



Transactions per Dollar

PostgreSQL in virtual and bare metal clouds

Oskari Saarenmaa

PostgresConf US 2019 - New York

Agenda

1. Introduction
2. Different clouds and cloud instances
3. Measuring the right things
4. What affects benchmark performance
5. Benchmark methodology and setup
6. Results
7. Q & A

This presentation was created by Aiven Ltd - <https://aiven.io>.
All content is owned by Aiven or used with owner's permission.



Speaker



@OskariSaarenmaa

- CEO, co-founder @ Aiven, a cloud company
- Previously: database consultant, software architect
- PostgreSQL user and supporter since 1999 (rel 6.4)



Aiven

- *Your data cloud*
- Based in Helsinki and Boston
- 8 data engines now available in 6 clouds and 80 regions, virtual and bare metal instances
- Launched a fully-managed PostgreSQL service in 2016
- First to offer the latest PostgreSQL features in a managed service



Google Cloud Platform



Clouds and cloud instances

The cloud provides quick, easy and sometimes cost-efficient access to infrastructure.

- New, interesting hardware available on-demand without long lead times
- Someone else covers the capital costs, relatively easy to try out different options
- Options vary from someone installing hardware based on requirements to immediate access to virtual machines, and recently to bare metal hardware over an API
- It's not just compute instances: options also exist for storage and network
 - Most clouds provide scalable storage over the network
 - Some also provide local NVMe storage devices, often in limited fashion

Cloud instances

Compute instances / nodes / servers set many constraints

- Most commonly, instances are virtual machines running on shared hardware
 - e.g. AWS EC2, Azure virtual machines, GCP instances, DigitalOcean droplets
- You share compute resources with other tenants, some clouds overprovision hardware so other users may have a dramatic impact on performance
- Nowadays also bare metal hardware available over an API
 - e.g. Packet.com and AWS *.metal instances
- You get billed for what you use, minute and byte based pricing on compute, network, storage (capacity and IOPS); prices per unit are tiny, but they add up
 - Discounts available through longer-term commitments

Measuring the right things

“Interesting”

- CPU count and model
- RAM type and size
- RAID array characteristics
- Kernel versions
- (microbenchmark looping
`select 1`)

Important

- Can it fit my production data set
 - Do I have an upgrade path?
- Can it run enough transactions to handle my sustained and peak loads?
- How much does it actually cost?
 - In terms of infrastructure and operations

The interesting things: what's PG performance made of

Hardware: Virtualization, CPU, storage IO, network

Software: tuned for the workload

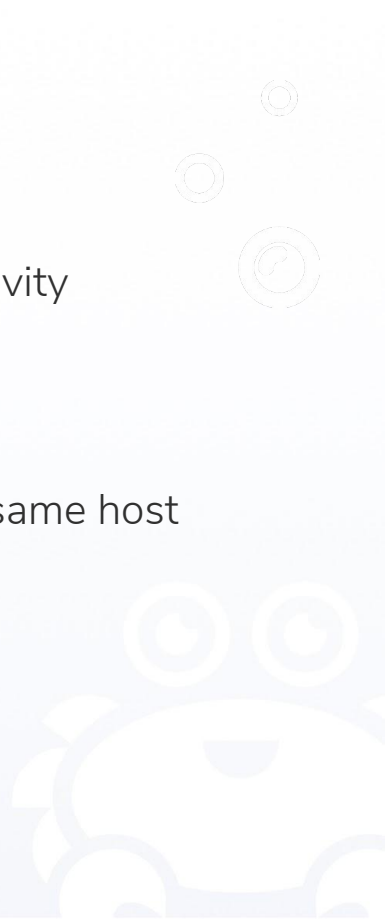
Network: application ↔ db connectivity and client ↔ application connectivity

Virtualization: adds overhead (but less than it used to)

Storage: the closer it is to the compute hardware the faster it goes

Other: data encryption, auxiliary processes (e.g. backup systems) on the same host

(and most importantly, your queries and schema)



The important things: what's PG performance made of

Capacity: Resources needed to host the workload

- In the past this meant estimating data set sizes and load and provisioning enough hardware ahead of production deployment
- Cloud makes this somewhat easier - faster to get new resources and usually simpler to scale up

Concurrency: PG doesn't scale very well to a large amount of concurrent clients

- Consider using a connection pooler to limit concurrency

Cloud: How many other tenants are competing for the same physical resources?

Cost: Usually, but not always, throwing more money at the problem makes it go away

Data gravity is real

- Cloud makes it easy to try out new systems
- It's often difficult to move production workload to a different cloud or region
 - Once you use a particular cloud, it's easier to deploy more resources there than to use an alternative provider for the next little thing
- Migrating any workload may be very expensive
 - Ingress (to the cloud) is usually free
 - Egress (from the cloud) can be very costly (\$0.15/GB for 1 PB of data adds up)
- Choose carefully, keep your options open

Benchmark setup

1. Provision a benchmarking host in the target cloud
 - PGBench from PostgreSQL 10.7 / Linux 4.20.11
 - 8 CPU threads
2. Provision a DB instance from the same cloud
 - 120 to 480 GB RAM per instance
 - PostgreSQL 11.2 / Linux 4.20.11
 - Local and network backed storage
 - **Encrypted storage, WAL-archiving enabled**
3. Initialize with a large dataset
 - Two data set sizes, ~450 GB and ~900 GB
4. PGBench 1 hour TPC-B benchmark

(repeat the test several times and pick the best result)



Benchmarks: 450 GB dataset

- Instances with 120 to 240 GB RAM, 16 to 32 vCPUs
- PGBench with 100 concurrent clients
- Different storage options including network and local storage devices
- Six different clouds, including virtual and bare metal instances



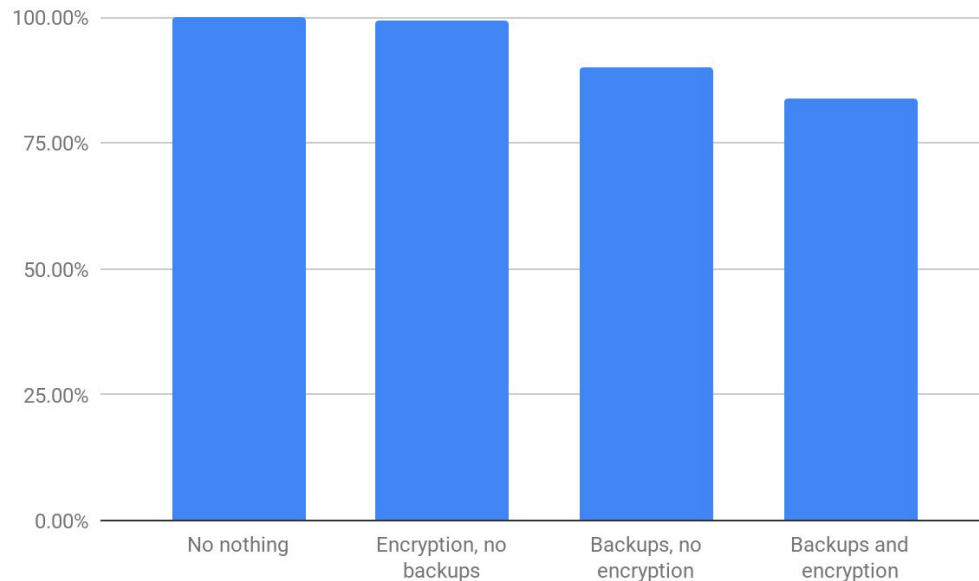
Google Cloud Platform



Overhead of encryption and backups

20 CPU / 120 GB instance in Google Cloud Iowa region

- Encryption: LUKS
 - Very low impact
- Backups:
 - PGHoard with GCS
 - wal_level = replica
- No backups:
 - wal_level = minimal
- 10% overhead for backups
 - Required for production



Block storage system options

Network disks

- + Persistent past node lifetime
- + Almost infinitely scalable
- Really slow, or
- Quite expensive (PrIOPS)
- Compete with others over limited IO bandwidth
- Not free of faults

Local disks

- + Fast
- + Potentially really fast
- + Cheap
- Available in limited sizes
 - (or not at all)
- Ephemeral
 - Node shuts down: data is gone
 - Test your backup & restore system

450 GB data set ~ 128 GB RAM instances

AWS

Local disks:

i3.xlarge
i3.8xlarge

Network disks:

m5.12xlarge

\$900 - \$1900

Azure

Local disks:

Standard L16

\$1000

DigitalOcean

Local disks:

128GB

\$640

Google Cloud

Local disks:

20x120
32x192

Network disks:

20x120
32x192

\$730 - \$1200

Packet

Local disks:

c1.xlarge
m1.xlarge

\$1200

UpCloud

Network disks:

128GB

\$640

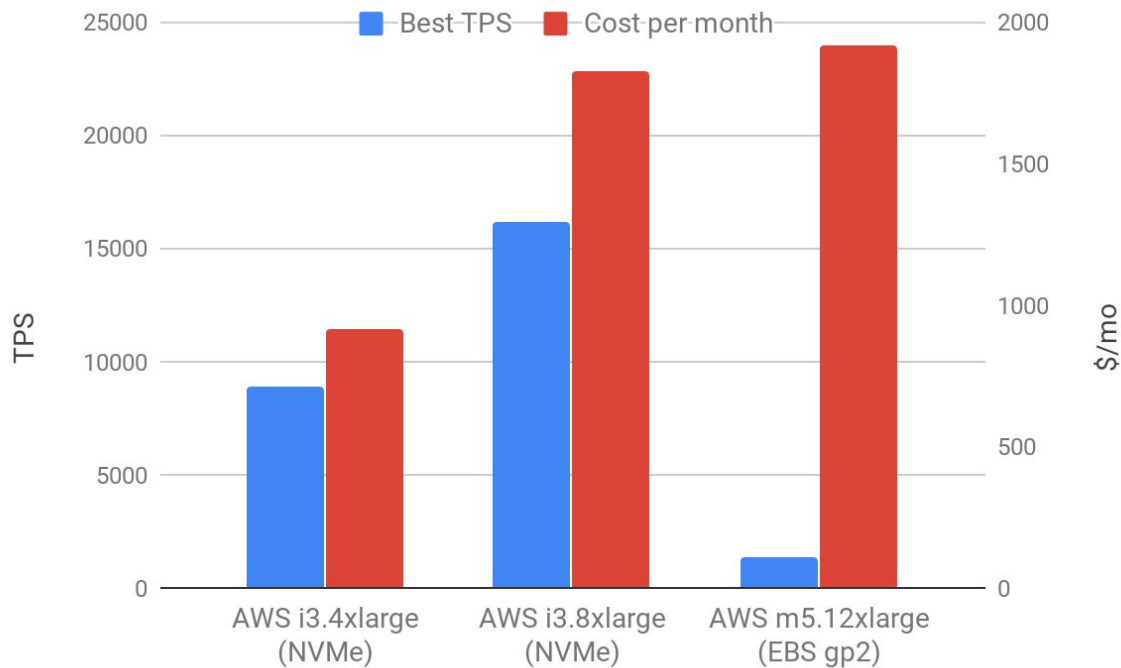
postgresql.conf settings

```
work_mem = 32MB  
shared_buffers = 20% of RAM  
max_wal_size = 32GB  
wal_level = replica
```

pgbench commands

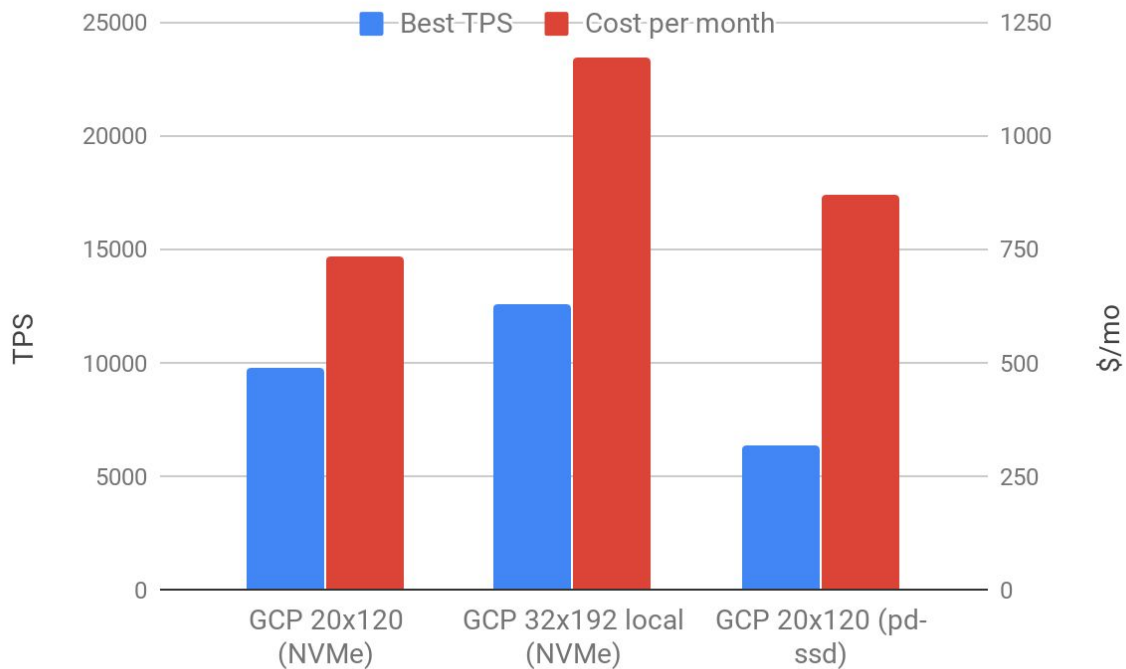
```
pgbench --initialize --scale=25000  
pgbench --jobs=8 --client=100 \  
--time=3600
```

450 GB data set in AWS



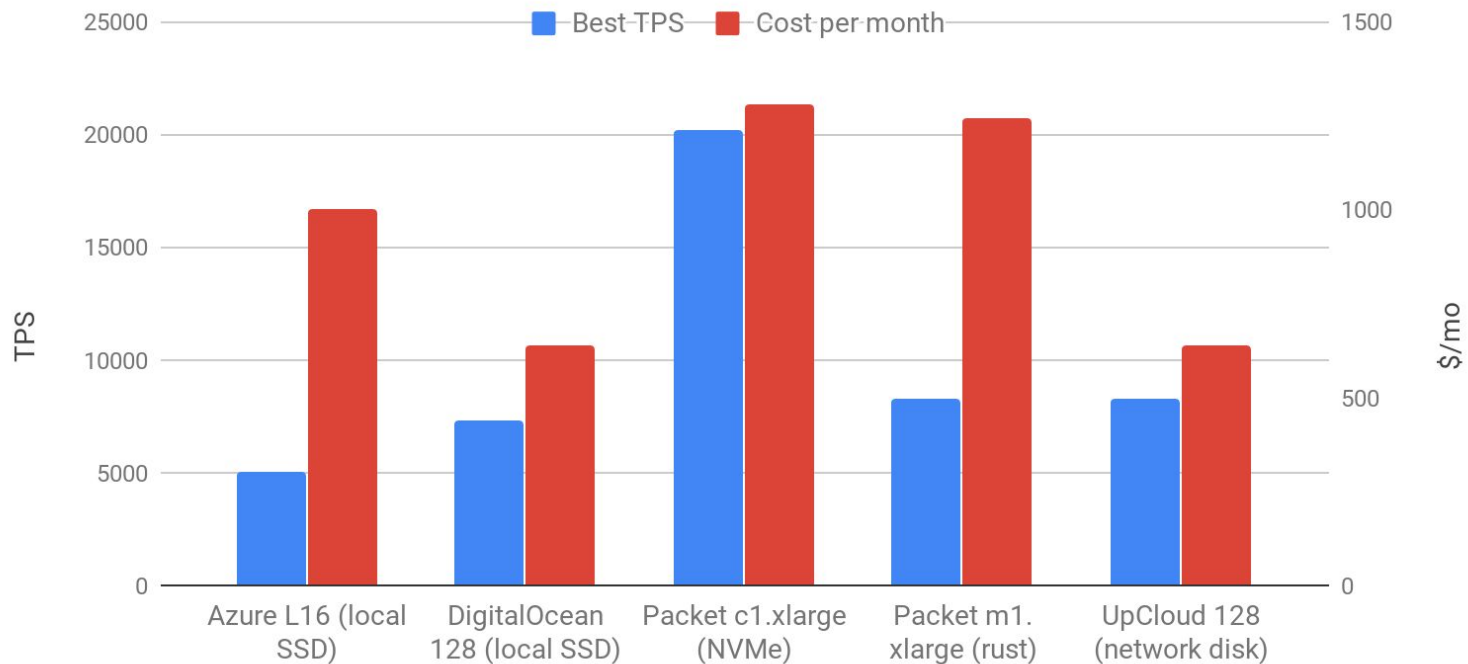
AWS i3.4xlarge and i3.8xlarge both provide 9 to 10 transactions per second at the cost of \$911 and \$1822/mo. Performance improves almost linearly with instance size. EBS gp2 backed m5.12xlarge is surprisingly slow.

450 GB data set in GCP



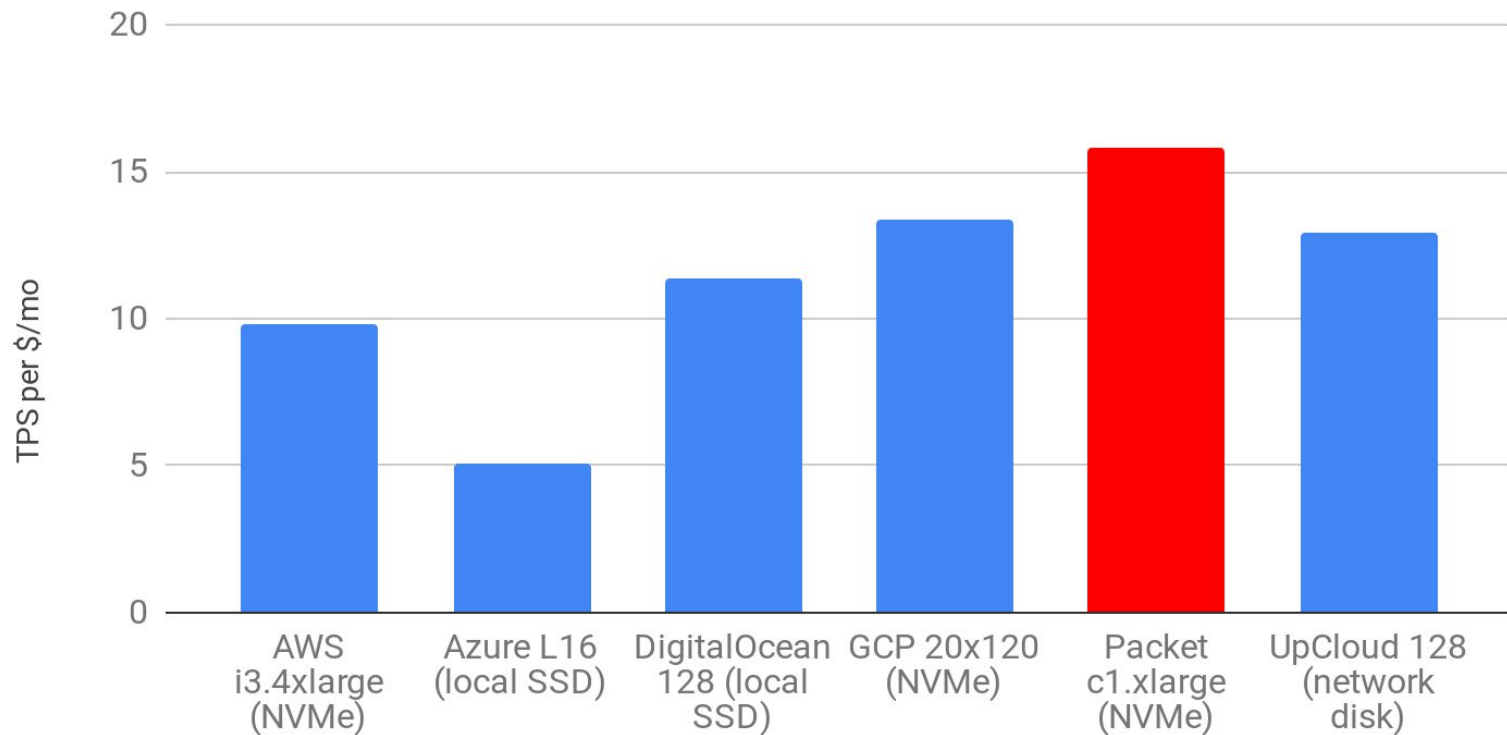
The slightly smaller Google Cloud instance offers better performance per dollar, but performance improves significantly with instance size. TPS/\$ numbers (11-13) are slightly higher than on AWS.

450 GB data set in other clouds



Bare metal instance with NVMe disk in Packet dominates, other instances offer similar performance at different price points.

Most cost efficient instances for 450 GB data set



Many of the 128 GB instances offer similar performance per dollar spent on our 450 GB data set benchmark. Packet wins, Azure loses and all other providers except UpCloud offer options for scaling higher.

Benchmarks: 900 GB dataset

- Instances with 120 to 480 GB RAM, 16 to 96 vCPUs
- PGBench with 200 concurrent clients
- Different storage options including network and local storage devices
- Six different clouds, including virtual and bare metal instances



Google Cloud Platform



900 GB data set ~ 256 GB RAM instances

AWS

Local disks:

i3.xlarge
i3.metal

Network disks:

m5.metal

\$1900 - \$3700

Azure

Local disks:

Standard L16
Standard L32

\$1000- \$2000

DigitalOcean

Local disks:

192GB

\$960

Google Cloud

Local disks:

32x192
n1-standard-96

Network disks:

32x192
n1-standard-96

\$1200 - \$2900

Packet

Local disks:

m1.xlarge
m2.xlarge

\$1200 - \$1400

UpCloud

Network disks:

128GB

\$640

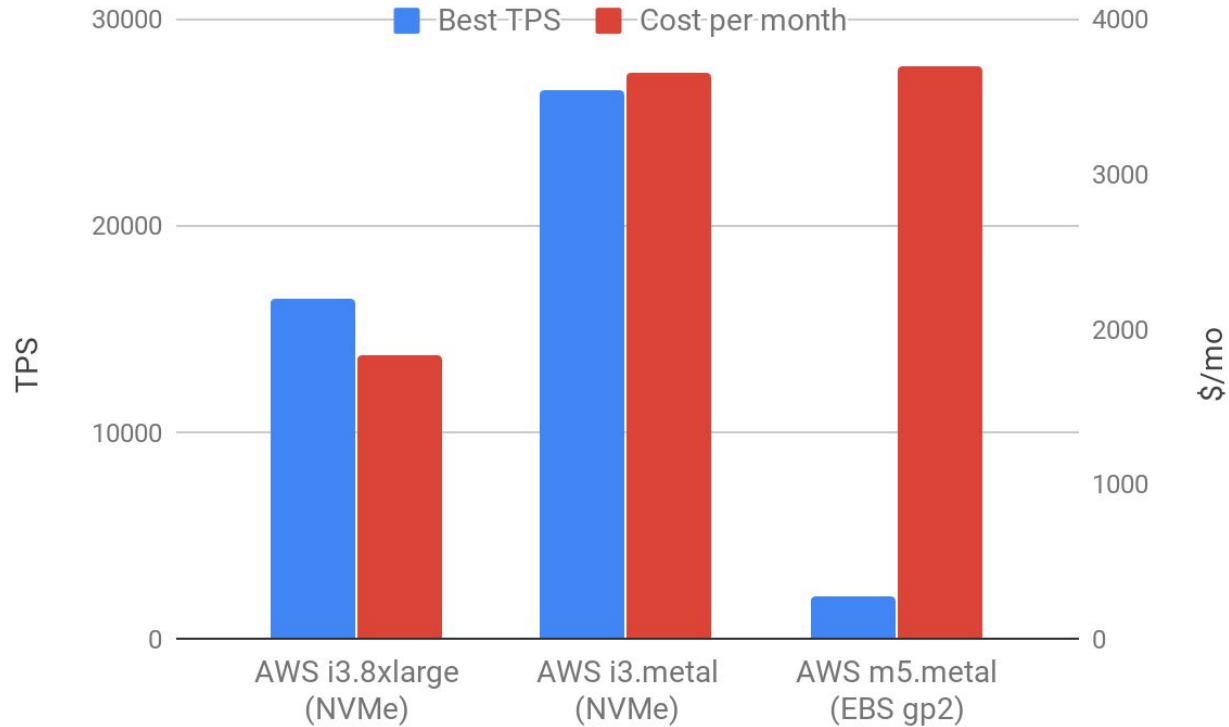
postgresql.conf settings

```
work_mem = 32MB  
shared_buffers = 20% of RAM  
max_wal_size = 64GB  
wal_level = replica
```

pgbench commands

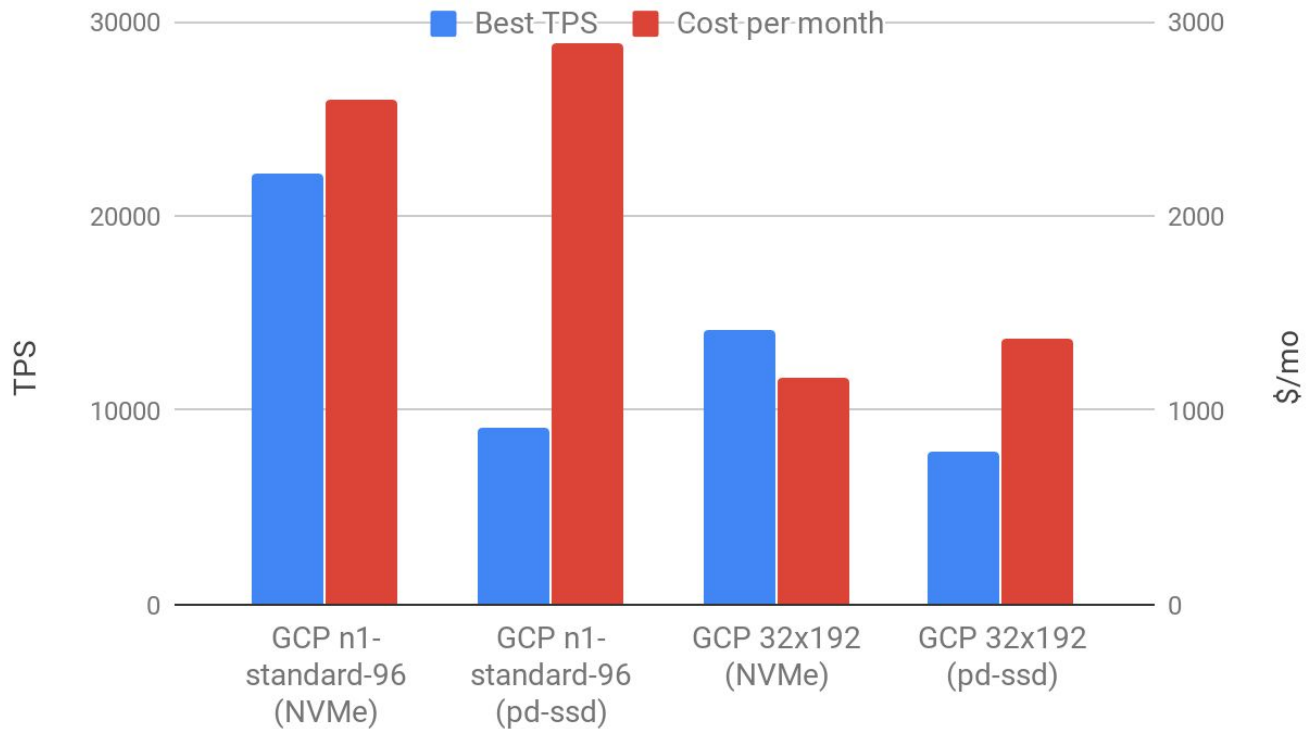
```
pgbench --initialize --scale=50000  
pgbench --jobs=8 --client=200 \  
--time=3600
```

900 GB data set in AWS



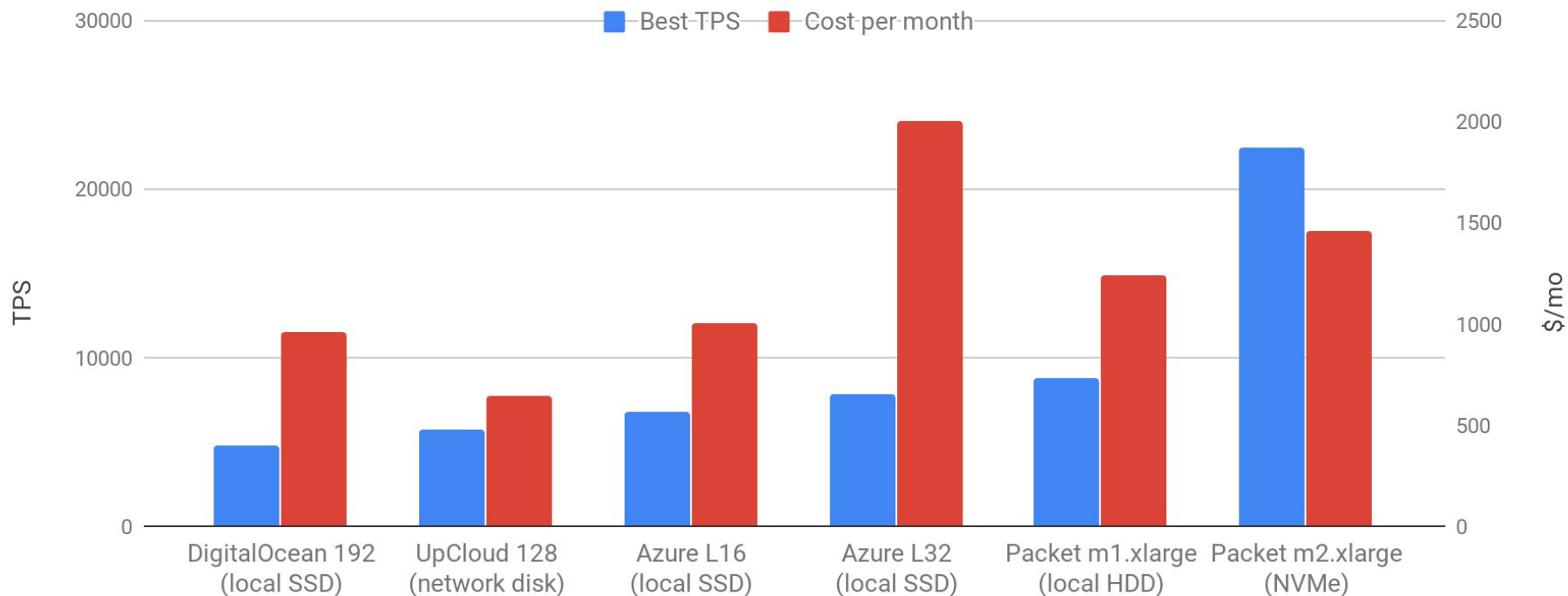
AWS i3.xlarge provides 9 tps per dollar at the cost of \$1800/mo. Larger local-NVMe backed instance improves performance but provides less value. The most expensive, EBS backed bare metal instance, m5.metal is again the loser.

900 GB data set in GCP



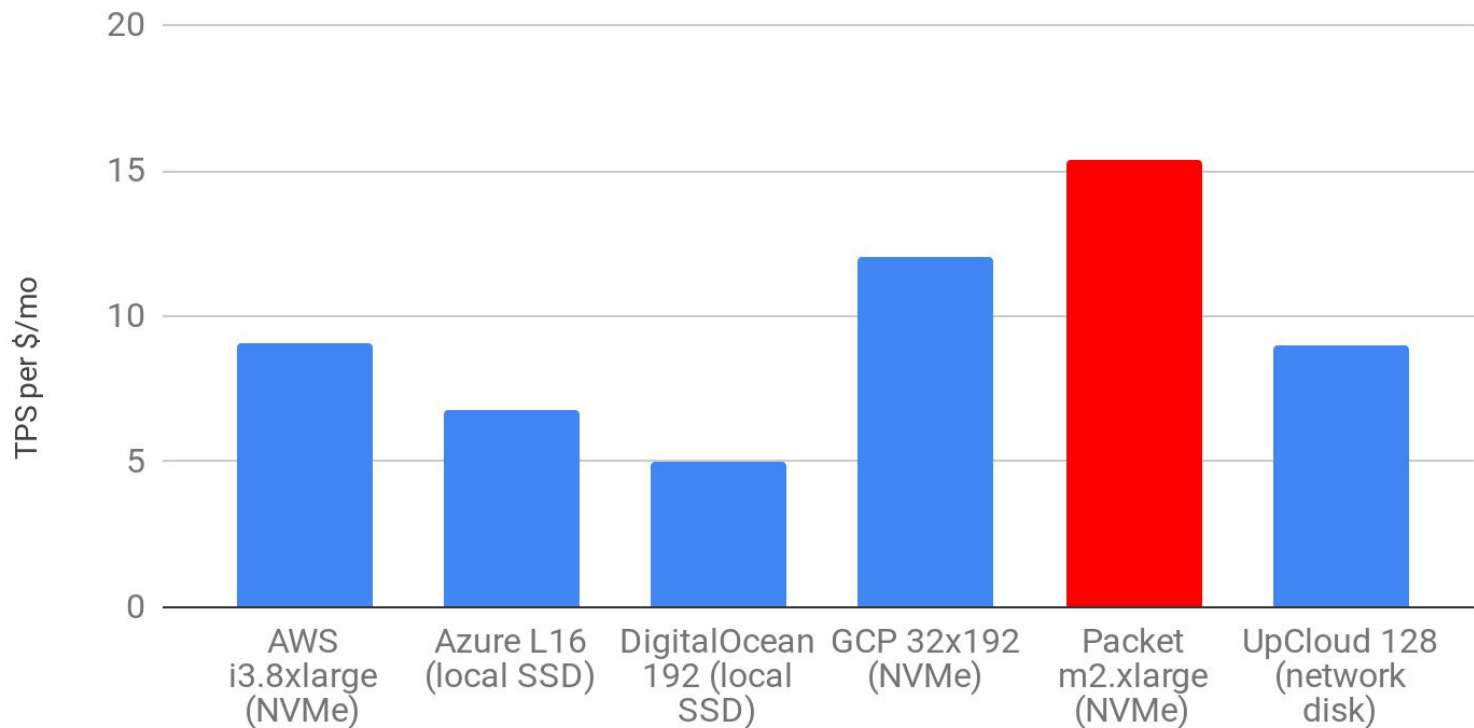
GCP 32x192 GB provides 12 tps per dollar at the cost of \$1400/mo. Larger local-NVMe backed instance improves performance but provides less value. The more expensive, pd-ssd backed instances perform and scale less well.

900 GB data set in other clouds



Bare metal instance with NVMe disk in Packet again dominates the picture, other instances offer similar performance at very different price points.

Most cost efficient instances for 900 GB data set



Many of the 256 GB instances offer similar performance per dollar spent on our 900 GB data set benchmark. Packet wins again with their topline bare metal server, many providers do not offer higher performing options.

Summary

- Performance varies greatly between different clouds and instances
- Storage devices appear to have the largest impact in an OLTP based benchmark
 - OLAP and batch workloads may have somewhat different results
- Larger instance sizes don't always offer better performance
- Test your workload on multiple instances before deciding what to use
- The most popular managed services may not offer all the options out there

Questions?


Cool t-shirts for the first ones to ask a question!

Continue the discussion at our booth at the conference and try out all these clouds and plans on <https://aiven.io>





Thanks!

 <https://aiven.io>

 @aiven_io

 @OskariSaarenmaa