DDN STORAGE

Accelerating High Performance Cluster Computing Through the Reduction of File System Latency

David Fellinger
Chief Scientist, DDN Storage

Accelerating HPC Applications

Traditionally, where has the focus been?

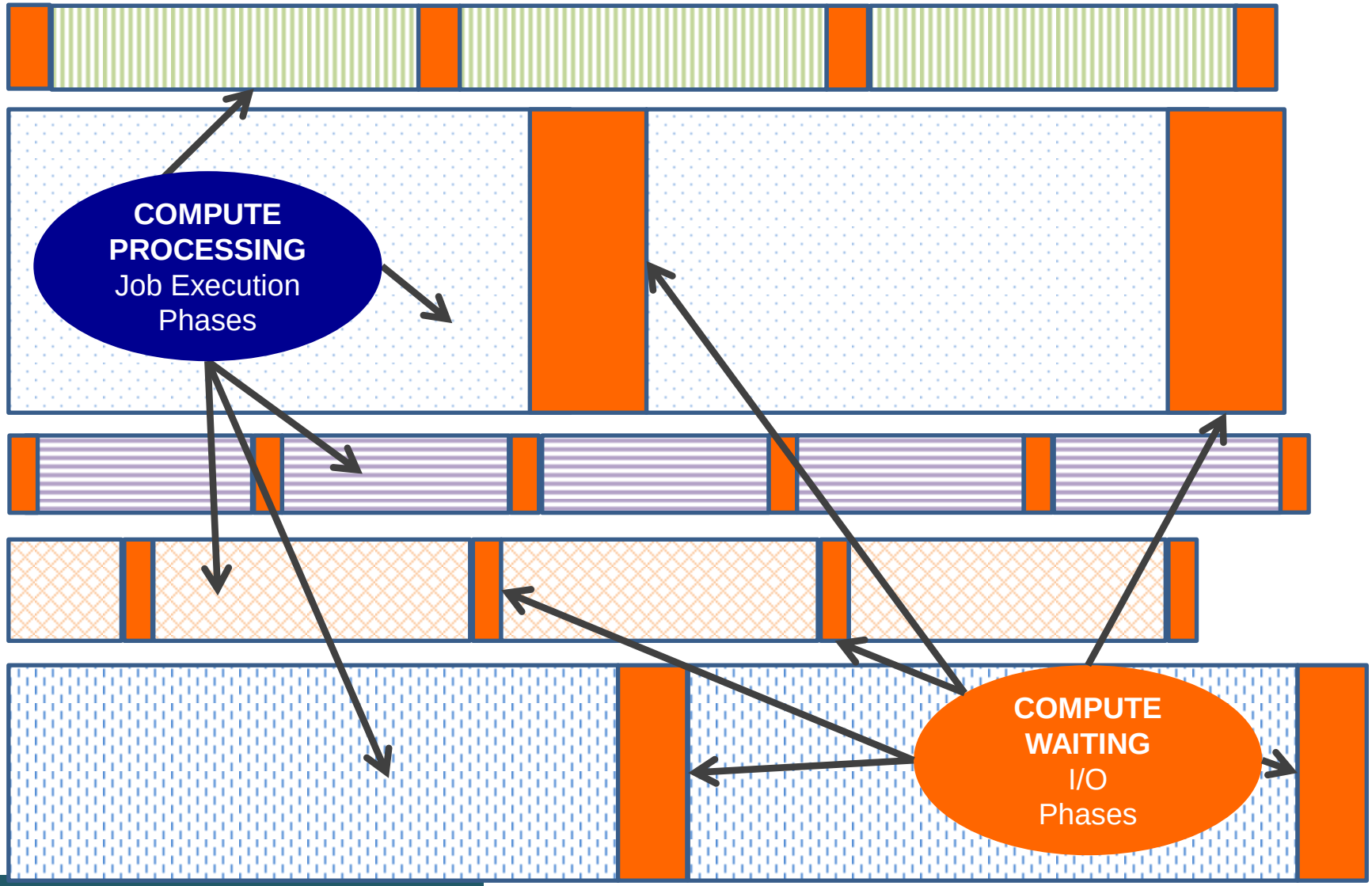
In large clusters, primarily on the cluster itself

- ✓ Lower latency interconnects
- ✓ More efficient message passing structures
- ✓ Higher performance processors & GPUs

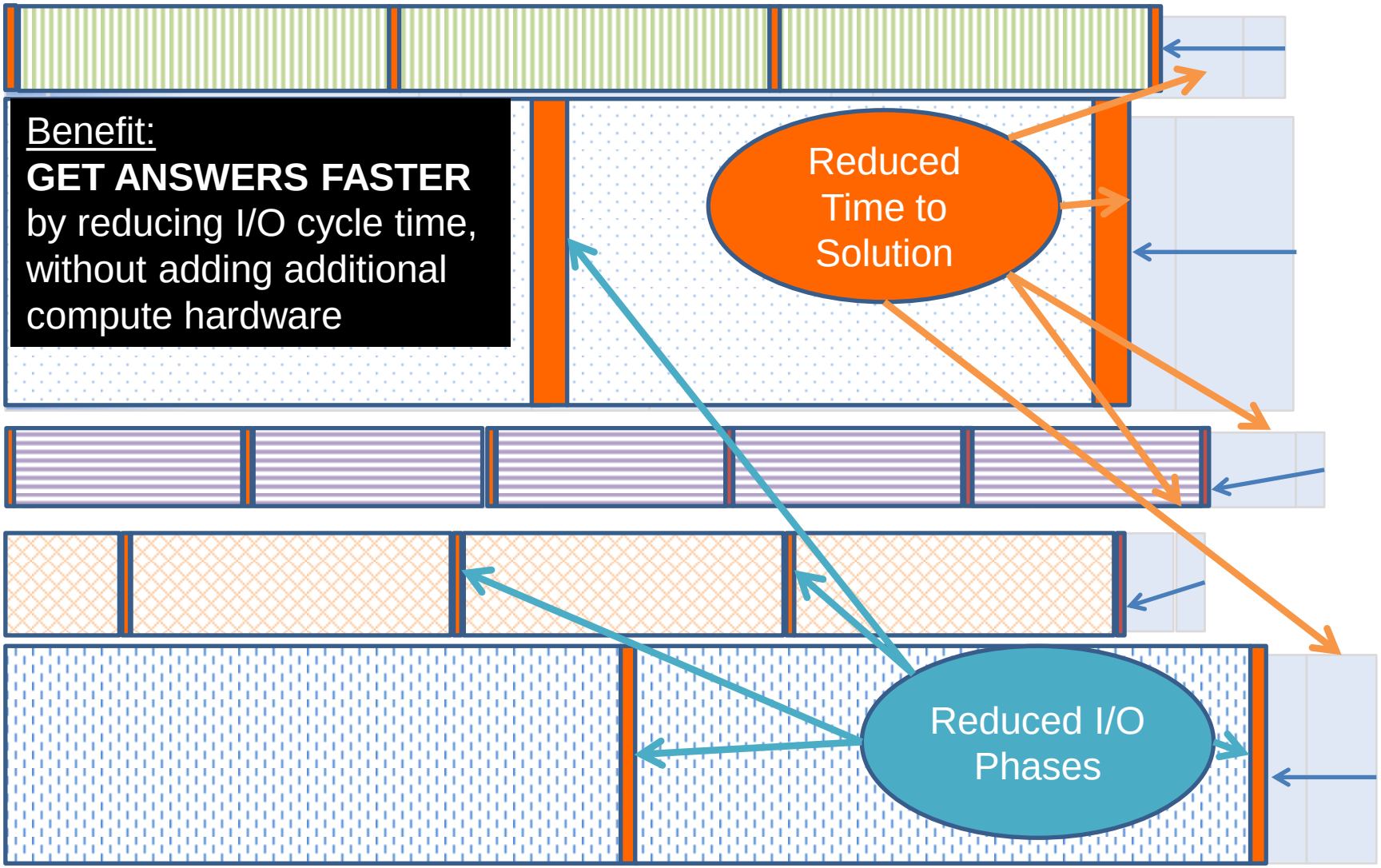
Also, research & study on processing techniques to achieve true parallel processing operations

- Symmetric multi-processing vs. Efficient message passing

Today's Challenge: Bulk I/O Latency

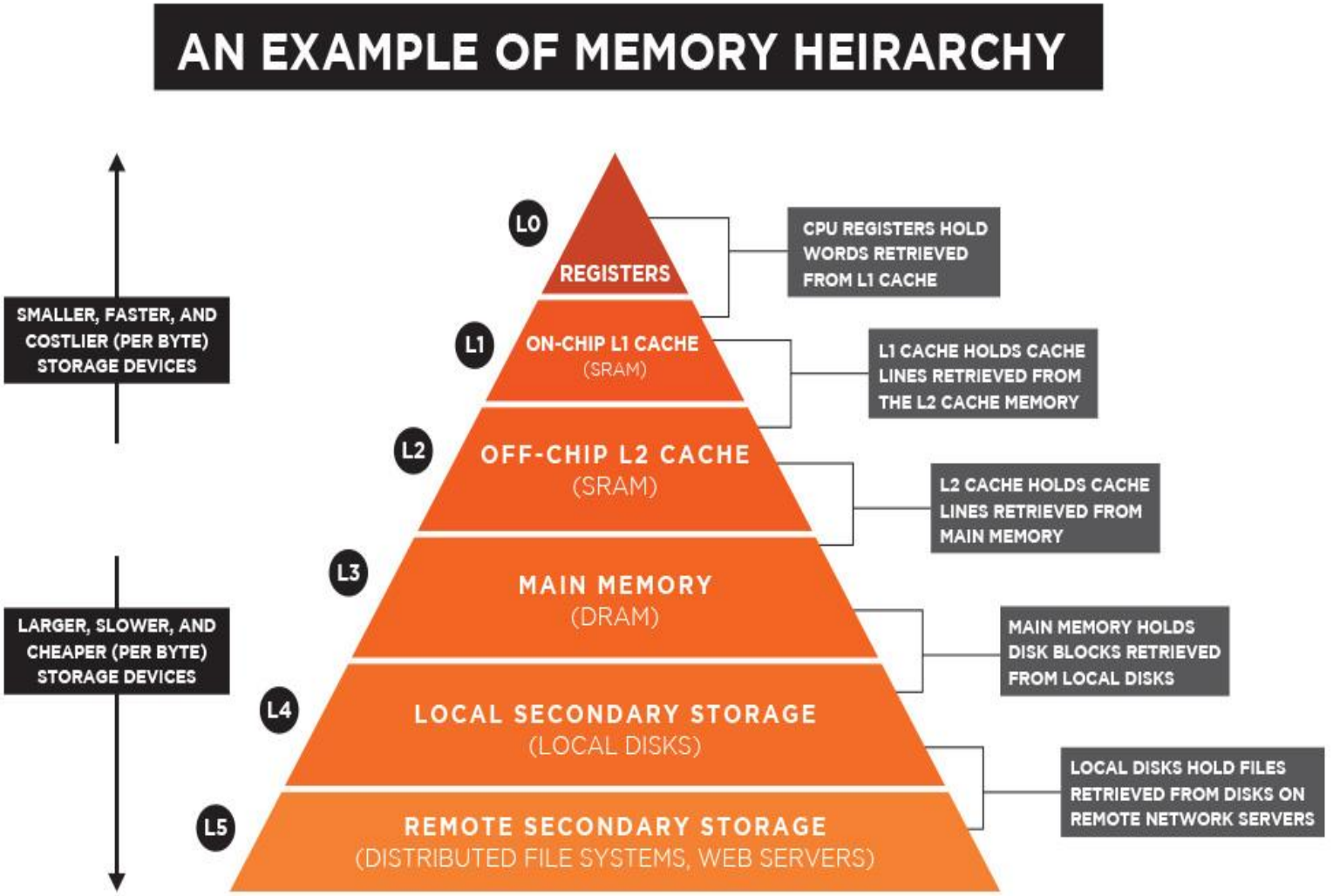


What's Needed? Compressed I/O!

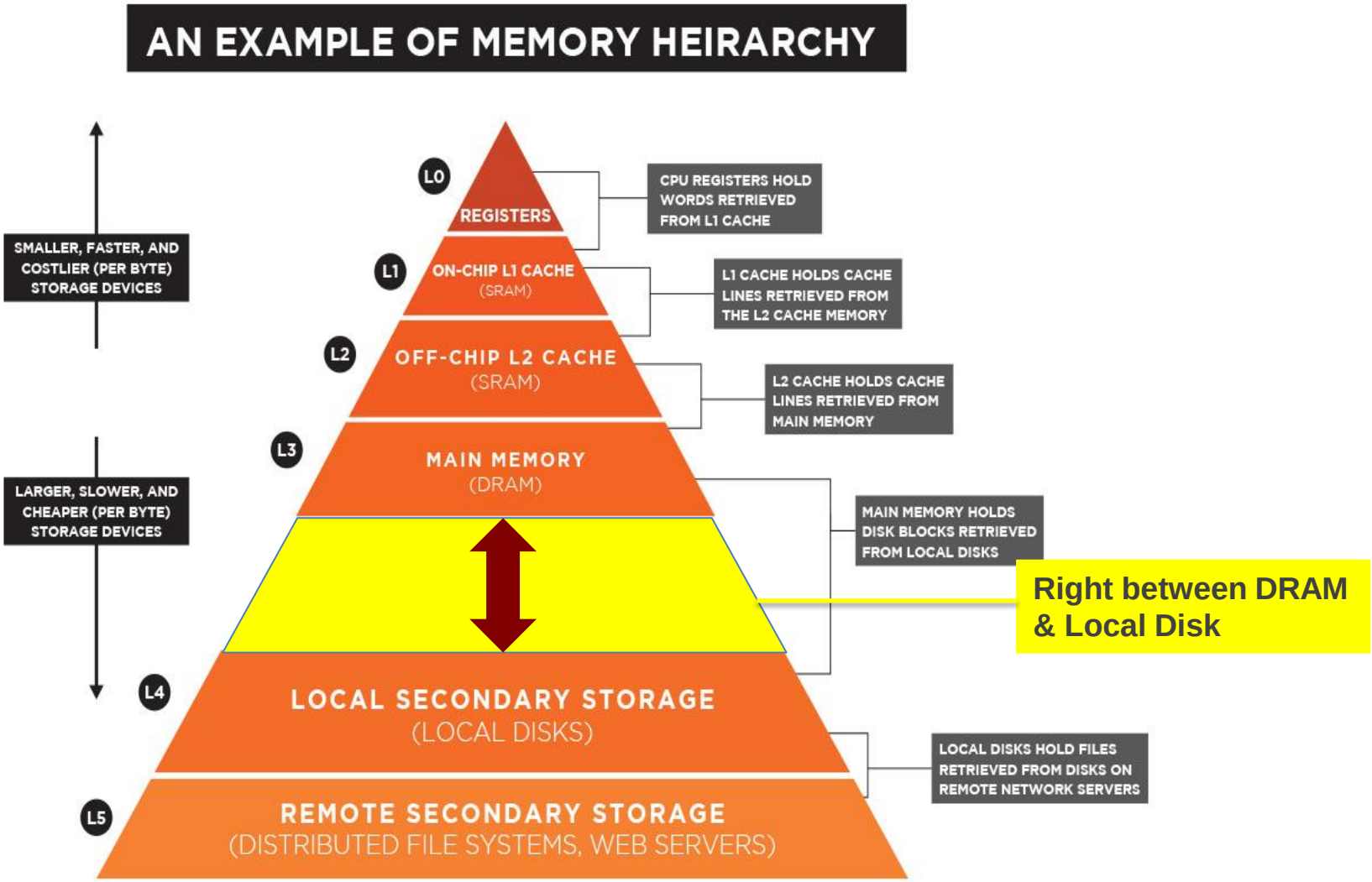


DDN STORAGE

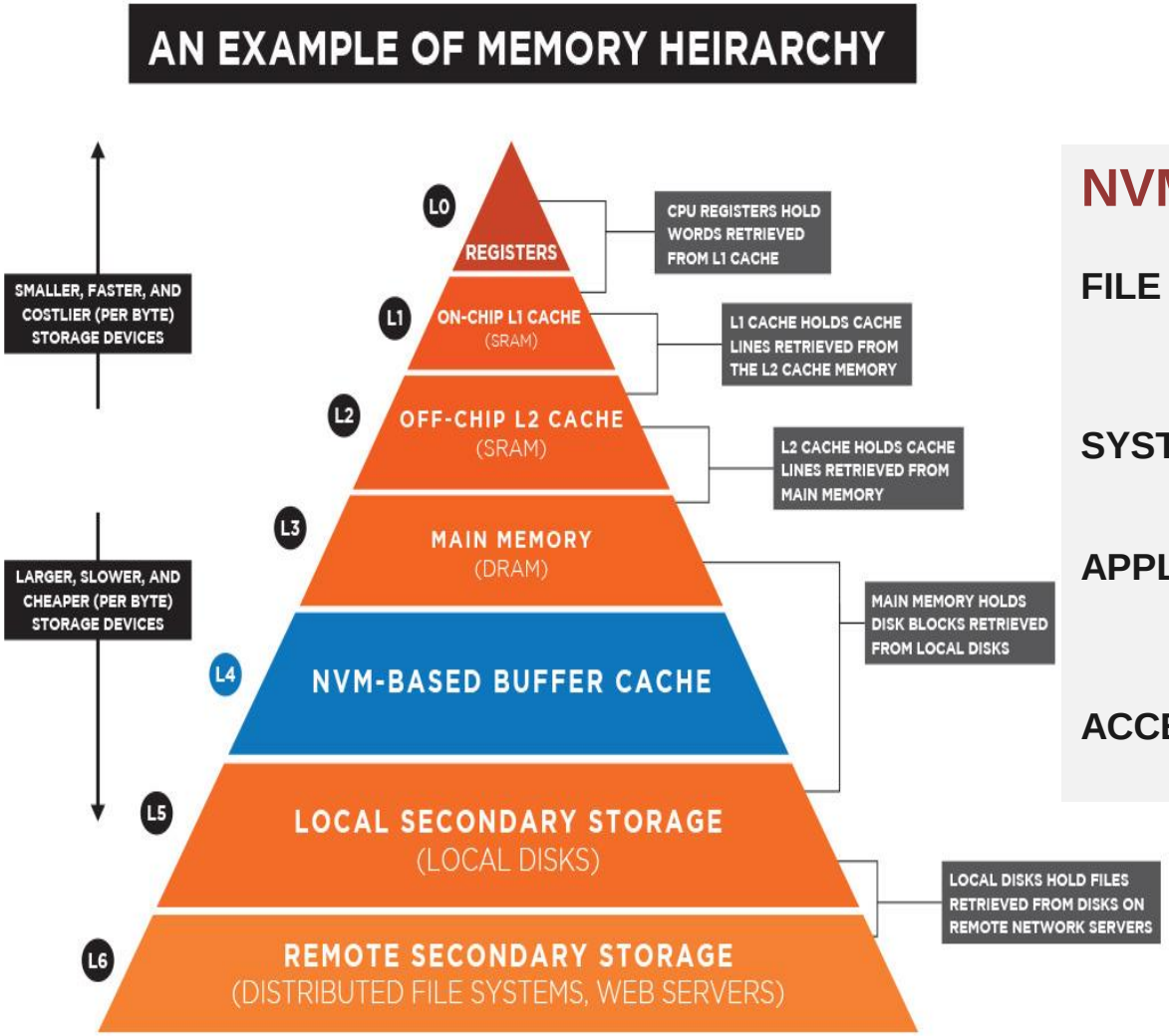
An Emerging Storage Tier is Needed



The Emerging Storage Tier's Placement



Utilizing The Emerging Storage Tier



NVM-BASED BUFFER / CACHE

FILE CACHE between DRAM & local disk
- Performance
- Capacity
- Cost

SYSTEM MANAGED Resource
- Shared PFS acceleration, pinned libraries, common datasets, etc.

APPLICATION MANAGED Resource
- Allocated on a per-job basis, dedicated to a specific job or application, etc.
- Application “co-design”

ACCELERATING parallel file system **I/O**
- POSIX, MPI, etc.

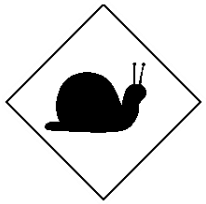
Is NVM Memory Hardware Enough?

Solid state storage offers high IOPS and low power, but in HPC . . .

File Systems Force Locks

Internal FS Operations Are Single Threaded

Large Operations Limited By Port Bandwidths

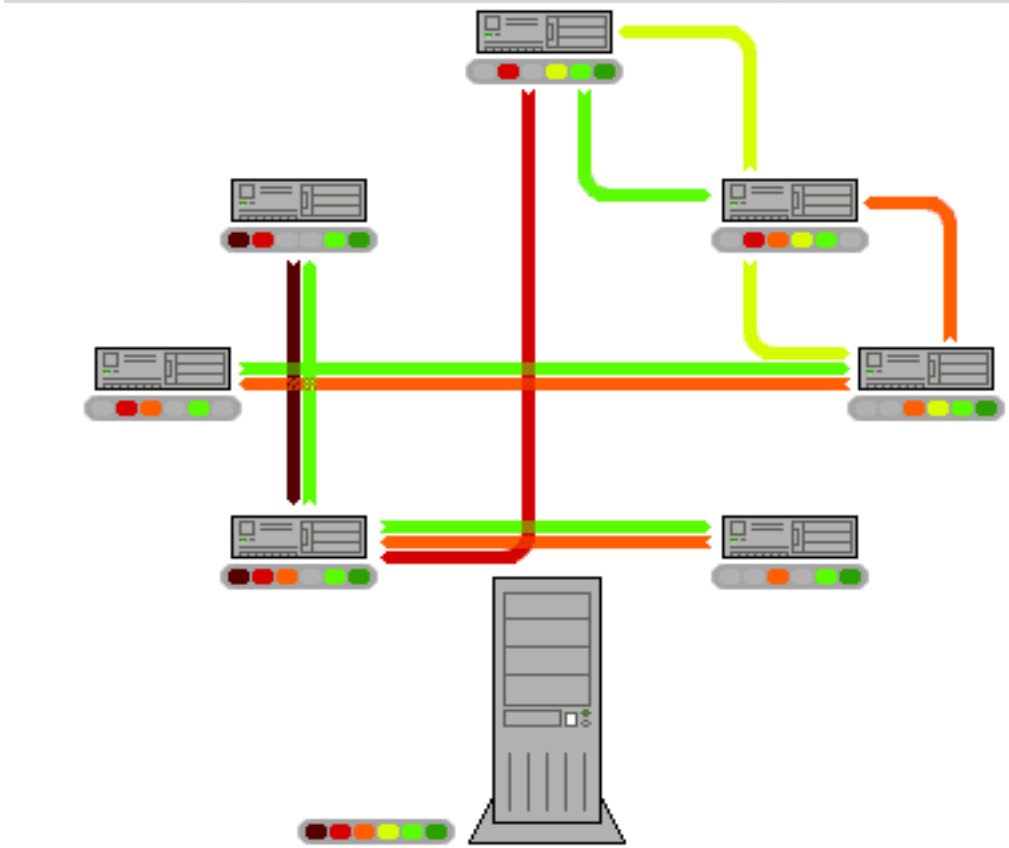


Many Operations Are Forced To Piecewise Sequential Constructs

Load Leveling Required To Extend SSD Life

SOFTWARE MUST BE DEVELOPED TO UTILIZE THE BENEFITS OF SSD HARDWARE

Other Industry Approaches

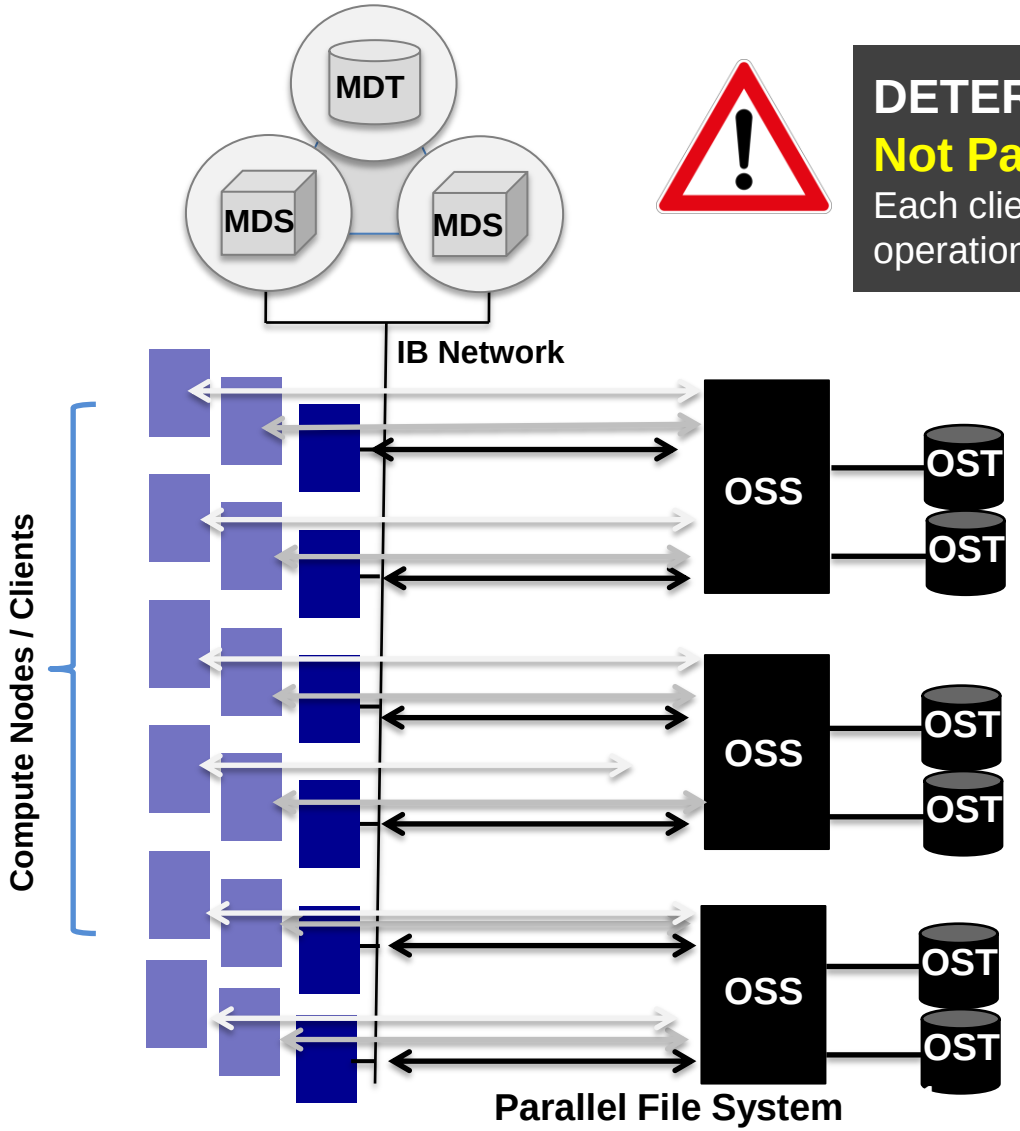


BEFORE:
Single FTP Server
(regardless of performance & availability)

NOW:
- **Utilizing parallelism** for network transfer efficiency
- **Data sharing**
- **P2P Grid**

Why is HPC living in the “Internet 80’s”?

Is a Parallel File System Really Parallel?



DETERMINISTIC WRITE SCHEMA

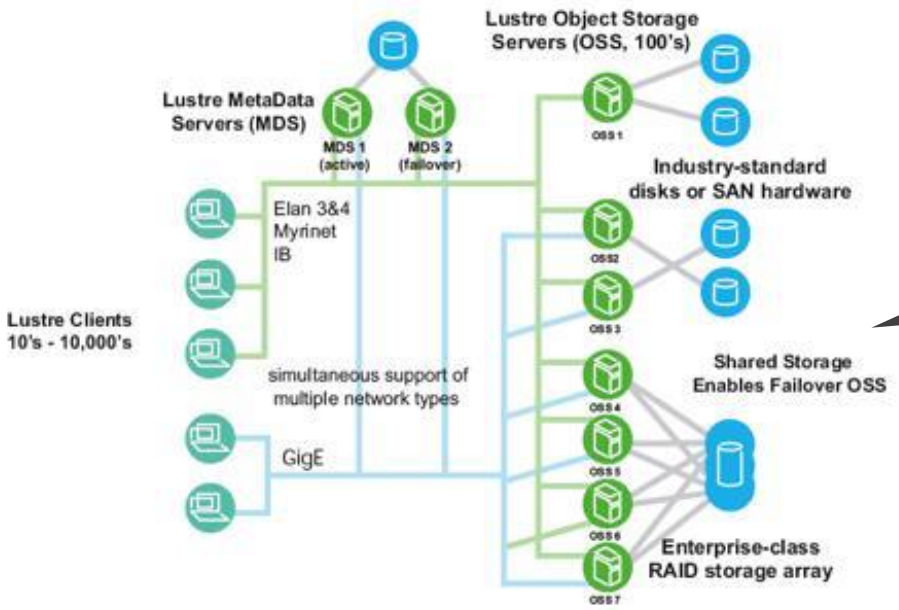
Not Parallel – It's a Bottleneck!

Each client competes to talk to a single OSS (serial operation)

What's Needed?

A **WRITE ANYWHERE FILE SYSTEM** is the next evolution

File Creation in Lustre®



But . . .
Where's the FATs?

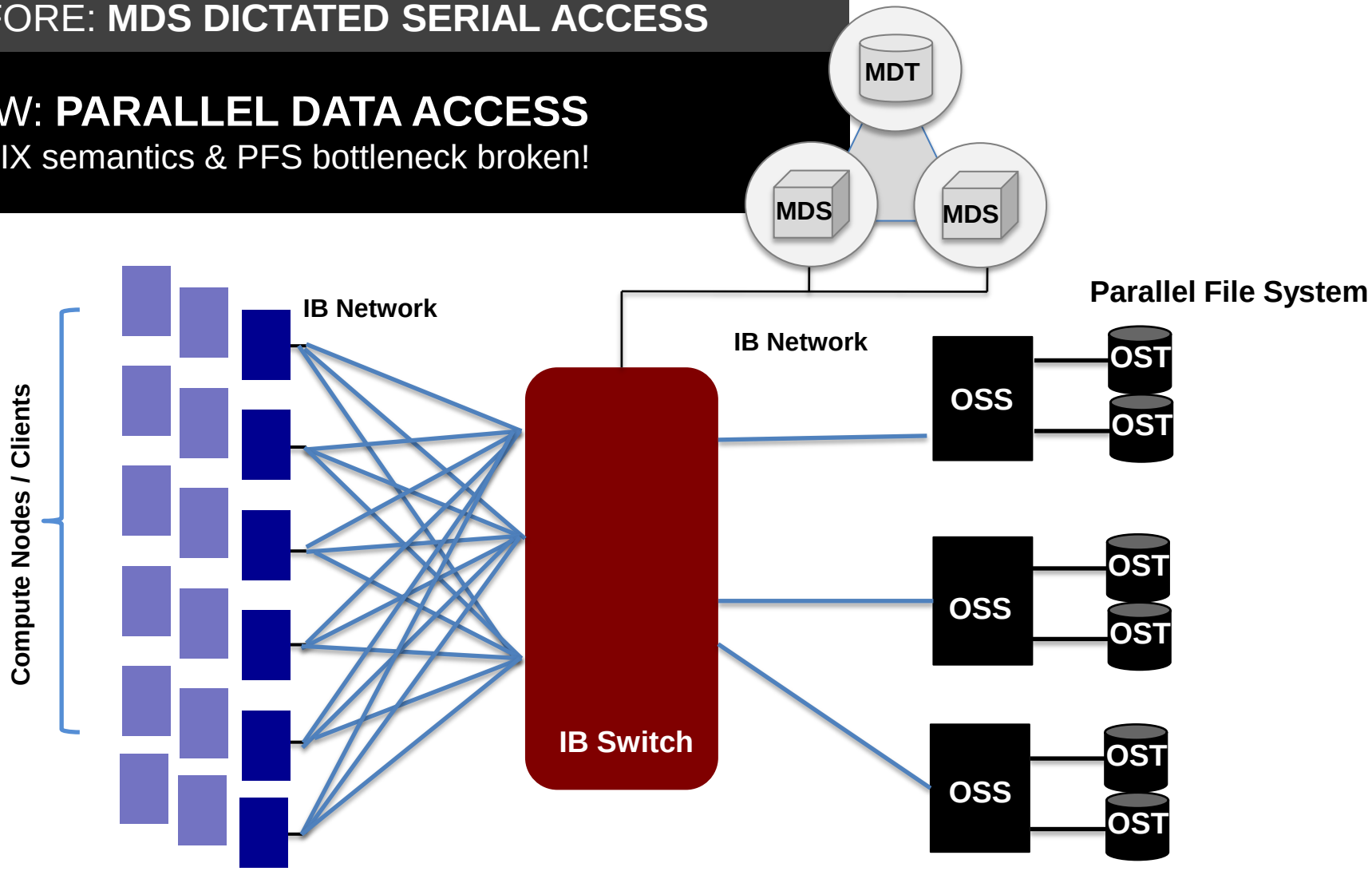
1. Ext4 extended write (I make a call to a Lustre client)
2. Redirect request to metadata server
3. Metadata server returns OST number & iNode number, execute locks
4. Begin classic write operation: iNode to file loc table EXT 4 gathers blocks from garbage collection to extent list
5. Metadata server assignment of handle - Then lock is released

IME Makes a Parallel File System Parallel

BEFORE: **MDS DICTATED SERIAL ACCESS**

NOW: **PARALLEL DATA ACCESS**

POSIX semantics & PFS bottleneck broken!

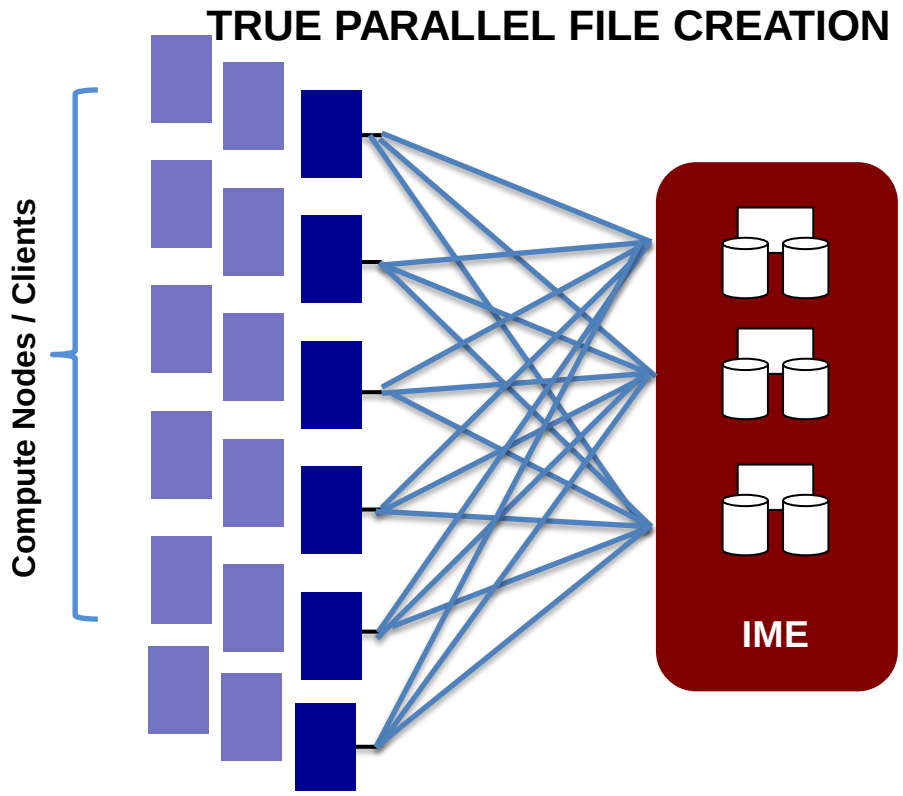


File Creation in IME®

The Magic of Write Anywhere

Now . . . **THE PFS IS PARALLEL!**

Every compute node can write file increments to every storage node along with metadata & erasure coding

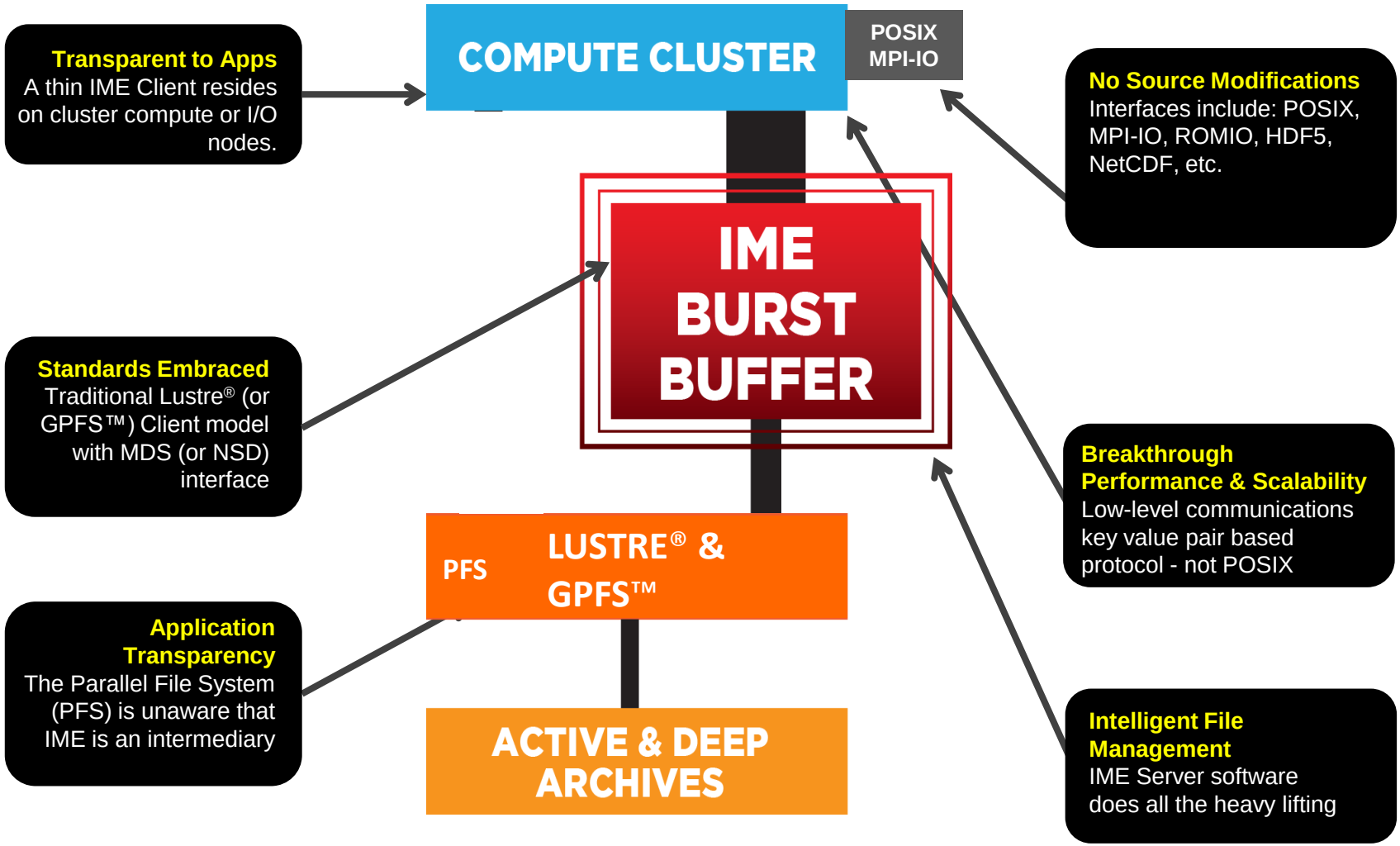


In IME, we've implemented a DHT.

Now, the **compute is sharing files** the same way Napster users do

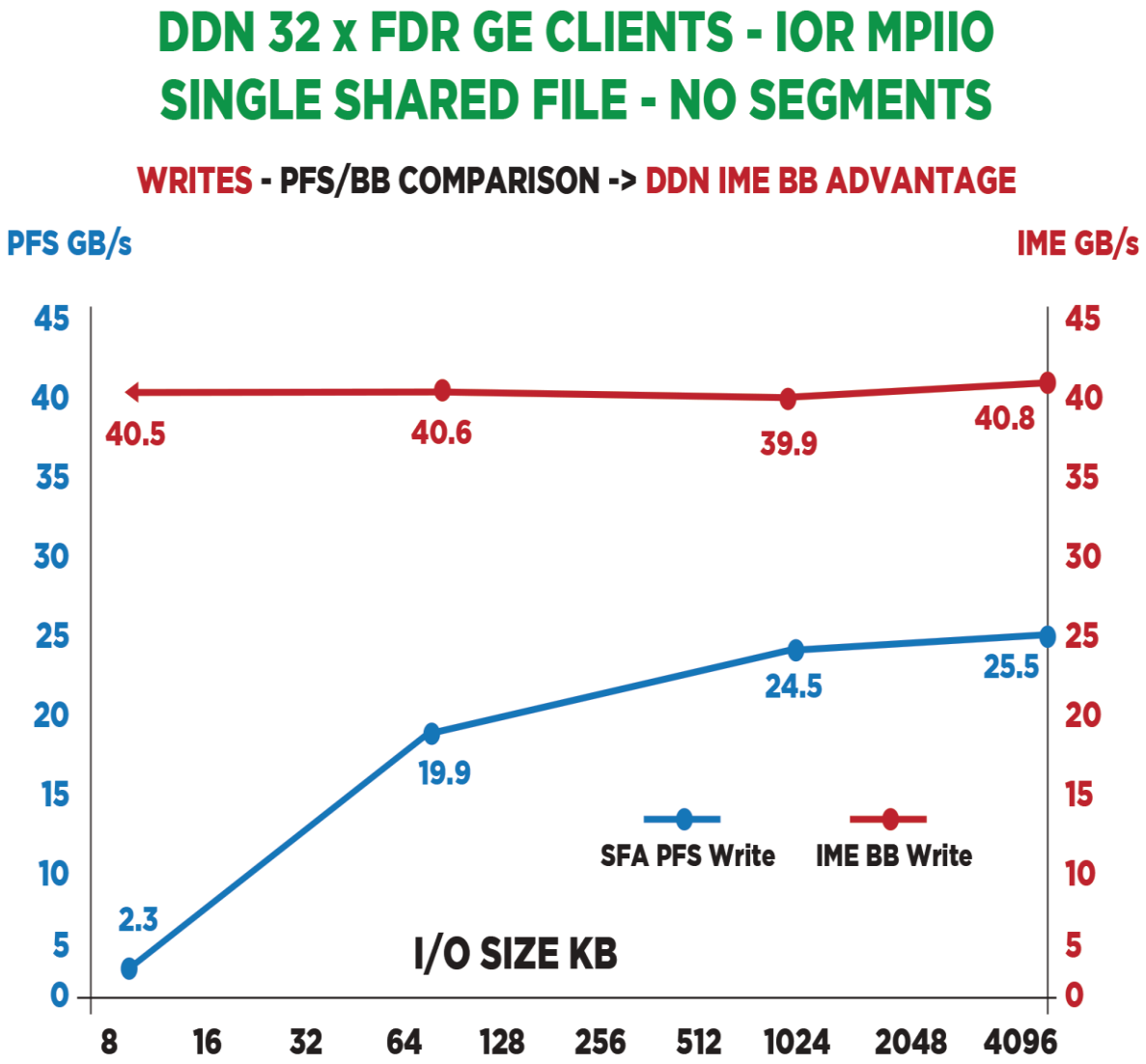
Introducing IME®

Technical Key Components and Operations



IME® Benchmarking

IOR MPI-IO



DDN STORAGE

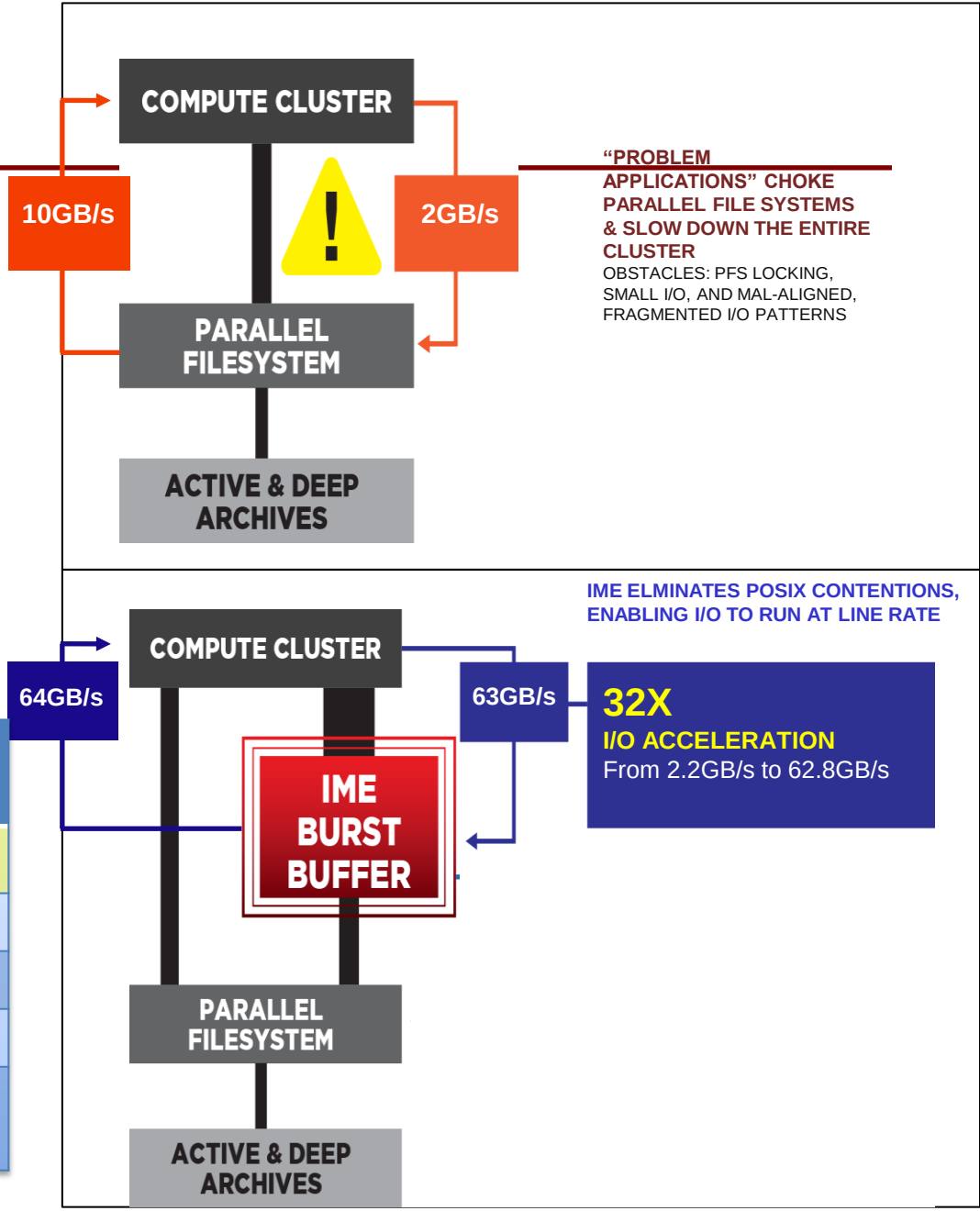
IME® Benchmarking

HACC_IO @TACC Results

Description:

HACC-IO is an HPC Cosmology Kernel

Particles per Process	Qty. Clients	IME Writes (GB/s)	IME Reads (GB/s)	PFS Writes (GB/s)	PFS Reads (GB/s)
34M	**128**	**62.8**	**63.7**	**2.2**	**9.8**
34M	256	68.9	71.2	4.6	6.5
34M	512	73.2	71.4	9.1	7.5
34M	1024	63.2	70.8	17.3	8.2
IME Acceleration		**3.7x-28x**	**6.5x-11x**		



Moral of the story

HPC Cluster = Large Group of Data Users

- ✓ Why haven't we learned . . .
  - What the internet p2p guys have known for a long time?
- ✓ Learn to share!
  - The precious resources of network bandwidth & storage

Thank You!

Keep in touch with us

David Fellinger
Chief Scientist,
DDN Storage
dfellinger@ddn.com

sales@ddn.com

@ddn_limitless

company/datadirect-networks

2929 Patrick Henry Drive
Santa Clara, CA 95054

1.800.837.2298
1.818.700.4000