Hardware Evaluation for FY2013 RACF Processor Farm Procurement

---

Chris Hollowell <hollowec@bnl.gov>
Tony Wong <tony@bnl.gov>
RHIC/ATLAS Computing Facility

Evaluation System Specfications

1. *Dell PowerEdge R420 (Sandy Bridge)*

2 8-core Intel Xeon E5-2470 CPUs@2.3 GHz 20 MB L3 cache
    32 logical cores (hyperthreading enabled)
64 GB 1600 MHz DDR3 RAM
4 SATA 3 Gbps 7200 RPM 1 TB 3.5” drives
    Western Digital WD1003FBYX-1 Rev 1V02; Software RAID0
Storage controllers:
    a. PERC H310
    b. PERC H710 with 512 MB cache
    c. Embedded SATA
1 active 1000baseT Ethernet NIC

2. *Dell PowerEdge R720xd (Sandy Bridge)*

2 8-core Intel Xeon E5-2660 CPUs@2.2 GHz 20 MB L3 cache
    32 logical cores (hyperthreading enabled)
64 GB 1600 MHz DDR3 RAM
Drive configurations:
    2 SSD 200 GB 2.5” drives for OS
        Samsung MZ-5EA2000-0D3 Rev 7D3Q; Hardware RAID0
    a. 6 SATA 6 Gbps 7200 RPM 3 TB 3.5” drives
        Hitachi HUA723030ALA640 Rev A6N0; Software RAID0
    b. 12 SATA 3 Gbps 7200 RPM 1 TB 3.5” drives
        Western Digital WD1002FBYS-18A680 Rev 0C06; Software RAID0
PERC H710 Storage Controller 512 MB cache
1 active 1000baseT Ethernet NIC

Evaluation System Specifications

3. Oracle Sun Fire X4170 M3 (Sandy Bridge)

2 8-core Intel Xeon E5-2690 CPUs@2.9 GHz 20 MB L3 cache
    32 logical cores (hyperthreading enabled)
64 GB 1600 MHz DDR3 RAM
1 SATA 6 Gbps SSD 100 GB 2.5” drive for OS
    Intel SA2BZ10 Rev 362
7 SAS 6 Gbps 10000 RPM 300 GB 2.5” drives
    Hitachi H106030SDSUN300G Rev A2B0; Software RAID0
Sun 6 Gbps SAS Storage Controller (LSI MR9261-8i)
1 active 1000baseT Ethernet NIC

4. HP ProLiant DL160 (Sandy Bridge)

2 6-core Intel Xeon E5-2630 CPUs@2.3 GHz 15 MB L3 cache
    24 logical cores (hyperthreading enabled)
64 GB 1333 MHz DDR3 RAM
Drive configurations:
    a. 4 SATA 6 Gbps 7200 RPM 2 TB 3.5” drives
        Hitachi HUA723020ALA640 Rev HPG3; Software RAID0
    b. 1 SATA 6 Gbps SSD 512 GB 3.5” drive for OS
        Crucial CT512M4SSD2 Rev 000F
       3 SATA 6 Gbps 7200 RPM 2 TB 3.5” drives
        Hitachi HUA723020ALA640 Rev HPG3; Software RAID0
Embedded SATA Controller (AHCI)
1 active 1000baseT Ethernet NIC

Evaluation System Specifications

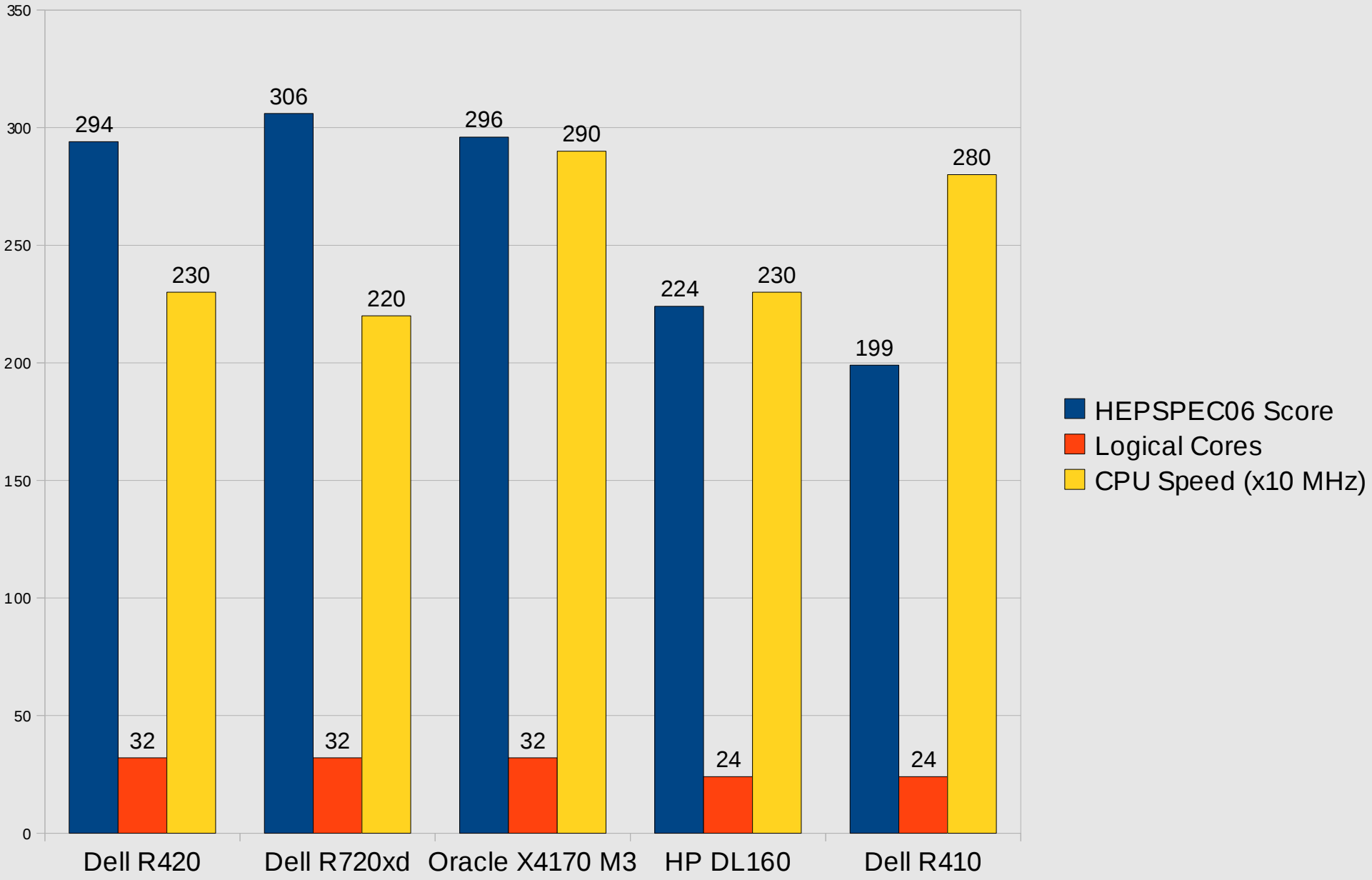
5. Dell PowerEdge R410 (Westmere)

From previous year's procurement: for comparative purposes only
2 6-core Intel Xeon X5660 CPUs@2.8 GHz 12 MB L3 cache
    24 logical cores (hyperthreading enabled)
48 GB 1333 MHz DDR3 RAM
4 3 Gbps 7200 RPM 1 TB SATA drives
    Western Digital WD1003FBYX-1 Rev 1V02; Software RAID0
SAS 6/iR disk controller
1 active 1000baseT Ethernet NIC

Note: all tests completed using 64-bit Scientific Linux 5

Systems #1-4: kernel 2.6.18-274.18.1
System #5: kernel 2.6.18-194.11.4

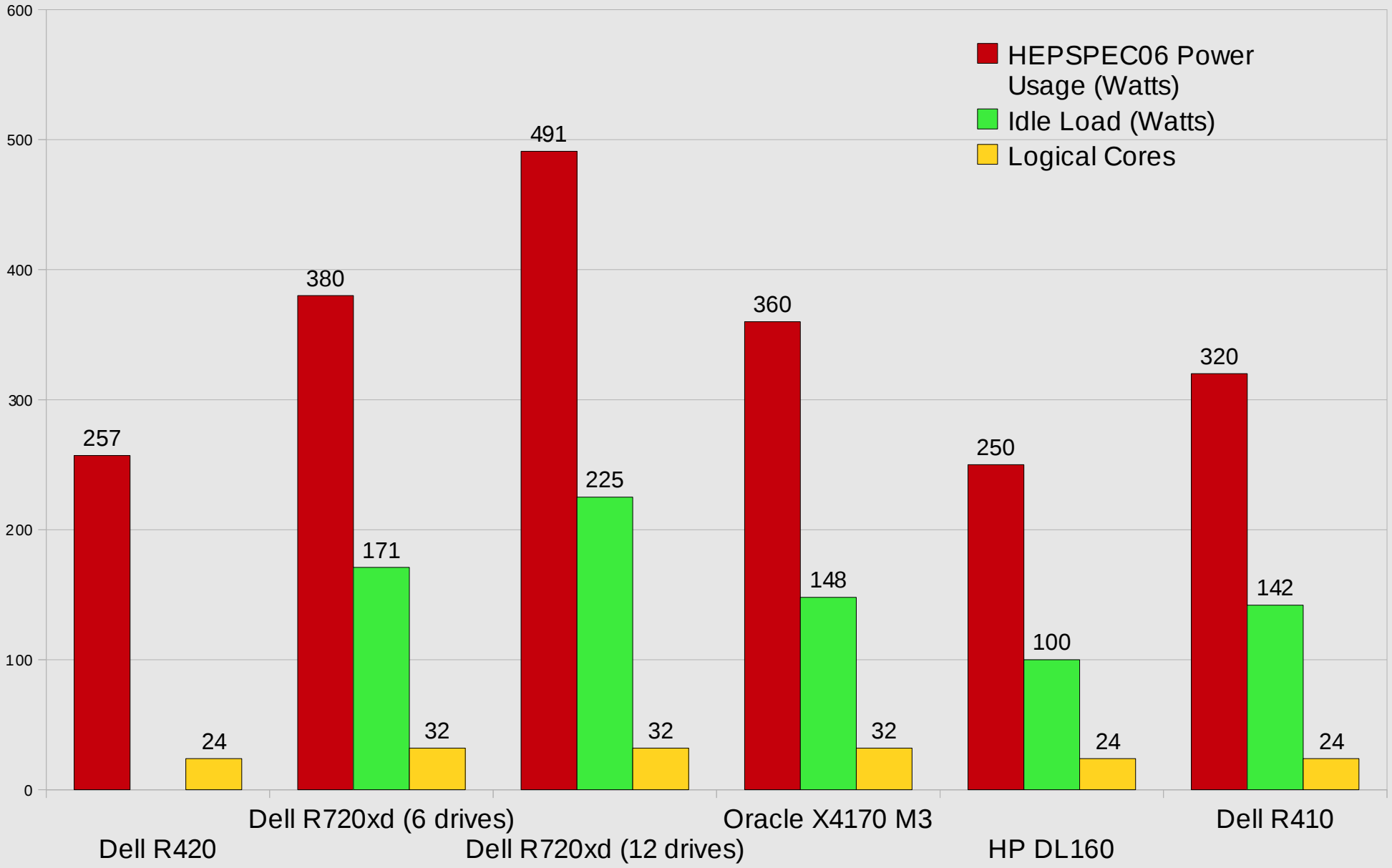
HEPSPEC06 Benchmark



HEPSPEC06 Benchmark

- Standard benchmark adopted by the HEP community for measuring CPU performance, approximating HEP software workloads
  - Based on a subset of the SPEC CPU2006 benchmark
  - Have encountered some cases where performance didn't correctly track ATLAS software performance
- Moving from 24-core based Westmere hosts to “equivalent” 32-core Sandy Bridge based machines increases HEPSPEC06 performance by ~30%
- E5-26XX series CPUs have an additional quickpath channel
- Unclear why the E5-2690 based Oracle system didn't significantly outperform the others
  - Vendor suggested a BIOS upgrade: this actually reduced HEPSPEC06 performance slightly

Power Utilization



NOTES

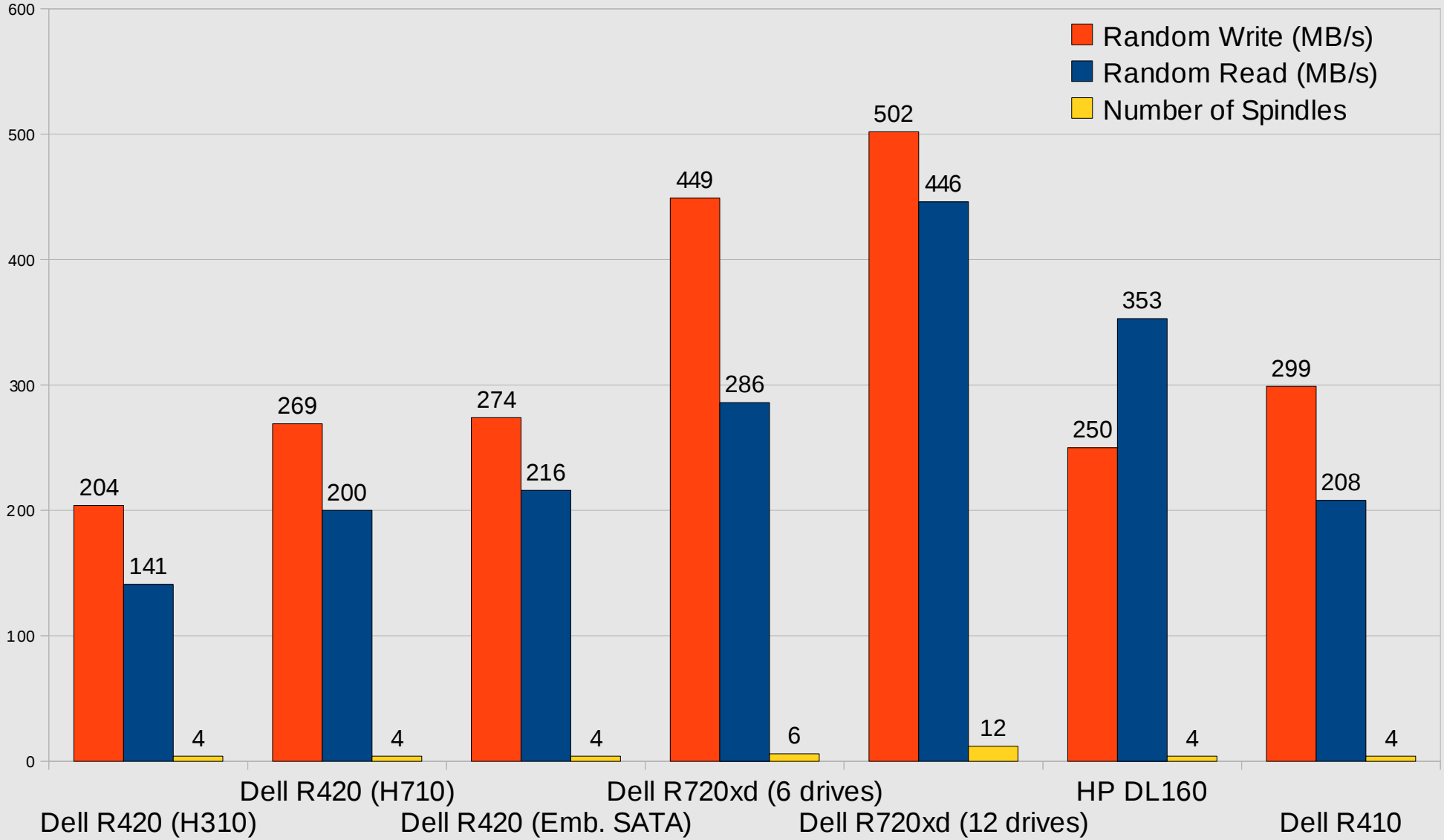
1. The power measurement listed for the Dell R420 is from a E5-2440 based host, rather than the E5-2470. The E5-2470 based R420 system may consume a bit more power.

Power Utilization

- Measurements made by powering servers via a portable power meter (operating at 110V)
- Power footprint largely unchanged by Sandy Bridge
  - Power utilization of servers based on mid-range Sandy Bridge CPUs is similar or lower than that of servers based on mid-range Westmere CPUs
- The power measurement listed for the Dell R420 is from a E5-2440 based host, rather than the E5-2470 tested a few months ago. The E5-2470 based system may consume a bit more power.
- The R720xd consumes considerably more power with 12 drives than with 6
  - Besides needing to power more drives, restricted frontal air flow likely increased fan speed, and therefore power consumption

SAS/SATA Drive RAID0 Arrays - bonnie++ I/O Benchmark

Multi-threaded (24) aggregate, buffering enabled, bonnie++ without options

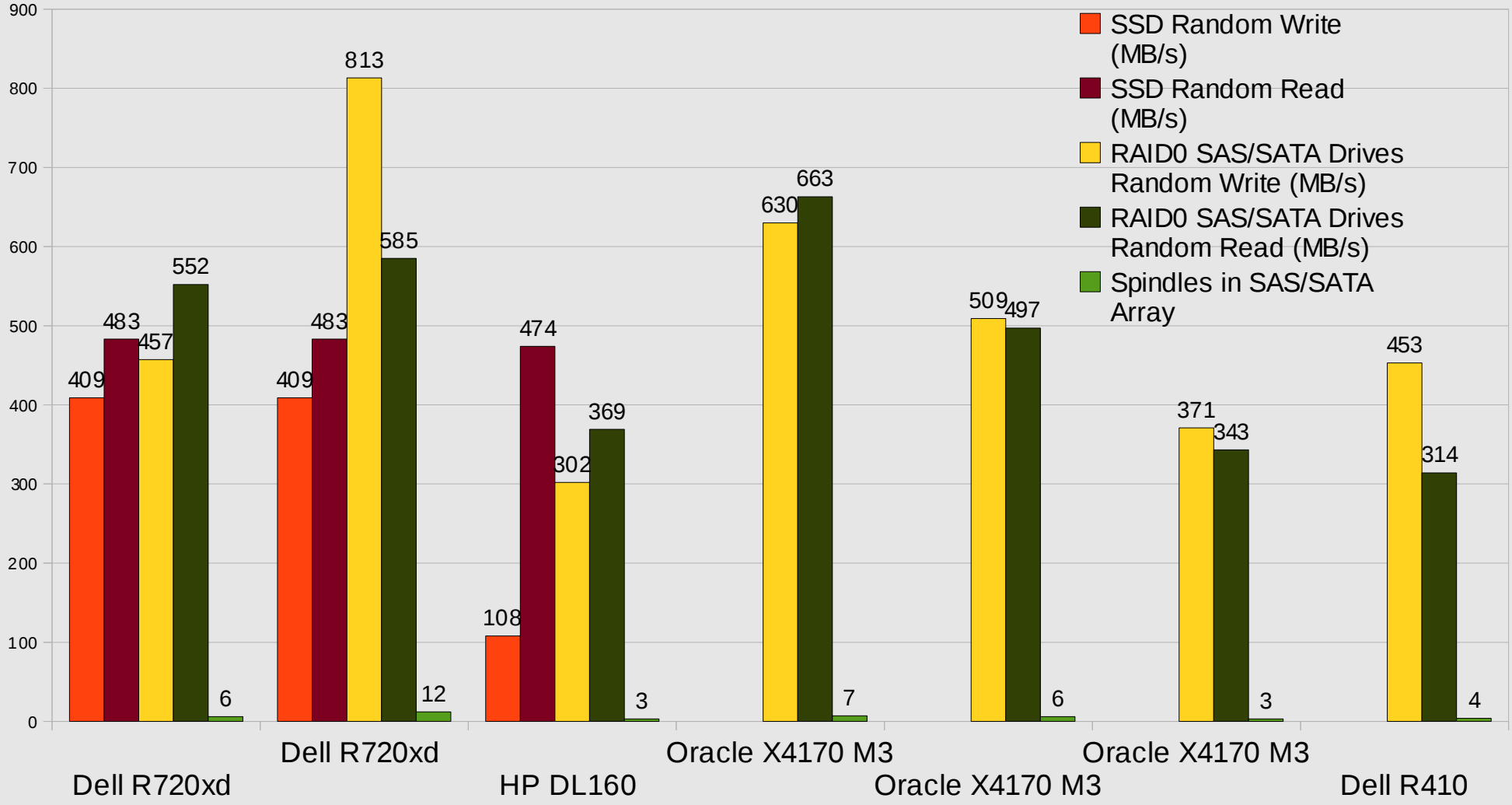


NOTES

1. Results for the Oracle X4170 M3 server are unavailable. The system's drives were not large enough to support this test.

bonnie++ I/O Benchmark – Reduced Filesize For SSD Testing

Multi-threaded (24) aggregate, buffering disabled, bonnie++ -b -r 2560 -s 5120

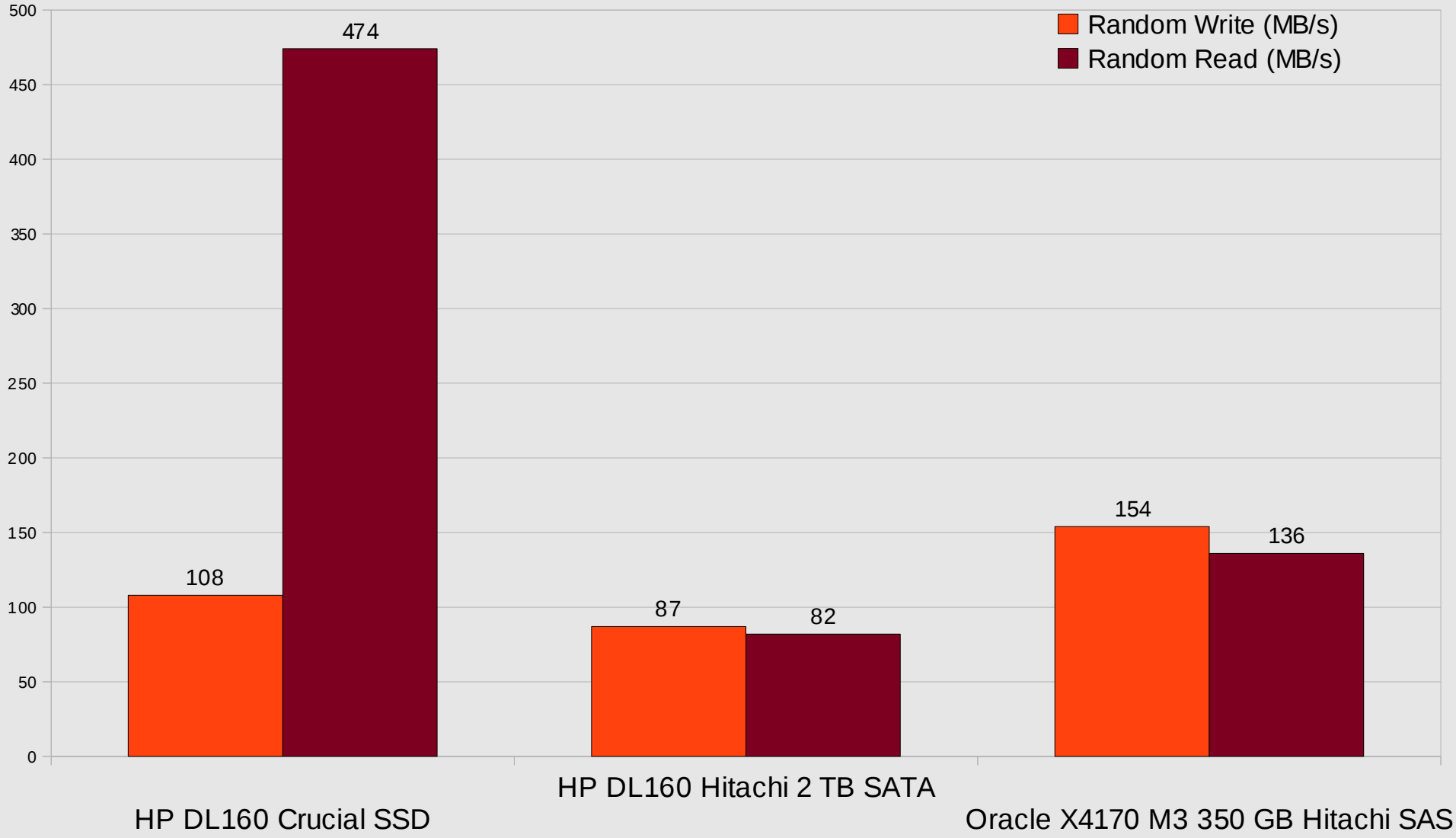


NOTES

1. *Reduced filesize necessary to support smaller SSD capacities.*
2. *Chosen filesize value x number of threads roughly equal to twice total RAM (24 x 5120 MB = 123 GB).*
3. *SSD data unavailable for Oracle host: 100 GB SSD too small for testing.*
4. *Dell R720xd system SSD results are for a hardware RAID0 volume consisting of two SSDs.*
5. *Results for the R420 unavailable. This host was tested a few months ago before SSDs were being evaluated*

bonnie++ I/O Benchmark – Single SSD vs Single SATA, SAS

Multi-threaded (24) aggregate, buffering disabled, bonnie++ -b -r 2560 -s 5120

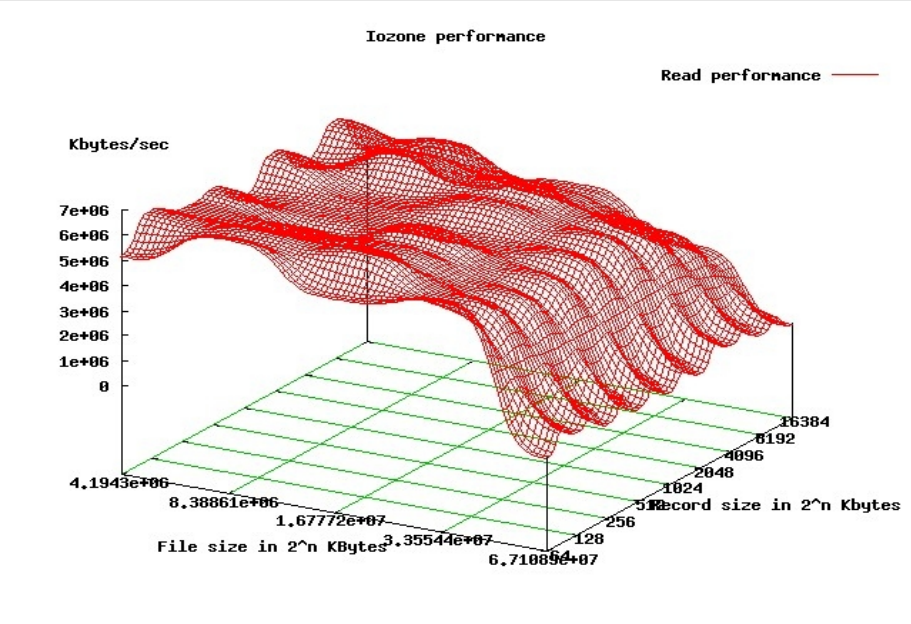


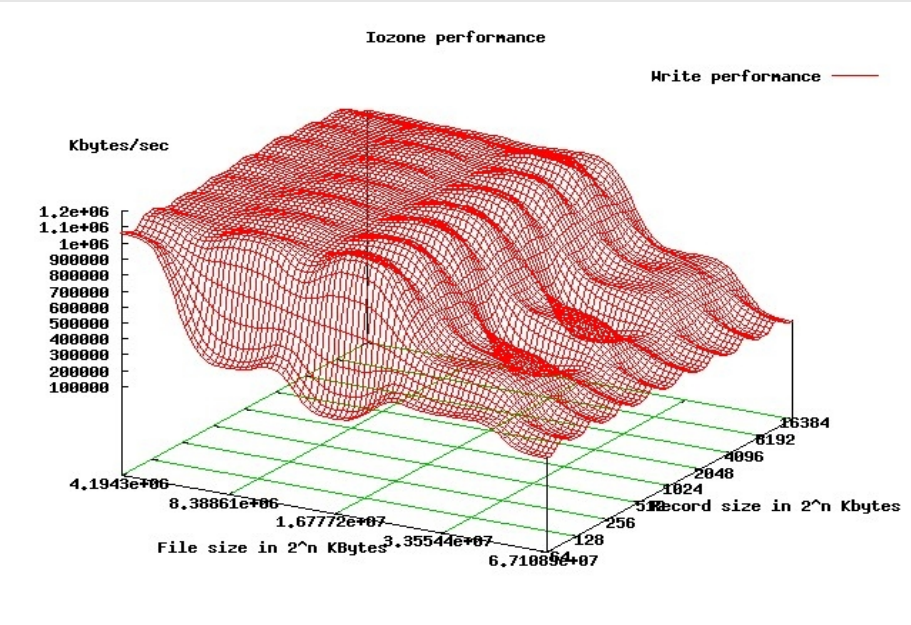
NOTES

1. SAS values were likely helped by the X4170's superior storage controller.

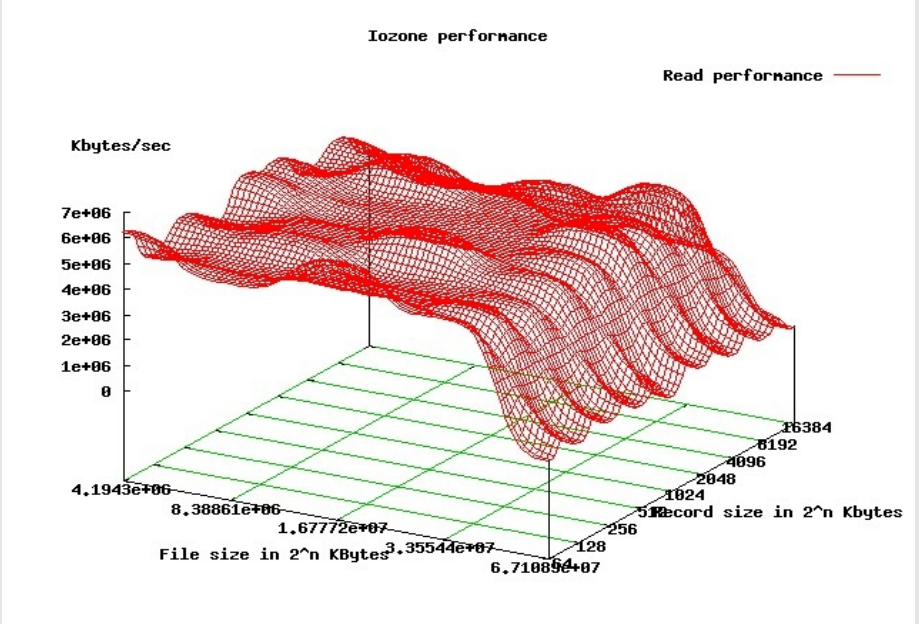
Iozone Benchmak - Single SSD Drive vs Single SAS Drive

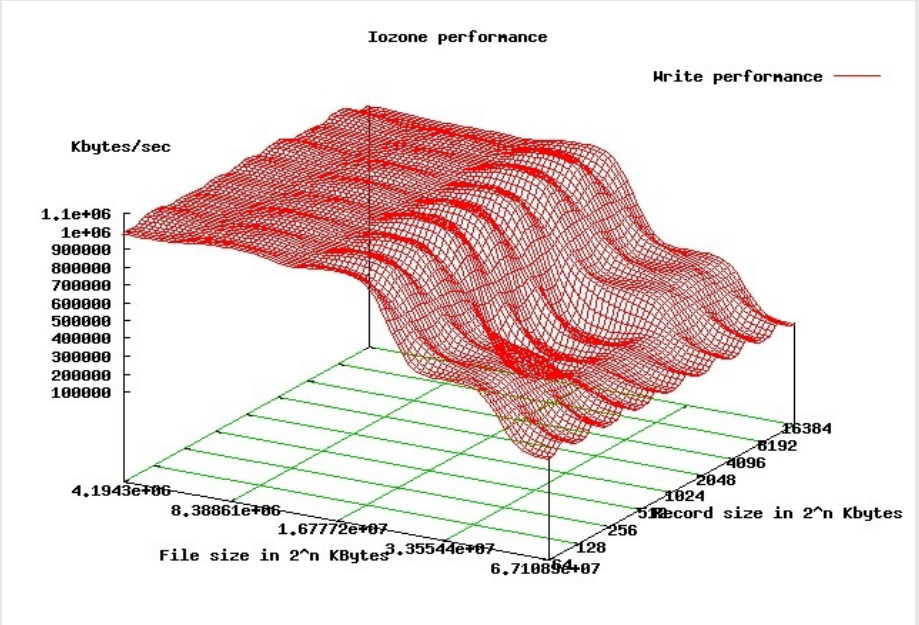
Oracle X4170 M3 Hitachi 350 GB SAS Drive





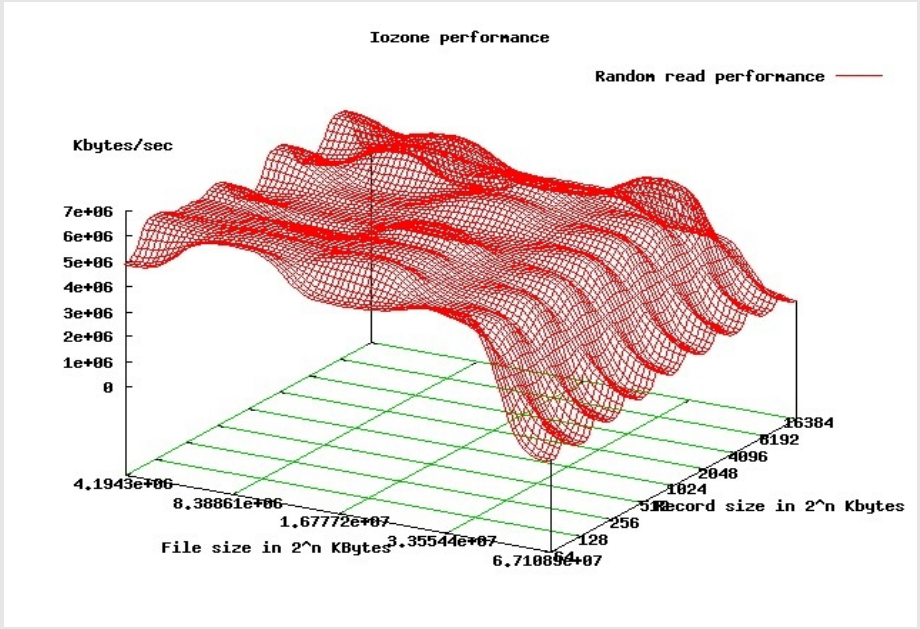
Oracle X4170 M3 Intel 100 GB SSD Drive

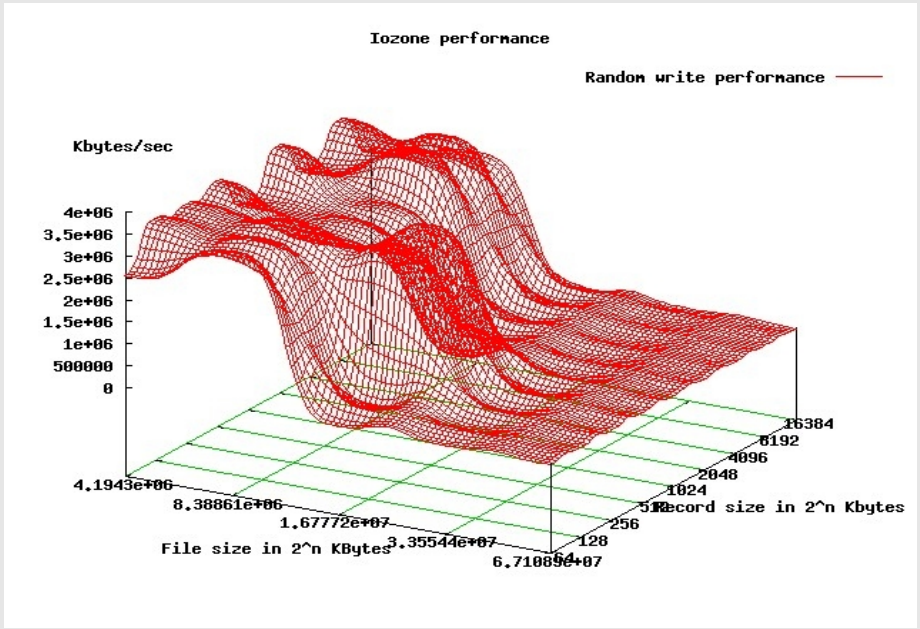




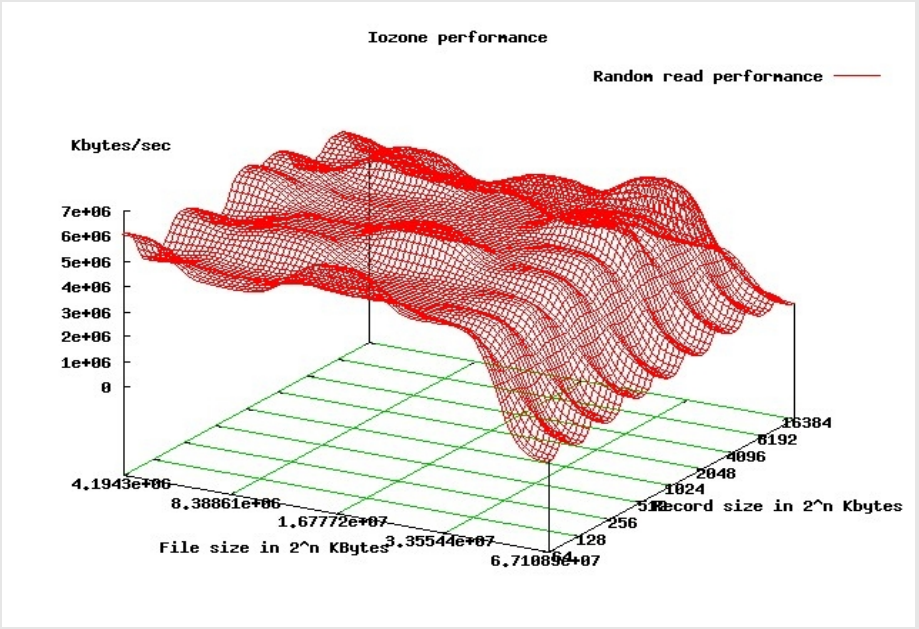
Iozone Benchmak - Single SSD Drive vs Single SAS Drive

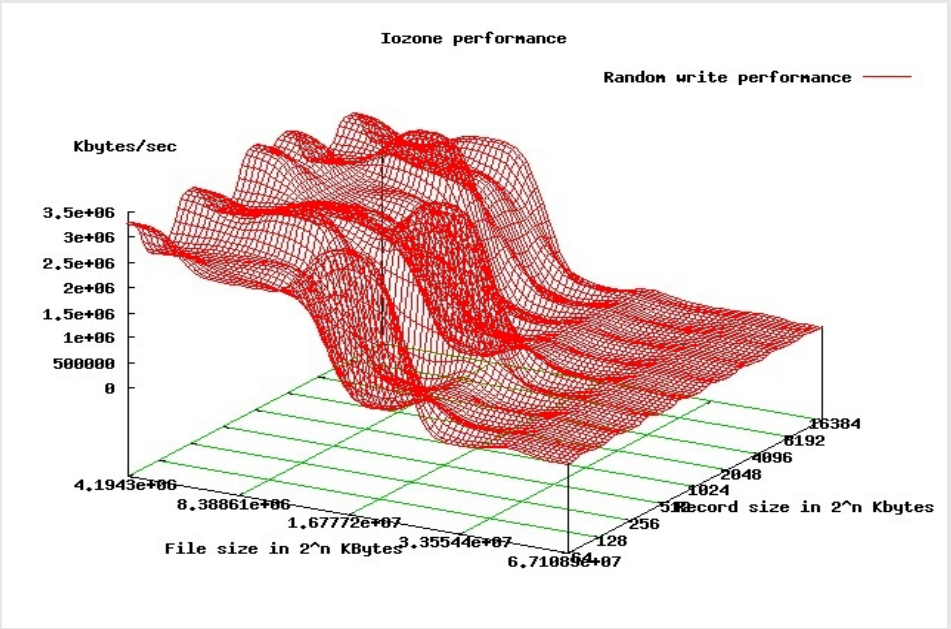
Oracle X4170 M3 Hitachi 350 GB SAS Drive





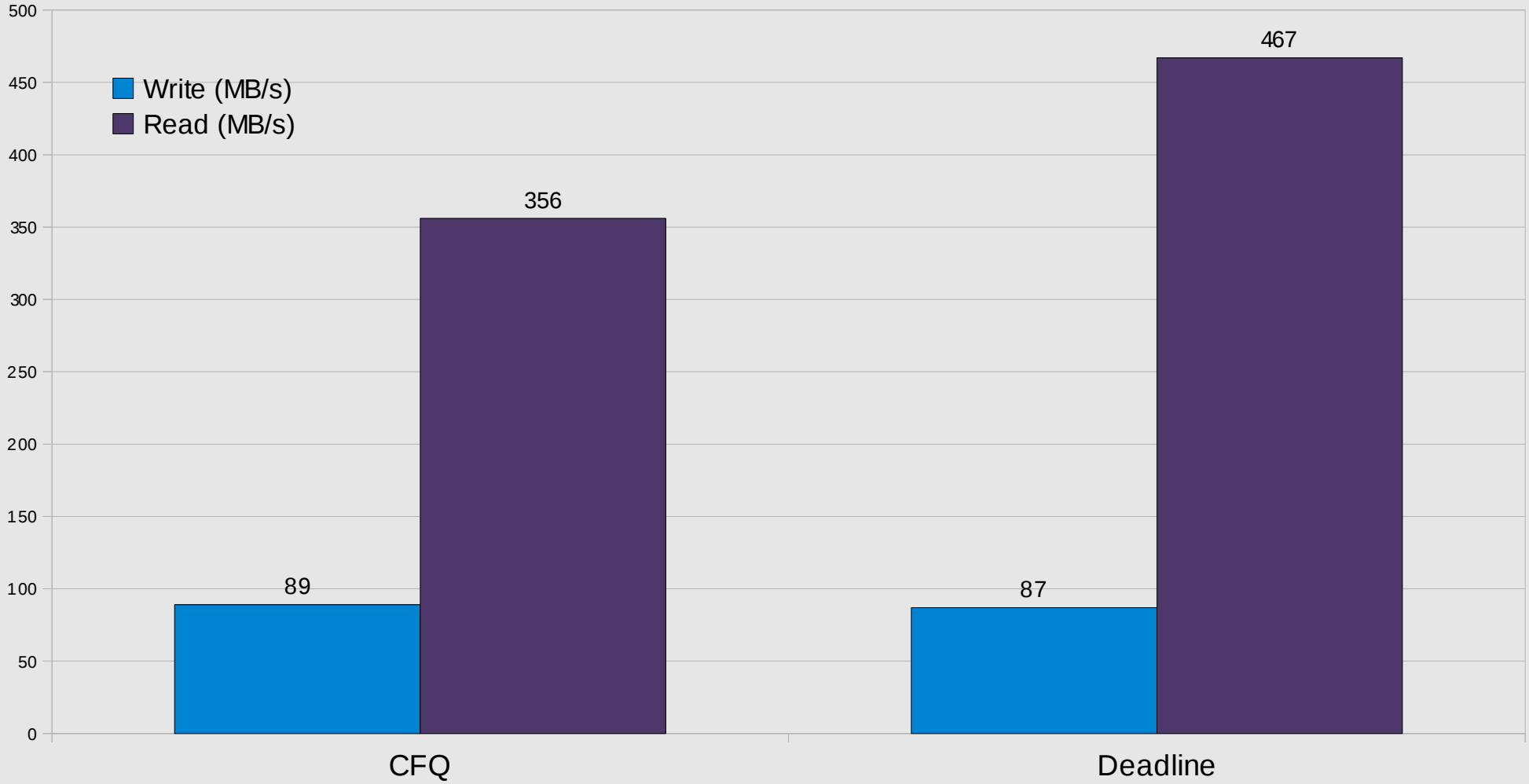
Oracle X4170 M3 Intel 100 GB SSD Drive





Effect Of Linux Kernel I/O Scheduler On SSD Performance

Multi-threaded (24) aggregate, buffering disabled, synchronized, bonnie++ -b -y -r 2560 -s 5120



NOTES

1. All tests run on the HP DL160 server with Crucial SSD.

2. CFQ – Completely Fair Queuing kernel block I/O scheduler. Attempts to fairly balance I/O access amongst processes. Implicitly implements anticipatory I/O scheduling which is optimized for standard drives.

3. The deadline kernel block I/O scheduler attempts to guarantee start service times for I/O requests. It is not anticipatory, and therefore a good scheduler for use with SSDs.

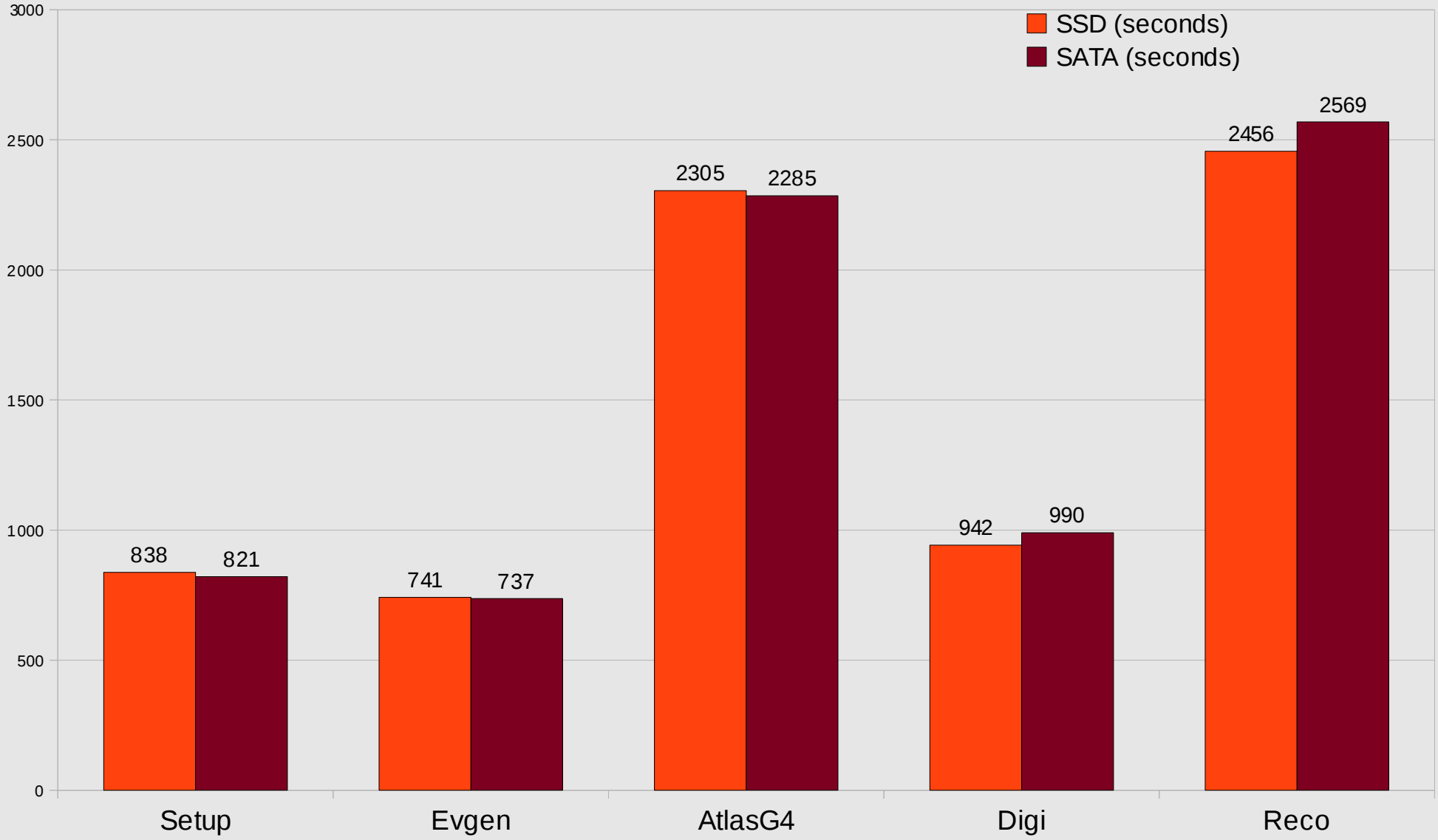
4. It's possible that other OS software/parameter modifications not explored during testing may improve SSD performance.

bonnie++ and Iozone Benchmarks

- Primarily interested in random I/O performance
  - Multiple jobs per system accessing the local drives simultaneously generates a random workload
- As expected, more spindles in a RAID0 array leads to better random I/O performance
  - Doubling the spindle count doesn't mean doubling performance, however
- SSD best-case performance for random I/O appears to be similar to 4 drives in a software RAID0 array
  - Unclear why the Crucial SSD performed poorly for random reads (~100 MB/s)
  - Oracle (Intel) SSD overall performance also somewhat disappointing
- SSD handles small I/O record sizes better than a SAS drive
- Deadline scheduler improved SSD read performance

Effect Of Moving AFS Cache To An SSD

Run times for the ATLAS software (24 threads) in AFS for both configurations

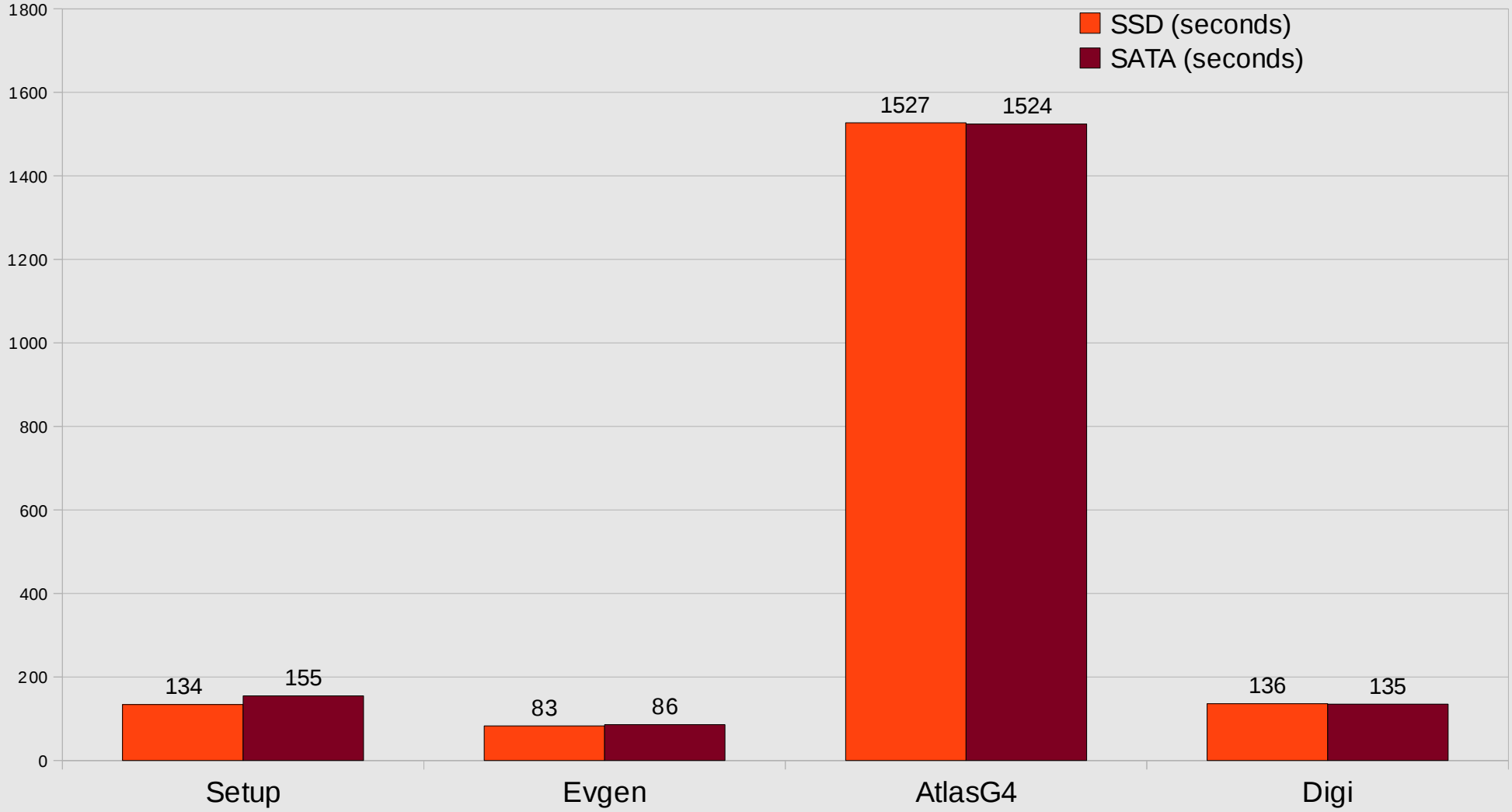


NOTES

1. *Thanks to Shuwei Ye <yesw@bnl.gov> for running the ATLAS software in these tests.*
2. *All tests run on the HP DL160 systems.*

Effect Of Moving CVMFS Cache To An SSD

Run times for the ATLAS software (24 threads) installed in CVMFS for both configurations



NOTES

1. *Thanks to Shuwei Ye <yesw@bnl.gov> for running the ATLAS software in these tests.*
2. *All tests run on the HP DL160 host.*
3. *Reco results not listed, as several of these processes crashed for unknown reasons during execution.*

Conclusions

1. We've completed our tests and are awaiting input from PHENIX and STAR. Their final performance measurements will provide guidance on upcoming Farm purchases in FY13.

2. Local I/O performance is correlated with the number of spindles. Therefore, a 2-U system filled with 3.5” drives provides much better I/O and more storage than 1-U, though one loses processing density per rack. The 1-U system filled with 3.5” drives doubles processing density per rack at the expense of I/O and (to some extent) storage capacity.

3. Historically, a 2-U server configured with maximum storage has cost ~50% more than a 1-U server configured in a similar manner, all other parameters (network, cpu, memory, etc) being equal.

4. Recommend purchasing systems based on dual 8-core (16-logical core) Sandy Bridge CPUs (32 logical cores total).

5. Recommend continuing to purchase multi-spindle (4+) SAS/SATA systems without SSDs. In a software RAID0 configuration, similar or better performance can be achieved with SAS/SATA drives along with much greater storage capacity, at less cost. Moving the AFS and CVMFS caches to SSD storage doesn't appear to be beneficial at this time.

Recommended Hardware Choices

	1-U	2-U	Comments
RHIC	4x3 TB	12x2 TB	3.5” SATA
ATLAS	4x1 TB 8x500 GB	xxx xxx	3.5” SATA 2.5” SATA