SCYLLA:
NoSQL at Ludicrous Speed

主讲人：ScyllaDB软件工程师 贺俊

Today we will cover:

- **Intro: Who we are, what we do, who uses it**
- Why we started ScyllaDB
- Why should you care
- How we made design decisions to achieve no-compromise performance and availability

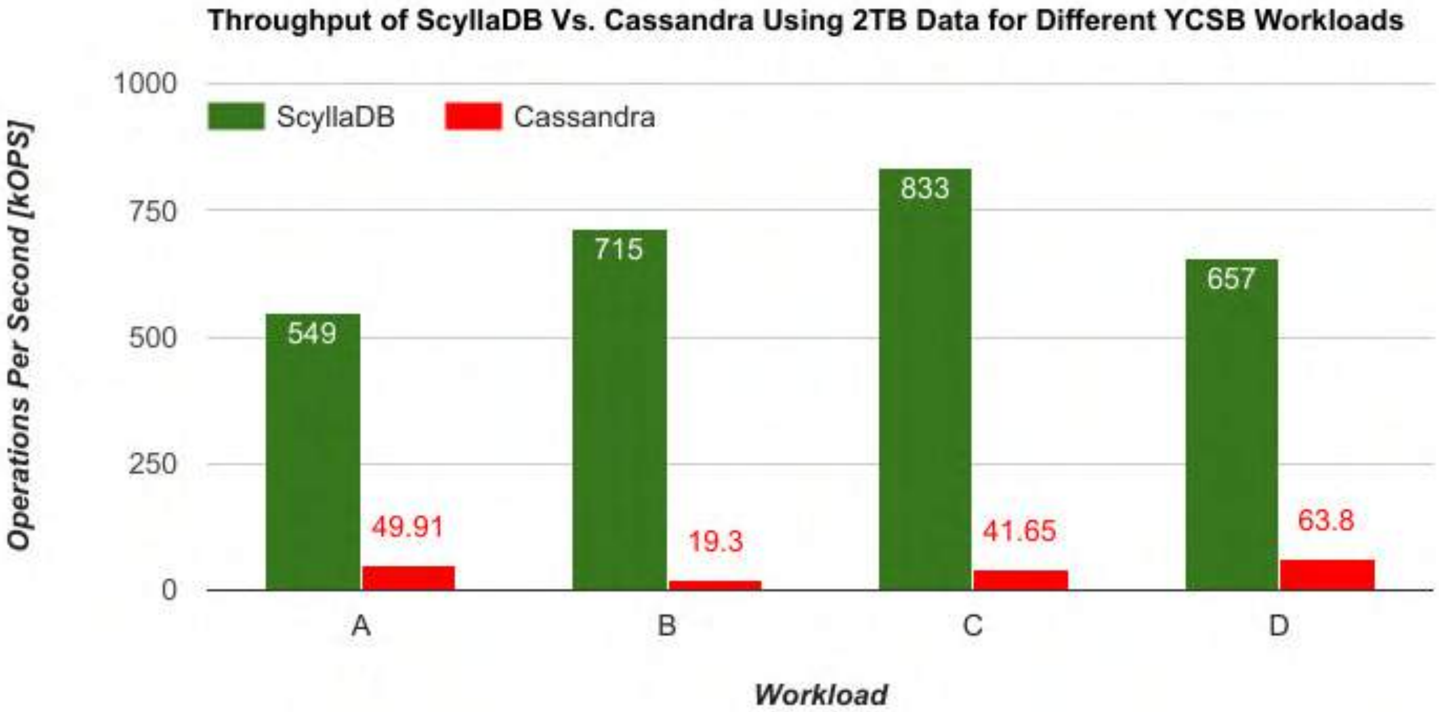
Introduction

- Founded by KVM hypervisor creators
- Q2 2014 - Pivot to the database world
- Q3 2015 - Decloak during Cassandra Summit 2015, Beta
- Q1 2016 - General Availability
- Q3 2016 - First Scylla Summit: 100+ Attendees
- Q1 2017 - Completed B round
- $25MM in funding
- HQs: Palo Alto, CA; Herzelia, Israel
- 42+ employees, hiring!

WD Western Digital · BESSEMER VENTURE PARTNERS · INNOVATION ENDEAVORS · magma Venture Partners · WiNG · QUALCOMM VENTURES · SAMSUNG VENTURE INVESTMENT

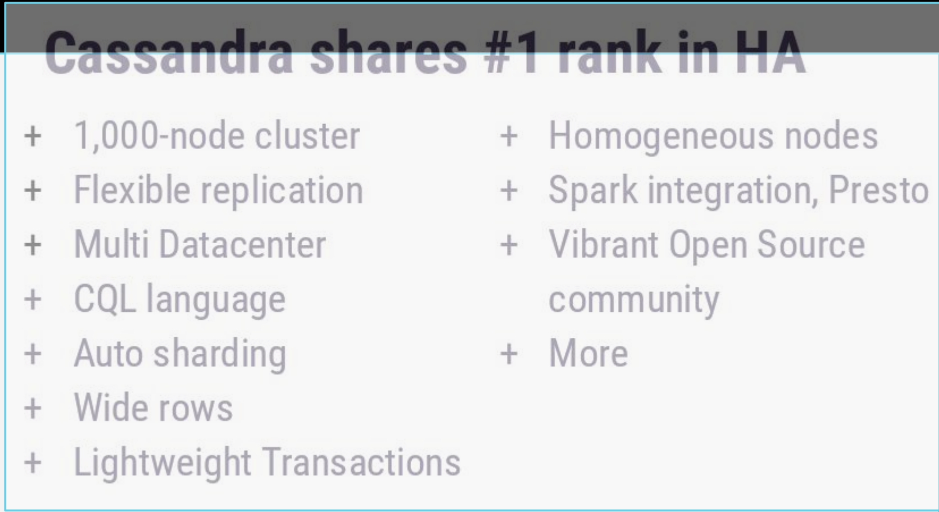
Why ?@#$%$$%^?

Scylla benchmark by Samsung



What we do: Scylla, towards the best NoSQL

- > 1 million OPS per node
- < 1ms 99% latency
- Auto tuned
- Scale up and out
- Open source
- Large community (piggyback on Cassandra)
- Blends in the ecosystem- Spark, Presto, time series, search, ..



Cassandra shares #1 rank in HA

- 1,000-node cluster
- Flexible replication
- Multi Datacenter
- CQL language
- Auto sharding
- Wide rows
- Lightweight Transactions
- Homogeneous nodes
- Spark integration, Presto
- Vibrant Open Source community
- More

Where Scylla is deployed?

AppNexus

IBM

SAMSUNG
SAMSUNG SDS

musical.ly

Kenshoo

COMPOSE
AN IBM COMPANY

Investing.com

cpi card group

xAd

D.A.Consortium

instaclustr

OLA

mparticle

LUCID

eniro

DigitalOcean

Outbrain

hulu

Dcard

MOGU
STREET

Today we will cover:

- **Intro: Who we are, what we do, who uses it**
- Why we started ScyllaDB
- Why should you care
- How we made design decisions to achieve no-compromise performance and availability

Why we started Scylla?

- Originally it was about performance/efficiency only
- Over time, we understood we can deliver more:
  - SLA between background and foreground tasks
  - Work well on any given hardware {back pressure}
  - Deliver consistent, low 99th percentile latency
  - Reduction in admin effort
  - Low latency under the face of failures (hot cache load balancing)
  - High observability

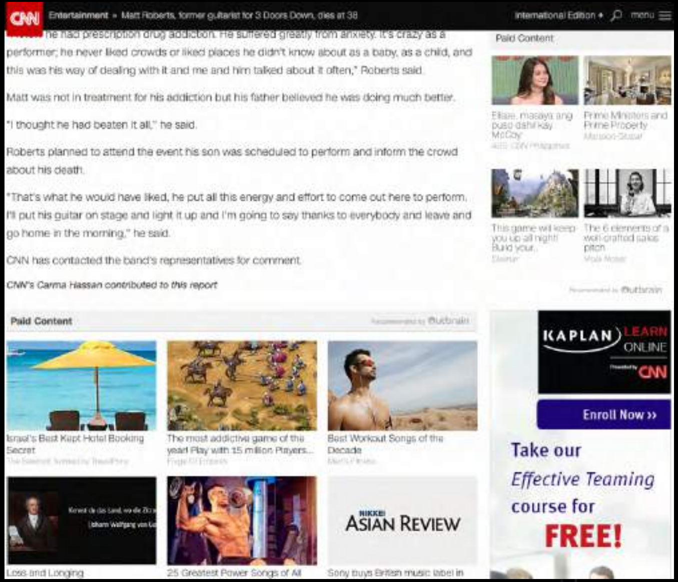
	Cassandra	Scylla
Throughput:	Cannot utilize multi-core efficiently	Scales linearly - shard-per-core
Latency:	High due to Java and JVM’s GC	Low and consistent - own cache
Complexity:	Intricate tuning and configuration	Auto tuned, dynamic scheduling
Admin:	Maintenance impacts performance	SLA guarantee for admin vs serving

Today we will cover:

- **Intro: Who we are, what we do, who uses it**
- Why we started ScyllaDB
- Why should you care
- How we made design decisions to achieve no-compromise performance and availability

Case study: Document column family

- Outbrain is the world’s largest content discovery platform.
- Over 557 million unique visitors from across the globe.
- 250 billion personalized content recommendations every month.



he had prescription drug addiction. He suffered greatly from anxiety. It's crazy as a performer; he never liked crowds or liked places he didn't know about as a baby, as a child, and this was his way of dealing with it and me and him talked about it often," Roberts said.

Matt was not in treatment for his addiction but his father believed he was doing much better.

"I thought he had beaten it all," he said.

Roberts planned to attend the event his son was scheduled to perform and inform the crowd about his death.

"That's what he would have liked, he put all this energy and effort to come out here to perform. I'll put his guitar on stage and light it up and I'm going to say thanks to everybody and leave and go home in the morning," he said.

CNN has contacted the band's representatives for comment.

CNN's Carma Hassan contributed to this report

Paid Content

Ellase, masaya ang puso dahil kay McCoy

Prime Ministers and Prime Property

This game will keep you up all night! Build your...

The 6 elements of a well-crafted sales pitch

Paid Content

Israel's Best Kept Hotel Booking Secret

The most addictive game of the year! Play with 15 million Players...

Best Workout Songs of the Decade

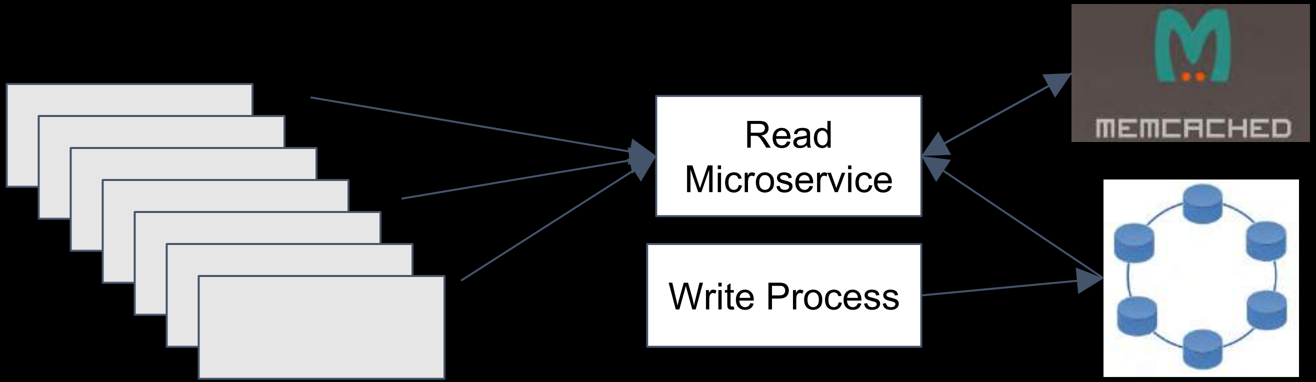
Loss and Longing

25 Greatest Power Songs of All

Sony buys British music label in

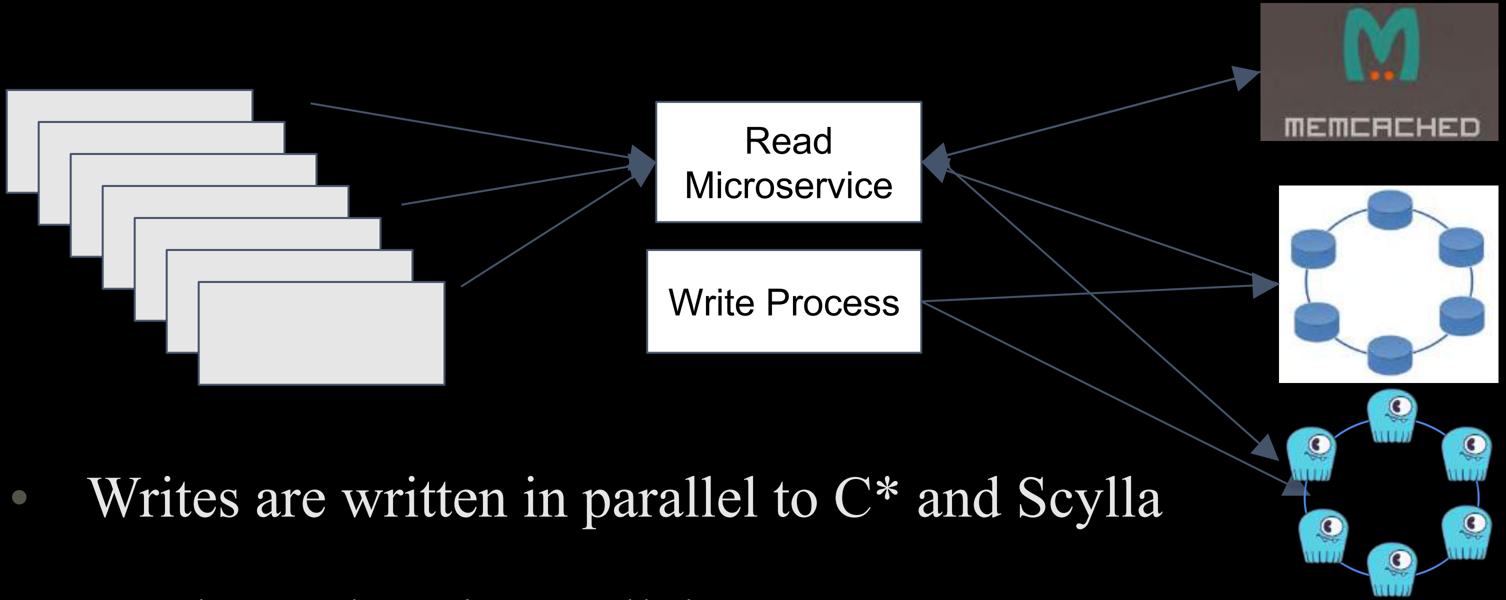
Take our *Effective Teaming* course for FREE!

Outbrain: Cassandra plus Memcache



- First read from memcached, go to Cassandra on misses.
- Pain: 1) Stale data from cache 2) Complexity 3) Cold cache -> C* gets full volume

Scylla/Cassandra side by side deployment

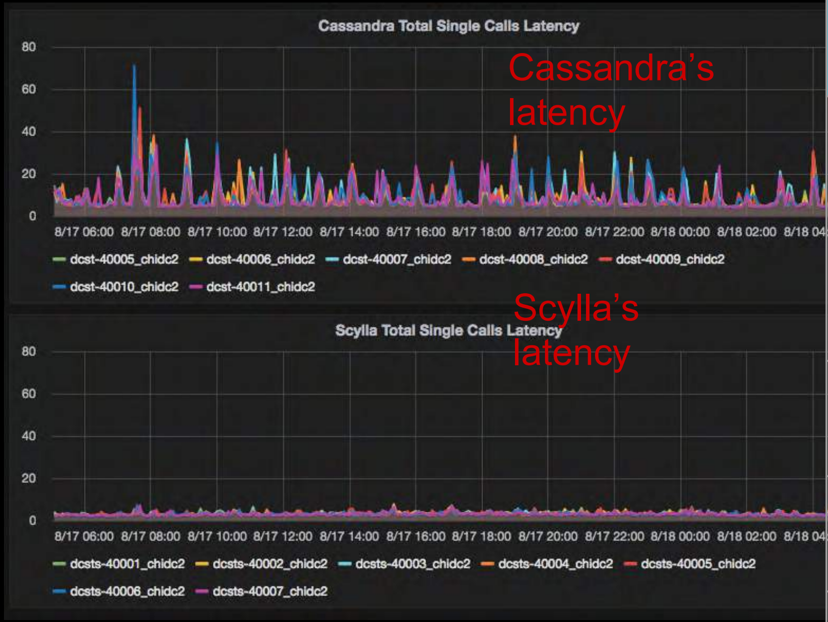


- Writes are written in parallel to C* and Scylla

- Reads are done in parallel:

1) Memcached + Cassandra 2) Scylla (no cache at all)

Scylla (w/o cache) vs Cassandra + Memcached



Cassandra’s latency

Scylla’s latency

	Scylla	Cassandra	Diff%
Requests/ Minute	**12,000,000**	**500,000 (memcache handles 11,500,000)**	**24X**
AVG Latency	**4 ms**	**8 ms**	**2X**
Max Latency	**8 ms**	**35 ms**	**3.5X**
Hardware	**9 machines**	**30+9 machines**	**4.3X**

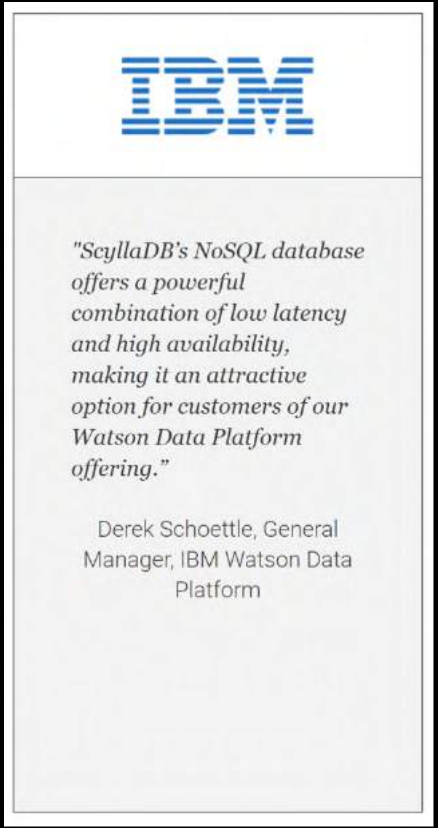
outbrain

TalkingData

What does it mean for a non Cassandra user?

- Throughput, latency and scale benefits
- Wide range of big data integration: {Kariosdb, Spark, JanusGraph, Presto, Kafka, Elastic}
- Best HA/DR in the industry.
- Stop using caches in front of the database
- Consolidate HBase, Redis, MySQL, Mongo and others

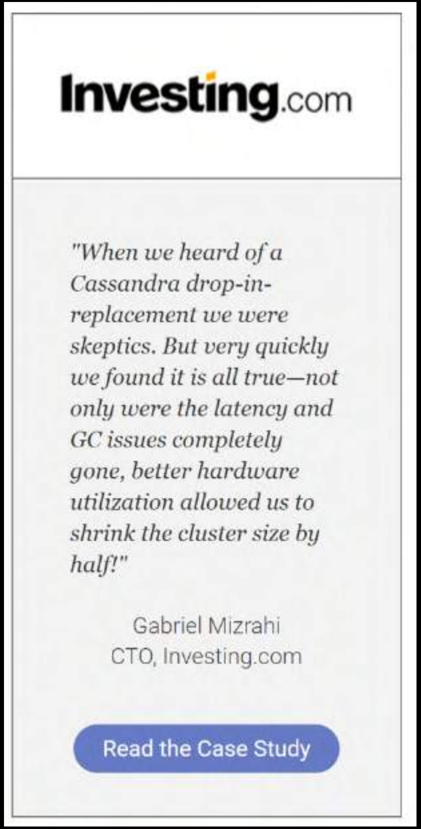
Assorted Quotes



IBM

"ScyllaDB's NoSQL database offers a powerful combination of low latency and high availability, making it an attractive option for customers of our Watson Data Platform offering."

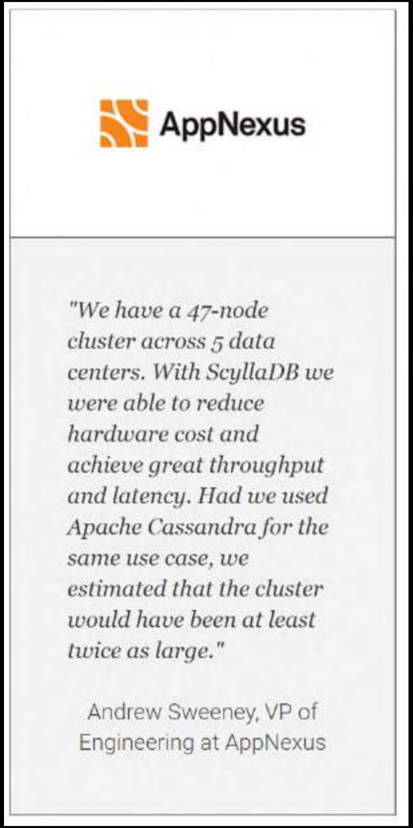
Derek Schoettle, General Manager, IBM Watson Data Platform



Investing.com

"When we heard of a Cassandra drop-in-replacement we were skeptics. But very quickly we found it is all true—not only were the latency and GC issues completely gone, better hardware utilization allowed us to shrink the cluster size by half!"

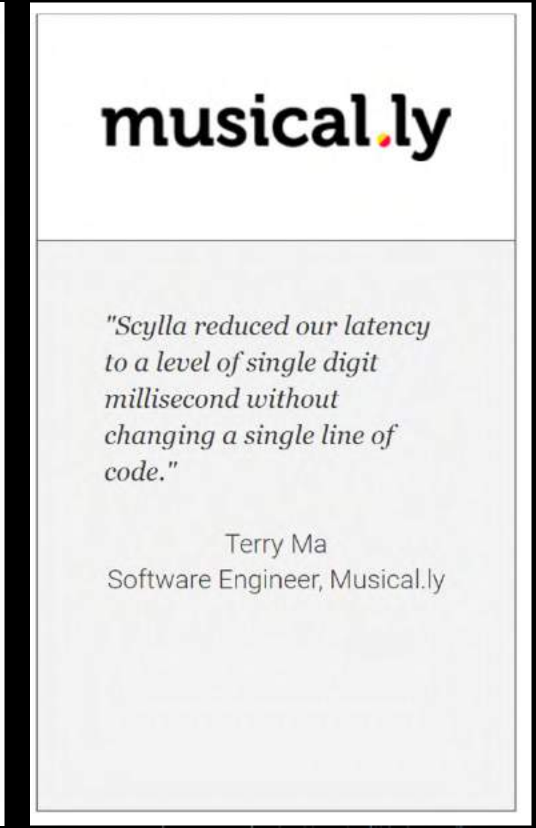
Gabriel Mizrahi
CTO, Investing.com



AppNexus

"We have a 47-node cluster across 5 data centers. With ScyllaDB we were able to reduce hardware cost and achieve great throughput and latency. Had we used Apache Cassandra for the same use case, we estimated that the cluster would have been at least twice as large."

Andrew Sweeney, VP of Engineering at AppNexus



musical.ly

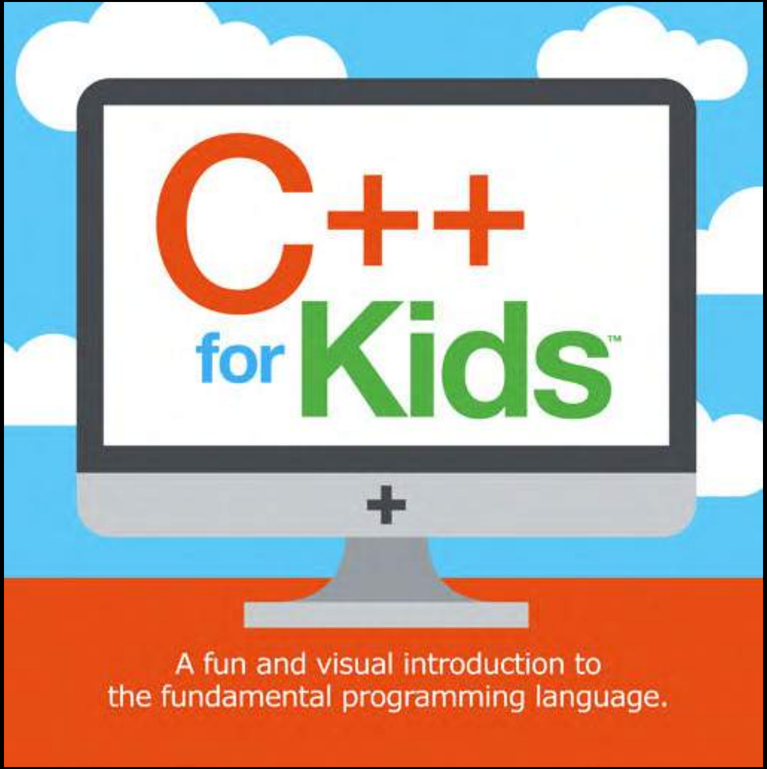
"Scylla reduced our latency to a level of single digit millisecond without changing a single line of code."

Terry Ma
Software Engineer, Musical.ly

Today we will cover:

- **Intro: Who we are, what we do, who uses it**
- Why we started ScyllaDB
- Why should you care
- How we made design decisions to achieve no-compromise performance and availability

Design decisions: #1 The trivials

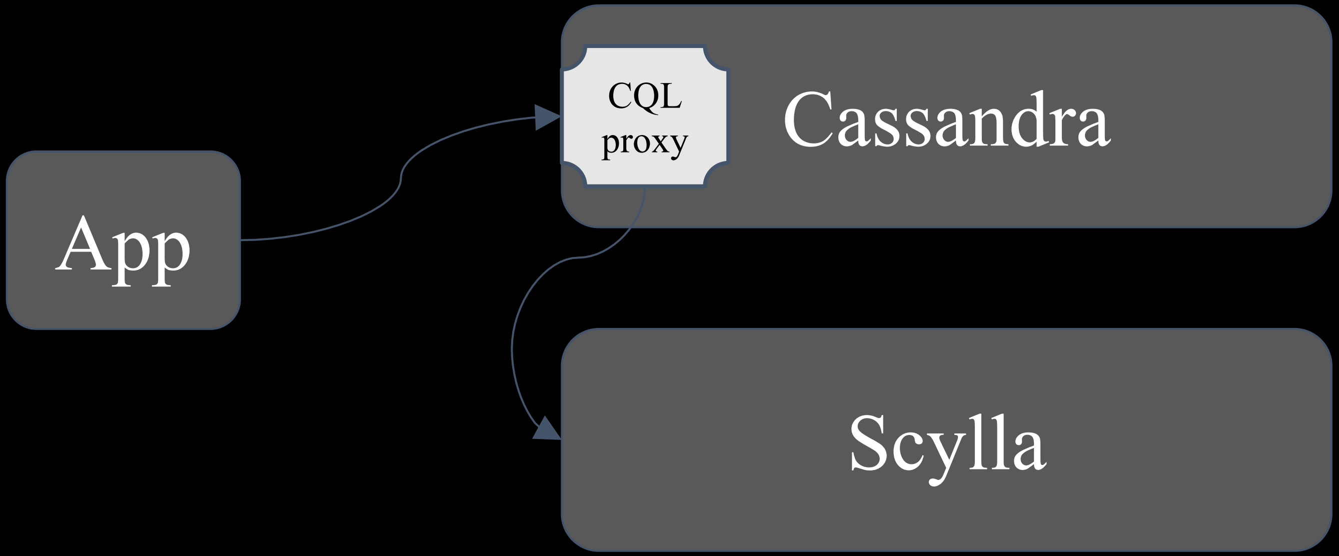

C++
for Kids™
A fun and visual introduction to
the fundamental programming language.

Design decisions: #2 Compatibility

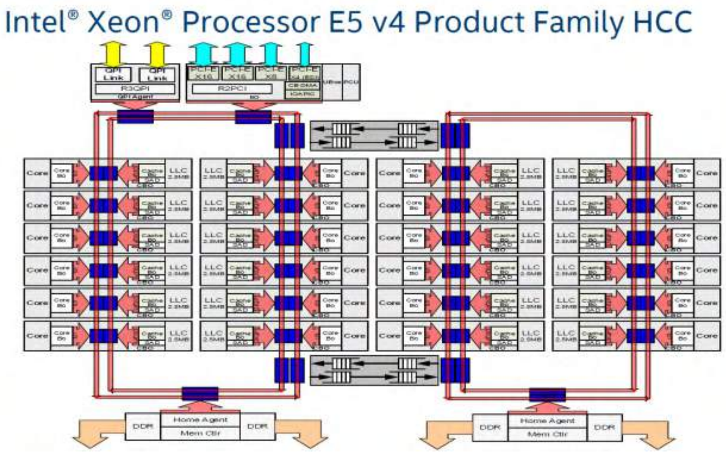
- SSTable file format
- Configuration file format
- CQL language
- CQL native protocol
- JMX management protocol
- Management

presto
Spark
KairosDB
JanusGraph

Double cluster - Migration w/o downtime



Design decisions: #3 All things async



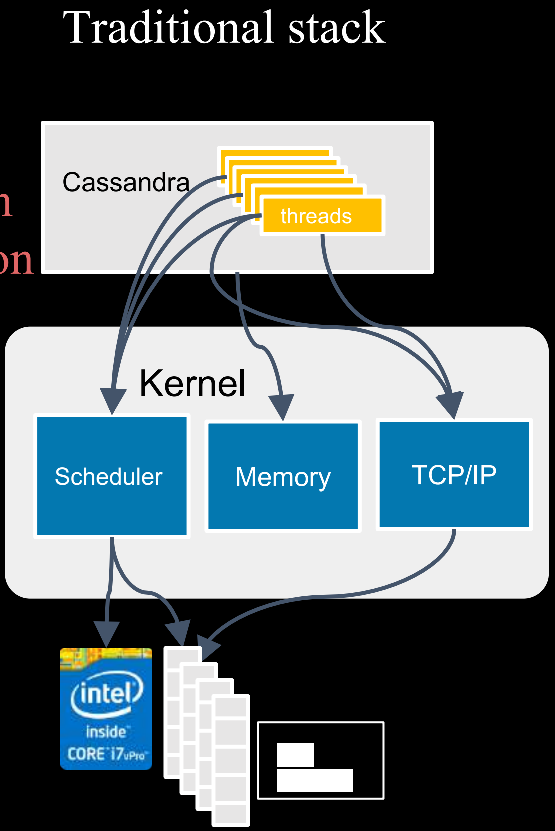
Design decisions: #4 Shard per core

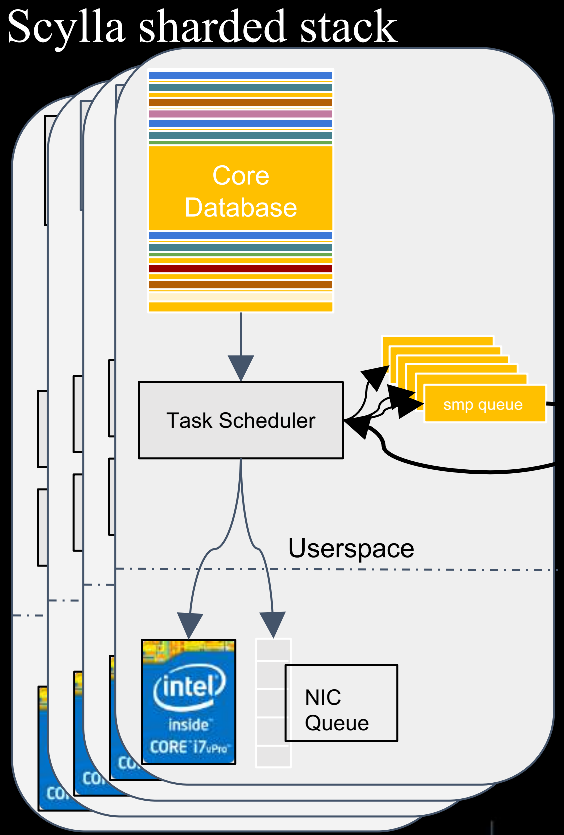
Threads

Shards

SCYLLA DB: Network Comparison

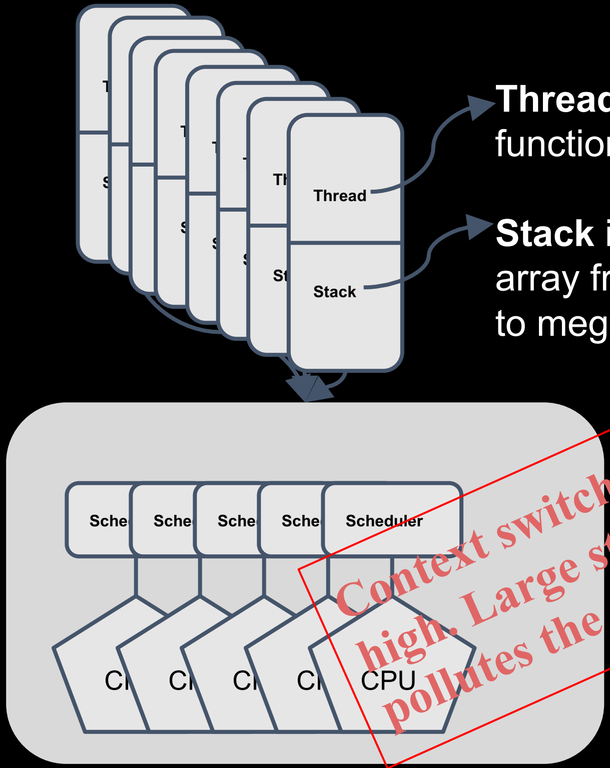
Lock contention
Cache contention
NUMA
unfriendly





No contention
Linear scaling
NUMA friendly

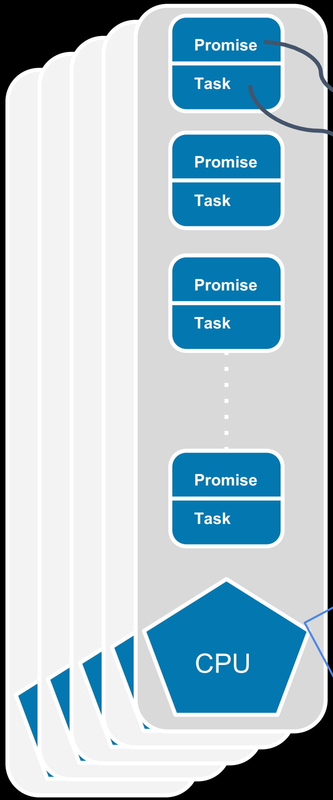
Scylla has its own task scheduler



Thread is a function pointer

Stack is a byte array from 64k to megabytes

Context switch cost is high. Large stacks pollutes the caches



Promise is a pointer to eventually computed value

Task is a pointer to a lambda function

No sharing, millions of parallel events

SCYLLA IS DIFFERENT

- □ Thread per core
- □ Lock-free
- □ Task scheduler
- □ Reactor programing
- □ C++14

- □ Multi queue
- □ Poll mode
- □ Userspace TCP/IP

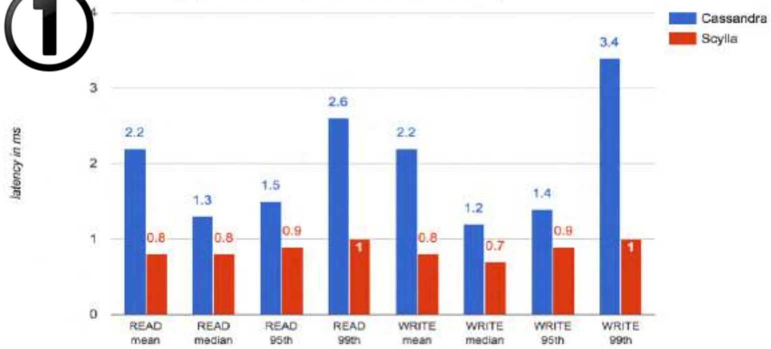
- □ NUMA friendly
- □ Log structured allocator
- □ Zero copy

- □ DMA
- □ Log structured merge tree
- □ DBaware cache
- □ Userspace I/O scheduler

Scylla vs C* latency by Kenshoo

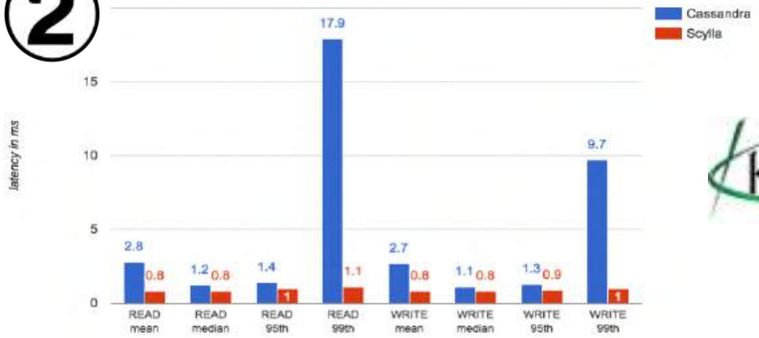
①

2K/sec ops (1.2K/sec write,0.8K/sec read, 1.5k value)



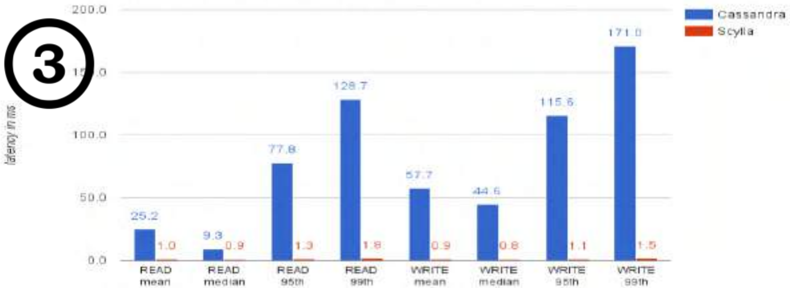
②

4K/sec ops (2.5K/sec write,1.5K/sec read, 1.5k value)



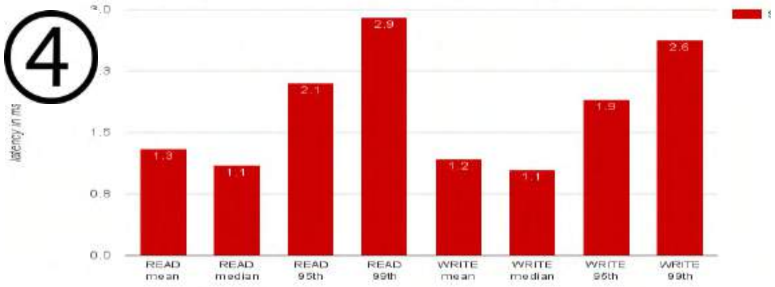
③

16K/sec ops (10K/sec write, 6K/sec read, 1.5k value)



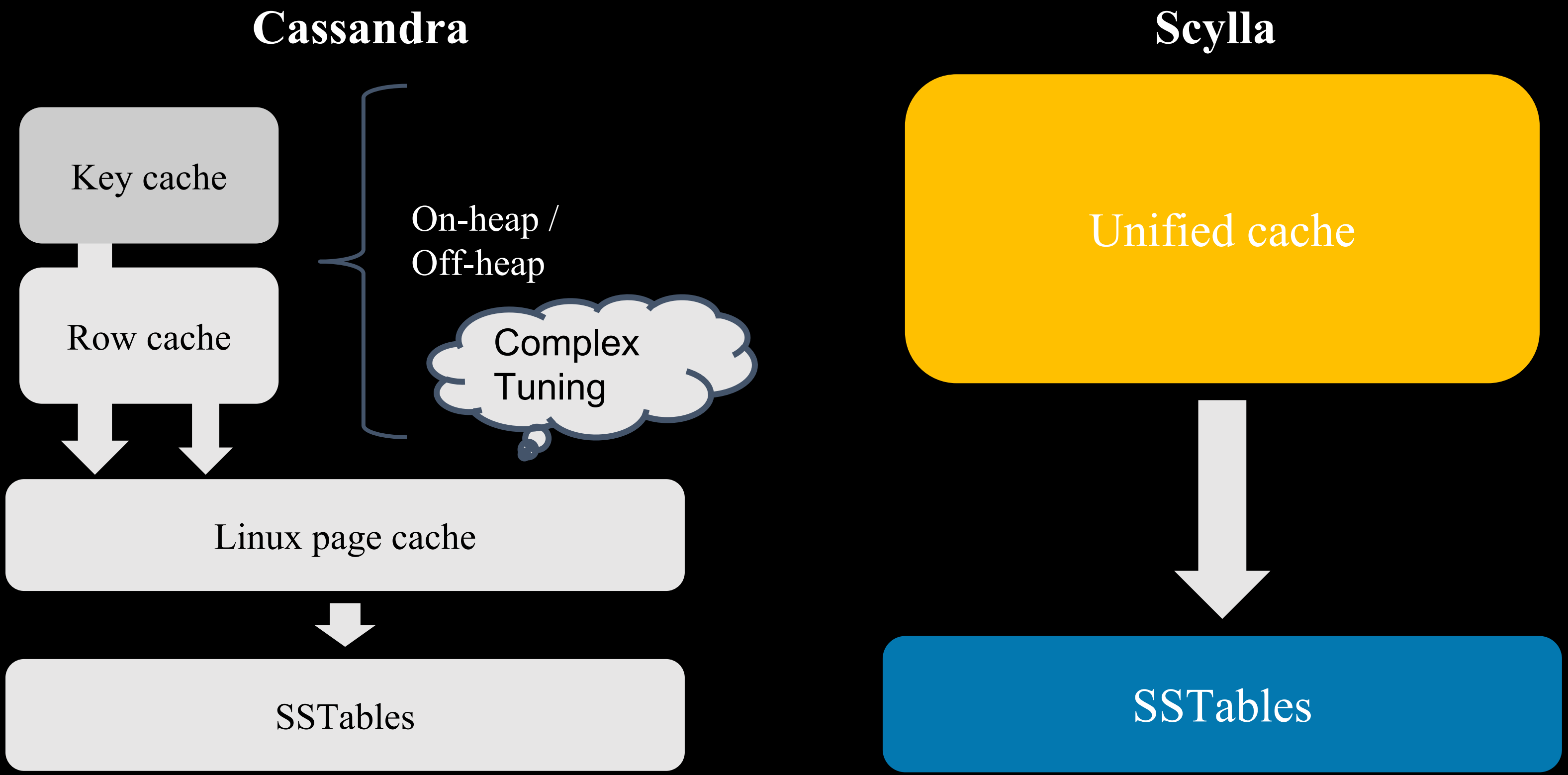
④

40K/sec ops (25K/sec write, 15K/sec read, 1.5k value)



Kenshoo

Design Decision: #5 Unified cache



Cassandra

Key cache

Row cache

On-heap /
Off-heap

Complex
Tuning

Linux page cache

SSTables

Scylla

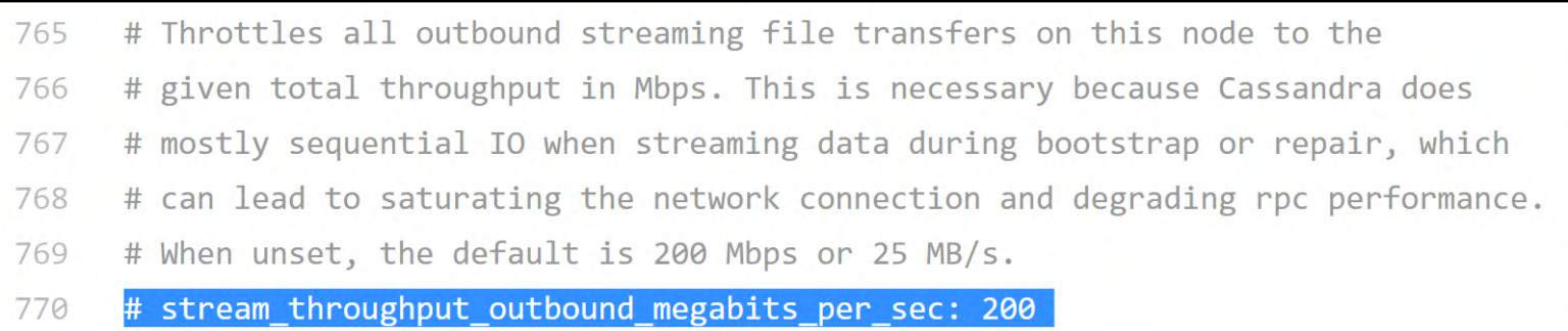
Unified cache

SSTables

TalkingData

Design decisions: #6 I/O scheduler

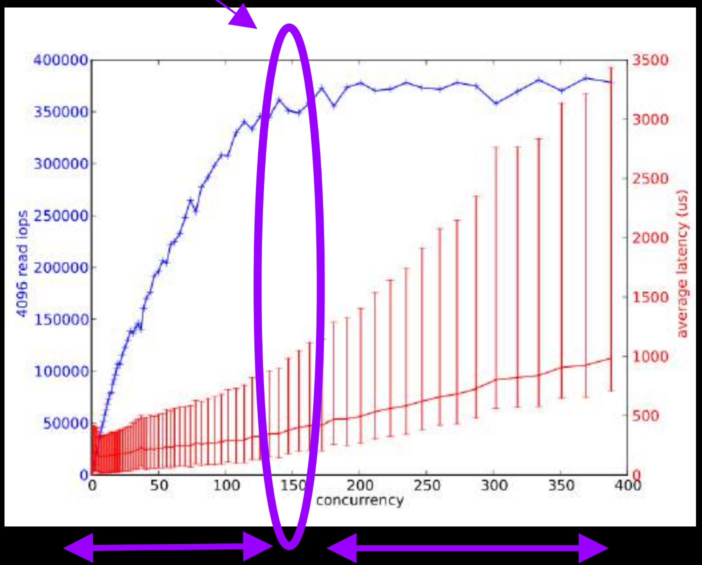
Cassandra Streaming configuration



```
# Throttles all outbound streaming file transfers on this node to the
# given total throughput in Mbps. This is necessary because Cassandra does
# mostly sequential IO when streaming data during bootstrap or repair, which
# can lead to saturating the network connection and degrading rpc performance.
# When unset, the default is 200 Mbps or 25 MB/s.
# stream_throughput_outbound_megabits_per_sec: 200
```

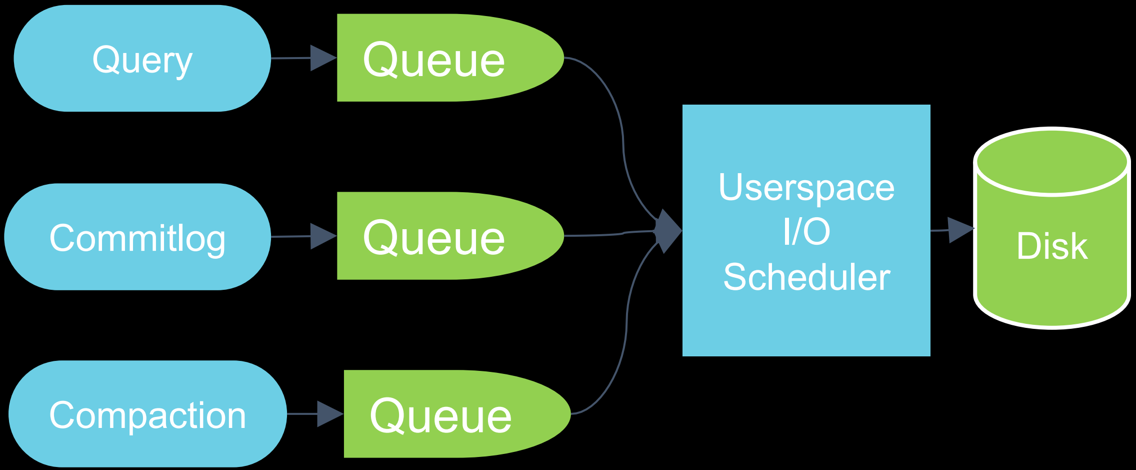
Scylla I/O Scheduling

Max useful disk concurrency

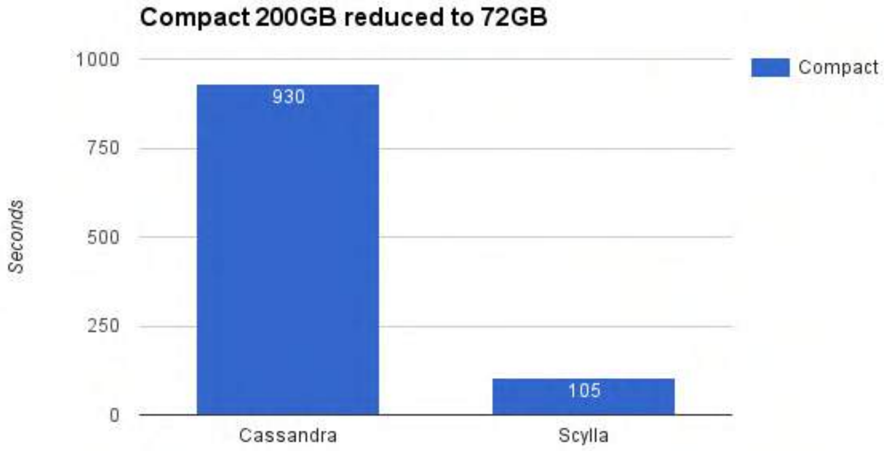


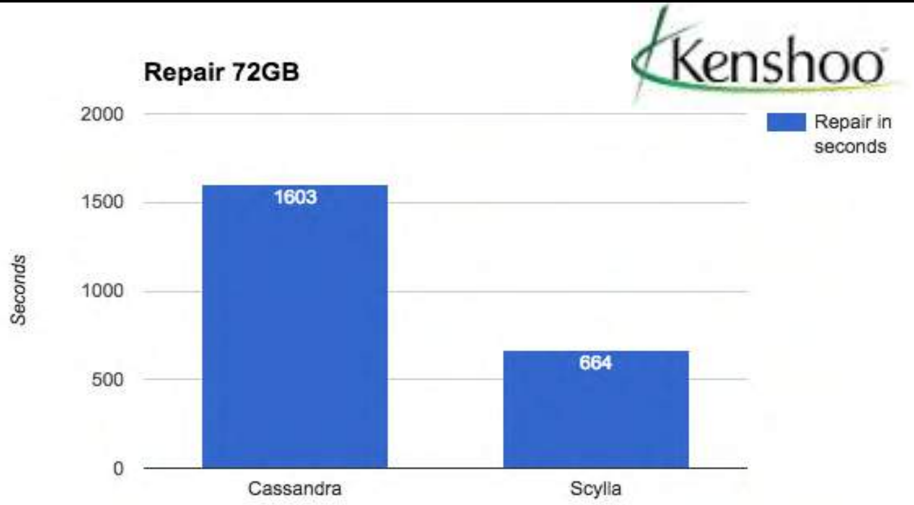
No queues

I/O queued in FS/device

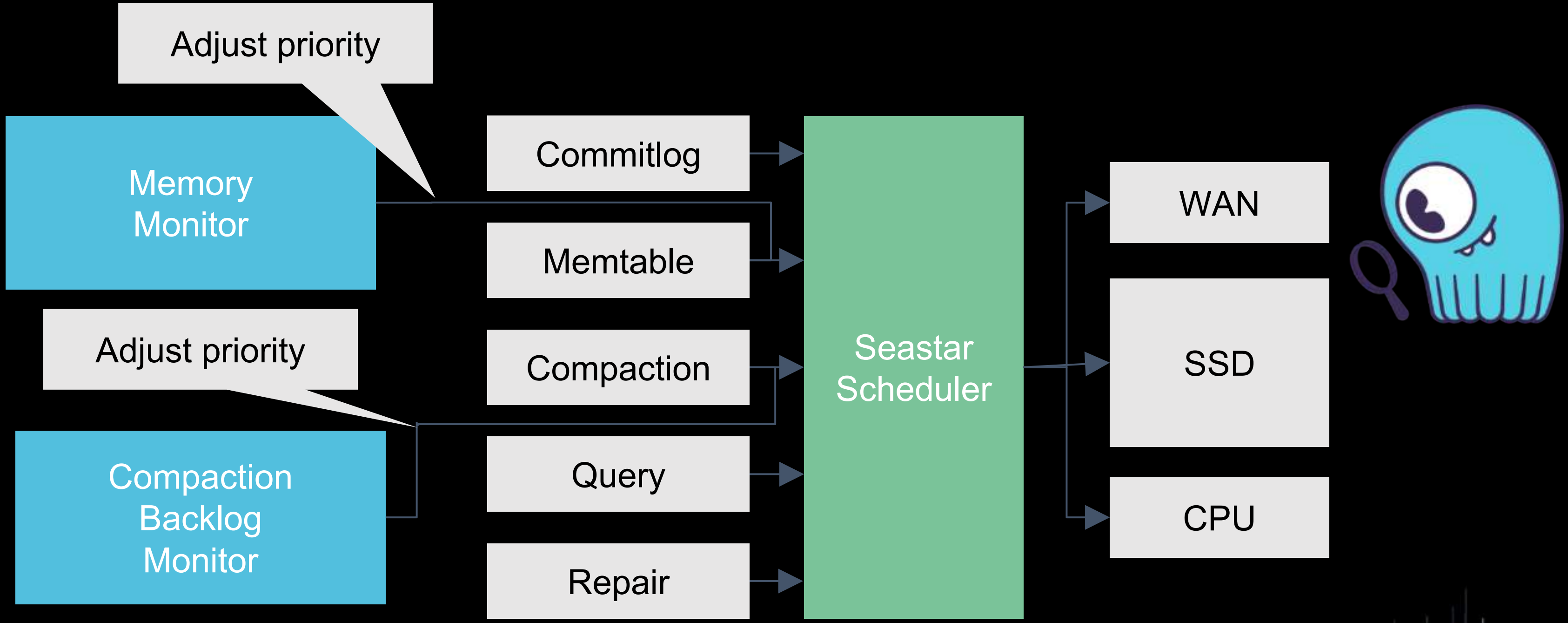


I/O scheduler result by Kenshoo

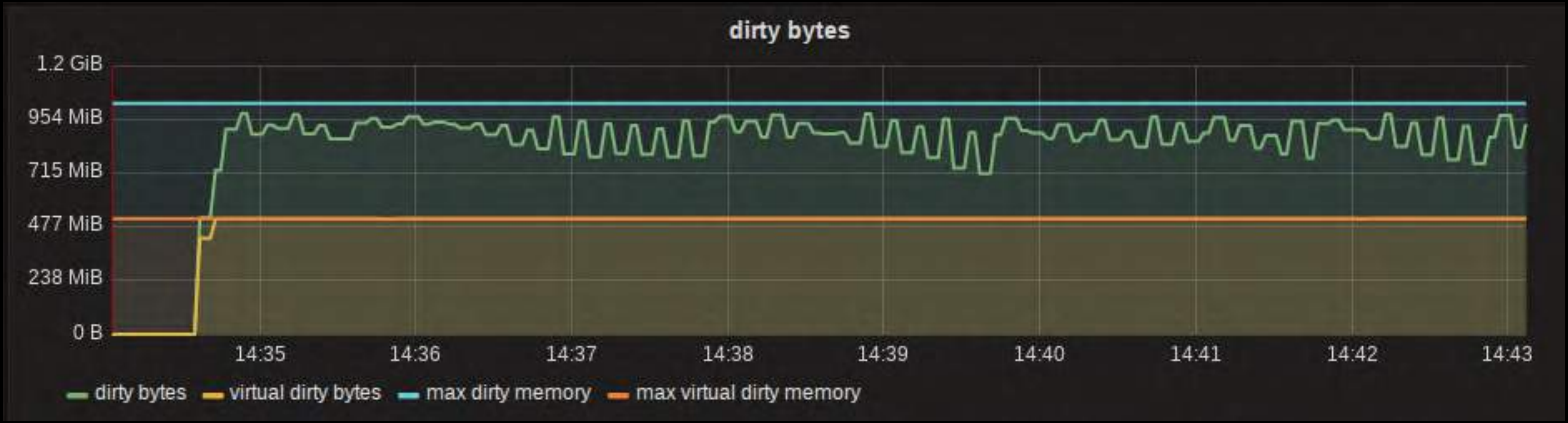

Compact 200GB reduced to 72GB

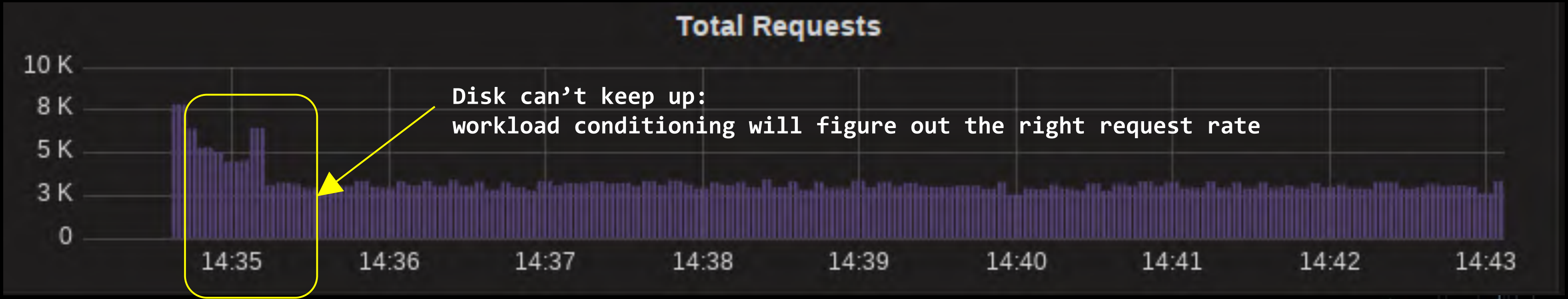

Repair 72GB

Design Decision: #7 Workload conditioning



Workload Conditioning in practice





Disk can’t keep up:
workload conditioning will figure out the right request rate

Upcoming releases

- Enterprise release, based on 1.6
- 1.7 - May 2017
  - Counters
  - New intra-node sharding algorithm
  - SStableloader from 2.2/3.x
  - Debian
- 2.0 – Sep 2017
  - Materialized views
  - Execution blocks (cpu cache optimization which boost performance)
  - Partial row cache (for wide row streaming)
  - Heat Weighted Load Balancing

Scylla Beyond Cassandra

Core database

Vertical

Horizontal

Q&A

Resources

dor@scylladb.com (@DorLaor)

avi@scylladb.com (@AviKivity)

github.com/scylladb/scylla

scylladb.com/blog

@scylladb

http://bit.ly/2oHAfok

youtube.com/c/scylladb

slideshare.net/ScyllaDB

T11 2017

THANKS

SCYLLA: NoSQL at Ludicrous Speed