HPC and Big Data: Better Together!

Kirill Malkin
Director of Storage Engineering
kmalkin@sgi.com

The SGI logos and SGI product names used or referenced herein are either registered trademarks or trademarks of Silicon Graphics International Corp. or one of its subsidiaries. All other trademarks, trade names, service marks and logos referenced herein belong to their respective holders. Any and all copyright or other proprietary notices that appear herein, together with this Legal Notice, must be retained on this presentation. The information contained herein is subject to change without notice.
Disclaimers & Notices

This presentation contains forward-looking statements and plans that are subject to risks and uncertainties. All products, dates and information are preliminary and subject to change without notice. SGI may choose not to make generally available any product or features discussed in this presentation.

SGI, the SGI logo, SGI UV, SGI ICE, SGI InfiniteData, NUMAlink, SGI InfiniteStorage, SGI ZCA and Zero Copy Architecture are trademarks or registered trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries. Intel and Xeon are registered trademarks of Intel Corporation. All other trademarks are property of their respective holders.
Big Data as part of Workflow

• Big Data technologies address analytical part of HPC workflows
• For example, Apache Spark was shown to aid climate analysis
  – Feature Extraction
  – Predictive Modeling
  – Operational Forecasting
• Basic Spark abstraction is RDD
  – Resilient Distributed Dataset
  – Filter, Map, Scatter, Gather
  – Various set operations supported
• One key challenge is how to load & share data
NSCI Strategic Objective 2

Executive Order

• “Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing”
  – Dynamic interaction between analysis and simulations
  – More agile and reusable HPC software portfolio
# HPC vs. Big Data

**HPC (Big Compute)**
- Focus on Compute
- Take data to compute
- MPI, PGAS
- Compiled code
- Network Latency
- Tends to produce data
- Shared storage
- Storage Bandwidth

**Big Data**
- Focus on Analysis
- Take compute to data
- Map/Reduce, Transform
- Interpreted code
- Network Bandwidth
- Tends to consume data
- Shared nothing storage
- Storage Latency
Different Approach to Storage
Comparing HPC and Big Data

HPC

Single large external filesystem, metadata server bottleneck

Big Data

Many small filesystems, independent metadata processing
Can HPC run Big Data?

- Short answer is Yes
- “Scaling Apache Spark on Lustre”
  - Work by LBNL presented at LUG 2016
  - HPC nodes are running Spark
  - Lustre is used as storage
- Poor scaling of metadata operations (e.g. file open) and data I/O limitations observed
- Takeaways:
  - Requires techniques to reduce metadata operations
  - Faster storage (e.g. Burst Buffer) is needed to address heavy I/O concurrency
  - More memory is needed
  - Long term data retention in Lustre
Integrating Big Data & HPC

- Deploy high-performance “Tier Zero” NVM that appears “local” to all nodes
  - Export NVM over fabrics
  - SGI’s Zero Copy Architecture
- Reduce and isolate metadata servers through use of dynamic POSIX-compliant namespaces instead of one giant PFS
- Further isolate metadata processing through node-local temporary storage and loopback filesystems for Big Data
  - Facilitates “shared nothing” storage
- Include nodes with more memory for analytics
- Represent HPC files as “data source” for Big Data
Integrating Storage & Compute
For HPC and Big Data

Dynamic Tier Zero namespace (POSIX)

Node-local loopback filesystem

Dynamic MDS

Large memory node

Dynamic Tier Zero namespaces spanning from one to many nodes, created per workflow
SGI® UV™ 300 Compute & Flash Capabilities

As shown at SC14

32 Socket SGI UV 300 System with 64 Intel NVMe Flash Cards

- Up to 30M IOPs
- Up to 200 GB/s

Interconnect Fabric

Node 1
- Memory
- NVMe Flash
- CPU
- UV

Node 2
- Memory
- NVMe Flash
- CPU
- UV

Node 31
- Memory
- NVMe Flash
- CPU
- UV

Node 32
- Memory
- NVMe Flash
- CPU
- UV

Graph data points:
- IOPS
- Expected IOPS
- GB/s
- Linear (IOPS)
- Linear (GB/s)
Is PFS fading away?

As long-term Storage

• PFS limits HPC and Big Data integration
  – Single namespace metadata is hard to scale
  – Single tier requires accelerators such as burst buffers
  – Maintenance or reconfiguration leads to complete outage
  – Relatively weak data protection (RAID-6)
  – Lack of geo replication

• What if we used PFS only for Tier Zero?
  – PFS becomes POSIX API with local cache
  – Data is pre-fetched into PFS prior to launching job, then results are saved
Leveraging S3 and Metadata
As backend for Tier Zero

- Store bulk of data in scale-out S3 object storage, on prem or in cloud
- Automatically stage and de-stage objects to namespaces just-in-time, based on workflow
- Capture file metadata and rich metadata in a distributed database
  - Treat them as Big Data too
- Object data in S3 are directly accessible by Big Data apps as data source
  - S3 objects can be available for external Big Data clusters
HPC/HPDA Data Tiering

S3 backend, on premises or in the cloud

Distributed Metadata repository

©2016 SGI
Workflow-aware Data Management

HPC Cluster

Job Scheduler

Just-in-time data staging and de-staging

Initial Input

Fast Parallel Sequential Data Movers

Tier Zero: Campaign Namespace

Mover

Mover

Mover

Metadata Repo

Namespace Change Capture

Campaign & Forever store

Tape

Seq Disk

Cloud

©2016 SGI
SGI® Zero Copy Architecture
Next Generation Approach

HPC : Scalable Compute Fabric
(All Nodes Have Access to All Data)

Local PCIe Data Fabric

CPU

NVM

NVM

NVM

NVM

NVM

NVM

CPU

Local PCIe Data Fabric

Infiniband or
Omnipath Fabric

Job Orchestration and Data Management Control

HPDA : Global Data Fabric
Single Address Space

In-Situ Visualization & Production Optimization

DMF : Capacity Mgmt
Intelligent Data Tiering

HPDA Applications

©2016 SGI
SGI Converged HPC/HPDA Components

HPC Cluster

SGI® Rackable® or SGI® ICE™ XA Cluster

ZCA Storage

Local

Local

HPDA

SGI UV 300

Capacity Tier

SGI® DMF™ Zero Watt Storage™, Object Storage, Cloud and Tape
Summary

• Combining HPC & Big Data is required to deliver workflow acceleration
• Leverage modern NVM in combination with S3 to deliver complete data management solution
• Zero-Copy Architecture provides sharing of Tier Zero storage without introducing fabric overhead or data-driven jitter
TOGETHER WE ARE BETTER