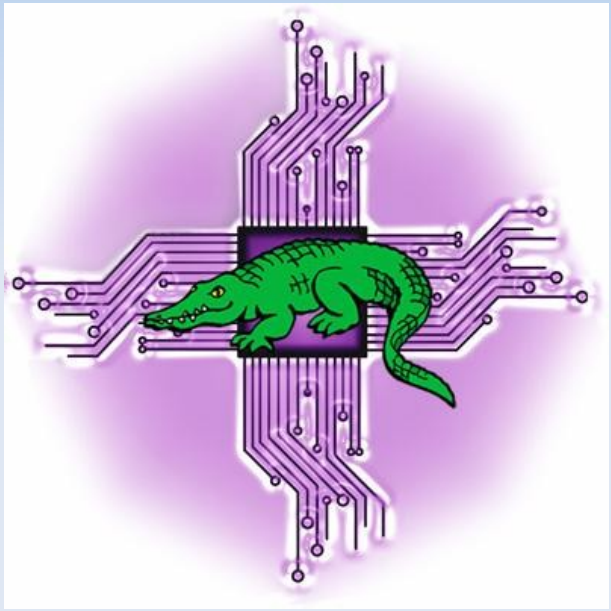
The Spartan HPC Story: From Small Scale Experimental to Top500 and Beyond



eResearch Australasia, October, 2024

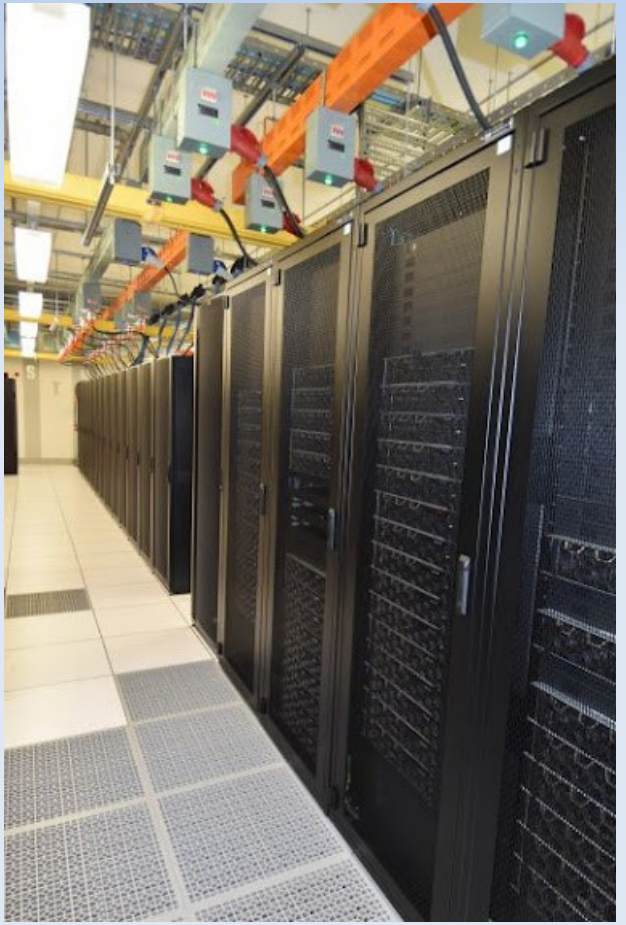
Lev Lafayette

lev.lafayette@unimelb.edu.au

HPC Services Team Leader

Previous Presentations

- **There have been many presentations and papers about the Spartan HPC since 2016.**
- **Presentations include: Multicore World, New Zealand, 2016, eResearch Australasia, (2016, 2020, 2021), Several European HPC centres (2016), Linux Users of Victoria (2016), OpenStack Summit (2016), HPC Advisory Council (2017, 2018), eResearch NZ (2021)**
- **Chapters, papers, etc include: The Crossroads of Cloud and HPC (2016), IEEE International Conference on e-Science (2017), Advances in Science, Technology and Engineering Systems Journal (2019), Cybersecurity and High Performance Computing Environments (2022).**
- **Spartan also has an very extensive HPC training and workshop programme and is part of the international HPC Certification Forum, which has resulted in several papers and presentations as well: eResearch Australasia (2017, 2020), Australia National Data Service workshop (2018), International Supercomputing Conference (2018), Journal of Computer Science Education (2019, 2020), HPC-AI Advisory Council (2019), eResearchNZ (2022), Supercomputing Asia (2024)**



Conception and Birth

- **Prior to Spartan, UniMelb's general purpose HPC was "Edward" (2011-2016) and the one prior to that was "Alfred". These followed the typical architecture for a small cluster (login node, management node, compute nodes, storage servers, NFS and fast interconnect.**
- **Actual analysis of job metrics revealed that 76% of tasks were single *core* and low memory. Limited budgetary resources (the laconic pun "spartan") meant the new system had a small traditional HPC layout ("physical partition", 200 cores) and and with virtual machines making about the majority of compute resources ("cloud partition", 3000 cores) from NeCTAR; cloudbursting with Slurm was also implemented.**
- **Other innovative technologies include RoCE for physical partition interconnect (outperforming another Melbourne-based cluster that used FDR14 Infiniband), CephFS as the file system, and EasyBuild and Lmod for software builds. Spartan was a high-throughput system and a cluster/cloud hybrid.**
- **Edward completed 375,000 jobs in 2015, Spartan complete over a million in its first year.**
- **Big promotional tour of New Zealand an Australian conferences, European HPC centres, OpenStack summit Barcelona in 2016.**

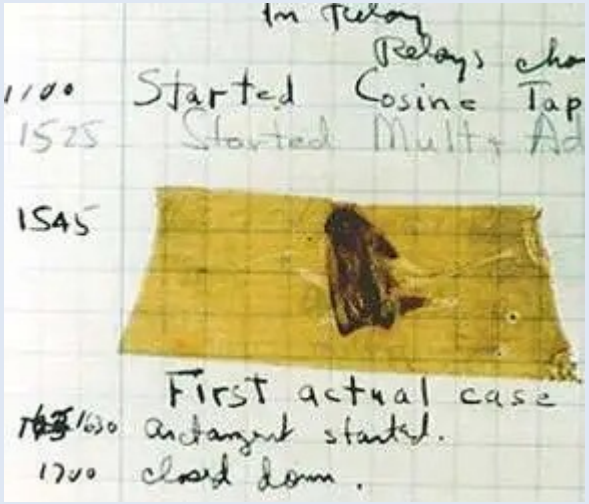
Spartan Launch, 2016

The GPU Partition and Tier One

- Spartan expanded with additional compute nodes from specialist projects, departments, and research agencies etc. By 2018 there was a big expansion with of 68 nodes and 272 nVidia P100 GPGPU cards, funded by a LIEF grant LE170100200 from the Australian Research Council with a consortium of Victorian Universities.

- Suddenly Spartan transformed from a successful innovative and experimental system to tier one system for Australia and probably with sufficient performance to be counted in the Top 500 (we estimated around #200).

- In the course of these early days some discoveries were made: (a) use of VMs had to be deployed on a one-to-one basis; the use of over-commit, typical in most cloud scenarios, could lead to time mismatch errors (b) the use of cloud VMs and physical machines in a hybrid manner was challenging to CephFS and (c) cloudbursting was not really necessary and the main use case would have been any SLAs; plus there was a cloudbursting bug in Slurm v16.



Visualisation & Reconfiguration

- **As Spartan grew, we introduced interactive visualisation nodes, initially with FastX and Open OnDemand. Notably, Conda would often caused a conflict with the former. An interesting and ongoing development is with the introduction of high-volume data for cryo-electron microscopy processing (2 from 138 instruments in the PetaScale Initiative but 2+ terabytes per day, c50% of the instrumentation total).**
- **An evaluation of the extensive training programme was conducted in 2022; typically we conduct between 25-30 training workshops per annum. Surprisingly, it turned out that at least 50% of job submissions were conducted by users a year after they had received training.**
- **A very substantial change occurred in 2023 when we decided it was time to do a major update of the RHEL operating system from a 7.x (which we first installed in 2015) to 9.x. This required re-installing most of the applications, but we also kept the RHEL7 installs as an Apptainer container for reproducibility (without multinode access).**
- **The upgrade also finally provided us an opportunity to partially test Spartan against the Top500; on the GPU partitions alone it was ranked #454 in November 2023. Spartan finally received its laurels!**
- **This year we've also added new a GPGPU partition with H100 ("Hopper") Tensor Core GPUs, plus additional compute nodes. Our Slurm workflow manager ran out of job IDs and flipped back to job 1 ! :)**

Spartan’s Top 500 Certificate

TOP 500 CERTIFICATE

The List.

Spartan GPU Partitions - PowerEdge R750XA, Xeon Gold 6326 16C 2.9GHz, NVIDIA Tesla A100 80G, 100G Ethernet

The University of Melbourne, Australia

is ranked

No. 454

among the World's TOP500 Supercomputers

with 2.14 PFlop/s Linpack Performance

in the 62nd TOP500 List published at the SC23

Conference on November 14, 2023.

Congratulations from the TOP500 Editors

Erich Strohmaier
NERSC/Berkeley Lab

Jack Dongarra
University of Tennessee

Horst Simon
NERSC/Berkeley Lab

Martin Meuer
Prometeus

The Future of Spartan

- **Compute nodes are already being upgraded to 400Gbps switching. Research network upgrade from 100Gbps to 400/800Gbps imminent, upgrading the HPC spine to 400Gbps capability next year.**
- **In the very near future water cooling is expected to be required within the next couple of years for some high-performance components, and an “all flash” conversion for storage to replace some or all spinning discs.**
- **More intensive HPC configuration and update automation to reduce maintenance windows and effort. Introduction of more MFA requirements, encryption, introducing our own CA and certificate service.**
- **Shorter funding cycles to match technology changes. More modular purchases.**
- **Improved retention of technical staff; further developments in depth and scope of user training and support.**


THANKS FOR WATCHING
& LISTENING PATIENTLY
makeameme.org