Science & Technology Facilities Council

Ceph vs Local Storage for Virtual Machine

26th March 2015

HEPiX Spring 2015, Oxford

Alexander Dibbo

George Ryall, Ian Collier, Andrew Lahiff, Frazer Barnsley

Background

- As presented earlier, we have developed a cloud based on OpenNebula and backed by Ceph storage
- We have
  - 28 hypervisors (32 threads, 128GB RAM, 2TB disk, 10GB networking)
  - 30 storage nodes (8 threads, 8GB RAM, 8 x 4TB disks, 10GB front and backend networks)
  - OpenNebula Headnode is virtual
  - Ceph monitors are virtual

What Are We Testing?

- The performance of Virtual Machines on Local Storage (hypervisor local disk) vs Ceph RBD storage
- How quick machines are to deploy and to become useable in different configurations
- How quickly management tasks can be performed (i.e. live migration)

What are we trying to find out?

- The performance characteristics of virtual machines on each type of storage
- How agile we can be with machines on each type of storage

Test Setup

- Virtual Machine – Our SL6 image, 1 CPU, 4GB of RAM, 10GB OS Disk and a 50GB Sparse Disk for Data
- 4 Different Configurations
  - OS on Ceph, Data on Ceph
  - OS Local, Data Local,
  - OS on Ceph, Data Local
  - OS Local, Data on Ceph
- 3 VMs of each configuration spread across the cloud for a total of 12 VMs
- The cloud is very lightly used as it is still being commissioned

How Are We Testing?

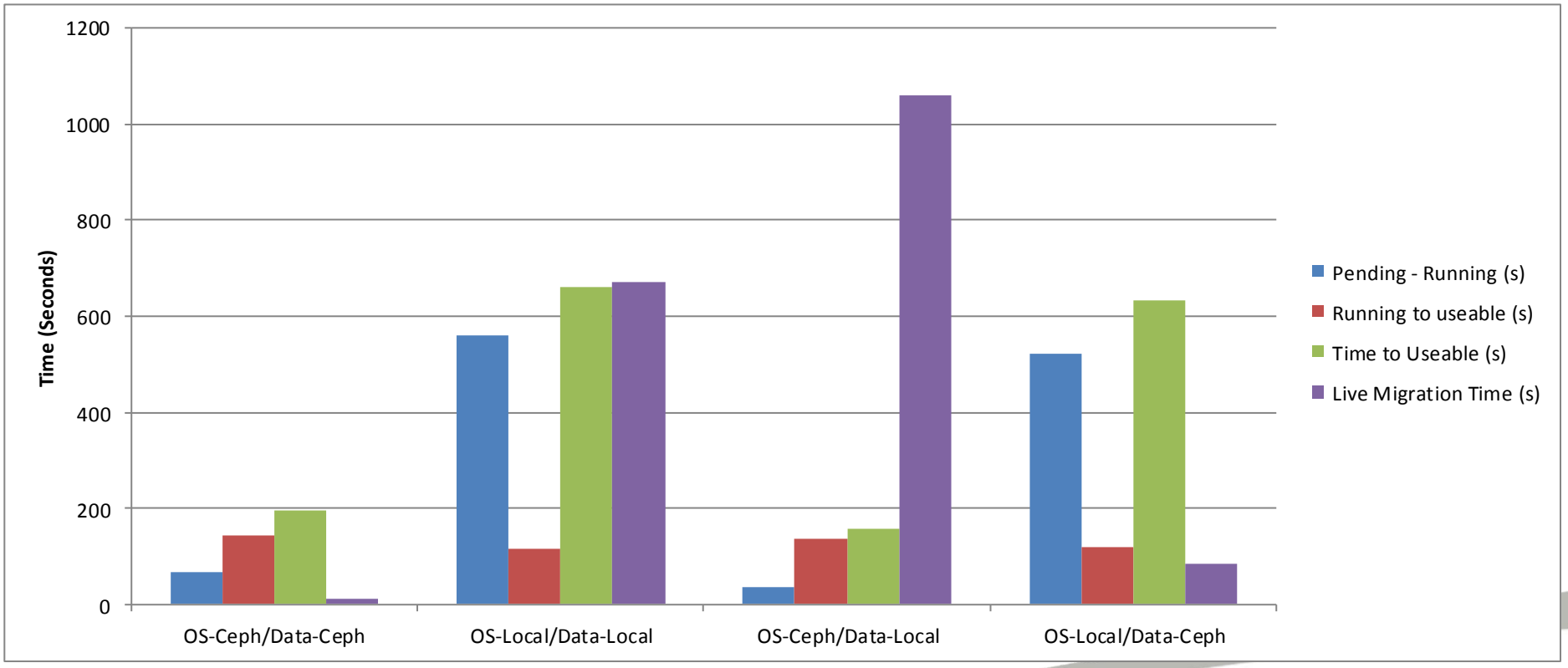
- Pending to Running (Time to deploy to Hypervisor)
- Running to useable (How long to boot)
- Pending to useable (Total of the above)
  - This is what users care about
- Live migration time
- IOZone Single Thread Tests (Read, ReRead, Write, ReWrite)
  - 6GB on OS Disk
  - 24GB on Data Disk
  - 3 VMs of each configuration throughout our cloud. 20 instance s of each test per VM

How Are We Testing?

- IOZone Aggregate Test – 12 Threads equal split mixed Read and Write (Read, ReRead, Write, ReWrite)
  - 0.5 GB per thread on the OS disk – 6GB total
  - 2 GB per thread on the Data disk – 24GB total
  - 3 VMs of each configuration throughout our cloud. 20 instance s of each test per VM (240 data points)

Results

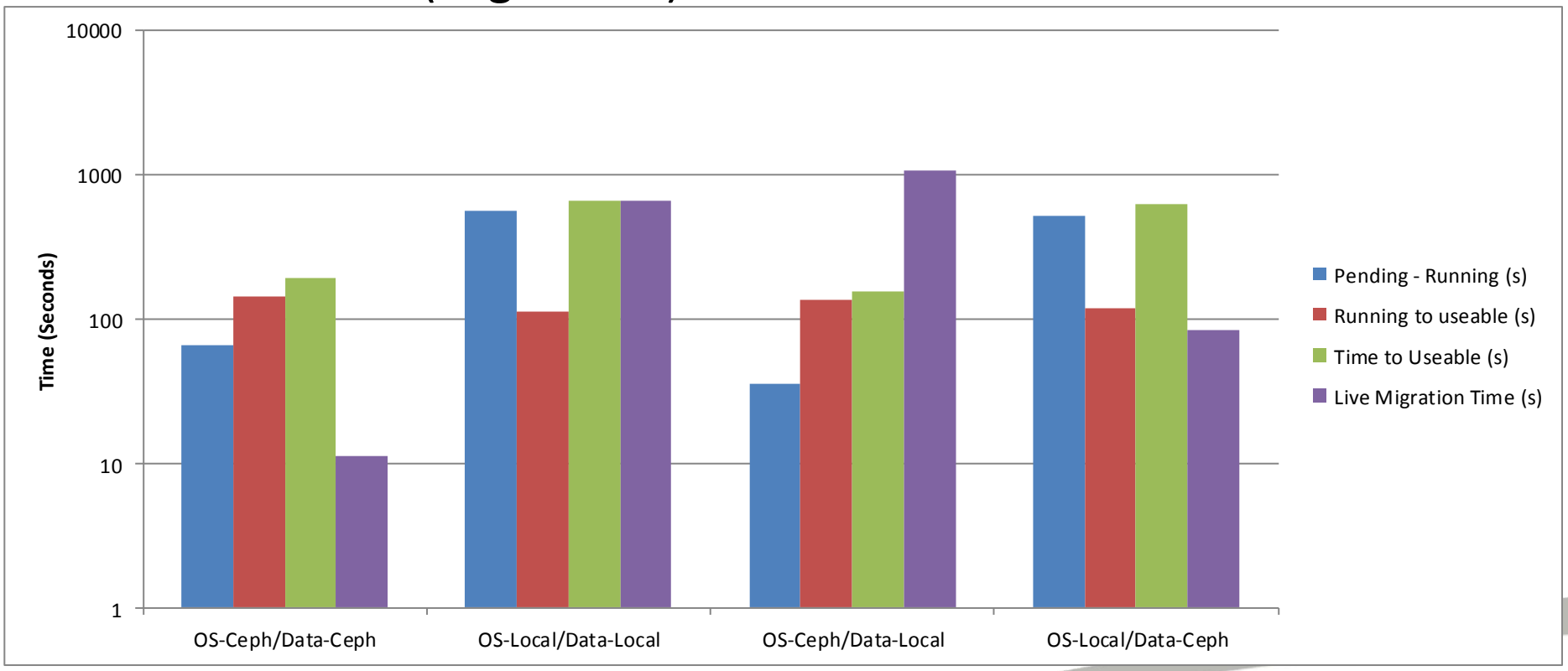
Launch Tests



Science & Technology Facilities Council

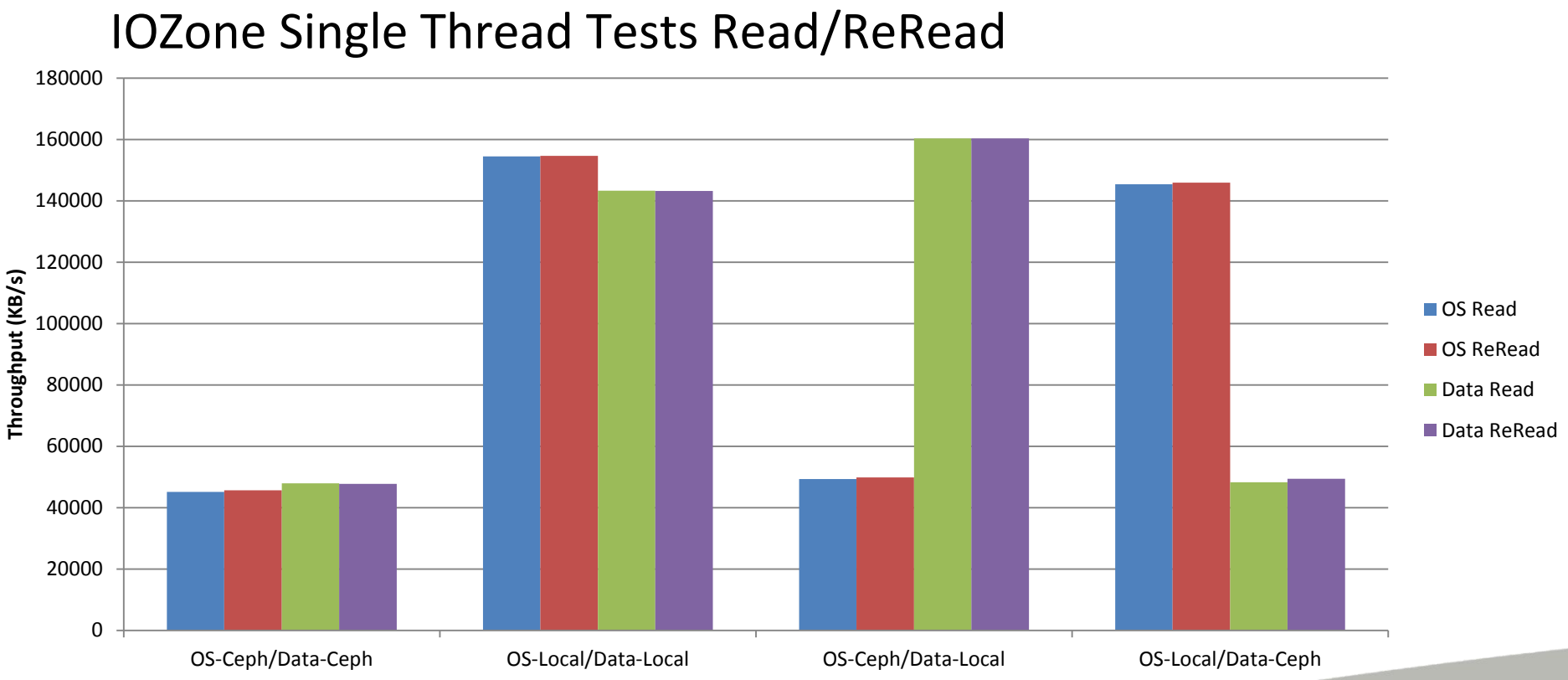
Results

Launch Tests (Log Scaled)



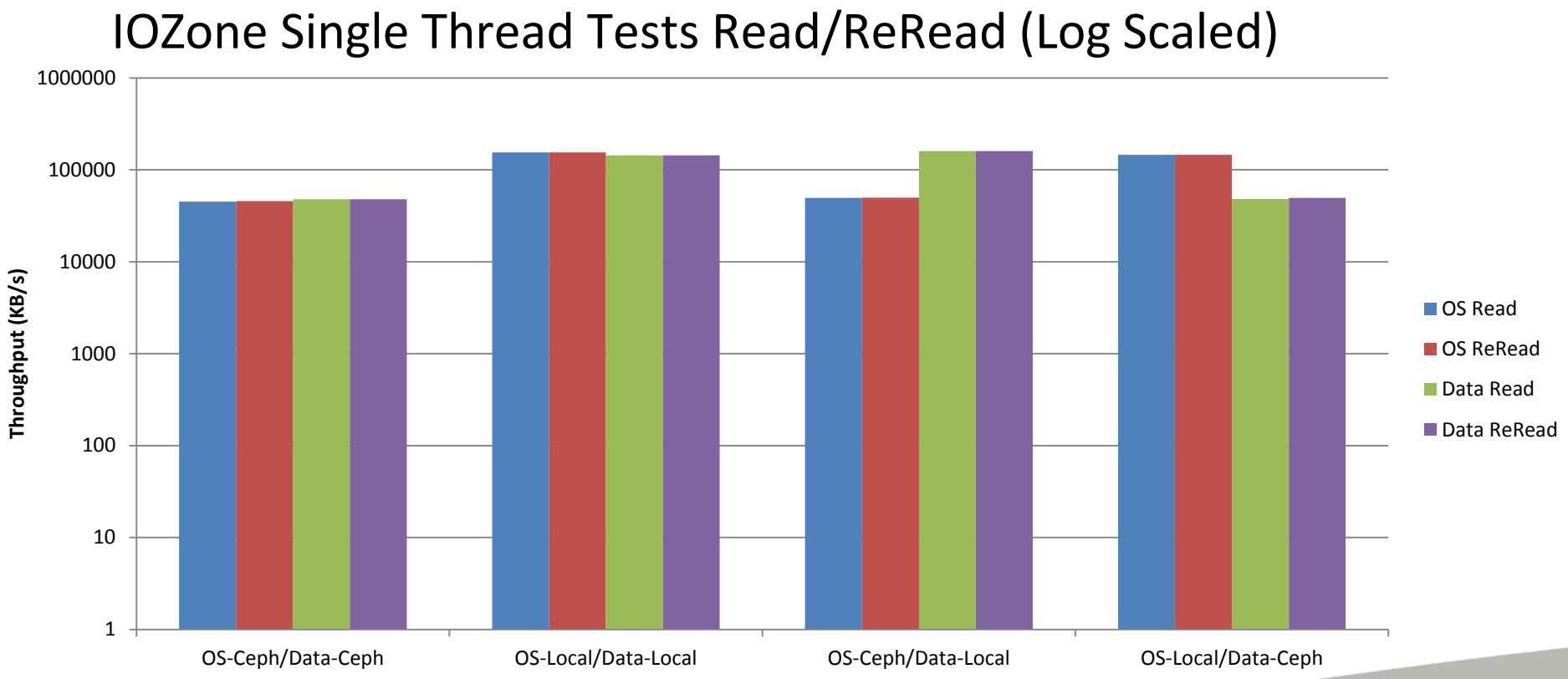
Science & Technology Facilities Council

Results



Science & Technology Facilities Council

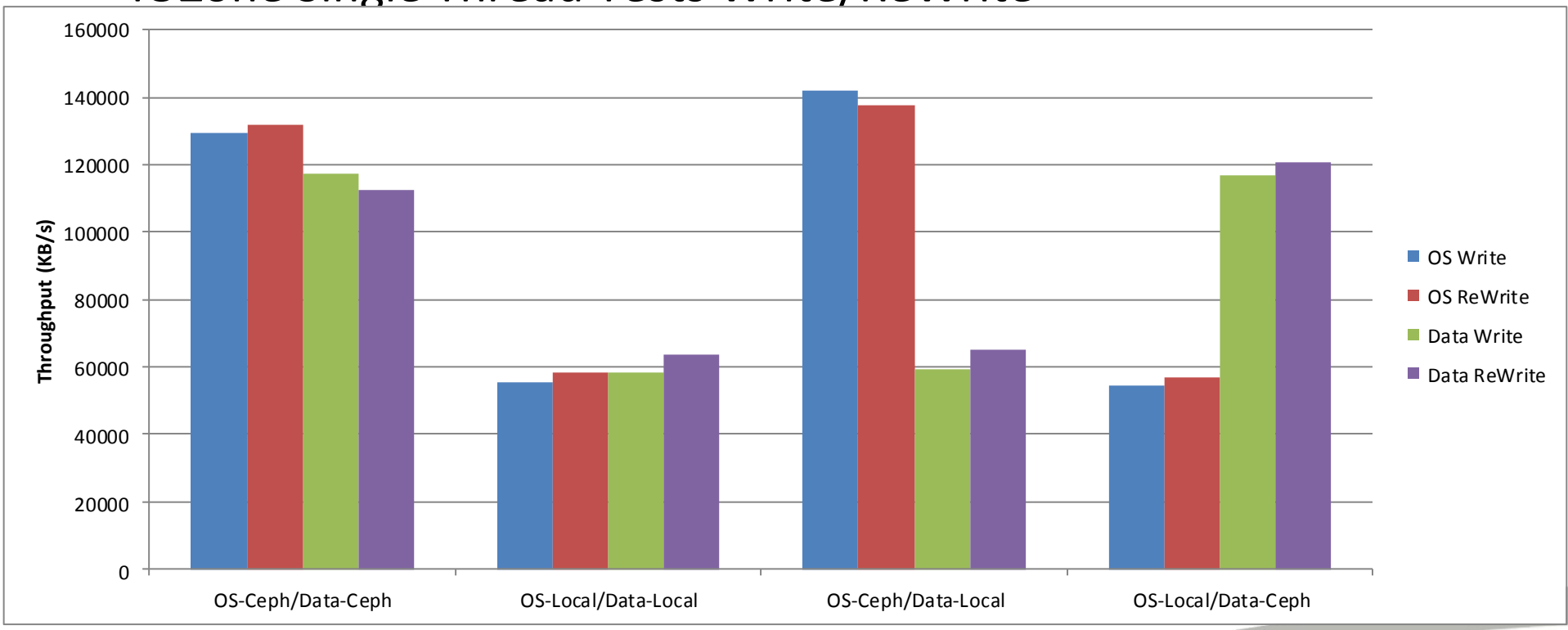
Results



Science & Technology Facilities Council

Results

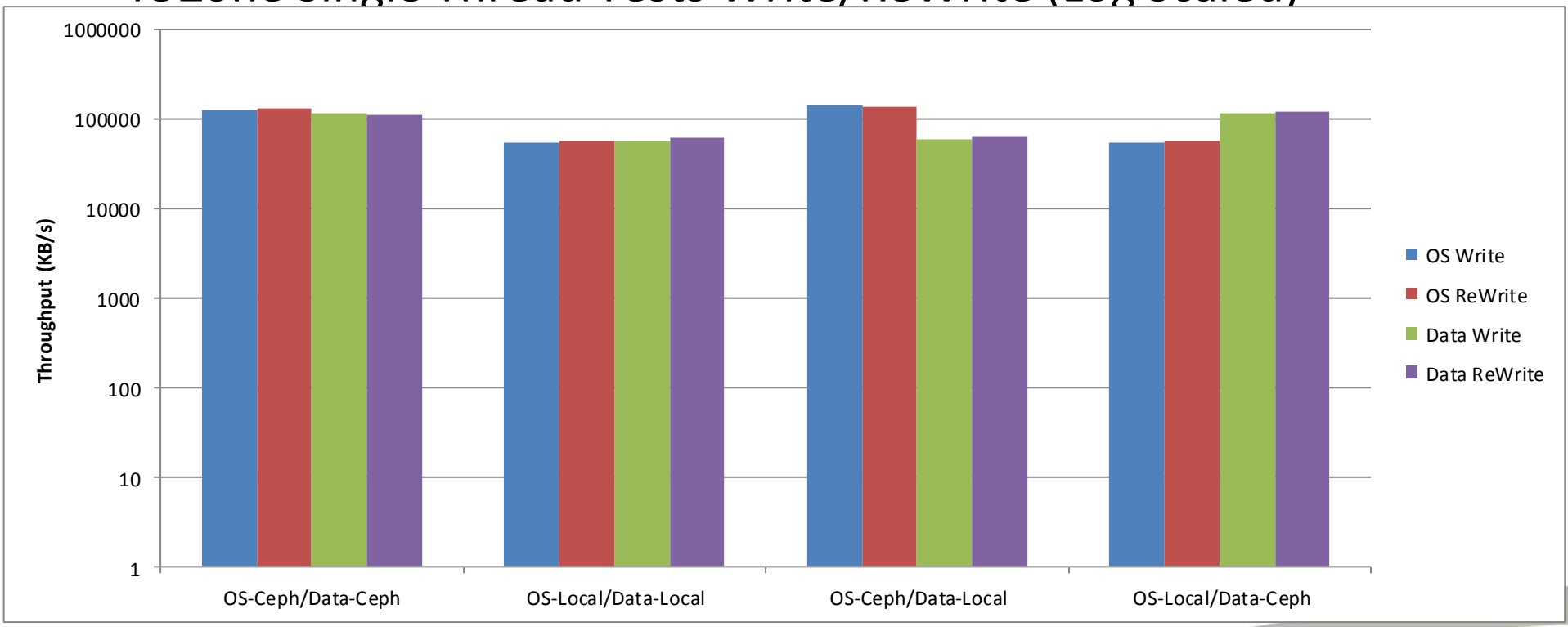
IOZone Single Thread Tests Write/ReWrite



Science & Technology Facilities Council

Results

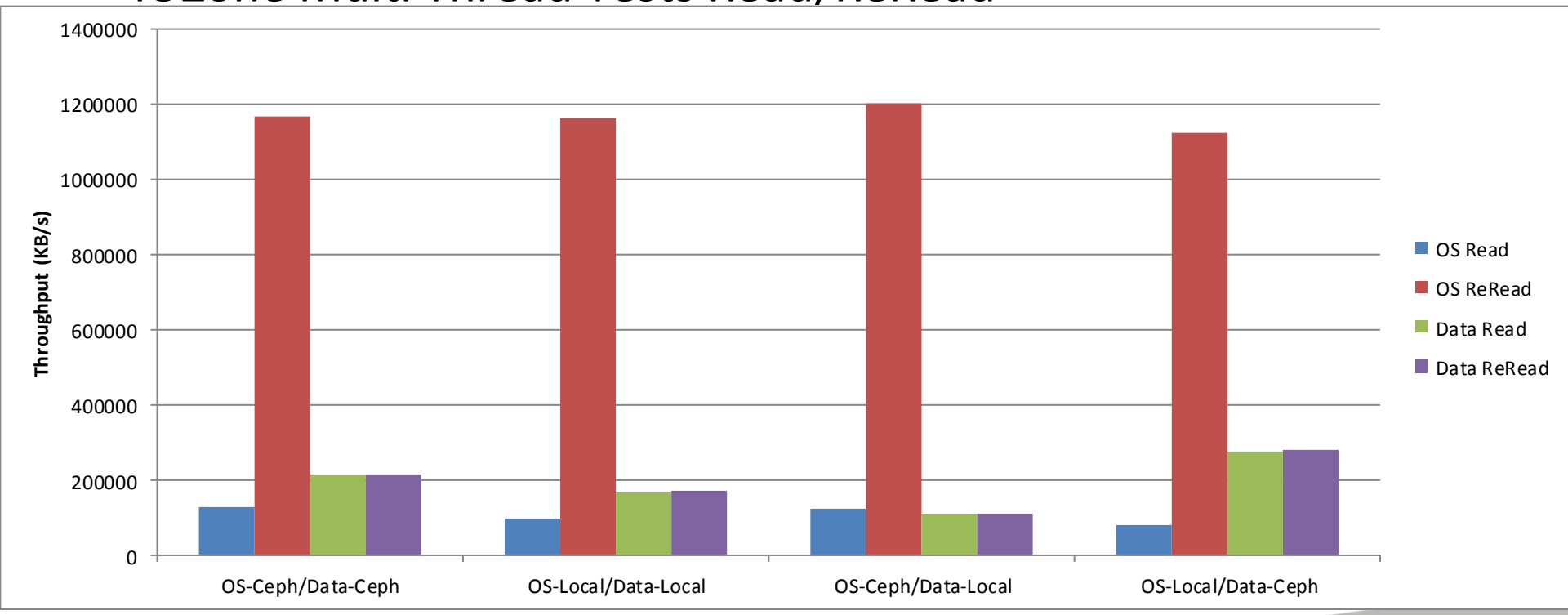
IOZone Single Thread Tests Write/ReWrite (Log Scaled)



Science & Technology Facilities Council

Results

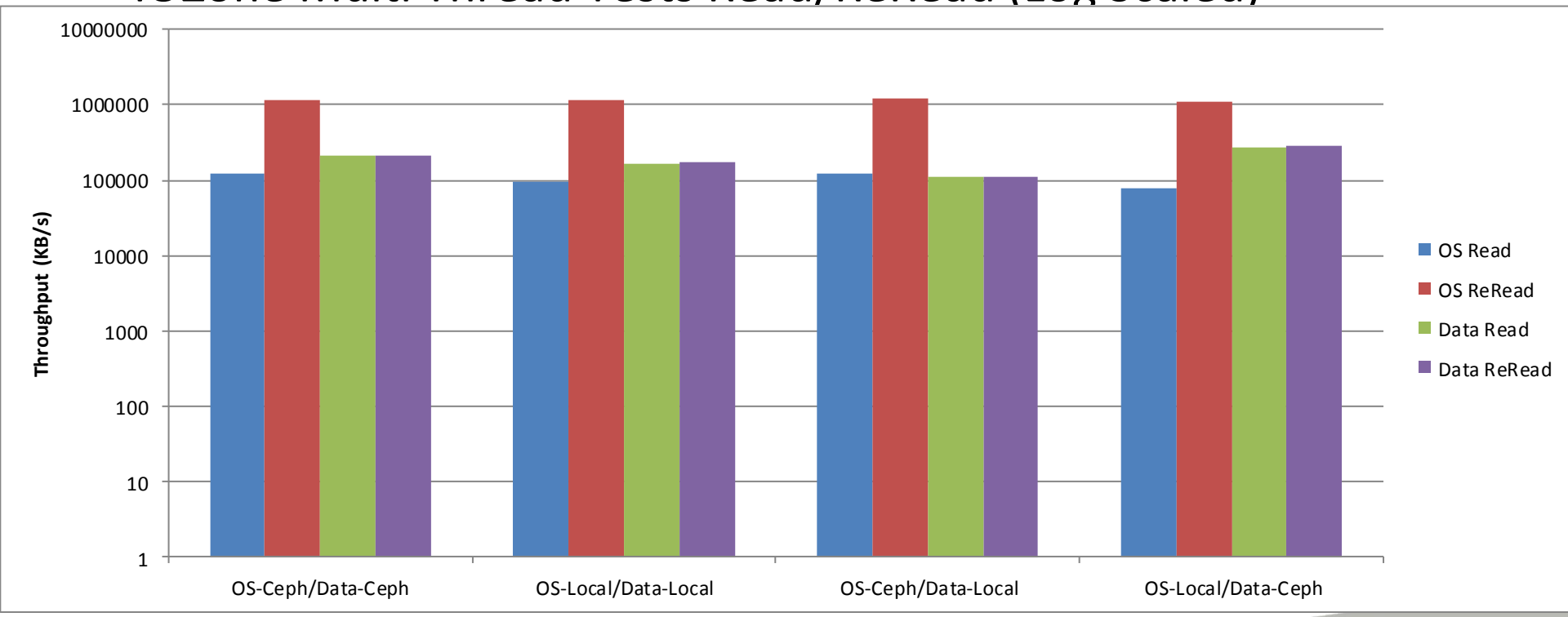
IOZone Multi Thread Tests Read/ReRead



Science & Technology Facilities Council

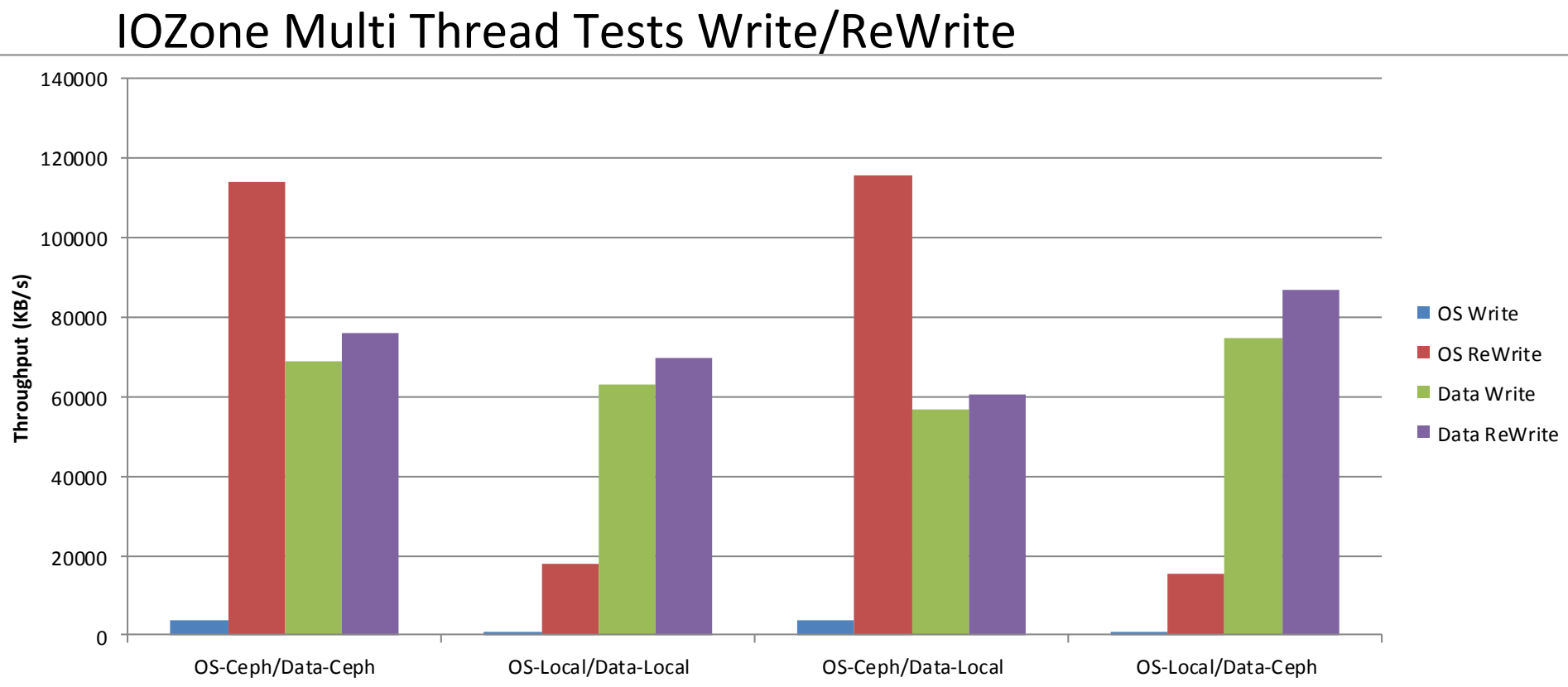
Results

IOZone Multi Thread Tests Read/ReRead (Log Scaled)



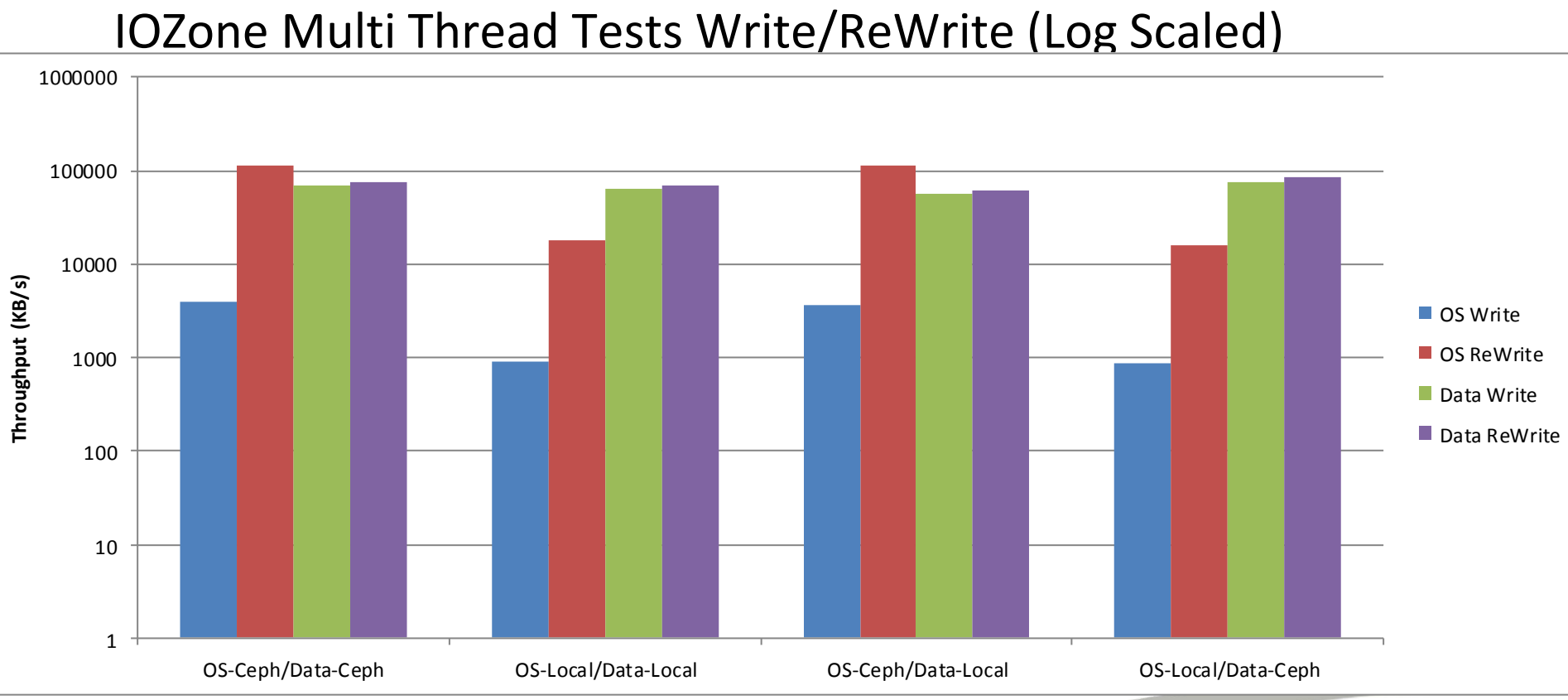
Science & Technology Facilities Council

Results



Science & Technology Facilities Council

Results



Science & Technology Facilities Council

Conclusions

- Local disk wins for single threaded read operations (such as booting the virtual machine)
- Ceph wins for single threaded write operations (large sequential writes)
- Ceph wins for both reads and writes for multi threaded operations

Why is this?

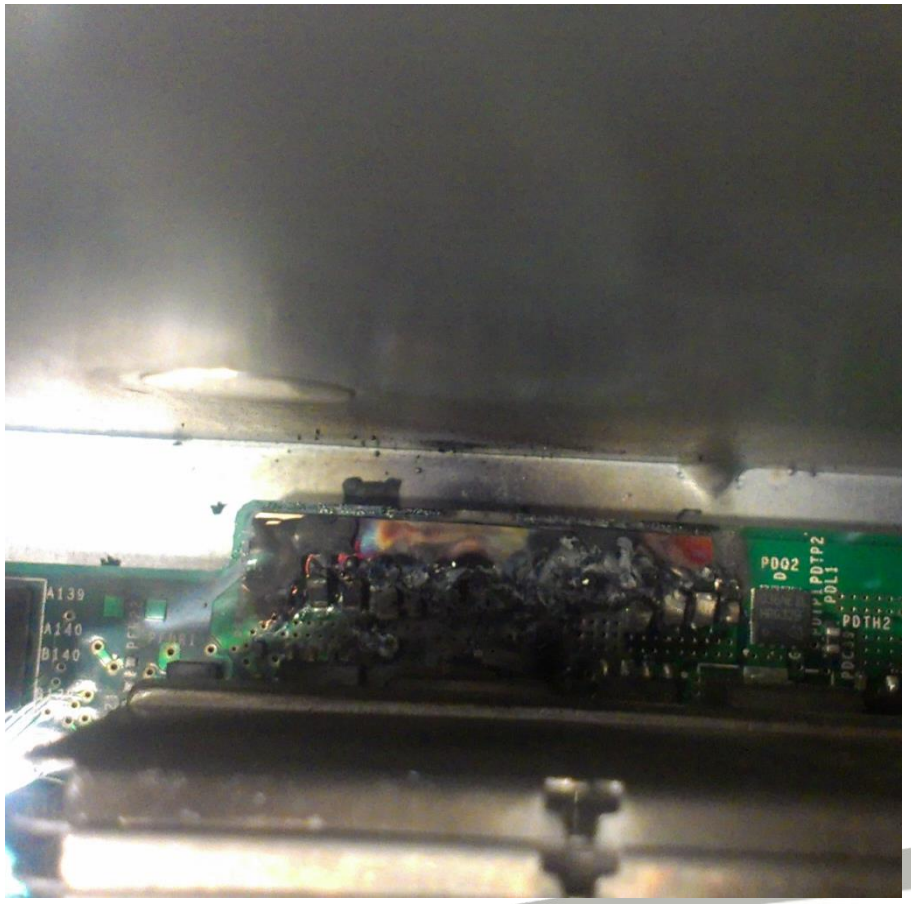
- Local disks have a maximum throughput which is very limited
- Due to the way RBD stripes data across the Ceph cluster the bottleneck here is the NIC on the hypervisor
  - In this case the NICs are 10Gb so to get equivalent performance would require a large RAID set in each hypervisor.

Further Work

- Test when the cloud is under more load
- Test using micro kernel VMs such as the µCernVM
- Test larger data sets

A Minor Issue

During the testing run we noticed that one of the storage nodes had dropped out of use.
After some investigation we found this ->
The testing, and the cloud as a whole, didn't skip a beat



Science & Technology Facilities Council

Any Questions?

Email: alexander.dibbo@stfc.ac.uk

Science & Technology
Facilities Council