outthink limits

Spectrum Scale Enhancements for CORAL

Sarp Oral, Oak Ridge National Laboratory
Gautam Shah, IBM

SC16 Salt Lake City, Utah | hpc matters.

What is CORAL

Collaboration of DOE Oak Ridge, Argonne, and Lawrence Livermore National Labs

- Established in early 2014 to leverage supercomputing investments, streamline procurement processes and reduce costs to develop supercomputers
  - “High-performance computing is an essential component of the science and technology portfolio required to maintain U.S. competitiveness and ensure our economic and national security”
    - U.S. Secretary of Energy Ernest Moniz
- Two new High Performance Computing (HPC) awards announced in November 2014
  - Both CORAL awards leverage the IBM Power Architecture, NVIDIA’s Volta GPU and Mellanox’s Interconnected technologies to advance key research initiatives for national nuclear deterrence, technology advancement and scientific discovery
    - Oak Ridge National Laboratory’s (ORNL’s) new system, Summit, is expected to provide at least five times the performance of ORNL’s current leadership system, Titan
    - Lawrence Livermore National Laboratory’s (LLNL’s) new supercomputer, Sierra, is expected to be at least seven times more powerful than LLNL’s current machine, Sequoia.

Source: http://energy.gov/articles/department-energy-awards-425-million-next-generation-supercomputing-technologies

CORAL Systems

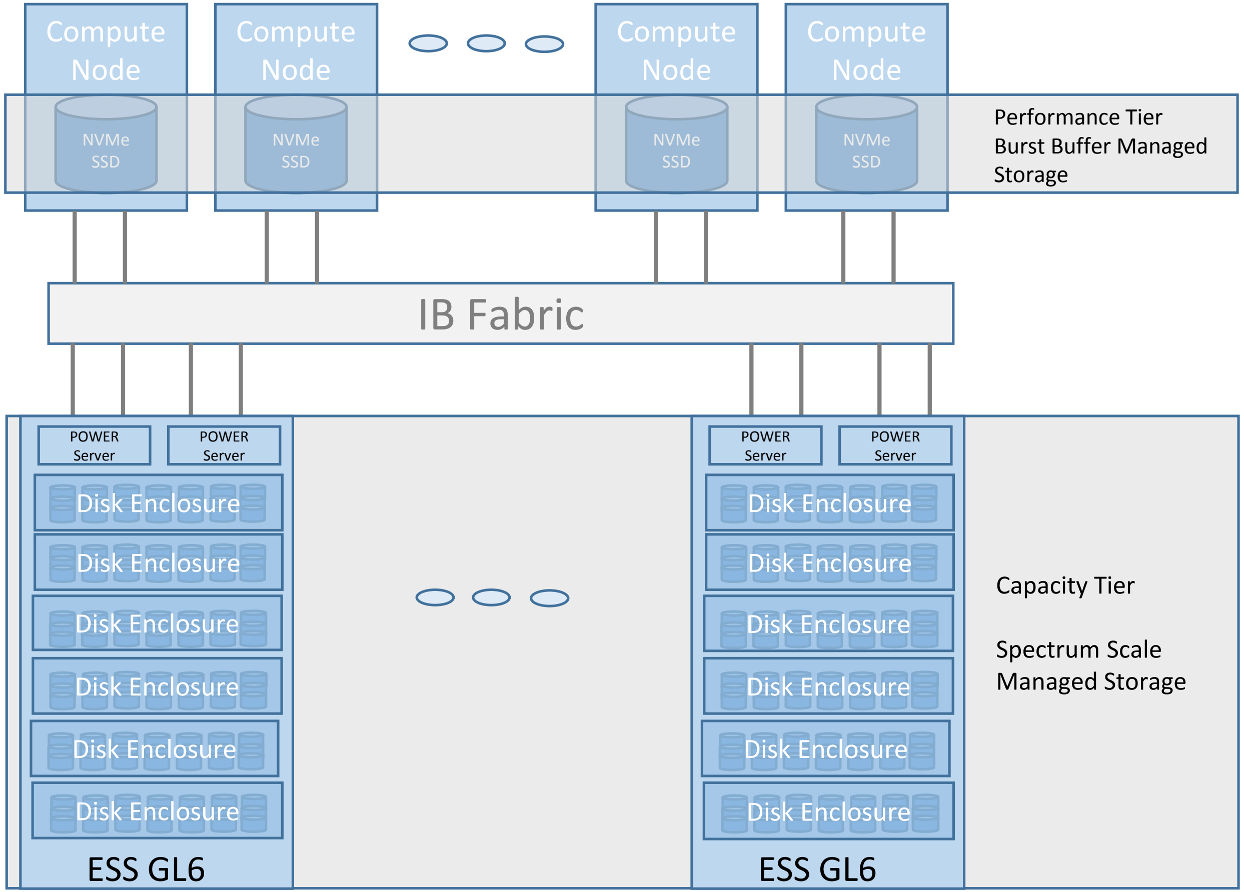
- **LLNL’s Sierra system**
  - ~4000 Power9 nodes with GPU acceleration
  - ~2.3 PB system memory (include DDR & HBM; does not include NVMe)
  - Dual-rail InfiniBand EDR fat tree network or better
  - ~120 PF
  - ~9 MW
- **ORNL’s Summit system**
  - ~4500 Power9 nodes with GPU acceleration
  - ~ 2.7 PB system memory (include DDR & HBM; does not include NVMe)
  - Dual-rail InfiniBand EDR fat tree network or better
  - ~200 PF peak
  - ~13 MW

CORAL System Storage Overview

Storage architecture

- Need for a burst buffer/performance tier
    - Lowers traditional spinning disk & lower power consumption
  - Node local NVMe SSD managed by Burst Buffer Software
- Capacity requirement – ESS Storage
- Performance/Scaling requirements – Spectrum Scale Software

CORAL Storage Overview



Burst Buffer Software

- Goals/Features
  - Support node-local checkpoints
    - SSD partitioned and formatted with standard Linux file system
  - Support staging data in and out of SSD
    - Provide a mechanism for pre- and post- job transfers for data staging via LSF
  - Build asynchronous file transfer service between SSD and GPFS
    - Software on compute node initiates transfer and can poll for completion
  - Avoid excessive data movement
  - Avoid performance jitter to running applications on compute node
    - Move data between compute node and ESS with NVMe over Fabrics for low performance impact
  - SSD wear awareness, health monitoring, and protection

Spider 3 @ OLCF

Spider 3 is a center-wide single namespace POSIX file system to serve all OLCF resources eliminating data islands, and enabling seamless data sharing between resources

- Built on IBM's Elastic Storage Server based on Power 9 Processor and uses Spectrum Scale (formerly known as GPFS) parallel filesystem technology utilizing GPFS Native RAID with 8+2 redundancy
- Provides a usable capacity of 250 PB
- Performs at an aggregate sequential peak read/write bandwidth of 2.5 TB/s
- Performs at an aggregate random peak read/write bandwidth of 2.2 TB/s
- Provides rich metadata performance; single directory parallel create rate of 50,000/s
- Provides rich interactive performance; @32 KiB I/O 2.6 million IOPs
- Disk-based, with tens of thousands of disks
- Connected to OLCF's SION 3 SAN with IB EDR
- Will also serve as the Summit Burst Buffer sink and source on the end-to-end I/O path

Spectrum Scale Enhancements for Scaling Namespace

The single namespace CFS will meet the following

- Single name space supporting 250 PB capacity
- Total number of files supported is 100 B
- Maximum file size equal to aggregate system memory
- 10 M files per directory

Enhancements needed in Spectrum Scale

- Improvements in fsck – time to run (including nodes to use), progress reporting, …
- Parallel virtual disk creation
- Reduce contention to allow more concurrency

Spectrum Scale Enhancements for Scaling Performance

Performance improvement are required to meet:

- Aggregate sequential peak read/write bandwidth of 2.5 TB/s
- Aggregate random peak read/write bandwidth of 2.2 TB/s
- Single directory parallel create rate of 50,000/s
- Interactive performance; @32 KiB I/O 2.6 million IOPs

Enhancements needed in Spectrum Scale:

- Performance counters to help uncover bottlenecks (mmfsadm dump iocounters/iocountercpu)
- Improve RPC communication (avoid global receive pool mutex by creating multiple pools of worker threads; Fast Condition Variable for some of the condition variables in RPC path; spread interrupt load across multitple IRQs using RDMA completion vectors)
- Improve parallelism of full track writes

Spectrum Scale Enhancements for Scaling Metadata Rate

Metadata requirements include:

- Single directory parallel create rate of 50,000/s
- Interactive performance; @32 KiB I/O 2.6 million IOPs

Enhancements needed in Spectrum Scale:

- Improve directory block management (avoid directory fragments, directory block split option,
- Avoid token manager revokes
- Pre-allocation of directory blocks (mmchattr)
- Smooth the filesystem sync work over the sync period

Collaboration for Success

We expect other challenges we have to overcome as we deliver/deploy the system and we are working together to anticipate and resolve these issues

- Impact of rebuild performance over the population size …
- Identify and eliminate “slow” disks so the system performance consistently

Thank you!

ibm.com/systems/hpc

Legal notices

Copyright © 2016 by International Business Machines Corporation. All rights reserved.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 1 0504- 785
U.S.A.

Information and trademarks

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

ITIL is a Registered Trade Mark of AXELOS Limited.

UNIX is a registered trademark of The Open Group in the United States and other countries.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.