HPCS I/O and Storage Issues

David Koester, Ph.D. (MITRE)
Henry Newman (Instrumental)

16 August 2005

HEC I/O Workshop

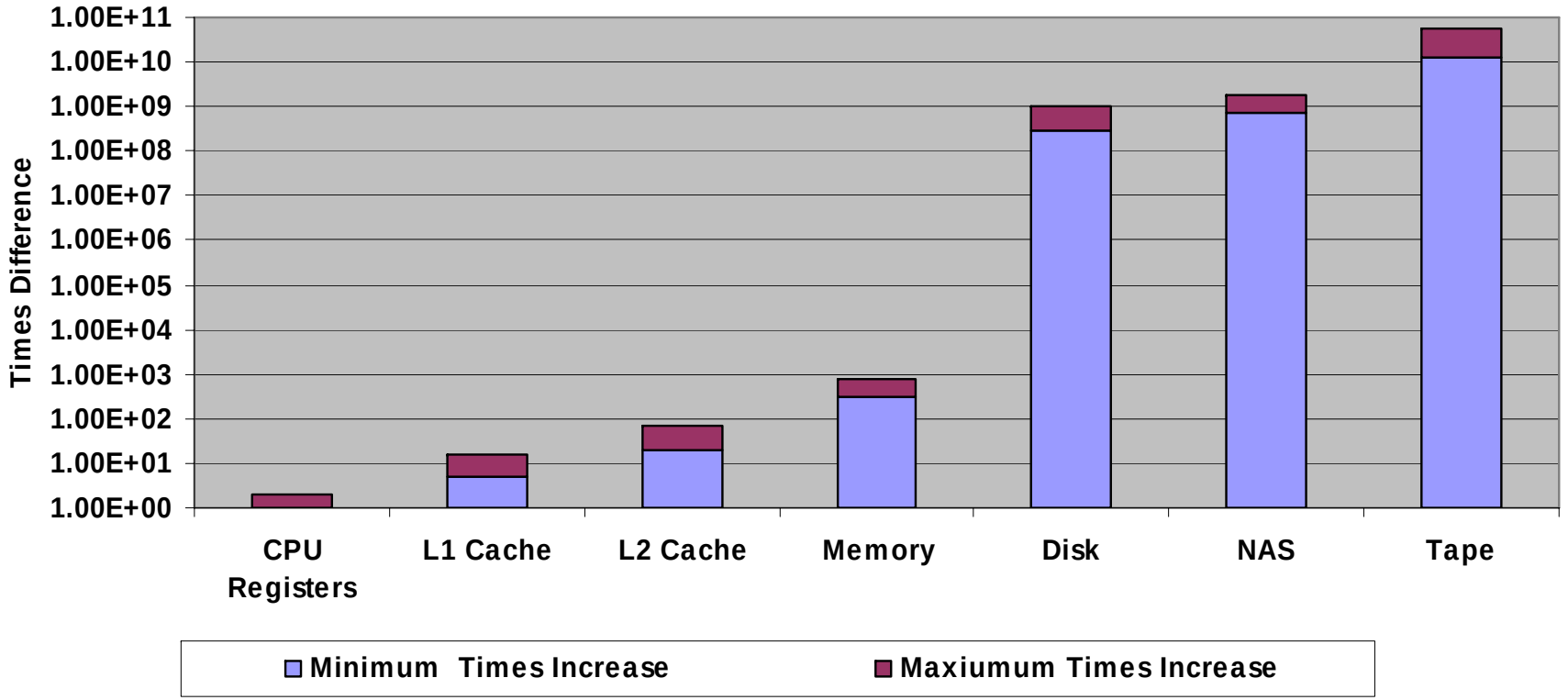
Outline

- Our view of the challenges
- Answer the following questions
  - High level summary of File Systems and I/O R&D
    - Emphasis on topics that will help DARPA HPCS
  - Areas that need more focus
  - Areas that have too much focus
  - If you were in our shoes with limited funds, what File Systems and I/O R&D would you fund?
- Summary

High level summary of File Systems and I/O R&D: Emphasis on topics that will help DARPA HPCS

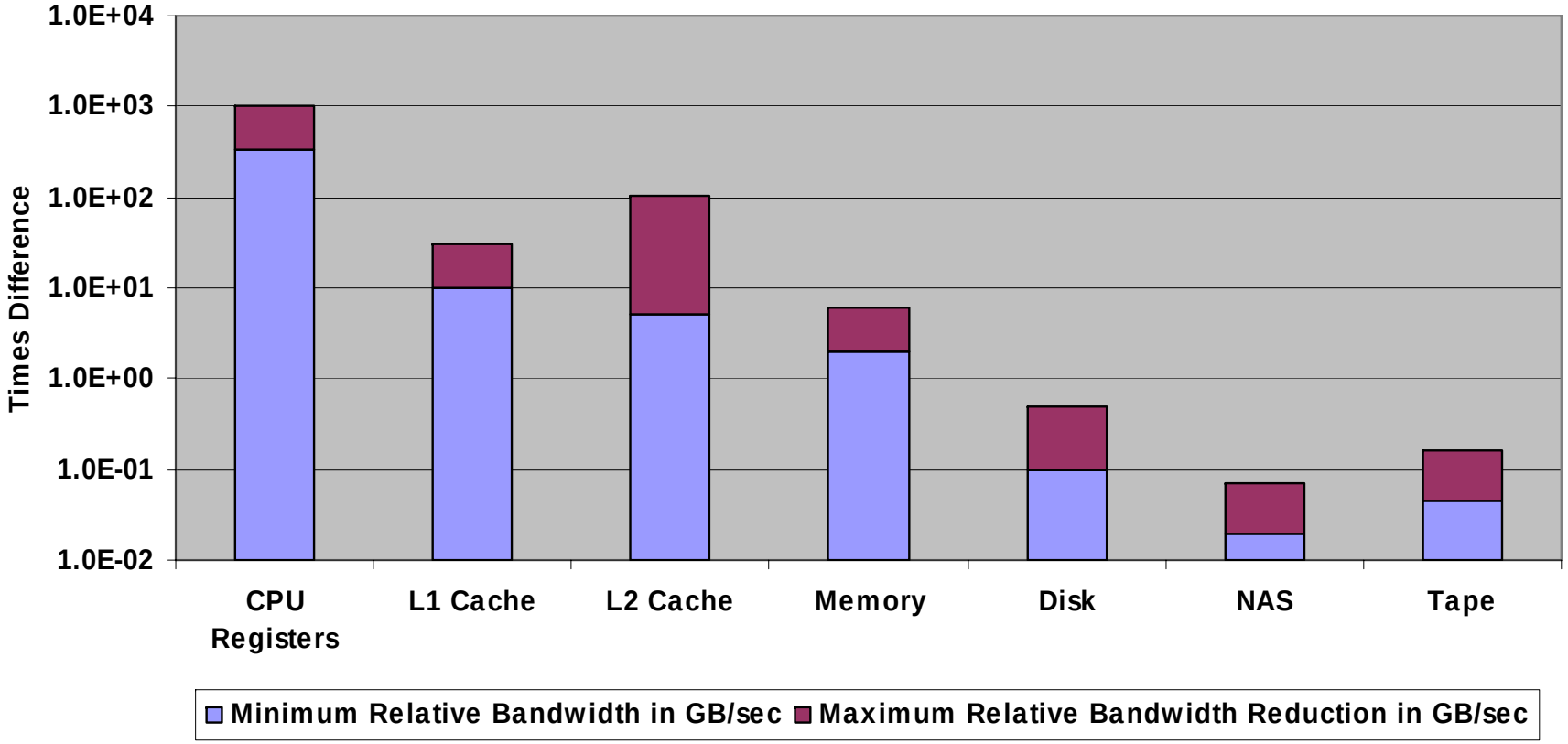
Information collected by DARPA HPCS

Relative I/O Latency



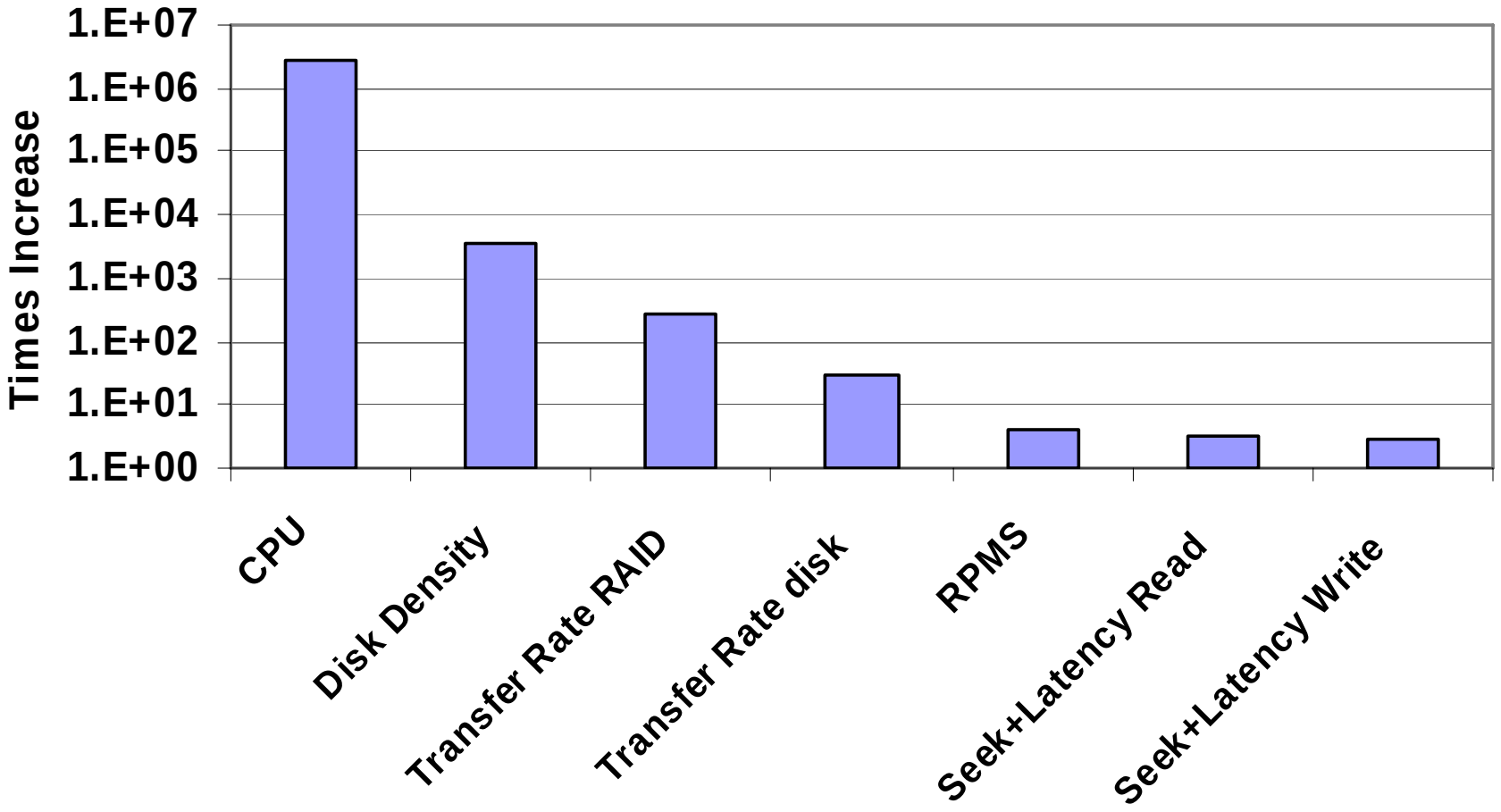
Note: Approximate values for various technologies and 12 order of magnitude

Relative I/O Bandwidth per CPU



Note: Approximate values for various technologies and 6 orders of magnitude

Performance Improvements for Storage 1977-2005



The Lessons of History

- Densities are growing but not nearly as fast as CPU performance
  - Seek, latency and transfer times are not going to change much over the next few years
- Transfer rate improvements lag Capacity increases significantly
  - I/O performance is comprised of a combination of
    - Raw media transfer rates
    - Access times (seek+latency)
    - File system efficiency
- Different applications and architectures are more sensitive to different combinations of 3 areas
  - You are not going to spin disks 5.76 B RPMs
  - You are not going to sustain 4.7 TB/sec to a single disk drive
  - You are not going to have a disk with 215 TB of capacity (soon)

Multi-Node File System Problems

- Though Lustre is working to solve one scaling problem it does not solve many others
- "Bandwidth" has different interpretations for different HPCS Mission Partners
  - File creations per second
  - Multiple streams of I/O from a single node
  - Multiple streams of I/O from multiple nodes
  - I/O from many nodes
- A single framework to achieve all "Bandwidth" interpretations may be difficult to achieve
  - May be mutually exclusive

Data Management is Hard

- Management of the hardware and software stack is not only hard for HPC environment but hard for everyone
  - Configuration control with all of the software options, firmware options and firmware release levels is mind numbing
- There is no good recognized framework to manage all of these and the wheel is constantly being reinvented by vendors
- Management of all the technologies is complex
  - Yet you must tune all aspects to get good performance

Data Locality

- Locality issues
  - Where is the data in memory
  - Knowing where data is on disk so read/write has minimal impact
- Locality issues are compounded by the enormous amount of software “in the middle”
  - Operating system
  - File System
  - Volume Management
  - Failover
  - Host bus adapters
  - etc.

I/O Parallelism

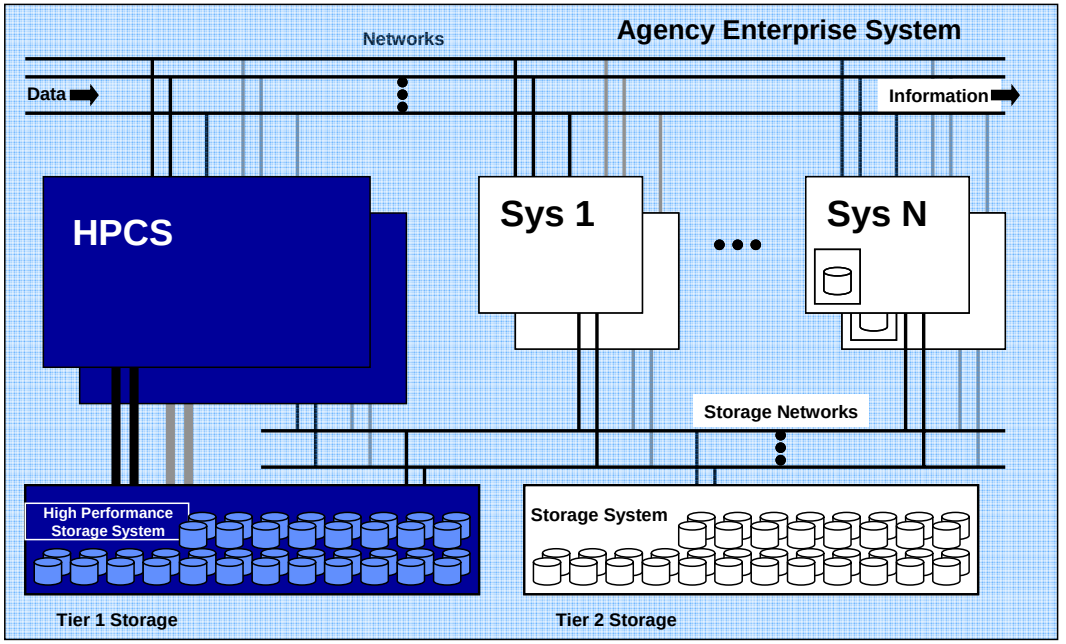
- Good performance will depend on many threads of I/O from many nodes
  - Storage technology is not going to change (fast enough) so this will be required
- Current file system technology will require changes to allow for this parallelism
- In fact it is not just the file system but the whole data path including applications and storage
- Standards for parallel I/O and storage must be developed that go far beyond MPI-I/O

HPCS I/O Challenges

An Envelope on HPCS Mission Partner Requirements

- 1 Trillion files in a single file system
  - 32K file creates per second
- 10K metadata operations per second
  - Needed for Checkpoint/Restart files
- Streaming I/O at 30 GB/sec full duplex
  - Needed for data capture
- Support for 30K nodes
  - Future file systems need low latency communication
- System will be part of enterprise architectures!

HPCS as part of a Mission Partner's Enterprise Architecture



Mission Partner Requirements Not Covered by HPCS

- All mission partners have a requirement for Data Lifecycle Management (DLM) — i.e., archived and managed
  - Some Mission Partners' applications require that the data be "very close" to the processing — i.e., time to access
    - DoD Agencies
    - Weather and Climate modelers
    - Others
  - Other Mission Partners' applications permit the data to have increased "time to access"
    - DOD/DOE/NASA
- With HPCS machines this problem may become even more challenging than today
  - Potentially large technology scaling
  - Potentially large increases in data volume
  - Potentially large increases in file counts
- Data lifecycle management issues will need to be addressed!

Areas that need to have more focus

What is missing

More Focus On

- Common interfaces for shared file systems in heterogeneous environments
- Data life cycle management and seamless integration into current HPC computational environments
- New technologies that will offer greater scalability
- Changes to standards to allow scaling without sacrificing data integrity
- Streaming I/O from single nodes not aggregate bandwidth
- Data alignment issues and RAID issues
- Small block vs. large block and fixed RAID issues
- File System Metadata
- HEC is more than clusters

Areas that have too much focus

What should be scaled back

What to De-Emphasize

- Nothing - we are not close to solving anyone's problem at the petascale
- Clusters do solve some organization's HPC problems but HEC is not just clusters
  - Some HPC problems cannot be solved with clusters

If you were in our shoes with limited funds, what File Systems and I/O R&D would you fund?

Thinking out of the box

Current Data Path View

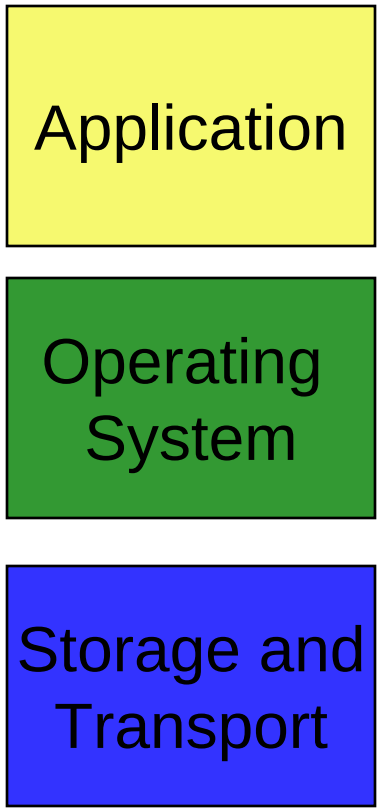
Data path today is the same as the data path 20 years ago

Evolutionary progress with Lustre and GPFS, but we still have the same limitations

Application Layer
Current POSIX system calls open/read/write/aio. Limited communication with OS layer

OS Layer
POSIX Atomic operations open/read/write/aio. No communication with physical layer

Physical Layer
Block based storage and limitations of 30+ year old technology



Latency Tolerance Data Path Last 30 Years

- Hardware evolution based on memory component latencies have caused dramatic shifts in design
  - Multi-threaded CPUs
  - Multiple levels of memory (L1, L2, L3, NUMA)
  - All of these changes are based on the need to hide latency when accessing memory as latency has increased as a function of CPU performance
- Application data path has not changed in a similar way to address latency
  - Storage latency has not changed much compared with CPU performance
  - File systems do not pass topology to block devices to impact latency
- Without addressing this latency, HPCS systems cannot scale

Applications and Hardware Changes

- While data path software changes are still limited, we have significant hardware changes
  - Hardware multi-thread
  - Multiple instruction and data paths
  - Cache coherency
  - NUMA memory
- Software changes allowing the user to control atomic behavior
  - Parallel programming via MPI, OPENMP and POSIX threads
    - Support for DMA communications specialized networks

It's all about Latency

- Applications changes such as multithreading are the same techniques used to address latency issues in the late 70s and 80s with vector based computers
  - Systems today are more efficient if they hide memory latency
  - I/O is more efficient today if they can hide latency by multi-threading the I/O
- The latency of the data path is growing
  - True for memory
  - True for I/O

Potential Research Area
Latency Optimized Data Path

Without end to end communication in the data path, latency tolerance optimization is not possible

Technologies such as OSD might provide framework to allow data path communications

Application Layer

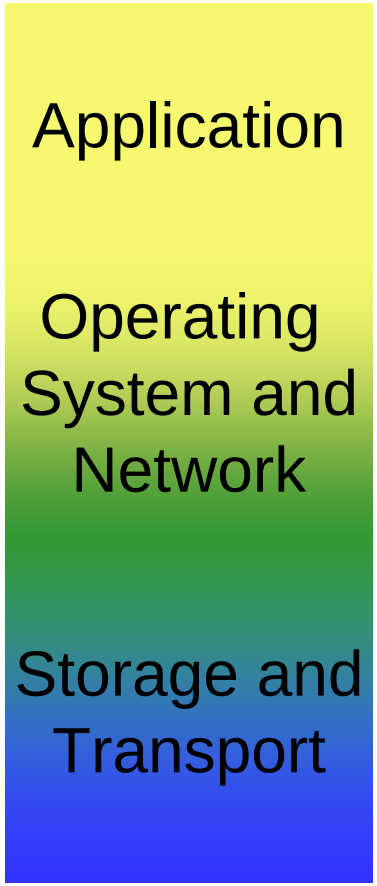
Changes to support new constructs for different types of latencies for data

OS Layer

OSD combined with networking constructs could address different latencies

Physical Layer

Given physical limitations of storage, optimizations must be done a higher level to impact the technology



Summary — Final Thoughts

- I/O challenges are daunting
- No proposed breakthrough hardware or software technology is expected to be available in the HPCS acquisition timeframe
- Mission Partner I/O requirements will grow with the acquisition of an HPCS System
- Revolutionary I/O technologies developments are need to address the growing technology gap
  - Both software and hardware