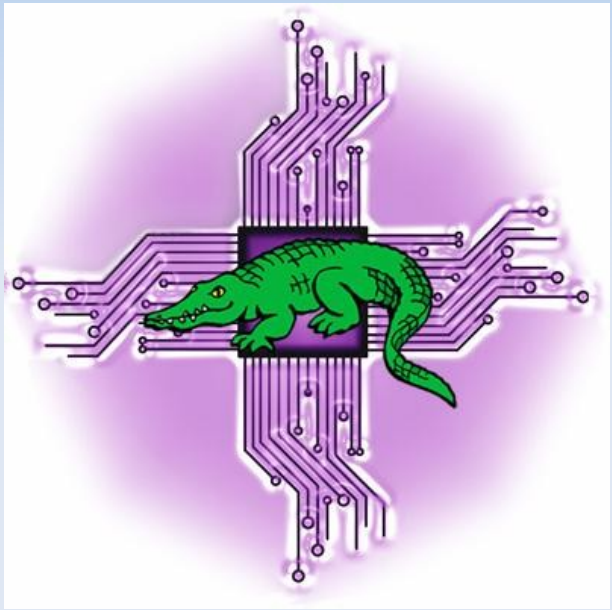
The Why and How of HPC-Cloud Hybrids with OpenStack



OpenStack Australia Day Melbourne
June 1 , 2017 and Telstra Melbourne
Brown Bag, June 5, 2017

lev@levlafayette.com

1.0 Management Layer

1.1 HPC for performance.

High-performance computing (HPC) is any computer system whose architecture allows for above average performance. Main use case refers to compute clusters with a teamster separation between head node and workers nodes and a high-speed interconnect acts as a single system.

1.2 Clouds for flexibility.

Precursor with virtualised hardware. Cloud VMs *always* have lower performance than HPC.

1.3 Hybrid HPC/Clouds.

University of Melbourne model, "the chimera". Cloud VMs deployed as HPC nodes and the Freiburg University Model, "the cyborg", HPC nodes deploying Cloud VMs.

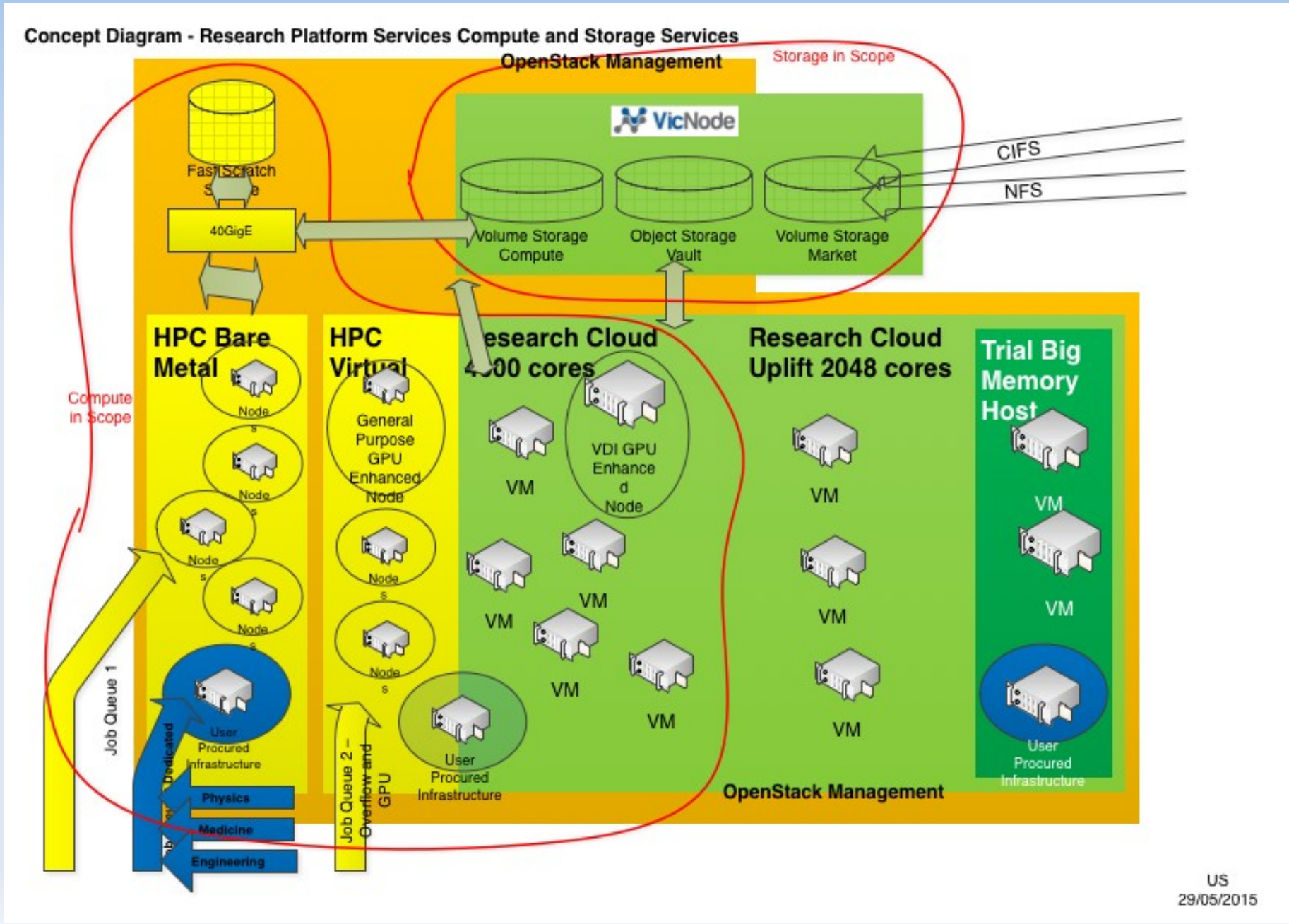
1.4 Reviewing user preferences and usage.

Users always want more of 'x'; real issue identified was queue times. Usage indicated a high proportion of single-node jobs.

1.5 Review and Architecture.

Review discussed whether UoM needed HPC; architecture was to use existing NeCTAR Research cloud with an expansion of general cloud compute provisioning and use of a smaller "true HPC" system on bare metal nodes.

1.0 Management Layer



2.0 Physical Layer

2.1 Physical Partitions.

"Real" HPC is a mere c276 cores, 21 GB per core. 2 socket Intel E5-2643 v3 E5-2643, 3.4GHz CPU with 6-core per socket, 192GB memory, 2x 1.2TB SAS drives, 2x 40GbE network. “Cloud” partitions is almost 400 virtual machines with over 3,000 2.3GHz Haswell cores with 8GB per core and . There is also a GPU partition with Dual Nvidia Tesla K80s (big expansion this year), and departmental partitions (water and ashley). Management and login nodes are VMs as is I/O for transferring data.

2.2 Network.

System network includes: Cloud nodes Cisco Nexus 10Gbe TCP/IP 60 usec latency (mpi-pingpong); Bare Metal Mellanox 2100 Cumulos Linux 40Gbe TCP/IP 6.85 usec latency and then RDMA Ethernet 1.15 usec latency. Later was superior to Infiniband FD14 control (1.17 usec).

2.3 Storage.

Mountpoints to home, projects (/project /home for user data & scripts, NetApp SAS aggregate 70TB usable) and applications directories across all nodes. Additional mountpoins to VicNode Aspera Shares. Applications directory currently on management node, needs to be decoupled. Bare metal nodes have /scratch shared storage for MPI jobs (Dell R730 with 14 x 800GB mixed use SSDs providing 8TB of usable storage, NFS over RDMA)., /var/local/tmp for single node jobs, pcie SSD 1.6TB.

3.0 Operating System and Scheduler Layer

3.1 Red Hat Linux.

Scalable FOSS operating system, high perfomance, very well suited for research application. In November 2016 of the Top 500 Supercomputers worldwide, every single machine used a "UNIX-like" operating system; and 99.6% used Linux.

3.2 Slurm Workload Manager.

Job schedulers and resource managers allow for unattended background tasks expressed as batch jobs among the available resources; allows multicore, multinode, arrays, dependencies, and interactive submissions. The scheduler provides for paramterisation of computer resources, an automatic submission of execution tasks, and a notification system for incidents.

Slurm (originally Simple Linux Utility for Resource Management), developed by Lawrence Livermore et al., is FOSS, and used by majority of world's top systems. Scalable, offers many optional plugins, power-saving features, accounting features, etc. Divided into logical partitions which correlate with hardware partitions.

3.3 Git, Gerrit, and Puppet.

Version control, paired systems administration, configuration management.

3.4 OpenStack Node Deployment.

Significant use of Nova (compute) service for provisioning and decommissioning of virtual machines on demand.

4.0 Application Layer

4.1 Source Code and EasyBuild.

Source code provides better control over security updates, integration, development, and much better performance. Absolutely essential for reproducibility in research environment. EasyBuild makes source software installs easier with scripts containing specified compilation blocks (e.g., configuremake, cmake etc) and specified toolchains (GCC, Intel etc) and environment modules (LMod). Modulefiles allow for dynamic changes to a user's environment and ease with multiple versions of software applications on a system.

4.2 Compilers, Scripting Languages, and Applications.

Usual range of suspects; Intel and GCC, for compilers (and a little bit of PGI), Python Ruby,

and Perl for scripting languages, OpenMPI wrappers. Major applications include: MATLAB,

Gaussian, NAMD, R, OpenFOAM, Octave etc.

Almost 1,000 applications/versions installed from source, plus packages.

4.3 Containers with Singularity.

A container in a cloud virtual machine on an HPC! Wait, what?

5.0 User Layer

5.1 Karaage.

Spartan uses its own LDAP authentication that is tied to the university Security Assertion Markup Language (SAML). Users on Spartan must belong to a project. Projects must be led by a University of Melbourne researcher (the "Principal Investigator") and are subject to approval by the Head of Research Compute Services. Participants in a project can be researchers or research support staff from anywhere. Karaage is Django-based application for user, project, and cluster reporting and management.

5.2 Freshdesk.

OMG Users!

5.3 Online Instructions and Training.

Many users (even post-doctoral researchers) require basic training in Linux command line, a requisite skill for HPC use. Extensive training programme for researchers available using andragogical methods, including day-long courses in “Introduction to Linux and HPC Using Spartan”, “Linux Shell Scripting for High Performance Computing”, and “Parallel Programming On Spartan”. Documentation online (Github, Website, and man pages) and plenty of Slurm examples on system.

6.0 Future Development

6.1 Cloudbursting with Azure.

Slurm allows cloudbursting via the powersave feature; successfully experiments (and bug discovery) within the NeCTAR research cloud.

Recent addition of Azure through same login node. Does not mount applications directory; wrap necessary data for transfer in script.

6.2 GPU Expansion

6.3 Test cluster (Thespian).

"Everyone has a test environment, some people also have a production and a test environment".

Test nodes already exist for Cloud and Physical partions. Replicate management and login nodes.

6.3 New Architectures

New architectures can be added to the system with separate build node (another VM) and with

software built for that architecture. Don't need an entirely new system.


THANKS FOR WATCHING
& LISTENING PATIENTLY
makeameme.org