



**MAKLEE**

software engineering  
solutions

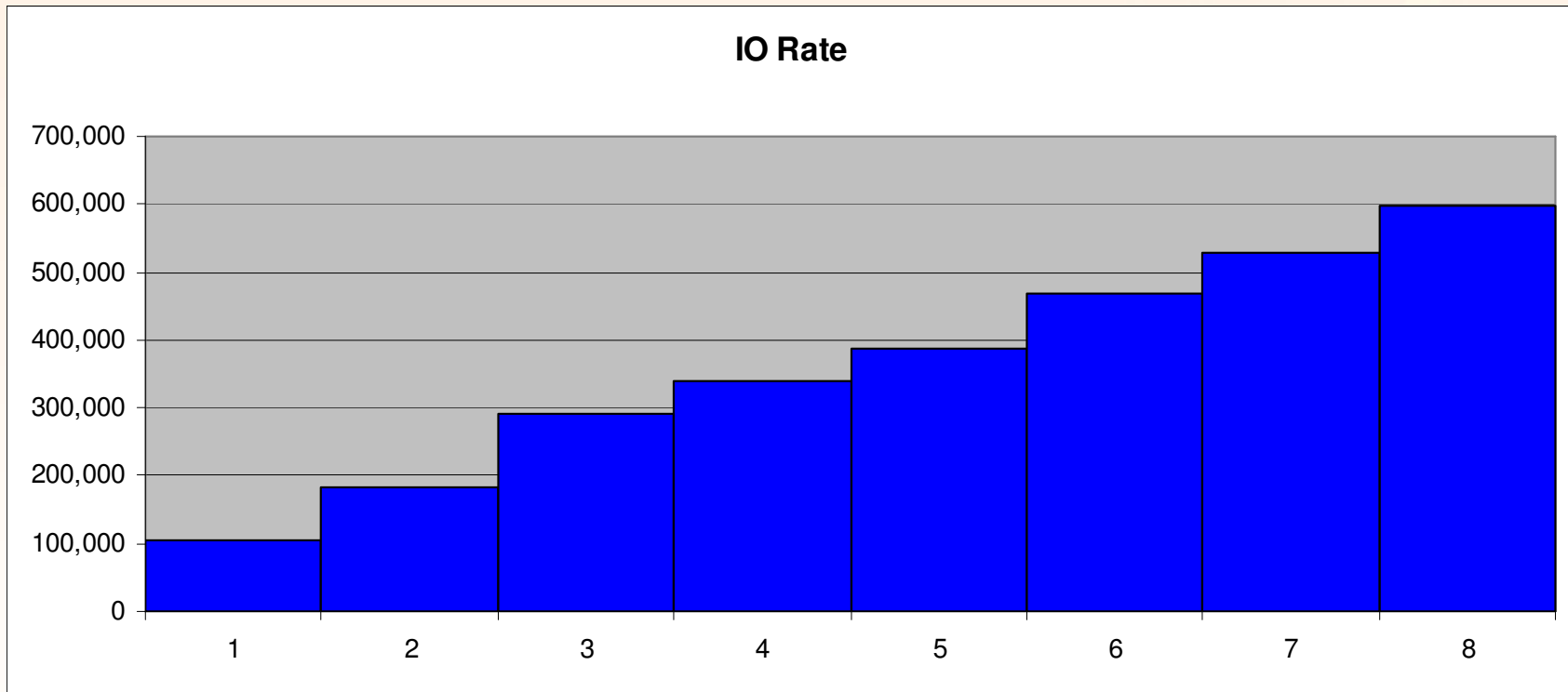
# FusionIO Performance

– Christian Moser  
Chief Technology Officer  
Maklee Engineering  
[cmos@maklee.com](mailto:cmos@maklee.com)

# Summary

- Goal
  - Use HP DL980 to measure IO rate, latency and throughput of FusionIO devices and configure for best performance
- Recommendations
  - Choose PCI slot with best bandwidth
  - Increase number of requests per block device
  - Distribute interrupts
  - Affinitize fct worker threads

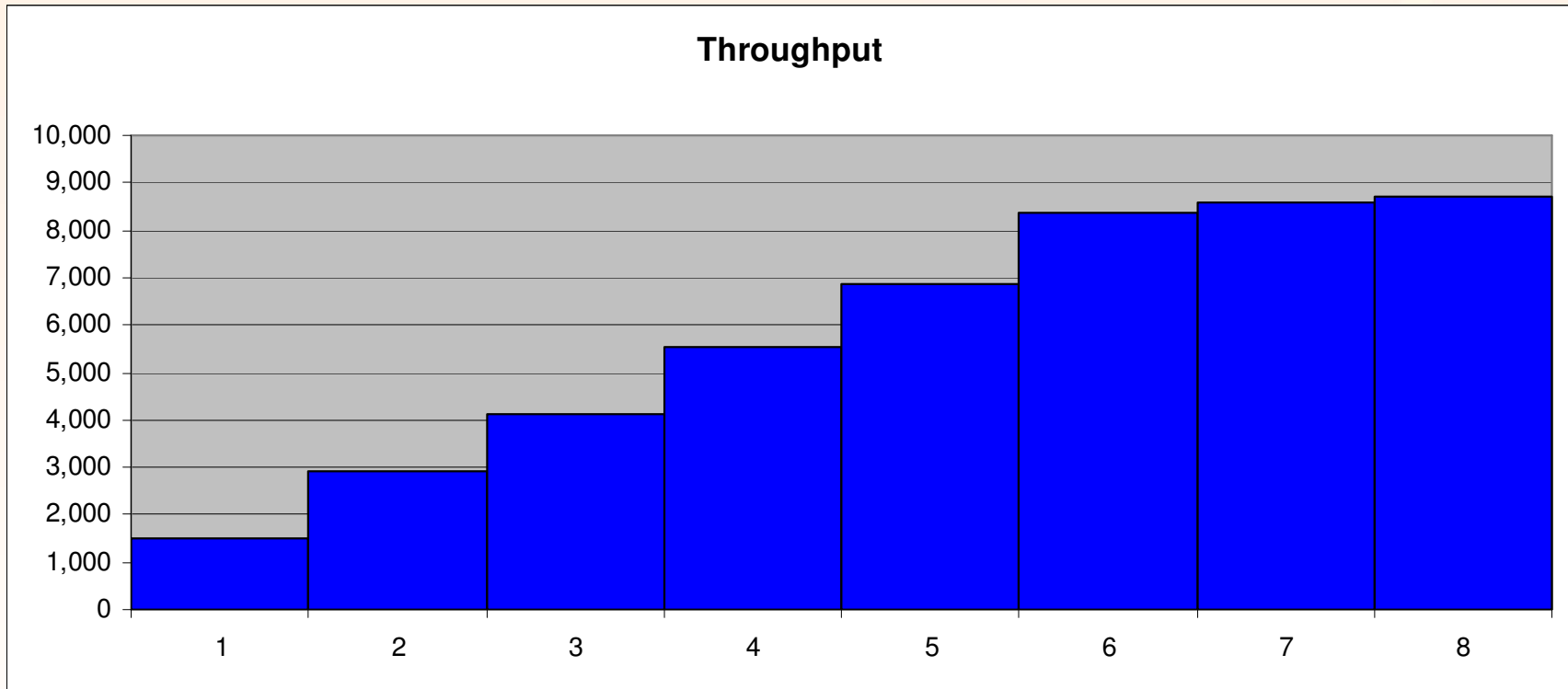
# IO Rate



**IO Rate per number of duos (in req/sec)  
(more is better)**



# Throughput



**Throughput per number of duos (in MB/sec)  
(more is better)**



# IO Rate and Throughput for each Duo

Slot	Type	Device	MB/sec	IO/sec
1	x4 PCI Express x8	fiog, fiqh	842	80,922
2	x8 PCI Express Gen 2 x16			
3	x8 PCI Express Gen 2 x16	fiog, fiop	1,468	99,166
4	x4 PCI Express Gen 2 x8	fioi, fioj	1,421	101,891
5	x8 PCI Express Gen 2 x16	fiom, fion	1,510	102,666
6	x8 PCI Express Gen 2 x16	fiok, fiol	1,516	103,299
7	x4 PCI Express Gen 2 x8	fioe, fiof	1,420	97,606
8	x4 PCI Express Gen 2 x8	fioc, fiod	1,427	104,398
9	x8 PCI Express Gen 2 x16	fioa, fiob	1,512	106,387
10	x4 PCI Express Gen 2 x8			
11	x8 PCI Express Gen 2 x16			

PCI slot location matters  
(each duo was tested separately)



# IO Rate

- Various combination of duos

Slot	Type	Device	IO/sec	IO/sec	IO/sec	IO/sec	IO/sec
1	x4	fiog, fioh	80,922			X	
2	x8 Gen 2						
3	x8 Gen 2	fioo, fiop	99,166			X	X
4	x4 Gen 2	fioi, fioj	101,891		X	X	
5	x8 Gen 2	fiof, fiom	102,666		X	X	X
6	x8 Gen 2	fiok, fiol	103,299		X	X	X
7	x4 Gen 2	fioe, fiog	97,606	X		X	X
8	x4 Gen 2	fioc, fiod	104,398	X		X	X
9	x8 Gen 2	fioa, fiob	106,387	X		X	X
10	x4 Gen 2						
11	x8 Gen 2						
<b>Total IO/sec</b>				<b>291,405</b>	<b>286,260</b>	<b>532,309</b>	<b>469,400</b>



MAKLEE

More than 500K IO/sec



# IO Throughput

- Various combination of duos

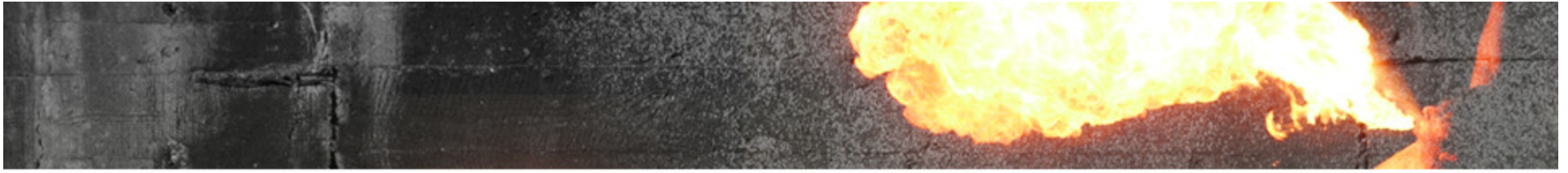
Slot	Type	Device	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec	MB/sec
1	x4	fiog, fioh	842			X	X				X		
2	x8 Gen 2												
3	x8 Gen 2	fioo, fiop	1,468			X	X	X		X	X		X
4	x4 Gen 2	fioi, fioj	1,421		X		X			X	X	X	X
5	x8 Gen 2	fiom, fion	1,510		X		X	X	X	X	X	X	X
6	x8 Gen 2	fiok, fiol	1,516		X		X	X	X	X	X	X	X
7	x4 Gen 2	fioe, fiof	1,420	X			X	X				X	X
8	x4 Gen 2	fioc, fiod	1,427	X			X	X	X			X	X
9	x8 Gen 2	fioa, fiob	1,512	X			X	X	X			X	X
10	x4 Gen 2												
11	x8 Gen 2												
<b>Total MB/sec</b>				<b>4,057</b>	<b>4,118</b>	<b>1,695</b>	<b>6,433</b>	<b>8,353</b>	<b>5,441</b>	<b>5,536</b>	<b>4,253</b>	<b>6,500</b>	<b>6,535</b>



MAKLEE

More than 8 GB/sec throughput





# *Technical Slides*



MAKLEE



# System Configuration

- Hardware
  - HP ProLiant DL980 G7 Server
  - Intel Xeon E7-4870 processors @2.4 GHz
  - 8 deca-core, hyperthreads disabled, 80 logical CPUs
  - 1 TB physical memory
- Operating System
  - Red Hat Enterprise Linux Server release 5.6 (Carthage)
  - Version 2.6.18-238.12.1.el5
  - 64-bit kernel (x86), 4 KB pagesize
- Storage Sub-System
  - Fusion ioDrive
  - 8 x HP 1.28TB MLC PCIe IO Accelerator

# PCI Configuration

- Use 'dmidecode -t 9' to display system slot information
  - Slot number and type
  - Bandwidth
    - x4 PCI-e 2.0 is limited to 4 x 5Gb/s => 20Gb/s => 2GB/s
    - x8 PCI-e 2.0 is limited to 8 x 5Gb/s => 40Gb/s => 4GB/s
- Use 'lspci' to find bus number and adapter and then match with dmidecode output
  - FusionIO drives are connected to PCI bridges
- Use 'fio-status -a' to display information about FusionIO adapters and devices
- Example
  - 3rd FusionIO duo is in system slot #7 (devices /dev/fioe and /dev/fiof)



# PCI Configuration (cont'd)

- Find PCI slot numbers

```
# dmidecode -t 9
[...]
Handle 0x0906, DMI type 9, 17 bytes
System Slot Information
  Designation: PCI-E Slot 7
  Type: x4 PCI Express Gen 2 x8
  Current Usage: In Use
  Length: Long
  Characteristics:
    3.3 V is provided
    PME signal is supported
  Bus Address: 0000:05:00.0
```

```
Handle 0x0907, DMI type 9, 17 bytes
System Slot Information
  Designation: PCI-E Slot 8
  Type: x4 PCI Express Gen 2 x8
  Current Usage: In Use
  Length: Long
  Characteristics:
    3.3 V is provided
    PME signal is supported
  Bus Address: 0000:0a:00.0
```

```
Handle 0x0908, DMI type 9, 17 bytes
System Slot Information
  Designation: PCI-E Slot 9
  Type: x8 PCI Express Gen 2 x16
  Current Usage: In Use
  Length: Long
  Characteristics:
    3.3 V is provided
    PME signal is supported
  Bus Address: 0000:0f:00.0
```

```
Handle 0x0909, DMI type 9, 17 bytes
System Slot Information
  Designation: PCI-E Slot10
  Type: x4 PCI Express Gen 2 x8
  Current Usage: Available
  Length: Long
  Characteristics:
    3.3 V is provided
    PME signal is supported
  Bus Address: 0000:14:00.0
```

```
Handle 0x090A, DMI type 9, 17 bytes
System Slot Information
  Designation: PCI-E Slot11
  Type: x8 PCI Express Gen 2 x16
  Current Usage: Available
  Length: Long
  Characteristics:
    3.3 V is provided
    PME signal is supported
  Bus Address: 0000:17:00.0
```

[...]



# PCI Configuration (cont'd)

- Find bus address of bridge where FusionIO device is connected to

```
# lspci -D
[...]
0000:05:00.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:06:04.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:06:05.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:06:06.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:08:00.0 Mass storage controller: Fusion-io ioDimm3 (rev 01)
0000:09:00.0 Mass storage controller: Fusion-io ioDimm3 (rev 01)
0000:0a:00.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:0b:04.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:0b:05.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:0b:06.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:0d:00.0 Mass storage controller: Fusion-io ioDimm3 (rev 01)
0000:0e:00.0 Mass storage controller: Fusion-io ioDimm3 (rev 01)
0000:0f:00.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:10:04.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:10:05.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:10:06.0 PCI bridge: PLX Technology, Inc. PEX 8616 16-lane, 4-Port PCI Express Gen 2 (5.0 GT/s)
0000:12:00.0 Mass storage controller: Fusion-io ioDimm3 (rev 01)
0000:13:00.0 Mass storage controller: Fusion-io ioDimm3 (rev 01)
0000:14:00.0 Fibre Channel: QLogic Corp. ISP2532-based 8Gb Fibre Channel to PCI Express HBA (rev 02)
0000:14:00.1 Fibre Channel: QLogic Corp. ISP2532-based 8Gb Fibre Channel to PCI Express HBA (rev 02)
0000:17:00.0 Fibre Channel: QLogic Corp. ISP2532-based 8Gb Fibre Channel to PCI Express HBA (rev 02)
0000:17:00.1 Fibre Channel: QLogic Corp. ISP2532-based 8Gb Fibre Channel to PCI Express HBA (rev 02)
[...]
```



# PCI Configuration (cont'd)

- Match PCI address of fct device (slot 7 has fct4 = /dev/fioe)

```
# fio-status -a
```

```
Found 16 ioDrives in this system with 8 ioDrive Duos  
Fusion-io driver version: 2.3.1 build 123
```

```
[...]
```

```
Adapter: ioDrive Duo
```

```
HP 1280GB MLC PCIe ioDrive Duo for ProLiant Servers, Product Number:641027-B21 SN:425140  
ioDrive Duo HL, PN:00190000107, Mfr:003, Date:20110620
```

```
[...]
```

```
PCIe negotiated link: 4 lanes at 5.00 Gbits/sec each, 2000 MBytes/sec total
```

```
Connected ioDimm modules:
```

```
fct4: HP 1280GB MLC PCIe ioDrive Duo for ProLiant Servers, Product Number:641027-B21 SN:424727  
fct5: HP 1280GB MLC PCIe ioDrive Duo for ProLiant Servers, Product Number:641027-B21 SN:424716
```

```
fct4 Attached as 'fioe' (block device)
```

```
HP 1280GB MLC PCIe ioDrive Duo for ProLiant Servers, Product Number:641027-B21 SN:424727  
ioDIMM 640, PN:00277100604, Mfr:003, Date:20110620
```

```
Located in slot 0 Upper of ioDrive Duo SN:425140
```

```
Powerloss protection: protected
```

```
PCI:08:00.0
```

```
Vendor:1aed, Device:1005, Sub vendor:103c, Sub device:176f
```

```
Firmware v5.0.6, rev 101583
```

```
640.00 GBytes block device size, 812 GBytes physical device size
```

```
Format: block, v300, 1,250,001,920 sectors, 512 bytes per sector
```

```
[...]
```

```
PCIe negotiated link: 4 lanes at 2.50 Gbits/sec each, 1000 MBytes/sec total
```

```
Internal temperature: 50.7 degC, max 54.1 degC
```

```
Board temperature: 44 degC
```



MAKLEE



# IO Requests

- Maximum size of IO request for FusionIO drives is 128 KB
  - Cannot be changed
  - Requires driver modification
- Default maximum number of outstanding IO requests per device is 128
  - Can be changed
  - Setting to 4096 will give a 10% performance boost

```
# cat /sys/block/fioa/queue/nr_requests
128
# echo 4096 > /sys/block/fioa/queue/nr_requests

# cat /sys/block/fioa/queue/max_sectors_kb
128
# cat /sys/block/fioa/queue/max_hw_sectors_kb
128
# echo 1024 > /sys/block/fioa/queue/max_sectors_kb
-bash: echo: write error: Invalid argument
```



# Interrupts

- Find out interrupt number for FusionIO devices

```
# cat /proc/interrupts | grep fct
      CPU0      CPU1      ...      CPU77      CPU78
66:    2349465    7985609    ...    2244421    2212862    IO-APIC-level    iodrive-fct6, iodrive-fct7
82:    3803303    3192953    ...    11566128    10975799    IO-APIC-level    iodrive-fct8
90:    5293857    2886859    ...      806999    8739663    IO-APIC-level    iodrive-fct9
106:   4742011    3372508    ...    5000348    3004360    IO-APIC-level    iodrive-fct10
114:   4486469    2804446    ...    5042572    5378870    IO-APIC-level    iodrive-fct11
130:   4641609    3145305    ...    4435407    2689418    IO-APIC-level    iodrive-fct12
138:   3568994    4286899    ...    2920571    2766660    IO-APIC-level    iodrive-fct13
154:   4083644    3878031    ...    3527673    2342842    IO-APIC-level    iodrive-fct14
162:   3025641    3541571    ...    2221282    2244549    IO-APIC-level    iodrive-fct15
177:   4446189    3434100    ...    2053803    5046910    IO-APIC-level    iodrive-fct0
185:   1547292    3981810    ...    2554187    3300409    IO-APIC-level    iodrive-fct1
201:   2225524    4640512    ...    1665550    7612210    IO-APIC-level    iodrive-fct2
209:   3083077    3442566    ...    3647490    3701438    IO-APIC-level    iodrive-fct3
225:   2085092    4944081    ...    1155726    3137496    IO-APIC-level    iodrive-fct4
233:   2402016    4522325    ...    2553983    6309903    IO-APIC-level    iodrive-fct5
```



## Interrupts (cont'd)

- Check if 'irqbalance' is running
  - # ps -ef | grep irqbalance
  - Irqbalance does a good job of distributing interrupts
  - Is not NUMA-aware in RHEL 5
- Need to kill irqbalance process if you plan to manually affinitized interrupts
- Monitor how interrupts are distributed
  - Collect sar data
  - # mpstat -P ALL 1 10
- Verify which CPU handles a given interrupt
  - Interrupts for device /dev/fioa are currently affinitized to CPU 39

```
# cat /proc/irq/177/smp_affinity
00000000,00000000,00000000,00000000,00000000,00000000,00000080,00000000
```



## Interrupts (cont'd)

- DL980 locality of processor sockets and I/O hub (IOH)
  - IOH for slots #1...#6 closest to processors #2 and #3
  - IOH for slots #7...#11 closest to processors #0 and #1
  - IOH for slots #12...#16 closest to processors #6 and #7
- Use the following commands to affinitize FusionIO device interrupts

```
echo 10 > /proc/irq/177/smp_affinity
echo 4000 > /proc/irq/185/smp_affinity
echo 20 > /proc/irq/201/smp_affinity
echo 8000 > /proc/irq/209/smp_affinity
echo 40 > /proc/irq/225/smp_affinity
echo 10000 > /proc/irq/233/smp_affinity
echo 1000000 > /proc/irq/66/smp_affinity
echo 2000000 > /proc/irq/82/smp_affinity
echo 4,00000000 > /proc/irq/90/smp_affinity
echo 4000000 > /proc/irq/106/smp_affinity
echo 8,00000000 > /proc/irq/114/smp_affinity
echo 8000000 > /proc/irq/130/smp_affinity
echo 10,00000000 > /proc/irq/138/smp_affinity
echo 10000000 > /proc/irq/154/smp_affinity
echo 20,00000000 > /proc/irq/162/smp_affinity
```



# fct Worker Threads

- Each FusionIO device has its own worker thread
  - Does the heavy lifting
  - Consumes lots of CPU cycles
  - Should run on a core of a processor close to the IOH with the corresponding FusionIO device
    - fct0-worker should be close to /dev/fioa
  - Find pid and then use taskset to affinitize worker thread

```
# ps -ef | grep fct
root      14773  8387   1 Sep02 ?          01:15:52 [fct0-worker]
root      17379  8387   1 Sep02 ?          01:03:34 [fct1-worker]
root      17477  8387   1 Sep02 ?          01:10:03 [fct2-worker]
root      17491  8387   1 Sep02 ?          01:00:39 [fct3-worker]
[...]
```

```
# taskset -p 14773
pid 14773's current affinity mask: ffffffff
```

```
# taskset -pc 0-19 14773
pid 14773's current affinity list: 0-79
pid 14773's new affinity list: 0-19
```



# ORION

- Use Oracle's IO utility to measure IO rate, latency and throughput
- Measure IO rate and latency
  - Use random 8K reads to simulate single block reads
  - # ./orion\_linux\_x86-64 -testname fio -run oltp -duration 10 -simulate raid0
- Measure throughput
  - Use random 1M large reads
  - # ./orion\_linux\_x86-64 -testname fio -run dss -duration 10 -simulate raid0
- However Orion has some internal limits and a single job cannot drive hard enough and saturate the FusionIO drives
  - For example asynch IO limit is 2048
- Solution is to run multiple Orion jobs in parallel



# ORION (cont'd)

- Scripts to run IO rate and throughput workload

```
# cat mbps.sh
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 2048
    -num_small 0 -size_large 1024 -type rand -duration 60 -simulate raid0 &
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 2048
    -num_small 0 -size_large 1024 -type rand -duration 60 -simulate raid0 &
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 2048
    -num_small 0 -size_large 1024 -type rand -duration 60 -simulate raid0 &
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 2048
    -num_small 0 -size_large 1024 -type rand -duration 60 -simulate raid0 &

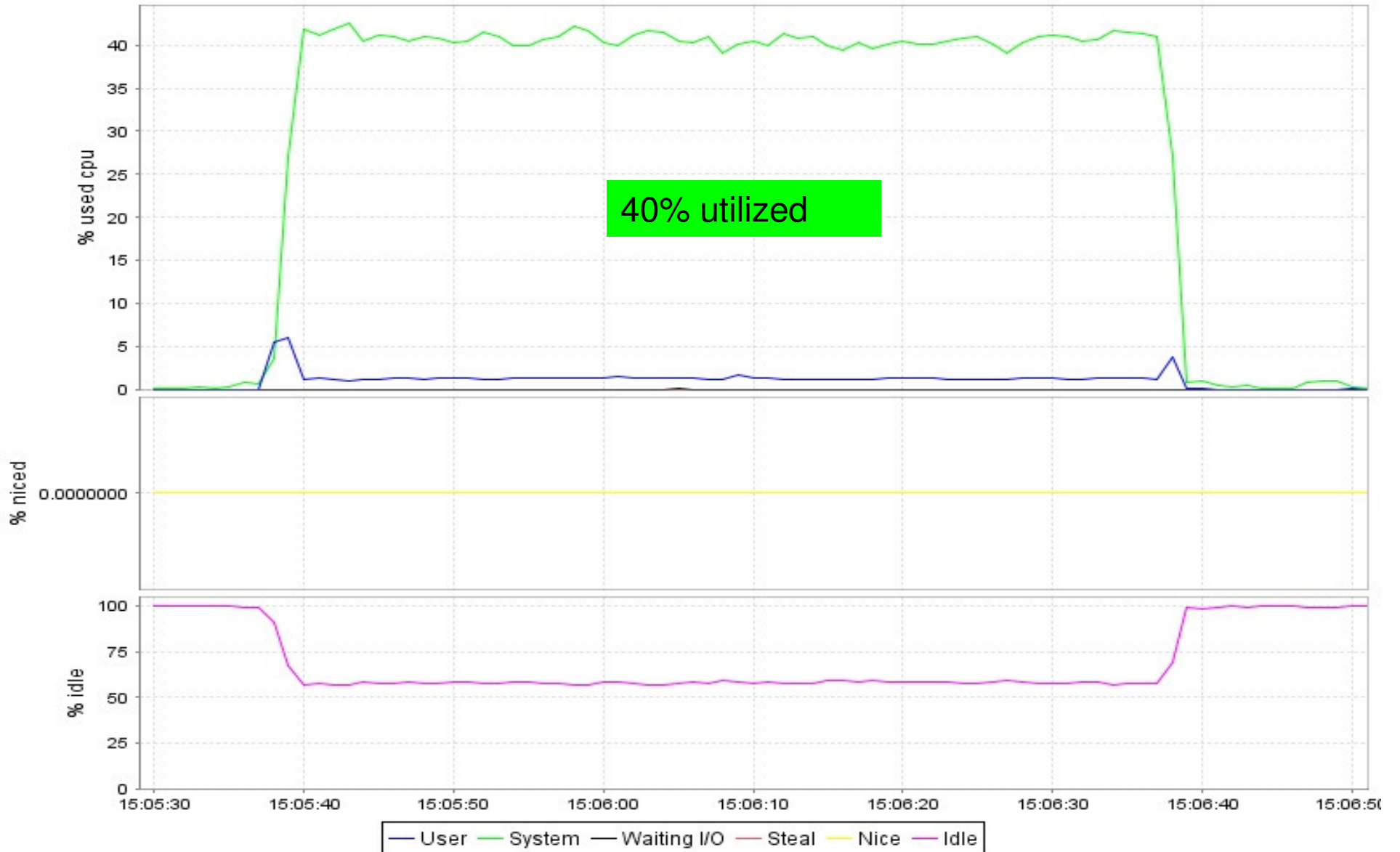
# cat iops.sh
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 0
    -num_small 2048 -size_small 8 -type rand -duration 60 -simulate raid0 &
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 0
    -num_small 2048 -size_small 8 -type rand -duration 60 -simulate raid0 &
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 0
    -num_small 2048 -size_small 8 -type rand -duration 60 -simulate raid0 &
./orion_linux_x86-64 -testname fio -run advanced -matrix point -num_large 0
    -num_small 2048 -size_small 8 -type rand -duration 60 -simulate raid0 &

# cat fio.lun
/dev/fioa
/dev/fiob
/dev/fioc
/dev/fiod
[...]
```



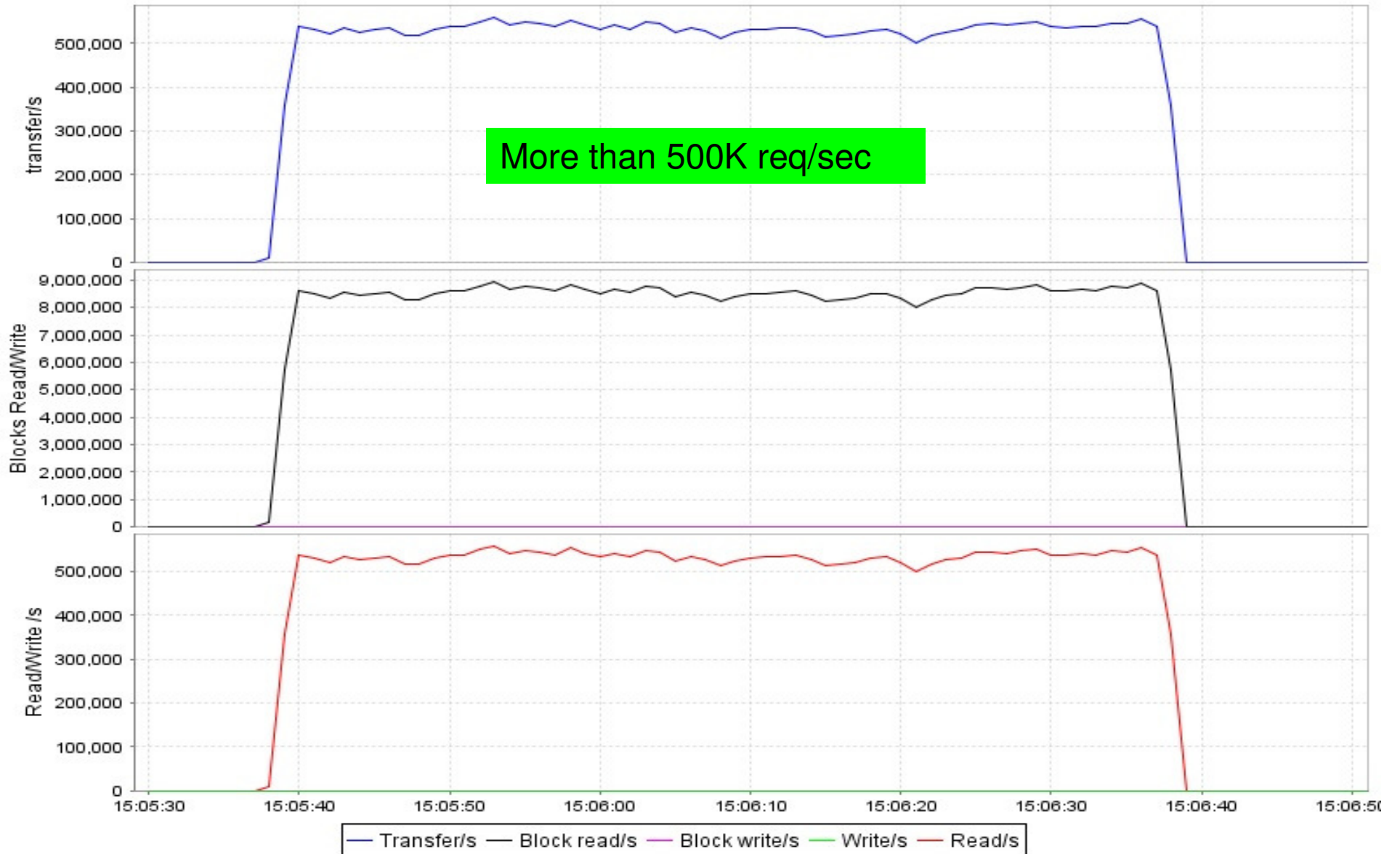
# IO Rate – CPU Utilization

CPU all for dl980g72



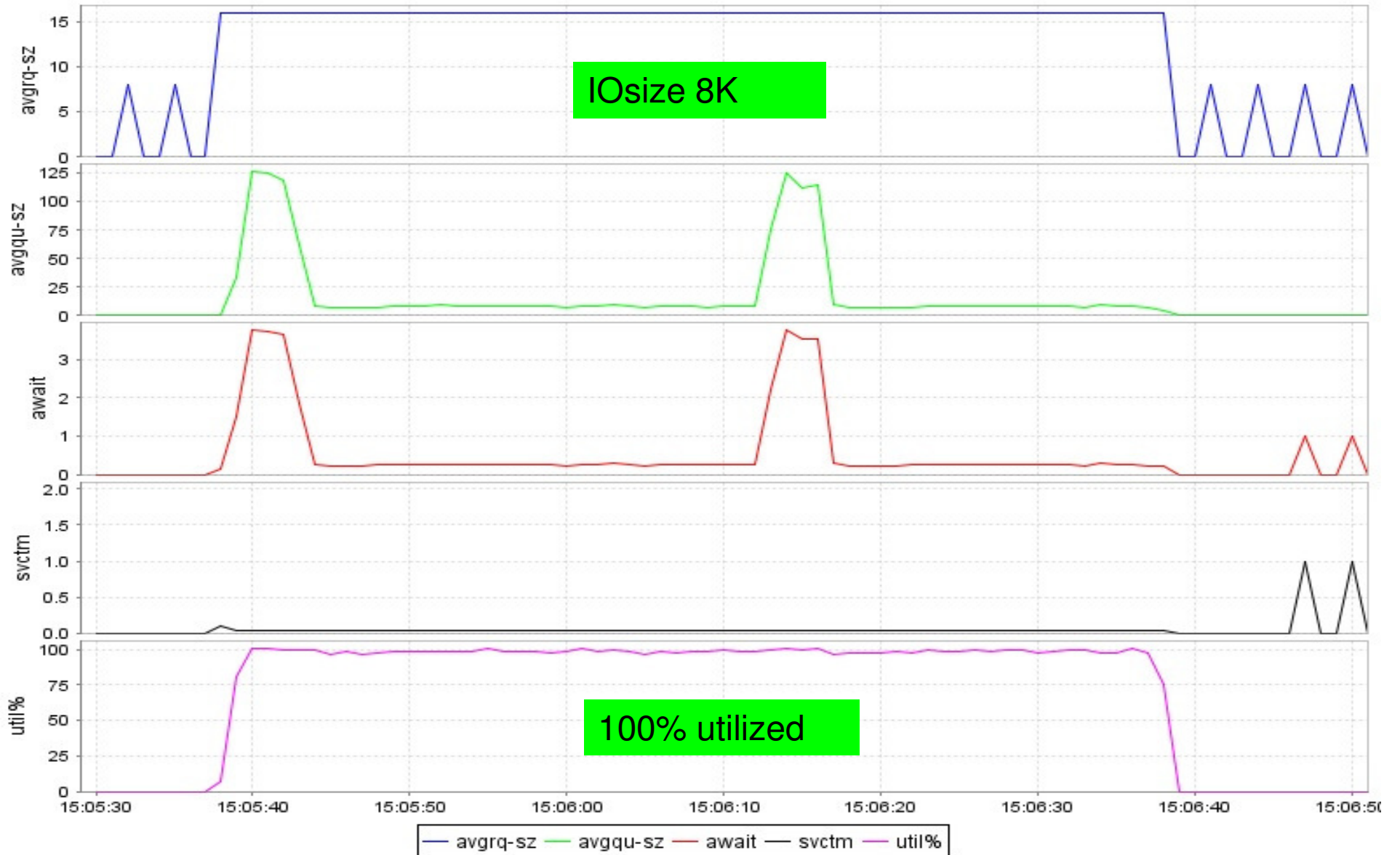
# IO Rate - Requests

I/O for dl980g72



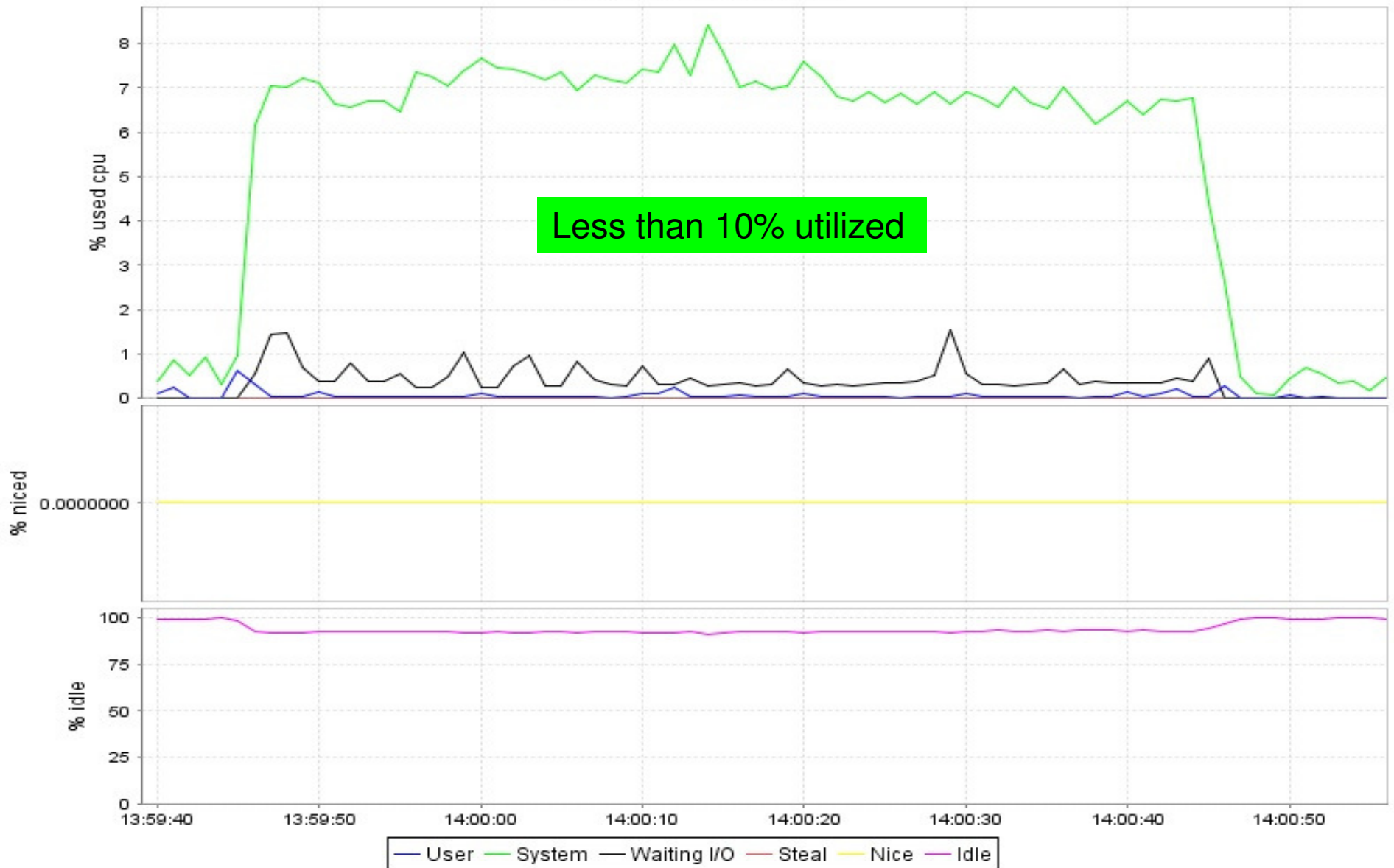
# IO Rate – FusionIO Device

Block Wait dev252-0 for dl980g72



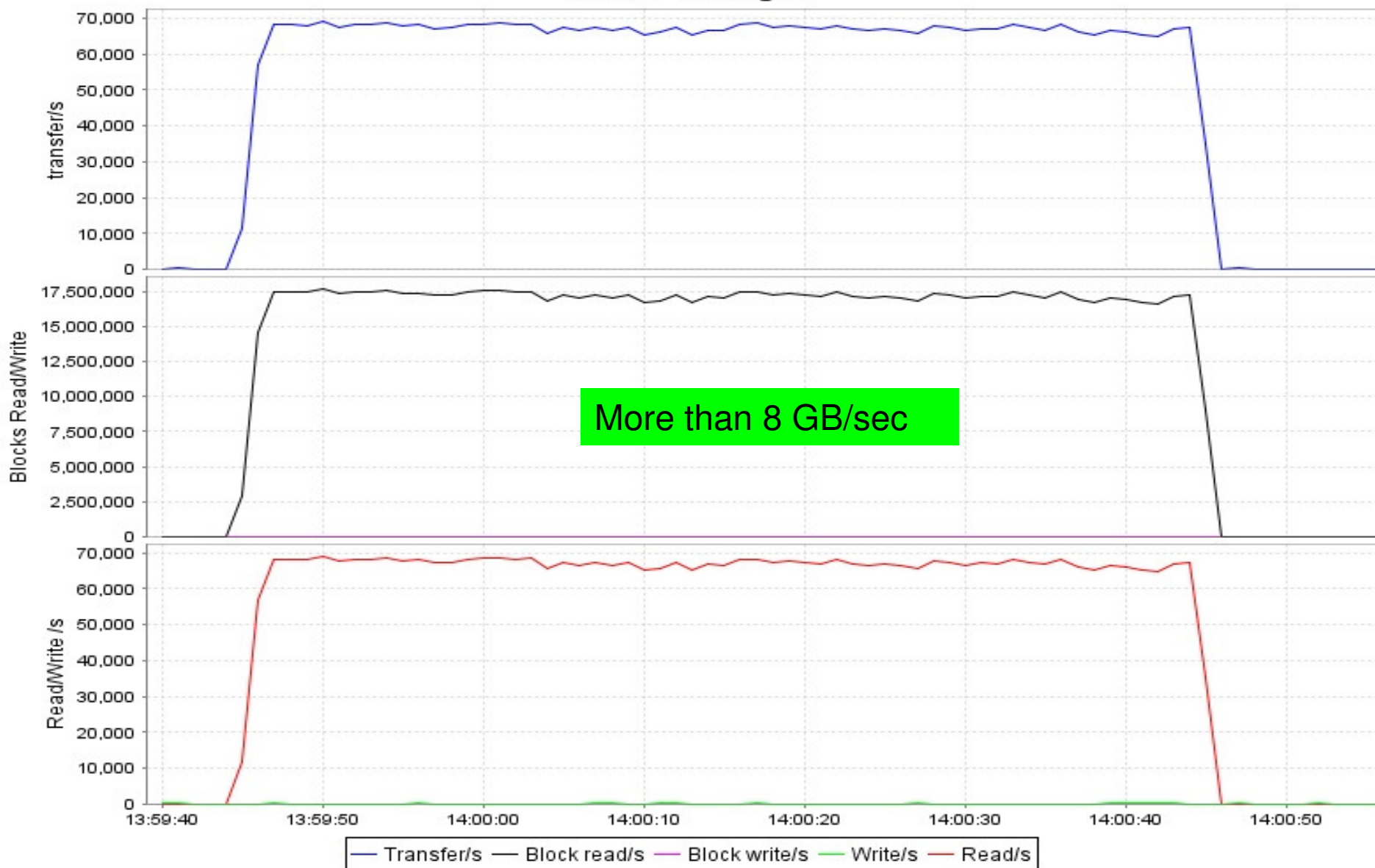
# Throughput – CPU Utilization

CPU all for dl980g72



# Throughput - Bandwidth

I/O for dl980g72



# Throughput – FusionIO Device

Block Wait dev252-0 for dl980g72

