STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

Deep Dive and Comparison of RAID Solutions for PCIe Gen5

Performance Analysis and Datapath Breakdown

Presented by Davide Villa and Sergei Platonov, XINNOR

Agenda

- **Who we are**
- **PCIe Gen5: great performance… if properly handled**
- **RAID benchmark with PCIe Gen5 SSD**
- **Conclusions**
- **Q&A**

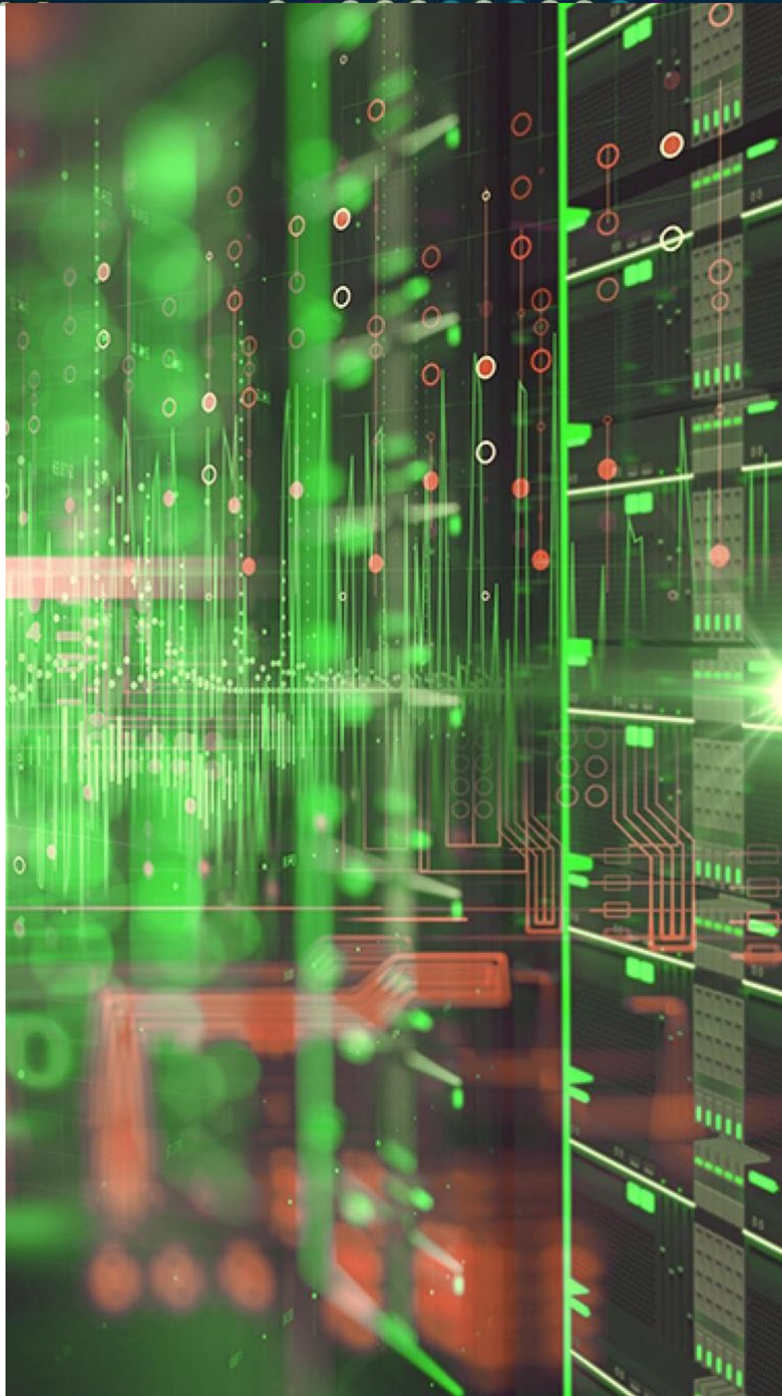
Who we are

About Xinnor

- Founded in Haifa, Israel, May 2022
- Background: 10+ years of experience with software RAID design and mathematical research
- Mission: to be the fastest RAID Engine
- Team: Around 40 people; >30 are accomplished mathematicians and industry talents from Global Storage OEMs
- >20 selling partners worldwide
- >100PB of end-customers data

Technology partners

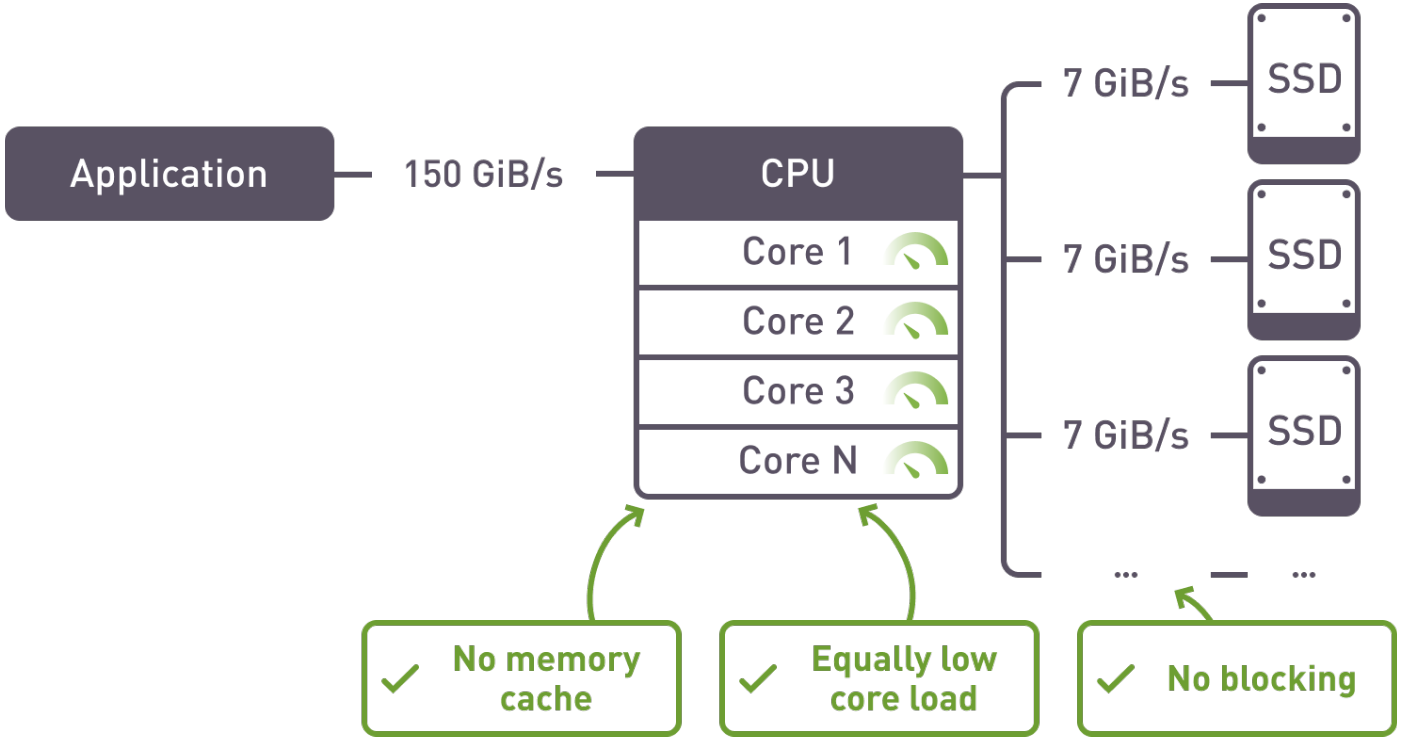
ATTO, Western Digital, ScaleFlux, KIOXIA, BeeGFS, ingrasys, TUXERA, LINBIT, NVIDIA, DapuStor



Xinnor's xiRAID unique architecture

CPU assisted RAID (AVX)

Lockless data path



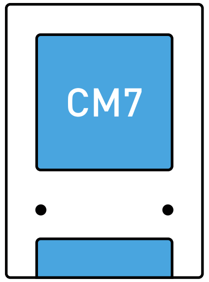
PCIe Gen5: great performance… if properly handled

PCIe Gen5: a new wave of modern Servers

- 4th Gen Intel and 3rd Gen AMD Epyc processors.
- 12-24 PCIe Gen5 drives.
- **Theoretically** capable of
  - >60 million IOPs
  - 300GB/s throughput.

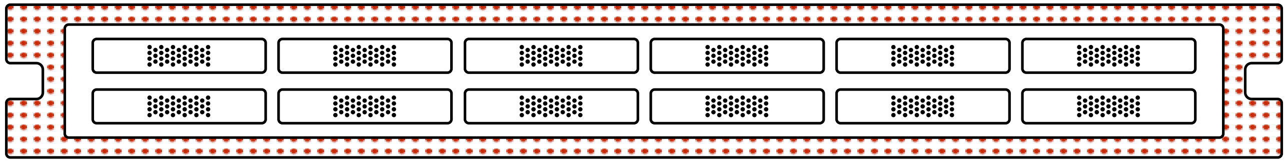
Warning: Fault Tolerance Needed!

Test Environment



12x
KIOXIA
CM7 PCIe Gen 5
NVMe SSD

FOXCONN ingrasys®



- CPU: Beaverton/Intel Xeon Gold 6430 (32Cores x2)
- Memory: 2TB (DDR5 4800 64GBx32)
- OS: Oracle Linux 8.8 (kernel 5.4.17-2136.322.6.2.el8uek.x86_64 and kernel-ml-6.5.1-1.el8)
- Benchmarking tools: fio, bdevperf

Test Environment

Single Drive Performance

Random Read:
2.7M IOps

Random Write:
0.3M IOps

Sequential Read:
14 GBps

Sequential Write:
7 GBps

Theoretical Performance with 12-Bay Chassis with RAID

Random Read:
>30M IOps

Random Write:
>3.9M IOps

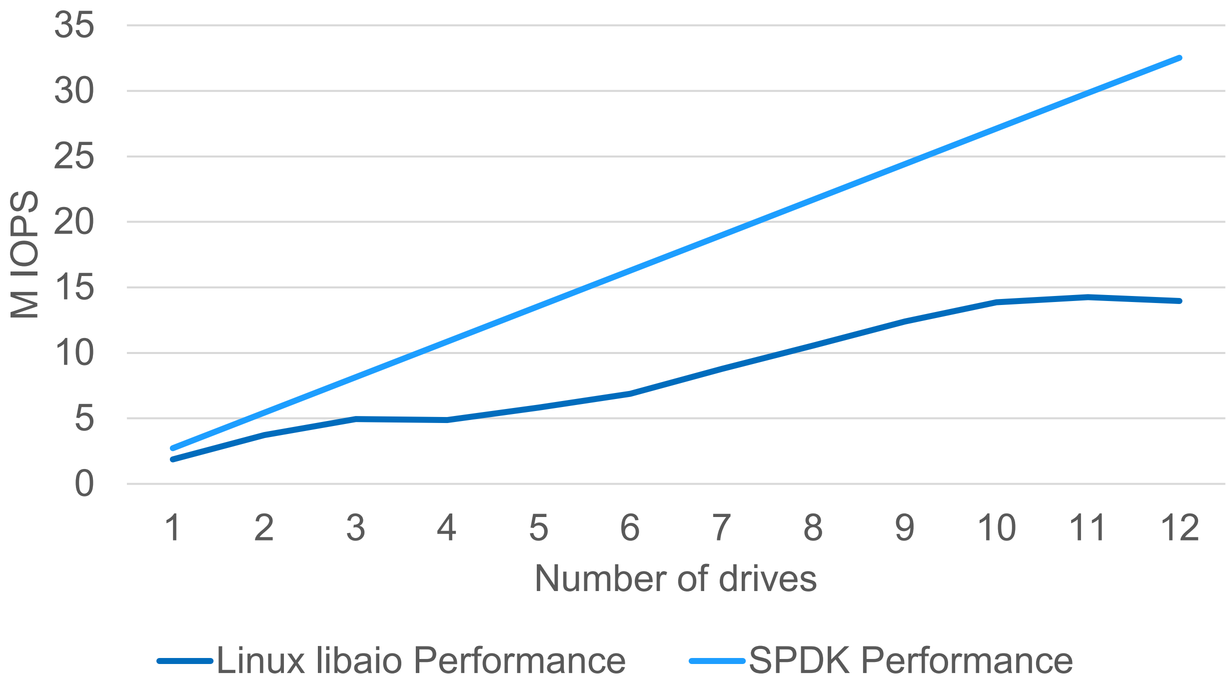
Sequential Read:
>150 GBps

Sequential Write:
>80 GBps

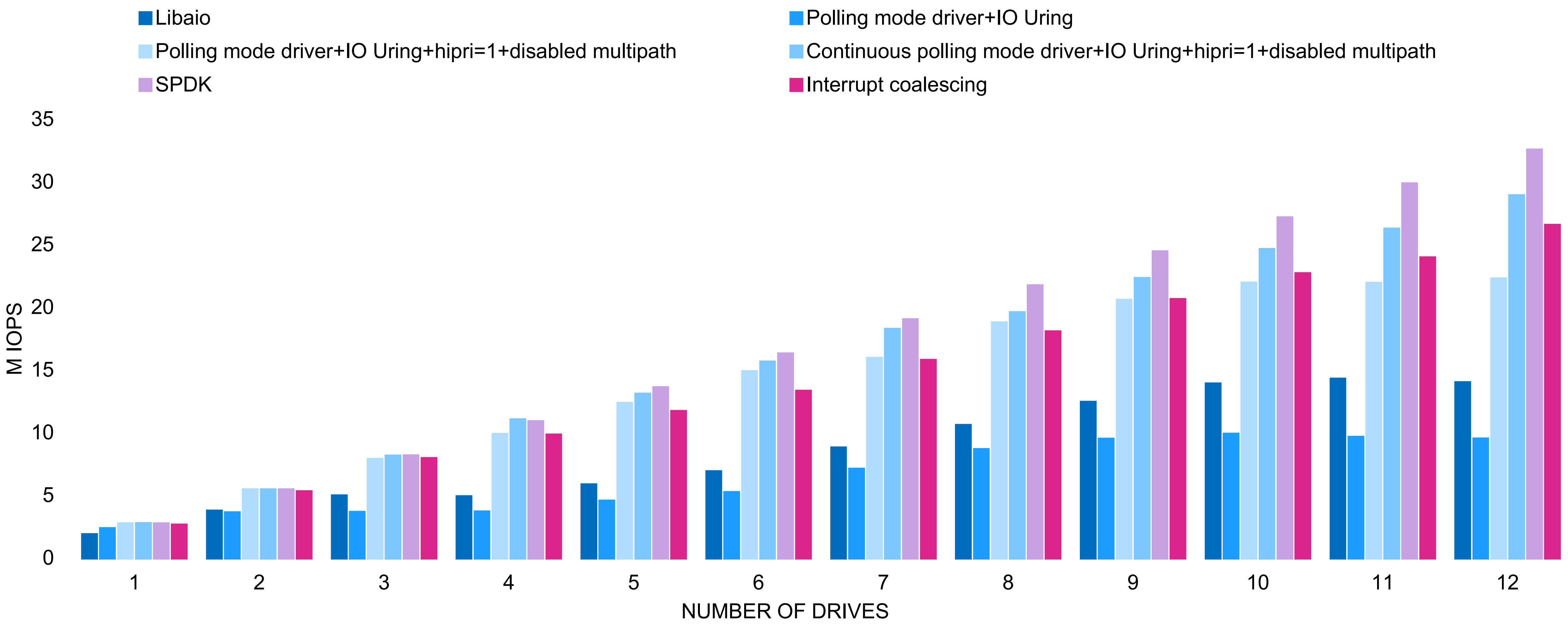
First challenge: performance scalability

4k Random Read
Single drive performance: 2.7M IOPS
Expected performance over 12 drives: > 30M IOPS

Reality:

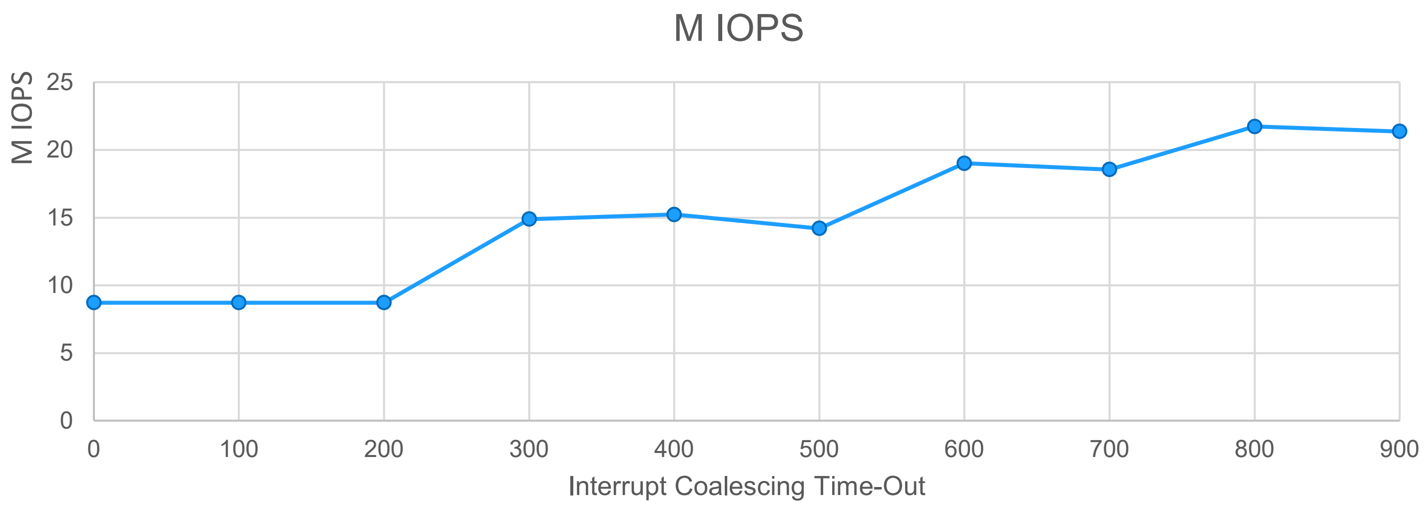


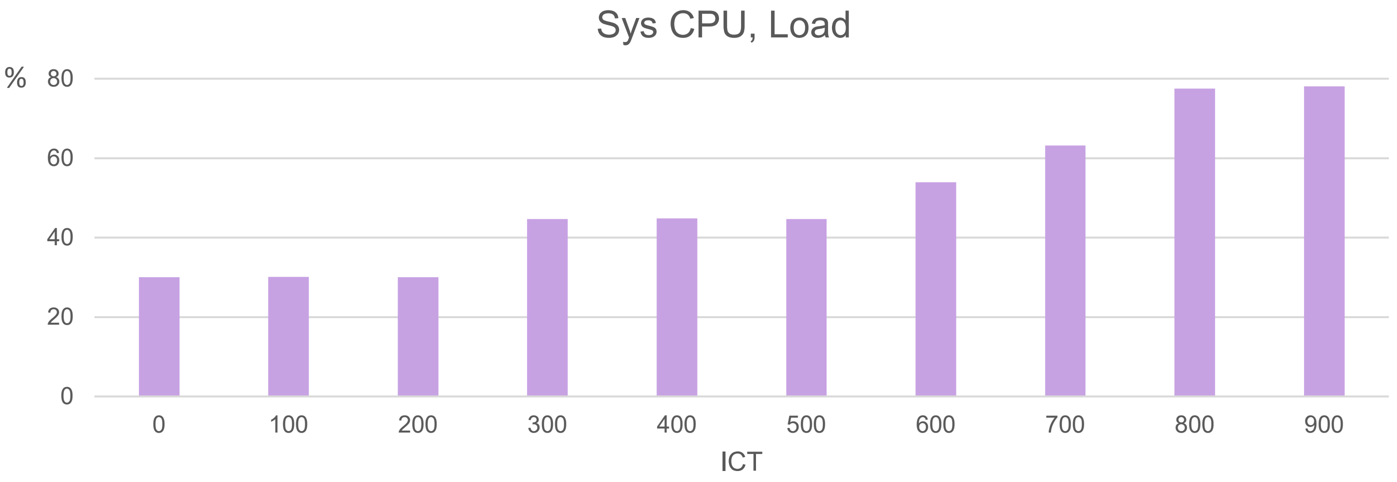
Trying different settings to find the optimal scenario

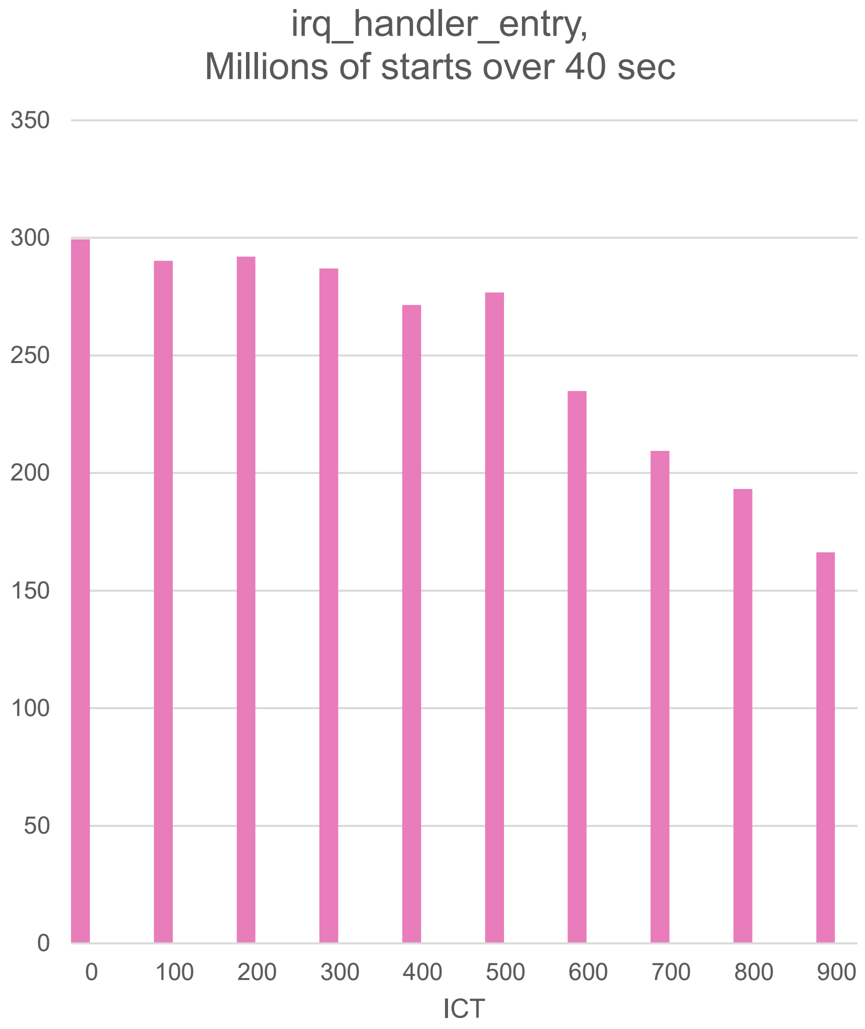


SDC 23

Interrupt coalescing: technical dive







Bad news ☹

- Interrupt coalescing is OKish only for high workloads:
  - QD = 16+ Number of Jobs = 16+
  - QD = 1 or Number of Jobs = 1: **Interrupt coalescing should be switched off!!!**

- Polling mode drivers and io_uring with hipri=1 can “eat” your CPU

- as well, SPDK is not the “REMEDY” for all the cases:
  - Great solution for VirtIO, vfio-user and NVMEoF networks, but…
  - …no support of Linux block devices, and…
  - …significant performance degradation with ublk target

RAID Benchmark with PCIe Gen5 SSD

RAID Engines under review

1. xiRAID (Linux kernel mode)
   - Kernel space driver: expose Linux block devices
   - User space functionality for management
2. xiRAID (Linux user space):
   - SPDK: supports export via VirtIO, vfio-user and NVMEoF
   - Evaluated with SPDK fio plugin
   - User space functionality for management
3. mdRAID (Linux kernel mode only)
   - Kernel 5.4
   - Kernel 6.5 - New
4. ~~RAID5F (Linux user space) – Intel SPDK RAID~~

   Not applicable due to lack of enterprise readiness

How to compare different RAIDs: workloads

1. **Random READ:**
   - in normal and degraded
2. **Random WRITE:**
   - in normal mode and degraded
3. **Sequential WRITE:**
   - in normal mode
   - Full stripe AND not-aligned sequential write
4. **Sequential READ:**
   - in normal and degraded

CPU consumption matters

How to compare: metrics

1. **RAID efficiency** = RAID performance / Raw drive performance
2. **RAID CPU efficiency** = RAID performance / CPU consumption

RAID engines comparison

3. **RAID relative CPU efficiency**

   (RAID Engine1 performance/ CPU consumption)
   (RAID Engine2 performance/ CPU consumption)
   If >1, RAID1 is better than RAID2

4. **RAID relative latency efficiency**

   (RAID Engine2 99,9% latency)
   (RAID Engine1 99,9% latency)
   If >1, RAID1 is better than RAID2

BASELINE definition

- BASELINE is NOT a single number,
- It is the theoretical RAID performance based on:
  - measured RAW drives performance in SPDK
  - Specific workload
- and taking into consideration the RAID penalty

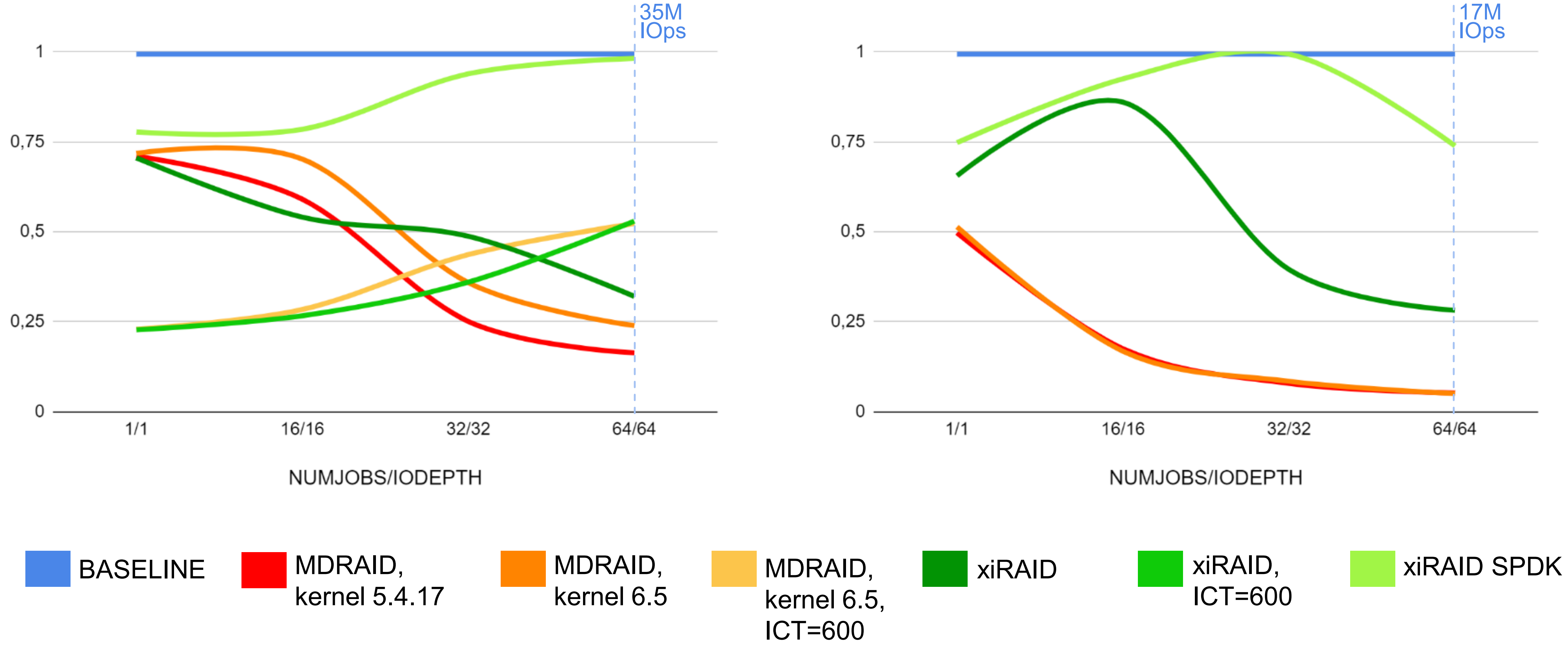
EXAMPLE: RANDOM READS BASELINE

N Jobs/IODepth	BASELINE, IOPs
1/1	40 966
16/16	9 514 053
32/32	23 557 220
64/64	34 982 233

Random Read RAID5x2. RAID Efficiency

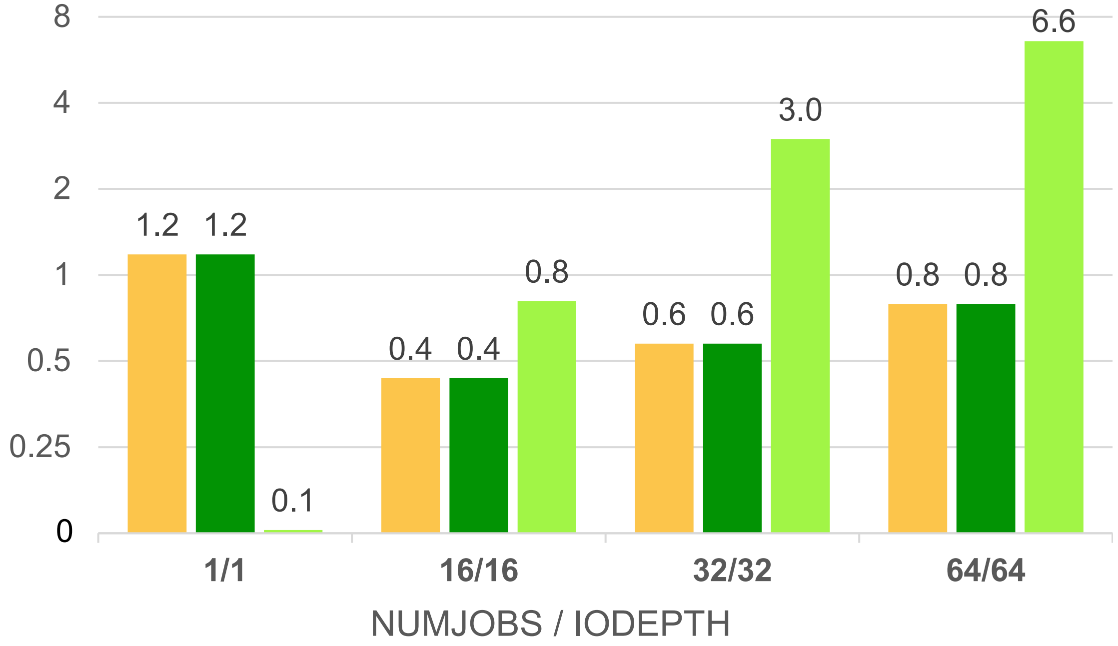
Normal operation

Degraded mode

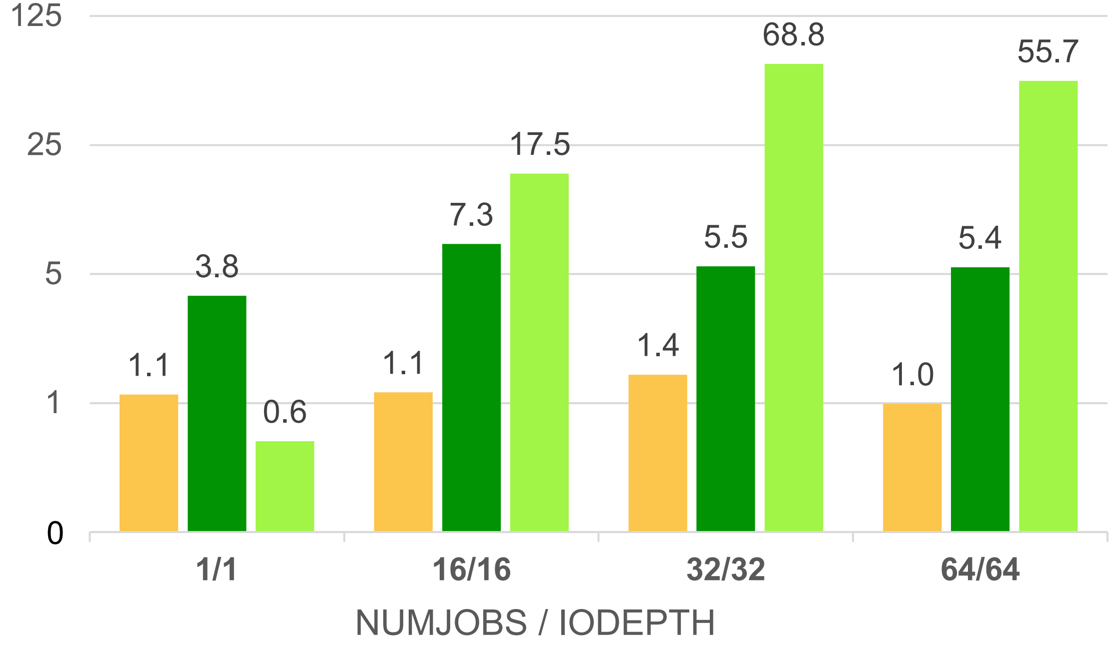


Random Read RAID5x2. RAID CPU relative efficiency
(in relation to MDRAID 5.4)

Normal operation



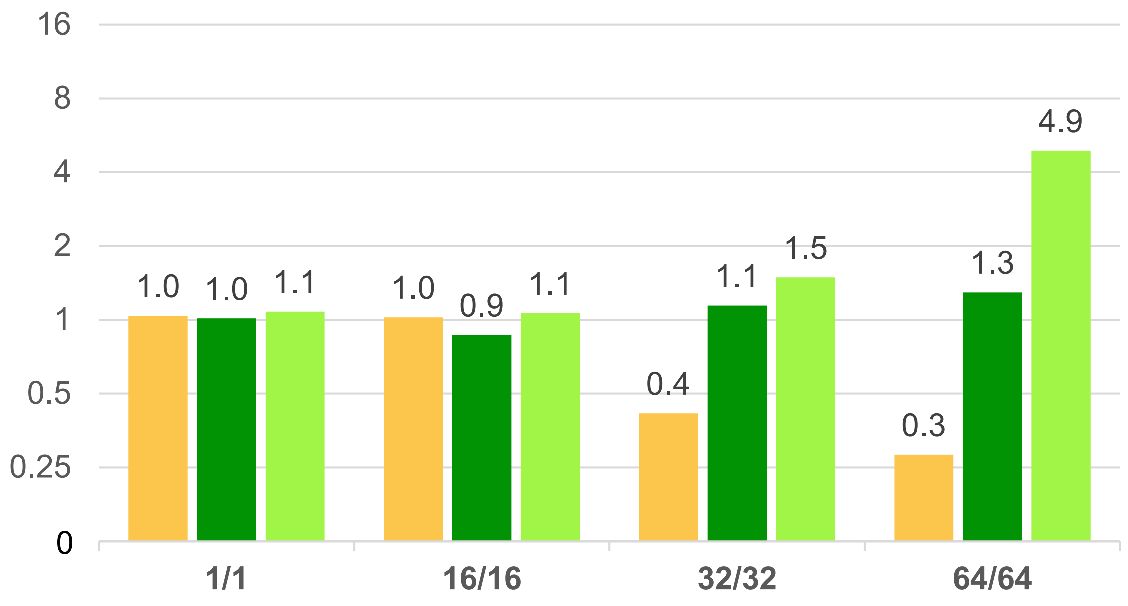
Degraded mode



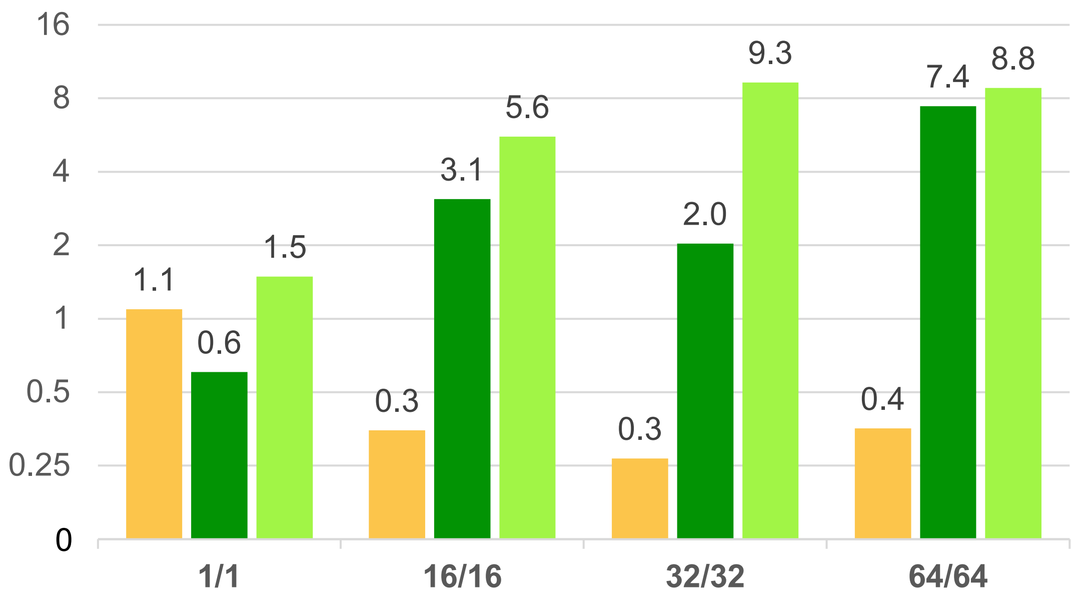
MDRAID, kernel 6.5 xiRAID xiRAID SPDK

Random Read RAID5x2. RAID relative latency efficiency
(in relation to MDRAID 5.4)

Normal operation



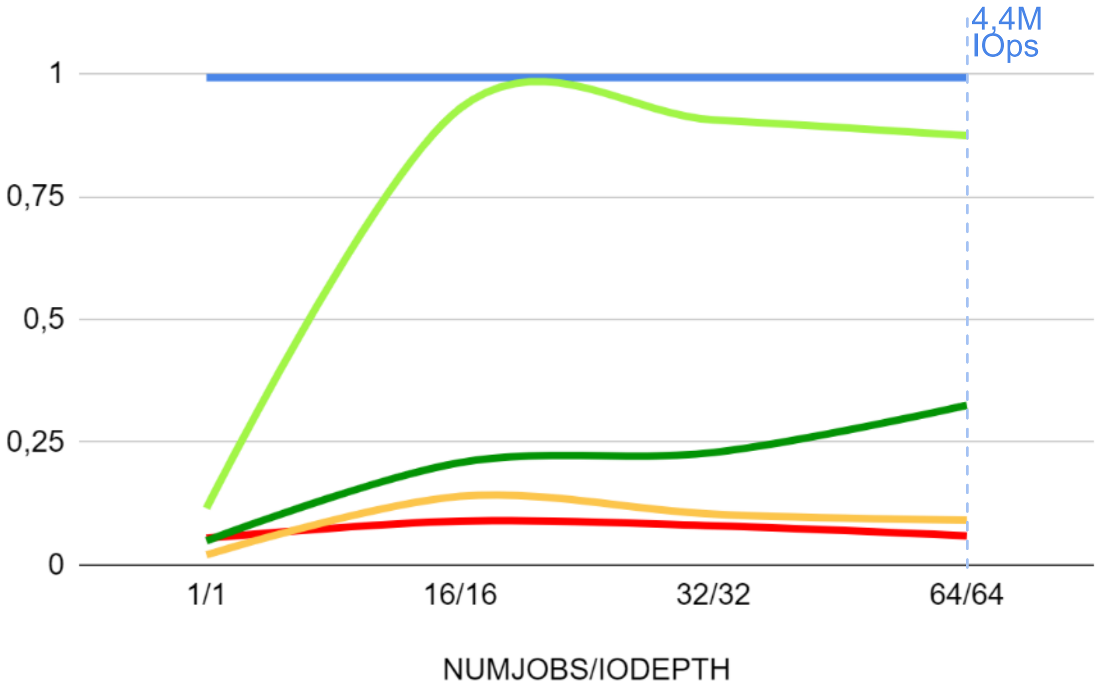
Degraded mode



MDRAID, kernel 6.5 xiRAID xiRAID SPDK

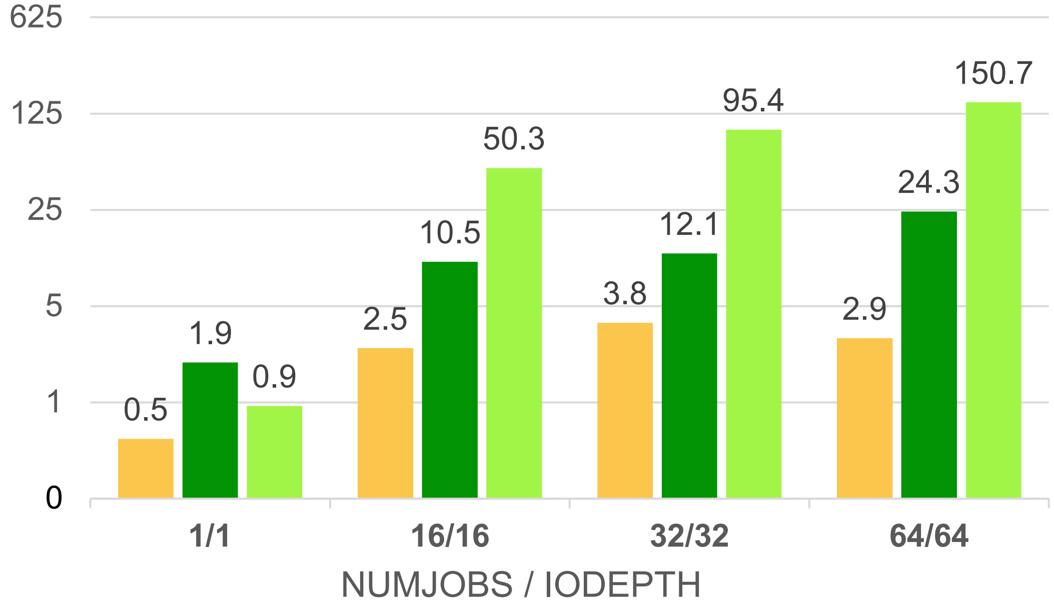
Random Write RAID5x2

RAID Efficiency



RAID Relative CPU Efficiency

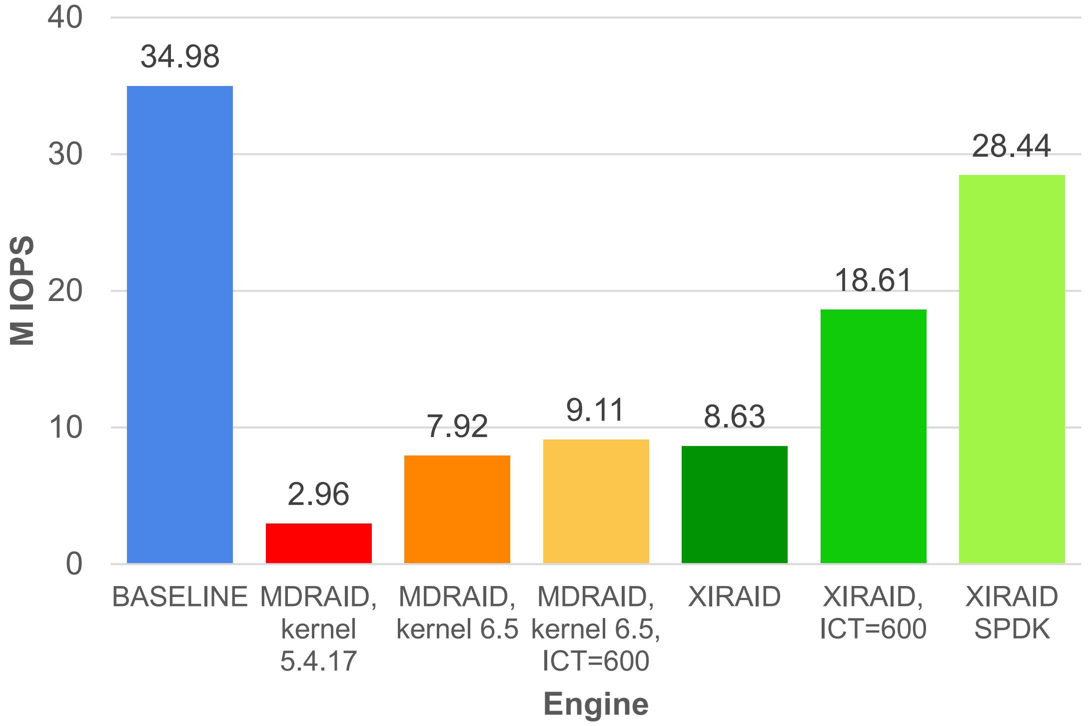
vs MDRAID 5.4



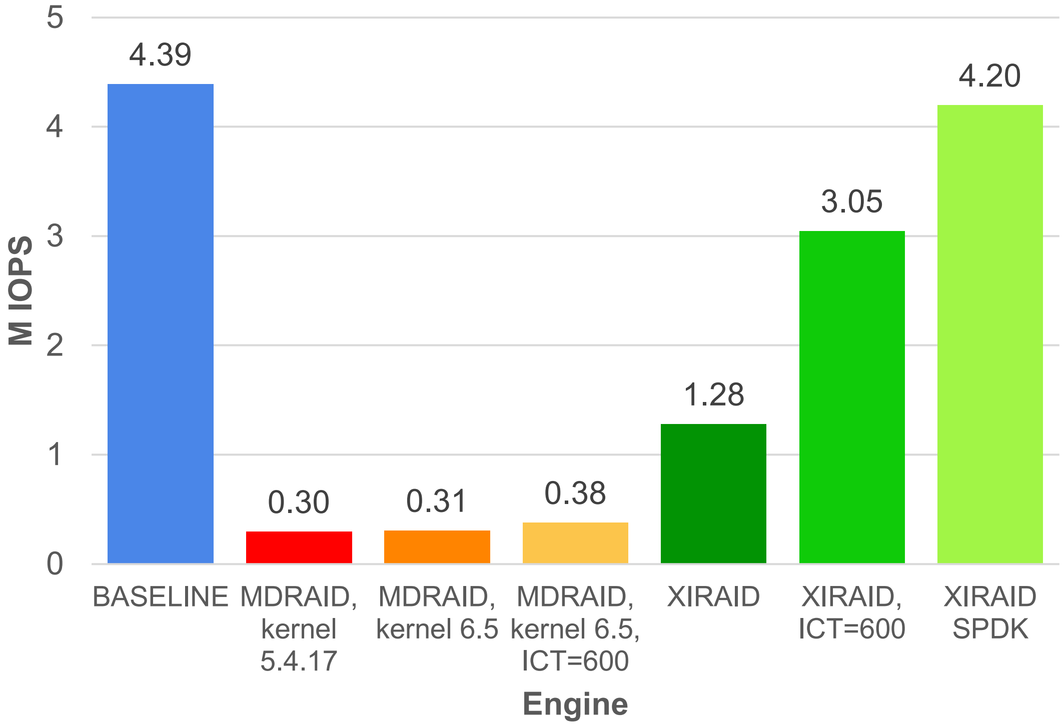
BASELINE | MDRAID, kernel 5.4.17 | MDRAID, kernel 6.5 | xiRAID | xiRAID SPDK

A Single RAID Scalability in RAID 5

Random Read

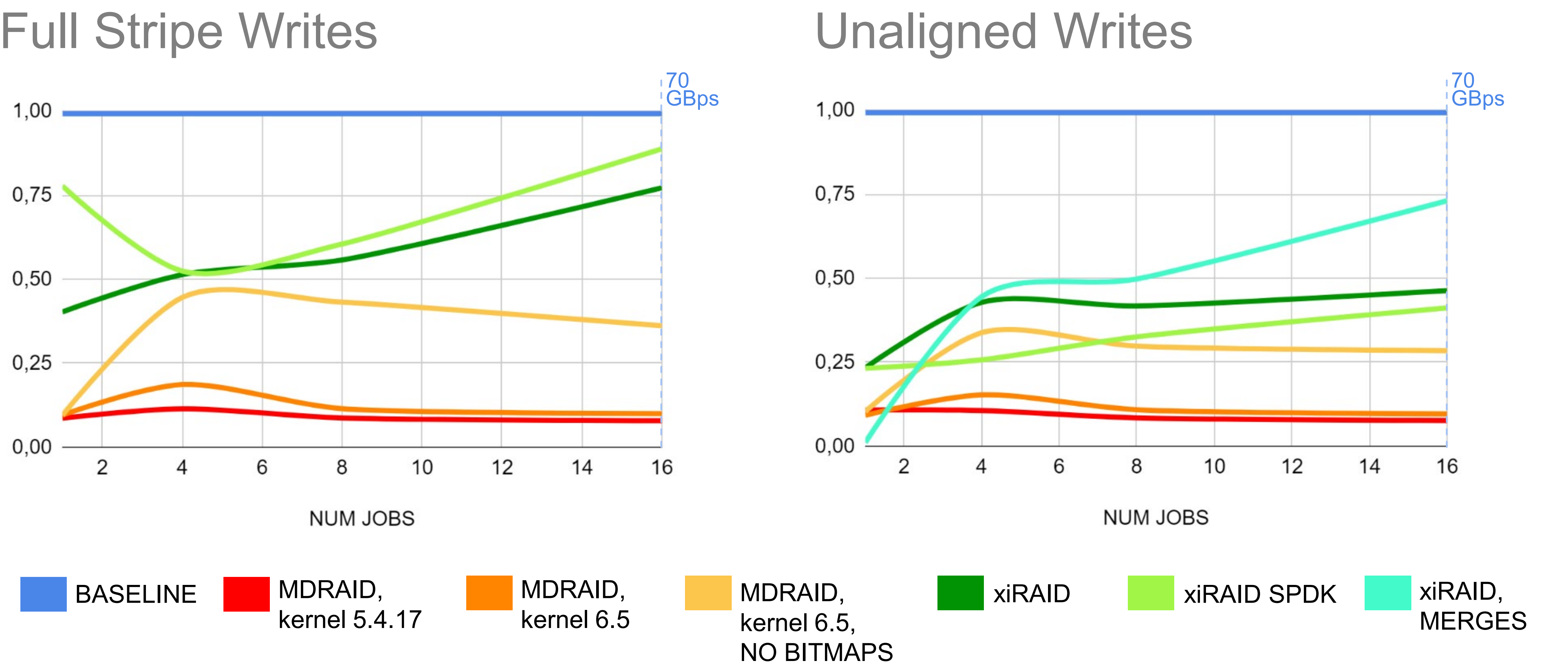


Random Write



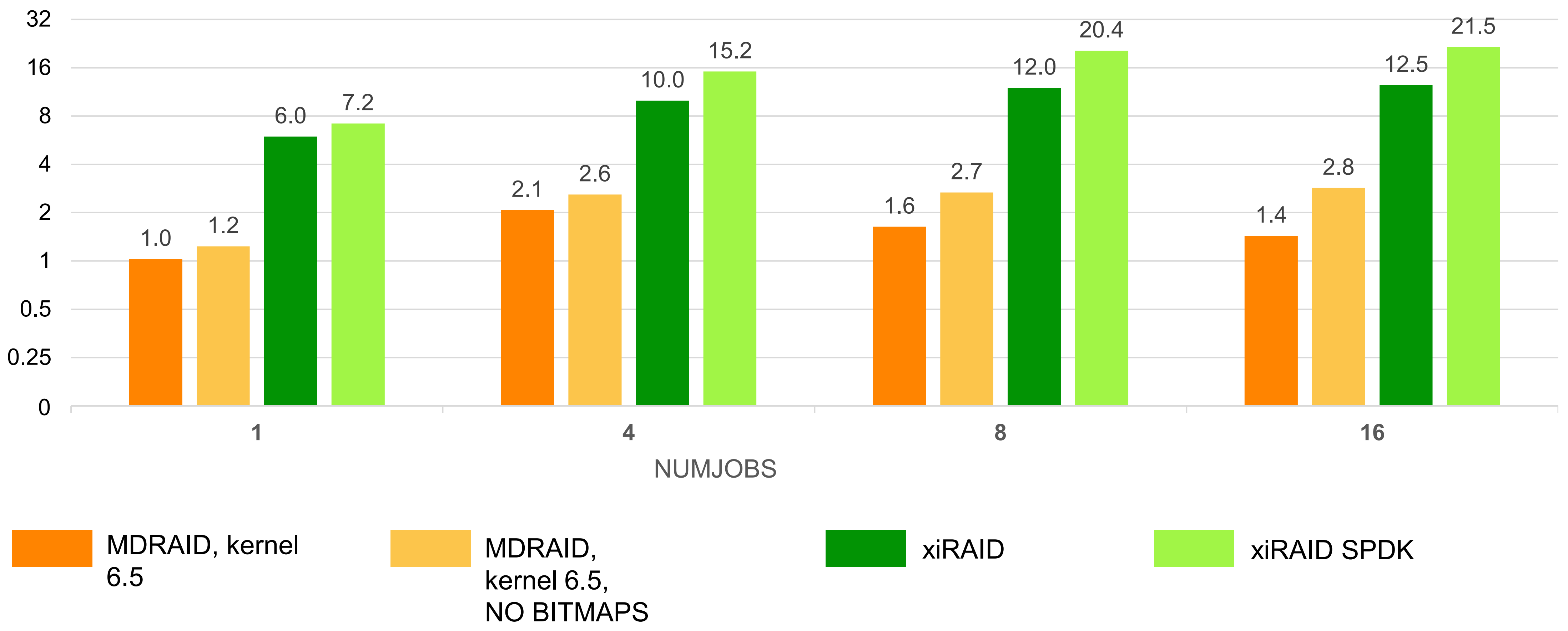
The maximum performance numbers achieved under growing workload
No NUMA NODE affinity
bs=4k

Sequential write RAID6 (10+2). RAID Efficiency



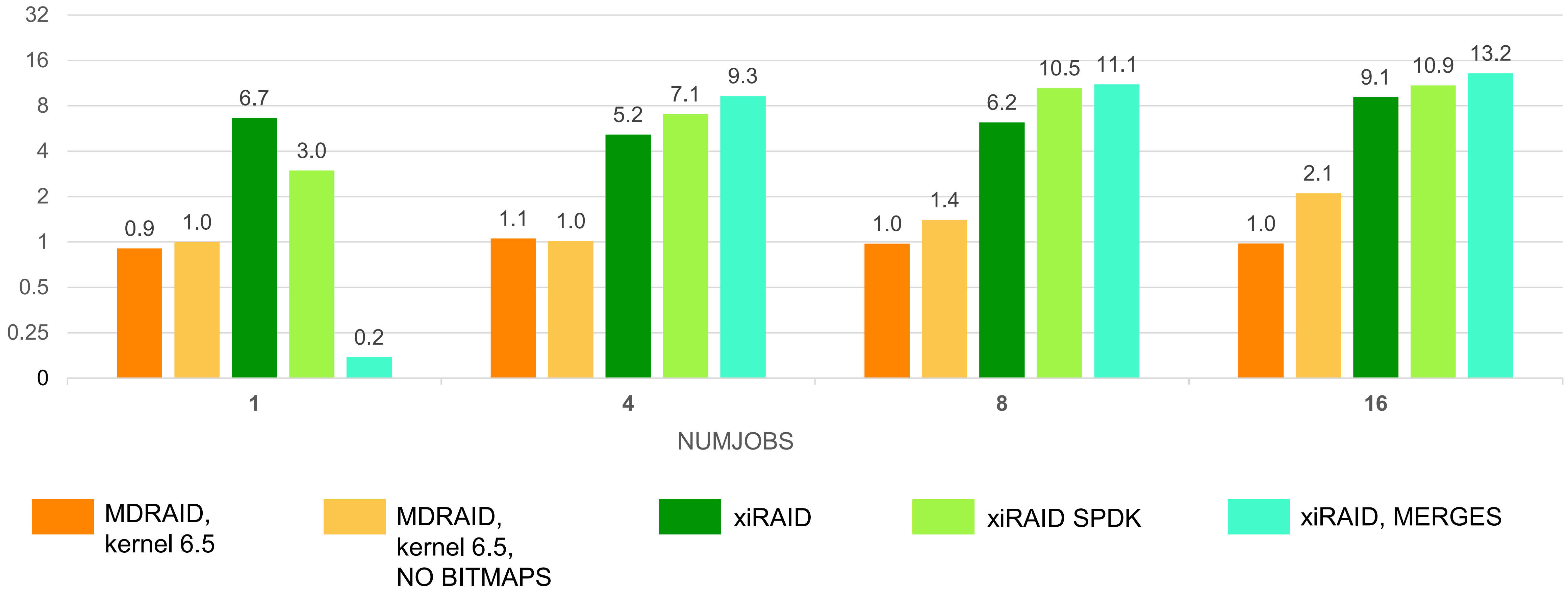
Sequential write RAID6. RAID CPU relative efficiency
(in relation to MDRAID 5.4)

Full Stripe Writes



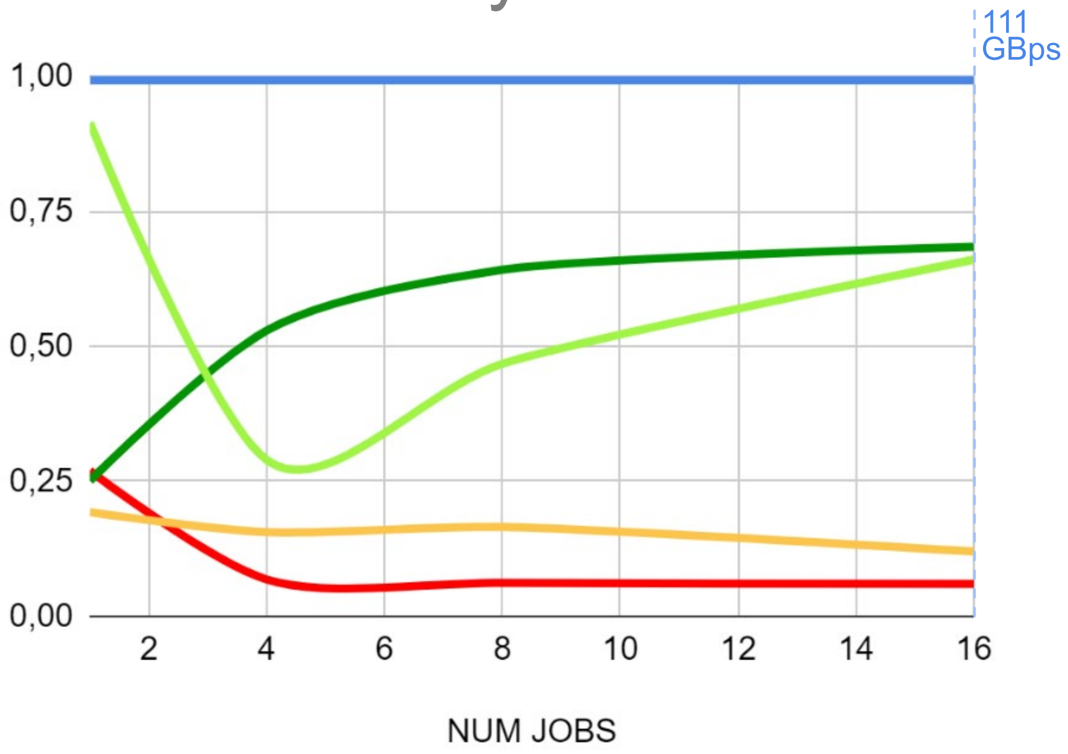
Sequential write RAID6. RAID CPU relative efficiency
(in relation to MDRAID 5.4)

Unaligned Writes

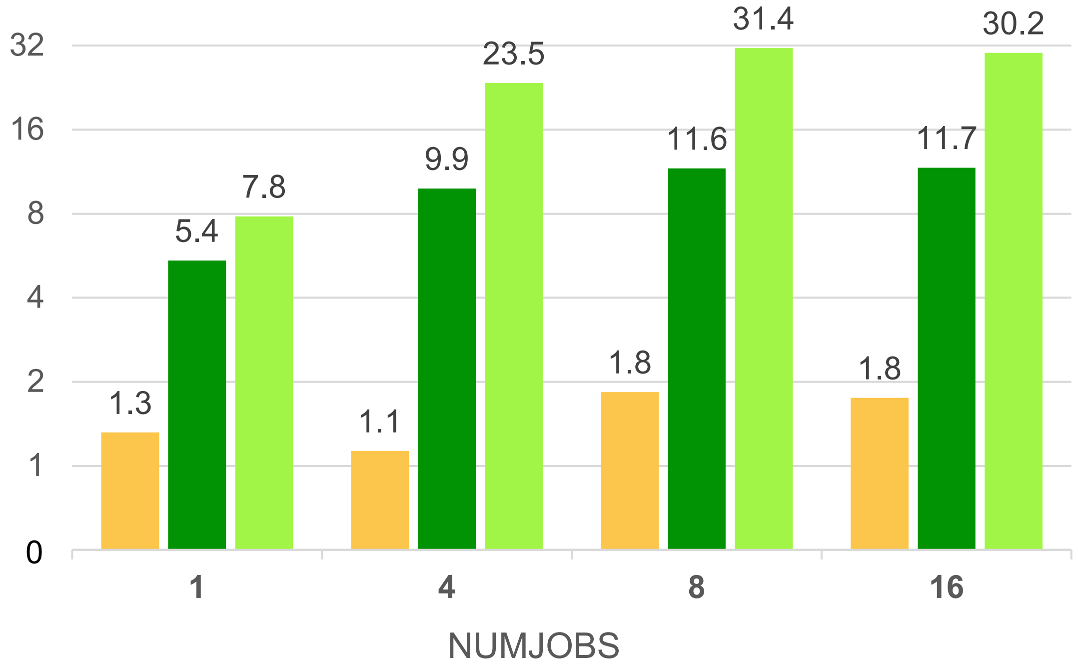


Sequential read Degraded RAID6 (10+2). RAID Efficiency and RAID CPU relative efficiency

RAID Efficiency



CPU Efficiency



BASELINE | MDRAID, kernel 5.4.17 | MDRAID, kernel 6.5 | xiRAID | xiRAID SPDK

Final considerations

Conclusions

1. Proper system tuning is critical to enable performance scalability on PCIe Gen5 environment
2. RAID benchmarks should look at multiple variables:
   - Normal and degraded mode
   - Different workloads
   - Performance vs CPU and latency efficiency
3. MDRAID 6.5 provides performance improvements in normal operations but not in degraded mode and sometimes at the expense of CPU and latency efficiency
4. For Block Devices, xiRAID (kernel) outperforms by multiple times MDRAID 6.5, particularly in degraded mode, random and sequential write and in CPU and latency efficiency.
5. In virtualized environments and NVMeoF, with xiRAID SPDK we can exploit almost full theoretical PCIe Gen5 performance

Q&A

Please take a moment to rate this session.

Your feedback is important to us.