



Gauging File System Performance: DICE Parallel File Systems Benchmarking and Evaluation

Tracey Wilson
DICE, Program Manager
CSC

twilso23@csc.com



Overview

- ◆ **Quick Look at the Issues, Benefits, and Goals**
- ◆ **Survey Results**
- ◆ **Benchmark Framework and Methodology Plan**
- ◆ **Next steps**
- ◆ **How to participate**

Project Team

- ◆ Tracey Wilson, CSC, Project Lead
- ◆ Lee Ward, Sandia National Lab
- ◆ Ed Wahl, Avetec
- ◆ Paul Buerger, Avetec
- ◆ Armen Ezekilian, Avetec

- ◆ DICE TAP Involvement
 - ◆ Dan Duffy, NASA Goddard
 - ◆ Matt Leininger, Lawrence Livermore National Lab

Parallel File System *Issues!*

- ◆ **Lack of Standardized metrics for performance**
 - ◆ Need exists for file system performance comparison
 - ◆ Need exists for defined benchmarks for local and remote file systems
- ◆ **Scaling:**
 - ◆ Scaling of current file systems has unpredictable performance impacts
- ◆ **Load:**
 - ◆ Performance varies with file systems with loads of 50-70%
 - ◆ Fragmented data is not accounted for

Benefits of Standardization

- ◆ **Standardized Metric**
 - ◆ One set of benchmarks for all parallel file systems
 - ◆ Results will be normalized for comparison
 - ◆ HPC Community needs non-biased benchmark
- ◆ **Increased understanding of the impacts on current and proposed upgrades to a center's file system:**
 - ◆ The Storage subsystem performance
 - ◆ Scaling
 - ◆ Load performance

Goals

- ◆ **Parameterized Benchmark Suite**
 - ◆ Simulate many different workloads
 - ◆ Expose more tuning/configuration options in the benchmarks
 - ◆ Suite Extensible so users can add new benchmarks
 - ◆ Have unique benchmarks for synthetic workloads
 - ◆ Transactions
 - ◆ Streaming
 - ◆ Random I/O
 - ◆ Various Read/Write Ratios
- ◆ Normalization is needed between differing architectures
 - ◆ Need a formula
 - ◆ Aging and fragmentation of data should be considered

First Task: Survey the Community

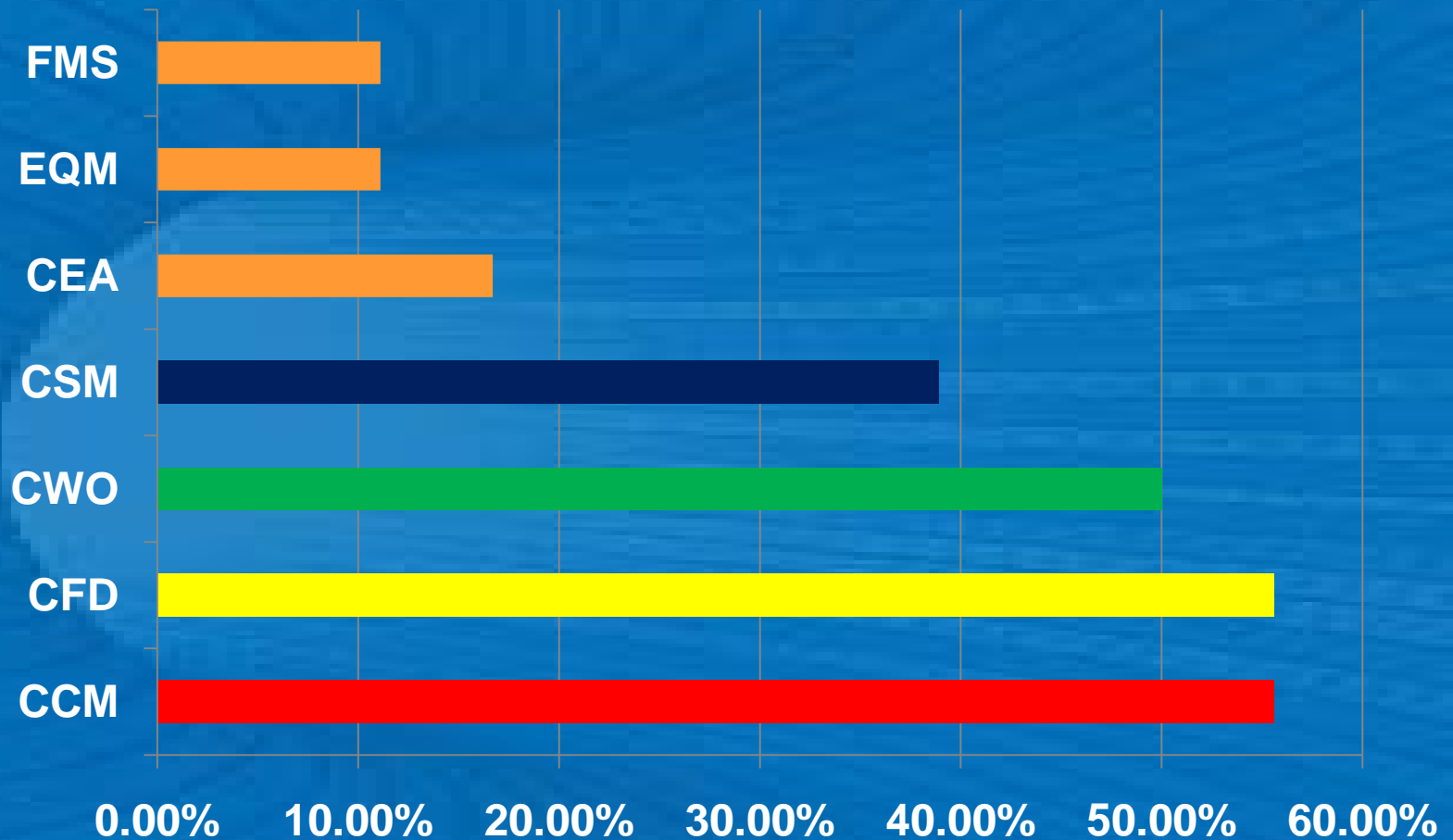
- ◆ DICE created a file system focused survey
- ◆ Targeted
 - ◆ HPC Data Center Managers
 - ◆ HPC Administrators
 - ◆ Storage Administrators
 - ◆ Archive Administrators
 - ◆ HPC Technology Architects
- ◆ Why?
 - ◆ Need to understand what community is measuring and why are they doing it

Survey Scope

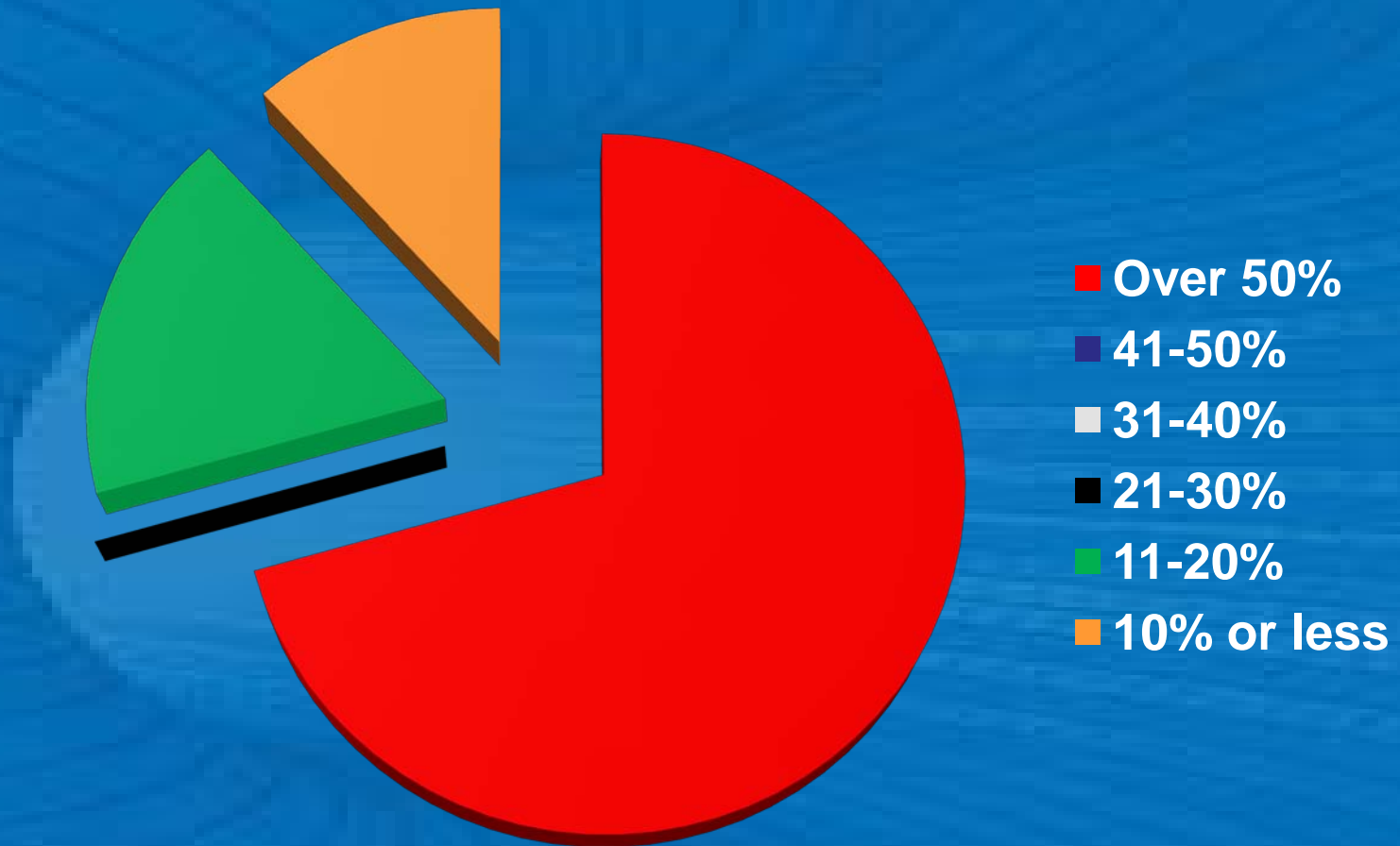
- ◆ Contained 29 questions focused on:
 - ◆ File system use
 - ◆ Benchmarking methods
 - ◆ I/O patterns and traces
 - ◆ What key metrics are valued?
 - ◆ Are there any metrics not covered today?
 - ◆ Are there any other efforts like this underway today?

..So, what did we learn?

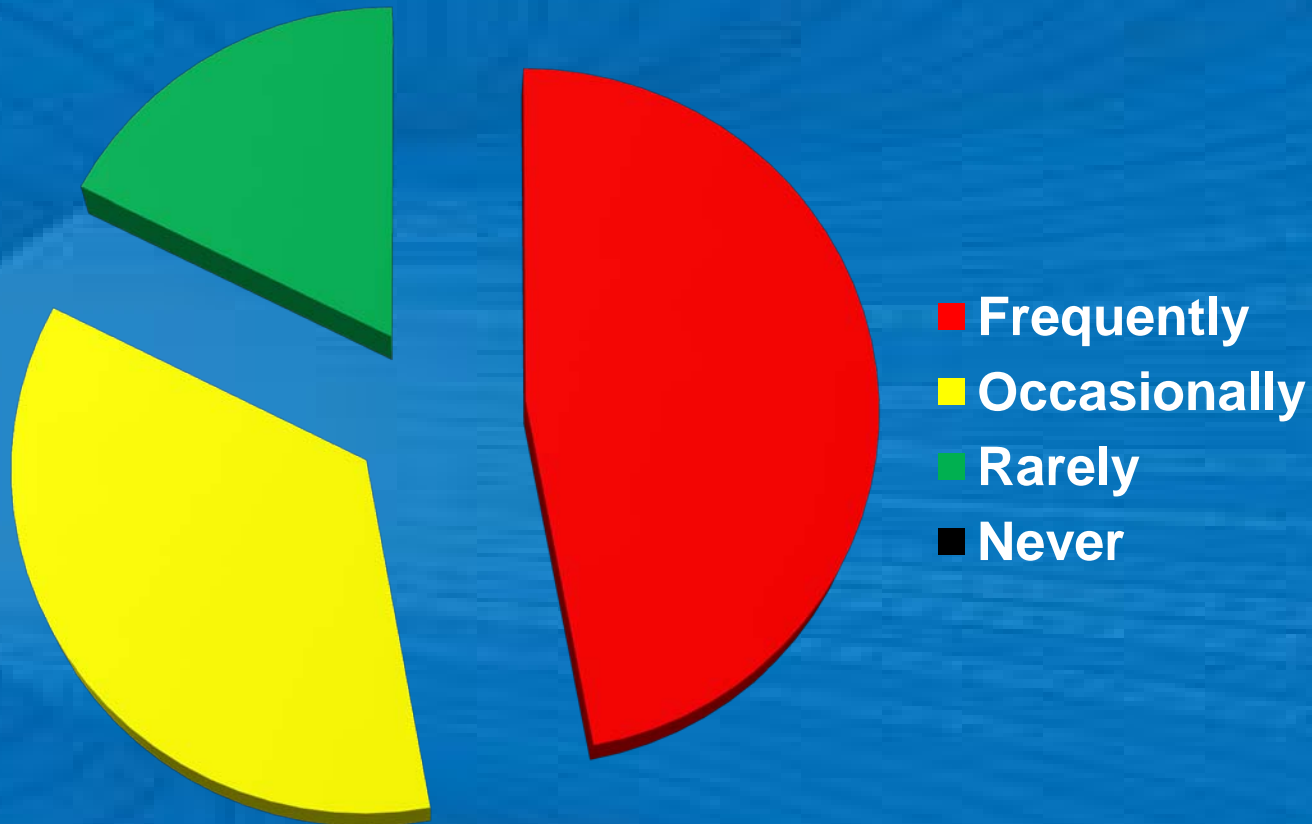
Primary Computational Areas



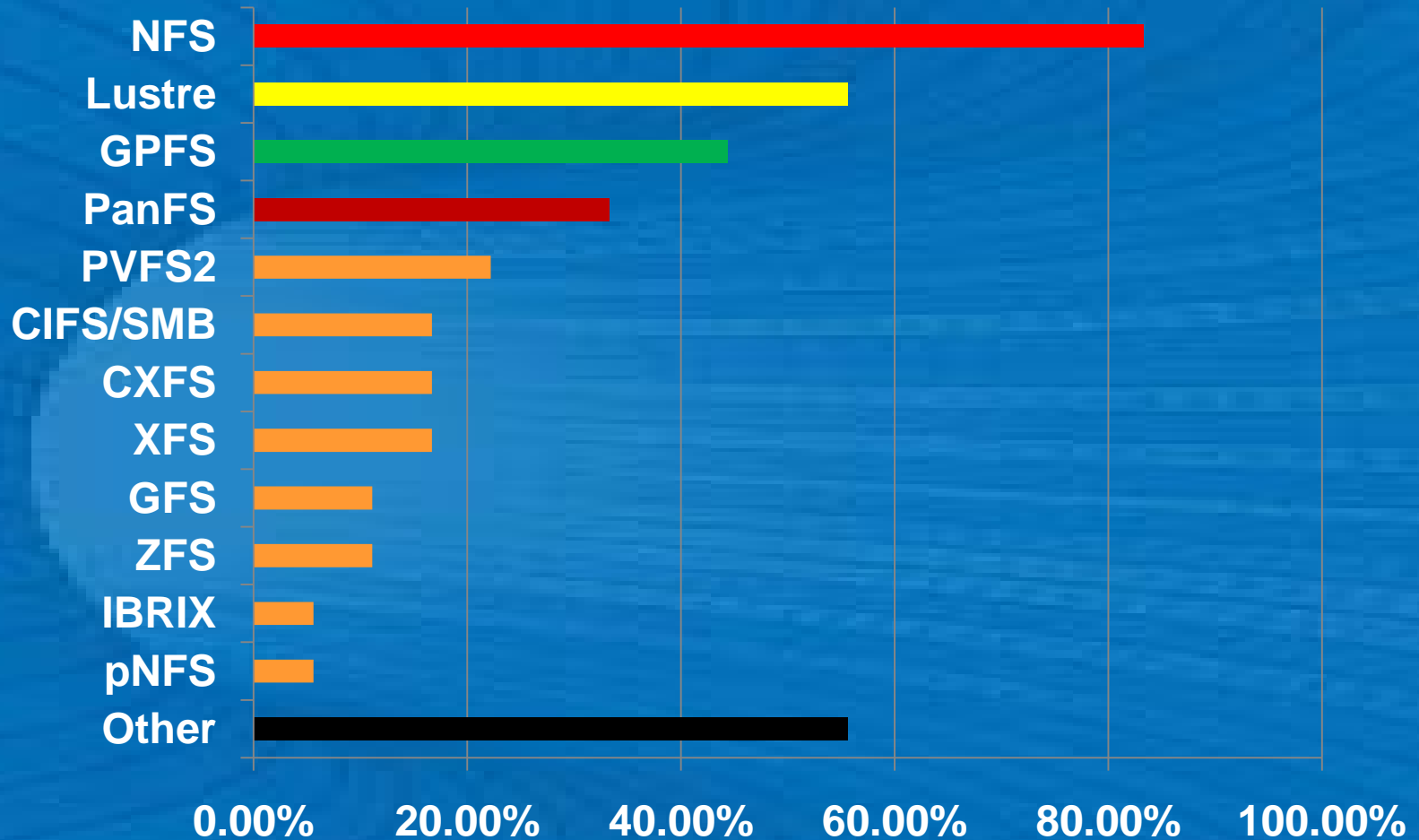
Percentage of Spinning Disks Used for Parallel File Systems



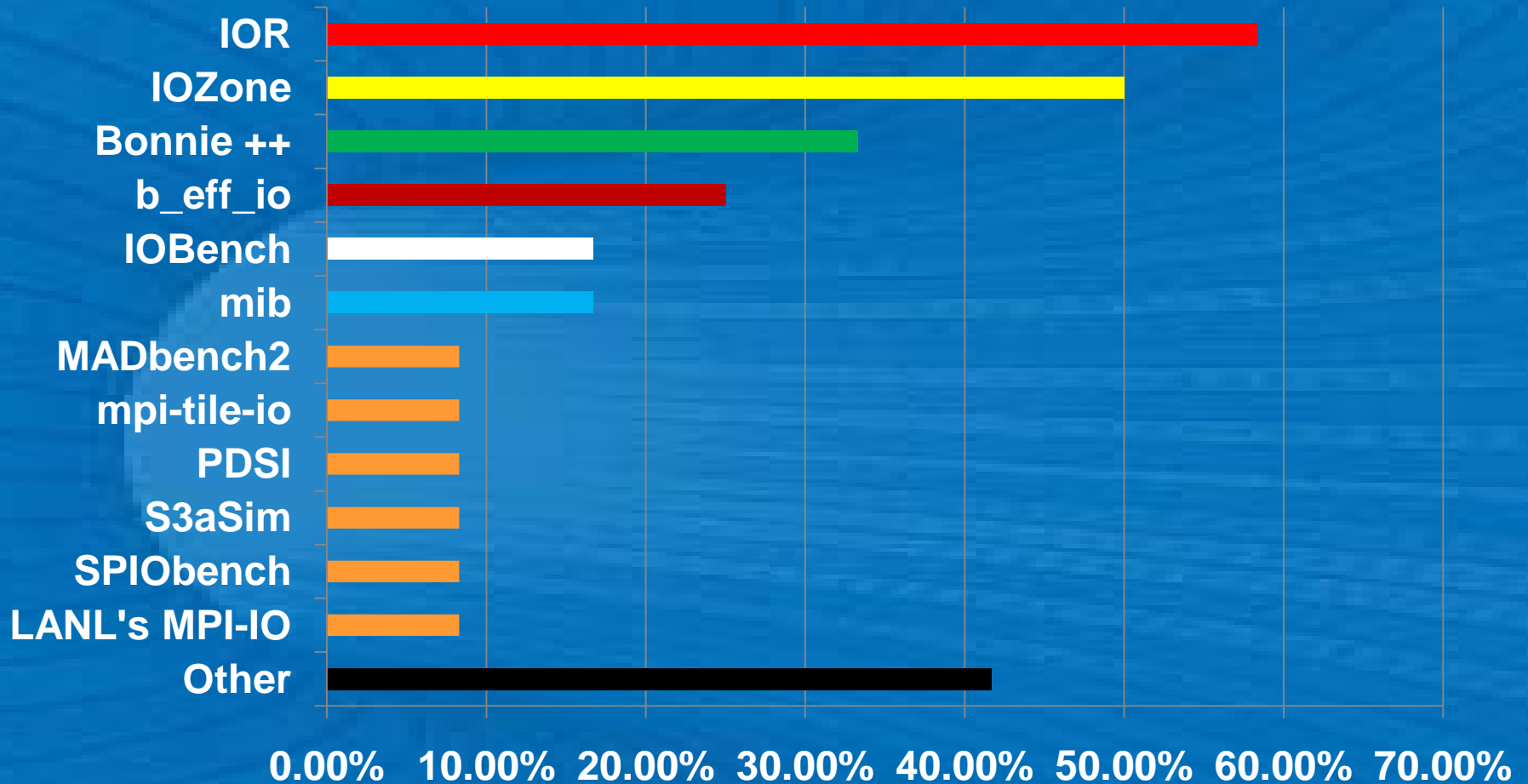
Benchmarking/Tuning Frequency of Storage Systems



Data Center Use of File Systems



Benchmarking Tools Used



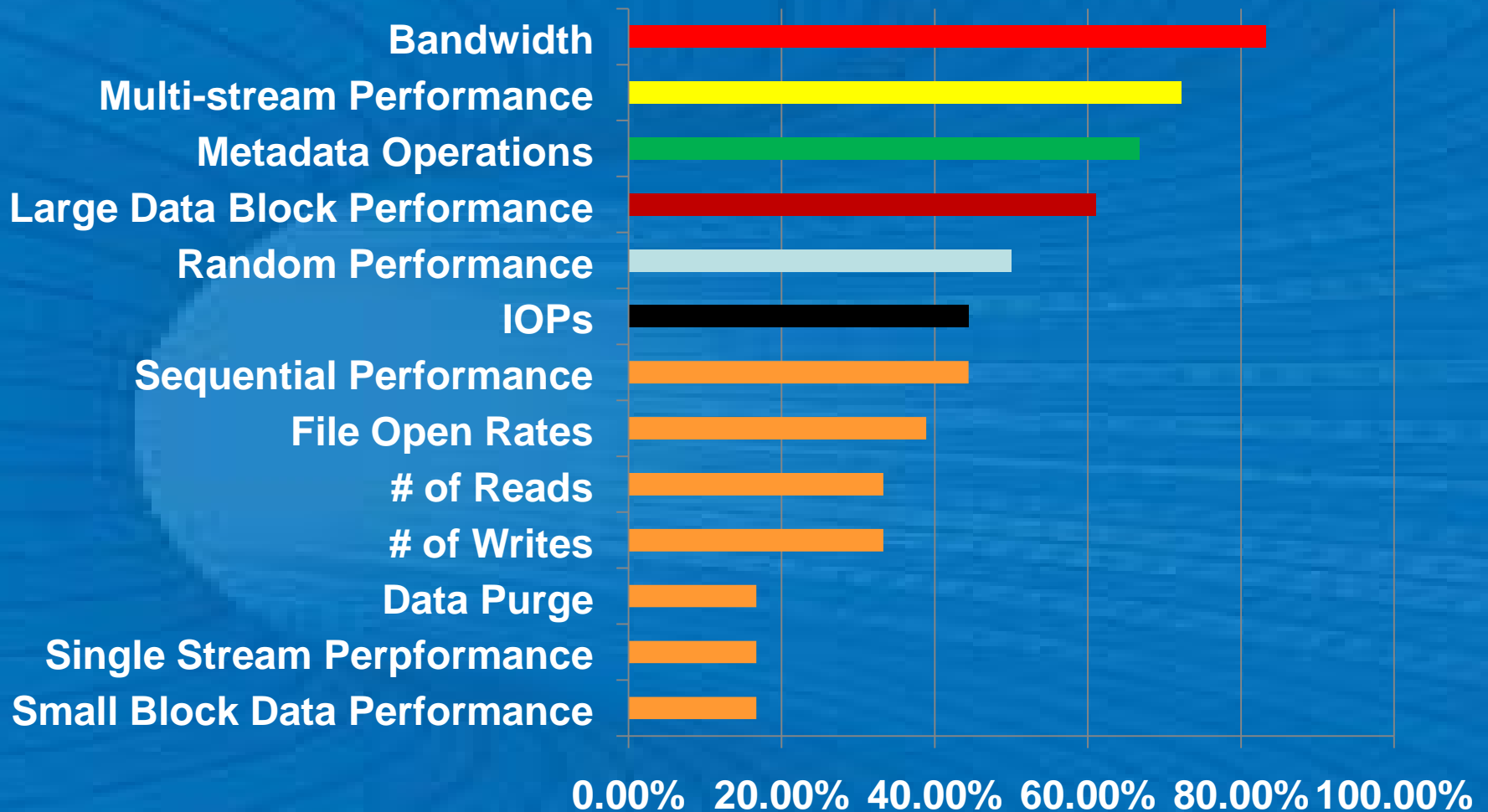
Benchmark Use

- ◆ Half use them for tracking and performance reference
- ◆ More than 40% use it for performance validation
- ◆ A small percentage use it for problem assessment and/or resolution
- ◆ No responses mention that it was mandated

When are Benchmarks Run?

- ◆ Over 60% state that they use them upon system arrival
- ◆ More than 55% say they use benchmarks during design and evaluation periods
- ◆ Defined Periods:
 - ◆ Monthly and Quarterly ~ 17%
 - ◆ Rarely or never ~ 10%
- ◆ Some use them for stress testing of changes

Most Important Performance Aspect



Single Most Important Metric

- ◆ **Aggregate throughput was the most recommended**
- ◆ **Close second – Metadata performance**
- ◆ **Parallel Random Access mentioned**
- ◆ **Correctness checking/integrity validation**
- ◆ **Synthesis of block level traces of actual workloads**

Issues Seen After Benchmarking and Tuning

- ◆ **Stability issues**
- ◆ **Sluggish metadata performance**
- ◆ **Degraded performance or poor reliability**
- ◆ **Data integrity issues – often seen with bit flips on physical media**
- ◆ **Features in the file system and storage media meant to enhance performance become bottlenecks**

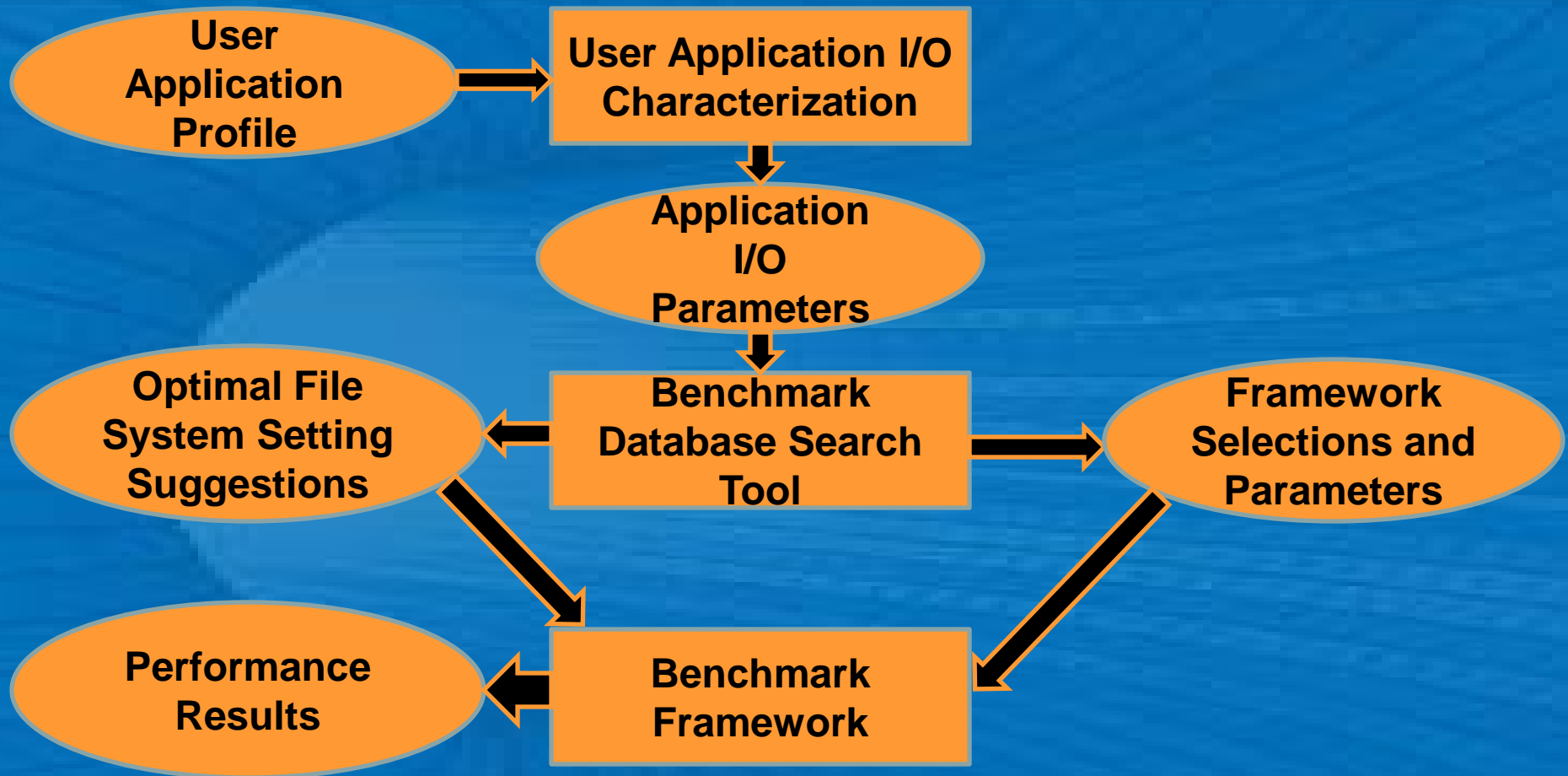
Techniques and Suggestions for Normalization

- ◆ Get away from max MB or GB/sec performance
- ◆ Look at performance measures
 - ◆ Per spindle
 - ◆ Per I/O node
 - ◆ Per switch
- ◆ Peak Sequential I/O rates
- ◆ Perhaps standardize one configuration and compare all others to that control

Need for Normalization

- ◆ **Normalization of different architectures**
 - ◆ **Over 60% say absolutely yes**
 - ◆ **About 25% are uncertain**
 - ◆ **Only a small portion disagree with the need**

Benchmarking Framework and Methodology



Part 1:

I/O Characterization

- ◆ DICE Collecting Trace and Benchmark results
- ◆ Parsing results to identify patterns
- ◆ First step to using the DICE Framework
 - ◆ I/O Characterization of application or workload
 - ◆ DICE will generate a catalogue of I/O characterizations based on inputs
 - ◆ Users can provide traces, select a similar pattern/workload, or provide their known parameters

Part 2:

Database Search Tool

- ◆ DICE will create a searchable database
- ◆ Will provide Benchmark Framework with
 - ◆ Ideal benchmarks to target selected metrics
 - ◆ Parameter settings based on I/O characterization of applications
 - ◆ Will also provide some suggested optimization parameters based on previous performance data

Part 2:

Database Search Tool

◆ What does this tool need?

Characterization of all benchmarks to be utilized

Characterization of known file systems

Characterization of I/O and storage sub-system components

Performance data from community and previous benchmark efforts

Received good Lustre Characterization information

- Scott Teige, High Performance Applications Group, Univ. of Indiana

Part 3:

Benchmark Framework

- ◆ In depth Framework tool to evaluate file systems
- ◆ Parameters passed to it from the search tool
- ◆ Will have list of potential benchmarks available to use
- ◆ Will be extensible to add new ones as needed
- ◆ Will include any DICE developed benchmarks for key file system criteria

Next Steps

- ◆ First design review of the development in May
- ◆ Will present this to the DICE Technical Advisory Panel
- ◆ Will begin to parse traces and benchmark data submitted
- ◆ Lots of Characterization work required
- ◆ Software development and scripting for framework and search tool
- ◆ 1st phase tool and report due Feb-March 2011

Next Steps con't

- ◆ **New DICE developed benchmarks**
 - ◆ Focus on Metadata and Aggregate throughput
- ◆ **Normalization**
 - ◆ Data collected and analyzed to build normalization
 - ◆ Will need more in I/O subsystem characterization
 - ◆ Initial concepts for normalization will be realized
 - ◆ Will be performing intensive testing utilizing the framework to achieve this goal

How to participate

- ◆ Survey still open

- ◆ <http://questionweb.com/51082>

- ◆ Web collection site

- ◆ Benchmark and Trace data

- ◆ <http://webproj.usa40.net>

- ◆ DICE Forum

- ◆ <http://www.diceprogram.org/news/forum.shtml>