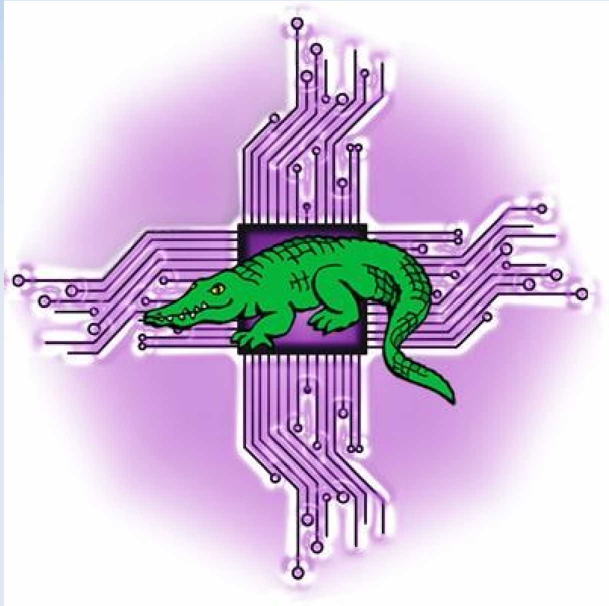
HPC/Cloud Hybrids for Efficient Resource Allocation and Throughput

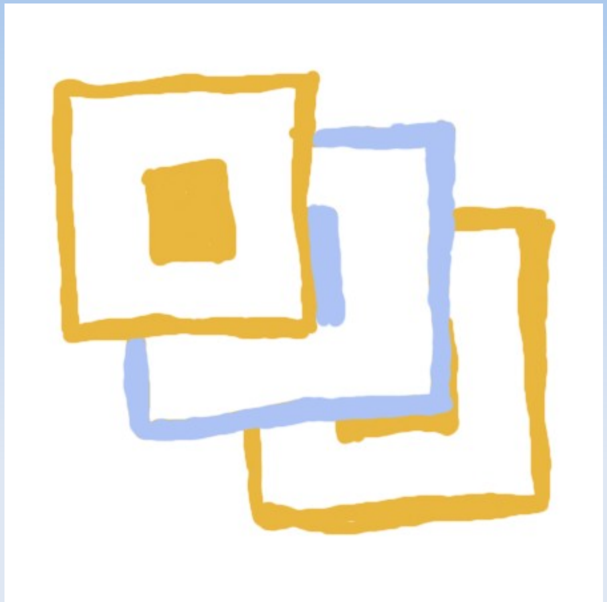


Multicore World, Wellington, New Zealand, Feb 2017

lev@levlafayette.com

It All Begins at Multicore World

At the last Multicore World “A Laconic HPC with an Orgone Accumulator” was presented – Spartan, the hybrid HPC/cloud system at the University of Melbourne; uses cloud architecture for single-node jobs and HPC architecture for multi-node jobs with the distribution based on profiling..



Spartan was launched in June 2016, and was subject to presentations at Linux User of Victoria (July), eResearch Australasia (October), Frankfurt University, University of Stuttgart, University of Freiburg, CERN, CINES (October), BSC-CNS (November) and the OpenStack Summit (November). People were *very interested* indeed in a HPC/Cloud hybrid for throughout and financial reasons.

Img: Multicore World logo from Twitter, Shed 6 from Seven Events

We Were Not Alone!

The Albert-Ludwigs-University Freiburg HPC centre also had a hybrid system. But from the opposite direction. Whereas the University of Melbourne's Spartan system treated a cloud resource as a separate partion for HPC-style job submissions the University of Freiburg used HPC compute nodes as a resource to launch cloud VMs.

The University of Freiburg has configured NEMO to virtual machines as standard compute jobs (VM jobs) through the resource manager (bare metal jobs) or inside a virtual machine (VM job) without partitioning the cluster into two parts.

NEMO is a CentOS 7 cluster with has 756 compute nodes (plus a few others) using Intel Broadwell for a total of 526 TFlops of performance and OmniPath interconnect. The name comes from the main use cases (Neuroscience, Elementary Particle Physics and Microsystems Engineering).

NEMO and Spartan Architecture

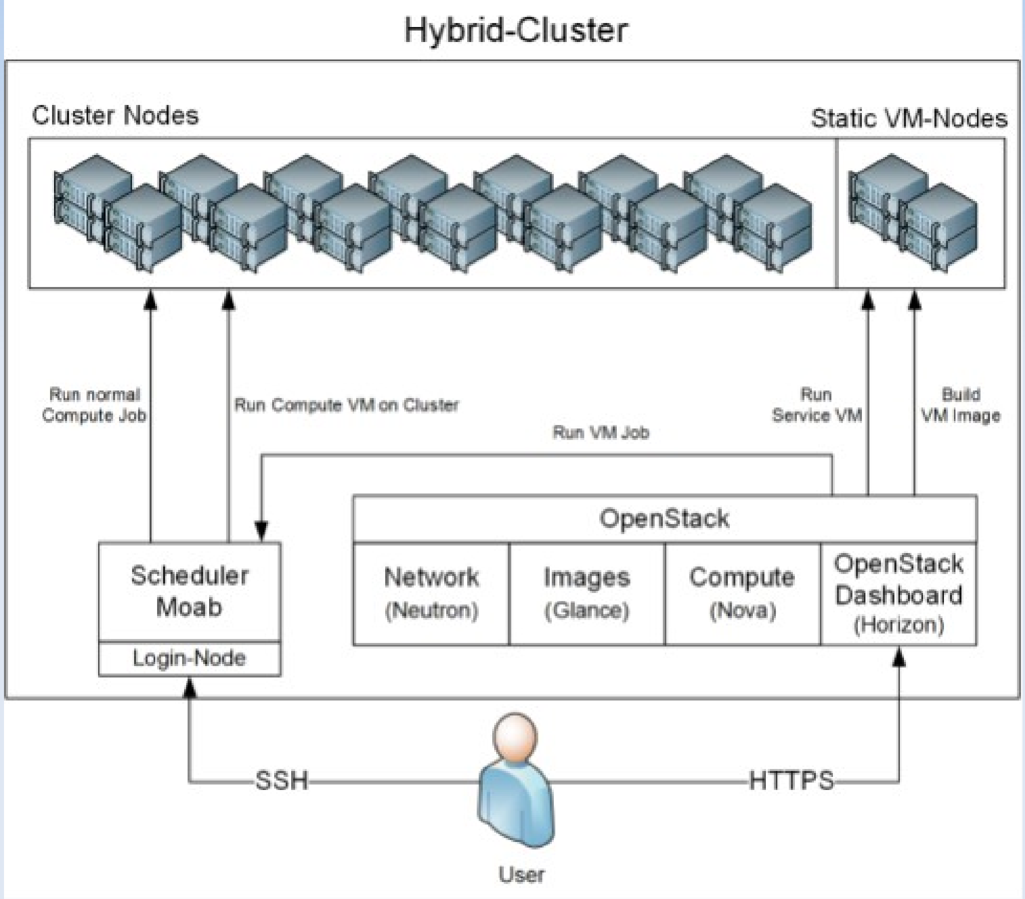
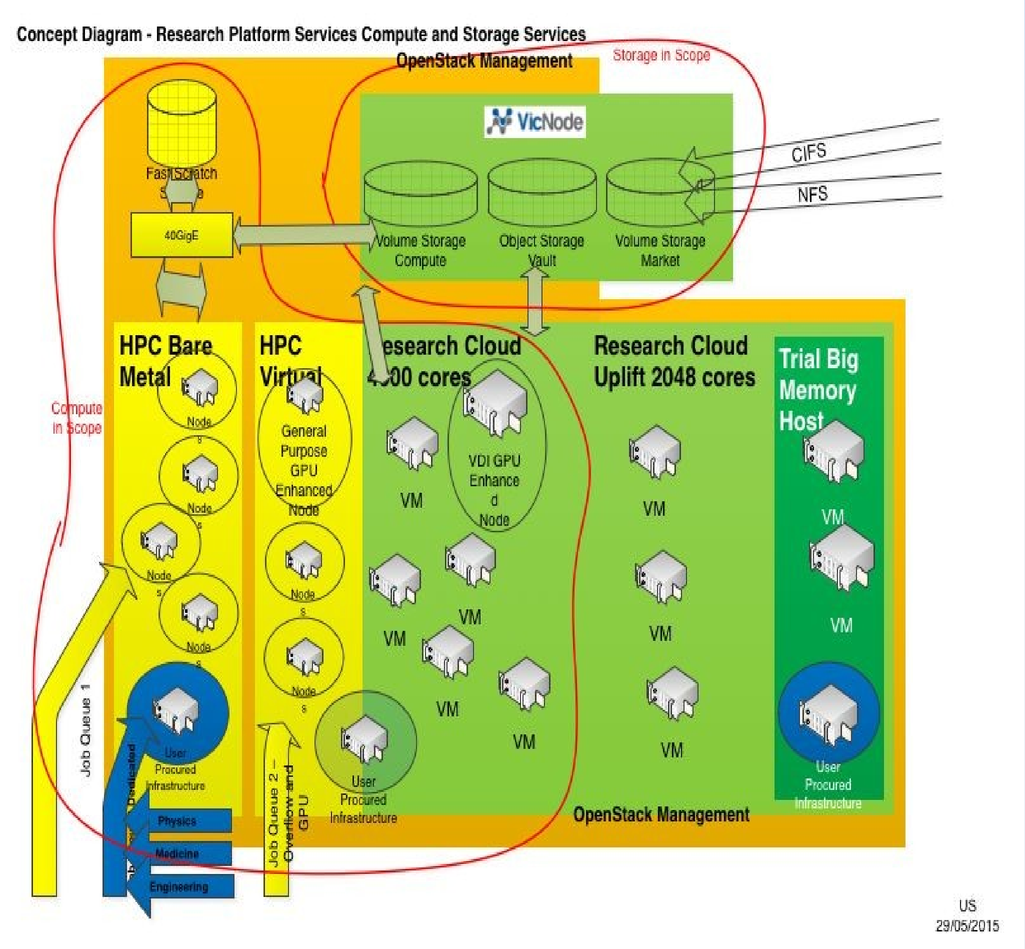


Fig. 3: OpenStack integration and user access



Resource Allocation and Use Cases

A financial and throughput justification for the different approaches is a matter of resource allocation. Unused clock cycles an expense, and any computational resource that is not being utilised is merely using electricity and generating heat. In Freiburg's situation making compute nodes available for cloud virtual machines allowed for backfilling. In Melbourne's case a surplus of cloud resources allowed for the opportunity of compute utilisation through the standard job submission framework.

A user and job profile justification means finding the right architecture and environment for particular tasks. For cloud jobs a specific software stack and operating system is often required, at the cost of decentralised set-up. The alternative, attaching an cloud resource as an HPC partition, loses the decentralisation – costing flexibility but saving time. In the latter example, deciding the appropriate allocation resources to ensure that the different queues are of an appropriate size and specification to provide the best possible resource allocation.

Img: from mimiandeunice.com

IF I STEAL YOUR COPY, YOU DON'T HAVE IT ANY MORE.
IF I MAKE A NEW COPY WE BOTH HAVE ONE!
IF YOU COMPETE WITH MY MONOPOLY, I DON'T HAVE IT ANY MORE!
WHAM!
mimiandeunice.com

Workflows

For the Freiburg case job submission to moab take the following form:

```
msub -l vm=cloudinit -l image=<id|name> myjob.moab
```

Or the user submits a VM job via msub to the Moab scheduler, e.g:

```
msub -l vm=pbs -l image=<id|name> myjob.moab
```

Or user is logs into the OpenStack web dashboard (Horizon) and starts a VM on the Cluster and the start VM command is send t the OpenStack API

In the Melbourne case, submission is simply a matter of allocating the right partition:

```
#!/bin/bash

#SBATCH -p cloud

#SBATCH --nodes=1

#SBATCH --ntasks=8
```

Future Developments

For Freiburg, the possibility to start cluster jobs within VMs enables some interesting features including:

Suspend/Resume of a VM: Implemented by mapping the Moab commands to OpenStack commands to pause/hibernate and resume the virtual machine. This can be used for preemption or maintenance. Instead of killing the job in case of a preemption the job can be paused and migrated to a “parking node” until it could run on the cluster again.

Live migration of a VM: It is possible to migrate a VM during runtime in OpenStack. A corresponding implementation in Moab gives the opportunity to migrate compute jobs during runtime to optimize the overall cluster utilization.

Archive compute environment: Research agencies may have requirements to archive working environments.

For Melbourne, a flexible partitioning system has allowed different features:

Configurations: The opportunity to build particular configurations of virtual machines and attach them to the HPC submission system has provided the performance of traditional HPC with a high-speed interconnect, and the flexibilty of cloud compute. In addition there is is the investigation of adding varied architectures on the same system, which will of course require alternative application installations.

Cloud Bursting: Having successfully conducted experiments within the NeCTAR research cloud, in the very near future the University of Melbourne system will be extended to include cloudburtsting to external providers (e.g., Amazon, Azure).

Cyborgs and Chimeras

The two models - HPC with Cloud VMs on Compute Nodes, and HPC with Compute Nodes as Cloud VMs - represent different hybrid systems to solve different problems. In effect, the University of Freiburg model provides a “cyborg”, where the HPC compute nodes are replaced with cloud virtual machines, whereas the University of Melbourne model provides a “chimera”, a multi-headed beast where the virtual machines have become new cloud nodes.

In the former case there was a desire to make existing compute nodes available to researchers for their particular configurations. In the latter case there was a desire to make virtual machines accessible to an HPC system to provide a cost-efficiencies and improved throughput. The two approaches illustrate the importance of HPC-Cloud hybrids in the provision of general purpose research computing. As computational tasks require additional architectural flexibility there is no doubt that more institutions worldwide will adopt the hybrid model.

The University of Freiburg and the University of Melbourne are co-sponsoring a BoF on HPC/Cloud hybrids at the International Supercomputing Conference in Frankfurt, June 18-22

Img: Terminiator movie and NichtElf on DeviantArt

Thanks

Thanks to the University of Melbourne for supporting the tour of the European HPC centres, and attendance to the OpenStack Barcelona Summit.

Thanks to the University of Melbourne for supporting this visit to New Zealand, and Multicore World.

Thanks to the staff of Freiburg University in particular Bernd Wiebelt, Michael Janczyk, Dirk von Suchodoletz, and Konrad Meier.

Thanks to the Spartan staff; Bernard Meade, Daniel Tosello, Linh Vu, and Greg Sauter.

Thanks to Nicolas and the Multicore World team for organising yet another great conference!

THANKS FOR WATCHING
& LISTENING PATIENTLY