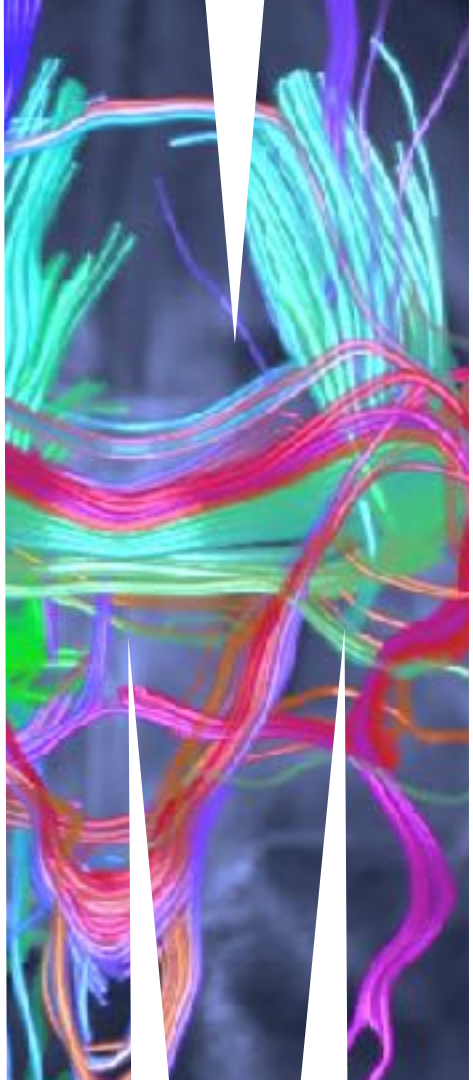
MONASH University

MONASH
eRESEARCH

Building a GPU-enabled OpenStack Cloud for HPC

Blair Bethwaite (and many others)

GROUP OF EIGHT AUSTRALIA



Monash eResearch Centre:

Enabling and Accelerating 21st Century Discovery through the application of advanced computing, data informatics, tools and infrastructure, delivered at scale, and built by with “co-design” principle (researcher + technologist)

Collect

Compute

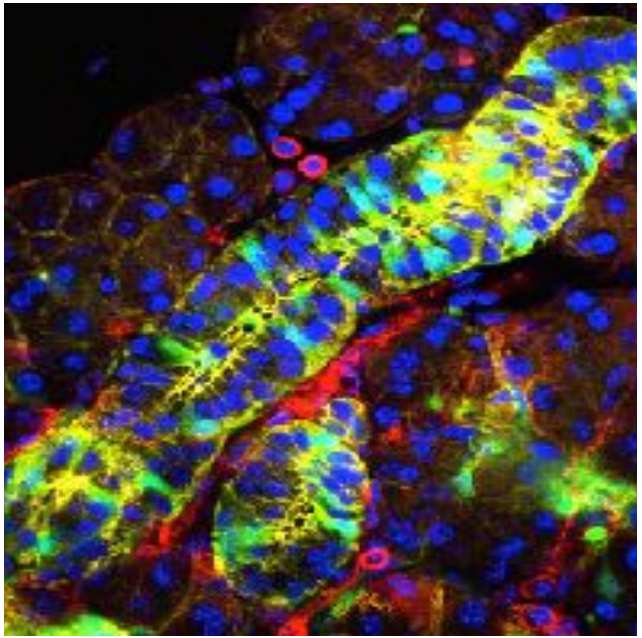
Comprehend

Collaborate

Communicate

Customise

ecosystem for life sciences HPC

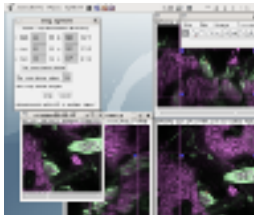


Imaging as a major driver of HPC for the life sciences

Instrument(s)

Experiment(s)

Command Line / Batch



Desktop Tools

Rich Web Tools

HPC

Databases and Reference Data

MONASH University

Professor Trevor Lithgow

ARC Australian Laureate Fellow

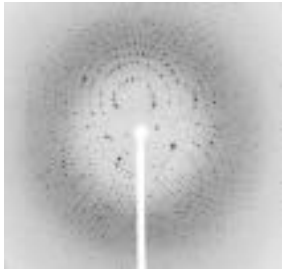
Discovery of new protein transport machines in bacteria, understanding the assembly of protein transport machines, and dissecting the effects of anti-microbial peptides on anti-biotic resistant “super-bugs”

FEI Titan Krios

Nationally funded project to develop environments for Cryo analysis

MMI Lattice Light Sheet

Nationally funded project to capture and preprocess LLS data

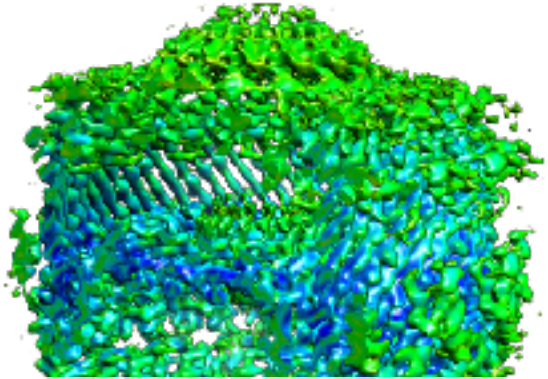


Synchrotron MX

Store.Synchrotron Data Management

MASSIVE M3

Structural refinement and analysis



Chamber details from the nanomachine that secretes the toxin that causes cholera.

Research and data by Dr. Iain Hay (Lithgow lab)

MASSIVE

Multi-modal Australian ScienceS Imaging and Visualisation Environment
Specialised Facility for Imaging and Visualisation

~$2M per year funded by partners and national project funding

Partners

Monash University
Australian Synchrotron
CSIRO

Affiliate Partners

ARC Centre of Excellence in Integrative Brain Function
ARC Centre of Excellence in Advanced Molecular Imaging

HPC

150 active projects
1000+ user accounts
100+ institutions across Australia

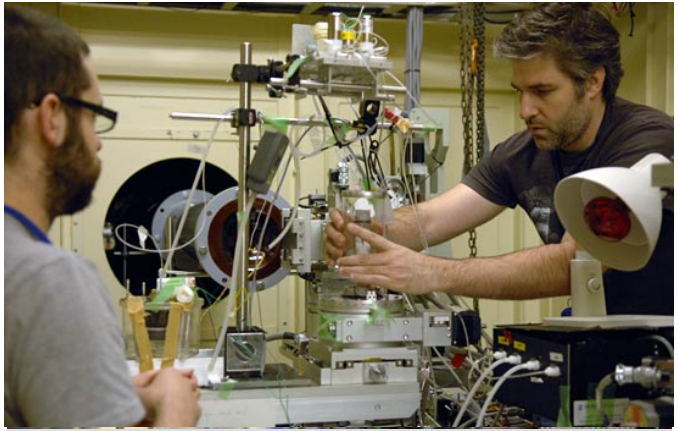
Interactive Vis

600+ users

Instrument Integration

Integrating with key Australian Instrument Facilities.
– IMBL, XFM
– CryoEM
– MBI
– NCRIS: NIF, AMMRF

Large cohort of researchers new to HPC



CT Reconstruction at the

Imaging and Medical Beamline Australian Synchrotron

Imaging and Medical Beamline

- Phase-contrast x-ray imaging, which allows much greater contrast from weakly absorbing materials such as soft tissue than is possible using conventional methods
- Two and three-dimensional imaging at high resolution (10 μm voxels)
- CT reconstruction produces multi-gigabyte volumes

Analysis:

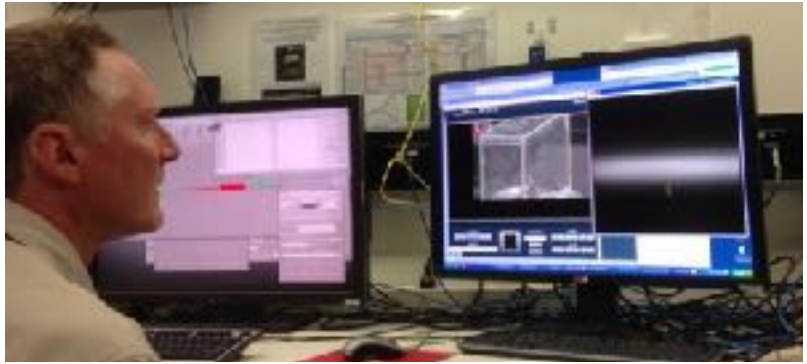
- Capture to M1 file system
- Easy remote desktop access through AS credentials
- Dedicated hardware to CT reconstruction
- CSIRO X-TRACT CT reconstruction software
- A range of volumetric analysis and visualisation tools
- Built on M1 and M2 (306 NVIDIA M2070s and K20s)

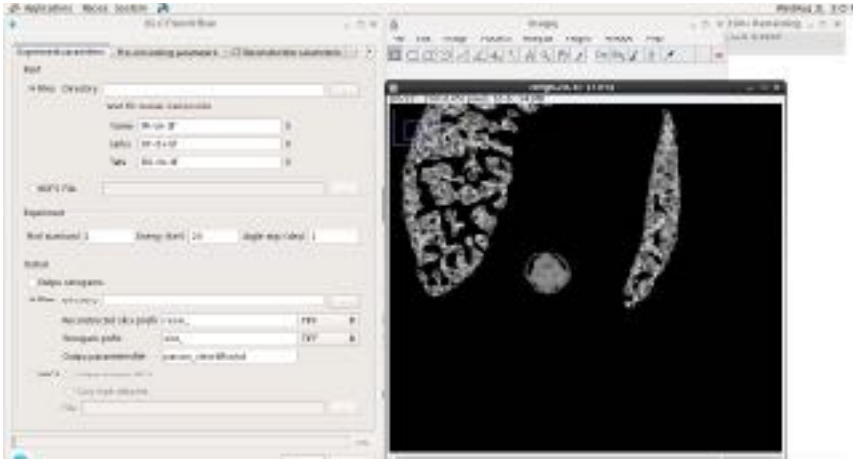
Data Management:

- Data to dedicated VicNode storage by experiment
- Available to researchers for at least 4 months after experiment
- Continued access to MASSIVE Desktop for analysis

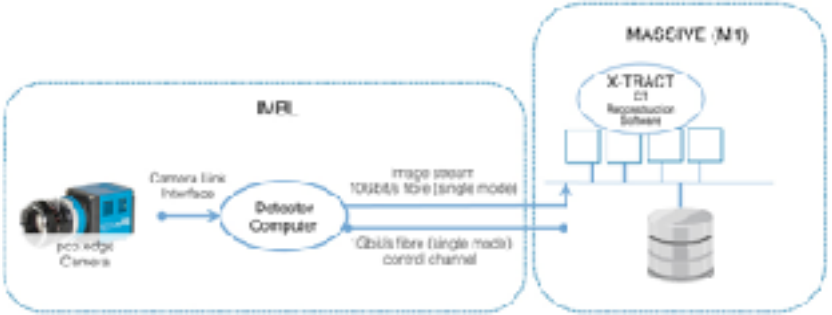
IMBL User View

Remote Desktop with Australian Synch credentials during and after experiment

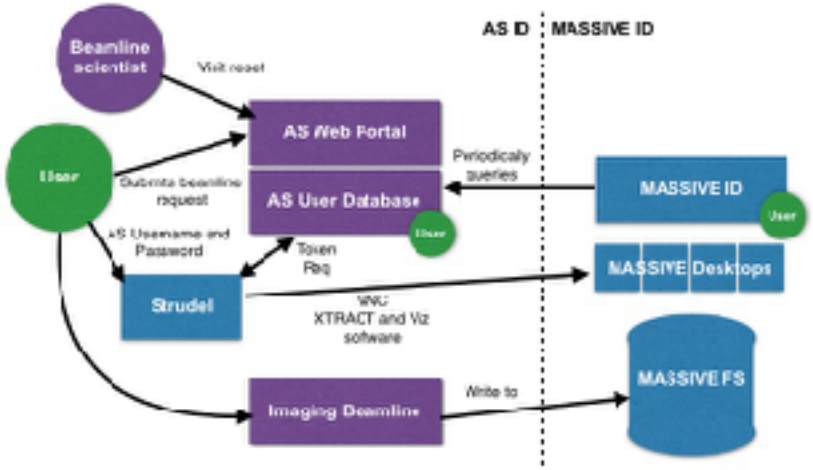




Hardware Layer Integration



Systems View



M3

M3 at Monash University
(including upcoming upgrade)

A Computer for
Next-Generation Data Science

2100 Intel Haswell CPU-cores
560 Intel Broadwell CPU-cores

NVIDIA GPU coprocessors for data processing and visualisation:

- 48 NVIDIA Tesla K80
- 40 NVIDIA Pascal P100 (16GB PCIe) (upgrade)
- 8 NVIDIA Grid K1 (32 individual GPUs) for medium and low end visualisation

A 1.15 petabyte Lustre parallel file system

100 Gb/s Ethernet Mellanox Spectrum
Supplied by Dell, Mellanox and NVIDIA

Alan Finkel
Australia's Chief Scientist

Steve Oberlin, Chief Technology Officer
Accelerated Computing, NVIDIA

M3

M3 is a little different

Priority on:
– File system in the first instance
– GPU and interactive visualisation capability

Hardware deployment through R@CMon (local research cloud team), provisioning via OpenStack
– Leverage
– Organisational
– Technical

Middleware deployment using “cloud” techniques
– Ansible “cluster in an afternoon”
– Shared software stack with other Monash HPC systems

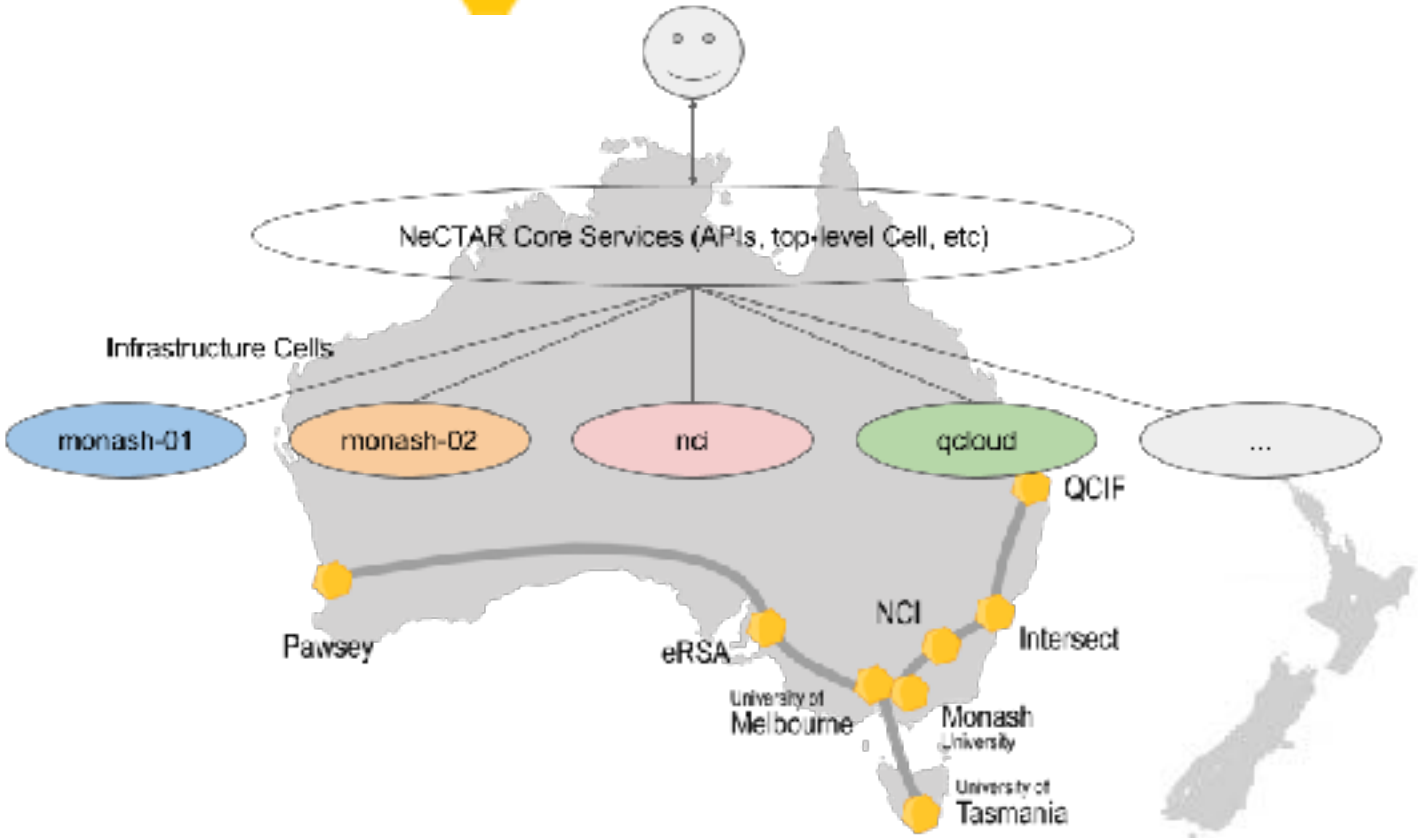
Expectations

– 24 gigabyte a second read (4x faster than M2)
– Scalable and extensible
– High end GPU and Desktop - K80
– Low and desktop - K1
– 4-way K80 boxes (8 GPUs) for dense compute-bound workloads
– Initially virtualised (KVM) for cloud-infrastructure flexibility, with bare-metal cloud-provisioning to follow late 2017

nectar

- UniMelb, as lead agent for Nectar, established first Node/site of the Research Cloud in Jan 2012 and opened doors to the research community
- Now eight Nodes (10+ DCs) and >40k cores around Australia
- Nectar established an OpenStack ecosystem for research computing in Australia
- M3 built as first service in a new “monash-03” zone of the Research Cloud focusing on HPC (computing) & HPDA (data-analytics)

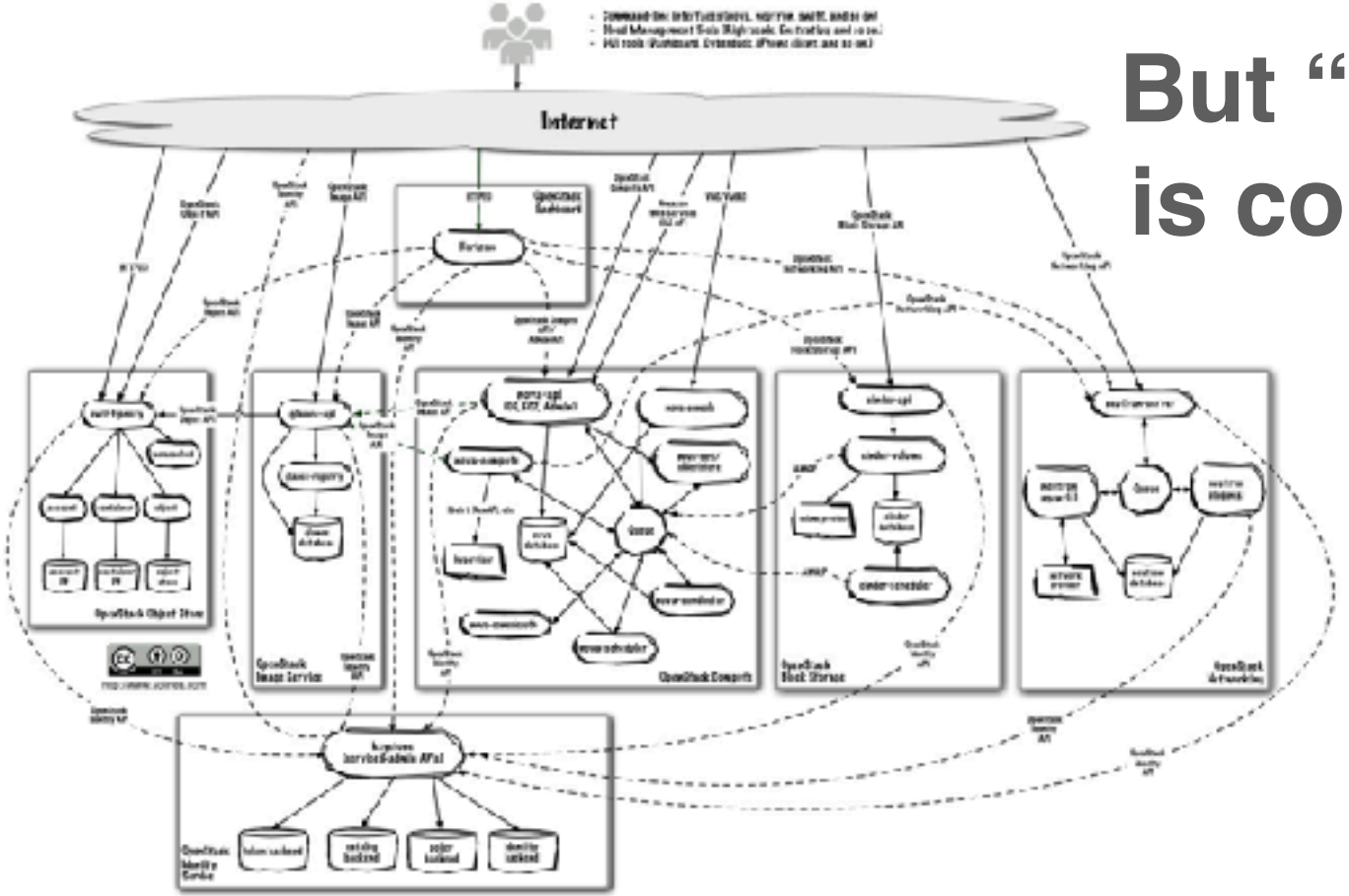
nectar


NeCTAR Core Services (APIs, top-level Cell, etc)
Infrastructure Cells
monash-01
monash-02
nci
qcloud
...
QCIF
Pawsey
eRSA
NCI
Intersect
University of Melbourne
Monash University
University of Tasmania

Why OpenStack

- Heterogeneous user requirements
  - same underlying infrastructure can be expanded to accommodate multiple distinct and dynamic clusters services (e.g. bioinformatics focused, Hadoop)
- Clusters need provisioning systems anyway
- Forcing the cluster to be cloud-provisioning and managed makes it easier to leverage other cloud resources e.g. community science cloud, commercial cloud
- OpenStack is a big focus of innovation and effort in the industry - benefits of association and osmosis
- Business function boundaries at the APIs

But “OpenStack is complicated”



Not so complicated

- http://www.openstack.org/software/sample-configs
- new navigator with maturity ratings for each project
- helps to deconvolute the Big Tent project model
- upcoming introduction of “constellations” - popular project combinations with new integrated testing

Virtualised HPC?!

- Discussed in literature for over a decade but little production adoption
  - Very similar requirements to NFV - and this is taking off in a big way over the last 12-18 months

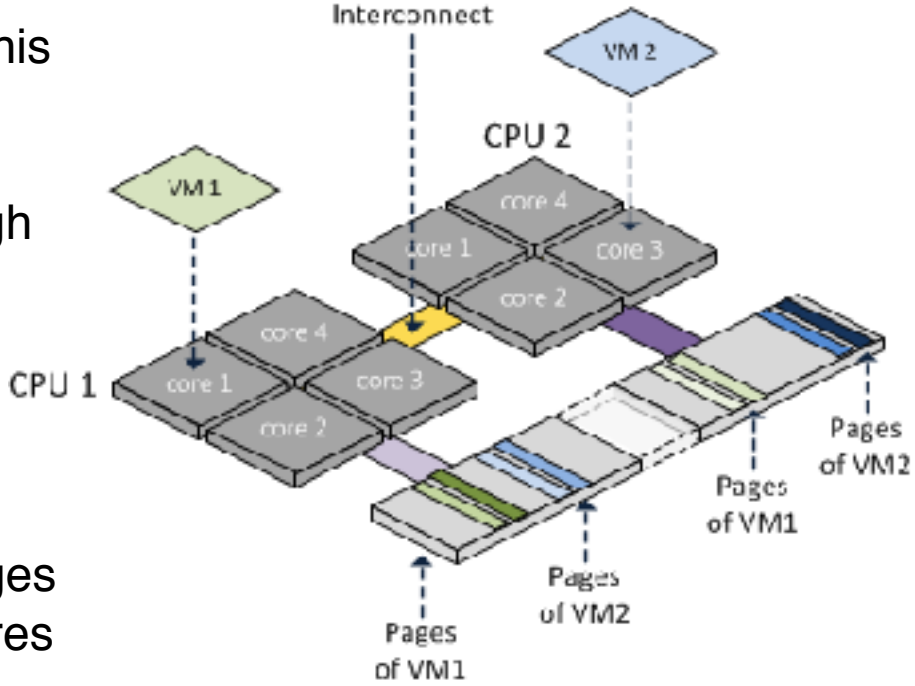
> "This study has also yielded valuable insight into the merits of each hypervisor. KVM consistently yielded near-native performance across the full range of benchmarks."
> Supporting High Performance Molecular Dynamics in Virtualized Clusters using IOMMU, SR-IOV, and GPUDirect [1]

> "Our results find MPI + CUDA applications, such as molecular dynamics simulations, run at near-native performance compared to traditional non-virtualized HPC infrastructure"
> Supporting High Performance Molecular Dynamics in Virtualized Clusters using IOMMU, SR-IOV, and GPUDirect [1]

[1] Andrew J. Younge, John Paul Walters, Stephen P. Crago, Geoffrey C. Fox

Key tuning for HPC

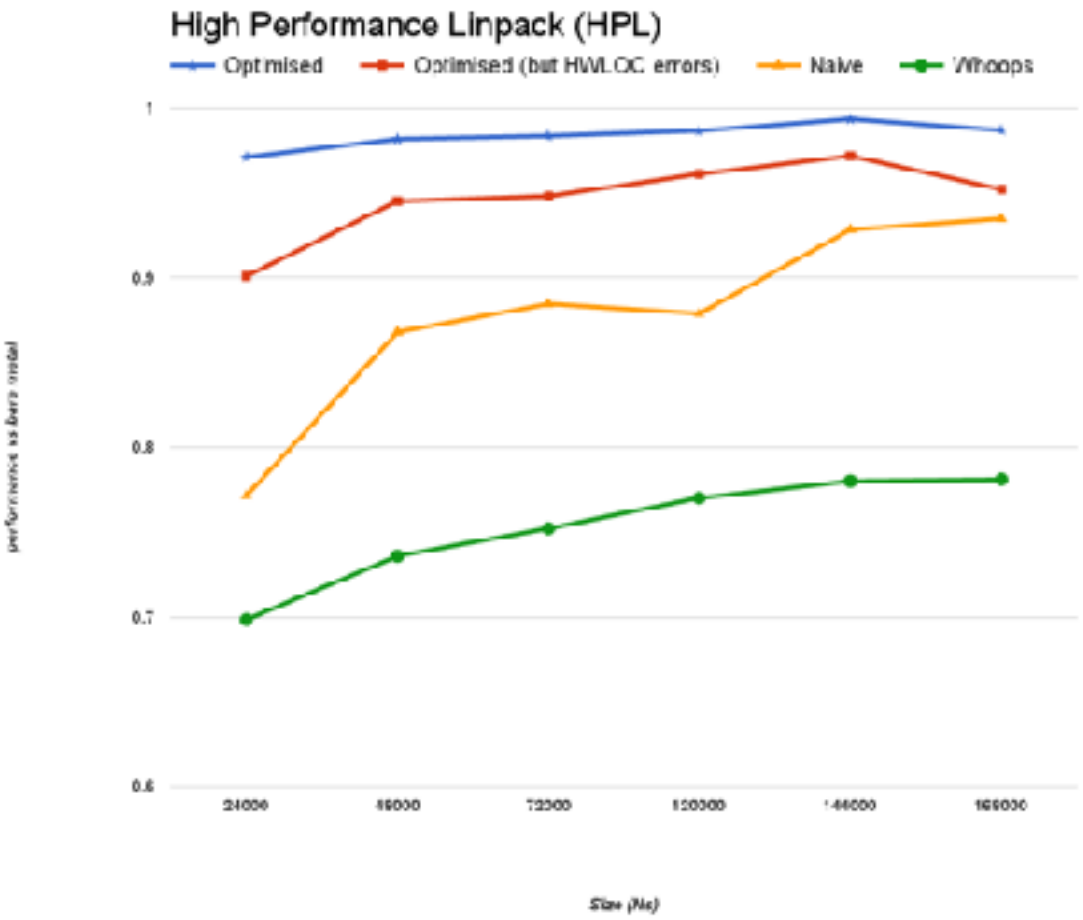
- With hardware features & software tuning this is very much possible and performance is almost native
    - CPU host-model / host-passthrough
    - Expose host CPU and NUMA cell topology
    - Pin virtual cores to physical cores
    - Pin virtual memory to physical memory
    - Back guest memory with huge pages
    - Disable kernel consolidation features

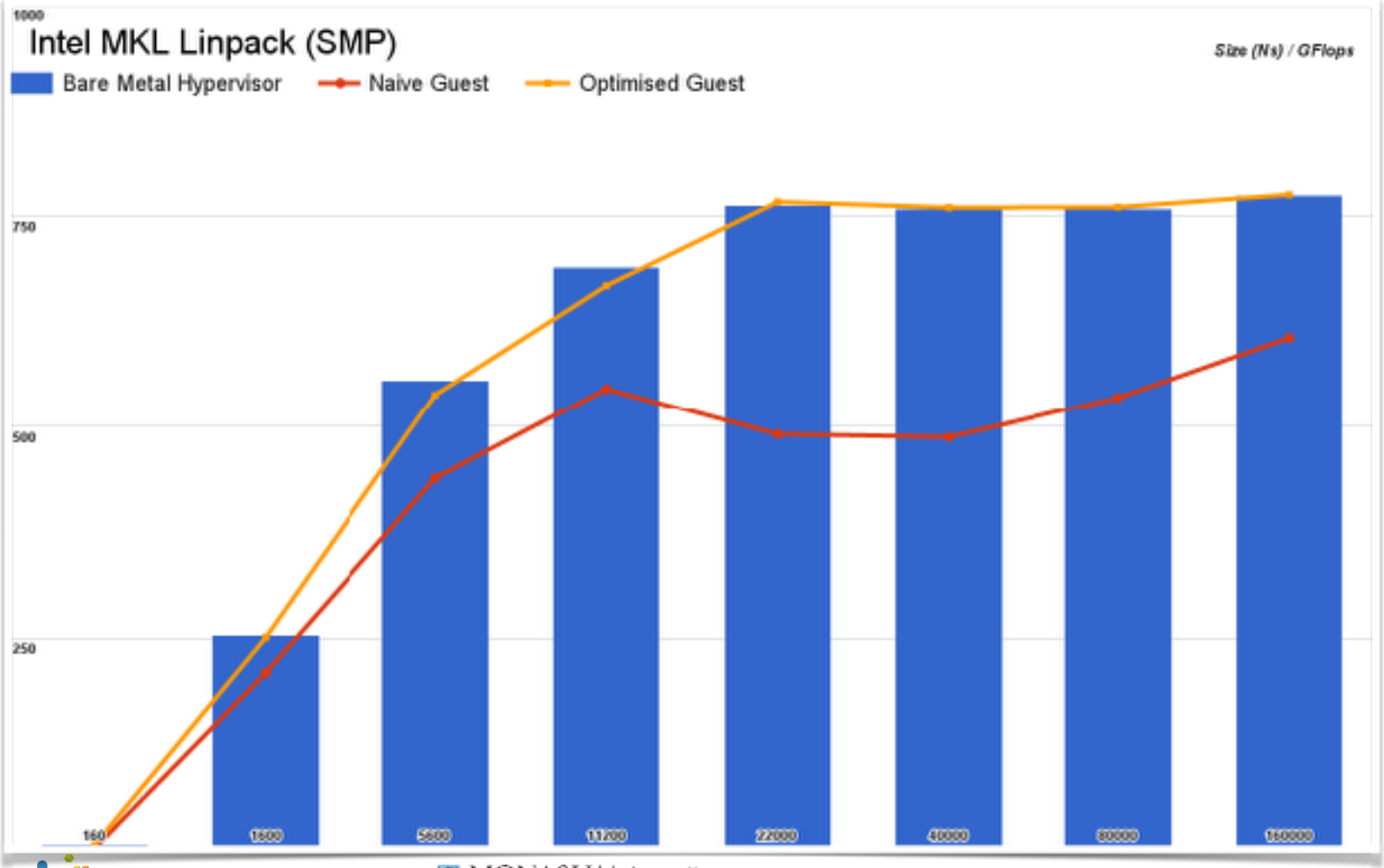


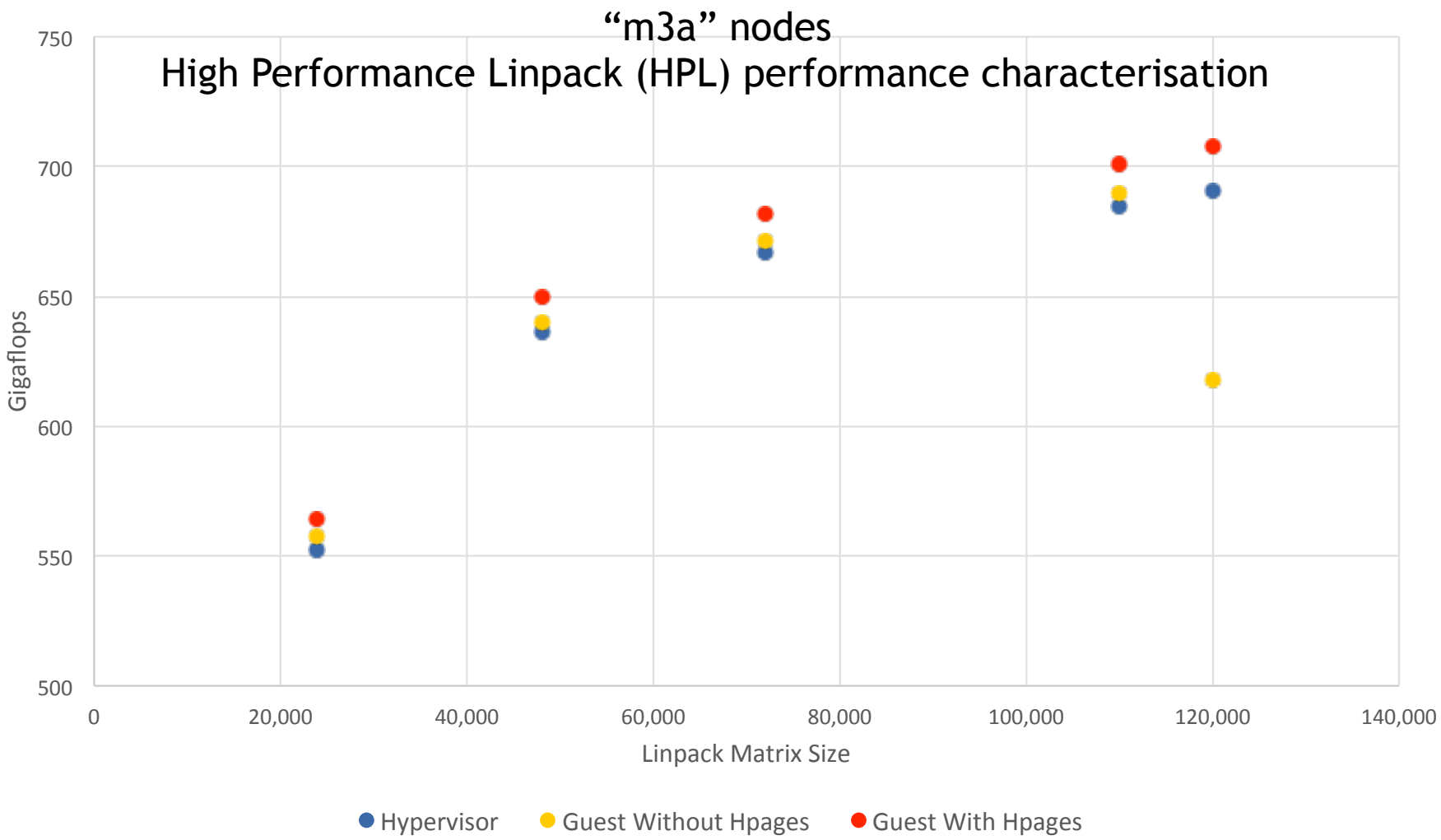
http://frankdenneman.nl/2015/02/27/memory-deep-dive-numa-data-locality/

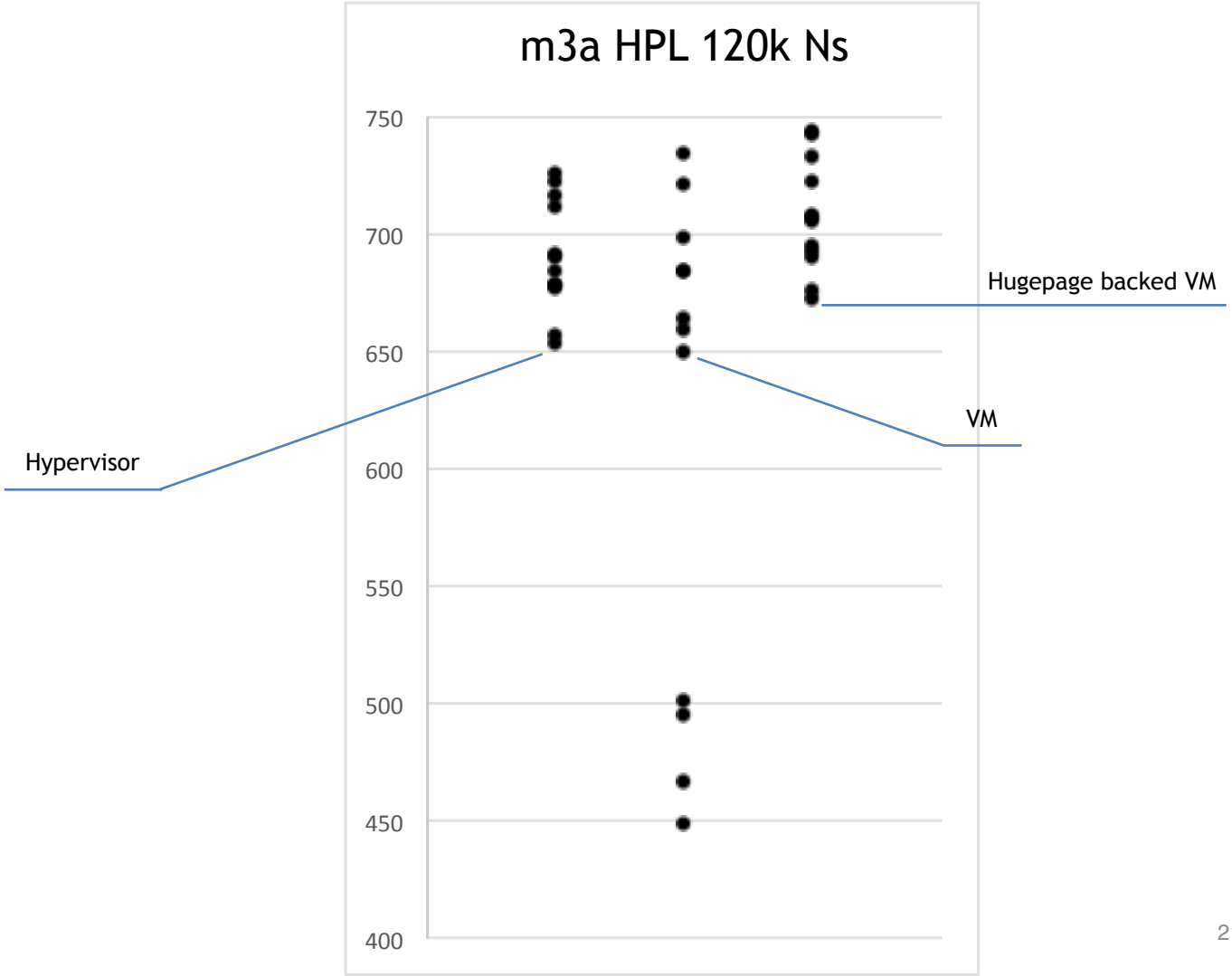
M3 Compute Performance Snapshot

- Linpack benchmarks from an “m3d” node:
    - Dell R730, 2x E5-2680 v3 (2x 12 cores, HT off), 256GB RAM, 2x NVIDIA K80 cards, Mellanox CX-4 50GbE DP
- High Performance Linpack and Intel Optimised Linpack
- Ubuntu Trusty host with Xenial kernel (4.4) and Mitaka Ubuntu Cloud archive hypervisor (QEMU 2.5 + KVM)
    - (Kernel samepage merging and transparent huge pages disabled)
- CentOS7 guest (3.10 kernel)
    - M3 large GPU compute flavor (“m3d”) - 24 cores, 240GB RAM, 4x K80 GPUs, 1x Mellanox CX-4 Virtual Function


High Performance Linpack (HPL)
Optimised
Optimised (but HWLOC errors)
Naive
Whoops
1
0.9
0.8
0.7
0.6
24000
48000
72000
120000
144000
168000
Size (Ns)


Intel MKL Linpack (SMP)
Size (Ns) / GFlops
Bare Metal Hypervisor
Naive Guest
Optimised Guest
1000
750
500
250
160
1600
5600
11200
22000
40000
80000
160000


"m3a" nodes
High Performance Linpack (HPL) performance characterisation
750
700
650
600
550
500
Gigaflops
0
20,000
40,000
60,000
80,000
100,000
120,000
140,000
Linpack Matrix Size
Hypervisor
Guest Without Hpages
Guest With Hpages


m3a HPL 120k Ns
750
700
650
600
550
500
450
400
Hypervisor
VM
Hugepage backed VM

GPU-accelerated OpenStack Instances

How-to?

1. Confirm hardware capability
    - IOMMU - Intel VT-d, AMD-Vi (common in contemporary servers)
    - GPU support
2. Prep nova-compute hosts/hypervisors
3. Configure OpenStack nova-scheduler
4. Create GPU flavor

GPU-accelerated OpenStack Instances

1. ~~Confirm hardware capability~~
2. Prep compute hosts/hypervisors
    1. ensure IOMMU is enabled in BIOS
    2. enable IOMMU in Linux, e.g., for Intel:

```
# in /etc/default/grub:
GRUB_CMDLINE_LINUX_DEFAULT="quiet splash intel_iommu=on iommu=pt rd.modules-load=vfio-pci”
~$ update-grub
```

    3. ensure no other drivers/modules claim GPUs, e.g., blacklist nouveau
    4. Configure nova-compute.conf pci_passthrough_whitelist:

```
~$ lspci -nn | grep NVIDIA
03:00.0 3D controller [0302]: NVIDIA Corporation Device [10de:15f8] (rev a1)
82:00.0 3D controller [0302]: NVIDIA Corporation Device [10de:15f8] (rev a1)
# in /etc/nova/nova.conf:
pci_passthrough_whitelist=[{"vendor_id":"10de", "product_id":"15f8"}]
```

GPU-accelerated OpenStack Instances

1. ~~Confirm hardware capability~~
2. ~~Prep compute hosts/hypervisors~~
3. Configure OpenStack nova-scheduler
    1. On nova-scheduler / cloud-controllers

```
# in /etc/nova/nova.conf:
pci_alias={"vendor_id":"10de", "product_id":"15f8", "name":"P100"}
scheduler_driver=nova.scheduler.filter_scheduler.FilterScheduler
scheduler_available_filters=nova.scheduler.filters.all_filters
scheduler_available_filters=nova.scheduler.filters.pci_passthrough_filter
.PciPassthroughFilter
scheduler_default_filters=RamFilter,ComputeFilter,AvailabilityZoneFilter,
ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter
```

GPU-accelerated OpenStack Instances

1. ~~Confirm hardware capability~~
2. ~~Prep compute hosts/hypervisors~~
3. ~~Configure OpenStack nova-scheduler~~
4. Create GPU flavor

```
~$ openstack flavor create --ram 122880 --disk 30
--vcpus 24 mon.m3.c24r120.2gpu-p100.mlx
~$ openstack flavor set mon.m3.c24r120.2gpu-p100.mlx
--property pci_passthrough:alias='P100:2'
```

GPU-accelerated OpenStack Instances

```
~$ openstack flavor show 56cd053c-b6a2-4103-b870-a83dd5d27ec1
+----------------------------+-------------------------------------------+
| Field                      | Value                                     |
+----------------------------+-------------------------------------------+
| OS-FLV-DISABLED:disabled   | False                                     |
| OS-FLV-EXT-DATA:ephemeral  | 1000                                      |
| disk                       | 30                                        |
| id                         | 56cd053c-b6a2-4103-b870-a83dd5d27ec1      |
| name                       | mon.m3.c24r120.2gpu-p100.mlx              |
| os-flavor-access:is_public | False                                     |
| properties                 | pci_passthrough:alias='P100:2,MlxCX4-VF:1' |
| ram                        | 122880                                    |
| rxtx_factor                | 1.0                                       |
| swap                       |                                           |
| vcpus                      | 24                                        |
+----------------------------+-------------------------------------------+
~$ openstack server list --all-projects --project d99… --flavor 56c…
+--------------------------------------+------------+--------+-----------------------------------+
| ID                                   | Name       | Status | Networks                          |
+--------------------------------------+------------+--------+-----------------------------------+
| 1d77bf12-0099-4580-bf6f-36c42225f2c0 | massive003 | ACTIVE | monash-03-internal=10.16.201.20 |
+--------------------------------------+------------+--------+-----------------------------------+
```

GPU Instances - rough edges

- Hardware monitoring
  - No OOB interface to monitor GPU hardware when it is assigned to an instance (and doing so would require loading drivers in the host)

- P2P (peer-to-peer multi-GPU)
  - PCIe topology not available in default guest configuration (no PCIe bus on legacy QEMU i440fx)
  - PCIe ACS (Access Control Services - forces transactions through the Root Complex which blocks/disallows P2P for security)

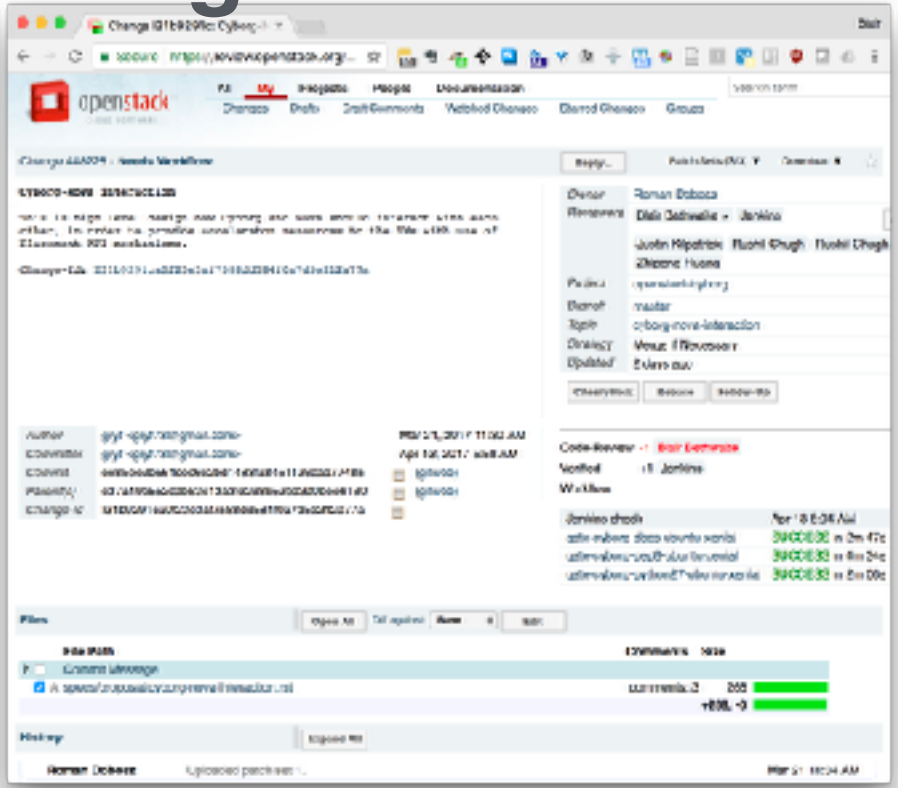
GPU Instances - rough edges

- PCIe security
  - Along similar lines as P2P issues
  - Compromised device could access privileged host memory via PCIe ATS (Address Translation Services)

- Common to use cloud images for base OS+driver versioning and standardisation, but new NVIDIA driver versions do not support some existing hardware (e.g. K1)
  - Requires multiple images or automated driver deployment/config

OpenStack Cyborg - accelerator management

… aims to provide a general purpose management framework for acceleration resources (i.e. various types of accelerators such as Crypto cards, GPUs, FPGAs, NVMe/NOF SSDs, ODP, DPDK/SPDK and so on) (https://wiki.openstack.org/wiki/Cyborg)



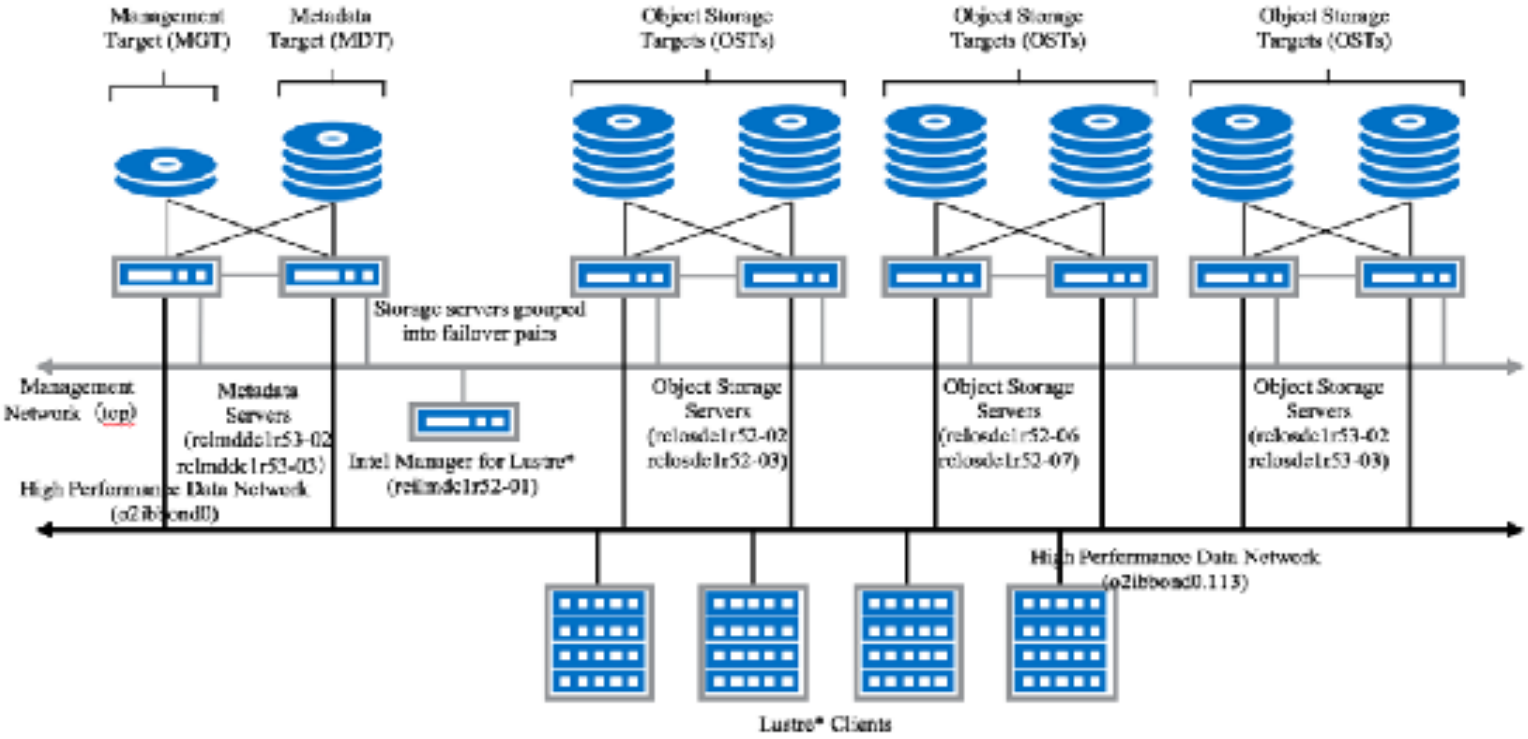
https://review.openstack.org/#/c/448228/

The Crossroads of Cloud and HPC: OpenStack for Scientific Research

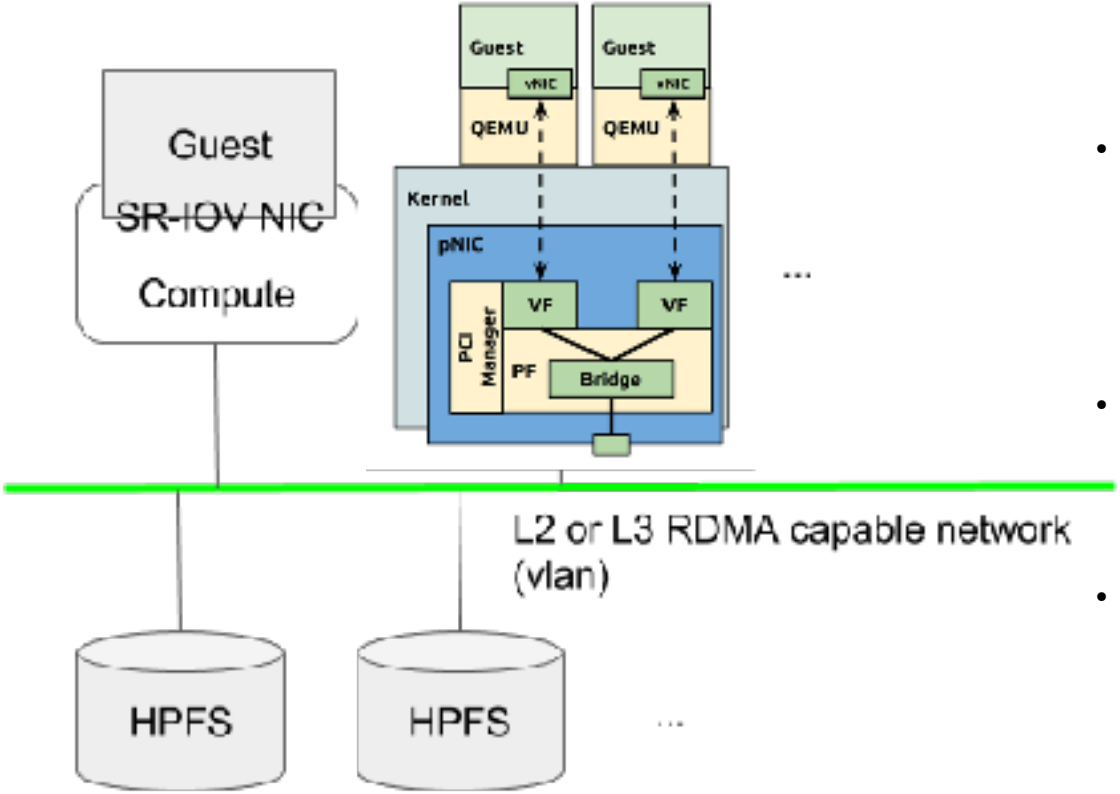
Exploring OpenStack cloud computing for scientific workloads

openstack.org

M3 Lustre


Management Target (MGT)
Metadata Target (MDT)
Object Storage Targets (OSTs)
Object Storage Targets (OSTs)
Object Storage Targets (OSTs)
Storage servers grouped into failover pairs
Management Network (ipo)
Metadata Servers (rclmddc1r53-02 rclmddc1r53-03)
High Performance Data Network (o2ibbond0)
Intel Manager for Lustre* (rclimdc1r52-01)
Object Storage Servers (rclosdc1r52-02 rclosdc1r52-03)
Object Storage Servers (rclosdc1r52-06 rclosdc1r52-07)
Object Storage Servers (rclosdc1r53-02 rclosdc1r53-03)
High Performance Data Network (o2ibbond0.113)
Lustre* Clients

M3 HPFS Integration



- special flavors for cluster instances which specify a PCI passthrough SRIOV vNIC
- hypervisor has NICs with VFs tied to data VLAN(s)
- data VLAN is RDMA capable so e.g. Lustre can use o2ib LNET driver

S7232
Processing the next generation of angstrom-scale microscopy

11:30am (now!) - room 210C

Application layers:

Open IaaS:

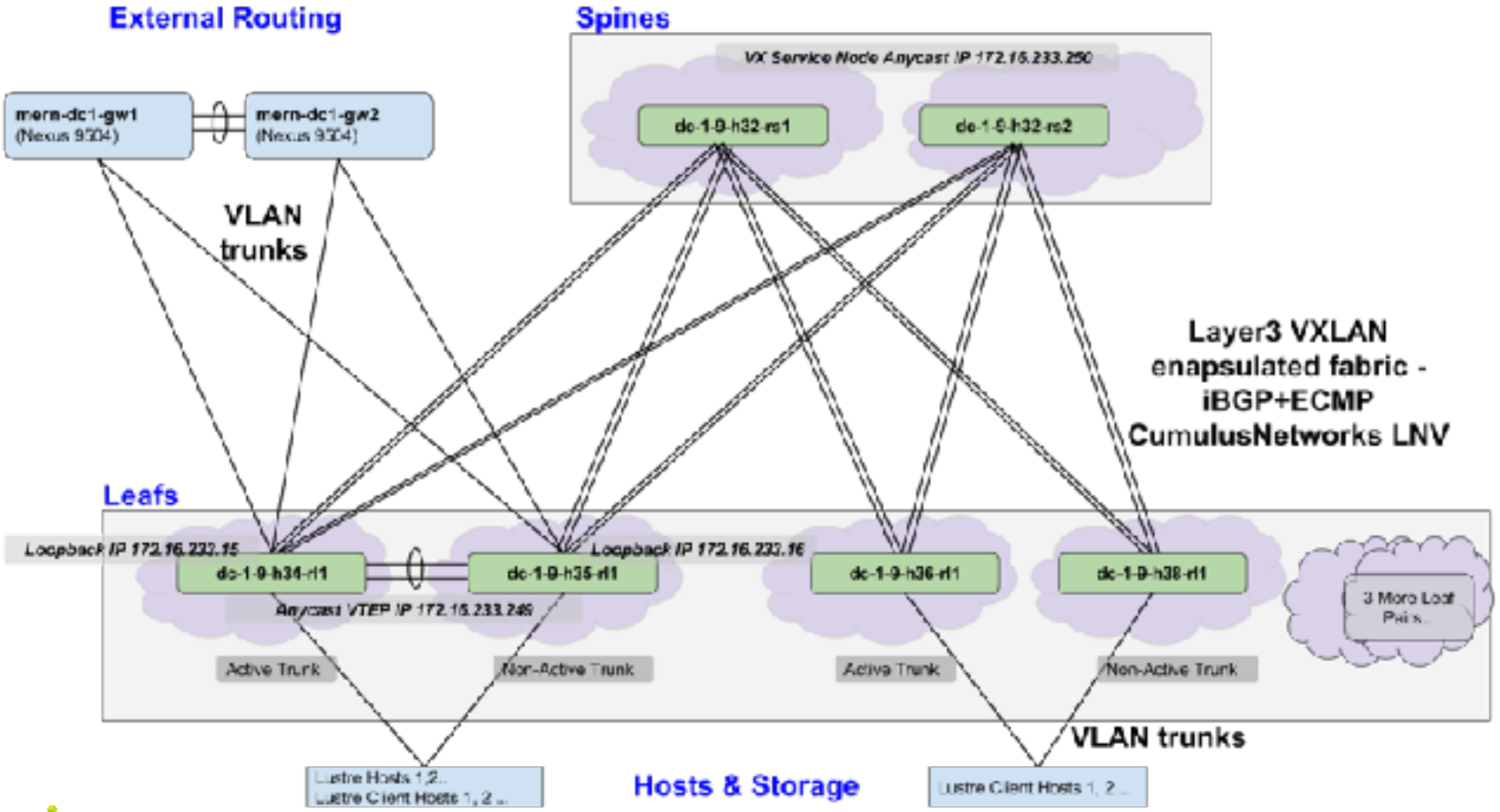
Technology:

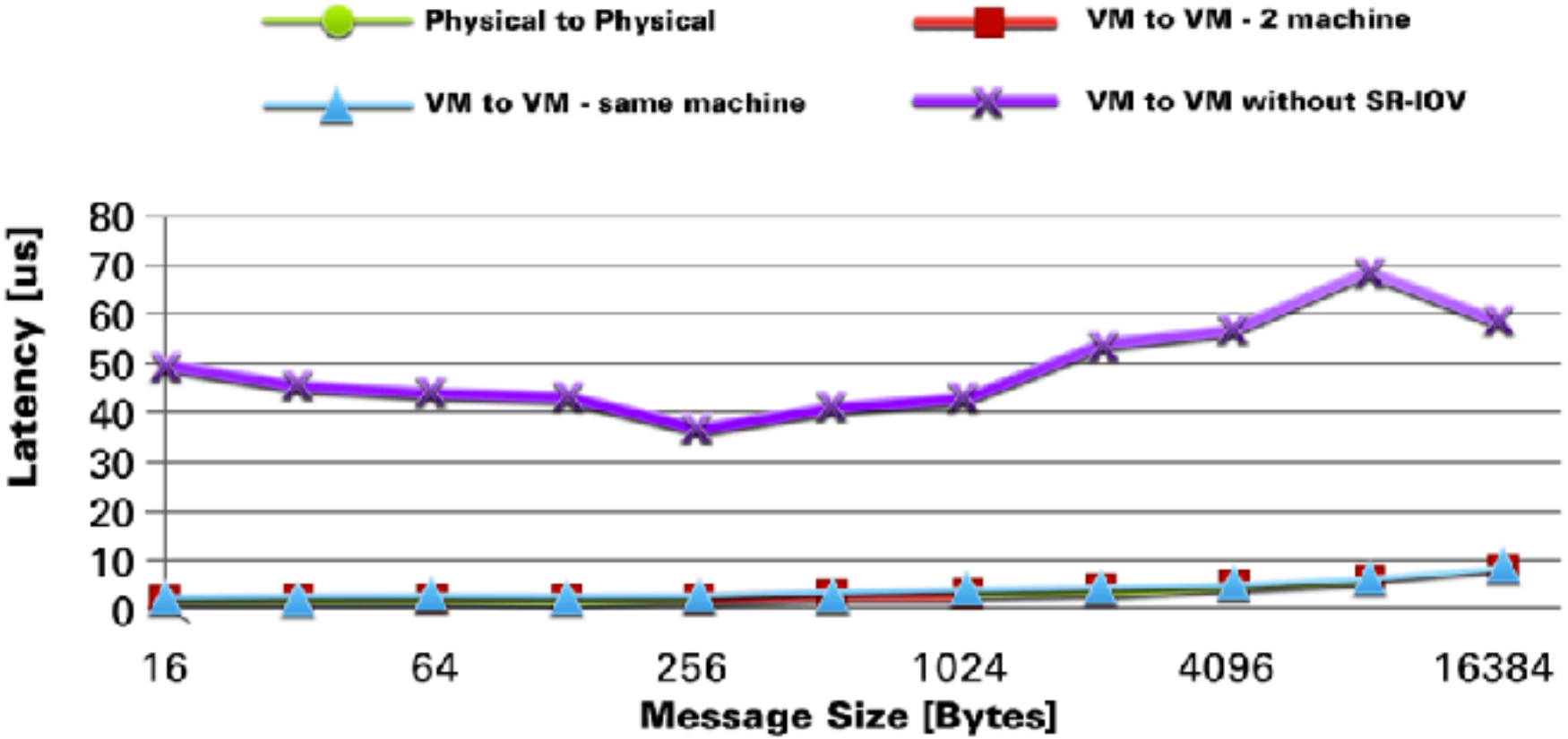
Australian Government
Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education

DEPARTMENT OF STATE DEVELOPMENT BUSINESS AND INNOVATION

Backups…

HPC-Cloud Interconnect





https://www.mellanox.com/related-docs/whitepapers/WP_Solving_IO_Bottlenecks.pdf