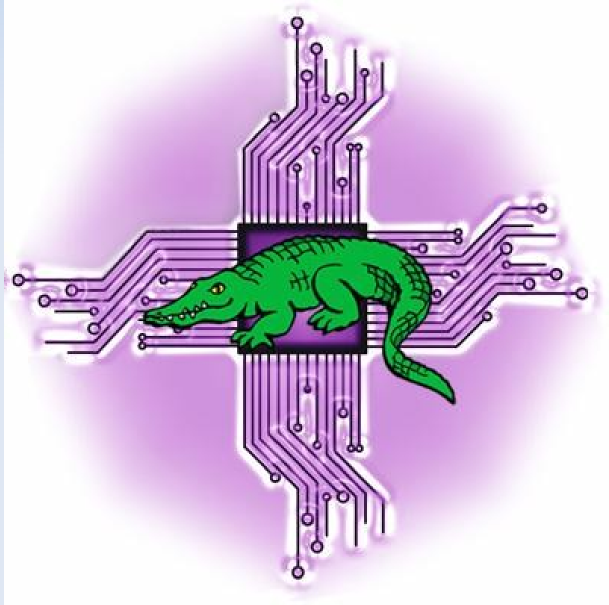
A Container On a Virtual Machine On an HPC?

Presentation to HPC Advisory Council



Perth, July 31-Aug 01, 2017

http://levlafayette.com

Necessary and Sufficient Definitions

- High Performance Computing: High Performance Computer (HPC) is any computer system whose architecture allows for above average performance. Typically implemented as a cluster, where two or more computers serve a single resource.

- Virtual Machine: An emulation of a computer system based on a system architectures and provides the same functionality as a physical computer. An emulation machine's performance must be less than the physical architecture that it is based on.

- Container: A form of virtualisation where multiple user-space environments are isolated and run within a single instance. From a user perspective, it provides much of the functionality of a virtual machine without the overhead of initialisation etc.

Advantages and Problems With Traditional HPC Environment

- High Performance Computing systems are the 'big iron' of the computational world. No other model provides such a single point for tightly coupled performance. For a genuinely high-performance computing system several components are required: (a) server-grade systems., (b) very high speed interconnect., (c) scalable and efficient operating system, (d) scheduling and resource management system, (e) source-code builds with optimisation.

- However HPC systems have cost issues with systems and especially interconnect. Conflicts in software build environments are not easy to resolve. Conflicts in resource allocation both in terms of management decisions for purchases, and queueing resources. Lack of training for researchers and difficulty in replication of compute outcomes.

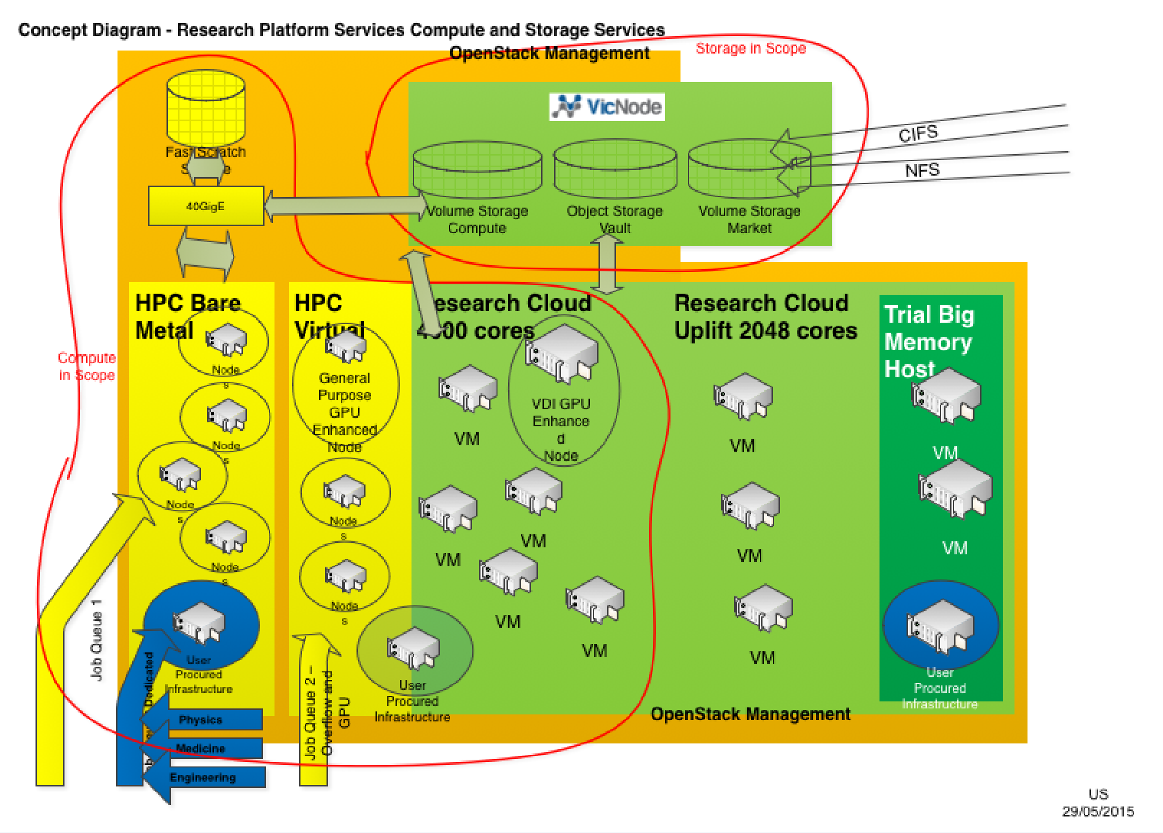
The Spartan Approach

- University of Melbourne had a traditional HPC system in operation from 2011 to 2016. Review of existing HPC facilities conducted in 2014-2015. High speed interconnect only used in a minority of jobs. User surveys indicated usual performance-enhancing hardware requests. Very minimal training program.

- Decision to incorporate existing under-utilised resources from NeCTAR OpenStack research cloud as needed. Smaller "bare metal" partition with high speed network (RoCE; Remote Direct Memory Access over Converged Ethernet). Login and management nodes are cloud VMs.

The Spartan Approach

- Core partitions separated into "physical" and "cloud", for multinode and single-node jobs; other partitions added for specialist projects and departments. Slurm for workload manager. Software installations via EasyBuild. Configuration management with Puppet, Git, and Gerrit ("paired systems administration"). Extensive training programme and specialist workshops.



Current Specifications and Future Development

- Overall system has been quite small; "physical" partition is 276 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 approximately 400 virtual machines with over 3,000 2.3GHz Haswell cores with 8GB per core. Over one million jobs run, 613 users, 303 projects.

- Also a small GPU partition; *massive* increase this coming year. Spartan will be developing from an small, innovative system acting as a stepping stone to peak national facilities (the latter role it will still perform), to a Top500 system operating at over 1000 teraflops.

Virtual Machines on an HPC

- An advantage of having cloud infrastructure was the flexibility to scale or reduce according to demand, and even to architecture subject to physical constraints and overhead. An early discovery was that virtual CPUs were not an option for computationally intensive tasks.

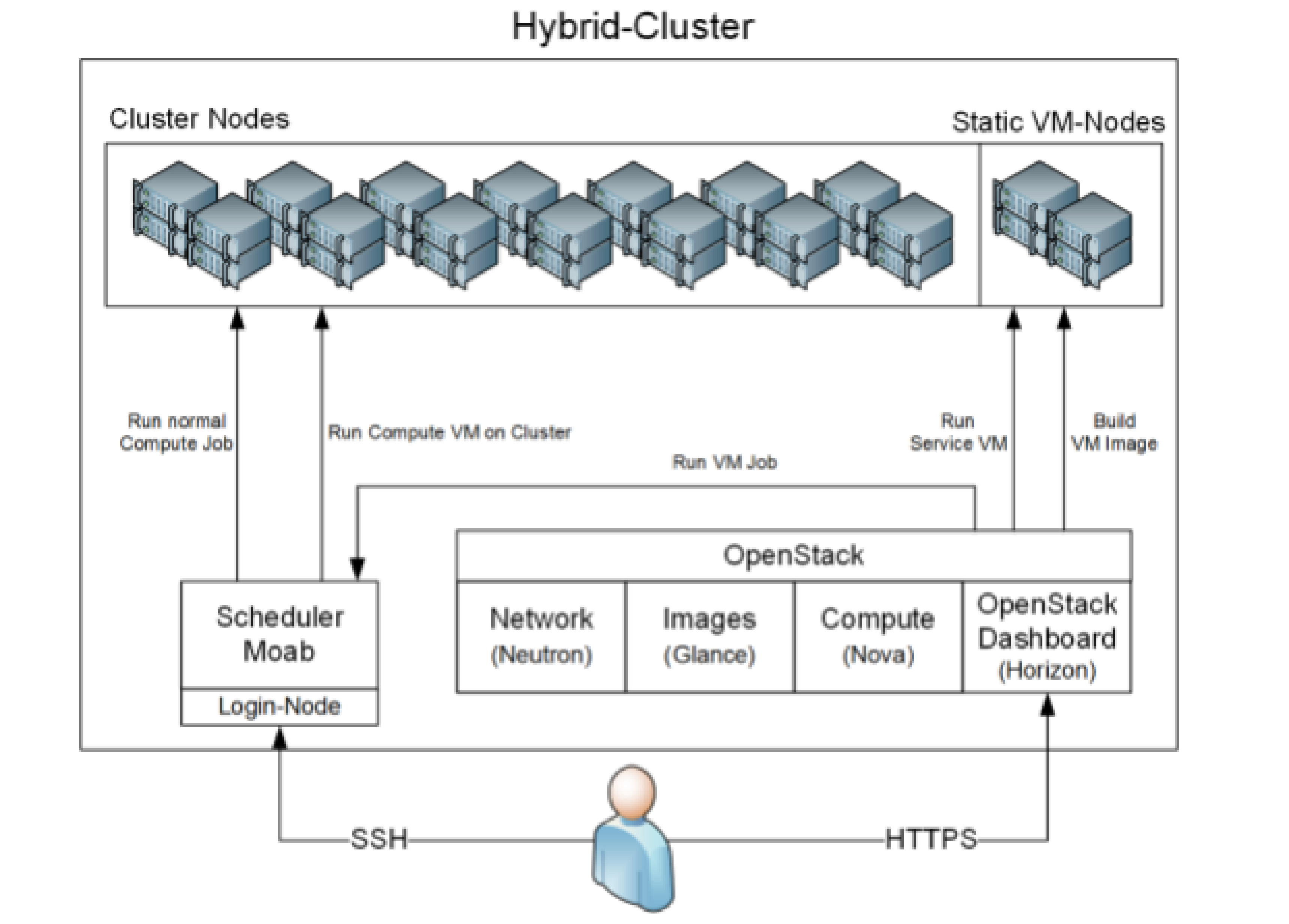
- Unlike usual cloud deployments, virtual machines on Spartan do not allow user determined images or superuser access. A single 'golden image' is used a the base for all cloud compute nodes using Nova service for deployment; for nearly all intents and purpose it is just like a physical compute node.

Virtual Machines on an HPC II

- An alternative model has been carried out by the University of Freiburg with their NEMO system. In that system unused HPC compute nodes are made available for cloud virtual machines according to user specifications. Multiple implementations are offered.

- A cloud VM via OpenStack can be deployed directly on an available compute node in a static VM queue through the Horizon dashboard; *or* the Horizon dashboard could submit a job to the Moab HPC scheduler when then launches a VM on a cluster node; *or* the user could submit a job Moab to create a VM on a cluster node.

Virtual Machines on an HPC III



Cloudbursting

- On Spartan local cloudbursting was tested with a Slurm feature (and a bug discovered, since fixed).

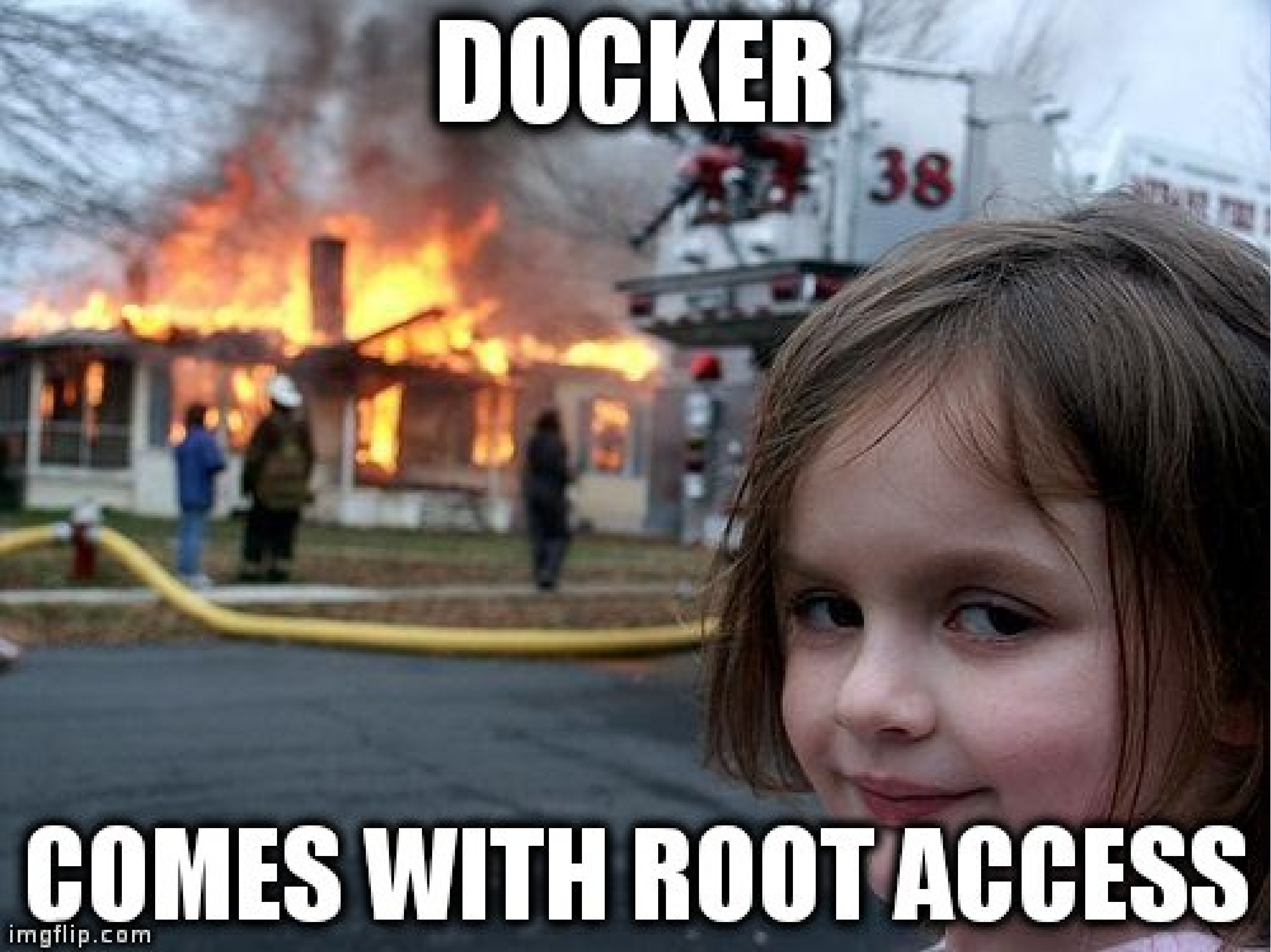
- Cloudbursting to external provides suffers from mount and latency issues. Only plausible solution is selective partial installations of software stacks with wrapper scripts for remote logins and copying data.

Containers on a Virtual Machine on a HPC

- “Look mum, we've made a human pyramid!”

- Spartan uses Singularity for containers, very effective in an research computing environment for single, complex, workflows (as opposed to the Docker orientation towards devops and microservices). Singularity can import and convert Docker images. Very good support for MPI, GPU, etc. See also NERSC's Shifter as an alternative.

- Prevents security escalation because a user inside a Singularity container is the same user as outside the container and has the same privileges. Recent security exploit (Stack Clash) affected setuid binaries; containers from Shifter and Singularity were protected; Docker was *not* protected.



DOCKER
38
COMES WITH ROOT ACCESS
imgflip.com

Future Architectures; Throughput and Diversity

- High Performance Computing is an absolute necessity for research as datasets increasing faster than computational capacity, demand will increase. However software applications and computational workflows have also become more diverse.

- Performance is *very* important. But even more important is *throughput*. Increased throughput, in some circumstances, is achieved by adopting architectures which are less than optimal in *abstract performance* but rather concentrate in *concrete performance*..

Acknowledgements and References

Thanks to the HPC Advisory Council and the Pawsey Supercomputing Centre for organising the conference. Thanks to the University of Melbourne and Bernard Meade for supporting my attendance here. Thanks to David Perry and Chris Samuel for technical advice relevant to this presentation.

Apon A., Ahalt S., Dantuluri V, et. al., "High Performance Computing Instrumentation and Research Productivity in U.S. Universities", Journal of Information Technology Impact, Vol 10, No 2, p87-98, 2010

Kurtzer, G. M., "Singularity: Containers for Science, Reproducibility, and High Performance Computing", HPC Advisory Council Standford University Conference, February 2017

Wilson, G., "High-Performance Computing Considered Harmful", 22nd International Symposium on High Performance Computing Systems and Applications, 2008

THANKS FOR WATCHING
& LISTENING PATIENTLY
makeameme.org