The University of Queensland
Australia
CREATE CHANGE

Evolving a campus-wide research storage capture, analysis and management strategy

A five year journey with SpectrumScale.

Jake Carroll, Chief Technology Officer, Research Computing Centre, The University of Queensland, Australia.

jake.carroll@uq.edu.au

UQ has been on a data fabric journey for five years.

This is a (very) abridged story of how things have panned out for us…


I have a lot of big problems.
DC3 PNs
Cell 1
Cell 2
Cell 3
CA
LH
DC3 PN Boutons
Confocal
ExLLSM
t = 0 s
t = 80 s
mEmerald - Lifeact
tagRFP - plasma membrane
ANDOR
NextSeq 500
Welcome
illumina
BRUKER
900
UltraStabilized
FEI

CAI

QBI

100’s of terabytes per day of data generated.

These are my “*problem children*”

IMB

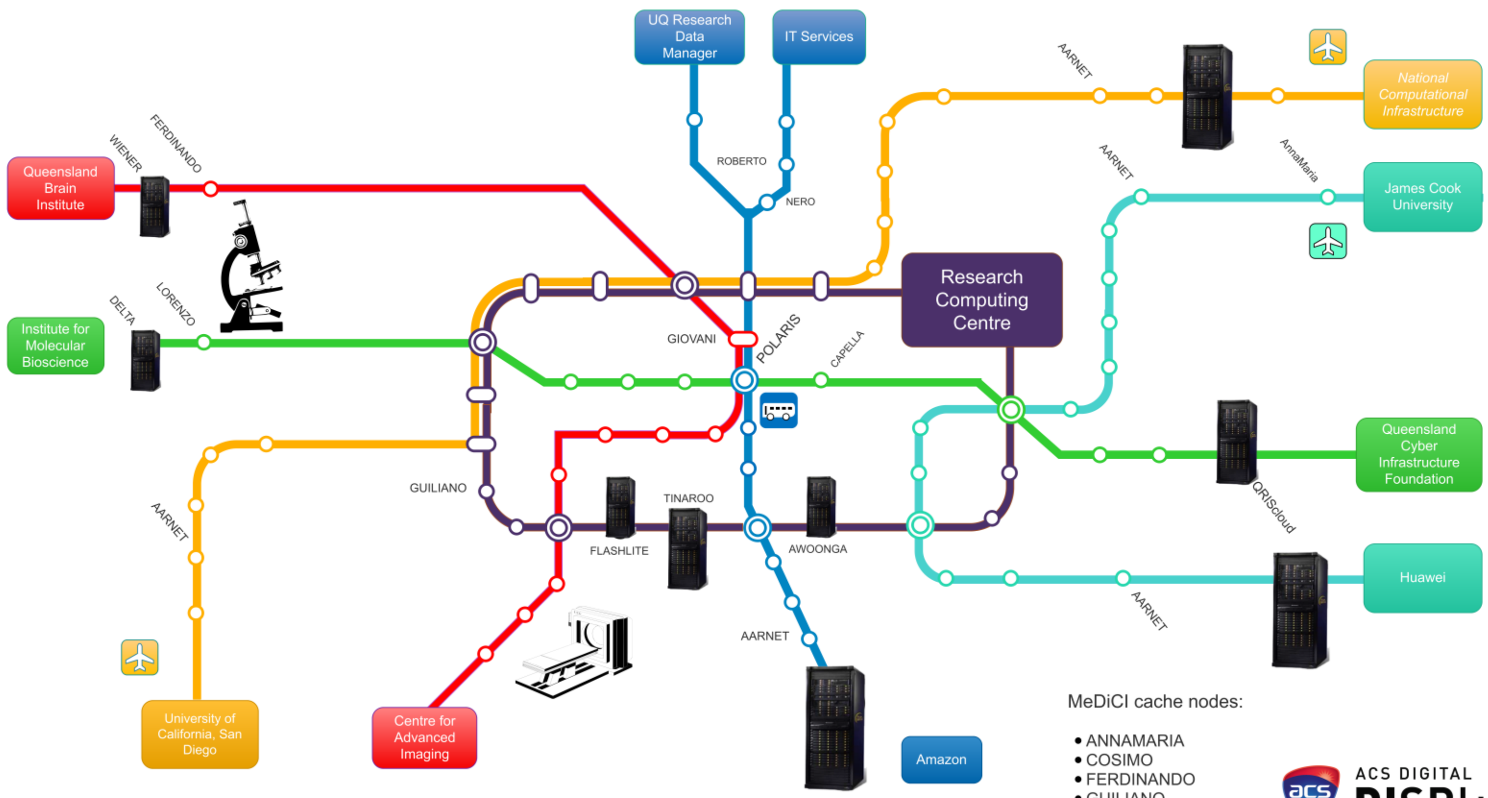
AIBN

CMM

MeDiCI (2019): UQ's Data Fabric of Choice

The University of Queensland's Metropolitan Data Caching Infrastructure (MeDiCI) provides seamless access to data regardless of where it is created, manipulated and archived.

Developed by the Research Computing Centre (RCC), MeDiCI holds copies of data on campus until it is not required for some time. Data is moved between on and off-campus storage on demand without user involvement. MeDiCI is underpinned by HPE DMF, DDN Grid Scalar storage, and IBM Spectrum Scale technologies.



MeDiCI cache nodes:

- ANNAMARIA
- COSIMO
- FERDINANDO
- GUILIANO
- GIOVANI
- CAPELLA
- NERO
- ROBERTO
- PIERO

ACS DIGITAL DISRUPTORS AWARDS

powered by aarnet | Hewlett Packard Enterprise | intel | The University of Queensland Australia | DDN Storage | IBM | HUAWEI | aws

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

MeDiCI – How to Withstand a Research Data Tsunami

December 12, 2017 by Rich Brueckner Leave a Comment

Connected high-speed data storage coming to University of Queensland

The Polaris Data Centre in Springfield, just outside of Brisbane.

22 Feb 2016 St Lucia - *The Research Computing Centre (RCC) at the University of Queensland (UQ) is building an experimental data fabric that connects on-campus computing hardware to data storage in QRIScloud, the Queensland node of Australia's research Cloud. Working with UQ's IT Services Division on a proof of concept project, RCC is building MeDiCI, the Metropolitan Data Caching Infrastructure, that supports data collection on campus, close to scientific instruments, data sources and local computers. These collections will be connected to the significant storage and processing capabilities of the Polaris Data Centre, where QRIScloud is sited.*

Polaris, located in Springfield, just outside of Brisbane, offers a state power back-up, air conditioning and security monitoring, providing a storage.

Whilst UQ is presently connected to Polaris through high-speed net separately and users are responsible for initiating data movement b

MeDiCI will deliver a hierarchical storage fabric with cache nodes or allows researchers on campus to access local, high-speed storage,

Importantly, data will be migrated automatically between campus ca pipes, providing an illusion of large data holdings on campus. Data using QRIScompute - QRIScloud's compute service - and/or UQ's n FlashLite, which is also housed at Polaris. This will allow unprecede

RCC is testing commercial off-the-shelf components, and expects to

Source: University of Queensland Research Computing Centre

itnews

Sign up to our free newsletter Bit BUSINESS IT

GOVERNMENT | SECURITY | FINANCE | TELCO | BENCHMARK AWARDS | DIGITAL NATION | LOG IN | SUBSCRIBE

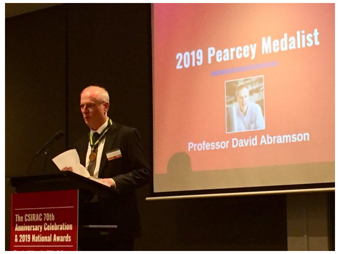
Qld Uni debuts its own data storage fabric

By Andrew Sadauskas
Mar 30 2016
6:15AM

2 Comments

Using on-campus cache to speed up off-site data access.

The University of Queensland has prototyped a new data storage fabric it hopes to roll out by the end of the year to improve data access between its main campus and its off-site data centre.



MeDiCI

Centralising research data storage and computation

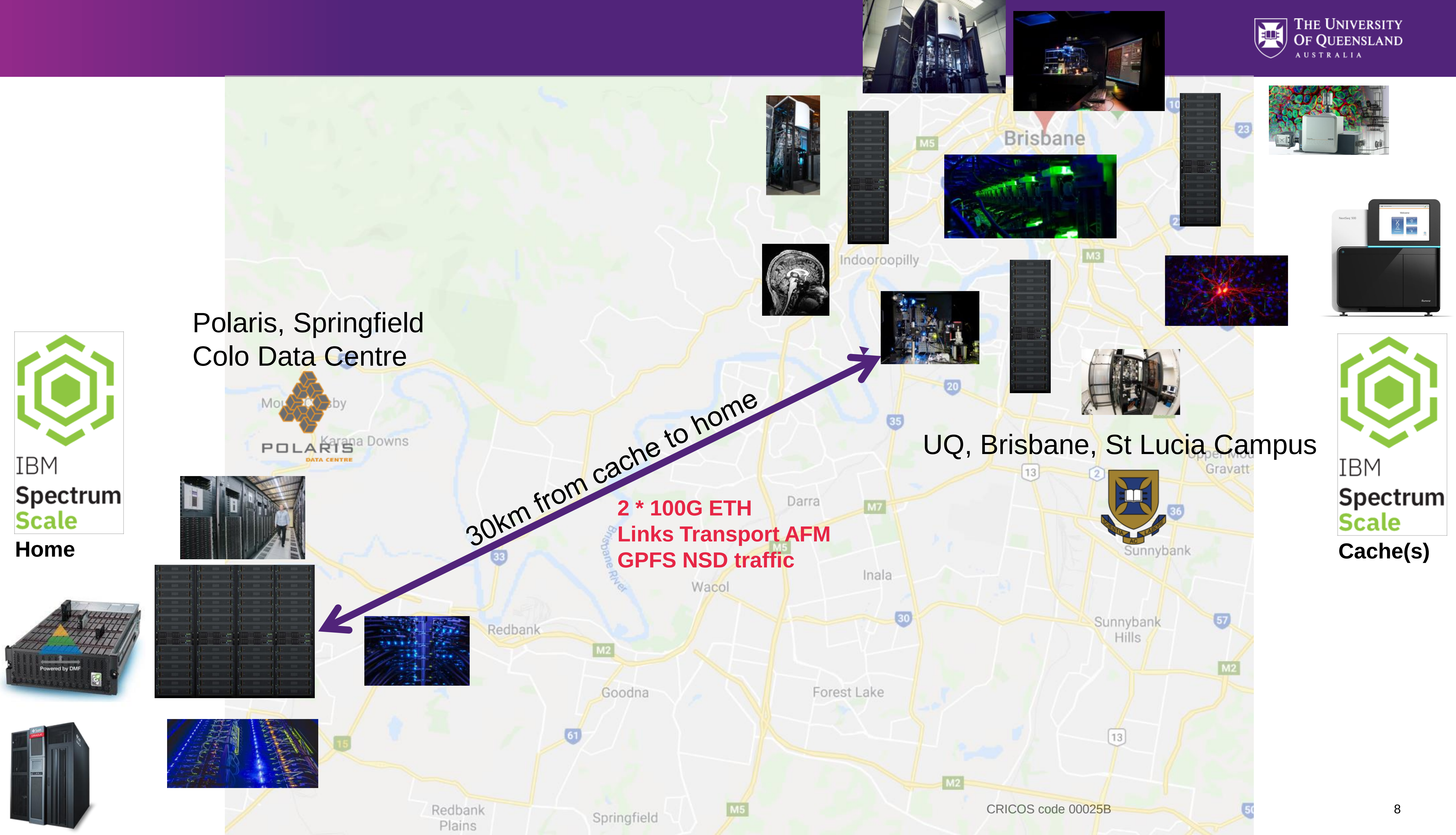
Distributed data is further from both the instruments that generate it, some of the computers that process it, and the researchers that interpret it.

Existing mechanisms manually move data

MeDiCI solves this by

- Augmenting the existing infrastructure,
- Implementing on campus caching
- Automatic data movement

Current implementation based on IBM Spectrum Scale (GPFS) and HPE DMF



THE UNIVERSITY OF QUEENSLAND AUSTRALIA
Polaris, Springfield
Colo Data Centre
POLARIS DATA CENTRE
IBM Spectrum Scale
Home
30km from cache to home
2 * 100G ETH
Links Transport AFM
GPFS NSD traffic
UQ, Brisbane, St Lucia Campus
IBM Spectrum Scale
Cache(s)
Powered by DMF
Brisbane
Indooroopilly
Darra
Inala
Wacol
Redbank
Goodna
Forest Lake
Sunnybank
Sunnybank Hills
Gravatt
Karana Downs
Redbank Plains
Springfield

Trying to avoid historical data-silo effects…

Stop trapping things at the building, instrument, data-centre and institutional level…

We’ve been able to do some amazing things with fabrics.

GRAPHICAL USER WEB BASED INTERFACE FOR BATCH PROCESSING OF IMAGES ON A LINUX BASED GPU HIGH PERFORMANCE CLUSTER

Hoang Anh Nguyen, Zane van Iperen, Jake Carroll, David Abramson [*]
Nicholas Condon, Mark Scott, James Springfield[**]
[*] Research Center for Computing
[**] Institute for Molecular Bioscience
University of Queensland, Queensland, Australia
E-mail: j.springfield@uq.edu.au

KEY WORDS: GPU, HPC, Cluster, Software, Image Processing, Deconvolution, Web, GUI, Big Data, Lattice Lightsheet, Andor Dragonfly.

The latest generation of sCMOS camera based microscopes such as the lattice lightsheet and Andor Dragonfly spinning disc confocal have placed growing demands on researchers to process, analyse and store huge datasets an order of magnitude greater than what was considered normal only a few years ago. Unfortunately the hardware and software systems which have been built to handle such large data (namely Linux based High performance computing, HPC), are typically managed by IT specialists, and are not considered layman friendly. Our goal, was to produce an intuitive Image Processing Portal that our core facility users, with no HPC experience could use with minimal support or training. We present a web portal that is capable of performing large scale, batch processing of microscopy images within a GPU based HPC. This portal provides intuitive web pages allowing users to login remotely from anywhere in the world to submit image processing jobs.

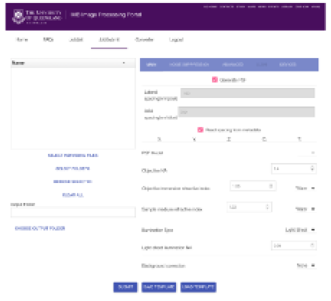
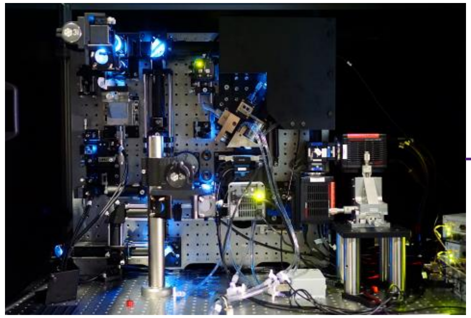


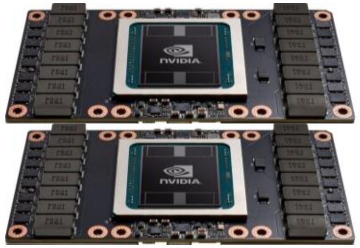
Figure 1: Batch Deconvolution

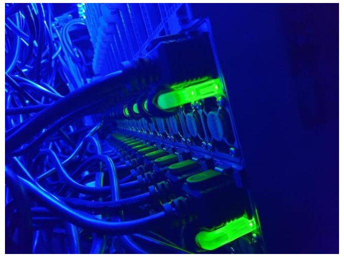
Utilising the portal to submit jobs, image deconvolution is performed using the Microvolution deconvolution engine [1] on our Wiener GPU cluster [2] via multiple Dell r740 nodes each containing two NVIDIA V100 GPUs.

Other utilities being developed for the Image Processing Portal include: A file format conversion page, which allows users to easily batch convert their datasets between image formats. This was originally developed to deal with large Andor Dragonfly “.ims” files which could not be opened in ImageJ/FIJI.

“One click” to HPC from LLSM



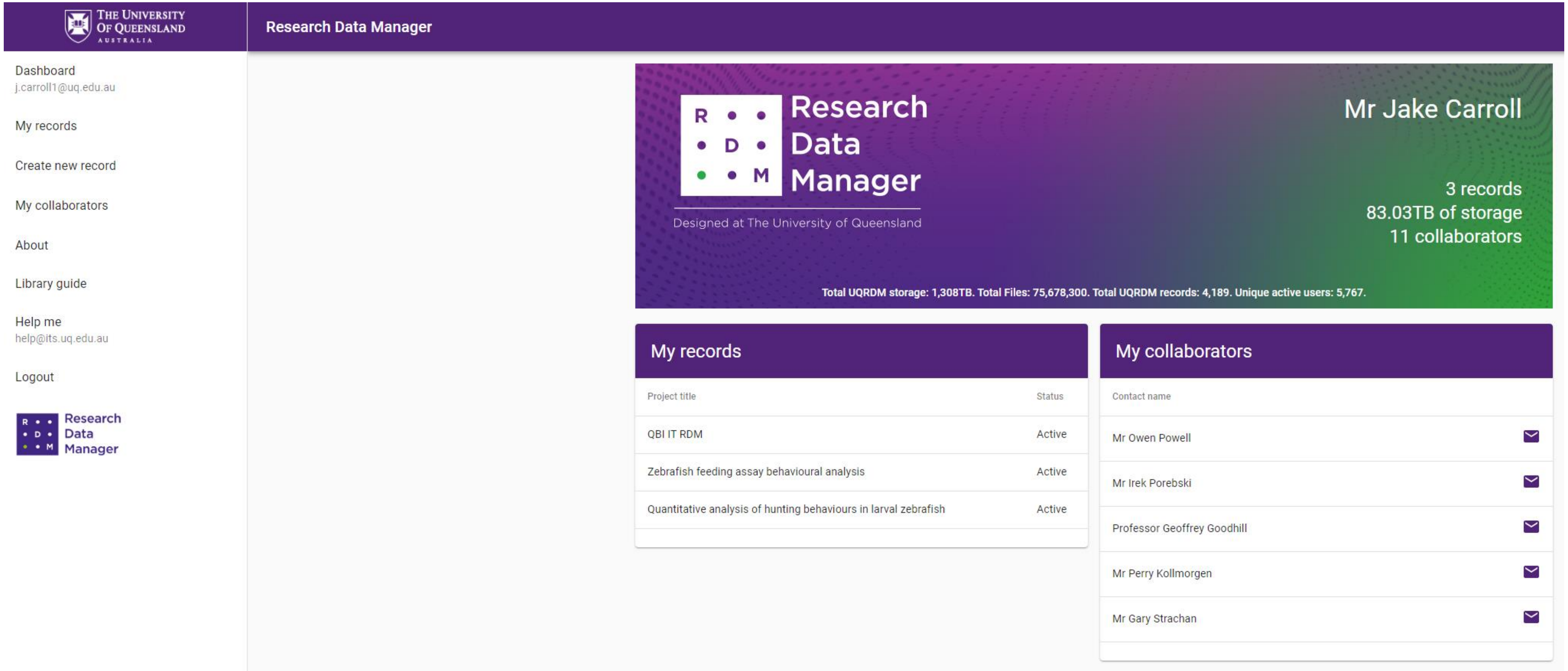




Supercomputing scale deconvolution via a friendly web portal – leveraging many HPC GPUs (Volta), multi GB/sec parallel filesystems and automatic data movement – without the researcher needing supercomputing expertise.

UQ’s Research Data Manager

https://rdm.uq.edu.au

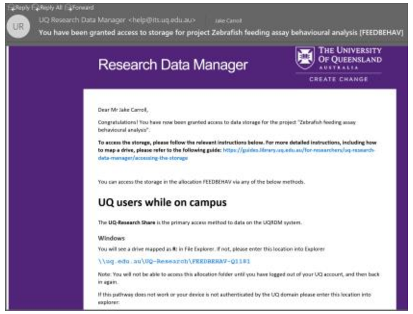


A typical workflow

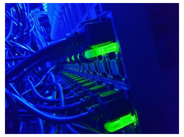
Research Data Manager

Designed at The University of Queensland

Researcher obtains a collection. *Qabcd*

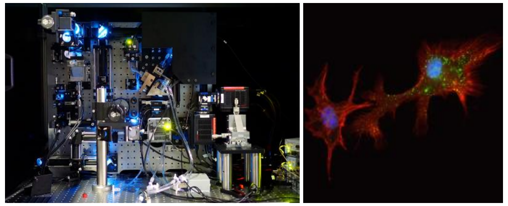


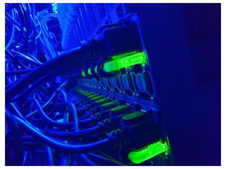
Researcher is sent an email explaining access instructions.

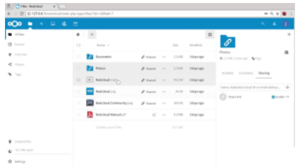


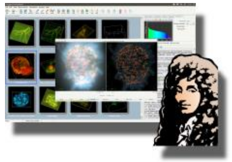
Qabcd then mounted on the fabric.

Researcher then acquires data and stores in Qabcd







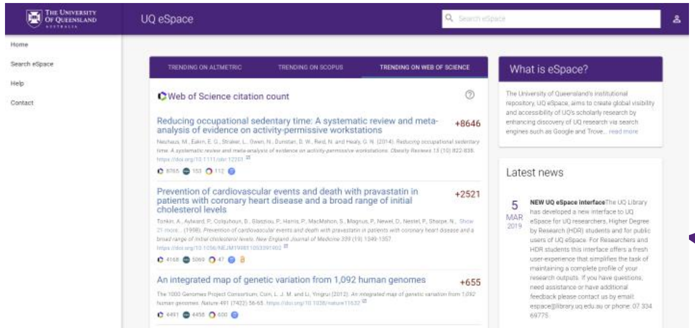


PYTORCH

nectar

FAST

It is already on the fabric. Can be accessed via CVL, supercomputers, other.



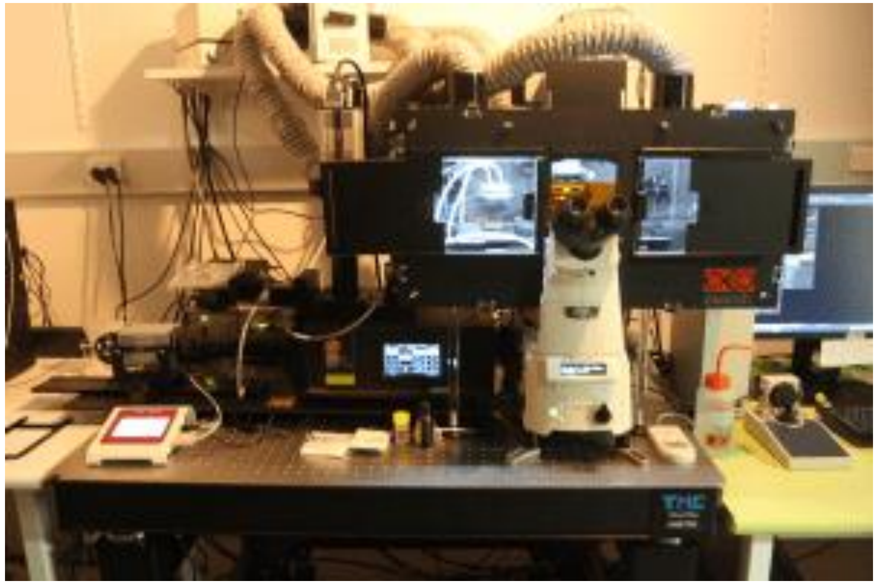
DOIs can be minted, published to UQ eSpace.
RDM can facilitate data linking back for durable URL.

Challenges in making our data fabric a reality.

So many things…

- Windows. SMB integration. Protocols. NFS stability.
- At the beginning, in 4.2.x – AFM didn't know how to map UUIDs from cache to home with two completely different mmname2uid and mmuid2name authentication spaces. The code was literally missing.
- AFM performance was…questionable. Resource utilisation was massive and variable. Many bugs, PMR's, eFixes later…
- NFS backend at "home" caused unusual timeouts/stalls in LoomHA [maybe some of you know what LoomHA is!] and created stubbing issues.
- Management of filesets was and still is hard. 1000's of filesets…
- Hardware variability – we started out down a rocky path and didn't really know "right" from "wrong" when it came to hardware sizing, options, appliances, OEMs. Made some mis-steps and learned a lot of hard and painful lessons.
- Virtualising gateways and protocols at our scale is a really bad idea…
- NSD VerbsRDMA on one site trying to transfer over ETH wire to "home" where we have mixed use at "cache" created odd "root map" bugs. Yeah, that wasn't fun. When you mix supercomputing with userland…
- Quirky interop stuff between 5.x caches and 4.x homes.
- Us disobeying the golden nfs4 acl vs POSIX vs "all" rules…
- IW vs SW resource utilisation differences laid to bare.
- The things that happen when you don't have enough meta-data IOPS in your NL-SAS spindles but try to do meta data intensive operations *anyway*….

The *scourge* of Windows….



Despite the baffling array of technology, lasers, sensors, CCD's, sCMOS, ultra sensitive mirrors, PWM control and automation, the unfortunately vast majority of our most special scientific instruments run **Windows. What is worse? The vendors won't let us touch them…**

Why is that bad?

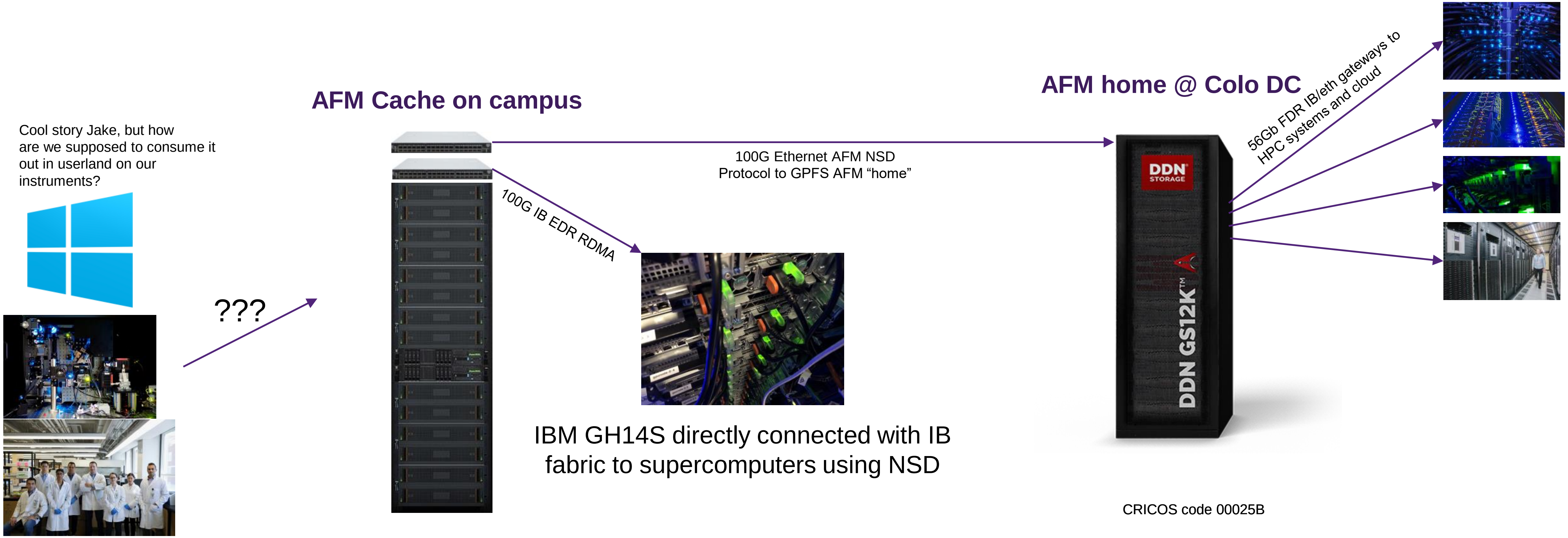
Try running SpectrumScale GPFS NSD POSIX client on one of these, in a huge enterprise environment, with locked down requirements from the vendor, a complex security regime and best of all – and active directory domain that says you cannot have a user called "root*" in the forest anywhere at all.

So, we are faced with one lonely choice. **SMB!**

*For all those playing at home, no, you can't use GPFS NSD POSIX client on windows without a user called "root" – which no sensible admin would ever allow.

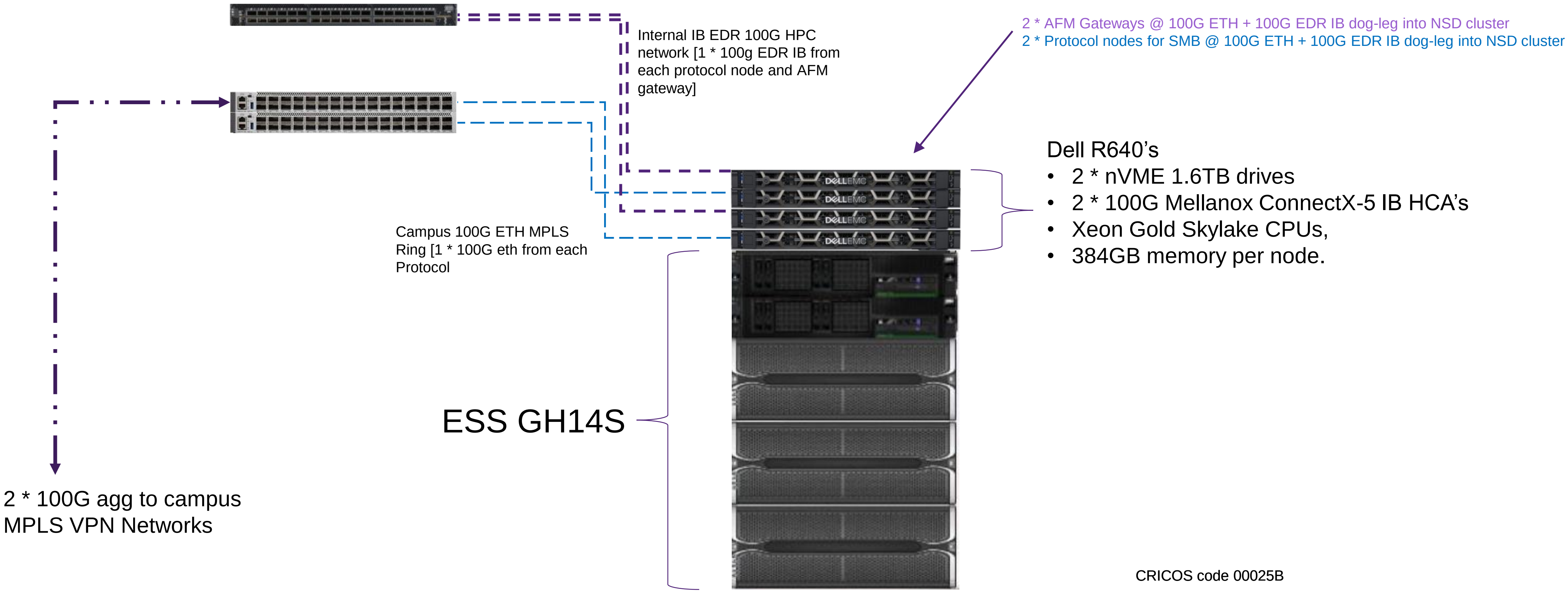
Instrument data was stuck in Windows workstations…

We knew we needed to somehow leverage the good from SpectrumScale up and out to our users at their desk, outside the HPC fabric. It was doing great things for our people with our freshly implemented IBM GH14S – but this was “enclosed” inside HPC.



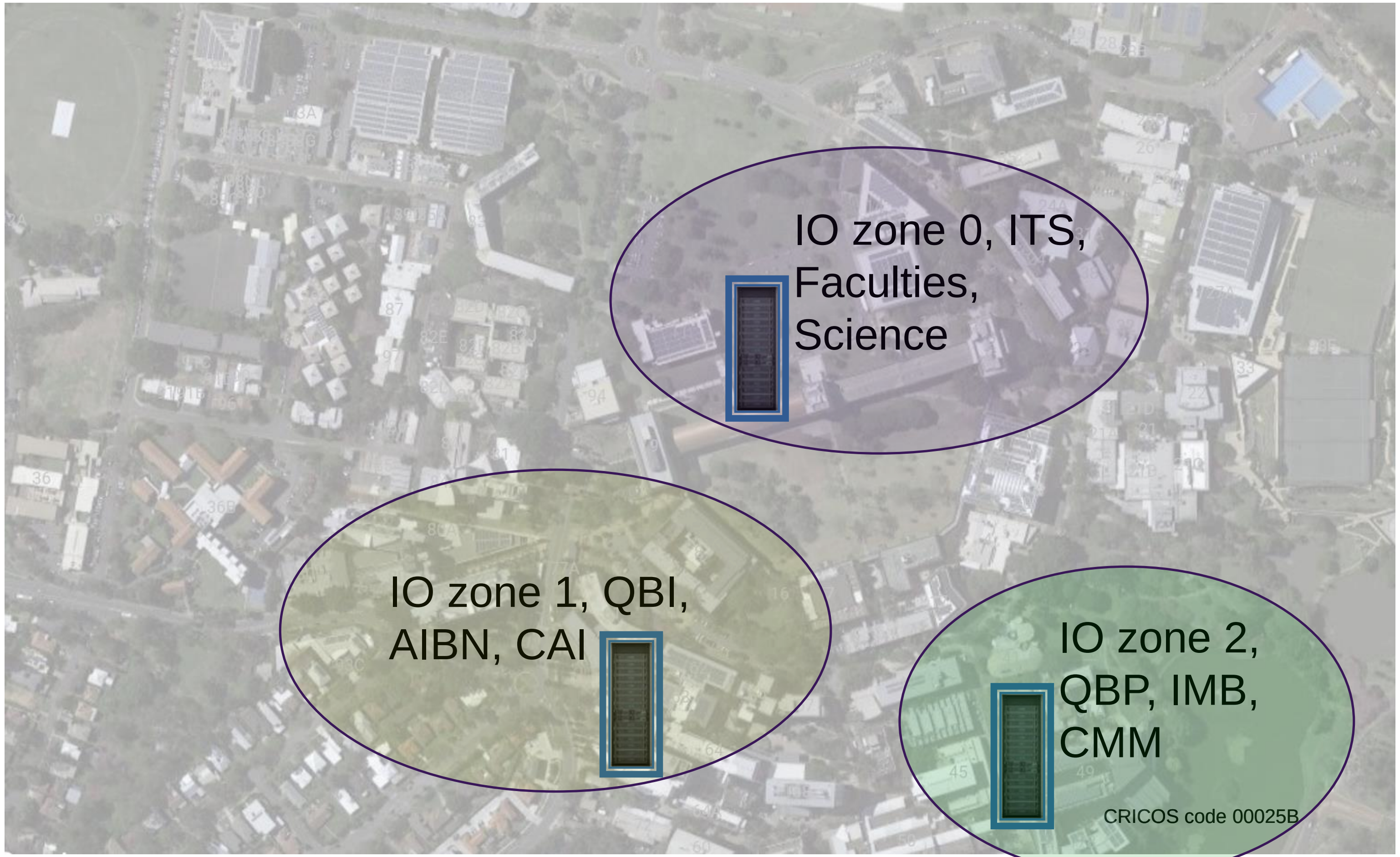
Spectrum Scale CES [Protocols]

We knew we had to build protocol nodes like no other protocol nodes had ever been built before, to handle the sheer weight of multiple instruments generating untold IO on a 100G instrument network.



A step further. Distributed IO zones around campus.

Cache and IO zone distribution around campus – putting caches near our instruments and IO intensive locations, using AFM, protocols and an SMB, to “knit” the fabric together…



Smart things with DNS, DFS and presentation of shares

We've been able to manipulate DNS and DFS a little bit to create a software defined set of IO zone hot spots around the campus where we can share the IO requirement out and bring up the share where it is needed, almost "elastically"…

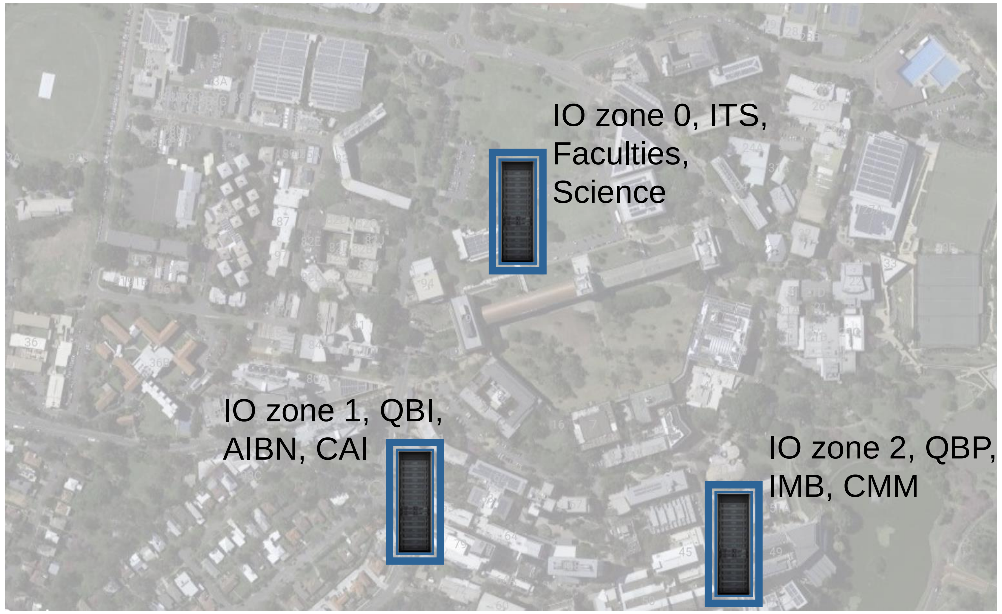
```
data.aibn.uq.edu.au
data.cai.uq.edu.au
data.imb.uq.edu.au
data.qbi.uq.edu.au
```

DNS addresses
Point at protocol nodes virtual clustered address space
DNS Round Robin…

DFS
Distributed File System

\\uq.edu.au\UQ-Inst-Gateway1
\\uq.edu.au\UQ-Inst-Gateway2

Windows DFS top level names for auto-mapping…



The power of multiple caches working together…

We have the ability to "duct" IO where we want, transparently, depending upon our C-name manipulation. We can also bring up the same filesets into different caches and push the IO around to different concentration points….

```
data.aibn.uq.edu.au
~~data.cai.uq.edu.au~~
data.imb.uq.edu.au
data.qbi.uq.edu.au
```

If a cache fails, we can simply re-cname to a different protocol cluster and have the fileset mounted. To the user, it is like connecting to the same unit and they cannot tell the difference. We are flushing to AFM home constantly, so when a *fileset* is rehomed to another cache, rarely is it the case that a user will see missing data.

DFS masks the share map, the cname and the required thought. It lets us string up **Filesets** onto any cache we want and have the user automatically resolve to it…

What is happening as we speak.

We have IO zone 1 up and running in building 79, AFM’ing back to home.

```
data.aibn.uq.edu.au
data.cai.uq.edu.au
data.imb.uq.edu.au
data.qbi.uq.edu.au
```

Currently doing this.

In a few days time, IO zone 2 will turn on, shifting imb and adding CMM for our large CryoEM load out…

```
data.imb.uq.edu.au
data.cmm.uq.edu.au
```

To then doing this.

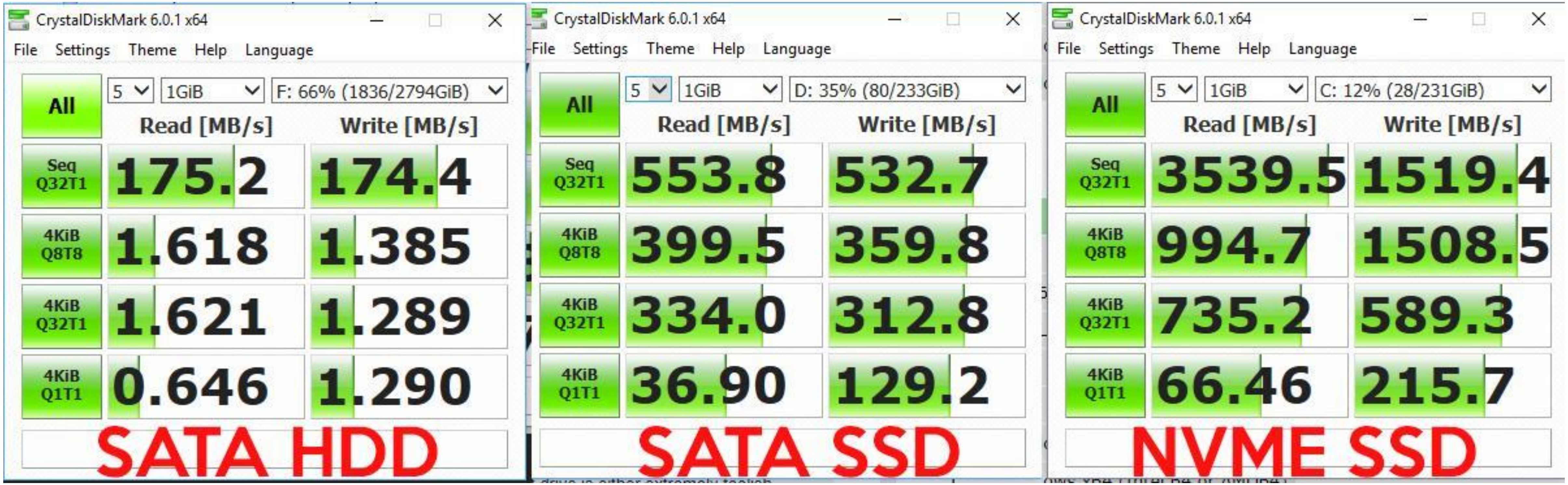
The challenges of this much data and the SMB protocol

samba

!=

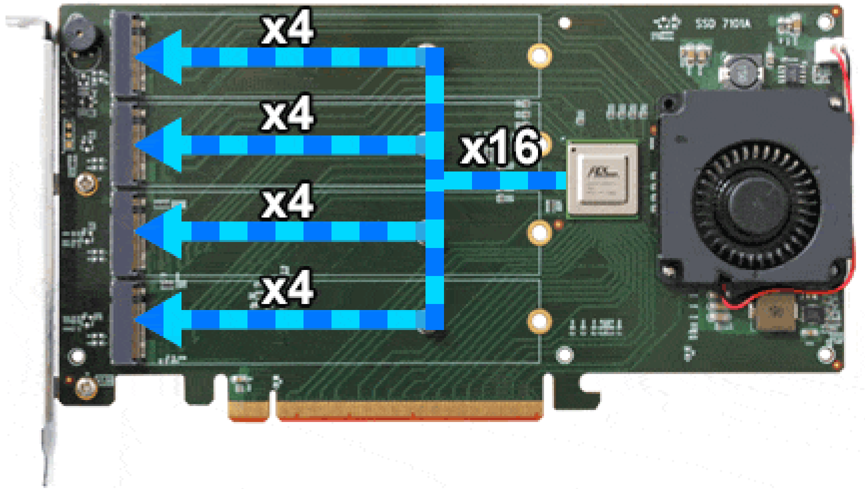
It is nice to have a 100G pervasive campus network for our fabric, but it is kind of irrelevant if the SMB stack isn't going to go much faster than 1.5GB/sec per stream…

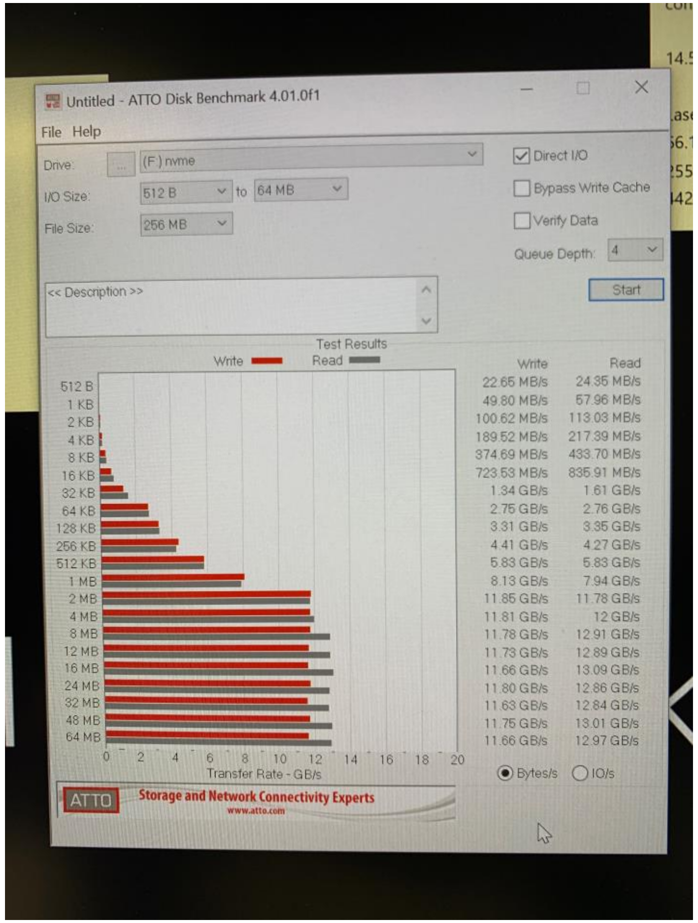
Workstations start to show their weaknesses (IO subsystems)



You can have all the network bandwidth in the world and even an amazing 100G ethernet card but it will all fall in a heap if your IO subsystem isn't up to the task on the localhost…

We are retrofitting with very high end NVME aggregation controllers just to keep our 10 and 100G pipes "full"…





	Write	Read
512 B	22.65 MB/s	24.35 MB/s
1 KB	49.80 MB/s	57.96 MB/s
2 KB	100.62 MB/s	113.03 MB/s
4 KB	189.52 MB/s	217.39 MB/s
8 KB	374.69 MB/s	433.70 MB/s
16 KB	723.53 MB/s	835.91 MB/s
32 KB	1.34 GB/s	1.61 GB/s
64 KB	2.75 GB/s	2.76 GB/s
128 KB	3.31 GB/s	3.35 GB/s
256 KB	4.41 GB/s	4.27 GB/s
512 KB	5.83 GB/s	5.83 GB/s
1 MB	8.13 GB/s	7.94 GB/s
2 MB	11.85 GB/s	11.78 GB/s
4 MB	11.81 GB/s	12 GB/s
8 MB	11.78 GB/s	12.91 GB/s
12 MB	11.73 GB/s	12.89 GB/s
16 MB	11.66 GB/s	13.09 GB/s
24 MB	11.80 GB/s	12.86 GB/s
32 MB	11.63 GB/s	12.84 GB/s
48 MB	11.75 GB/s	13.01 GB/s
64 MB	11.66 GB/s	12.97 GB/s

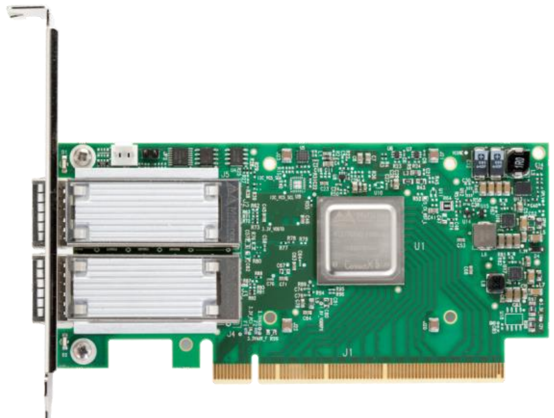
You can have all the network bandwidth in the world and even an amazing 100G ethernet card but it will all fall in a heap if your IO subsystem isn't up to the task on the localhost…

…and then there is the NIC itself. It isn’t as simple as *“all 10G cards are made equal!”*

Intel x540 t2

Intel x550 t2

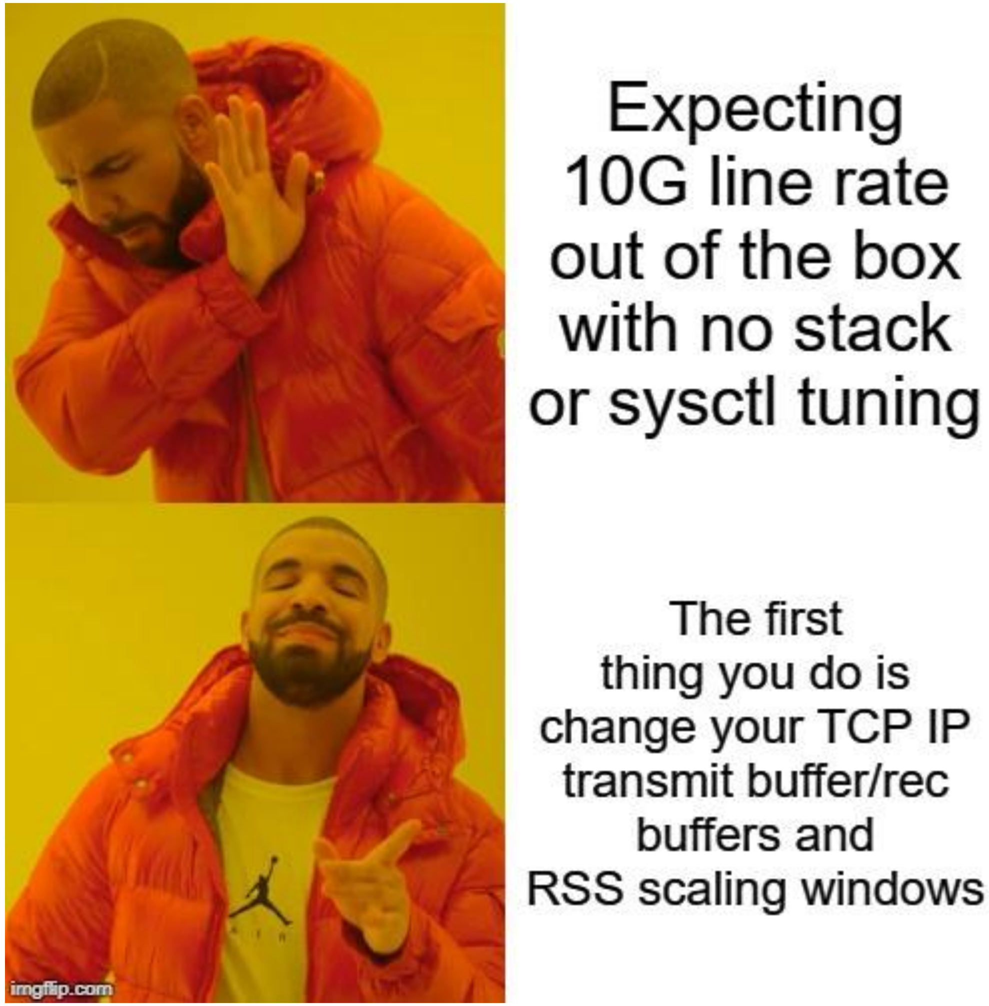
Qlogic 10GbE QP 3xxx



Mellanox 100G CX-5

10G and 100G NICs vary wildly in their capability, tunability and performance with the Windows (and Linux) network stack.

Evil tuning required

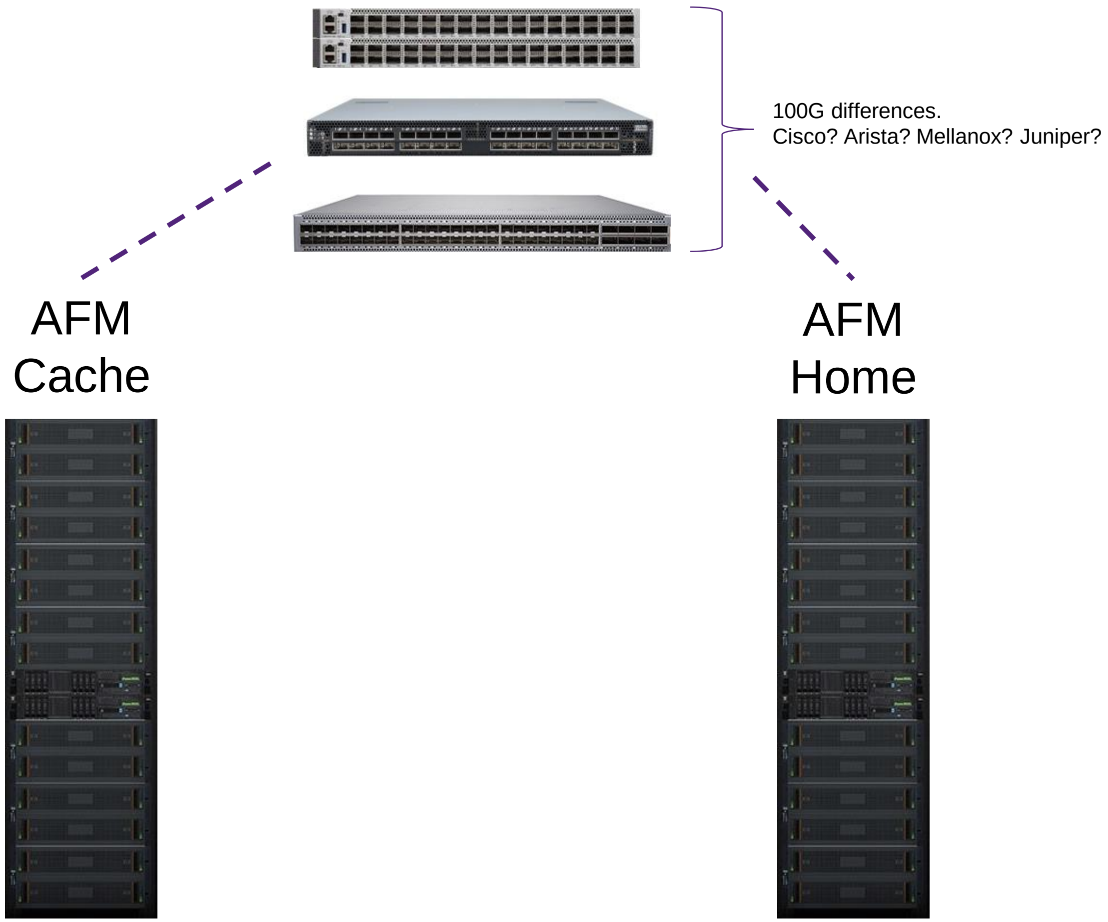


If you don’t tune your max RSS queues, TCP/IP transmit and receive buffers _AND_ make sure your network is MTU clean 9000, you’ve virtually no hope of seeing > a few hundred MB/sec with the SMB stack the way it is, much less anywhere near 100GbE class transfers.

MTU9000 “clean” across a network can be difficult as you may not have the control of your entire network.

Become good friends with your network engineering team or suffer the painful consequences of a high value, under performing network environment!

The network is on fire (always) with AFM



Spectrum Scale AFM can “fill pipes”. It is notorious (in our experience) for creating mass ethernet fabric reverberations, problems and shows problematic points in networks incredibly quickly. This includes:

- Buffer shock scenarios.
- Port buffer exhaustion.
- Backplane IO fairness contention.
- Port overcommit scenarios.
- MTU cleanliness issues.
- Bandwidth “bottom out” and traffic over-run.
- Discards. Frame drops. QoS failures.

You must consider very carefully not just the 100G era in your switching but WHAT 100G technology you use!

A missive from my network team, one morning…

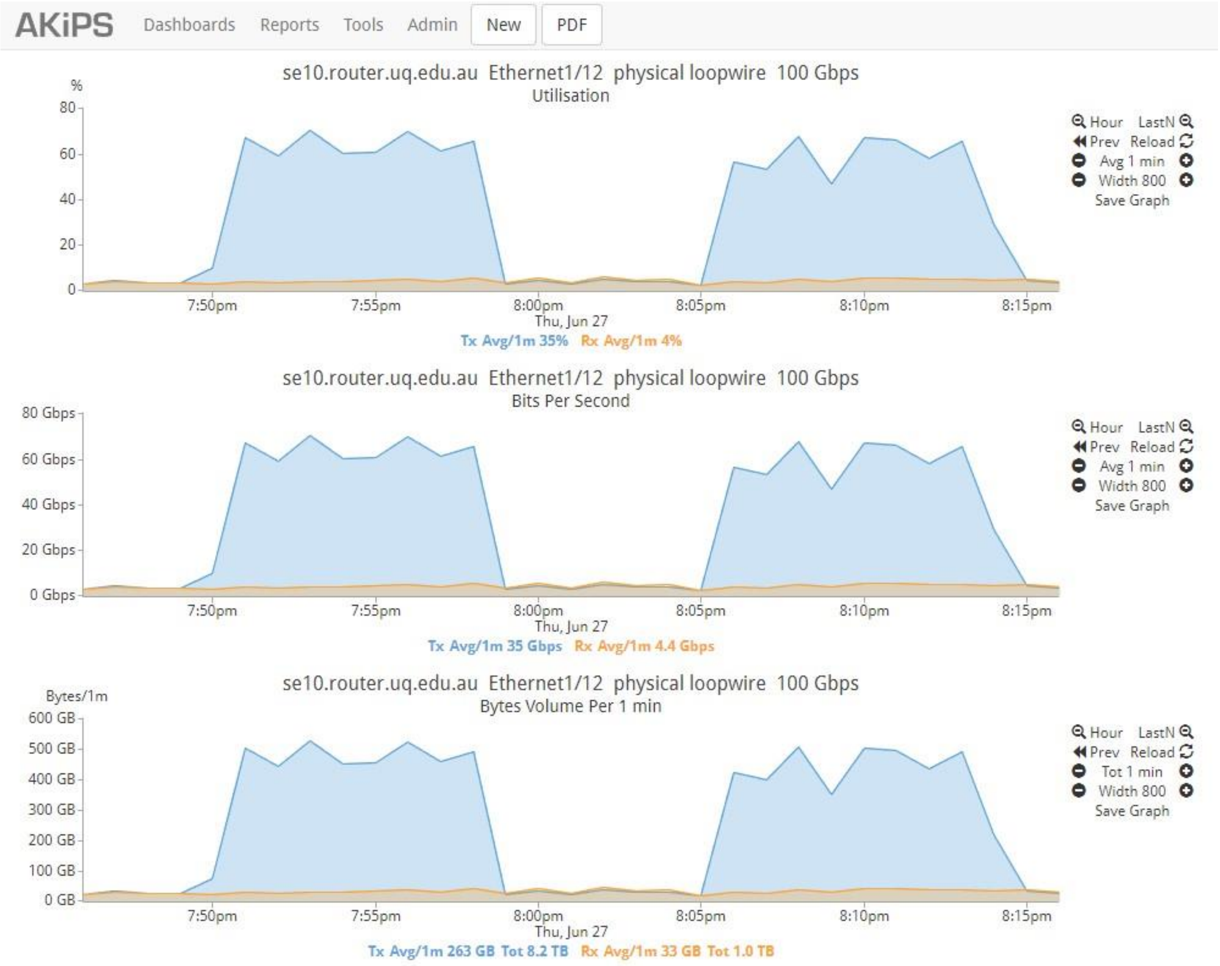
"Hey mate.

So, love your work. Heard you just turned on that big black rack with the flash and the disk all in one. Something about a next generation MeDiCI node, yeh? ~~NameRedacted~~ put the Krone cassette in the TOR for ~~NameRedacted~~ to plug in the 100G LC-LC cables out of the Mellanox SN2700's to the uplink. You'll see the grey tab Cisco optic on the other side in the chassis hanging free.

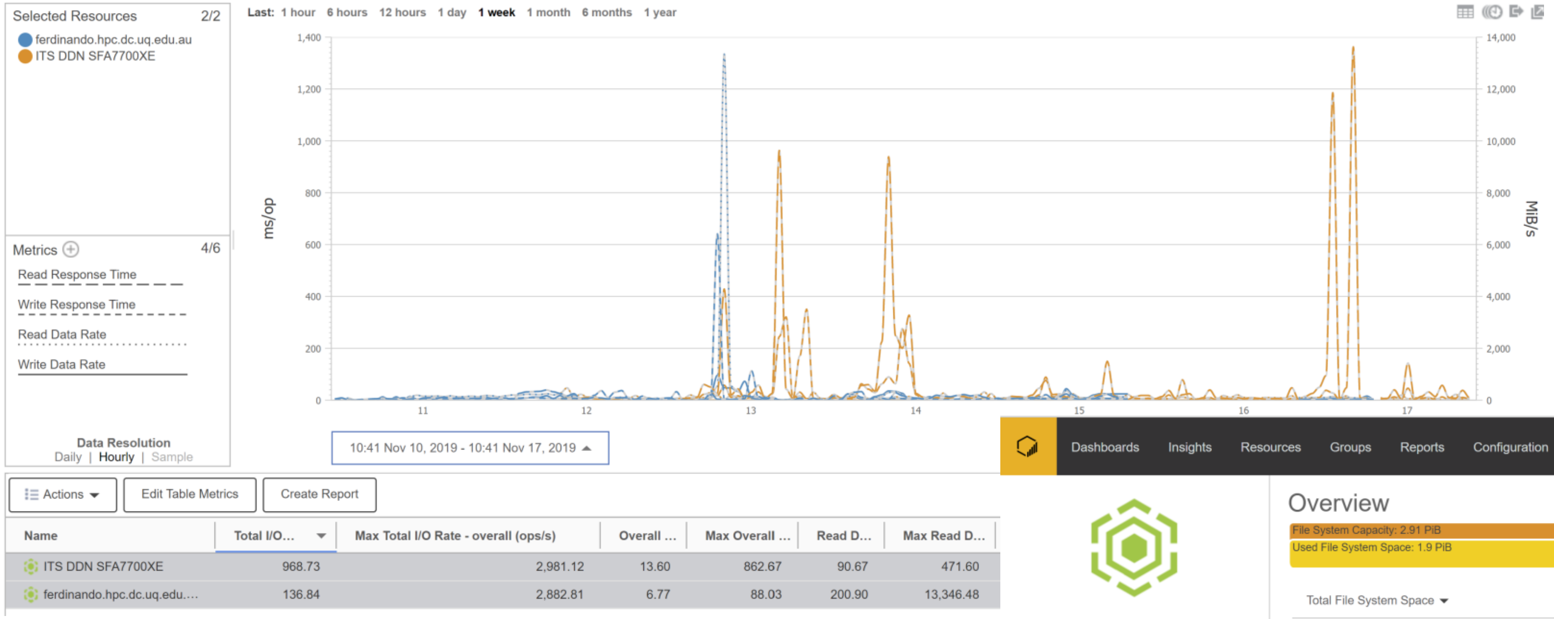
Can you do me a favour though? Whatever you did last night – can you please not do that again in business hours? You actually created a DoS-like behaviour on the entire ~~NameRedacted~~ side of the network.

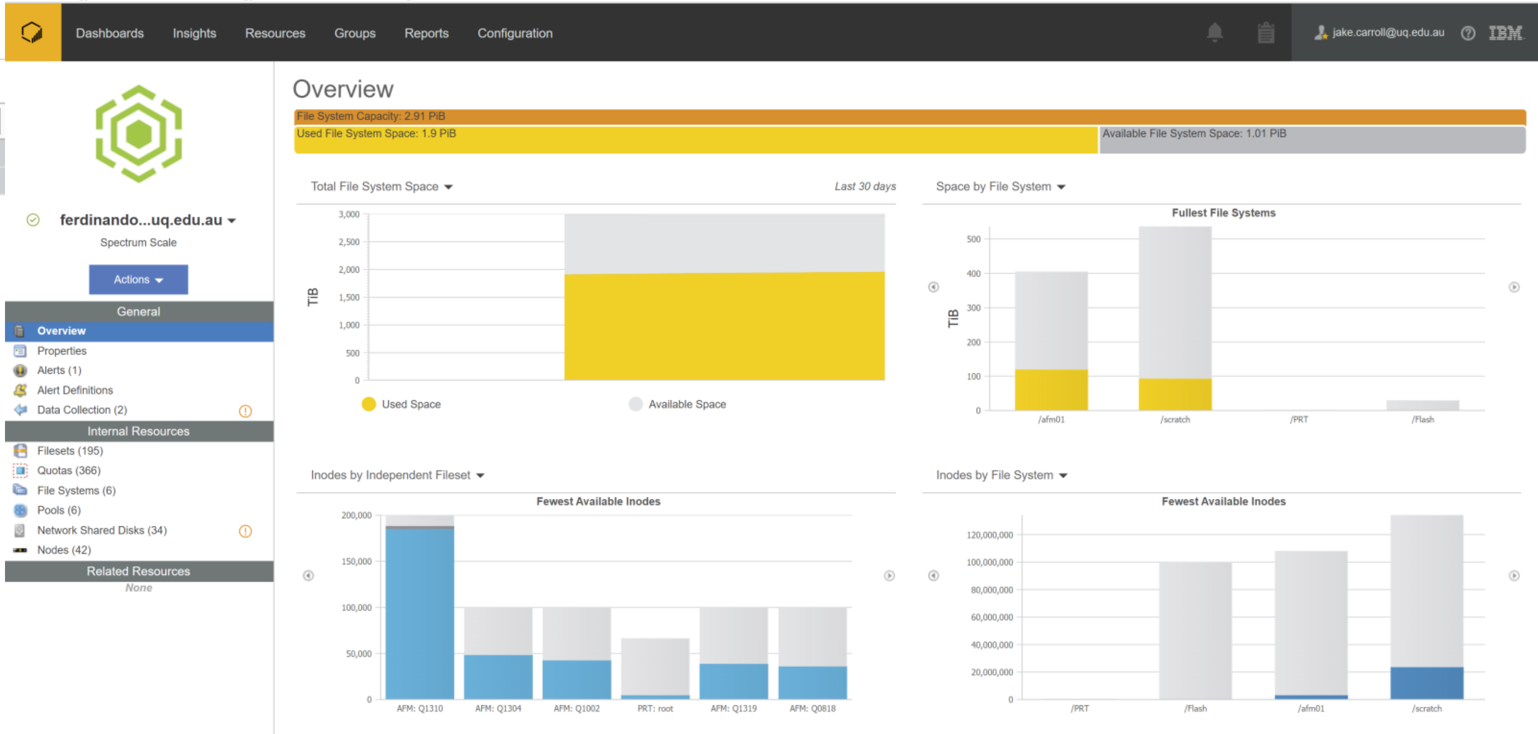
Thanks mate.

~~NameRedacted~~.



Metrics. IBM Storage Insights Pro





A bit of work to be done to make Spectrum Scale a first class citizen inside this platform. Lots of little things missing, stuff that hasn't been thought about – and things that plain don't make sense for file/SDS based storage infrastructure. Clearly had "block" in mind when the tools were built…

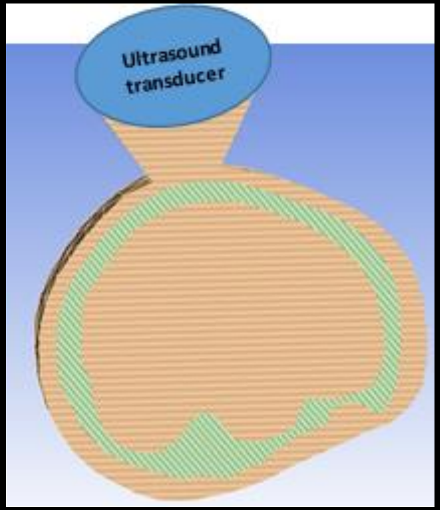
This all results in enormous trial and error required. You can't "project manage" this level of uncertainty and product variability. You must accept change and pivots, constantly.

Rockstars…

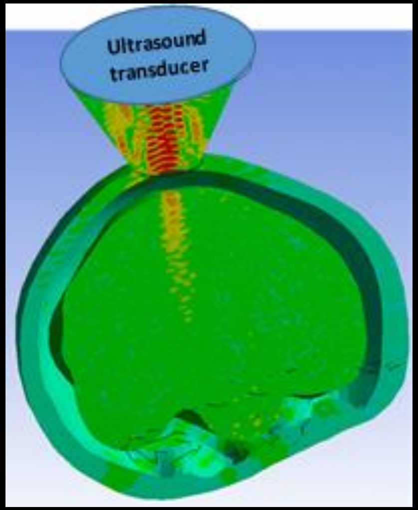
Using simulation to understand ultrasonic propagation through the skull bone...to treat Alzheimer's disease.

- **To predict the delivered ultrasound energy through the human skull.**

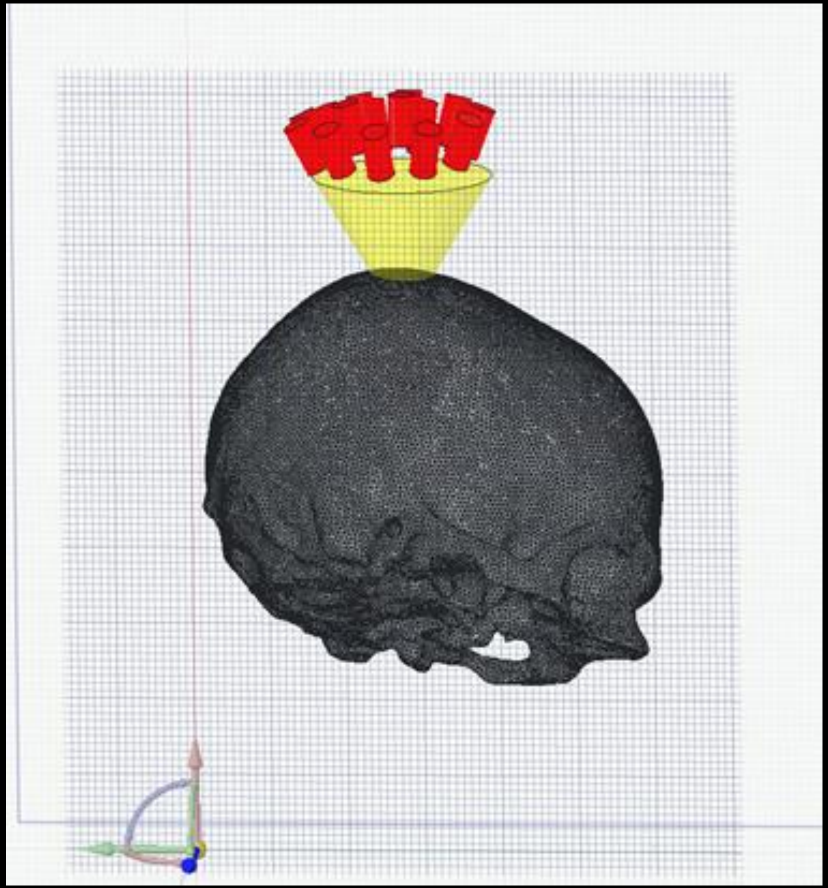
Human CT scan



Computational domain



Human skull simulation

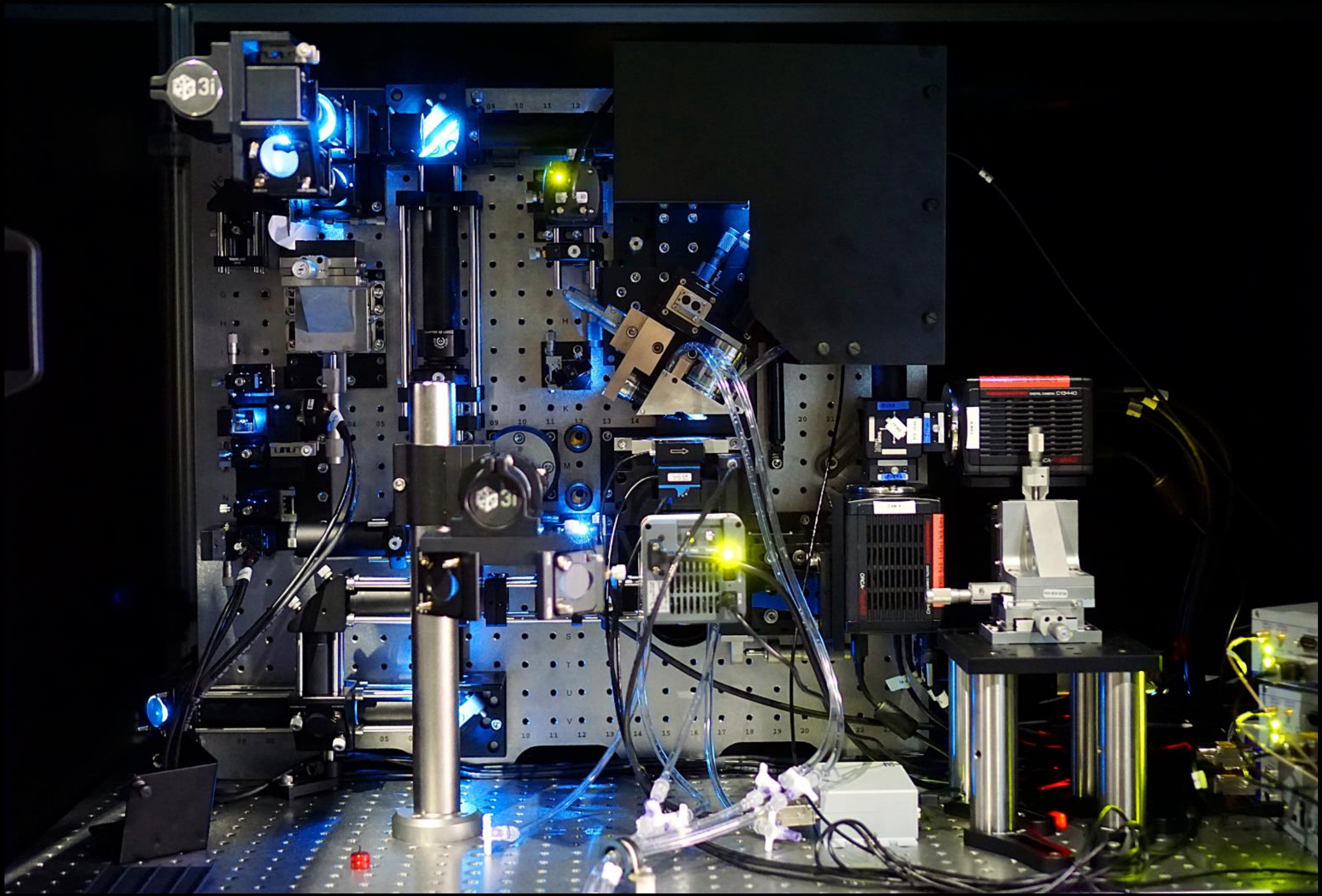


Predicted treatment envelope

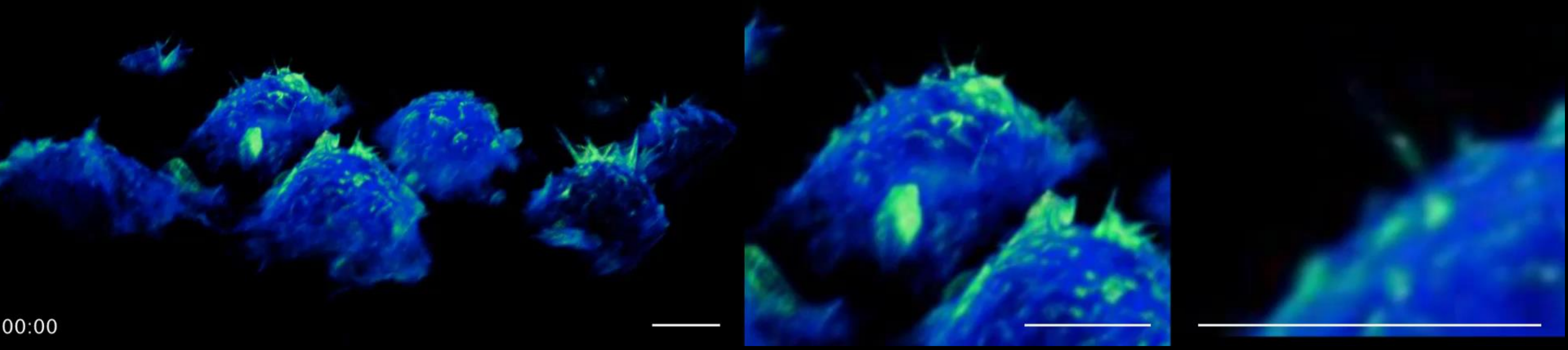
- 61 million surface-sweep required to be solved to simulate a human skull
- Using 22 nodes, 8TB of memory, 500,000 CUDA cores @ 110GB/sec.
- Solution completed in 19 hours, 7 days faster than all previous supercomputing facilities in Australia attempting solvers of this scale.

Credit: Odabaee, M. (2018)

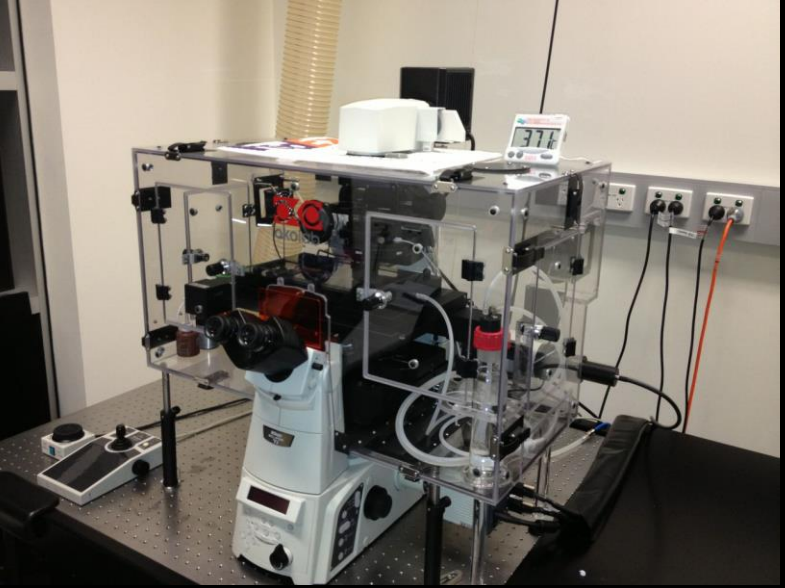
UQ IMB LLSM (Lattice Light Sheet Microscope)



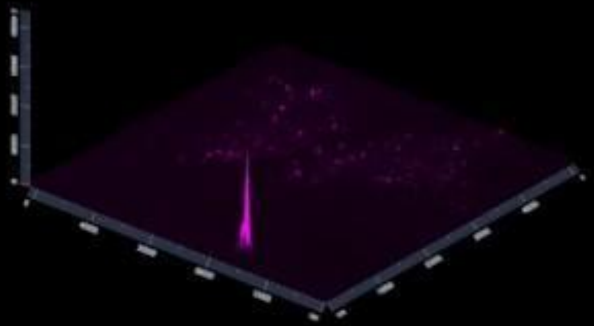
Credit: Hamilton, N. (2018)



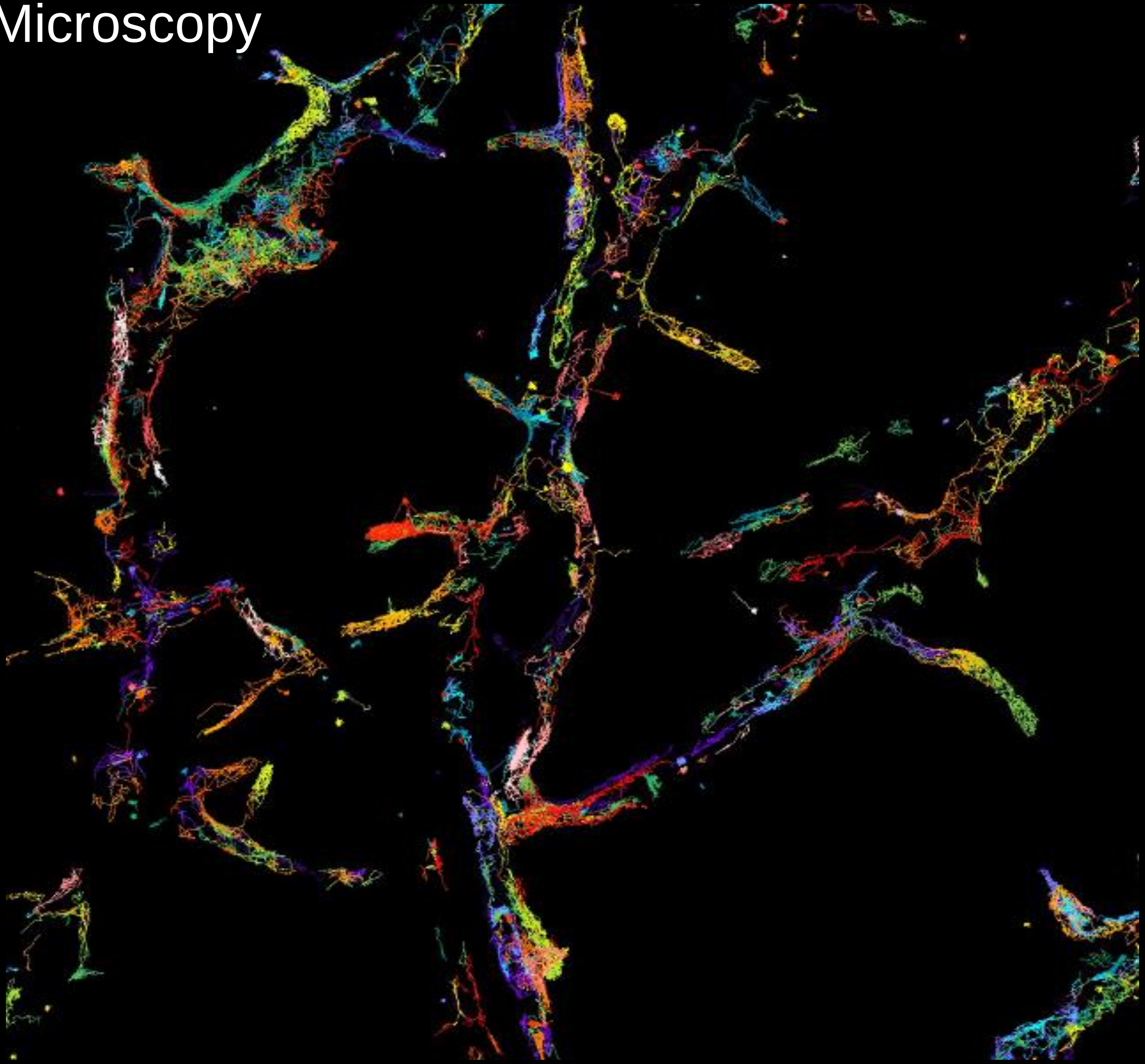
Condon, N., Heddleston, J., Chew, T., Luo, L., McPherson, P., Ioannou, M., . . . Wall, A. (2018). Macropinosome formation by tent pole ruffling in macrophages. *The Journal of Cell Biology, 217*(11), 3873-3885.



Super Resolution Imaging

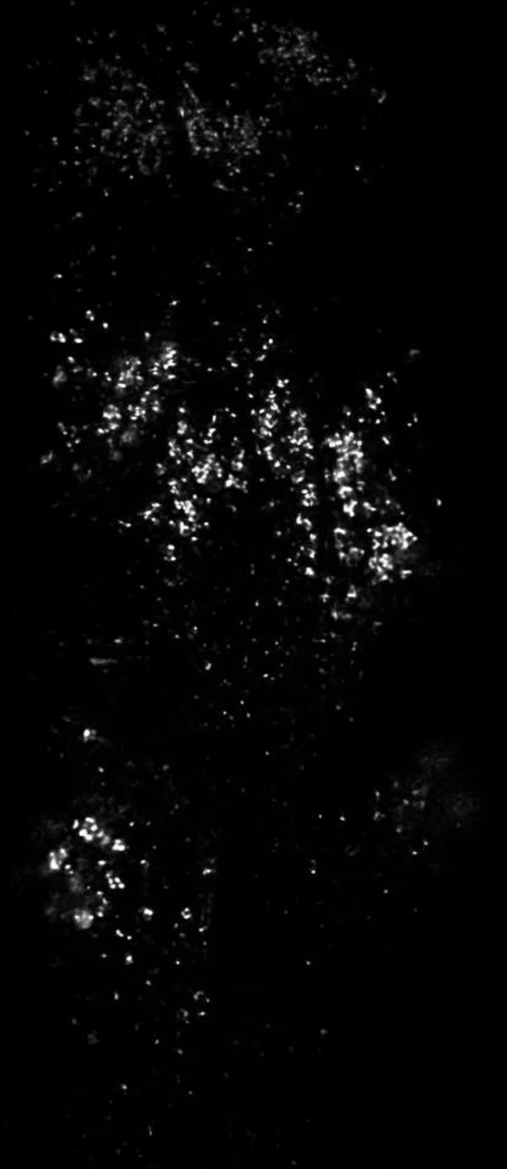


Super Resolution Microscopy



40,000 8k 32 bit images in 5 minutes

Deconvolution in action…

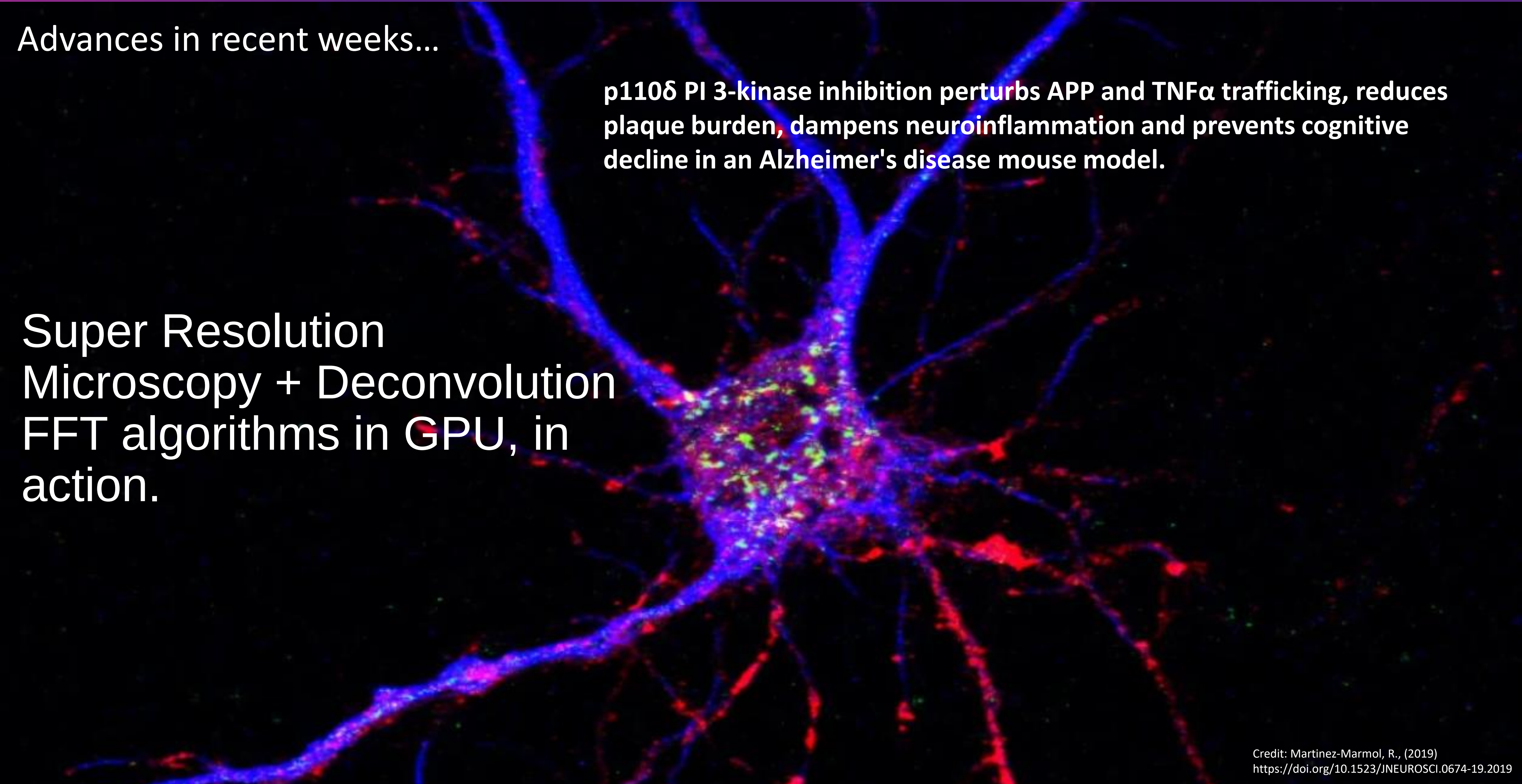


Endosomes "eating" cells – an immune system at work!

* Raw acquisition volume: **400GB**
* Acquisition time: 40 minutes
* Sustained IOPS to disk – 145,000 in acquisition, reassembly on-the-fly.
* Processing time deconv on CPU: 19 hours (2 * Xeon Skylake 6132's).
* Processing time on 6 * nVidia Volta's @ Wiener: 8 minutes.

Credit: Condon N. (2019)

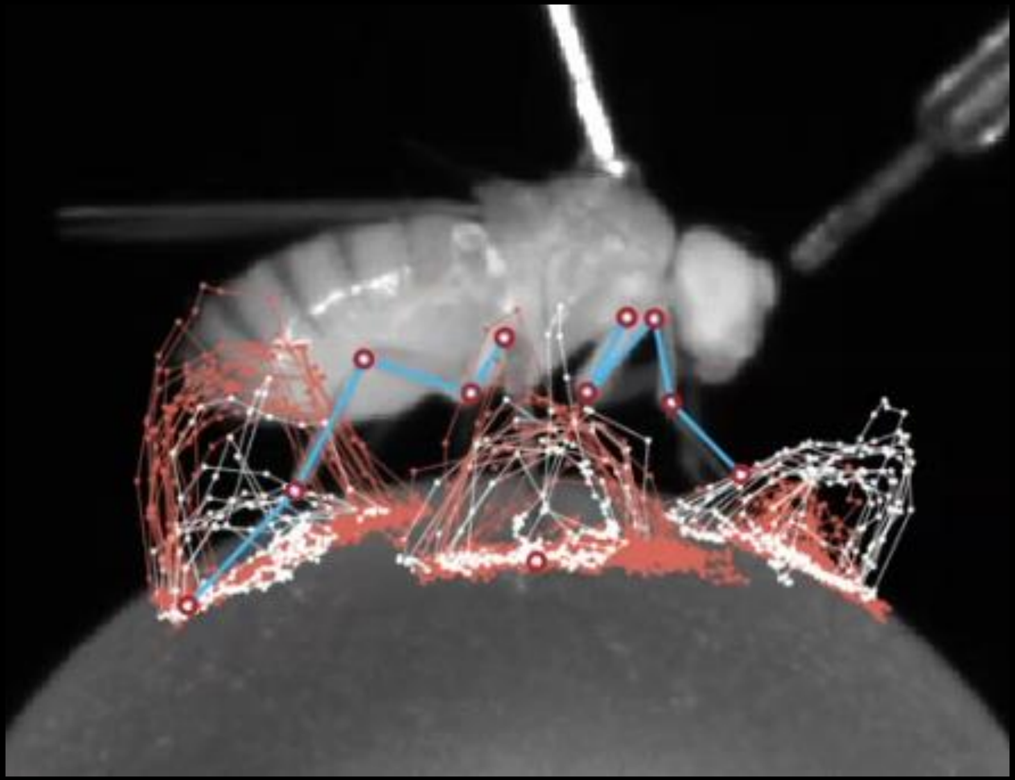
Advances in recent weeks…



p110δ PI 3-kinase inhibition perturbs APP and TNFα trafficking, reduces plaque burden, dampens neuroinflammation and prevents cognitive decline in an Alzheimer's disease mouse model.

Super Resolution Microscopy + Deconvolution FFT algorithms in GPU, in action.

Credit: Martinez-Marmol, R., (2019)
https://doi.org/10.1523/JNEUROSCI.0674-19.2019



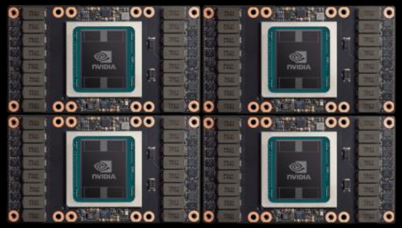
1000+ fps of trajectory data from Drosophila Fly movement decomposed and modelled using 3 * nVidia Volta GPU’s stream processing from 8 * 8k ultra high speed cameras simultaneously to SpectrumScale.

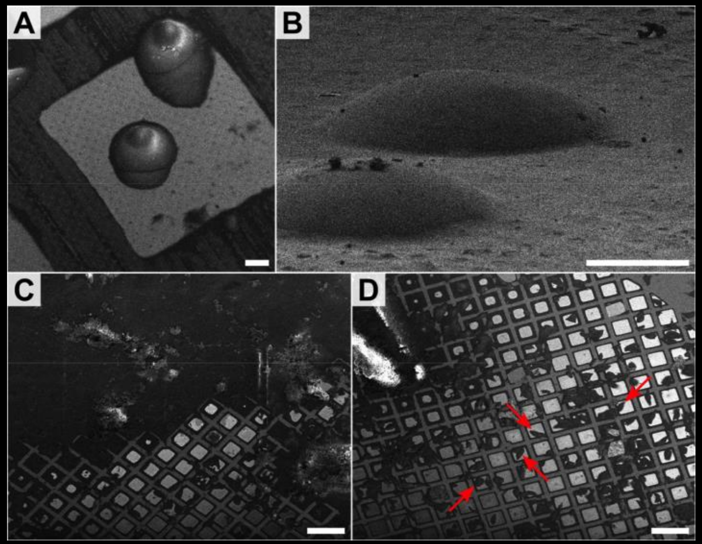
Allowing us answer questions like what exact role each neuron plays in a complex sequence of movements like backward walking or turning and decipher the general principles of how motor programs are generated by neural circuits.

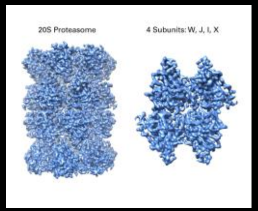
Credit Dickson B., Feng, K., (2019)

Single electron direct detection infrastructure. Gatan K3 EM sCMOS.

NSD connected single particle tracking acceleration host (quad nVidia Voltas)



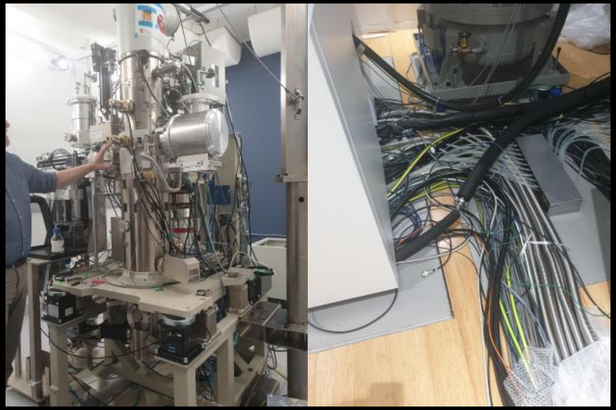




AFM Cache

5760 x 4092 (24 megapixels) per frame @ 1,500 frames per second

KriOS Cryo-EM workflow. The epitome of the data fabric in action.



4 * 100G QSFP28 eth → FPGA

NSD connected single particle tracking acceleration host (quad nVidia Voltas)

UnBur
MotionCorr2
CryoSPARC

100G NSD
Linux

AFM
Cache

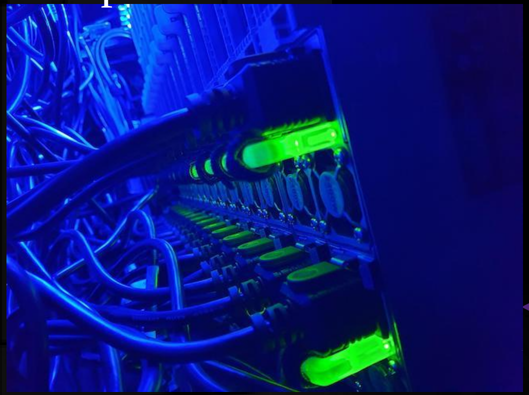
Tinaroo Intel CPU super: Particlepicker[MPI]

AFM
Home

100G NSD
Linux

FlashLite CPU super: CrYOLO

Wiener nVidia Volta GPU super: Relion 3.1 SPT



NSD POSIX Client
IB EDR/HDR

NSD AFM over ETH @ 100G

Q and A

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Thank you

Jake Carroll | Chief Technology Officer
Research Computing Centre
jake.carroll@uq.edu.au
07 3346 6407

facebook.com/uniofqld
Instagram.com/uniofqld
twitter.com/RCCUQ
facebook.com/rccuq

https://rcc.uq.edu.au

Credits and thank you to…

- Professor David Abramson, Director UQ RCC
- Rob Moffatt, CIO, UQ
- Irek Porebski, Senior Systems Engineer, UQ QBI
- Michael Mallon, Senior Cloud Systems Engineer, RCC, UQ
- Leslie Elliott, Research Infrastructure Manager, ITS, UQ
- Stephen Bird, Service Dev Manager, QCIF.
- Andrew Beattie, IBM, AU.
- Ulf Troppens, IBM, DE.
- Venkat Puuvada, IBM, IN.