goto; conference aarhus

October 1-3, 2012
gotocon.com

Apache Cassandra As A BigData Platform

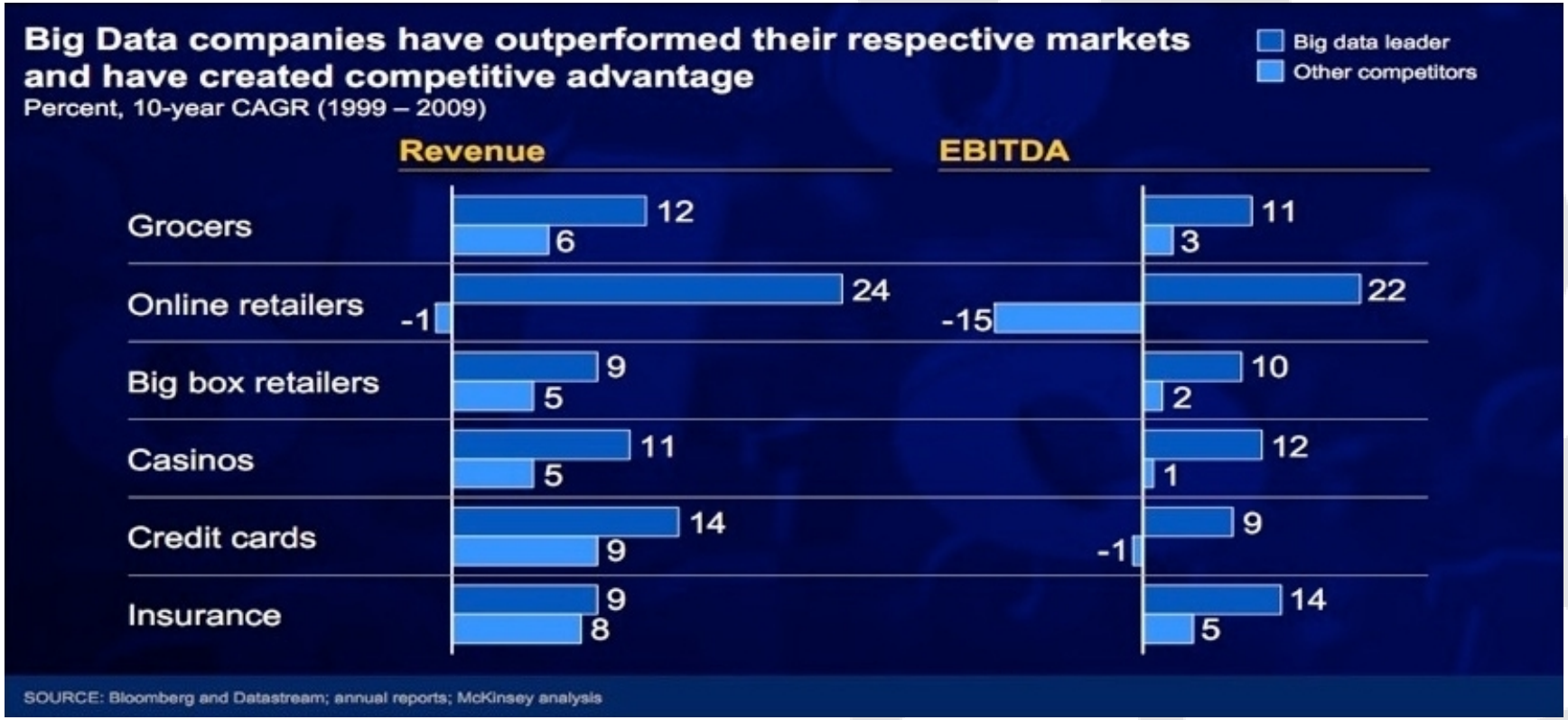
Matthew F. Dennis // @mdennis

DATASTAX

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

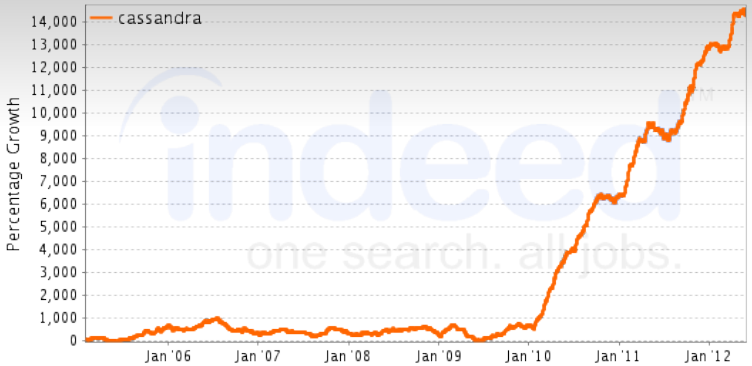
Why Does BigData Matter?

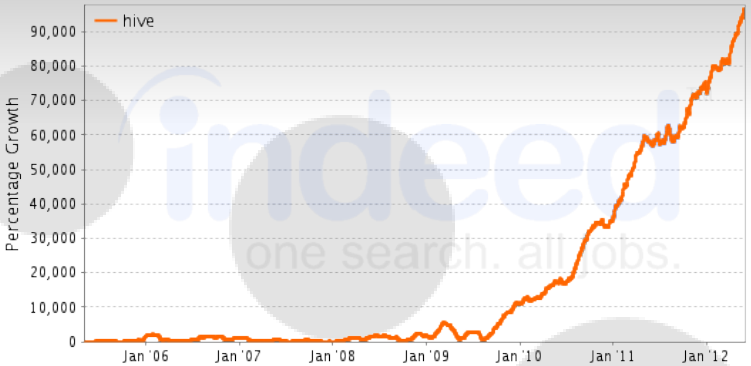
Effective Use of BigData Leads To Success

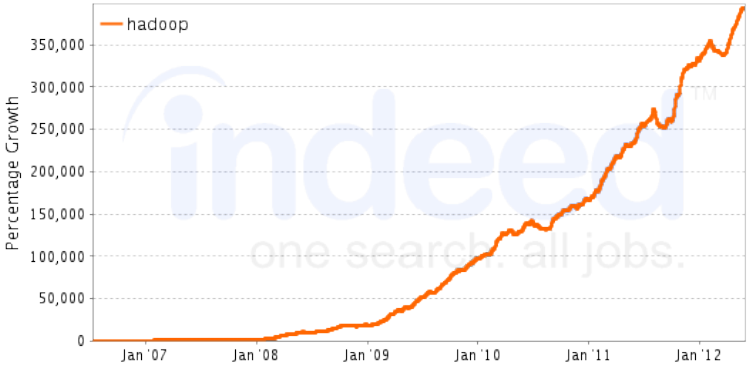


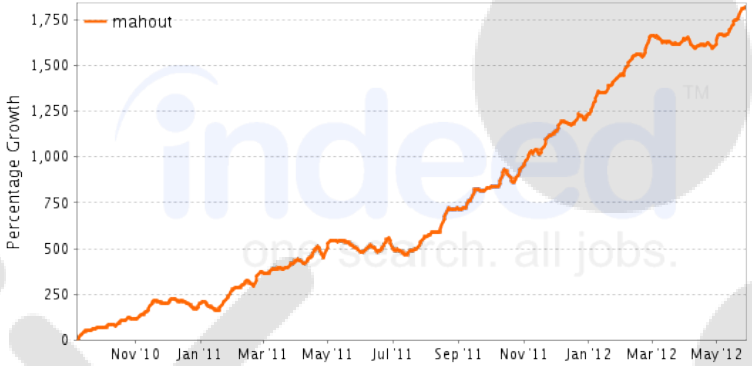
And The Trends Continue …

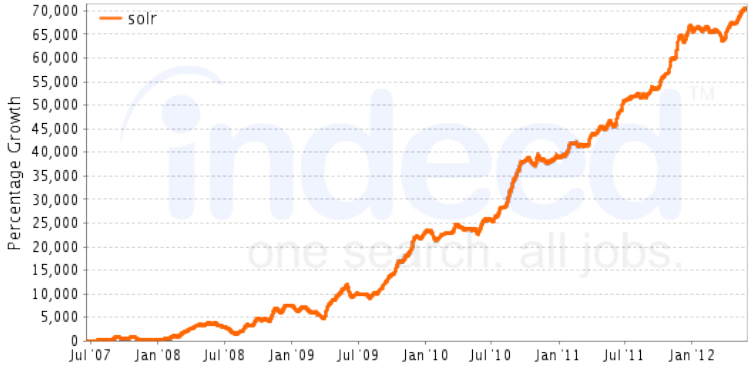
(according to indeed.com job trends)

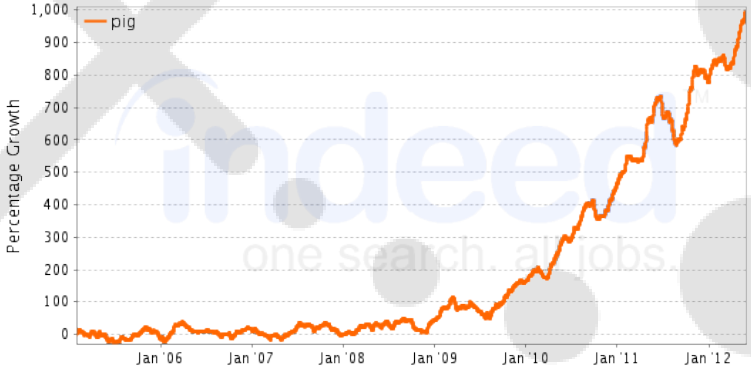












Requirements Of A BigData Platform

(*necessary* but not *sufficient* requirements)

- Performance
- Scalability
- Availability

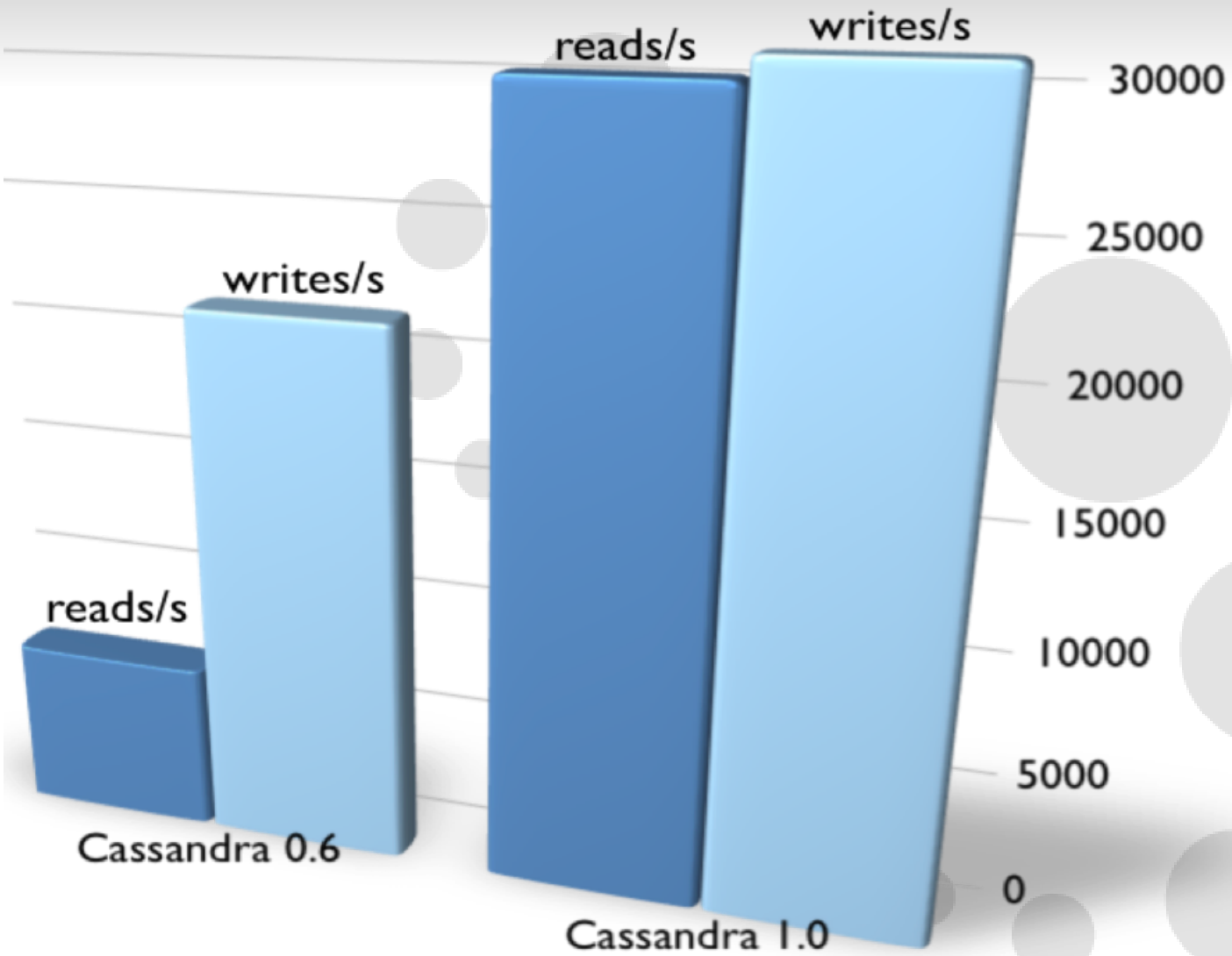
Measuring Performance

- throughput (in ops/sec)
- latency (for a single request)

Uncommon Cassandra Performance Features

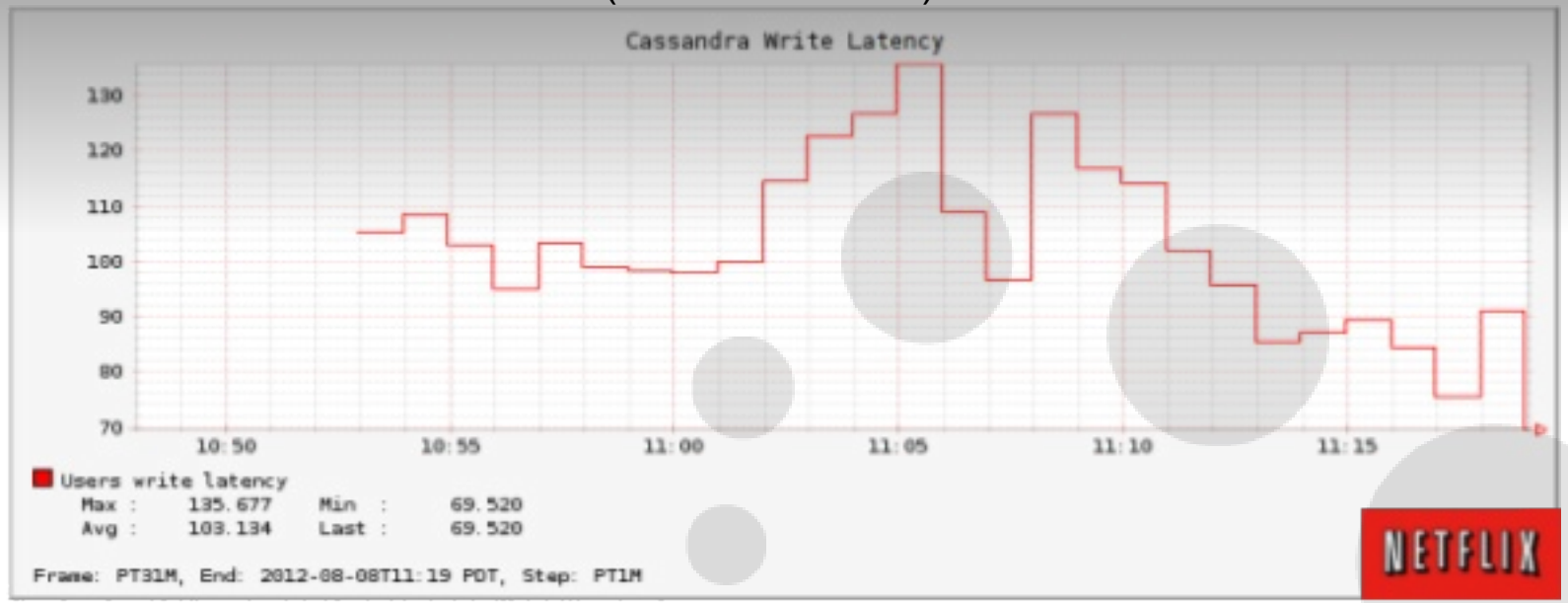
- No Locking (in the fast path)
- Log Structured Storage Engine
- Highly Parallel
- No BTrees
- On Disk Compression
- No “Master” Nodes

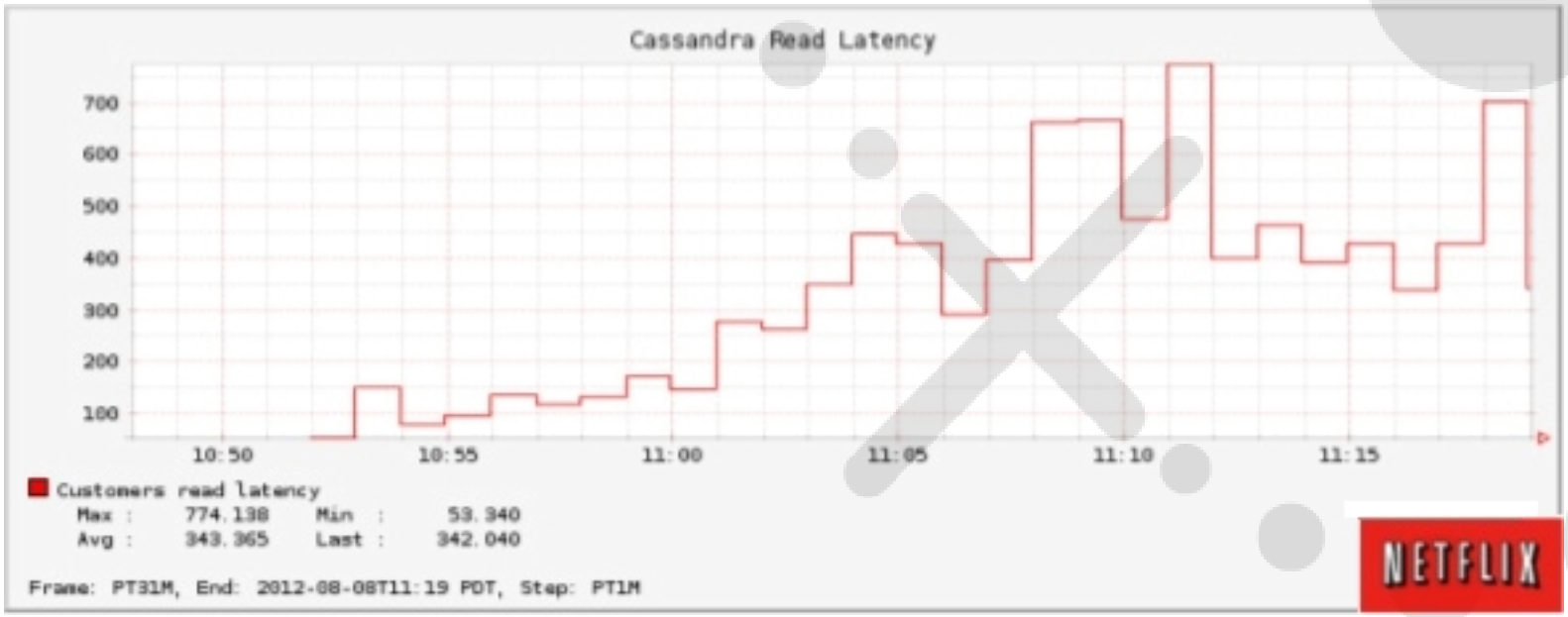
Cassandra Throughput



Cassandra Latency

(in microseconds)





http://www.slideshare.net/adrianco/cassandra-performance-on-aws

Cassandra + SSD

But can you get such low latency and high throughput for **random reads** from disk?

Yes, with Cassandra + SSDs
(SSD latency is usually only ~100 us)

A Random Note About C* and SSDs

- Cassandra can use cheap consumer grade **MLC** SSDs (~$1.00 USD / GB)
- no in-place updates results in far fewer erase cycles on the drive which results in the drive lasting longer
- Compared to nearly all other databases, consumer SSDs last ~10x longer on C*
- to put it another way MLC drives last about as long with C* as enterprise SLC drives last with most other databases

http://www.anandtech.com/show/5518/a-look-at-enterprise-performance-of-intel-ssds

Why Not On Rotational Disks Too?

- rotational disks **require** ~8ms per seek
- note that this is a **HW limitation**, an absolute upper limit (for that HW)
- **no** system can do better than the seek time when randomly retrieving data from disk (and most do far worse)

What About Writes/Updates?

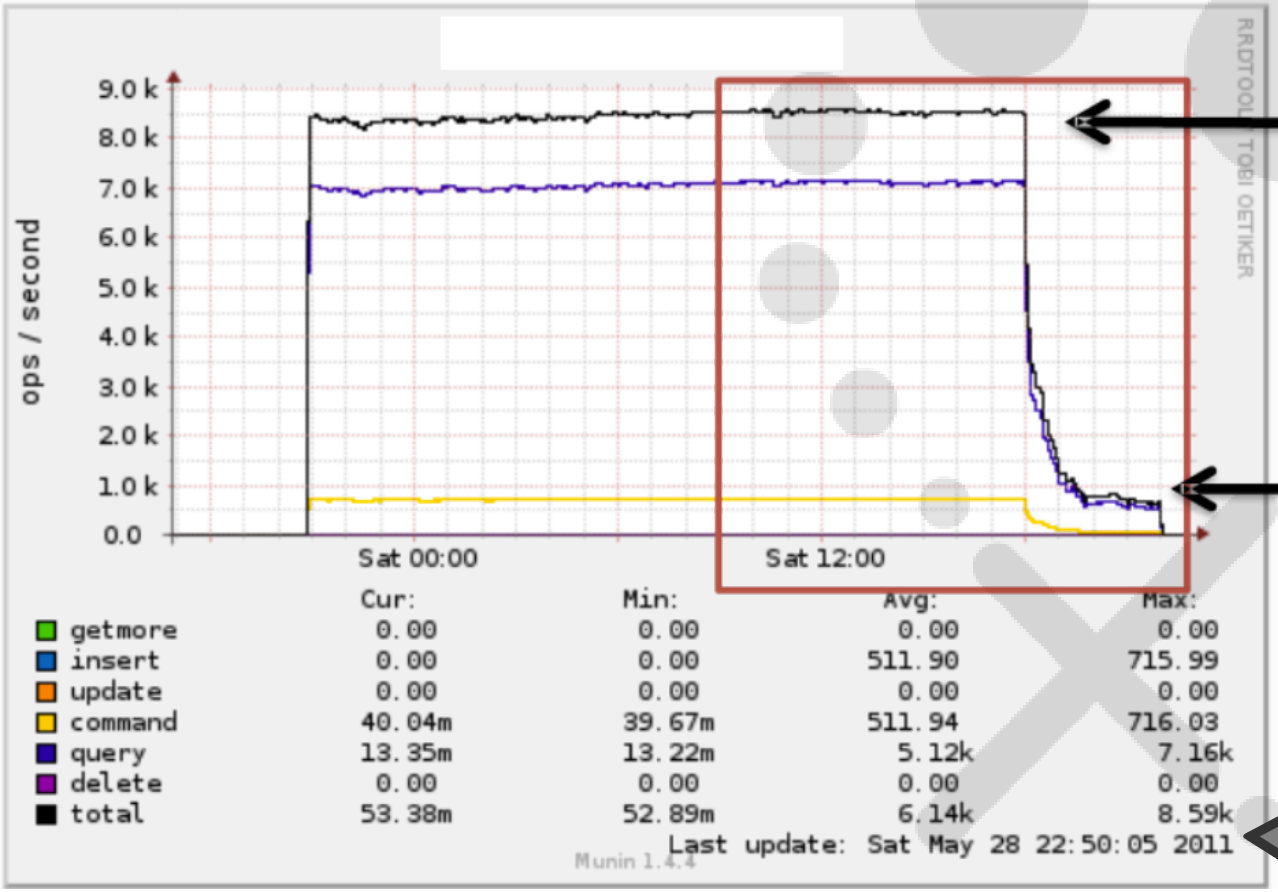
- **all** write I/O in Cassandra is sequential
- no global write lock
- no BTrees
- compare to MySQL, BerkeleyDB, MongoDB, Oracle, et cetera which either lock (sometimes with a global lock) and/or generate random writes for updates (and/or inserts)
- locking is not the only way to handle concurrency !!!

Larger Than Memory Datasets

- write performance degrades only marginally as the dataset outgrows memory; Cassandra exhibits essentially no change in latency or throughput

- read performance degrades gracefully and is relative to the percent of data in memory

Most Systems Do Not Behave That Way



In RAM

Not In RAM

old, but ...

Performance Compared to HBase

- 10x better read throughput
- 8x better write throughput
- 8x better read latency
- 10x better write latency
  (even when HBase was running without durability)

University of Toronto, Canada
Middleware Systems Research Group, et al
38th International Conference On Very Large Data Bases
http://vldb.org/pvldb/vol5/p1724_tilmannrabl_vldb2012.pdf

And We’re Not Even Finished Yet ...

- native CQL transport layer
- unified off-heap row and key caches
- vnodes
- no read-before-write on secondary indexes
- native JBOD support
- straight up profiler driven optimizations

Measuring Scalability

If performance is measured in throughput and latency, then scalability is the stability of latency as throughput increases (or the stability of latency and throughput as “load” increases); essentially scalability is how well a system handles growth

Linear Scalability

If for all values of X, Y, Z, C and N:

latency=X
throughput=Y
"load"=**Z**
nodes=**N**

latency=X
throughput=Y
"load"=**CZ**
nodes=**CN**

Then: the system is perfectly linearly scalable with respect to "load"

Linear Scalability

If for all values of X, Y, Z, C and N:

latency=X
throughput=**Y**
"load"=Z
nodes=**N**

latency=X
throughput=**CY**
"load"=Z
nodes=**CN**

Then: the system is perfectly linearly scalable with respect to throughput

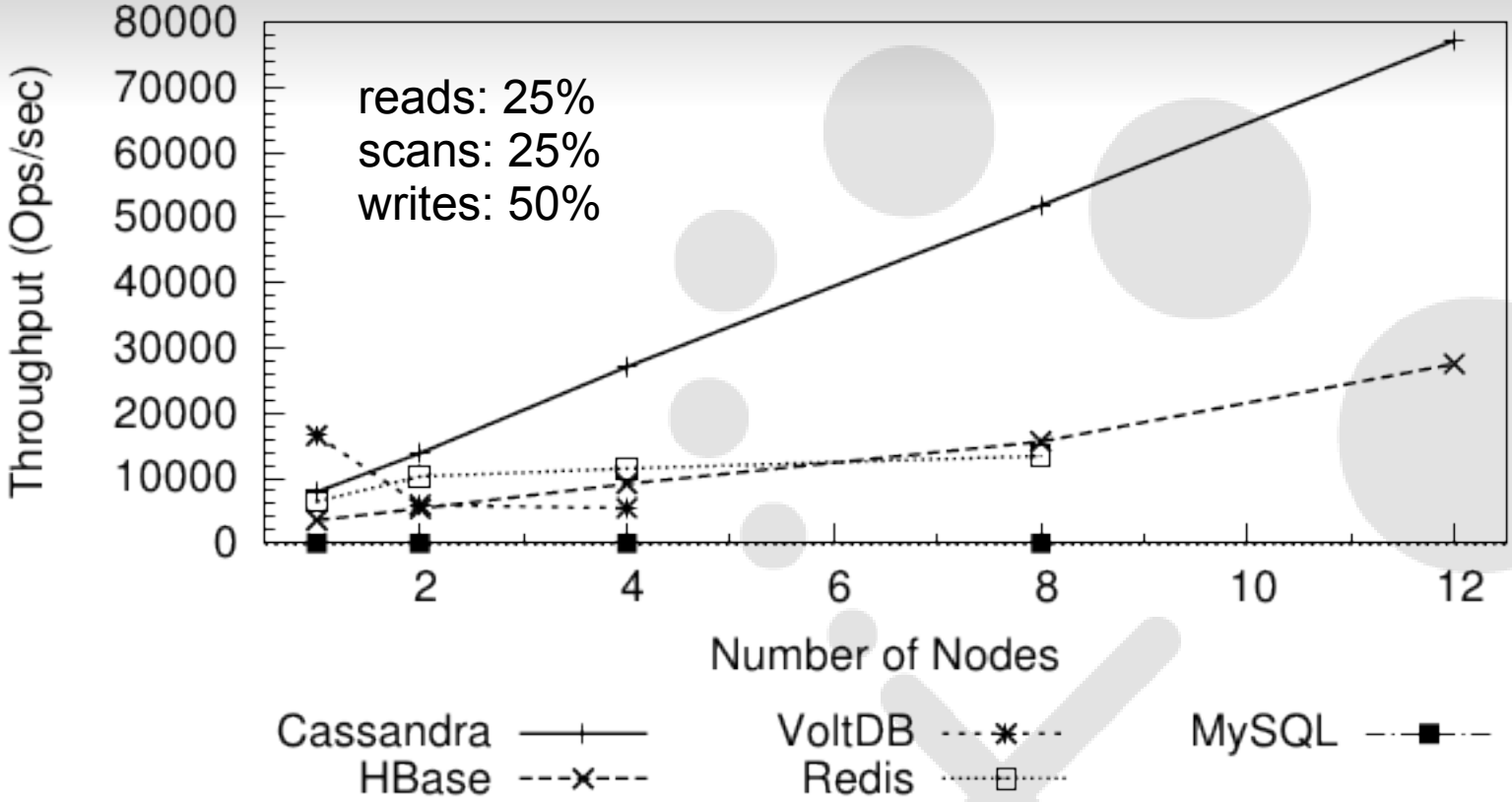
So, How Does Cassandra Stack Up To That Definition?

Cassandra Scalability

"In terms of scalability, there is a **clear winner throughout our experiments**. Cassandra achieves the highest throughput for the maximum number of nodes in all experiments with [nearly linear] increasing throughput from 1 to 12 nodes."

University of Toronto, Canada
Middleware Systems Research Group, et al
38th International Conference On Very Large Data Bases
http://vldb.org/pvldb/vol5/p1724_tilmannrabl_vldb2012.pdf

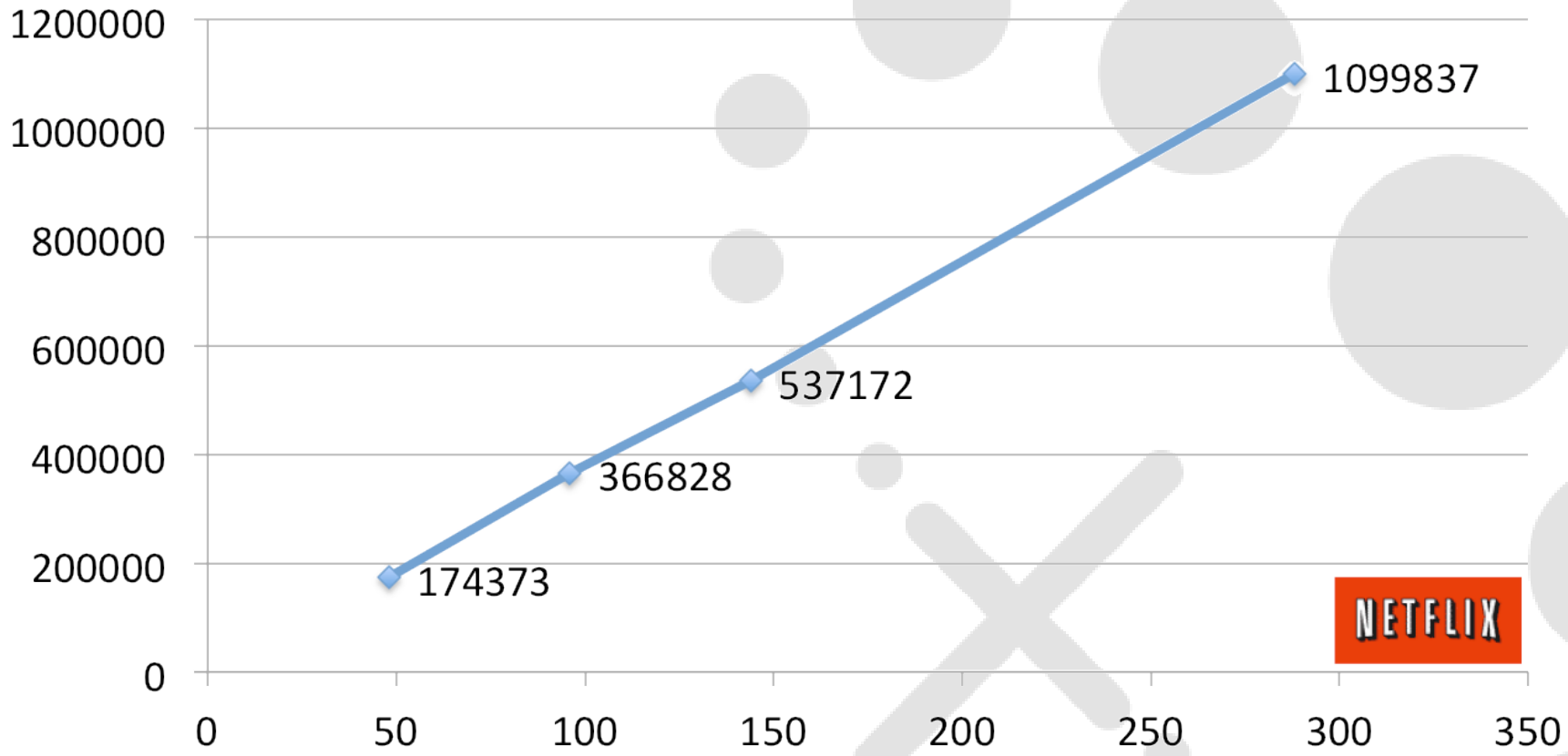
Cassandra Scalability



University of Toronto, Canada
Middleware Systems Research Group, et al
38th International Conference On Very Large Data Bases
http://vldb.org/pvldb/vol5/p1724_tilmannrabl_vldb2012.pdf

Cassandra Scalability

(client operations per second at RF=3)



http://techblog.netflix.com/2011/11/benchmarking-cassandra-scalability-on.html

Requirements Of A BigData Platform

✓ Performance

✓ Scalability

- Availability – performance and scalability are easy if you ignore availability and/or assume failures never happen

Measuring Availability

availability is measured by the amount of downtime a system has over a given period of time

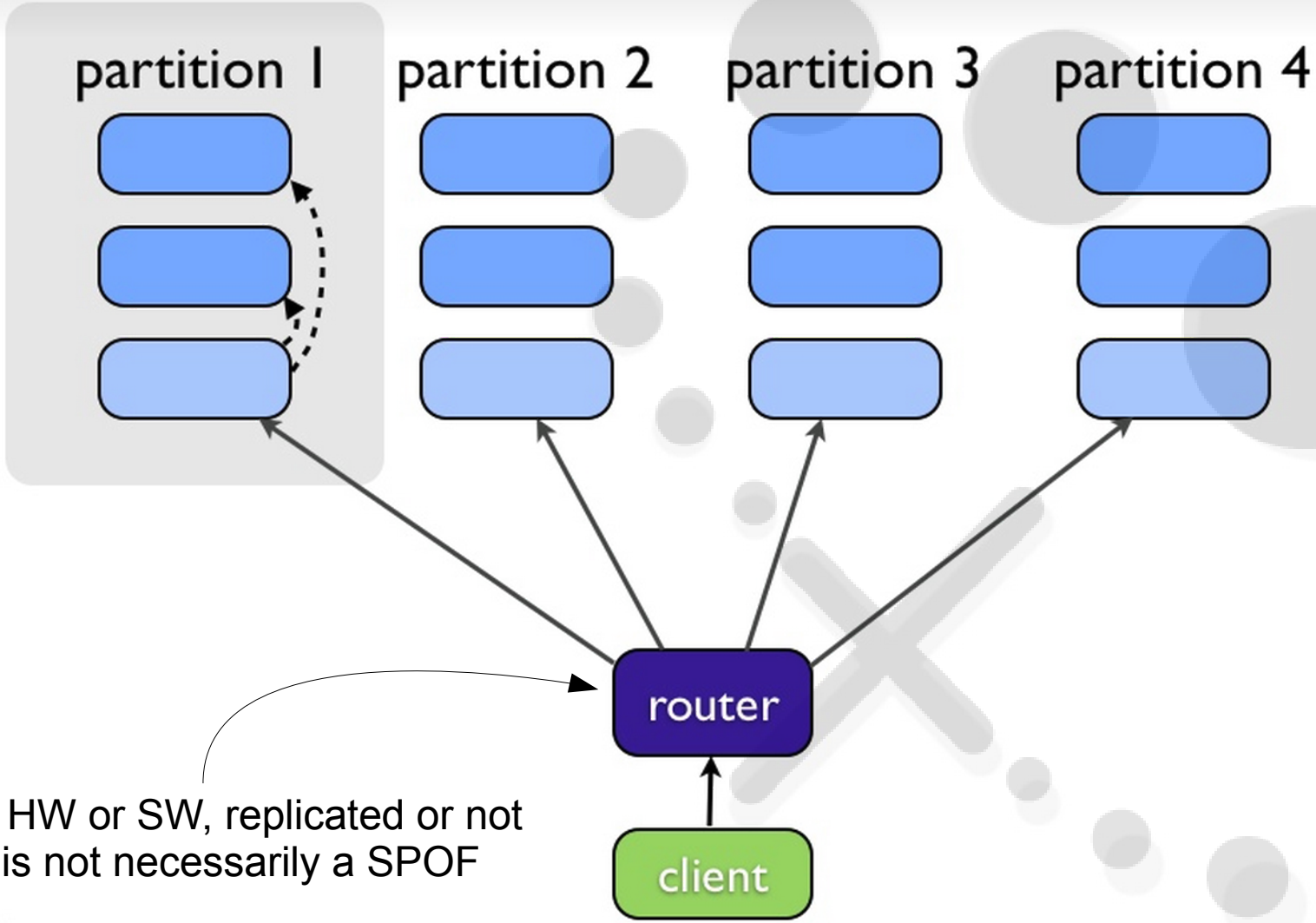
Common Causes of Downtime In Large Scale Systems

- Component Failure (disk)
- Machine Failure (NIC, cpu, power supply)
- Rack Failure (router, switch, UPS, AC)
- Site Failure (power grid, natural disaster, war, coup)

The Common Theme In The Solutions?

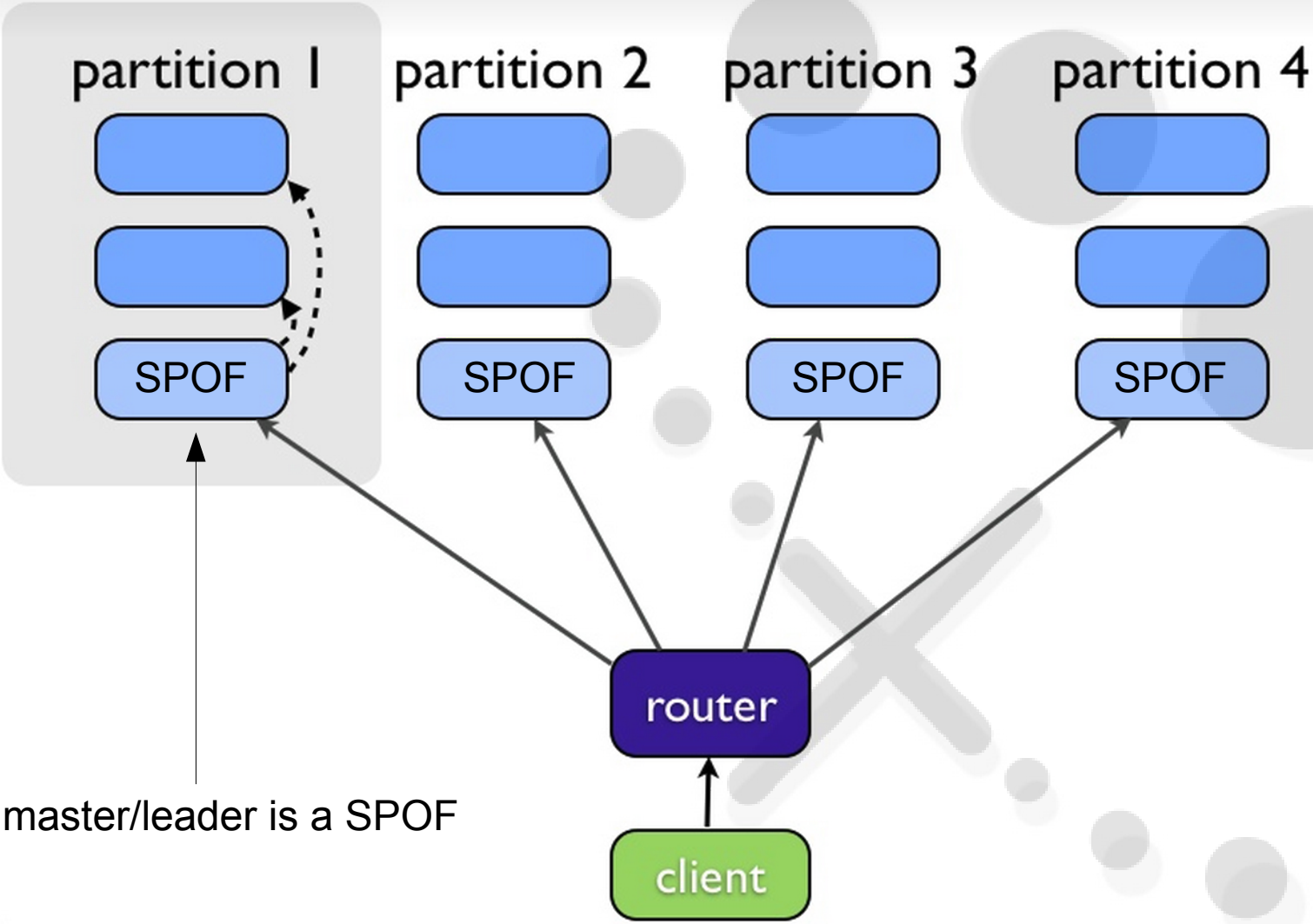
Replication

Legacy Replication



can be HW or SW, replicated or not
this is not necessarily a SPOF

Legacy Replication



the master/leader is a SPOF

Thoughts On Availability (and legacy replication)

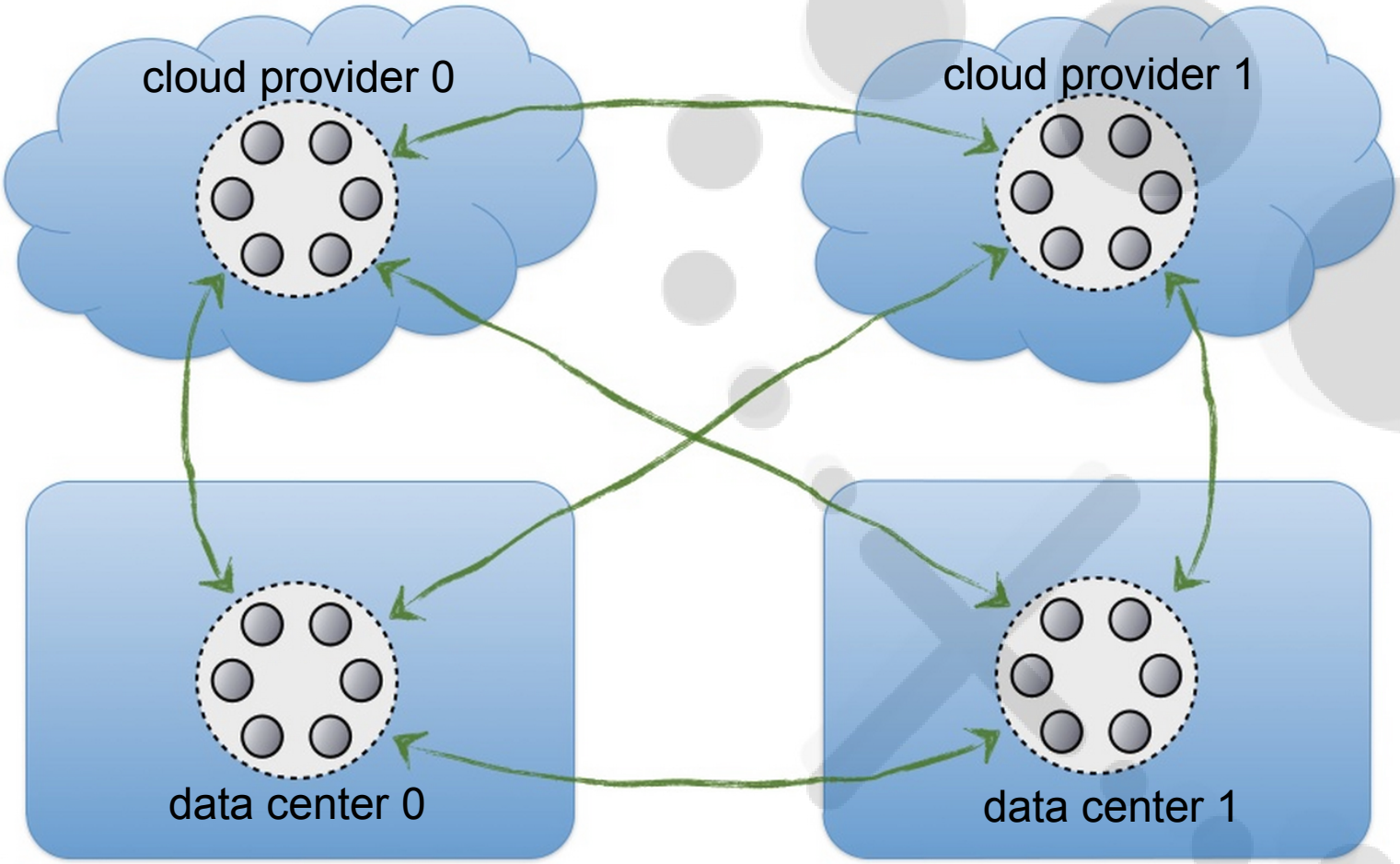
“High availability implies that a single fault will not bring down your system. Not ‘we’ll recover quickly.’”

-- Ben Coverston, DataStax

“The biggest problem with failover is that you’re almost never using it until it really hurts. Its like backups that you never test.”

-- Rick Branson, Instagram

Cassandra Replication



*complicated networking be damned, at least it is *possible* with Cassandra ...

More Thoughts On Availability (and legacy replication)

mdennis
@mdennis

Follow

"active/passive", "shared" and "standby" are not phrases found in the description of actual "high availability" systems

Reply Retweet Favorite

Bill de hÓra
@dehora

Follow

Coming to the conclusion that #cassandra is kind of indestructible. "Robust" doesn't do it justice.

Reply Retweet Favorite

Eric Florenzano
@ericflo

Follow

"Cassandra ... dealt with the loss of one third of its regional nodes without any loss of data or availability." techblog.netflix.com/2012/07/lesson... - Nice!

Reply Retweet Favorite

Aaron Turner
@synfinatic

Follow

took me 10hrs to notice a #cassandra node had a hw failure because everything just kept working. #sweet

Reply Retweet Favorite

Continuous Global Availability ...

... requires more than being able to recover from faults, it requires being able to tolerate the faults without downtime in the first place

if you care about

continuous global availability

then you must serve reads and writes from

multiple geographical locations

there is no alternative

Cassandra As A BigData Platform

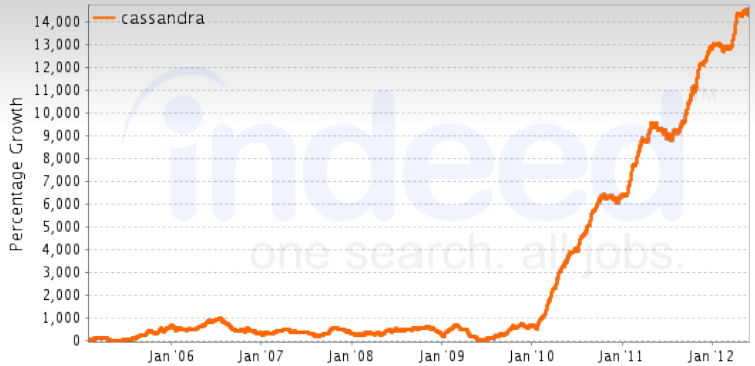
- ✓ Performance
- ✓ Scalability
- ✓ Availability

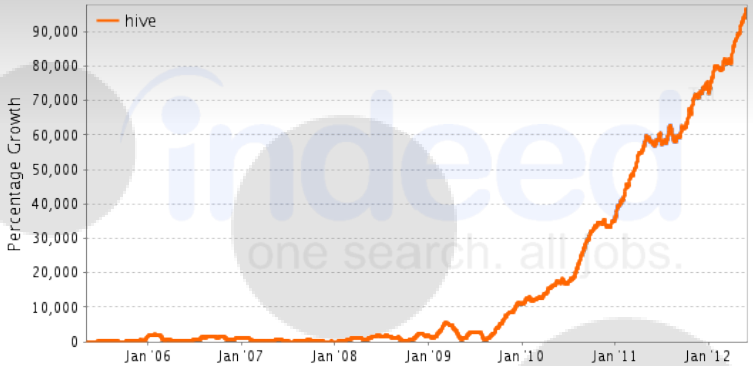
And now back to our trends ...

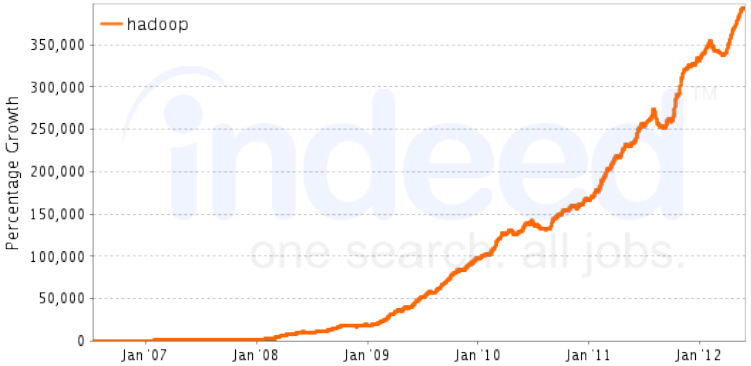
Public Cassandra Users Early 2011

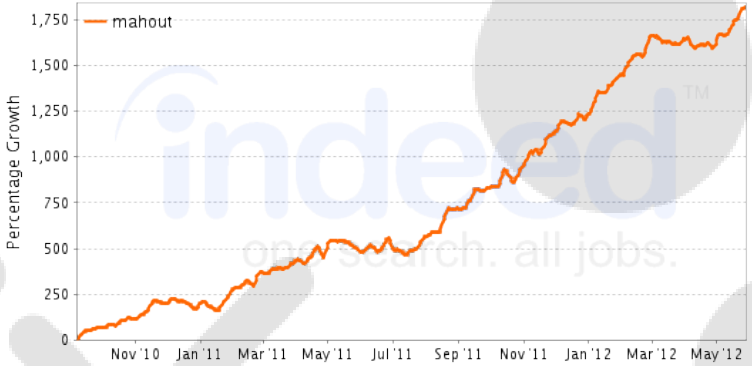
Public Cassandra Users Mid 2012

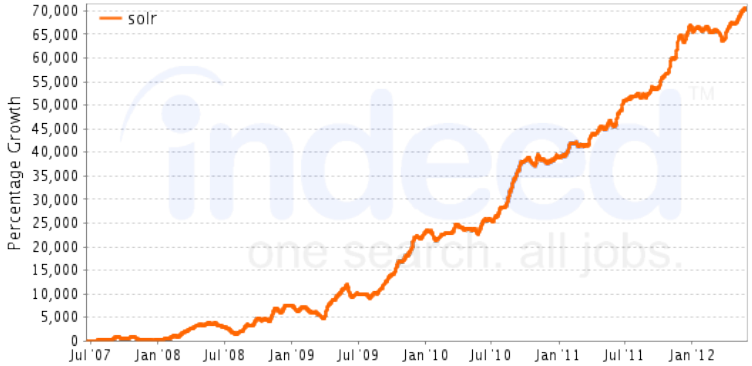
DataStax Enterprise

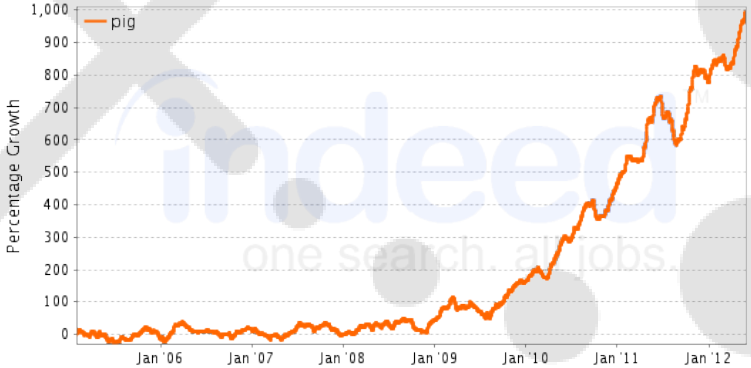




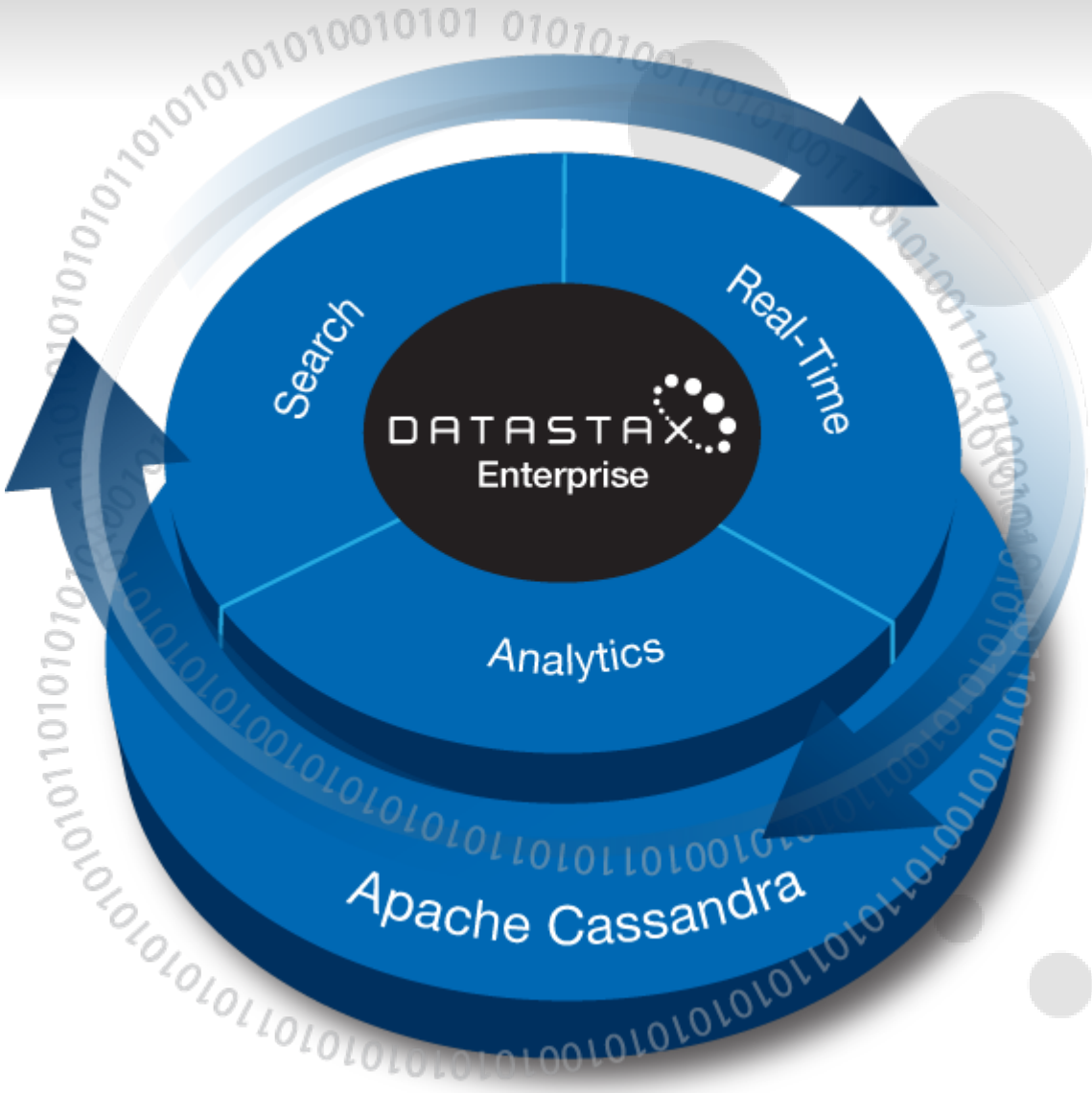








DataStax Enterprise



Q?

Matthew F. Dennis // @mdennis
http://slideshare.net/mattdennis

DATASTAX

Thank You!

Matthew F. Dennis // @mdennis
http://slideshare.net/mattdennis

A Quick Side Note ...

Cassandra Replication Follows The Dynamo Model *

http://www.allthingsdistributed.com/2007/10/amazons_dynamo.html

Read It!

*Cassandra is not a strict reimplementation of dynamo