Thumpers au CC-IN2P3

Centre de Calcul de l'IN2P3
X. Canehan, L. Tortay

Utilisation
ZFS
Performances

dapnia
cea
saclay

CNRS
CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

À voir

- Acquisition et utilisation des Thumpers
- ZFS de long en large
- Tests et Perfs
- À l'usage...

Utilisation au CC-IN2P3

- 108 en production :
  - 65 machines pour dCache
  - 19 machines pour Xrootd
  - 12 machines pour HPSS
  - 11 machines SRB
  - 1 développement (pour pièces)
- 38 machines juste câblées, toutes pour dCache

Appels d'offres

- Réponse de Sun à 400 To utiles en 2006 : 47 X4500
- Appel d'offre 2007 pour 1600 To utiles (dont une option de 400 To) : 85 X4500
- Achat de 14 machines supplémentaires en fin d'année 2007.
- 146 X4500 : la plus grosse installation mondiale ?

Appel d'offres 2006, détails

- 400 TBytes
  - Xrootd
    - 1920 IOps (30 MBytes/s) random read (16 kB blocks) & 80 MB/s write per server (1 MByte blocks)
    - minimum de 10 TB par serveur
    - read/write ratio : 95/5
  - dCache : profil d'IO moins précis, gros blocs
- DAS, petits SANs, SAN non obligatoire
- Serveurs avec simple ou double (trunked) Gbit/s Ethernet

Thumpers 2006

- 47 Thumpers
  - 35 dCache
  - 8 Xrootd
  - 2 HPSS test, 1SRB, 1 Spare
- 8 X4500 remplacent et étendent le service de 20 serveurs Xrootd

Appel d'offres 2007, détails

- 1.2 PB à 1.6 PB pour Xrootd et dCache
  - Disque et serveurs
  - DAS, ~ 20 TB par serveur
  - Xrootd : 1920 IOps (30 MBytes/s) random read (16 kBytes blocks) & 80 MBytes/s write par server (1 MByte blocks) par serveur
  - read/write ratio: 95/5

Thumpers 2007

- Total de 16 armoires de 9 machines, 17ème incomplète

Thumpers et ZFS

- Thumpers conçus pour fonctionner avec ZFS : 6 contrôleurs « basiques » gèrent 8 disques chacun
- Utilisation possible avec Solaris et SVM (configuration HPSS)
- Quelques machines avec Linux et LVM2

ZFS au Centre de Calcul de l'IN2P3

- Première présentation à LISA 2003 (San Diego) parmi les nouveautés de Solaris X
- Disponibilité réelle 2005 (OpenSolaris)
- Au CC-IN2P3
  - ZFS est toujours associé aux X4500
  - Première utilisation 2006

ZFS, "The last word in file systems"

- Système de fichiers journalisé avec des fonctionnalités de LVM
- Raison d'être : disques capacitifs peu chers, volumes de données importants
- Besoin de :
  - Fiabilité
  - Simplicité de configuration et utilisation
  - « Scalabilité »

ZFS

- Fiabilité : RAID-Z, *end-to-end data integrity*, *parity check on read*, *scrub*, *copy-on-write*
- Simplicité : deux commandes pour gérer les volumes et les systèmes de fichiers
- Scalabilité : système de fichiers 128 bits (zettaoctets), mais en pratique :
  - 2^64 octets maxi (16 Eo) par fichier et système de fichiers
  - 2^78 octets maxi (256 Zo) pour un *pool*

ZFS - Concepts (1)

- *pools* : ensemble de périphériques (*vdevs*) sur lesquels sont distribuées les données (*dynamic striping*)
- *vdevs* : disques simples (entiers ou pas), volumes logiques (RAID non ZFS sous-jacent) ou redondance ZFS (RAID-1, RAID-Z, RAID-Z2)
- *RAID-Z* : RAID-5 à largeur de bande variable
- *RAID-Z2* : RAID-Z double parité (RAID-6)

ZFS - Concepts (2)

- Tous les systèmes de fichiers d'un pool sont répartis sur les vdevs du pool (*striping*)
- Contrairement aux LVM classiques : l'allocation des blocs aux systèmes de fichiers est totalement gérée par ZFS
- Tous les systèmes de fichiers d'un pool partagent l'espace du pool

ZFS - Fonctionnalités (1)

- ACLs évoluées type NFSv4/CIFS
- Transactions
- Compression des données (par système de fichiers)
- Exports NFS « auto-magiques »
- Héritage des propriétés
- « Copy on Write » (COW)
- « Parity check on read » : plus exactement « Checksum check on read »
- « End-to-end data integrity »

ZFS - Fonctionnalités (2)

- Points de montages & systèmes de fichiers
- RAID-Z
- *Resilvering* : correction des ***données*** détectées endommagées lors des scrubs (ou après une panne)
- *Scrub* : vérification de l'intégrité des ***données***
- *Snapshots* : habituels, moins courant : *rollback*
- Transactions : données sur disque *toujours* cohérentes
- Quotas : habituels, moins courant : réservation
- *zvols*

ZFS - Fonctionnalités apparues ultérieurement

- *Hot-spares*
- RAID-Z2
- *Ditto blocks* : méta-données en double ou plus, maintenant étendu aux données
- *Snapshot send/receive*
- Clones (promotion)
- zvols *iSCSI*

ZFS - Exemples (1)

- ```
  zpool create -m /home homes \
        mirror c0t0d0 c1t0d0 c2t0d0
        mirror c0t1d0 c1t1d0 c2t1d0
        mirror c0t2d0 c1t2d0 c2t2d0
  ```
- `zfs create homes/loic`
- `zfs create homes/xavier`
- `zfs set reservation=1T homes/loic`
- `zfs set quota=100M homes/xavier`
  ```

ZFS - Exemples (2)

```
# zpool create -f data \
    raidz c0t0d0 c1t0d0 c4t0d0        c6t0d0 c7t0d0 \
    raidz c0t1d0 c1t1d0 c4t1d0 c5t1d0 c6t1d0 c7t1d0 \
    raidz c0t2d0 c1t2d0 c4t2d0 c5t2d0 c6t2d0 c7t2d0 \
    raidz c0t3d0 c1t3d0 c4t3d0 c5t3d0 c6t3d0 c7t3d0 \
    raidz c0t4d0 c1t4d0 c4t4d0        c6t4d0 c7t4d0 \
    raidz c0t5d0 c1t5d0 c4t5d0 c5t5d0 c6t5d0 c7t5d0 \
    raidz c0t6d0 c1t6d0 c4t6d0 c5t6d0 c6t6d0 c7t6d0 \
    raidz c0t7d0 c1t7d0 c4t7d0 c5t7d0 c6t7d0 c7t7d0
#
```

```
% zpool list; zfs list; df -h /data; df -hFzfs
NAME                    SIZE    USED   AVAIL    CAP  HEALTH     ALTROOT
data                   20.8T    197K   20.8T     0%  ONLINE     -
NAME                   USED  AVAIL  REFER  MOUNTPOINT
data                   154K  16.9T  63.9K  /data
Filesystem             Size   Used  Available Capacity  Mounted on
data                    17T    63K        17T     1%    /data
Filesystem             Size   Used  Available Capacity  Mounted on
data                    17T    63K        17T     1%    /data
%
```

ZFS - Exemples (3)

- ```
  # zfs create data/loic; zfs list; zfs set reservation=2T data/loic; zfs list;
  df -hFzfs
  NAME                    USED  AVAIL  REFER  MOUNTPOINT
  data                    205K  16.9T  64.6K  /data
  data/loic              40.7K  16.9T  40.7K  /data/loic
  NAME                    USED  AVAIL  REFER  MOUNTPOINT
  data                   2.00T  14.9T  64.6K  /data
  data/loic              40.7K  16.9T  40.7K  /data/loic
  Filesystem              Size   Used  Available Capacity  Mounted on
  data                     17T    65K        15T     1%    /data
  data/loic                17T    40K        17T     1%    /data/loic
  ```

- ```
  # zfs set mountpoint=/loic data/loic
  % zfs list; df -hFzfs
  NAME                    USED  AVAIL  REFER  MOUNTPOINT
  data                   2.00T  14.9T  63.0K  /data
  data/loic              40.7K  16.9T  40.7K  /loic
  Filesystem              Size   Used  Available Capacity  Mounted on
  data                     17T    63K        15T     1%    /data
  data/loic                17T    40K        17T     1%    /loic
  %
  ```

ZFS - Exemples (4)

```
# zfs set quota=3T data/loic
% zfs list; df -hFzfs
NAME                    USED  AVAIL  REFER  MOUNTPOINT
data                   2.00T  14.9T  63.0K  /data
data/loic              40.7K  3.00T  40.7K  /loic
Filesystem              Size   Used  Available Capacity  Mounted on
data                     17T    63K        15T     1%    /data
data/loic               3.0T    41K       3.0T     1%    /loic
%
```

```
# zfs unmount /loic
# zfs destroy data/loic
% zfs list; df -hFzfs
NAME                    USED  AVAIL  REFER  MOUNTPOINT
data                    152K  16.9T  62.4K  /data
Filesystem              Size   Used  Available Capacity  Mounted on
data                     17T    62K        17T     1%    /data
%
```

ZFS - Bonnes pratiques RAID

- Bonnes pratiques habituelles
- Sun déconseille les vdevs de plus de 9 disques
- ZFS préfère les *JBODs*
- Pour les entrées/sorties dominées par les écritures ou gros blocs : RAID-Z ou RAID-Z2
- Pour les entrées/sorties dominées par les lectures de petits blocs : mirroirs (RAID-1)

ZFS – Corruptions de données

- Aucune détectée sur les X4500
- ZFS à DESY :
  - aucune corruption sur quelques dizaines de Thumpers
  - détection et correction de corruptions sur d'autres types de matériel (cartes RAID Areca)
- Tests de corruption active au CC
  - détection et correction comme attendues

ZFS et Thumpers

- Comportent 6 contrôleurs, gérant chacun 8 disques
- Trouver le bon équilibre entre :
  - sécurité des données
  - fiabilité globale
  - espace utile
  - performances
- Sur les X4500, la configuration proposée par Sun est très équilibrée

ZFS au CC-IN2P3

- Principalement dCache et Xrootd :
  - Un seul système de fichiers POSIX par pool
  - pas d'export NFS ou iSCSI, pas de zvols
  - très peu de quotas, pas de réservations
  - pas d'ACLs
  - généralement RAID-Z (6x5+P + 2x4+P) sans hot-spare, aussi RAID-Z (5x8+P+ S)
- HPSS : configuration « petits fichiers »
  - GNU/Linux
  - Solaris 10

~30 configurations testées

- 34 configurations testées
  - peu avec GNU/Linux et LVM2
  - quelques unes avec Solaris Volume Manager
  - beaucoup avec ZFS
- Tests effectués avec un benchmark maison
  - simule une charge classique CC-IN2P3
  - reconstruit une utilisation disque classique

Pourquoi tant de tests ?

Configuration 1

Disks	Controllers c5	c4	c7	c6	c1	c0
t7d0	v8	v8	v8	v8	v8	v8
t6d0	v7	v7	v7	v7	v7	v7
t5d0	v6	v6	v6	v6	v6	v6
t4d0	*Sys2*	v5	v5	v5	v5	v5
t3d0	v4	v4	v4	v4	v4	v4
t2d0	v3	v3	v3	v3	v3	v3
t1d0	v2	v2	v2	v2	v2	v2
t0d0	*Sys1*	v1	v1	v1	v1	v1

Balance load on controllers and minimize impact of a single failing controll
Unbalanced raidz vdev: 6x 5+P, 2x 4+P

Configuration 2

Disks	Controllers c5	c4	c7	c6	c1	c0
t7d0	v1	v2	v3	v4	v5	v6
t6d0	v1	v2	v3	v4	v5	v6
t5d0	v1	v2	v3	v4	v5	v6
t4d0	*Sys2*	v2	v3	v4	v5	v6
t3d0	v1	v2	v3	v4	v5	v6
t2d0	v1	v2	v3	v4	v5	v6
t1d0	v1	v2	v3	v4	v5	v6
t0d0	*Sys1*	v2	v3	v4	v5	v6

Balance load on controllers, no resilience to controller failure
Unbalanced raidz vdevs: 5x 7+P, 1x 5+P

Configuration 33

Disks	Controllers c5	c4	c7	c6	c1	c0
t7d0	v1	v1	v1	v1	v1	v1
t6d0	v1	v1	v1	v2	v2	v2
t5d0	v2	v2	v2	v2	v2	v2
t4d0	v3	v3	v3	v3	v3	v3
t3d0	v3	v3	v3	v4	v4	v4
t2d0	v4	v4	v4	v4	v4	v4
t1d0	v5	v5	v5	v5	v5	v5
t0d0	*Sys1*	v5	v5	v5	*Spare*	*Sys2*

Mediocre load balance on controllers, no resillience to controller failure
Balanced raidz vdevs: 5x 8+P, 1 spare
Alternative layout for Configuration 6, better balanced, second system disk moved to c

Configuration 10

Disks	Controllers c5	c4	c7	c6	c1	c0
t7d0	v2	v2	v2	v2	v2	v2
t6d0	v2	v2	v2	v2	v2	v2
t5d0	v2	v2	v2	v2	v2	v2
t4d0	*Sys2*	v2	v2	v2	v2	v2
t3d0	v1	v1	v1	v1	v1	v1
t2d0	v1	v1	v1	v1	v1	v1
t1d0	v1	v1	v1	v1	v1	v1
t0d0	*Sys1*	v1	v1	v1	v1	v1

Almost balanced load on controllers, maximise usable disk space, only 2 security disks
Balanced raidz vdevs: 2x 22+P

Configuration 15

Disks	Controllers c5	c4	c7	c6	c1	c0
t7d0	v1	v2	v3	v4	v5	v6
t6d0	v6	v1	v2	v3	v4	v5
t5d0	v5	v6	v1	v2	v3	v4
t4d0	*Sys2*	v7	v8	v9	v10	v11
t3d0	v12	v13	v14	v15	v7	v8
t2d0	v9	v10	v11	v12	v13	v14
t1d0	v15	*Spare*	v7	v8	v9	v10
t0d0	*Sys1*	v11	v12	v13	v14	v15

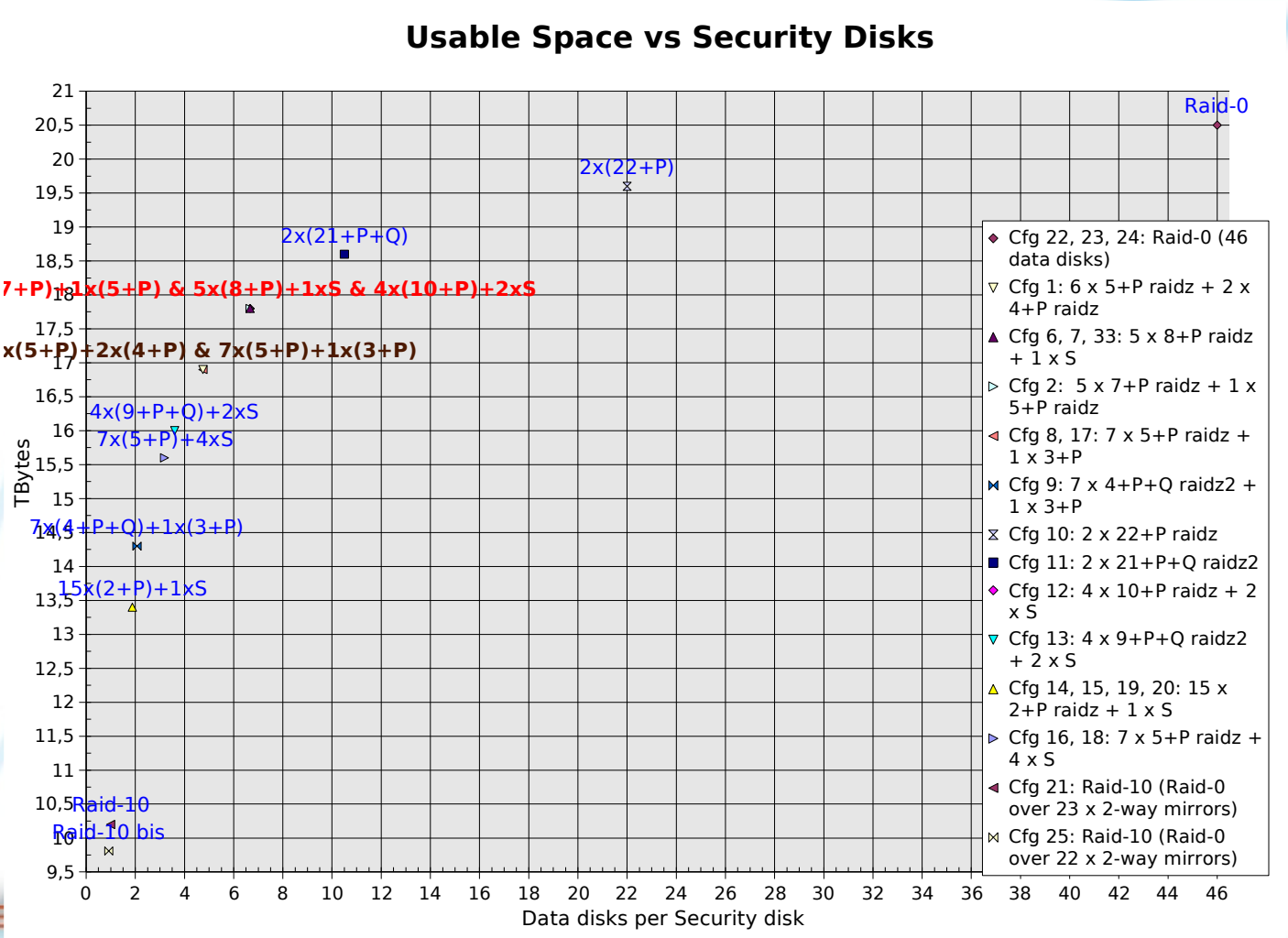
Random read performance oriented configuration, usable space minimisation (16 secu
Balanced raidz vdevs: 15x 2+P, 1 spare
Good resilience to single controller failure
Alternative layout for Configuration 14

Configuration 23

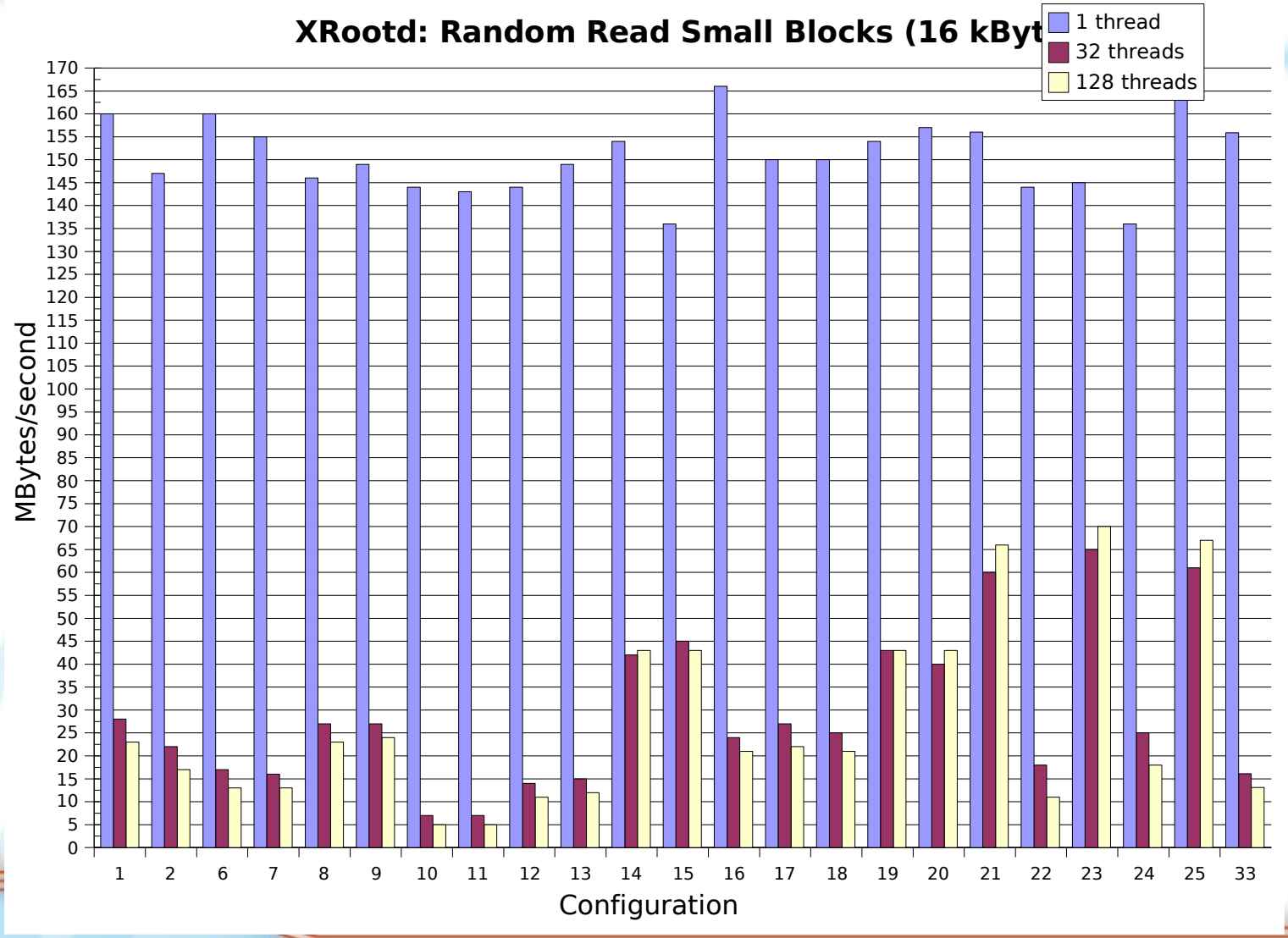
Disks	Controllers c5	c4	c7	c6	c1	c0
t7d0	v1	v2	v3	v4	v5	v6
t6d0	v7	v8	v9	v10	v11	v12
t5d0	v13	v14	v15	v16	v17	v18
t4d0	v19	v20	v21	v22	v23	v24
t3d0	v25	v26	v27	v28	v29	v30
t2d0	v31	v32	v33	v34	v35	v36
t1d0	v37	v38	v39	v40	v41	v42
t0d0	*Sys1*	v43	v44	v45	v46	*Sys2*

Usable space maximisation (0 security disks)
Single zpool made of the 46 data disks as independant vdevs
Second system disk moved to c0
No raidz or raidz2

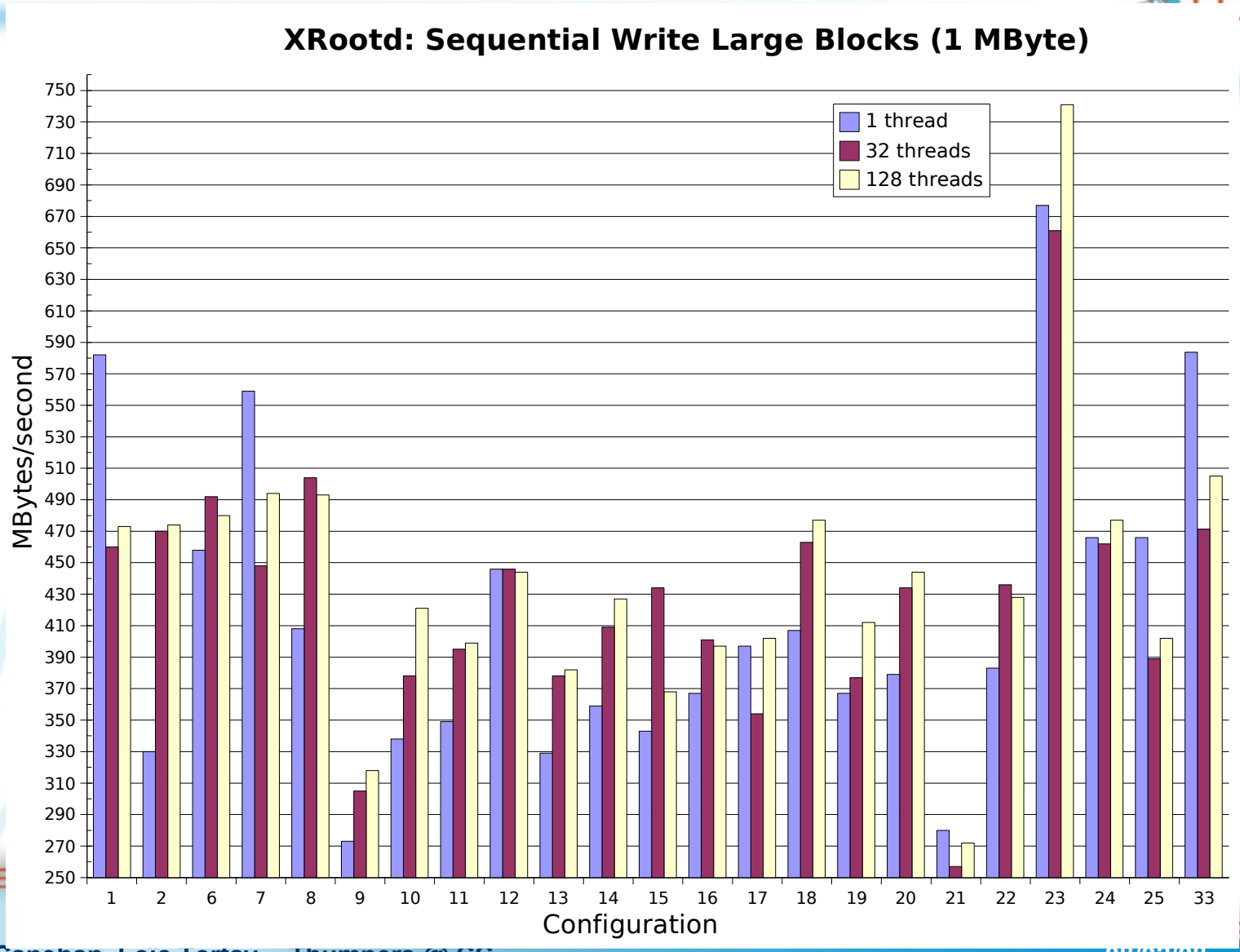
Espace ou sécurité



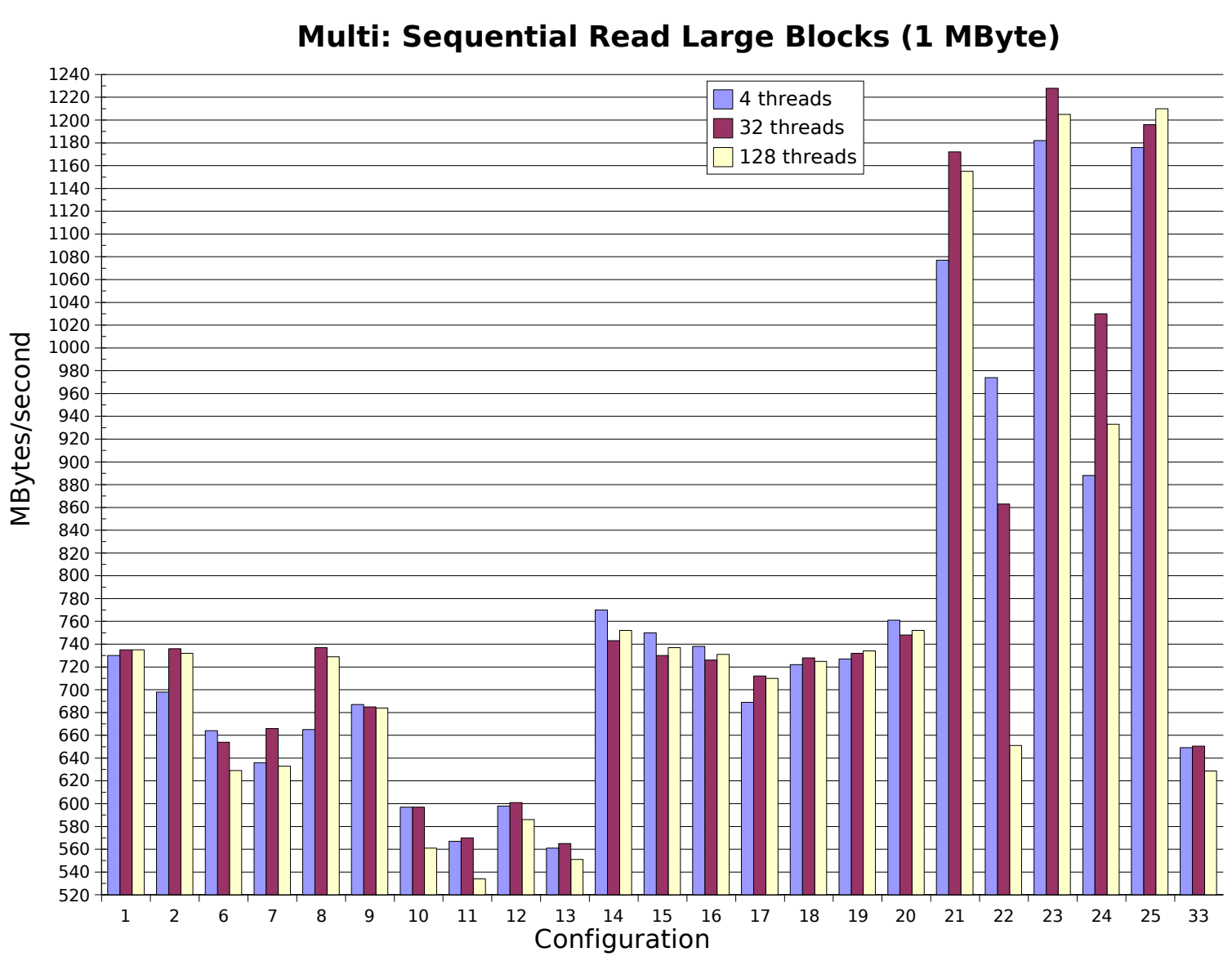
Lecture aléatoire, 16 KB



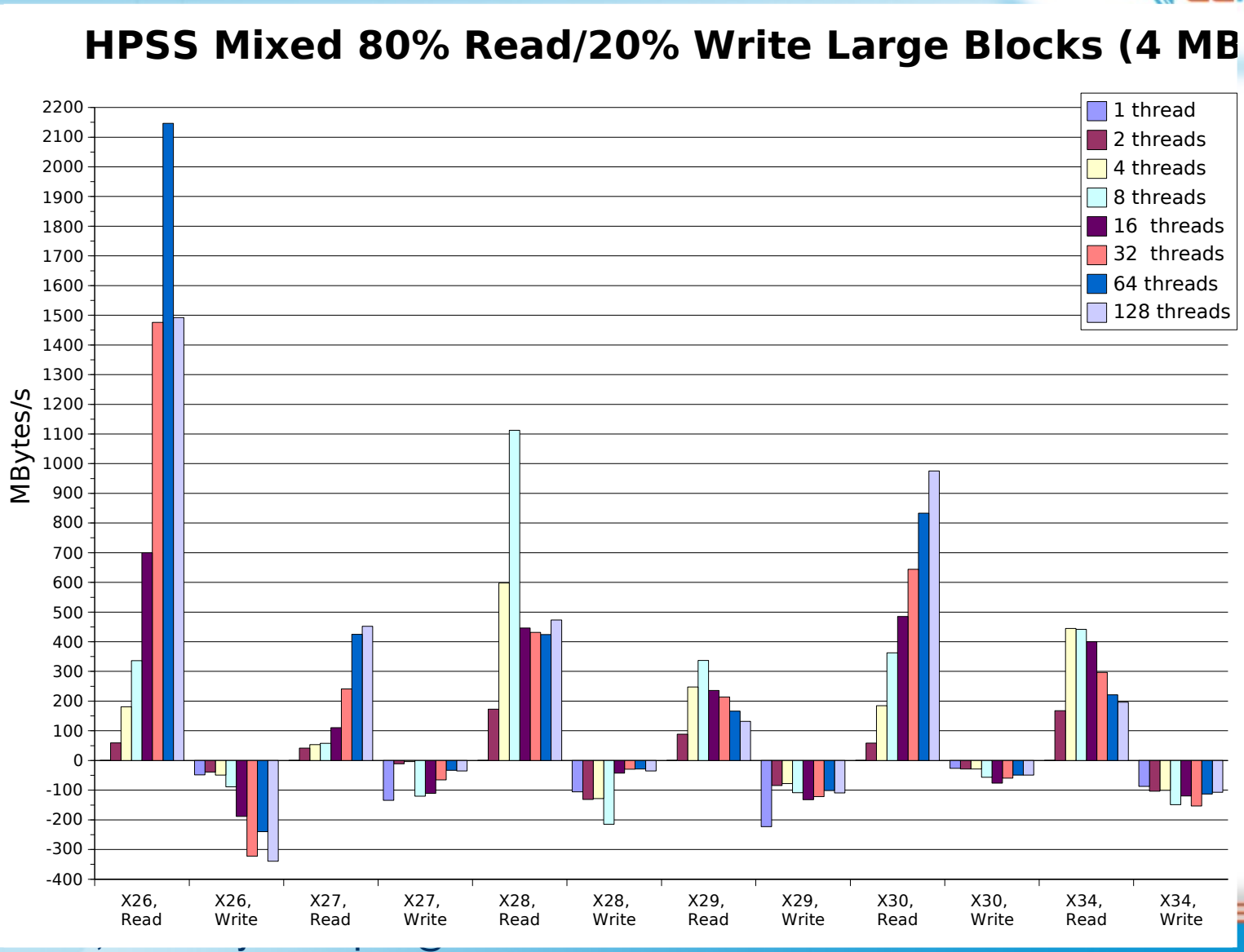
Écriture séquentielle, 1MB



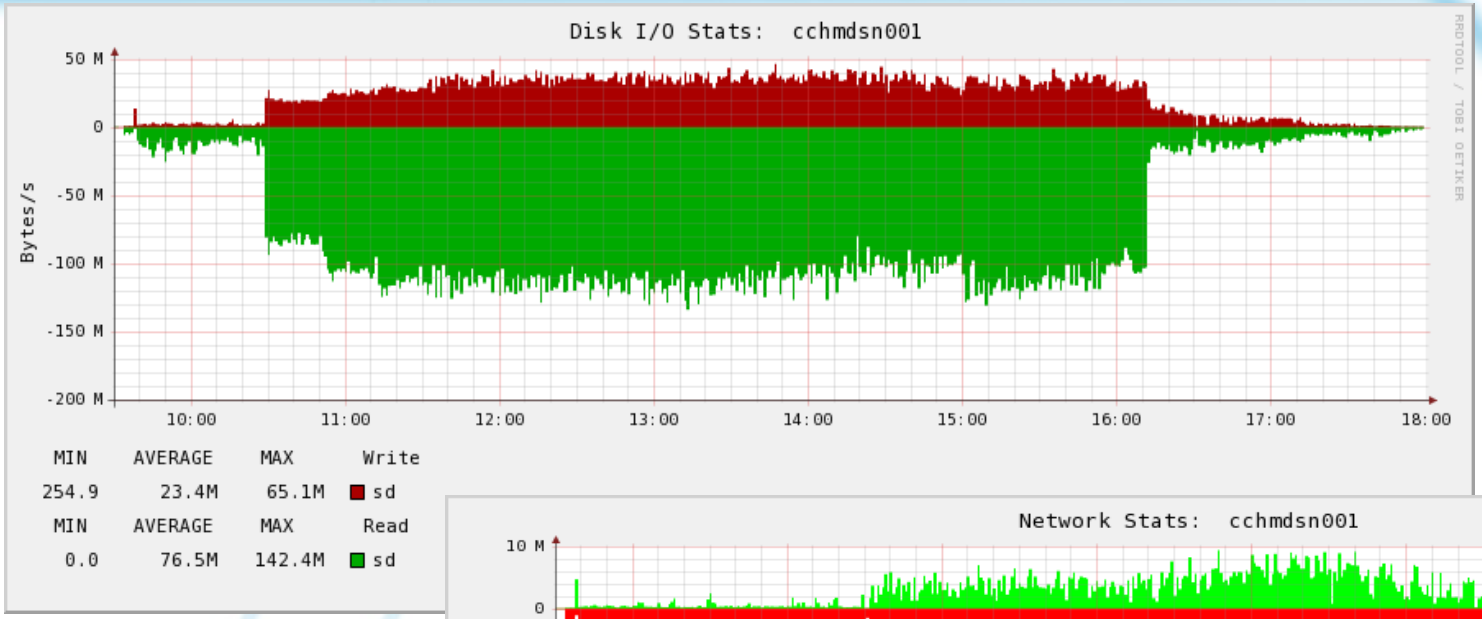
Lecture séquentielle, 1MB

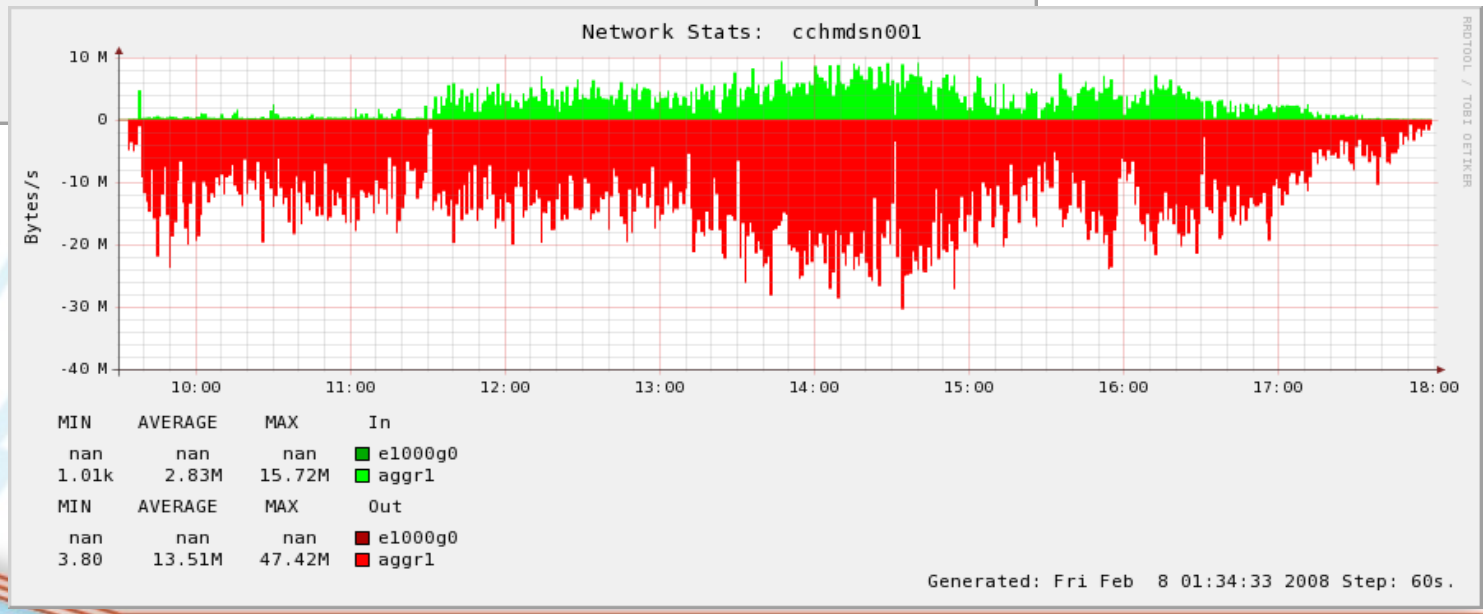


SVM ou LVM ?


HPSS Mixed 80% Read/20% Write Large Blocks (4 MB
MBytes/s
2200
2100
2000
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
-100
-200
-300
-400
1 thread
2 threads
4 threads
8 threads
16 threads
32 threads
64 threads
128 threads
X26, Read
X26, Write
X27, Read
X27, Write
X28, Read
X28, Write
X29, Read
X29, Write
X30, Read
X30, Write
X34, Read
X34, Write

HPSS en production





Conclusions tests

- Préférence pour Solaris
- Espace utile : 16,9 To à 17,8 To
- ZFS fournit à ces machines
  - Capacité
  - Performance
  - Sécurité

Thumpers, très bons côtés

- Installation aisée via Jumpstart
- Excellente intégration Solaris 10
- Matériel fiable :
  - ~ 40 incidents matériels ouverts
  - 2 RAM, 2 CPU, 1 alim, 3 ILOM
  - le reste pour des disques
- Pour 7008 disques, taux de renouvellement largement acceptable

Thumpers, petits soucis

- Besoin de trouver le bon interlocuteur au support
- Mises à jour système critiques mais nécessaires
  - Live upgrade peut-être une solution
- Brique de stockage minimale à 20 To
  - partitionnement de ressources
- Comportement des disques à plus long terme

Thumpers, Soucis plus sérieux

- Machines bruyantes, a fortiori lorsque elles sont nombreuses : 82 dB annoncés, 87 mesurés
- Boot sur 2 disques imposés
- ILOM buggué
  - support SOL (IPMI v2)
  - bug obligeant un reboot

Thumpers, à suivre

- Tuning OS
  - IO profiling
  - Dtrace
- Ethernet 10 Gbit/s
- GPFS
  - serveur sur Linux
  - iSCSI
- Thumper V2 !

Conclusions

Bonnes machines, dont nous tirons plus que ce que nous avions prévu.

Possibilité d'affiner la configuration selon les critères

Capacité

Performance

Sécurité