Stephan Seitz <s.seitz@heinlein-support.de>

Hochverfügbares iSCSI Storage

mit ceph

Vers. 12/2015 iscsi timeout ergänzt

Stephan Seitz <s.seitz@heinlein-support.de>

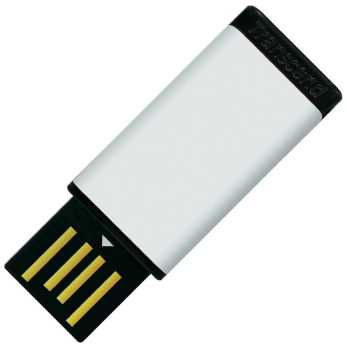
Wer ist die Heinlein Support GmbH?

- Wir bieten seit 20 Jahren Wissen und Erfahrung rund um Linux-Server und E-Mails
- IT-Consulting und 24/7 Linux-Support mit 20 Mitarbeitern
- Eigener Betrieb eines ISPs seit 1992
- Täglich tiefe Einblicke in die Herzen der IT aller Unternehmensgrößen

Stephan Seitz <s.seitz@heinlein-support.de>

Ziel:
Verbinden klassischer und moderner Storagekonzepte

+



Stephan Seitz <s.seitz@heinlein-support.de>

Intro:

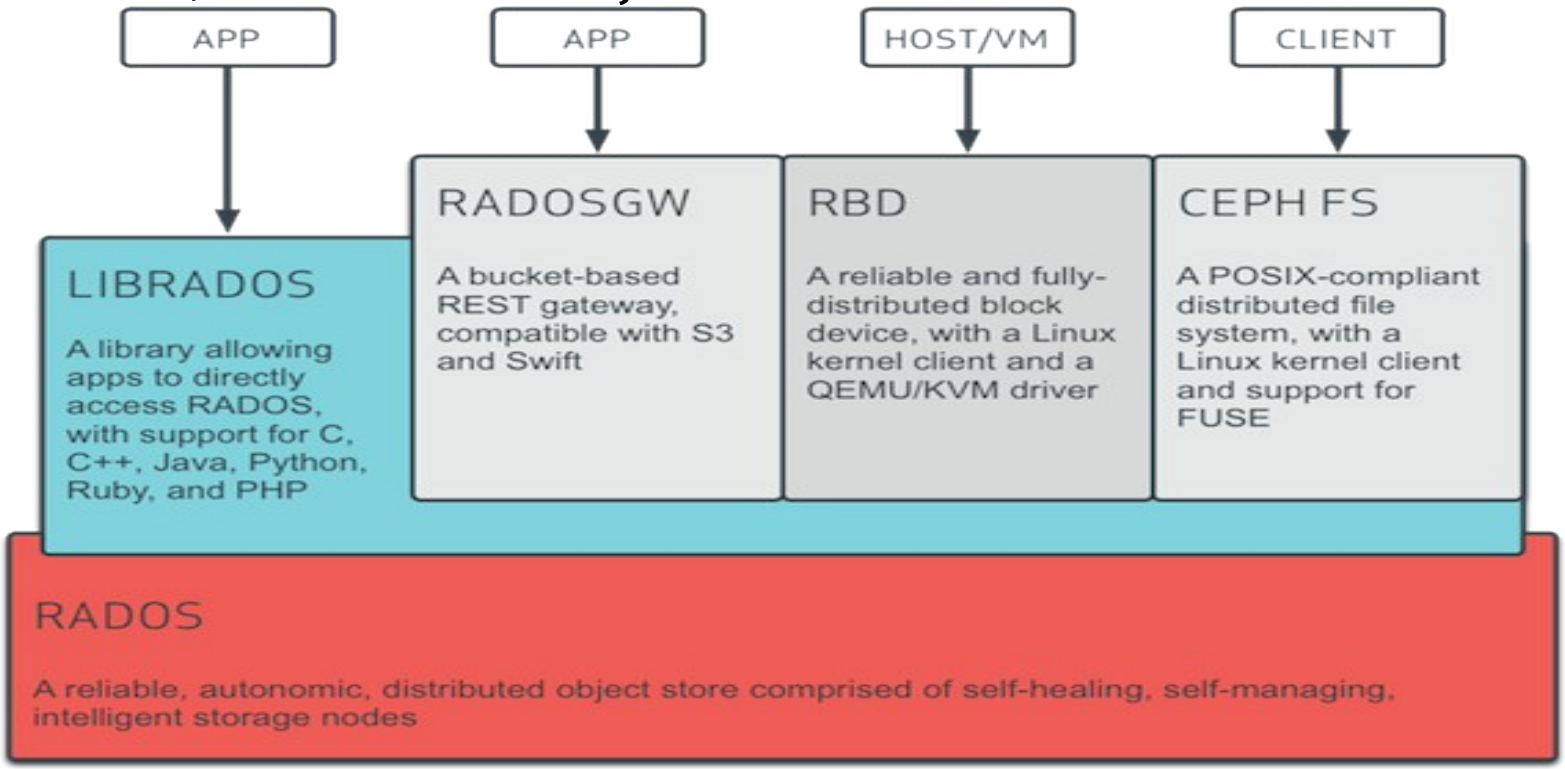
Ein kurzer Blick auf die Techniken

Stephan Seitz <s.seitz@heinlein-support.de>

Ein Blick auf ceph

Ein Blick auf ceph

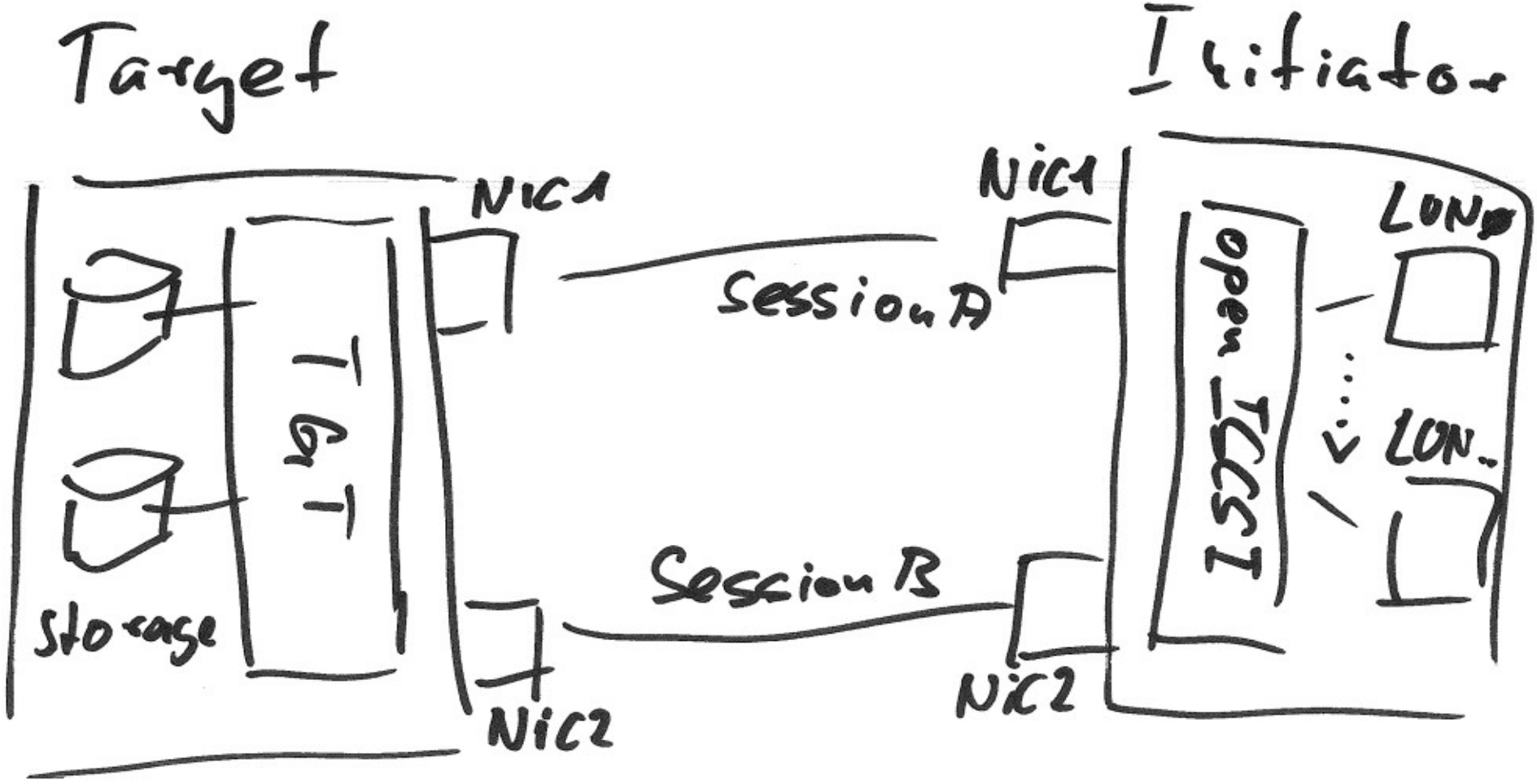
- Flexibler, skalierbarer Object-Store



- „THE FUTURE OF STORAGE™“

Stephan Seitz <s.seitz@heinlein-support.de>

Ein Blick auf iSCSI



Ein Blick auf iSCSI

- Einfach zu etablieren
- Software-defined Blockstorage
- Etablierte Mechanismen für Failover
  - ALUA
  - auch proprietäre Pfadpriorisierungen (EMC, NetApp, ...)

Warum dann ceph + iSCSI ?

- Verbindung von modernen und klassischen Storagekonzepten
  - Nicht jede Plattform unterstützt Rados Block Devices (RBD)
  - iSCSI dagegen fast immer
  - CephFS benötigt ebenfalls client-seitige Unterstützung und ist als Dateisystem nicht unbedingt ein Ersatz für Block-Storage
  - iSCSI Targets sind (in Software) üblicherweise nicht über Portale mehrerer Hosts zu betreiben
    - Hochverfügbarkeit nur über Hardware, d.h. mehrere Controller-Units an einer gemeinsamen Backplane
  - Linux TGT besitzt native RBD-Unterstützung (Userland, benötigt bs_rbd, in Mainline seit 02/2013)

Stephan Seitz <s.seitz@heinlein-support.de>

Hands On!

Implementierung und Test

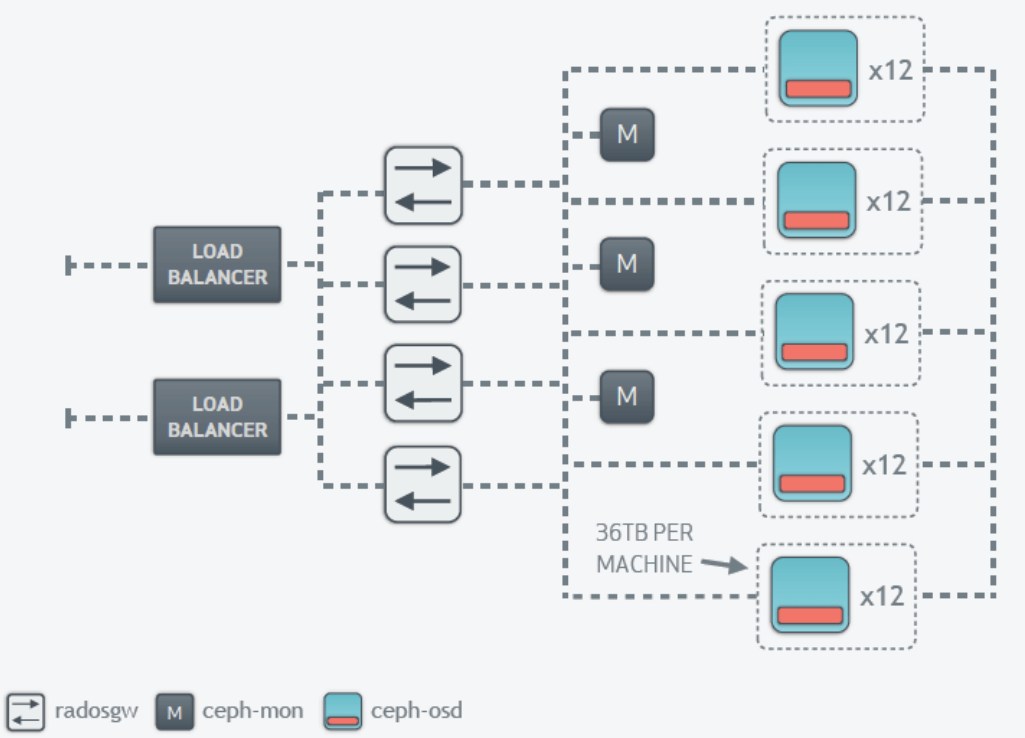
Stephan Seitz <s.seitz@heinlein-support.de>

Prerequisites

- Laufender ceph-cluster
  - Getestet mit Ceph Firefly 0.80.5 und Ceph Giant 0.87
- zwei oder mehr Hosts als ceph ↔ iSCSI „Bridge“
  - Idealerweise getrennte NIC für Backing Store (ceph) und iSCSI Portale
  - Getestet mit Ubuntu 14.04 LTS / ceph 0.80.5 sowie ceph 0.87 und tgt 1.0.43
- Hosts / Cluster / Applikationen / etc. mit dem Bedarf an skalierbarem Storage und „legacy“ Möglichkeiten

Konzept

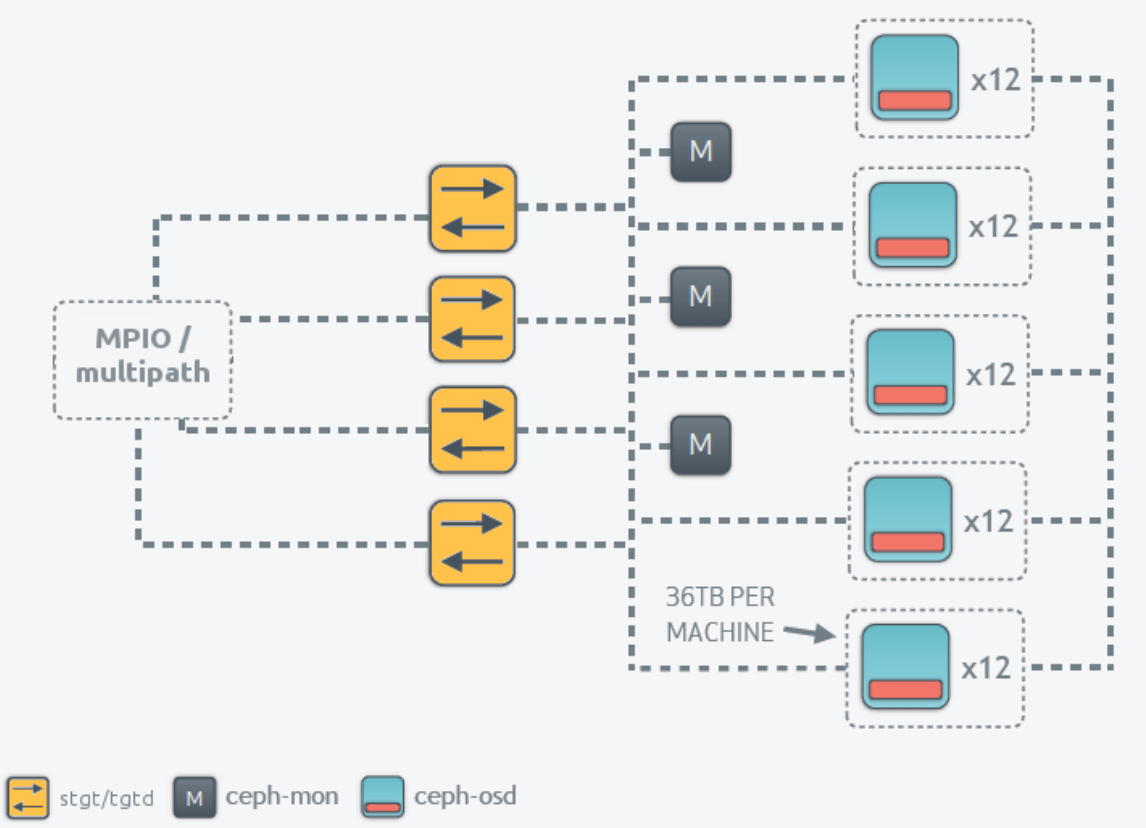
- Ceph kann parallel auf Objekte in OSD zugreifen



- Hier mit multiplen RadosGW Instanzen und S3/HTTP Loadbalancer

Konzept

→ Paralleler Zugriff funktioniert auch mit TGT iSCSI Target



Vorbereitung der ceph ↔ iSCSI „Bridge“

- Paketinstallation am Beispiel Ubuntu 14.04 LTS

```
tgt1 # aptitude install ceph tgt
```

- Anbindung an den ceph-Cluster (Konfiguration)

```
tgt1 # cat /etc/ceph/ceph.conf
[global]
mon_host = 192.168.100.247,192.168.100.249
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx

[client]
rbd_cache = false
# wird auf den „Bridge“ Hosts zwingend benötigt
# alternativ auch in [global] für den gesamten Cluster
```

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Anbindung an den ceph-Cluster (Authentisierung)

```
tgt1 # cat /etc/ceph/ceph.client.admin.keyring
[client.admin]
    key = QXVZBiNWUF2uEhAAnixKgvkpQkWvxQHWSBmr7g==
```

  - via **ceph auth** kann natürlich auch ein eigener Account für tgt verwendet werden.

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Erzeugen eines Ceph-RBD, die Basis für eine spätere LUN

```
tgt1 # rbd create --size 3072 ceph-tgt-shared
tgt1 # rbd info ceph-tgt-shared
rbd image 'ceph-tgt-shared':
    size 3072 MB in 768 objects
    order 22 (4096 kB objects)
    block_name_prefix: rb.0.56c7.238e1f29
    format: 1
```

  - Im Beispiel wird ein 3 GB grosses RBD „ceph-tgt-shared“ erzeugt.
  - Ceph-typisch ist der Platz nicht belegt, sondern reserviert.
  - *Das RBD wird nicht über Kernel-RBD verwendet und sollte daher auch nicht gemapped werden!*
  - Reminder: Der RBD Cache muss für tgt deaktiviert sein!

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Neues iSCSI Portal erzeugen

```
tgt1 # tgtadm --lld iscsi --op new --mode target \
 --tid 1 \
 -T iqn.2005-10.de.heinlein-hosting:storage.s0.ceph.shared.tgt
```

  - Aufbau des *IQN*: iqn.YYYY-MM.NAMING_AUTHORITY:UNIQUE_NAME

- LUN 1 mit neuem RBD als Storage Backend erzeugen

```
tgt1 # tgtadm --lld iscsi --mode logicalunit --op new \
 --tid 1 --lun 1 --bstype rbd --backing-store ceph-tgt-shared \
 --bsopts "conf=/etc/ceph/ceph.conf;id=admin"
```

- Portal / Target binden

```
tgt1 # tgtadm --lld iscsi --op bind --mode target --tid 1 \
 -I ALL
```

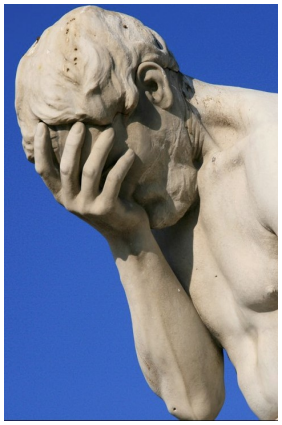
Stephan Seitz <s.seitz@heinlein-support.de>

Vorbereitung der ceph ↔ iSCSI „Bridge“

- Wiederholen des ceph- und tgt-Setups auf weiteren Hosts
  - ! Die tgt Konfiguration kann (Stand Version 1.1x) via tgt-admin nicht vollständig erzeugt werden.

    Der übliche Weg
    tgt-admin -dump > /etc/tgt/conf.d/targets.conf
    enthält nicht die Parameter bstype und bsopts.

    Je nach Umgebung: Eigene INIT-Skripte, evtl. **/etc/rc.local**
  - Alternativ: Targetkonfiguration schreiben.



Vorbereitung der ceph ↔ iSCSI „Bridge“

- Eigene Targetkonfiguration schreiben:

```
tgt1 # cat /etc/tgt/conf.d/ceph-tgt-shared.conf

default-driver iscsi # better located in /etc/tgt/target.conf

<target iqn.2005-10.de.heinlein-hosting:storage.s0.ceph.shared.tgt>
    driver iscsi
    bs-type rbd
    backing-store rbd/ceph-tgt-shared
</target>
```

  - Die Konfiguration auf alle iSCSI-“Bridge“ Hosts identisch erzeugen.

- Ab hier stehen zwei+ iSCSI Targets mit einem Ceph-RBD als LUN bereit
  - Der Storage-Teil ist vollständig :)

Einbinden der iSCSI Targets

- Paketinstallation am Beispiel Ubuntu 14.04 LTS

```
client # apt-get install open-iscsi multipath-tools
client # service open-iscsi start
client # service multipath-tools start
```

- Standard recovery timeout von 120 auf 5 Sekunden setzen

```
client # grep node.session.timeo.replacement_timeout
/etc/iscsi/iscsid.conf
node.session.timeo.replacement_timeout = 5 # default 120
```

  - Der Wert ist für jede (später) aktive iSCSI Session unter `/sys/class/iscsi_session/session*/recovery_tmo` R/W gespeichert.
  - lässt sich ebenfalls via iscsiadm setzen:

```
iscsiadm -m node -T $target \
    -o update \
    -n node.session.timeo.replacement_timeout -v 5
```

Stephan Seitz <s.seitz@heinlein-support.de>

Einbinden der iSCSI Targets

- iSCSI Targets finden

```
client # iscsiadm -m discoverydb -t sendtargets \
 --portal 192.168.100.88 --discover
```

  - Für jeden iSCSI-Host (jedes Portal) wiederholen

```
client # iscsiadm -m node
192.168.100.88:3260,1 iqn.2005-10.de.heinlein-hosting:st..
192.168.100.91:3260,1 iqn.2005-10.de.heinlein-hosting:st..
```

  - Kurzer Check.
    Jeder iSCSI-Host (target) sollte hier mit seiner Portal-IP und identischer IQN stehen.

Multipath-Konfiguration

→ MP-Konfiguration festlegen - Variante 1: Fokus auf Performance

```
client # cat /etc/multipath.conf
defaults {
        user_friendly_names     yes
        polling_interval        2
        path_grouping_policy    group_by_serial
        features                "1 queue_if_no_path"
        path_checker            directio
        rr_min_io               100
        failback                immediate
        no_path_retry           5 # 5 retries, evtl. „queue“
}
[...]
```

→ Alternativ auch jede WWID einzeln in multipaths { multipath { ...

Multipath-Konfiguration

- MP-Konfiguration festlegen - Variante 2: Fokus auf Verfügbarkeit

```
client # cat /etc/multipath.conf
defaults {
        user_friendly_names     yes
        polling_interval        2
        path_grouping_policy    failover
        features                "1 queue_if_no_path"
        path_checker            directio
        rr_min_io               100
        failback                immediate
        no_path_retry           fail # evtl. 5 (retries)
}
[...]
```

  - Alternativ auch jede WWID einzeln in multipaths { multipath { ...

Stephan Seitz <s.seitz@heinlein-support.de>

Multipath-Konfiguration und iSCSI Login

- MP-Konfiguration übernehmen …

```
client # echo reconfigure | multipathd -k
```

  - Alternativ auch multipath-tools neu starten
- … und in den iSCSI Portalen anmelden

```
client # iscsiadm -m node --login
Logging in to [iface: default, target: iqn.2005-10.de.he..
Logging in to [iface: default, target: iqn.2005-10.de.he..
Login to [iface: default, target: iqn.2005-10.de.heinlei..
Login to [iface: default, target: iqn.2005-10.de.heinlei..
```

  - für jedes Portal, d.h. jeden iSCSI Host sollte der Login erfolgreich sein
- … fast Fertig. → iSCSI timeouts prüfen und evtl. anpassen

Multipath vs. iSCSI → Ein Blick auf Timeouts

- iSCSI replacement_timeout prüfen

```
client # iscsiadm -m node -T iqn.2005-10.de.heinlein-hosting:st..
[…]
node.session.timeo.replacement_timeout = 144
[…]
```

  - steht hier auf 144 Sekunden, d.h. knapp 2,5 Min Wartezeit für den multipathd …

- iSCSI replacement_timeout anpassen

```
client # iscsiadm -m node -T iqn.2005-10.de.heinlein-hosting.. \
-o update \
-n node.session.timeo.replacement_timeout \
-v 5 # 5 Sekunden kann multipathd existieren, darf auch 2 Sek.
```

  - Aktive Sessions werden dadurch nicht verändert. Hier hilft das sysfs
    `echo 5 > /sys/class/iscsi_session/session`***N***`/recovery_tmo`

    Dank an Oliver Dzombic für diesen Hinweis!

Abschliessender Test

- MP-Konfiguration der iSCSI-LUN prüfen

```
client # multipath -ll
ceph-tgt-shared (3300000100000001) dm-3 IET,VIRTUAL-DISK
size=3.0G features='1 queue_if_no_path' hwhandler='0'
wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 28:0:0:1 sdf 8:80 active ready running
  `- 29:0:0:1 sdg 8:96 active ready running
```

  - je nach **/etc/multipath.conf** existiert **/dev/mapper/mpath0** oder der konfigurierte Alias.
    In diesem Beispiel **/dev/mapper/ceph-tgt-shared**

- Fertig!

Weitere sinnvolle Initiator: Hypervisors...

- Erfolgreicher iSCSI-Initiator Test mit XenServer 6.5
  - [x] Multipath aktivieren (Am einfachsten via XenCenter)
  - identisch dem vorgestellten Ubuntu 14.04 LTS:
    - alle Portale via iscsiadm discover finden
    - Die **`/etc/multipath.conf`** um Device Vendor „`IET*`“ und Product „`VIRTUAL*`“ erweitern. Auch hier path_grouping_policy **`group_by_serial`** bzw. **`failover`** eintragen.
    - multipathd neustarten bzw. „reconfigure“
    - Ein neues SR über **ein** Portal hinzufügen.
    - Die automatisch generierte Beschreibung des SR erwähnt nur das eine Portal, ein Blick auf die CLI zeigt, dass multipath auf allen Portalen erfolgreich verwendet wird. Beim ersten SR kann es notwendig sein, in alle Portale einzuloggen:

```
iscsiadm -m node -T iqn.2005-10.de.heinlein-hosting:st.. \
--login
```

Abschliessende Überlegungen

- Linux TGT muss nicht zwingend iSCSI over IP anbieten
  - RBD sollten auch problemlos via FCoE oder iSER (via RDMA) exportiert werden
  - Export via FC-HBA muss definitiv getestet werden
- Resize?!?
  - Ist nicht trivial, konkret:
    - RBD vergrössern
    - Alle Portale beenden und **ein** Portal starten, danach die restlichen Portale
    - iscsiadm Portal LUN neu einlesen, bzw. „discover“ als harter Weg
    - iscsiadm „logout“ und „login“
    - Multipathd neustarten oder „reconfigure“
    - Eigentlichen Inhalt (pv, partition, Filesystem, etc...) vergrössern
  - Eigentlich kein Unterschied zum Aufwand mit Hardware-SAN ...

Stephan Seitz <s.seitz@heinlein-support.de>

Abschliessende Überlegungen

- Kernel RBD ist performanter
  - Kann via multipath Split-Brain und Inkonsistenz verursachen.
    tgt zeigt im Userland ausreichend gute Performancewerte
  - I/O Caches müssen deaktiviert werden
    - damit bricht ein Grossteil des Performancevorteils des Kernel RBD weg
    - ist konfigurativ nicht vollständig deaktivierbar
- LIO ist das neuere Target-Framework
  - stgt kann 100% im Userland laufen, LIO ist stark Kernelbasiert
  - Mike Christie / RedHat arbeitet bereits am LIO Framework
    - Datensynchronisation ist noch ungelöst für PREEMPT/ABORT, ALUA-Status, etc.
- FC Target
  - qla2xxx in stgt ist Proof-Of-Concept / qla4xxx nicht enthalten
  - Benötigt Hardware zum testen

- Natürlich und gerne stehe ich Ihnen jederzeit mit Rat und Tat zur Verfügung und freue mich auf neue Kontakte.
  - Stephan Seitz
  - Mail: s.seitz@heinlein-support.de
  - Telefon: 030/40 50 51 – 44

- Wenn's brennt:
  - Heinlein Support 24/7 Notfall-Hotline: 030/40 505 - 110

Stephan Seitz <s.seitz@heinlein-support.de>

Soweit, so gut.

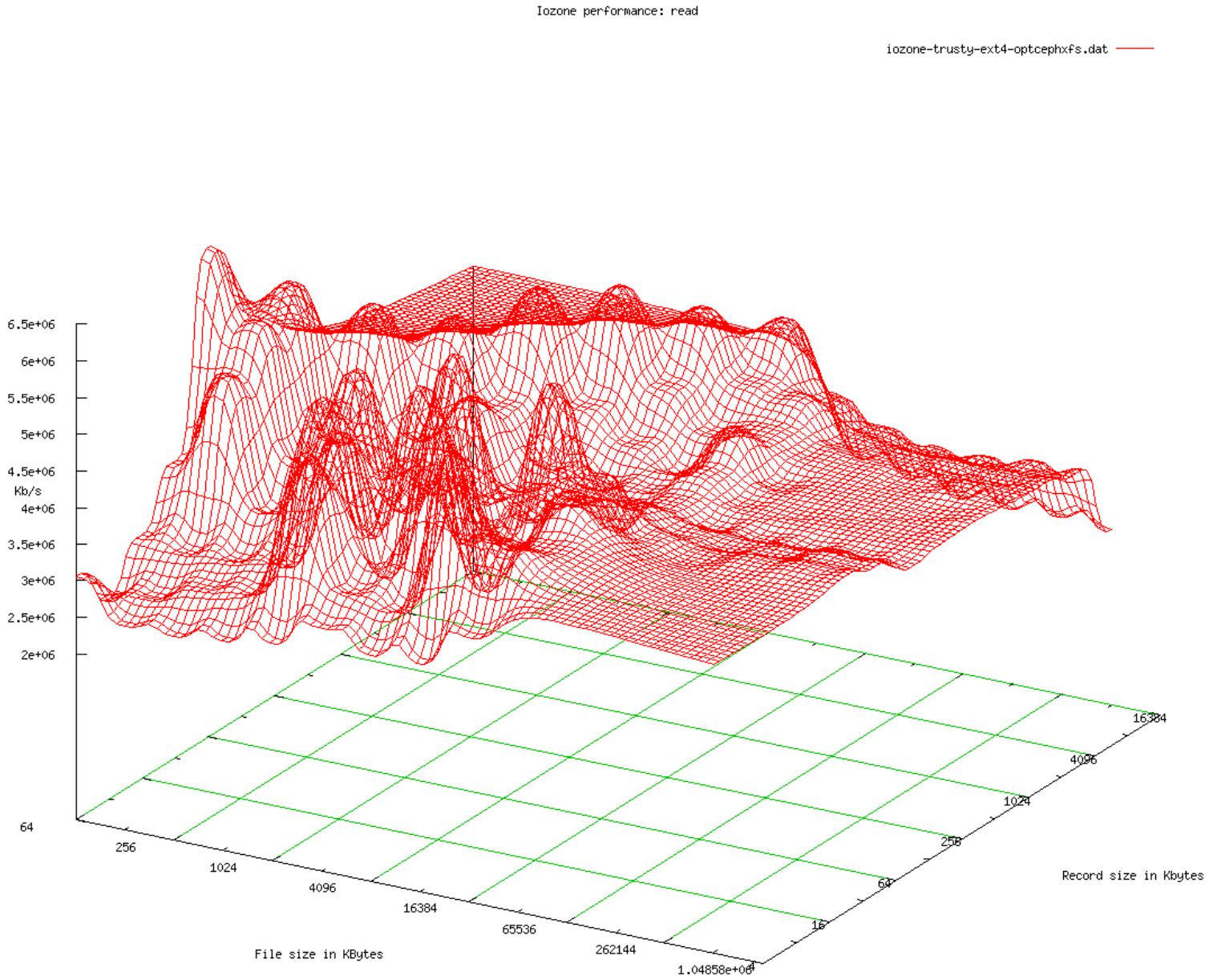
Gleich sind Sie am Zug:
Fragen und Diskussionen!

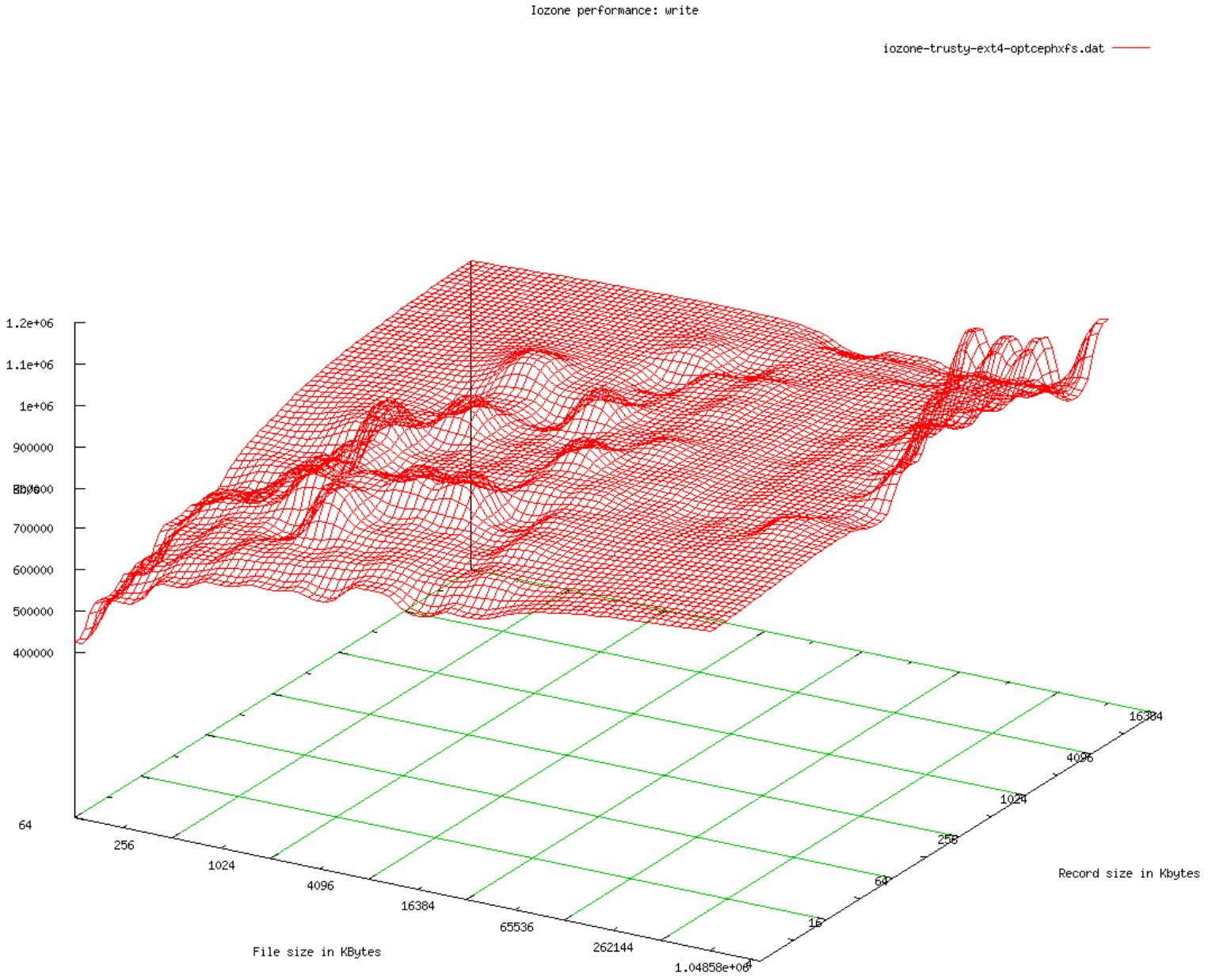
Stephan Seitz <s.seitz@heinlein-support.de>

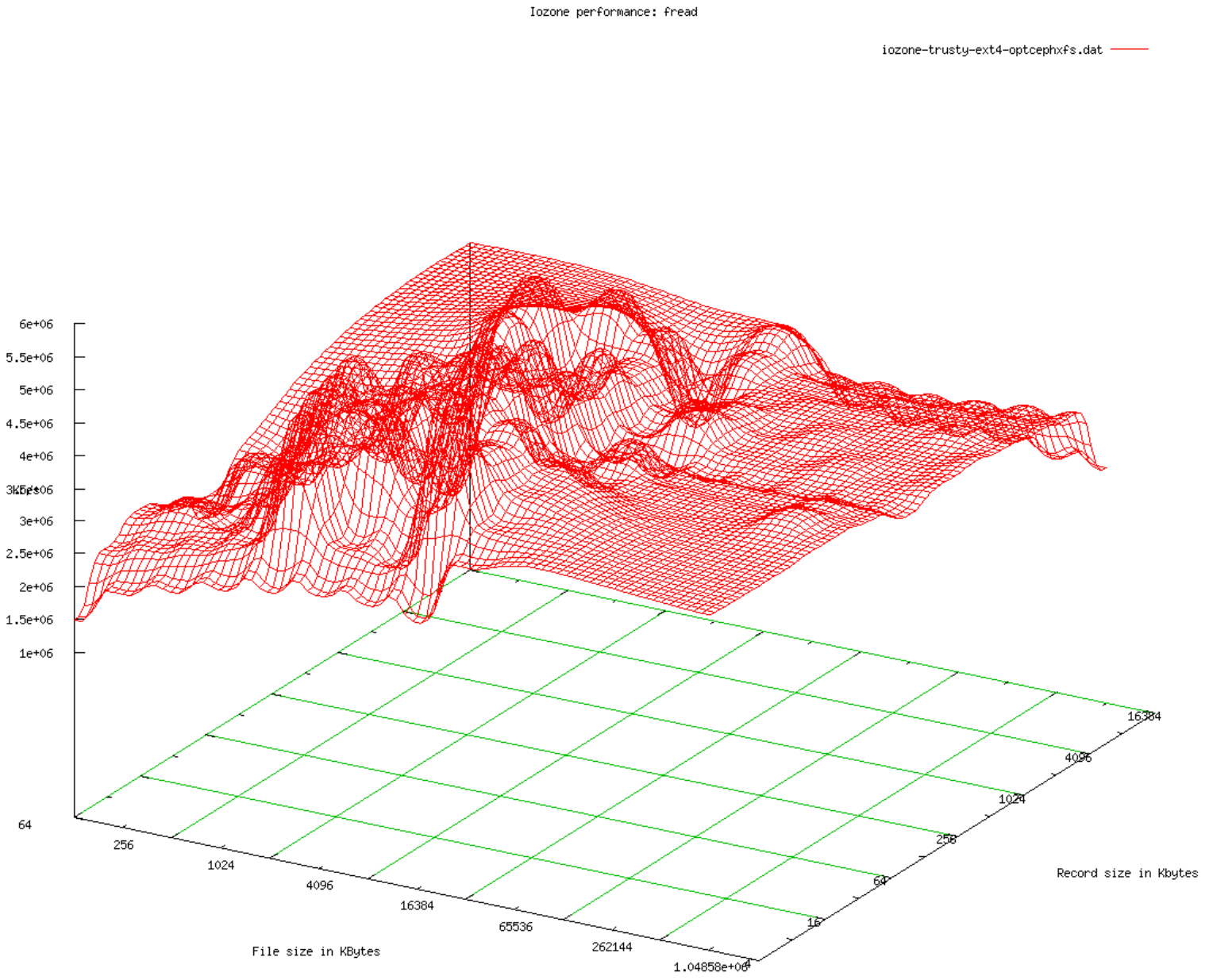
Performance!

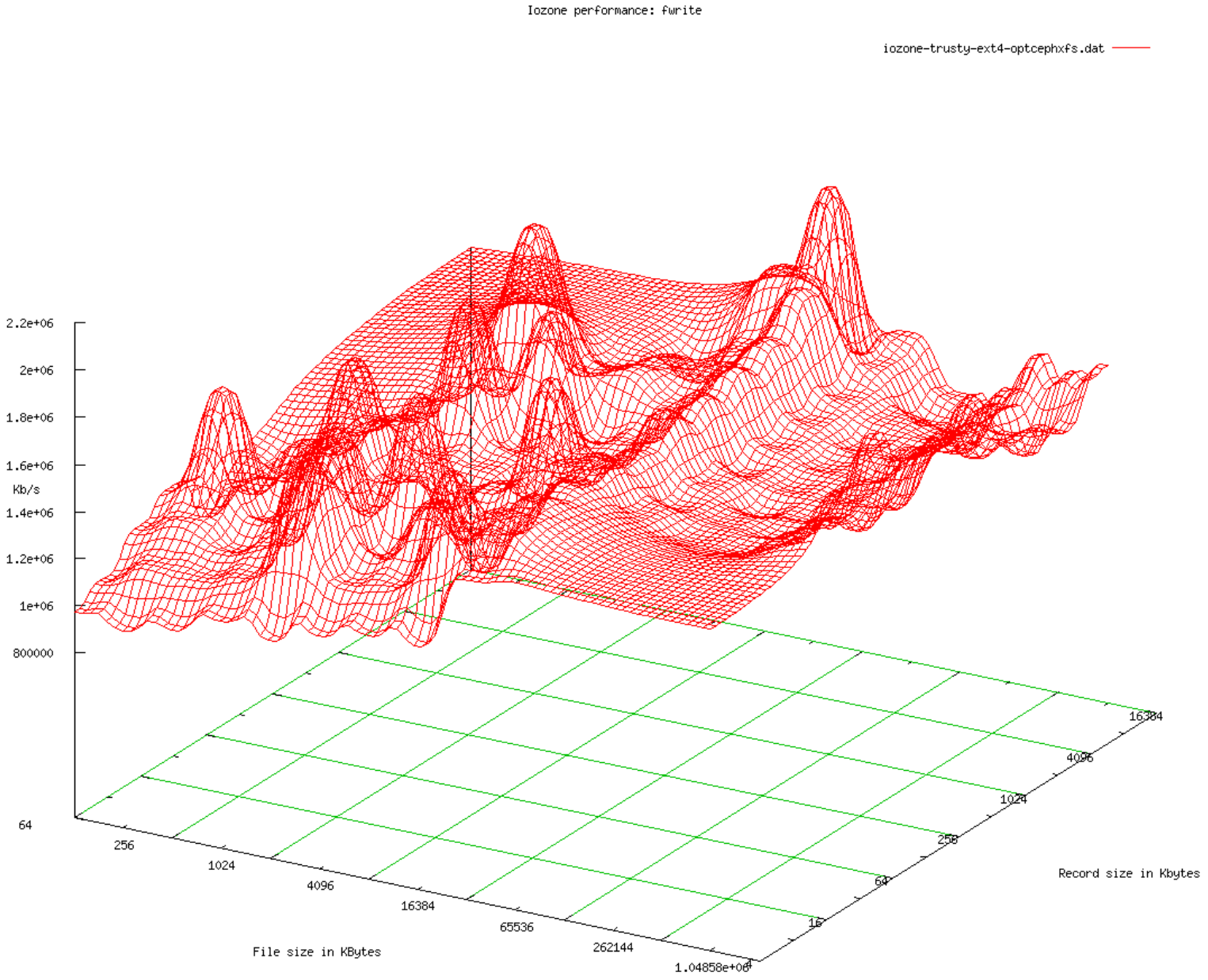
24 OSD - 3 Nodes - 3 Journal SSD
2 iSCSI Bridge Nodes
10 GBit/s Ceph - 10 GBit/s iSCSI
XenServer 6.5 - Ubuntu 14.04 VM

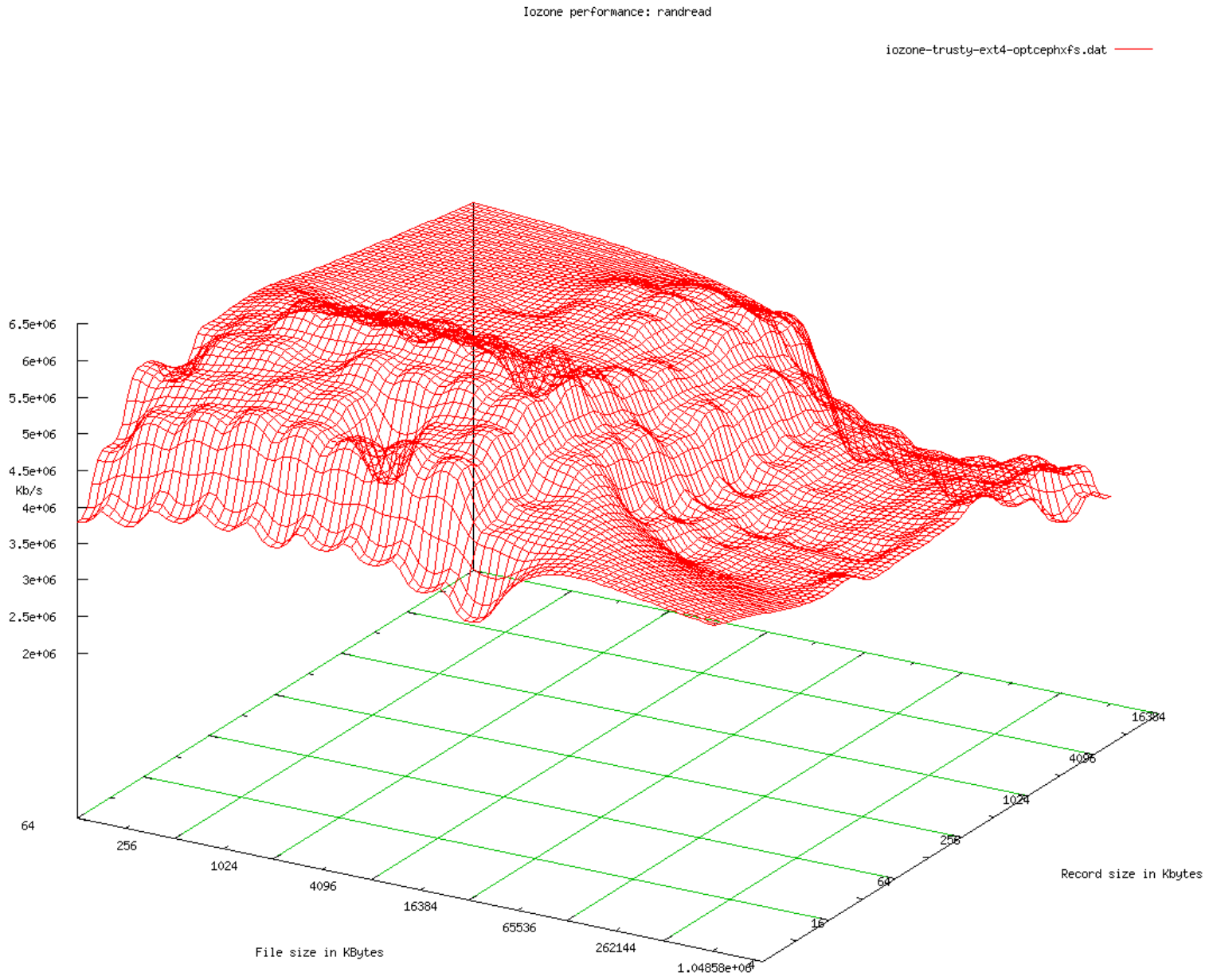
Messwerte via iozone3 in einer VM

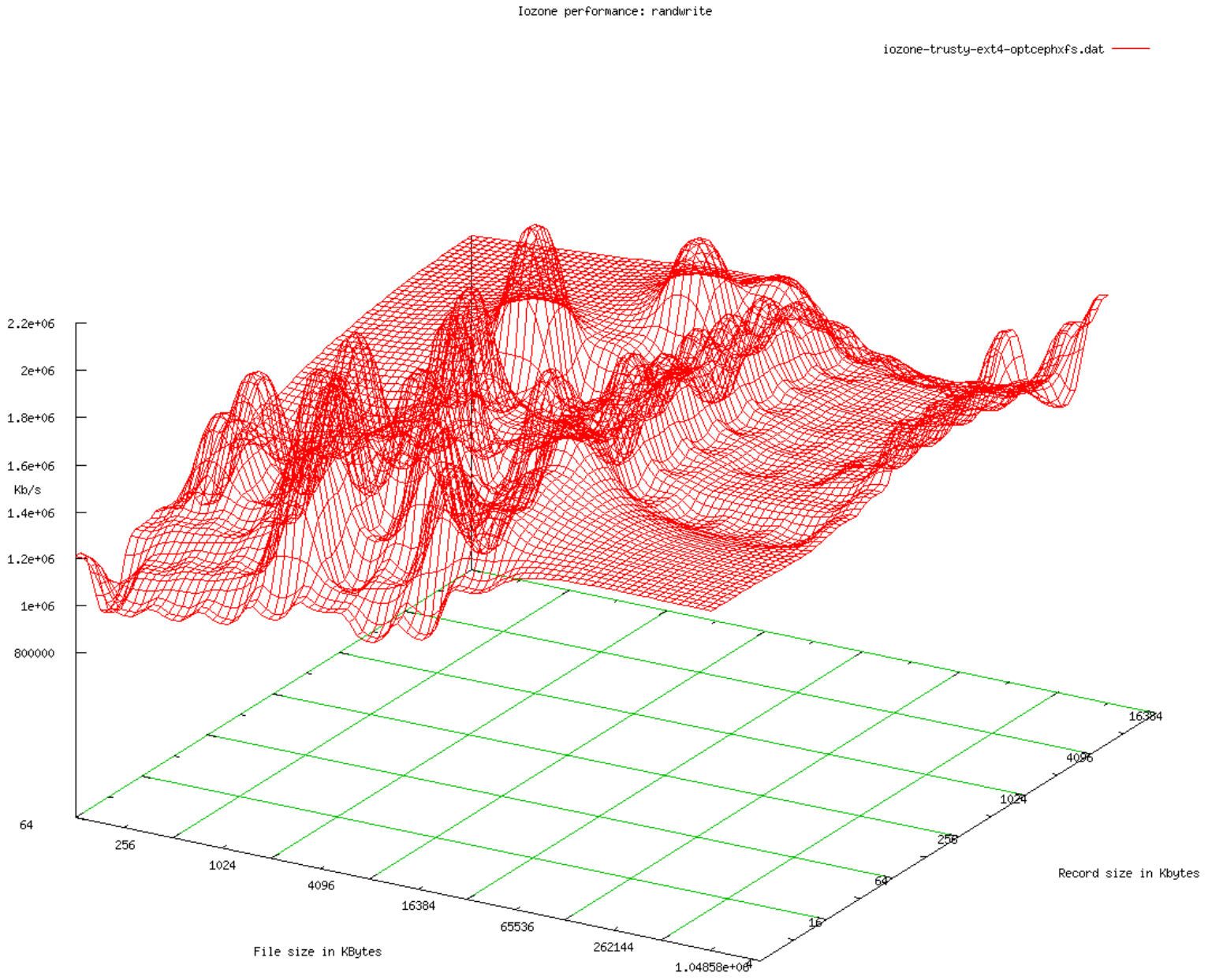

Iozone performance: read
iozone-trusty-ext4-optcephxfs.dat
6.5e+06
6e+06
5.5e+06
5e+06
4.5e+06
Kb/s
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
16384
4096
1024
256
64
16
4
Record size in Kbytes

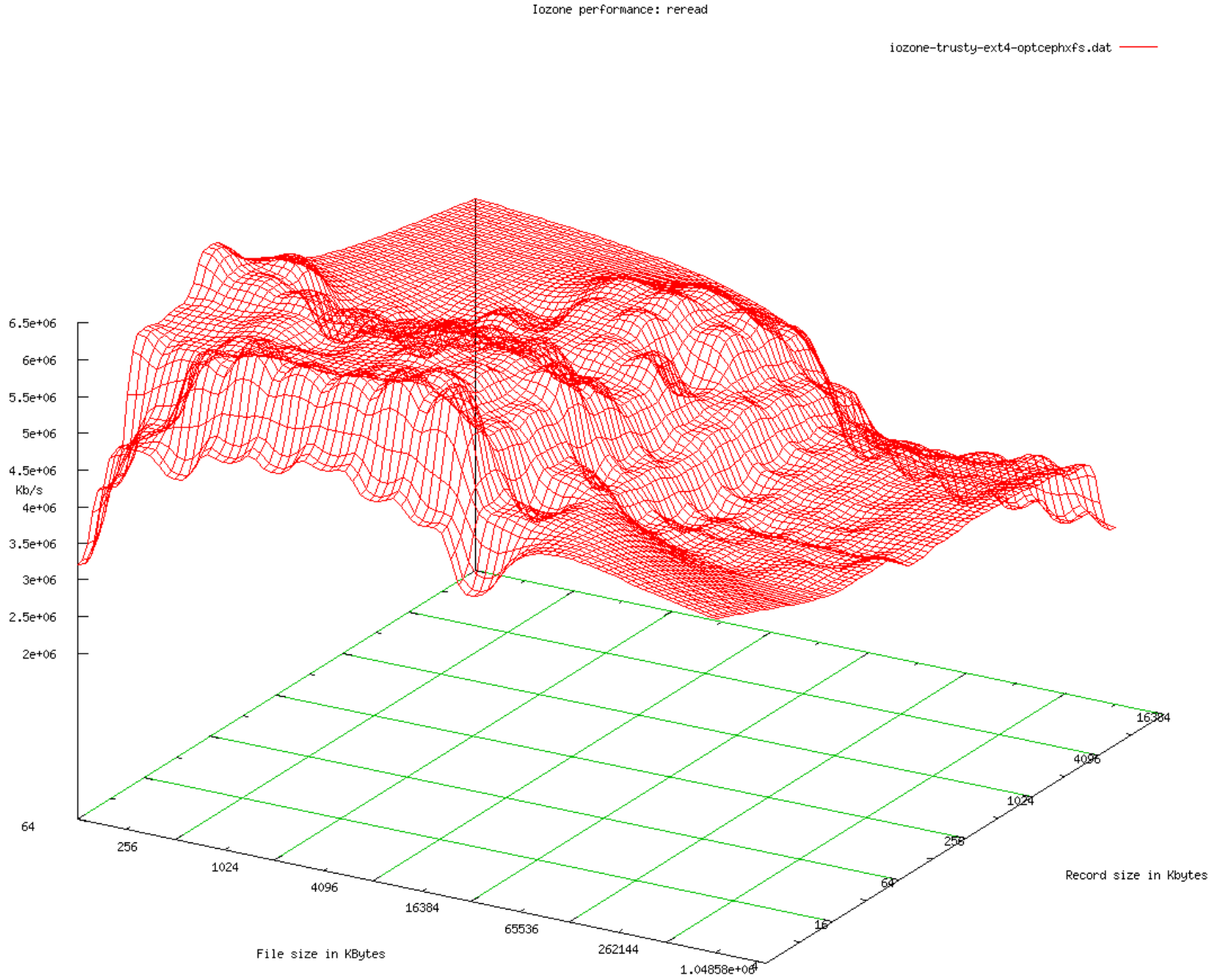

Iozone performance: write
iozone-trusty-ext4-optcephxfs.dat
1.2e+06
1.1e+06
1e+06
900000
800000
700000
600000
500000
400000
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes

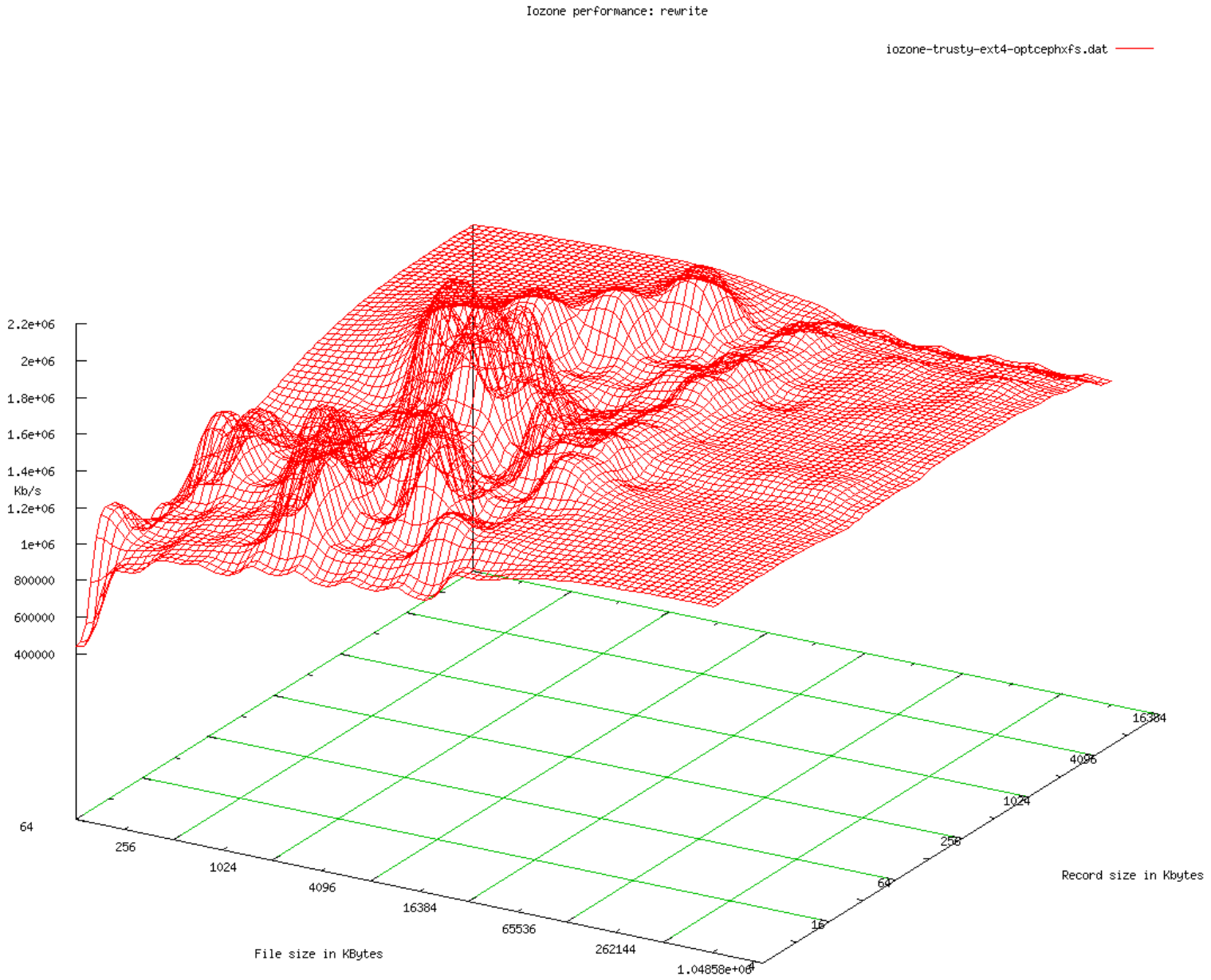

Iozone performance: fread
iozone-trusty-ext4-optcephxfs.dat
6e+06
5.5e+06
5e+06
4.5e+06
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
1.5e+06
1e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes

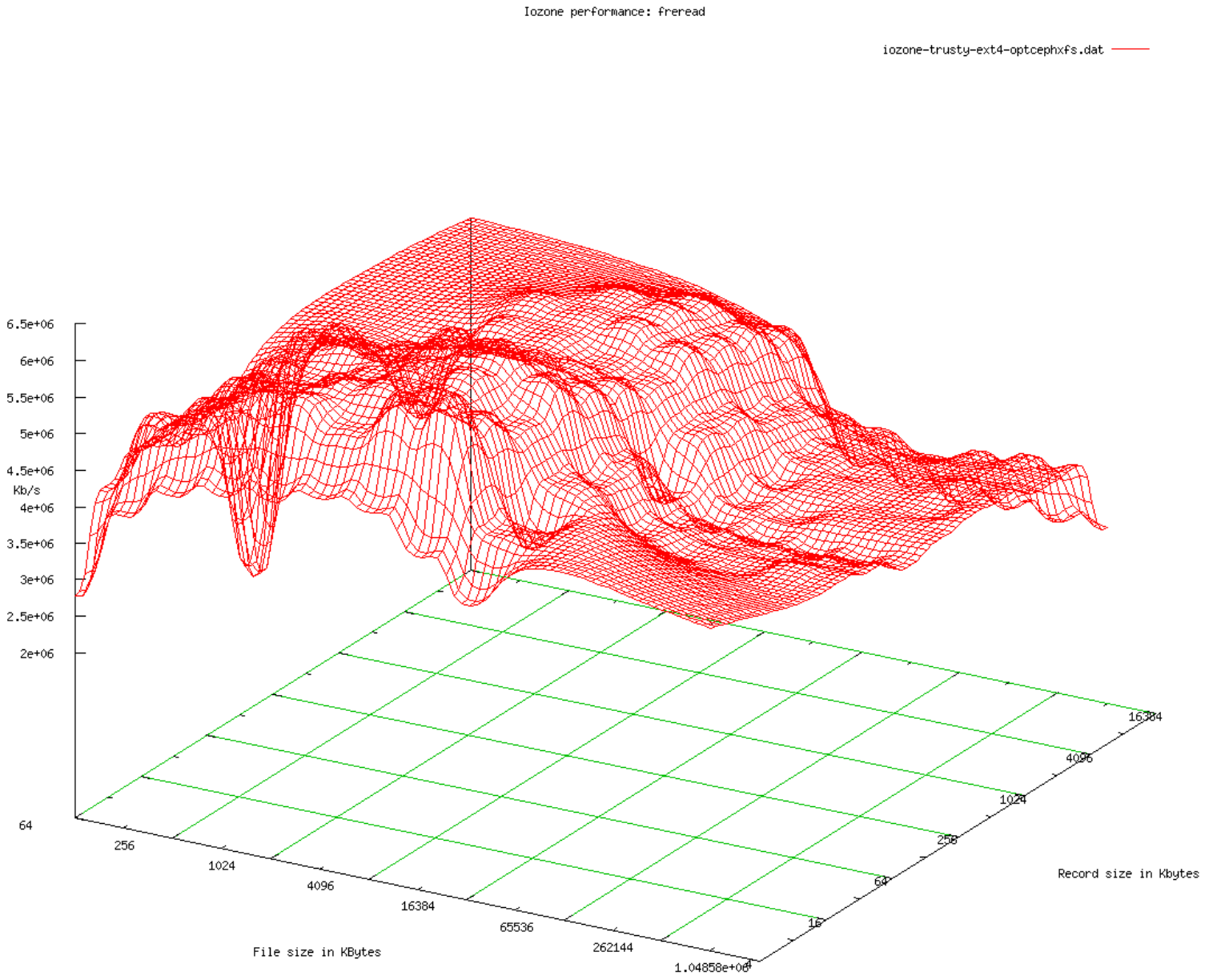

Iozone performance: fwrite
iozone-trusty-ext4-optcephxfs.dat
2.2e+06
2e+06
1.8e+06
1.6e+06
Kb/s
1.4e+06
1.2e+06
1e+06
800000
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes

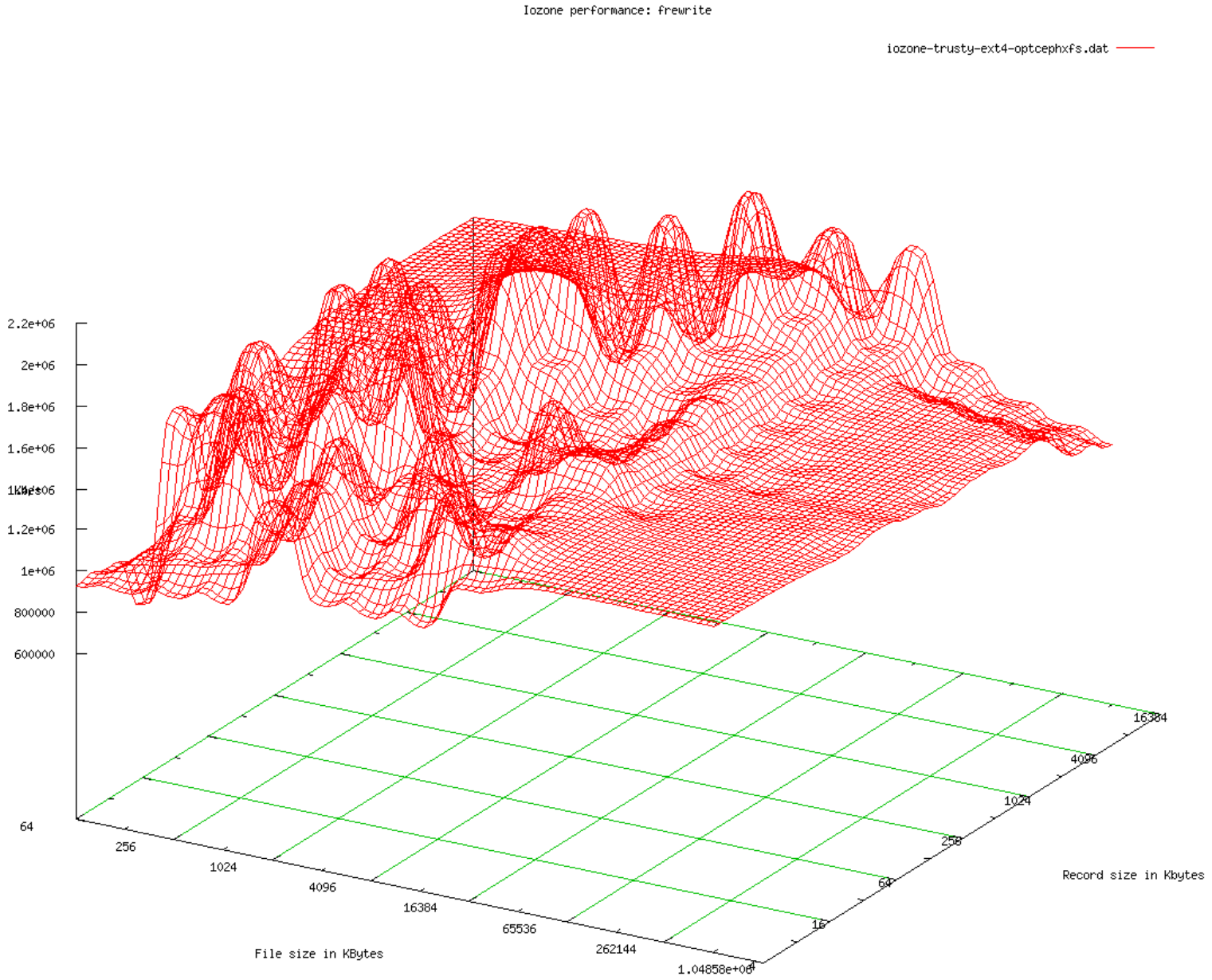

Iozone performance: randread
iozone-trusty-ext4-optcephxfs.dat
6.5e+06
6e+06
5.5e+06
5e+06
4.5e+06
Kb/s
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes

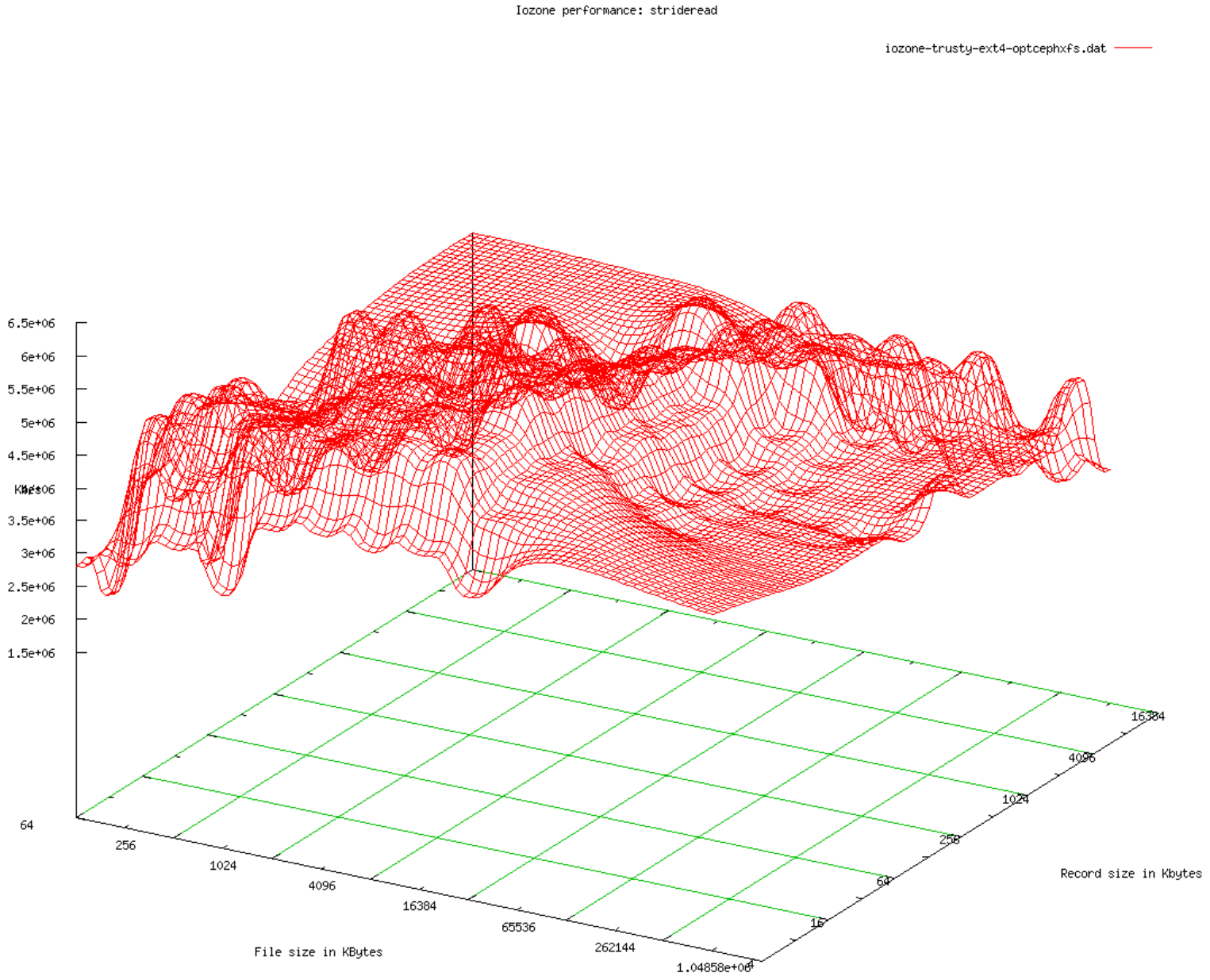

Iozone performance: randwrite
iozone-trusty-ext4-optcephxfs.dat
2.2e+06
2e+06
1.8e+06
1.6e+06
Kb/s
1.4e+06
1.2e+06
1e+06
800000
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes

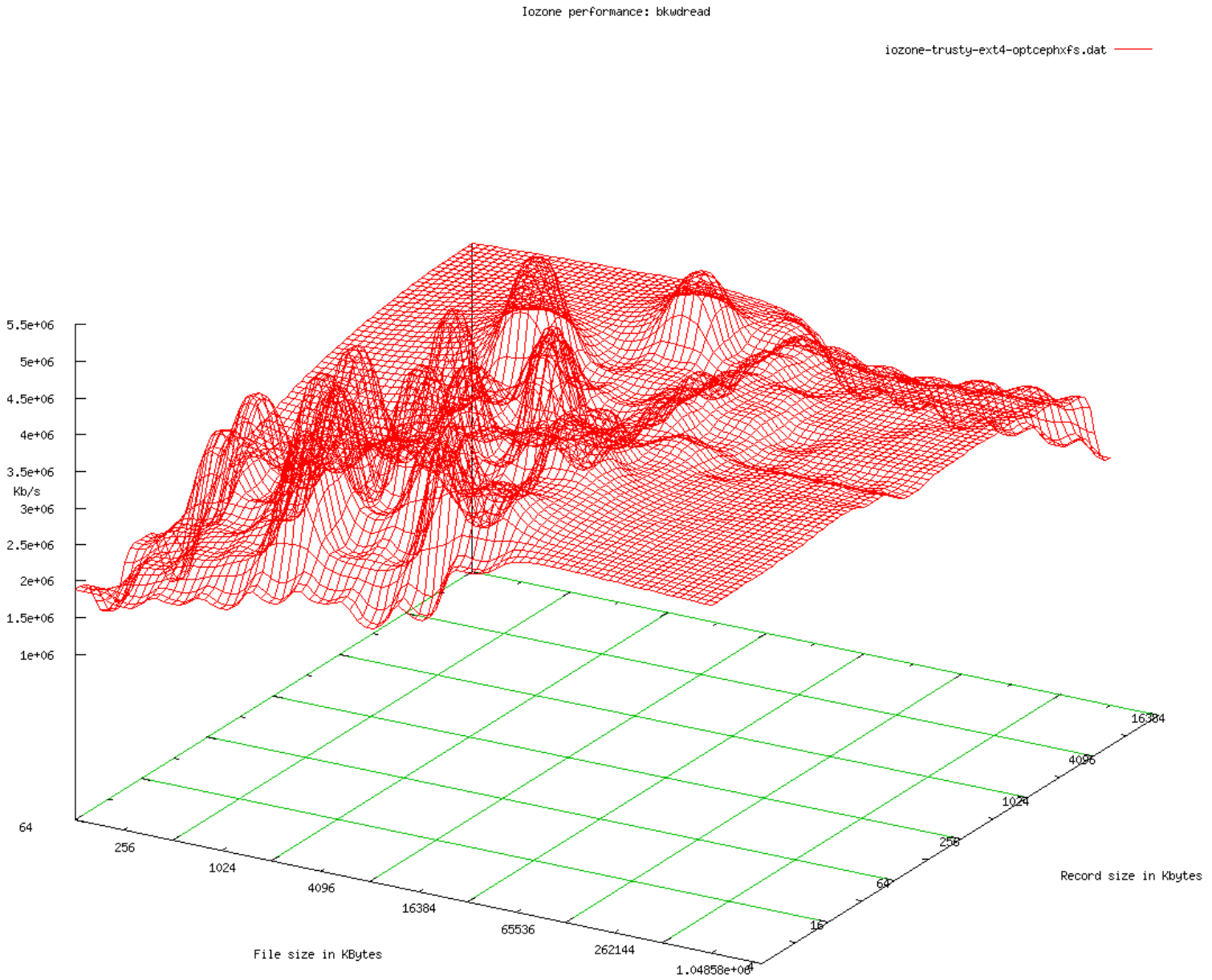

Iozone performance: reread
iozone-trusty-ext4-optcephxfs.dat
6.5e+06
6e+06
5.5e+06
5e+06
4.5e+06
Kb/s
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
16
64
256
1024
4096
16384
Record size in Kbytes

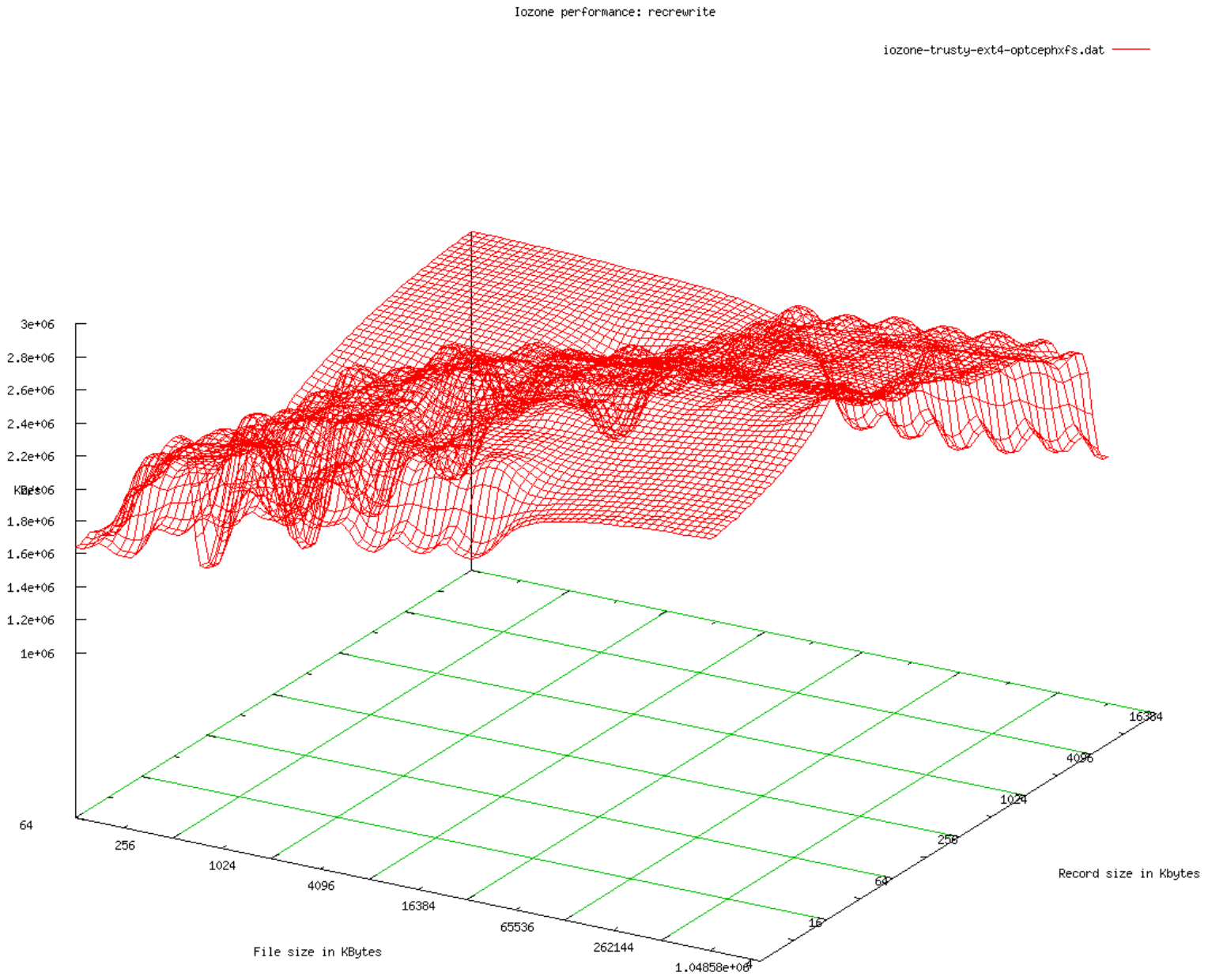

Iozone performance: rewrite
iozone-trusty-ext4-optcephxfs.dat
2.2e+06
2e+06
1.8e+06
1.6e+06
1.4e+06
Kb/s
1.2e+06
1e+06
800000
600000
400000
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes


Iozone performance: fread
iozone-trusty-ext4-optcephxfs.dat
6.5e+06
6e+06
5.5e+06
5e+06
4.5e+06
Kb/s
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes


Iozone performance: frewrite
iozone-trusty-ext4-optcephxfs.dat
2.2e+06
2e+06
1.8e+06
1.6e+06
1.4e+06
1.2e+06
1e+06
800000
600000
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
16
64
256
1024
4096
16384
Record size in Kbytes


Iozone performance: strideread
iozone-trusty-ext4-optcephxfs.dat
6.5e+06
6e+06
5.5e+06
5e+06
4.5e+06
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
1.5e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes


Iozone performance: bkwdread
iozone-trusty-ext4-optcephxfs.dat
5.5e+06
5e+06
4.5e+06
4e+06
3.5e+06
Kb/s
3e+06
2.5e+06
2e+06
1.5e+06
1e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes


Iozone performance: recrewrite
iozone-trusty-ext4-optcephxfs.dat
3e+06
2.8e+06
2.6e+06
2.4e+06
2.2e+06
2e+06
1.8e+06
1.6e+06
1.4e+06
1.2e+06
1e+06
64
256
1024
4096
16384
65536
262144
1.04858e+06
File size in KBytes
4
16
64
256
1024
4096
16384
Record size in Kbytes

Referenz

- ceph - Storage Cluster Quick Start (ceph-deploy)
  - http://ceph.com/docs/master/start/quick-ceph-deploy/
- Linux SCSI target framework
  - http://stgt.sourceforge.net/
- Generic SCSI Target Subsystem for Linux
  - http://scst.sourceforge.net/
- Multipath I/O
  - http://en.wikipedia.org/wiki/Multipath_I/O
  - http://en.wikipedia.org/wiki/Linux_DM_Multipath
  - /usr/share/doc/multipath-tools/examples/multipath.conf.annotated.gz

Wir suchen neue Kollegen für:

Helpdesk, Administration, Consultanting!

Wir bieten:
Spannende Projekte, Kundenlob, eigenständige Arbeit, keine Überstunden, Teamarbeit

...und natürlich: Linux, Linux, Linux...

http://www.heinlein-support.de/jobs

Stephan Seitz <s.seitz@heinlein-support.de>

Heinlein Support hilft bei allen Fragen rund um Linux-Server

HEINLEIN AKADEMIE

Von Profis für Profis: Wir vermitteln die oberen 10% Wissen: geballtes Wissen und umfangreiche Praxiserfahrung.

HEINLEIN HOSTING

Individuelles Business-Hosting mit perfekter Maintenance durch unsere Profis. Sicherheit und Verfügbarkeit stehen an erster Stelle.

HEINLEIN CONSULTING

Das Backup für Ihre Linux-Administration: LPIC-2-Profis lösen im CompetenceCall Notfälle, auch in SLAs mit 24/7-Verfügbarkeit.

HEINLEIN ELEMENTS

Hard- und Software-Appliances und speziell für den Serverbetrieb konzipierte Software rund ums Thema eMail.