Fast, Faster, Fastest

Object Storage, Cloud Block Storage, and NVMe SSD in Analytic Databases

Robert Hodges &
Diego Nieto

Let's make some introductions

Robert Hodges

Database geek with 30+ years on DBMS systems. Day job: Altinity CEO

Diego Nieto

Software engineer on Altinity with interests in Databases and Python

ClickHouse support and services including Altinity.Cloud
Authors of Altinity Kubernetes Operator for ClickHouse
and other open source projects

Questions for our talk today

What are the main forms of cloud storage?

How do they work and how can we measure performance?

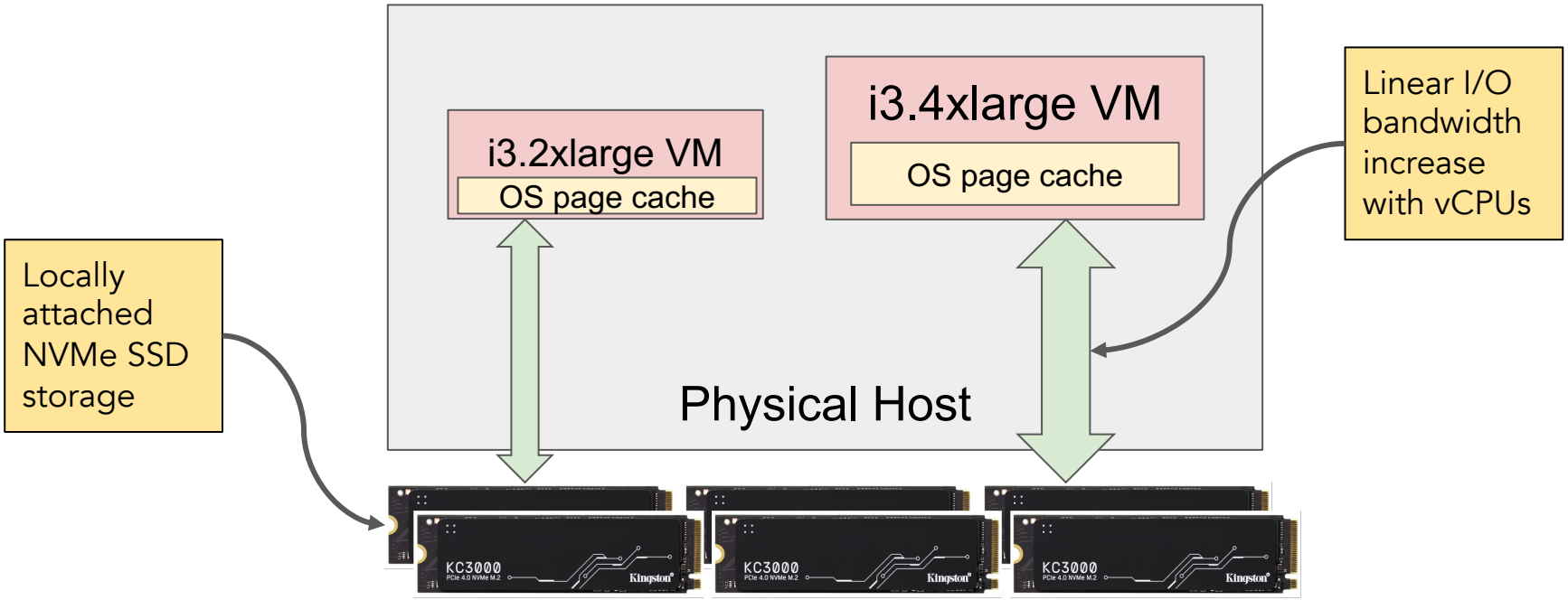
What is an analytic database?

How does storage affect analytic database performance?

> It’s the not the destination, It's the journey.
>
> — Ralph Waldo Emerson, Self-Reliance

Exploring cloud storage performance

How does NVMe SSD storage work in clouds?



Sounds fast! How do I measure it?

kioperf

A simple program to measure disk & object storage speed

```
./kioperf disk --operation=write \
  --size 512 --threads=4 \
  --iterations=50 --files=50 --fsync \
  --dir-path /data/test --csv
sudo sync
sudo echo 3 > /proc/sys/vm/drop_caches
./kioperf disk --operation=read \
  --threads=4 --iterations=500 \
  --files=50 --dir-path /data/test
```

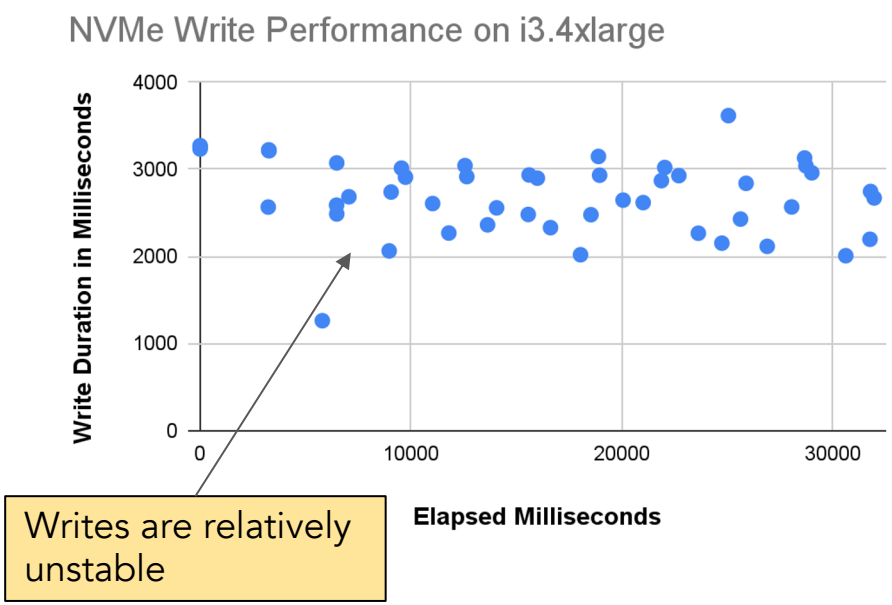
WRITE TEST

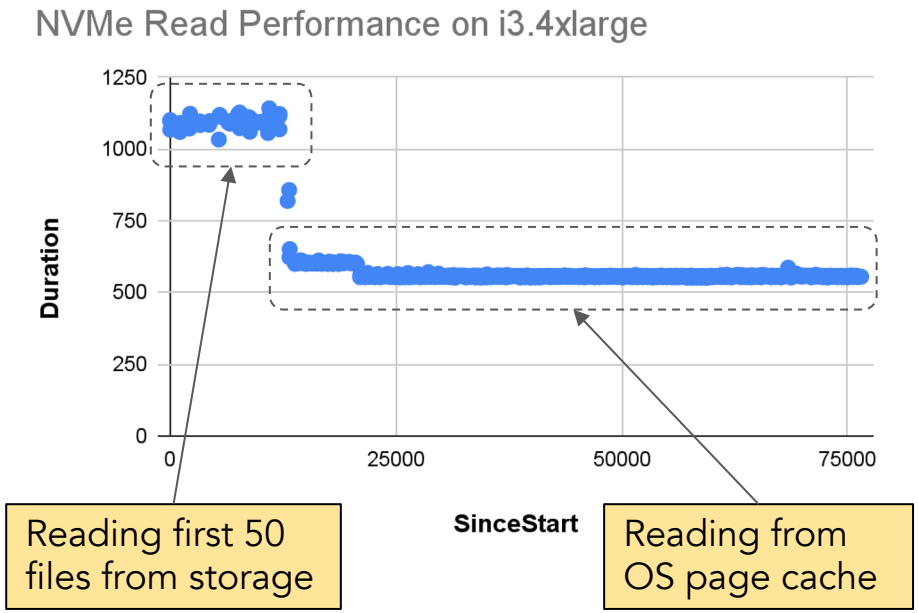
Write 50 512MiB files using 4 threads

READ TEST

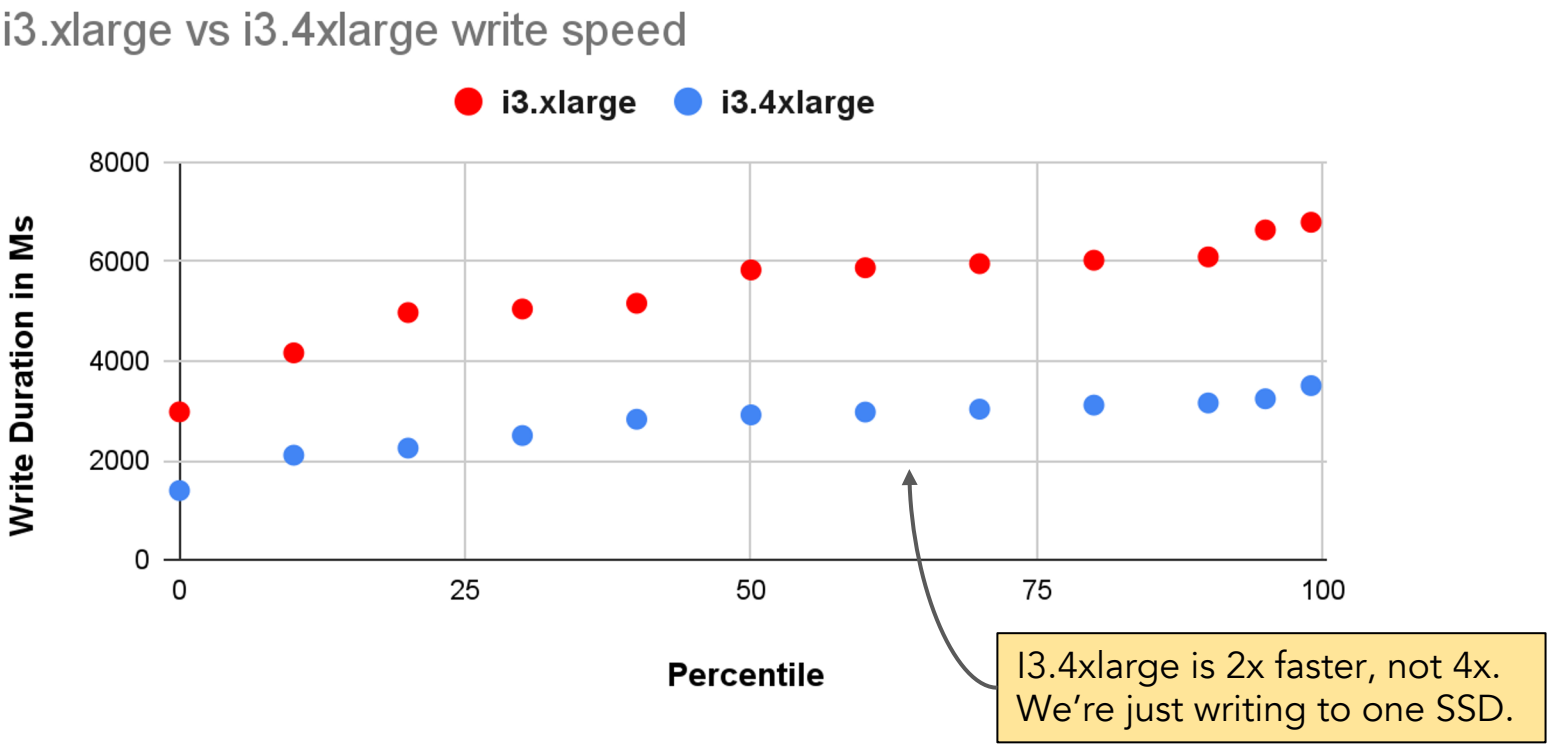
Read 50 512MiB files 500 times using 4 threads

NVMe SSD performance on AWS i3.4xlarge instances

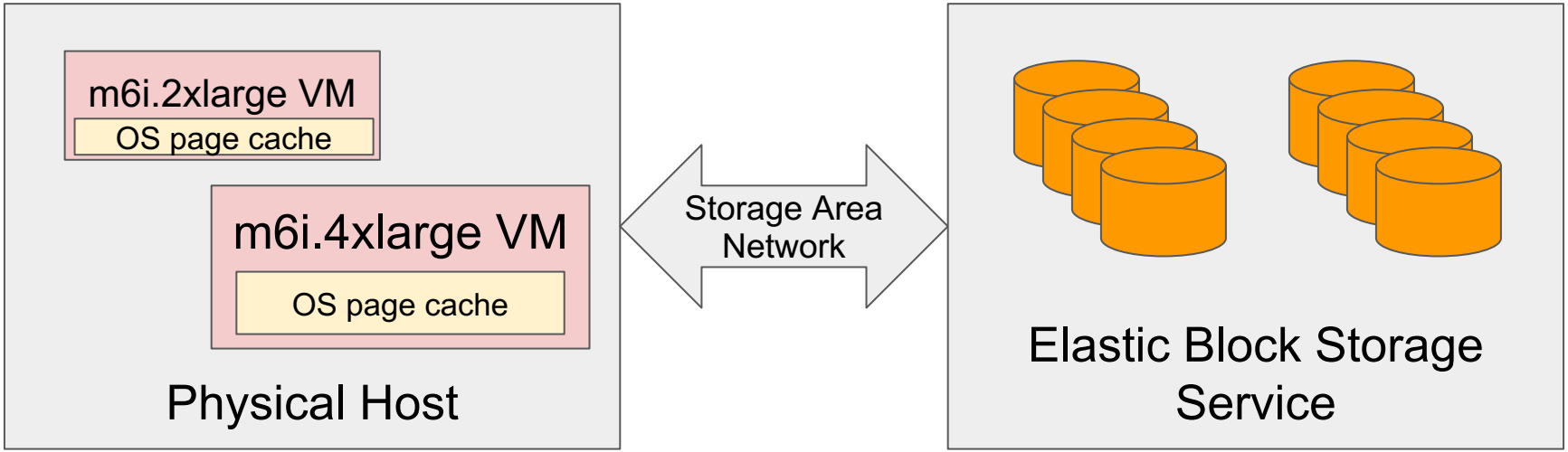




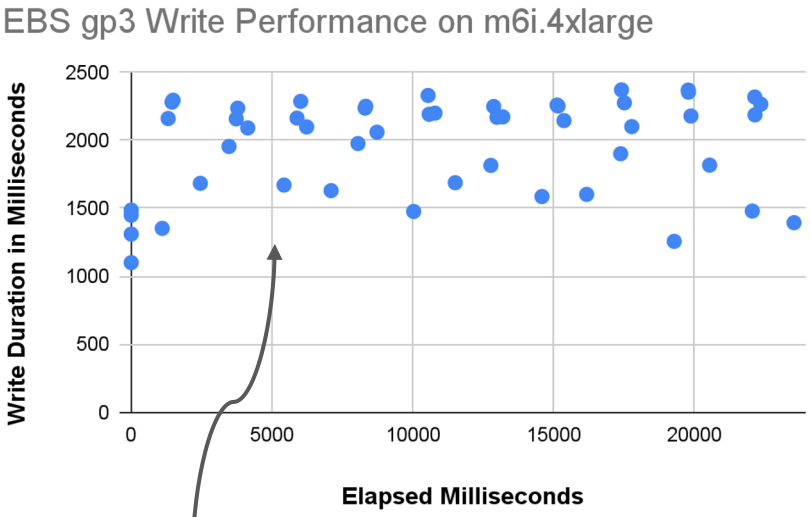
Smaller VM types get less storage bandwidth!



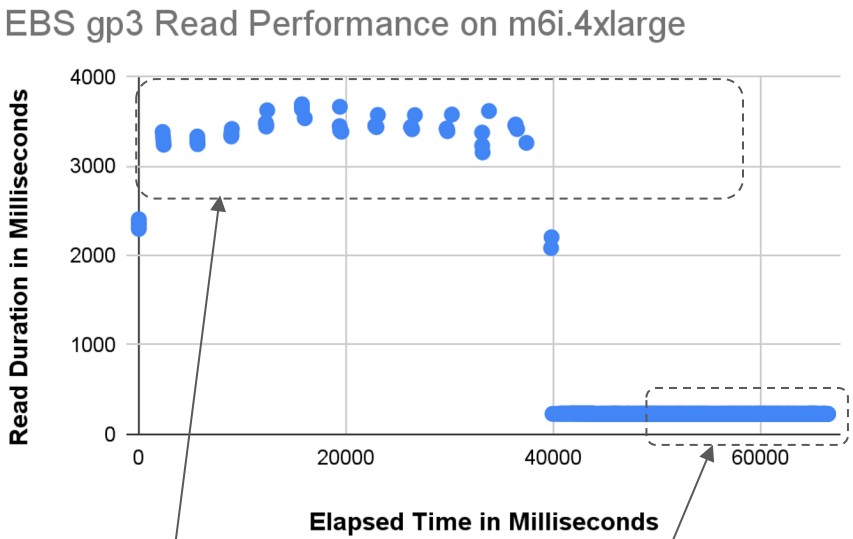
How does cloud block storage work?



Let's look at EBS performance!



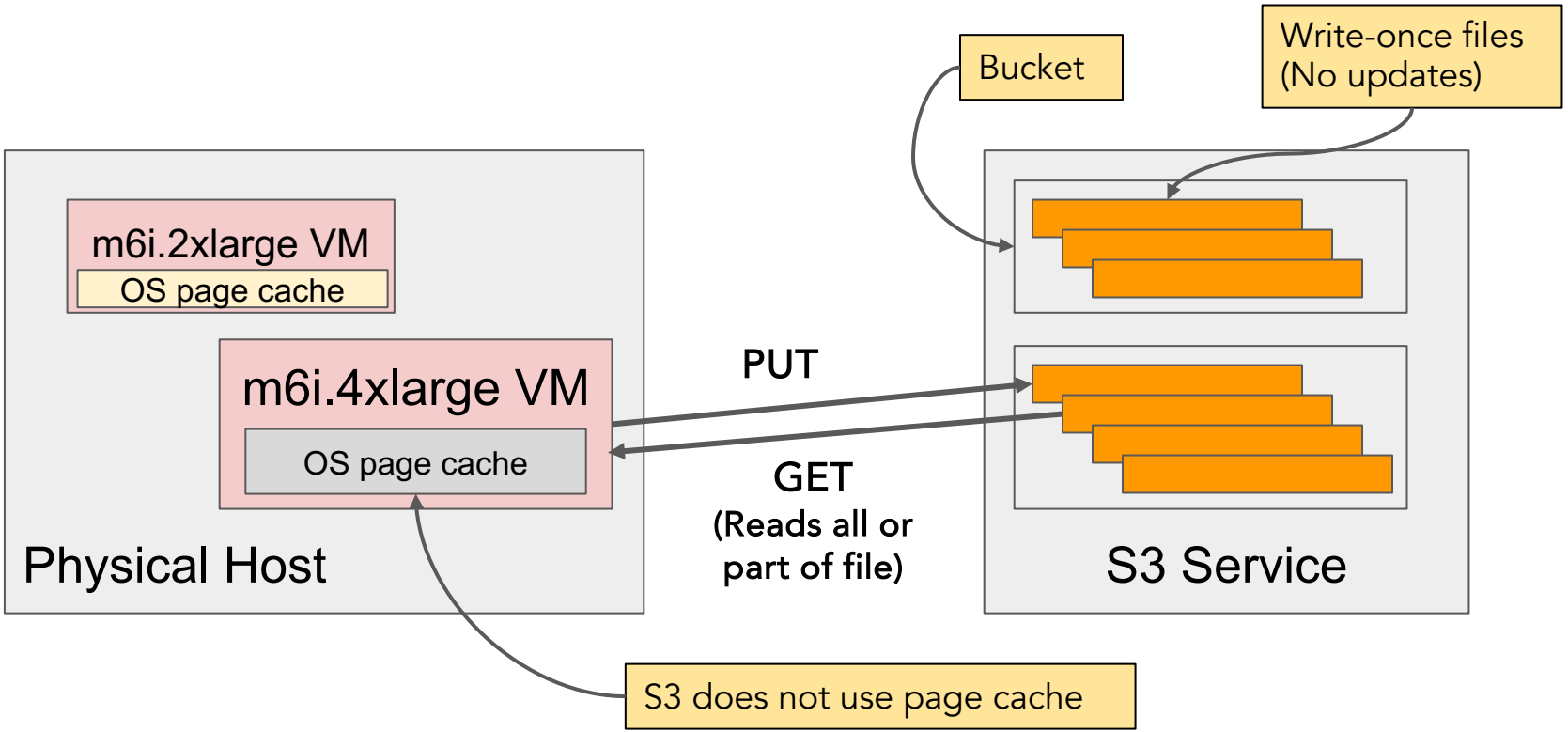
Writes compare well to NVMe SSD



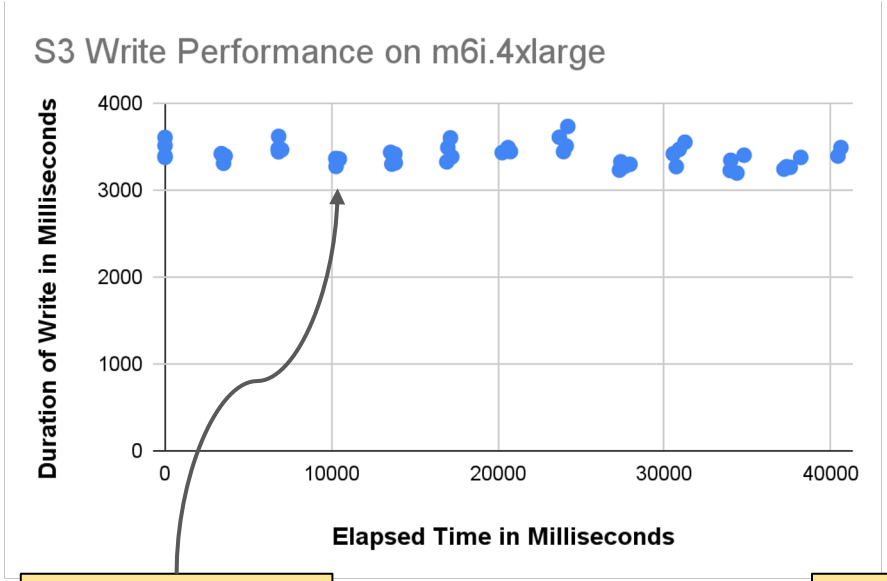
Read from storage slower than NVMe SSD

Page cache reads are faster though!

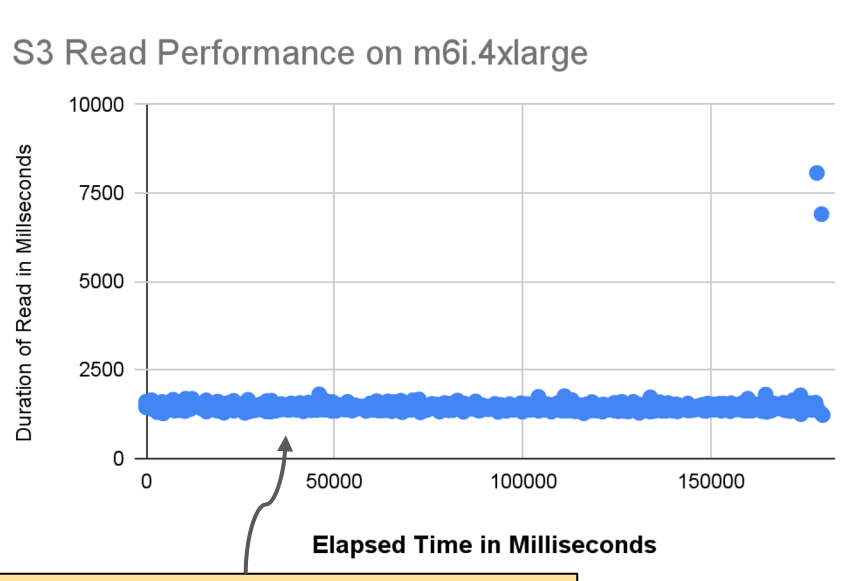
S3 object storage works on files, not blocks



How does S3 perform??

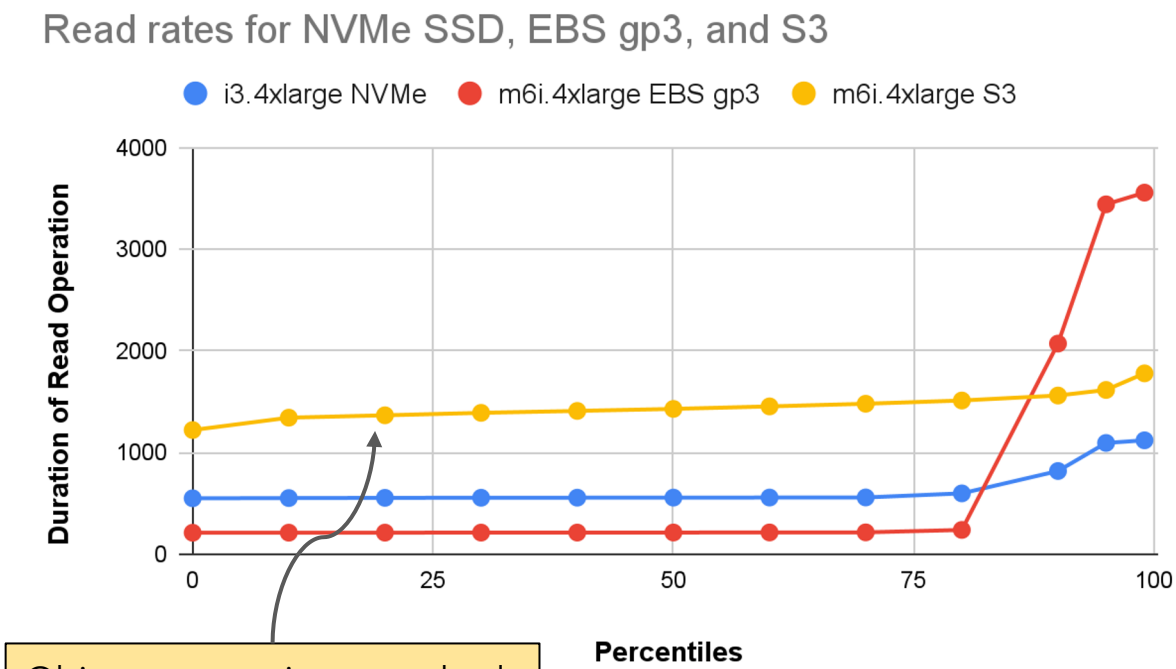


Slower than EBS writes



Fast and stable from storage but much slower than OS page cache

Comparison of reads for NVMe/EBS/S3



Block storage 90% percentile and above reads directly from storage so it's slow

Object storage is not cached locally. It's more consistent

Analytic databases and storage

Meet ClickHouse. It's a real-time analytic database

Understands SQL

Runs on bare metal to cloud

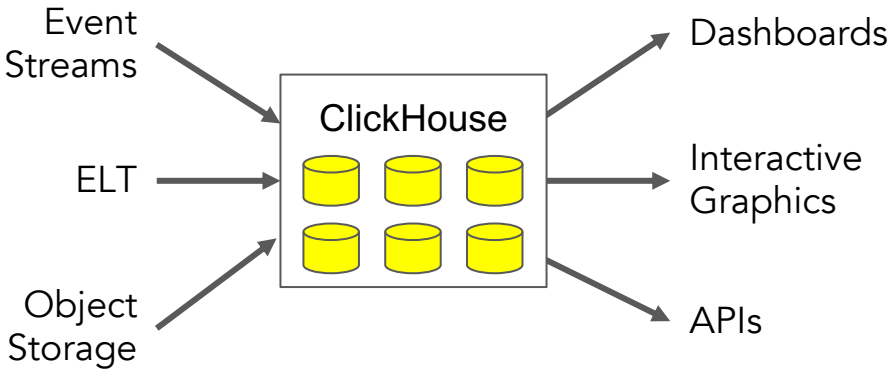
Shared nothing architecture

Stores data in columns

Parallel and vectorized execution

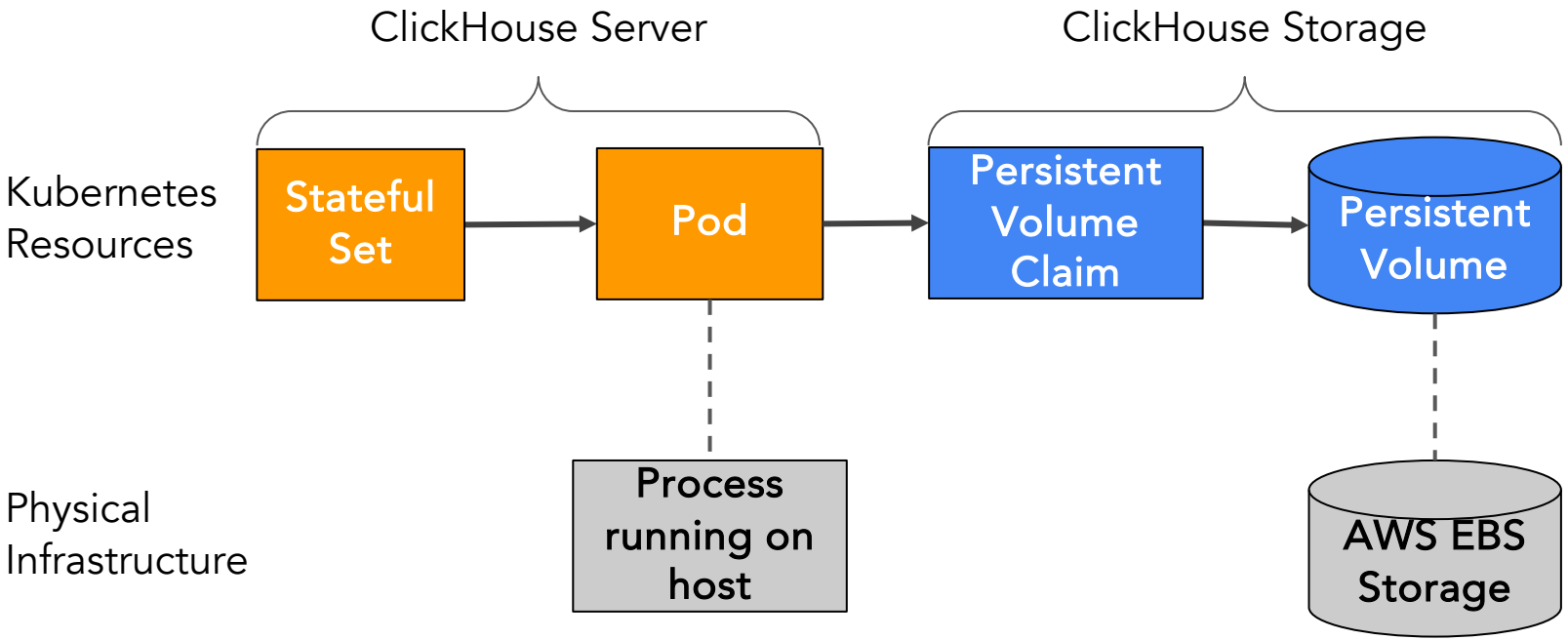
Scales to many petabytes

Is Open source (Apache 2.0)

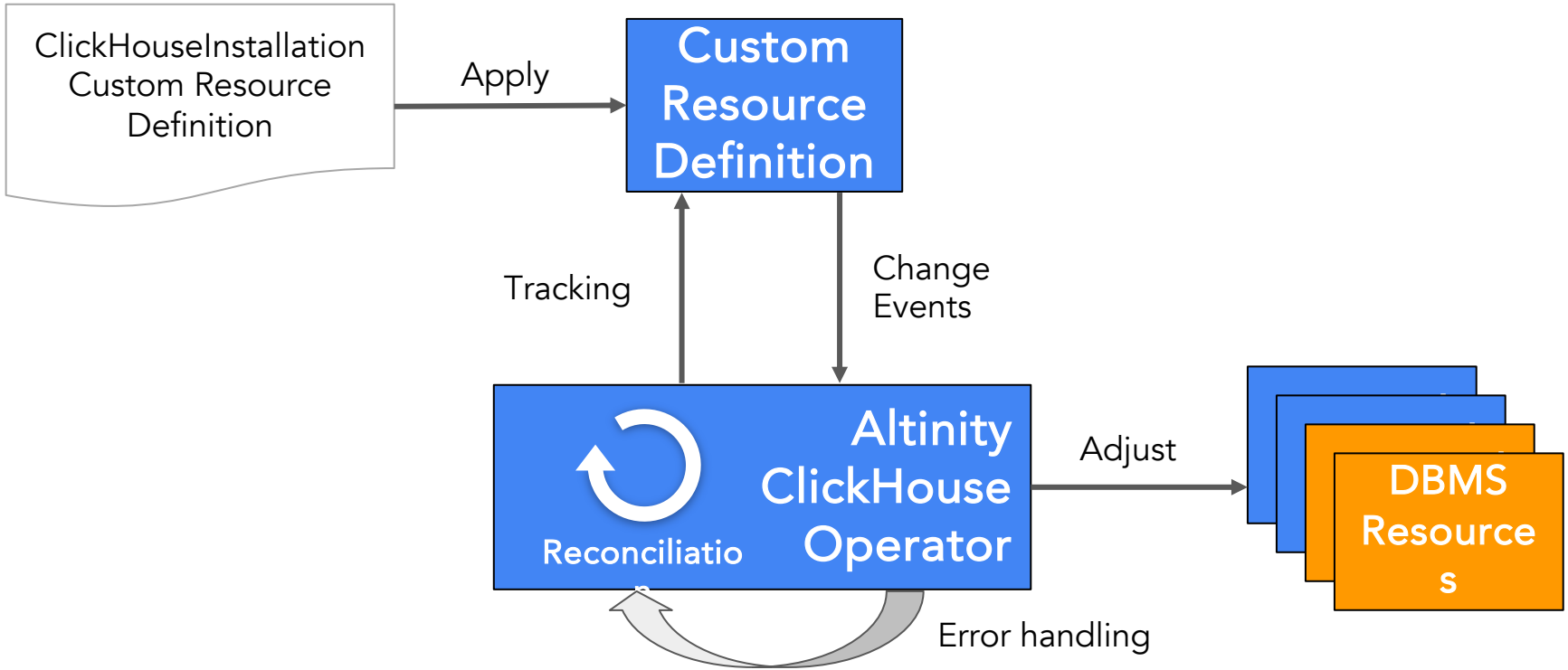


It's the core engine for low-latency analytics

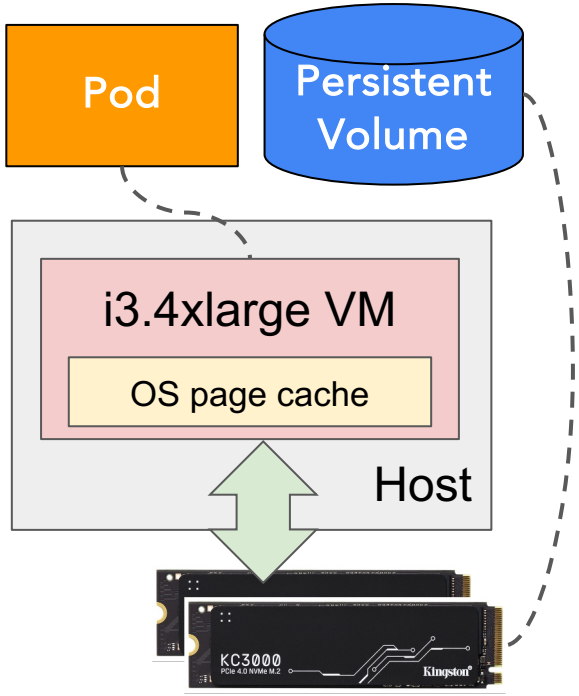
We like to run ClickHouse on Kubernetes



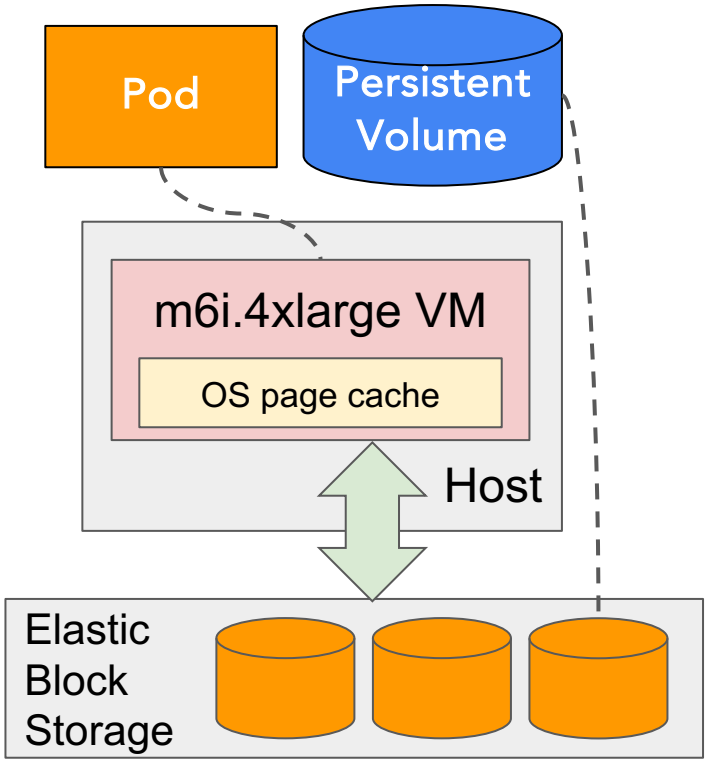
The Altinity ClickHouse Operator sets it up



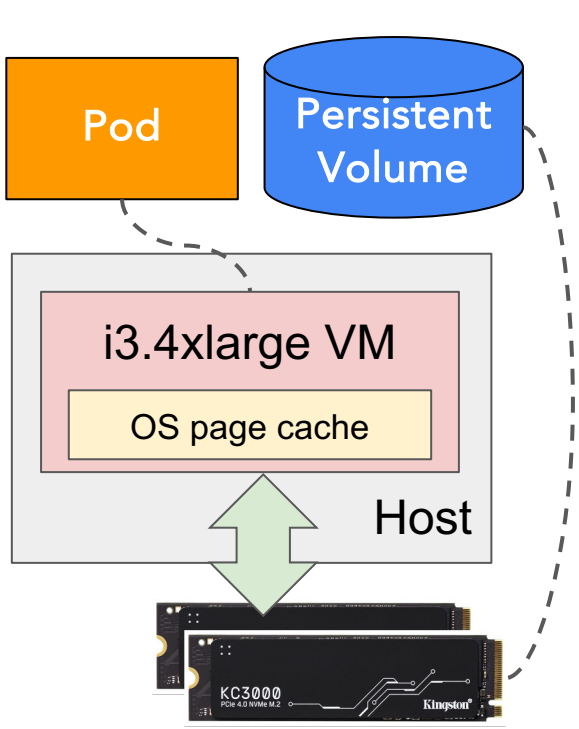
So really what we have is the following



OR

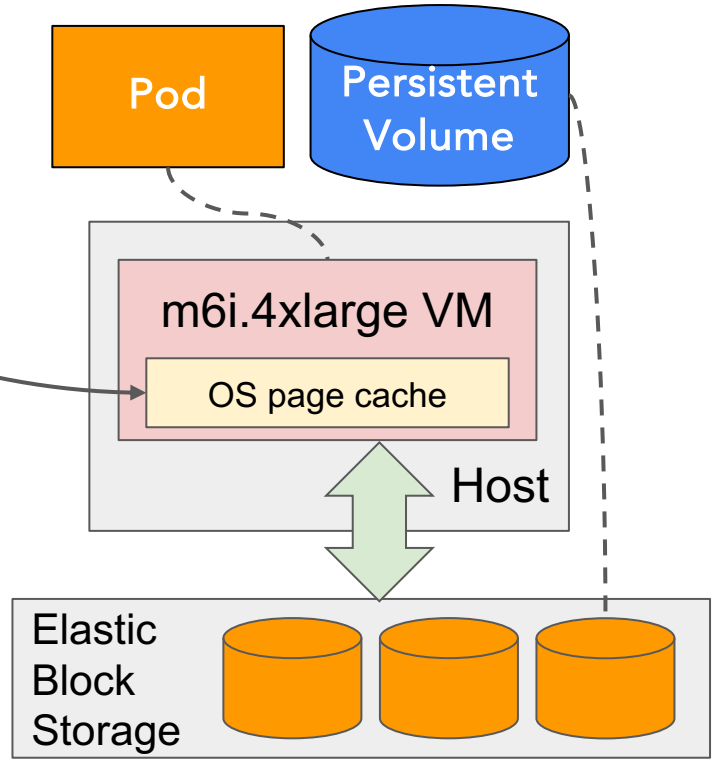


So really what we have is the following



CAVEAT!! Containers don't own the page cache

OR



Let's measure it!

Introducing **ClickBench**

A Benchmark for Analytical Databases
By Alexei Milovidov

- Realistic e-commerce data
- ~100M rows in on table (15 GiB on disk)
- 43 queries run 3 times each
- Sets up and runs in 20 minutes

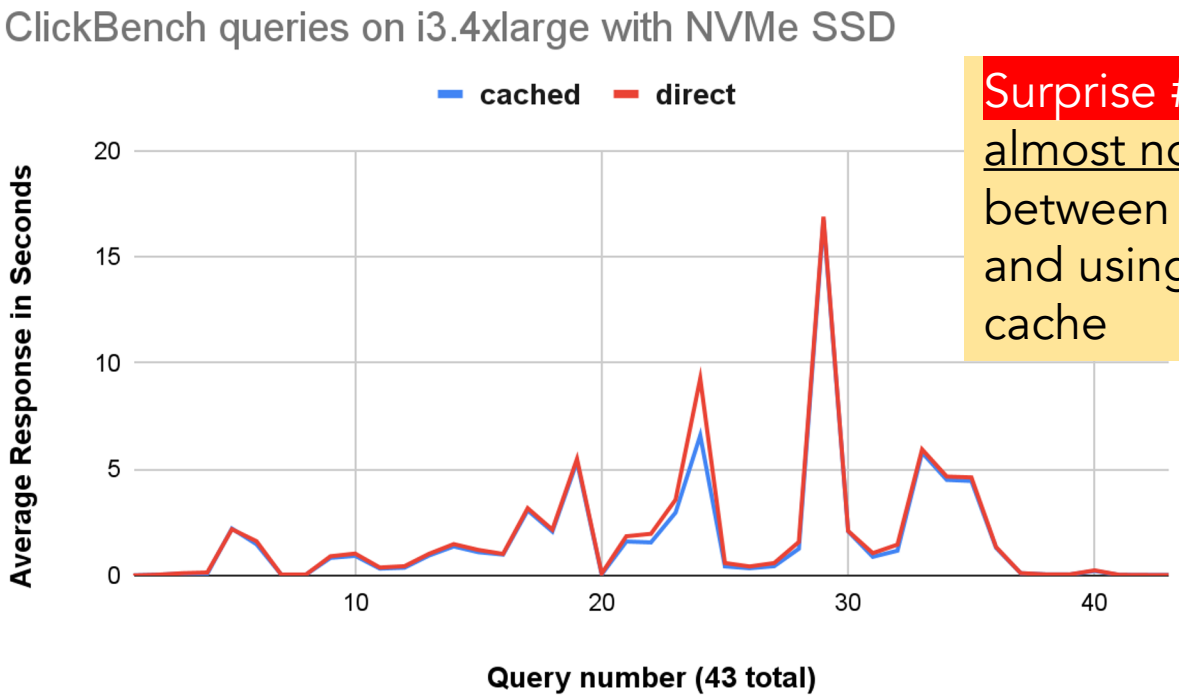
https://github.com/ClickHouse/ClickBench

We adapt ClickBench to Kubernetes

ClickBench run 1: Force direct I/O to check storage speed

ClickBench run 2: Allow page cache

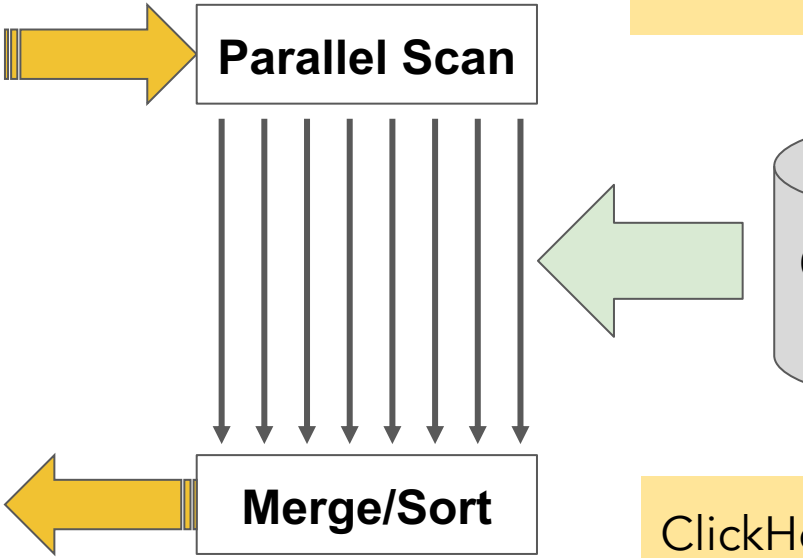
Test results for ClickHouse on NVMe SSD



Surprise #1: There's almost no difference between direct I/O and using the page cache

Why isn't there a visible difference? Let's investigate!

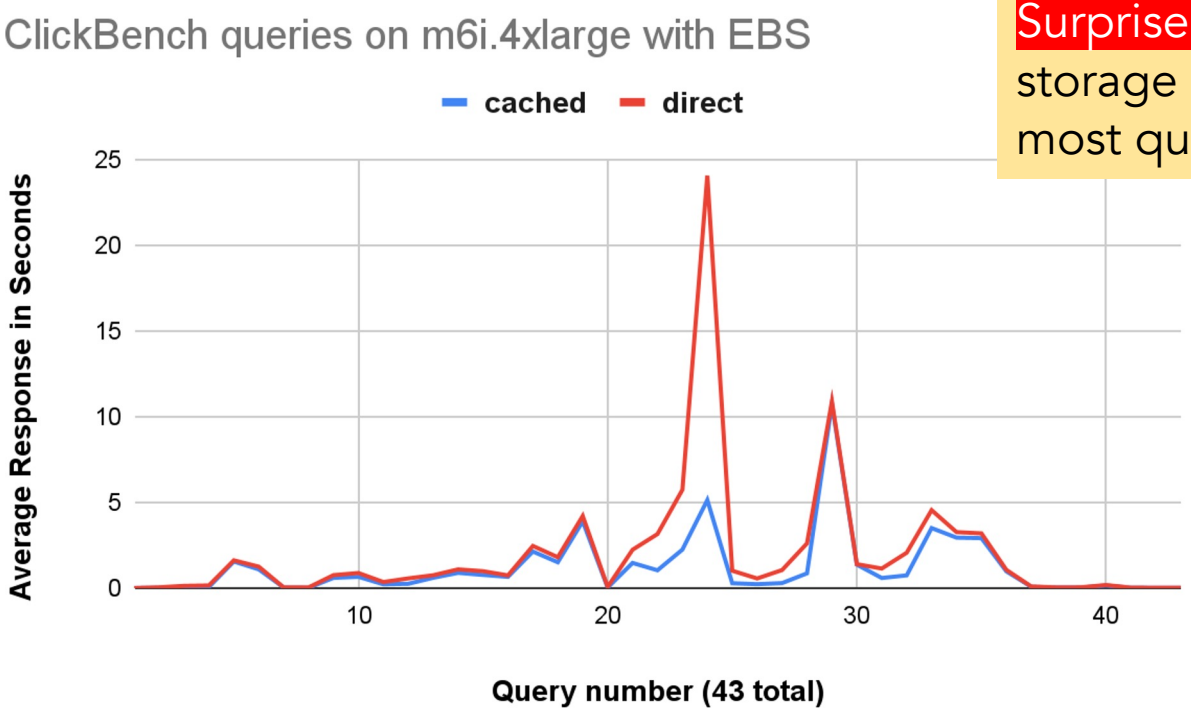
Answer:
ClickBench queries are dominated by compute!



ClickHouse minimizes I/O

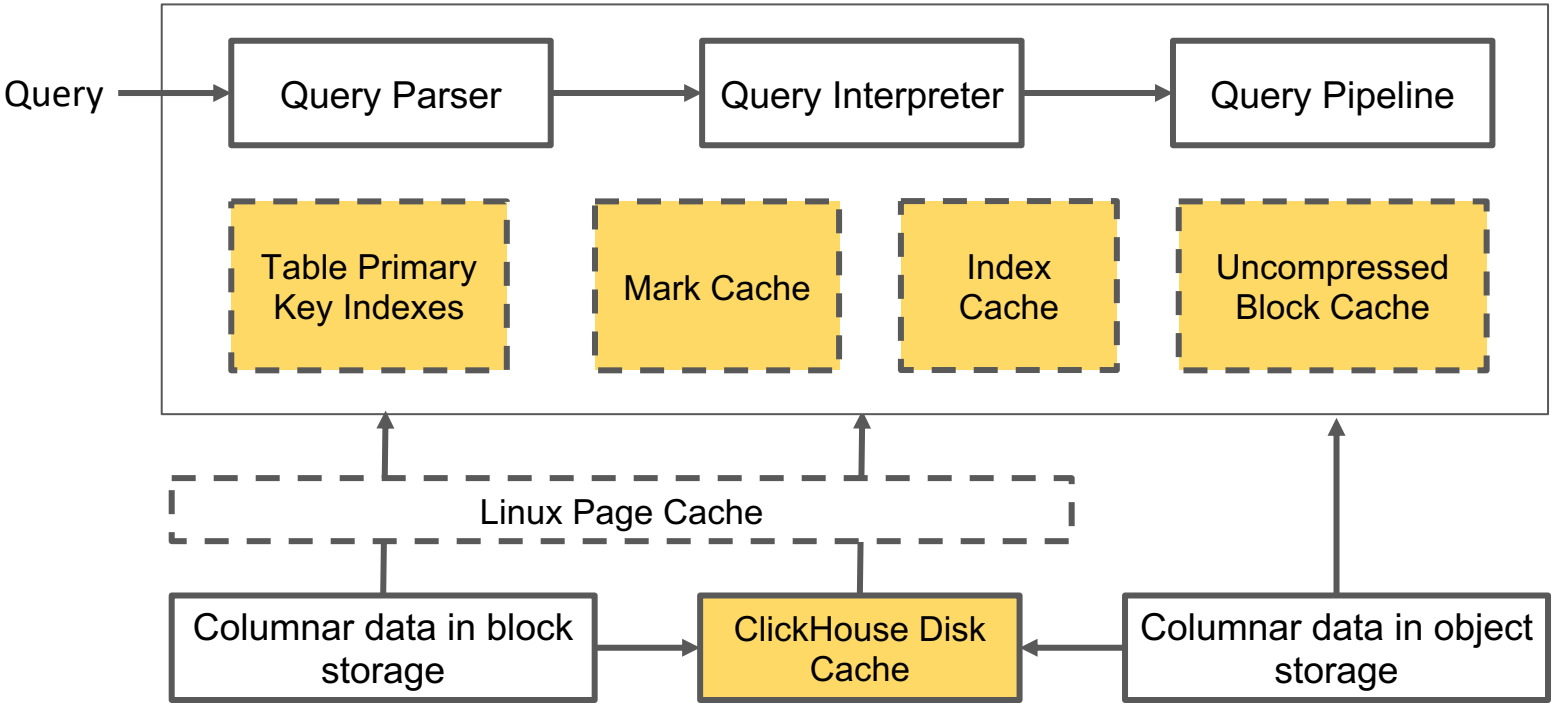
ClickHouse reads in parallel

Test results for ClickHouse on EBS

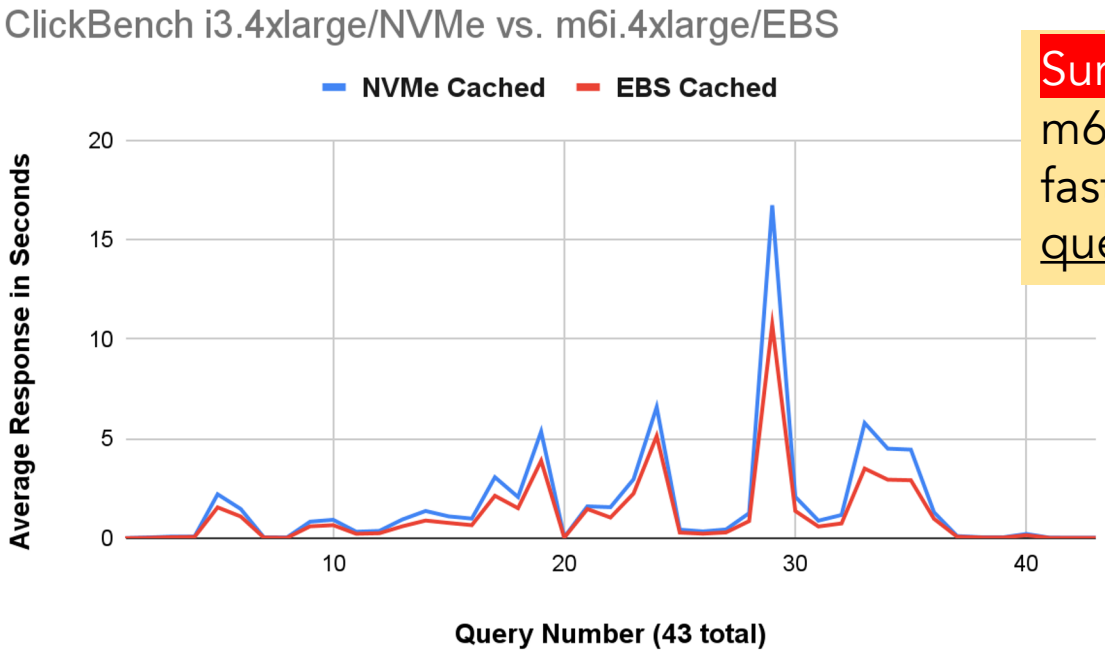


Surprise #2: EBS storage is fast on most queries, too

ClickHouse has caches to speed up direct I/O

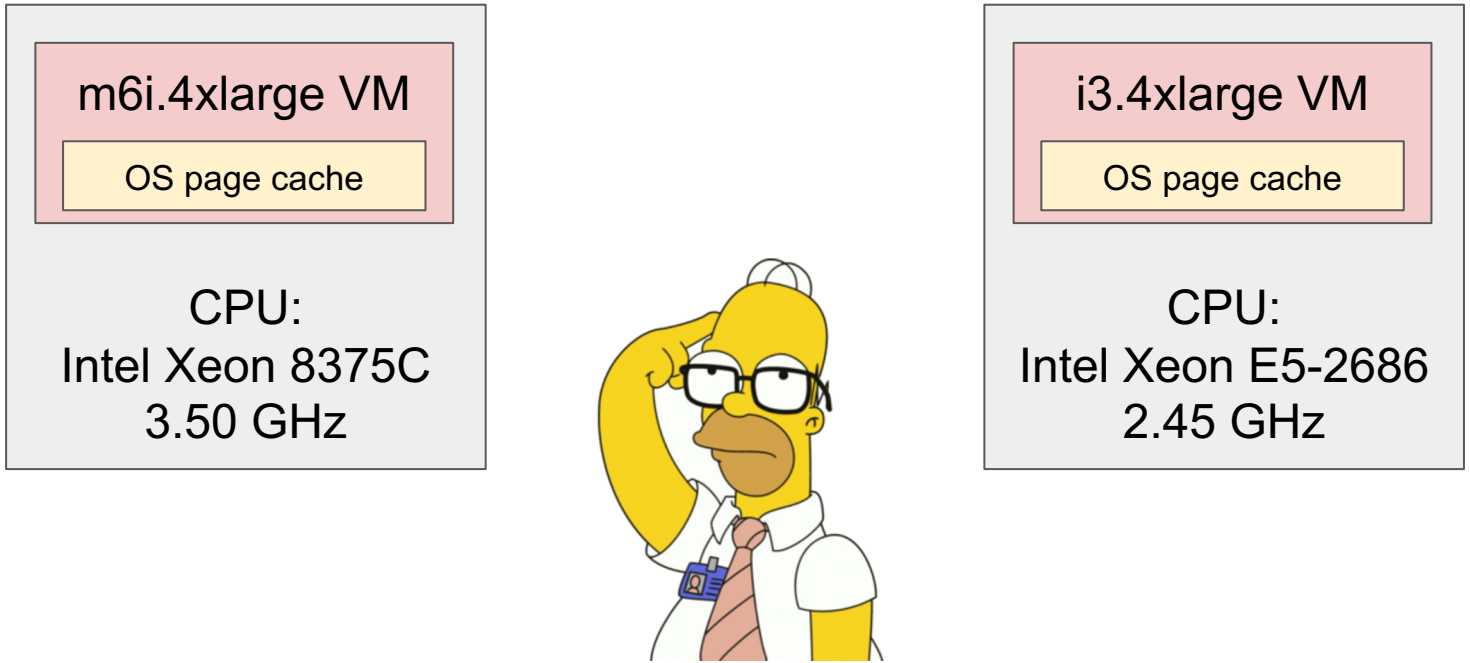


Comparing cached query response for NVMe and EBS

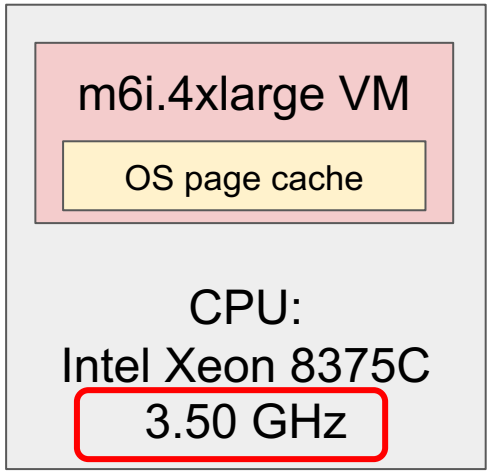


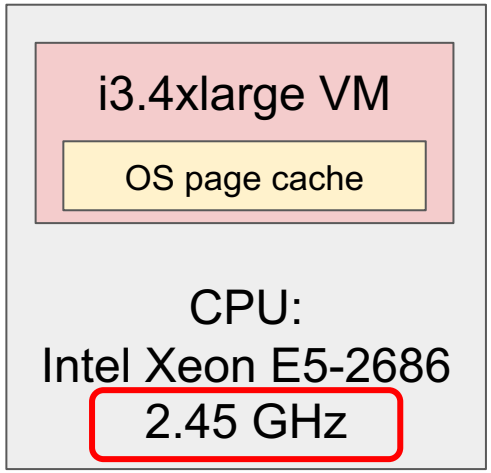
Surprise #3: m6i.4xlarge is faster in every query

Let's investigate why the EBS host is faster



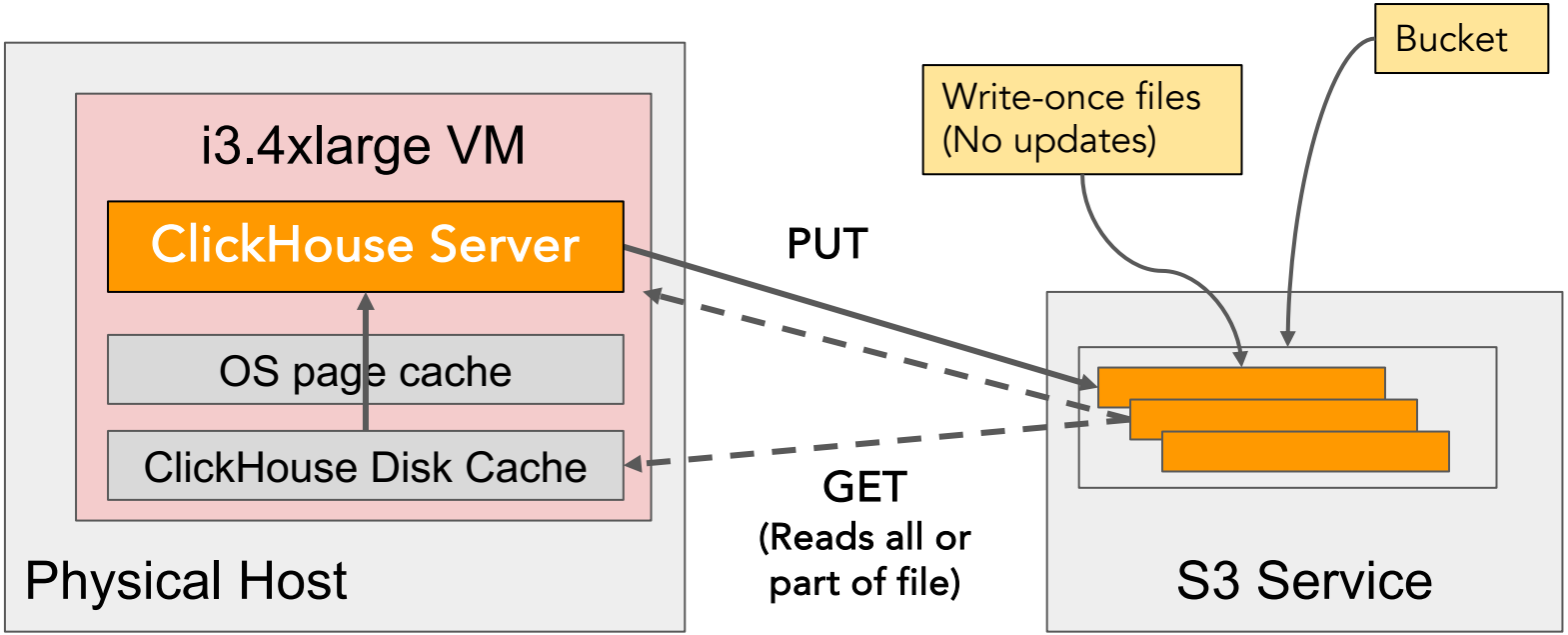
Let's investigate why the EBS host is faster



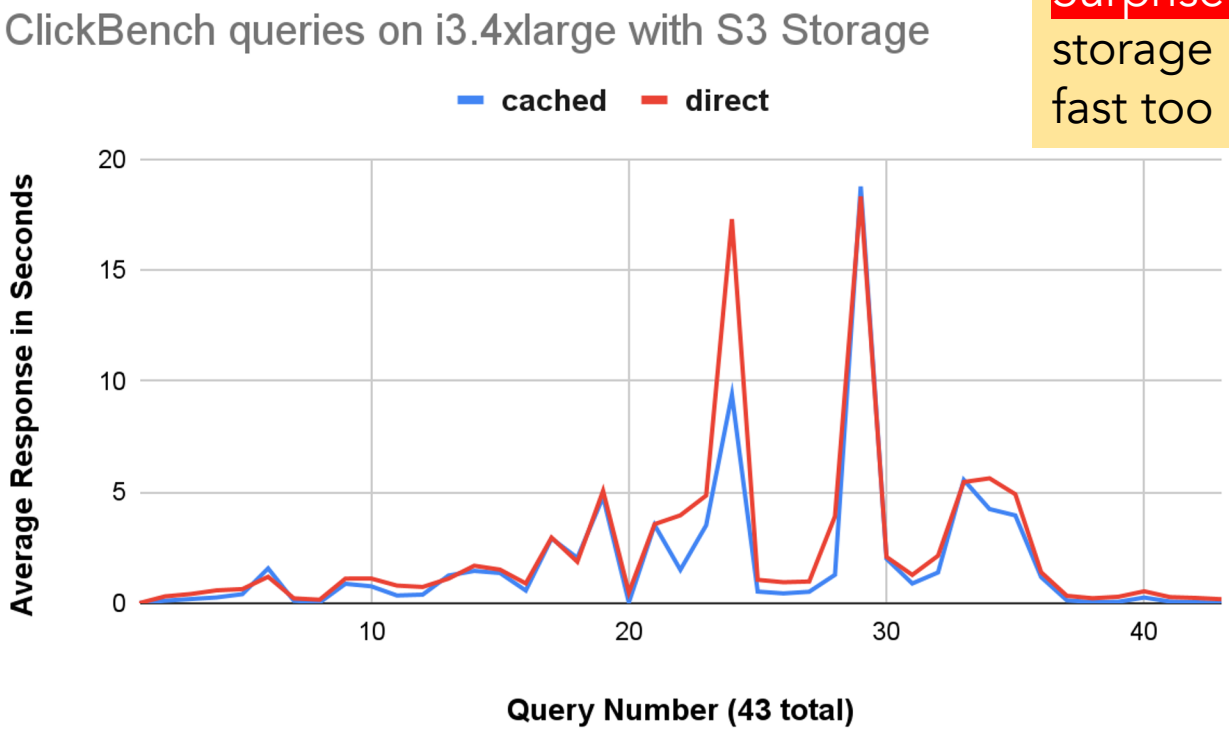


D'Oh! 39% faster!

Here's how object storage works in Altinity.Cloud

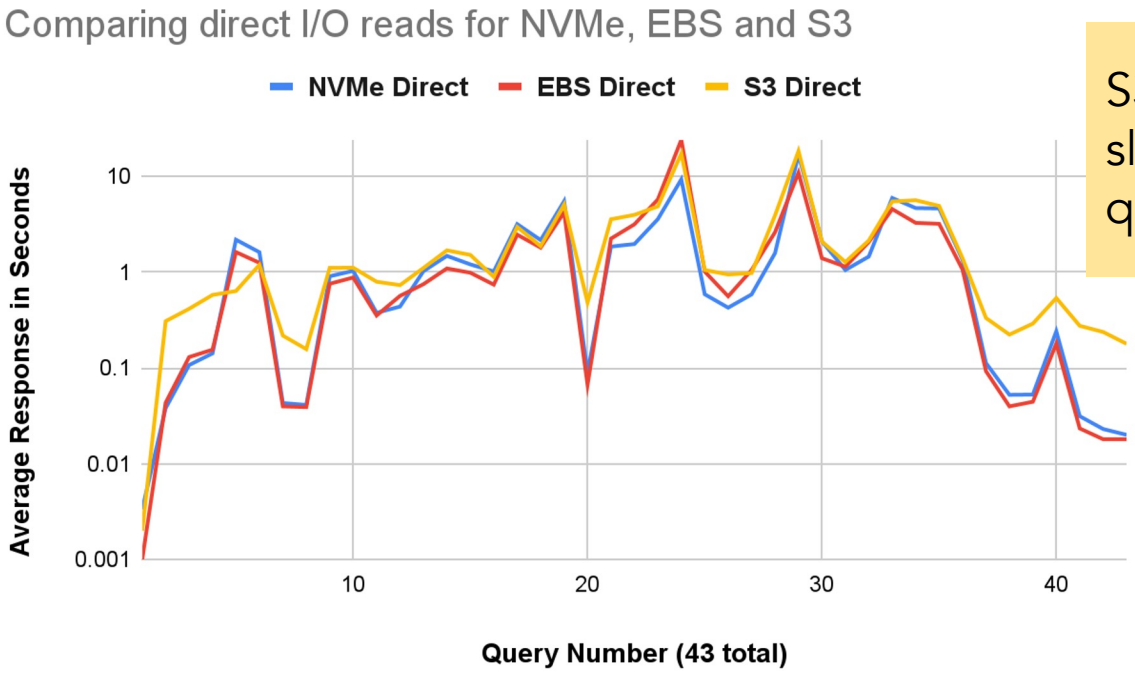


Test results for ClickHouse on S3



Surprise #4: Object storage is pretty fast too

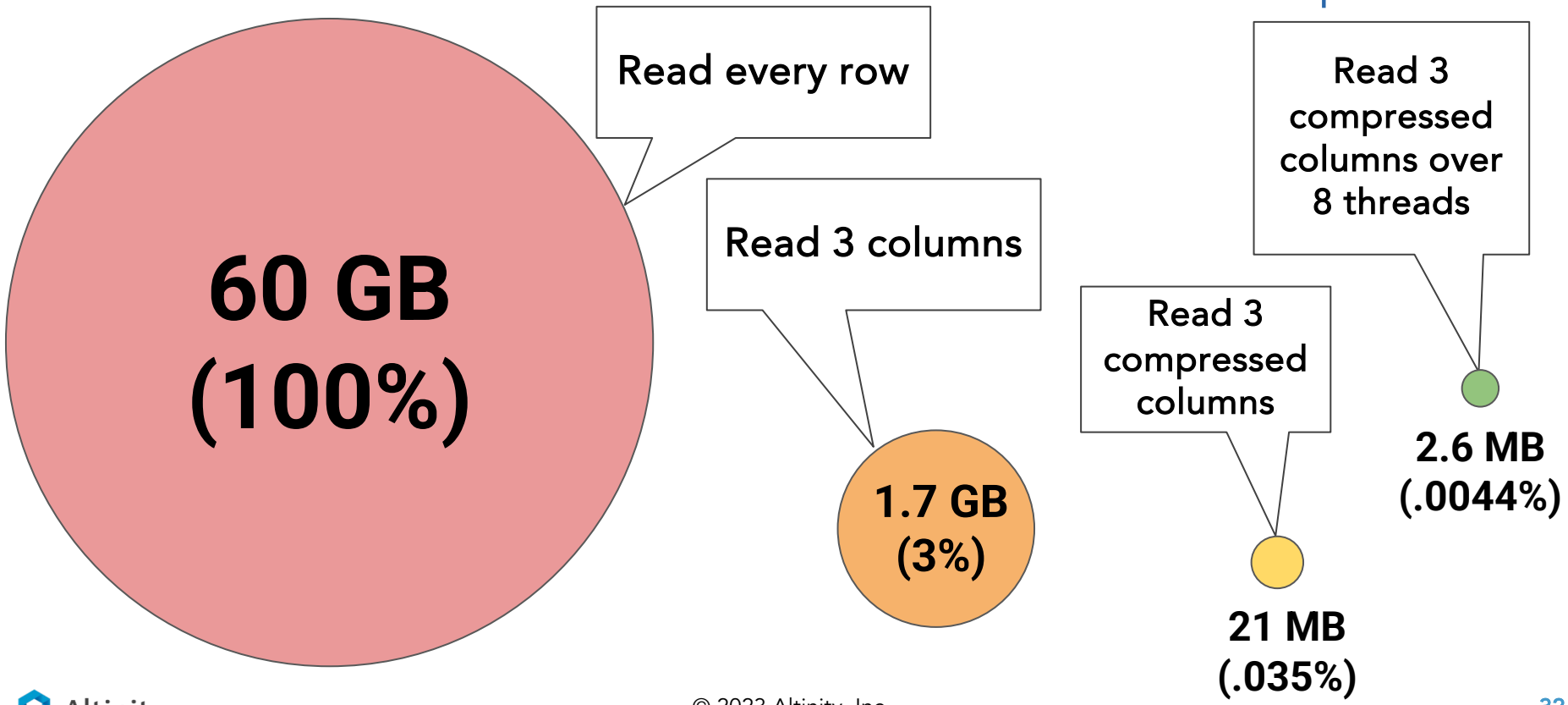
Comparing direct query response for NVMe, EBS, and S3



S3 reads are slower on small queries

Conclusion

In the best cases, ClickHouse is insensitive to I/O speed



Not everything is a best case

Server start-up on large systems	→	NVMe SSD can be 100x faster than EBS
Table scans on billions of rows	→	Shard your data or it will be slooooow
Guaranteed low latency apps	→	If you want answers fast put data in block storage

So what did we learn today?

- NVMe SSD is not necessarily the fastest game in town
- VM size affects I/O bandwidth in AWS
- Analytic databases like ClickHouse make I/O small
- Compute can dominate on small datasets
- Know your caches
- Know who shares what
- A slow CPU makes “storage” access slow
- Object storage has slow time to first block

Databases are complicated.

Don’t trust anybody.

Test it yourself.

More information!

- Testing tools
  - ClickBench - https://github.com/ClickHouse/ClickBench
  - FIO - https://fio.readthedocs.io
  - Sysbench - https://github.com/akopytov/sysbench
  - "ioperf" - https://github.com/hodgesrm/ioperf
- ClickHouse Documentation
  - https://clickhouse.com/docs/en/intro
- Altinity YouTube channel
  - A Day in the Life of a ClickHouse Query
- Altinity Blog – https://altinity.com/blog
- Altinity Operator for ClickHouse
  - https://github.com/Altinity/clickhouse-operator

Thank you! Questions?

Altinity.Cloud

Altinity Support

Altinity Stable Builds

https://altinity.com