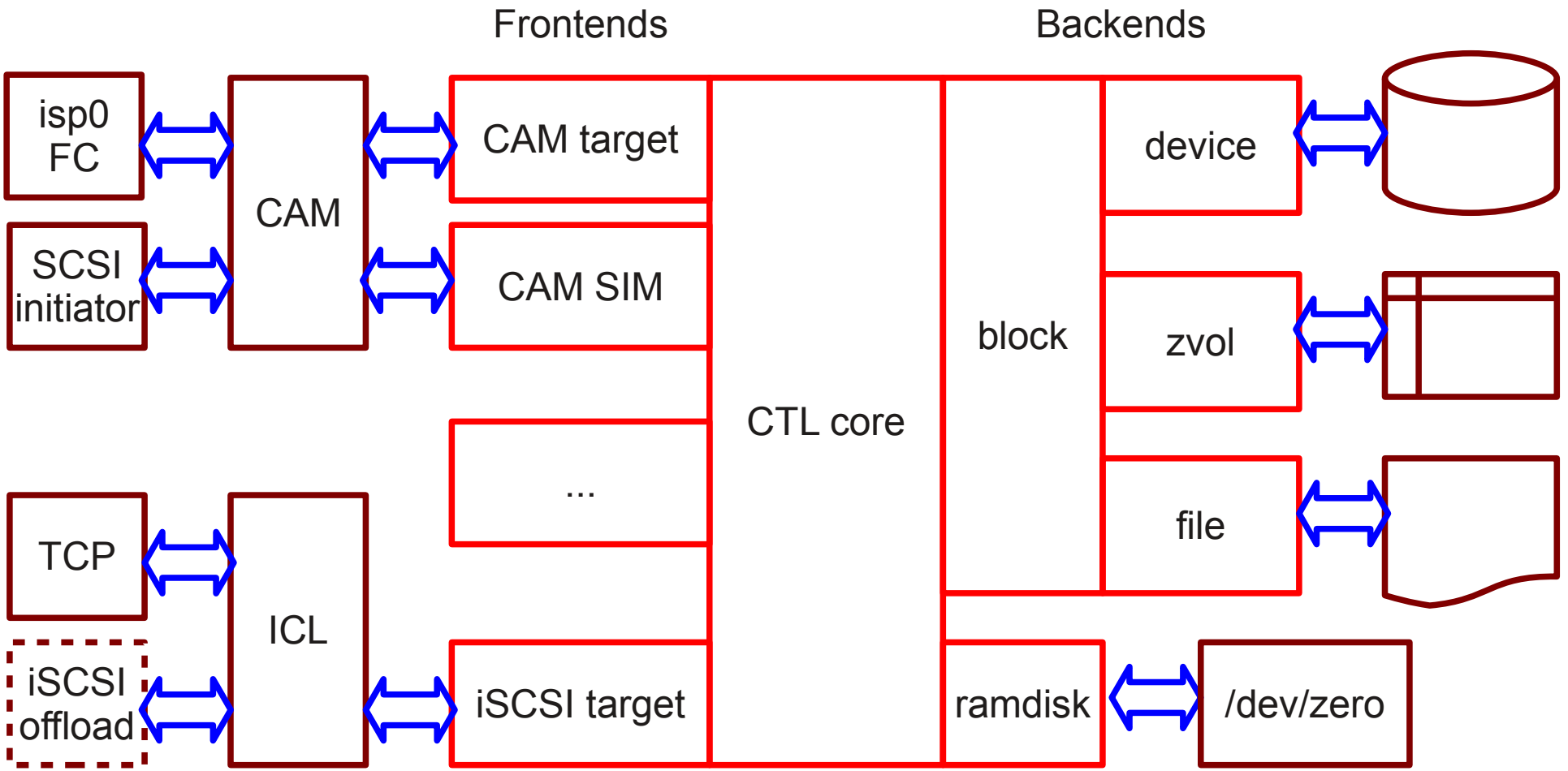
Functional and fast SCSI target with CTL and ZFS

Alexander Motin <mav@FreeBSD.org>
iXsystems, Inc.

RuBSD'2014

CTL – CAM Target Layer



CTL functional improvements
... for intelligent performance

CTL got support for storage acceleration:

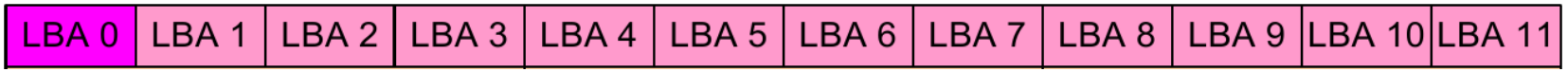
- VMware VAAI Block
- VMware VAAI Thin Provisioning
- Microsoft Offloaded Data Transfer (ODX)

… and more.

CTL functional improvements

Basic SCSI disk:

- READ CAPACITY(10)
  - Get block size and number of blocks



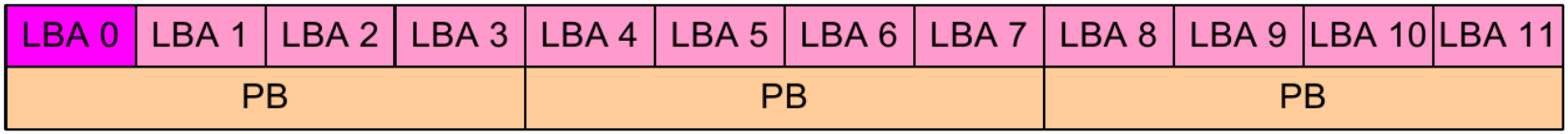
- READ
  - Read range of blocks
- WRITE
  - Write range of blocks

```
# diskinfo -v /dev/da0
/dev/da0
        512             # sectorsize
        107374182400    # mediasize in bytes (100G)
        209715200       # mediasize in sectors
```

CTL functional improvements

Advanced Format (512e) SCSI disks:

- READ CAPACITY(16)
  - Get physical block size and offset




```
# diskinfo -v /dev/da0
/dev/da0
        512             # sectorsize
        107374182400    # mediasize in bytes (100G)
        209715200       # mediasize in sectors
        8192            # stripesize
        0               # stripeoffset
```

CTL functional improvements

Basic thin-provisioned disks:

- Block Limits VPD page
  - Get UNMAP block size

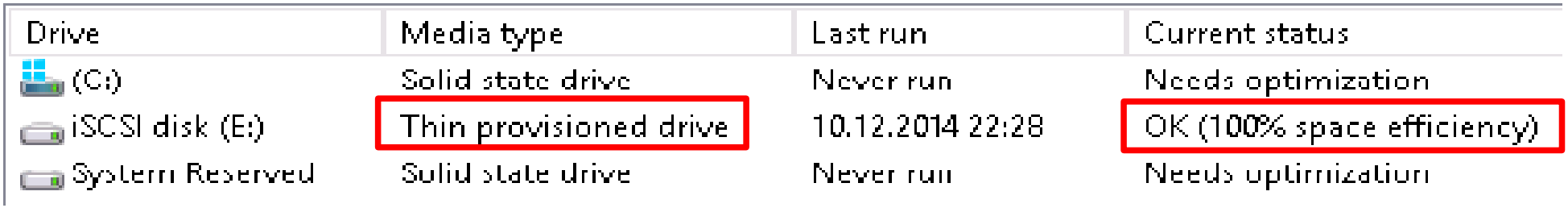
  - Get UNMAP parameters limitations
- Logical Block Provisioning VPD VAAI TP Reporting
  - Get Supported UNMAP commands
- WRITE SAME with UNMAP
  - Unmap sequential range of blocks
- UNMAP VAAI Unmap
  - Unmap arbitrary list of blocks
- Proper overflow error reporting VAAI TP Stun

CTL functional improvements

Featured thin-provisioned disk:

- GET LBA STATUS
  - Get provisioning status of specific block(s)

Windows defrag



Drive	Media type	Last run	Current status
(C:)	Solid state drive	Never run	Needs optimization
iSCSI disk (E:)	Thin provisioned drive	10.12.2014 22:28	OK (100% space efficiency)
System Reserved	Solid state drive	Never run	Needs optimization

CTL functional improvements

Featured thin-provisioned disk:

- Logical Block Provisioning log page
  - Get space usage statistics

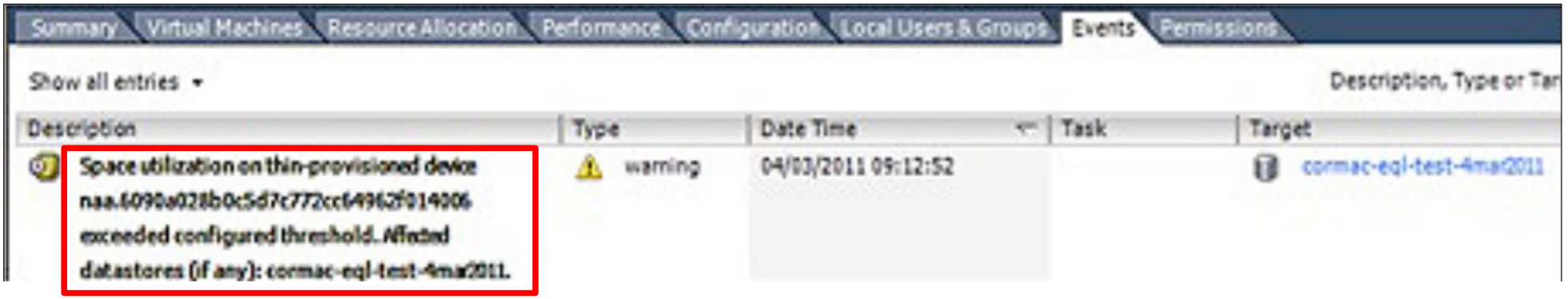
```
# sg_logs -p 0x0c da20
    FREEBSD   CTLDISK            0001
Logical block provisioning page  [0xc]
    Available LBA mapping threshold resource count: 2554830
    Scope: not dedicated to lu
    Used LBA mapping threshold resource count: 91
    Scope: dedicated to lu
```

CTL functional improvements

Featured thin-provisioned disk:

- Logical Block Provisioning mode page
  - Get/set space usage threshold notifications

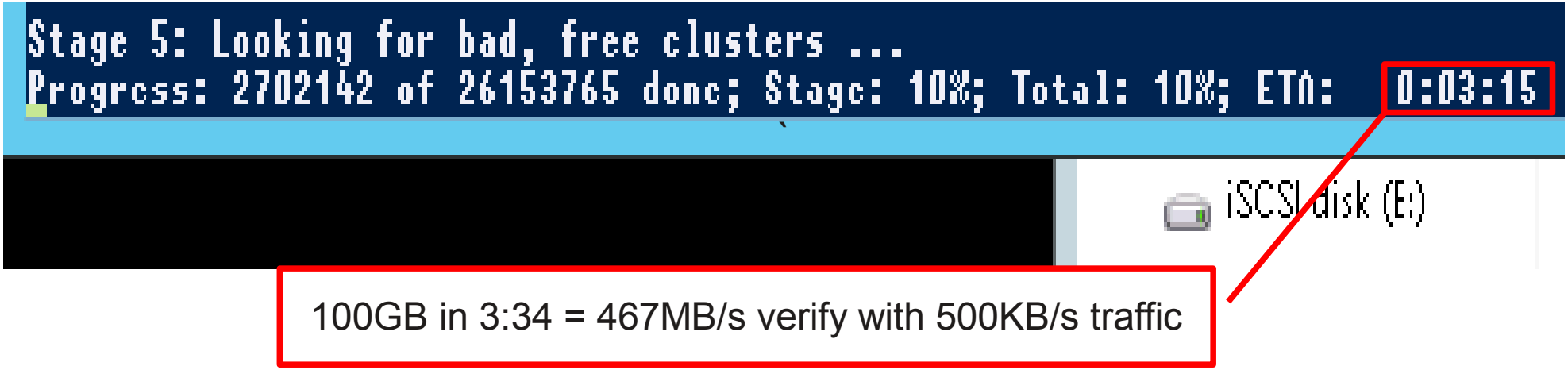
VAAI TP Space Threshold Warning



CTL functional improvements

Basic offloaded disk:

- COMPARE AND WRITE — VAAI ATS
  - Atomic operations to avoid LUN reservations
- VERIFY — Windows chkdsk
  - Verify data without transfer



- WRITE SAME — VAAI Zero
  - Write disk with pattern or zeroes

vSphere 5.5 Eager Zero creating 40GB VM

Recent Tasks

Name	Target	Status	Details	Start Time	Completed Time
Relocate virtual machine	40GB Virtual Machine	Completed		08.12.2014 23:22:18	08.12.2014 23:22:48
Create virtual machine	mavhome	Completed		08.12.2014 23:21:14	08.12.2014 23:21:24
Rescan VMFS	192.168.4.101	Completed		08.12.2014 23:19:29	08.12.2014 23:19:29
Rescan all HBAs	192.168.4.101	Completed		08.12.2014 23:19:23	08.12.2014 23:19:29

40GB in 10 seconds – 4GB/s disk write

packets	errs	idrops	bytes	packets	errs	bytes	colls
2	0	0	180	3	0	362	0
3480	0	0	1740864	2475	0	7050330	0
11317	0	0	7552298	11087	0	1135586	0
9083	0	0	6164486	9068	0	935036	0
7201	0	0	4619810	6947	0	706314	0
8516	0	0	5724648	8413	0	869216	0
7525	0	0	4979642	7435	0	757080	0
6648	0	0	4210176	6440	0	648980	0
7046	0	0	4452684	6831	0	673058	0
8027	0	0	5081110	7793	0	770006	0
2674	0	0	1499828	2339	0	1503434	0
1	0	0	66	1	0	294	0

4-7MB/s network traffic

CTL functional improvements

Featured offloaded disk v1 (XCOPY ala SPC-3):

- RECEIVE COPY OPERATING PARAMETERS
  - Get supported parameters
- XCOPY
  - Copy data from source to destination
- RECEIVE COPY STATUS
  - Check status of copy operation

VAAI Extended Copy

vSphere 5.5 migrating 40GB VM

Recent Tasks

Name	Target	Status	Details	Start Time	Completed Time
Relocate virtual machine	40GB Virtual Machine	Completed		08.12.2014 23:22:18	08.12.2014 23:22:48
Create virtual machine	mavhome	Completed		08.12.2014 23:21:14	08.12.2014 23:21:24
Rescan VMFS	192.168.4.101	Completed		08.12.2014 23:19:29	08.12.2014 23:19:29
Rescan all HBAs	192.168.4.101	Completed		08.12.2014 23:19:23	08.12.2014 23:19:29

40GB in 30 seconds – 1.3GB/s copy

packets	errs	idrops	bytes	packets	errs	bytes	colls
1077	0	0	755914	881	0	102582	0
940	0	0	702936	784	0	86260	0
972	0	0	719384	793	0	95526	0
978	0	0	735044	816	0	89620	0
942	0	0	731644	796	0	96156	0
983	0	0	720014	805	0	88126	0
1003	0	0	722550	821	0	97134	0
997	0	0	706698	820	0	88444	0
939	0	0	702870	786	0	86200	0
965	0	0	705802	792	0	94980	0
959	0	0	689950	775	0	84898	0
837	0	0	624938	682	0	76456	0

700KB/s network traffic

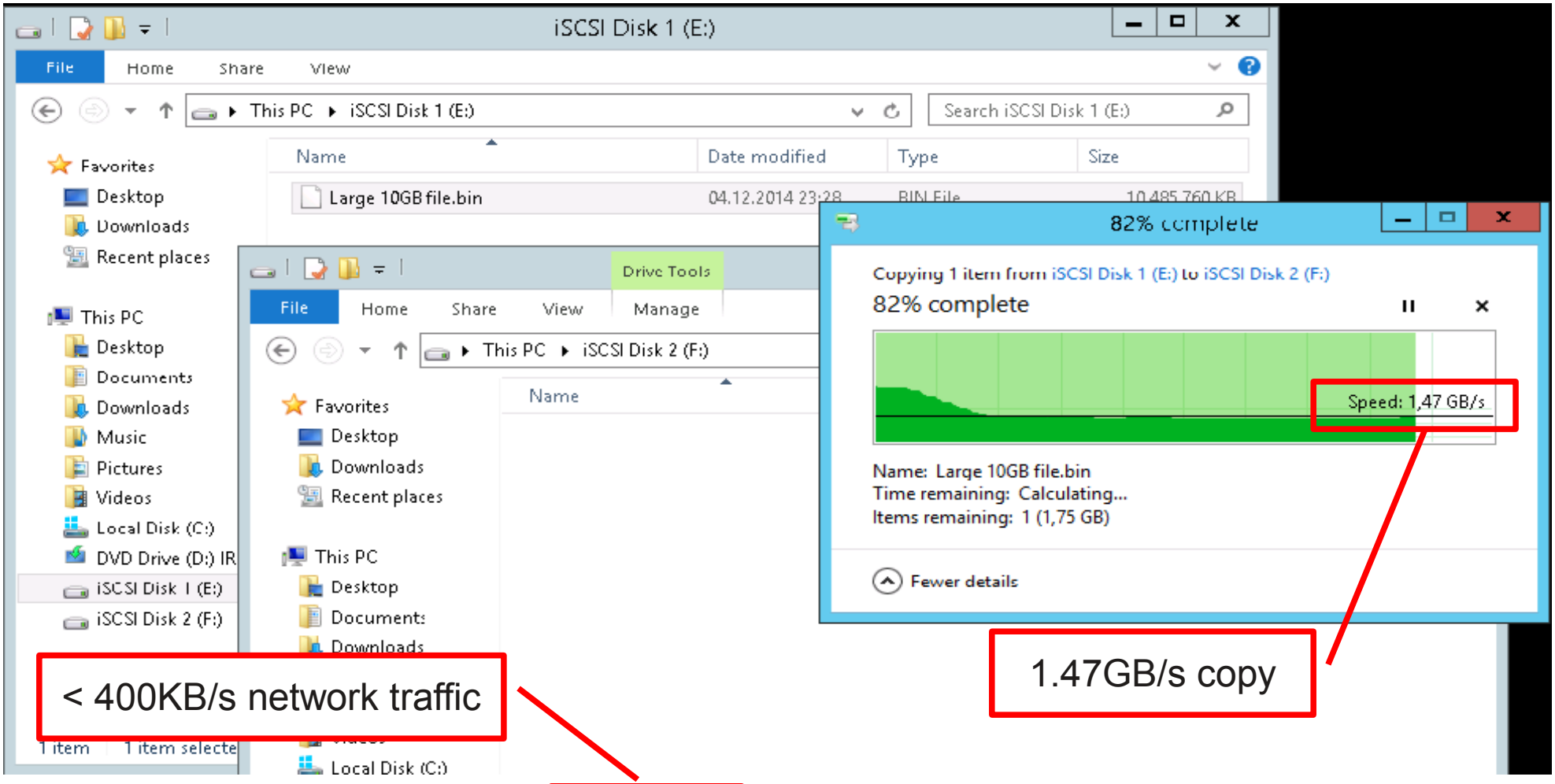
CTL functional improvements

Featured offloaded disk v2 (XCOPY SPC-4/Lite):

- Third Party Copy VPD page
  - Get supported parameters
- POPULATE TOKEN
  - Read data from source to token
- RECEIVE ROD TOKEN INFORMATION
  - Get token
- WRITE USING TOKEN
  - Write data from token to destination

Microsoft ODX

Windows 2012R2 copying 10GB file



packets	errs	idrops	bytes	packets	errs	bytes	colls
26	0	0	16416	26	0	2732	0
270	0	0	381074	152	0	9764	0
28	0	0	17724	28	0	2936	0
27	0	0	17070	27	0	2834	0

CTL functional improvements

Support for ...

Test Unit Ready
Request Sense
Format Unit
Read(6)
Write(6)
Inquiry
Mode select(6)
Reserve(6)
Release(6)
Mode sense(6)
Start stop unit
Read capacity(10)
Read(10)
Write(10)
Write and verify(10)
Verify(10)
Synchronize cache(10)
Read defect data(10)
Write buffer
Read buffer
Write same(10)
Unmap
Log sense
Mode select(10)
Reserve(10)
Release(10)
Mode sense(10)
Persistent reserve in, read keys
Persistent reserve in, read reservation
Persistent reserve in, report capabilities
Persistent reserve in, read full status
Persistent reserve out, register
Persistent reserve out, reserve
Persistent reserve out, release
Persistent reserve out, clear
Persistent reserve out, preempt
Persistent reserve out, preempt and abort
Persistent reserve out, register and ignore existing key
Extended copy(LID1)
Extended copy(LID4)
Populate token
Write using token
Copy operation abort
Receive copy status(LID1)
Receive copy operating parameters
Receive copy failure details(LID1)
Receive copy status(LID4)
Receive ROD token information
Report all ROD tokens
Read(16)
Compare and write
Write(16)
Write and verify(16)
Verify(16)
Synchronize cache(16)
Write same(16)
Write atomic(16)
Read capacity(16)
Get LBA status
Report luns
Report target port groups
Report supported operation codes
Report supported task management functions
Report timestamp
Read(12)
Write(12)
Write and verify(12)
Verify(12)
Read defect data(12)

… 69 commands

CTL performance improvements
... for brute force performance

CTL got long list of optimizations:

- Multiple threads instead of single;
- Per-LUN and per-queue locks instead of single;
- Status and data transfer coalescing;
- UMA zones instead of own allocator;
- Reduced per-LUN memory use;
- Many optimizations to iSCSI and isp(4) code.

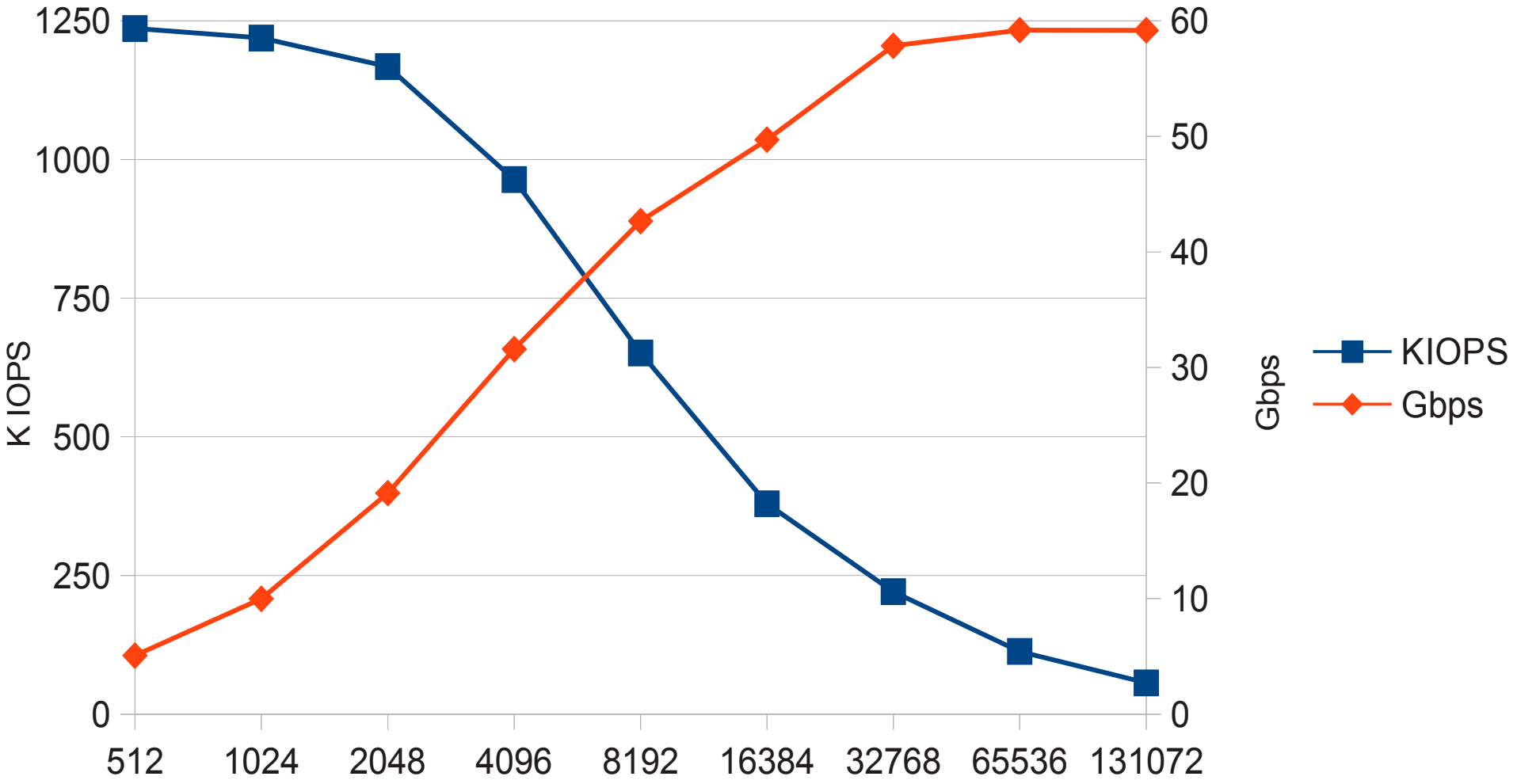
Peak iSCSI IOPS/Throughput

Test setup:

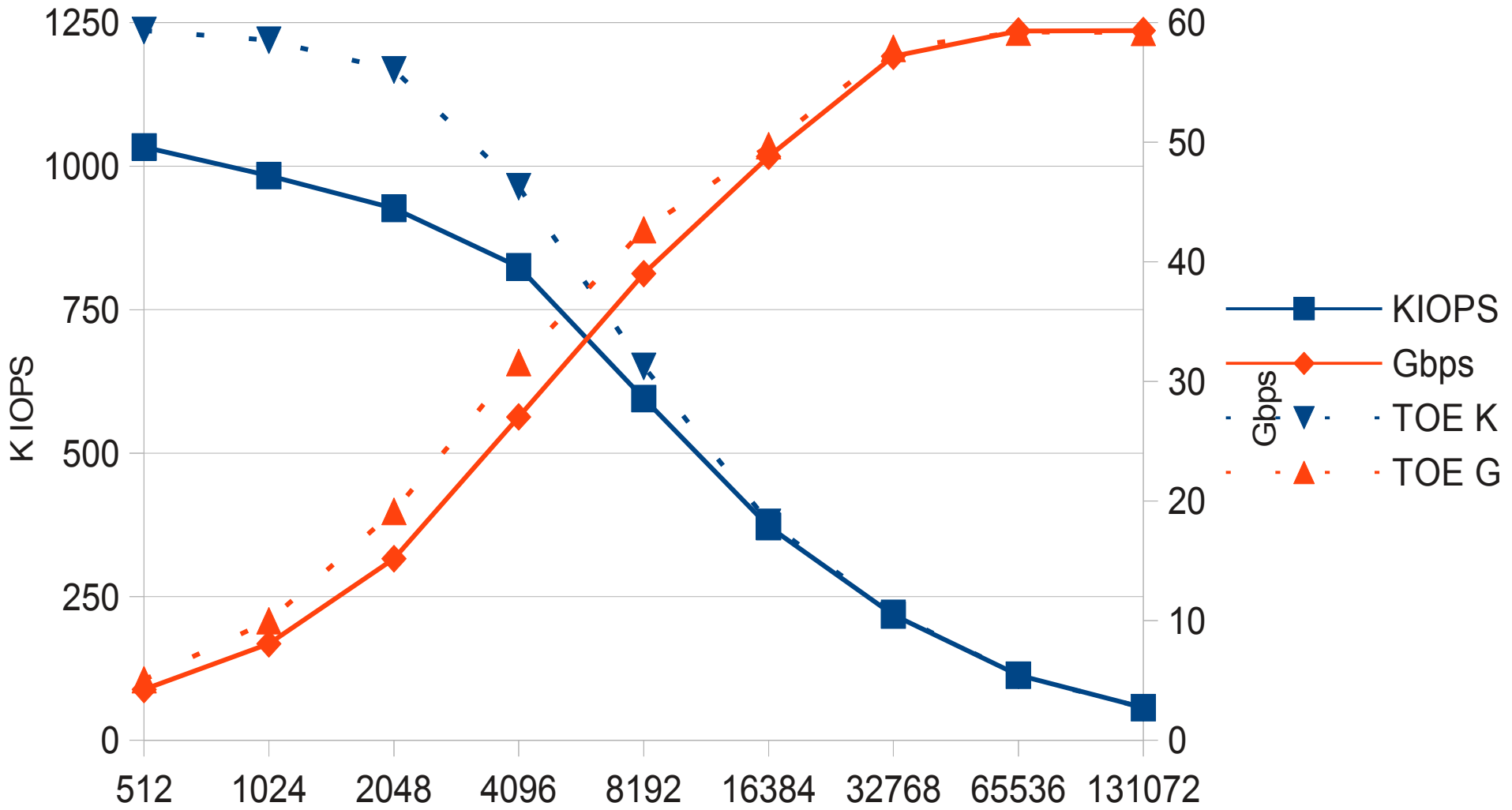
- Target:
  - 2xXeon E5-2690v2 @ 3.00GHz (40 SMT cores)
  - 256GB RAM
  - 1xChelsio T580-LP-CR 40Gbps NIC
  - 1xChelsio T520-CR 2x10Gbps NIC
  - 20xIntel 520/530 Series SSD
  - 6x100GB ZVOL-backed iSCSI LUNs
- Initiators:
  - 3xCore i7 desktop machines with 10/40Gbps NICs
  - 2xiSCSI connections per NIC.

Test: Multi-threaded linear read from all 6 LUNs through each iSCSI connection with different block sizes.

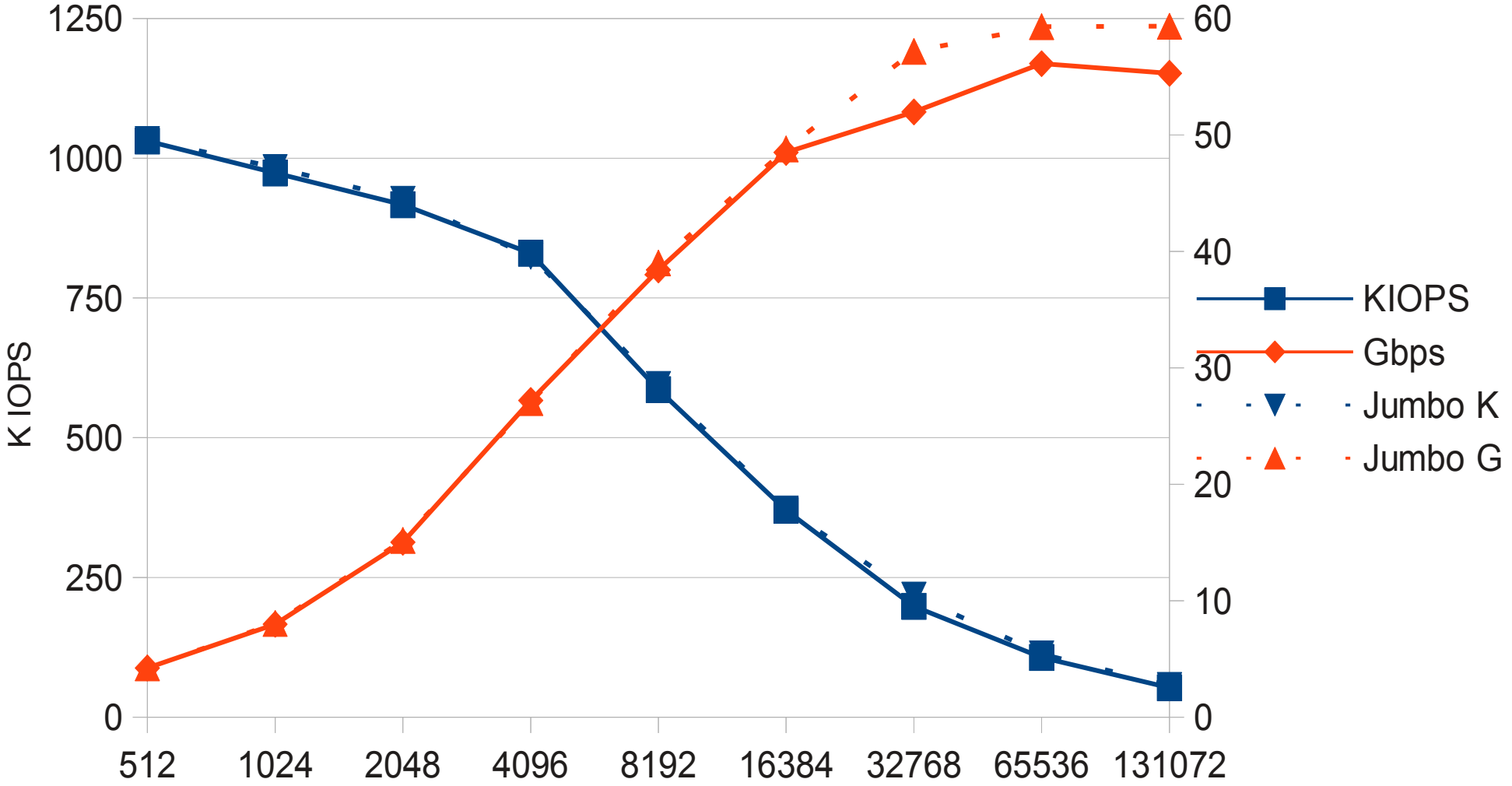
Peak iSCSI IOPS/Throughput
Test 1: Full acceleration (TOE + Jumbo Frames)



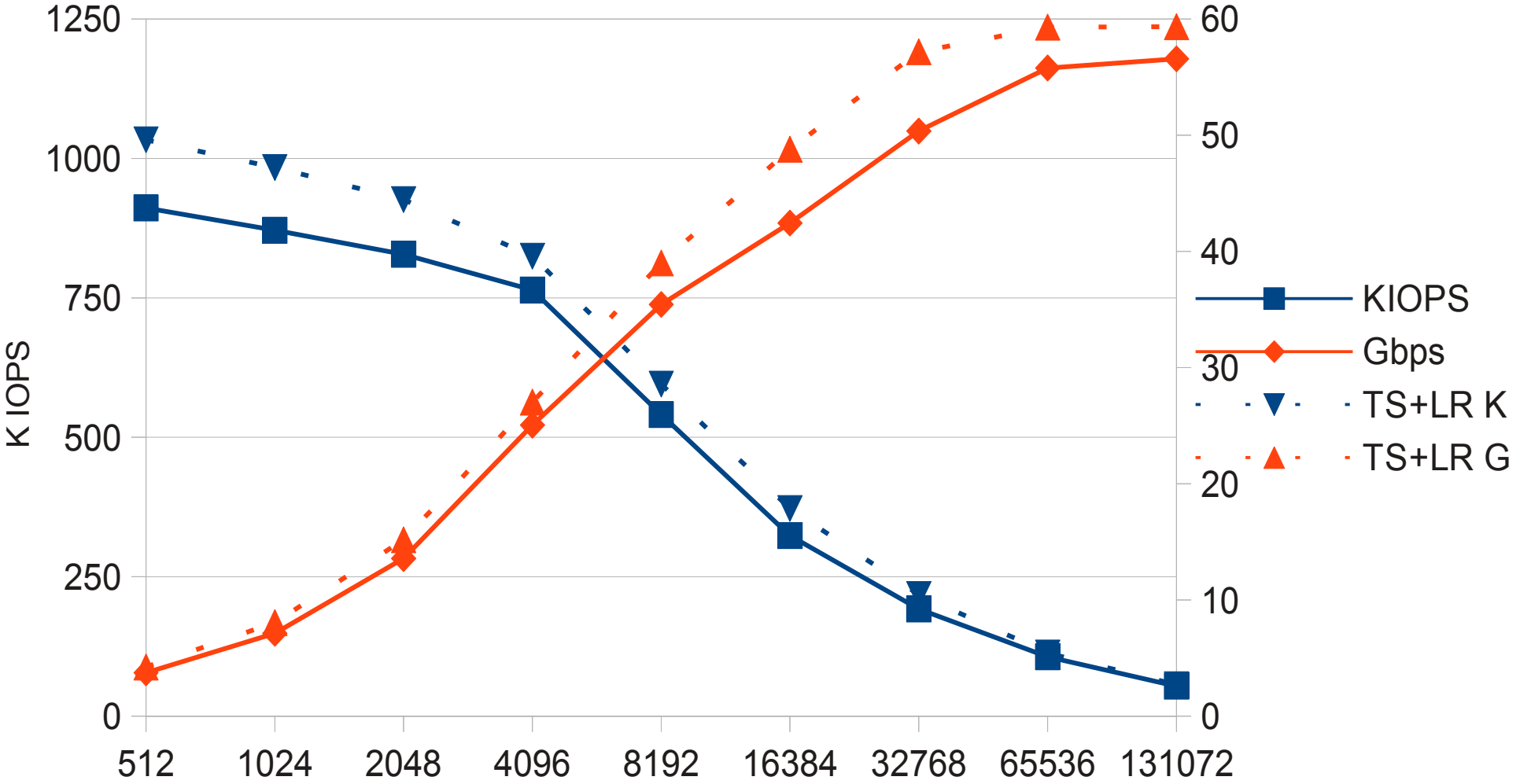
Peak iSCSI IOPS/Throughput
Test 2: T1 - TOE (TSO + LRO + Jumbo Frames)



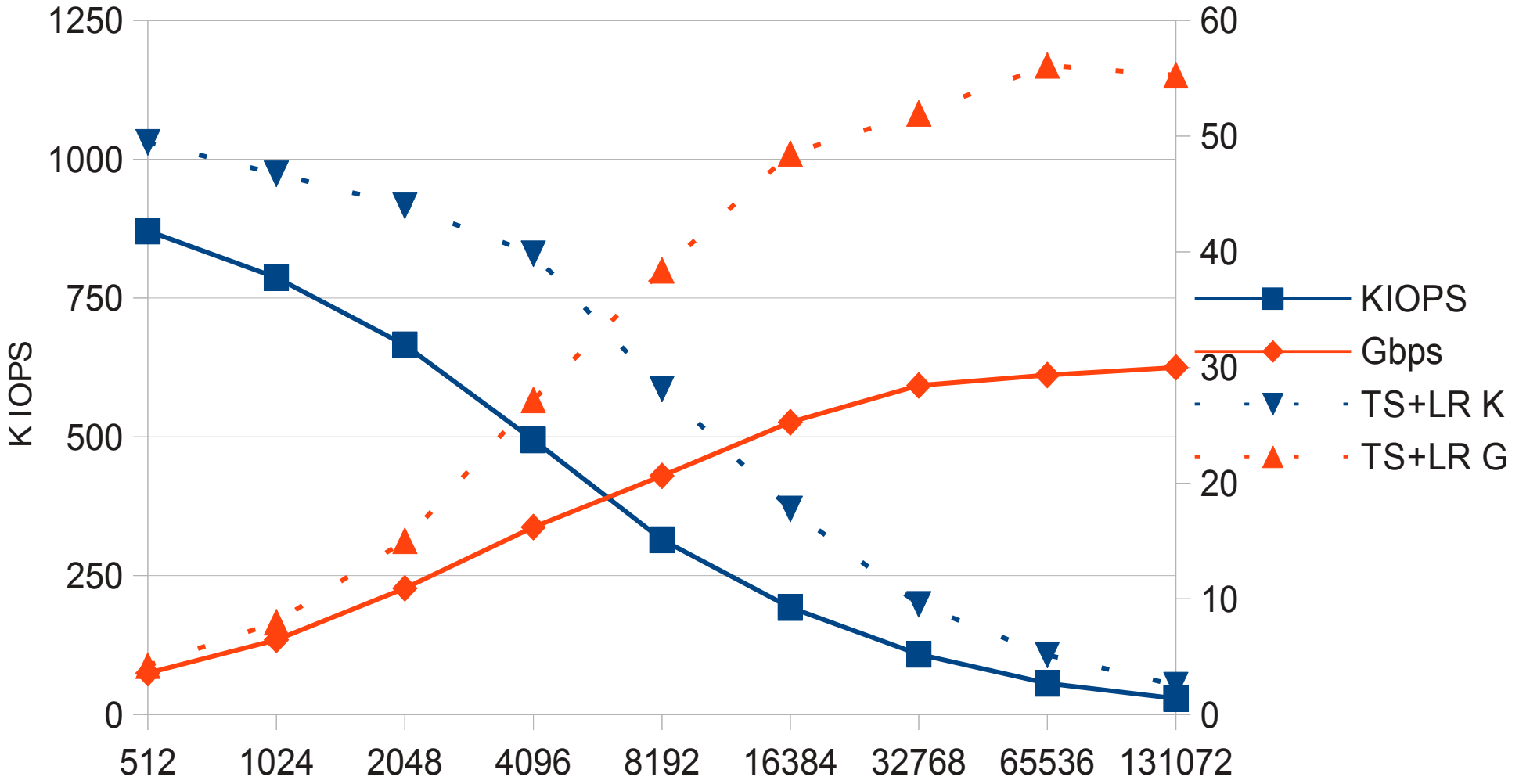
Peak iSCSI IOPS/Throughput
Test 3: T2 - Jumbo (TSO + LRO)



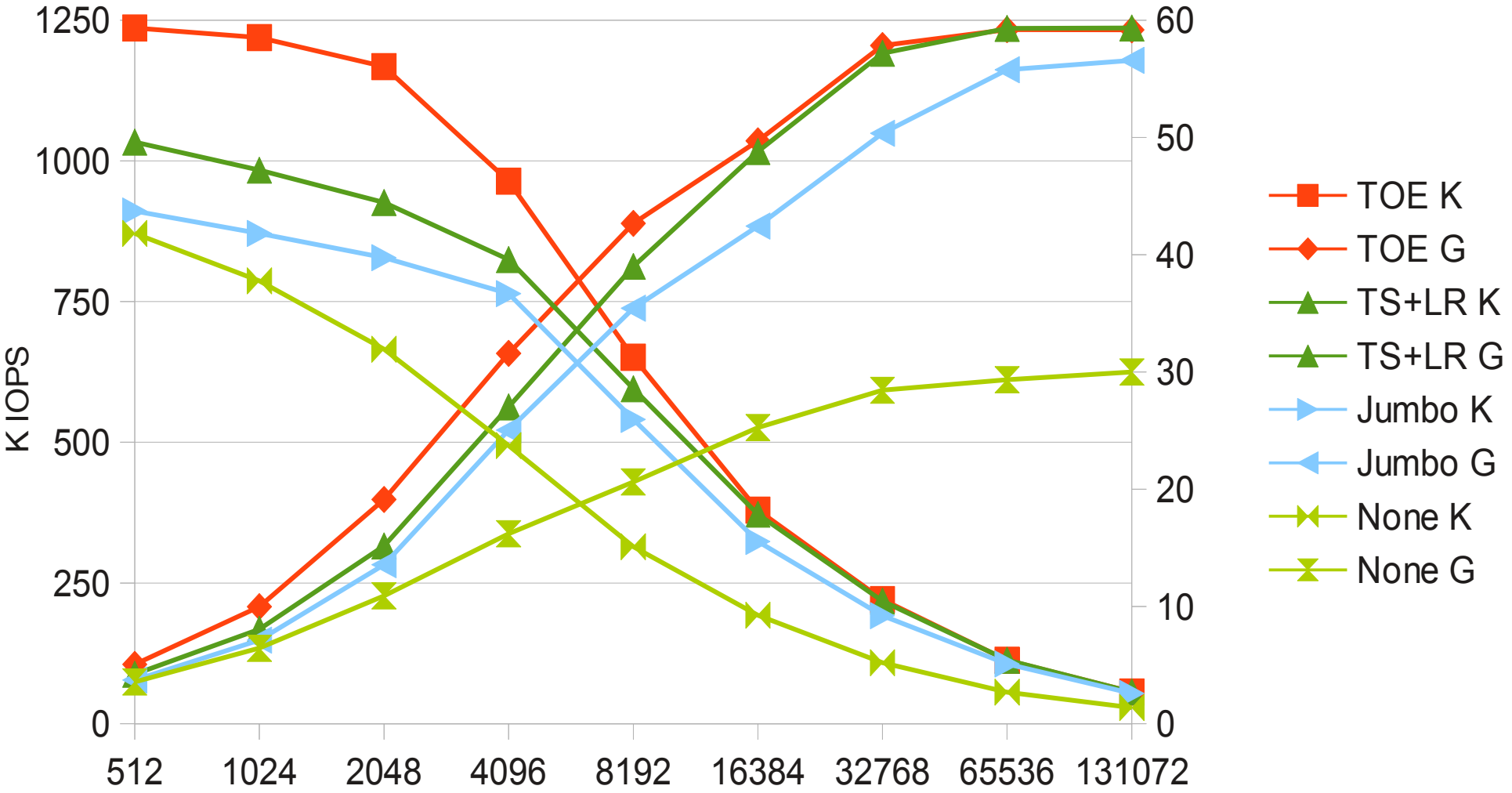
Peak iSCSI IOPS/Throughput
Test 4: T2 - TSO - LRO (Jumbo Frames)



Peak iSCSI IOPS/Throughput
Test 5: T3 - TSO - LRO (No acceleration)



Peak iSCSI IOPS/Throughput Summary graph



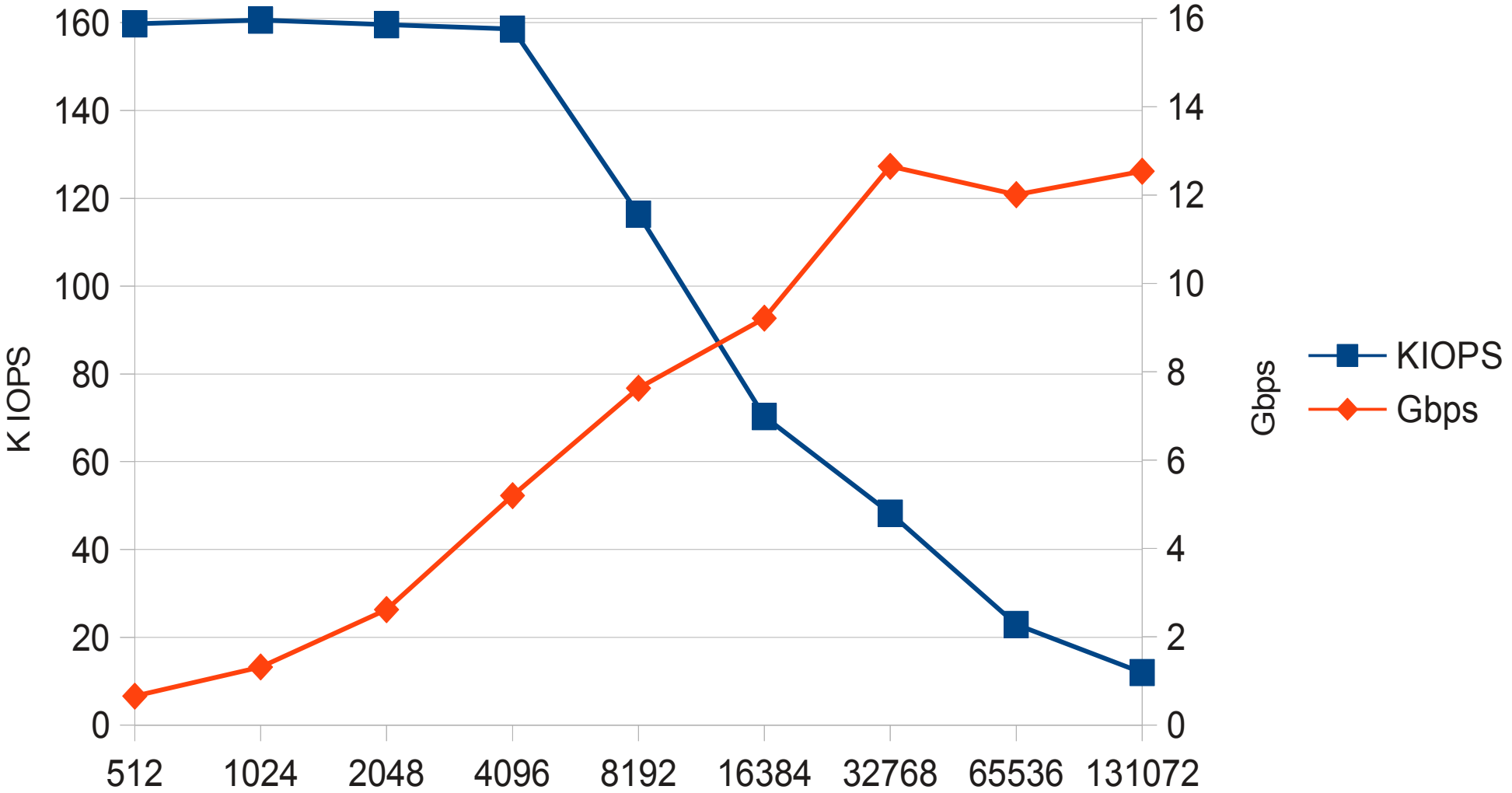
Peak Fibre Channel IOPS/Throughput

Test setup:

- Target:
  - 2xXeon E5-2690v2 @ 3.00GHz (40 SMT cores)
  - 256GB RAM
  - 1xQlogic QLE2562 2x8Gbps FC HBA
  - 20xIntel 520/530 Series SSD
  - 6x100GB ZVOL-backed iSCSI LUNs
- Initiator:
  - Core i7 desktop machine
  - 1xQlogic QLE2562 2x8Gbps FC HBA

Test: Multi-threaded linear read from all 6 LUNs through each FC port with different block sizes.

Peak Fibre Channel IOPS/Throughput



Further improvements

Plans and/or wishes:

- Implement XCOPY without copy (manual dedup)
- Implement XCOPY between hosts
- Recreate HA/clustering support
- Improve ZFS prefetch
- Improve Fibre Channel support

FreeBSD®

FreeNAS®

Thank you!