DMF & Tiering Update

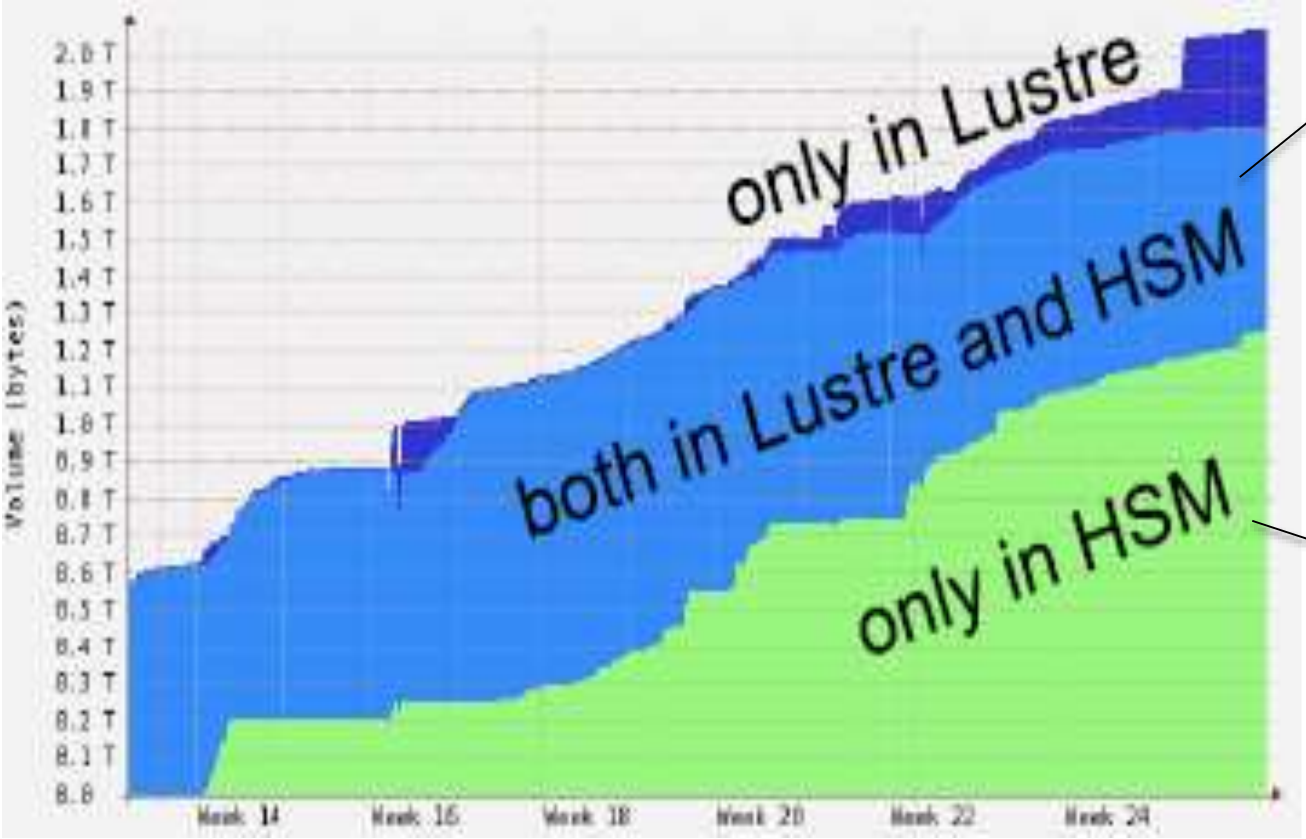
Kirill Malkin
Director of Storage Engineering
September 2015

What's New in DMF?

- Data Migration Facility (DMF)
  - Data management solution for HPC/HPDA
  - Over 20 years of data management
  - High performance parallel data movement
  - 0.5EB of storage, some sites over 100PB
  - Potential to grow by a magnitude over next 2 years
- DMF for Lustre
  - Transparent integration with Lustre HSM
  - Eliminates intermediate copy operation
- Tiered Data on UV
  - Data tiering solution for Flash on UV
  - High-density cold storage on disks

Seamless Tiered Data Management

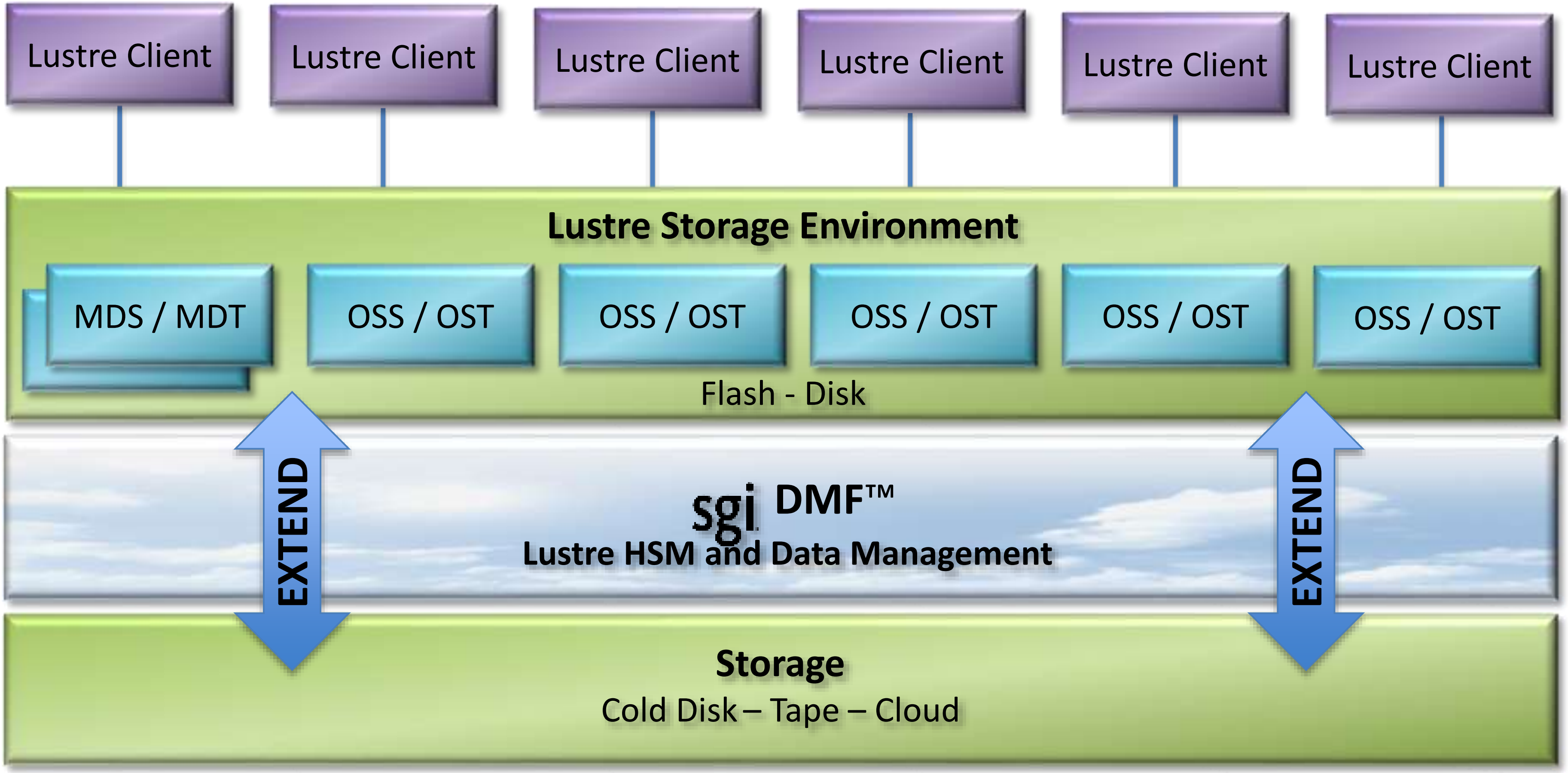
The Result: Infinitely Scalable Lustre



The most recent and active data is “live” in Lustre and mirrored to the HSM, but ALL DATA APPEARS ONLINE to users

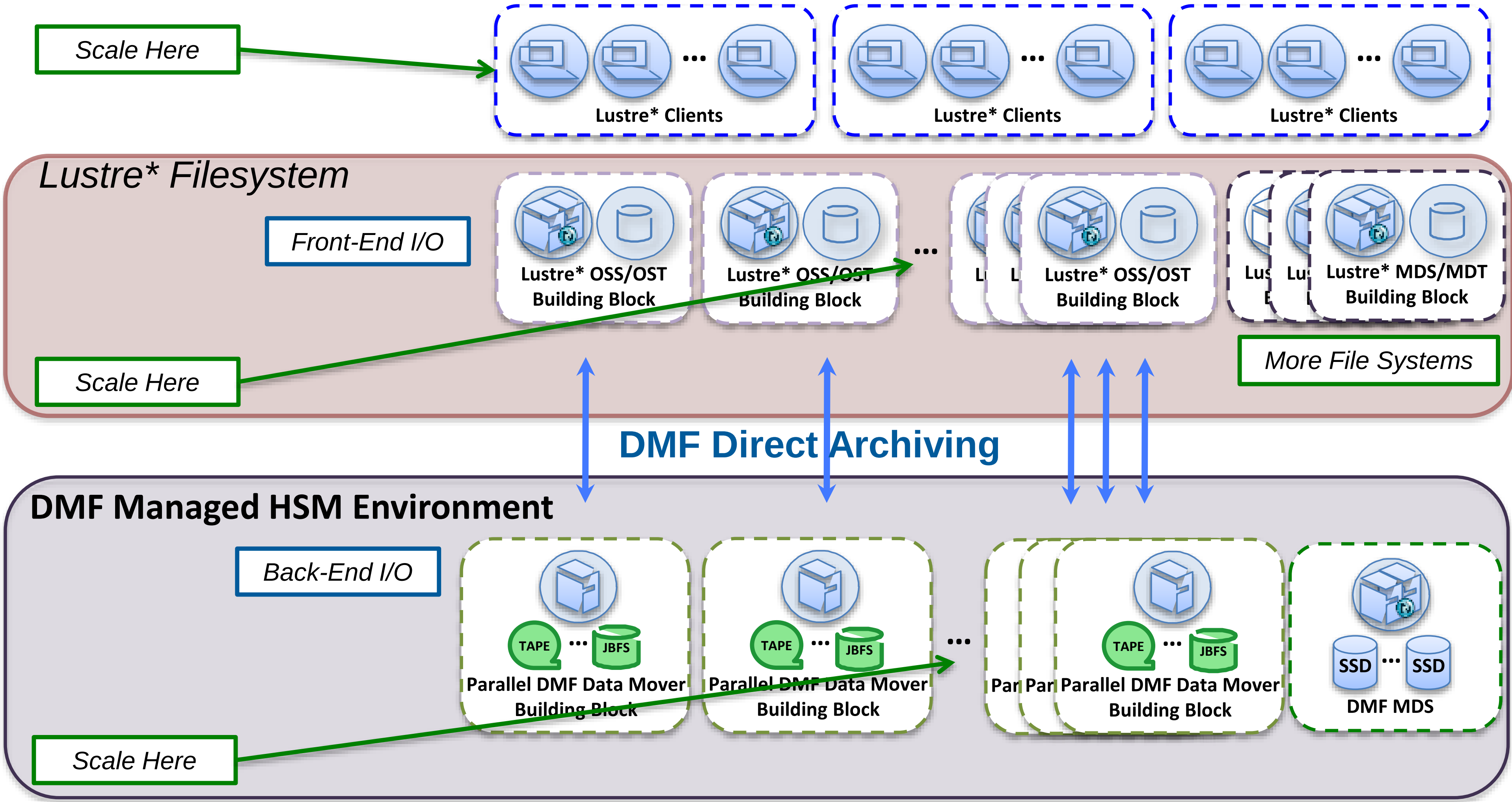
“Overflow” data is stored and protected within DMF on various cost-correct media types

Seamless Tiered Data Management



Scalability without compromise

Capacity, Performance & Reliability



* = Some names and brands may be claimed as the property of others

UV 300 & Flash: Converged Compute & Storage Platform

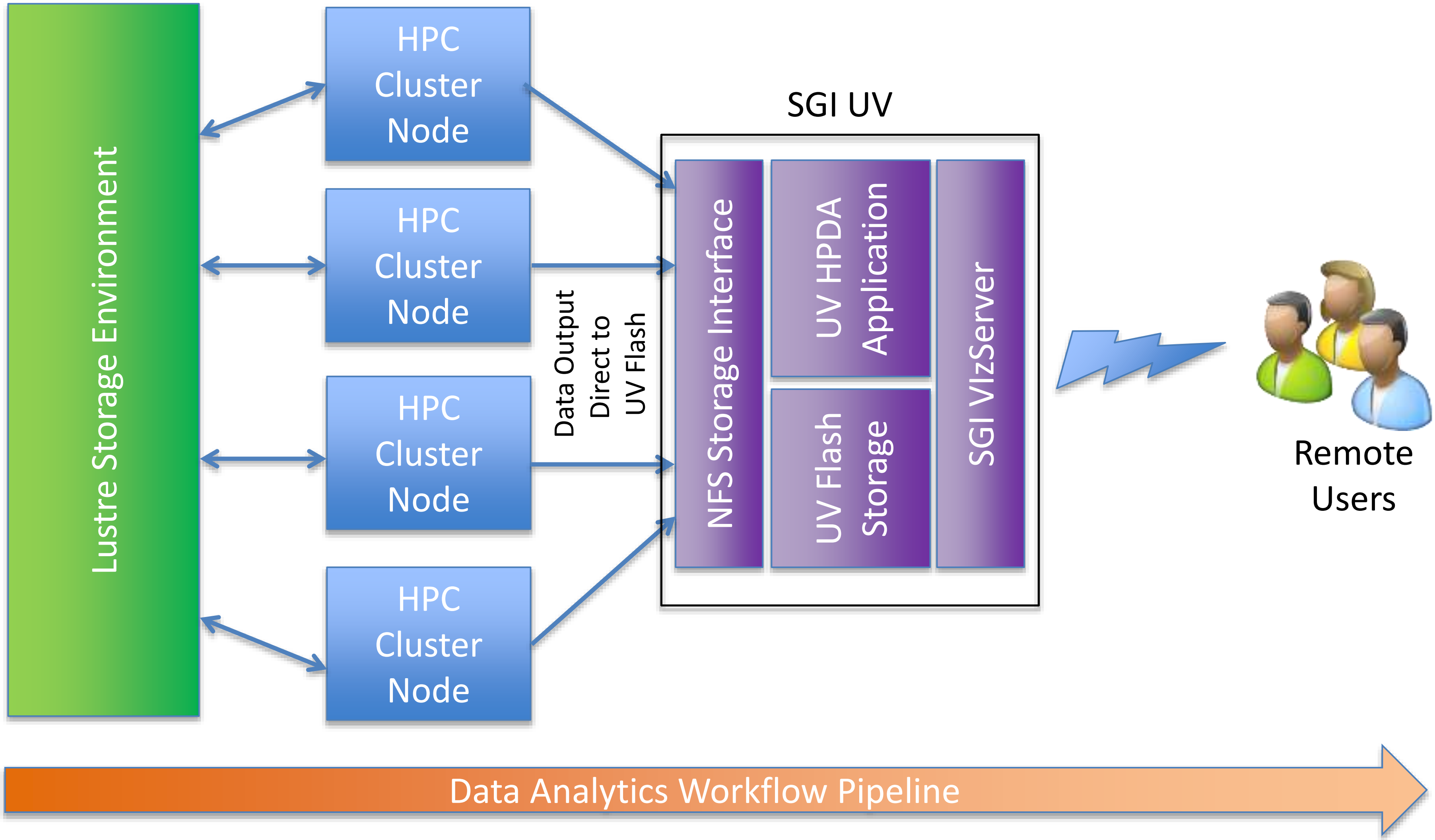
- Single scale-up SMP system delivers
  - Compute:
    - Up to 480 Intel Xeon™ cores (32 sockets)
  - Memory:
    - Up to 24TB DRAM
  - Non-volatile Memory:
    - Up to 128TB FLASH
    - Up to 30 Million IOPs
    - Up to 200GB/s
- All in one rack
- No switches, no arrays, no problems

Intel P3700
400K IOPS
2.4GB/s

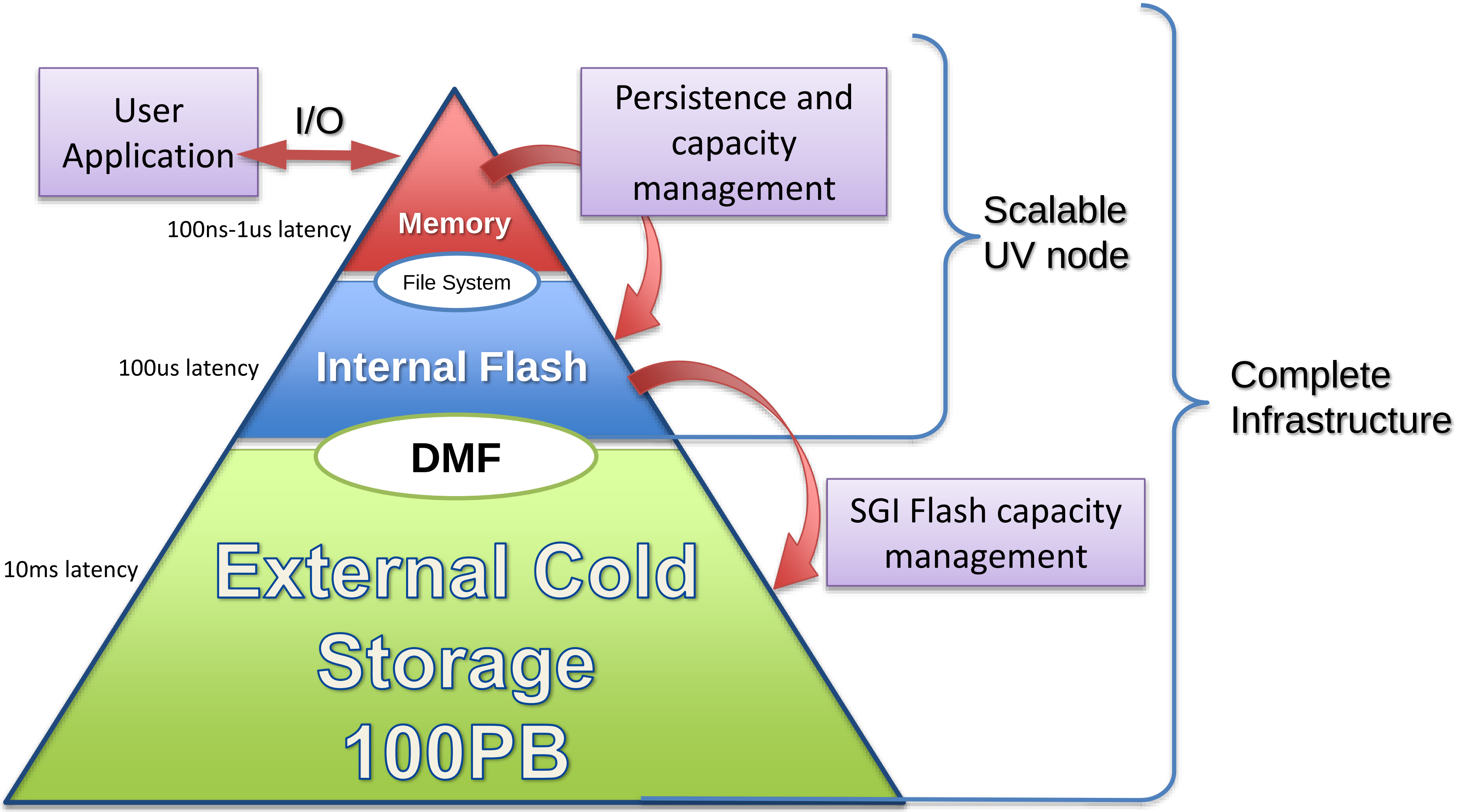
Optimized Code Examples

- Direct, parallel I/O
  - Leveraging multi-queue capability of NVMe
  - Deploys asynchronous I/O & threads
  - Moving data in parallel on multiple sockets
- uvcp
  - Replaces standard cp command
  - Up to 50x improvement
  - 25 GB/s copy with 32 NVMe cards
    - 50GB/s combined
- uvmem
  - Quick loading of files into memory
  - 35-40 GB/s with 32 NVMe cards

Customer Deployment Example



UV Tiered Data Concept



UV Tiered Data Components

Phase 1

Compute:
- 32 Sockets Intel Xeon™

Memory:
- 24TB DRAM

Non-volatile Memory:
- 128TB Flash NVMe
- 30 Million IOPS
- 200 GB/s

UV300 with PCIe Flash

High performance at 1/5 array cost

Phase 2

Transparent Data Extension:
Great performance and cost

Leverages DMF Software:
Hybrid solution

Cold Storage on Disk

Infinite capacity at cost of media

UV Tiered Data Solutions

- Data intensive HPC/HPDA applications running on UV
- High-performance Lustre OSS/OST with DMF for Lustre servicing compute cluster
- Alternative fabric / namespace server, such as NFS over RDMA
- Distributed databases such as Apache Spark running on UV
- Application virtualization on UV (OpenStack, VDI)
  - HPC customers increasingly request OpenStack
- Databases (PostgreSQL) on UV

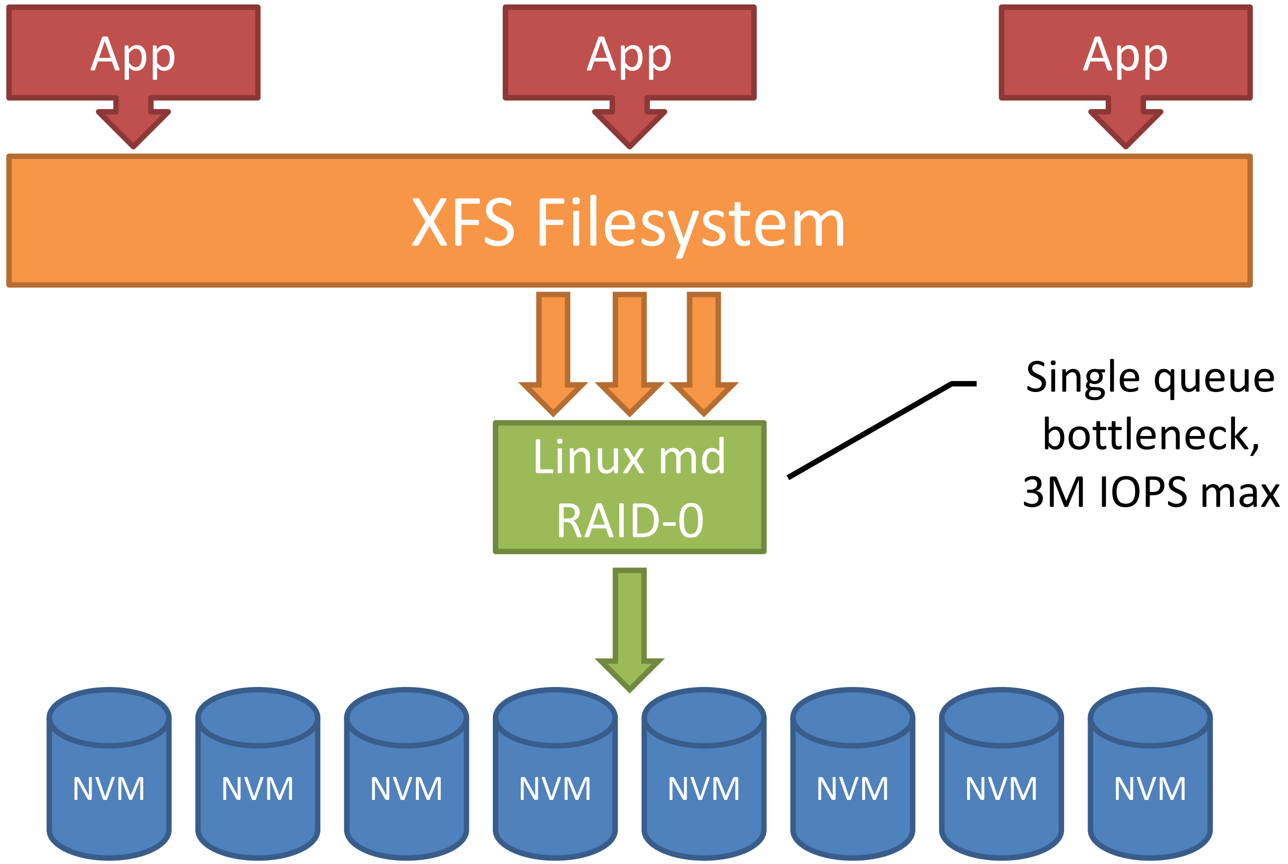
Data Management for HPC

Future Direction

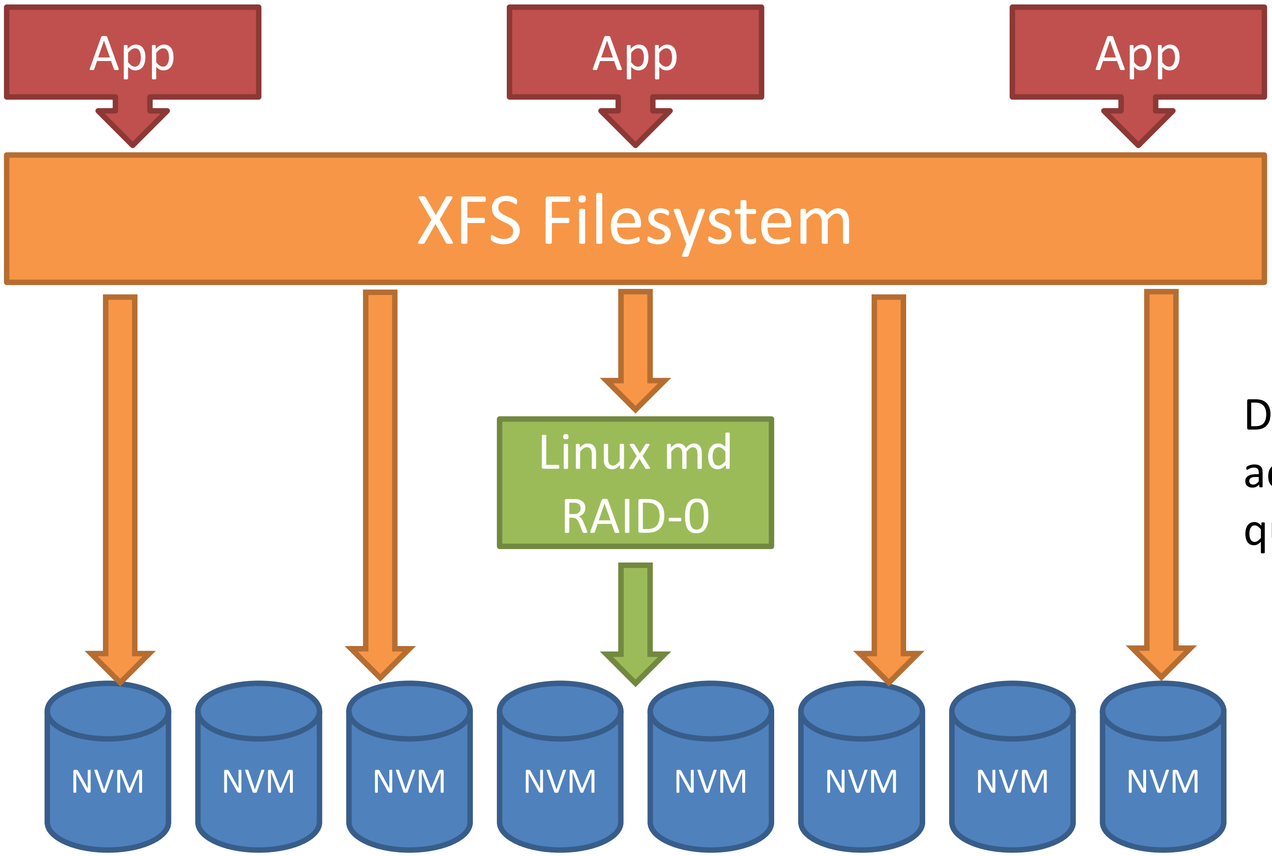
sgi

Performance – Standard XFS

3M read IOPS max



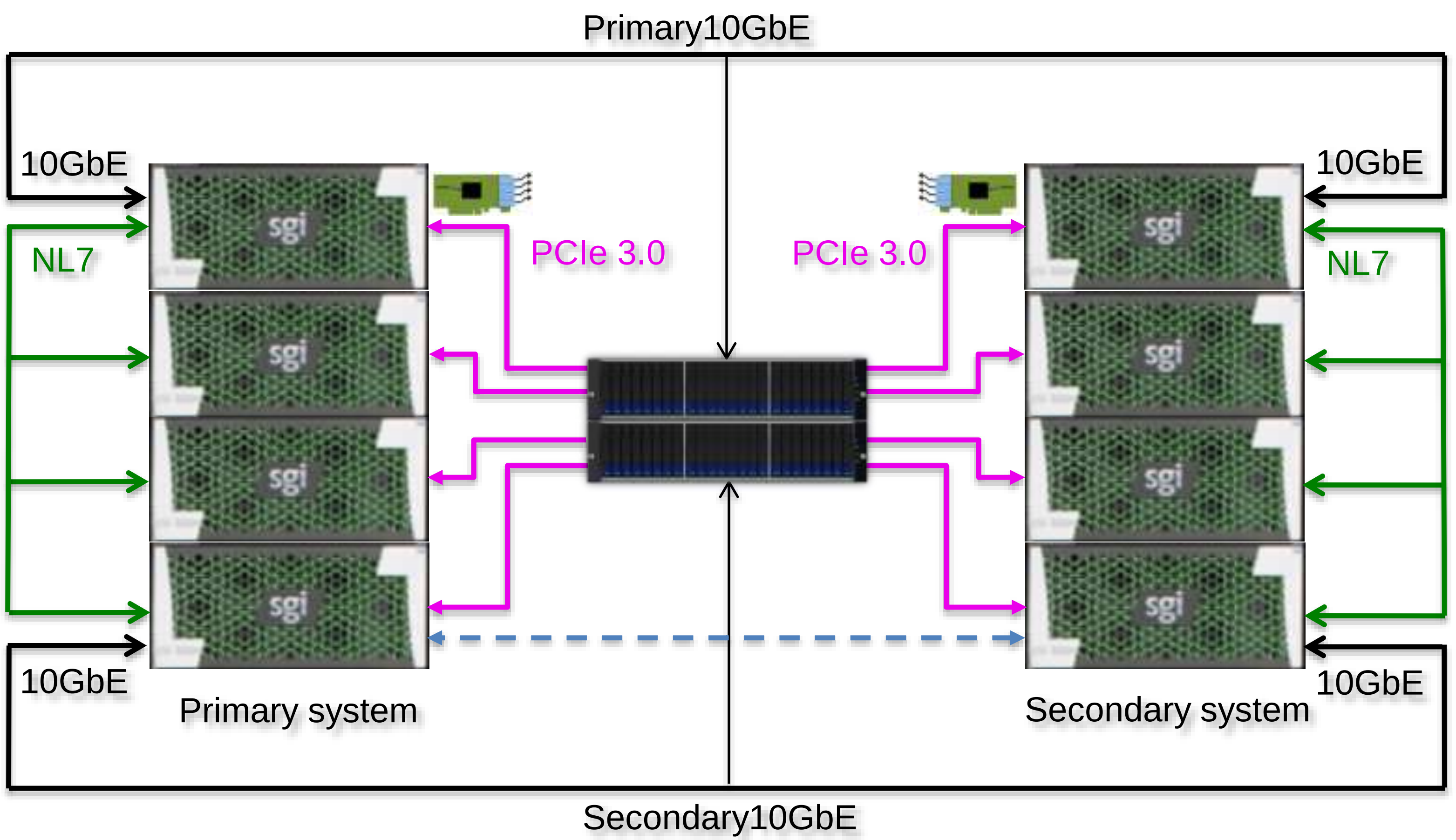
SGI's Enhanced XFS
13M read IOPS (32 cards)



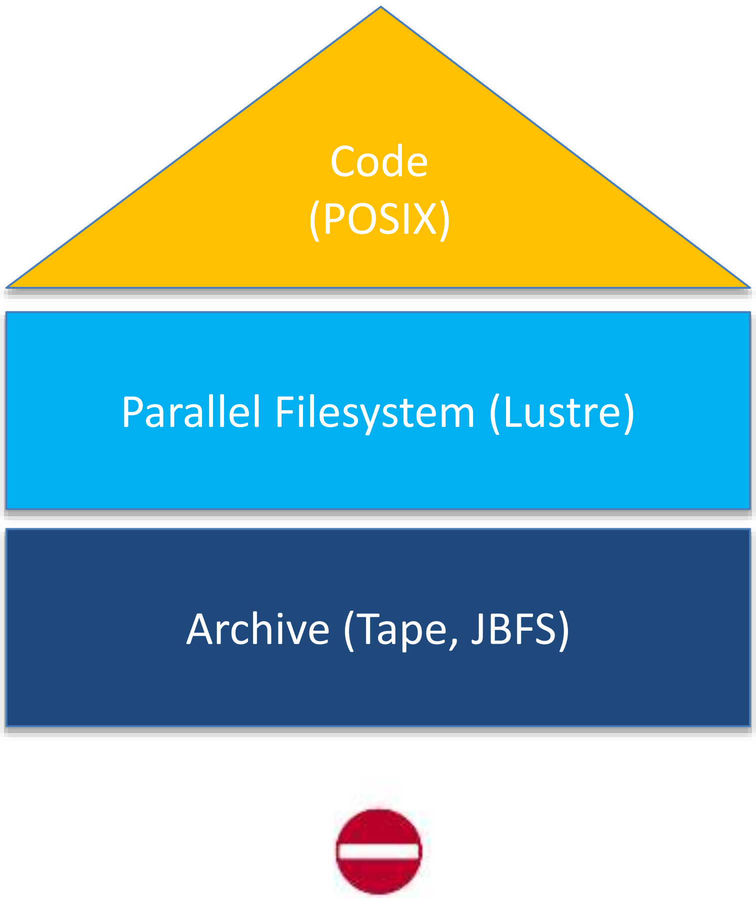
Direct parallel access to NVMe queues

Demonstrated at ISC 15

HA System Concept



Data Tiering with DMF Today



Codes require POSIX
Buffered or Direct I/O
MPI communications

High streaming performance
Limited namespace (1B)
Redundancy required

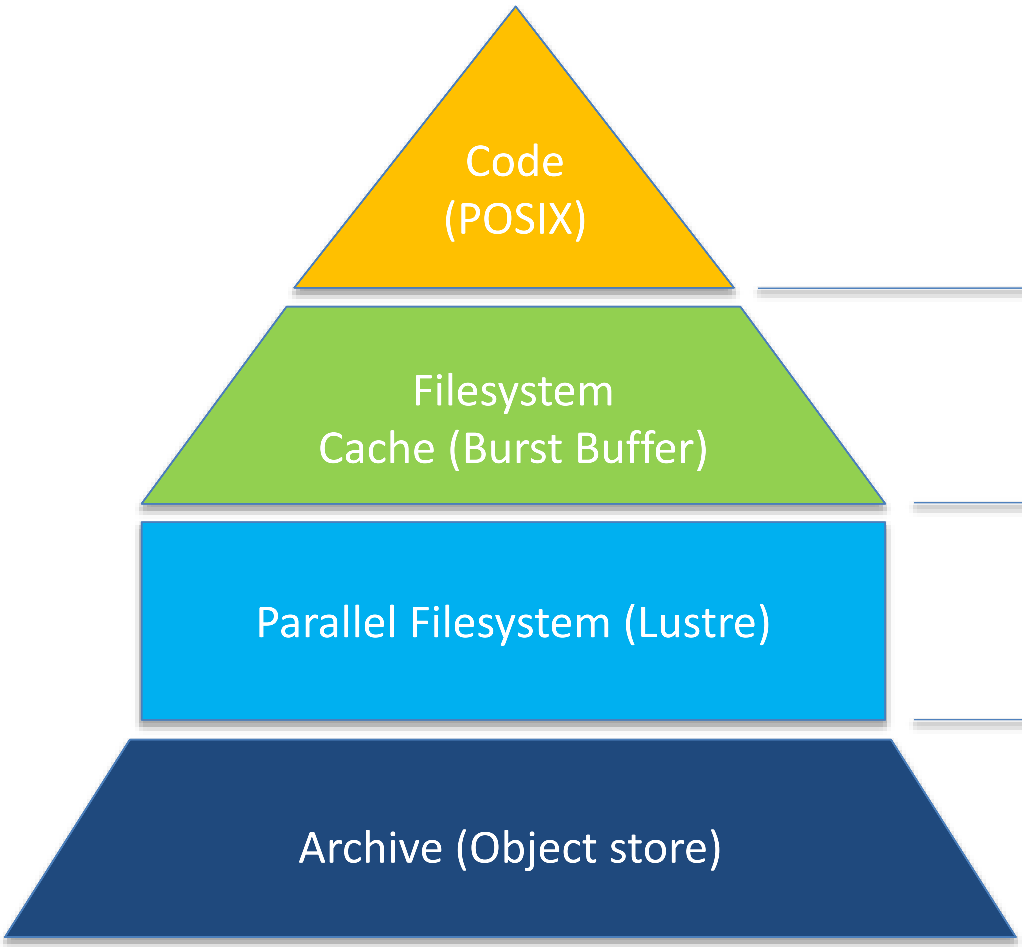
Good streaming performance
Limited namespace (1B)
Mirroring only

No multisite support
No replication to cloud
Manual archive replication only

Dynamic Namespaces

- What if we could construct namespaces (i.e. filesystems) on the fly and present them to codes?
- Initially all files are offline. We use DMF-like recall triggered by scheduler to bring data into target tier (e.g. Flash)
- Then we run HPC job
- Scheduler triggers migration, then backs up namespace

Data Tiering for HPDA



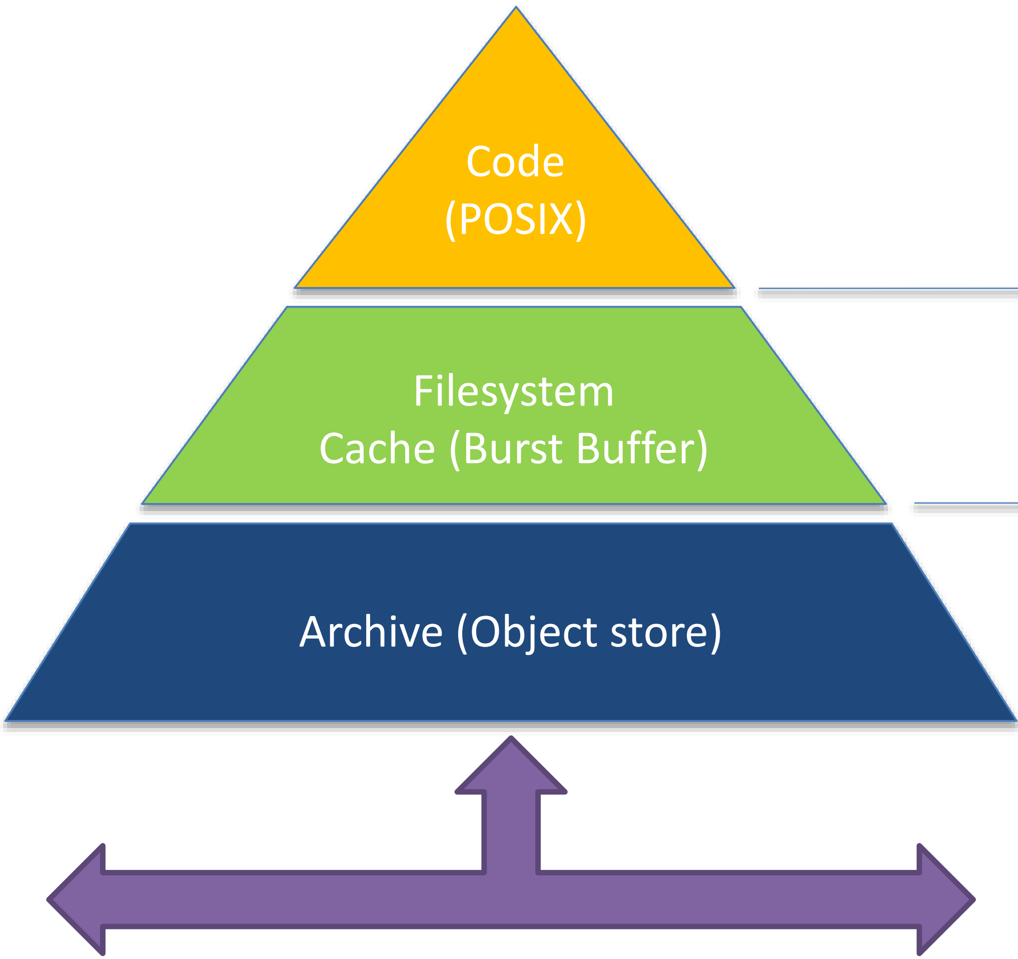
Codes require POSIX
Buffered or Direct I/O
MPI communications

Highest random I/O performance
Limited namespace (up to 100M)
Flexible or no data redundancy

High streaming performance
Medium namespace (1B)
Redundancy required

Good streaming performance
Very large namespace (over 10B)
Erasure codes and DR

Future Data Tiering for HPDA



Codes require POSIX
Buffered or Direct I/O
MPI communications

Highest random I/O performance
Limited namespace (up to 100M)
Flexible or no data redundancy

Good streaming performance
Very large namespace (over 10B)
Erasure codes and DR

Third-party archive integration
Geographic replication & Backup
For example, using DDN WOS

Summary and Discussion

Kirill Malkin
kmalkin@sgi.com