



# PARALLEL FILE SYSTEM SURVEY

Tracey Wilson  
Al Stutz  
Dr. Paul Buerger

Roger Panton  
Michelle Parker  
Armen Ezekielian

File system performance is a key component of user application efficiency in high performance computing (HPC) data centers. Storage administrators and data center directors must choose wisely to select the correct file system to meet their users' requirements and properly adjust settings for optimal performance. *But which file system is the best for each situation?* Today, there is no direct correlation between parallel file systems (PFS) and many of the decisions are based upon vendor marketing. Avetec's HPC Research Division, DICE Program (Data Intensive Computing Environment), in partnership with the Department of Energy's Sandia National Laboratories, is conducting a project to develop a comprehensive benchmarking framework for evaluating file systems. This project also seeks to develop a normalization factor between different file system architectures. HPC community involvement is key to providing a viable tool and understanding all the factors for an accurate comparison, especially for file systems at large scale.

## I. BACKGROUND & SURVEY OBJECTIVES

In support of this project, DICE surveyed 27 HPC data center representatives concerning parallel file systems, benchmarks and I/O measurement, and trace collection.

The survey was a web-based survey conducted in late Spring 2010. An invitation to participate in the survey was provided to:

- HPC Data Center Managers,
- HPC Administrators,
- Storage Administrators,
- Archive Administrators and
- HPC Technology Architects.

## II. KEY FINDINGS

**1. Data centers storing large amounts of data are more likely to store data online.** Over 88 percent of the respondents have more than 500 Terabytes of stored data (54 percent have more than 2 Petabytes) and of these, 58 percent are managing between 10 million and 1 billion files of online storage.

**2. Most respondents (77 percent) have more than 50 percent of their spinning discs dedicated to parallel file systems.** Data centers storing the most data also experienced the most growth in storage requirements.

**3. Over 84 percent of the centers tune and evaluate their storage system** (42.3 percent do so frequently and 42.3 percent occasionally). Fifteen percent rarely tune and these are primarily centers with less than 2 Petabytes of stored data.

**4. Over 70 percent use file system or storage subsystem benchmarking tools and the frequency of doing so is situational-based, not calendar-based,** e.g., upon system arrival, problem assessment and resolution, design or evaluation, etc.. The most frequently cited file systems used are NFS and Lustre. The most frequently cited benchmarking tools are IOR, IOZone and Bonnie. The overarching reasons for using a particular benchmarking tool were for performance validation and tracking/performance reference.

**5. Bandwidth is considered the most important aspect critical to performance metrics followed by multi-stream performance and metadata operations.** The most commonly used file systems were NFS (74 percent) and GFFS (48% use).

**6. Over 40 percent regularly validate configuration parameters using real-time traces of applications** (an additional 15 percent were uncertain). Over 22 percent regularly validate trace results to assess overall system health/performance. The most commonly cited tool used for application profiling was *strace*. Some of the centers use tools (e.g., VAMPIR or Darshan), while some conduct manual comparisons.

**7. Issues still encountered after benchmarking and tuning** included jitter, poor reliability, rebuild degraded operations, undersized I/O systems halting entire system, sluggish/delayed response time and system bottlenecks, data integrity, file system resiliency and throughput.

**8. Sixty-three percent of the respondents agree that there is a need for a normalization approach in benchmarking.**

**9. Over 59 percent of the respondents indicated interest in participating in the PFS research effort.** Refer to page 22 for actual respondent comments as to How they would be willing to participate.

**10. Only 19 percent of the respondents are willing to provide actual trace data and an additional 44 percent are unsure.** The primary reason was privacy issues/sensitivity of data as well as the "hassle factor" due to workload, time involved and the need to get permission.

The next section of this report provides results and analysis for each question along with actual comments provided by the respondents.

### III. SURVEY SUMMARY by QUESTION

#### I. Data Center & Respondent Profile

This section summarizes the role of the respondents and characteristics of their data center including computational technical areas (CTAs), file size and annual growth in terabytes and number of files.

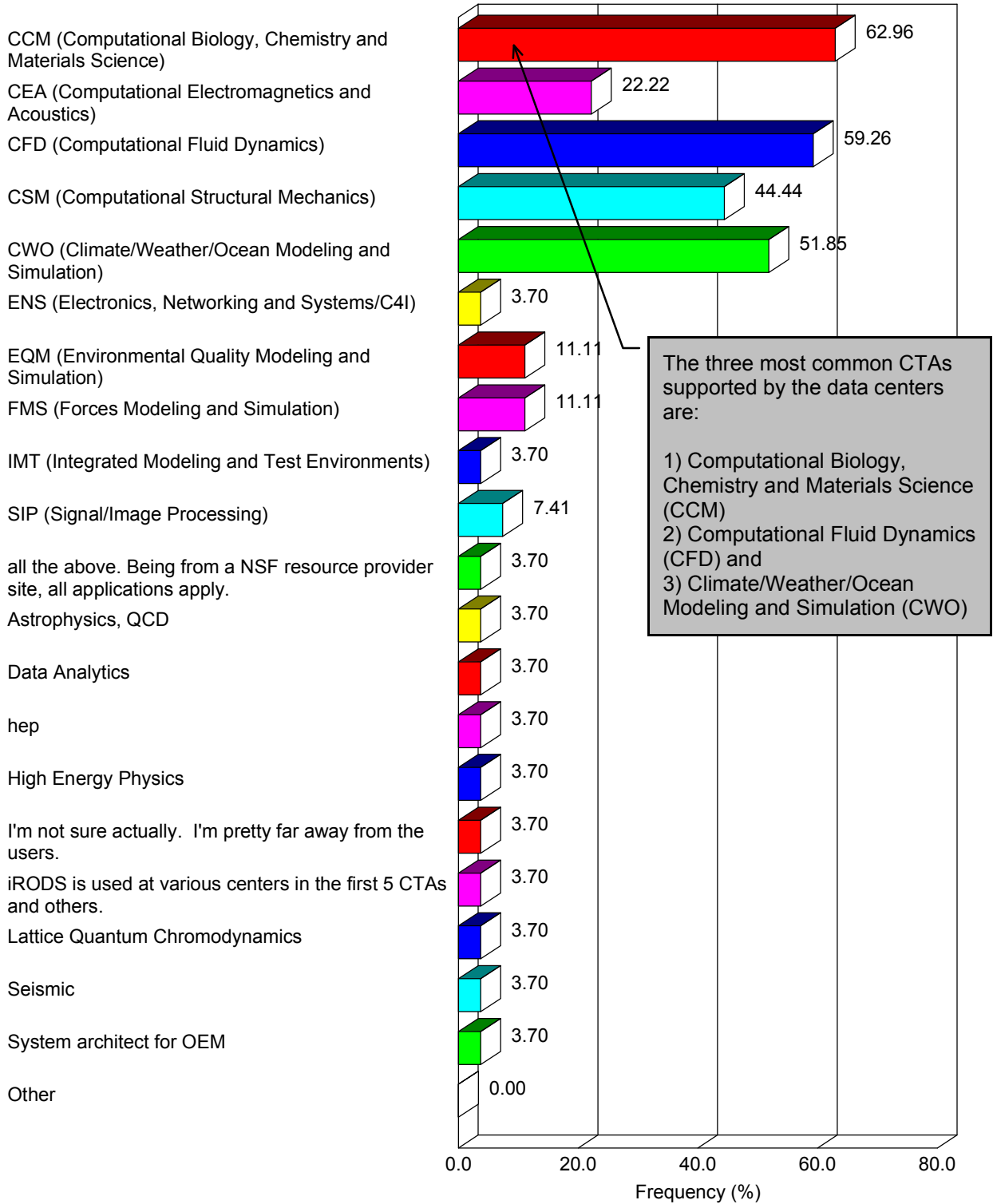
#### Respondent's Role in File Systems Management

Role in storage subsystems and file systems management			
	Counts	Percents	Percents
			0 100
Design	4	14.8%	
Evaluate	3	11.1%	
Influencer and/or Selection Authority	4	14.8%	
End User	0	0.0%	
Operational Management	8	29.6%	
Senior Management	4	14.8%	
Design, Evaluate, User and influencer all with about equal weight	1	3.7%	
I design, evaluate, select and then senior manage the projects that need storage solutions and file systems.	1	3.7%	
I/We design software (iRODS) that runs on top of file systems	1	3.7%	
principal investigator for government research contracts that buy and operate storage	1	3.7%	
Other	0	0.0%	
Totals	27	100.0%	

The most common role of the respondents was Operational Management.

## Primary Computational Technical Areas (CTAs) Data Center Supports

Primary computational technical areas (CTAs)



\* Note: Multiple answer percentage-count totals not meaningful.

### Size of Stored Data

Amount of stored data		
	Percents	Percents
		0 100
500 Terabytes or less	11.5%	
501 Terabytes - 2 Petabytes	34.6%	
Over 2 Petabytes	53.8%	
Totals	100.0%	

Over 88 percent of the respondents have greater than 500 Terabytes of stored data with approximately 54 percent storing more than 2 Petabytes.

### Size of Stored Data Compared to CTAs the Data Centers Support

	Overall	Amount of stored data		
		Over 2 Petabytes 53.8%	501 Terabytes - 2 Petabytes 34.6%	500 Terabytes or less 11.5%
Primary computational technical areas (CTAs)				
CCM (Computational Biology, Chemistry and Materials Science)	63.0%	57.1%	88.9%	33.3%
CEA (Computational Electromagnetics and Acoustics)	22.2%	21.4%	11.1%	66.7%
CFD (Computational Fluid Dynamics)	59.3%	64.3%	66.7%	33.3%
CSM (Computational Structural Mechanics)	44.4%	50.0%	44.4%	33.3%
CWO (Climate/Weather/Ocean Modeling and Simulation)	51.9%	57.1%	55.6%	33.3%
EQM (Environmental Quality Modeling and Simulation)	11.1%	14.3%	11.1%	0.0%
FMS (Forces Modeling and Simulation)	11.1%	7.1%	22.2%	0.0%
SIP (Signal/Image Processing)	7.4%	7.1%	11.1%	0.0%
Other	40.7%	50.0%	22.2%	33.3%
Totals	*	*	*	*

\* Note: Multiple answer percentage-count totals not meaningful.

	Overall	Number of files on online storage		
		10M - 1 Billion (B) 79.2%	More than 1B 12.5%	Fewer than 10 Million (M) 8.3%
Amount of stored data				
500 Terabytes or less	11.5%	5.3%	0.0%	50.0%
501 Terabytes - 2 Petabytes	34.6%	36.8%	0.0%	50.0%
Over 2 Petabytes	53.8%	57.9%	100.0%	0.0%
Totals	100.0%	100.0%	100.0%	100.0%

### Number of Online Storage Files

Number of files on online storage		
	Percents	Percents
		0 100
Fewer than 10 Million (M)	8.3%	
10M - 1 Billion (B)	79.2%	
More than 1B	12.5%	
Totals	100.0%	

Data centers that are storing larger amounts of data are storing data online. Of the 54 percent who have over 2 Petabytes of stored data, 58 percent of these respondents are managing between 10 million and 1 billion (B) files on online storage.

### Spinning Files Dedicated to Parallel File Systems

Percentage of spinning discs dedicated to parallel file systems		
	Counts	Percents
		0 100
10% or Less	2	
11% - 20%	3	
21% - 30%	0	
31% - 40%	0	
41% - 50%	1	
Over 50%	20	
Totals	26	

Most of the respondents have greater than 50 percent of their spinning discs dedicated to parallel file systems.

### Annual Growth in NUMBER OF FILES

Average annual growth: Number of files	
40	1
100000	2
750000	1
1000000	1
2000000	3
5000000	2
10000000	1
15000000	1
50000000	1
80000000	1
90000000	1
150000000	1
200000000	3
Totals	19
Mean	53313160.00
Median	5000000.00

### Annual Growth in TERABYTES

Average annual growth: In Terabytes of Storage	
50	1
75	1
100	2
150	1
168	1
250	1
268	1
300	2
440	1
500	1
900	1
1000	1
2000	2
3000	2
3500	1
4300	1
8192	1
Totals	21
Mean	1456.81
Median	440.00

The average growth in storage requirements was provided in both Terabytes and number of files. Results are summarized below:

Average Annual Growth in **Terabytes**: 1,456 Terabytes with a median value of 440 Terabytes. The largest growth was cited at 8192 Terabytes.

Average Annual Growth in **Number of Files**: 53,313,160 files with a median value of 5 million files.

### Growth in TERABYTES by Overall Size of Stored Data

	Overall		Amount of stored data						
			Over 2 Petabytes 53.8%		501 Terabytes - 2 Petabytes 34.6%		500 Terabytes or less 11.5%		
Average annual growth: In Terabytes of Storage									
50	4.8%	1	0.0%	0	11.1%	1	0.0%	0	
75	4.8%	1	0.0%	0	11.1%	1	0.0%	0	
100	9.5%	2	0.0%	0	11.1%	1	100.0%	1	
150	4.8%	1	0.0%	0	11.1%	1	0.0%	0	
168	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
250	4.8%	1	0.0%	0	11.1%	1	0.0%	0	
268	4.8%	1	0.0%	0	11.1%	1	0.0%	0	
300	9.5%	2	0.0%	0	22.2%	2	0.0%	0	
440	4.8%	1	0.0%	0	11.1%	1	0.0%	0	
500	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
900	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
1000	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
2000	9.5%	2	18.2%	2	0.0%	0	0.0%	0	
3000	9.5%	2	18.2%	2	0.0%	0	0.0%	0	
3500	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
4300	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
8192	4.8%	1	9.1%	1	0.0%	0	0.0%	0	
Totals	100.0%	21	100.0%	11	100.0%	9	100.0%	1	
Mean	1456.81		← 2596.36		214.78		100.00		

Data centers storing the most data also experienced the most growth in storage requirements, i.e., data centers with over 2 Petabytes of data experienced more growth in storage requirements (average of 2,596 Terabytes) than those with 501 Terabytes to 2 Petabytes of data, which experienced an average of 215 Terabytes growth in storage requirements.

### Growth in NUMBER OF FILES by Overall Size of Stored Data

	Overall	Amount of stored data					
		Over 2 Petabytes 53.8%		501 Terabytes - 2 Petabytes 34.6%		500 Terabytes or less 11.5%	
Average annual growth: Number of files							
40	5.3% 1	8.3% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
100000	10.5% 2	16.7% 2	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
750000	5.3% 1	0.0% 0	16.7% 1	0.0% 0	0.0% 0	0.0% 0	
1000000	5.3% 1	0.0% 0	16.7% 1	0.0% 0	0.0% 0	0.0% 0	
2000000	15.8% 3	16.7% 2	16.7% 1	0.0% 0	0.0% 0	0.0% 0	
5000000	10.5% 2	0.0% 0	16.7% 1	100.0% 1	0.0% 0	0.0% 0	
10000000	5.3% 1	0.0% 0	16.7% 1	0.0% 0	0.0% 0	0.0% 0	
15000000	5.3% 1	8.3% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
50000000	5.3% 1	8.3% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
80000000	5.3% 1	0.0% 0	16.7% 1	0.0% 0	0.0% 0	0.0% 0	
90000000	5.3% 1	8.3% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
150000000	5.3% 1	8.3% 1	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
200000000	15.8% 3	25.0% 3	0.0% 0	0.0% 0	0.0% 0	0.0% 0	
Totals	100.0% 19	100.0% 12	100.0% 6	100.0% 1	100.0% 1	100.0% 1	
Mean	53313160.00	75766672.00	16458333.00	5000000.00			

As shown in the previous table and this table, data centers with the most amount of stored data also experienced the largest growth in storage requirements, both in "Terabytes" and "Number of Files."

### How Often Tune & Evaluate Storage System

Frequency of tuning and evaluating storage system	Amount of stored data				
	Overall	Over 2 Petabytes	501 Terabytes - 2 Petabytes	500 Terabytes or less	Percents
					0 100
Frequently	42.3%	50.0%	33.3%	33.3%	
Occasionally	42.3%	42.9%	44.4%	33.3%	
Rarely	15.4%	7.1%	22.2%	33.3%	
Never	0.0%	0.0%	0.0%	0.0%	
Totals	100.0%	100.0%	100.0%	100.0%	

Over 84 percent of the data centers tune and evaluate their storage system either frequently or occasionally. Fifteen percent rarely tune and these are primarily data centers with less than 2 Petabytes of stored data.

## II. File System & Benchmarking

To assess current practices in benchmarking, respondents were asked questions about type, frequency and importance of metrics and benchmarking tools. They were also asked what performance metrics are the most important to benchmarking. Summary statistics are provided below:

- 1) Over 70 percent of the respondents use file system or storage subsystem benchmarking tools.
- 2) The frequency in which benchmarking is used is not calendar-based, but situation based, i.e., for those who do use benchmarking tools, they use it during design or evaluation periods, upon system arrival and problem-assessment resolution.
- 3) The most frequently cited file systems used at the data centers are NFS and Lustre.
- 4) The most frequently cited benchmarking tools used are IOR, IOZone and Bonnie.
- 5) The overarching reason for using a particular benchmarking tool were performance validation and tracking/performance reference. None of the respondents indicated that any tool was "mandated."
- 6) Respondents cited *bandwidth* as the most important aspect critical to performance metrics followed by *multi-stream performance* and *metadata operations*.
- 7) The most cited benchmarking tool to which this would be applied was aggregate bandwidth followed by metadata operations and multi-stream performance.

### Data Centers' Use of Benchmarking Tools

2. Does your data center use file system or storage subsystem benchmarking tools?			
	Counts	Percents	Percents
			0 100
Yes	19	70.4%	
No	8	29.6%	
Totals	27	100.0%	
Mean	1.70		

## File Systems Used at Data Center

	Overall	2. Does your data center use file system or storage subsystem benchmarking tools?	
		Yes 70.4%	No 29.6%
File systems used at Data Center			
CIFS/SMB	14.8%	10.5%	25.0%
CXFS	29.6%	31.6%	25.0%
GPFS	48.1%	57.9%	25.0%
Lustre	59.3%	68.4%	37.5%
NFS	74.1%	84.2%	50.0%
PanFS	29.6%	36.8%	12.5%
pNFS	11.1%	10.5%	12.5%
PVFS2	18.5%	15.8%	25.0%
Redhat GFS	11.1%	5.3%	25.0%
StorNext	7.4%	10.5%	0.0%
XFS	25.9%	31.6%	12.5%
ZFS	7.4%	10.5%	0.0%
Other	25.9%	21.1%	37.5%
Totals	*	*	*

Of the 70 percent who do use file system or storage subsystem benchmarking tools, the most common file systems used are NFS, Lustre and GPFS.

\* Note: Multiple answer percentage-count totals not meaningful.

## Benchmarking Tools Used (File System or Storage Subsystem)

Benchmarking tools being used

- 12 IOR
- 12 IOZone
- 7 Bonnie
- 5 IOBench
- 3 b\_eff\_io
- 2 mib
- 2 NAS PB-IO
- 2 SPIObench
- 1 MADbench2
- 1 mpi-tile-io
- 1 PARKBENCH
- 1 PDSI
- 1 S3aSim
- 1 Actual end-user workloads
- 1 flash I/O benchmark
- 1 fs\_test.
- 1 in house custom
- 1 LANL's MPI-IO test, Terasort, GridMix
- 1 Locally developed synthetic benchmark for random access
- 1 mdtest, fdtree
- 1 mdtest, LANL fs\_test
- 1 xdd
- 1 xdd and local tool called havoc
- 1 xdd, iometer, SPC-1/2,HNQ
- 1 xdd, md\_test and benchmarks created in house
- 0 FileBench
- 0 HPIO

The most common benchmarking tools used are IOR and IOZone.

- 0 PIORAW
- 0 Other

**Users of Benchmarking: *Frequency of Use by Type of Benchmarking Tool***

	Respondents Who Use Benchmarking Tools				
	Overall	Frequency of Running Benchmarks			
		Upon system arrival	Problem-assessment resolution	During design or evaluation periods	Quarterly
		26.4%	24.5%	22.6%	7.5%
Benchmarking tools being used					
b_eff_io	15.8%	7.1%	7.7%	16.7%	25.0%
Bonnie	36.8%	42.9%	38.5%	50.0%	25.0%
IOBench	26.3%	21.4%	15.4%	8.3%	75.0%
IOR	63.2%	57.1%	53.8%	58.3%	50.0%
IOZone	63.2%	57.1%	61.5%	58.3%	50.0%
mib	10.5%	7.1%	7.7%	16.7%	0.0%
NAS PB-IO	10.5%	14.3%	15.4%	8.3%	0.0%
SPIObench	10.5%	14.3%	7.7%	8.3%	0.0%
Other	63.2%	57.1%	69.2%	66.7%	50.0%
Totals	*	*	*	*	*

	Respondents Who Use Benchmarking Tools			
	Frequency of Running Benchmarks			
	Semi-annually	Monthly	Rarely	Other
	7.5%	5.7%	1.9%	3.8%
Benchmarking tools being used				
b_eff_io	0.0%	0.0%	0.0%	0.0%
Bonnie	50.0%	33.3%	0.0%	0.0%
IOBench	25.0%	0.0%	100.0%	0.0%
IOR	75.0%	66.7%	0.0%	50.0%
IOZone	100.0%	66.7%	100.0%	50.0%
mib	0.0%	0.0%	0.0%	0.0%
NAS PB-IO	25.0%	0.0%	0.0%	0.0%
SPIObench	0.0%	33.3%	0.0%	0.0%
Other	75.0%	66.7%	0.0%	100.0%
Totals	*	*	*	*

\* Note: Multiple answer percentage-count totals not meaningful.

## Benchmarking Tools Used by Computational Technical Area (CTA)

	Benchmarking tools being used				
	b_eff_io	Bonnie	IOBench	IOR	IOZone
Overall	15.8%	36.8%	26.3%	63.2%	63.2%
Primary computational technical areas (CTAs)					
CCM (Computational Biology, Chemistry and Materials Science) 20.0%	15.4%	38.5%	30.8%	69.2%	69.2%
CFD (Computational Fluid Dynamics) 18.8%	15.4%	38.5%	30.8%	69.2%	61.5%
CWO (Climate/Weather/Ocean Modeling and Simulation) 16.5%	20.0%	50.0%	20.0%	70.0%	60.0%
CSM (Computational Structural Mechanics) 14.1%	10.0%	30.0%	30.0%	60.0%	60.0%
CEA (Computational Electromagnetics and Acoustics) 7.1%	25.0%	25.0%	25.0%	75.0%	50.0%
EQM (Environmental Quality Modeling and Simulation) 3.5%	0.0%	66.7%	33.3%	100.0%	100.0%
FMS (Forces Modeling and Simulation) 3.5%	0.0%	50.0%	50.0%	50.0%	50.0%
SIP (Signal/Image Processing) 2.4%	0.0%	0.0%	50.0%	100.0%	50.0%
ENS (Electronics, Networking and Systems/C4I) 1.2%	0.0%	0.0%	0.0%	100.0%	100.0%
IMT (Integrated Modeling and Test Environments) 1.2%	0.0%	0.0%	0.0%	100.0%	0.0%
all the above. Being from a NSF resource provider site, all applications apply. 1.2%	0.0%	0.0%	100.0%	100.0%	100.0%
Astrophysics, QCD 1.2%	100.0%	100.0%	0.0%	100.0%	0.0%
Data Analytics 1.2%	0.0%	0.0%	0.0%	100.0%	100.0%
hep 1.2%					
High Energy Physics 1.2%	0.0%	100.0%	0.0%	0.0%	0.0%
I'm not actually. I'm pretty far away from the users. 1.2%	0.0%	0.0%	0.0%	0.0%	0.0%
iRODS is used at various centers in the first 5 CTAs and others. 1.2%					
Lattice Quantum Chromodynamics 1.2%	100.0%	100.0%	0.0%	0.0%	100.0%
Seismic 1.2%					
System architect for OEM 1.2%	0.0%	0.0%	0.0%	100.0%	100.0%
Other 0.0%					

	Benchmarking tools being used				
	mib	NAS PB-IO	SPIObench	Other	Totals
Overall	10.5%	10.5%	10.5%	63.2%	*
Primary computational technical areas (CTAs)					
CCM (Computational Biology, Chemistry and Materials Science) 20.0%	7.7%	15.4%	15.4%	53.8%	*
CFD (Computational Fluid Dynamics) 18.8%	15.4%	15.4%	15.4%	53.8%	*
CWO (Climate/Weather/Ocean Modeling and Simulation) 16.5%	20.0%	20.0%	10.0%	70.0%	*
CSM (Computational Structural Mechanics) 14.1%	10.0%	10.0%	20.0%	60.0%	*
CEA (Computational Electromagnetics and Acoustics) 7.1%	0.0%	0.0%	25.0%	50.0%	*
EQM (Environmental Quality Modeling and Simulation) 3.5%	33.3%	33.3%	33.3%	33.3%	*
FMS (Forces Modeling and Simulation) 3.5%	50.0%	0.0%	0.0%	100.0%	*
SIP (Signal/Image Processing) 2.4%	0.0%	0.0%	50.0%	0.0%	*
ENS (Electronics, Networking and Systems/C4I) 1.2%	0.0%	0.0%	0.0%	100.0%	*
IMT (Integrated Modeling and Test Environments) 1.2%	100.0%	0.0%	0.0%	100.0%	*
all the above. Being from a NSF resource provider site, all applications apply. 1.2%	0.0%	0.0%	0.0%	100.0%	*
Astrophysics, QCD 1.2%	0.0%	0.0%	0.0%	0.0%	*
Data Analytics 1.2%	0.0%	0.0%	0.0%	100.0%	*
hep 1.2%					*
High Energy Physics 1.2%	0.0%	0.0%	0.0%	100.0%	*
I'm not sure actually. I'm pretty far away from the users. 1.2%	0.0%	0.0%	0.0%	100.0%	*
iRODS is used at various centers in the first 5 CTAs and others. 1.2%					*
Lattice Quantum Chromodynamics 1.2%	0.0%	0.0%	0.0%	0.0%	*
Seismic 1.2%					*
System architect for OEM 1.2%	0.0%	0.0%	0.0%	100.0%	*
Other 0.0%					*

\* Note: Multiple answer percentage-count totals not meaningful.

**Users of Benchmarking: Reason for Using by Type of Benchmarking Tool**

	Respondents Who Use Benchmarking Tools			
	Overall	Why use these benchmarks		
		Performance Validation	Tracking and performance reference	
		57.9%	31.6%	
Benchmarking tools being used				
b_eff_io	15.8% 3	27.3% 3	0.0% 0	
Bonnie	36.8% 7	36.4% 4	33.3% 2	
IOBench	26.3% 5	36.4% 4	16.7% 1	
IOR	63.2% 12	72.7% 8	50.0% 3	
IOZone	63.2% 12	72.7% 8	50.0% 3	
mib	10.5% 2	0.0% 0	16.7% 1	
NAS PB-IO	10.5% 2	18.2% 2	0.0% 0	
SPIObench	10.5% 2	9.1% 1	16.7% 1	
Other	63.2% 12	45.5% 5	83.3% 5	
Totals	* *	* *	* *	

	Respondents Who Use Benchmarking Tools			
	Why use these benchmarks			
	Problem assessment and/or resolution	All of the above except mandated - with about equal emphasis		
	5.3%	5.3%		
Benchmarking tools being used				
b_eff_io	0.0% 0	0.0% 0		
Bonnie	0.0% 0	100.0% 1		
IOBench	0.0% 0	0.0% 0		
IOR	100.0% 1	0.0% 0		
IOZone	0.0% 0	100.0% 1		
mib	100.0% 1	0.0% 0		
NAS PB-IO	0.0% 0	0.0% 0		
SPIObench	0.0% 0	0.0% 0		
Other	100.0% 1	100.0% 1		
Totals	* *	* *		

\* Note: Multiple answer percentage-count totals not meaningful.

## Aspects of Performance Metrics Critical to Data Center

Most important metric's performance aspects			
	Counts	Percents	0 Percents 100
# of Reads	9	33.3%	<div style="width: 33.3%;"></div>
# of Writes	9	33.3%	<div style="width: 33.3%;"></div>
Bandwidth	21	77.8%	<div style="width: 77.8%;"></div>
Data Deletion	3	11.1%	<div style="width: 11.1%;"></div>
Data Purge	3	11.1%	<div style="width: 11.1%;"></div>
File Open Rates	8	29.6%	<div style="width: 29.6%;"></div>
IOPs	11	40.7%	<div style="width: 40.7%;"></div>
Large Block Data Performance	16	59.3%	<div style="width: 59.3%;"></div>
Metadata Operations	18	66.7%	<div style="width: 66.7%;"></div>
Multi-stream Performance	18	66.7%	<div style="width: 66.7%;"></div>
Random Performance	14	51.9%	<div style="width: 51.9%;"></div>
Sequential Performance	11	40.7%	<div style="width: 40.7%;"></div>
Single Stream Performance	6	22.2%	<div style="width: 22.2%;"></div>
Small Block Data Performance	8	29.6%	<div style="width: 29.6%;"></div>
Other	3	11.1%	<div style="width: 11.1%;"></div>
Totals	27	n/a	

Bandwidth is considered the most important performance metric for a data center followed by metadata operations and multi-stream performance.

## Most Important Metric or Statistical Measure for Benchmarking

5. Currently, no single community benchmark has all the metrics the HPC community requires. Please describe the single most important statistical measure or metric you would recommend to enhance benchmarking.

- A benchmark that synthesizes block level traces of actual workloads
- Metadata performance (create, unlink, stat)
- This is a silly question. There is not one most important measure. We have multiple important applications. I'd say that the two most important measures are N-N write bandwidth and N-1 small strided unaligned write bandwidth.
- parallel radom access
- Real world applications of importance that are not likely to change with new hardware.
- Aggregate Throughput, and the sustained maximum multi-client throughput that adds up to the same number as the aggregate
- Correctness checking, i.e. make sure what you wrote out is the expected data, that the operations completed, etc.
- Aggregate bandwidth from all client nodes
- aggregate read/write throughput from O(100) clients
- It depends on the application. I would look at data movement metrics from device to defice.
- aggregate bandwidth
- parallel reads and writes from mulitple hosts and using libraries such has HDF or netcpdf
- Small block data performance, but also random performance
- After benchmarking all theses years we have look at single large block sequential IO performance as key, but real workload is there are 100 jobs doing different things, therefore 50% read/write of random IO with N steams with different block sizes (i.e. 20 streams each with a different block size) performance would be more realistic of what the users do on the systems.
- Operations per second for varying file/block sizes.

- Well as your question put it no single benchmark has them all and this is no single that is most important to us
  - a mixed workload full campus resource. We require a balance in small/large block random and sequential io/iops/bandwidth/throughput/metadata etc.
- random access.stream speed

### III. Trace Data & Results

#### Validate Configuration Parameters using Real-time Traces

Regularly validate configuration parameters using real-time traces of applications			
	Counts	Percents	Percents
			0 100
Yes	11	40.7%	
Uncertain	4	14.8%	
No	12	44.4%	
Totals	27	100.0%	
Mean	1.96		

#### Compare Trace Results to Assess Overall System Health/Performance

	Overall		Regularly validate configuration parameters using real-time traces of applications			
	27		Yes 40.7%, 11	Uncertain 14.8%, 4	No 44.4%, 12	
Compare trace results against anticipated results/modeling tools						
Yes	22.2%	6	45.5%	5	0.0%	0
Uncertain	22.2%	6	27.3%	3	75.0%	3
No	55.6%	15	27.3%	3	25.0%	1
Totals	100.0%	27	100.0%	11	100.0%	4
Mean	1.67		2.18		1.75	
						1.17

## Trace Tools Currently Used for Application Profiling

Trace tools currently using for application profiling on storage subsystems and file systems:	
Do Not Use Trace Tools	27.3%
io_profile	9.1%
IBM HPCT	4.5%
OProfile	4.5%
strace	31.8%
TAU	13.6%
collectl	4.5%
custom macros	4.5%
Darshan	4.5%
Locally developed test harness	4.5%
not sure I have a name for other tools	4.5%
PMaC	4.5%
Vampir	4.5%
Other	0.0%
Totals	*

The most common trace tool currently being used is "strace." Refer to the comments below for an explanation of how respondents compare trace results.

\* Note: Multiple answer percentage-count totals not meaningful.

## How Data Center Compares Trace Results to Modeling Tools or Anticipated Results

Please describe HOW your center compares trace results against anticipated results or modeling tools.

Regularly validate configuration parameters using real-time traces of applications = Yes

- We use VAMPIR to find bottlenecks in performance
- Against prior runs, runs using different hardware or software, and published results from other installations.
- We capture I/O characterization of most runs using the Darshan tool and assess effectiveness of I/O system and approach used by applications.
- Accurate within 15%
- Manual comparison of results to expected (past) results.

Regularly validate configuration parameters using real-time traces of applications = Uncertain

[None]

Regularly validate configuration parameters using real-time traces of applications = No

- We run a suite of regression tests periodically (generally after each maintenance period) and insert the results into a database. This is done primarily to detect performance regressions or improvements in the OS software.

## Other Performance Measurement or Modeling Tools Data Centers Using

10. In addition to benchmarks and traces, please describe any performance measurement and/or modeling tools your data center uses for storage subsystems and file systems.

- Performance under degraded mode operations.
- Locally developed test harness
- Ganglia
- see above.
- None
- We have a few I/O intensive applications that we use to validate performance after system maintenances to ensure consistent performance
- None
- PSiNS Tracer
- Depends on the filesystems and OS, XFS - topio, sar, topdisk, Lustre - ltop, and local tools.
- Performance (bandwidth and iops) data from backend RAID controller hardware.
- Applications - gaussian 03/09, vasp, and various others as well as homebrew benchmarking apps.

## Issues Still Encountered After Benchmarking and Tuning

11. Regardless of the best tuning practices for a data center or application, some storage systems may still exhibit less than desirable performance in certain areas. After benchmarking and tuning of your systems, please describe




- jitter, rebuild degraded operations, poor reliability
- The I/O subsystem in our primary resource is rather undersized compared to the other capabilities of the machine. Very I/O intensive jobs which use a significant fraction of the machine (>25%) can effectively grind the entire system to a halt. As a result, we've taken to running these very I/O intensive jobs in isolation after maintenance periods.
- Delayed response time due to sluggish metadata server performance
- Stability.
- Interactions between file system and enterprise storage features (e.g., scrubbing). Bottlenecks in system network.
- Data correctness and integrity. I'm thinking of bit flips on the physical media or on the way to being stored. This may not exactly be solely a file system issue, but the file system can help in this area by end to end data checks and checks on data at rest.  
Other file system issues we are facing are; quality of service - can the file system make service guarantees, there are data patterns that are just bad for all parallel file systems like all processes writing to a single file in a strided manner.
- Delays and unexplained random failures of some filesystem operations, or of the file server..
- Large hurdle dealing with tuning of Lustre. Almost need expert system to advise on best approach for given system.
- Lustre still struggles with a few codes (Blast, Blast3d?) that generate large volumes of small files
- Configuring and tuning file system for all types of IO (large and small files).
- File system resiliency and time to repair after an interrupt are metrics not previously captured.  
When an application is encountered which performs differently on multiple file systems or file system types, we have not found a good tool for comparing behavior to highly where (or when) exactly the performance diverges. Note that not all tracing tools have been evaluated yet.
- Throughput - multiple streams. Slow metadata (eg. ls hangs).

#### IV. Interest in Normalizing Tools and Participation in PFS Research

Respondents were asked a few questions to assess interest in participating in the Parallel File System (PFS) research as well as the need for normalization in benchmarking results across various architectures. Sixty-three percent of the respondents see the need for normalization and over 59 percent of the respondents are interested in participating in the PFS research effort.

However, only 18.5 percent are willing to provide actual trace data and an additional 44.4 percent are unsure. See page 21 for supporting comments.

#### See Need for a Normalization Approach in Benchmarking




See the need for an approach that would normalize benchmark results across various architectures			
	Counts	Percents	Percents
			0 100
Yes	17	63.0%	
Uncertain	5	18.5%	
No	5	18.5%	
Totals	27	100.0%	
Mean	2.44		

#### Benchmark Normalizing Techniques Data Centers ARE USING or CONSIDERING

13. Please describe any techniques your data center has used, or is considering, for normalizing benchmarking results between parallel file system configurations (hardware and software).

- Getting away from max MB/GB/sec. not appropriate and does not relate to actual IO patterns.
- Performance per spindle. Performance per IO node. Performance per switch.
- No mechanical normalization, except total price, or perhaps number of spindles, is useful without a careful, situation specific analysis.
- None today, other than comparing single client and aggregate throughputs
- Normalizing against peak sequential I/O rates.
- We don't have any methods yet.
- We attempt to test filesystems on the same hardware test stand and with the same access patterns
- Generally normalized over aggregate bandwidth
- This would be good since we have systems with different architectures and different parallel file systems.
- We have not yet accomplished "normalization". Rather, tests are run on idle file systems which use far less than peak capabilities of each file system, in order to most nearly achieve apples-to-apples comparisons.
- We currently use none - but it would be nice to be able to do so when evaluating.

## Interest in Participating in Parallel File System (PFS) Research

Interest in further participation:			
	Counts	Percents	Percents
			0 <span style="float: right;">100</span>
Yes	16	59.3%	
Uncertain	9	33.3%	
No	2	7.4%	
Totals	27	100.0%	

Over 59 percent of the respondents are interested in participating further in the parallel file research effort. See actual respondent comments below about how they would like to participate.

Please describe how you would like to participate in this Parallel File System research effort.

- to help my customers design better solutions with more realistic expectations.
- We have access to a number of very I/O intensive user applications which might be able to be used as benchmarks.
- I'd like to make sure that proposed benchmarks fit LANL needs. One issue that we've seen with some benchmarks is that they report the bandwidth of the fastest rank; what's meaningful however is the bandwidth of the slowest rank.
- Possibly providing data, input on benchmarks, and open to other participation.
- Help develop i/o use cases apropos to high energy physics.
- I would be interested in evaluating different HW and SW to find a "best fit" approach for parallel file systems. I would also be interested in participating in designing efficient file systems that could handle both small and large I/O patterns.
- Through the DICE node at AFRL.
- Review or testing would be fine.
- Participate in team evaluations, discussions, panels, workshops, etc.
- Review and provide comments on results.
- Hardware onsite integrated into the environment with staff running benchmarks and providing results.
- I would definitely be able to test and evaluate beta software in order to give feedback on bugs, usability, and effectiveness. Providing data and test codes is subject to Government approval for release, but may be possible.
- providing data or serving on a sub-team.

## Willingness to Provide Actual Application Traces

	Would consider providing actual application traces to the community			
	Yes	Uncertain	No	Totals
Overall	18.5% 5	44.4% 12	37.0% 10	100.0% 27
Role in storage subsystems and file systems management				
Operational Management 29.6%	12.5% 1	75.0% 6	12.5% 1	100.0% 8
Design 14.8%	0.0% 0	75.0% 3	25.0% 1	100.0% 4
Influencer and/or Selection Authority 14.8%	0.0% 0	50.0% 2	50.0% 2	100.0% 4
Senior Management 14.8%	0.0% 0	0.0% 0	100.0% 4	100.0% 4
Evaluate 11.1%	100.0% 3	0.0% 0	0.0% 0	100.0% 3
Design, Evaluate, User and influencer all with about equal weight 3.7%	100.0% 1	0.0% 0	0.0% 0	100.0% 1
I design, evaluate, select and then senior manage the projects that need storage solutions and file systems. 3.7%	0.0% 0	0.0% 0	100.0% 1	100.0% 1
I/We design software (iRODS) that runs on top of file systems 3.7%	0.0% 0	0.0% 0	100.0% 1	100.0% 1
principal investigator for government research contracts that buy and operate storage 3.7%	0.0% 0	100.0% 1	0.0% 0	100.0% 1
End User 0.0%	0	0	0	% 0
Other 0.0%	0	0	0	% 0

## Reasons for Not or Uncertain About Providing Trace Data

15b. If no or uncertain, please describe the primary reason why you may not consider providing actual application traces to the research task force.

- LSI does not own the data. several sites are very keen to giving this up in DOD and DOE.
- User privacy issues.
- It is a lot more work than most expect, and real traces have no assured interpretation without a lot of communication with the source of the traces.
- Depends on whether it interferes with operations and/or operational team capacity.
- Headache to deal with getting permission, etc.
- NDA&#12288;System
- Need to be authorized by HPCMP
- Don't have any. :-)
- Time and effort involved in obtaining traces
- It will depend on the load the trace activities will place on systems, networks and support staff.
- May be limited by releasability of the data.
- I would have to check with our security team, and not sure what traces you need.
- Sensitivity of data (FOUO at least)
- Potential for unauthorized release of sensitive data.
- Need to get approval and need to make sure data is not sensitive and likely need to know there are only US citizens looking at the data.
- Release of this data is subject to Government approval.
- HIPPA, proprietary data/applications, and limited resources for participation in such programs.

Full Name of Your Organization:

Willing to be recognized as a participants in this study: = Yes

- LSI
- National Institute for Computational Sciences, University of Tennessee
- P&G
- Texas Advanced Computing Center
- Lawrence Livermore National Lab
- AFRL DSRC
- Maui High Performance Computing Center DSRC
- NASA Ames Research Center
- Raytheon & NOAA
- University of Louisville

Willing to be recognized as a participants in this study: = No