

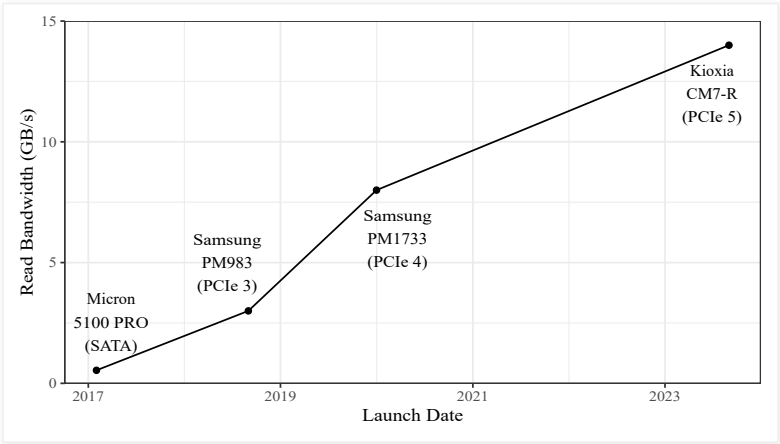
# Database Architects

A blog by and for database architects.

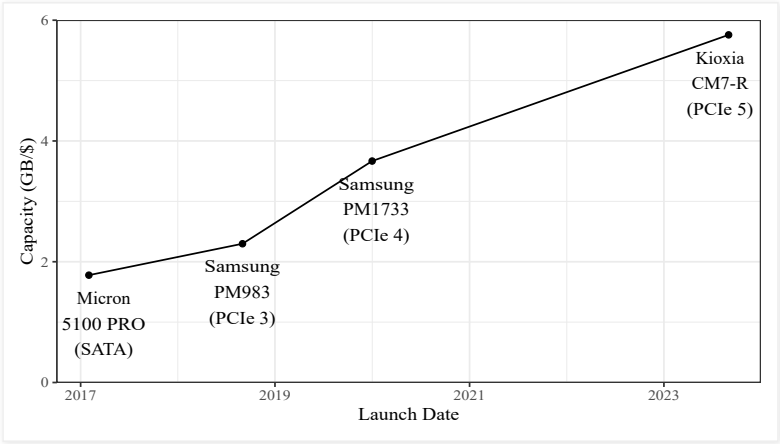
Monday, February 19, 2024

## SSDs Have Become Ridiculously Fast, Except in the Cloud

In recent years, flash-based SSDs have largely replaced disks for most storage use cases. Internally, each SSD consists of many independent flash chips, each of which can be accessed in parallel. Assuming the SSD controller keeps up, the throughput of an SSD therefore primarily depends on the interface speed to the host. In the past six years, we have seen a rapid transition from SATA to PCIe 3.0 to PCIe 4.0 to PCIe 5.0. As a result, there was an explosion in SSD throughput:



At the same time, we saw not just better performance, but also more capacity per dollar:



The two plots illustrate the power of a commodity market. The combination of open standards (NVMe and PCIe), huge demand, and competing vendors led to great benefits for customers. Today, top PCIe 5.0 data center SSDs such as the [Kioxia CM7-R](#) or [Samsung PM1743](#) achieve up to 13 GB/s read throughput and 2.7M+ random read IOPS. Modern servers have around 100 PCIe lanes, making it possible to have a dozen of SSDs (each usually using 4 lanes) in a single server at full bandwidth. For example, in our lab we have a single-socket Zen 4 server with 8 Kioxia CM7-R SSDs, which achieves 100GB/s (!) I/O bandwidth:

### Contributors

- [Peter Boncz](#)
- [Thomas Neumann](#)
- [Viktor Leis](#)

### Blog Archive

- ▼ [2024](#) (1)
  - ▼ [February](#) (1)
    - SSDs Have Become Ridiculously Fast, Except in the ...
- [2023](#) (3)
- [2022](#) (7)
- [2021](#) (2)
- [2020](#) (4)
- [2019](#) (3)
- [2018](#) (2)
- [2017](#) (2)
- [2016](#) (2)
- [2015](#) (7)
- [2014](#) (11)

Total DISK READ:		101.70 G/s		Total DISK WRITE:		0.00 B/s	
Current DISK READ:		101.71 G/s		Current DISK WRITE:		0.00 B/s	
TID	PRIO	USER	DISK READ	DISK WRITE	COMMAND		
1	be/4	root	0.00 B/s	0.00 B/s	system --deserialize=56		
2	be/4	root	0.00 B/s	0.00 B/s	[kthreadd]		
3	be/0	root	0.00 B/s	0.00 B/s	[rcu_gp]		
4	be/0	root	0.00 B/s	0.00 B/s	[rcu_par_gp]		
5	be/0	root	0.00 B/s	0.00 B/s	[slub_flushwq]		
6	be/0	root	0.00 B/s	0.00 B/s	[netns]		
8	be/0	root	0.00 B/s	0.00 B/s	[kworker/0:0H-events_highpri]		
11	be/0	root	0.00 B/s	0.00 B/s	[mm_percpu_wq]		
12	be/4	root	0.00 B/s	0.00 B/s	[rcu_tasks_kthreadd]		
13	be/4	root	0.00 B/s	0.00 B/s	[rcu_tasks_rude_kthreadd]		
14	be/4	root	0.00 B/s	0.00 B/s	[rcu_tasks_trace_kthreadd]		
15	be/4	root	0.00 B/s	0.00 B/s	[ksoftirqd/0]		
16	be/4	root	0.00 B/s	0.00 B/s	[rcu_preempt]		
17	rt/4	root	0.00 B/s	0.00 B/s	[migration/0]		
18	rt/4	root	0.00 B/s	0.00 B/s	[idle_inject/0]		
19	be/4	root	0.00 B/s	0.00 B/s	[cpuhp/0]		
20	be/4	root	0.00 B/s	0.00 B/s	[cpuhp/1]		
21	rt/4	root	0.00 B/s	0.00 B/s	[idle_inject/1]		
22	rt/4	root	0.00 B/s	0.00 B/s	[migration/1]		
23	be/4	root	0.00 B/s	0.00 B/s	[ksoftirqd/1]		
25	be/0	root	0.00 B/s	0.00 B/s	[kworker/1:0H-events_highpri]		
26	be/4	root	0.00 B/s	0.00 B/s	[cpuhp/2]		
27	rt/4	root	0.00 B/s	0.00 B/s	[idle_inject/2]		
28	rt/4	root	0.00 B/s	0.00 B/s	[migration/2]		
29	be/4	root	0.00 B/s	0.00 B/s	[ksoftirqd/2]		
31	be/0	root	0.00 B/s	0.00 B/s	[kworker/2:0H-events_highpri]		
32	be/4	root	0.00 B/s	0.00 B/s	[cpuhp/3]		
keys: any: refresh g: quit i: ionice o: active p: proc a: accum							
sort: r: asc left: DISK READ right: COMMAND home: TID end: COMMAND							
CONFIG_TASK_DELAY_ACCT and kernel.task_delayacct sysctl not enabled in kernel, cannot determine SWAPIN and IO %							

AWS EC2 was an early NVMe pioneer, launching the [i3 instance](#) with 8 physically-attached NVMe SSDs in early 2017. At that time, NVMe SSDs were still expensive, and having 8 in a single server was quite remarkable. The per-SSD read (2 GB/s) and write (1 GB/s) performance was considered state of the art as well. Another step forward occurred in 2019 with the launch of [i3en instances](#), which doubled storage capacity per dollar.

Since then, several NVMe instance types, including i4i and im4gn, have been launched. Surprisingly, however, the performance has not increased; seven years after the i3 launch, we are still stuck with 2 GB/s per SSD. Indeed, the venerable i3 and i3en instances basically remain the best EC2 has to offer in terms of IO-bandwidth/\$ and SSD-capacity/\$, respectively. Personally, I find this very surprising given the SSD bandwidth explosion and cost reductions we have seen on the commodity market. At this point, the performance gap between state-of-the-art SSDs and those offered by major cloud vendors, especially in read throughput, write throughput, and IOPS, is nearing an order of magnitude. (Azure's top NVMe instances are only slightly faster than AWS's.)

What makes this stagnation in the cloud even more surprising is that we have seen great advances in other areas. For example, during the same 2017 to 2023 time frame, EC2 network bandwidth exploded, increasing from 10 Gbit/s (c4) to 200 Gbit/s (c7gn). Now, I can only speculate why the cloud vendors have not caught up on the storage side:


- One theory is that EC2 intentionally caps the write speed at 1 GB/s to avoid frequent device failure, given the total number of writes per SSD is limited. However, this does not explain why the read bandwidth is stuck at 2 GB/s.
- A second possibility is that there is no demand for faster storage because very few storage systems can actually exploit tens of GB/s of I/O bandwidth. See our [recent VLDB paper](#). On the other hand, as long as fast storage devices are not widely available, there is also little incentive to optimize existing systems.
- A third theory is that if EC2 were to launch fast and cheap NVMe instance storage, it would disrupt the cost structure of its other storage service (in particular EBS). This is, of course, the classic innovator's dilemma, but one would hope that one of the smaller cloud vendors would make this step to gain a competitive edge.

Overall, I'm not fully convinced by any of these three arguments. Actually, I hope that we'll soon see cloud instances with 10 GB/s SSDs, making this post obsolete.

Posted by [Viktor Leis](#) at 9:00AM

19 comments:

Anonymous February 20, 2024 at 6:42 PM  
Related Hackernews Discussion: <https://news.ycombinator.com/item?id=39443679>  
[Reply](#)

 AdamK February 20, 2024 at 7:27 PM

Cloud providers buy only high capacity drives, so they have less transfer speed per TB of storage available (compared to commodity drives). If the drive is shared between multiple VMs, throughput is shared between them, and they have to obey SLAs. This could be also a way to manage wear of the drives.

[Reply](#)

Anonymous [February 20, 2024 at 9:19 PM](#)

I've seen speeds of 60k iops and 7 GB/s speeds on akamai cloud/linode. Even in the 5\$ nanodes. Varies depending on the class of system you get, you get a random Zen 2 or Zen 3 class core and the better disks are on the Zen 3 instances.

Still pretty slow for databases compared to bare metal. The fractional vCPUs they sell are comparable with the disk difference.

Cloud resources are a pretty bad deal right now and don't reflect the gains from the last 3 years - which have been huge on the CPU side too.

[Reply](#)

Anonymous [February 20, 2024 at 9:20 PM](#)

Interesting. I can say that locally on my workstation, SSDs are beneficial for searching with tools like FileSearchEX, where one has to load the contents in memory of thousands of files over and over to find keywords. But I wonder, as the HN article states, is the reason because of a protocol in front of the actual drives?

[Reply](#)

Anonymous [February 20, 2024 at 11:12 PM](#)

Cloud just means someone else's hardware and your virtual machine shares I/O with everyone. If you want faster, use your own hardware or keep syncing your cloud instance around until you find hardware without noisy neighbors.

[Reply](#)

[Replies](#)

Anonymous [February 21, 2024 at 1:04 AM](#)

The important thing here is that even on the AWS metal instances which are supposed to give you one entire host and even if you choose an instance that has LOCAL NVMe SSDs, those are still ridiculously slow.

---

[Reply](#)

Akash [February 21, 2024 at 12:51 AM](#)

SSD's on all the clouds are Network SSD's and will never be performant as local disk on bare metal. Network will always be slower than local disk. It's just a matter of time when these network volumes will fall behind in numbers compared to local SSD's.

[Reply](#)

[Replies](#)

Anonymous [February 21, 2024 at 1:09 AM](#)

This is not talking about EBS. This is talking about the "instance-local" SSDs.

Also Networking on AWS is now faster than the supposed LOCAL NVMe SSDs that you are supposed to get with the instance types offering them. You can read at 80 GBit/s from S3 if you choose an instance with a lot of network bandwidth but only at 64 GBit/s from "local NVMe SSD" if you have an instance with 4 of them and put them in RAID 0... This is clearly ridiculous.

---

[Reply](#)

Anonymous [February 21, 2024 at 11:53 AM](#)

I don't think the higher ups at the Big Three Cloud Providers are ready for making the investment for the benefits that faster SSDs provide.

[Reply](#)

Anonymous [February 21, 2024 at 1:24 PM](#)

FWIW the value of cloud has gone so far away thanks to the vendor/ecosystem lock-in effect. The big three doesn't have to improve performance or lower prices to compete anymore, when so many companies have swallowed the "devops" buzzpill over having competent system admins and infra engineers..

[Reply](#)



Stuart [February 21, 2024 at 2:40 PM](#)

With a 10+ billion record data file, I could never get Postgres for Azure to match bare metal without spending thousands per month—even when the "bare metal" was a corporate laptop with NVMe running the Windows version of Postgres.

[Reply](#)



Mark Callaghan [February 21, 2024 at 3:23 PM](#)

I look forward to more results on this topic. I probably need to re-read your recent paper(s) on what we can get from modern storage. The issues for me include:

- \* QoS - high write rates can mean high GC rates which bring higher variance from flash GC stalls
- \* QoS, v2 - with an LSM, high write rates also mean high TRIM rates and the ability of a device to support TRIM varies, so your DBMS might have to learn this at run time or there will be stalls during TRIM. FusionIO spoiled us a long ago -- TRIM was always fast. Modern devices don't spoil us.
- \* backup/restore are hard with high write rates. If restoring from a snapshot then you need to somehow replay minutes or hours of logs to catch up
- \* replication -- both network capacity and the ability to replay the log on a replica are a challenge. There has been some recent work on parallel replication apply. We need more.

[Reply](#)

Anonymous [February 21, 2024 at 4:32 PM](#)

Hi Mark. Your points are all related to writes and make a lot of sense for write-heavy OLTP workloads. However, I don't see why cloud vendors don't offer higher read speeds. Viktor

[Reply](#)

[Replies](#)



Mark Callaghan [February 21, 2024 at 4:55 PM](#)

Faster reads would be nice. We might never learn why without working there. Is it network capacity? The perf overhead on VMs? A bottleneck in EBS?

---

[Reply](#)

Anonymous [February 21, 2024 at 7:11 PM](#)

Why don't you press Control + to a font size which suits you? The font size in this blog is quite normal.

[Reply](#)

Anonymous [February 22, 2024 at 1:12 AM](#)

All NVMe access in recent AWS instances is managed ("and secured") by the Nitro backplane running a custom ARM chipset. It's not networked SSD as another commenter suggests but it's also not true local SSD. The timing of Nitro introduction and stalled SSD bandwidth is suggestive. I doubt the Nitro backplane has the capability to emulate NVMe at 100gbps. <https://aws.amazon.com/blogs/hpc/bare-metal-performance-with-the-aws-nitro-system/>

[Reply](#)

Anonymous [February 22, 2024 at 12:33 PM](#)

As a sales from a public cloud provider(non big three) I can confidently say the amount of clients we have that need higher then 2gb/s storage is very very low, and in my 5 years here I have maybe seen 3 people who ask about iops details and SLA. In addition I have always heard our tech team be very cautious about providing higher speed storage, mostly due to fear of network congestion. Bare metal nvme machines seems like the only way to get really high speeds at our cloud at the moment, but this has uptime impact due to being a single machine.

[Reply](#)

[Replies](#)



**Mark Callaghan** [February 22, 2024 at 3:17 PM](#)

I don't doubt that is true for many clients but the big 3 cloud vendors all compete (sell) block storage solutions based on high perf and high QoS - Azure Premium Storage, EBS Provisioned IOPs, Google Hyperdisk. Alas all of these are network attach storage and the focus of this post might be limited to local attach NVMe.

---

[Reply](#)

**Chris Craft (Exadata PM)** [February 23, 2024 at 10:00 PM](#)

Exadata runs in the Cloud and takes full advantage of NVMe performance with full redundancy and persistence of data even during failures and without downtime. Exadata is available in Oracle Cloud, in a data center of your choosing using Exadata on-premises or Cloud@Customer, and in Azure. Make sure to look at Exadata. It's ridiculously fast.

[Reply](#)



Enter Comment

[Home](#)

[Older Post](#)

Subscribe to: [Post Comments \(Atom\)](#)

---

Simple theme. Powered by [Blogger](#).