OLCF
Oak Ridge Leadership Computing Facility

Technical Overview of the OLCF's Next Generation Parallel File System

Galen Shipman, Doug Reitz, James Simmons,
David Dillow, Doug Fuller, Youngjae Kim, Jason Hill, and Sarp Oral

Presented by: Sarp Oral, PhD

Transitioning I/O to next gen computing

- From Jaguar to Titan
  - Number of cores: 224K → 300K
  - Memory: 300 TB → 600 TB
  - Peak Performance: 2.2 PFlops → 10-20 Pflops
  - Proprietary Interconnect: SeaStar2+ → Gemini
  - Peak egress I/O (over IB): (192 x 1.5 GB/s) → (384-420 x 2.8-3 GB/s)

More capable platform for science → more demanding I/O requirements to deliver the science

Starting from Spider …

- Spider → Next gen parallel file system
- Designing, deploying, and maintaining Spider was a trail blazer
  - No ready available solution at the time of design or deployment
  - Novel architecture
- Center-wide shared file system approach
  - Eliminating islands of data
  - Decoupled file system from compute and analysis platforms
  - Rolling or partial upgrades possible with no down time
  - *Single-point of failure*

Spider availability

- Scheduled Availability (SA)
  - % of time a designated level of resource is available to users, excluding scheduled downtime for maintenance and upgrades

System	Scheduled Availability (SA)			
	2010 Target	2010 Actual	2011 Target	2011 Actual
Widow1	95.0%	99.7%	95.0%	99.26%
Widow2	NIP	NIP	95.0%	99.93%
Widow3	NIP	NIP	95.0%	99.95%

- Widow1
  - **100% availability in 8 of the 12 months of 2011** with SA of 99.26% over the entire year
- Availability and reliability surpassed our expectations

Next gen file system will also be center-wide shared architecture

New Architecture

- Target numbers for next gen parallel file system
  - 1 TB/s file system-level well-formed I/O performance
  - 240 GB/s file system-level random I/O performance
  - Capacity will be based on the selected storage media
    - Expected to be 9-20 PB
  - Availability: >95%
    - Expected availability will be similar of Spider's

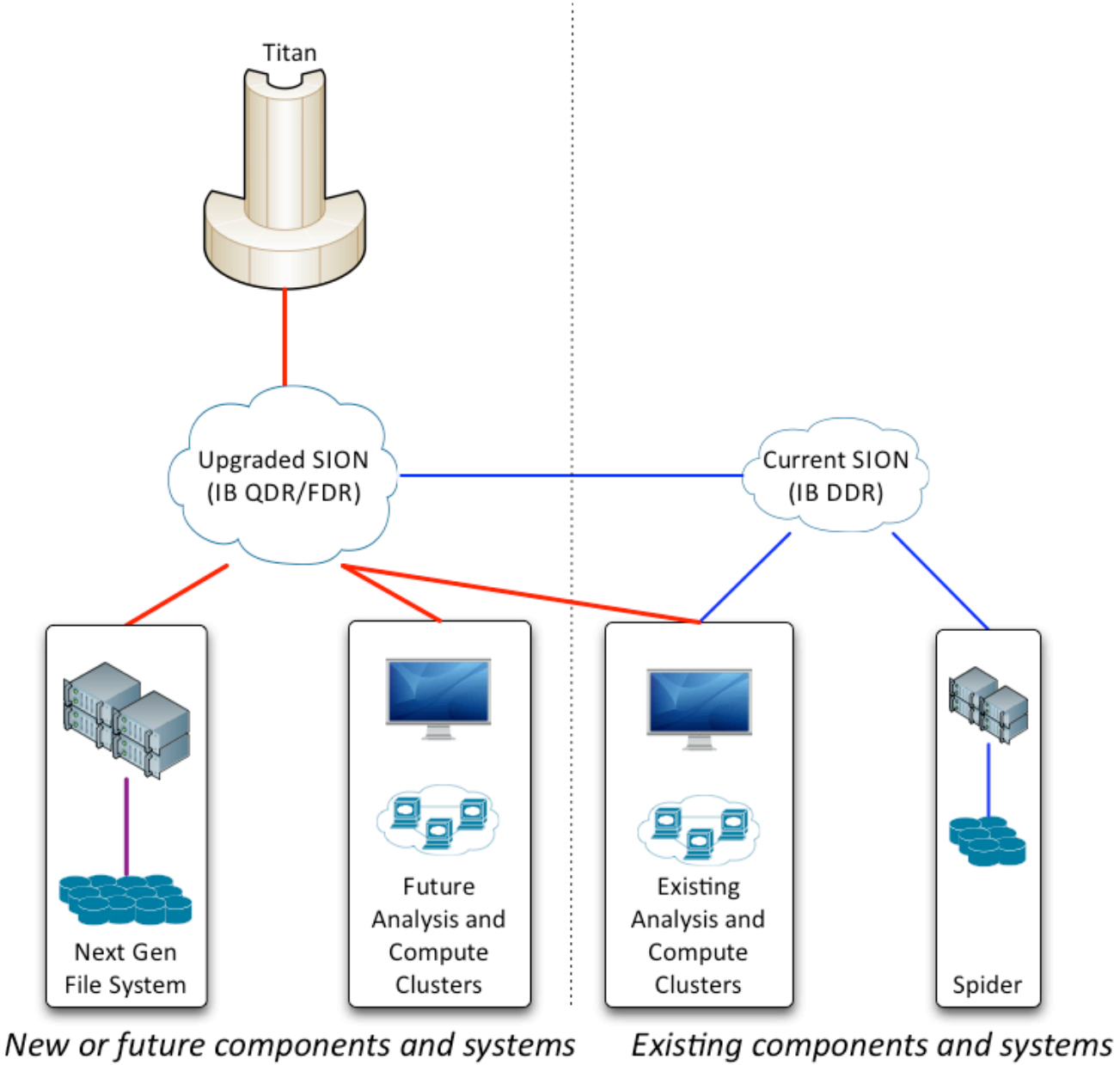
Architecture

- Expected storage and network architecture
  - Will be built using scalable building blocks (SSU)
  - Host-side connectivity: IB FDR or QDR
    - SION tech refresh and upgrade
  - Disk-side connectivity: FC, IB, SAS, …
    - Agnostic of the host-side

Another advantage of decoupled parallel file system architectures

  - Next gen file system and Spider will be online concurrently
    - Spider will be connected to the upgraded SION through existing SION
    - Spider EOL expected to be 2014

Architecture



Lustre for next gen parallel file system

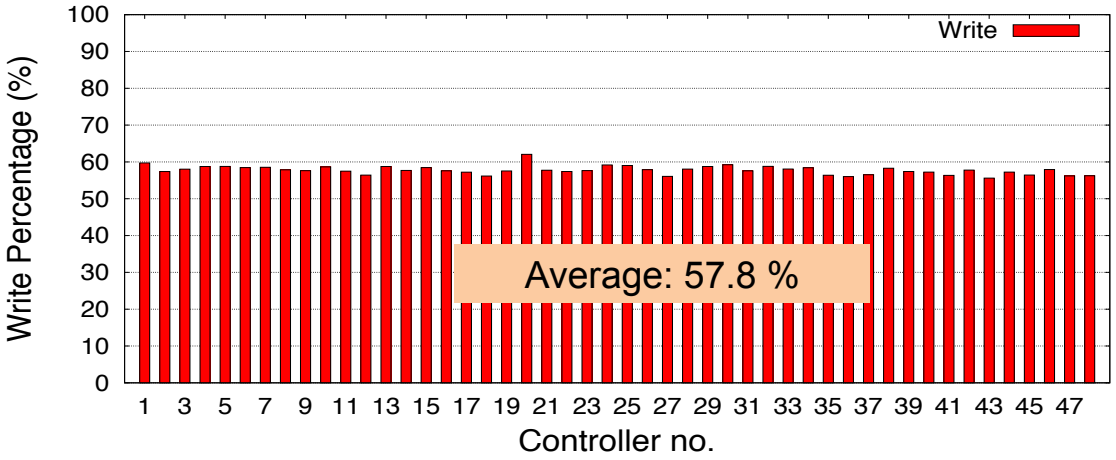
- Lustre v. 2.2 or later will be used
  - Improved metadata performance
    - pDirOps (2.2)
    - Async glimpse lock (statahead issue)
    - *DNE and SMP scaling*
  - Scalability improvements (2.2)
    - Imperative recovery
    - Wide-striping
    - Portals RPC thread pool
    - *NRS*

Working with Whamcloud to harden and stabilize 2.2

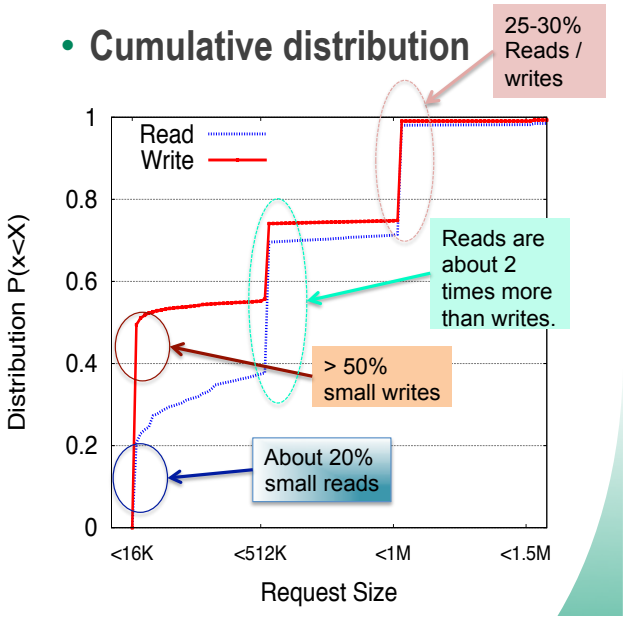
Scheduled down-times can be used to harden 2.2 and test future Lustre features, bug fixes, and improvements

I/O Workload Characterization

- "Workload characterization of a leadership class storage cluster"
  - http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5668066



42.2% Read requests ➔ still significantly high!!!



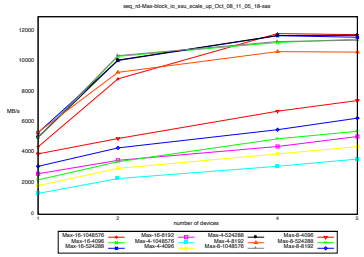
Next gen file system

can not only be optimized for checkpointing

should support mixed workloads

Procurement

- Acquisition process
  - Open procurement
  - Timetable: TBD (2012-2013 timeframe)
- Procurement benchmarks
  - Publicly available
    - http://www.olcf.ornl.gov/wp-content/uploads/2010/03/olcf3-benchmark-suite.tar.gz
  - Block I/O
    - Libaio based, fair-lio as I/O engine
    - Single host single LUN
    - Single host all LUNs
    - SSU all LUNs – healthy
    - SSU all LUNs – degraded



  - File system I/O
    - Obdfilter-survey based
    - Tested against Lustre v1.8