NVM Express (NVMe):
An Overview and Performance Study

Chris Hollowell <hollowec@bnl.gov>
RHIC/ATLAS Computing Facility
Brookhaven National Laboratory

What is NVMe?

NVMe - NVM Express - Non-Volatile Memory Express

- An industry standard for attaching SSDs (NAND flash) directly to the PCIe bus

Eliminates latency and bandwidth limitations imposed by SAS/SATA storage controllers optimized for traditional rotating media

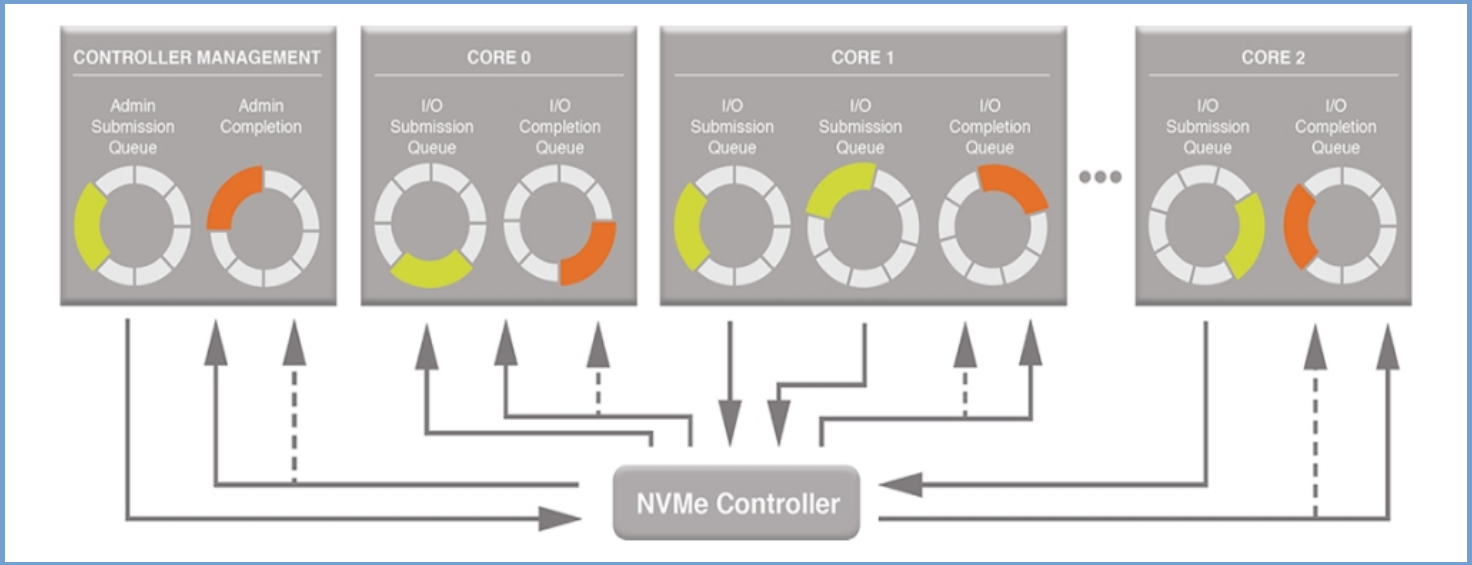
Architected for highly parallel access

- Support for up to 64k hardware I/O queues, with up to 64k commands per queue
- Excellent for parallel I/O operations in servers with ever-increasing processor core counts

Supported in the Linux kernel since 3.3

- Backported to RHEL/SL 6 (kernel 2.6.32) in the 6.5 release
- Uses “Multi-Queue Block IO Queuing” (blkmq) rather than the standard kernel block I/O schedulers (noop, cfq, deadline) to support parallel hardware queue architecture

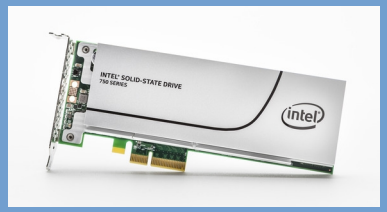
What is NVMe? (Cont.)



NVMe Command Queue Architecture (from nvmexpress.org)

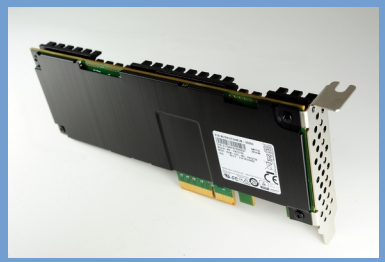
Several vendors manufacturing NVMe hardware:

Intel, Crucial, Samsung, etc.



Sizes over 3 TB available

Available as PCIe add-in cards, or a 2.5” SFF-8639/U.2 form factor with a physically SAS-like connector for drive backplanes



Cost is still fairly high:

400 GB drive ~$400+

1 TB drive >$1000

Comparison to Fusion-IO Devices

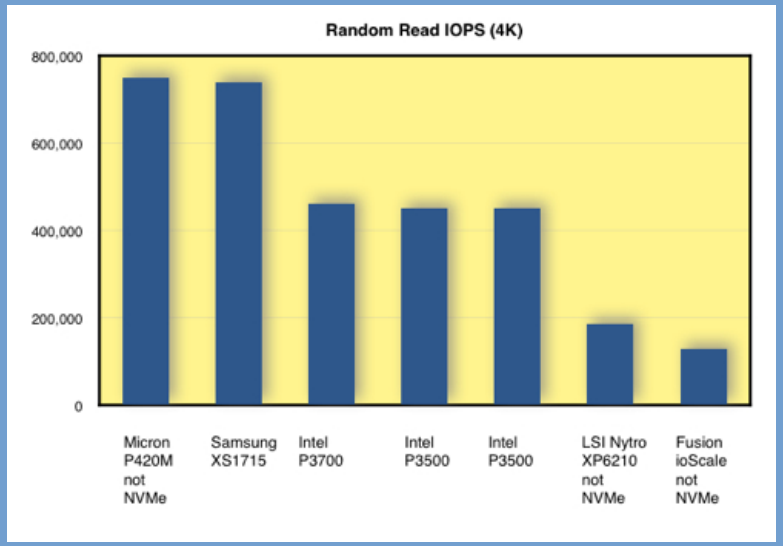
Fusion-IO has offered PCIe-connected SSD storage (ioDrive) for a number of years: how is NVMe different?

NVMe interface/protocol is an industry standard
- No need for proprietary OS drivers

Commoditization of NVMe makes the technology significantly more affordable

SFF-8639/U.2 form factor NVMe drives are in a familiar 2.5” physically SAS-like form which can be used on a backplane, and easily hotplugged
- Helps to reduce downtime when replacing failed devices

Performance of NVMe drives can be better than traditional Fusion-IO devices



NVMe and Fusion-IO Read IOPs (from “The Register”)

NVMe Evaluation

- Test Configuration
  - Dell PowerEdge R630
    - 2 Intel Xeon 2650v3 2.3 GHz CPUs (32 logical cores total)
    - PERC H730 (1 GB cache) storage controller
    - 64 GB (8x8 GB) 2133 MHz DDR4 DIMMs
    - 2 300 GB Dell 400-AEEH 15K RPM 6 Gbps SAS 2.5” drives
    - 2 400 GB Sasmsung/Dell MZWEI400HAGM NVMe 2.5” drives
      - SFF-8639/U.2 form factor – front loading
    - SSDs:
      - Samsung MZ-7PD512 512 GB 6 Gbps SATA 2.5” drive
      - Crucial CT1024M550SSD1 1 TB 6 Gbps SATA 2.5” drive
    - Most tests performed with EXT4
    - Scientific Linux 6
      - Kernel 2.6.32-504.3.3.el6

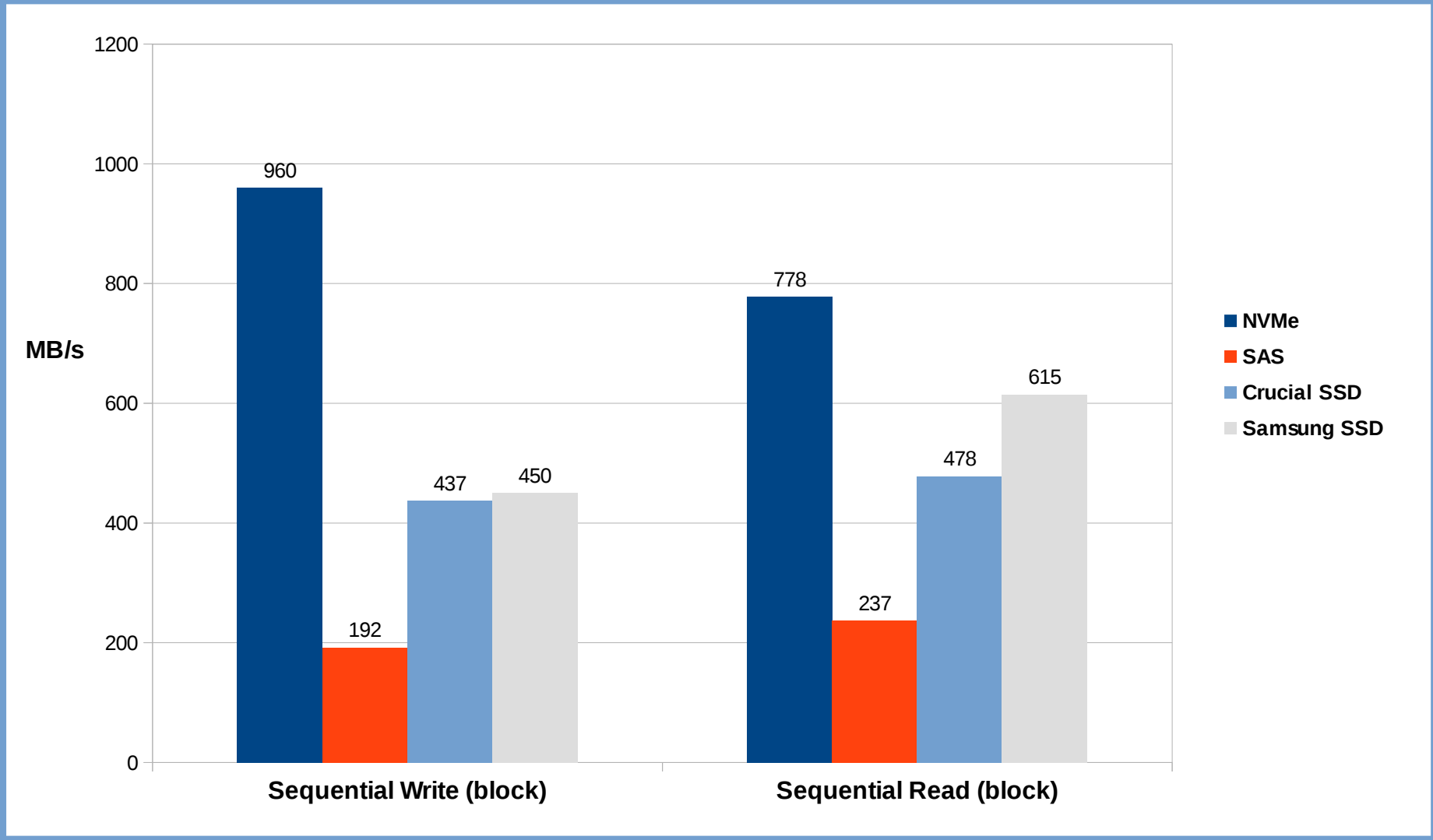
- Benchmarks
  - CFQ I/O scheduler used with SAS
  - Deadline I/O scheduler used with SSDs
  - Blkmq scheduling used with NVMe
    - No other scheduling options available

NVMe Evaluation (Cont.)

Benchmarks (Cont.)

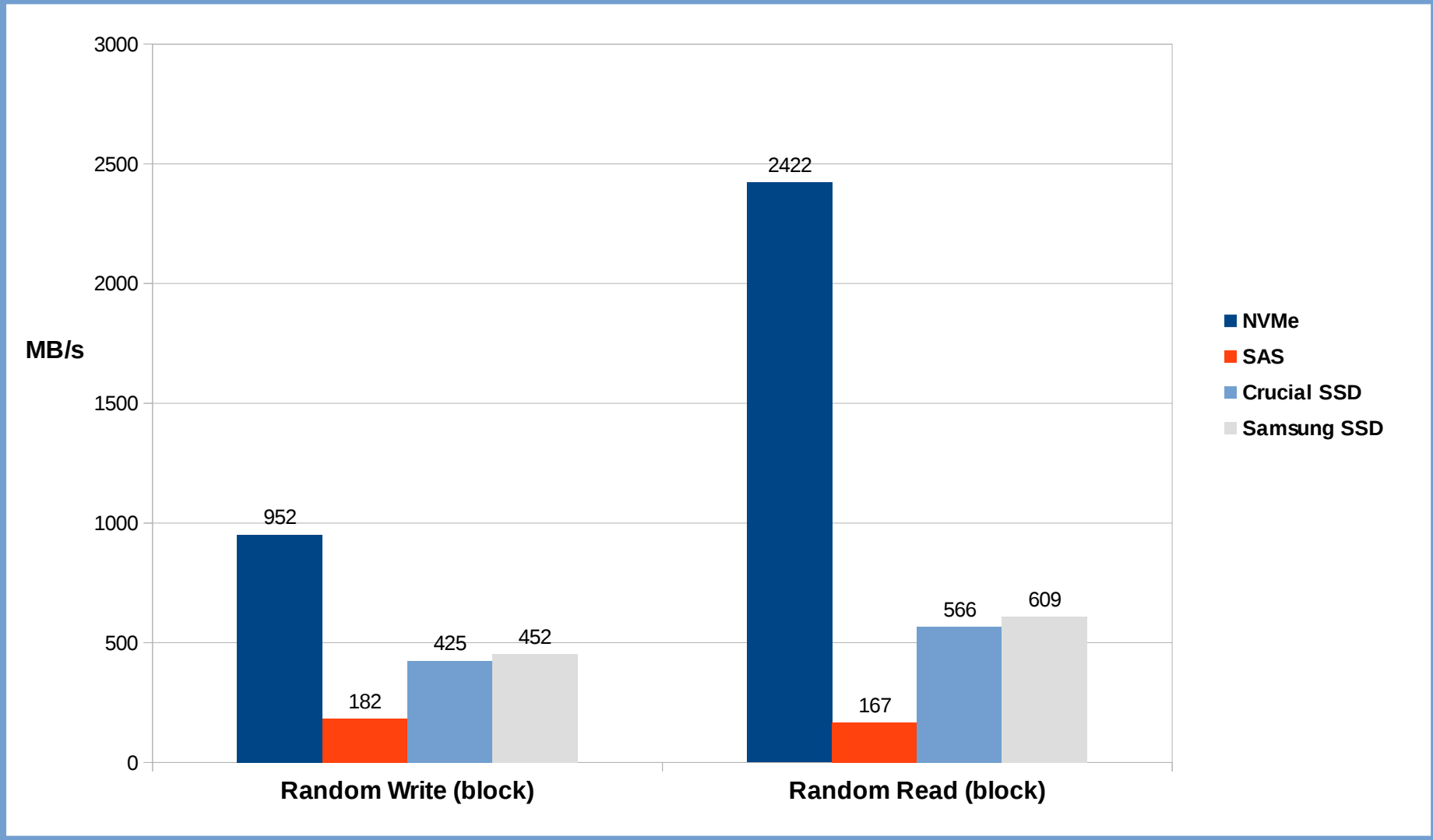
- bonnie++
  - http://www.coker.com.au/bonnie++/
  - Single and synchronized multi-process tests run
  - Primarily interested in sequential I/O tests results
    - Multiple processes performing sequential I/O creates a somewhat randomized workload
      - Likely a good simulation of the workload in our batch processing environment, in particular because our batch jobs often use a stage in/out to/from local scratch I/O model
- IOzone
  - http://www.iozone.org
  - Primarily interested in random I/O performance
- Pgbench
  - Interested in testing NVMe as the backend storage for PostgreSQL, potentially for use with dCache

Bonnie++ - Single Process



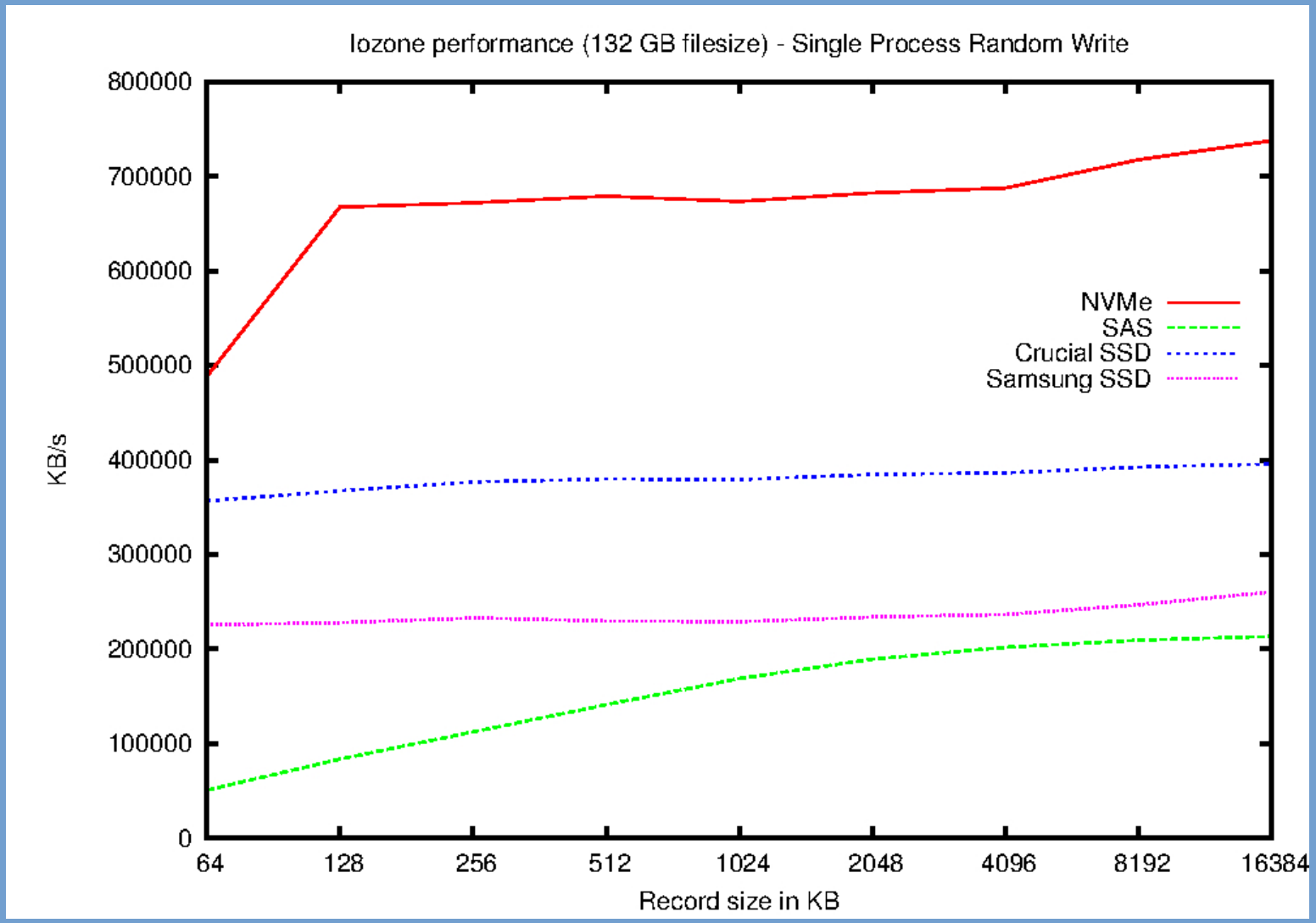
bonnie++ with default options

Bonnie++ - 32 Processes

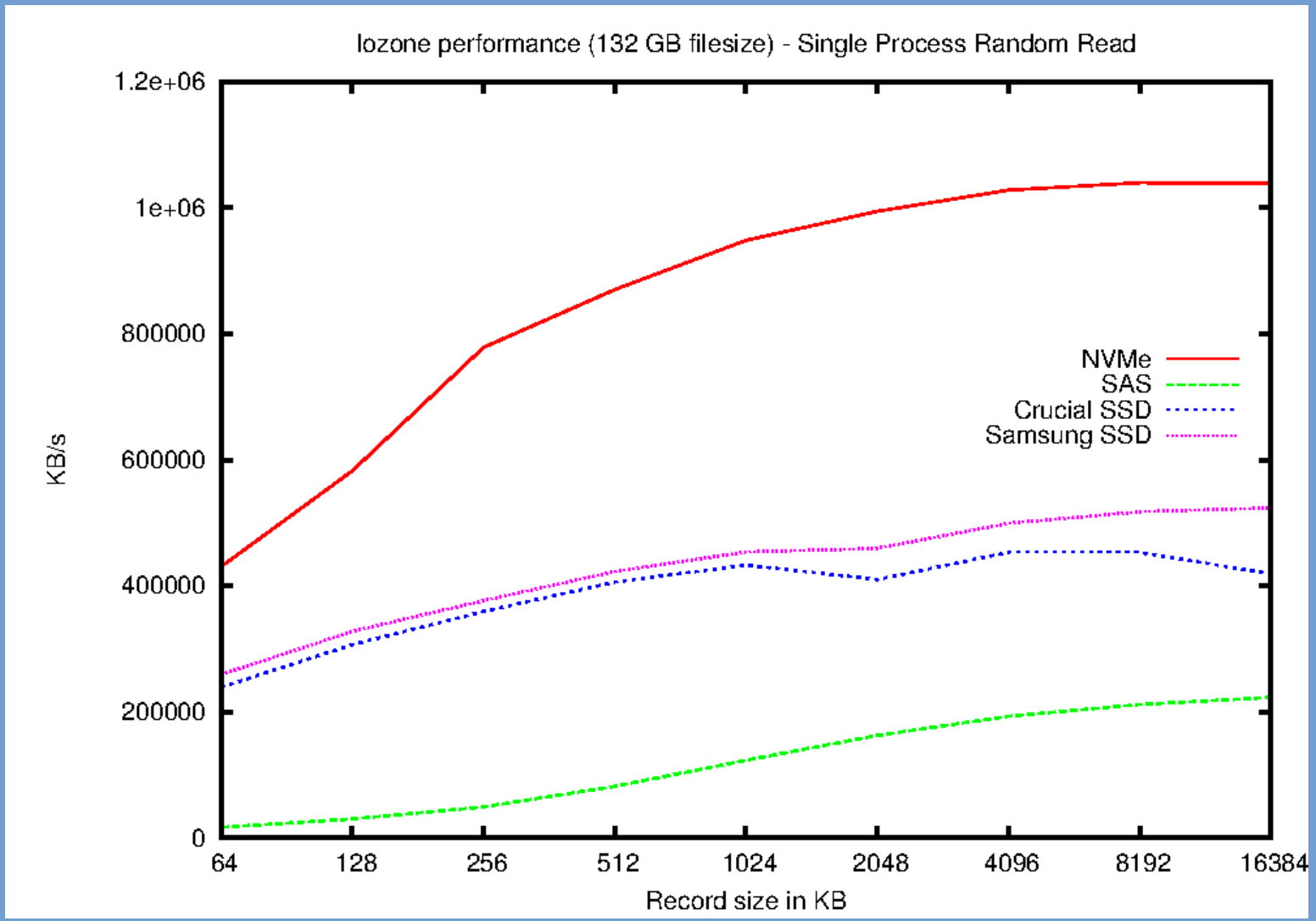


32 synchronized parallel bonnie++ processes: bonnie++ -y -r 2560 -s 8120
Parallelism creates a randomized workload

IOzone – Random Write

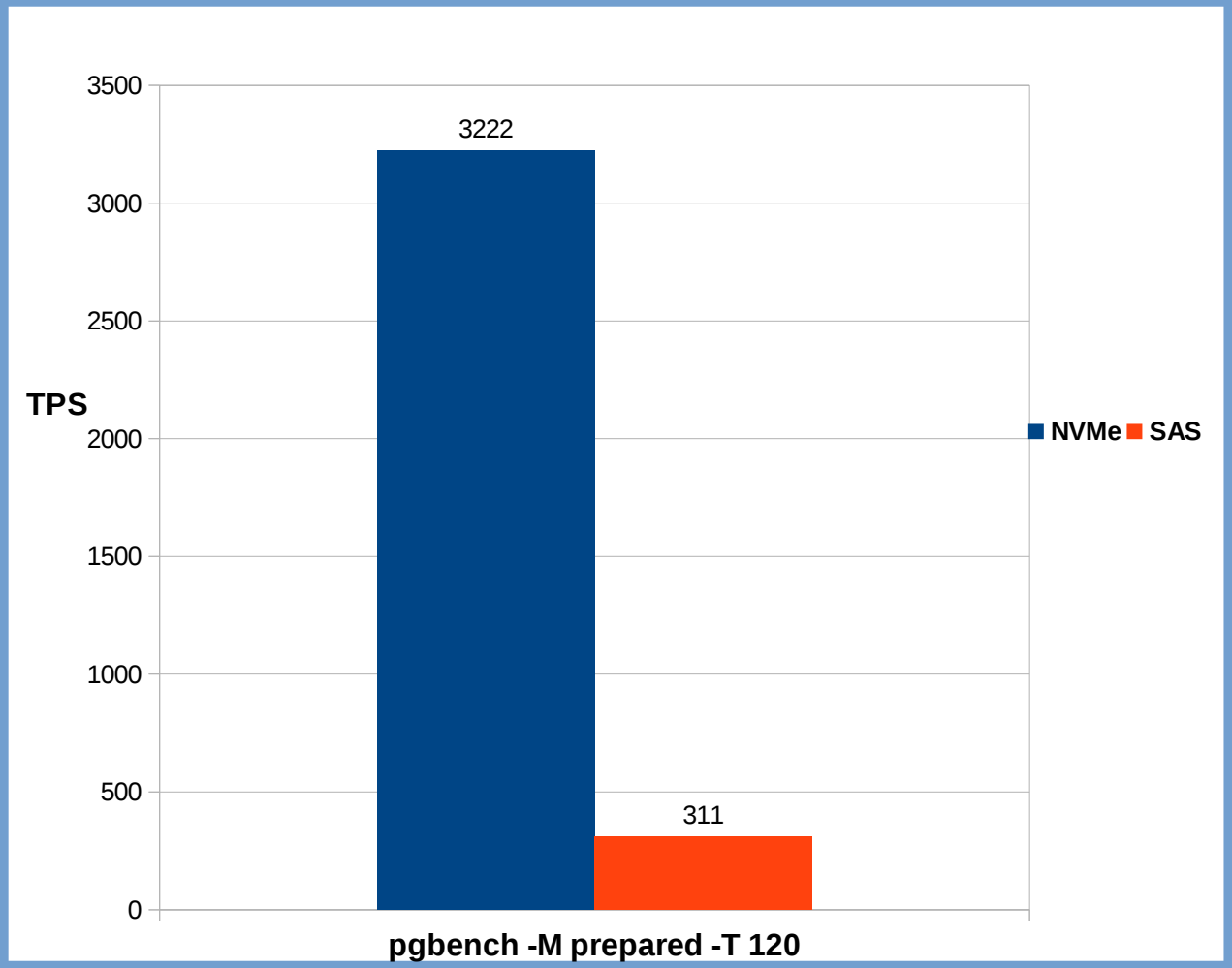


Iozone – Random Read

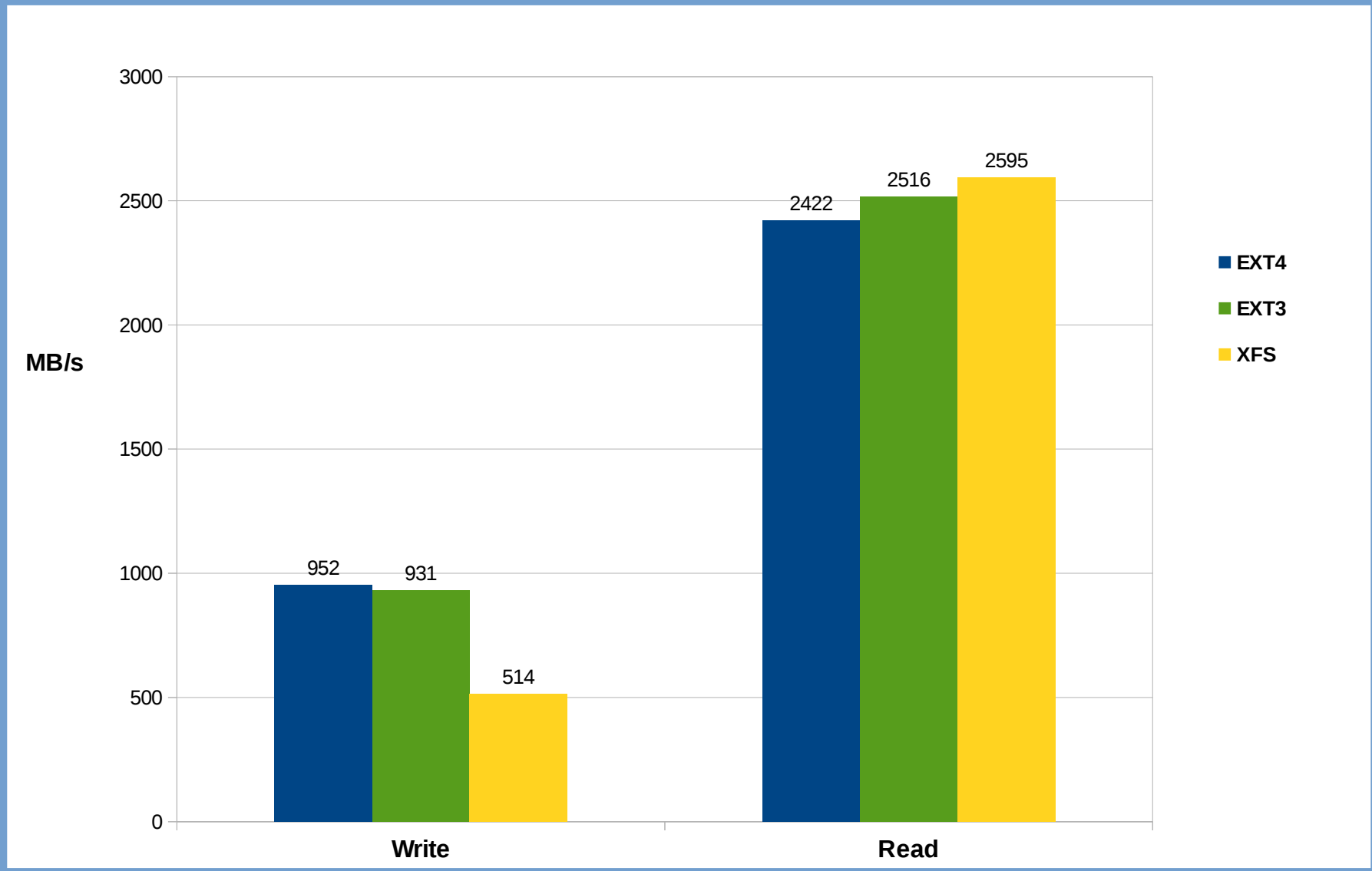


Pgbench

Postgres server benchmark - database stored on the NVMe device



NVMe Filesystem Peformance Comparison



32 synchronized parallel bonnie++ processes: bonnie++ -y -r 2560 -s 8120

Conclusions

NVMe drives eliminate latency and bandwidth limitations imposed by SAS/SATA storage controllers which are optimized for traditional rotating media

NVMe technology available today can provide impressive I/O performance

- Typically saw a 100% or more performance improvement for NVMe over traditional SSDs in our sequential and random I/O benchmarks

- It was not uncommon to see the NVMe drive perform ten times better than the SAS drive benchmarked, particularly with smaller record sizes, and with random I/O tests

High density (3 TB+) NVMe drives are available, making this a viable storage alternative to traditional drives and SSDs

- Unfortunately, still a relatively expensive option
- May change in the future
  - As this commoditized hardware becomes increasingly commonplace, expect the cost to drop

Acknowledgments

Thanks to Shawn Hoose, a student intern at RACF this summer, for his work in conducting this study

Thanks to Dell, Costin Caramarcu (RACF), Tejas Rao (RACF) Alexandr Zaytsev (RACF) for providing evaluation equipment, and additional assistance