Using ZFS to manage Grid storage and improve middleware resilience

Robert Currie rob.currie@ed.ac.uk The University of Edinburgh

THE UNIVERSITY *of* EDINBURGH

GridPP
UK Computing for Particle Physics

Introduction

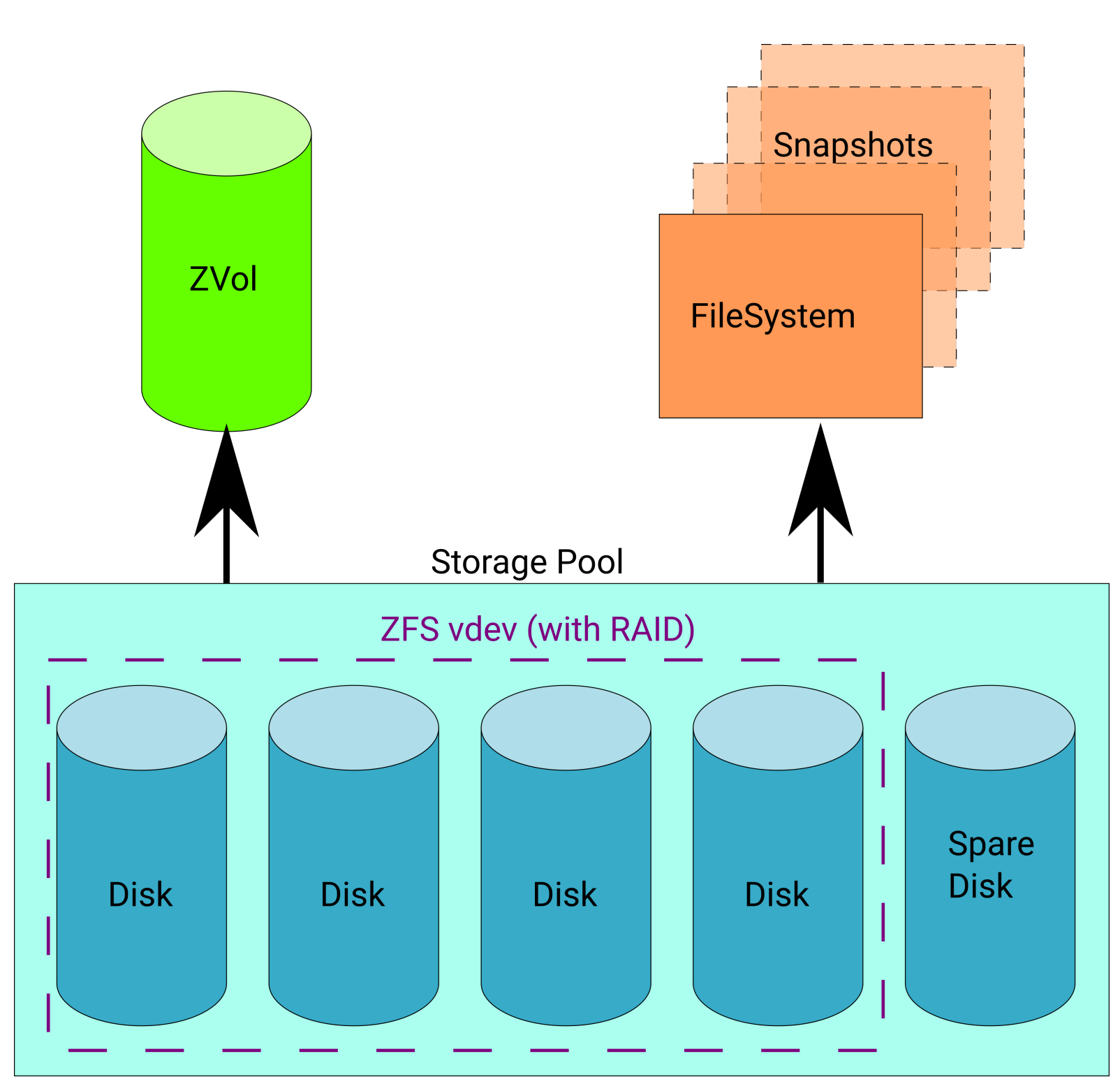
What is ZFS

ZFS is originally a storage managment solution developed by Sun Microsystems in the early 2000's for the Solaris platform. This solution differs from most other storage systems as it mixes volume managment, software raid and filesystems into the same piece of software.
ZFS is particularly attractive due to it's emphasis on data integrity and reliability.

Useful ZFS Terms

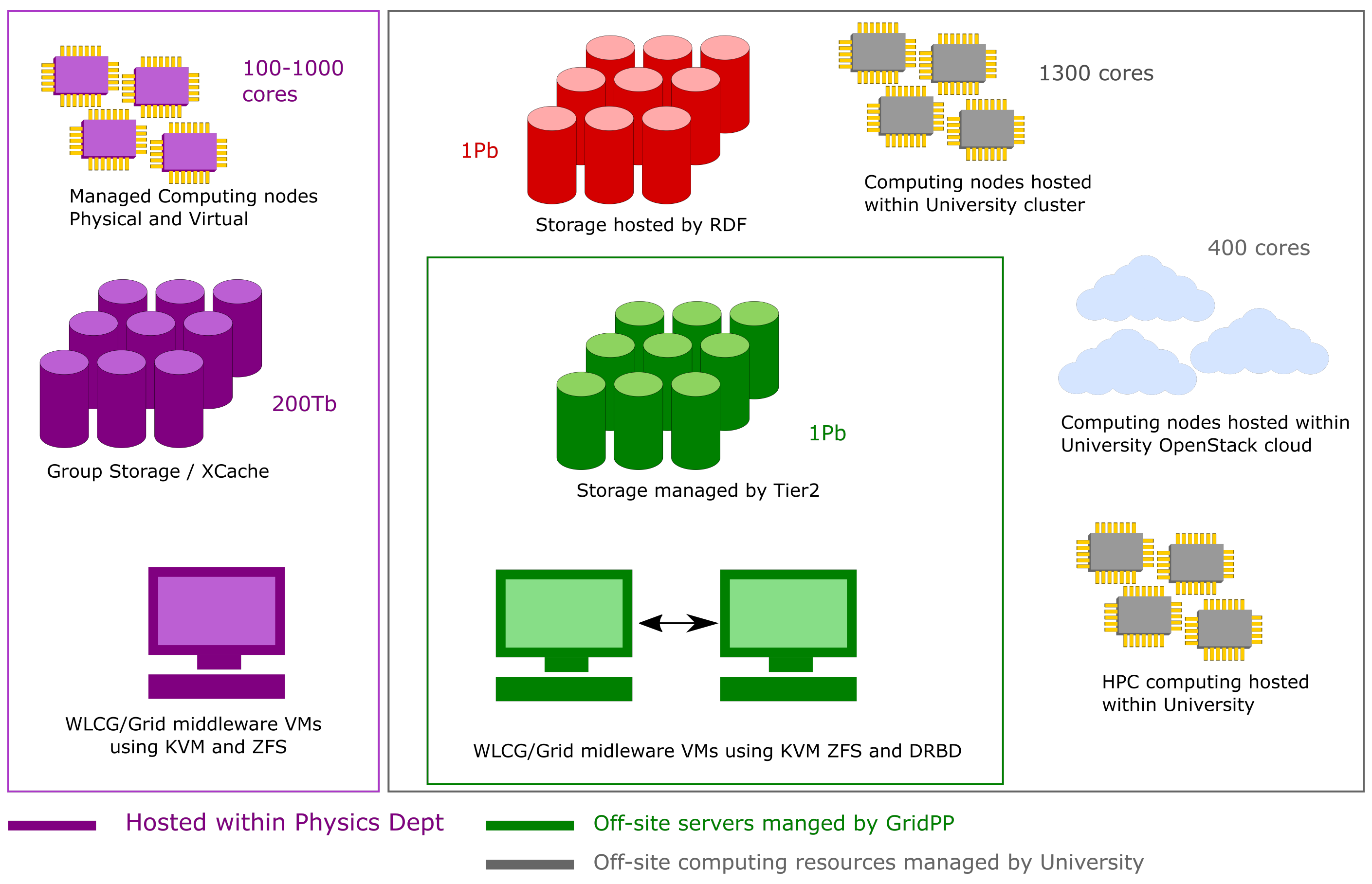
RAID-Z n	This is a variation on RAID-5 which has 'n' disk redundancy.
vdev	A virtual device which is composed of a set of disks. A vdev may be configured in different RAID configurations.
pool	A pool is a storage space composed of multiple vdev.
ZVol	This is a virtual block device created on top of a ZFS pool.
ARC	Adaptive Replacement Cache is the caching system within ZFS to cache data on-disk.
L2ARC	This is a device which has is used to cache data for low latency access.
ZIL	ZFS Intent Log, this is the internal log which is similar to a journal in other filesystems.

ZFS Structure



Edinburgh Tier2 WLCG Site

The WLCG Tier2 site at Edinburgh is built atop a combination of dedicated and shared computing resources hosted by members of GridPP and University staff.


100-1000 cores
Managed Computing nodes Physical and Virtual
200Tb
Group Storage / XCache
WLCG/Grid middleware VMs using KVM and ZFS
1Pb
Storage hosted by RDF
1300 cores
Computing nodes hosted within University cluster
400 cores
Computing nodes hosted within University OpenStack cloud
1Pb
Storage managed by Tier2
WLCG/Grid midleware VMs using KVM ZFS and DRBD
HPC computing hosted within University
Hosted within Physics Dept
Off-site servers manged by GridPP
Off-site computing resources managed by University

Edinburgh WLCG Tier2

ZFS backed storage plays a critical role at Edinburgh managing both the GridPP storage as well as the shared networked storage required to run Grid computing tasks on the university computing resources.

For more information see the presentation: "*Many hands make light work: Experiences from a shared resource WLCG Tier-2 computing site*" by A. Washbro, et al. Track 3 Distributed computing.

Using ZFS

Storage Administration Tier2

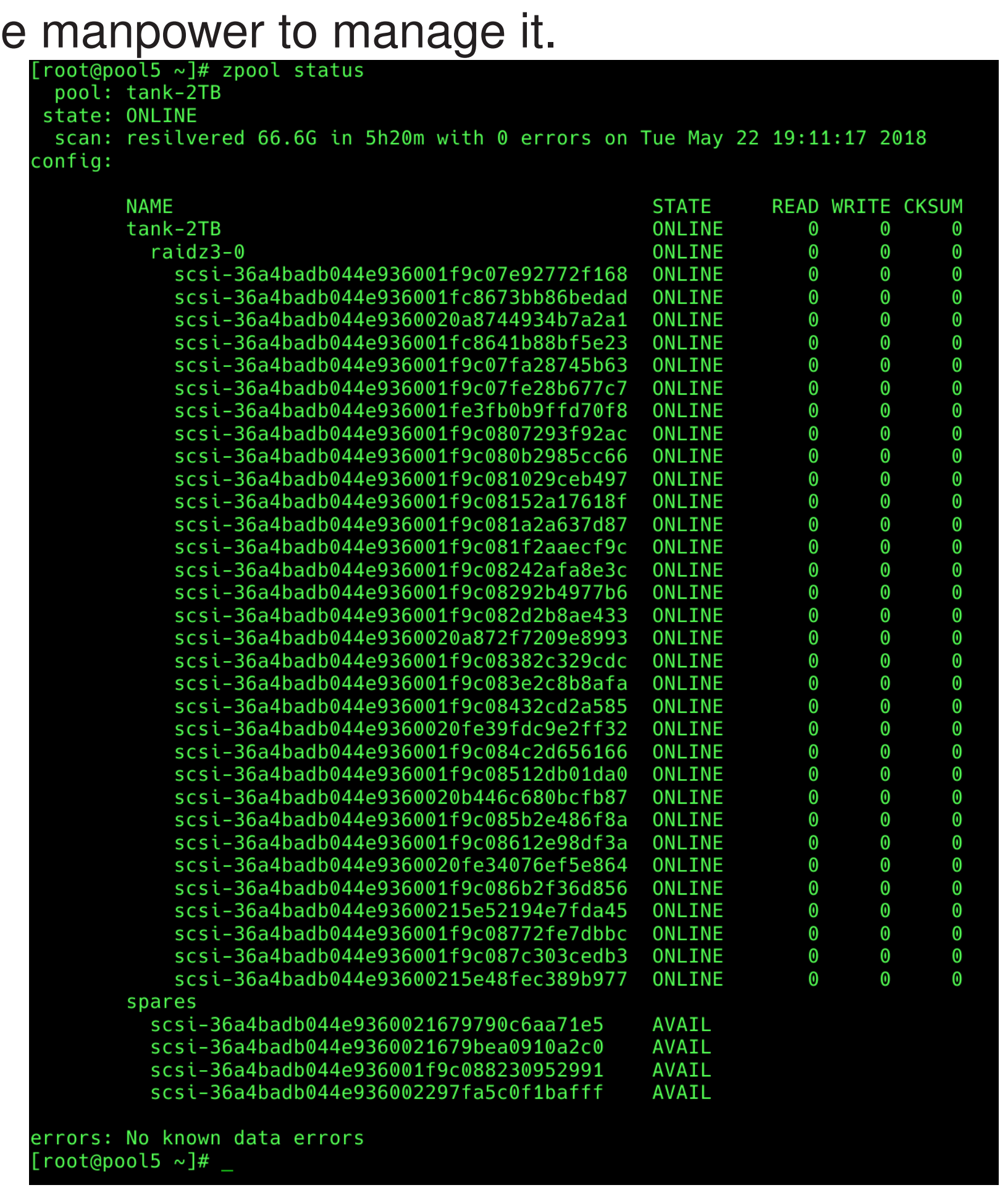
At the WLCG Tier 2 in Edinburgh we have replaced our existing hardware RAID managed storage with a ZFS based solution. This Storage is configured across 10 servers each managing 36 disks.

Best practice with ZFS suggests not to mount too many disks within the one vdev due to performance concerns. In practice managing many disks with 1 vdev can prove attractive as it can help reduce the complexity of a setup, and therefore manpower to manage it.

Due to the constraints of older RAID controllers our ZFS pools are constructed from an array of RAID-0 drives each containing 1 disk. This is the closest configuration to a HBA we can build.

To improve the performance of ZFS when managing pools containing large numbers of disk it is required to tune certain ZFS module parameters.

In particular this includes raising the amount of operations which can be performed by each vdev. Due to the large amount of disks managed by single vdevs the storage can cope with more iops compared to pools with smaller numbers of drives.



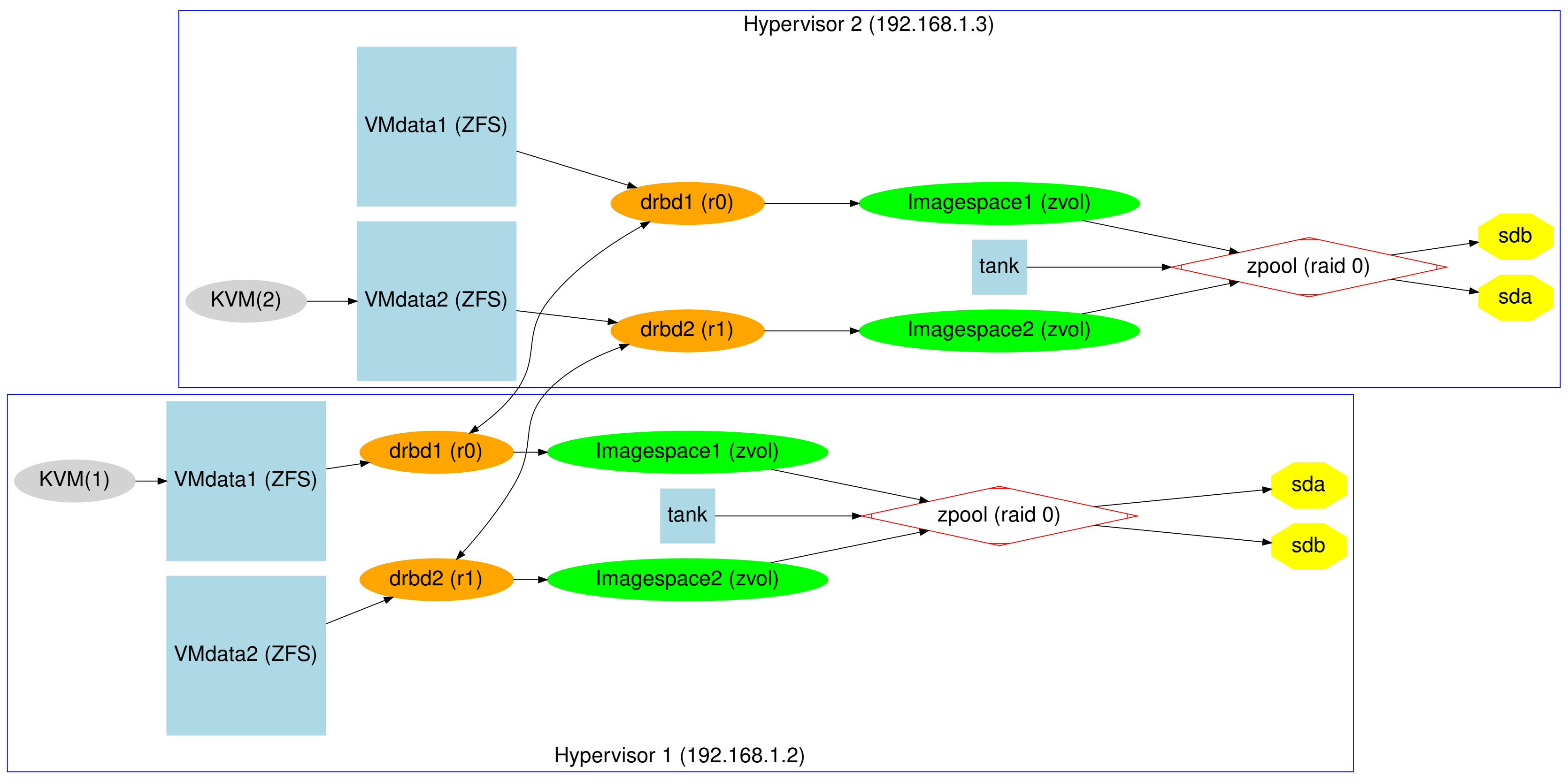
```
[root@pool5 ~]# zpool status
  pool: tank-2TB
 state: ONLINE
  scan: resilvered 66.6G in 5h20m with 0 errors on Tue May 22 19:11:17 2018
config:

        NAME                                        STATE     READ WRITE CKSUM
        tank-2TB                                    ONLINE       0     0     0
          raidz3-0                                  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c07e92772f168  ONLINE       0     0     0
            scsi-36a4badb044e936001fc8673bb86bedad  ONLINE       0     0     0
            scsi-36a4badb044e9360020a8744934b7a2a1  ONLINE       0     0     0
            scsi-36a4badb044e936001fc8641b88bf5e23  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c07fa28745b63  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c07fe28b677c7  ONLINE       0     0     0
            scsi-36a4badb044e936001fe3fb0b9ffd70f8  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c0807293f92ac  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c080b2985cc66  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c081029ceb497  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08152a17618f  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c081a2a637d87  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c081f2aaecf9c  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08242afa8e3c  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08292b4977b6  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c082d2b8ae433  ONLINE       0     0     0
            scsi-36a4badb044e9360020a872f7209e8993  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08382c329cdc  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c083e2c8b8afa  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08432cd2a585  ONLINE       0     0     0
            scsi-36a4badb044e9360020fe39fdc9e2ff32  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c084c2d656166  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08512db01da0  ONLINE       0     0     0
            scsi-36a4badb044e9360020b446c680bcfb87  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c085b2e486f8a  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08612e98df3a  ONLINE       0     0     0
            scsi-36a4badb044e9360020fe34076ef5e864  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c086b2f36d856  ONLINE       0     0     0
            scsi-36a4badb044e93600215e52194e7fda45  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c08772fe7dbbc  ONLINE       0     0     0
            scsi-36a4badb044e936001f9c087c303cedb3  ONLINE       0     0     0
            scsi-36a4badb044e93600215e48fec389b977  ONLINE       0     0     0
        spares
          scsi-36a4badb044e9360021679790c6aa71e5    AVAIL
          scsi-36a4badb044e9360021679bea0910a2c0    AVAIL
          scsi-36a4badb044e936001f9c088230952991    AVAIL
          scsi-36a4badb044e936002297fa5c0f1bafff    AVAIL

errors: No known data errors
[root@pool5 ~]# _
```

Building a Redundant Hypervisor

In order to build a hypervisor with maximum resiliance and reliability we decided to cluster the storage from 2 HV machines using ZFS and DRBD.
This configuration exploits ZFS's ability to create virtual block devices (ZVOL) atop a ZFS filesystem which can be used in a normal way with DRBD to offer failover redundancy at the host and filesystem level.



Redundant Hypervisor managed with ZFS and DRBD

With 2 Hypervisors set up with a redundant server configuration it has been demonstrated that a guest VM can be launched from backup incase of a server failure in under 5minutes with minimal data loss.
An important change was to limit the write cache for ZFS to reduce the possibility of service disruption due to large incoming writes.

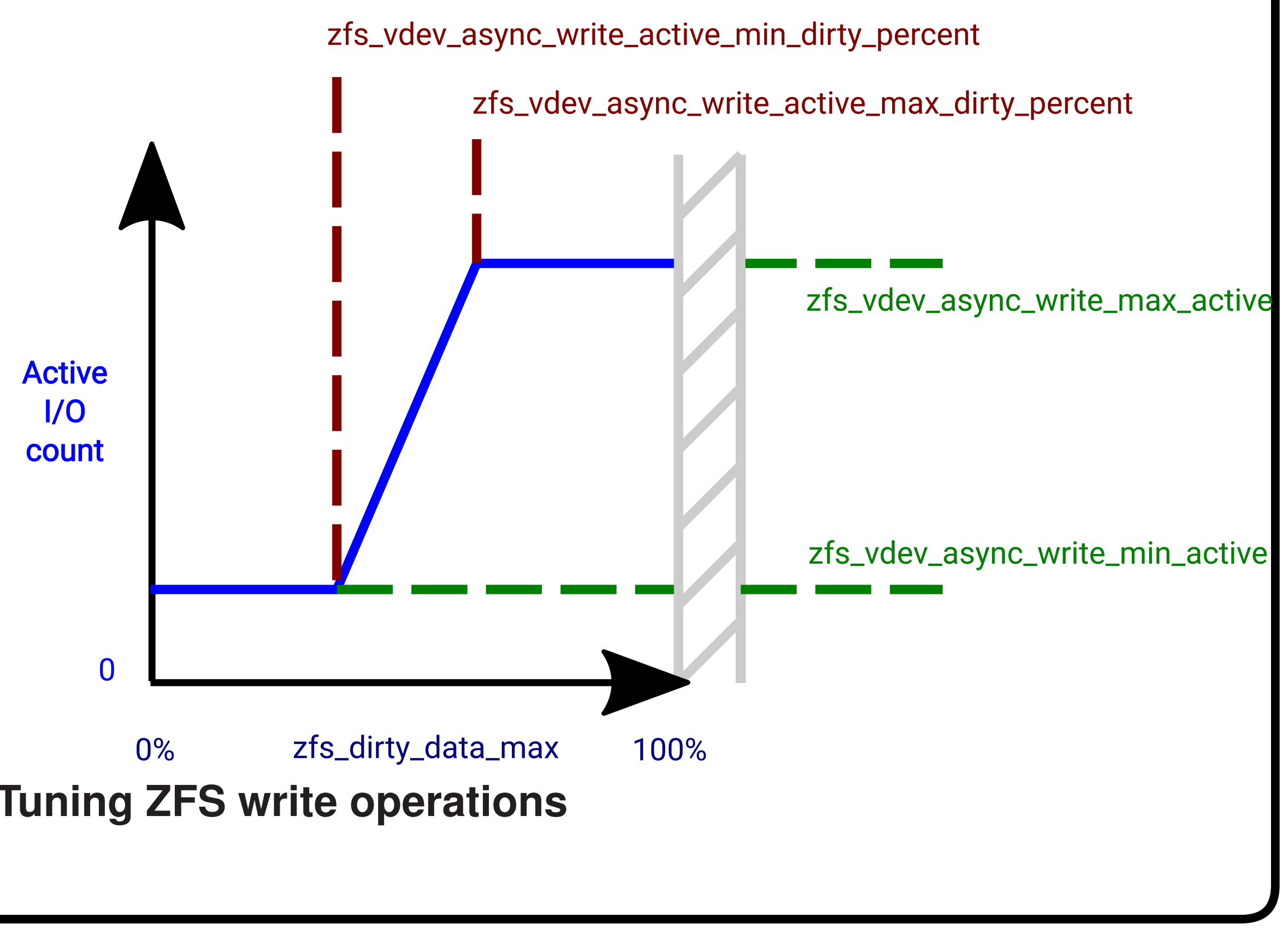
Best Practices

Configuring ZFS

ZFS provides a lot of options for tuning the filesystem performance as well as different tools for providing debugging information.
In order to have efficient read/write operations ZFS buffers data that is read/written to disk in ECC RAM. This offers a large performance boost as hot data can be effectively cached as well as write operations being streamlined before the data is actually written to disk.

Parameters described in the diagram on the right impact the number of I/O operations that ZFS has running in parallel. These can be configured dynamically at runtime on the host system.

Tuning ZFS module parameters can increase the number of iops in environments with a small amount of RAM relative to the storage or environments where the buffer size has to be controlled.



Tuning ZFS write operations

Conclusions

Within the LCG Tier2 at Edinburgh we have been able to replace our hardware RAID managment interfaces with ZFS based RAID-Z3 solutions.
This has brought us several advantages:

- More control over aging storage with failing drives.
- Better performance over pure hardware based solutions.
- Reliable storage for network sharing across different compute resources.
- Combined ZFS + DRBD to provide redundancy at the server level on hypervisor hosts.
- Built a highly reliable and redundant framework to host a Tier2 VMs.
- Better alerts and handling of silent corruption on disk.
- Improve out of the box performance due to tuning ZFS configuration.
- Doubled disk throughput compared to out of the box ZFS installation.