AEROSPIKE

Flash-Optimized, High-Performance NoSQL Database for All

Khosrow Afroozeh
Engineer
Aerospike

Aerospike – Built for the Age of Scale

- The Gold Standard
  6 of top 16
  powered by Aerospike

  (after Google, FB, from BuiltWith.com)

Top in Advertising · Week beginning Feb 10th 2014

Name	10k	100k	Million
Google Adsense	⬆3,191	⬆22,305	⬆172,178
DoubleClick.Net	⬆4,407	⬆28,215	⬆168,902
AppNexus	⬆1,595	⬆7,987	⬆30,936
Google Adsense for Search	⬇452	⬇4,363	⬇26,126
Google Remarketing	⬆758	⬆5,106	⬆25,302
Google Publisher Tag	⬆1,404	⬆5,724	⬆18,071
Openads/OpenX	⬆847	⬆4,261	⬆16,030
Turn	⬆887	⬆4,091	⬆12,725
Facebook Exchange FBX	⬆741	⬆3,627	⬆11,878
Rubicon Project	⬆809	⬆3,610	⬆11,404
eXelate	⬆558	⬆2,662	⬆11,390
AdRoll	⬆478	⬆3,078	⬆10,695
BlueKai	⬆814	⬆3,503	⬆9,783
Pubmatic	⬆720	⬆3,144	⬆9,490
Casale Media	⬆574	⬆2,832	⬆9,307
Rocket Fuel	⬆714	⬆3,267	⬆9,006
The Trade Desk	⬇521	⬇2,599	⬇8,538
Yield Manager	⬇421	⬇1,784	⬇8,253
Simpli.fi	⬆304	⬆1,778	⬆8,079
X Plus One	⬆454	⬆2,157	⬇7,924
Dstillery	⬆397	⬆1,962	⬇7,796
Chango	⬆454	⬆2,194	⬆6,845

appnexus

Extreme Speed

…we process many terabytes of data daily across our global data centers at rates in excess of one million requests per second.

Mike Nolet – CTO

bluekai

Internet Scale

We are now the largest online data exchange and respond to requests 2 trillion times a month using Aerospike as our foundation.

Alex Hooshmand, co- founder & Chief Strategy Officer & SVP Operations

adMarketplace.COM
Performance-Driven Pay Per Click Advertising

100% Uptime

For us, this is the top metric of success, and that's what we've achieved with the Aerospike real-time database.

Mike Yudin – co-founder & CTO

TINKOFF DIGITAL

Cost Effective

Aerospike's performance with the ability to reduce maintenance, support and hardware costs make it a truly attractive data management solution.

K. Kruglov - CTO

tapad

Simple Operations

Aerospike makes upgrading simple. There's no planning required. You can take servers down and still have the system running.

Dag Liodden – co-founder & CTO

[x+1]
Make Every Interaction Count

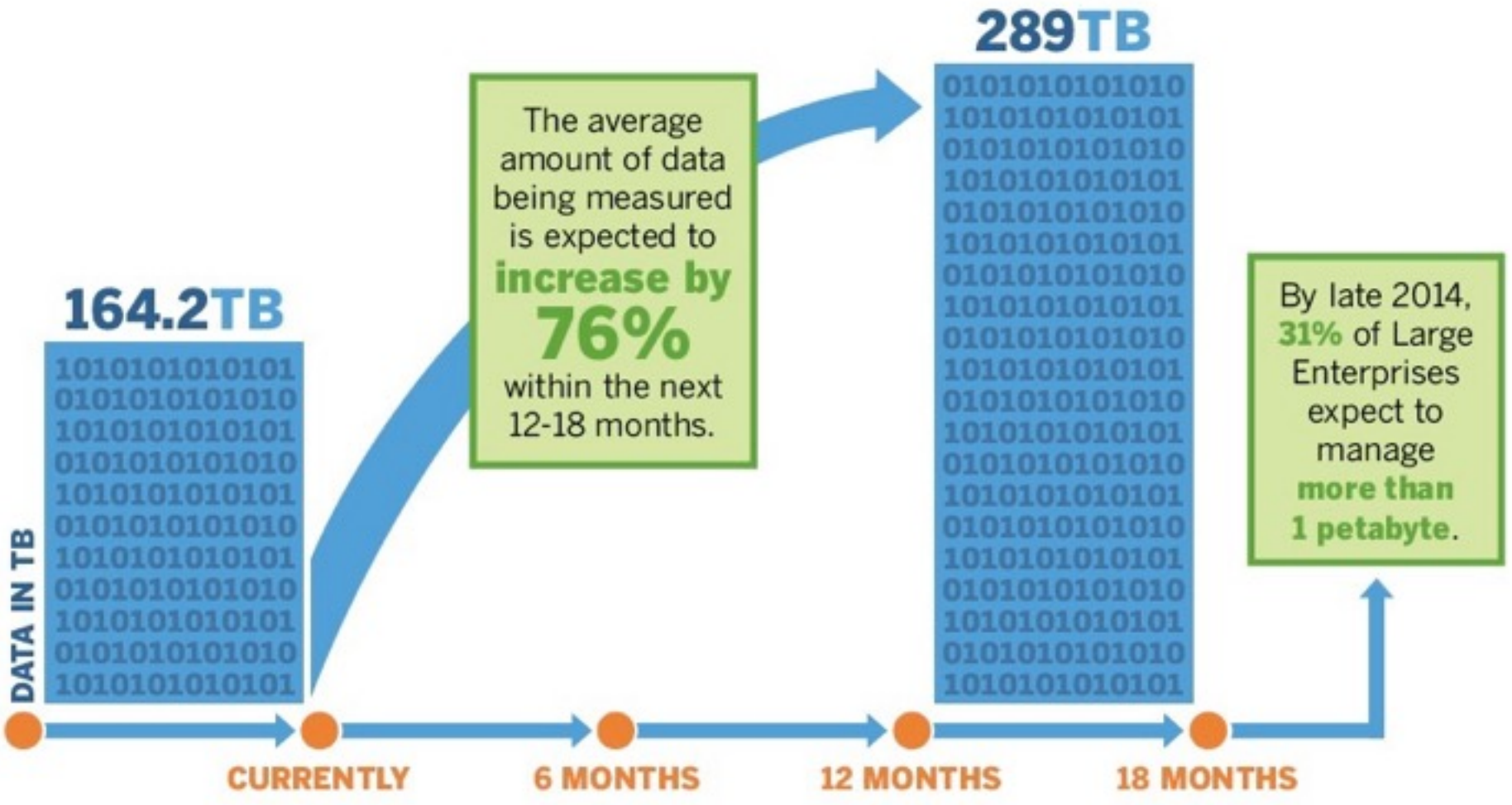
The Right Choice

Providing fast reliable access to data in real-time is not easy to do. Aerospike has proven that our choice to buy, not build, was the right decision.

Pat DeAngelis – CTO

BIG DATA: IS IT GOING TO HAPPEN TO ME?

Every Business is Demanding "Internet Scale"



- Image: "Great Migrations in IT – Cloud, Big Data and the Race for Web-Scale IT. It's All About Business Agility." blog post by Mike Jochimsen at Emulex Labs 1.22.14.

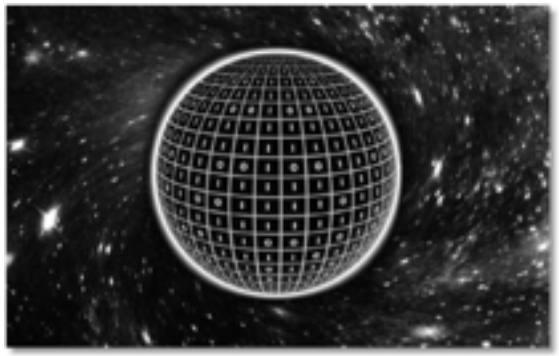
BUT…

… I don't have that much data!

- Acquire it! It's not like the technology to manage it doesn't exist.

AEROSPIKE mongoDB cassandra

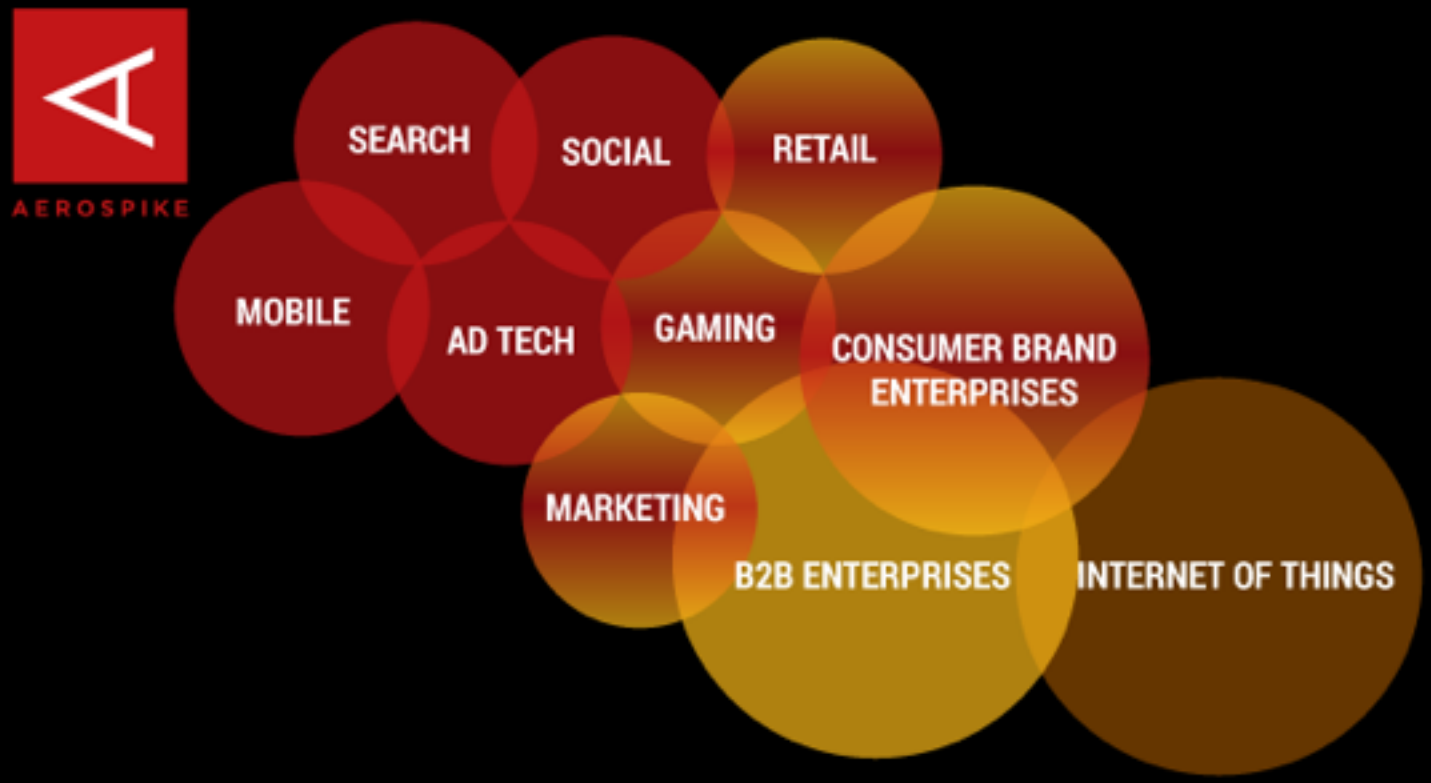
- Data provides more insight into trends, if not behavior.
- Information behaves like mass: It attracts more information!



Typical Deployment

- Last Year
  - 8 core Xeon
  - 24G RAM
  - 400G SSD (SATA)
  - 30,000 read TPS, 20,000 write TPS
  - 1.5K object size / 200M objects
  - 4 to 40 node clusters
- This Year
  - 16 core Xeon
  - 128G RAM
  - 2T~4T SATA / PCIe (12 s3700 / 4 P320h)
  - 100,000 read TPS, 50,000 write TPS
  - 3K object size / 1B objects
  - 4 to 20 node cluster

REAL-TIME INTERACTIONS ARE EVERYWHERE



RT-Advertising
- Targeted Ads
- Search Retargeting
- Offer Performance Management

RT-Marketing
- Omni-channel Marketing
- Real-time Pricing
- In-store Inventory Optimization

RT-Interactions
- Infinite Scroll Recommendations
- Location-based Services
- Mass-customized Digital Properties

RT-Intel & Control
- Inter-enterprise Customer Service
- Sensor Monitoring & Response
- Real-time Control Fabrics

Internet Of Things

North American RTB speeds & feeds

- 100 millisecond or 150 millisecond ad delivery
  - De-facto standard set in 2004 by Washington Post and others
- North America is 70 to 90 milliseconds wide
  - Two or three data centers
- Auction is limited to 30 milliseconds
  - Typically closes in 5 milliseconds
- Winners have more data, better models – in 5 milliseconds

North American RTB speeds & feeds

- 1 to 6 billion cookies tracked
  - Some companies track 200M, some track 20B
- Each bidder has their own data pool
  - Data is your weapon
  - Recent searches, behavior, IP addresses
  - Audience clusters (K-cluster, K-means) from offline Hadoop
- “Remnant” from Google, Yahoo is about 0.6 million / sec
- Facebook exchange: about 0.6 million / sec
- “other” is 0.5 million / sec

Currently more than 2.0M / sec in North American

PERFORMANCE → PERSONALIZATION → PROFITS

AGE OF CUSTOMER = READ/WRITE PATTERN

CONTEXT

- **IDENTITY**
  - SessionIDs, Cookies, DeviceIDs, ip-Addr
- **ATTRIBUTES**
  - Demographic, geographic
- **BEHAVIOR**
  - Presence, swipe, search, share..
  - Channels – web, phone, in-store..
  - Services – frequency, sophistication
- **SEGMENTS**
  - Attitudes, values, lifestyle, history..
- **TRANSACTIONS**
  - Payments, campaigns

BIG DATA: EMERGING ARCHITECTURE

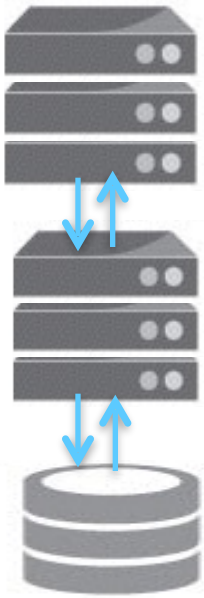
Add-a-Layer Architecture

CONTENT
DELIVERY NETWORK

LOAD BALANCER

Request routing and sharding

APP SERVERS

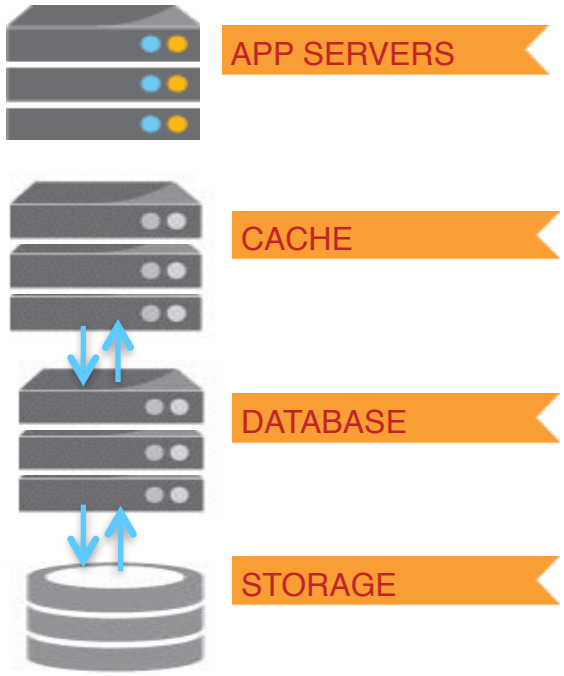


CACHE

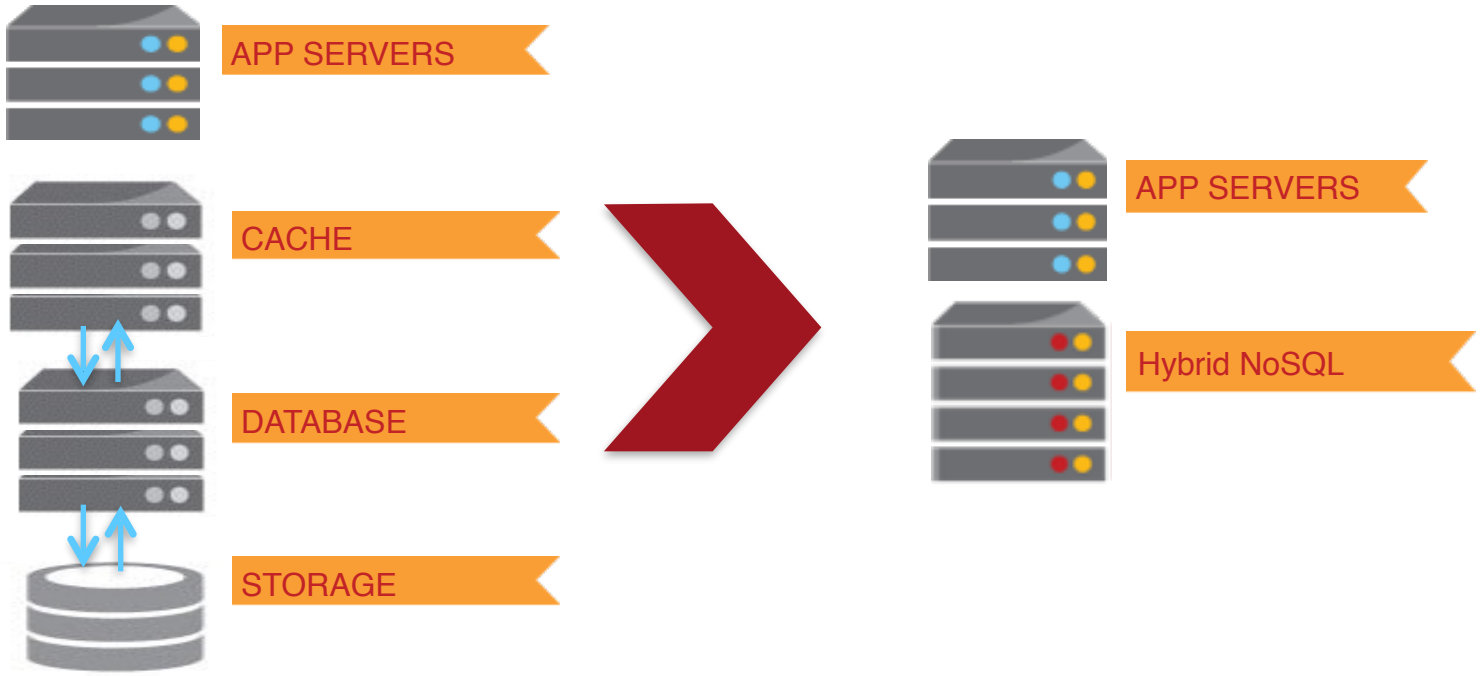
DATABASE

STORAGE

Minimalism Makes a Comeback

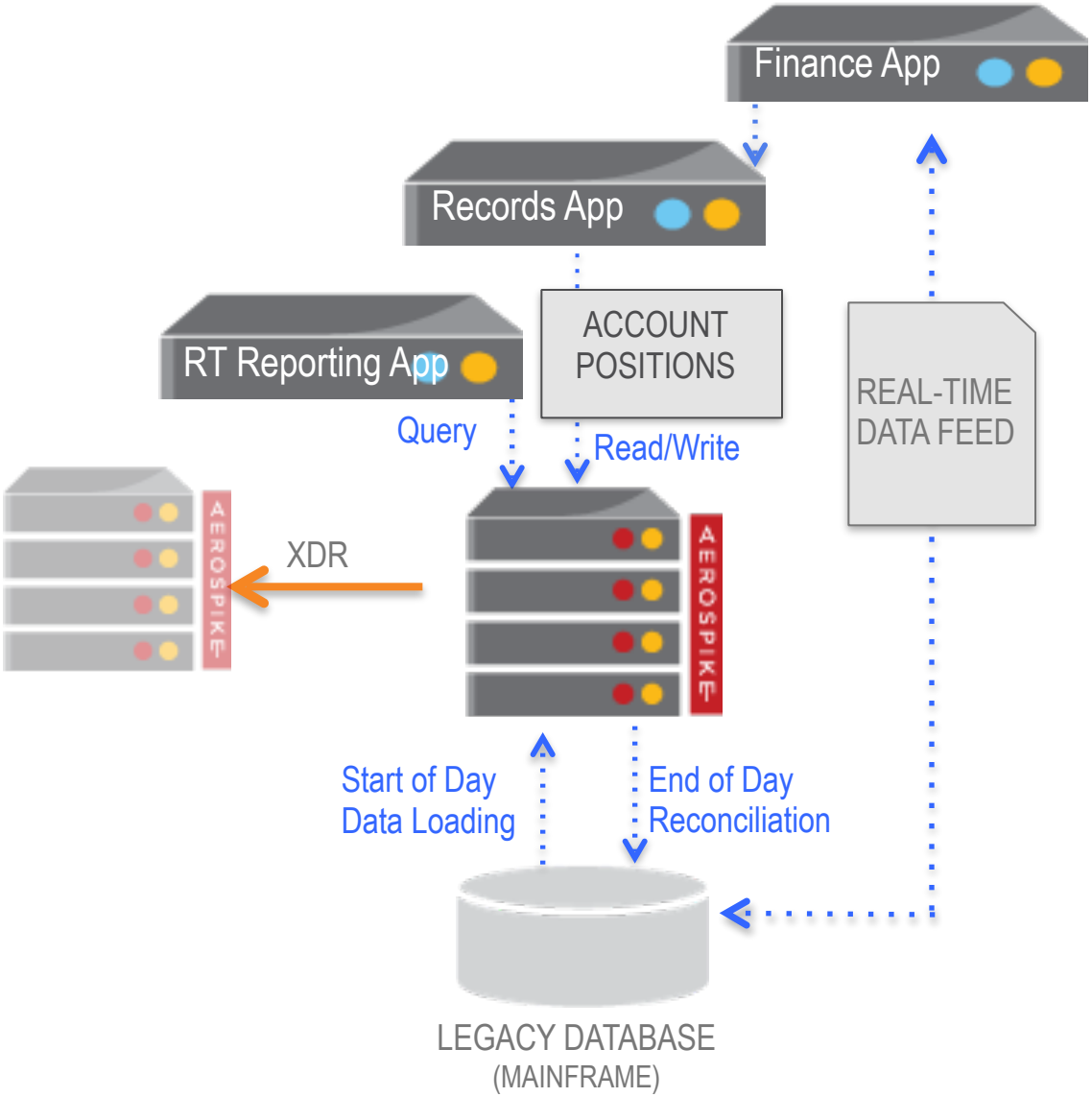


Minimalism Makes a Comeback



BIG DATA: EDGE DATABASES

Financial Services – Intraday Positions

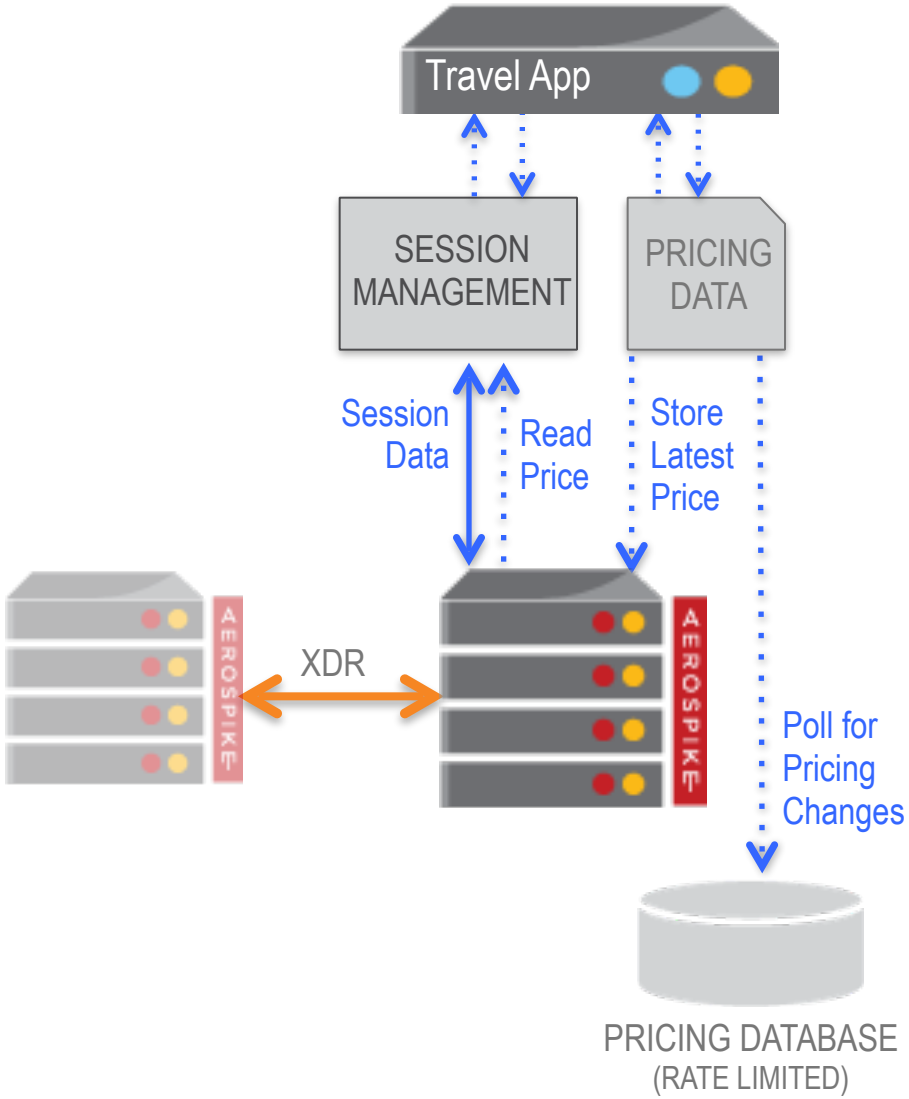


10M+ user records

Primary key access

1M+ TPS planned

Travel Portals



Airlines forced interstate banking

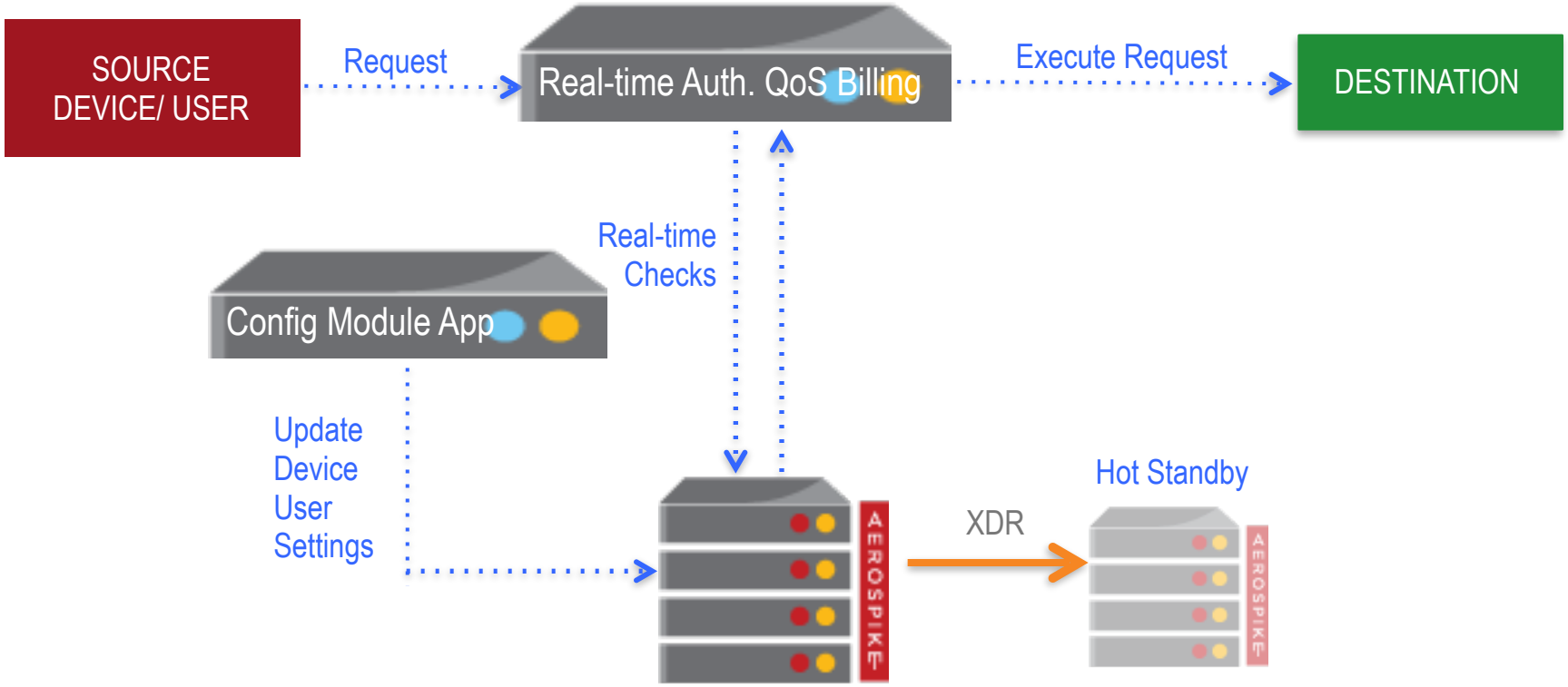
Legacy mainframe technology

Multi-company reservation and pricing

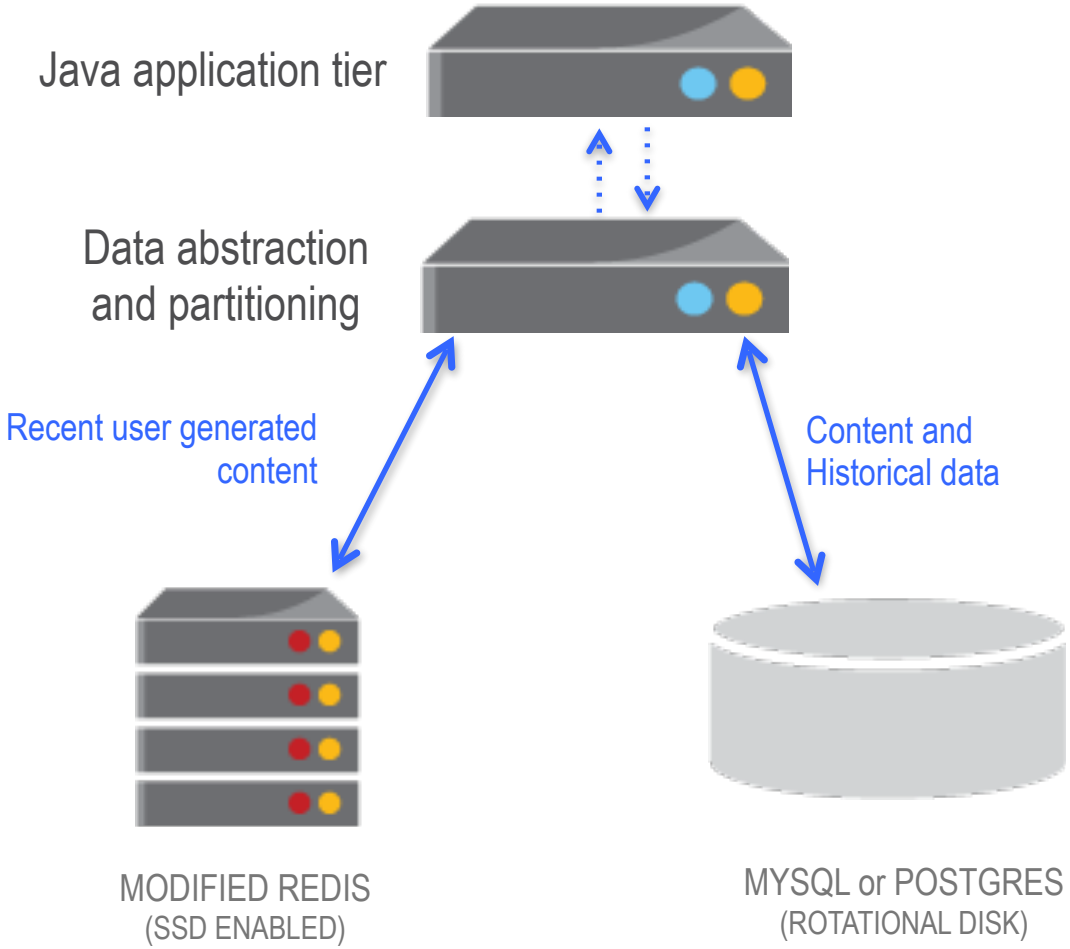
Requirement: 1M TPS allowing overhead

QOS & Real-Time Billing for Telcos

- In-switch Per HTTP request Billing
  - US Telcos: 200M subscribers, 50 metros
- In-memory use case



Social Media

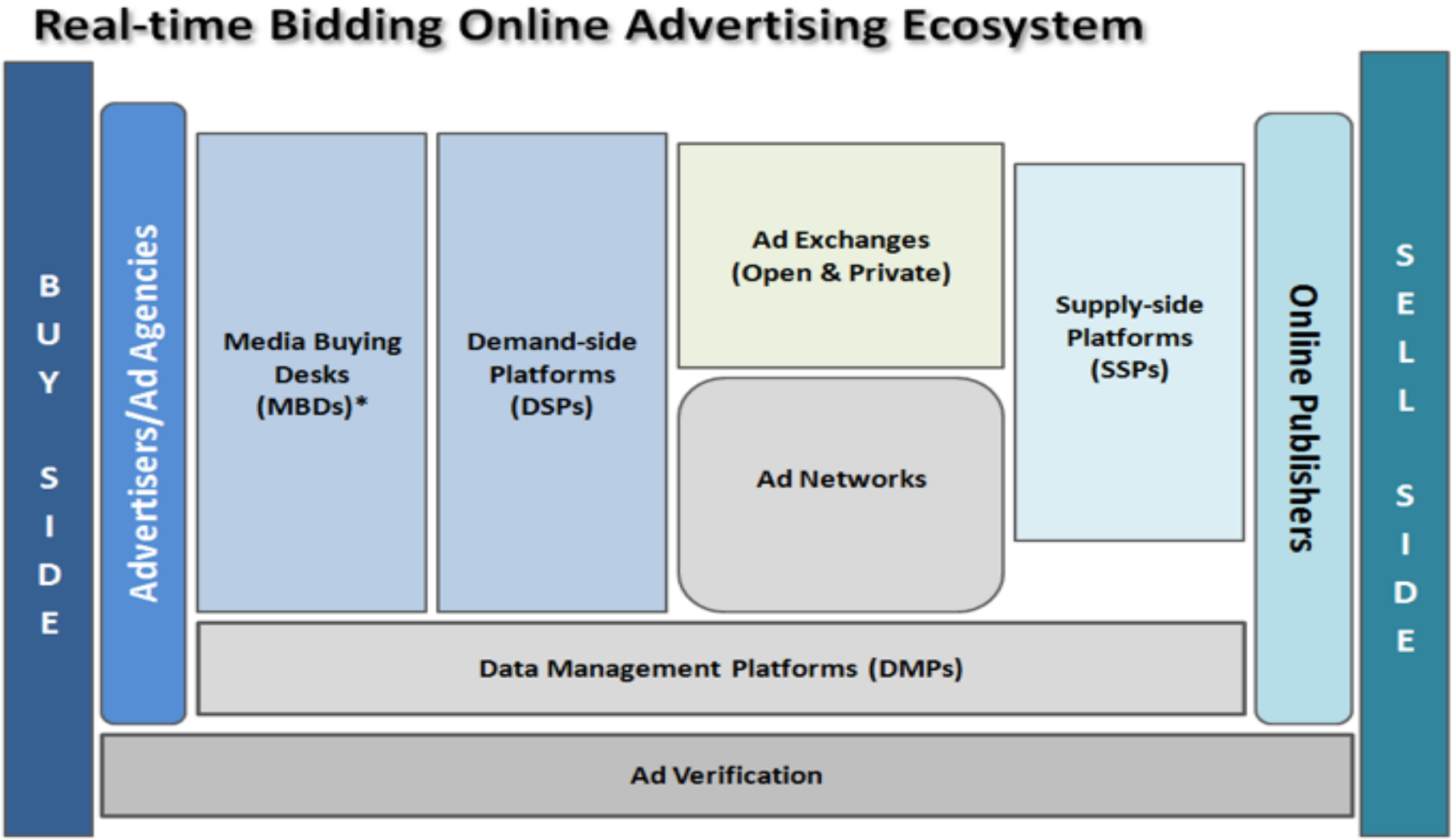


新浪微博 weibo.com

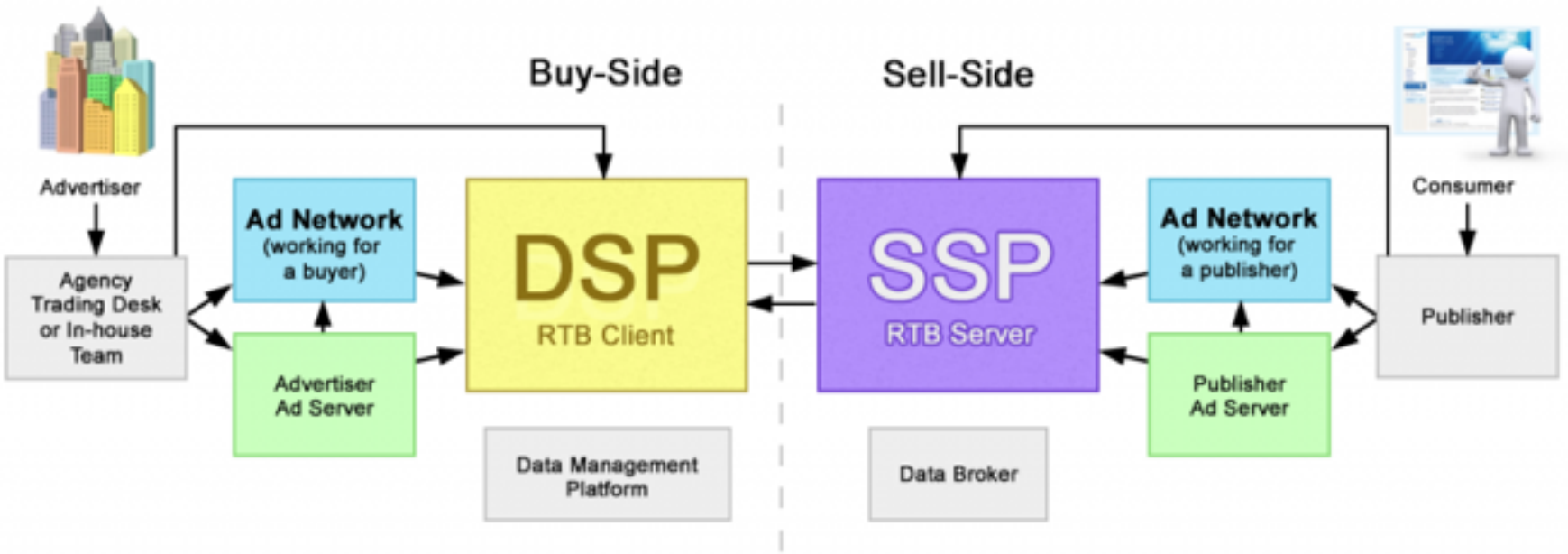
Alibaba.com

Tencent 腾讯

Real-time bidding



Real-time bidding



Modern Scale-Out Architecture

CONTENT
DELIVERY NETWORK

LOAD BALANCER

Simple stateless

APP SERVERS

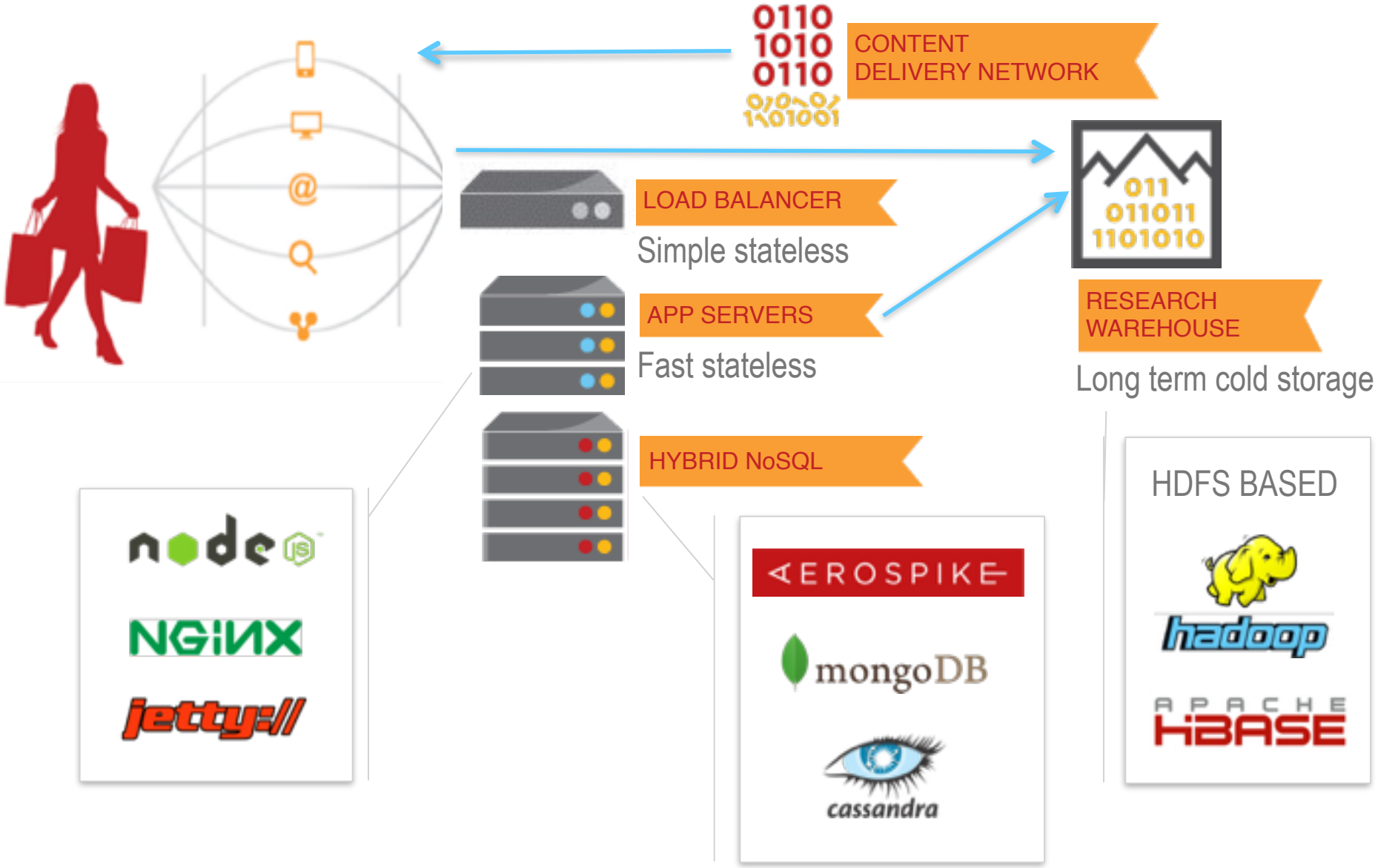
Fast stateless

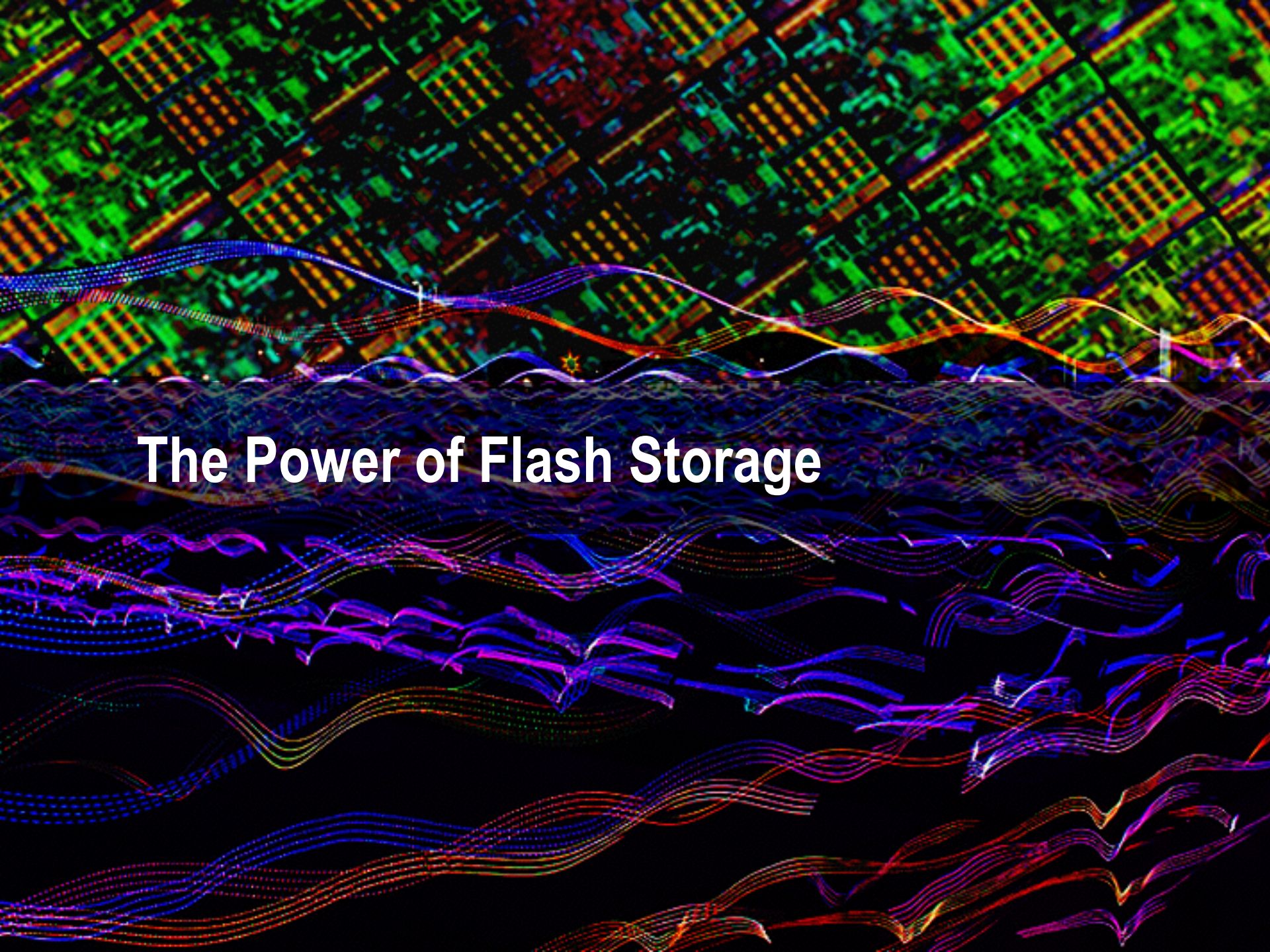
HYBRID NoSQL

RESEARCH
WAREHOUSE

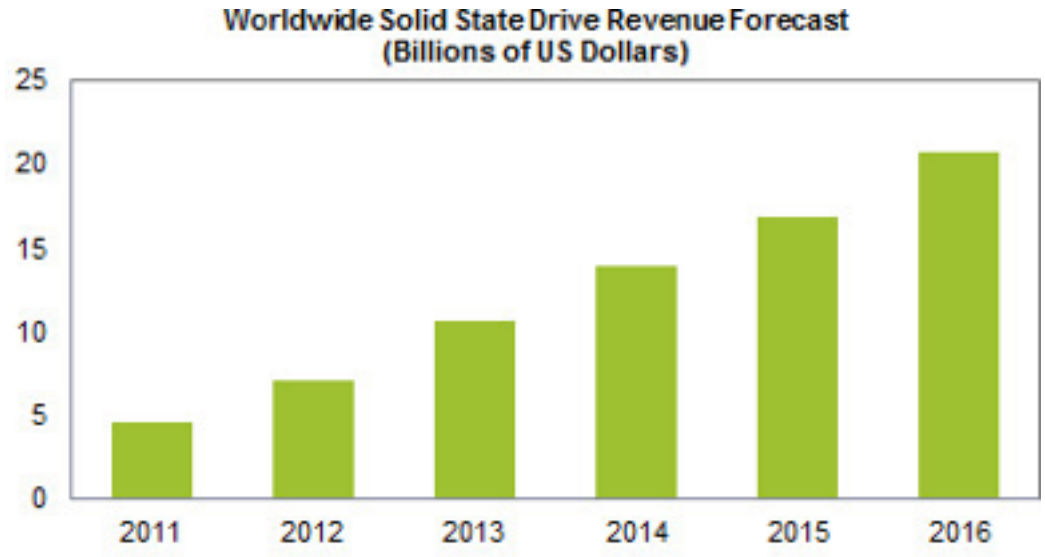
Long term cold storage

Modern Scale-Out Architecture




The Power of Flash Storage

Flash Storage Proven and Growing

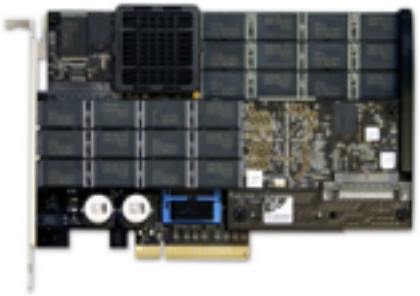


Source: IHS iSuppli Research, January 2013

+

=

$200+M



Facebook and Apple bought *at least* $200+M in FusionIO cards in 2012

(55% of $440M revenue estimate, reported in quarterly FusionIO earnings)

Everyone wants that “facebook architecture”

Aerospike's Flash Experience

- Flash Knowledge
  - ACT benchmark http://github.com/aerospike/act
  - Read-write benchmark results back to 2011

- All clouds support flash now
  - New EC2 instances
  - Google Compute
  - Internap, Softlayer, GoGrid…

- Write durability usually not a problem with modern flash
  - Durability is high (5 "drive writes per day" for 5 years, etc)

Aerospike's Flash Experience

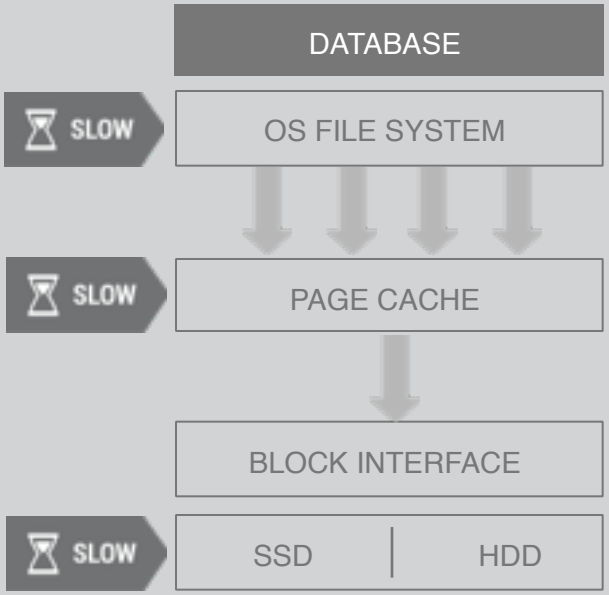
- Densities increasing
  - 100GB 2 years ago → 800GB today
  - SATA vs PCI-E
  - Appliances: 50T per 1U this year
- Prices still dropping: perhaps $1/GB next year
- Intel P3700 results
  - 250K per device @ $2.5 / GB
  - Old standard: Micron P320h 500K @ $8 / GB
- "Wide SATA"
  - 20 SATA drives
  - LSI "pass through mode"
  - 250K+ per server

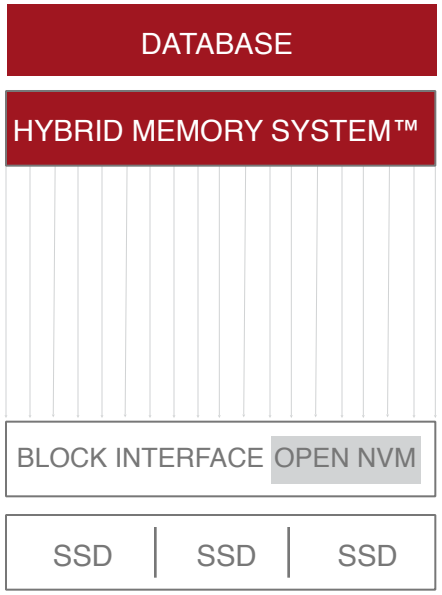
FLASH OPTIMIZED HIGH PERFORMANCE

- Direct device access
- Large Block Writes
- Indexes in DRAM
- Highly Parallelized
- Log-structured FS “copy-on-write”
- Fast restart with shared memory

Ask me. I’ll look up the answer and then tell it to you.

Ask me and I’ll tell you the answer.





Flash Big Data Economics

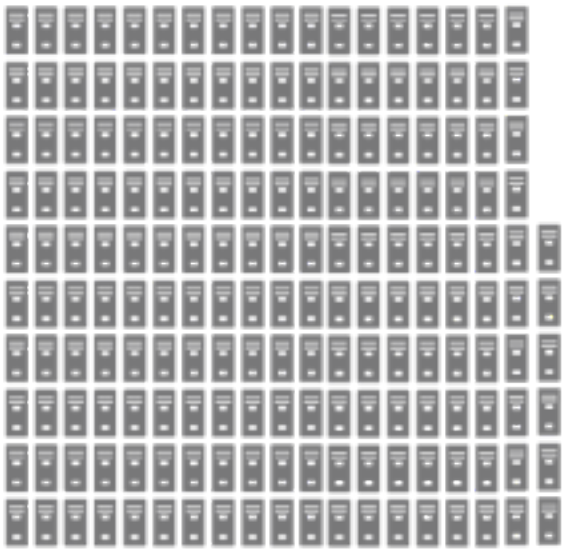
10x FASTER

&

10x FEWER

SERVERS REQUIRED

Actual Customer Analysis
99% < 1ms
500K TPS
10TB Storage
2x Replication

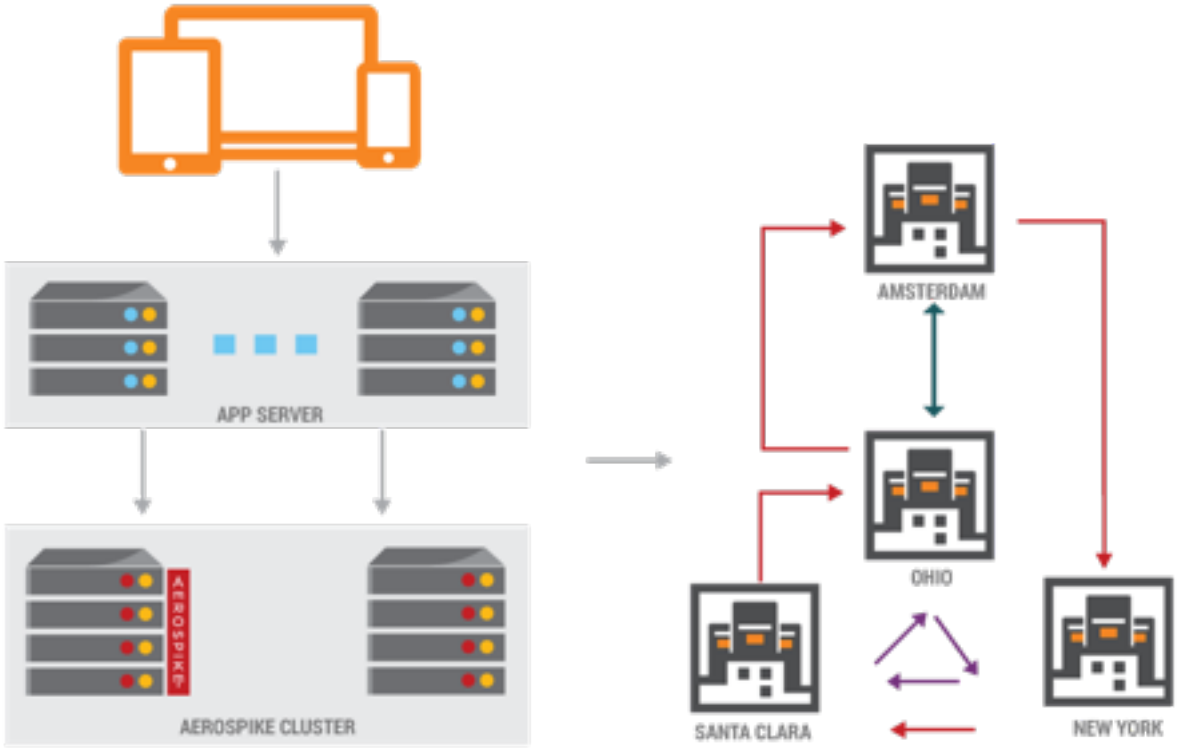


OTHER DATABASES

	DRAM & HDD	SSD & DRAM
Storage per server	180 GB (on 196 GB server)	2.8 TB (4 x 700 GB)
TPS per cluster	500,000	500,000
Cost per server	$8,000	$11,000
Server costs	$1,488,000	$154,000
Power/server	0.9 kW	1.1 kW
Power (2 years) $0.12 per kWh ave. US	$352,2002	$32,400
Maintenance(2 years) $3600/server	$670,000	$5042
Total	**$2,510,000**	**$236,800**

ARCHITECHTURE

Architecture – The Big Picture



1) **No Hotspots**
– DHT simplifies data partitioning

2) Smart Client – **1 hop** to data, no load balancers

3) **Shared Nothing** Architecture, every node identical

4) Single row **ACID**
– synch replication in cluster

5) Smart Cluster, **Zero Touch**
– auto-failover, rebalancing, rack aware, rolling upgrades..

6) Transactions and long running tasks prioritized real-time

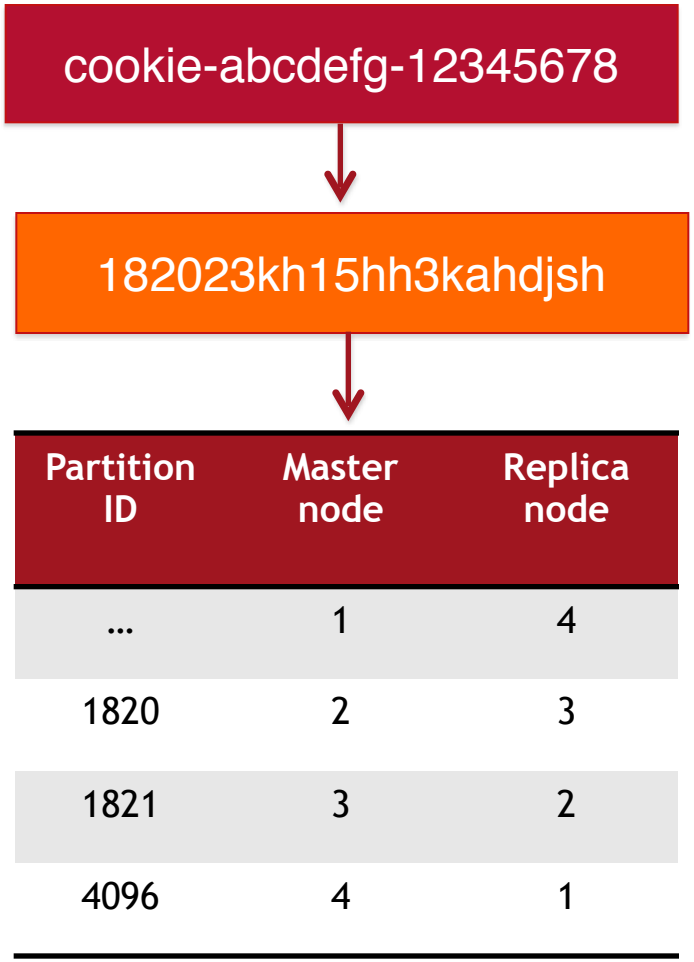
7) **XDR** – sync replication across data centers ensures **Zero Downtime**

8) **Scale linearly** as data-sizes and workloads increase

9) Add capacity with **no service interruption**

ROBUST DHT TO ELIMINATE HOT SPOTS

How Data Is Distributed (Replication Factor 2)



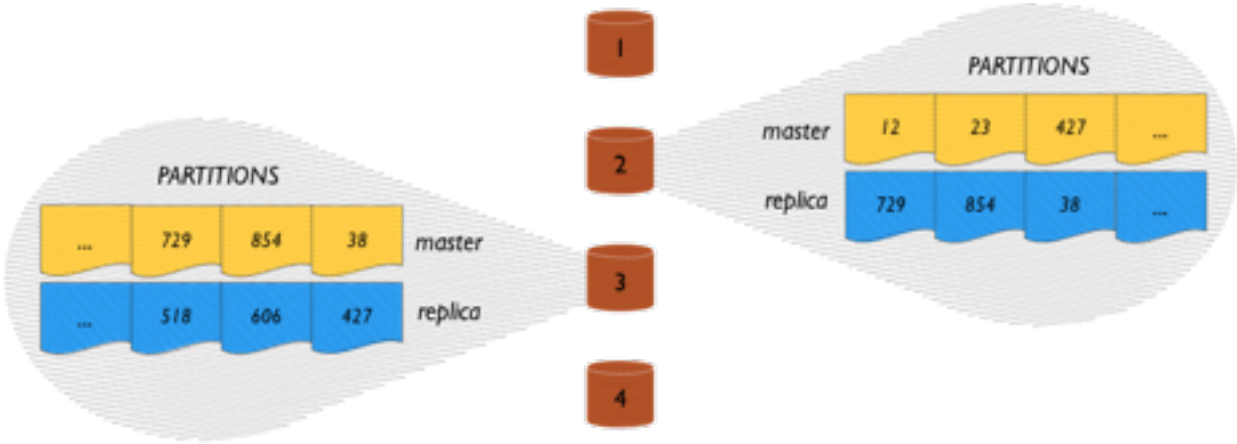
Partition ID	Master node	Replica node
...	1	4
1820	2	3
1821	3	2
4096	4	1

- Every key is hashed into a 20 byte (fixed length) string using the RIPEMD160 hash function
- This hash + additional data (fixed 64 bytes) are stored in RAM in the index
- Some bits from this hash value are used to compute the **partition id**
- There are 4096 partitions
- Partition id maps to node id based on cluster membership

Data Distribution

Data is **distributed evenly** across nodes in a cluster using the Aerospike Smart Partitions™ algorithm.

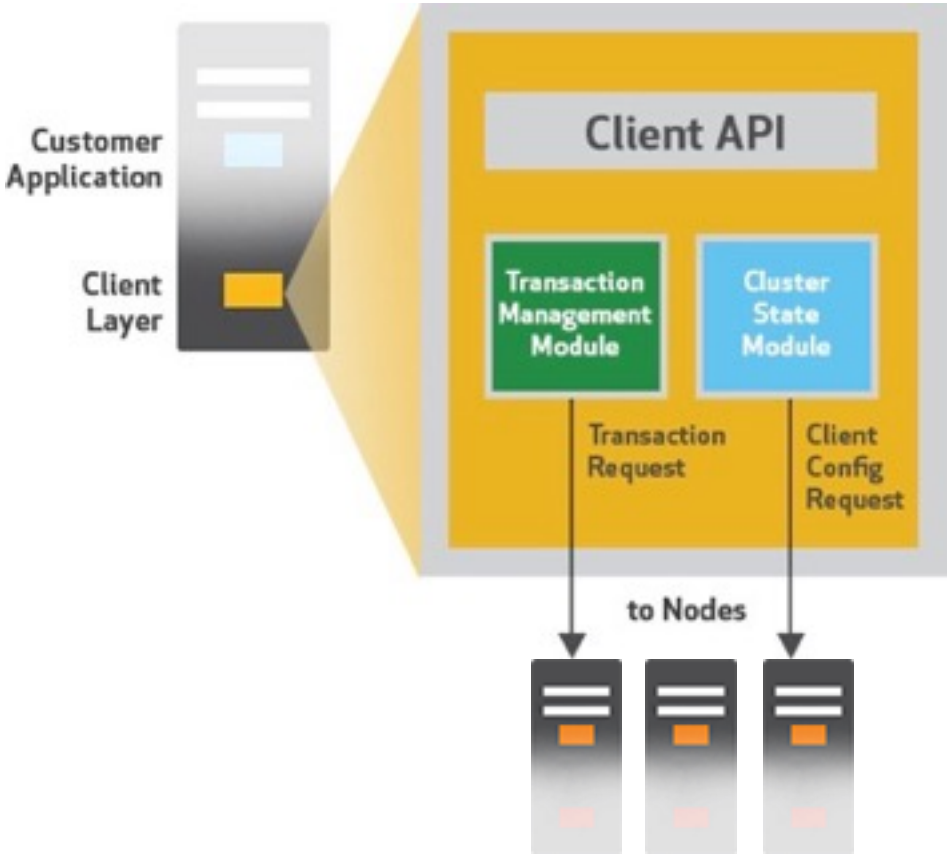
- Even distribution of
  - Partitions across nodes
  - Records across Partitions
  - Data across Flash devices
- Primary and Replica Partitions



INTELLIGENT CLIENT TO MAKE APPS SIMPLER

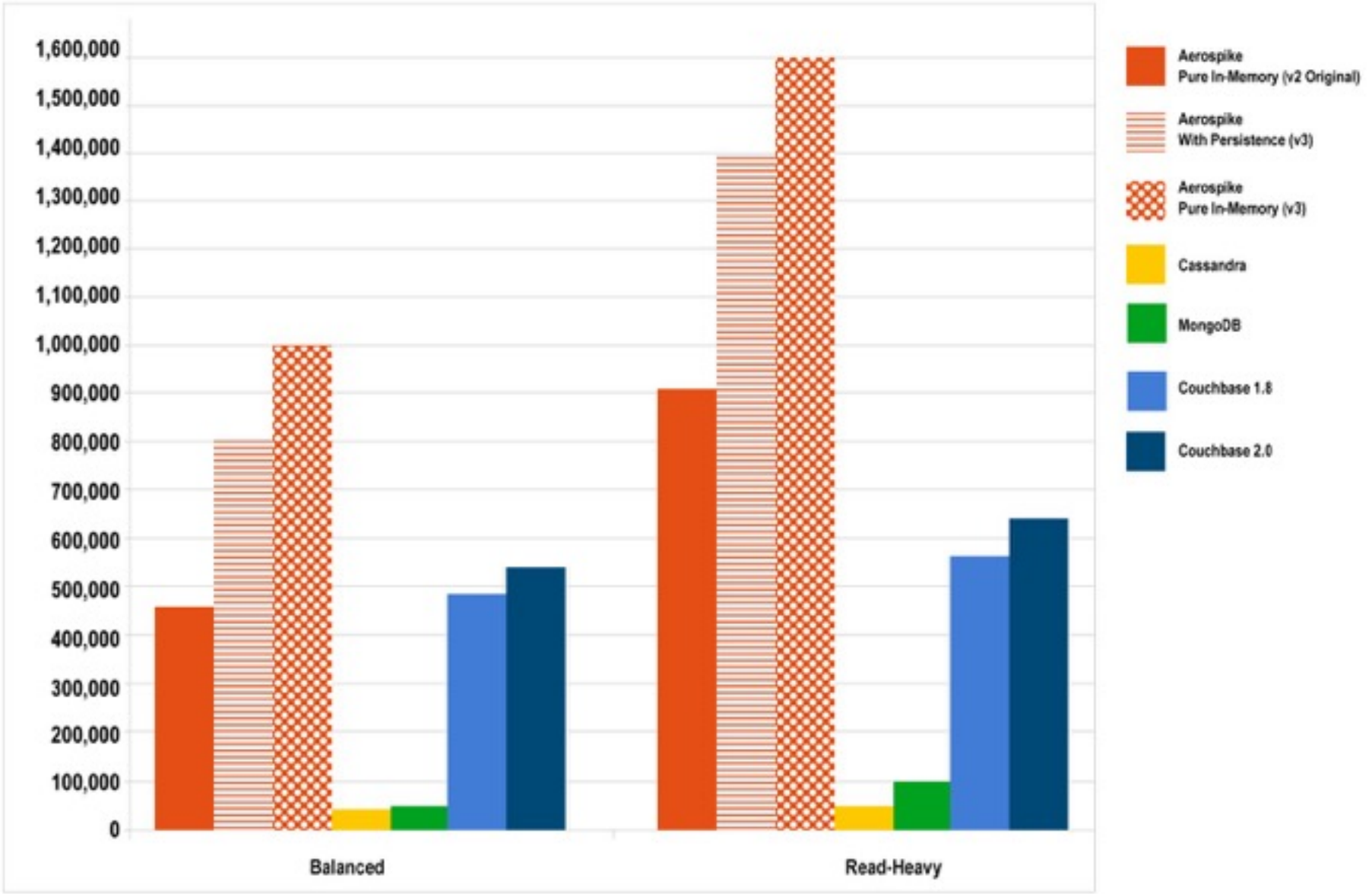
Shield Applications from the Complexity of the Cluster

- Implements Aerospike API
  - Optimistic row locking
  - Optimized binary protocol
- Cluster tracking
  - Learns about cluster changes, partition map
- Transaction semantics
  - Global Transaction ID
  - Retransmit and timeout
- Linear scale
  - No extra hop
  - No load balancers

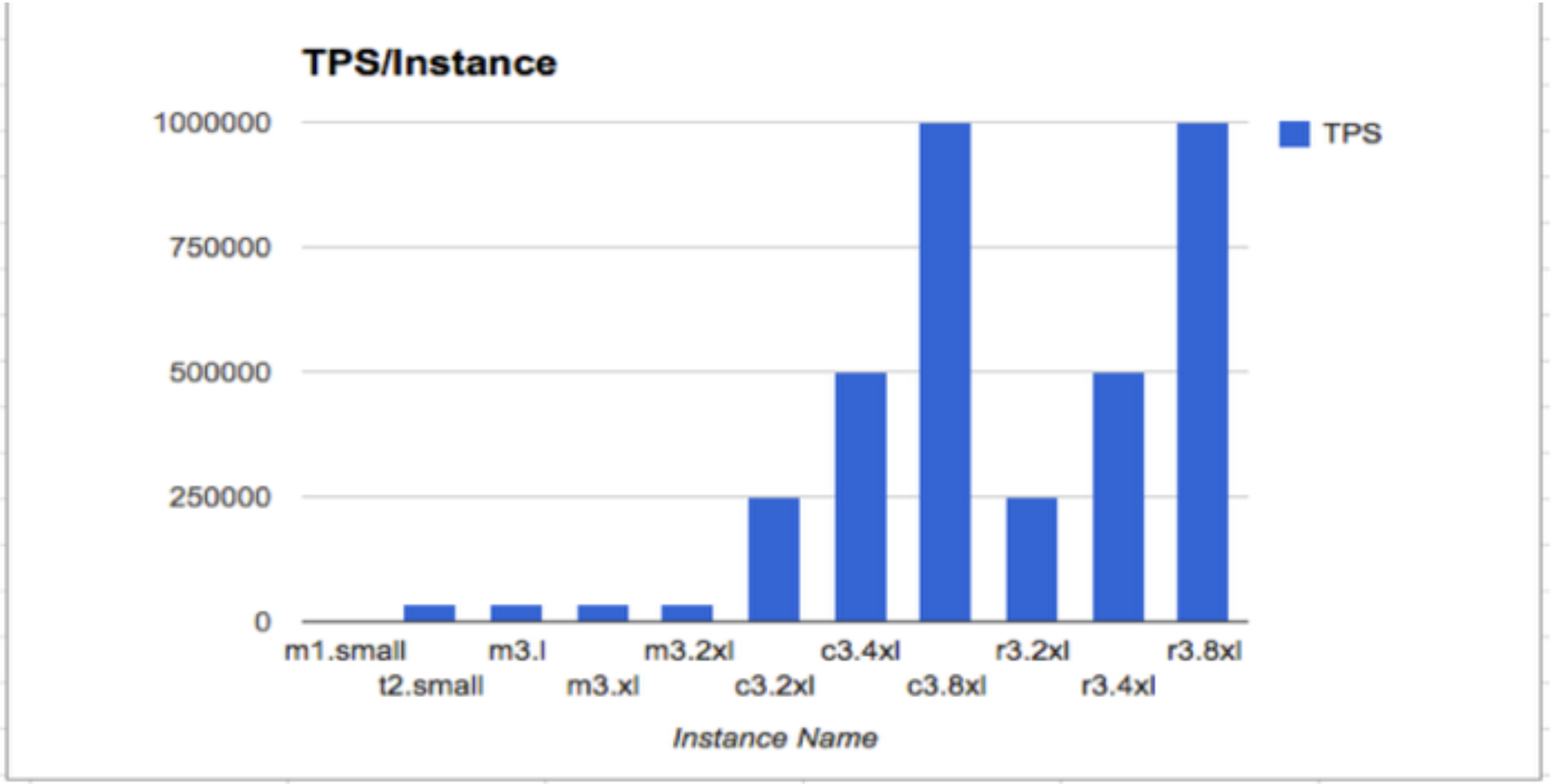


Single Server YCSB Performance

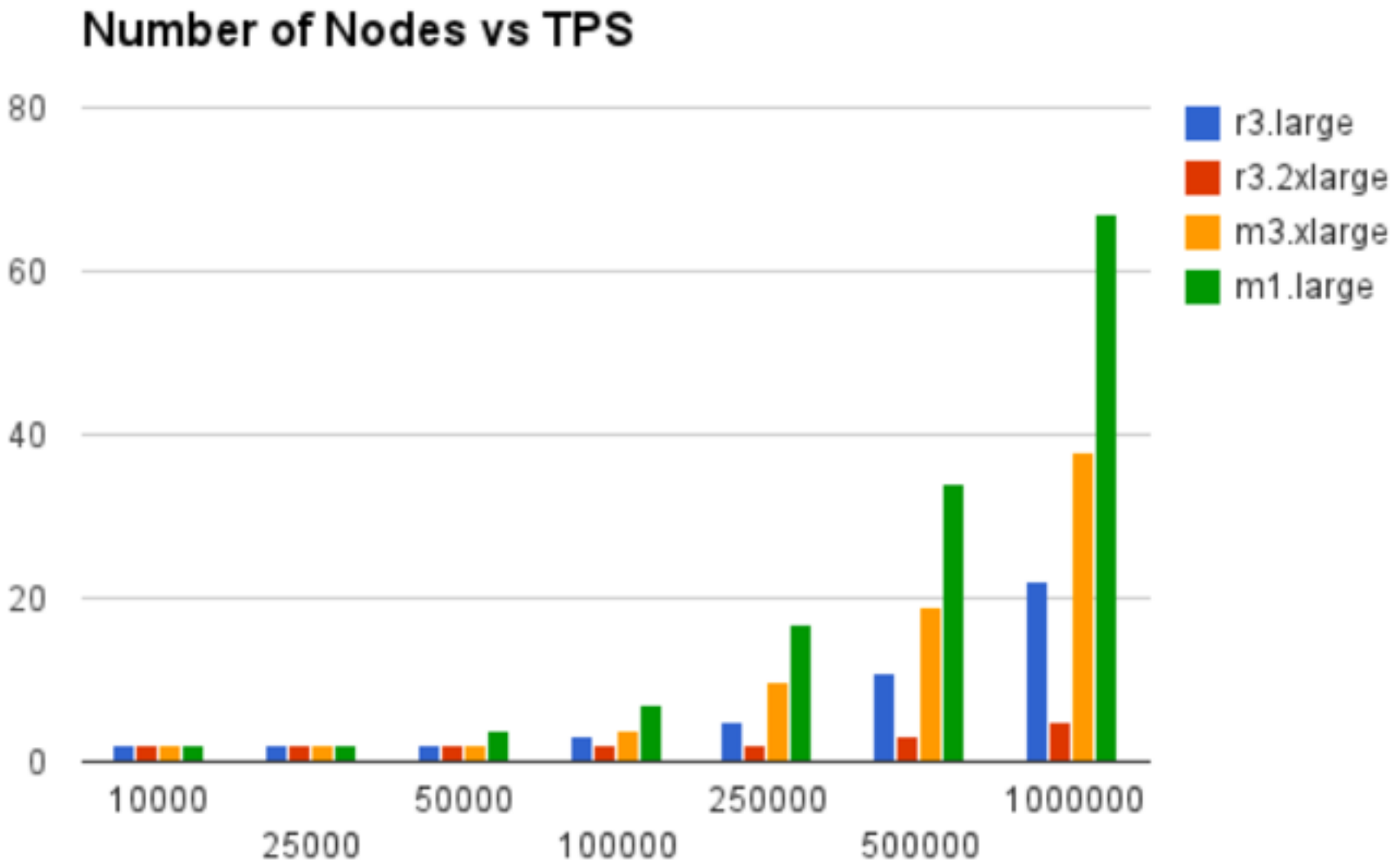
YCSB Benchmark Test 3, Fig 5: Updated with Aerospike 3 numbers



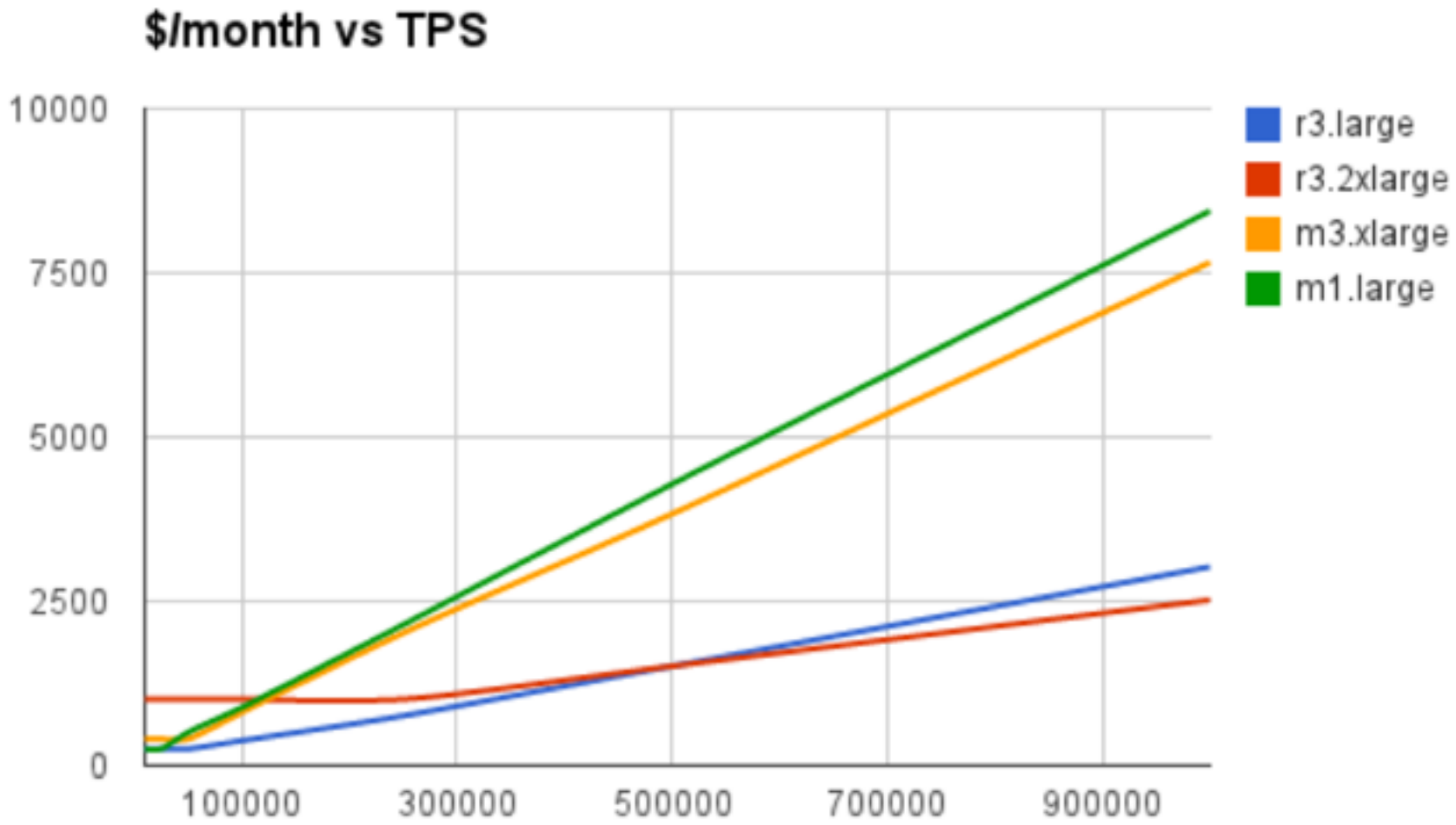
Amazon EC2 results



Amazon EC2 results



Amazon EC2 results



IMPLEMENTATION MATTERS

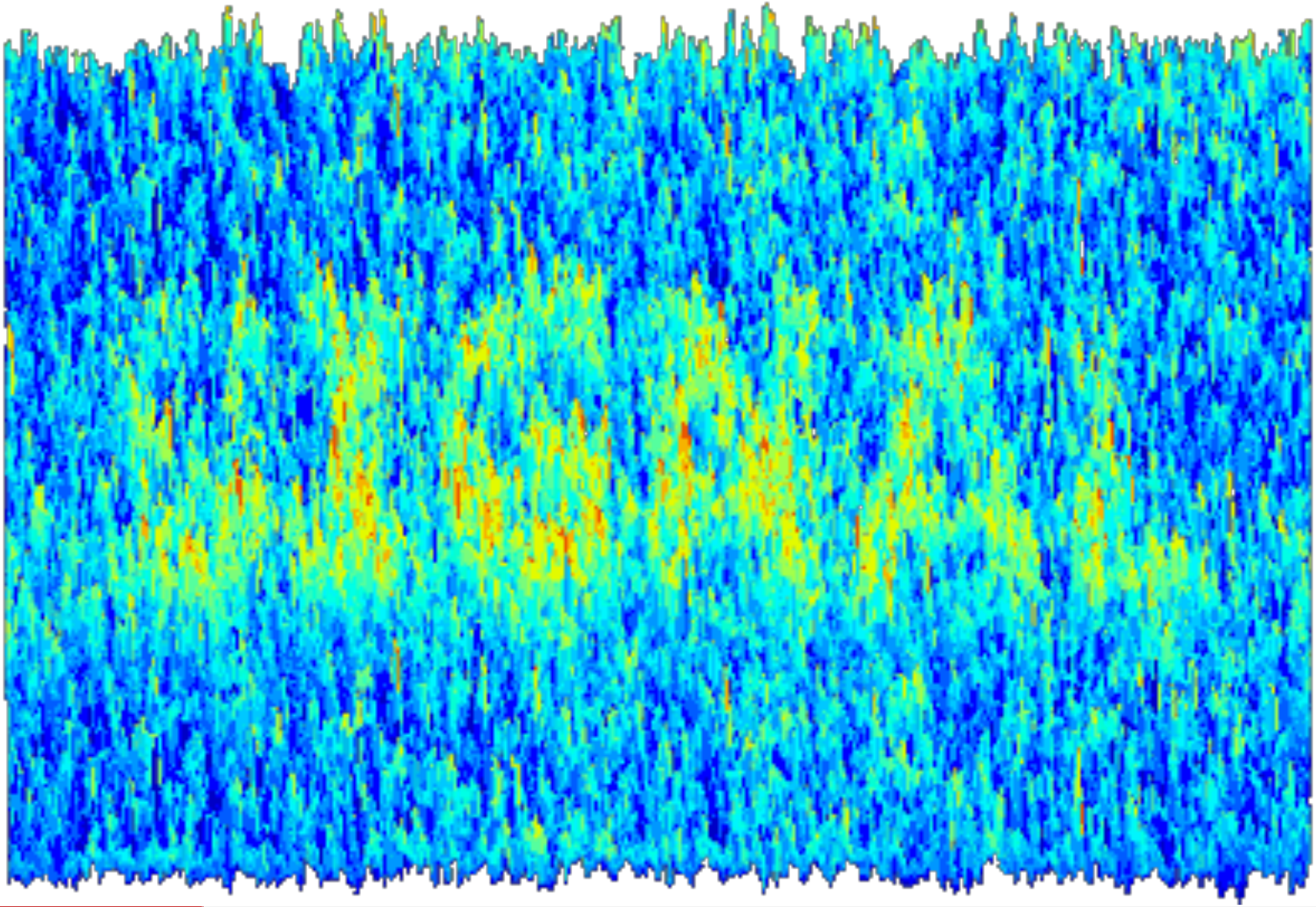
Implementation Matters

1. Optimize Key-Value code paths
   - No hot spots (e.g., robust DHT)
   - Scales up easily (e.g., easy to size)
   - Avoids points of failure (e.g., single node type)
   - Binary protocol
2. Code in C
   - Read() / Write() / Linux AIO (don't trust a library)
   - Multithreading
   - Direct device access
3. Memory allocation matters
   - Stack-based allocators
   - Own stack allocator
   - JEMalloc for pools (less memory fragmentation, SMP optimized)

Implementation Matters

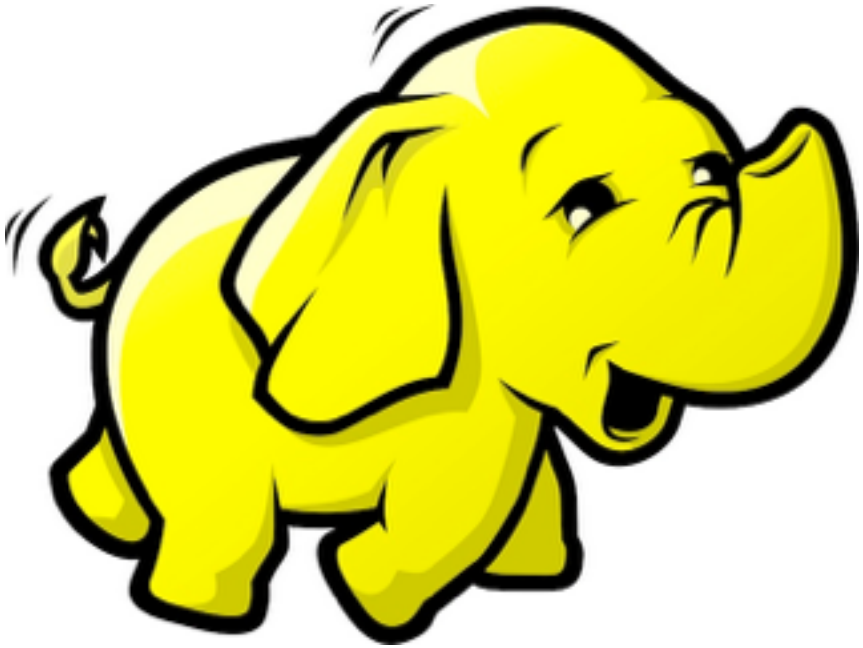
4. Masters in a shared nothing system
   - Fast cluster organization
   - Fast transaction capabilities
   - Can be CP or AP - and resolve data accurately
5. Client libraries are hard (so we do it for you)
   - Fast stable connection pools are hard
   - API design matters
   - Slow languages need Aerospike more

Hot Analytics - Signal in Noise



ANALYTICS - TECHNOLOGIES

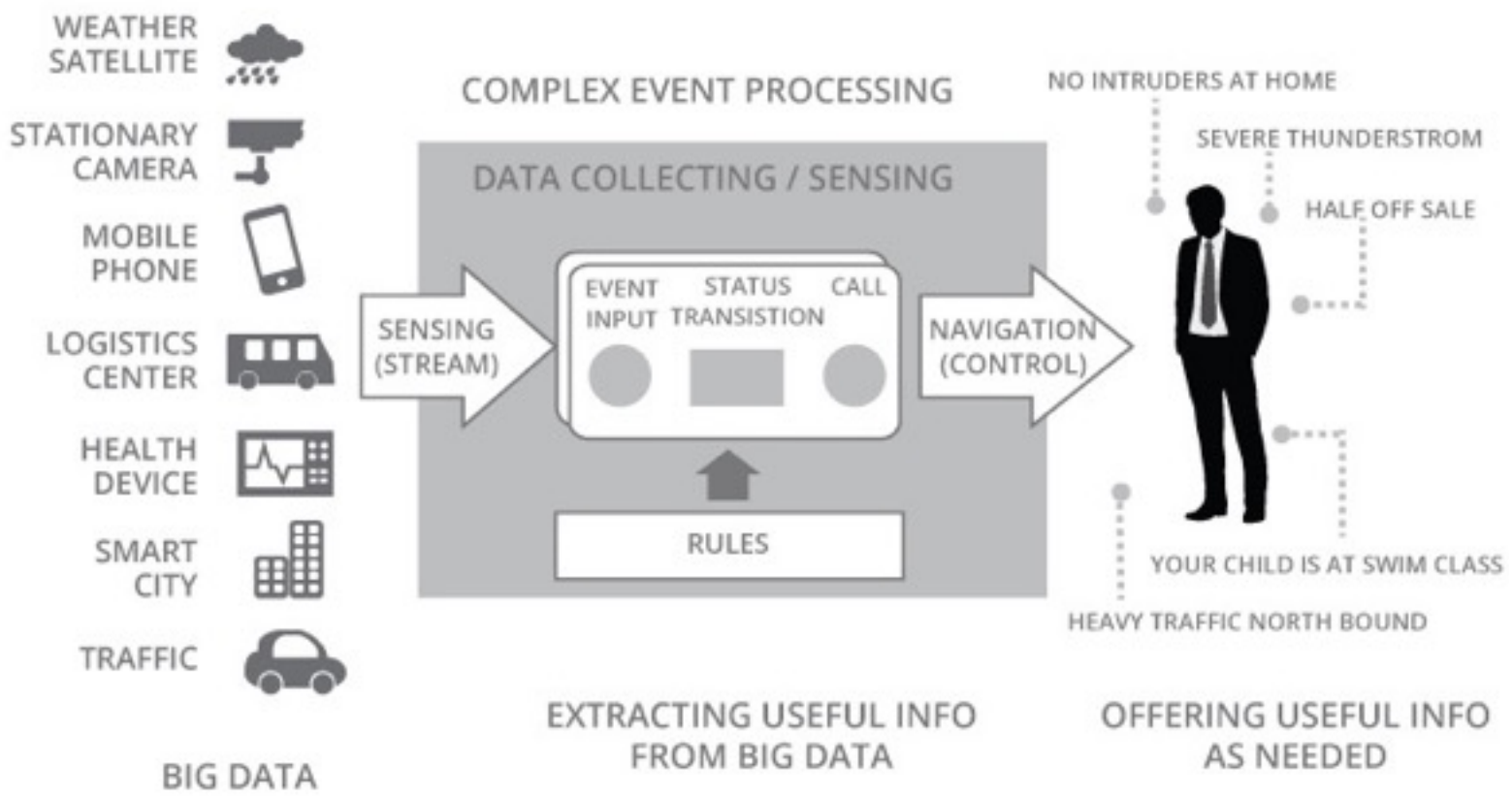
Hadoop



Large and capable, but not fast.

http://www.aerospike.com/community/labs/

Complex Event Processing (CEP)



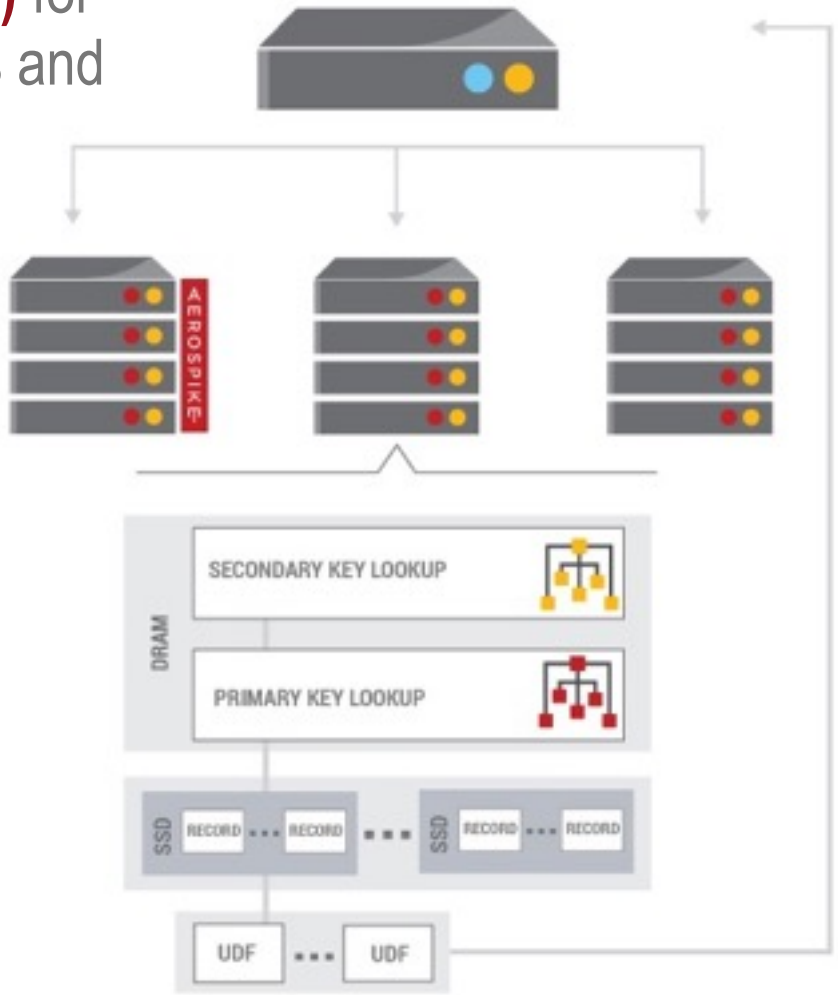
Spark

Key Challenges

- Handle extremely high rates of read/write transactions with concurrent real-time analytics
- Avoid hot spots
  - On a node
  - An index
  - A key
- Pre-qualify data to be processed in Map Reduce
- Maximize parallelism
- Minimize programmer complexity
- In **Real-time**

Queries + User Defined Functions = Real-Time Analytics

User Defined Functions (UDFs) for real-time analytics and aggregations



STREAM AGGREGATIONS

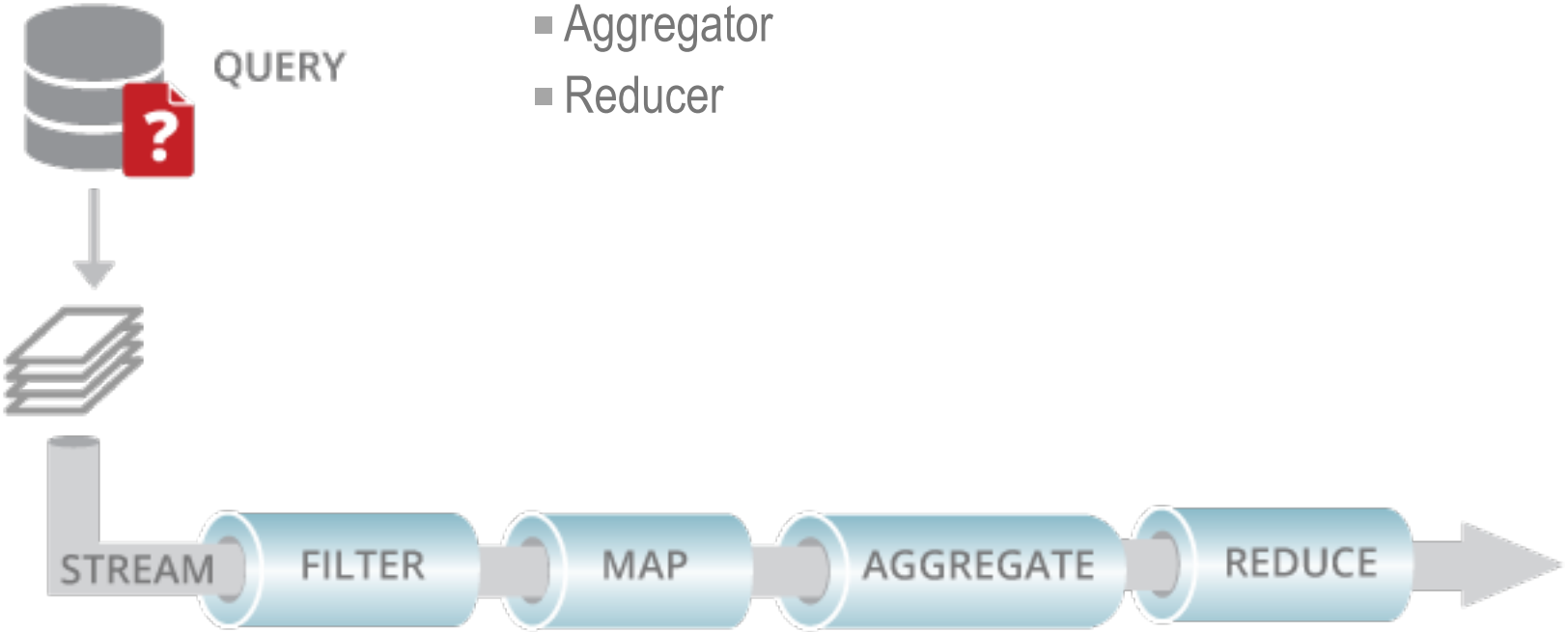
(INDEXED MAP-REDUCE)

Pipe Query results through UDFs

- Filter, Transform, Aggregate.. Map, Reduce

Conceptual Stream Processing

- Output of a query is a **Stream**
- Stream flows through
  - Filter
  - Mapper
  - Aggregator
  - Reducer



Hot Analytics Scnario – Airline Late Flights

Data

- Airline flights in the USA January 2012
- 1,050,000 flight records

Task

- On a specific date
  - Which Airline had late flights?
  - How many flights?
  - How many were late?
  - Percentage late flights?

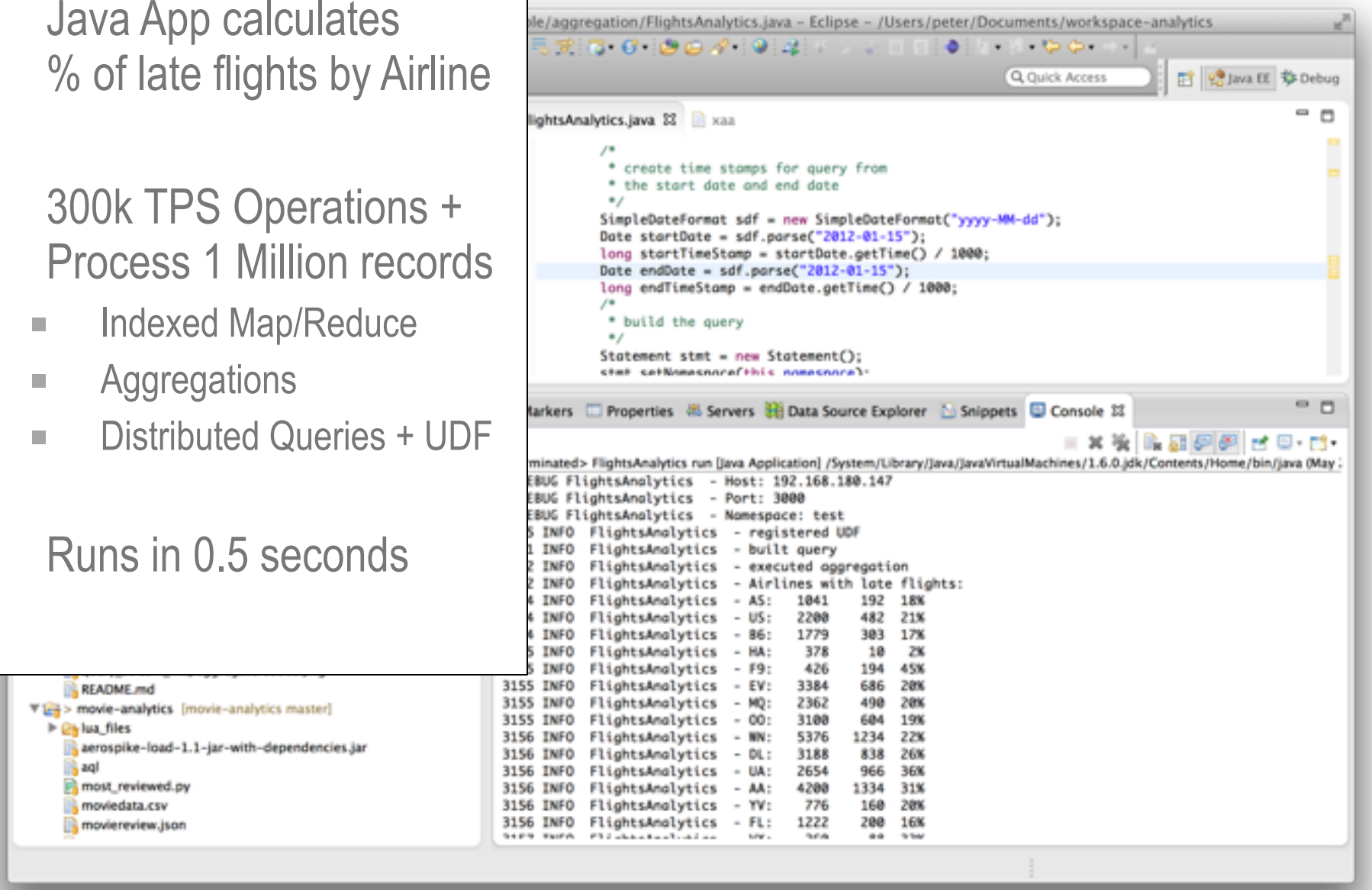
Performance Requirements

- Results in < 1 Sec
- No impact on production transaction performance (300K TPS)

GitHub Repo - https://github.com/aerospike/flights-analytics

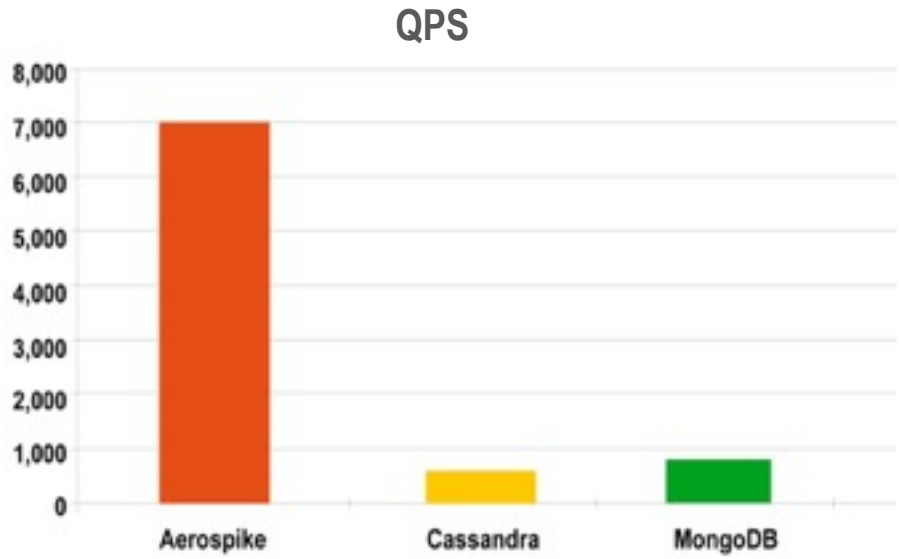
Operations (300k TPS) + Analytics (Indexed Map/Reduce)

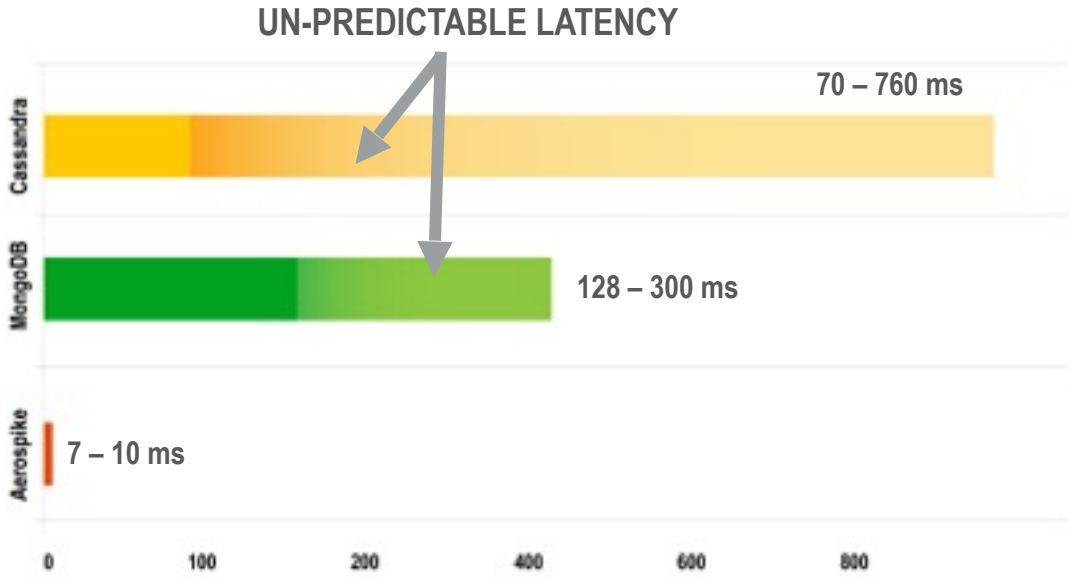
- Java App calculates % of late flights by Airline
- 300k TPS Operations + Process 1 Million records
  - Indexed Map/Reduce
  - Aggregations
  - Distributed Queries + UDF
- Runs in 0.5 seconds

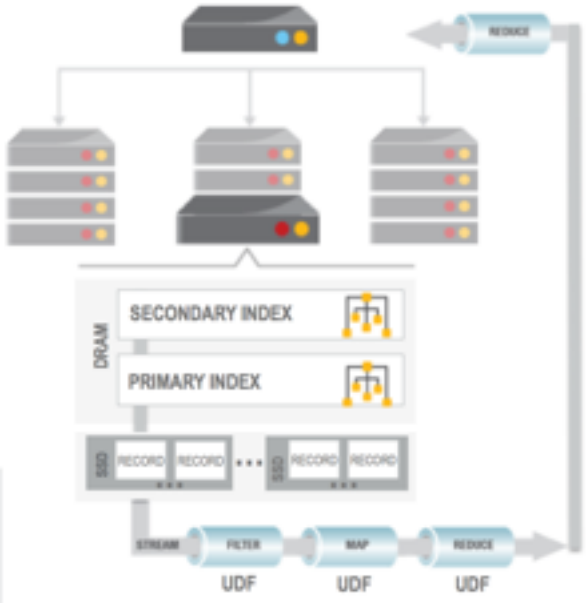


Key-Value with Analytics

Add basic analytics capability to improve measurements and metrics for your highest velocity data







SUMMARY

- Support for Popular Languages and Tools
  - AQL and Aerospike Client in C, Java, C#, Go, Node, Ruby, Python, …
- Complex Data Types
  - Nested documents (map, list, string, integer)
  - Large (Stack, Set, List) Objects
- Queries
  - Single Record
  - Batch multi-record lookups
  - Equality and Range
  - Aggregations and Map-Reduce
- User Defined Functions
  - In-DB processing
- Aggregation Framework
  - UDF Pipeline
  - MapReduce
- Time Series Queries
  - Just 2 IOPs for most r/w (*independent of object size)*

Aerospike: The Trusted In-Memory NoSQL

Performance

- Over 20 trillion transactions per month
- 99% of transactions < 2 ms
- 150K TPS per server

Scalability

- Billions of Internet users
- Clustered Software
- Maintenance without downtime
- Scale up & scale out

Reliability

- 50 customers; zero down-time
- Immediate Consistency
- Rapid Failover; Data Center Replication

Price/Performance

- Makes impossible projects affordable
- Flash-optimized
- 1/10 the servers required


Open Source
Straight Ahead



Speed + Scale + Reliability =

AEROSPIKE

The power of 3

Questions and Answers

@parshua

khosrow@aerospike.com