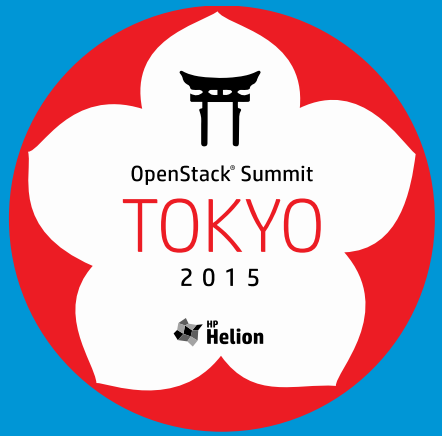
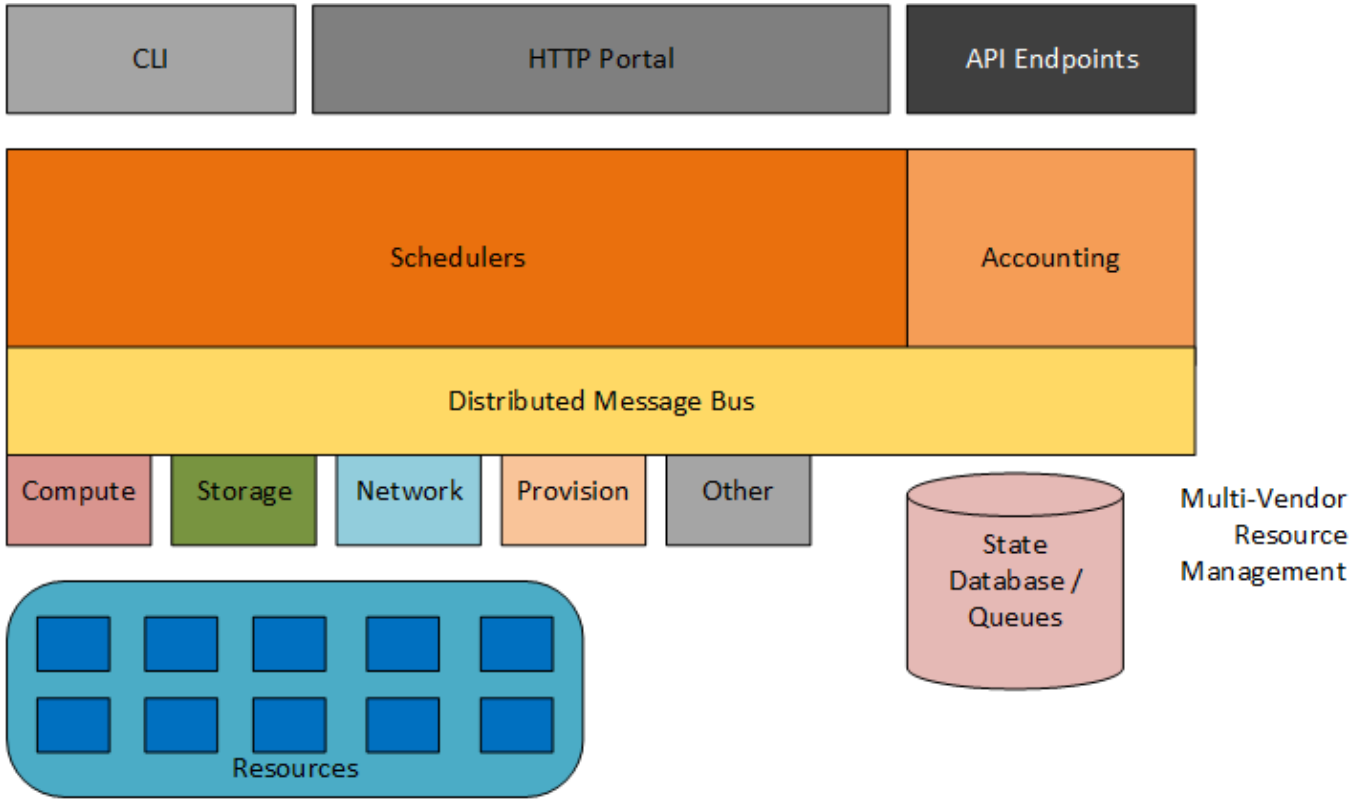
HPC on OpenStack® Use Cases

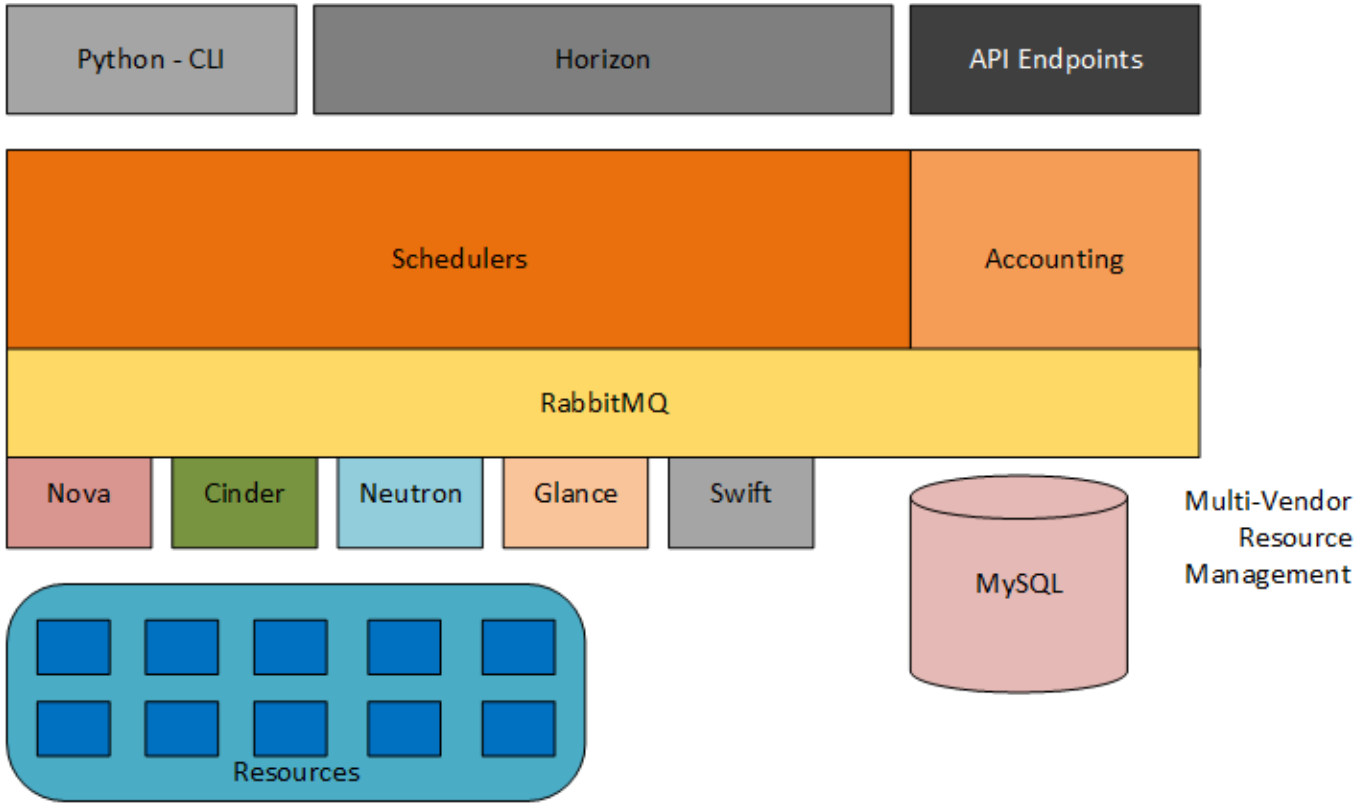


Glyn Bowden, Chief Technologist
Eric Lajoie, Snr. Solution Architect
HP Helion Professional Services

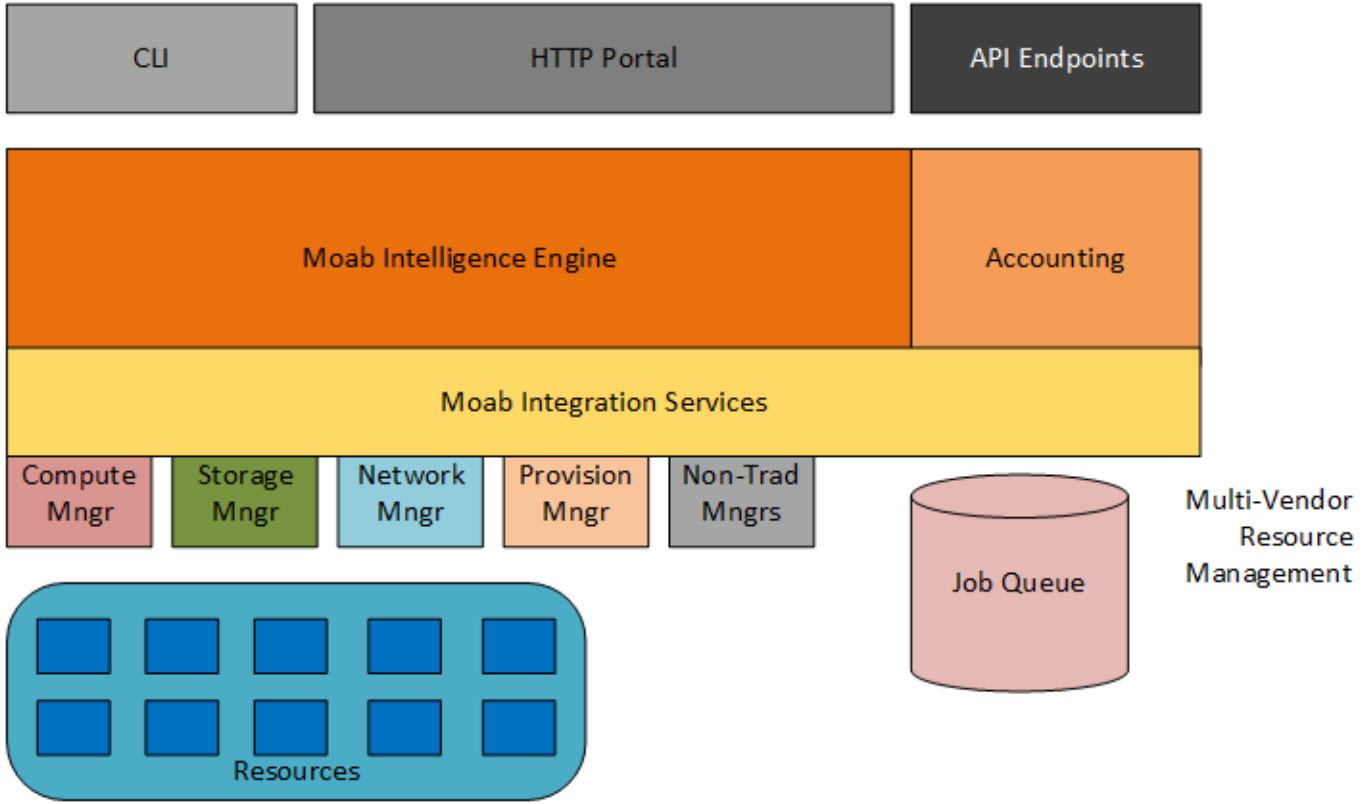
Identify the Architecture…



OpenStack Architecture



Moab HPC Suite



HPC and Cloud – Opposing Forces

HP Helion

Cloud

- Share Everything
- Generic Workloads
- Loosely Coupled
- Many small workloads

HPC and Cloud – Opposing Forces

HP Helion

Cloud

- Share Everything
- Generic Workloads
- Loosely Coupled
- Many small workloads

HPC

- Share Nothing
- Specific, Niche Workloads
- Tightly Coupled (RDMA)
- Few Large Distributed Workload

But the same…

HP Helion

Cloud

- Highly Distributed
- Large Storage Pools
- Resource Management Key
- Performance Management

HPC

- Highly Distributed
- Large Storage Pools
- Resource Management Key
- Performance Management

Background On HPC

HP Helion

- Two types of HPC
  - Analytics
    - Big Data Sets
    - Simple Operations repeated many times
    - Aggregation of results
  - Computationally Intensive
    - Smaller Data Sets
    - Very complex algorithms that need to be broken down
    - Sequential processing and summary
    - Often Latency Sensitive (RDMA, Lustre) or Bandwidth Sensitive (High Volume Filesystems, Large Files)

Background On HPC

HP Helion

- Two types of Computational HPC
  - Batch Processing
    - Loosely coupled
    - Embarrassingly Parallel
    - Limited / No shared resources during jobs
  - Realtime / Grid Computing
    - Tightly Coupled
    - Requires High Performance Networking for Remote Direct Memory Access (RDMA)
    - Usually a high performance shared file system is required
    - CPU and Memory Architecture more critical than batch

The Multi-Tenancy Challenge

HP Helion

- HPC Schedulers and Frameworks have been working very well for years, why use OpenStack?
- Tenancy and Security have always been an issue!
- Most HPC environments, if you have access to the bastion host, you tend to have access to everything. Security usually a secondary thought.
- Single, flat but high performance networks mean mixed jobs. Insecure.
- Initially a problem for Genomics. European and US orgs have laws about security of human data, and human DNA is included!
- How to have different teams use the cluster without sharing the data.
- Same can be true in pharmaceuticals and test data.

The Multi-Tenancy Solution

HP Helion

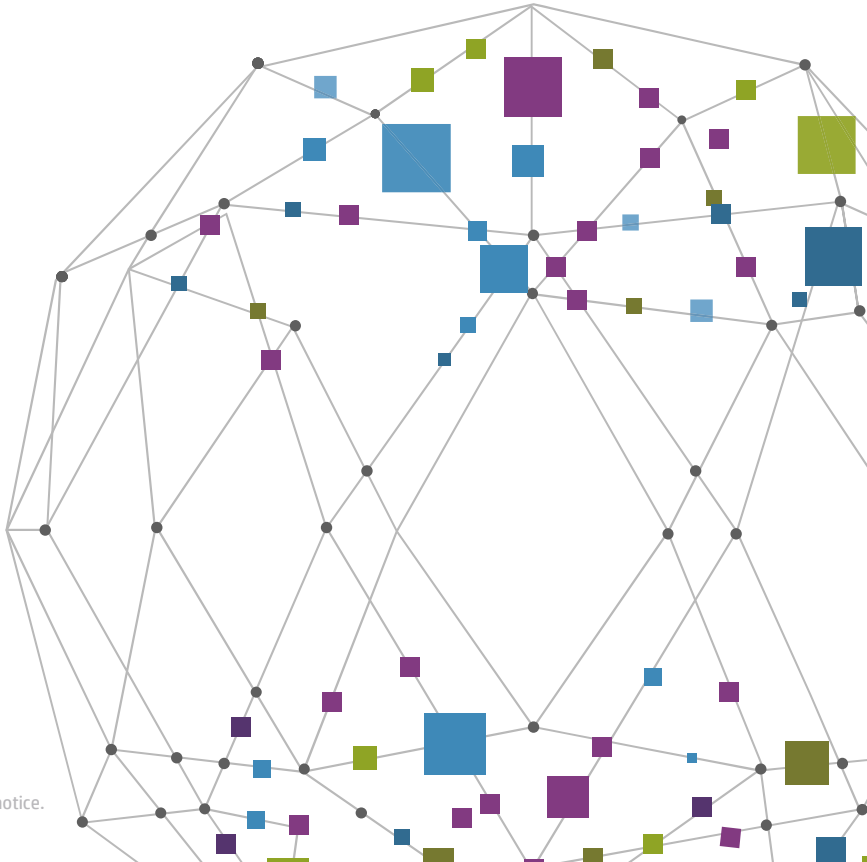
- OpenStack allows the tenants to leverage the same golden images for processing but use separated tenant networks
- Tenant based storage and provider VLANs means the data doesn't mix
- Better scheduling options within Nova now allow tenants to not compete
- Use of either Availability Zones or Cells can provide complete segregation even at compute management layer.

Federation

- Allows for multiple separate clouds to communicate, share or trust same authentication
- Useful for abstracting cloud models. Share API standards and Authentication model, everything else can be different.
- Every cloud owner can manage their own authentication domains but leverage others.

Compute

HP Helion



Compute

HP Helion

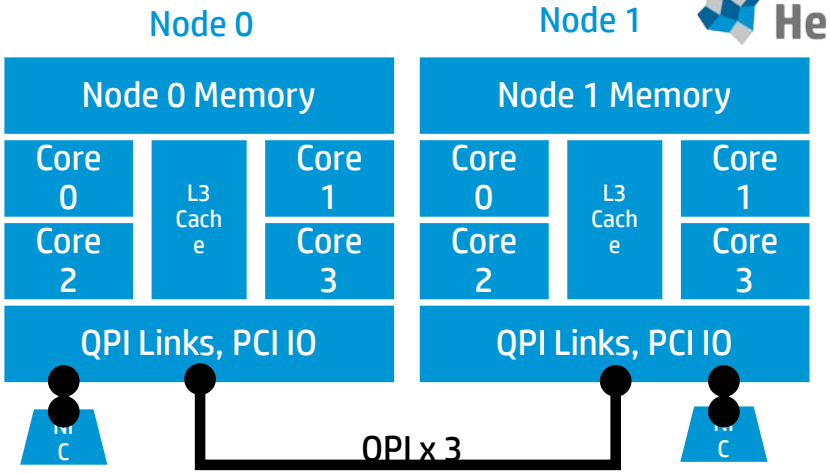
- Multi-Tenancy
  - Regions
    - Separate set of API endpoints per region.
    - Reduce impact of one tenants activities on another
    - Independent tenant scaling
    - Requires more hardware (or careful management!)
  - Cells
    - Single API endpoint
    - Nova-centric
    - Hierarchical tree of Cells
    - Each cell runs own set of Nova services except API which is shared

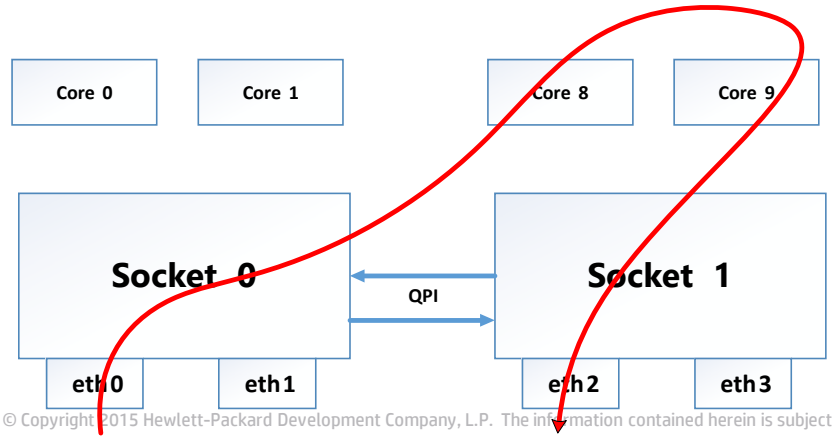
Compute cont.

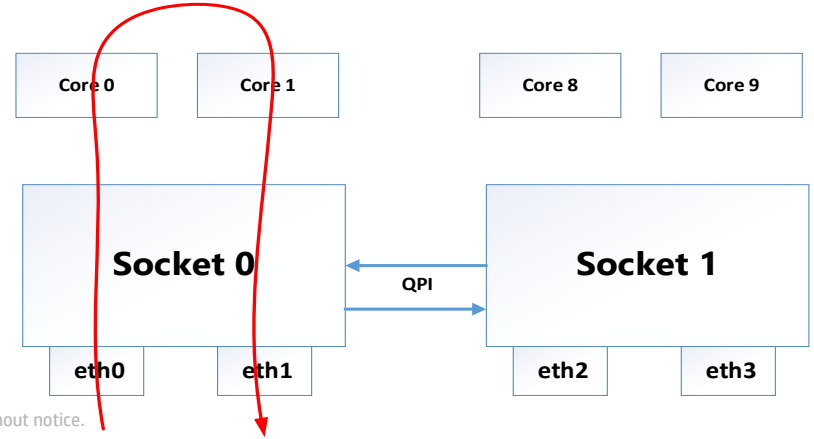
- Low Latency (Consistently)
  - RDMA over Infiniband
  - Remote Procedure Calls between nodes (RPC)
  - Dataset Shared amongst many nodes
  - Non Uniform Memory Access (NUMA) and Pinning
    - Cache Efficiency and Memory localisation
    - vCPU -> pCPU Pinning (libvirt support April 14)
  - CPU Affinity and Hyper-threading
  - DPDK Framework
- CPU Offload
  - DPDK Framework
  - NVMe Based SSD (Samsung 950 Pro)
  - GPU PCI Passthrough
  - NPIV
  - SR-IOV

NUMA & Pinning

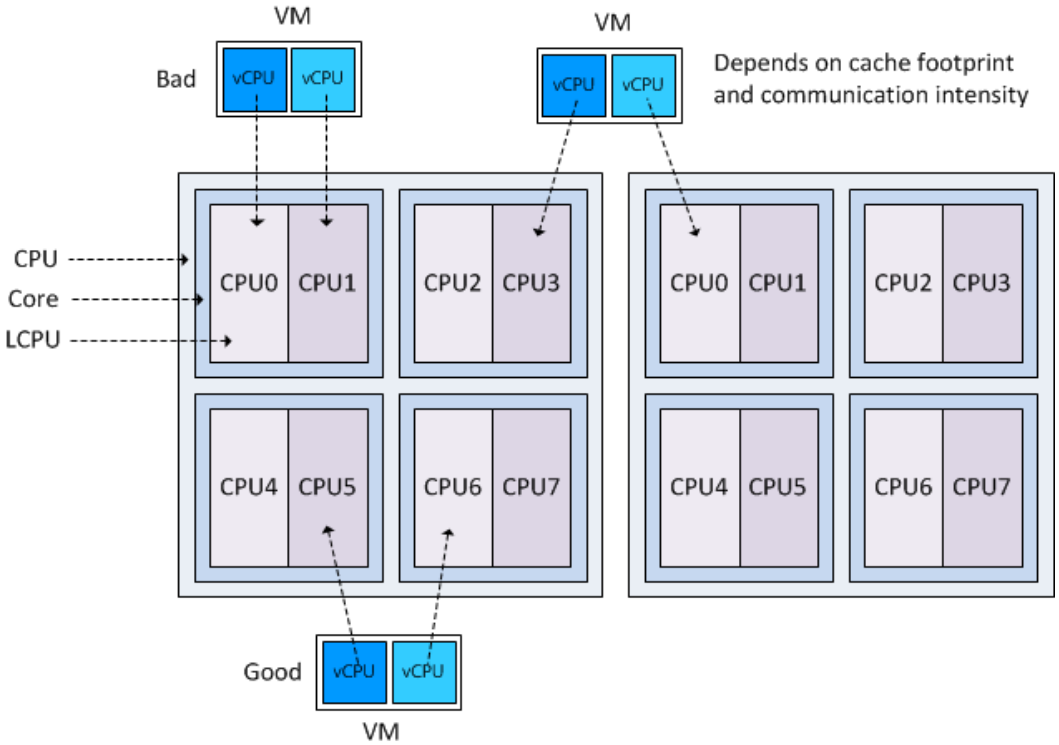
- Place VMs into your host in an ordered efficient fashion
- Exposing resources in an efficient way
- Glance image hint used for NUMA







Hyper-Threading and CPU Affinity



Fast Low Latancy Networking

HP Helion

DPDK 2.0 Requirements w/ vhost-user

- ✓ OVS 2.4
- ✓ Hugepages (2M and 1G)
- ✓ Realtime Kernel (3.19)
- ✓ NUMA
- ✓ QEMU w/ NUMA (libnuma)
- ✓ IOMMU (VT-d)
- ✓ VFIO (UIO as option is HW does not have IOMMU)

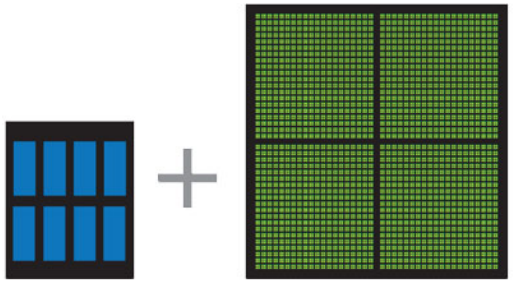
SR-IOV (not all use cases work)

VirtIO in VMs (no change needed in VM to hit performance numbers)

CPU and GPU

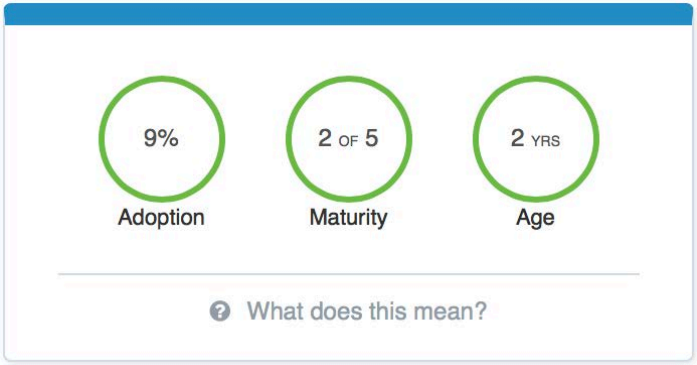
HP Helion

- GPU = Graphics Processing Unit
- Pioneered by NVidia in 2007
- Race to render more polygons for better gaming experience
- Architecture led to increased cores with dedicated GPU memory
- GPU Cores are massively parallel compared to CPU cores which number in the 10's at most.
- NVidia created CUDA (Compute Unit Device Architecture)
- CUDA designed for General Compute on GPUs
- Workload shared between GPU and CPU



Compute on Baremetal

HP Helion



- Ironic provides OpenStack Baremetal
- Integrated project as of Kilo
- Allows baremetal servers to be provisioned by Nova
- Dedicated resources, no sharing
- Hardware Agnostic (vendors responsible for own drivers for lights own management etc).
- Image based booting, not abstraction free!

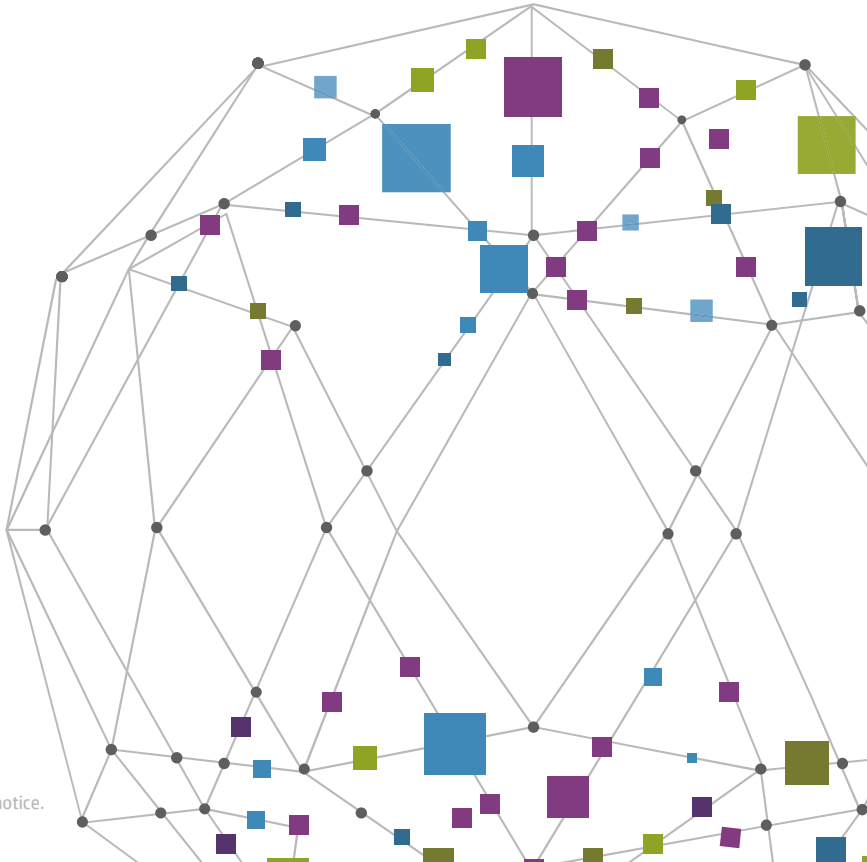
OpenStack Liberty Advances for HPC Compute

HP Helion

- VirtIO Queue Scaling
  - Hw_vif_multiqueue_enabled flag on image.
  - Provides enhanced network performance for guests with >1 vCPU or large packet sizes
- Support for InfiniBand SR-IOV for libvirt virtualization
  - Allows for increased performance of InfiniBand interfaces direct to VMs.
- different_cells scheduler filter for Nova
  - Allows anti-affinity based on Cells
- Neutron QoS API
  - Ability to provide per-port Quality of service configuration

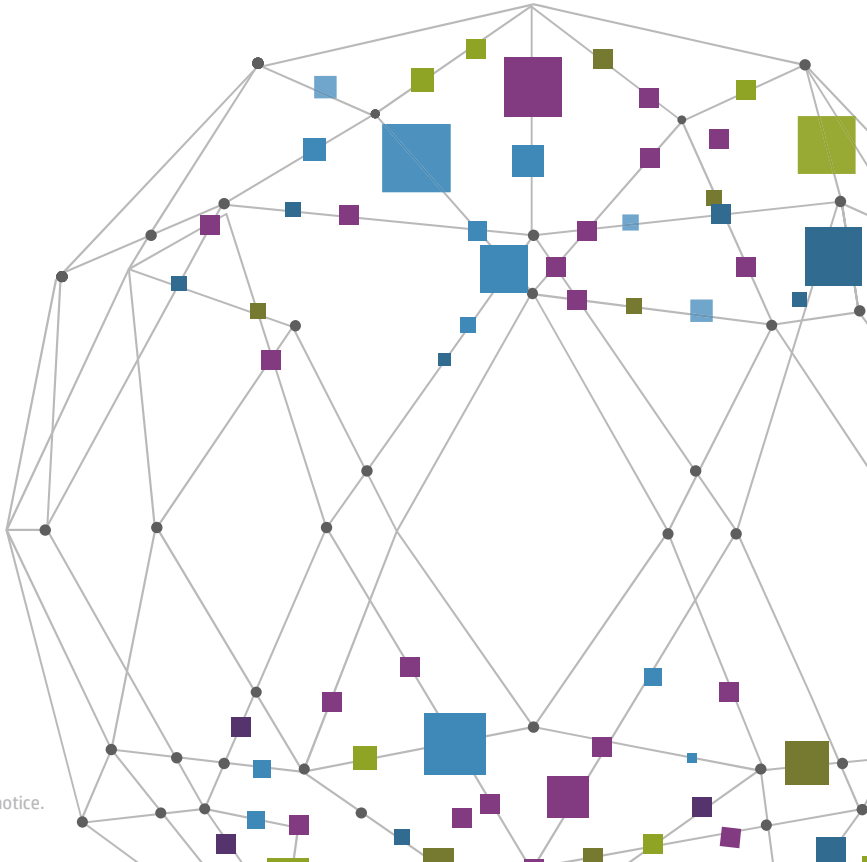
Storage

HP Helion



Ephemeral Storage

HP Helion



Ephemeral Storage

What is it?

Persists only as long as the VM exists

Usually located locally on the compute server

HPC Use Cases?

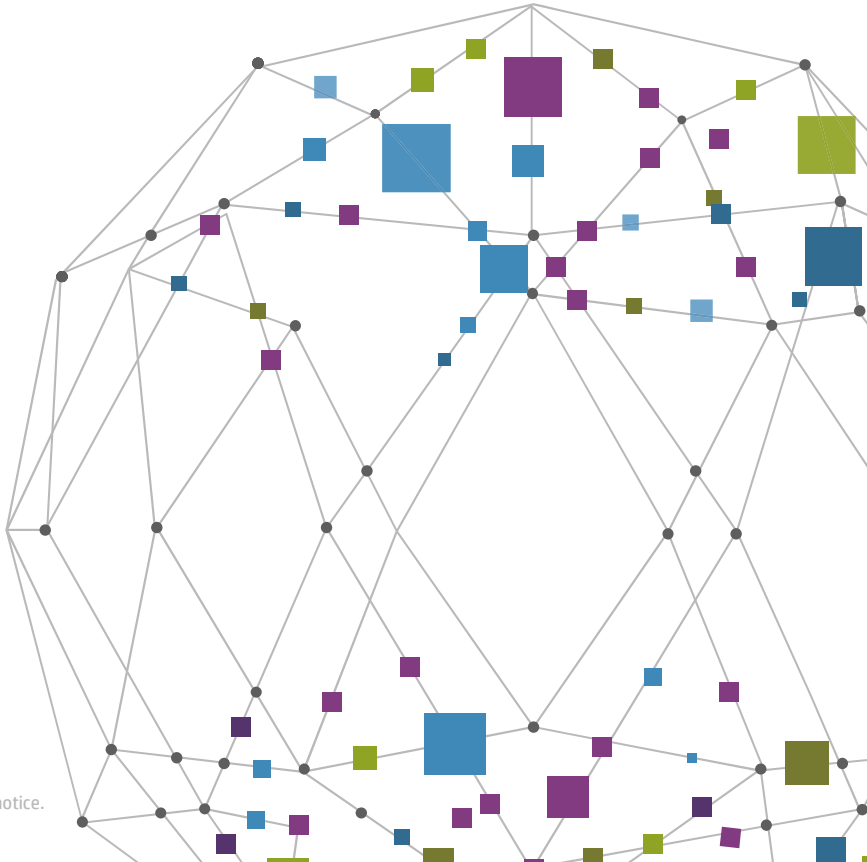
User scratch space

Work scratch space

Operating Environment

Block Storage

HP Helion



Block Storage

What is it?

Persistent, non-shared block storage

Can be provided by many sources, SAN Arrays, Local Disk etc.

HPC Use Cases?

Supporting Databases

High performance scratch space

Block Storage

OpenStack Project is CINDER

A Large Disk Array attached by Fibre Channel SAN to all of the compute nodes

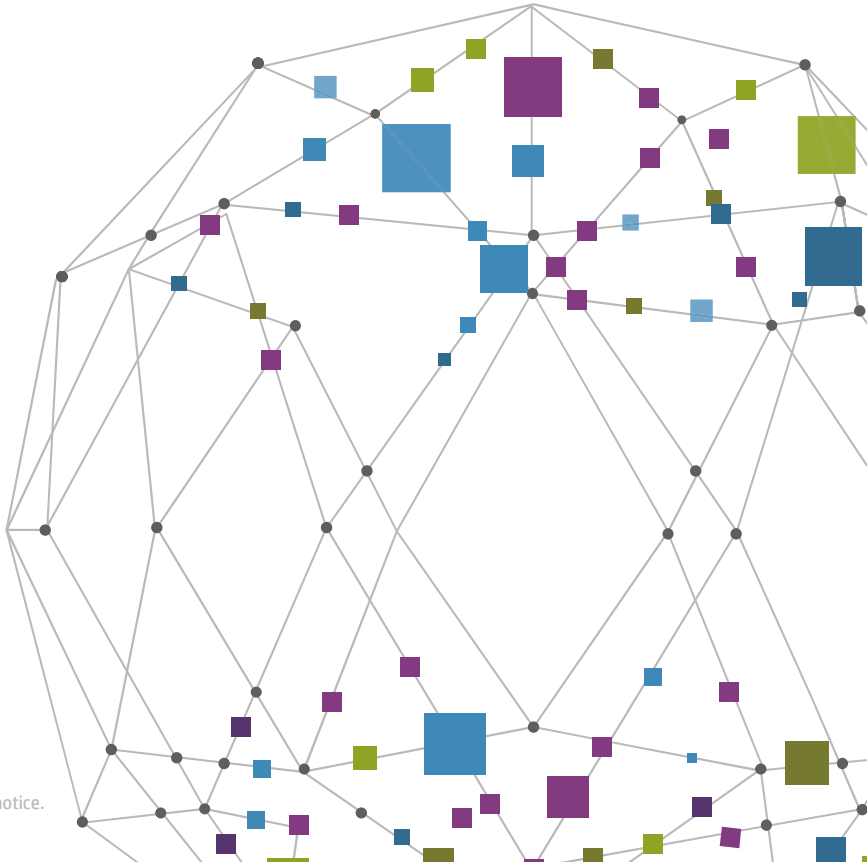
OpenStack uses Cinder drivers to create, mount and protect LUNs for the guests.

Guest is responsible for creating a file system on those LUNs

Not shared with other guests

Object Storage

HP Helion



Object Storage

What is it?

Persistent, scalable storage pools

Access using a REST based API

Not bound to an individual Guest

HPC Use Cases?

Centralised Data Lakes

Archives / Backups of source data

Object Storage

OpenStack Project is Swift

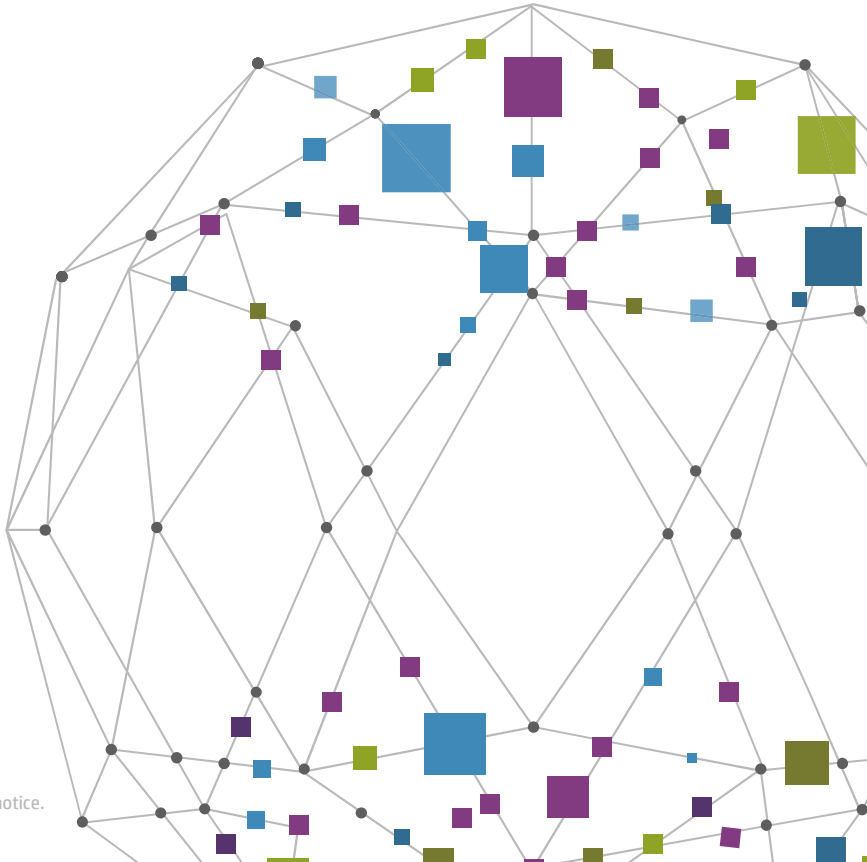
Usually uses large pools of local disk attached directly to the object servers

Uses metadata to index the data and locate object blocks from unique identifiers

Lots of plugins for the various analytics engines that are expanding the adoption of object as a centralized data lake

File Storage

HP Helion



File Storage

What is it?

Shared, persistent storage

Uses standard POSIX file system methods to access data

HPC Use Cases?

User Home Directories

Shared Project Data

Scale out file systems!

File Storage

OpenStack Project is Manila

Manage the creation of storage pools on the provider service

Create the shares and apply the correct permissions

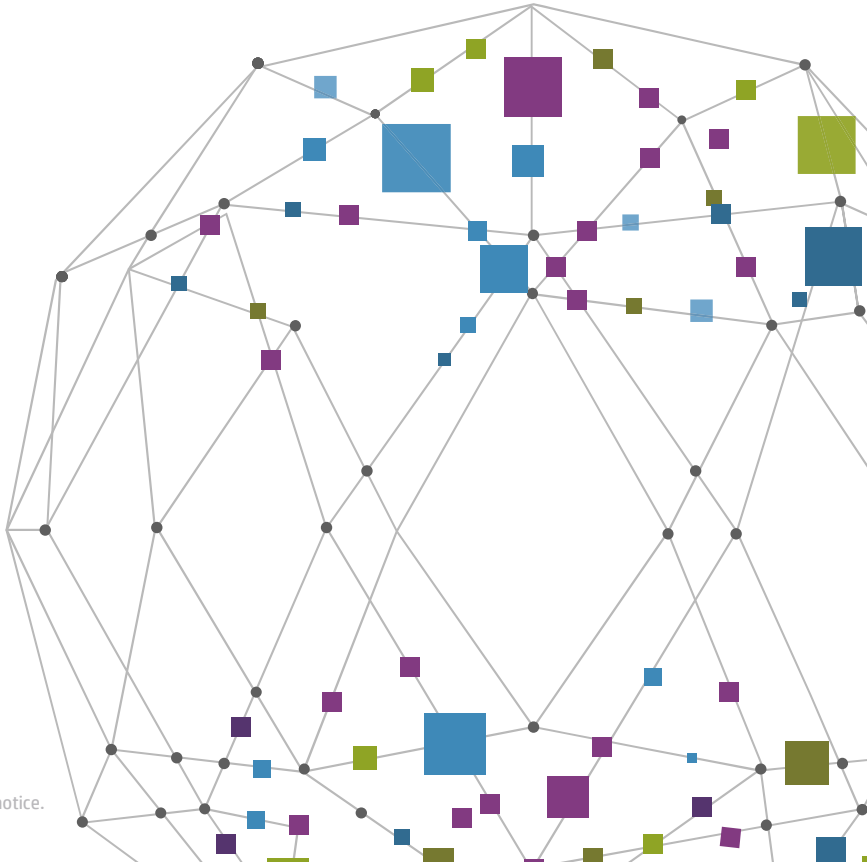
Mount those shares within the guests that need them

Can be NFS or CIFS based today

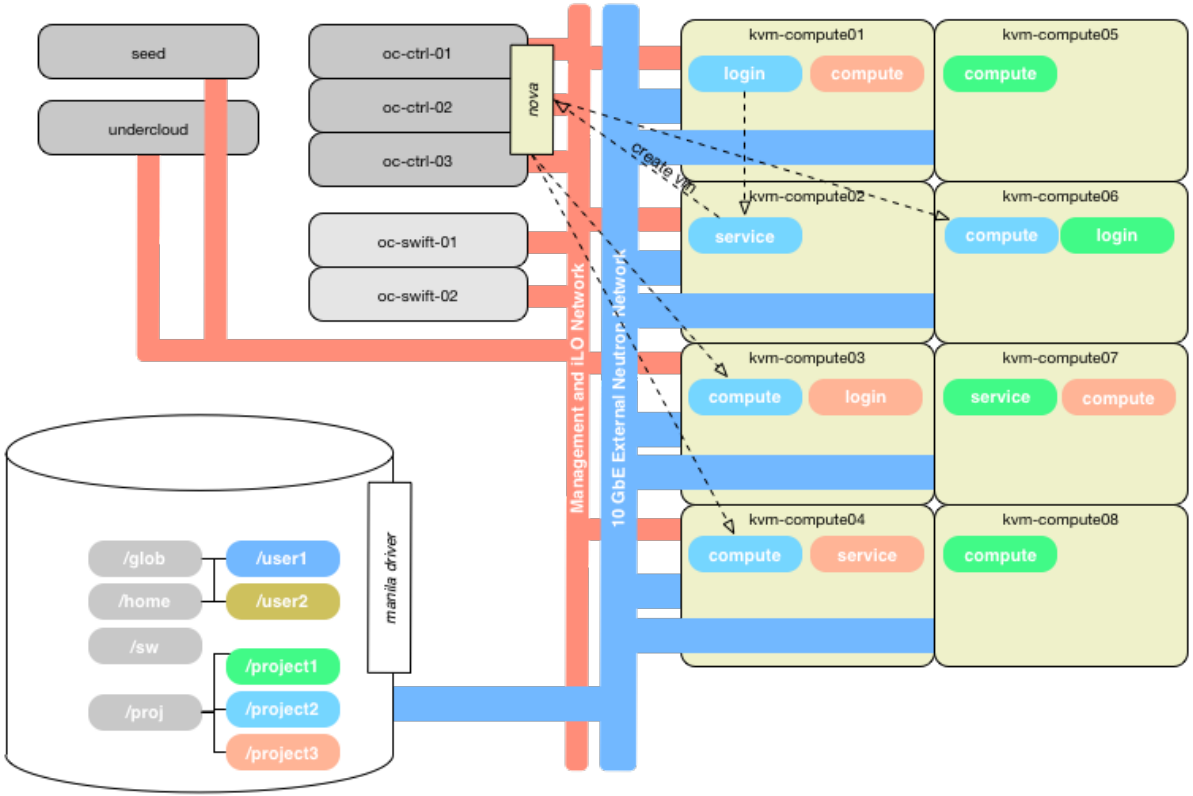
Plugin driven

Use Case - Lustre

HP Helion



Storage Use Case



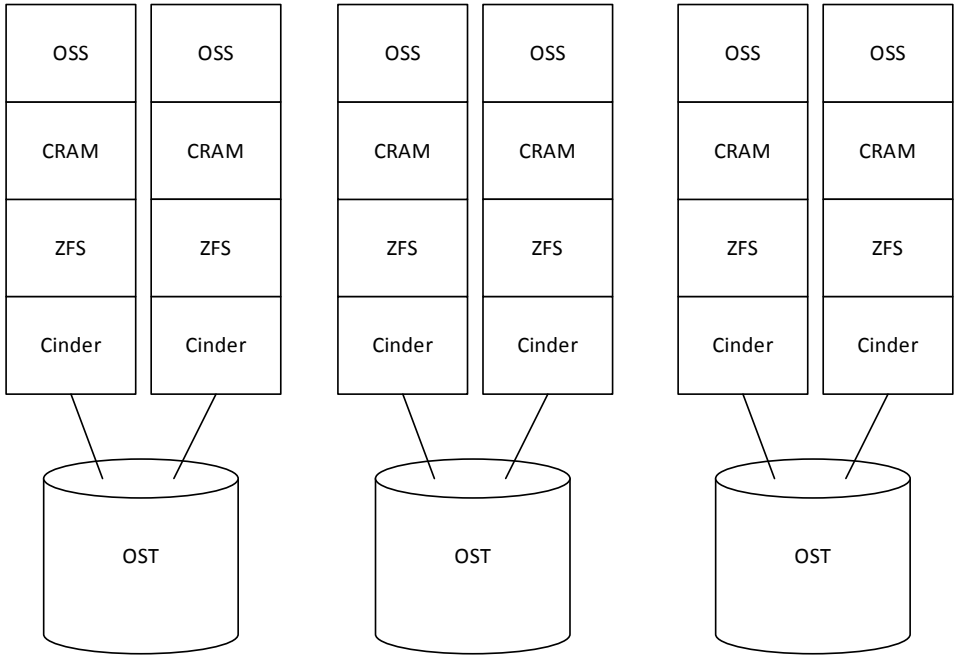
What is Lustre

HP Helion

- Massively Parallel Filesystem made up of key components…
  - MGS – Management Server
  - MGT – Management Target
  - MDS – Meta Data Server
  - MDT – Meta Data Target
  - OSS – Object Storage Server
  - OST – Object Storage Target
- 1 File can be spread over up to 2000 objects
- With ldiskfs, each of those each object can be up to 16 TB
- That's 31.25 PB (Yes PETA bytes) for a single file using ldiskfs
- Up to 4 Billion files per MDT
- Up to 4096 MDTs!

Layering of Services

Example for Genome Sequencing



Why ZFS?

HP Helion

- Lustre has limited data protection.
  - RAID 0
  - Protection from Physical Infrastructure
- Scale Out – Easy, Scale UP – Hard
- ZFS has healing, snapshots (not necessarily a good idea here) and scale up!
- ZFS Cache Pools for Meta-Data or even data sets, huge acceleration potential
- Scale…

Lustre and ZFS File Limits

	LDISKFS
Object Size	16 TB
Maximum File Size	21.25 PB
Max Files per MDT	4 Billion
Max MDTs	4096

Lustre and ZFS File Limits

HP Helion

	LDISKFS	ZFS
Object Size	16 TB	256 PB
Maximum File Size	21.25 PB	8 EB (2^63)
Max Files per MDT	4 Billion	4 Billion
Max MDTs	4096	4096

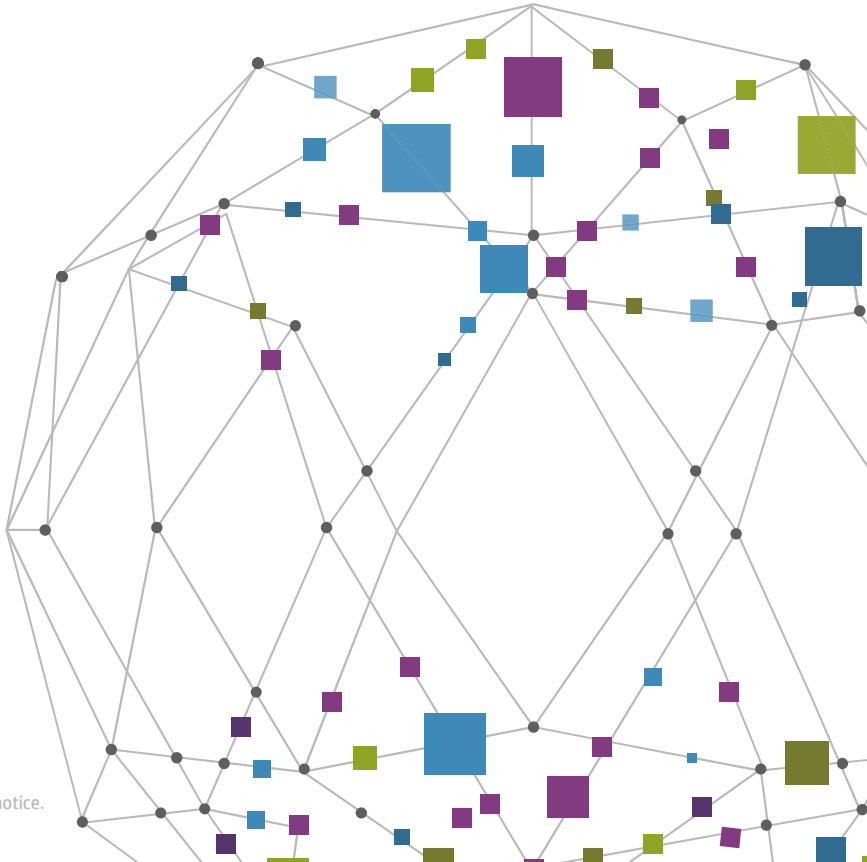
Work in progress – Lustre as a Service

HP Helion

- Lustre can be for high bandwidth and low latency
- Low latency challenging in virtual environment
- High Bandwidth, easier (not simple though)
- Use OpenStack tools to provision Lustre Components
- Build small scale, segregated clusters for multi-tenancy
- Export via NFS with Manila on private networks
- Include ZFS and Compression
- Use Cinder Block for the shared storage element

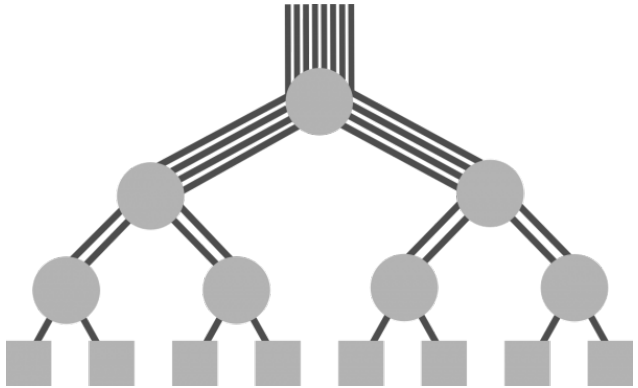
Networking

HP Helion



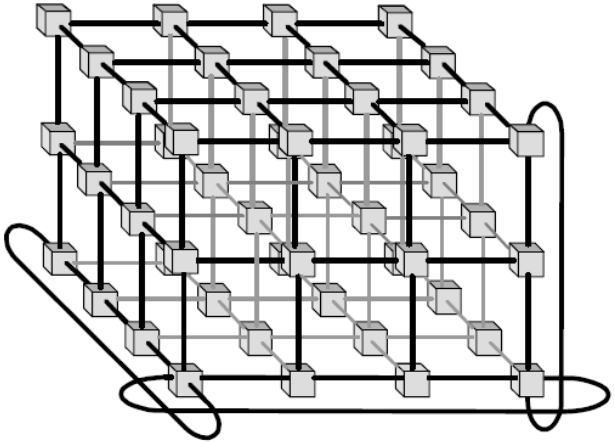
Fat Tree Network Topology

- Simplest and most generic is Fat Tree
- Execution intelligence to pass local workloads to neighbours to reduce RTT.
- Looks very similar to standard core / edge network topology but can be many more layers depending on number of compute nodes and upstream bandwidth requirements
- ISL count increase the further up the tree you are to allow for aggregated bandwidth increases.



3-Dimensional (wrap around) Torus

- Used in Low Latency Grid computing where RDMA is major factor
- Features in a lot of GPU compute grids to where RDMA is becoming more prominent
- Designed for nearest neighbour workloads
- Low cost vs. Fat Tree due to reduced switch requirements if using direct
- Means network switching can be brought closer to the CPU by using HTX technology.
- 2D needs 4 port cards, 3D needs 6 port cards! (EXTOLL)



Provider VLAN

HP Helion

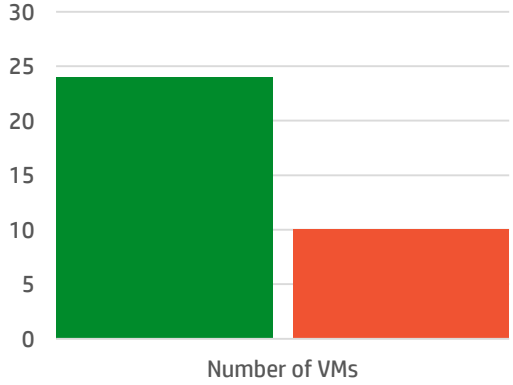
Network Considerations

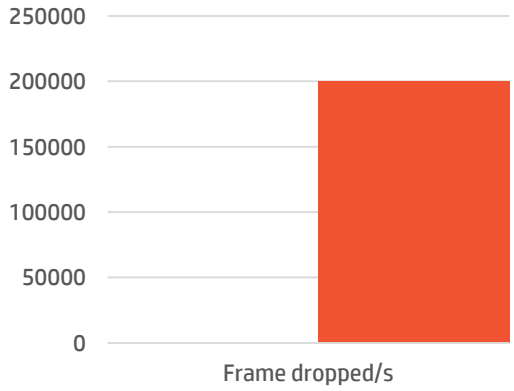
- 802.1q direct to compute host
- OVS 2.4 with user space virtIO support

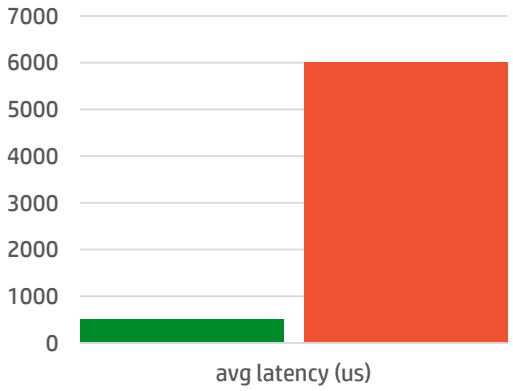
High Performance VM Considerations

- SR-IOV east-west vs. north-south
- MSI-X Issues with old VM kernels (Rx queue interrupts broken)

DPDK With and Without E-HOS

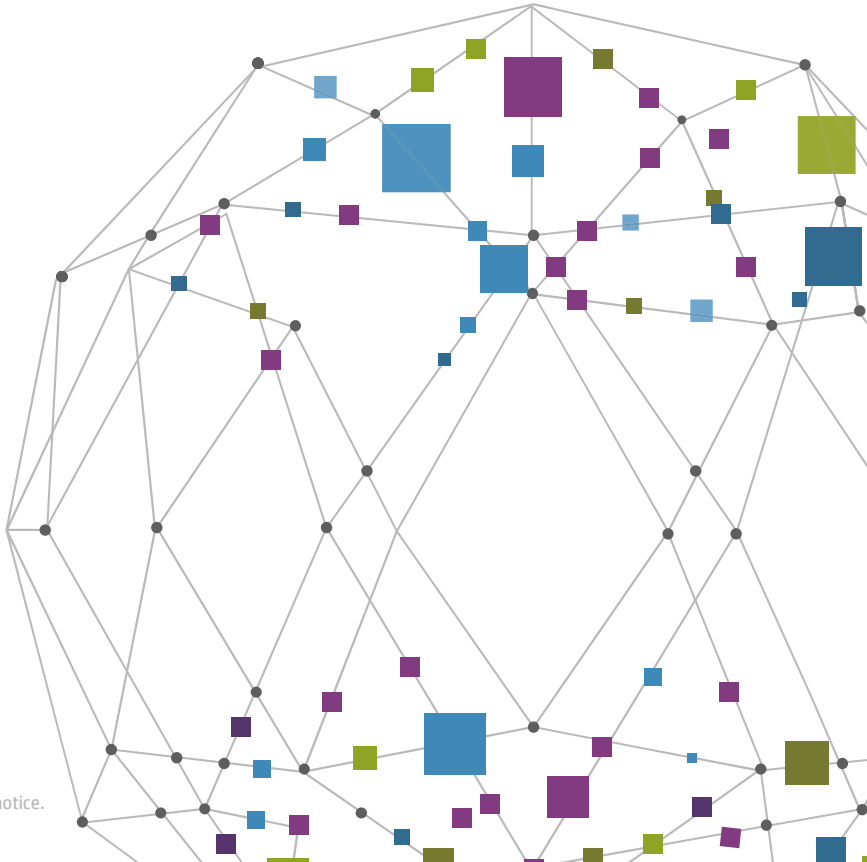






Summary

HP Helion



Summary

- Initial interest of HPC on OpenStack is being driven by tenancy requirements
- Managing flexible HPC resources has been tricky, OpenStack makes that easier for HPCaaS
- Many areas are needed to work well together for success, OpenStack Community beginning to address that as we have seen.
- HPC on OS is a reality and many are pushing the boundaries and committing back to the community.

HP Helion

Questions

どうもありがとう

Glyn Bowden - @bowdengl – gjb@hpe.com

Eric Lajoie – eric.lajoie@hpe.com