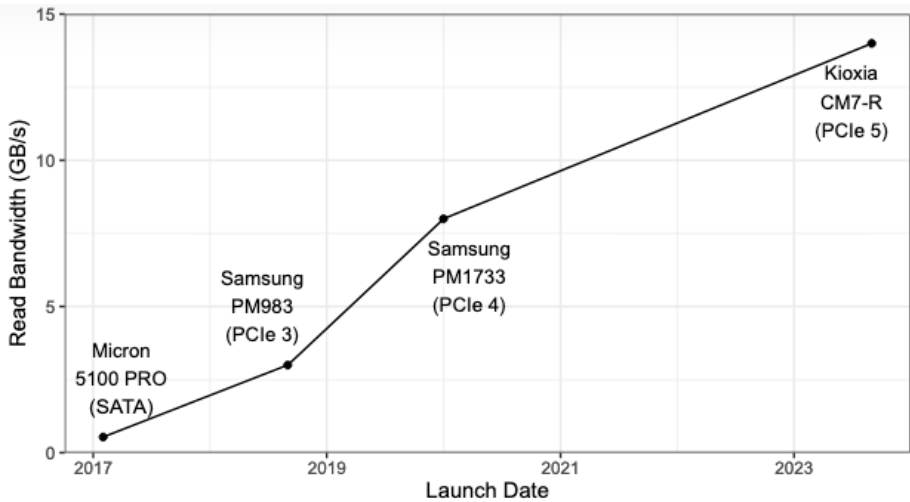


知乎



赞同 2



分享

HN 论坛里网友吵翻了 | SSD硬件速度飙升，唯独云存储未能跟上



小孩姐

每个开发者都想知道的云原生和数据库技术

关注她

你赞同过 TA 的内容

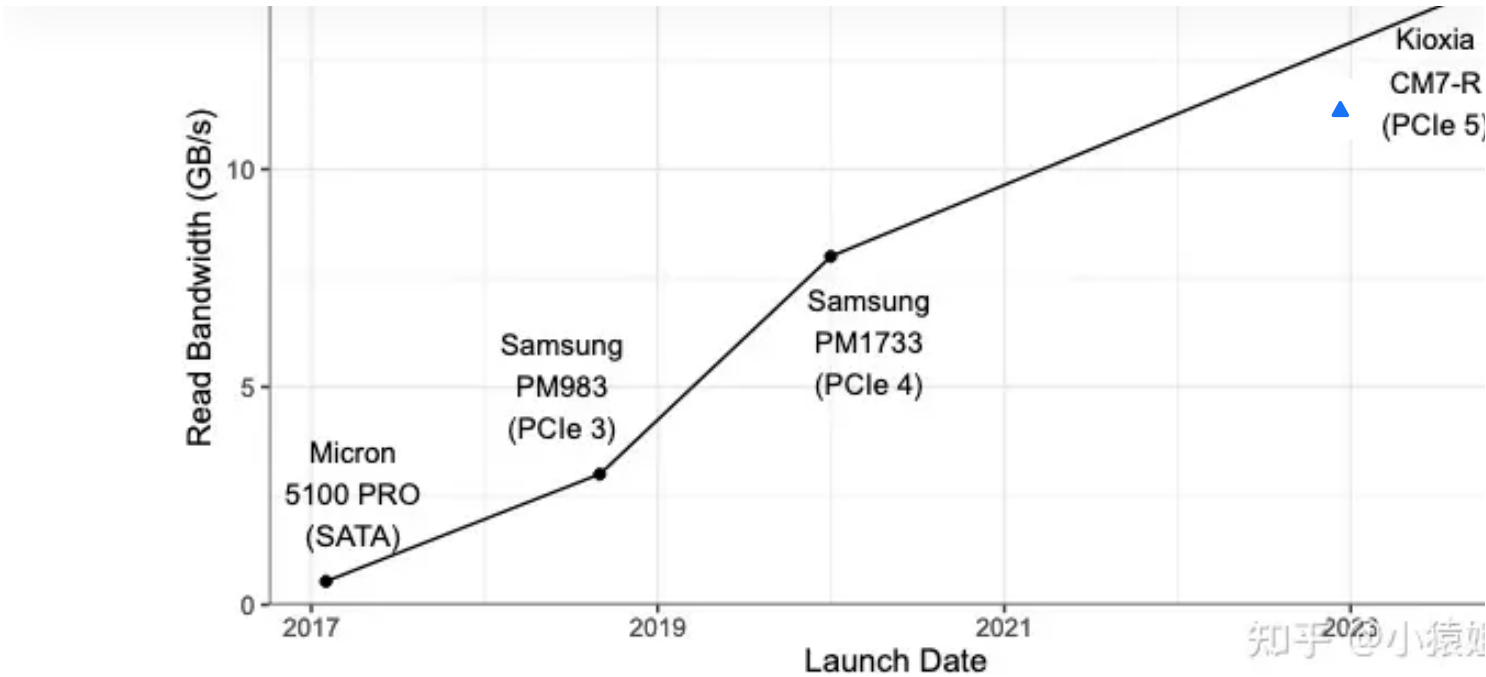
这几天，Hacker News 上的有个帖子下面的网友吵翻了。原帖是发表在 Database Architects 这个博客网站的。作者主要观点是认为基于闪存的固态硬盘（SSD）场景下已取代了磁盘。SSD 的吞吐量取决于与主机的接口速度，随着 PCIe 接口的升级，SSD 的吞吐量也增长。而云供应商的存储性能提升却相对缓慢，仍停留在过去的速度上。最先进的固态硬盘和主要云供应商提供的固态硬盘之间的性能差距，尤其是在读取吞吐量、写入吞吐量和 IOPS 方面，正在逼近一个数量级。究其原因，要么是擦写寿命到期引起故障吧，也可能是缺乏对更快的存储的需求，或者担心扰乱其他存储服务的成本结构。但是这三个点并不足够解释云存储的速度仍然落后的原因。感兴趣的同学可以看看。但是比起原文，原文引发的讨论更有趣。

原文如下

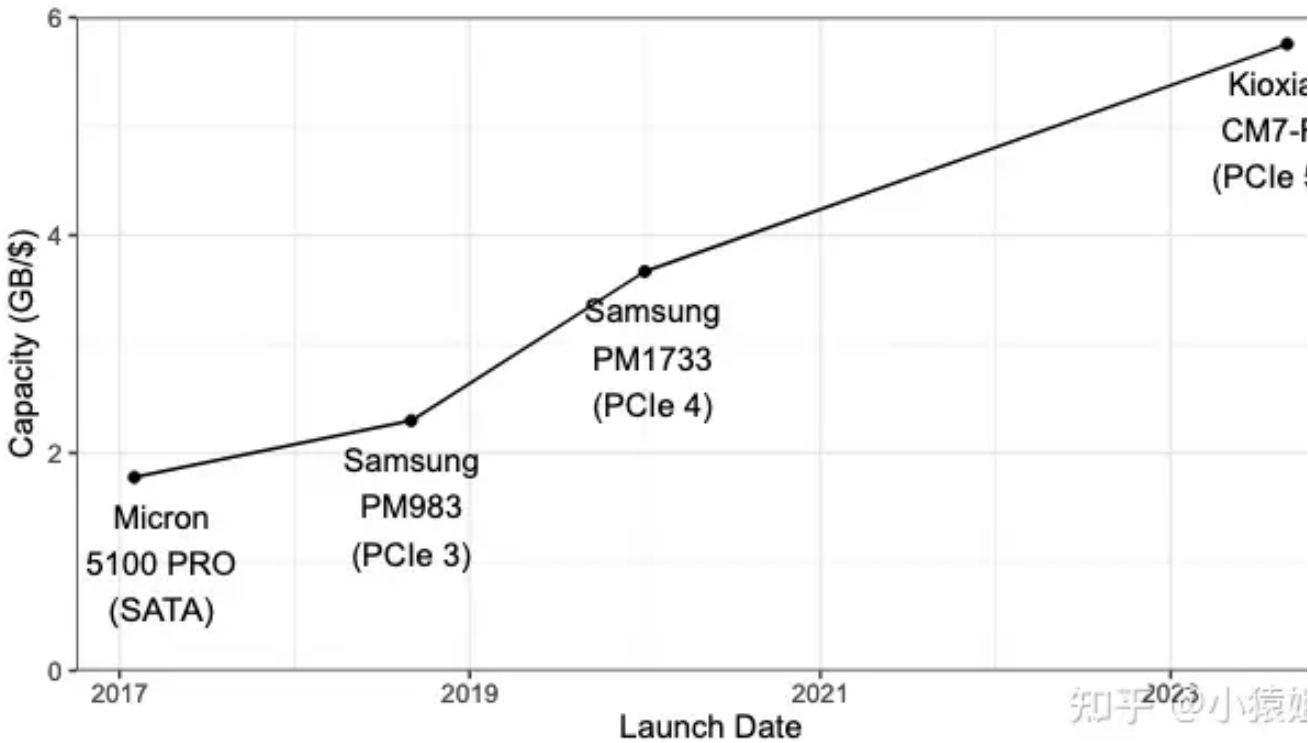
《SSD 硬件速度飙升，唯独云存储未能跟上》

近年来，基于闪存的固态硬盘（SSD）在大多数存储使用情况下已大幅取代了磁盘。每个 SSD 由许多独立的闪存芯片组成，每个芯片都可以并行访问。假设 SSD 的吞吐量主要取决于主机的接口速度。在过去的六年中，我们看到了从 SATA 过渡到 PCIe 3.0 再到 PCIe 4.0 再到 PCIe 5.0 的快速进化。SSD 吞吐量爆发

知乎

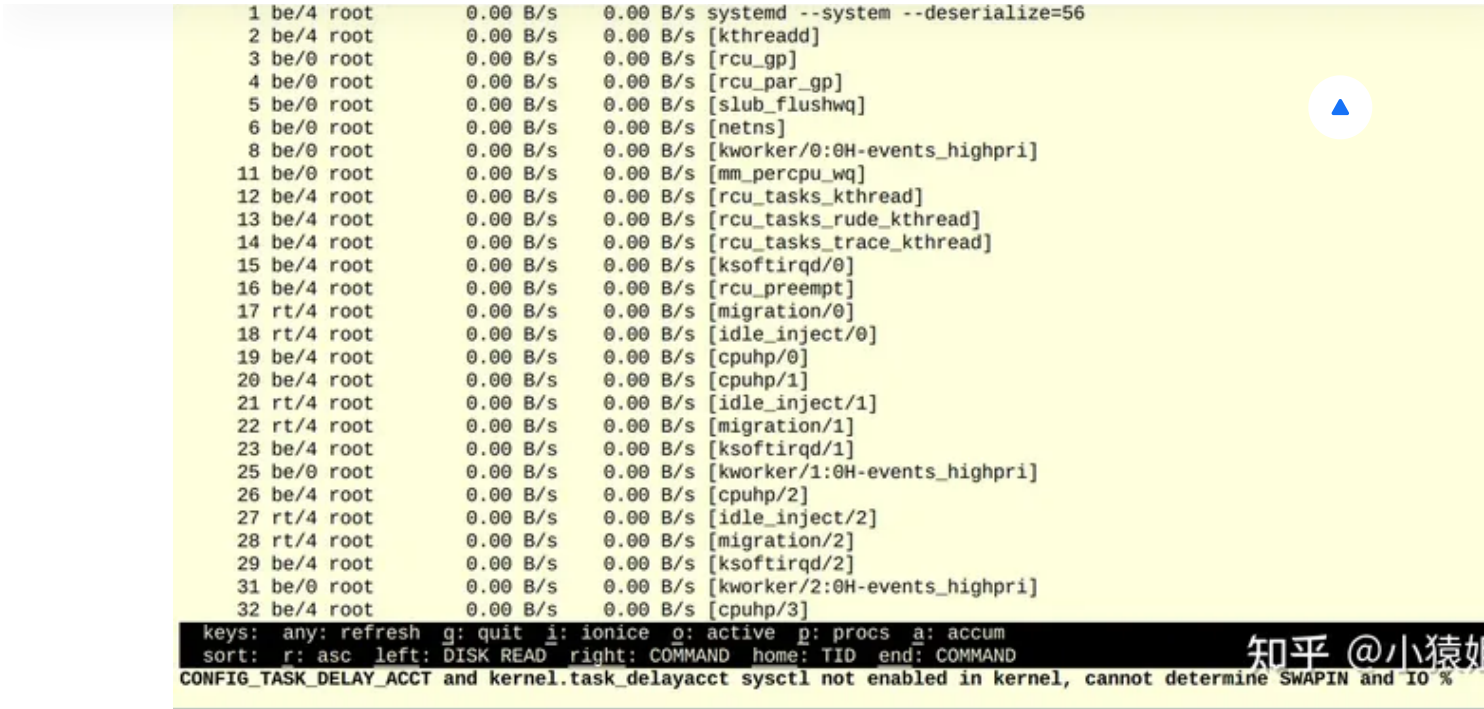


同时，变好的不仅仅是性能，还有成本。



开放标准（如 NVMe 和 PCIe）、巨大的需求加上厂商的内卷，为客户带来了巨大的好处。如今，顶级的 PCIe 5.0 数据中心固态硬盘（SSD），如 Kioxia CM 实现了高达 13 GB/s 的读取吞吐量和超过 270 万的随机读 IOPS。现代服务器拥有约 100 个 PCIe 通道，使得在单个服务器上使用数十个 SSD（每个通常使用为可能。例如，在我们的实验室中，我们有一台单插槽 Zen 4 服务器，配备了 8 个 Kioxia CM7-R SSD，实现了高达 100GB/s 的 I/O 带宽！

知乎



AWS EC2 是 NVMe 技术的先驱，在 2017 年初就推出了 i3 实例，配备了 8 个物理连接的 NVMe 固态硬盘。当时，NVMe 固态硬盘仍然很昂贵。单个固态硬盘写入（1GB/s）性能也被认为是当时的最先进技术。在 2019 年，又迈出了一步，推出了 i3en 实例，成本降了一倍。

自那时以来，已经推出了几种 NVMe 实例类型，包括 i4i 和 im4gn。然而，性能并没有增加；在 i3 推出七年后，我们仍然在每个固态硬盘 2GB/s 的速度上停滞不前。实例仍然是 AWS 在 IOPS 和 SSD 容量/\$ 方面提供的最佳选择。我个人认为，考虑到我们在商品市场上看到的固态硬盘带宽爆发和成本降低，这种停滞不前，固态硬盘和主要云供应商提供的固态硬盘之间的性能差距，尤其是在读取吞吐量、写入吞吐量和 IOPS 方面，正在逼近一个数量级。（Azure 的顶级 NVMe 快。）

更难以理解的是，云计算在其他领域有巨大的进步。例如，在 2017 年到 2023 年同期，EC2 网络带宽爆发增长，从 10 Gbit/s（c4）增加到 200 Gbit/s（c7g）。云供应商在存储方面没有赶上步伐的原因：

- 一种说法是，考虑到每个固态硬盘的总写入次数有限，EC2 有意将写入速度限制在 1GB/s，以避免频繁的设备故障。然而，这并不能解释为什么读取带宽停滞不前。
- 第二种可能性是，没有对更快存储的需求，因为很少有存储系统实际上可以利用数十 GB/s 的 I/O 带宽。请参阅我们最近的 VLDB 论文。只要快速存储设备没有系统的动力也很小。
- 第三，如果 EC2 推出快速且便宜的 NVMe 实例存储，它将扰乱其他存储服务（特别是 EBS）的成本结构。这当然是经典的创新者困境，但人们希望较小的进步，以获得竞争优势。

总的来说，我对这三个说法都不完全信服。我希望我们很快能看到配备 10GB/s 固态硬盘的云实例，使得这篇文章的内容过时。

Hacker News 网友的精彩评论

这篇文章发到 HackerNews，就立即上了首页，并引发了激烈的讨论。大多数网友表述支持文章观点，并从不同角度声援原文。

“确实如此”篇

这位网友认为云计算的问题比原文提及的更严重，因为 IOPS 才是关键。他指出了云盘 IOPS 不足和没有暴露出存储层次结构供上层业务（例如数据库）进行优

知乎

Last time (about a year ago) I ran a couple random IO benchmarks against a storage optimized instances and the random IOPs behavior is closer to a large spinning RAID array than SSDs if the disk size is over some threshold.

IIRC, What it looks like is that there is a fast local SSD cache with a couple hundred GB of storage and the rest is backed by remote spinning media.

Its one of the many reasons I have a hard time taking cloud optimization seriously, the lack of direct tiering controls means that database/etc style workloads are not going to optimize well and that will end up costing a lot of money.

So, maybe it was the instance types/configuration I was using, but <shrug> it was just something I was testing in passing.

有网友结合自己的实际经历, 认为如果每年的云计算费用已经超过了 5000 万美元, 在本地环境中运行大规模私有云可能比云计算更具成本效益, 在大数据和互联网公司已经采取了这种方式。

看来, 国际上, 对于下云的思考和实践比国内要迅速和果断的多。

▲ vinay_ys 1 day ago | prev | next [-]

Well, as hardware becomes more and more powerful, what's possible in a small footprint becomes bigger and bigger. And distributed software for disaggregated storage is becoming more accessible. You put these two together, running on-prem footprints at the scale of 50-100M\$ capex makes a lot of sense. In my personal experience, at this scale, (if your cloud bill compute+storage+local-network is \$50M+/year), you can get 2-4x more on-prem private-cloud capacity for the same money. Of course, this only makes sense if you have an in-house software engineering team already and the marginal cost of adding another 50-100 engineers to build and operate this is strategically valuable to your business.

In big data AI space, this is exactly what's happening with the top 20th to 100th companies in the world right now.

还有网友说道, 如果对 IOPS 和带宽有较高要求, 且不需要弹性扩展能力, 更便宜的专用服务器难道不香吗?

▲ fabioyy 1 day ago | prev | next [-]

it is not worth to use cloud if you need a lot of iops/bandwidth
heck, its not worth for anything besides scalability
dedicated servers are wayyyy cheaper

有人把话题发散开来, 猜想这会不会是因为云厂商到目前为止还没有采用最新硬件比如新的 CPU?

▲ eisa01 1 day ago | prev | next [-]

Would this be a consequence of the cloud providers not being on the latest technology CPU-wise?

At least I have the impression they are lagging, eg., still offering things like: z1d: Skylake (2017) <https://aws.amazon.com/ec2/instance-types/z1d/> x2i: Cascade Lake (2019) and Ice lake (2021) <https://aws.amazon.com/ec2/instance-types/x2i/>

I have not been able to find instances powered by the 4th (Q1 2023) or 5th generation (Q4 2023) Xeons?

We solve large capacity expansion power market models that need as fast single-threaded performance as possible coupled with lots of RAM (32:1 ratio or higher ideal). One model may take 256-512 GB RAM, but not being able to use more than 4 threads effectively (interior point algorithms have very diminishing returns past this point)

Our dispatch models do not have the same RAM requirement, but you still wish to have the fastest single-threaded processing available (and then parallelize)

[reply](#)

▲ zokier 1 day ago | parent | next [-]

AWS was offering Sapphire Rapids instances before those CPUs became even publicly available

<https://aws.amazon.com/about-aws/whats-new/2022/11/introducing-sapphire-rapids/>

[reply](#)

▲ deadmutex 1 day ago | parent | prev | next [-]

You can find Intel Sapphire Rapids powered VM instances on GCE

不过有网友回复说 AWS 其实已经提供了相关服务, 比如intel最新一代CPU "Sapphire Rapids".

知乎

▲ Ericson2314 11 hours ago | prev | next [-]

The cloud really is a scam for those afraid of hardware
reply

知乎 @小猿姐

手撕 bench 篇

当然也有较真的网友真的做了 bench 测试，用数据来支持文章观点，有以下发现：

- Azure 网络延迟约为 85 微秒。
- AWS 网络延迟约为 55 微秒。
- 两者都可以做得更好，但仅限于特殊情况，例如 HPC 集群中的 RDMA NIC。
- 跨 VPC 或跨 VNET 基本相同。有些人说它非常慢，但我在测试中没有看到这一点。
- 由于不可避免的光速延迟，跨区时间为 300-1200 微秒。
- 两个云的虚拟机到虚拟机带宽均超过 10 Gbps (>1 GB/s)，即使对于最小的两个 vCPU 虚拟机也是如此！
- Azure Premium SSD v1 延迟大约在 800 到 3,000 微秒之间变化，这比网络延迟要差很多倍。
- Azure Premium SSD v2 延迟约为 400 到 2,000 微秒，这并没有好多少，因为：
 - Azure 中的本地 SSD 缓存比远程磁盘快得多，我们发现Premium SSD v1 几乎总是比Premium SSD v2 快，因为后者不支持缓存。
 - 同样在 Azure 中，本地 SSD缓存和本地临时磁盘的延迟均低至 40 微秒，与现代笔记本电脑 NVMe 驱动器相当。我们发现，切换到最新一代 VM 并打开可以让数据库一键加速.....而且没有丢失数据的风险。

我们研究了两个云中的各种本地 SSD VM机型，例如 Lasv3 系列，正如文章提到的，性能增量并没有让我大吃一惊，但数据丢失风险让这些不值一提。

分析原因篇

对于文章结尾提到带宽被限制了的问题，也有网友找来OSDI的一篇论文《mClock: Handling Throughput Variability for Hypervisor IO Scheduling》，说在不同虚拟机之间进行公平的 I/O 调度的算法，可以帮助避免“噪声邻居”问题，即一个虚拟机的高负载会影响同一物理主机上其他虚拟机的性能）可以解决

也有网友从成本角度出发为原文总结的理由引入新观点，SSD 存储因为数据擦写放大效应会带来寿命下降的问题，而且云厂商维护海量的SSD硬件，并对它们维护成本。

▲ cogman10 1 day ago | prev | next [-]

There's a 4th option. Cost.

The fastest SSDs tend to also be MLC which tend to have much lower write life vs other technologies. This isn't unusual, increasing data density generally also makes it easier to increase performance. However, it's at the cost that the writes are typically done for a block/cell in memory rather than for single bits. So if one cell goes bad, they all fail.

But even if that's not the problem, there is a problem of upgrading the fleet in a cost effective mechanism. When you start introducing new tech into the stack, replacing that tech now requires your datacenters to have 2 different types of hardware AND for the techs swapping drives to have a way to identify and replace that stuff when it goes bad

reply

知乎 @小猿姐

来自 Oracle Cloud Infra 的网友也补充了自己的看法，他认为可能是缺乏具体的需求反馈。

知乎

We offer faster NVMe drives in instances. Our E4 Dense shapes ship with SAMSUNG MZWLJ7T6HALA-00AU3, which supports Sequential Reads of 7000 MB/s, and Sequential Write 3800 MB/s.

From a general perspective, I would say the _likely_ answer to why AWS doesn't have faster NVMe's at the moment is likely be lack of specific demand. That's a guess, but that's generally how things go. If there's not enough specific demand and being in through TAMs and the like for faster disks, upgrades are likely to be more of an after-thought, or reflecting supply chain.

I know there's a tendency when you engineer things, to just work around, or work with the constraints, and grumble among your team, but it's incredibly invaluable if you can make sure your account manager knows what shortcomings you're faced with and work around.

解决方案（趁机广告）篇

有网友（疑似微软员工）说，AWS 和 Azure 面对的问题是“虚拟化成本”。Azure 的下一代“Azure Boost”虚拟化技术，VM 操作系统内核直接与硬件对话程序，单个虚拟机的 IOPS 高达 380 万，性能将得到大幅提升。

▲ jiggawatts 1 day ago | prev | next [-]

There's a lot of talk about cloud network and disk performance in this thread. I recently benchmarked both Azure and AWS and found that:

- Azure network latency is about 85 microseconds.
- AWS network latency is about 55 microseconds.
- Both can do better, but only in special circumstances such as RDMA NICs in HPC clusters.
- Cross-VPC or cross-VNET is basically identical. Some people were saying it's terribly slow, but I didn't see that in my tests.
- Cross-zone is 300-1200 microseconds due to the inescapable speed of light delay.
- VM-to-VM bandwidth is over 10 Gbps (>1 GB/s) for both clouds, even for the *smallest* two vCPU VMs!
- Azure Premium SSD v1 latency varies between about 800 to 3,000 microseconds, which is many times worse than the network latency.
- Azure Premium SSD v2 latency is about 400 to 2,000 microseconds, which isn't that much better, because:
- Local SSD *caches* in Azure are so much faster than remote disk that we found that Premium SSD v1 is almost always faster than Premium SSD v2 because the latter doesn't support caching.
- Again in Azure, the local SSD "cache" and also the local "temp disks" both have latency as low as 40 microseconds, on par with a modern laptop NVMe drive. We found that switching to the latest-gen VM SKU and turning on the "read caching" for the data disks was the magic "go-fast" button for databases... without the risk of losing out data.

We investigated the various local-SSD VM SKUs in both clouds such as the Lasv3 series, and as the article mentioned, the performance delta didn't blow my skirt up, but the data loss risk made these not worth the hassle.

[reply](#)

▲ computerdork 1 day ago | parent | next [-]

Interesting. And would you happen to have the numbers on the performance of the local SSD? Is it's read and write throughput to the level of modern SSD's?

[reply](#)

知乎

laptop, let alone a high-end server.

I'm not an insider and don't have any exclusive knowledge, but from reading a lot about the topic my impression is that issue in both clouds is the virtualization overheads.

That is, having the networking or storage go through *any* hypervisor software layer is what kills the performance. I've seen similar numbers with on-prem VMware, Xen, and Nutanix setups as well.

Both clouds appear to be working on next-generation VM SKUs where the hypervisor network and storage functions are offloaded into 100% hardware, either into FPGAs or custom ASICs.

"Azure Boost" is Microsoft's marketing name for this, and it basically amounts to both local and remote disks going through an NVMe controller directly mapped into the memory space of the VM. That is, the VM OS kernel talks *directly* to the hardware, bypassing the hypervisor completely. This is shown in their documentation diagrams:

<https://learn.microsoft.com/en-us/azure/azure-boost/overview>

They're claiming up to 3.8M IOPS for a single VM, which is 3-10x what you'd get out of a single NVMe SSD stick, so... not too shabby at all!

Similarly, Microsoft Azure Network Adapter (MANA) is the equivalent for the NIC, which will similarly connect the VM OS directly into the network, bypassing the hypervisor software.

I'm not an AWS expert, but from what I've seen they've been working on similar tech (Nitro) for years.

[reply](#)

▲ computerdork 1 day ago | root | parent | next [-]

Makes a lot of sense! Yeah, seems like for the OP's performance-issue, you pretty much have the reason why it's happening (VM overhead) and solutions for it (bypassing the software layer using custom hardware like Azure Boost).

Thanks for the info!

[reply](#)

知乎 @小猿姐

除了对原文总结的原因进行补充, 也有网友分享了自己的存储解决方案:

有选择混合方案的:

▲ spintin 1 day ago | prev | next [-]

I'm going with hybrid:

- 2011 X-25E 64GB (2W write and almost nothing read/idle) at 100.000 writes per bit for OS

- 2021 PM897 3.7TB (2.3 Watt (read) | 3 Watt (write) | 1.4 Watt (idle) down from the PM983 (8.7 Watt (read) | 10.6 Watt (write) | 1.4 Watt (idle)) for DB.

This way I can get the most robust solution, with largest DB at lowest power. They are both in a 8-core Atom Mini-ITX board at 25W TDP.

[reply](#)

知乎 @小猿姐

也有还没调研好方案, 求推荐针对小团队的 AWS 平替. 热心网友 (各其他厂员工) 当场安利一波, 比如 Hetzner、Entrywan、Supabase等。

▲ teaearylgraycold 1 day ago | prev | next [-]

What's a good small cloud competitor to AWS? For teams that just need two AZs to get HA and your standard stuff like VMs, k8s, etc.

[reply](#)

▲ catherinacodes 1 day ago | parent | next [-]

Hetzner and Entrywan are pure-play cloud companies with good prices and support. Hetzner is based in Germany and Entrywan is based in the US.

[reply](#)

▲ madars 1 day ago | root | parent | next [-]

Hetzner has a reputation for locking accounts for "identity verification" (Google "hetzner kyc" or "hetzner identity verification"). Might be worthwhile to go through a reseller just to avoid downtime like that.

[reply](#)

▲ infamia 1 day ago | root | parent | prev | next [-]

Thanks for mentioning Entrywan, they look great from what I can tell on their site. Have you used their services? If so, curious about your experiences with them.

[reply](#)

▲ catherinacodes 1 day ago | root | parent | next [-]

My irc bouncer and two kubernetes clusters are running there. So far the service has been good.

[reply](#)

知乎 @小猿姐

知乎

it to be absolutely genius.

It basically allows me to forego having to make a server for the CRUD operations so I can focus on the actual business implications. My REST API is automatically managed for me (mostly with lightweight views and functions) and all of my other logic is either spread out through edge functions or in a separate data store (Redis) where I perform the more CPU intensive operations related to my business.

There's some rough edges around their documentation and DX but I'm really loving it so far.

[reply](#)

▲ pritambarhate 1 day ago | parent | prev | next [-]

Digital Ocean. But they don't have concept of multi az as far as I know. But they have multiple data centers in same region. But I am not aware if there is any private networking between DCs in the same region.

[reply](#)

▲ infecto 1 day ago | parent | prev | next [-]

AWS is pretty great and I think reasonably cheap, I would include any of the other large cloud players. The amount saved is just not reasonable enough for me, I would rather work with something that I have used over and over with.

Along with that, ChatGPT has knocked down most of the remaining barriers I have had when permissions get confusing in one of the cloud services.

[reply](#)

反对的声音也有--但很少

由于作者提出来的数据是个事实，

反对者使用 AWS 博客提供的数据质疑了作者文章中说的 “(AWS) 在 i3 发布七年后，每个 SSD 仍保持 2 GB/s 的速度”。

▲ zokier 1 day ago | prev | next [-]

> 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.

AWS marketing claims otherwise:

Up to 800K random write IOPS
Up to 1 million random read IOPS
Up to 5600 MB/second of sequential writes
Up to 8000 MB/second of sequential reads

<https://aws.amazon.com/blogs/aws/new-storage-optimized-amazo...>

当然，这组宣传数据立即引发了质疑：有个别较真网友拿刚测试好的数据直接打脸：单个 SSD 是 2.7 GB/s，虽然确实比 2GB/s 好一丢丢，但是在 4202 年来也没什么好看。

知乎

Just tested a 47.52xlarge.

```

$ lsblk
NAME        MAJ:MIN RM  SIZE RO TYPE MOUNTPOINTS
loop0        7:0      0   24.9M  1 loop /snap/amazon-ssm-agent/7628
loop1        7:1      0   55.7M  1 loop /snap/core18/2812
loop2        7:2      0   63.5M  1 loop /snap/core20/2015
loop3        7:3      0  111.9M  1 loop /snap/lxd/24322
loop4        7:4      0   40.9M  1 loop /snap/snapd/20290
nvme0n1      259:0     0    8G    0 disk
├─nvme0n1p1  259:1     0   7.9G    0 part /
├─nvme0n1p14 259:2     0    4M    0 part
└─nvme0n1p15 259:3     0  106M    0 part /boot/efi
nvme2n1      259:4     0   3.4T    0 disk
nvme4n1      259:5     0   3.4T    0 disk
nvme1n1      259:6     0   3.4T    0 disk
nvme5n1      259:7     0   3.4T    0 disk
nvme7n1      259:8     0   3.4T    0 disk
nvme6n1      259:9     0   3.4T    0 disk
nvme3n1      259:10    0   3.4T    0 disk
nvme8n1      259:11    0   3.4T    0 disk

```

Since nvme0n1 is the EBS boot volume, we have 8 SSDs. And here's the read bandwidth for one of them:

```

$ sudo fio --name=bla --filename=/dev/nvme2n1 --rw=read --iodepth=128 --
ioengine=libaio --direct=1 --blocksize=16m
  bla: (g=0): rw=read, bs=(R) 16.0MiB-16.0MiB, (W) 16.0MiB-16.0MiB, (T) 16.0MiB-16.0MiB
ioengine=libaio, iodepth=128
  fio-3.28
Starting 1 process
^Cbs: 1 (f=1): [R(1)][0.5%][r=2704MiB/s][r=169 IOPS][eta 20m:17s]

```

So we should have a total bandwidth of $2.7 \times 8 = 21$ GB/s. Not that great for 2024.

知乎 @小猿姐

还有一类反对者说，一方面是对高速存储没有太多需求，另一方面，限制 I/O 速度可以允许虚拟化层在 I/O 栈上实现一些功能代码。而这些操作在“原始”了，无法实现。

▲ bombcar 1 day ago | prev | next [-]

I think the obvious answer is there's not much demand, and keeping it "low" allows trickery and funny business with the virtualization layer (think: SAN, etc) that you can't do with "raw hardware speed".

还有人觉得这些速度上的差异对普通使用者没什么区别，文章是从 prosumer ("Prosumer" 是 "professional" 和 "consumer" 两个词的组合，用来描述专业人士) 论。

▲ 0cf8612b2e1e 1 day ago | prev | next [-]

Serious question, for a consumer does it make any sense to compare SSD benchmarks? I assume the best and worst models give a user an identical experience in 99% of cases, and it is only prosumer activities (video? sustained writes?) which would differentiate them.

[reply](#)

▲ wmf 1 day ago | parent | next [-]

Yeah, that's pretty much the case. Cheap SSDs provide good enough performance for desktop use.

[reply](#)

知乎 @小猿姐

结尾

知乎

针对这个问题，你怎么看呢？欢迎在留言区跟我们互动，也欢迎扫码入群聊。

参考文章链接

- [databasearchitects.blogspot.com...](#)
- [news.ycombinator.com/it...](#)

End

KubeBlocks 已发布 v0.8.0 (KubeBlocks v0.8.0 发布! Component API 让数据库引擎组装更简单!) ! KubeBlocks v0.8.0 推出了 Component API, 让数加简单. Addon 机制也有了重大改进, 数据库引擎的 helm chart 从 KubeBlocks repo 中拆分出去, 从此数据库引擎或者版本的变动已与 KubeBlocks 发版解耦. Pika、Clickhouse、OceanBase、MySQL、PostgreSQL、Redis 等均有功能更新, 快来试试看!

小猿姐诚邀各位体验 KubeBlocks, 也欢迎您成为产品的使用者和项目的贡献者. 跟我们一起构建云原生数据基础设施吧!

官网: [The control plane for your cloud-native data infrastructure | KubeBlocks](#)

GitHub: [GitHub - apecloud/kubeblocks: KubeBlocks is an open-source control plane that runs and manages databases, message queues and other on K8s](#)

Get started: [Try out KubeBlocks in 5 minutes on laptop | KubeBlocks](#)

关注小猿姐，一起学习更多云原生技术干货。

发布于 2024-02-23 13:57 · IP 属地浙江

云计算 SSD



发布一条带图评论吧



还没有评论，发表第一个评论吧

推荐阅读

知乎



SSD硬盘多久才能写死？长江存储科普TBW寿命：能用80年

快科技



4T SSD只需2000大洋？YES！

钱乎



SSD选购解读
有缓存就不掉速？固态硬盘SSD缓存向选购解读

浅色月

发表于浅色月硬盘...



PCIe NVMe快？它的优势

养猫的哈士...