Investigating Spectrum Scale File System (GPFS) with Spark and (Spectrum) Conductor

by Chris Schlipalius

Senior Systems Administrator & Team Lead, Storage Infrastructure – The Pawsey Supercomputing Centre - Data Team

Where is and who is Pawsey?

- Located on the West Coast of Australia – Perth (yes we have many white sharks).

PAWSEY supercomputing centre

Background: Pawsey CSIRO Datacenter

- One of two National Supercomputing Centers in Australia

(the other is NCI in Canberra near Eastern Side of Australia).

PAWSEY
supercomputing centre

What do we do?

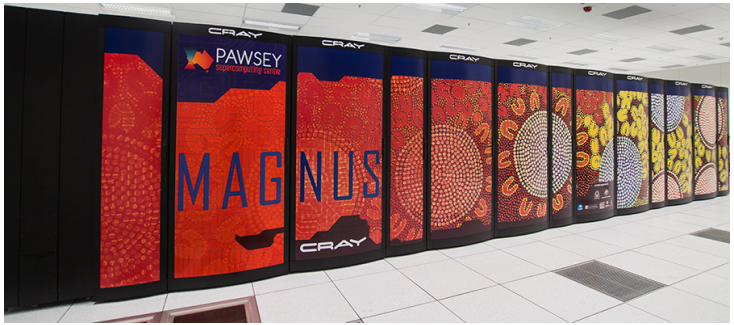
- We provide services to scientific researchers in Australia – our areas of expertise are in Visualisation, Supercomputing, Data Storage and Services, Training and Outreach.
- Types of projects:
  - Radio Astronomy (Pre-SKA, MWA/GLEAM, CASDA)
  - Life Sciences
  - Medicine
  - Geoscience and Physics….

  https://www.pawsey.org.au/research/

PAWSEY supercomputing centre

Compute resources at Pawsey?

We have two Crays (XC40 – Magnus and XC30 – Galaxy)



- a SGI/HPE UV2000 and now,
- a new Advanced Technology Cluster from HPE (called Athena):
  - Eight C2112-4KNL nodes with Intel Xeon Phi 7210 processors
  - Eleven C1102-GP8 nodes with four NVIDIA Tesla P100 SXM2 GPUs

PAWSEY
supercomputing centre

Oh and our Crays are groundwater-cooled (22°C circuit), our water pumps are PV-powered 208kW

https://www.pawsey.org.au/pawsey-centre/geothermal-cooling-system/

PAWSEY
supercomputing centre

GPFS Storage Infrastructure

- 5.6PB usable (2.3PB used) of disk across 3TB SATA Hitachi and HGST 4TB NL-SAS - DDN SFA12K-40X, 8x External NSD servers from DDN
- GPFS v4.2.1.0 (DDN GRIDScaler)
- 107PB of Tape backup (SL8500, 1000+ tapes running compression on the T10000D drive with Tivoli Storage Manager server v7.1.1.100 – 8x NSD Server Clients)
- Mellanox FDR IB fabric
- Data Portal - Mediaflux on SGI Server HA cluster (2100 series, each with 512GB RAM, Intel Xeon 4 socket 8 core, XFS and CXFS filesystems)

What do we store and for whom?

- Data for projects of national significance –

  Ranging from a few GB in size to multiple PB for Radio Astronomy (including Pre-SKA) data files - both raw and product, related to supercompute processing, and standalone data.

- Research Data Services node (GPFS) and General Science (CXFS and DMF – HSM).
- Data Portal (Mediaflux)is used to front-end GPFS (and CXFS and DMF for most of our users)

The challenge…?

- Develop new data services for data analytics, in particular, Spark with third party notebook support as there is a growing user need for Spark workload support and resourcing
- Use compute on our Open Stack on-premises cloud (Nimbus) as this is POC and users are already using Spark in hosted VM's.
- Utilise existing managed data collections (workflow support for efficiency of data movement and management)
- Utilise the currently unused 1.3PB of 3TB SATA II disk in our Spectrum Scale file system
- Plus, all the other good stuff:-
  - ✓ User & group management,
  - ✓ Orchestrator,
  - ✓ Scheduling,
  - ✓ Interfaces.

PAWSEY supercomputing centre

Why Apache Spark?

- Open Source
- Has Orchestrators - Mesos, Yarn, Conductor
- Suits containerization and our Openstack Nimbus cloud
- In-memory and parallel distributed processing
- Apps are built with Scala, Python and Java
- Workload use cases:
  - SparkSQL
  - Spark Streaming
  - Machine Learning
  - Graph Processing
  - Deep Learning (on certain platforms)

Challenges of running Spark effectively

- Per-user-use in single VM's is not efficient- VM sprawl
- No integrated access to parallel file system-resident data collections.
- We need an orchestrator for scale out, but,

  not all Orchestrators are created equally!

So what came next?

Research:

- Looked at STAC benchmarks for Orchestrators
- Looked into Spectrum Scale (GPFS)- integrated Orchestrators

We found Spectrum Conductor!

PAWSEY
supercomputing centre

What is Spectrum Conductor?

Part of Spectrum Computing and it's tightly integrated with Storage (Spectrum Scale

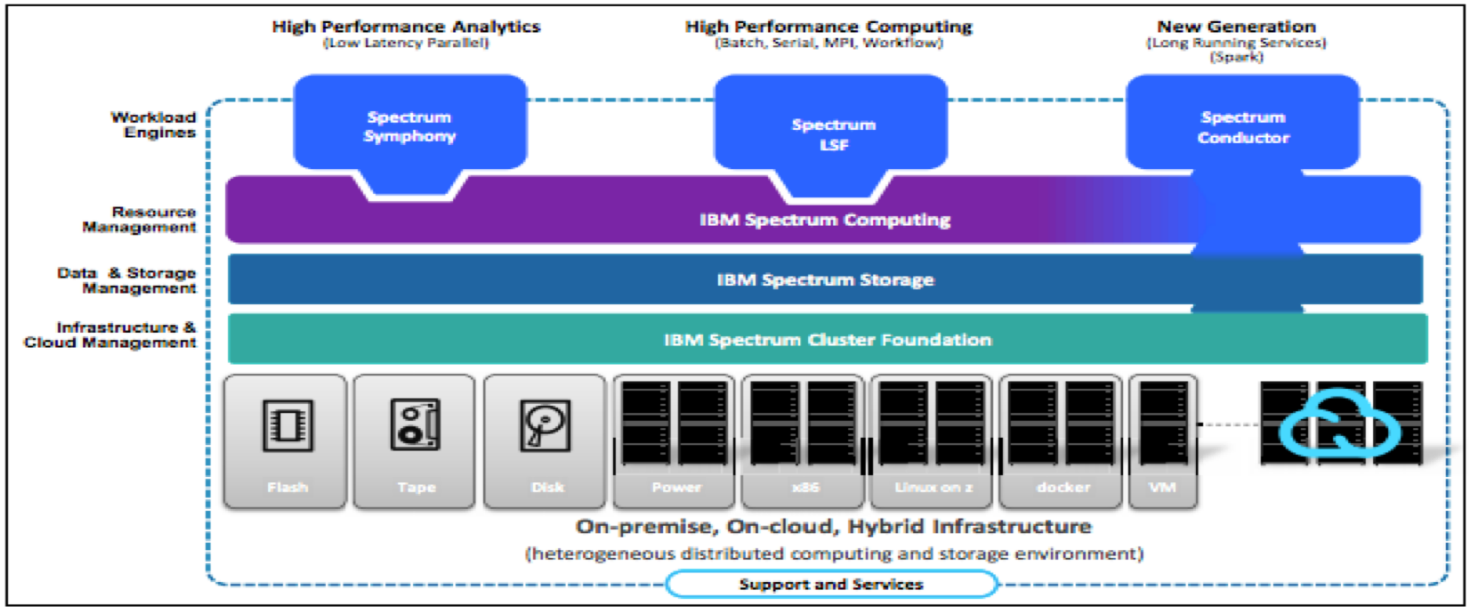


Figure 3 IBM software defined infrastructure solution

PAWSEY supercomputing centre

Why are we looking at Conductor with Spark (CwS)

- It is the **most efficient Orchestrator** (EGO) – vs Apache Mesos or Hadoop Yarn (software-defined infrastructure) as tested by STAC – top Orchestrator in Spark Multi-tenancy Benchmark testing.
- CwS accelerates results, increases resource usage and is an Enterprise class-solution for Spark workloads and management

- Jupyter Notebook support
- SpectrumScale CES OR FPO (File Placement Optimizer) – more space-efficient than HDFS
- GPU support – CUDA, OpenCL

PAWSEY supercomputing centre

Hang-on, but what's all this about Spectrum Scale (GPFS) integration?

- Shuffle (this default file algorithm is disabled as we have GPFS)
- HA (Master node and shared filesystem)
- Monitor (GUI for GPFS)

Next, the POC setup…

- Upgraded Spectrum Scale (GPFS) v4.2.3 on test DDN SFA12KX
- Setup LDAP service accounts
- Installed Conductor with Spark (CwS) orchestrator from IBM on our two Dell test NSD servers and one KVM VM
- Add LDAP service config - development environment 389 LDAP integration,
- Set up Cluster Export Services (CES) using Network File System (NFSv4) on two Test NSD Servers (DDN),
- Deployed 10x “Jumbo” CentOS VM’s 1GBE, 16xvCPU, 40GB RAM each on AMD64,
- Add in Enterprise SSL Certs for secure client and Master slave node

Ironspark – console

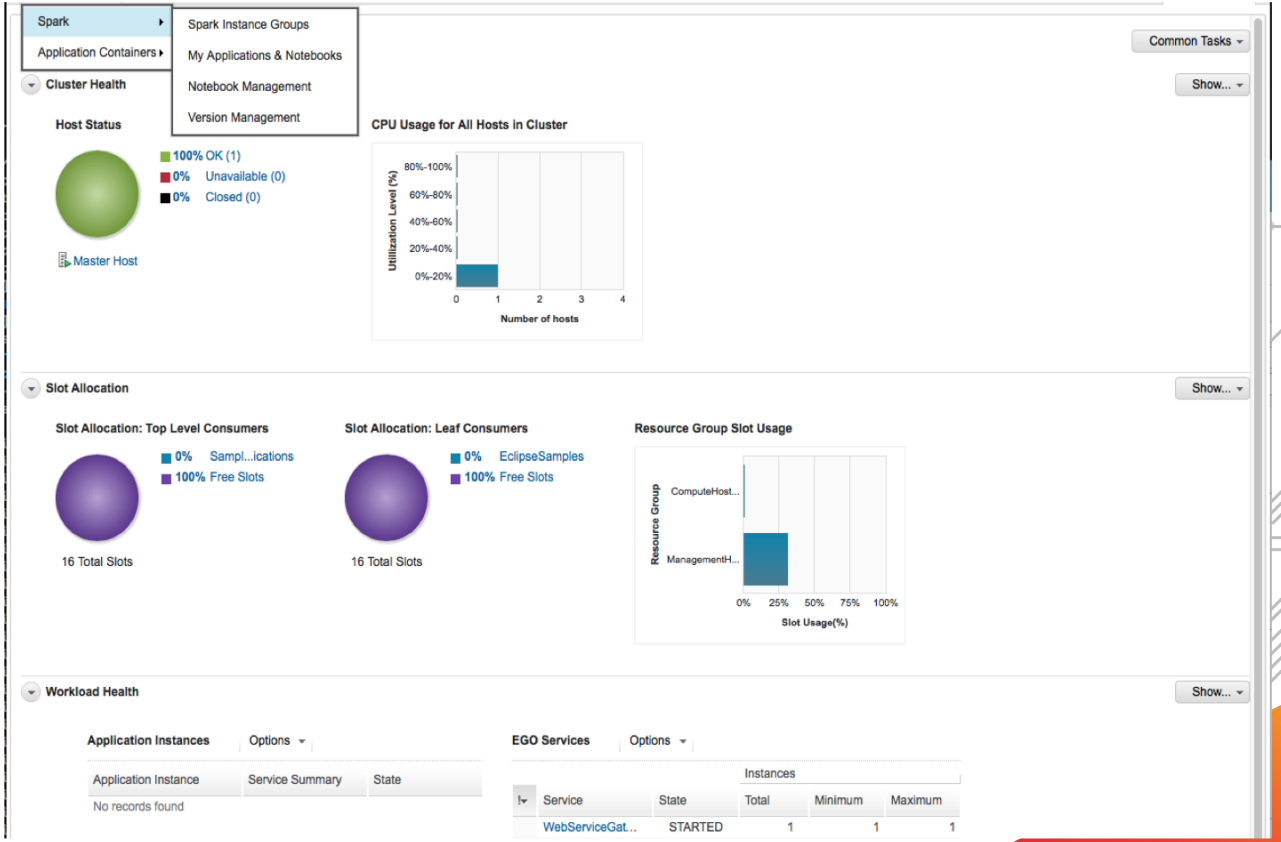
Application Submission Interfaces:

1. Web GUI
2. Cmd line
3. Notebook – Jupyter

Scenarios:

1. multiple jobs on single spark instance
2. multiple jobs on multiple spark instances

Console



PAWSEY supercomputing centre

Future work

- More use cases – discover, define and execute
- Scale out on more nodes, both physical (bare-metal provisioned) and virtual
- Testing with differently-sized Nimbus VM's
- Profiling IO and Sparkbench benchmarking of each setup
- Developing and working with real-world use cases (in-train, Data61 genomics, Business Risk model entity relationships)
- Workflow integration with PAWSEY Data Portal and content stores

PAWSEY supercomputing centre

Conclusion

- We can easily deploy and manage multiple spark instances through IBM CwS,
- Excellent GPFS integration and we are working towards using existing Data Collections “in-place” (WORM) and looking to workflow or process product with Mediaflux for project sharing (thus solving the *scratch* dilemma),
- On-premises/physical machine, or in the cloud, or on virtual machines – *somewhere* (e.g. Nimbus Open Stack), suits secure data services
- Enterprise-grade – LDAP user management and SSL-CA
- Conductor speeds up **any** type of Spark workload (vs. other Spark orchestrators),
- Thanks to my team member - Jeffery Jiang!

PAWSEY supercomputing centre

Recommended reading

https://www.slideshare.net/kahmed0610/experiences-in-delivering-spark-as-a-service

http://www.redbooks.ibm.com/redpapers/pdfs/redp5379.pdf

https://www.ibm.com/support/knowledgecenter/en/SSZU2E_2.1.0/installing/install_upgrade.html

https://www.ibm.com/support/knowledgecenter/en/SSZU2E_2.1.0/installing/install_roadmap_s.html

STAC

Questions?

PAWSEY
supercomputing centre

Logs

https://support.pawsey.org.au/documentation/display/DATA/CwS+on+Nimbus+Installation (file permission part took a long time, it is undocumented.)

https://support.pawsey.org.au/documentation/display/DATA/GPFS+-+CES+Configuration (access mode is important, it wasn't mentioned in manual, but it is Must)

PAWSEY supercomputing centre