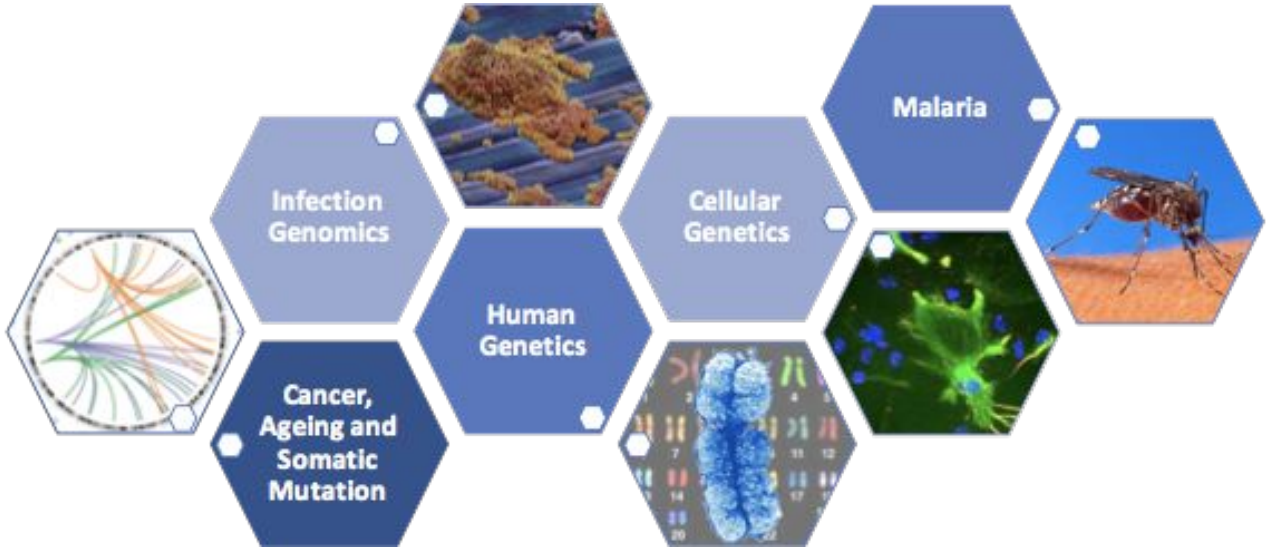
WELLCOME GENOME CAMPUS

wellcome trust sanger institute

Experiences in providing secure multi-tenant Lustre access to OpenStack

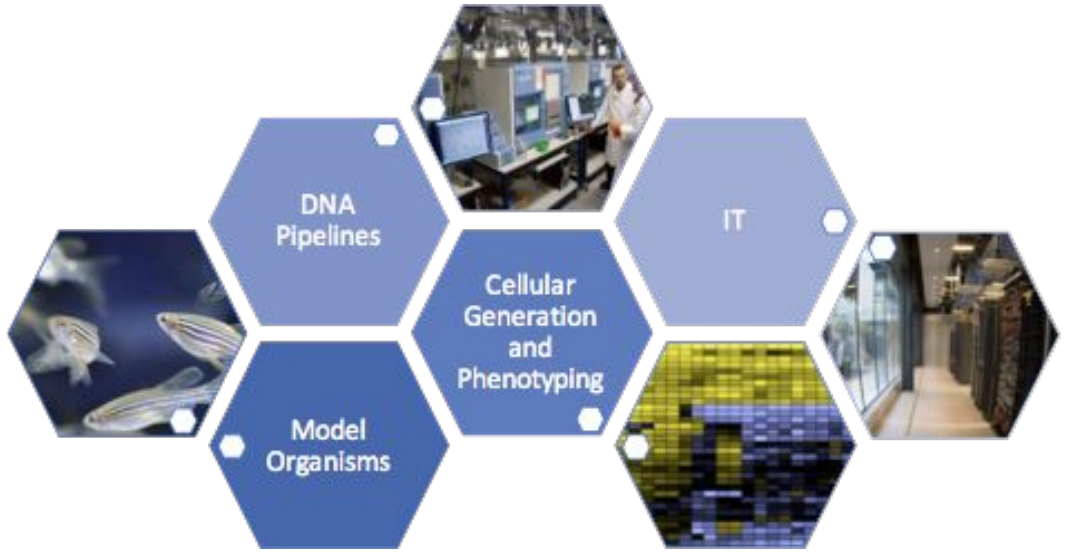
Dave Holland <dh3@sanger.ac.uk>
Wellcome Trust Sanger Institute

Sanger Science



Scientific Research Programmes

Core Facilities



WELLCOME GENOME CAMPUS

wellcome trust sanger institute

The Sanger Institute
Traditional HPC Environment

LSF 9

~10,000 cores in main compute farm

~10,000 cores across smaller project-specific farms

15PB Lustre high performance storage

Limited security - “isolation” is based on POSIX file permissions

Limited flexibility - no root access, incompatible software dependencies etc

Pipelines and stacks are complex, and scientific reproducibility is hard

HPC and Cloud computing are complementary

Traditional HPC

- Highest possible performance
- A mature and centrally managed compute platform
- High performance Lustre filesystems for data intensive analysis

Flexible Compute

- Full segregation of projects ensures data security
- Developers no longer tied to a single stack
- Reproducibility through containers / images and infrastructure-as-code

But there's a catch or two...

- Large number of traditional/legacy pipelines
  - They require a performant shared POSIX filesystem, while cloud workloads support object stores
- We do not always have the source code or expertise to migrate
- We need multi-gigabyte per second performance
- The tenant will have root
  - and could impersonate any user, but Lustre trusts the client's identity assertions, just like NFSv3
- The solution must be simple for the tenant and administrator

WELLCOME GENOME CAMPUS

Our OpenStack History

wellcome trust sanger institute

2015

- Training and experiments with RHOSP6 (Juno)
- December: pilot "beta" system on cobbled-together hardware

2016

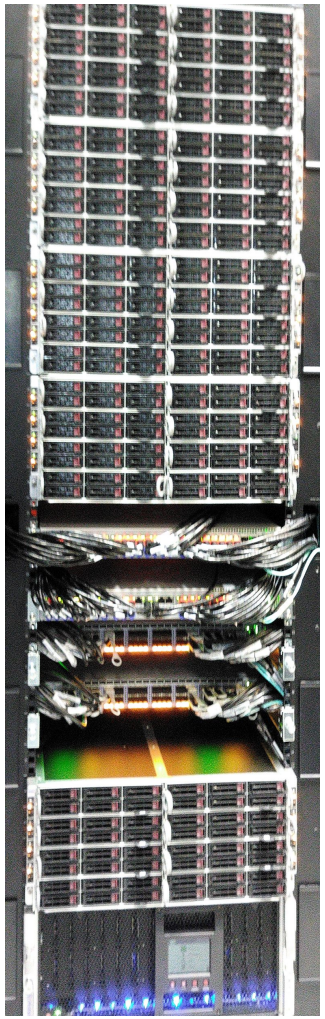
- Science-as-a-Service service for biotech spin-out customers
- July: Kilo "gamma" system for internal scientists. "Proper" Ceph storage.
- September: full scale hardware installation

2017

- January: Production “delta” system opened to early adopters (RHOSP8, Liberty)
- February: Sanger Flexible Compute Environment announced
- August: test deployments of next iteration “epsilon” (RHOSP10, Newton)

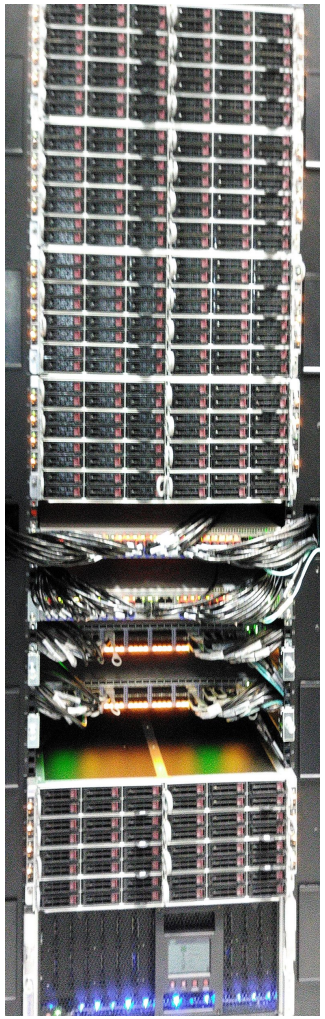
Production OpenStack

- 107 Compute nodes (Supermicro) each with:
  - 512GB of RAM, 2x 25Gbit/s network interfaces,
  - 1x 960GB local SSD, 2x Intel E52690v4 (14 cores @ 2.6Ghz)
- 6 Control nodes (Supermicro) allows 2 versions side by side
  - 256 GB RAM, 2x 100 GB/s network interfaces,
  - 1x 120 GB local SSD, 1x Intel P3600 NVMe (/var)
  - 2x Intel E52690v4 (14 cores @ 2.6Ghz)
- Total of 53 TB of RAM, 2996 cores, 5992 with hyperthreading
- RedHat OSP8 (“Liberty”) deployed with Triple-O



Ceph Storage Layer

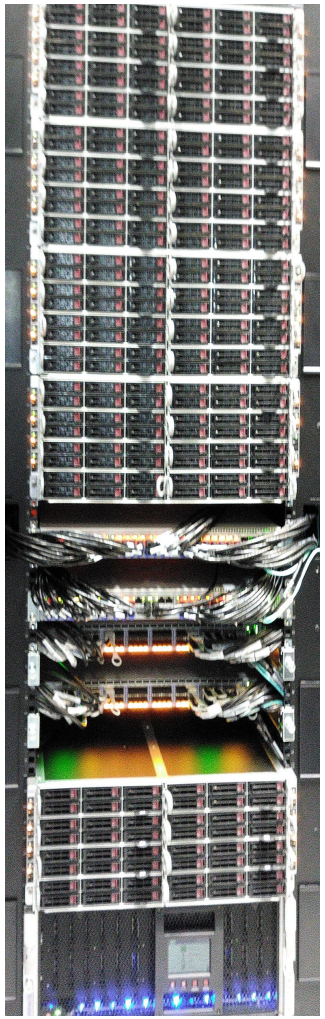
- ~~9~~ **42** Storage nodes (Supermicro) each with:
  - 512GB of RAM, 2x Intel E52690v4 (14 cores @ 2.6Ghz)
  - 2x 100Gbit/s network interfaces,
  - 60x 6TB SAS discs, 2 system SSD, 4TB of Intel P3600 NVMe used for journal.
- Ubuntu Xenial
- ~~3PB of disc space, 1PB usable~~. **Now 14PB, ~4.5PB usable!**
- Single node (1.3 GBytes/sec write, 200 MBytes/sec read)
- Ceph benchmarks imply 7 GBytes/second.
- Rebuild traffic of 20 GBytes/second.



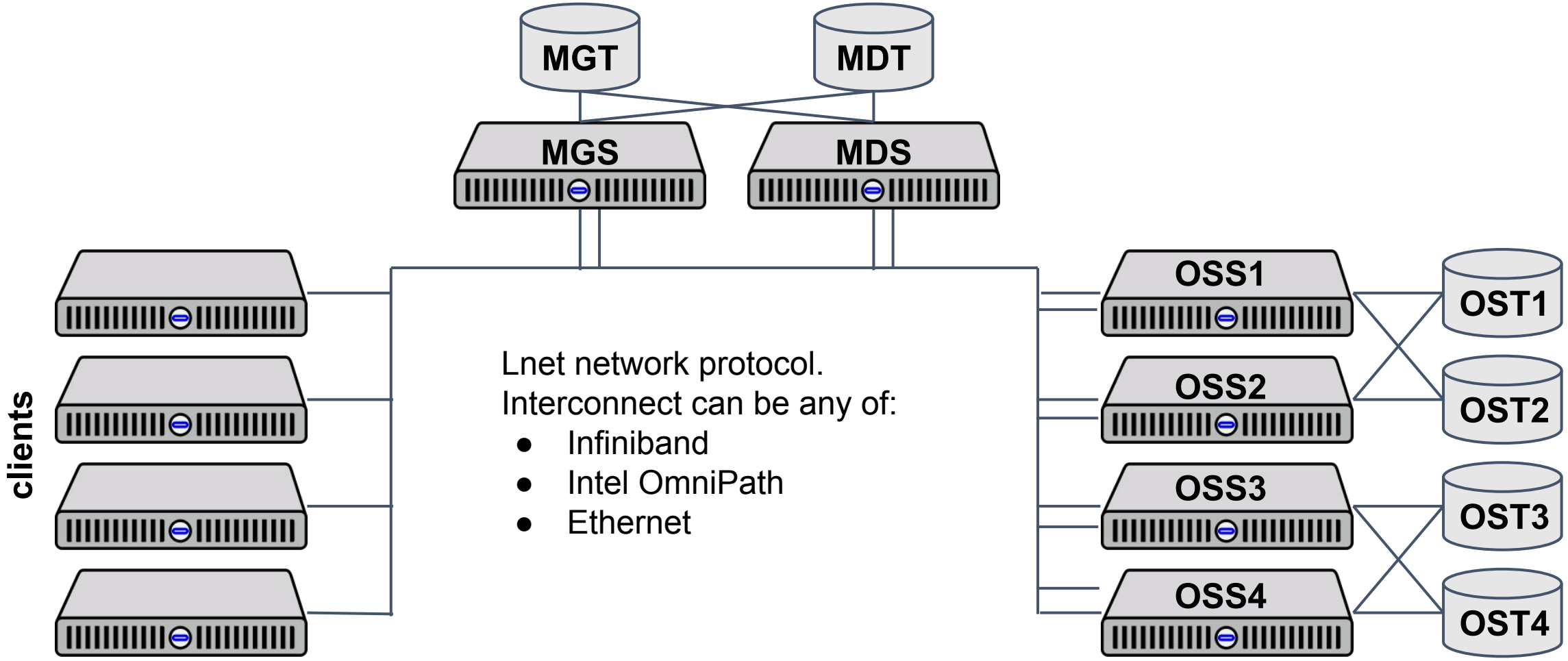
Networking

- 3 Racks of equipment, 24 KW load per rack.
- 10 Arista 7060CX-32S switches.
  - 1U, 32 * 100Gb/s -> 128 * 25Gb/s .
  - Hardware VXLAN support integrated with OpenStack*
  - Layer two traffic limited to rack, VXLAN used inter-rack.
  - Layer three between racks and interconnect to legacy systems.
  - All network switch software can be upgraded without disruption.
  - True Linux systems.
  - 400 Gb/s from racks to spine, 160 Gb/s from spine to legacy systems.

(* VXLAN in ml2 plugin not used in first iteration because of software issues)



Crash course: Lustre Architecture



Lnet network protocol.
Interconnect can be any of:

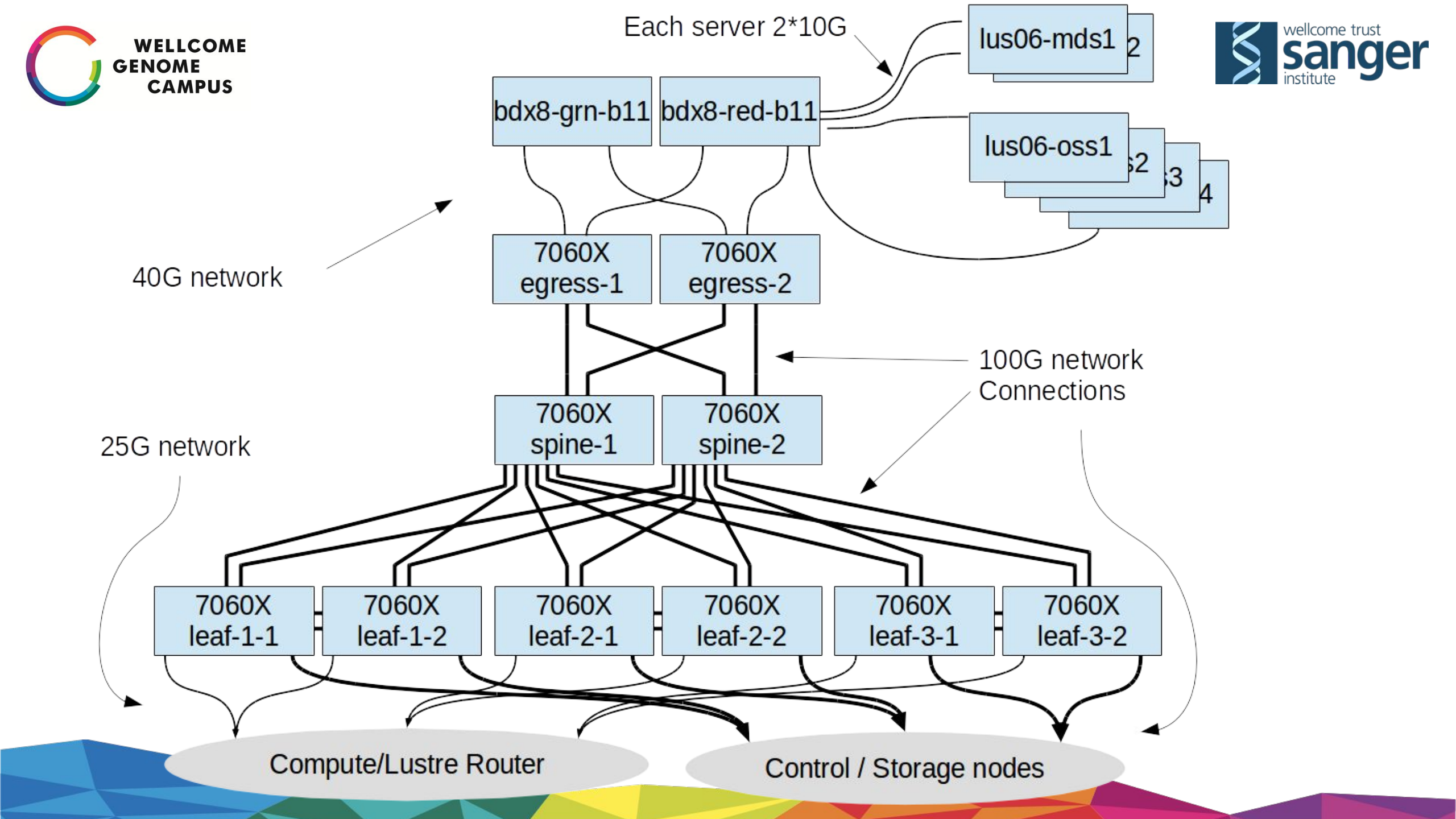
- Infiniband
- Intel OmniPath
- Ethernet

Lustre hardware

6+ year old hardware

- 4x Lustre object storage servers
  - Dual Intel E5620 @ 2.40GHz
  - 256GB RAM
  - Dual 10G network
  - lustre: 2.9.0.ddnsec2
  - https://jira.hpdd.intel.com/browse/LU-9289 (landed in 2.10)
- OSTs from DDN SFA-10k
  - 300x SATA, 7200rpm , 1TB spindles

We have seen this system reach 6 GByte/second in production


Each server 2*10G
lus06-mds1
bdx8-grn-b11
bdx8-red-b11
lus06-oss1
40G network
7060X
egress-1
7060X
egress-2
100G network
Connections
7060X
spine-1
7060X
spine-2
25G network
7060X
leaf-1-1
7060X
leaf-1-2
7060X
leaf-2-1
7060X
leaf-2-2
7060X
leaf-3-1
7060X
leaf-3-2
Compute/Lustre Router
Control / Storage nodes

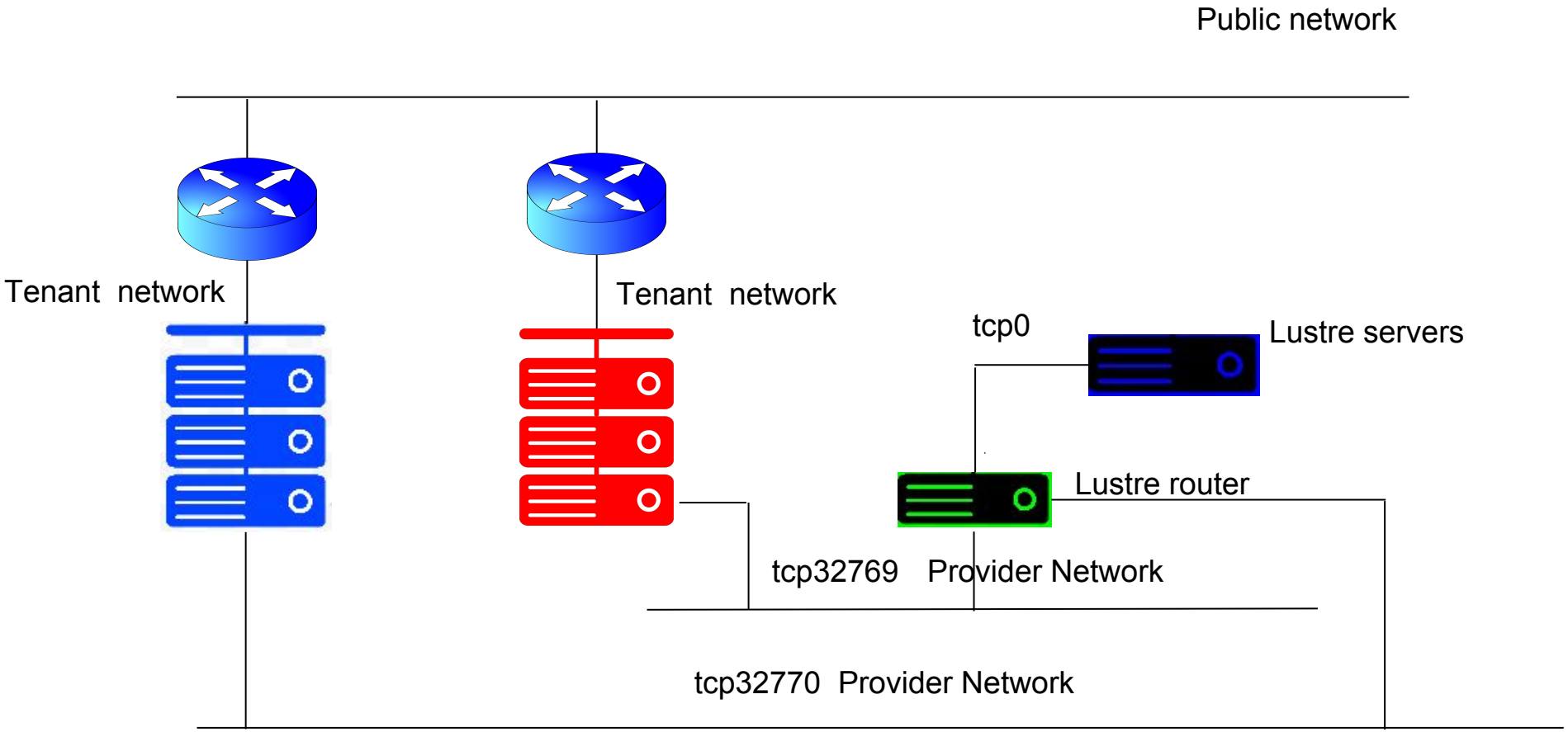
Lustre 2.9 features

- Each tenant's I/O can be squashed to their own unique UID/GID
- Each tenant is restricted to their own subdirectory of the Lustre filesystem

It might be possible to treat general access outside of OpenStack as a separate tenant with:

- a UID space reserved for a number of OpenStack tenants
- only a subdirectory exported for standard usage

Logical layout



Lustre server
Per-tenant UID mapping

Allows UIDs from a set of NIDs to be mapped to another set of UIDs

These commands are run on the MGS:

```
lctl nodemap_add ${TENANT_NAME}
lctl nodemap_modify --name ${TENANT_NAME} --property trusted --value 0
lctl nodemap_modify --name ${TENANT_NAME} --property admin --value 0
lctl nodemap_modify --name ${TENANT_NAME} --property squash_uid --value ${TENANT_UID}
lctl nodemap_modify --name ${TENANT_NAME} --property squash_gid --value ${TENANT_UID}
lctl nodemap_add_idmap --name ${TENANT_NAME} --idtype uid --idmap 1000:${TENANT_UID}
```

Lustre server: Per-tenant subtree restriction

Constrains client access to a subdirectory of a filesystem.

```
mkdir /lustre/secure/${TENANT_NAME}
chown ${TENANT_NAME}  /lustre/secure/${TENANT_NAME}
```

Set the subtree root directory for the tenant:

```
lctl set_param -P nodemap.${TENANT_NAME}.fileset=/${TENANT_NAME}
```

Lustre server: Map nodemap to network

Add the tenant network range to the Lustre nodemap

```
lctl nodemap_add_range --name ${TENANT_NAME} --range \
[0-255].[0-255].[0-255].[0-255]@tcp${TENANT_UID}
```

And this command adds a route via a Lustre network router. This is run on all MDS and OSS (or the route added to `/etc/modprobe.d/lustre.conf`)

```
lnetctl route add --net tcp${TENANT_UID} --gateway ${LUSTRE_ROUTER_IP}@tcp
```

In the same way a similar command is needed on each client using TCP

OpenStack: Network configuration

```
neutron net-create LNet-1 --shared --provider:network_type vlan \
--provider:physical_network datacentre --provider:segmentation_id \
${TENANT_PROVIDER_VLAN_ID}

neutron subnet-create  --enable-dhcp --dns-nameserver 172.18.255.1 --no-gateway \
--name LNet-subnet-1 --allocation-pool start=172.27.202.17,end=172.27.203.240 \
172.27.202.0/23 ${NETWORK_UUID}

openstack role create LNet-1_ok
```

For each tenant user that needs to create instances attached to this Lustre network:

```
openstack role add --project ${TENANT_UUID} --user ${USER_UUID} ${ROLE_ID}
```

OpenStack policy

Simplify automation by minimial change to `/etc/neutron/policy.json`

```
"get_network": "rule:get_network_local"
```

`/etc/neutron/policy.d/get_networks_local.json` then defines the new rule:

```
{
 "get_network_local": "rule:admin_or_owner or rule:external or
 rule:context_is_advsvc or rule:show_providers or ( not
 rule:provider_networks and rule:shared )"
}
```

OpenStack policy

`/etc/neutron/policy.d/provider.json` is used to define networks and their mapping to roles, and allow access to the provider network.

```
{
    "net_LNet-1": "field:networks:id=d18f2aca-163b-4fc7-a493-237e383c1aa9",
  "show_LNet-1": "rule:net_LNet-1 and role:LNet-1_ok",
  "net_LNet-2": "field:networks:id=169b54c9-4292-478b-ac72-272725a26263",
  "show_LNet-2": "rule:net_LNet-2 and role:LNet-2_ok",
  "provider_networks": "rule:net_LNet-1 or rule:net_LNet-2",
  "show_providers":    "rule:show_LNet-1 or rule:show_LNet-2"
}
```

Restart Neutron - can be disruptive!

Physical router configuration

- Repurposed Nova compute node
  - RedHat 7.3
  - Lustre 2.9.0.ddnsec2
  - Mellanox ConnectX-4 (2*25GbE)
  - Dual Intel E5-2690 v4 @ 2.60GHz
  - 512 GB RAM

Connected in a single rack so packets from other racks will have to transverse the spine. No changes from default settings.

Client virtual machines

- 2 CPU
- 4 GB RAM
- CentOS Linux release 7.3.1611 (Core)
- Lustre: 2.9.0.ddnsec2
- Two NICs
  - Tenant network
  - Tenant-specific Lustre provider network

Testing procedure - vdbench

http://bit.ly/2rjRuPP The Oracle download page (version 5.04.06)

Creates a large pool of files on which tests are later run.

Sequential and Random IO, block sizes of 4k,64k,512k,1M,4M,16M.

Each test section is run for 5 minutes.

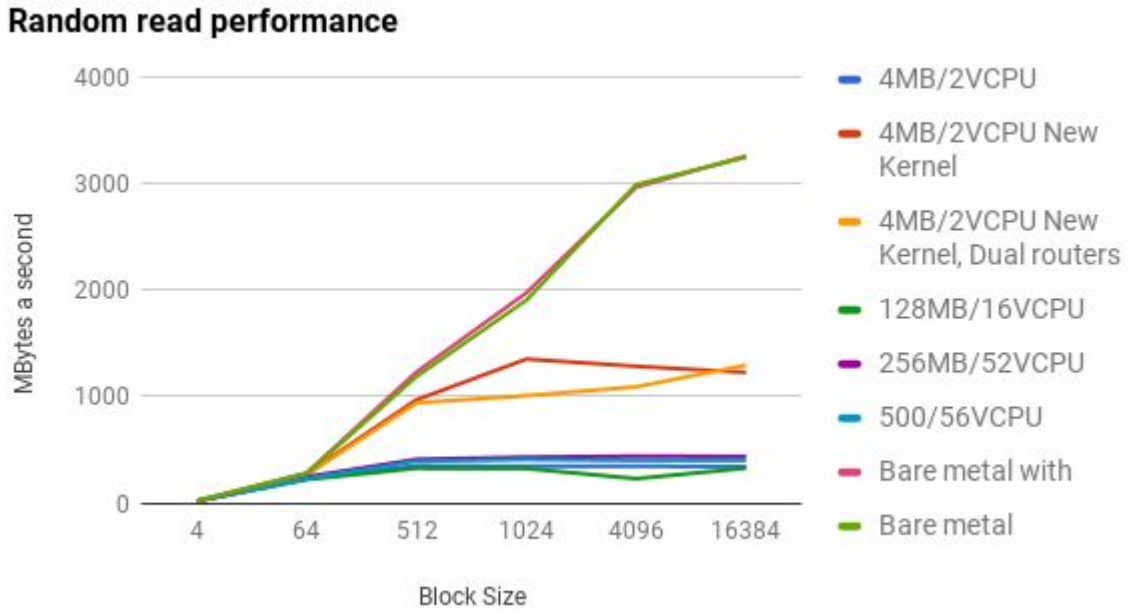
Threads are used to increase performance.

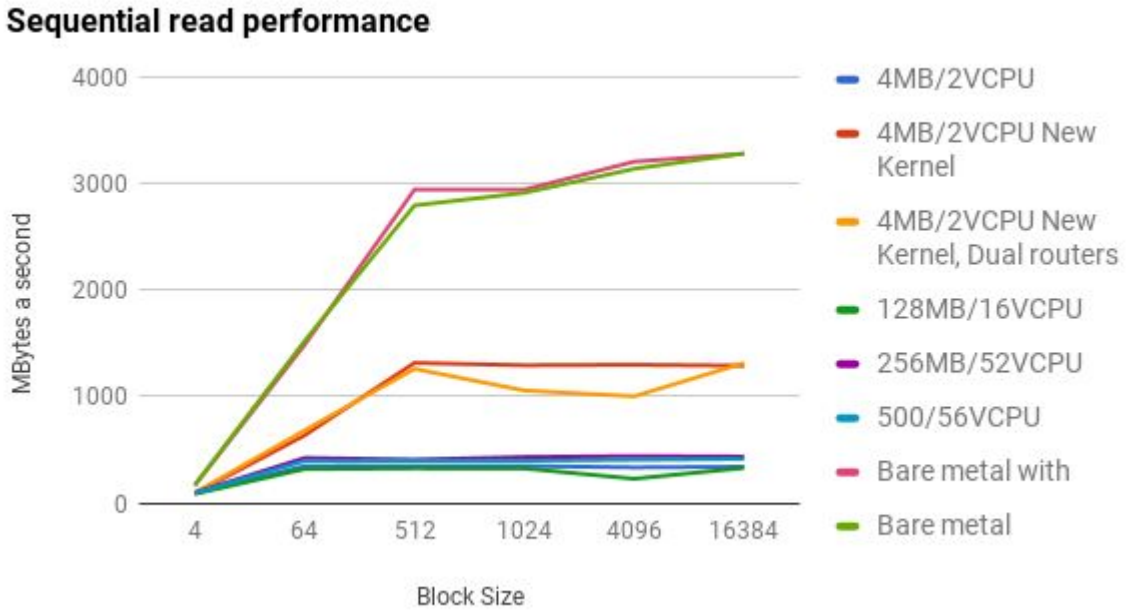
No performance tuning attempted.

Single client read performance

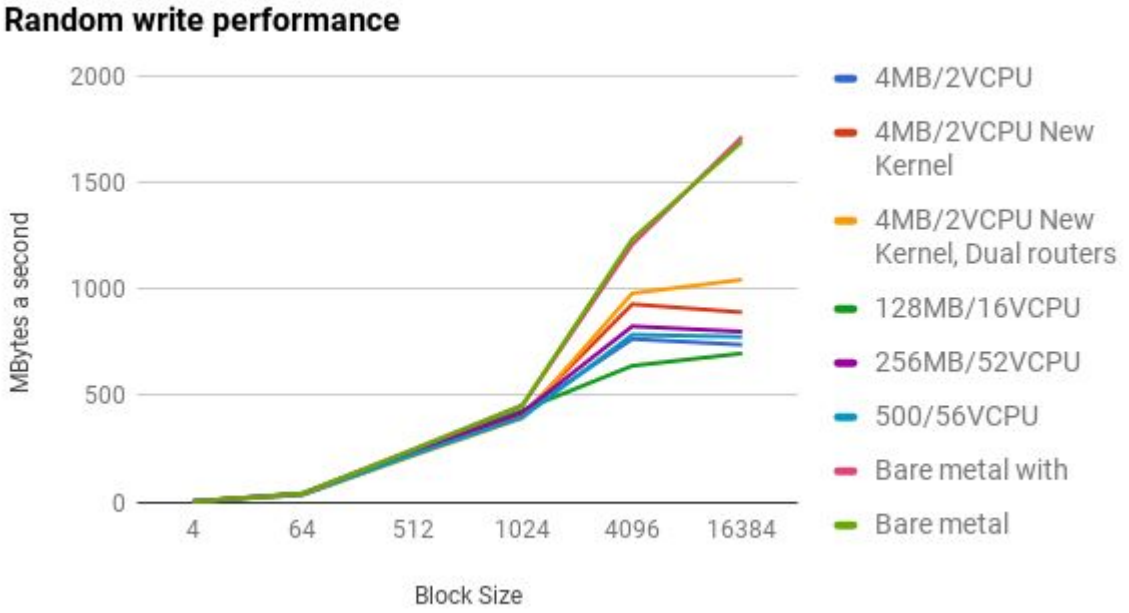
Filesets and uid mapping have no effect

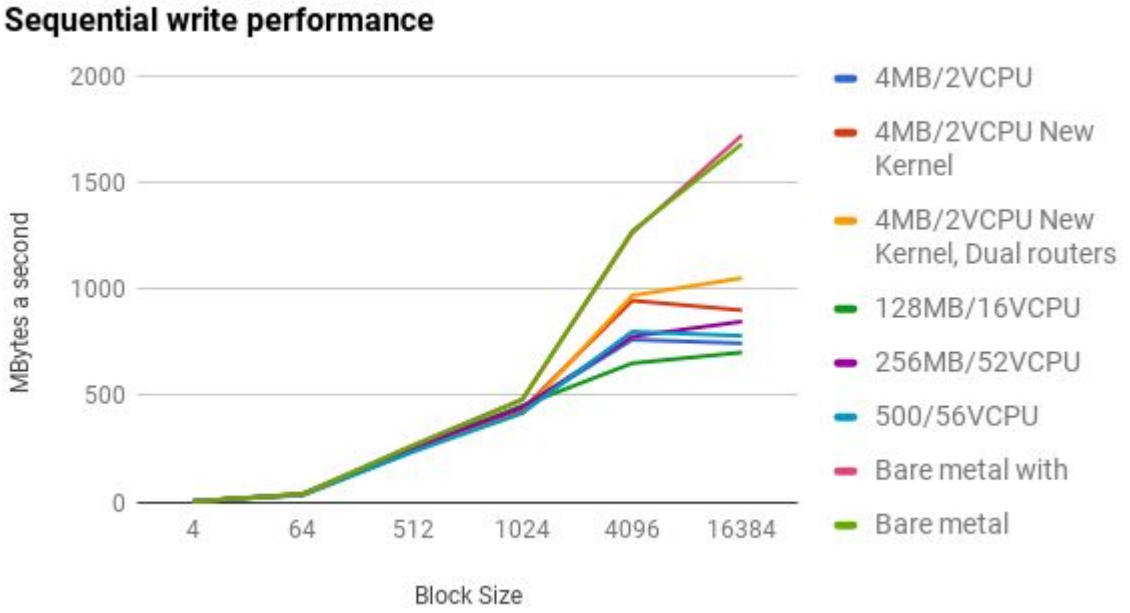
Instance size has little effect





Single client write performance





Single client performance

Filesets and UID mapping overhead is insignificant.

Read performance (Virtual machine, old kernel)≅ 350 MB/s

Write performance (Virtual machine, old kernel)≅ 750 MB/s

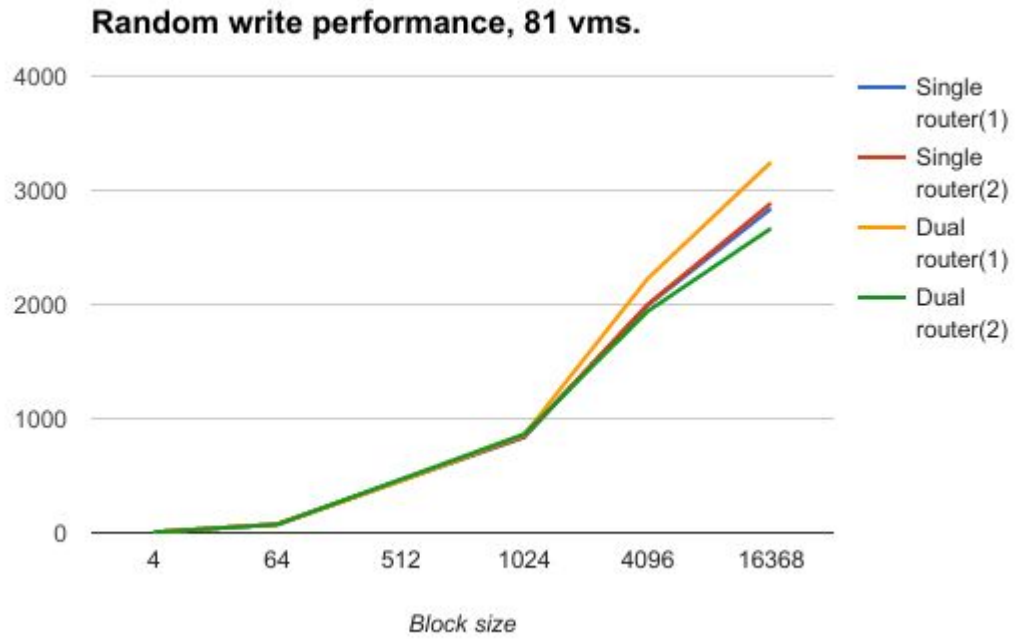
Read performance (Virtual machine, new kernel)≅ 1300 MB/s

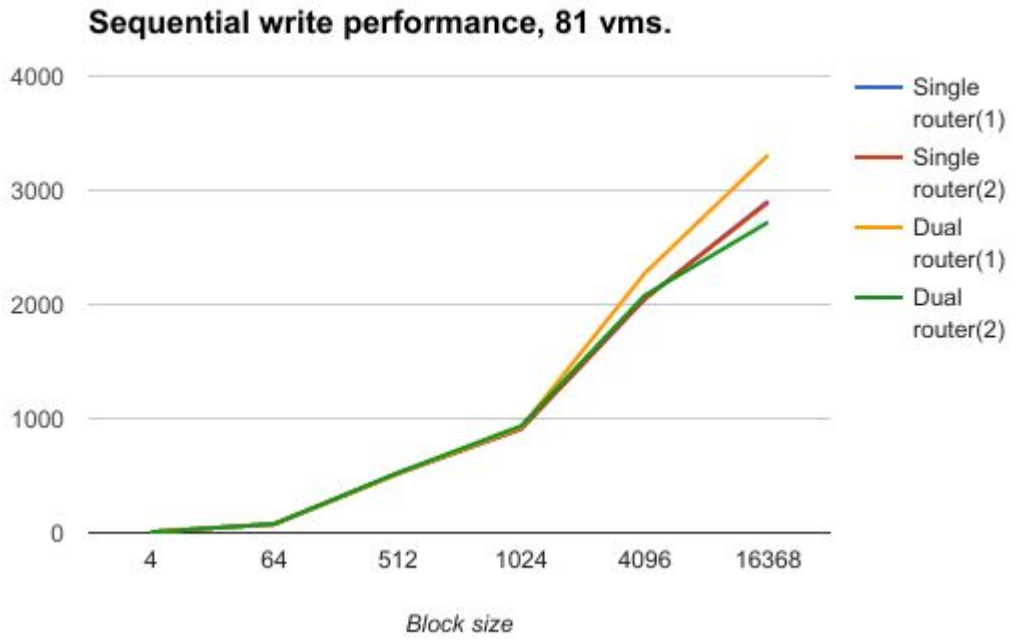
Write performance (Virtual machine, new kernel)≅ 950 MB/s

Read performance (Physical machine)≅ 3200 MB/s

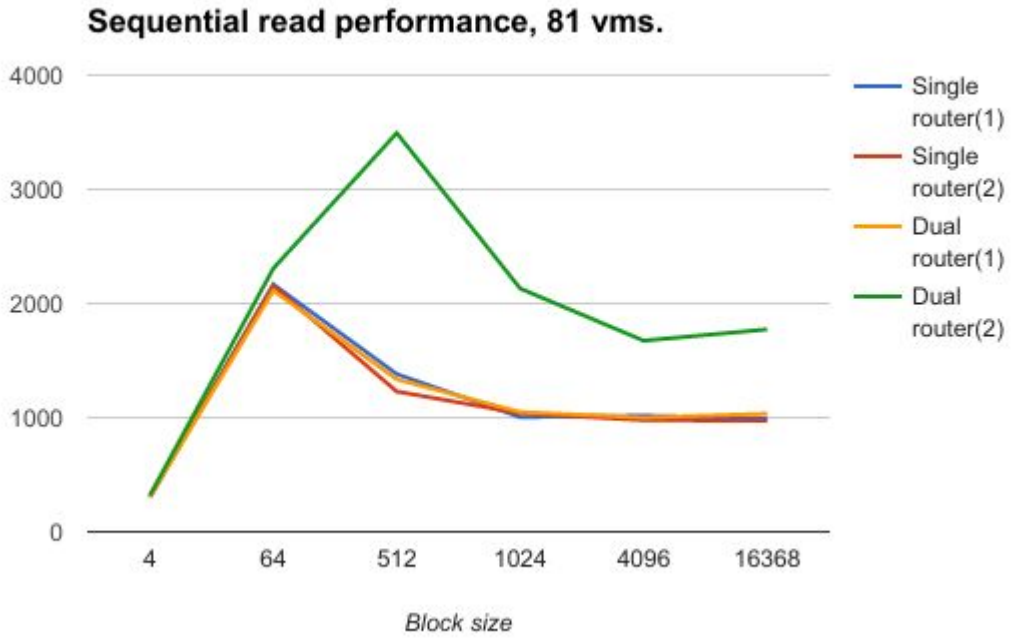
Write performance (Physical machine)≅ 1700 MB/s

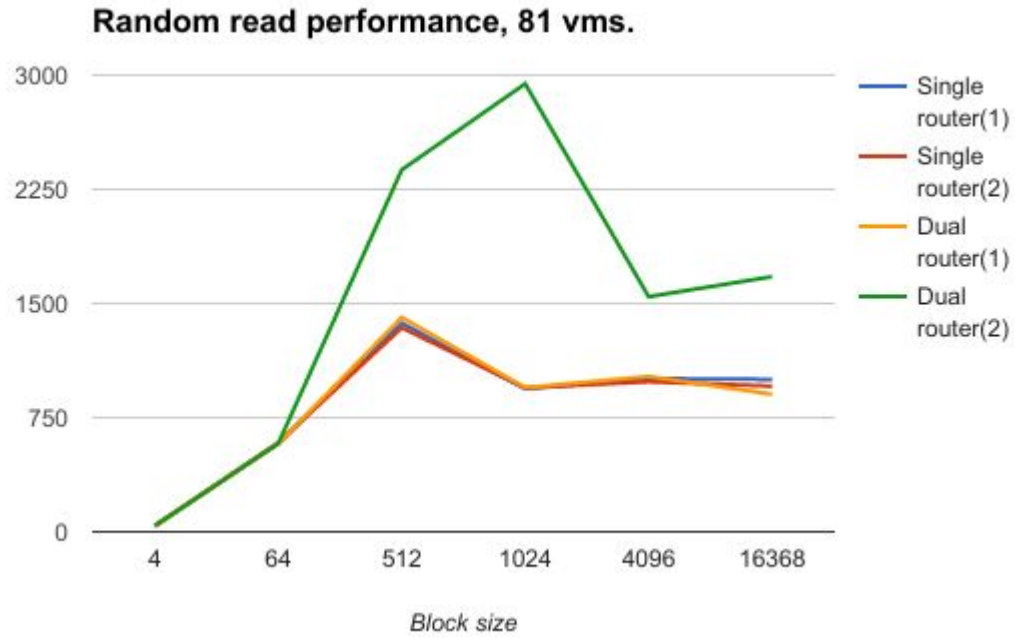
Multiple VMs, aggregate write performance, metal LNet routers





Multiple VMs, aggregate read performance, metal LNet routers

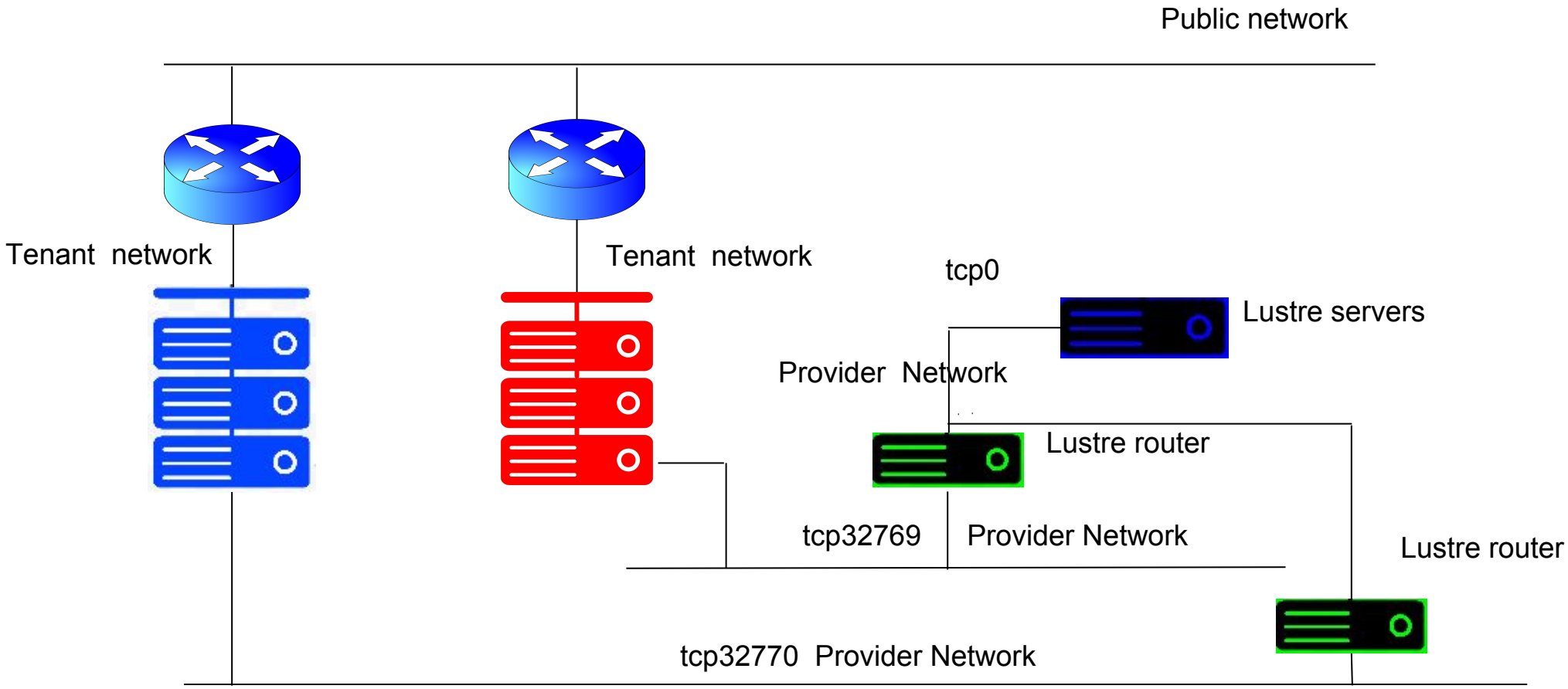




Virtualised Lustre routers

- We could see that bare metal Lustre routers gave acceptable performance
- We wanted to know if we could virtualise these routers
  - Each tenant could have their own set of virtual routers
  - Fault isolation
  - Ease of provisioning routers
  - No additional cost
- Increases east-west traffic, but that’s OK.

Logical layout



Improved security

As each tenant has its own set of Lustre routers:

- The traffic to a different tenant does not go to a shared router
- A Lustre router could be compromised without directly compromising another tenant's data - the filesystem servers will not route data for @tcp1 to the router @tcp2
- Either a second Lustre router or the Lustre servers would need to be compromised to intercept or reroute the data

Port security...

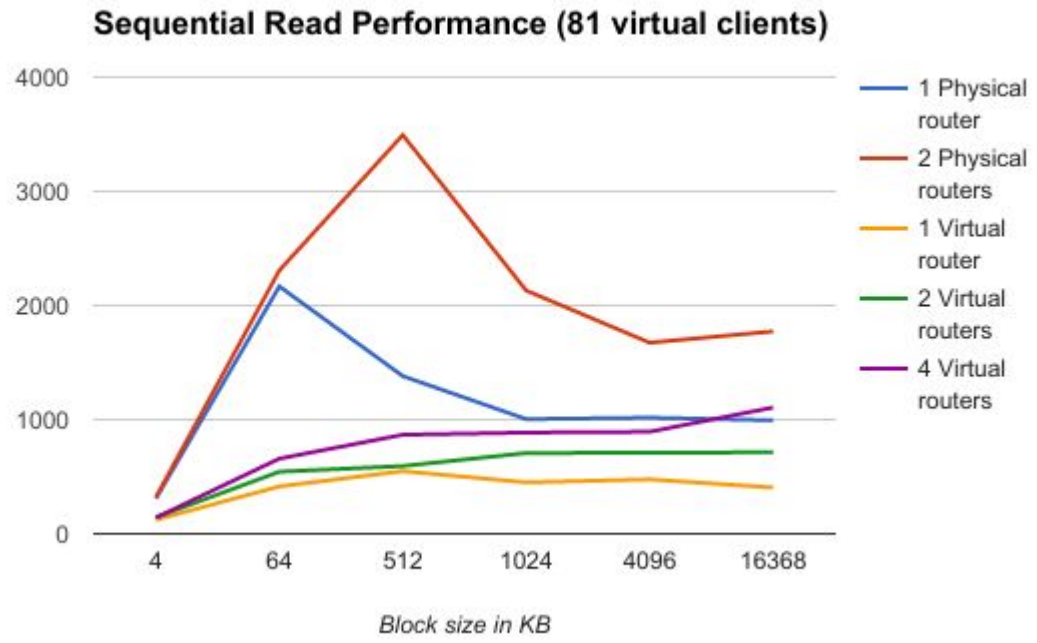
The routed Lustre provider network (tcp32769 etc) required that port security was disabled on the virtual Lustre router ports.

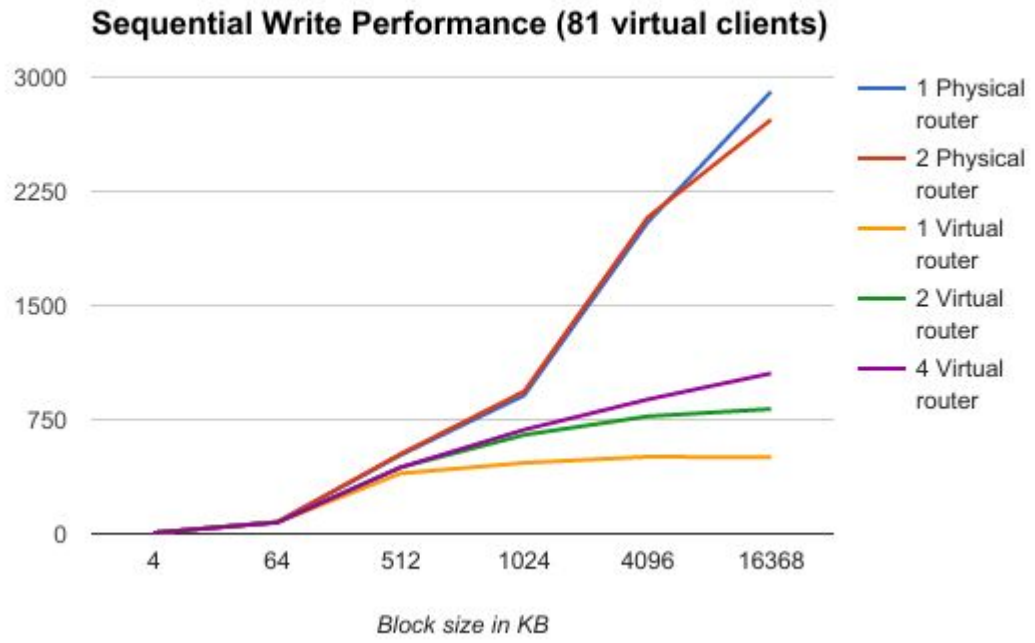
```
neutron port-list | grep 172.27.70.36 | awk '{print $2}'

08a1808a-fe4a-463c-b755-397aedd0b36c

neutron port-update --no-security-groups 08a1808a-fe4a-463c-b755-397aedd0b36c

neutron port-update 08a1808a-fe4a-463c-b755-397aedd0b36c \
 --port-security-enabled=False
```

http://kimizhang.com/neutron-ml2-port-security/

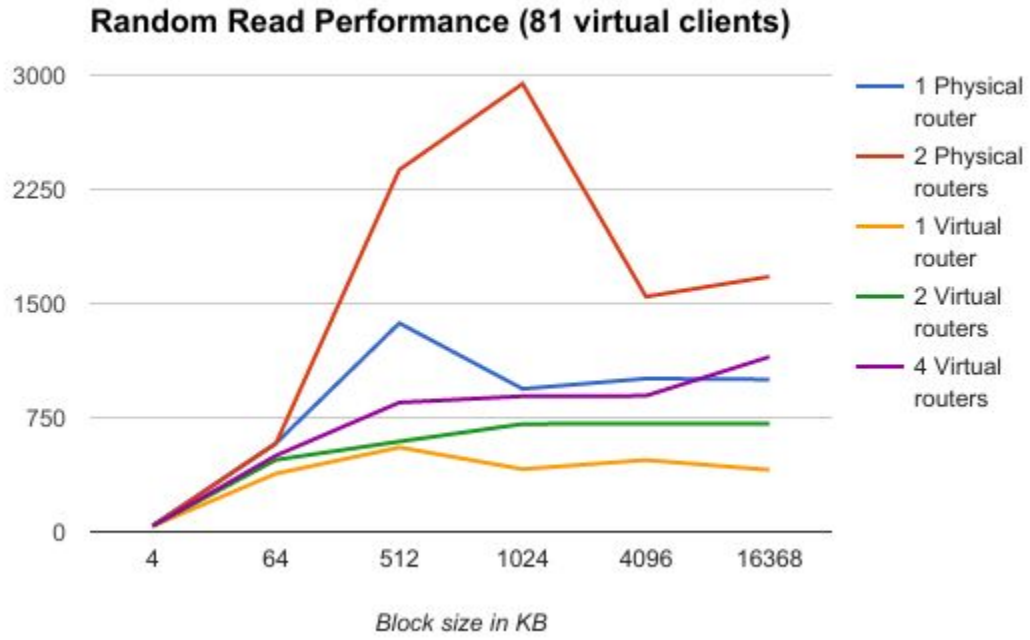
This is due to a race condition in the Liberty release, which can be avoided by adding the Lustre provider network interface when the instance is created.

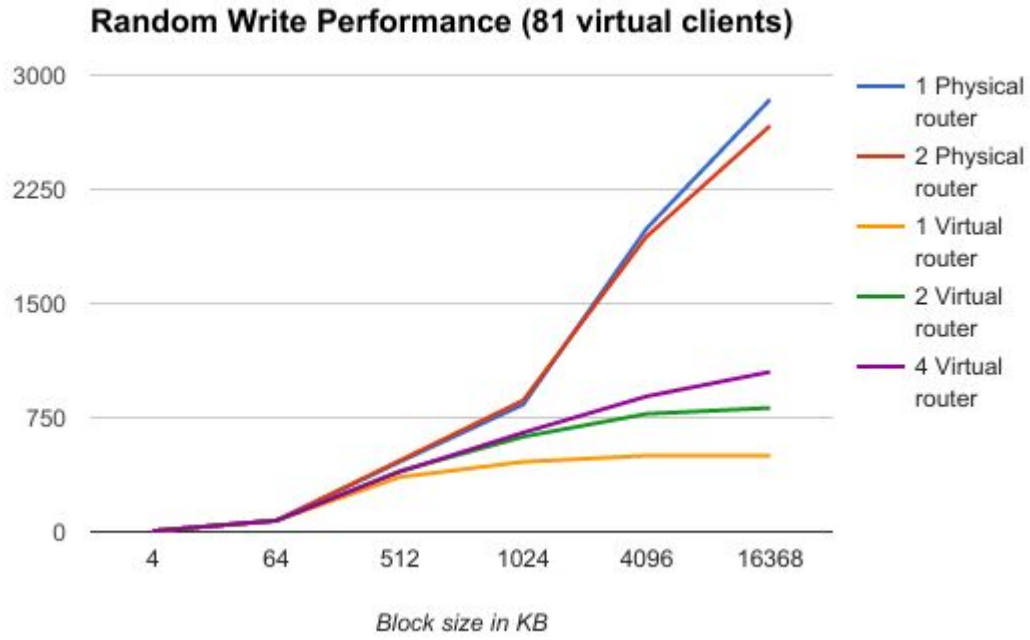
Virtual Lnet routers: Sequential performance





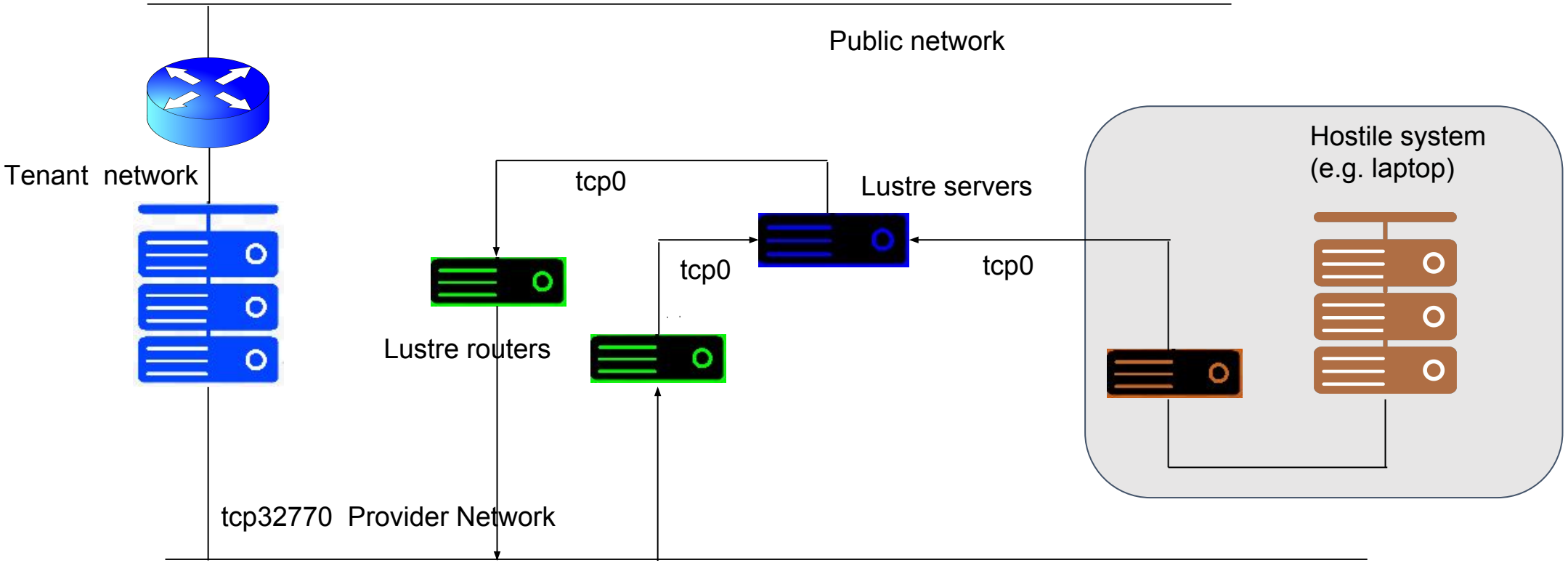
Virtual Lnet routers: Random Performance





Asymmetric routing?

http://tldp.org/HOWTO/Adv-Routing-HOWTO/lartc.kernel.rpf.html



Conclusions

- Follow our activities on http://hpc-news.sanger.ac.uk
- Isolated POSIX islands can be deployed with Lustre 2.9+
- Performance is acceptable
- Lustre routers require little CPU and memory
- Physical routers work and can give good locality for network usage
- Virtual routers work, can scale and give additional security benefits

- Next steps:
  - Improve configuration automation
  - Check port security issue is fixed in Newton
  - Improve network performance (MTU, OpenVSwitch etc).

Acknowledgements

DDN: Sébastien Buisson, Thomas Favre-Bulle, Richard Mansfield, James Coomer
Sanger Informatics Systems Group: Pete Clapham, **James Beal**, John Constable, Helen Cousins, Brett Hartley, **Dave Holland**, **Jon Nicholson**, **Matthew Vernon**