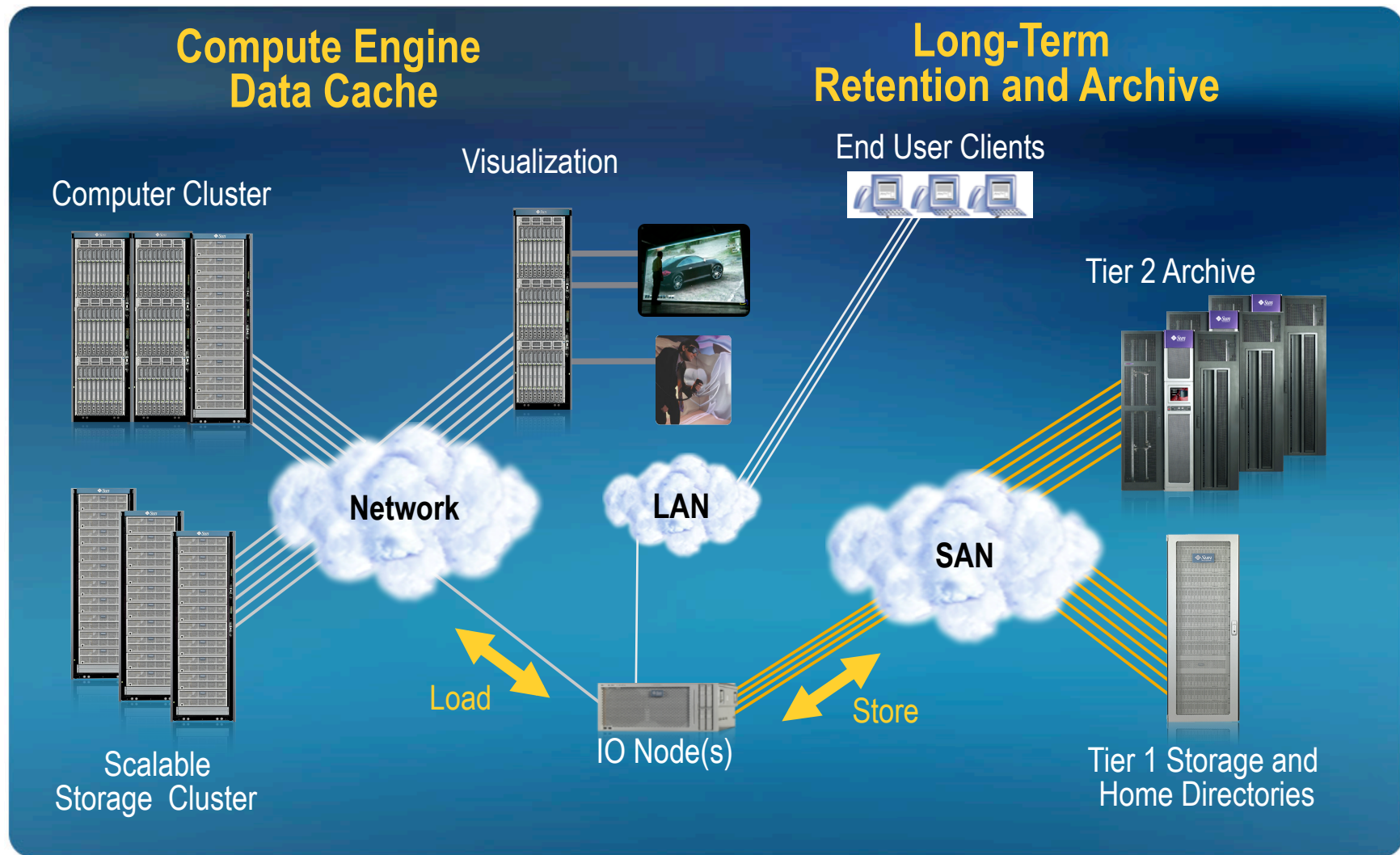Sun Microsystems

HPC Storage Ecosystem
NFS + SAN = Small to medium systems or small data

HPC Storage Ecosystem
Lustre + NFS + SAN = Large systems or large data

Compute Engine Data Cache

Long-Term Retention and Archive

Computer Cluster

Visualization

End User Clients

Tier 2 Archive

Network

LAN

SAN

Load

Store

IO Node(s)

Scalable Storage Cluster
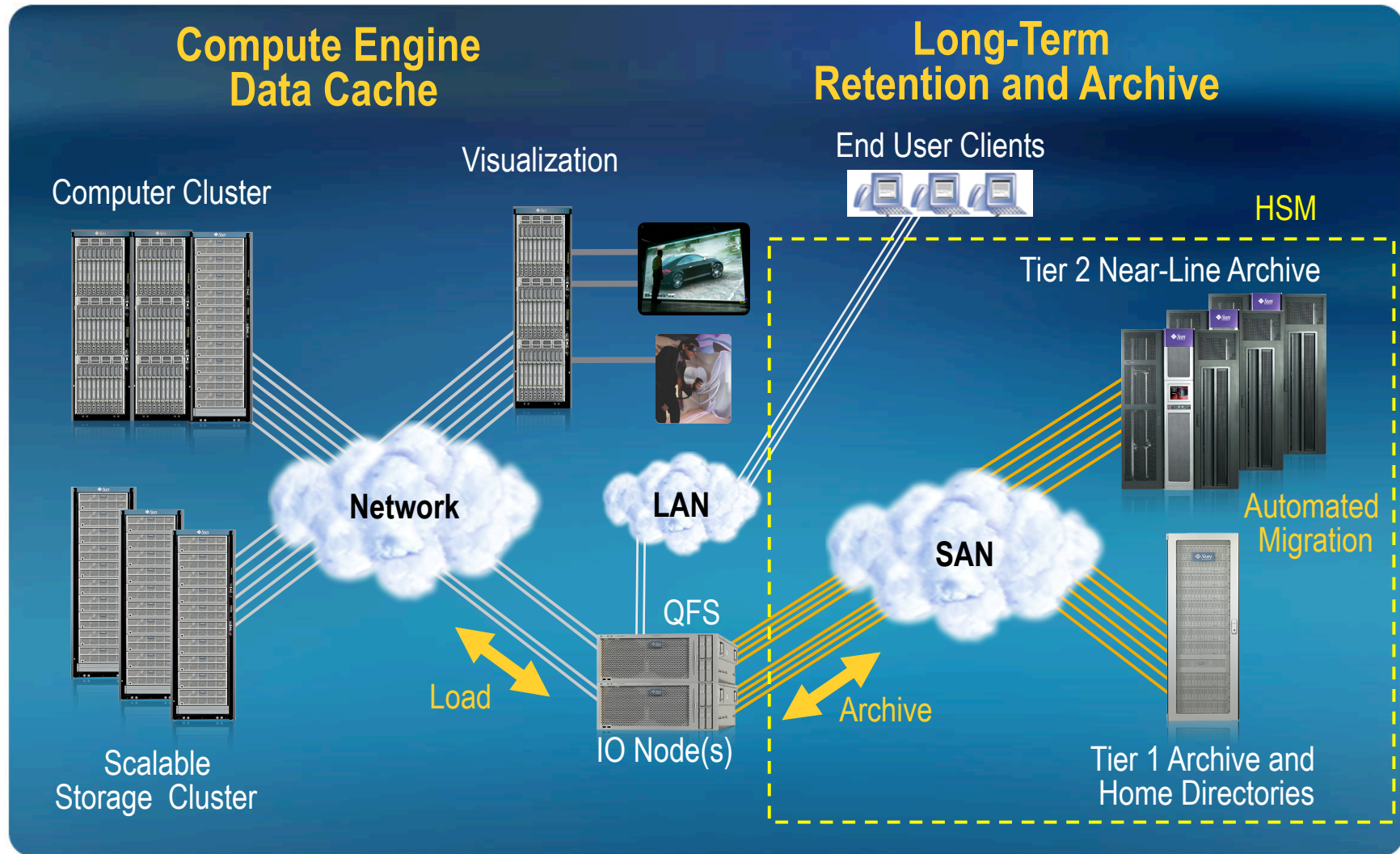
Tier 1 Storage and Home Directories

Hierarchical Storage Management (HSM)
Ex. Sun Storage Archive Manager (SAM) + QFS
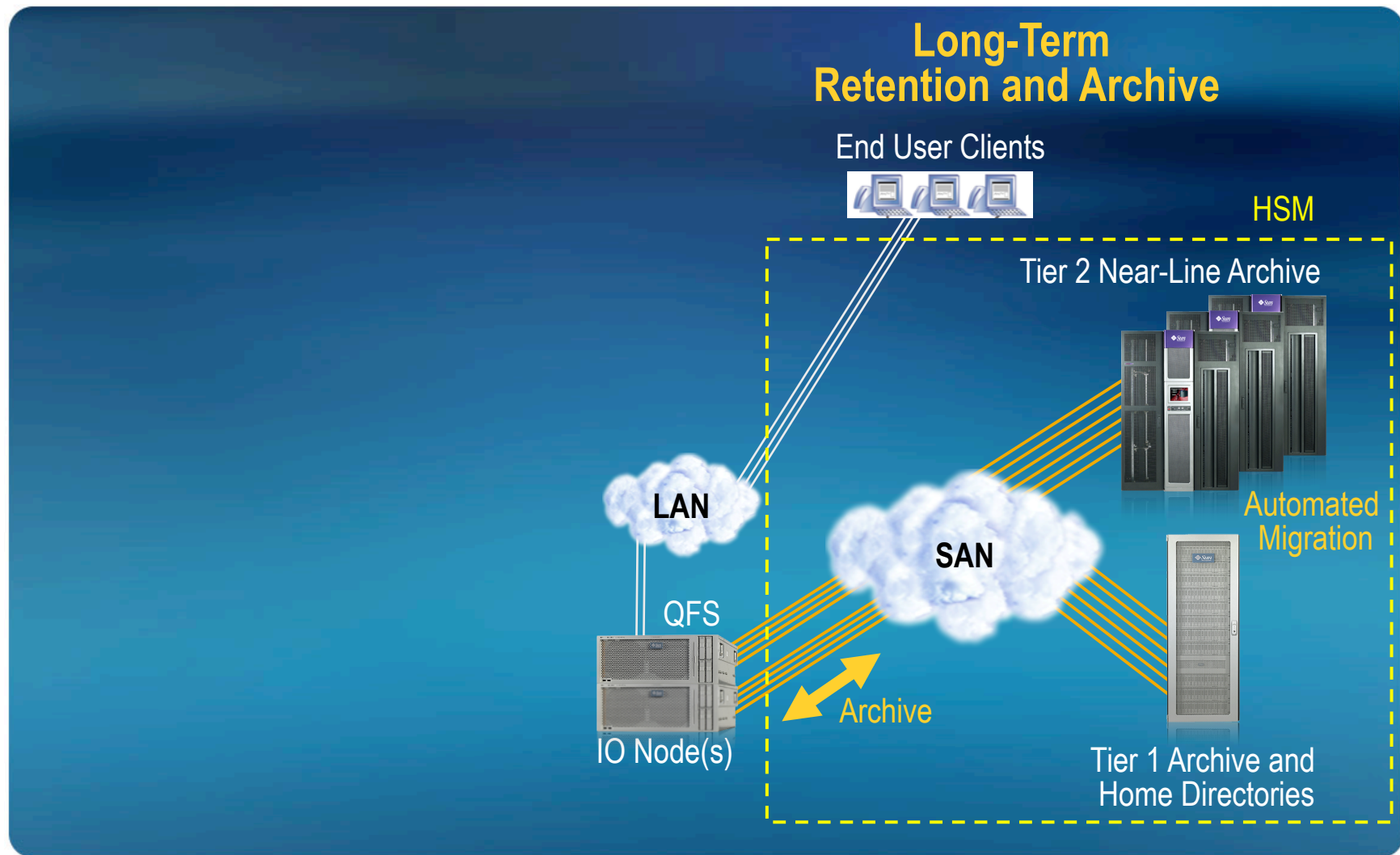
HPC Storage Ecosystem

Lustre + NFS + HSM = Large systems or large data with archiving

Compute Engine Data Cache

Long-Term Retention and Archive

Visualization

End User Clients

Computer Cluster

HSM

Tier 2 Near-Line Archive

Network

LAN

Automated Migration

SAN

QFS

Load

Archive

Scalable Storage Cluster

IO Node(s)

Tier 1 Archive and Home Directories

HPC Storage Ecosystem – In the Future

Lustre + HSM = Large systems or large data with archiving
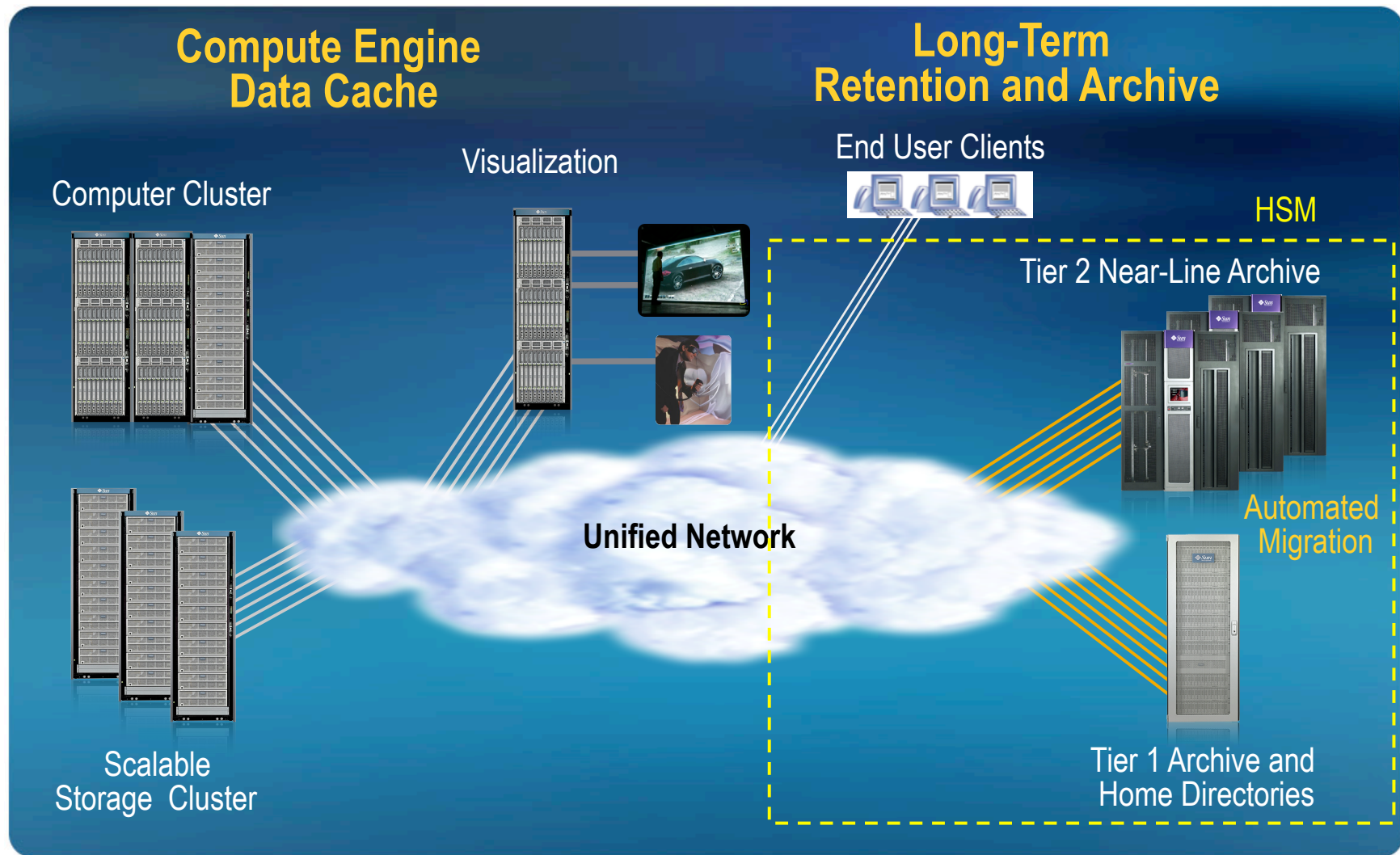
The Changing Storage Landscape

Where to Store Data?
Optimization Trade-Off

Cost Effective Performance

SSDs are 70X more cost effective

- Enterprise HDD
  - > 180 Write IOPS
  - > 320 Read IOPS
  - > 300 GB
  - > ~18W
- $ per IOPS: 2.43

- Enterprise SSD
  - > 7,000 Write IOPS
  - > 35,000 Read IOPS
  - > 32GB
  - > ~3W
- $ per IOPS: 0.04

ZFS Hybrid Pool Example

- 4 Xeon 7350 Processors (16 cores)
- 32GB FB DDR2 ECC DRAM
- OpenSolaris with ZFS

Configuration A:

(7) 146GB 10,000 RPM SAS Drives

Configuration B:

(1) 32G SSD ZIL Device

(1) 80G SSD Cache Device

(5) 400GB 4200 RPM SATA Drives

ZFS Hybrid Pool Example

Legend: Configuration A (yellow), Configuration B (blue)

Read IOPs: 3.2x
Write IOPs: 11%
Cost: 4%
Storage Power (Watts): 4.9x
Raw Capacity (TB): 2x
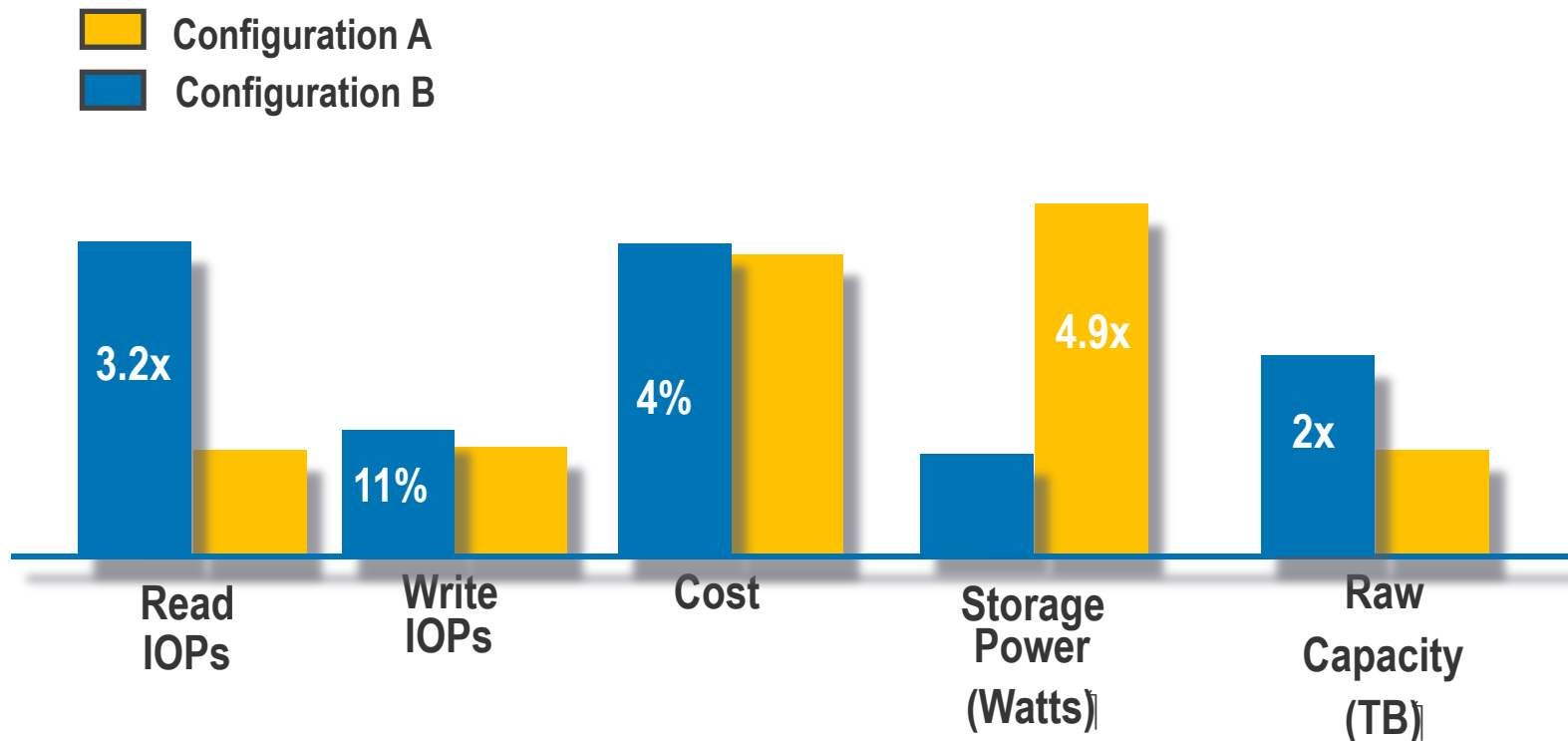
- If NVRAM were used, Hybrid wins on cost, too
- For large configs (e.g. 48T–750T+) cost is entirely amortized

References

- Implementing the Lustre File System with Sun Storage

  > http://wikis.sun.com/display/BluePrints/Implementing+the+Lustre+File+System+with+Sun+Storage

- Sun Storage and Archive Solution for HPC

  > http://wikis.sun.com/display/BluePrints/Sun+Storage+and+Archive+Solution+for+HPC

Lustre Architecture

MDS disk storage containing
Metadata Targets (MDT)

OSS storage with
Object Storage Targets (OST)

Object Storage Servers (OSS)
1-1000's

MDS 1
(active)

MDS 2
(standby)

OSS 1

Commodity Storage

Elan

Myrinet

InfiniBand

Lustre Clients
1 - 100,000

Simultaneous
support of multiple
network types

OSS 2

OSS 3

Shared storage
enables failover OSS

OSS 4

Router

OSS 5

OSS 6

GigE

OSS 7

Enterprise-Class
Storage Arrays and
SAN Fabric

= failover

Sun
microsystems