Achieving the ultimate performance with KVM

Boyan Krosnov
Open Infrastructure Summit
Shanghai 2019

StorPool & Boyan K.

- NVMe software-defined storage for VMs and containers
- Scale-out, HA, API-controlled
- Since 2011, in commercial production use since 2013
- Based in Sofia, Bulgaria
- Mostly **virtual disks for KVM**
- … and bare metal Linux hosts
- Also used with VMWare, Hyper-V, XenServer
- Integrations into **OpenStack/Cinder, Kubernetes Persistent Volumes**, CloudStack, OpenNebula, OnApp

StorPool DISTRIBUTED STORAGE

Why performance

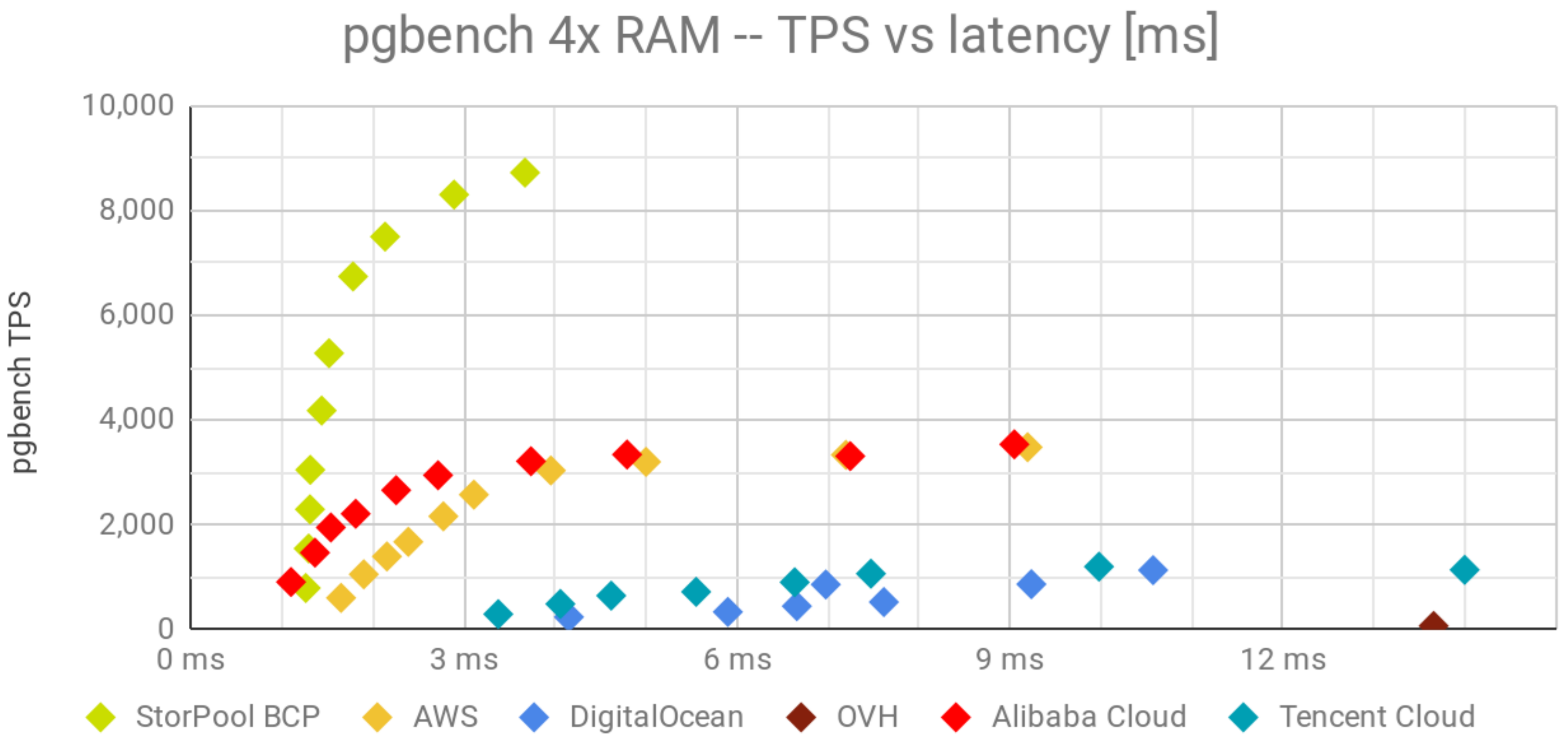
- Better application performance -- e.g. time to load a page, time to rebuild, time to execute specific query

- Happier customers (in cloud / multi-tenant environments)

- ROI, TCO - Lower cost per delivered resource (per VM) through higher density

StorPool
DISTRIBUTED STORAGE

Why performance



StorPool
DISTRIBUTED STORAGE

Agenda

- Hardware
- Compute - CPU & Memory
- Networking
- Storage

Compute node hardware

Usual optimization goal

- lowest cost per delivered resource
- fixed performance target
- calculate all costs - power, cooling, space, server, network, support/maintenance

Example: cost per VM with 4x dedicated 3 GHz cores and 16 GB RAM

Unusual

- Best single-thread performance I can get at any cost
- 5 GHz cores, yummy :)

StorPool DISTRIBUTED STORAGE

Compute node hardware

	Brand	Model	release date	ark.intel.com status	release price ($)	Cores	TDP (W)	All-Core Turbo Clock (GHz)	Selected 1S or 2S or 4S ?	Total $ per core	Total $/GHz
2	Gold	6222V	May 2019	Launched	$1,600	20	115	2.4	2S	$290	$121
3	Silver	4216	April 2019	Launched	$1,002	16	100	2.7	2S	$296	$110
4	Gold	6230	April 2019	Launched	$1,894	20	125	2.8	2S	$308	$110
5	Gold	6230T	May 2019	Launched	$1,988	20	125	2.8	2S	$313	$112
6	Gold	6230N	May 2019	Launched	$2,046	20	125	2.9	2S	$316	$109
7	Gold	5220	April 2019	Launched	$1,555	18	125	2.7	2S	$317	$117
8	Gold	6262V	May 2019	Launched	$2,900	24	135	2.5	2S	$317	$127
9	Gold	5220T	May 2019	Launched	$1,727	18	105	2.7	2S	$321	$119
10	Gold	5218T	May 2019	Launched	$1,349	16	105	2.7	2S	$321	$119
11	Gold	5218N	April 2019	Launched	$1,375	16	105	3.0	2S	$323	$108
12	Gold	5218	April 2019	Launched	$1,273	16	125	2.8	2S	$323	$115
13	Gold	5218B	April 2019	Launched	$1,273	16	125	2.8	2S	$323	$115
14	Gold	6238	May 2019	Launched	$2,612	22	140	2.8	2S	$329	$118
15	Gold	6238T	April 2019	Launched	$2,742	22	125	2.7	2S	$331	$123
16	Silver	4214	April 2019	Launched	$694	12	85	2.7	2S	$333	$123
17	Gold U	6209U	May 2019	Launched	$1,350	20	125	2.8	1S	$339	$121
18	Silver	4214Y	April 2019	Launched	$768	12	85	2.7	2S	$340	$126
19	Gold	5220S	May 2019	Launched	$2,000	18	125	2.7	2S	$343	$127
20	Gold	6252	April 2019	Launched	$3,655	24	150	2.8	2S	$354	$126
21	Gold U	6210U	May 2019	Launched	$1,500	20	150	3.2	1S	$355	$111
22	Gold	6252N	May 2019	Launched	$3,984	24	150	3.0	2S	$368	$123
23	Silver	4210	April 2019	Launched	$501	10	85	2.7	2S	$371	$137
24	Gold	6248	April 2019	Launched	$3,072	20	150	3.2	2S	$378	$118
25	Gold	6240	April 2019	Launched	$2,445	18	150	3.3	2S	$378	$115

StorPool DISTRIBUTED STORAGE

Compute node hardware

Intel

lowest cost per core:

- Xeon Gold 6222V - 20 cores @ 2.4 GHz

lowest cost per 3GHz+ core:

- Xeon Gold 6210U - 20 cores @ 3.2 GHz
  - Xeon Gold 6240 - 18 cores @ 3.3 GHz
- Xeon Gold 6248 - 20 cores @ 3.2 GHz

AMD

- EPYC 7702P - 64 cores @ 2.0/3.35 GHz - lowest cost per core
- EPYC 7402P - 24 cores / 1S - low density
  - EPYC 7742 - 64 cores @ 2.2/3.4GHz x 2S - max density

StorPool DISTRIBUTED STORAGE

Compute node hardware

Form factor

from

to

Compute node hardware

- firmware versions and BIOS settings
- Understand power management -- esp. C-states, P-states, HWP and “bias”
  - Different on AMD EPYC: "power-deterministic", "performance-deterministic"
- Think of rack level optimization - how do we get the lowest total cost per delivered resource?

StorPool DISTRIBUTED STORAGE

Agenda

- Hardware
- Compute - CPU & Memory
- Networking
- Storage

StorPool
DISTRIBUTED STORAGE

Tuning KVM

RHEL7 Virtualization_Tuning_and_Optimization_Guide link

https://pve.proxmox.com/wiki/Performance_Tweaks

https://events.static.linuxfound.org/sites/events/files/slides/CloudOpen2013_Khoa_Huynh_v3.pdf

http://www.linux-kvm.org/images/f/f9/2012-forum-virtio-blk-performance-improvement.pdf

http://www.slideshare.net/janghoonsim/kvm-performance-optimization-for-ubuntu

… but don’t trust everything you read. Perform your own benchmarking!

StorPool DISTRIBUTED STORAGE

CPU and Memory

Recent Linux kernel, KVM and QEMU
… but beware of the bleeding edge
E.g. qemu-kvm-ev from RHEV (repackaged by CentOS)

tuned-adm virtual-host
tuned-adm virtual-guest

StorPool
DISTRIBUTED STORAGE

CPU

Typical

- (heavy) oversubscription, because VMs are mostly idling
- HT
- NUMA
- route IRQs of network and storage adapters to a core on the NUMA node they are on

Unusual

- CPU Pinning

Understanding oversubscription and congestion

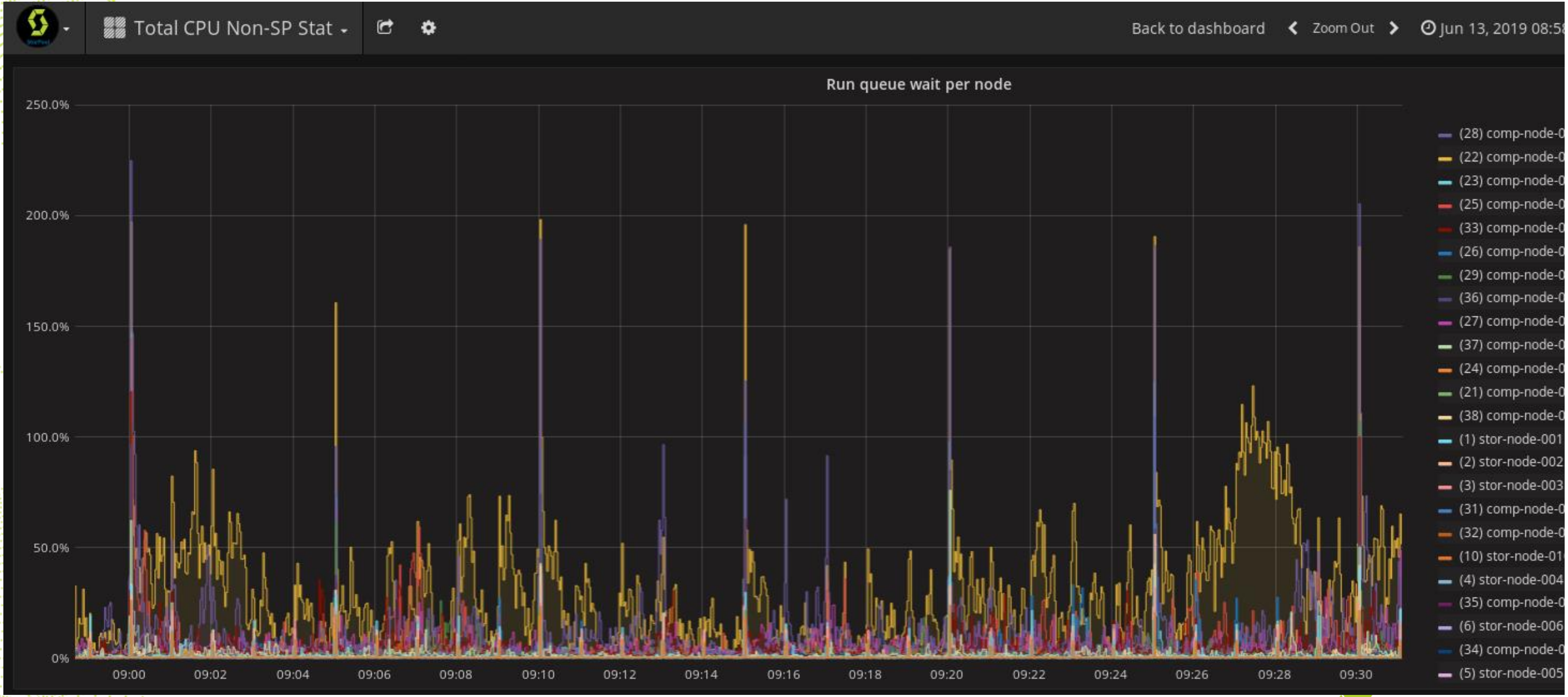
Linux scheduler statistics: linux-stable/Documentation/scheduler/sched-stats.txt

```
Next three are statistics describing scheduling latency:
     7) sum of all time spent running by tasks on this processor (in jiffies)
     8) sum of all time spent waiting to run by tasks on this processor (in 
jiffies)
     9) # of timeslices run on this cpu
```

20% CPU load with large wait time (bursty congestion) is possible
100% CPU load with no wait time, also possible

Measure CPU congestion!

Understanding oversubscription and congestion



StorPool
DISTRIBUTED STORAGE

Discussion

Memory

Typical

- Dedicated RAM
- huge pages, THP
- NUMA
- use local-node memory if you can

Unusual

- Oversubscribed RAM
- balloon
- KSM (RAM dedup)

StorPool DISTRIBUTED STORAGE

Discussion

Agenda

- Hardware
- Compute - CPU & Memory
- Networking
- Storage

StorPool
DISTRIBUTED STORAGE

Networking

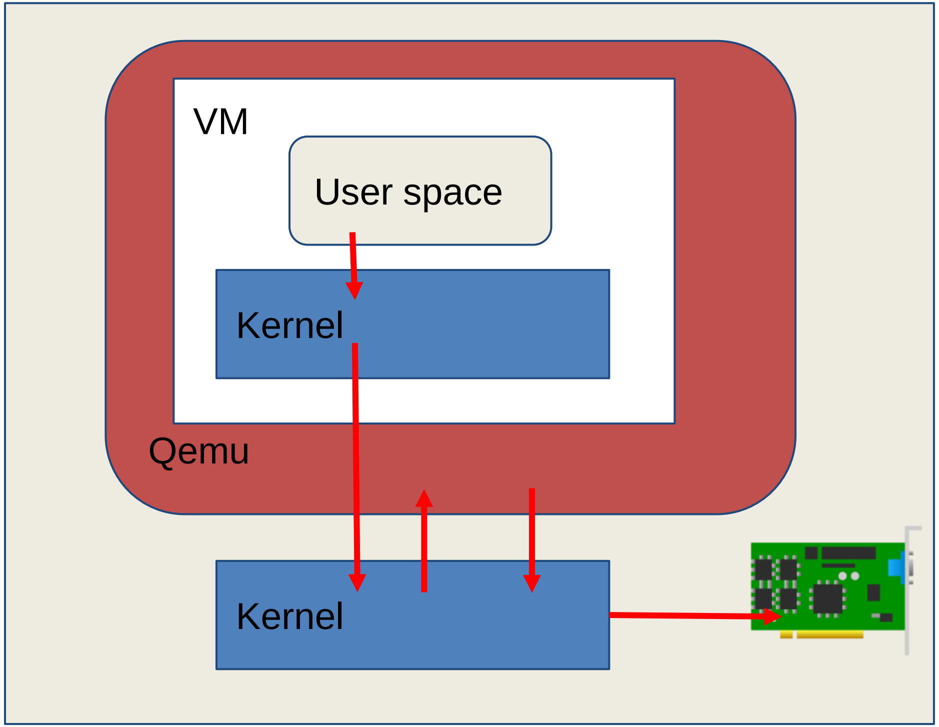
Virtualized networking
Use virtio-net driver
regular virtio vs vhost_net

Linux Bridge vs OVS in-kernel vs OVS-DPDK

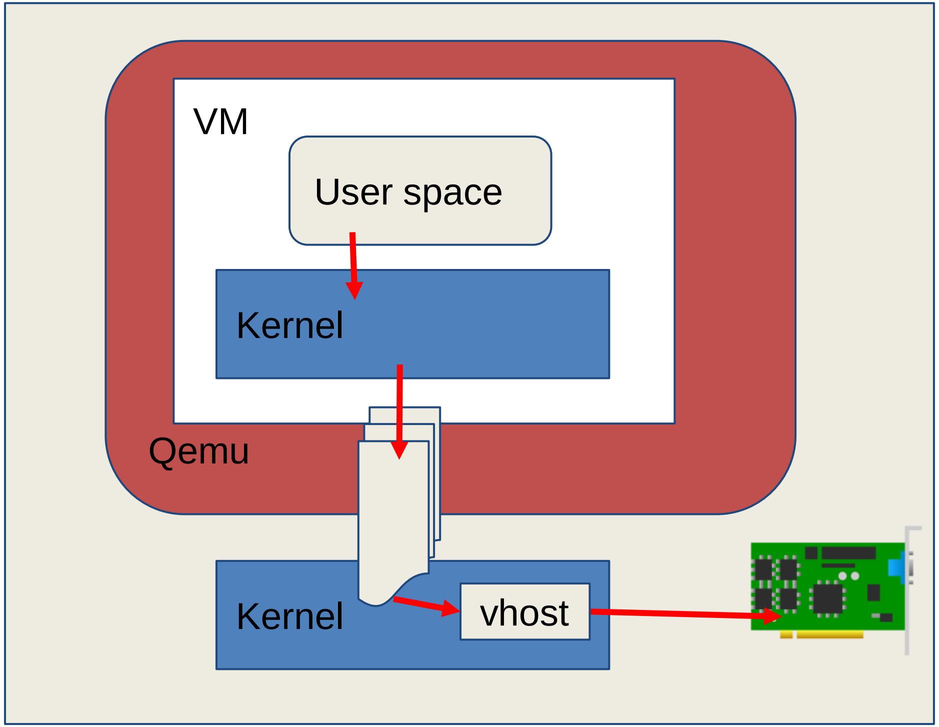
Pass-through networking
SR-IOV (PCIe pass-through)

StorPool DISTRIBUTED STORAGE

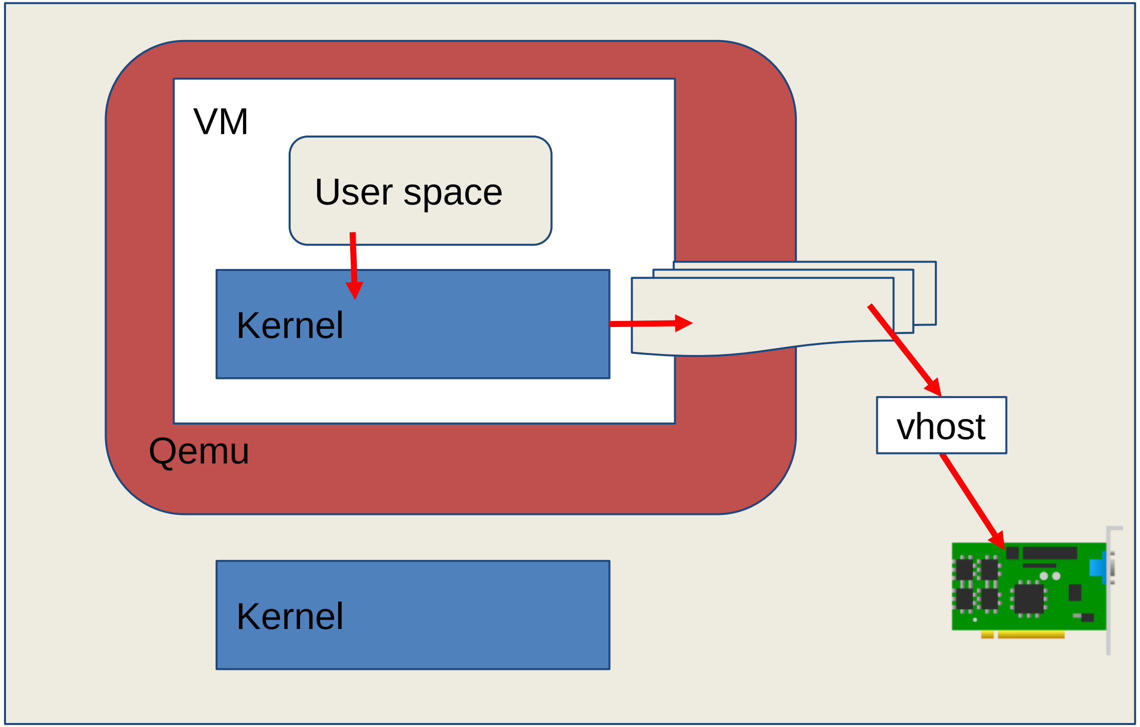
Networking - virtio



Networking - vhost

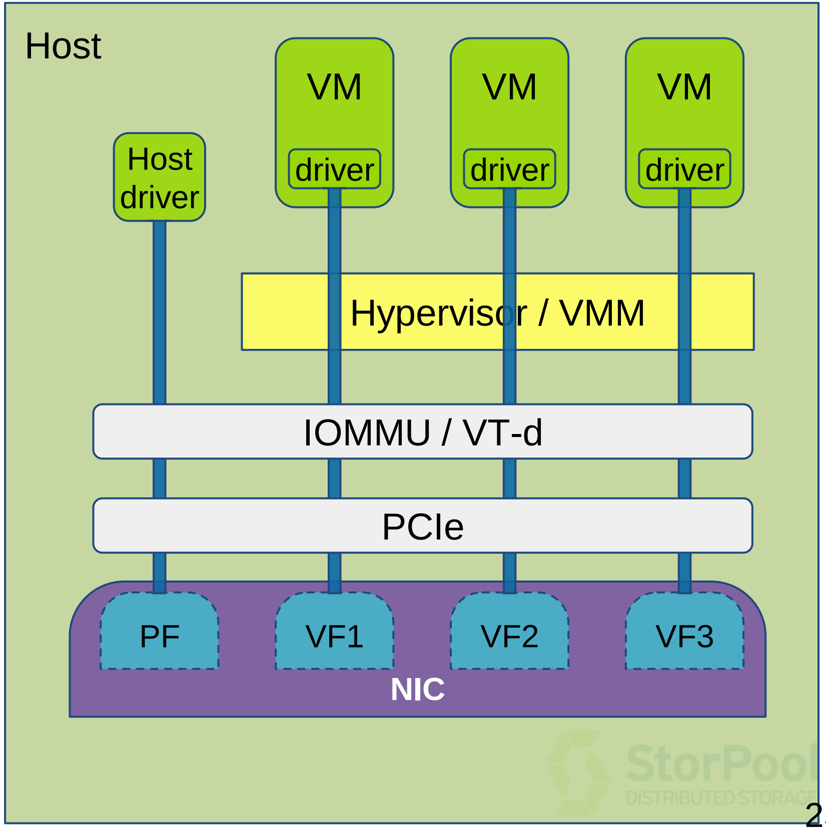


Networking - vhost-user



Networking - PCI Passthrough and SR-IOV

- Direct exclusive access to the PCI device
- SR-IOV - one physical device appears as multiple virtual functions (VF)
- Allows different VMs to share a single PCIe hardware



Discussion

Agenda

- Hardware
- Compute - CPU & Memory
- Networking
- Storage

StorPool
DISTRIBUTED STORAGE

Storage - virtualization

Virtualized

- cache=none -- direct IO, bypass host buffer cache
- io=native -- use Linux Native AIO, not POSIX AIO (threads)
- virtio-blk vs virtio-scsi
- virtio-scsi multiqueue
- iothread

vs. Full bypass

- SR-IOV for NVMe devices

StorPool DISTRIBUTED STORAGE

Storage - vhost

Virtualized with host kernel bypass

vhost

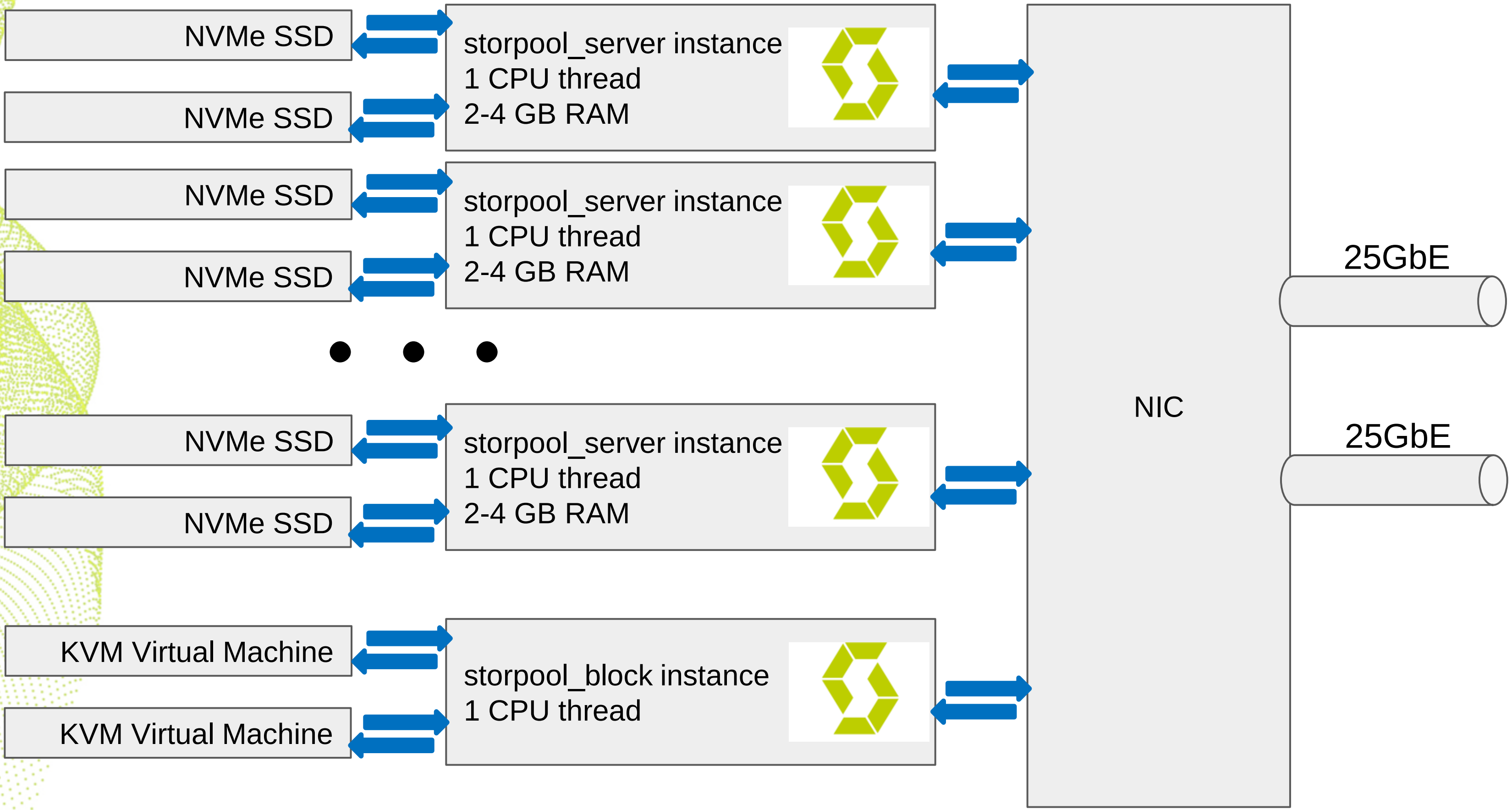
before:

guest kernel -> host kernel -> qemu -> host kernel -> storage system

after:

guest kernel -> storage system

StorPool DISTRIBUTED STORAGE



- Highly scalable and efficient architecture
- Scales up in each storage node & out with multiple nodes

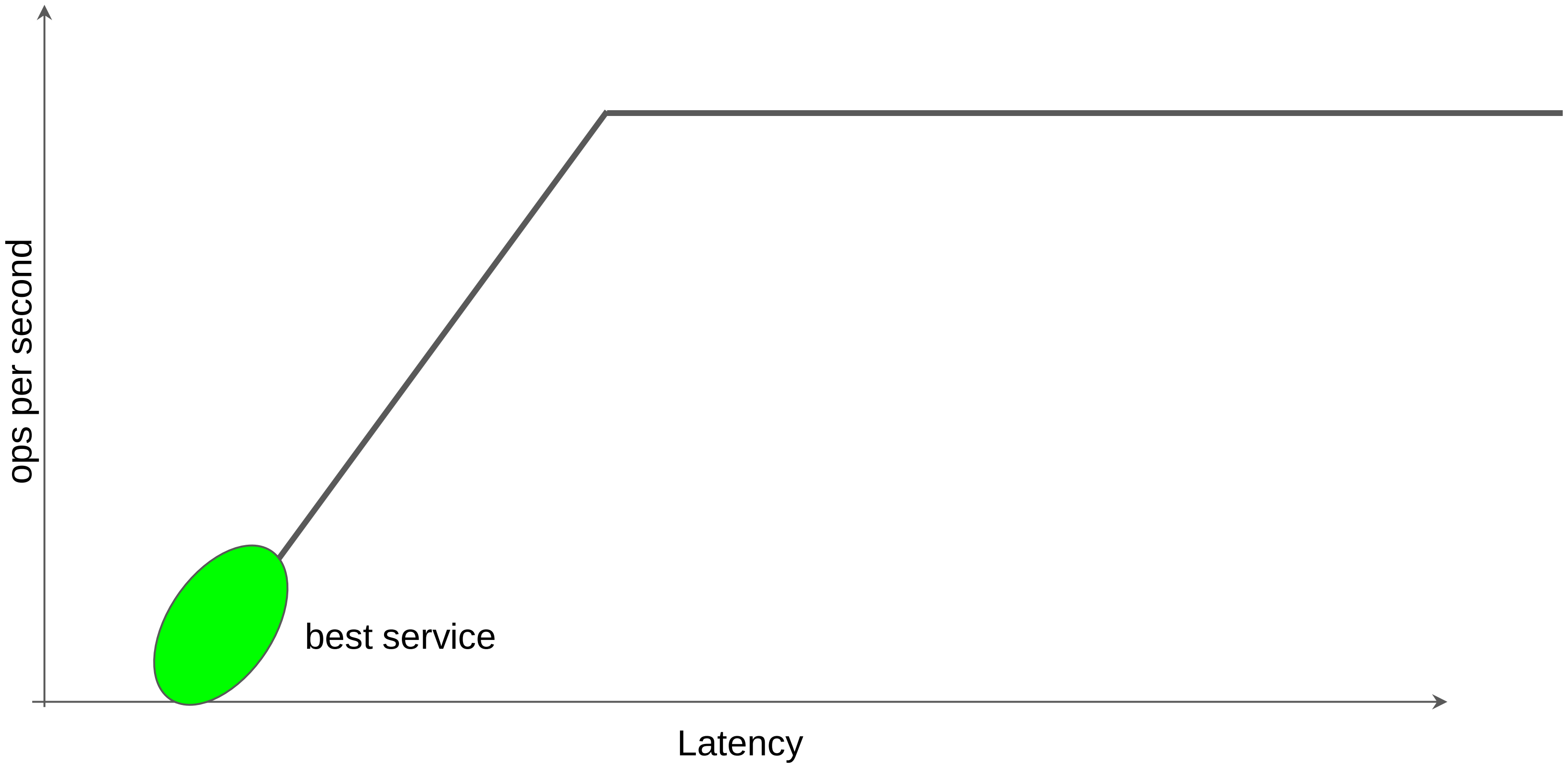
StorPool DISTRIBUTED STORAGE

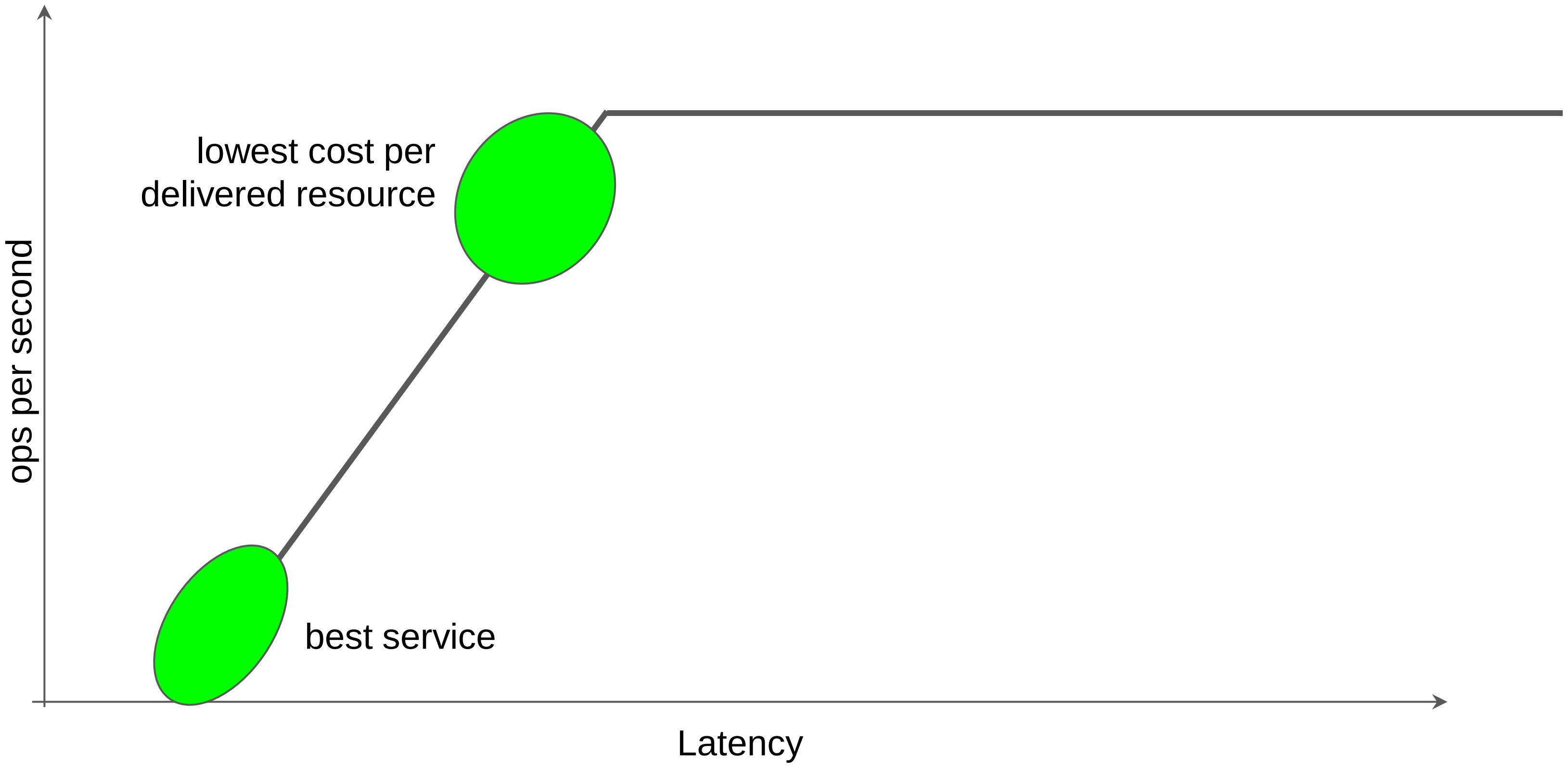
Storage benchmarks

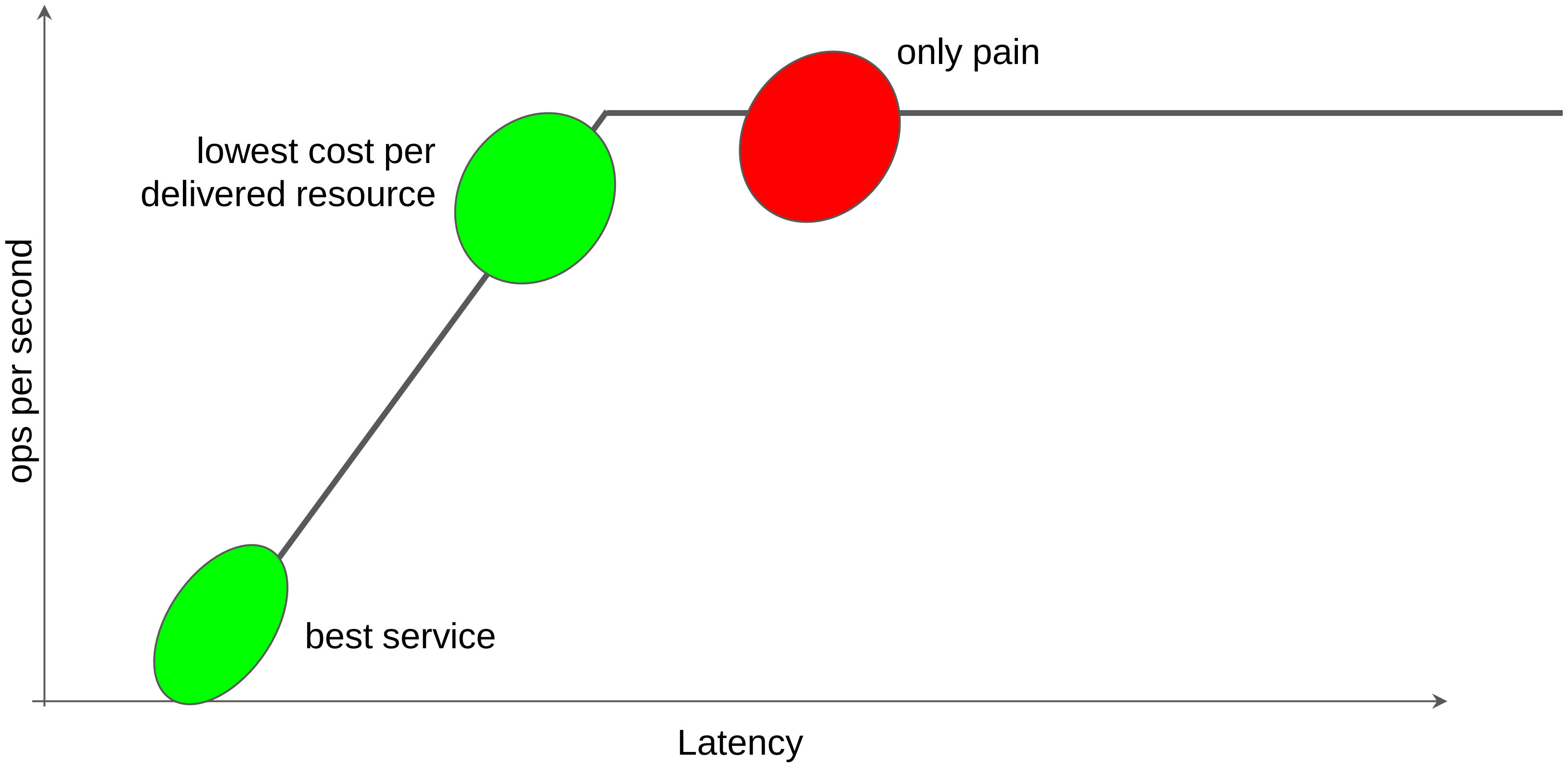
Beware: lots of snake oil out there!

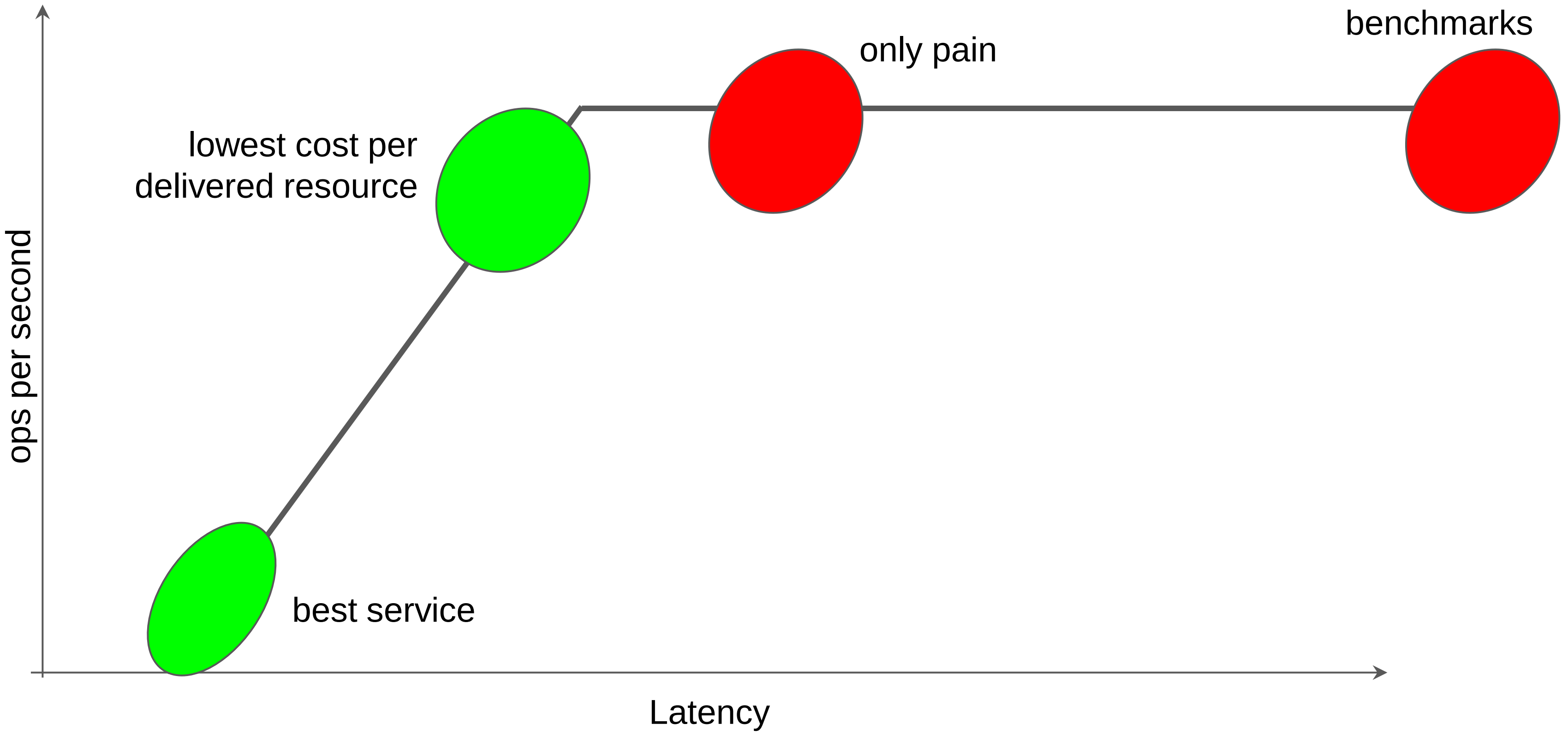
- performance numbers from hardware configurations totally unlike what you'd use in production
- synthetic tests with high iodepth - 10 nodes, 10 workloads * iodepth 256 each. (because why not)
- testing with ramdisk backend

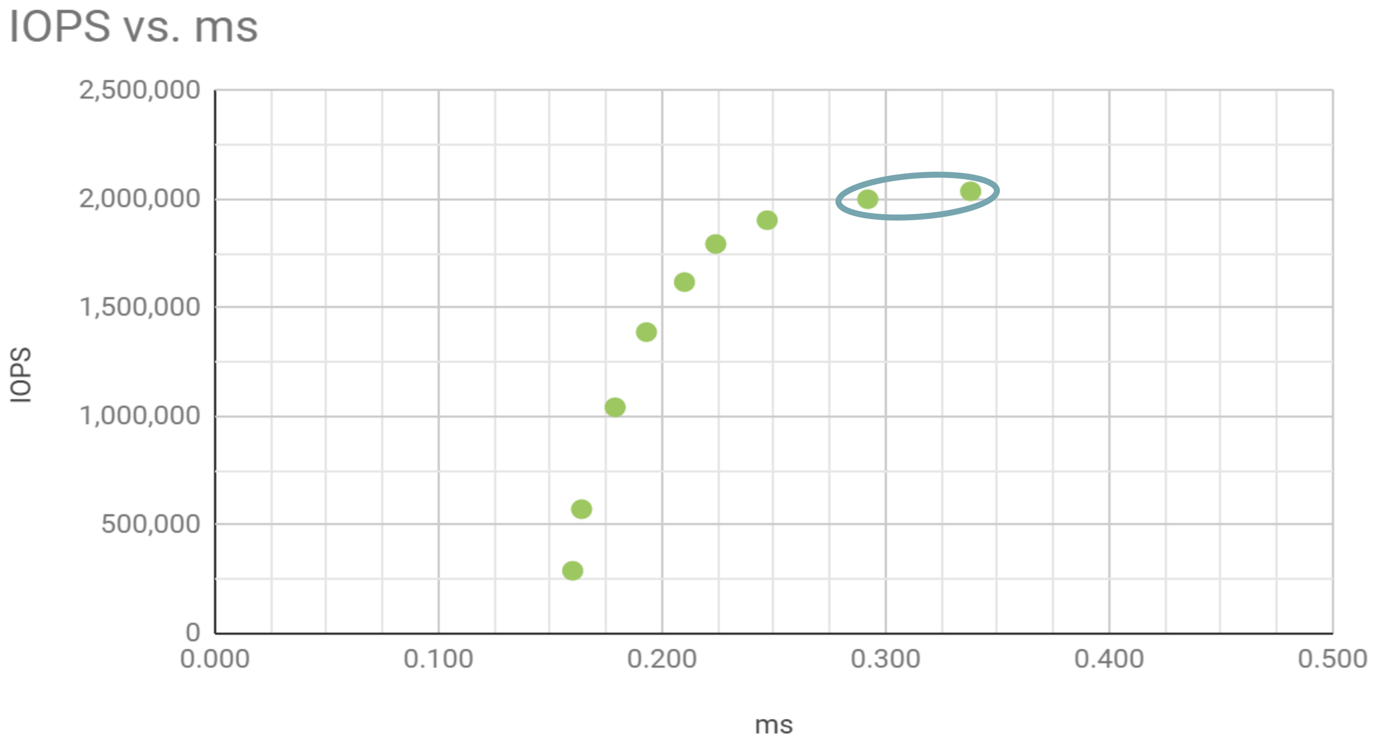
- synthetic workloads don't approximate real world (example)


ops per second
best service
Latency


lowest cost per
delivered resource
best service
ops per second
Latency


only pain
lowest cost per
delivered resource
ops per second
best service
Latency

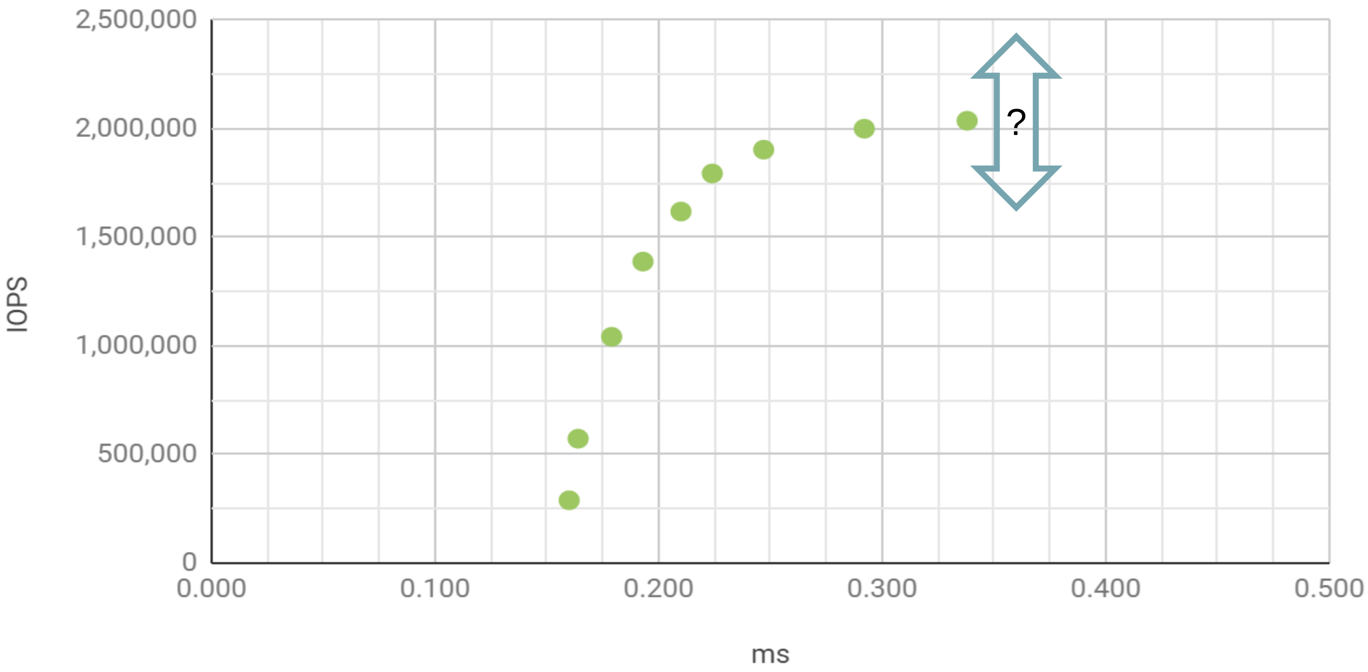

benchmarks
only pain
lowest cost per
delivered resource
ops per second
best service
Latency



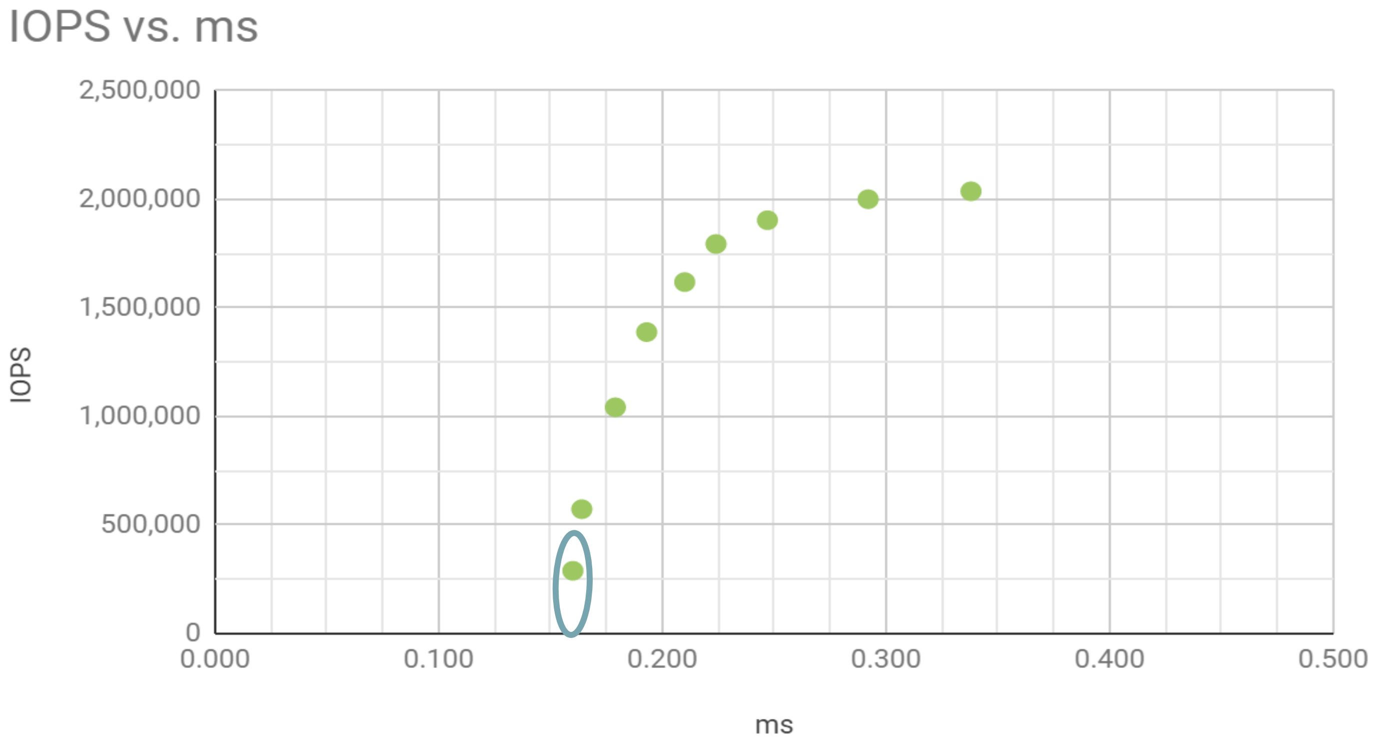
example1: 90 TB NVMe system - 22 IOPS per GB capacity

example2: 116 TB NVMe system - 48 IOPS per GB capacity

IOPS vs. ms

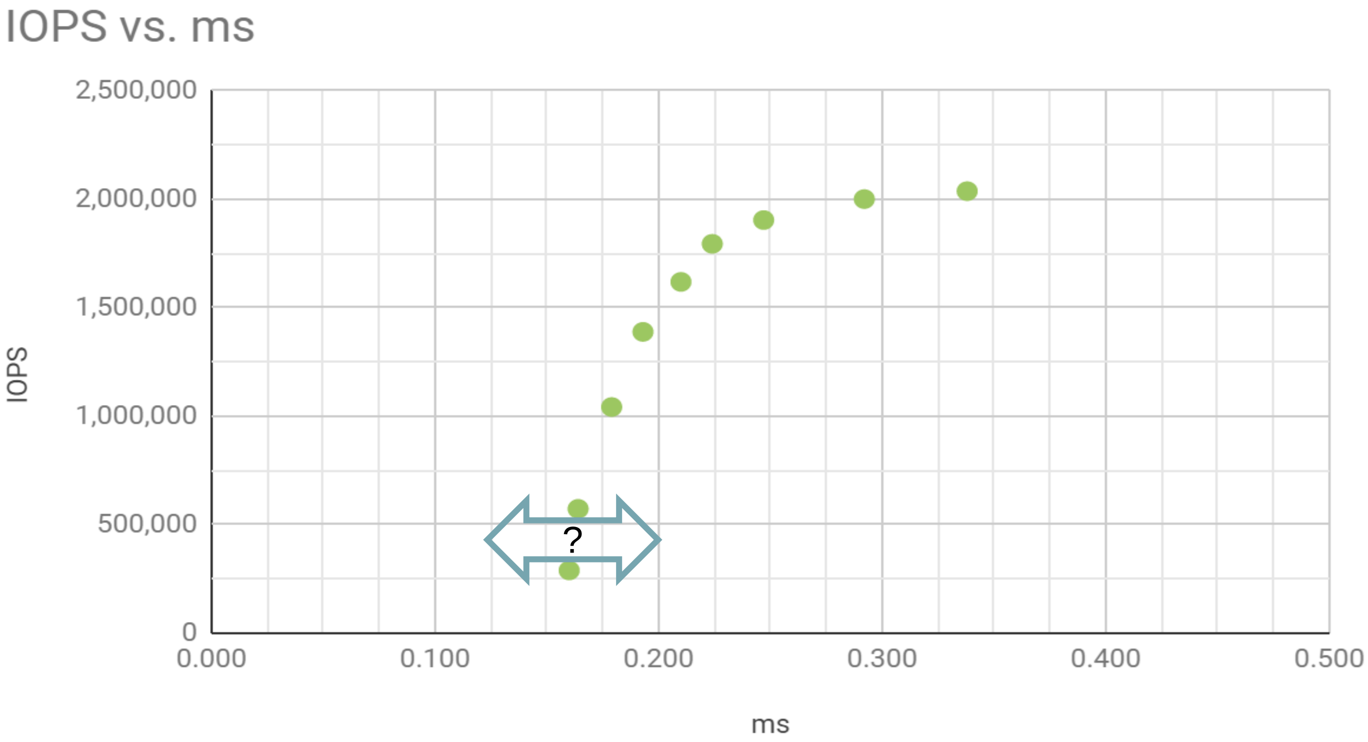


StorPool
DISTRIBUTED STORAGE



IOPS vs. ms
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
IOPS
0.000
0.100
0.200
0.300
0.400
0.500
ms

Real load


IOPS vs. ms
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
IOPS
0.000
0.100
0.200
0.300
0.400
0.500
ms
?

Discussion

Thank you!

Boyan Krosnov
bk@storpool.com
@bkrosnov

www.storpool.com
@storpool