# Spectrum Scale with Spectrum Conductor

## Improving Spark – using Spectrum Conductor on Spectrum Scale

*The Pawsey Supercomputing Centre is an unincorporated joint venture between*

CSIRO, Curtin University, ECU Edith Cowan University, Murdoch University, The University of Western Australia

*and proudly funded by*

NCRIS National Research Infrastructure for Australia, An Australian Government Initiative

Government of Western Australia

PAWSEY supercomputing centre

# Introduction to Pawsey Supercomputing Centre


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

# National Reach


ECU Edith Cowan University

Curtin University

The University of Western Australia

Murdoch University

Pawsey Supercomputing Centre

ICRAR

Department of Mines and Petroleum

Department of Biodiversity, Conservation and Attractions

Government of Western Australia

Landgate

The University of Adelaide

CSIRO

DHI

The University Of Queensland Australia

Australian National University

NCI

UNSW Australia

The University of Sydney

Swinburne University of Technology

RMIT University

The University of Melbourne

Monash University

Pawsey Supercomputing Centre

# Supporting Australian Researchers

**1600** researchers

**190** projects

**45** staff

## MAKING TOMORROW HAPPEN, TODAY

through

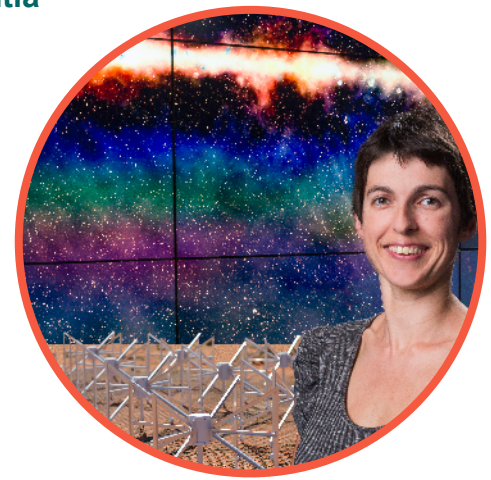
- supercomputing
- data
- visualisation
- training and consulting

# Accelerating Scientific Outcomes

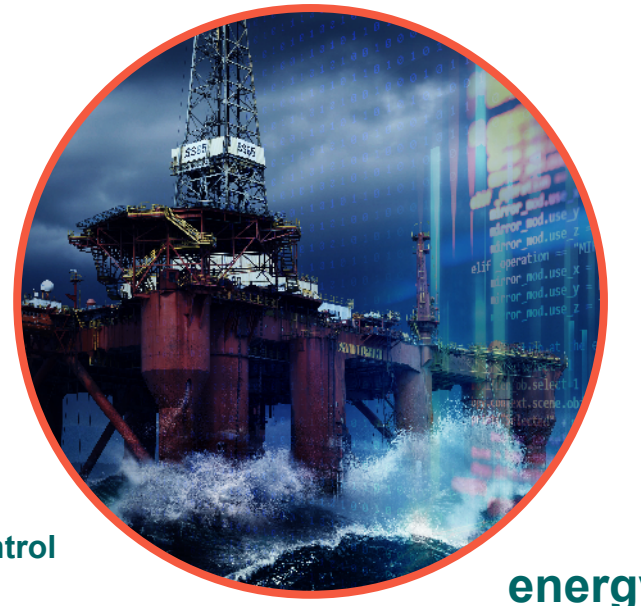
**health**
Combating Alzheimer's and dementia

**190**
projects

**resources**
Protecting Perth's aquifers



**agriculture**
AI for targeted weed control



**energy**
Understanding the ocean to improve offshore designs

**astronomy**
A panoramic view of the universe in colour

PAWSEY supercomputing centre

# Compute resources at Pawsey

PAWSEY
supercomputing centre


# Magnus Supercomputer

35,712 cores, 1.09 PFLOPS ,
Aries dragonfly interconnect

PAWSEY
supercomputing centre


9,440 CPU cores
64 K20X GPUs
Aries dragonfly interconnect

# Galaxy Supercomputer

PAWSEY
supercomputing centre

Crays are groundwater-cooled (22°C circuit)

Cooling water pumps are powered by an 208kW photovoltaic array set

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

PAWSEY
supercomputing centre


20 visualization nodes
44 Pascal GPUs for GPU computing
80 Xeon Phi nodes for manycore jobs
1 TB large memory nodes
2,240 CPU cores for serial codes
FDR/EDR Infiniband interconnect

# Zeus Supercomputer

PAWSEY
supercomputing centre


3000 Cores, OpenStack,
Sahara, Volta GPUs
Nimbus Research Cloud
PAWSEY
supercomputing centre

65 PB Migrating Disk and Tape

# Data Storage

PAWSEY
supercomputing centre

# SFA12K DDN (2012 – 2015)

- 5.6PB usable (2.8PB 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, 16x T10000D drives 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 (SGI 2100 series, each w 512GB RAM, Intel Xeon 4 socket 8 core, XFS and CXFS filesystems)

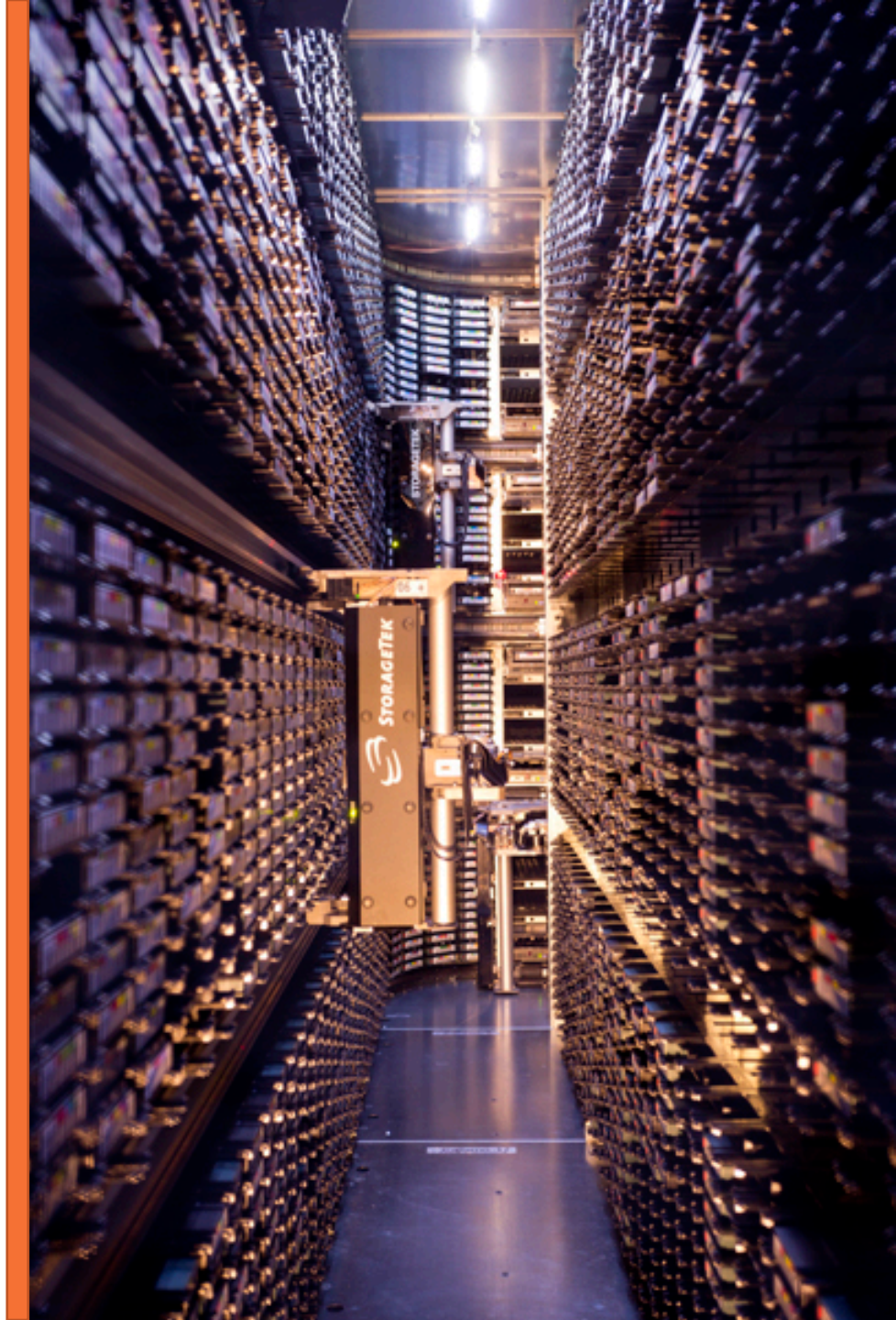
PAWSEY supercomputing centre

# Emerging topics in HPC

Machine Learning

Containers

Data management




PAWSEY
supercomputing centre

# That's the challenge – new Data Services

- Data Analytics (M/L)
- Utilizing existing Research Data Collections – Data Centric or Data Lake model - no moving of data!
- Using "Open Source" DA tools and languages and Notebooks
- Making it easy to:
  - Merge job results with existing data project directories
  - Manage jobs – no need to be your own Sys Admin as a researcher
  - Utilise existing Parallel FS (not Ceph as it's for VM qcow virt disks)

# But, what does Data want?

- Explore options that are manageable with only a few staff and that best accelerate research outcomes.
- Best “packing of jobs” onto existing host nodes (Nimbus users are already using Spark on VM’s – albeit inefficiently).
- Exploit new GPU resources coming online (Deep Learning Impact)
- Use existing parallel file systems  - GPFS and exisiting research data sets and collections – utilise the unused FS space (1.3PB)
- DA workload management of files and output for users
- User & group management, Orchestrator, Scheduling, Interfaces.

# Why Apache Spark?

Open Source
Has Orchestrators  - Mesos, Yarn, Conductor
Potentially suits containerization – we can test this on our Openstack Nimbus cloud
In-memory and parallel distributed processing
Scala, Python and Java
Workload use cases:

- SparkSQL
- Spark Streaming
- Machine Learning
- Graph Processing
- Deep Learning  - Power9 and GPU (investigating)

# What is Spectrum Conductor?

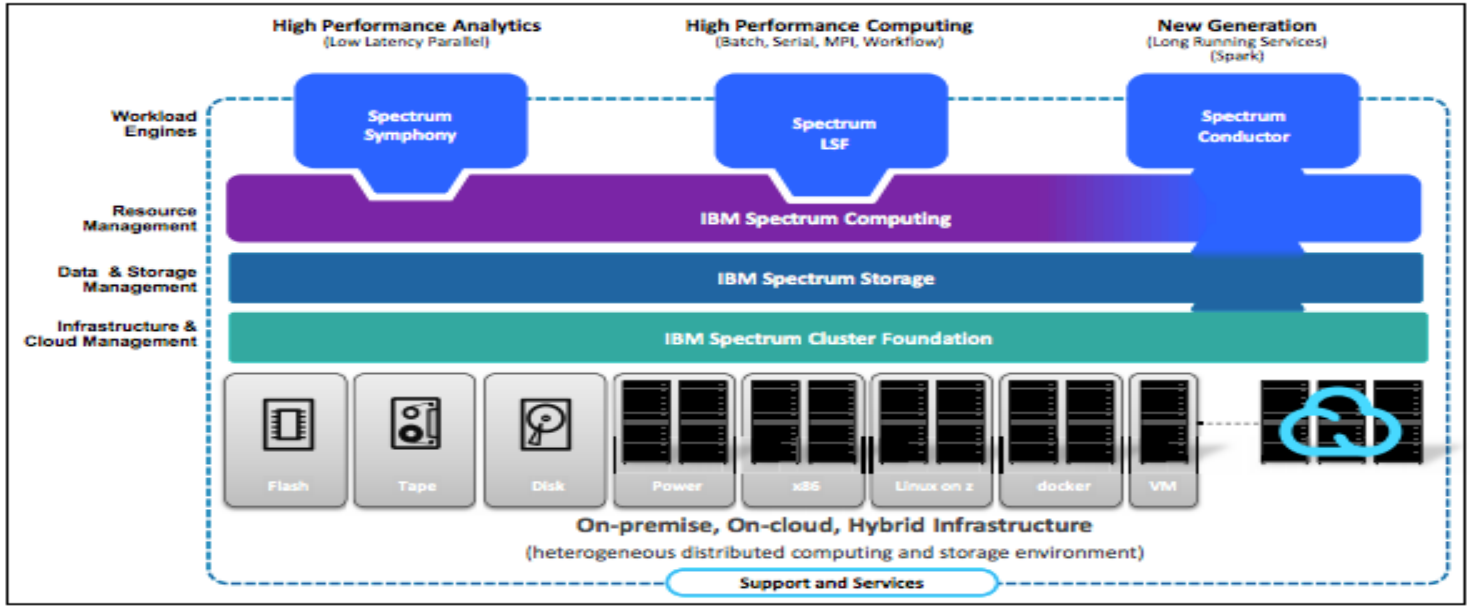
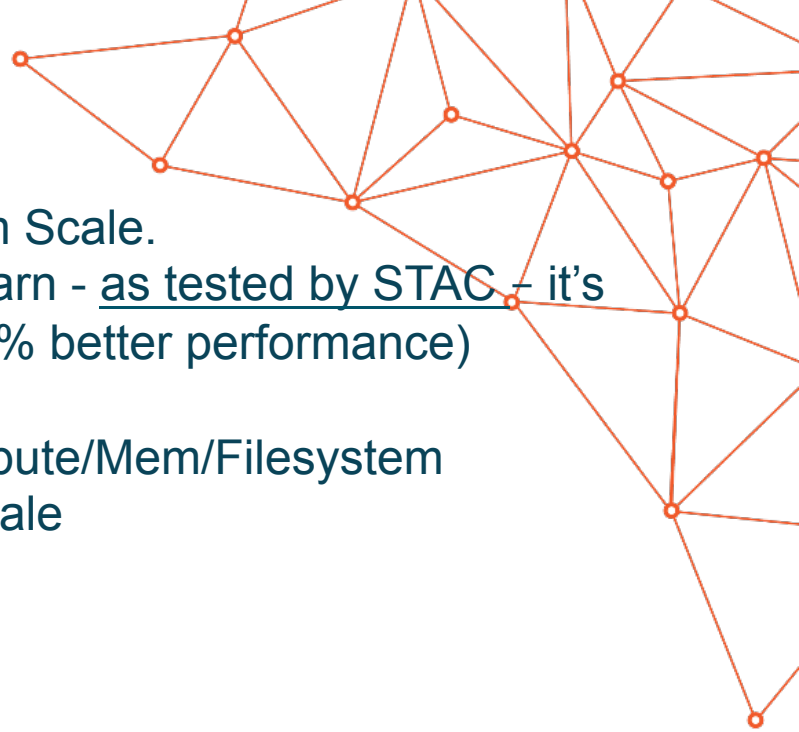



*Figure 3 IBM software defined infrastructure solution*

PAWSEY supercomputing centre



- Conductor is part of Spectrum Computing and it is integrated with Spectrum Scale.
- It is the most efficient Orchestrator (EGO) – vs Apache Mesos or Hadoop Yarn - as tested by STAC – it's the top Orchestrator in Spark Multi-tenancy Benchmark testing. (at least 30% better performance)

- It is efficient, it's better at increasing utilization of existing resources – Compute/Mem/Filesystem
- Shuffle  - default file algorithm is disabled  - Conductor is integrated with Scale
- It's HA (Master node and shared filesystem)
- Comes with it's own Monitor

- Enterprise class-solution for Spark workloads and management
- Jupyter Notebook support
- Supports SpectrumScale CES OR FPO (File Placement Optimizer) – more space-efficient than HDFS
- Best integrations with Spectrum Scale – Client
- Has GPU support – CUDA, OpenCL

# Journal

- Upgraded Spectrum Scale (GPFS) v4.2.3 on test DDN SFA12KX
- Setup LDAP service accounts
- Initially we installed Conductor with Spark (CwS) orchestrator 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(AMD64), 40GB RAM.
- Then we deployed Master and three Slave nodes on Test SGI Server HW CentOS (tested Ubuntu but had issues so we changed)
- Add in Enterprise SSL Certs for secure client and Master slave node
- LDAP Migration

Issues:

- AMD support and OS version support for troubleshooting with IBM
- NFS performance

# “Population” - Scale Genomics Use Case

Data61 Use case

Big, extremely highly-dimensional data.

Suitable for multi-gene research (whole genome), examining multi-gene causes of common complex diseases of diabetes and cancer.

We compared our test environment results accuracy with a DataBricks Spark workload case.

But we had to rework the VariantSpark code example Java and Shell code.

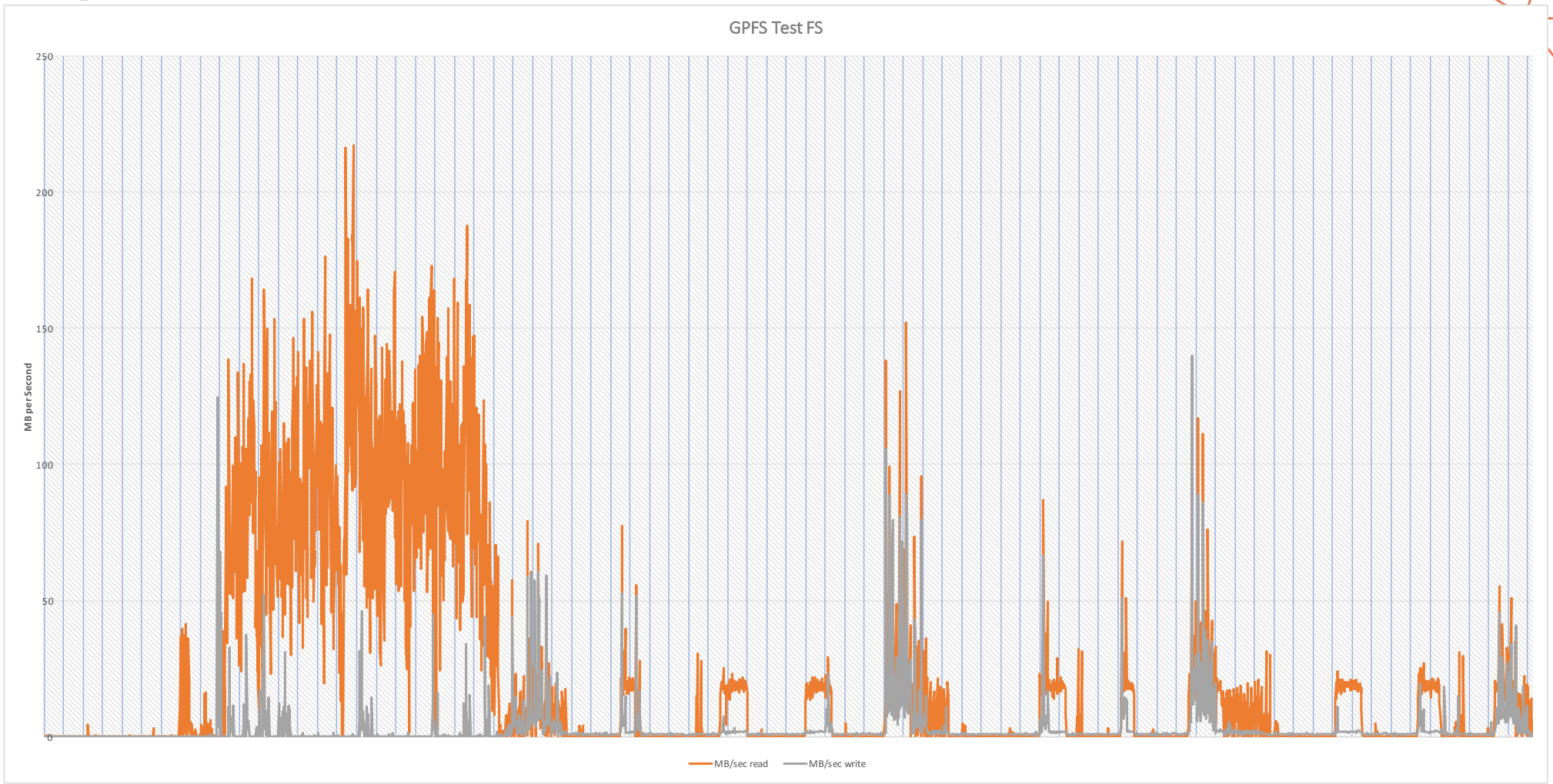
Hipster genotype data set 16MB vcf file.  Half million variables. 1092 rows and labels.

We ran a Random Forest Walk with 500 trees generated to find the 20 most important features.
Random forest oob accuracy: 0.136, took 288.191s

What matters? Speed? Well yes, but more so - accuracy – i.e. out of bag and run time (seconds).

What we found was even though we have older test HW and FS on prem – we were told by Data61 our results are comparable to DataBricks.
(Thank you GPFS and Conductor).

# mmperf IO Profile



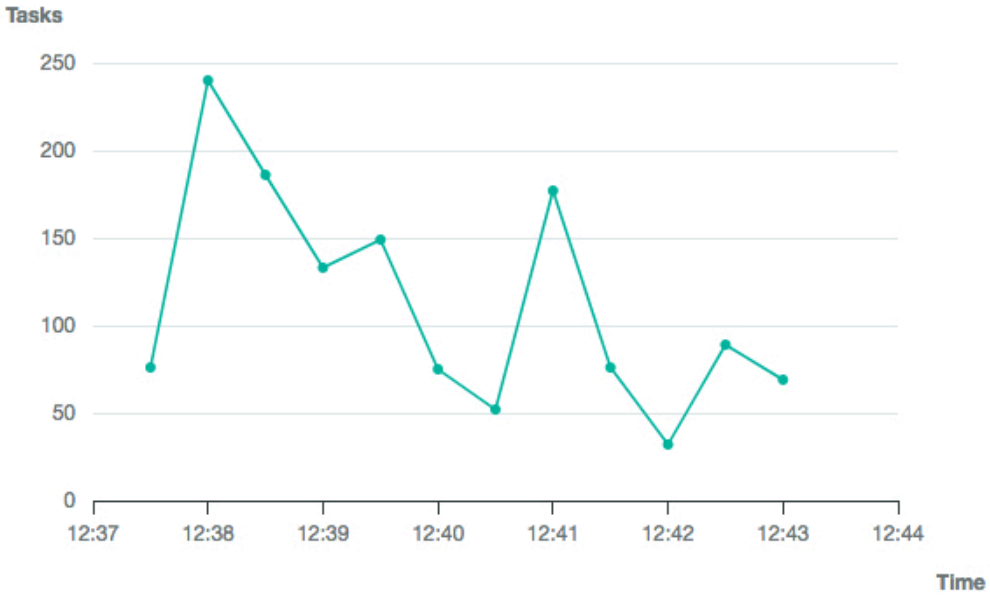

PAWSEY supercomputing centre

# Population Scale Genomics runs

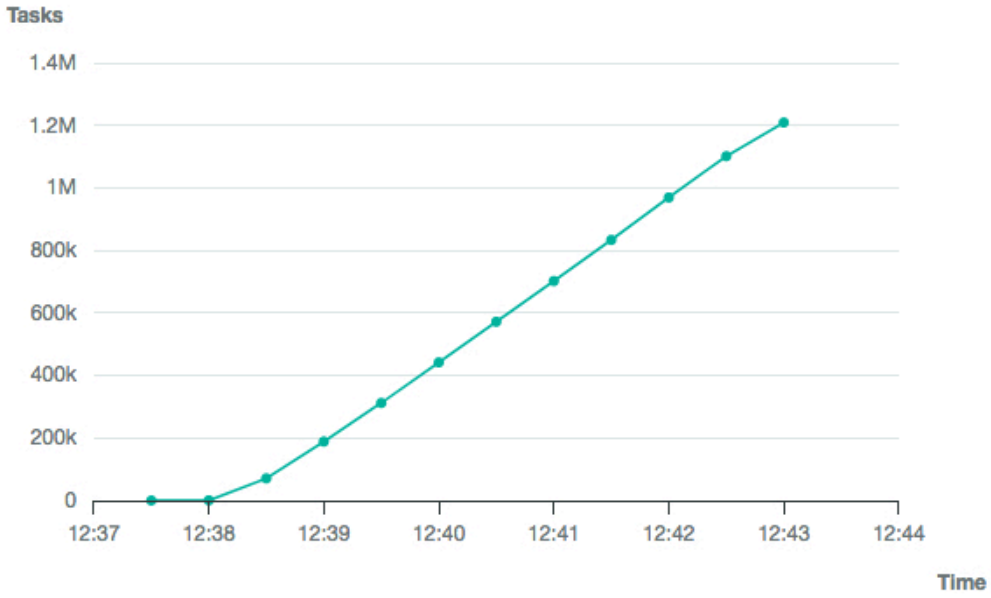
Entire time period

7/6/2018 12:37:27 PM to 7/6/2018 12:43:19 PM

## Running tasks




## Completed tasks




## Task durations

Click the values in the Median Task Duration column to view distribution for completed tasks.

| Stage ↑ | Median Task Duration | Executor with Longest Task |
|---|---|---|
| 0 | 0.58s | 0-a2f7f752-4673-4bb7-a9b1-eb66fb27f29d(0.58s) |
| 1 | 11.77s | 3-6e5555ab-abb3-4f26-9be1-91e1bd65ede8(15.33s) |
| 2 | 6.26s | 0-a2f7f752-4673-4bb7-a9b1-eb66fb27f29d(9.44s) |

PAWS supercomputing

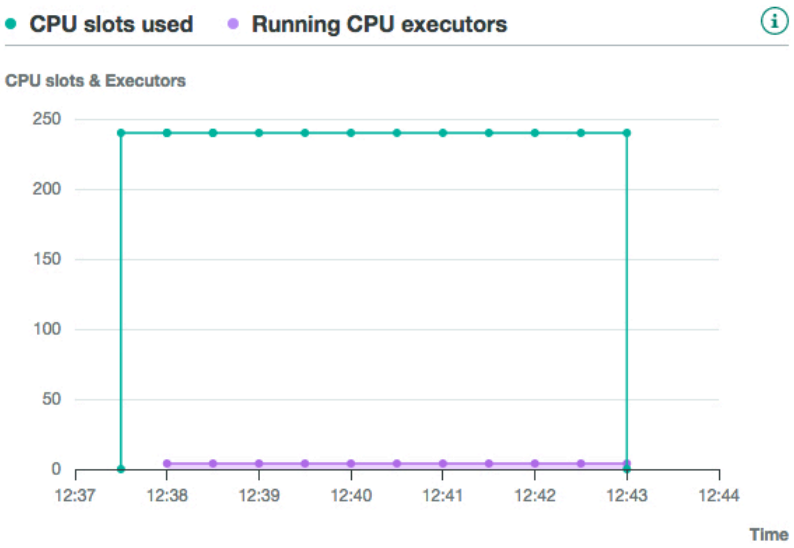
# Population Scale Genomics runs continued

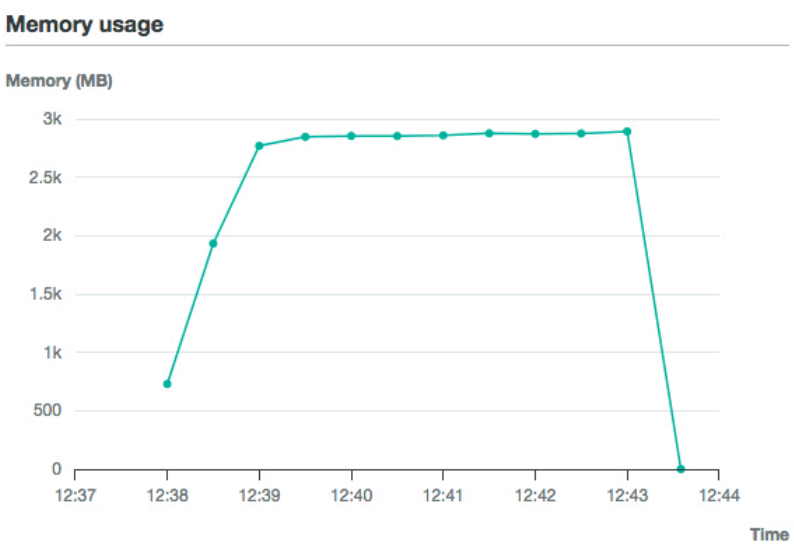
Finished app-20180706123727-0000-6220e366-e20b-4201-83b4-cfcb781126dd

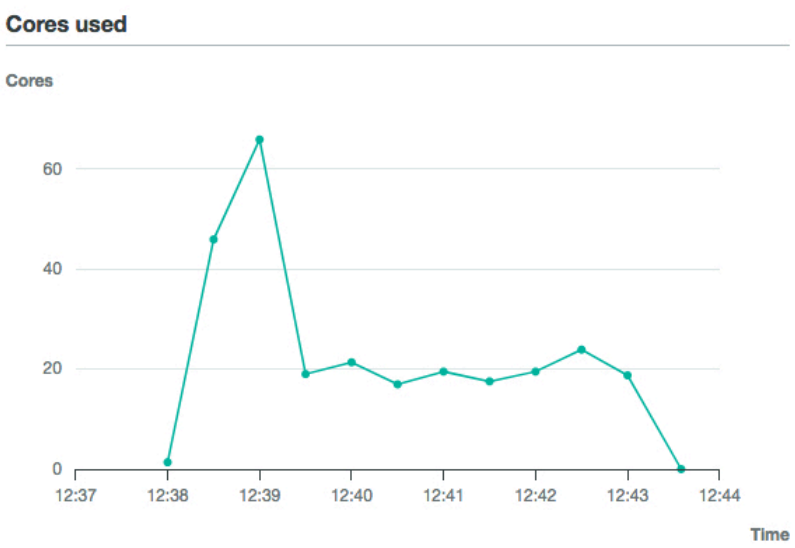
## ImportanceCmd

Overview | Drivers and Executors | Performance | Resource Usage

Entire time period ▾ 7/6/2018 12:37:27 PM to 7/6/2018 12:43:19 PM




PAWSEY supercomputing centre

# *Ironspark* – console

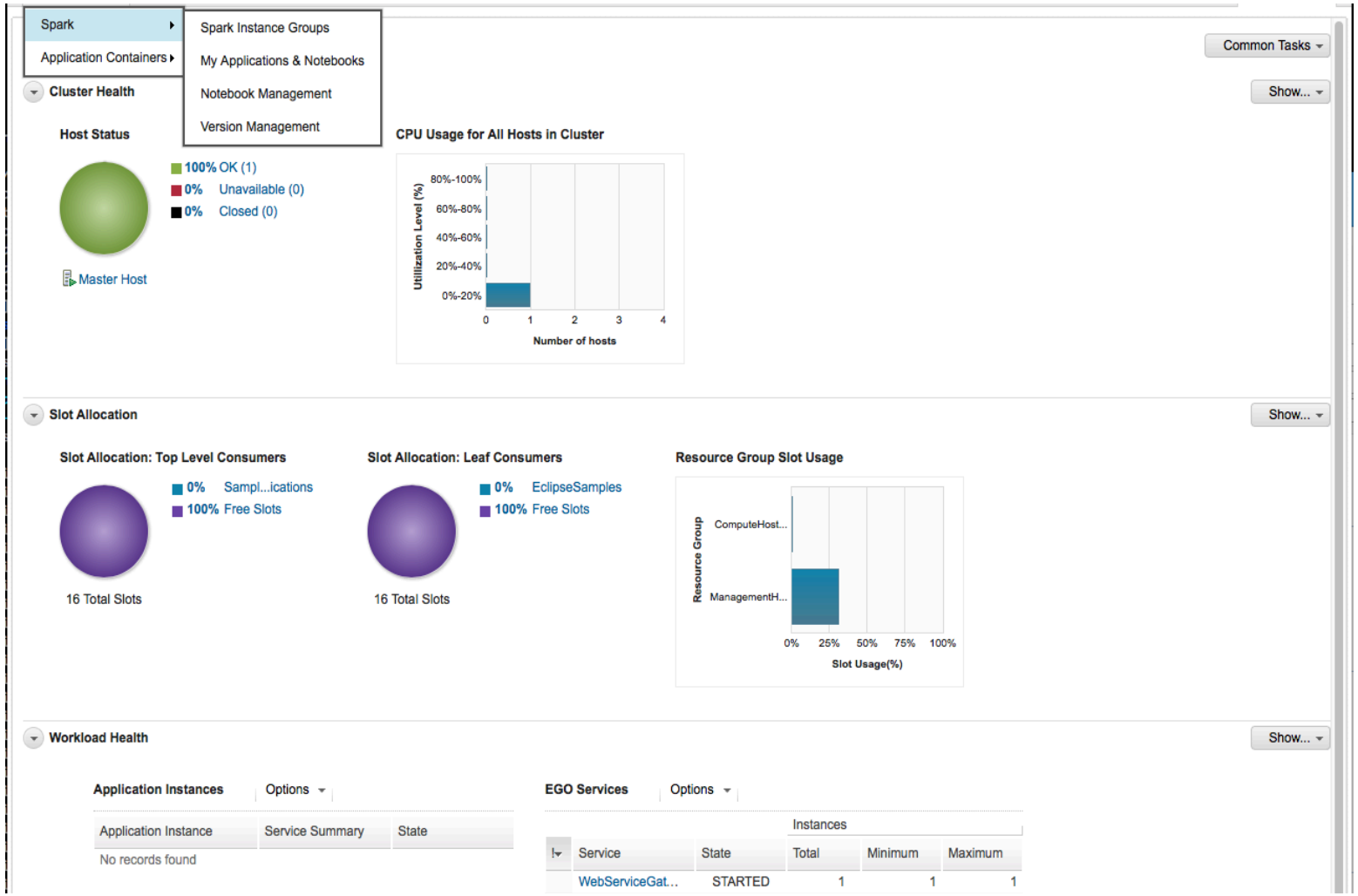
Application Submission Interfaces:

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

Scenarios:

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

# IronSpark Console




PAWSEY supercomputing centre

# So, where does this leave us?

## Is Conductor on Scale worthy of a new data service?

- Our results are good so far – we want more use cases – especially those that reuse existing unstructured data sets - maybe a ML example for DFN
- We need to build the value case if we want to launch this as a full, new service
- Spark is useful, but it is not necessarily super popular with our researchers - so we are looking at Deep Learning Impact – maybe with our new GPU's.
- Profiling IO on GPFS is ongoing – for HW specifications. We need larger Structured Data examples in the 100'sGB-100'sTB size, reading a small structured data file is not hard.

  What we have run is very fast but not a "difficult" workflow for GPFS.
- Ongoing work to integrate Conductor data workflow with our Data Portal (product integration, ILM Policy, maybe Watcher – V5 GPFS)

## Recommendations:

- If you can, use physical nodes.
- Use CES carefully – we found NFS didn't "cut it" – so we now use GPFS Client (10GBE).

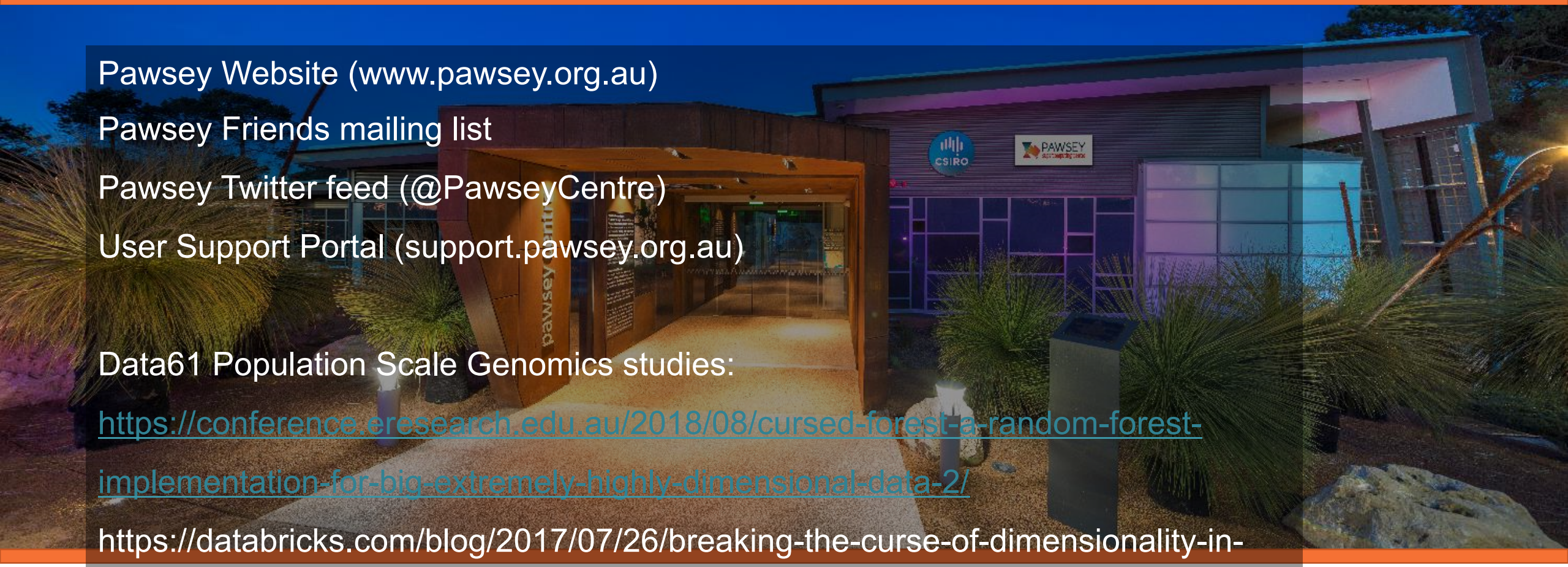
## Applicability:

- Multi-dimensional data set analysis methods may have applications to other very large, multi-dimensional data sets – perhaps Radio Astronomy data (we have a lot of that – over 30PB)…
- There's a lot of competition with other Containerised options for HPC, what makes a difference here is integration with the Data Lake. Ubiquity, Watcher.

## Do you want to test?

- Maybe on BlueMix - https://git.ng.bluemix.net/ibmcws-spark-samples/conductor-bluemix-schematics

# Find out more




Pawsey Website (www.pawsey.org.au)

Pawsey Friends mailing list

Pawsey Twitter feed (@PawseyCentre)

User Support Portal (support.pawsey.org.au)

Data61 Population Scale Genomics studies:

https://conference.eresearch.edu.au/2018/08/cursed-forest-a-random-forest-implementation-for-big-extremely-highly-dimensional-data-2/

https://databricks.com/blog/2017/07/26/breaking-the-curse-of-dimensionality-in-genomics-using-wide-random-forests.html

PAWSEY supercomputing centre

# Questions? Where to go next?