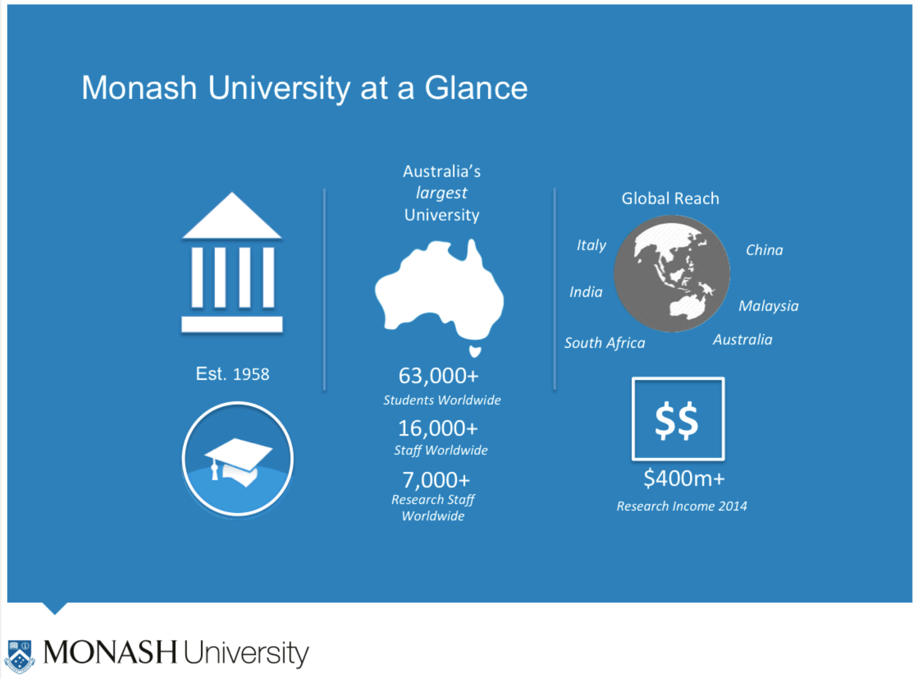
RED HAT
SUMMIT

Ceph as Monash University’s research data engine

Dr Steve Quenette / Blair Bethwaite / Rafael Lopez
Deputy Director / Lead, R@Cmon / Devops engineer
3rd May 2017

Part 1 – The context (Steve)

Monash University



Monash University at a Glance

Est. 1958

Australia's *largest* University

63,000+
Students Worldwide

16,000+
Staff Worldwide

7,000+
Research Staff Worldwide

Global Reach

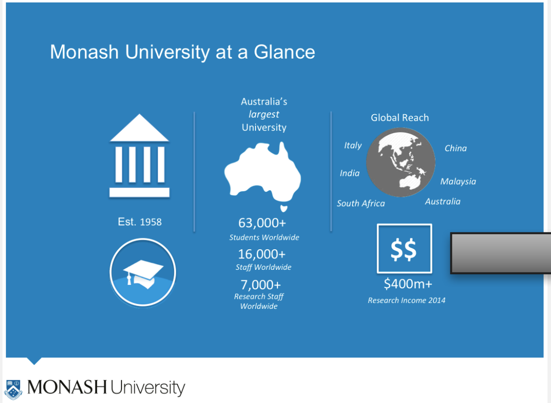
Italy
India
South Africa
China
Malaysia
Australia

$$

$400m+
Research Income 2014

MONASH University

Monash eResearch Centre



Monash University at a Glance

Est. 1958

Australia's *largest* University

63,000+ *Students Worldwide*

16,000+ *Staff Worldwide*

7,000+ *Research Staff Worldwide*

Global Reach

Italy *China* *India* *Malaysia* *South Africa* *Australia*

$$

$400m+ *Research Income 2014*

MONASH University

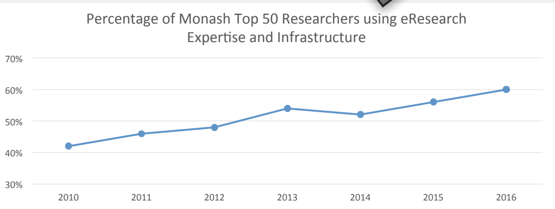
2,000+ researchers use our capabilities: expertise, tools and infrastructure

Another **4,000+** within Monash and around Australia indirectly using our eResearch services

Monash topped both ARC and NHMRC – eResearch underpins are large portion of these grants

Top 6 FOR Codes

09	Engineering	**25.3%**
06	Biological Sciences	**24.4%**
11	Medical and Health Sciences	**13.8%**
03	Chemical Sciences	**11.8%**
02	Physical Sciences	**8.7%**
04	Earth Sciences	**6.1%**



Percentage of Monash Top 50 Researchers using eResearch Expertise and Infrastructure

63+ million
CPU-core hours p.a. of computing time for Monash researchers

Monash University is the largest user of national merit allocated supercomputing time

39+ peak instruments integrated at Monash

National Instrument integration program:
60+ instruments across Australia ($250M+ capital)

5x oversubscription

$3.4m p.a. of research cloud access from contribution of $250k p.a.

10+ petabytes of research storage

Break free from the stereotype

Excerpt from the dedication in Terry Pratchett’s book “Guards! Guards!” …

IT services *HPC center* *the comp. sci. postdoc*

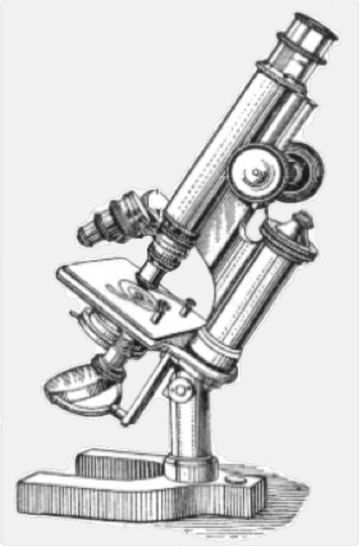
“They may be called the ~~Palace Guard,~~ the ~~City Guard, or~~ the ~~Patrol.~~ Whatever the name, their purpose in any work of heroic fantasy is ~~identical~~: is, round Chapter Three (or ten minutes into the film) ~~to rush into the room,~~ attack the hero one at a time, and be ~~slaughtered~~. No one ever asks them if they wanted to.

seconds after the first ticket *research* *professor / clinician / …*

This book is dedicated to those fine men.”

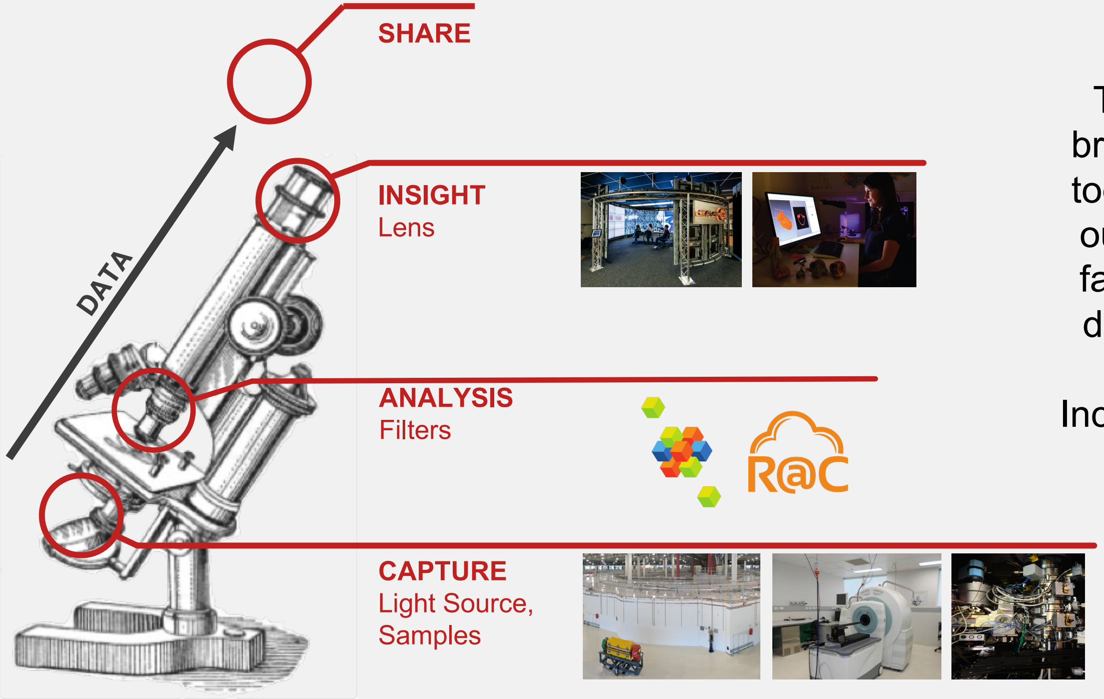
Our business

Can be best described by the humble microscope



The microscope for 21st century discovery

To discover reliably, repeatedly, first…



The working of the brass, that holds it all together, is no longer outsourced and pre-fabricated but rather democratised to the researcher. Increasingly, the brass is software.

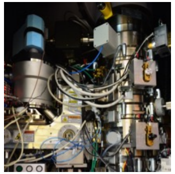
Teaser: Fundamental science

The “microscope” analogy using many “microscope”-like things…

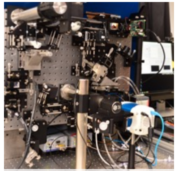
MONASH University

Professor Trevor Lithgow
ARC Australian Laureate Fellow

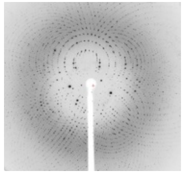
Discovery of new protein transport machines in bacteria, understanding the assembly of protein transport machines, and dissecting the effects of anti-microbial peptides on anti-biotic resistant “super-bugs”



FEI Titan Krios
Nationally funded project to develop environments for Cryo analysis

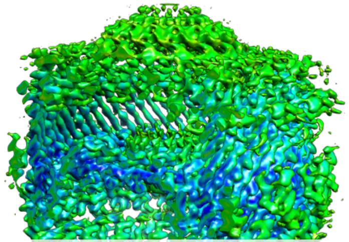


MMI Lattice Light Sheet
Nationally funded project to capture and preprocess LLS data



Synchrotron MX
Store.Synchrotron Data Management

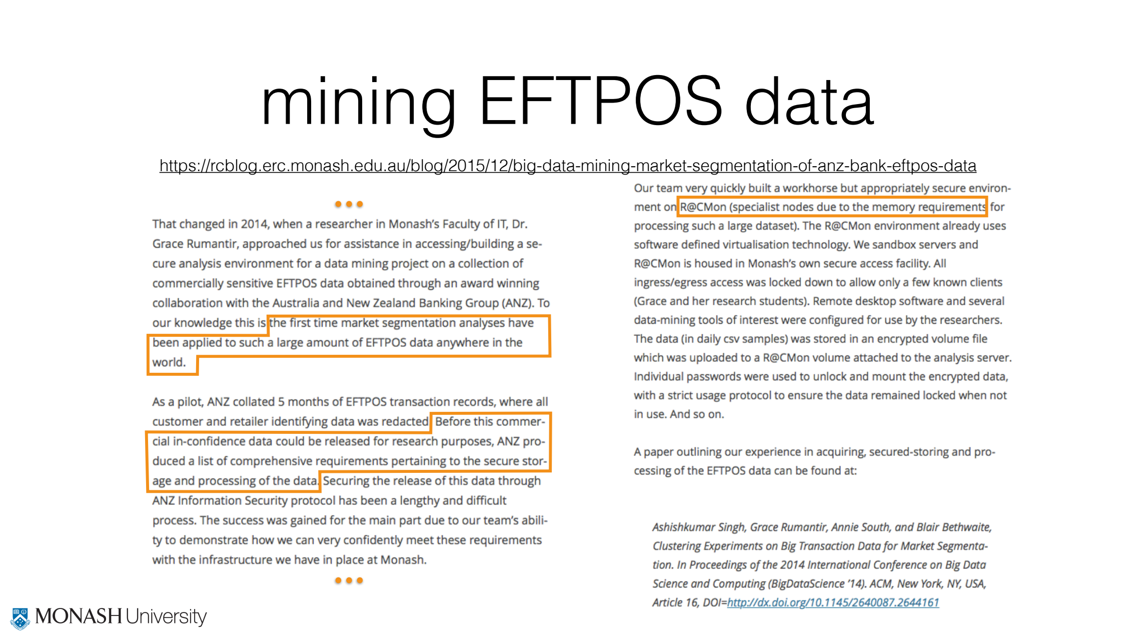
MASSIVE M3
Structural refinement and analysis



Chamber details from the nanomachine that secretes the toxin that causes cholera.
Research and data by Dr. Iain Hay (Lithgow lab)

Teaser: Applied science

The “microscope” embodies the application of novel technologies



mining EFTPOS data

https://rcblog.erc.monash.edu.au/blog/2015/12/big-data-mining-market-segmentation-of-anz-bank-eftpos-data

• • •

That changed in 2014, when a researcher in Monash’s Faculty of IT, Dr. Grace Rumantir, approached us for assistance in accessing/building a secure analysis environment for a data mining project on a collection of commercially sensitive EFTPOS data obtained through an award winning collaboration with the Australia and New Zealand Banking Group (ANZ). To our knowledge this is the first time market segmentation analyses have been applied to such a large amount of EFTPOS data anywhere in the world.

As a pilot, ANZ collated 5 months of EFTPOS transaction records, where all customer and retailer identifying data was redacted. Before this commercial in-confidence data could be released for research purposes, ANZ produced a list of comprehensive requirements pertaining to the secure storage and processing of the data. Securing the release of this data through ANZ Information Security protocol has been a lengthy and difficult process. The success was gained for the main part due to our team’s ability to demonstrate how we can very confidently meet these requirements with the infrastructure we have in place at Monash.

• • •

Our team very quickly built a workhorse but appropriately secure environment on R@CMon (specialist nodes due to the memory requirements for processing such a large dataset). The R@CMon environment already uses software defined virtualisation technology. We sandbox servers and R@CMon is housed in Monash’s own secure access facility. All ingress/egress access was locked down to allow only a few known clients (Grace and her research students). Remote desktop software and several data-mining tools of interest were configured for use by the researchers. The data (in daily csv samples) was stored in an encrypted volume file which was uploaded to a R@CMon volume attached to the analysis server. Individual passwords were used to unlock and mount the encrypted data, with a strict usage protocol to ensure the data remained locked when not in use. And so on.

A paper outlining our experience in acquiring, secured-storing and processing of the EFTPOS data can be found at:

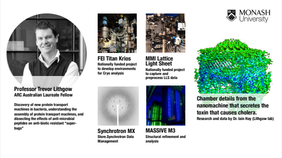
> *Ashishkumar Singh, Grace Rumantir, Annie South, and Blair Bethwaite, Clustering Experiments on Big Transaction Data for Market Segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing (BigDataScience ’14). ACM, New York, NY, USA, Article 16, DOI=http://dx.doi.org/10.1145/2640087.2644161*

MONASH University

Final render frames from “How Man Found Fire” 2016 ©MCLA & Taungurung Dolodanin-dat Animation Group.

The challenge

Many problems



MONASH University

Professor Trevor Lithgow
ARC Australian Laureate Fellow

Discovery of new protein transport machines in bacteria, understanding the assembly of protein transport machines, and dissecting the effects of anti-microbial peptides on anti-biotic resistant "super-bugs"

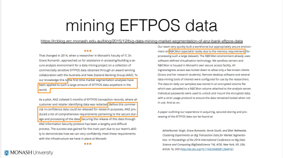
FEI Titan Krios
Nationally funded project to develop environments for Cryo analysis

MMI Lattice Light Sheet
Nationally funded project to capture and preprocess LLS data

Synchrotron MX
Store.Synchrotron Data Management

MASSIVE M3
Structural refinement and analysis

Chamber details from the nanomachine that secretes the toxin that causes cholera.
Research and data by Dr. Iain Hay (Lithgow lab)



mining EFTPOS data

https://rcblog.erc.monash.edu.au/blog/2015/12/big-data-mining-market-segmentation-of-anz-bank-eftpos-data

• • •

That changed in 2014, when a researcher in Monash's Faculty of IT, Dr. Grace Rumantir, approached us for assistance in accessing/building a secure analysis environment for a data mining project on a collection of commercially sensitive EFTPOS data obtained through an award winning collaboration with the Australia and New Zealand Banking Group (ANZ). To our knowledge this is the first time market segmentation analyses have been applied to such a large amount of EFTPOS data anywhere in the world.

As a pilot, ANZ collated 5 months of EFTPOS transaction records, where all customer and retailer identifying data was redacted. Before this commercial in-confidence data could be released for research purposes, ANZ produced a list of comprehensive requirements pertaining to the secure storage and processing of the data. Securing the release of this data through ANZ Information Security protocol has been a lengthy and difficult process. The success was gained for the main part due to our team's ability to demonstrate how we can very confidently meet these requirements with the infrastructure we have in place at Monash.

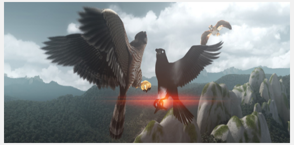
• • •

Our team very quickly built a workhorse but appropriately secure environment on R@CMon (specialist nodes due to the memory requirements for processing such a large dataset). The R@CMon environment already uses software defined virtualisation technology. We sandbox servers and R@CMon is housed in Monash's own secure access facility. All ingress/egress access was locked down to allow only a few known clients (Grace and her research students). Remote desktop software and several data-mining tools of interest were configured for use by the researchers. The data (in daily csv samples) was stored in an encrypted volume file which was uploaded to a R@CMon volume attached to the analysis server. Individual passwords were used to unlock and mount the encrypted data, with a strict usage protocol to ensure the data remained locked when not in use. And so on.

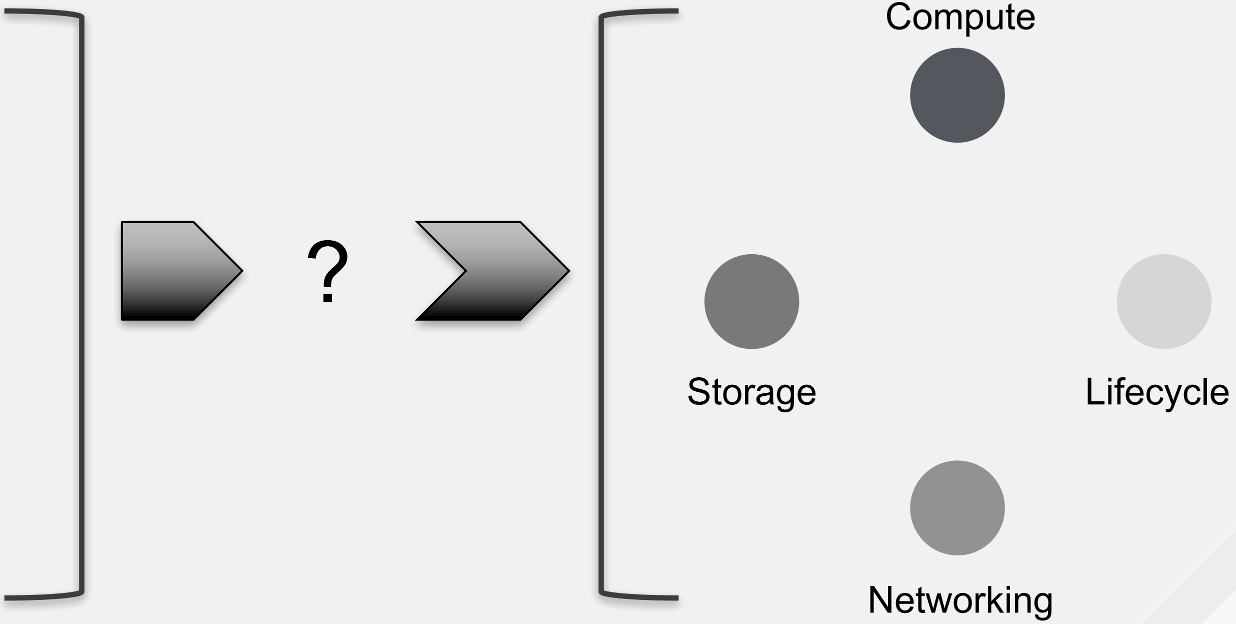
A paper outlining our experience in acquiring, secured-storing and processing of the EFTPOS data can be found at:

Ashishkumar Singh, Grace Rumantir, Annie South, and Blair Bethwaite, Clustering Experiments on Big Transaction Data for Market Segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing (BigDataScience '14). ACM, New York, NY, USA, Article 16, DOI=http://dx.doi.org/10.1145/2640087.2644161

MONASH University



● ● ●



?

Resources at scale

Compute

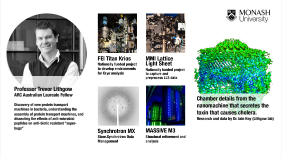
Storage

Lifecycle

Networking

The challenge – compute perspective

Many problems



MONASH University

Professor Trevor Lithgow
ARC Australian Laureate Fellow

Discovery of new protein transport machines in bacteria, understanding the assembly of protein transport machines, and dissecting the effects of anti-microbial peptides on anti-biotic resistant "super-bugs"

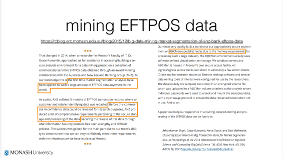
FEI Titan Krios
Nationally funded project to develop environments for Cryo analysis

MMI Lattice Light Sheet
Nationally funded project to capture and preprocess LLS data

Synchrotron MX
Store.Synchrotron Data Management

MASSIVE M3
Structural refinement and analysis

Chamber details from the nanomachine that secretes the toxin that causes cholera.
Research and data by Dr. Iain Hay (Lithgow lab)



mining EFTPOS data

https://rcblog.erc.monash.edu.au/blog/2015/12/big-data-mining-market-segmentation-of-anz-bank-eftpos-data

•••

That changed in 2014, when a researcher in Monash's Faculty of IT, Dr. Grace Rumantir, approached us for assistance in accessing/building a secure analysis environment for a data mining project on a collection of commercially sensitive EFTPOS data obtained through an award winning collaboration with the Australia and New Zealand Banking Group (ANZ). To our knowledge this is the first time market segmentation analyses have been applied to such a large amount of EFTPOS data anywhere in the world.

As a pilot, ANZ collated 5 months of EFTPOS transaction records, where all customer and retailer identifying data was redacted. Before this commercial in-confidence data could be released for research purposes, ANZ produced a list of comprehensive requirements pertaining to the secure storage and processing of the data. Securing the release of this data through ANZ Information Security protocol has been a lengthy and difficult process. The success was gained for the main part due to our team's ability to demonstrate how we can very confidently meet these requirements with the infrastructure we have in place at Monash.

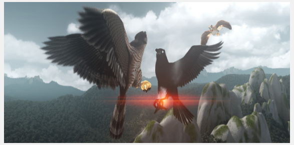
•••

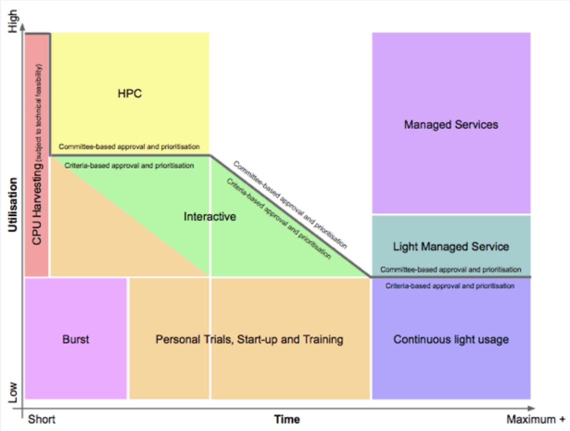
Our team very quickly built a workhorse but appropriately secure environment on R@CMon (specialist nodes due to the memory requirements for processing such a large dataset). The R@CMon environment already uses software defined virtualisation technology. We sandbox servers and R@CMon is housed in Monash's own secure access facility. All ingress/egress access was locked down to allow only a few known clients (Grace and her research students). Remote desktop software and several data-mining tools of interest were configured for use by the researchers. The data (in daily csv samples) was stored in an encrypted volume file which was uploaded to a R@CMon volume attached to the analysis server. Individual passwords were used to unlock and mount the encrypted data, with a strict usage protocol to ensure the data remained locked when not in use. And so on.

A paper outlining our experience in acquiring, secured-storing and processing of the EFTPOS data can be found at:

*Ashishkumar Singh, Grace Rumantir, Annie South, and Blair Bethwaite, Clustering Experiments on Big Transaction Data for Market Segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing (BigDataScience '14). ACM, New York, NY, USA, Article 16, DOI=*http://dx.doi.org/10.1145/2640087.2644161

MONASH University

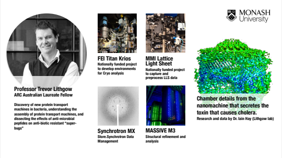



High
Utilisation
Low
CPU Harvesting (subject to technical feasibility)
HPC
Committee-based approval and prioritisation
Criteria-based approval and prioritisation
Interactive
Committee-based approval and prioritisation
Criteria-based approval and prioritisation
Managed Services
Light Managed Service
Committee-based approval and prioritisation
Criteria-based approval and prioritisation
Burst
Personal Trials, Start-up and Training
Continuous light usage
Short
Time
Maximum +

Compute

The challenge – storage perspective

Many problems



MONASH University

Professor Trevor Lithgow
ARC Australian Laureate Fellow

Discovery of new protein transport machines in bacteria, understanding the assembly of protein transport machines, and dissecting the effects of anti-microbial peptides on anti-biotic resistant "super-bugs"

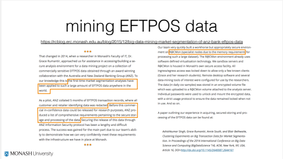
FEI Titan Krios
Nationally funded project to develop environments for Cryo analysis

MMI Lattice Light Sheet
Nationally funded project to capture and preprocess LLS data

Synchrotron MX
Store.Synchrotron Data Management

MASSIVE M3
Structural refinement and analysis

Chamber details from the nanomachine that secretes the toxin that causes cholera.
Research and data by Dr. Iain Hay (Lithgow lab)



mining EFTPOS data

https://rcblog.erc.monash.edu.au/blog/2015/12/big-data-mining-market-segmentation-of-anz-bank-eftpos-data

•••

That changed in 2014, when a researcher in Monash's Faculty of IT, Dr. Grace Rumantir, approached us for assistance in accessing/building a secure analysis environment for a data mining project on a collection of commercially sensitive EFTPOS data obtained through an award winning collaboration with the Australia and New Zealand Banking Group (ANZ). To our knowledge this is the first time market segmentation analyses have been applied to such a large amount of EFTPOS data anywhere in the world.

As a pilot, ANZ collated 5 months of EFTPOS transaction records, where all customer and retailer identifying data was redacted. Before this commercial in-confidence data could be released for research purposes, ANZ produced a list of comprehensive requirements pertaining to the secure storage and processing of the data. Securing the release of this data through ANZ Information Security protocol has been a lengthy and difficult process. The success was gained for the main part due to our team's ability to demonstrate how we can very confidently meet these requirements with the infrastructure we have in place at Monash.

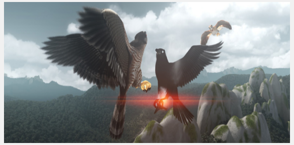
•••

Our team very quickly built a workhorse but appropriately secure environment on R@CMon (specialist nodes due to the memory requirements) for processing such a large dataset). The R@CMon environment already uses software defined virtualisation technology. We sandbox servers and R@CMon is housed in Monash's own secure access facility. All ingress/egress access was locked down to allow only a few known clients (Grace and her research students). Remote desktop software and several data mining tools of interest were configured for use by the researchers. The data (in daily csv samples) was stored in an encrypted volume file which was uploaded to a R@CMon volume attached to the analysis server. Individual passwords were used to unlock and mount the encrypted data, with a strict usage protocol to ensure the data remained locked when not in use. And so on.

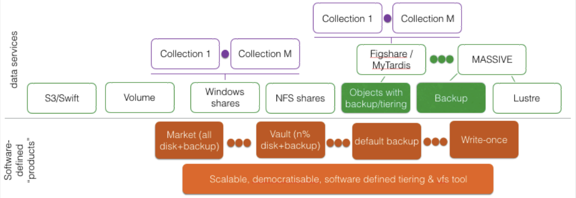
A paper outlining our experience in acquiring, secured-storing and processing of the EFTPOS data can be found at:

Ashishkumar Singh, Grace Rumantir, Annie South, and Blair Bethwaite, Clustering Experiments on Big Transaction Data for Market Segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing (BigDataScience '14). ACM, New York, NY, USA, Article 16, DOI=http://dx.doi.org/10.1145/2640087.2644161

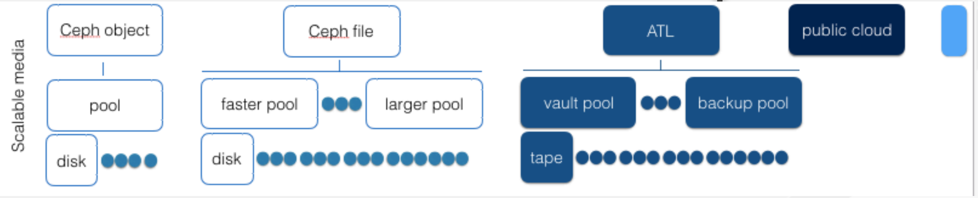
MONASH University



●●●



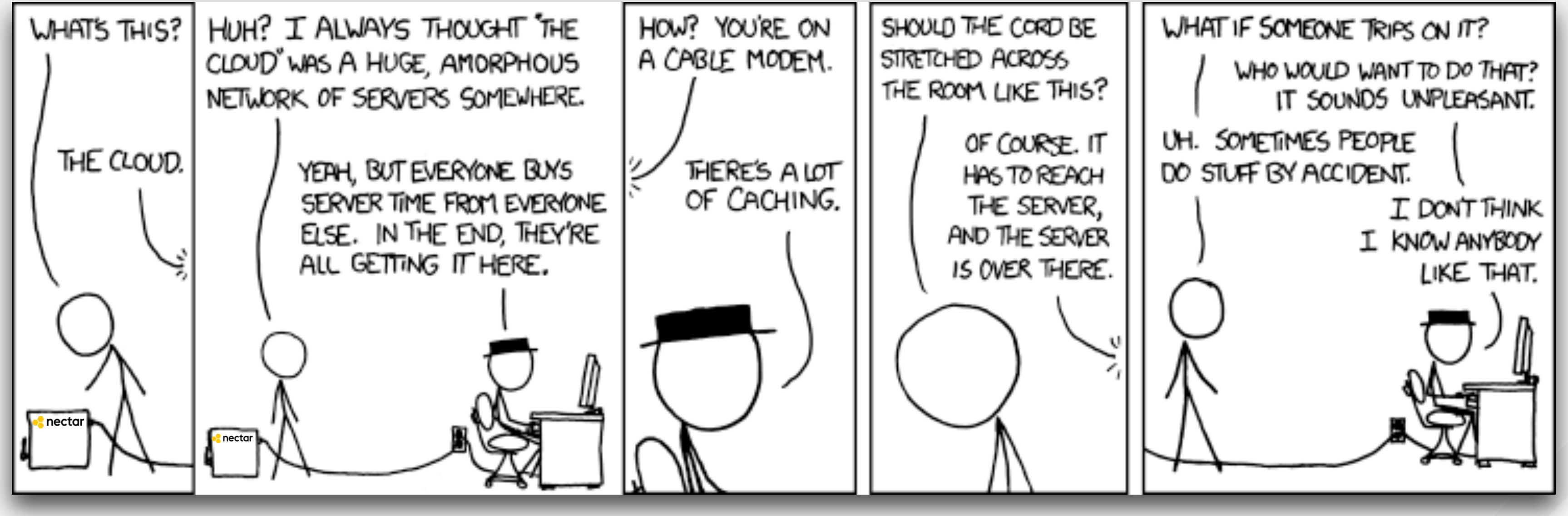
Storage



Part 2: Ceph (Blair)

Ceph@Monash, some history

It all started with the Cloud


WHAT'S THIS?
THE CLOUD.
nectar
HUH? I ALWAYS THOUGHT "THE CLOUD" WAS A HUGE, AMORPHOUS NETWORK OF SERVERS SOMEWHERE.
YEAH, BUT EVERYONE BUYS SERVER TIME FROM EVERYONE ELSE. IN THE END, THEY'RE ALL GETTING IT HERE.
nectar
HOW? YOU'RE ON A CABLE MODEM.
THERE'S A LOT OF CACHING.
SHOULD THE CORD BE STRETCHED ACROSS THE ROOM LIKE THIS?
OF COURSE. IT HAS TO REACH THE SERVER, AND THE SERVER IS OVER THERE.
WHAT IF SOMEONE TRIPS ON IT?
WHO WOULD WANT TO DO THAT? IT SOUNDS UNPLEASANT.
UH. SOMETIMES PEOPLE DO STUFF BY ACCIDENT.
I DON'T THINK I KNOW ANYBODY LIKE THAT.

https://xkcd.com/908/ (Nectar logo added)

Speaking of accidents

- In early 2013 R@CMon started with Monash's first zone of the Nectar research cloud
- Our own local cloud = awesome!
- Hang on… where do we store all the things?
- No persistent storage other than Swift provided by Nectar (expected to come from other funding sources)
- But users don't know how to use object storage and had (still don't really have) any significant application base that can effectively utilise it
- Enter Cuttlefish!
- "monash-01" Cinder zone backed by Ceph available mid 2013
- First of many Ceph deployments within the Nectar federation

Show and tell: “monash-01”

(No, we’re no good with names)

- The hardware - repurposed Swift servers:
    - 8x Dell R720xd (colo osds & mons x5) - 24TB/node
        - 12x 2TB 7.2k NL-SAS (12x RAID0, PERC H710p)
        - 2x E5-2650(2GHz), 32GB RAM
        - 20GbE (Intel X520 DP), VLANs for back/front-end
- Deployed on Ceph Cuttlefish release
- This was enough to make us trust it and want more!

Show and tell: “monash-02”

- 27x Dell R720xd (**3x virtualised mons on separate hardware**)
    - 9x 4/6TB 7.2k NL-SAS (9x RAID0, PERC H710p) – 36/54TB/node
    - **3x 200GB Intel DC S3700 SSDs** (journals and future cache)
    - **1x E5-2630Lv2 (2.4GHz)**, 32GB RAM
    - **2x 10GbE (Mellanox CX-3 DP)**, back/front-end active on alternate ports (different ToR switches)
- Currently Ceph Jewel on Ubuntu Trusty, 2 replicas, ~1PB

Show and tell: “market”

- 3x Dell R320 (mons)
- 4x Dell R720xd (cache tier) - 18TB/node
  - 20x 900GB 10k SAS (20x RAID0, PERC 710p) - rgw hot tier
  - 4x 400GB Intel DC S3700 SSDs (journals for rgw hot tier)
  - 2x E5-2630v2 (2.6GHz), 128GB RAM
  - 56GbE (Mellanox CX-3 DP), VLANs for back/front-end

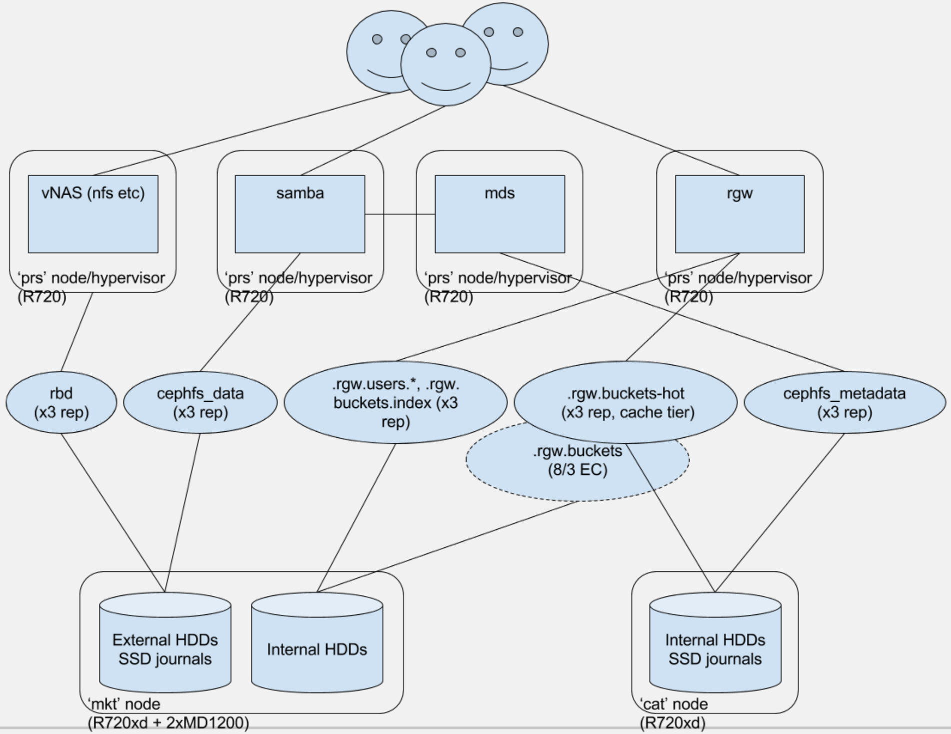
Show and tell: “market”

- 33x Dell R720xd + 66 MD1200 (2 per node) -**144TB/node**
- 8x 6TB 7.2k NL-SAS (8x RAID0, PERC H710p) - rgw EC cold tier
  - 24x 4TB 7.2k NL-SAS (24x RAID0, PERC H810) - rbds go here
  - 4x 200GB Intel DC S3700 SSDs (journals for rbd pool)
  - 2x E5-2630v2 (2.6GHz), 128GB RAM
  - *20GbE (Mellanox CX-3 DP)*, VLANs for back/front-end
- Ceph Hammer on RHEL Maipo

show and tell: “marketv2”

- 2016 expansion
- 18x Dell R730xd -**128TB/node**
- 16x 8TB 7.2k NL-SAS (PERC H330)
- 1x Intel P3700 NVMe 400GB (journals)
- *25GbE (Mellanox CX-3 DP),* VLANs for back/front-end

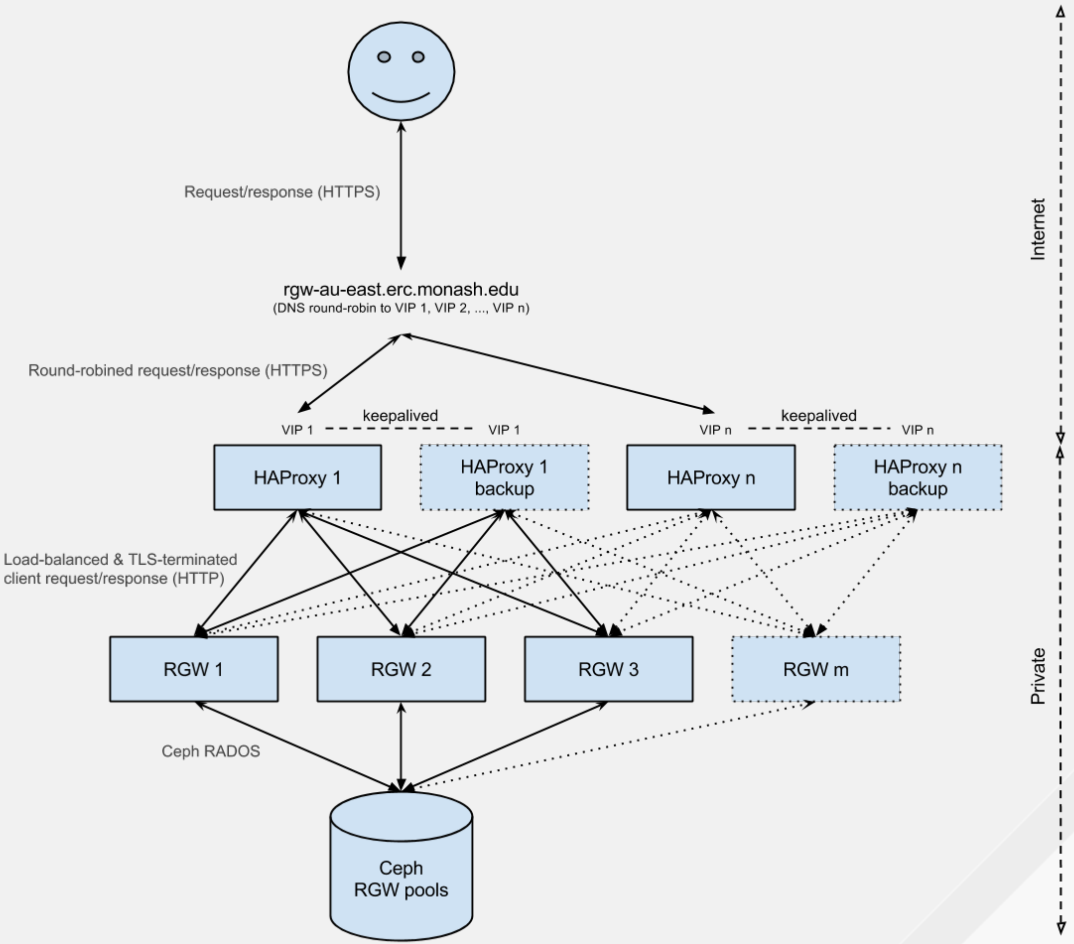
market: logical-to-physical layout


vNAS (nfs etc)
samba
mds
rgw
'prs' node/hypervisor (R720)
'prs' node/hypervisor (R720)
'prs' node/hypervisor (R720)
'prs' node/hypervisor (R720)
rbd (x3 rep)
cephfs_data (x3 rep)
.rgw.users.*, .rgw.buckets.index (x3 rep)
.rgw.buckets-hot (x3 rep, cache tier)
cephfs_metadata (x3 rep)
.rgw.buckets (8/3 EC)
External HDDs SSD journals
Internal HDDs
'mkt' node (R720xd + 2xMD1200)
Internal HDDs SSD journals
'cat' node (R720xd)

- DNS round-robin provides initial HA request fanout

- HAproxys handle load-balancing and SSL/TLS termination and security control point.
- Scale in pairs with keepalived pairing providing redundancy and HA via Virtual/floating IP address (VIP) failover.

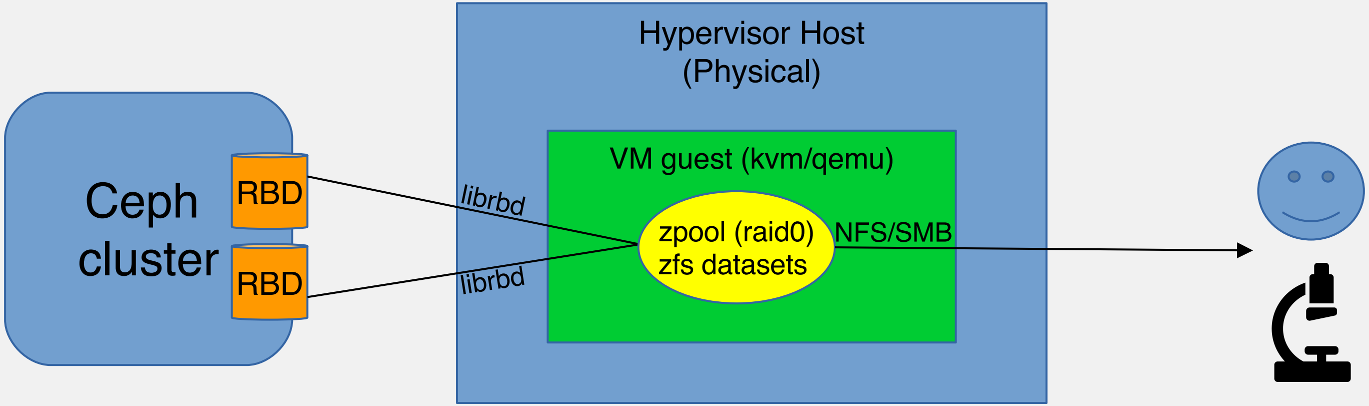
- RGW instances handle actual client/application protocol (S3, Swift, etc) traffic.
- Scale as needed.



Part 3: File storage service over Ceph (Raf)

(The good, the bad, the ugly)

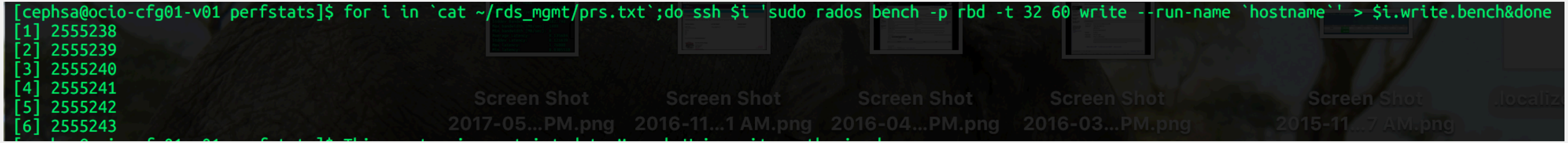
What are we talking about exactly?



Ceph pros – Perf (rados)

The GOOD

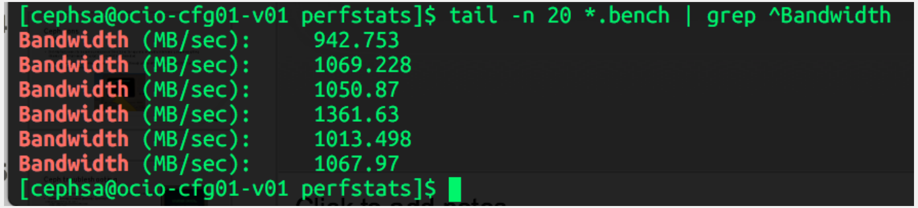
- Nice bandwidth, Eg. Rados bench from 6 simultaneous clients:



```
[cephsa@ocio-cfg01-v01 perfstats]$ for i in `cat ~/rds_mgmt/prs.txt`;do ssh $i 'sudo rados bench -p rbd -t 32 60 write --run-name `hostname`' > $i.write.bench&done
[1] 2555238
[2] 2555239
[3] 2555240
[4] 2555241
[5] 2555242
[6] 2555243
```

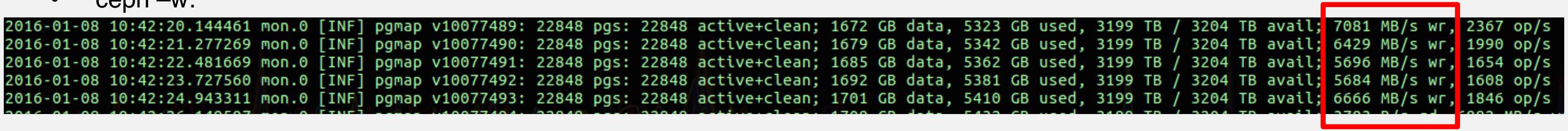
Screen Shot 2017-05...PM.png
Screen Shot 2016-11...1 AM.png
Screen Shot 2016-04...PM.png
Screen Shot 2016-03...PM.png
Screen Shot 2015-11...7 AM.png

- Results:



```
[cephsa@ocio-cfg01-v01 perfstats]$ tail -n 20 *.bench | grep ^Bandwidth
Bandwidth (MB/sec):      942.753
Bandwidth (MB/sec):      1069.228
Bandwidth (MB/sec):      1050.87
Bandwidth (MB/sec):      1361.63
Bandwidth (MB/sec):      1013.498
Bandwidth (MB/sec):      1067.97
[cephsa@ocio-cfg01-v01 perfstats]$
```

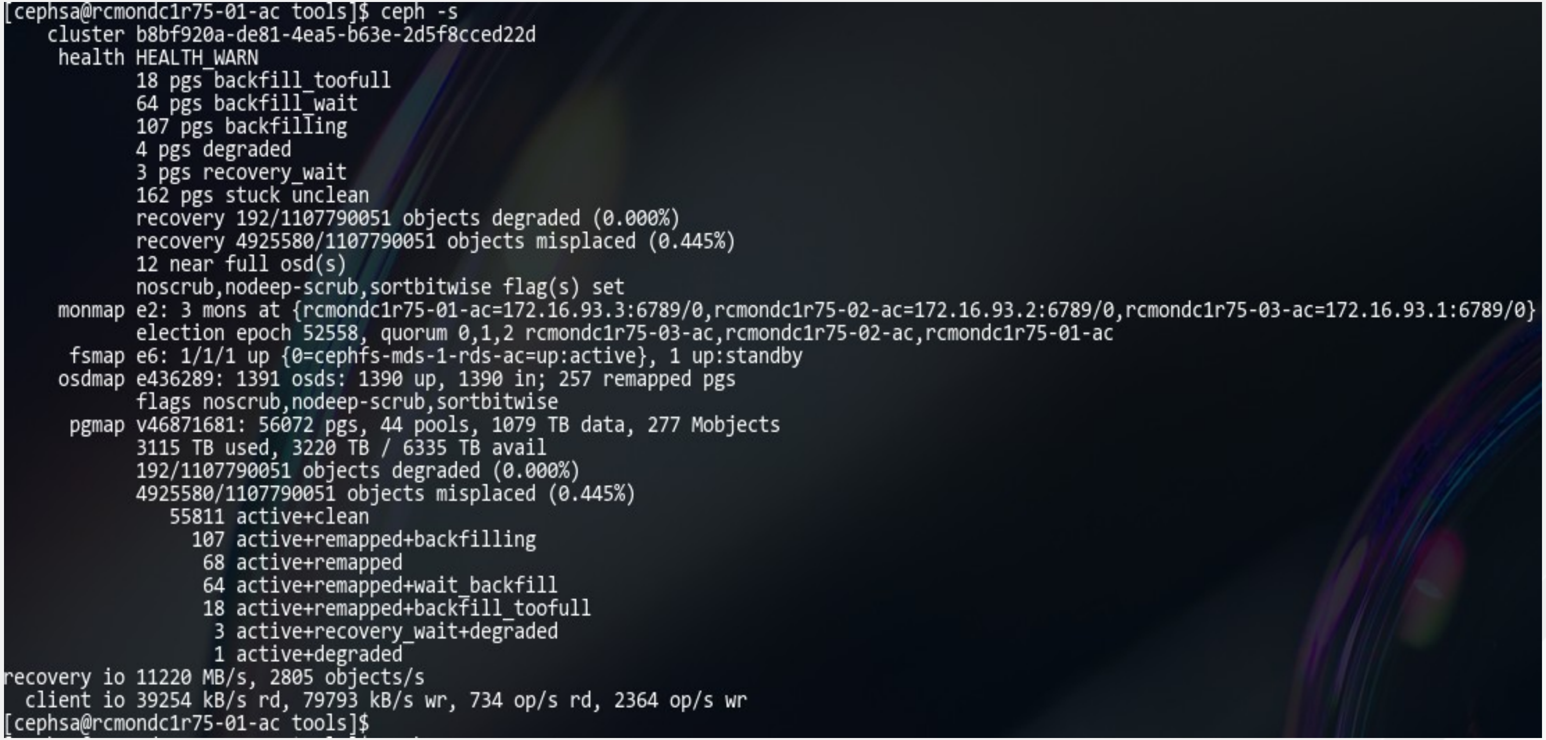
- ceph –w:



```
2016-01-08 10:42:20.144461 mon.0 [INF] pgmap v10077489: 22848 pgs: 22848 active+clean; 1672 GB data, 5323 GB used, 3199 TB / 3204 TB avail; 7081 MB/s wr, 2367 op/s
2016-01-08 10:42:21.277269 mon.0 [INF] pgmap v10077490: 22848 pgs: 22848 active+clean; 1679 GB data, 5342 GB used, 3199 TB / 3204 TB avail; 6429 MB/s wr, 1990 op/s
2016-01-08 10:42:22.481669 mon.0 [INF] pgmap v10077491: 22848 pgs: 22848 active+clean; 1685 GB data, 5362 GB used, 3199 TB / 3204 TB avail; 5696 MB/s wr, 1654 op/s
2016-01-08 10:42:23.727560 mon.0 [INF] pgmap v10077492: 22848 pgs: 22848 active+clean; 1692 GB data, 5381 GB used, 3199 TB / 3204 TB avail; 5684 MB/s wr, 1608 op/s
2016-01-08 10:42:24.943311 mon.0 [INF] pgmap v10077493: 22848 pgs: 22848 active+clean; 1701 GB data, 5410 GB used, 3199 TB / 3204 TB avail; 6666 MB/s wr, 1846 op/s
```

Ceph pros – Perf (recovery)

THE GOOD



```
[cephsa@rcmondc1r75-01-ac tools]$ ceph -s
    cluster b8bf920a-de81-4ea5-b63e-2d5f8cced22d
     health HEALTH_WARN
            18 pgs backfill_toofull
            64 pgs backfill_wait
            107 pgs backfilling
            4 pgs degraded
            3 pgs recovery_wait
            162 pgs stuck unclean
            recovery 192/1107790051 objects degraded (0.000%)
            recovery 4925580/1107790051 objects misplaced (0.445%)
            12 near full osd(s)
            noscrub,nodeep-scrub,sortbitwise flag(s) set
     monmap e2: 3 mons at {rcmondc1r75-01-ac=172.16.93.3:6789/0,rcmondc1r75-02-ac=172.16.93.2:6789/0,rcmondc1r75-03-ac=172.16.93.1:6789/0}
            election epoch 52558, quorum 0,1,2 rcmondc1r75-03-ac,rcmondc1r75-02-ac,rcmondc1r75-01-ac
      fsmap e6: 1/1/1 up {0=cephfs-mds-1-rds-ac=up:active}, 1 up:standby
     osdmap e436289: 1391 osds: 1390 up, 1390 in; 257 remapped pgs
            flags noscrub,nodeep-scrub,sortbitwise
      pgmap v46871681: 56072 pgs, 44 pools, 1079 TB data, 277 Mobjects
            3115 TB used, 3220 TB / 6335 TB avail
            192/1107790051 objects degraded (0.000%)
            4925580/1107790051 objects misplaced (0.445%)
               55811 active+clean
                 107 active+remapped+backfilling
                  68 active+remapped
                  64 active+remapped+wait_backfill
                  18 active+remapped+backfill_toofull
                   3 active+recovery_wait+degraded
                   1 active+degraded
recovery io 11220 MB/s, 2805 objects/s
  client io 39254 kB/s rd, 79793 kB/s wr, 734 op/s rd, 2364 op/s wr
[cephsa@rcmondc1r75-01-ac tools]$
```

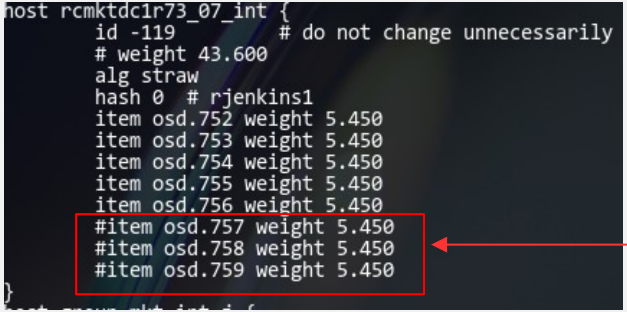
Ceph recovers like a boss. This is recovery from one down and out OSD.

→

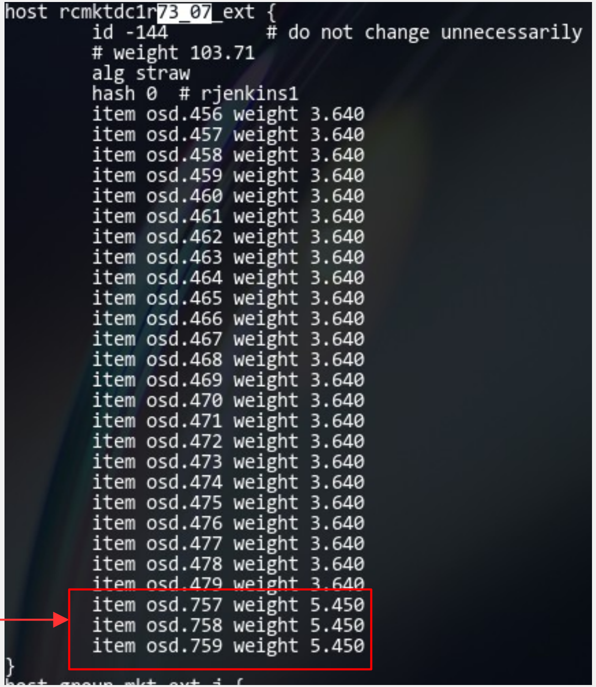
Ceph pros – CRUSH ftw

THE GOOD

- Easy to expand and add nodes/capacity to pools
- Flexible – eg. we reallocated hundreds of TB of capacity from rgw to rbd by moving items within CRUSH map



```
host rcmktdc1r73_07_int {
        id -119         # do not change unnecessarily
        # weight 43.600
        alg straw
        hash 0  # rjenkins1
        item osd.752 weight 5.450
        item osd.753 weight 5.450
        item osd.754 weight 5.450
        item osd.755 weight 5.450
        item osd.756 weight 5.450
        #item osd.757 weight 5.450
        #item osd.758 weight 5.450
        #item osd.759 weight 5.450
}
host group mkt_int_i {
```

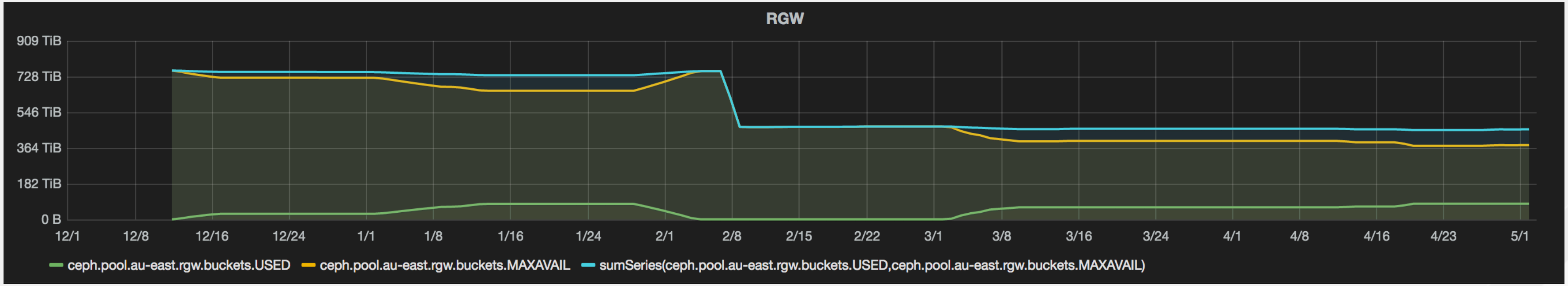


```
host rcmktdc1r73_07_ext {
        id -144         # do not change unnecessarily
        # weight 103.71
        alg straw
        hash 0  # rjenkins1
        item osd.456 weight 3.640
        item osd.457 weight 3.640
        item osd.458 weight 3.640
        item osd.459 weight 3.640
        item osd.460 weight 3.640
        item osd.461 weight 3.640
        item osd.462 weight 3.640
        item osd.463 weight 3.640
        item osd.464 weight 3.640
        item osd.465 weight 3.640
        item osd.466 weight 3.640
        item osd.467 weight 3.640
        item osd.468 weight 3.640
        item osd.469 weight 3.640
        item osd.470 weight 3.640
        item osd.471 weight 3.640
        item osd.472 weight 3.640
        item osd.473 weight 3.640
        item osd.474 weight 3.640
        item osd.475 weight 3.640
        item osd.476 weight 3.640
        item osd.477 weight 3.640
        item osd.478 weight 3.640
        item osd.479 weight 3.640
        item osd.757 weight 5.450
        item osd.758 weight 5.450
        item osd.759 weight 5.450
}
host group mkt_ext_i {
```

Ceph pros – CRUSH ftw

THE GOOD

- Easy to expand and add nodes/capacity to pools
- Flexible – eg. we reallocated hundreds of TB of capacity from rgw to rbd by moving items within CRUSH map



Ceph + Virtual filers

THE GOOD

- librbd integrates flawlessly with libvirt/kvm (NB: update max open files in libvirt to allow for the large number of OSDs!)
- Performance is OK too:

FIO config

```
[global]
ioengine=rbd
clientname=admin
pool=rbd
invalidate=1
ramp_time=5
runtime=30
time_based
direct=1

[write-rbd2-4m-depth16]
rbdname=fio-bench-rbd-2
bs=4m
iodepth=16
rw=write
Stonewall

[read-rbd2-4m-depth16]
rbdname=fio-bench-rbd-2
bs=4m
iodepth=16
rw=read
stonewall
```

Single client results (these scale with more clients)

```
Run status group 5 (all jobs):
  WRITE: io=24600MB, aggrb=839176KB/s, minb=839176KB/s, maxb=839176KB/s, mint=30018msec, maxt=30018msec
Run status group 7 (all jobs):
   READ: io=35412MB, aggrb=1175.7MB/s, minb=1175.7MB/s, maxb=1175.7MB/s, mint=30122msec, maxt=30122msec
```

Ceph + Virtual filers (why ZFS?)

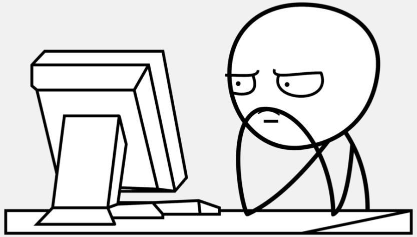
THE GOOD

- CephFS was not quite ready and we needed something to quickly replace aging, out-of-support hardware filers
- Easy to manage hundreds of shares (zfs datasets), eg. inherited parameters, quotas
- Easy to grow
- Use phys host SSD as the slog for sync performance
- Compression (as much as 3.0x on some datasets)
- Deduplication (haven’t had a good use case yet)
- Could have used btrfs but zfs easier and more intuitive to use, proven itself on other platforms, some experience in ops team with Oracle ZFS on Solaris and NAS physical appliance

Ceph cons – operational challenge

THE BAD

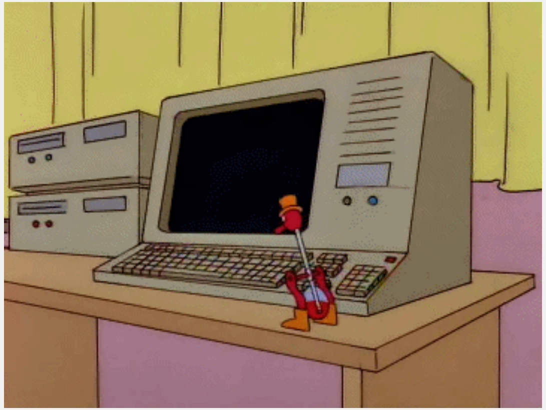
- Heterogeneous hardware configurations = more processes + docs
- Difficult for newcomers - need experience, CRUSH and cluster architecture knowledge to do even basic operations
- Eg. To replace a bad disk you need to know/work out things like:
  - ➢ How to remove associated OSD from cluster (5-6 commands)
  - ➢ How to deploy new OSD – can't just insert new disk in and forget about it
  - ➢ Does the server have more than one controller?
  - ➢ Is the disk internal to the server or on a JBOD?
  - ➢ Does the controller use virtual disks?
  - ➢ Should the new OSD use an SSD journal? (in our cluster it is based on what pool it backs, ec pool uses internal journals) What was the old journal partition?
  - ➢ Manual CRUSH bucket assignment and manual weight setting`
- All the above could probably be solved by DevOps (and time ☹)
- Firmware upgrades still challenging on so many servers



Ceph cons

THE BAD

- CRUSH data distribution not ideal at real world PG numbers, have to manually intervene regularly using OSD reweighting (thanks CERN for reweight script)
- No great management/monitoring tools out of the box, though plenty of info from admin sockets, third party plugins, tools, scripts etc = more time needed to properly productionise
- Once again, DevOps could probably solve these gripes



Ceph troubleshooting

THE UGLY

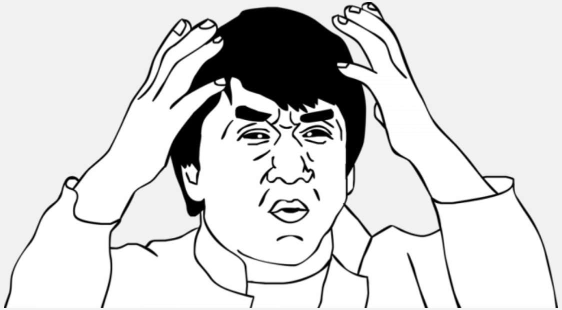
- Ceph status and health output is not easy to digest when there is a problem!
- Some warnings should probably be errors

```
[cephsa@rcmondc1r75-01-ac ~]$ ceph -s
    cluster b8bf920a-de81-4ea5-b63e-2d5f8cced22d
     health HEALTH_WARN
            24 pgs degraded
            4 pgs recovery_wait
            24 pgs stuck unclean
            100 requests are blocked > 32 sec
            recovery 34/1049607262 objects degraded (0.000%)
            2 near full osd(s)
            1/1391 in osds are down
            noscrub,nodeep-scrub,sortbitwise flag(s) set
     monmap e2: 3 mons at {rcmondc1r75-01-ac=172.16.93.3:6789/0,rcmondc1r75-02-ac=172.16.93.2:6789/0,rcmondc1r75-03-ac=172.16.93.1:6789/0}
            election epoch 52558, quorum 0,1,2 rcmondc1r75-03-ac,rcmondc1r75-02-ac,rcmondc1r75-01-ac
      fsmap e6: 1/1/1 up {0=cephfs-mds-1-rds-ac=up:active}, 1 up:standby
     osdmap e434984: 1391 osds: 1390 up, 1391 in
            flags noscrub,nodeep-scrub,sortbitwise
      pgmap v46240327: 56072 pgs, 44 pools, 1051 TB data, 271 Mobjects
            3060 TB used, 3273 TB / 6334 TB avail
            34/1049607262 objects degraded (0.000%)
               56048 active+clean
                  20 active+degraded
                   4 active+recovery_wait+degraded
  client io 80735 kB/s rd, 18639 kB/s wr, 1095 op/s rd, 955 op/s wr
[cephsa@rcmondc1r75-01-ac ~]$ ceph health detail
HEALTH_WARN 24 pgs degraded; 4 pgs recovery_wait; 24 pgs stuck unclean; 100 requests are blocked > 32 sec; 1 osds have slow requests; recovery 34/1049610873 objects degraded (0.000%); 2 near full
osd(s); 1/1391 in osds are down; noscrub,nodeep-scrub,sortbitwise flag(s) set
pg 27.bb is stuck unclean for 2013.439356, current state active+recovery_wait+degraded, last acting [650,664,707,24,628,770,736,634,756,776,714]
pg 27.5c is stuck unclean for 2051.826987, current state active+degraded, last acting [650,627,697,667,779,755,636,690,728,601,715]
…
…
…
```

Ceph troubleshooting

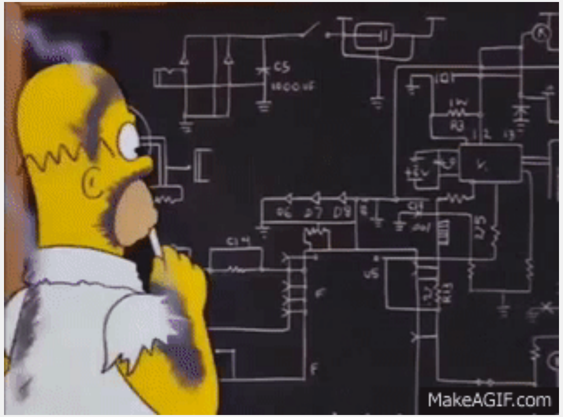
THE UGLY

- Ceph can be hard to troubleshoot
- Many places where things can go wrong and virtually limitless knobs to turn for tuning (this can be good too!) – network, storage devices, operating system/kernel, Ceph application (osd, filestore etc), filesystem…
- A short list of items that have caused varying degrees of pain:
  - networking (vlan IFs not using same mac as HW)
  - kernel parameters
  - CPU frequency (low BW)
  - filestore (slow reqs with default merge/split settings)
  - default backfill/recovery settings (slow perf during recovery)
  - selinux (slow reqs)
  - ulimits
  - Linux net tunables (tcp_sack with stretched cluster)



A few lessons learnt

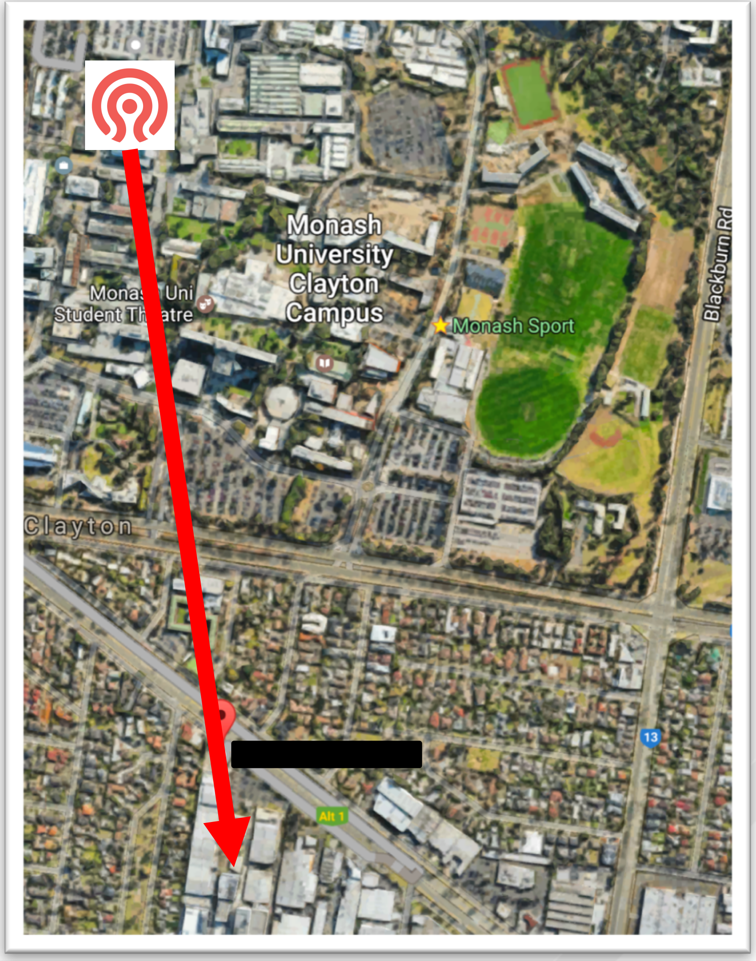
- Getting effective monitoring in place early can assist with identifying source of problems
- Stress your cluster before putting real data onto it, write many objects as perf can change as cluster fills
- Check performance periodically
- Put VM root disks on a pool that has it's own dedicated OSDs, or at least not in a busy pool serving data
- Investing more resources into DevOps early will save a lot of time on ops later – we are behind in this regard



The DC move

THE TASK

- Get out of old on-campus DC facility, move to commercial facility down the road: ~2kms as the crow flies
- Take the primary active research data storage for the university offline for days, in the middle of grant-writing season..?! Not acceptable.
- Can we operate temporarily with a “stretched” cluster?



The DC move

THE GOOD

- New switching fabric built and VLANs stretched over new DC and old core routers
- Physically moved 40+ servers/~5PB from one DC to another without end users noticing!
- Done piecemeal by setting ‘noout’, moving 1 part of the failure domain (crush leaf) at a time (custom crush bucket type called hostgroup = pair of servers)
- User-facing services: NAS servers and RGW live-migrated
- Provided network is in good shape, Ceph can easily operate as an inter-DC cluster!
- Trucks have poor RTT but amazing window size - overall effective bandwidth of ~200TB/day!

The DC move

THE BAD

- Getting off plane whilst on leave to see this:

- After much head scratching we found a major bug in new Cisco 9K router’s fabric module – “filtering” ARP replies between some host pairs

- Discovered tcp_sack is very important with extra network hop latency between OSDs

```
root@rcceph02-03-ac:~# ceph -s
  cluster f85befce-ca0d-4928-96aa-df385d7b67db
  health HEALTH_ERR
  122 pgs are stuck inactive for more than 300 seconds
  902 pgs backfill_wait
  91 pgs backfilling
  10088 pgs degraded
  122 pgs down
  122 pgs peering
  19 pgs recovering
  2 pgs recovery_wait
  122 pgs stuck inactive
  10088 pgs stuck unclean
  10069 pgs undersized
  recovery 42027838/114137365 objects degraded (36.822%)
  recovery 3677865/114137365 objects misplaced (3.222%)
  recovery 23/38048863 unfound (0.000%)
  crush map has legacy tunables (require bobtail, min is firefly)
  monmap e2: 3 mons at {rcceph02-01-ac=172.16.85.248:6789/0,rcceph02-02-ac=172.16.8
  election epoch 7862, quorum 0,1,2 rcceph02-03-ac,rcceph02-02-ac,rcceph02-01-ac
  fsmap e1120: 0/0/1 up
  osdmap e204752: 243 osds: 144 up, 144 in; 3727 remapped pgs
  pgmap v64021317: 13072 pgs, 15 pools, 143 TB data, 37157 kobjects
  273 TB used, 381 TB / 654 TB avail
  42027838/114137365 objects degraded (36.822%)
  3677865/114137365 objects misplaced (3.222%)
  23/38048863 unfound (0.000%)
  9073 active+undersized+degraded
  2862 active+clean
  902 active+undersized+degraded+remapped+wait_backfill
  122 down+peering
  91 active+undersized+degraded+remapped+backfilling
  19 active+recovering+degraded
  2 active+recovery_wait+undersized+degraded
  1 active+undersized+degraded+remapped
recovery io 3586 MB/s, 909 objects/s
  client io 11958 B/s rd, 3847 kB/s wr, 2 op/s rd, 969 op/s wr
```

Ceph @ Monash future

- We are testing CephFS for production use as a highly available (CTDB, ganesha) file storage service with clients connecting via NFS/SMB and CephFS kernel driver/fuse client
- Find a solution to HSM CephFS data
- Ceph-ansible – we have only used ceph-deploy until now, supplemented with parallel ssh tools to manage node deployment, upgrades, ops
- OpenStack-ify presentation layer for easier and consistent management
- Rados classes for data-processing

RED HAT SUMMIT

THANK YOU

plus.google.com/+RedHat

facebook.com/redhatinc

linkedin.com/company/red-hat

twitter.com/RedHatNews

youtube.com/user/RedHatVideos

RED HAT
SUMMIT

LEARN. NETWORK.
EXPERIENCE
OPEN SOURCE.

#redhat #rhsummit

redhat.