The State and Needs of IO Performance Tools

Scalable Tools Workshop
Lake Tahoe, CA

August 6–12, 2017

Elsa Gonsiorowski
Greg Becker

LLNL-PRES-735910
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Outline

Motivating Example

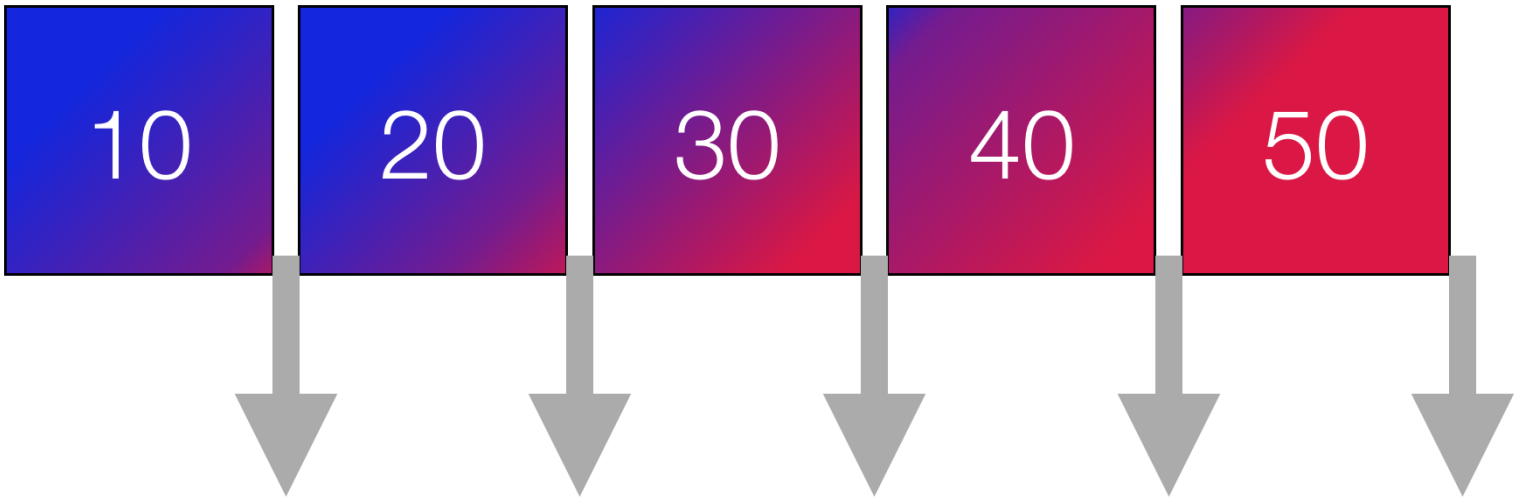
IO vs Compute Performance History

Measuring I/O Performance

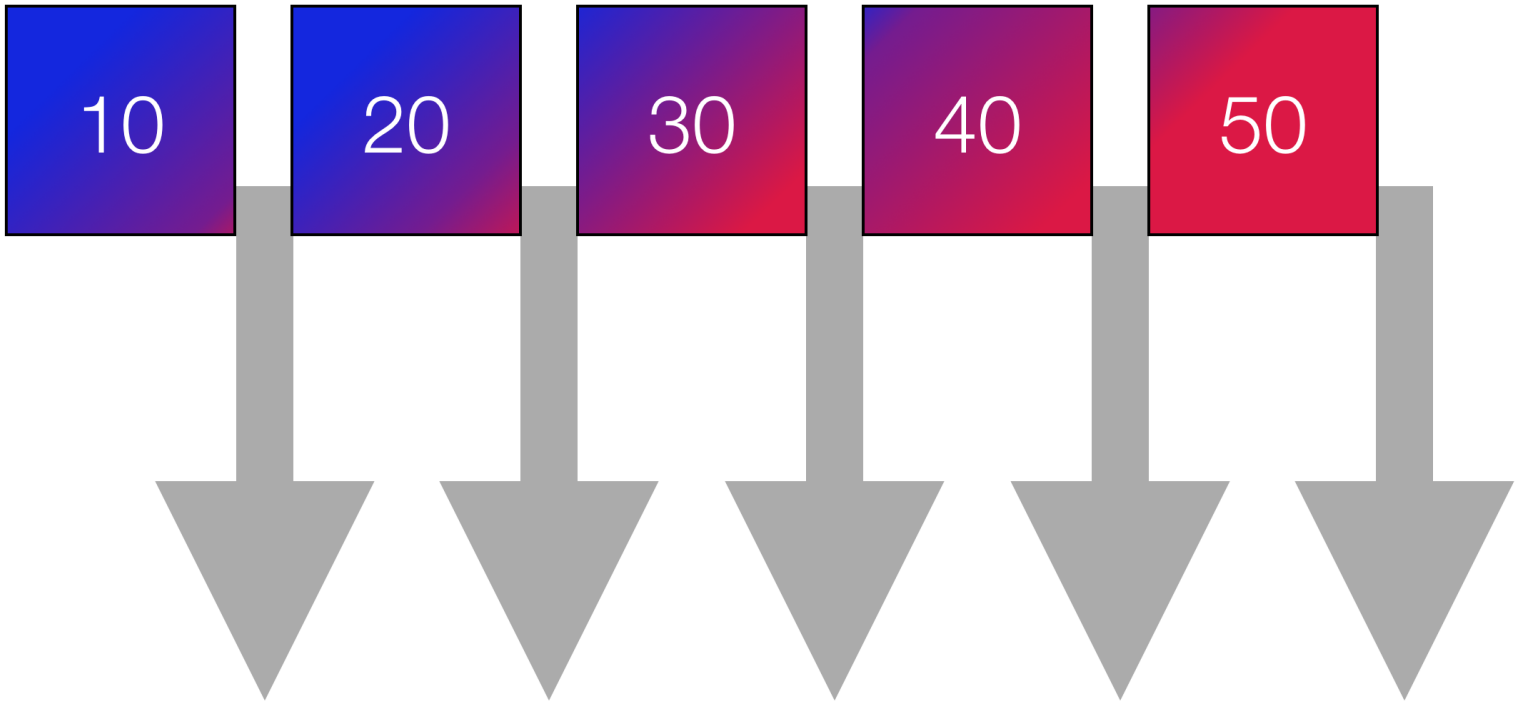
The I/O Stack

Questions from Applications

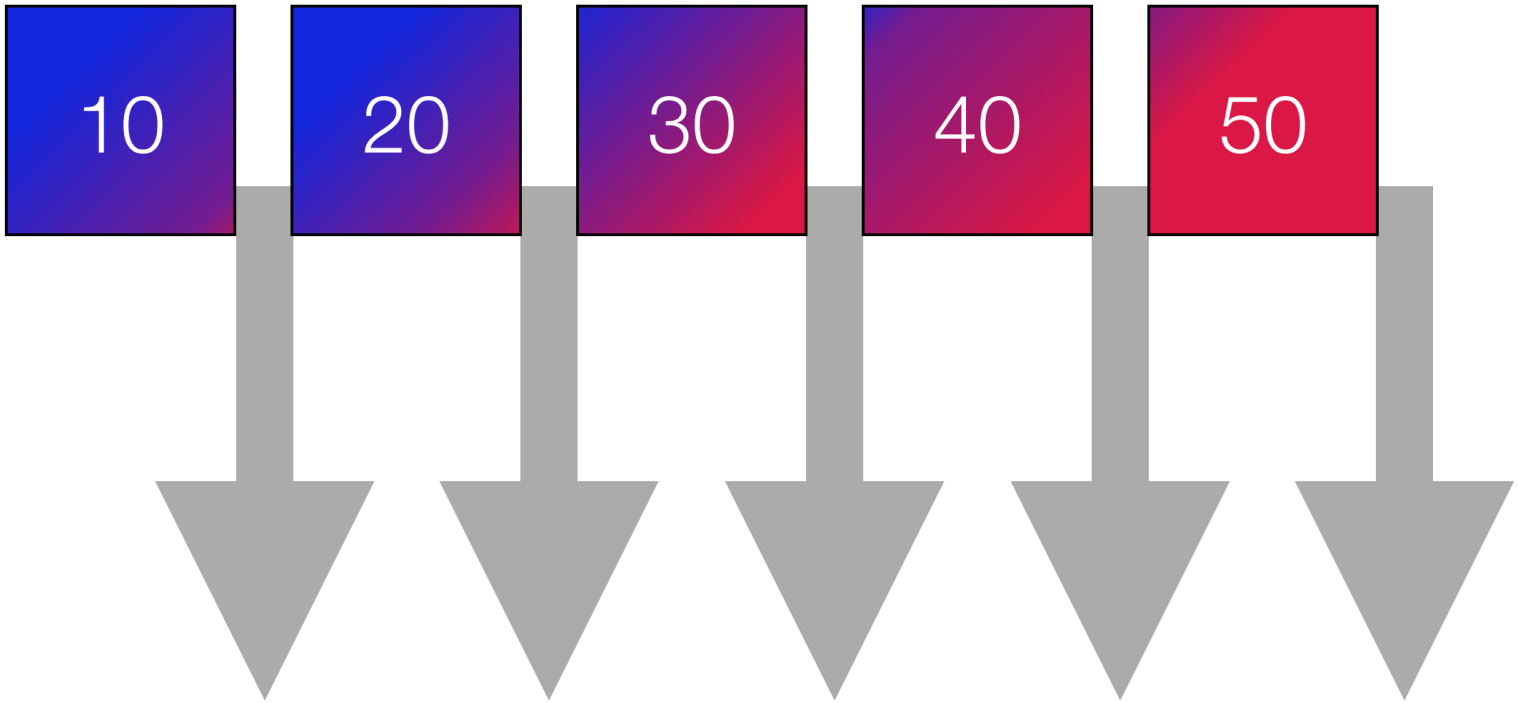
Simulation Output



Simulation Output



Simulation Output

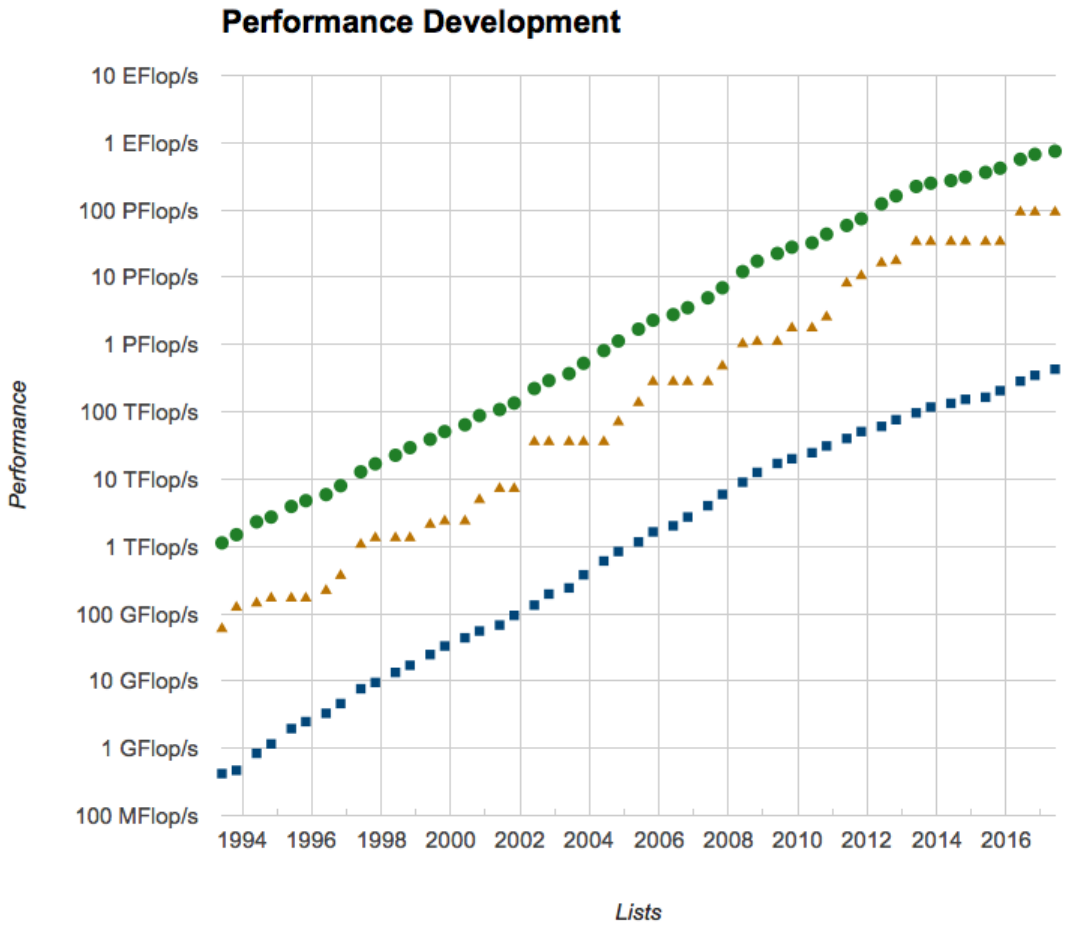


I/O Performance hasn't changed

Motivation

As computation performance increases
I/O must be re-evaluated.

Top-500 History



Initial IO-500 Effort

Site	PFLOPs	Peak IO (GiB/s)[1]
KAUST, SAU	7.2	1955.78
JCAHPC, JP	24.91	1918.52
RIKEN, JP	10.62	1510.85
NCSA, US	13.4	1158
LLNL, US	20.1	1000
NSCG, CN	59.6	1000
ORNL, US	27.1	1000

[1] vi4io.org

Challenges for IO-500

Which metrics matter?

Challenges for IO-500

- Storage capacity

Challenges for IO-500

- Storage capacity
- Storage hierarchy

Challenges for IO-500

- Storage capacity
- Storage hierarchy
- Performance / bandwidth

Challenges for IO-500

- Storage capacity
- Storage hierarchy
- Performance / bandwidth
- In-system memory size

Challenges for IO-500

- Storage capacity
- Storage hierarchy
- Performance / bandwidth
- In-system memory size
- Metadata performance

Challenges for IO-500

- Storage capacity
- Storage hierarchy
- Performance / bandwidth
- In-system memory size
- Metadata performance

Challenges for IO-500

- Storage capacity
- Storage hierarchy
- Performance / bandwidth
- In-system memory size
- Metadata performance

Easy to "game" the system

Draft IO-500

- Two workloads: IO and Metadata
- Two measurements: Easy and Hard

Draft IO-500

IO-Easy: IOR

Large, sequential IO on unique POSIX files

IO-Hard: IOR

Small, random IO on a shared POSIX file

MD-Easy: mdtest

Unique directories, empty files

MD-Hard: MD-REAL-IO

Complex metadata, 3900 byte file

Measuring I/O Performance

- Benchmarking
- Proxy Applications
- Profiling

Benchmarking

- IOR
- mdtest
- IO_Bench
- MPI Tile IO
- b_eff_io
- SPIOBENCH
- iozone
- MADbench2

Mainly testing POSIX interface, with some MPI-IO.

Proxy Applications

- MACSio
- HACC_IO / GenericIO

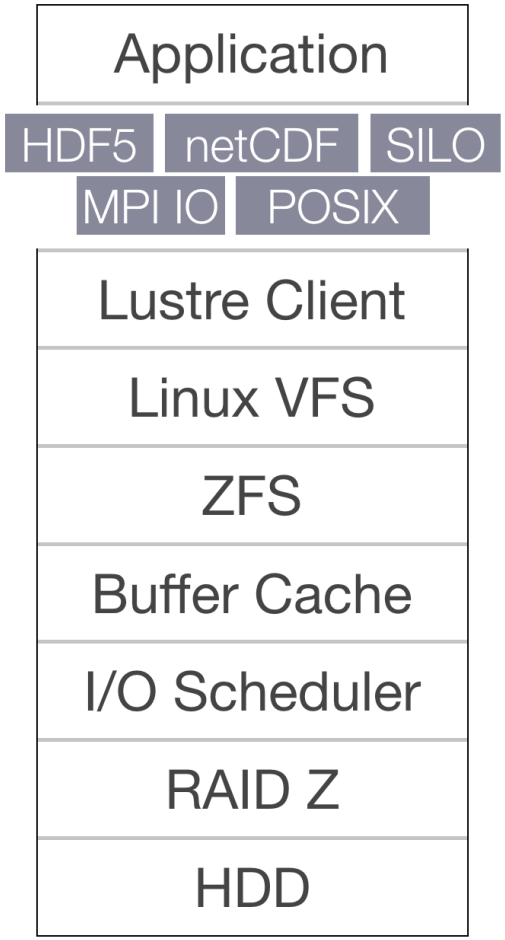
Profiling

- Darshan
- Vampir

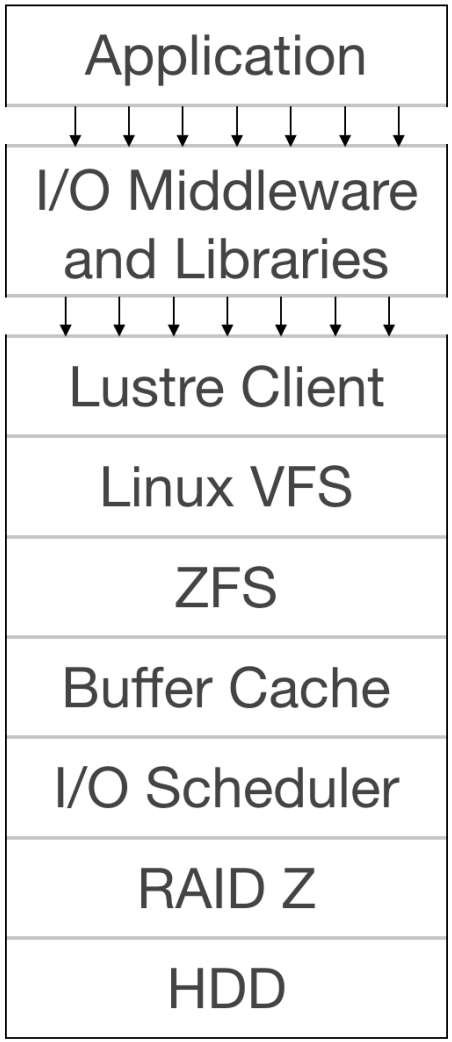
The I/O Stack

John Bent, Seagate

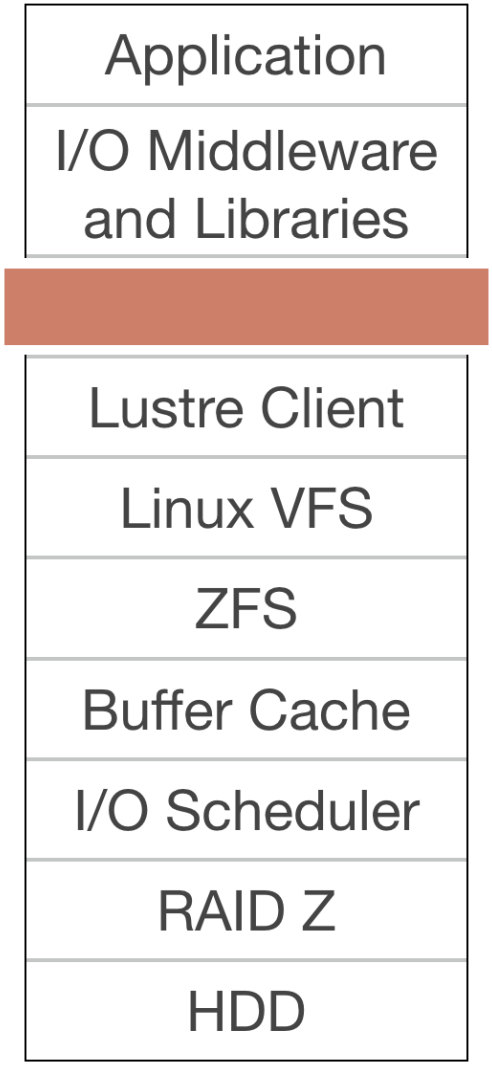
The I/O Stack


Application
HDF5
netCDF
SILO
MPI IO
POSIX
Lustre Client
Linux VFS
ZFS
Buffer Cache
I/O Scheduler
RAID Z
HDD

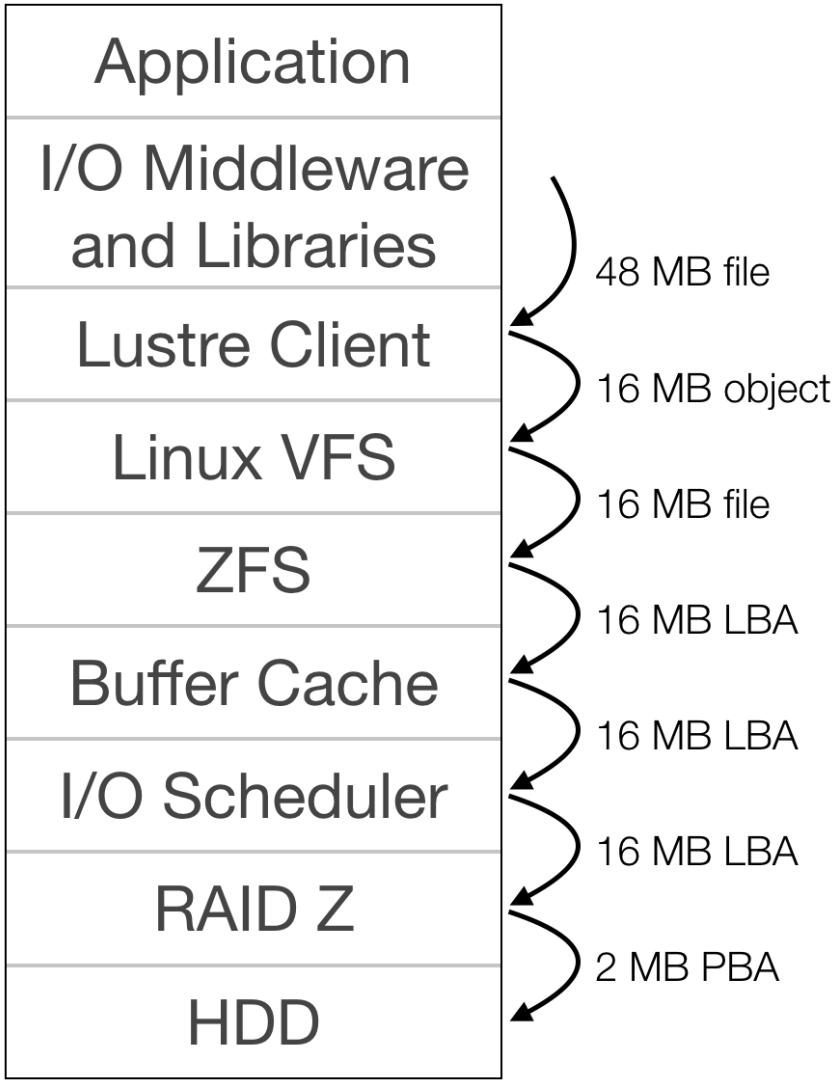
The I/O Stack


Application
I/O Middleware and Libraries
Lustre Client
Linux VFS
ZFS
Buffer Cache
I/O Scheduler
RAID Z
HDD

The I/O Stack


Application
I/O Middleware and Libraries
Lustre Client
Linux VFS
ZFS
Buffer Cache
I/O Scheduler
RAID Z
HDD

The I/O Stack



John Bent, Seagate

Questions from Applications

1. Where do we fall in the I/O envelope?
2. Parameters to achieve best performance?
3. How do we best use new storage tiers?

Current examples and some unposed questions

Where do we fall in the I/O Envelope?

Given:

- Peak system I/O performance
- Current application performance
- I/O pattern or trace
- . . . other details?

Answer:

- Where is the application losing performance?
- What will gains can be made?

Where do we fall in the I/O Envelope?

Current Examples

- Use IOR and mdtest to measure peak system performance
- I/O Specific proxy application
- Lots of work

Where do we fall in the I/O Envelope?

Unposed Questions

- What is the point of this I/O?
- Could this use-case be achieved in a more efficient way?
- How do we enable in-situ or co-situ processes?

High-level questions

Parameters to achieve best performance?

Given:

- Tuning of peak performing benchmark
- Current application I/O

Answer:

- What file system settings need to be tuned?
- Is metadata a bottleneck / file locking?

Parameters to achieve best performance?

Current Examples

- None.
- Validation of simulation models with counters, no analysis of real applications

Parameters to achieve best performance?

Unposed Questions

- Can any of this be detected at a lower level?
- Automatic tuning of the file system during a workload
- How can this drive future procurements?

Lower level and inter-level questions

How do we best use new Storage Tiers?

Given:

- Scientific need
- System limitations

Answer:

- Which I/O patterns perform best
- Resiliency models

How do we best use new Storage Tiers?

Current Examples

- Defensive I/O Assumption
  - Optimal checkpoint interval
  - SCR with system-specific configuration
- Lossy compressions
  - HDF5 ZFP Compression

How do we best use new Storage Tiers?

Unposed Questions

- Interactions between resource schedulers and application
  - pre-stage / post-stage
  - dynamic job allocation resources
- What is the scientific need? How much precision is needed?
- Work flows to manage data movement

Questions requiring full-stack knowledge

Thank you