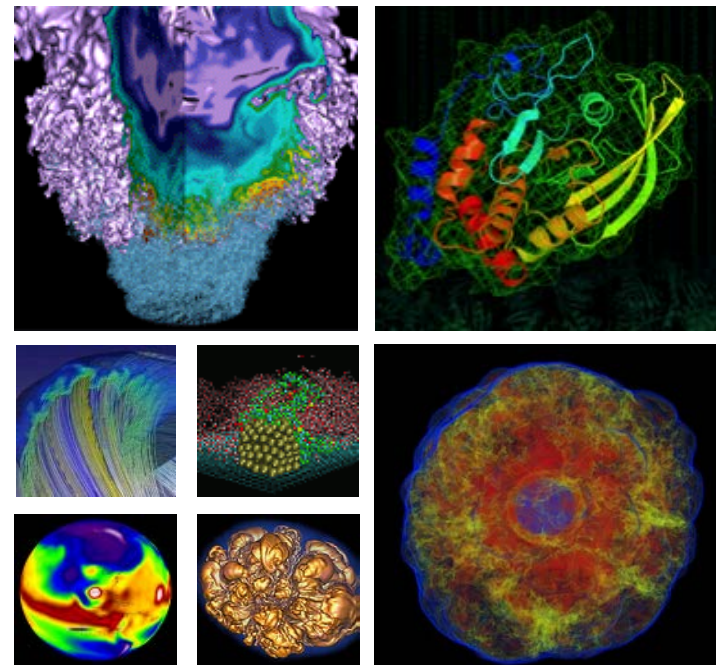


Architecting a 30 PB all - flash file system

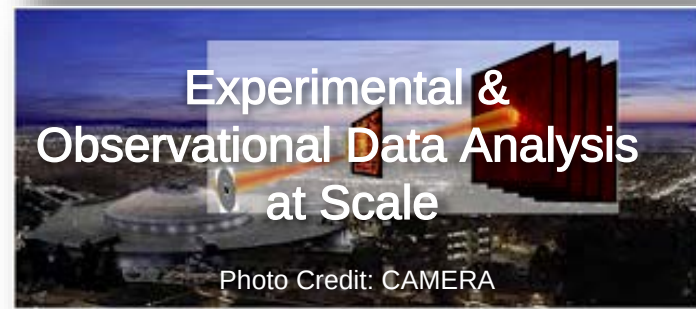
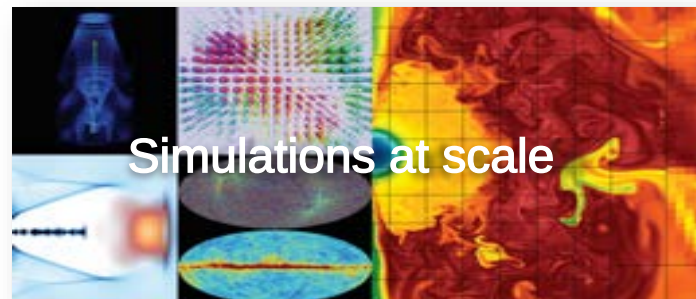


Kirill Lozinskiy
Glenn K. Lockwood et al.

NERSC @ Berkeley Lab (LBNL)



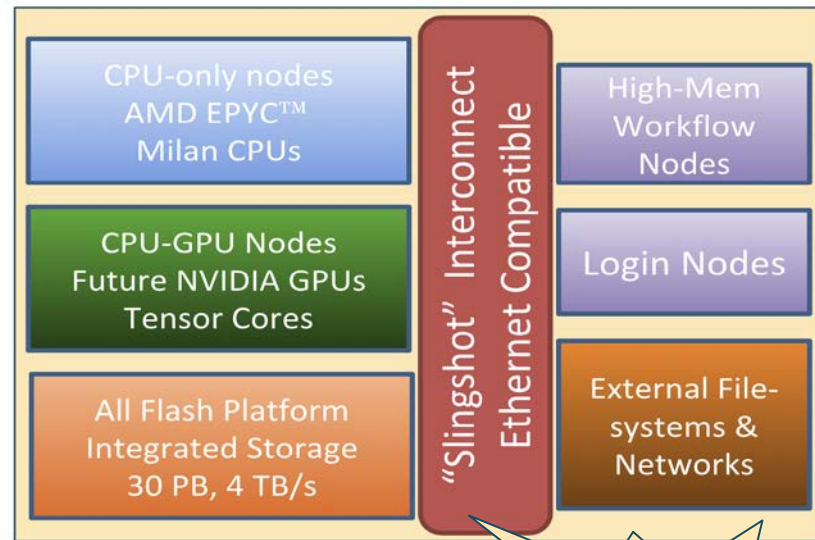
- NERSC is the mission HPC computing center for the DOE Office of Science
- HPC and data systems for the broad Office of Science community
- 7,000 Users, 870 Projects, 700 Codes
- >2,000 publications per year
- 2015 Nobel prize in physics supported by NERSC systems and data archive
- Diverse workload type and size:
 - Biology, Environment, Materials, Chemistry, Geophysics, Nuclear Physics, Fusion Energy, Plasma Physics, Computing Research
- New experimental and AI -driven workloads



NERSC's 2020 System, Perlmutter



- Designed for both large scale simulation and data analysis from experimental facilities
- Overall 3x to 4x capability of Cori
- Includes both NVIDIA GPU - accelerated and AMD CPU -only nodes
- Slingshot Interconnect
- Single Tier, All -Flash Lustre scratch filesystem



NERSC-9's All -Flash Architecture

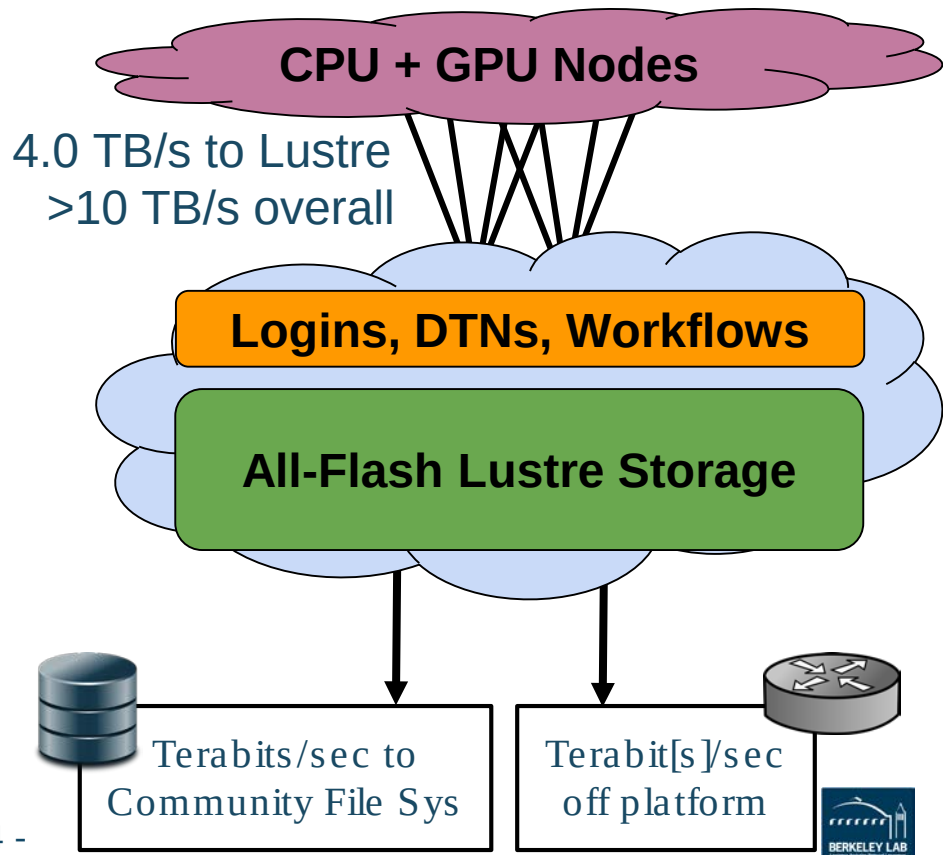


Fast across many dimensions

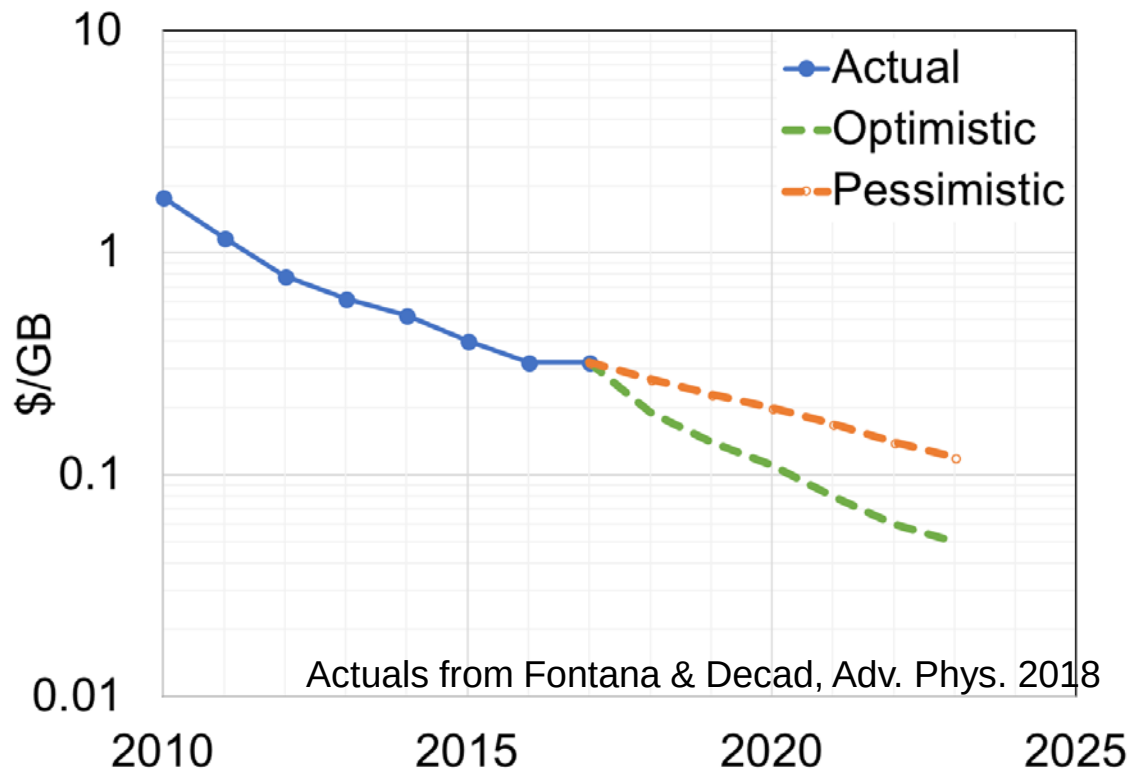
- 30 PB usable capacity
- ≥ 4 TB/s sustained bandwidth
- $\geq 7,000,000$ IOPS
- $\geq 3,200,000$ file creates/sec

Integrated network, separate groups

- Storage/logins remain up when compute is down
- No LNET routers between compute and storage



Myth: only DOE can afford all -flash



Myth: only DOE can afford all -flash



Data management policy

A measure of time between purge cycles
or time after which files are eligible for purging

Reference
system capacity
change

Minimum capacity of
Perlmutter scratch

$$C^{\text{new}} = \text{SSI} \cdot \left(\frac{\lambda_{\text{purge}}}{\text{PF}} \right) \cdot \left(\frac{\partial C^{\text{ref}}}{\partial t} \right)$$

Sustained System Improvement
3x - 4x output capacity over Cori

Desired capacity to
be reclaimed

Change in time

Myth: only DOE can afford all -flash



$\left(\frac{\partial C^{\text{ref}}}{\partial t}\right)$ Mean daily growth projected for Perlmutter at 133 TB/day

$\left(\frac{\lambda_{\text{purge}}}{\text{PF}}\right)$ Data retention policy for Perlmutter is at $\text{time} > 28$ days

- OK to purge after that time
- Each purge aims to remove or migrate 50% of the total capacity

SSI Anticipated 3x to 4x sustained system improvement

C^{new} Minimum Perlmutter capacity is between 22 PB and 30 PB

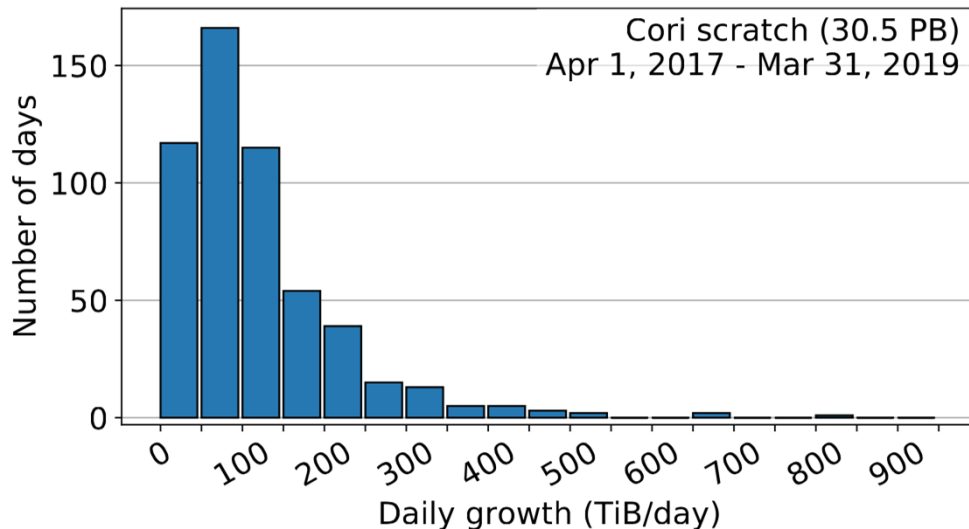


Figure 1 - Distribution of daily growth of Cori's scratch

Myth: Need high -endurance drives for HPC



Drive Writes Per Day required for Perlmutter

File System Writes Per Day of Cori's total write volume

Perlmutter parity blocks

$$DWPD^{new} = SSI \cdot FSWPD^{ref} \cdot \left(\frac{D + P}{D} \right) \cdot WAF$$

Sustained System Improvement
3x - 4x output capacity over Cori

Perlmutter data blocks

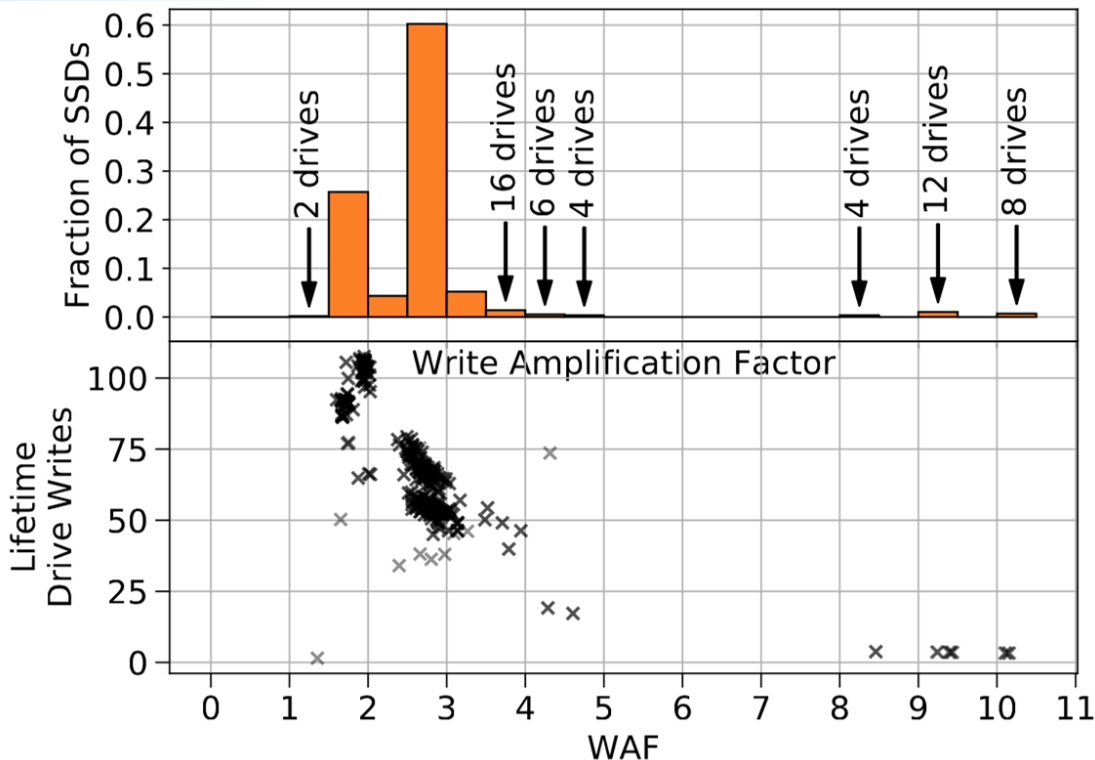
Write Amplification Factor

Myth: Need high -endurance drives for HPC



Measurements from Cori's burst buffer after 3.4 years in service

- WAF bottom quartile: 2.68
- WAF upper quartile: 3.17

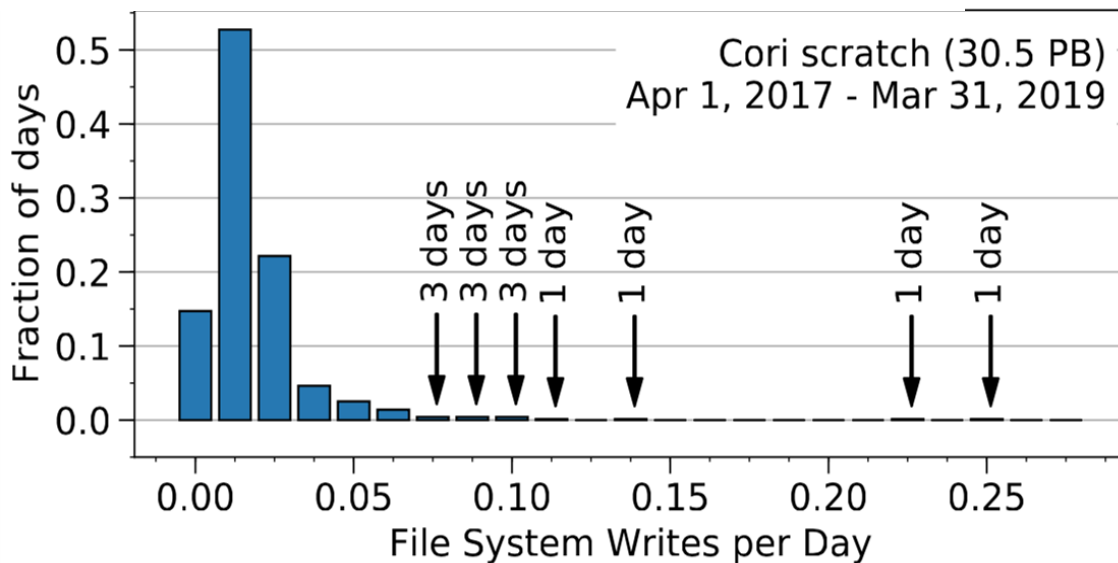


Myth: Need high -endurance drives for HPC



$$\text{DWPD}^{\text{new}} = \text{SSI} \cdot \text{FSWPD}^{\text{ref}} \cdot \left(\frac{D + P}{D} \right) \cdot \text{WAF}$$

- SSI: **3x – 4x**
- Mean FSWPD on 30.5 PB file system: **0.024**
- D, P = **8+2** or **10+2**
- WAF = **2.68 – 3.17**
- DWPD needed: **0.23 – 0.38**



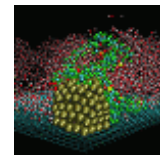
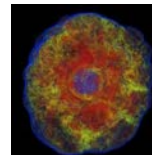
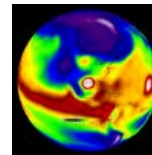
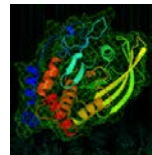
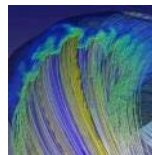
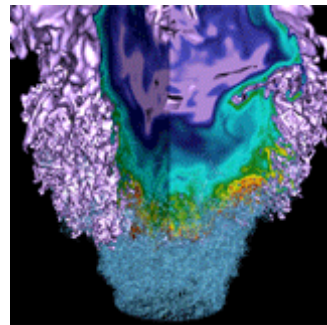
Myth: Lustre is terrible



	New all-flash FS #1	New all-flash FS #2	Lustre*
Read bandwidth	1 TB/s	4 TB/s	4 TB/s
Write bandwidth	0.75 TB/s	1.5 TB/s	4 TB/s
Read IOPS	15 MIOPS	300 MIOPS	7 MIOPS
Write IOPS	3 MIOPS	30 MIOPS	7 MIOPS
Usable capacity	40 PB	30 PB	30 PB
Maturity	GA < 2 years	GA < 2 years	GA > 16 years
Openness	Open protocols, closed source	Closed protocols, closed source	GPL

* All Lustre numbers are lower bounds. Other numbers derived from reference architectures. **ALL NUMBERS ARE SPECULATIVE.**

Metadata Configuration



U.S. DEPARTMENT OF
ENERGY

Office of
Science



Metadata Configuration

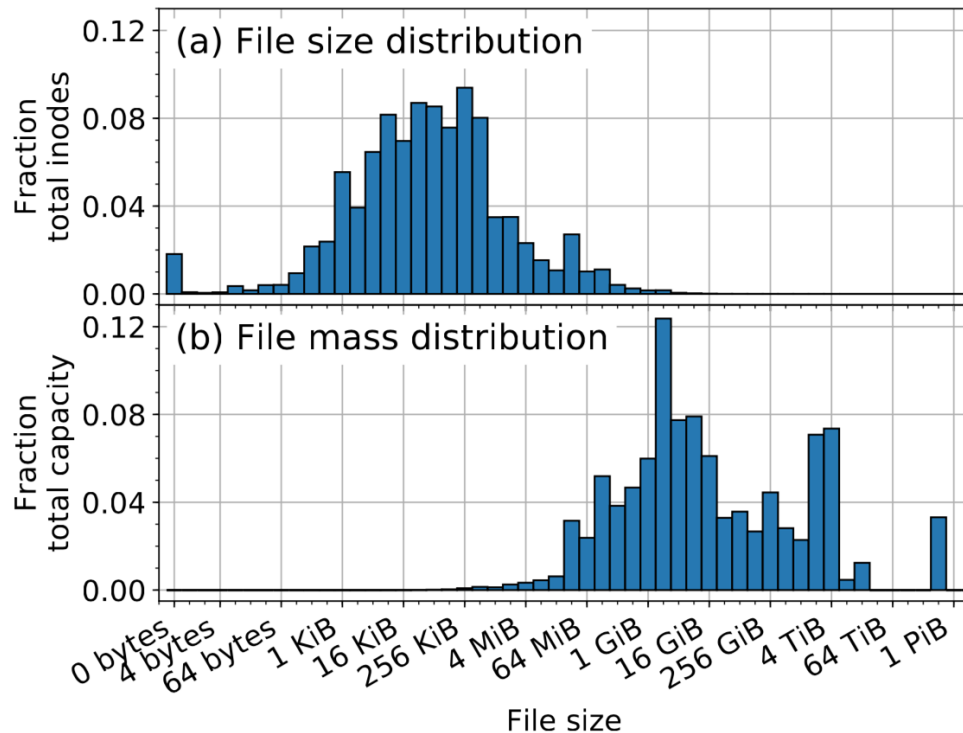


Figure 4 - Probability distribution of file size and file mass on Cori's file system in January 2019

95% of the files comprise only 5% of the capacity used

MDT capacity for a new system is a function of the *expected* file size distribution

- Average file size alone is not enough because HPC file size distribution skews towards small files
- Small changes to the mean file size could represent a significant change to where the optimal DOM size threshold should be



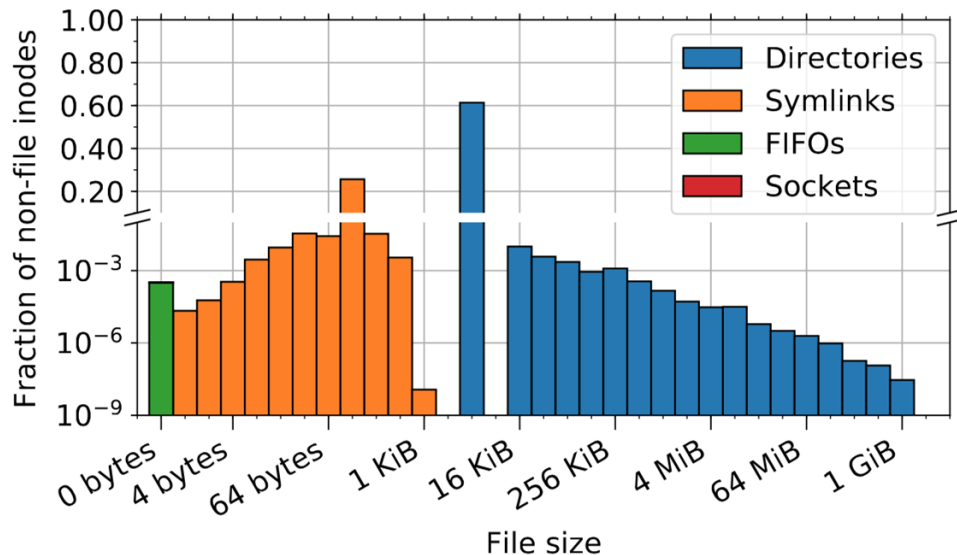
Metadata Configuration



Figure 5 - Probability distribution of inode sizes on Cori's file system in January 2019

MDT Capacity Required for Inodes

- Lustre reserves 4 KiB of MDT capacity per inode
- BUT Directories with millions of files are significantly larger
- Most extreme case is 1 GiB in size for 8 million child inodes



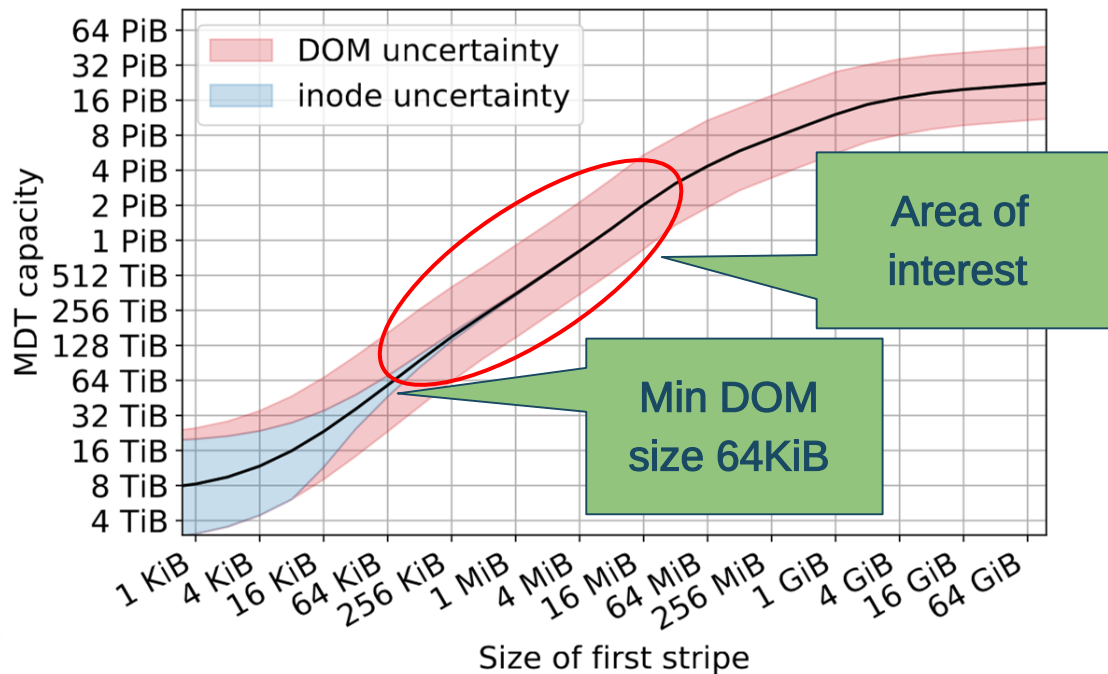
Metadata Configuration



Figure 6 - Required MDT capacity as a function of DOM threshold

Shaded area bounded by the minimum and maximum estimated requirements dictated by the DOM component and the inode capacity component of MDT capacity

- At a very small DOM threshold, the large number of small files does not consume much MDT space
- At a very large DOM threshold, the great majority of files are stored entirely within the MDT, and only a small number of very large files dictates a higher MDT capacity





Thank You
(and we're hiring)