Giving Cloud Computing an Edge

- Alasdair Lumsden
- al@everycity.co.uk
- blogs.everycity.co.uk/alasdair

EveryCity = Managed Hosting on Cloud Infrastructure

What is "**Cloud**"?

- For us, Cloud = Fully Virtualised Infrastructure

Why Cloud?

One word: Manageability.

Resize CPU, Memory, Disk easily

Rapid provisioning of new servers

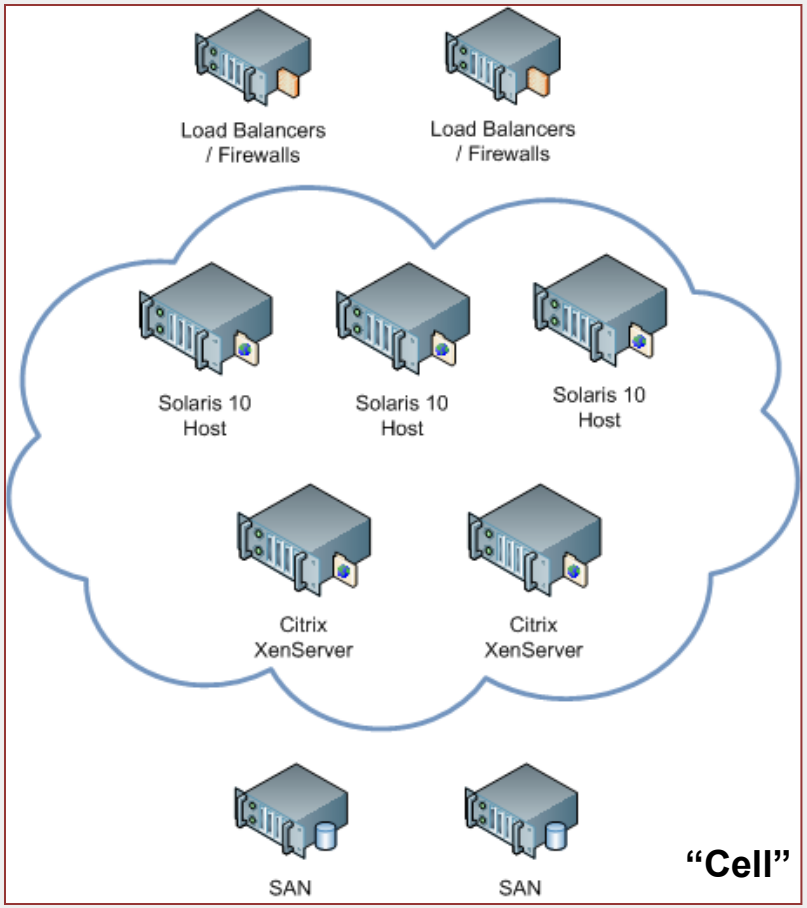
Move servers around

Clone Servers instantly

Respawning after hardware failure

Reduces Server Sprawl

Physical View:



Technologies:

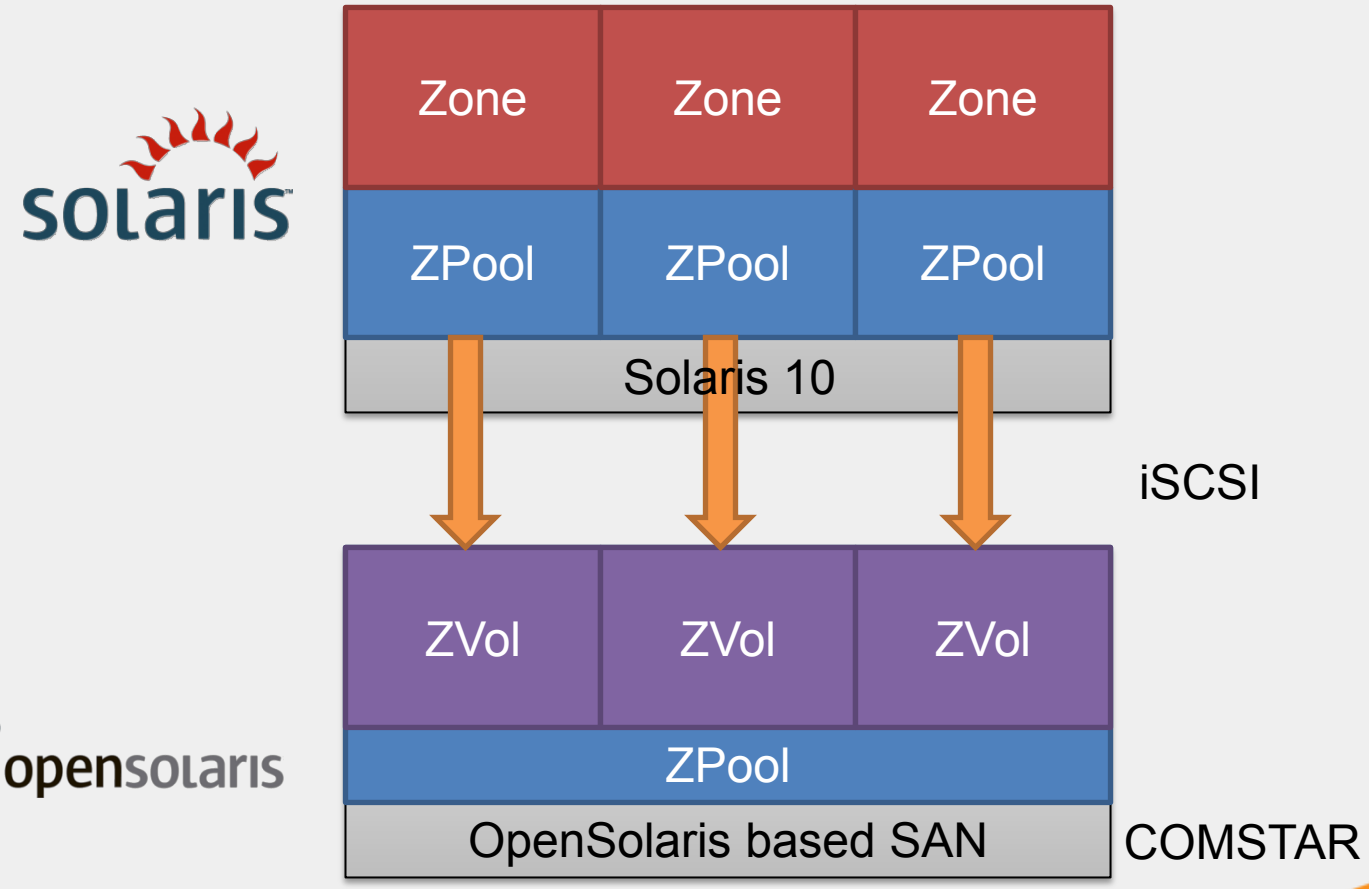
Zeus ZXTM Load Balancers

Primarily Solaris 10 Hosts running Zones

Citrix XenServer for Windows & Linux Requirements

ZFS based SANs including Sun Unified Storage Arrays

Logical View:



Why Solaris Zones Rock:

Superlight weight virtualisation

Update on Attach

Portability

Speed of Booting

Very fast provisioning

DTrace

Resource Management

RBAC

Scripting and manageability

Auditing

Service Management Framework

Why ZFS Rocks:

Without ZFS, much of what we are doing simply wouldn't be possible.

ZFS:

- Allows rapid cloning of **running** servers
- Allows server snapshots
- Massively simplify backups
- Allows mirroring across SANs – data portability, tiered storage
- Compression saves space AND increases performance
- Checksums all data. You know your data is correct on disk.

ZFS Checksums can save your bacon:

NAME	STATE	READ	WRITE	CKSUM	
pool01	DEGRADED	0	0	0	
raidz1-0	ONLINE	0	0	0	
c11t3d0	ONLINE	0	0	4	2.50K repaired
c10t3d0	ONLINE	0	0	0	
c13t3d0	ONLINE	0	0	4	1.50K repaired
c7t1d0	**REMOVED**	**0**	**0**	**0**	
c8t3d0	ONLINE	0	0	5	1K repaired
c7t3d0	ONLINE	0	0	4	2K repaired
c10t2d0	ONLINE	0	0	3	1K repaired
c13t2d0	ONLINE	0	0	2	1K repaired
c11t6d0	ONLINE	0	0	3	1K repaired
c8t2d0	ONLINE	0	0	16	7K repaired
c7t2d0	ONLINE	0	0	4	2.50K repaired
raidz1-1	DEGRADED	0	0	0	
c11t7d0	ONLINE	0	0	6	64K repaired
c10t7d0	**DEGRADED**	**0**	**0**	**58**	**too many errors**
c13t7d0	ONLINE	0	0	4	3.50K repaired
c12t7d0	ONLINE	0	0	3	7K repaired
c8t7d0	ONLINE	0	0	2	4.50K repaired
c7t7d0	ONLINE	0	0	4	11.5K repaired
c10t6d0	ONLINE	0	0	4	11K repaired
c13t6d0	ONLINE	0	0	8	86K repaired
c12t6d0	ONLINE	0	0	0	
c8t6d0	ONLINE	0	0	2	1K repaired
c7t6d0	ONLINE	0	0	2	2.50K repaired
raidz1-2	DEGRADED	0	0	0	
c11t5d0	ONLINE	0	0	1	9K repaired
c10t5d0	ONLINE	0	0	1	13K repaired
c13t5d0	ONLINE	0	0	2	1.50K repaired
c12t5d0	ONLINE	0	0	1	1K repaired
c8t5d0	**DEGRADED**	**0**	**0**	**135**	**too many errors**
c7t5d0	ONLINE	0	0	2	1.50K repaired
c10t4d0	ONLINE	0	0	8	44K repaired
c13t4d0	ONLINE	0	0	3	5K repaired
c12t4d0	ONLINE	0	0	3	2K repaired
c8t4d0	ONLINE	0	0	2	6.50K repaired
c7t4d0	ONLINE	0	0	2	13.5K repaired

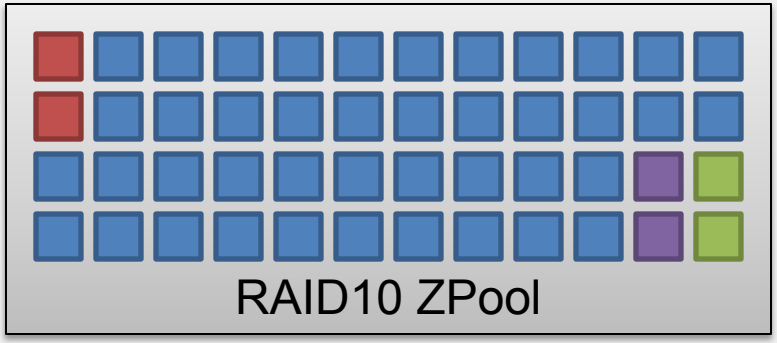
```
# iostat -En | grep c7t1d0
c7t1d0          Soft Errors: 1 Hard Errors: 127488 Transport Errors: 0
```

A refreshing change from this sort of thing (Linux Host):

```
# dmesg
<snip>
sdb: Mode Sense: 04 00 80 00
sdb: assuming drive cache: write through
sdb: Spinning up
disk.....................................................................................................not
responding...
sdb : READ CAPACITY failed.
sdb : status=1, message=00, host=0, driver=08
sd: Current: sense key: Not Ready
    Add. Sense: No additional sense information

sdb: Write Protect is on
sdb: Mode Sense: 04 00 80 00
sdb: assuming drive cache: write through
sdb: Spinning up
disk.....................................................................................................not
responding...
sdb : READ CAPACITY failed.
sdb : status=1, message=00, host=0, driver=08
sd: Current: sense key: Not Ready
    Add. Sense: No additional sense information

sdb: Write Protect is on
sdb: Mode Sense: 04 00 80 00
sdb: assuming drive cache: write through
 sdb:<3>Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
Dev sdb: unable to read RDB block 0
Buffer I/O error on device sdb, logical block 0
Buffer I/O error on device sdb, logical block 0
 unable to read partition table
```

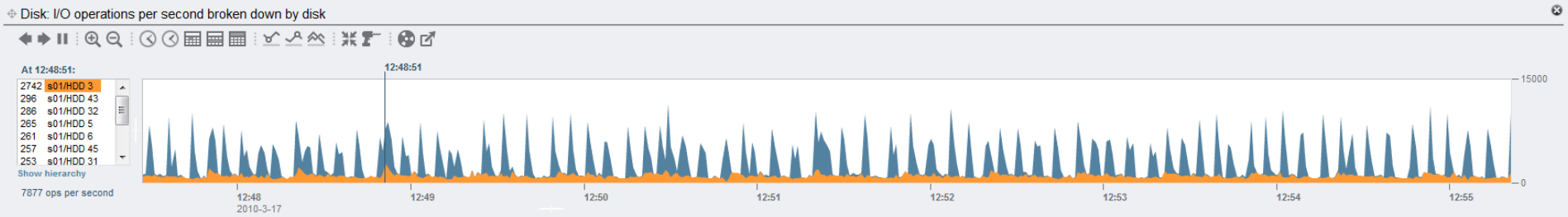
ZFS Performance : Hybrid Storage Pools

- 2 SATA Boot Disks
- 42 SATA Data Disks
- 2 ZIL/SLOG Devices
- 2 Hot Spare SATA Disks




```
NAME                                                STATE     READ WRITE CKSUM
        pool-0                                              ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C83E36d0                           ONLINE       0     0     0
            c0t5000CCA214C83ED0d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C84E3Ad0                           ONLINE       0     0     0
            c0t5000CCA214C85B2Dd0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C85B34d0                           ONLINE       0     0     0
            c0t5000CCA214C85B38d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C86CE6d0                           ONLINE       0     0     0
            c0t5000CCA214C86D2Fd0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C86D24d0                           ONLINE       0     0     0
            c0t5000CCA214C86D26d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C86D46d0                           ONLINE       0     0     0
            c0t5000CCA214C869AAd0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C871AAd0                           ONLINE       0     0     0
            c0t5000CCA214C871B7d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C871C8d0                           ONLINE       0     0     0
            c0t5000CCA214C871CDd0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C871E0d0                           ONLINE       0     0     0
            c0t5000CCA214C871E1d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C871FDd0                           ONLINE       0     0     0
            c0t5000CCA214C87478d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C875AFd0                           ONLINE       0     0     0
            c0t5000CCA214C875B0d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C875DAd0                           ONLINE       0     0     0
            c0t5000CCA214C875DCd0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C875EAd0                           ONLINE       0     0     0
            c0t5000CCA214C8718Dd0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C8723Ad0                           ONLINE       0     0     0
            c0t5000CCA214C84605d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C8756Ed0                           ONLINE       0     0     0
            c0t5000CCA214C85606d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C85627d0                           ONLINE       0     0     0
            c0t5000CCA214C85629d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C85684d0                           ONLINE       0     0     0
            c0t5000CCA214C86626d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C86981d0                           ONLINE       0     0     0
            c0t5000CCA214C86991d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C87189d0                           ONLINE       0     0     0
            c0t5000CCA214C87217d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C87231d0                           ONLINE       0     0     0
            c0t5000CCA214C87238d0                           ONLINE       0     0     0
          mirror                                            ONLINE       0     0     0
            c0t5000CCA214C87269d0                           ONLINE       0     0     0
            c0t5000CCA214C87271d0                           ONLINE       0     0     0
        logs                                                ONLINE       0     0     0
          c0tATASTEC2EUSIOPS018GBYTESSTS000004A38d0         ONLINE       0     0     0
          c0tATASTEC2EUSIOPS018GBYTESSTS0000086B3d0         ONLINE       0     0     0
        spares
          c0t5000CCA214C87274d0                             AVAIL
          c0t5000CCA214C87461d0                             AVAIL
```

ZFS Performance : ZFS Intent Log



ZFS collects writes together and flushes them in transaction groups, to stay consistent on disk. (Hence the sawtooth above)

However synchronous writes need to be committed to disk immediately. These get sent to the ZFS Intent Log.

Placing the ZFS Intent Log on Solid State Disks allows these to synchronous writes to return far more quickly, providing more IOPS to applications such as databases.

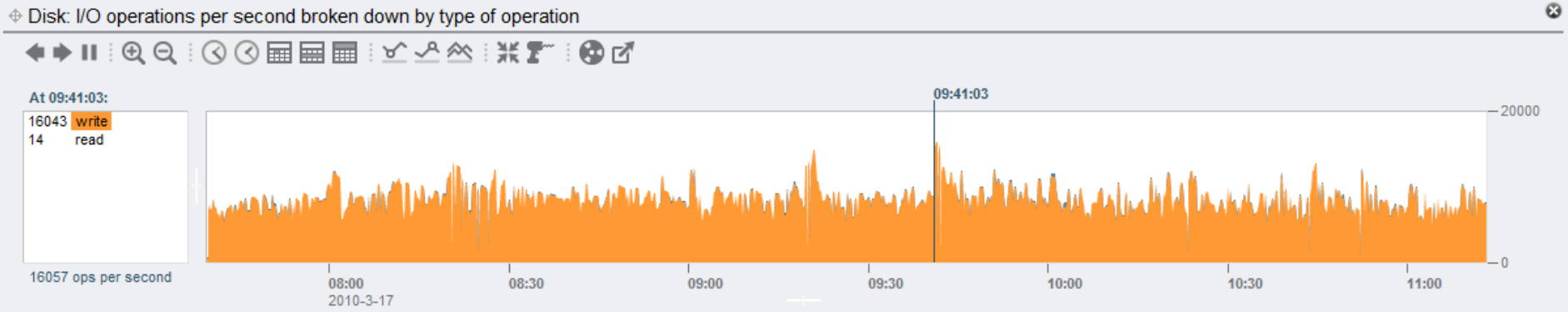
ZFS Performance : Adaptive Replacement Cache

ZFS Arc is an in-memory cache of Most Recently Used + Most Frequently Used data

Will grow to utilise all free memory (a good thing)

Our physical Zone servers have 32GB of ram, and since Zones pool memory, we have ample free memory for the ARC.

As a result, the workload on our SANs is over 95% write biased!



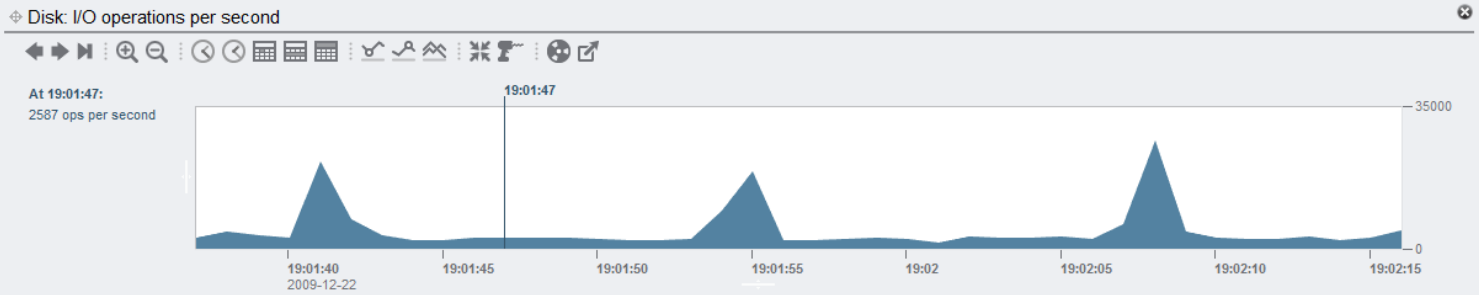
ZFS Performance : ARC Summary from a real Zone Server

```
System Memory:
         Physical RAM:  32759 MB
         Free Memory :  2332 MB
         LotsFree:      499 MB

ZFS Tunables (/etc/system):

ARC Size:
         Current Size:             15456 MB (arcsize)
         Target Size (Adaptive):   16294 MB (c)
         Min Size (Hard Limit):    3966 MB (zfs_arc_min)
         Max Size (Hard Limit):    31735 MB (zfs_arc_max)

ARC Size Breakdown:
         Most Recently Used Cache Size:          76%    12452 MB (p)
         Most Frequently Used Cache Size:        23%    3841 MB (c-p)

ARC Efficency:
         Cache Access Total:             1521248875
         Cache Hit Ratio:      99%       1516157567     [Defined State for buffer]
         Cache Miss Ratio:      0%       5091308        [Undefined State for Buffer]
         REAL Hit Ratio:       96%       1470780935     [MRU/MFU Hits Only]

         Data Demand   Efficiency:    99%
         Data Prefetch Efficiency:    98%

...
```

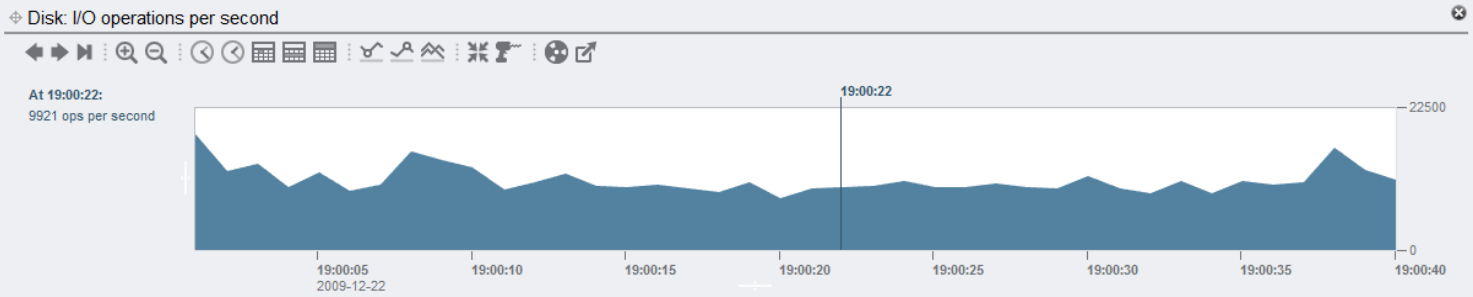
Ben Rockwood's Arc Summary Script available from:
http://cuddletech.com/arc_summary/

ZFS Performance : SAN Gotchas

Good:



Very very very bad:



ZFS Performance : SAN Gotchas (continued)

The biggest danger when running a SAN is running out of **write IOPS capacity**.

Performance will be consistently high, until suddenly it falls off a cliff.

Worse, if requests come in faster than can be serviced, queues form, leading to iSCSI timeouts and everything from that SAN stalling.

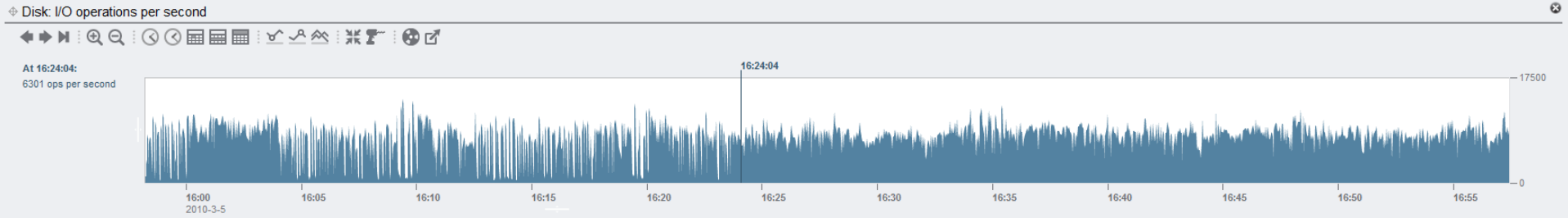
The ZIL can absorb peaks of IOPS, but these still have to make it back to disk.

You need to determine the **sustained** IOPS capacity your disks (not your ZIL). You then need to ensure you stay well, well well below this level.

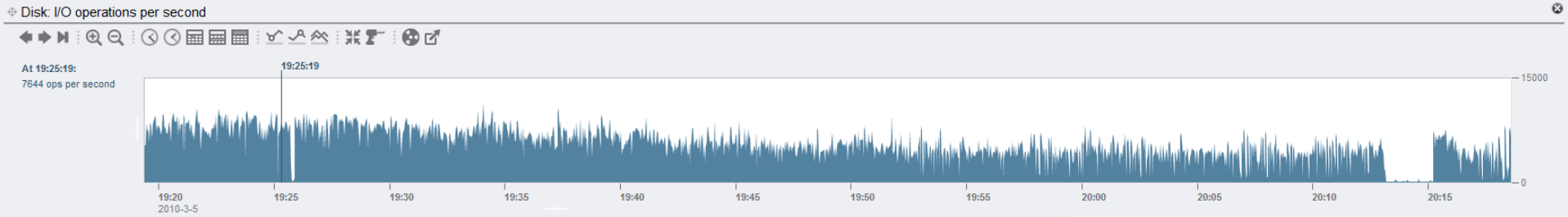
*** You also need enough spare IOPS to resilver your array after a disk failure! ***

ZFS Performance : SAN Gotchas (continued)

When ZFS hits 80% disk usage, disk IOPS shoot up, performance drops:



Deleting data restores performance (but be careful with recursive snapshot deletion!)



2010.Q1 Sun Unified Storage release and OpenSolaris snv_129 should help:

http://blogs.sun.com/dlutz/entry/oltp_improvements_in_2010_q1
http://bugs.opensolaris.org/bugdatabase/view_bug.do?bug_id=6869229

Things to watch out for:

iSCSI initiator buggy prior to Solaris 10 update 7

Zpool Scrubs/resilvers on large arrays slow prior to OpenSolaris snv_129 (bug 6678033)

Large MySQL instances (8GB+) inside Zones can suffer performance issues

System call to get free memory inside a zone can be slow on busy servers (echo 1 | sort)

iSCSI Timeouts can lock ZPools

Odd disk failure modes can go undetected and result in poor performance

OpenSolaris: The Future

Having extensively used OpenSolaris for over a year now, it is in our opinion production ready for a large number of uses.

It's rapidly maturing across the board, and has more new features than you can shake a stick at.

We're currently testing, evaluating and specifying our next generation platform for release this summer, and our intention is to utilise OpenSolaris 2010.03.

The potential for a fully virtualised cloud infrastructure (virtual routers, virtual switches, virtual load balancers) is becoming a reality.

The OpenSolaris team are very quick to respond to bugs filed on http://defect.opensolaris.org and paid support is available from Sun^H^HOracle

OpenSolaris: Killer Features

xVM has matured to a stable, usable state, and we're trialling this to replace Citrix XenServer

Solaris 10 Branded Zones allow a seamless migration to OpenSolaris

Crossbow provides rate limiting, and delivers a full network stack to each Zone, allowing firewalling, routing, snooping, etc.

IPS is a big win (Although why is there still no Exim package?!)

Hoping linux26 Branded Zones mature to production-ready status

VRRP Project of significant interest

Virtual switches via rbridges also of interest

Resources:

We run a UK OpenSolaris IPS Mirror:

pkg set-authority -m http://pkg.osol.mirror.everycity.co.uk opensolaris.org

My Twitter: http://twitter.com/alasdairsolaris

Blogs Worth Reading:
http://blogs.everycity.co.uk/alasdair
http://bugs.opensolaris.org (!)
http://www.c0t0d0s0.org
http://www.cuddletech.com/blog/
http://letsgetdugg.com (Solaris User)
http://chrisgerhard.wordpress.com/
http://blogs.sun.com/dlutz/ (Fishworks Engineer)
http://blogs.sun.com/bmc/ (Fishworks Engineer)
http://blogs.sun.com/pmonday/ (Fishworks Engineer)
(The list goes on...)