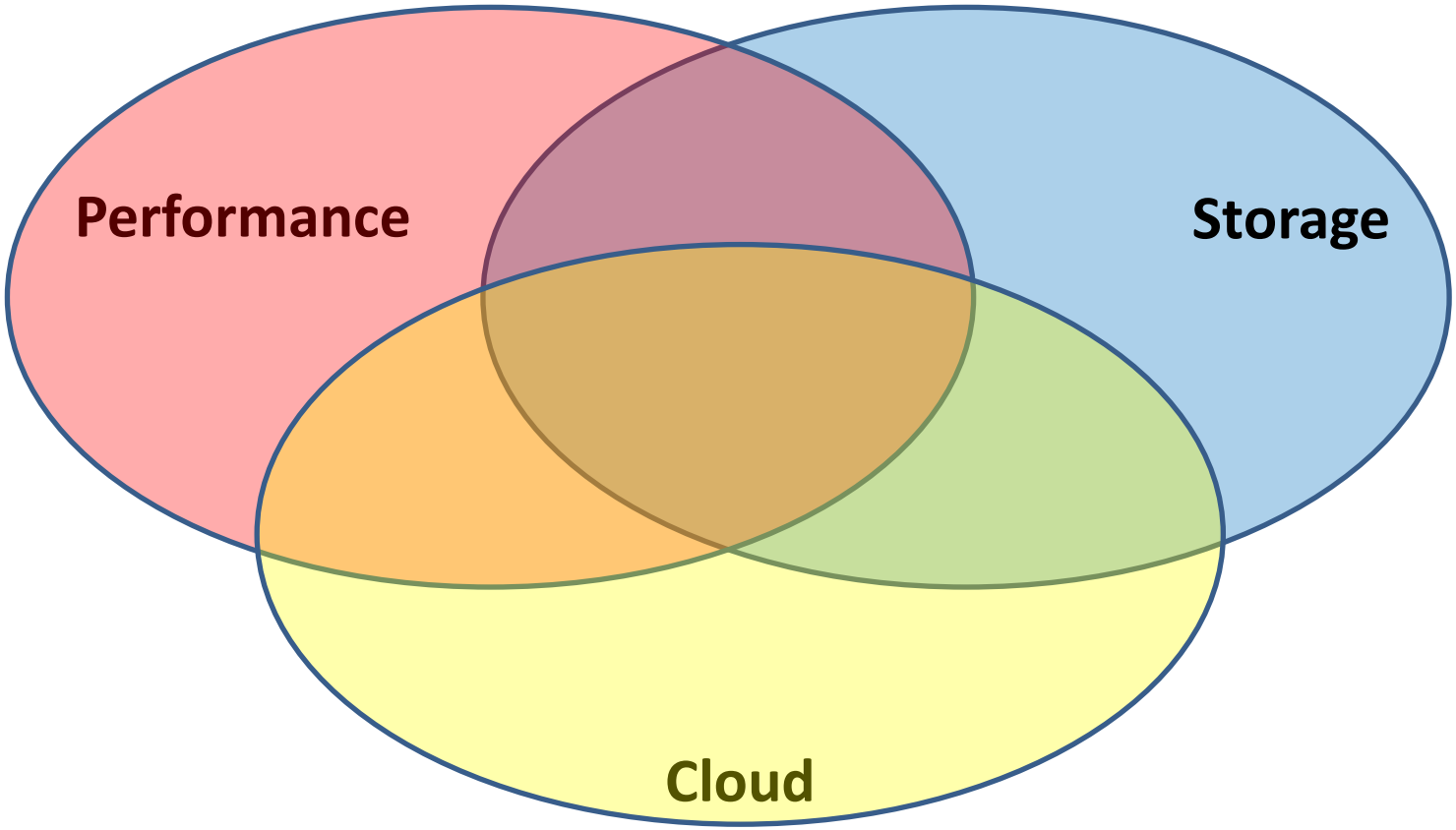
LISA '13
27th Large Installation System Administration Conference
November 3–8, 2013 • Washington, D.C.

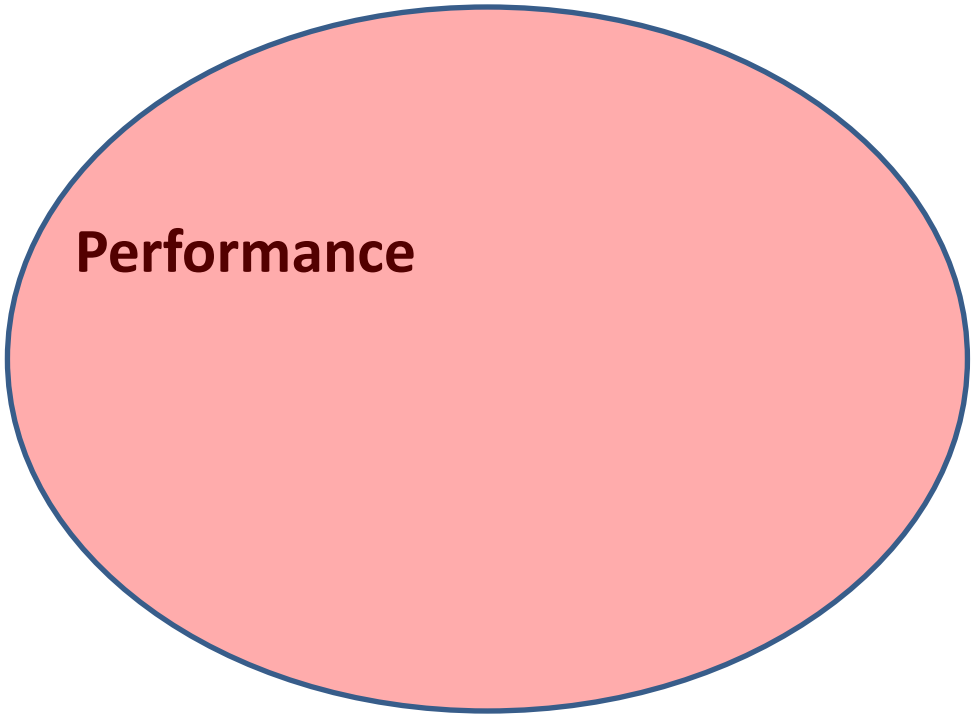
Measuring Storage Performance in the Cloud

Jeff Darcy
GlusterFS & Red Hat

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Overview





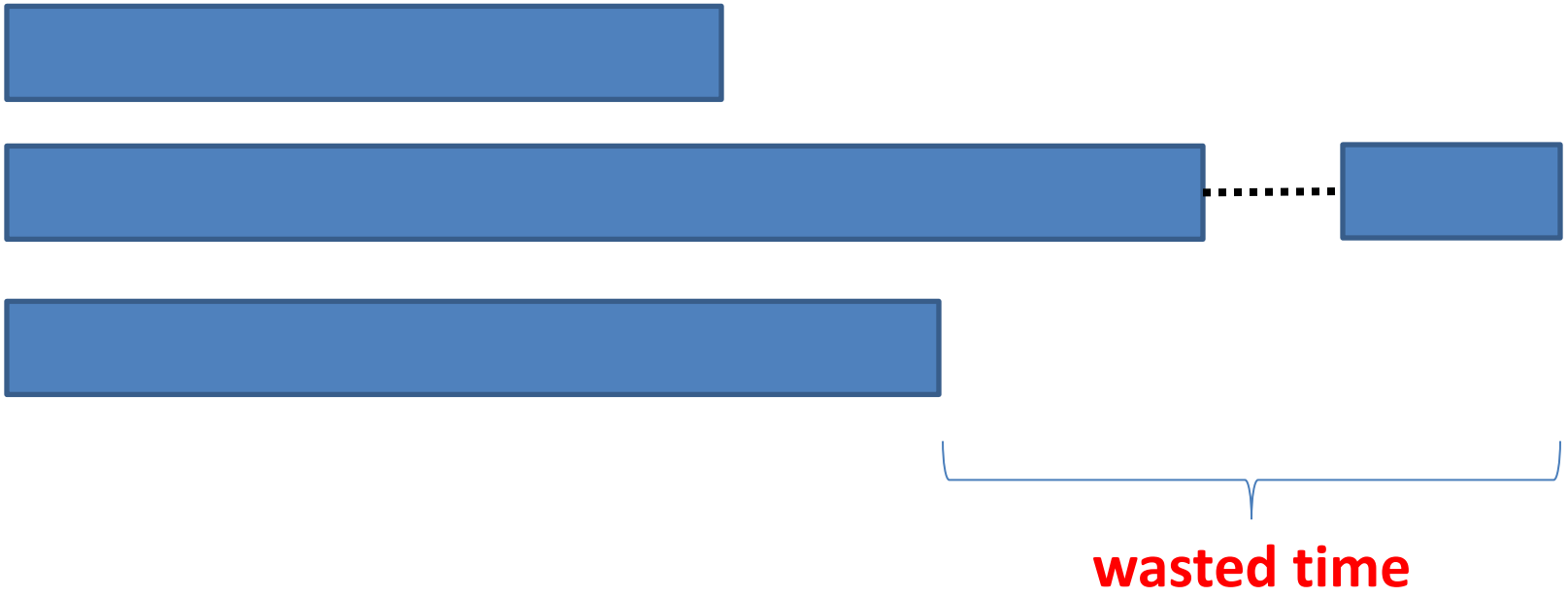
Types of Performance

	Bandwidth	Throughput	Latency	Latency Variation
Network	Gb/second	PPS	Milliseconds (average)	99th percentile
Storage	GB/second	IOPS	Milliseconds (average)	99th percentile

Often improves with thread count

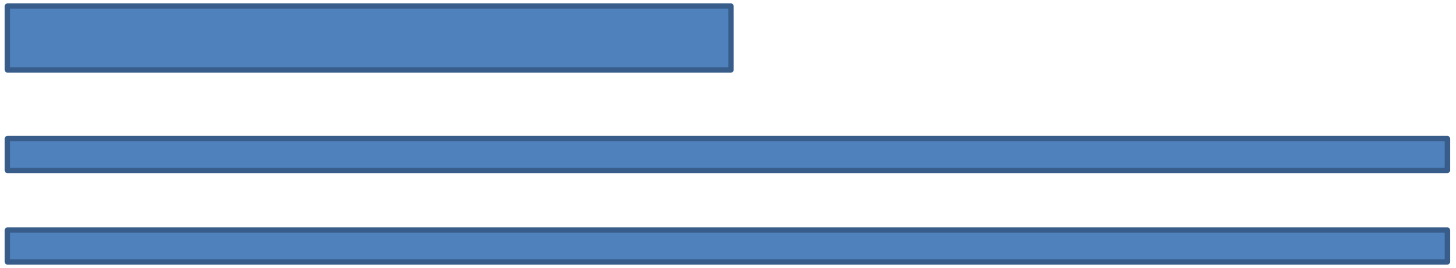
Often deteriorates with thread count

Tail at Scale = Fail



- If a request hits ten systems, with 10x latency 1% of the time, average latency **doubles**

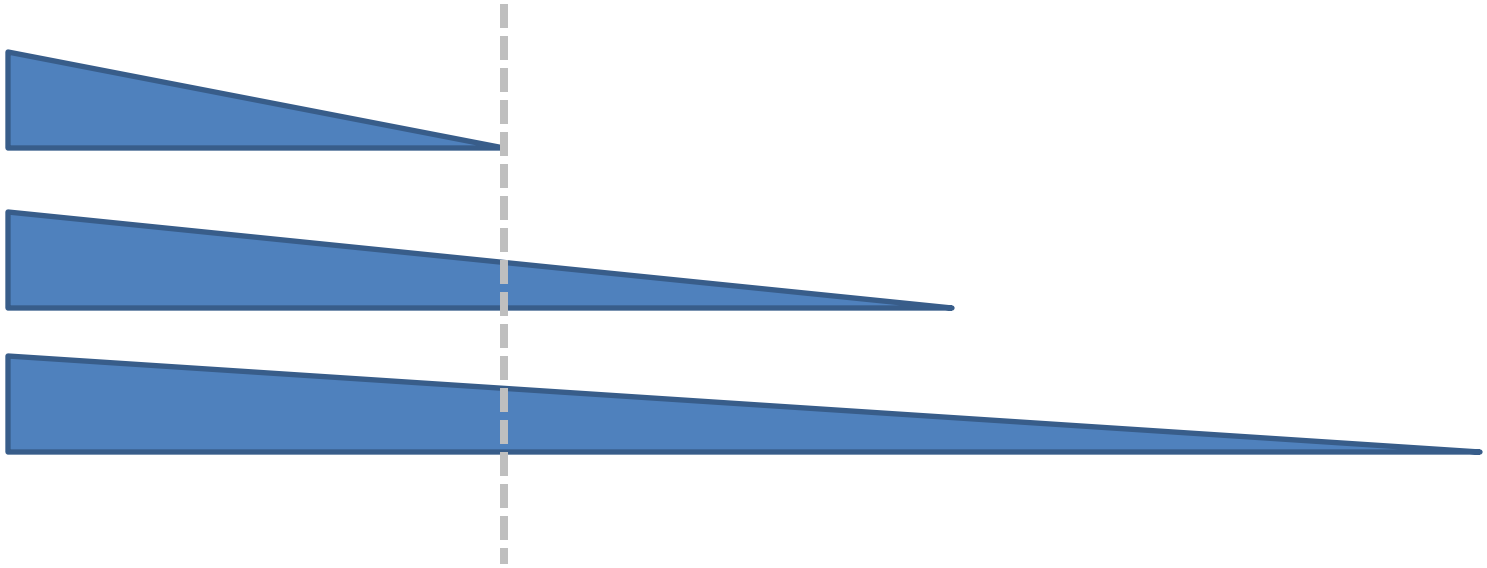
Aggregating Data



Global start/end = 200MB/s = WRONG

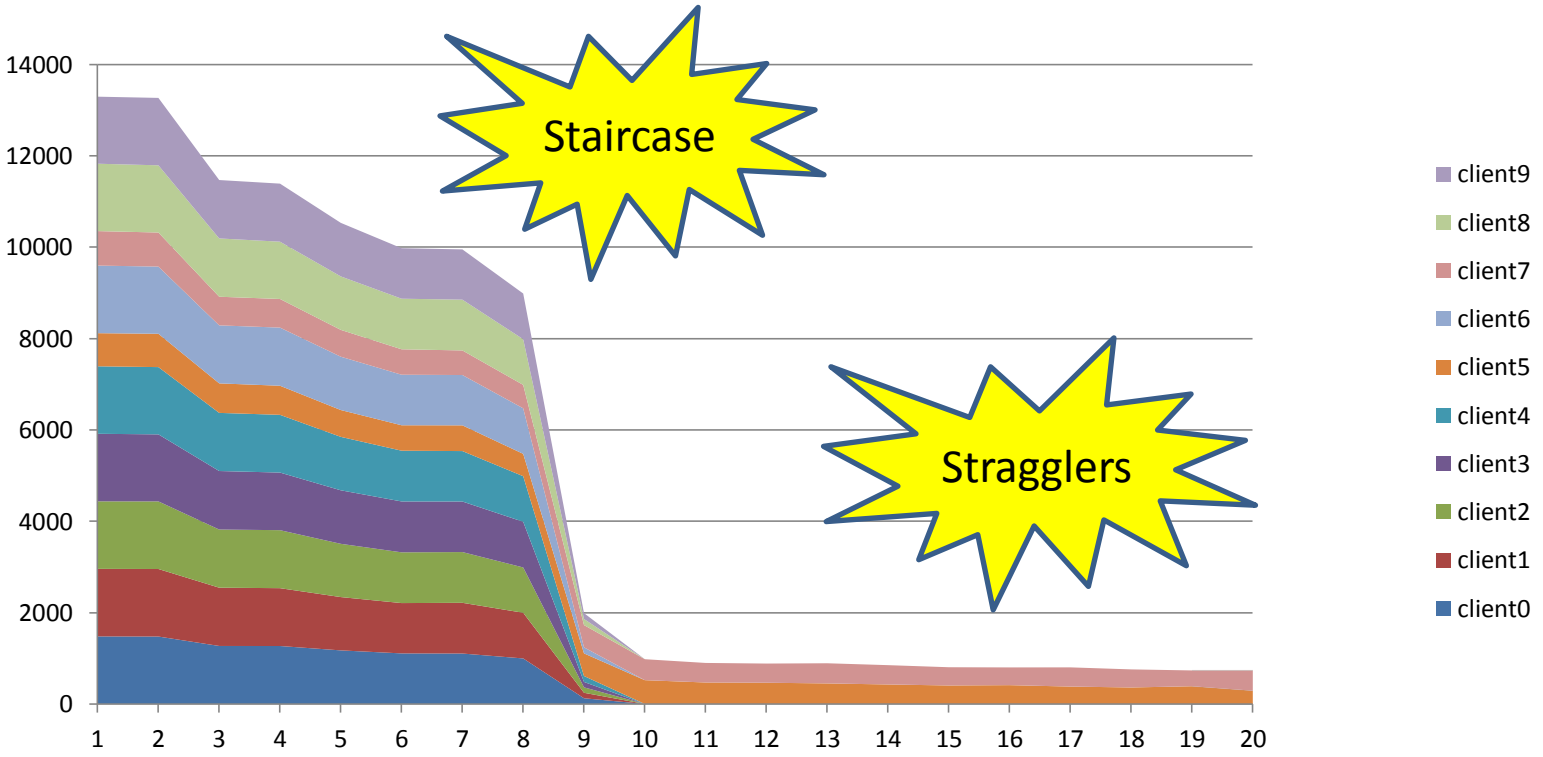
Additive = 183MB/s = WRONG

Aggregating Data

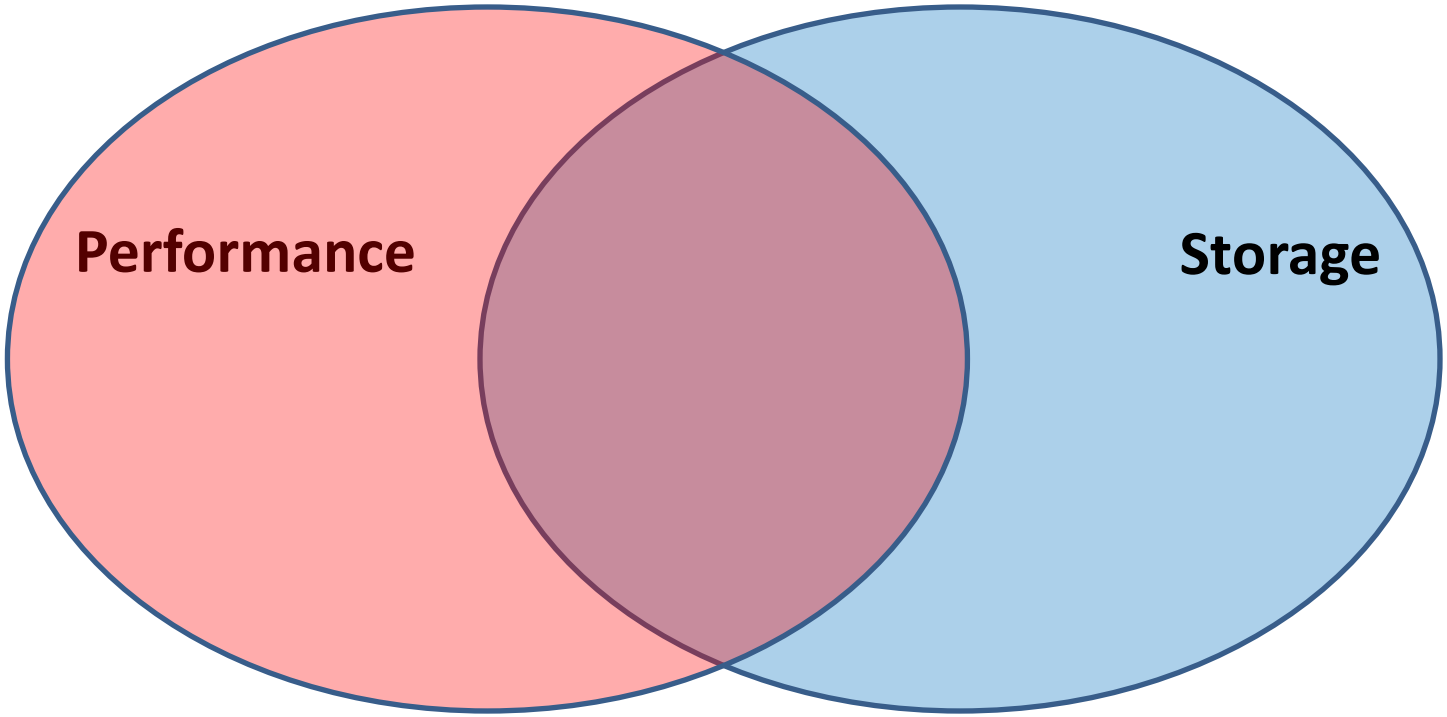


Stonewalling = 243MB/s = REALLY WRONG

Aggregating Data



- If your data is "X per second" then **measure per second**


Performance
Storage

Storage Performance Factors

Small requests vs. large requests

Data vs. metadata

Cached/buffered vs. synchronous

Read vs. write

Data Testing: iozone

- Can test block and file storage
- Data only
  - Sequential/random, O_SYNC/O_DIRECT, AIO, …
  - Could be better w.r.t. spatial distribution
  - Limited support for cluster testing
- "Stonewalling" by default
- Zillions of command-line options
- fio is very similar

iozone options

-c include close	-e include fsync
-o use O_SYNC	-O report ops/second
-r record size	-s file size
-l thread count	-C show child stats

+ 67 more

-i 0	sequential write
-i 1	sequential read
-I 9	random pwrite

Sample iozone output

Children see throughput for 4 rewriters = 49124.63 ops/sec
Parent sees throughput for 4 rewriters = 47810.32 ops/sec
Min throughput per process = 10874.55 ops/sec
Max throughput per process = 13628.46 ops/sec
Avg throughput per process = 12281.16 ops/sec
Min xfer = 831.00 ops

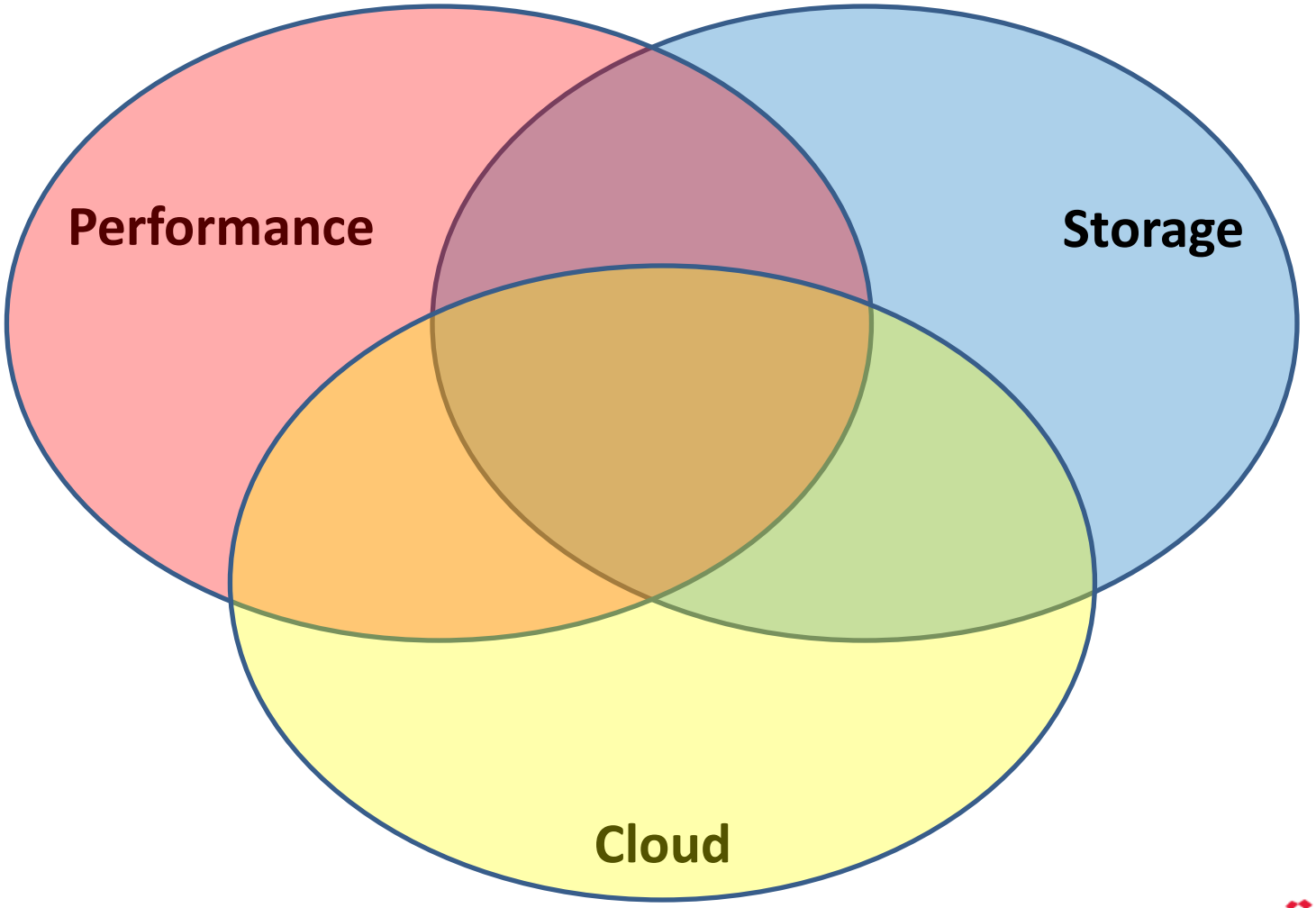
Child[0] xfer count = 1024.00 ops, Throughput = 13628.46 ops/sec
Child[1] xfer count = 950.00 ops, Throughput = 12620.25 ops/sec
Child[2] xfer count = 920.00 ops, Throughput = 12001.36ops/sec
Child[3] xfer count = 831.00 ops, Throughput = 10874.55 ops/sec

Metadata Testing

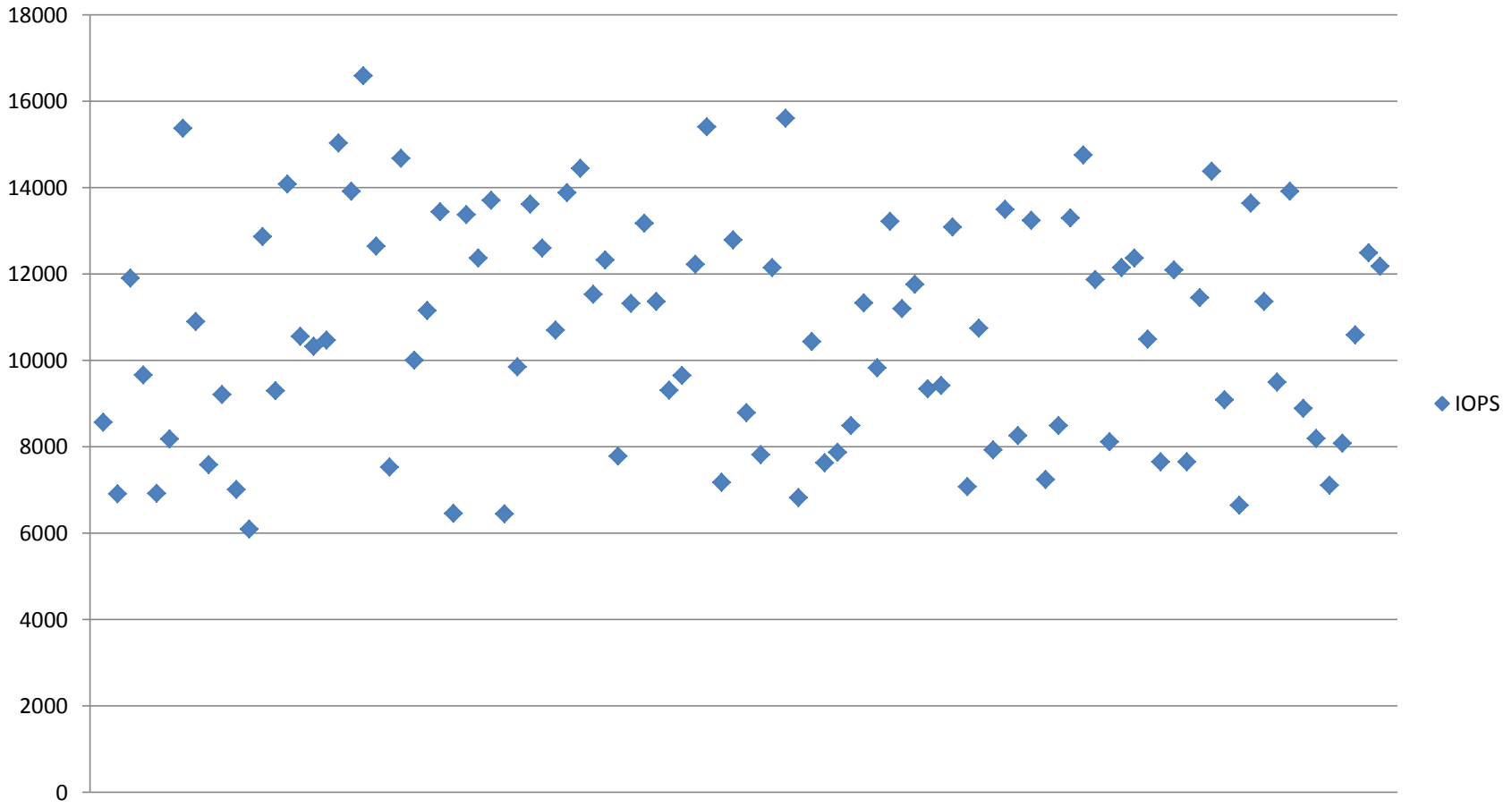
- Filebench
  - Workload Model Language
  - both data and synchronization
- Dbench
  - trace replay
- Somebody really needs to do better!

Object Testing

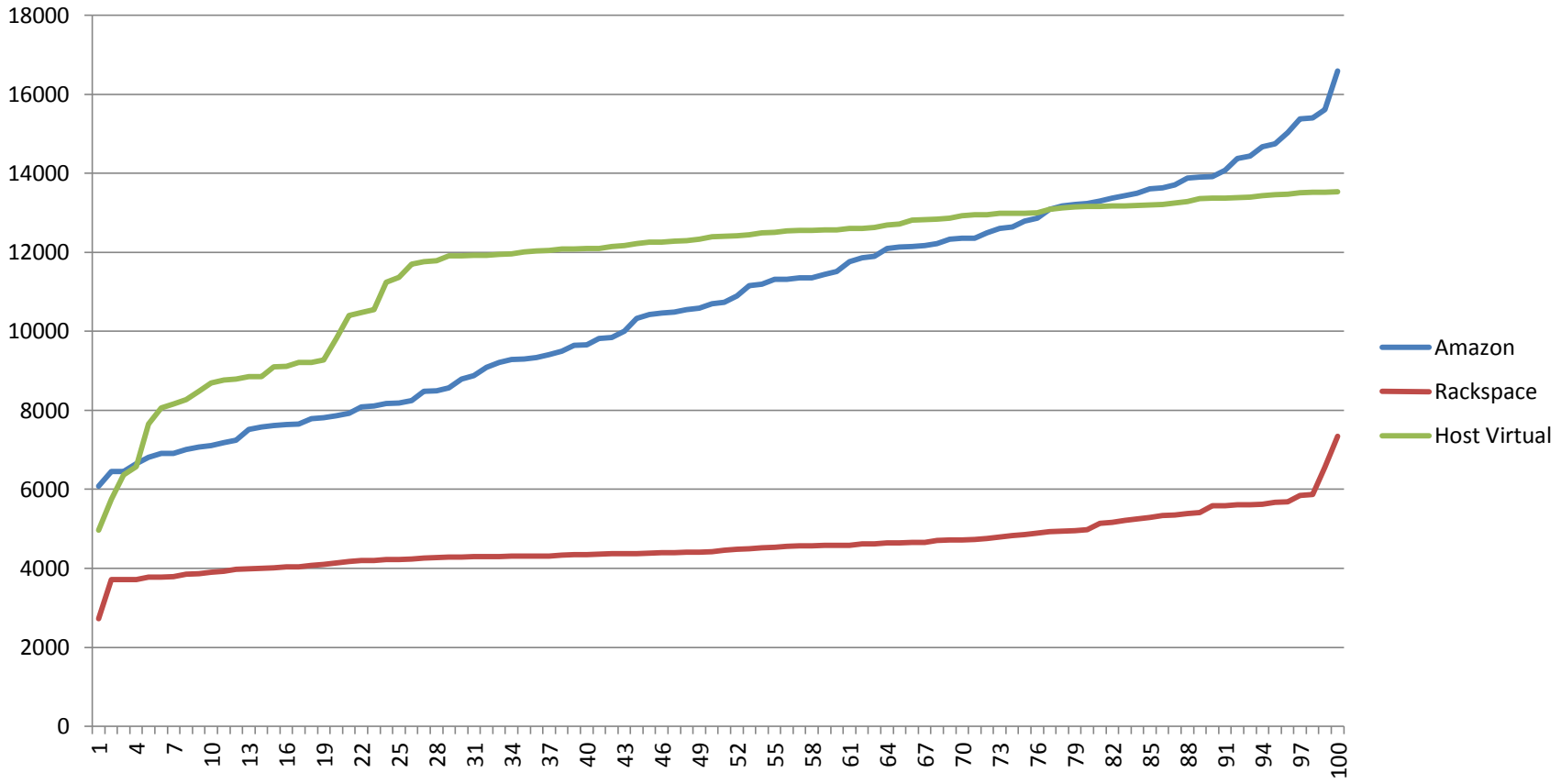
- COSBench
  - Object read, write, delete
  - Java + XML(ish) + Windows-style .ini files
  - Parallel and even distributed
  - still…
- Somebody needs to do **much** better!


Performance
Storage
Cloud

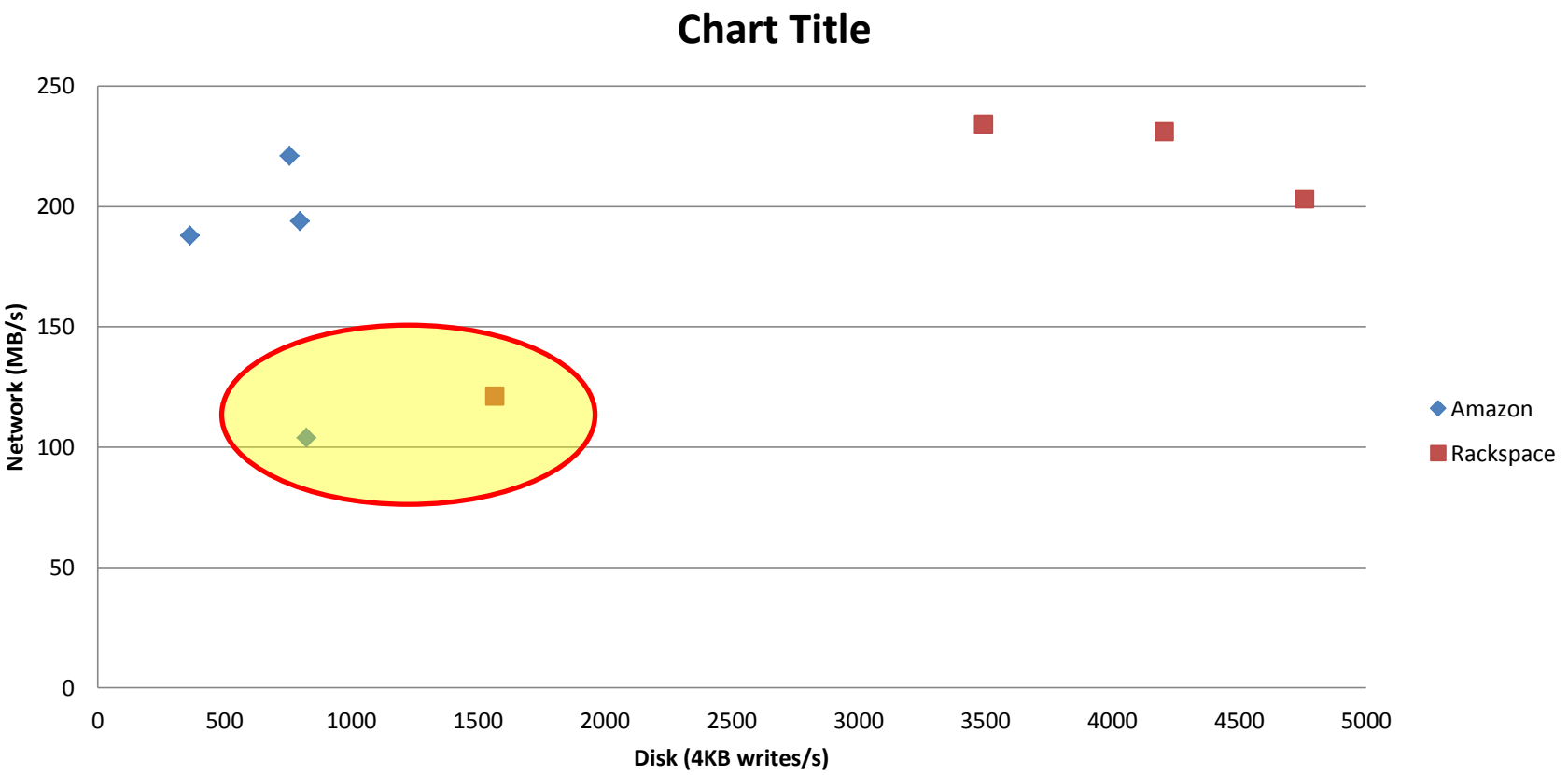
Noisy Neighbors



Performance Distributions



Performance Ratios



Other Problems

- "Clunker" instances
  - Netflix: kill and start a new one
- Network inconsistency
  - same host vs. same switch vs. ???
- Cheating
  - ignore O_SYNC, non-battery-backed cache

Conclusions

- Massive variability is what makes this hard
  - test many types, many instances, many times
  - reduce other variables (e.g. workloads)
  - automate, automate, automate
- Think in terms of probabilities instead of averages
- Use mathematical models or simulation to determine appropriate "insurance level"

Modeling Example

- Goal: 99% probability of 100K IOPS
- Same data as above

Provider	Ideal	From Model	Ratio
Amazon	7	13	1.86
Rackspace	14	28	2.00
Host Virtual	8	11	1.38

http://hekafs.org/index.php/2013/05/performance-variation-in-the-cloud/

Thank You!

jeff@pl.atyp.us