Have your cake, and eat it too

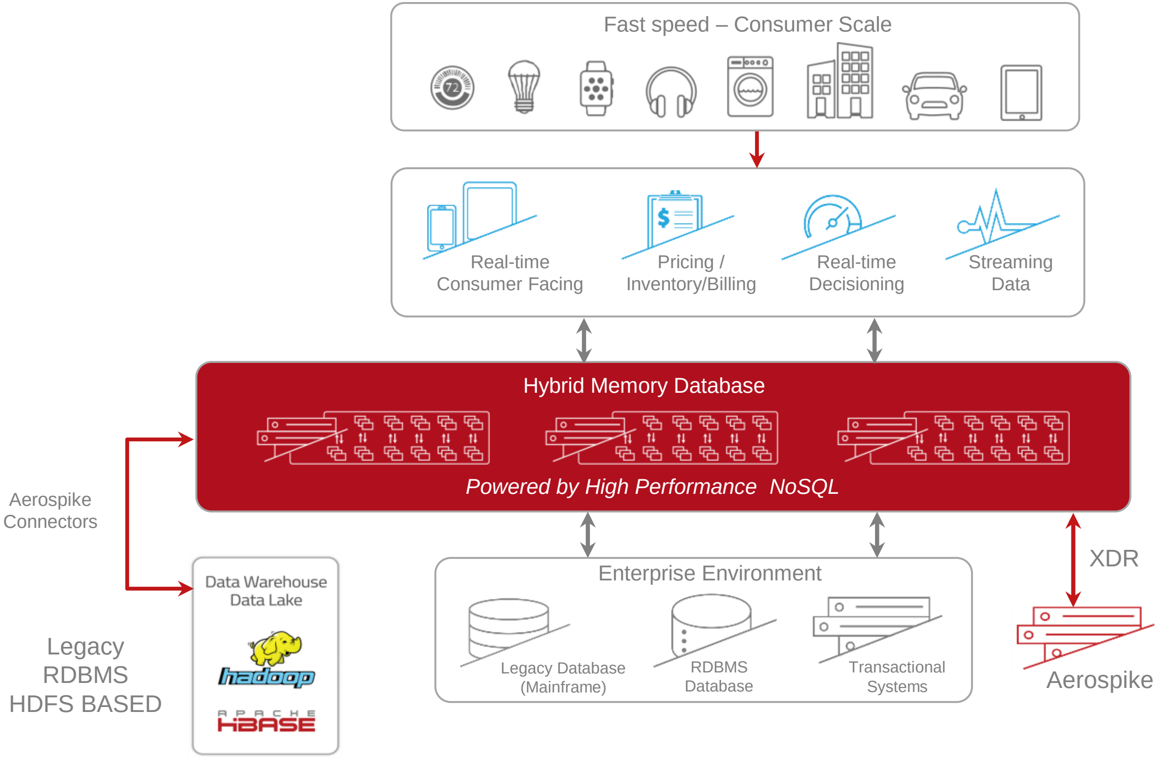
Strong Consistency and High Performance

Brian Bulkowski, CTO & Founder
March 7, 2018

Qcon London

Aerospike in a nutshell

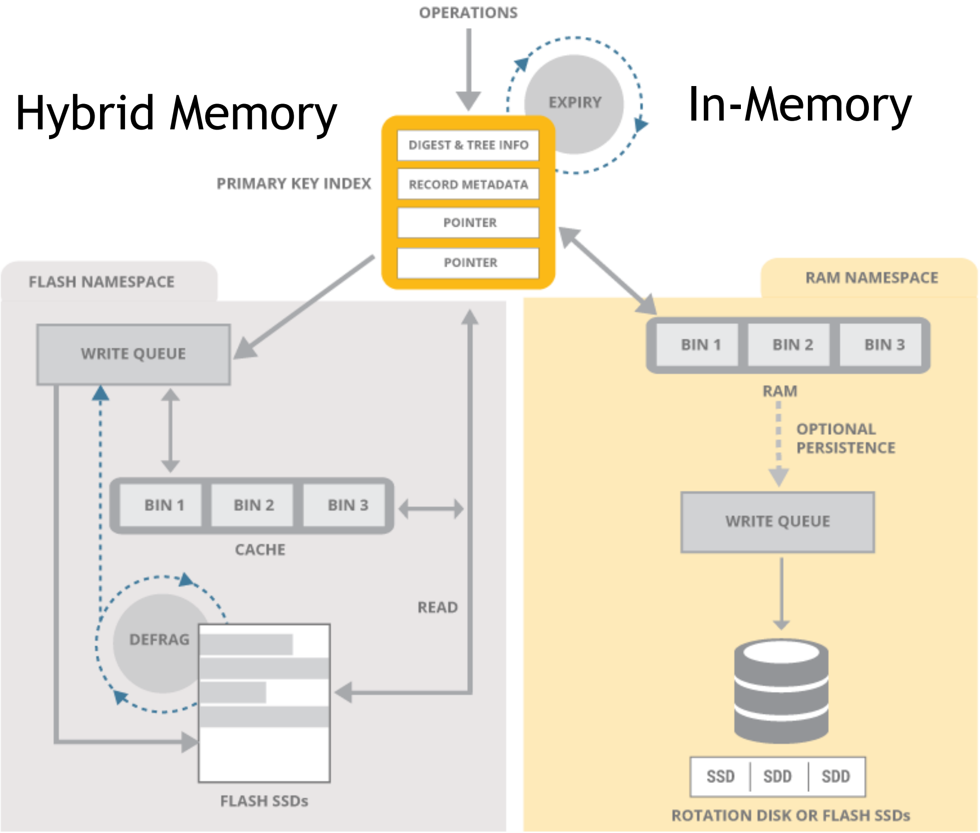
Hybrid Memory Enables Digital Transformation



Benefits:

- Simplicity
- Maintainability
- Durability
- Consistency
- Scalability
- Cost ($)
- Data Lag Reduced

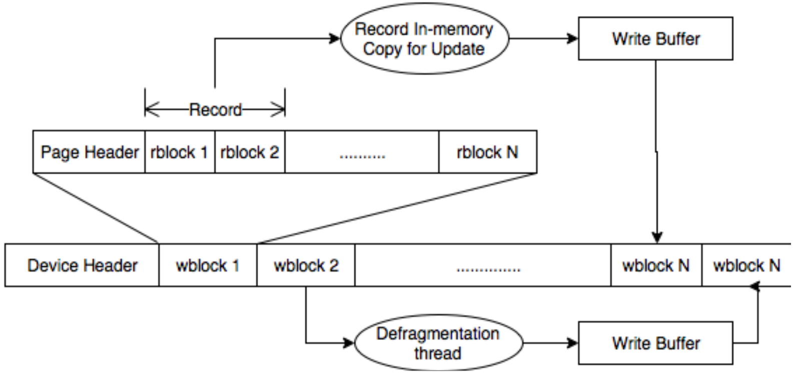
Aerospike Storage Architecture (HMA+)



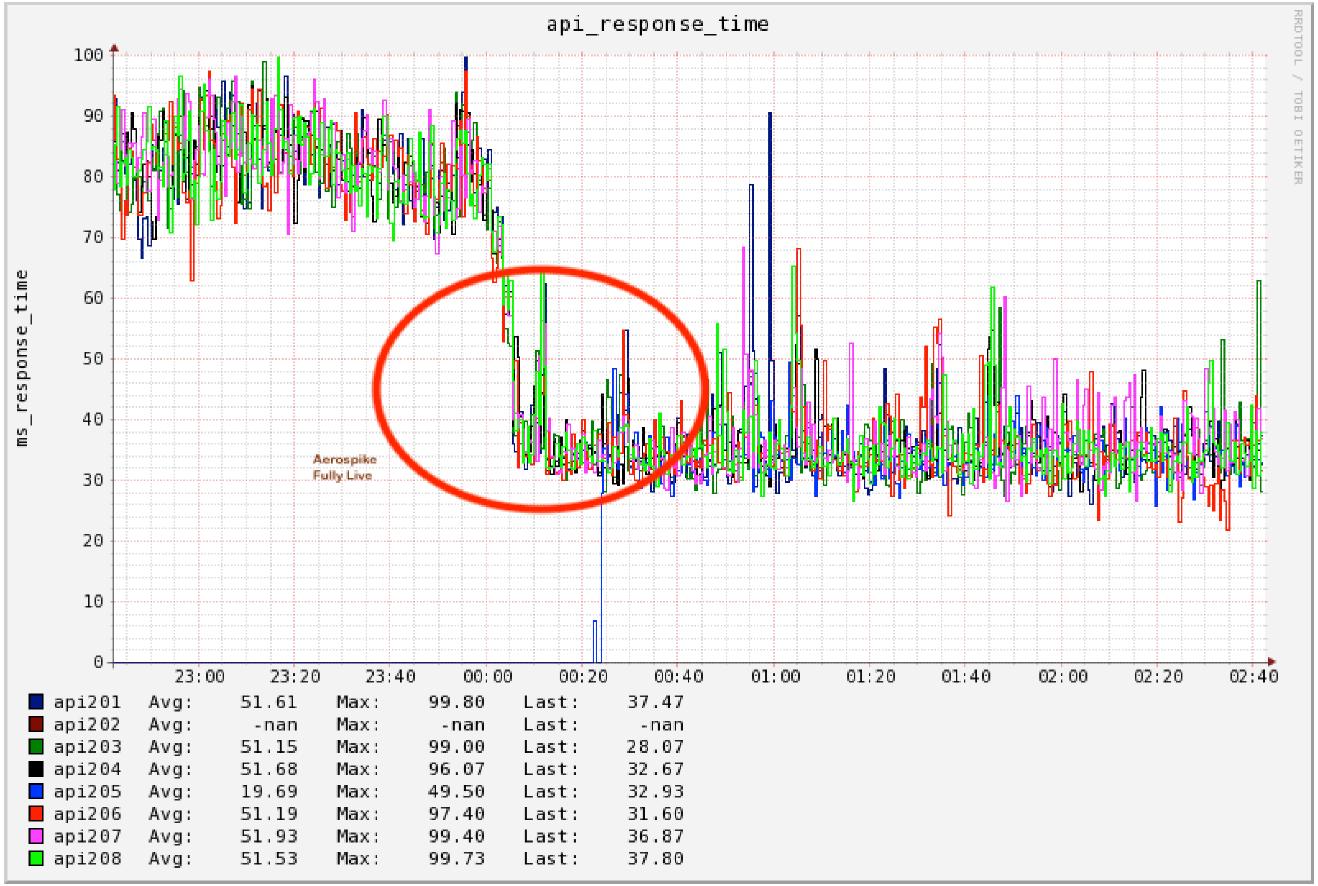
Highlights

1. Direct device access
2. Large Block Writes
3. Indexes in DRAM
4. Highly Parallelized
5. Log-structured FS “copy-on-write”
6. Fast restart with shared memory

Storage Layout



SLA: Aerospike versus Cassandra



🟦	api201	Avg:	51.61	Max:	99.80	Last:	37.47
🟥	api202	Avg:	-nan	Max:	-nan	Last:	-nan
🟩	api203	Avg:	51.15	Max:	99.00	Last:	28.07
⬛	api204	Avg:	51.68	Max:	96.07	Last:	32.67
🟦	api205	Avg:	19.69	Max:	49.50	Last:	32.93
🟥	api206	Avg:	51.19	Max:	97.40	Last:	31.60
🟪	api207	Avg:	51.93	Max:	99.40	Last:	36.87
🟩	api208	Avg:	51.53	Max:	99.73	Last:	37.80

Features

Hybrid Memory Architecture

Indexes in DRAM and data in Flash provide storage guarantees, and uniquely unlock Flash’s performance.

Data structure & Document indexing

Easy-to-program data structures such as lists, maps, which improve performance and allow indexed lookups on field values.

Non Volatile Memory (NVM)

Allows you to store large volumes of data with extremely high performance & consistency. Novel Flash-first algorithms provide unique performance benefits.

Enterprise Security

Full transport encryption. Support for encrypted storage and in-database transparent data encryption. Authentication, access control, exception logging.

Primary Key Consistency

4.0 provides Strong Consistency on primary key access, with Jepsen test results. Commit to device support for demanding correctness environments.

High Availability (AP mode)

Used traditionally for high performance machine-to-machine analytic systems, such as fraud detection, pattern matching, and real-time pricing.

Real-time Analytic Framework

Deep integration with Spark to allow SQL and ML based tools to operate seamlessly and efficiently

Geographic Replication (XDR)

Multiple datacenters stay in sync through high performance replication.

Case Studies: HMA - Lower TCO & better SLA

Customer	Situation	Problem	Hybrid Memory System
Trading Account Account Status, Trades, Risk	DB2+Gemfire cache	150 Servers growing to 1000	Single cluster – 12 servers
Fraud Detection	2 ORCL RAC clusters + Terracotta cache	System Stability & missing SLA's	3 Clusters – 20 Servers each
User Integrity Checking for Internet Transactions	DataStax/Cassandra	168 DataStax Servers growing to 450+	30 Servers – 2 clusters
Customer 360 and Rich Consumer Application	Green Field / Oracle / X.500	Largest Telco needs "MyService" application, integrated customer DB	15 Servers – 2 clusters
Telco Device and User Access	ORCL Coherence / DataStax Cassandra	Existing SOE solutions unstable & Costly	5 successful POC's
Telco Revenue Assurance	DataStax/Cassandra PostgreSQL + cache	Hundreds of cache & Cassandra Servers Scalability challenges	Significant reduction of server footprint – global deployment

Vertical Focus / Horizontal Expansion

ADTECH

appnexus
Aol.
BrightRoll
TAPAD
NIELSEN MARKETING CLOUD
FEDERATED MEDIA
madvertise
adMarketplace.COM
Performance-Driven Pay Per Click Advertising
DataXu
convertmedia
APPLOVIN
adform
ADMETA
acuity
AdIQuity
coxdigitalsolutions
bluecava
CROSS-SCREEN STARTS HERE
theTradeDesk
exponential
Advertising Intelligence
Cazamba
InMobi
Chango
digital turbine
yashi
IGNITIONONE
MARKETING TECHNOLOGY. SIMPLIFIED.
CONVERSANT
rubicon
PROJECT
komli
sitescout
Fringe81
[x+1]

ECOMMERCE

ebay
WILLIAMS-SONOMA, INC.
snapdeal.com
bookmyshow
ZETTATA

GAMING / BETTING

CURSE
GREE
Riot
GAMES
CyberAgent.
DRAFTKINGS

FINANCIAL

PayPal
ThreatMetrix.

TELECOM

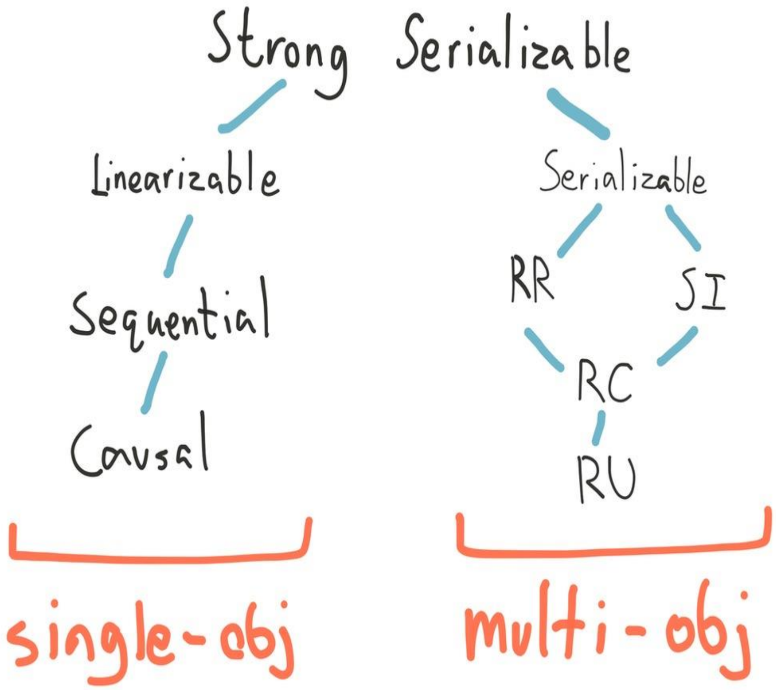
Alcatel·Lucent
Hewlett Packard
Enterprise
neustar
airtel
KYOCERA
NOKIA
So-net
NTT

TECHNOLOGY

FORENSIQ
ShareThis
KAYAK
PubMatic
f4m
new media solutions
Clicktale
Adobe
AmerisourceBergen

Strong Consistency
High Performance

What is Strong Consistency?



Strong Consistency Concepts

Strong Consistency: Data viewed immediately after an update will be the same for all observers of the entity

Linearizability: Provides a **real-time** (i.e., **wall-clock**) **guarantee** on **reads/writes** on a **single object (no stale reads)**

Sequential Consistency: All processes see **shared accesses** in the same order. Accesses are **not** ordered in **real-time**

Causal Consistency: All processes see only **causally-related** shared accesses in the same order.

DuDe,
Where's my Data?

How data can be lost

Data Location Updates: Server that should hold the data has changed, and not everyone is informed

Asynchronous Replication: A crash occurs before a write has been applied to enough servers

Buffered Writes: A crash occurs before data is written to persistent storage

Clock Problems: A subsequent update is applied to a server with a clock in the past

Bugs: A correct architecture, poorly implemented

Not enough performance

“Safe” but not enough write throughput

Queues back up, error codes are returned,
and you’ve got nowhere to put the data

“Safe” but impractical

Why doesn’t NoSQL talk about ACID?

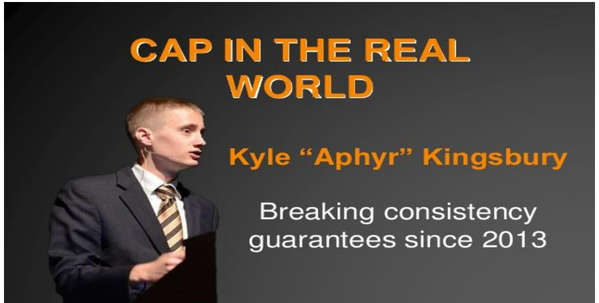
Atomicity: Multi-record transactions are all or nothing	**NO**: NoSQL is (mostly) parallel, single-record operations
Consistency: All states and constraints are maintained	**YES:** The only constraint is the record update
Isolation: All transactions are executed as if there was single sequential application timeline	**NO**: A single application timeline is not practical or desired at Internet scale
Durability: Writes survive power losses, crashes, errors	**YES:** Durability matters, and also network partitions (CAP)

Does your database lose data?

Jepsen (Kyle Kingsbury) to the rescue!

Independent, Open Source Testing

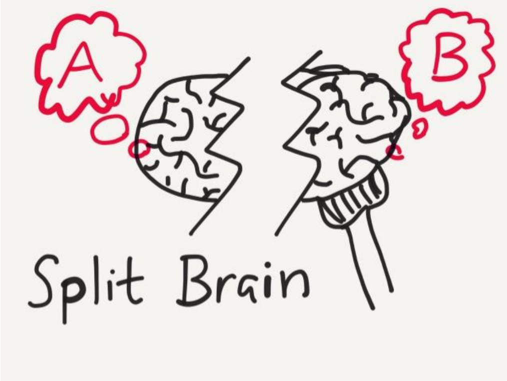
http://jepsen.io/

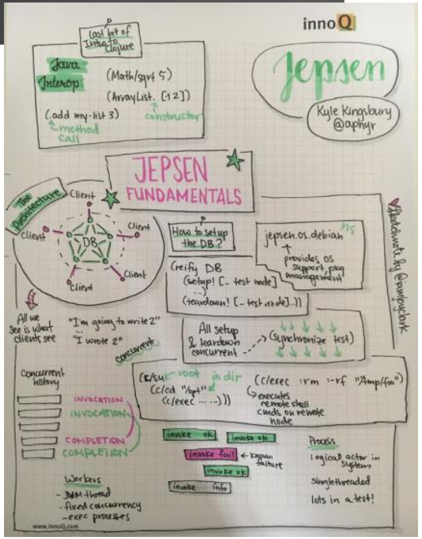


CAP IN THE REAL WORLD

Kyle “Aphyr” Kingsbury

Breaking consistency guarantees since 2013





Jepsen Results

Cassandra (2013): No

Redis (2013): No

Aerospike (2015): No

Mongo (2017) : Yes!

Cockroach (2017): Probably?

Aerospike (2018): Watch this space!

Note: *Jepsen is not pass fail!*

It's a discussion of the product claims vs reality.

You have to read and understand.

Which is hard.

Aerospike 4.0
Strong Consistency

Aerospike 4.0 Strong Consistency with Hybrid Memory

STRONG CONSISTENCY

AEROSPIKE

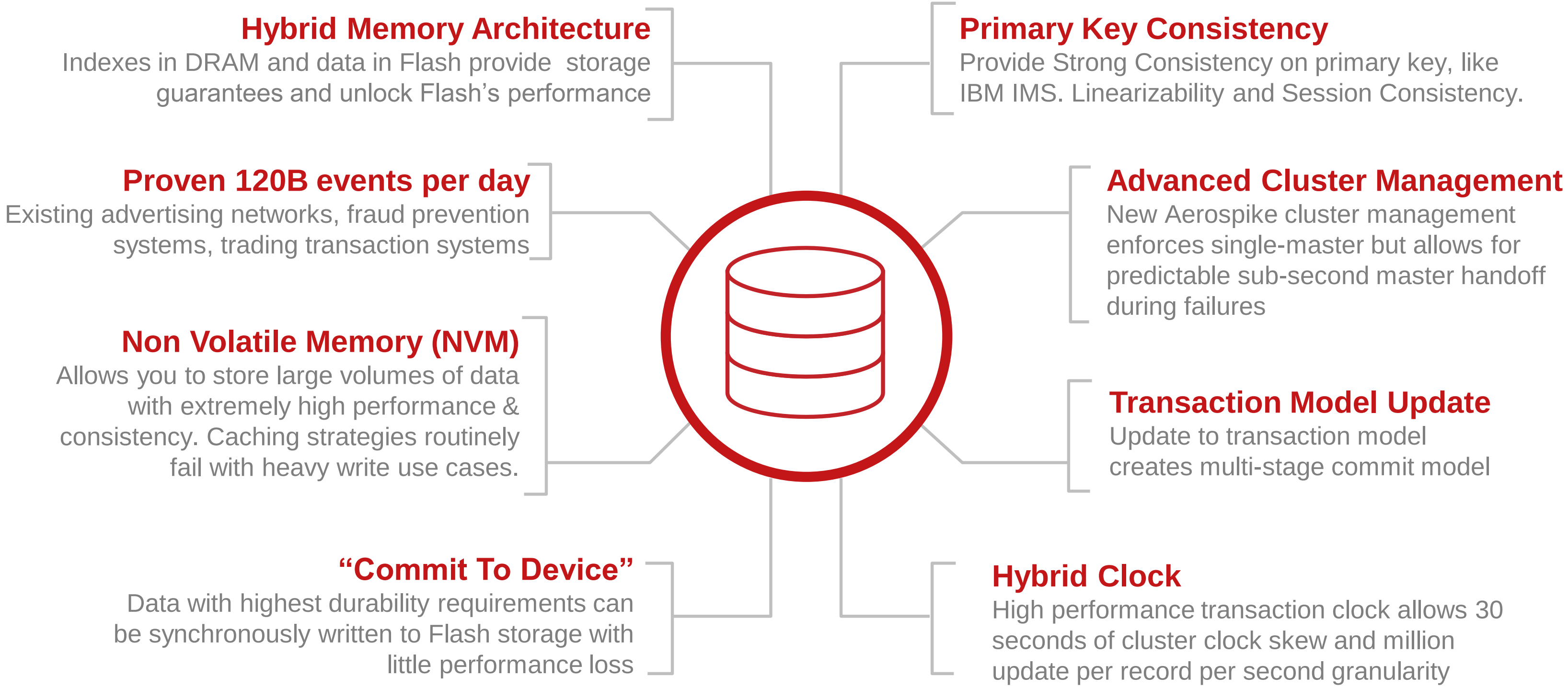
HIGH PERFORMANCE

Early Access

Fast System of Record
Enterprise System of Engagement

Vital Customer Experiences
In-flight Analytics – Risk & Fraud
Transformative Customer Services

How did we achieve this?



Hybrid Memory Architecture
Indexes in DRAM and data in Flash provide storage guarantees and unlock Flash's performance

Proven 120B events per day
Existing advertising networks, fraud prevention systems, trading transaction systems

Non Volatile Memory (NVM)
Allows you to store large volumes of data with extremely high performance & consistency. Caching strategies routinely fail with heavy write use cases.

"Commit To Device"
Data with highest durability requirements can be synchronously written to Flash storage with little performance loss

Primary Key Consistency
Provide Strong Consistency on primary key, like IBM IMS. Linearizability and Session Consistency.

Advanced Cluster Management
New Aerospike cluster management enforces single-master but allows for predictable sub-second master handoff during failures

Transaction Model Update
Update to transaction model creates multi-stage commit model

Hybrid Clock
High performance transaction clock allows 30 seconds of cluster clock skew and million update per record per second granularity

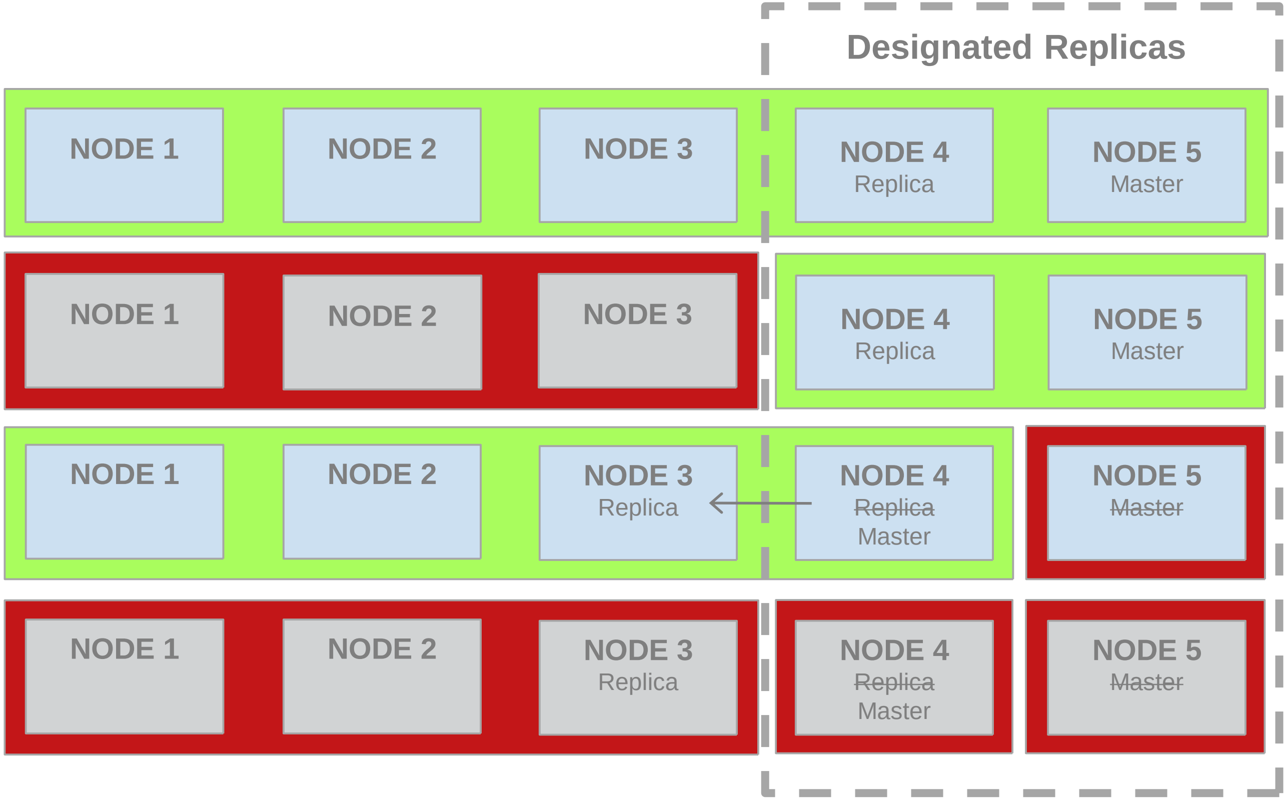
Performance?

It’s great!

	Linearize SC	Session SC	Availability (AP)
Read TPS	1,500,000	4,700,000	4,800,000
Write TPS	370,000	1,200,000	1,200,000

(5 node cluster, bare metal, DRAM data, 10 byte objects stress transaction system)

Master & Replica Availability and Promotion



Cluster Healthy

SPLIT – Rule 1

All designated replicas in a subcluster, and the data

SPLIT – Rule 2

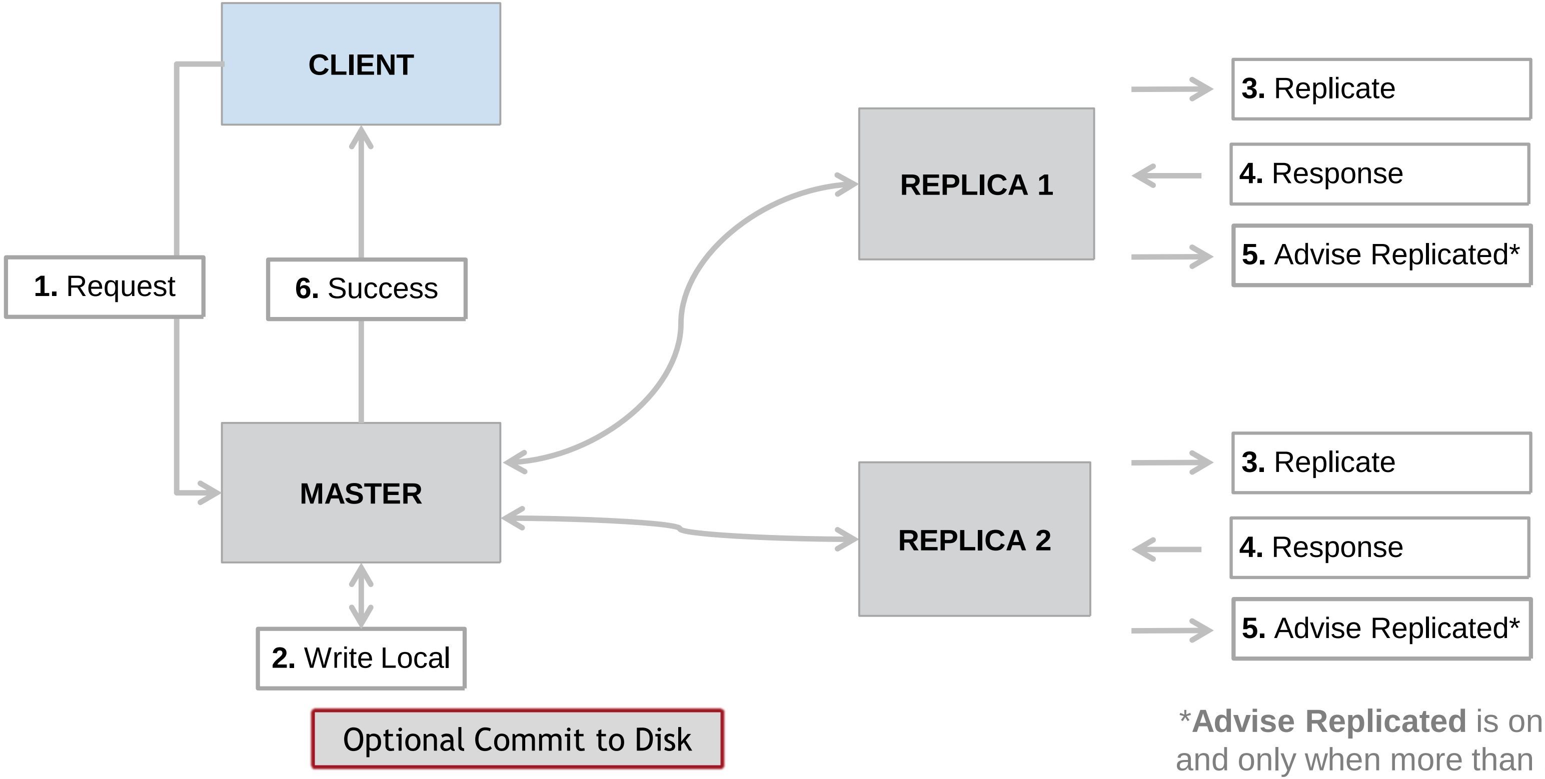
One designated replica, in a majority subcluster, and the data

SPLIT – Unavailable

Majority has no designated replicas, minorities don't have all replicas

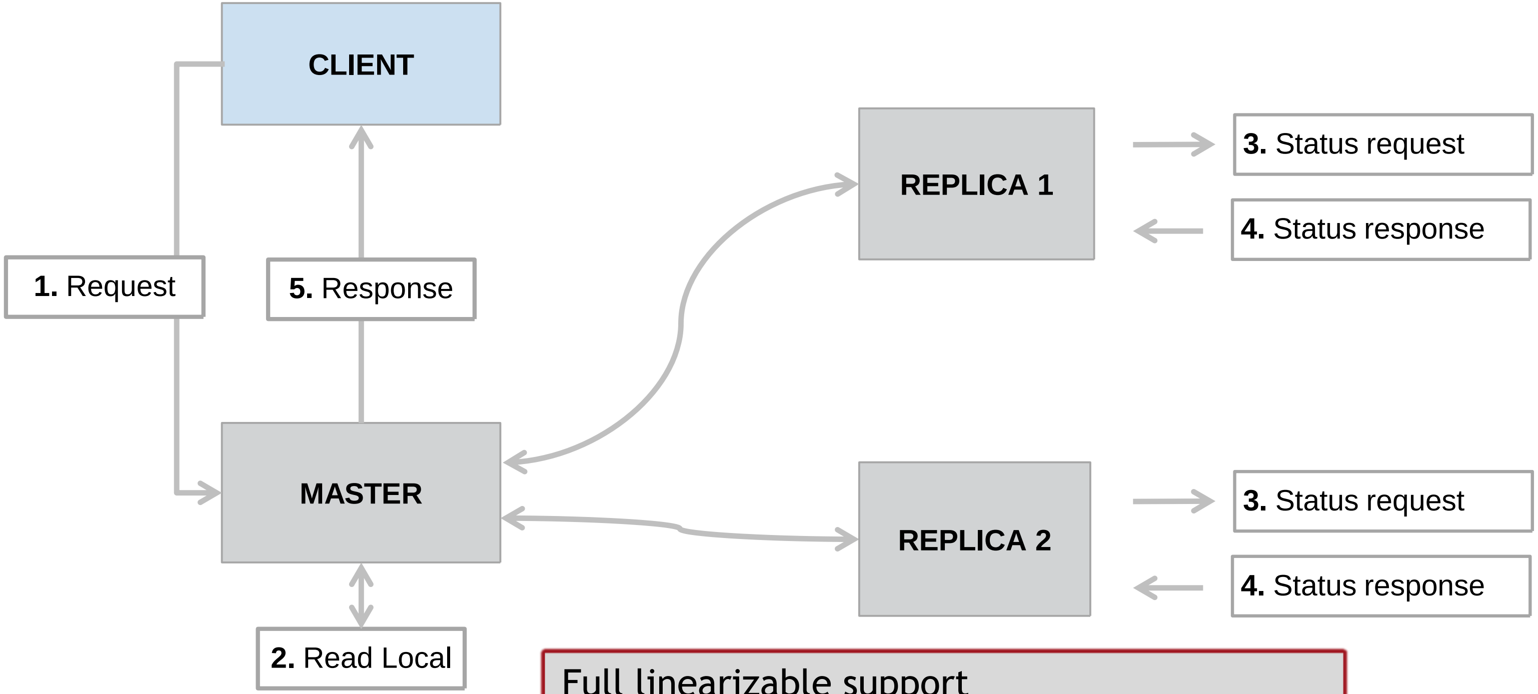
Example applies to an individual partition

Record Replication



***Advise Replicated** is one way, and only when more than 1 copy

Read Pattern - Linearize



Full linearizable support
No stale reads possible
Extra network round-trip

Aerospike NoSQL Database

Aerospike Features

Hybrid Memory Architecture

Indexes in DRAM and data in Flash provide storage guarantees, and uniquely unlock Flash's performance.

Data structure & Document indexing

Easy-to-program data structures such as lists, maps, which improve performance and allow indexed lookups on field values.

Non Volatile Memory (NVM)

Clustered with Native Flash optimizations allow Flash-economics data with DRAM performance. Proven millions of TPS per server, 50T++ per 2U server, commodity hardware.

Enterprise Security

Full transport encryption. Support for encrypted storage and in-database transparent data encryption. Authentication, access control, exception logging.

Primary Key Consistency

4.0 provides Strong Consistency on primary key access, with Jepsen test results. Commit to device support for demanding correctness environments.

High Availability (AP mode)

Used traditionally for high performance machine-to-machine analytic systems, such as fraud detection, pattern matching, and real-time pricing.

Real-time Analytic Framework

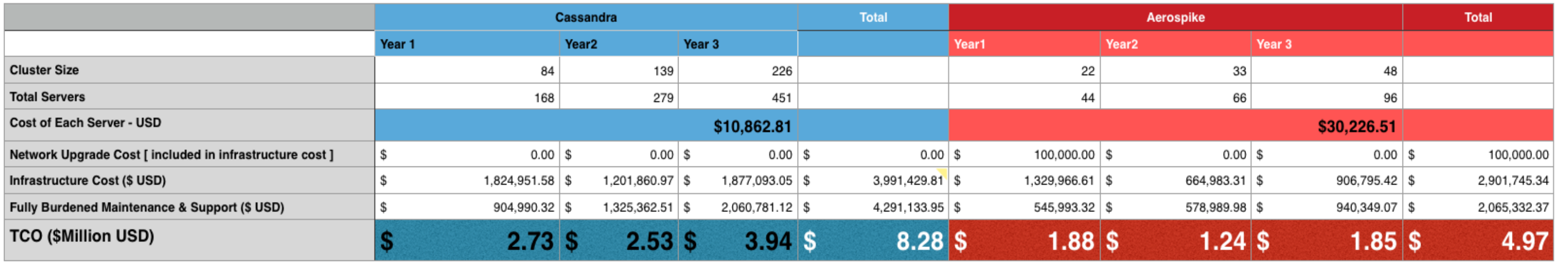
Deep integration with Spark to allow SQL and ML based tools to operate seamlessly and efficiently

Geographic Replication (XDR)

Multiple datacenters stay in sync through high performance replication.

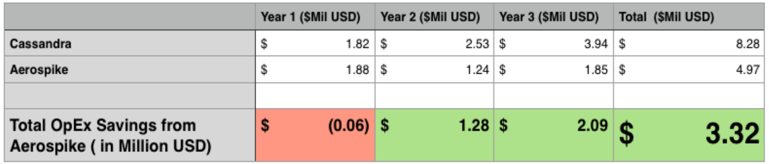
Faster Means Fewer Servers, More Opportunity

TCO - Summary



	Cassandra			Total	Aerospike			Total
	Year 1	Year2	Year 3		Year1	Year2	Year 3	
Cluster Size	84	139	226		22	33	48	
Total Servers	168	279	451		44	66	96	
Cost of Each Server - USD	$10,862.81				$30,226.51			
Network Upgrade Cost [included in infrastructure cost]	$ 0.00	$ 0.00	$ 0.00	$ 0.00	$ 100,000.00	$ 0.00	$ 0.00	$ 100,000.00
Infrastructure Cost ($ USD)	$ 1,824,951.58	$ 1,201,860.97	$ 1,877,093.05	$ 3,991,429.81	$ 1,329,966.61	$ 664,983.31	$ 906,795.42	$ 2,901,745.34
Fully Burdened Maintenance & Support ($ USD)	$ 904,990.32	$ 1,325,362.51	$ 2,060,781.12	$ 4,291,133.95	$ 545,993.32	$ 578,989.98	$ 940,349.07	$ 2,065,332.37
TCO ($Million USD)	$ 2.73	$ 2.53	$ 3.94	$ 8.28	$ 1.88	$ 1.24	$ 1.85	$ 4.97

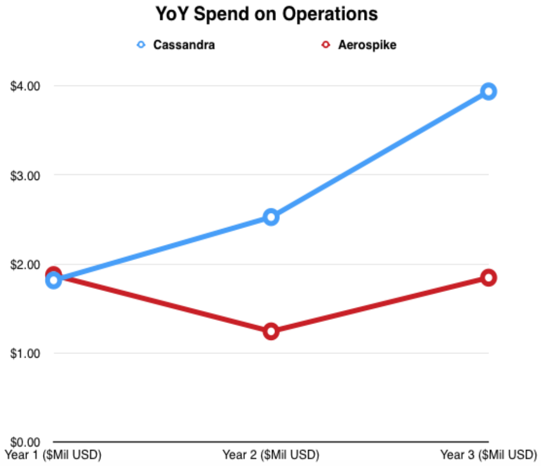
Aerospike OpEx Savings Calculator



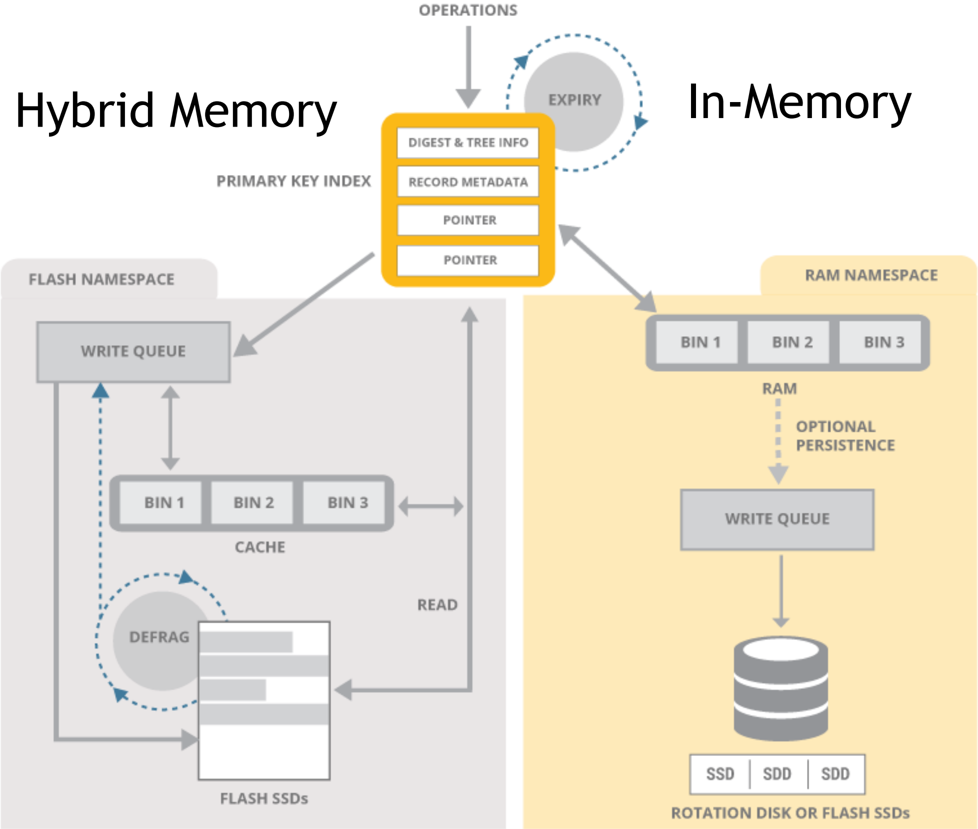
	Year 1 ($Mil USD)	Year 2 ($Mil USD)	Year 3 ($Mil USD)	Total ($Mil USD)
Cassandra	$ 1.82	$ 2.53	$ 3.94	$ 8.28
Aerospike	$ 1.88	$ 1.24	$ 1.85	$ 4.97
Total OpEx Savings from Aerospike (in Million USD)	$ (0.06)	$ 1.28	$ 2.09	$ 3.32

Note:

1. We assume 50% of 168 servers in operations are “Sunk Cost” and not part of TCO calculation
2. Total Cassandra infrastructure cost of $3.99M reflects reduction in #1 above
3. TCO does not include cage rent, power, cooling costs - which will further improve Aerospike OpEx savings
4. Network upgrade cost for Aerospike is included in cost of Aerospike
5. Calculation done with following storage assumptions
   a. Year 1 - 85B Keys, Year 2 - 130B keys, Year 3 - 195B keys



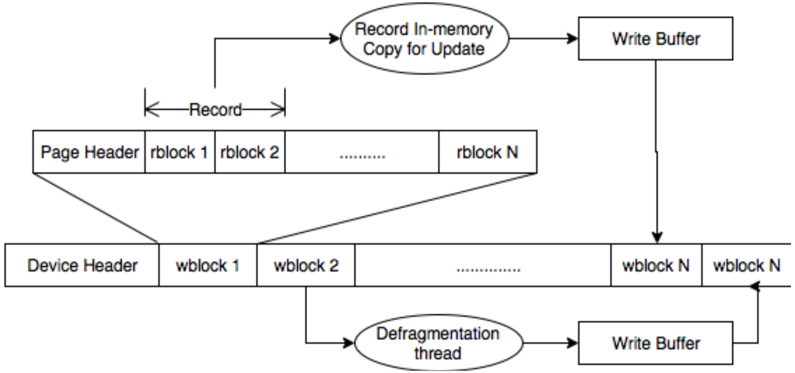
Aerospike Hybrid Memory



Highlights

1. Primary Key O(1)
2. Indexes in DRAM
3. Direct Device access
4. Large Block Writes
5. Fast restart with shared memory

Storage Layout



Aerospike Deployment - Wide range of Choices

Deployment Approach

When to Consider?

Bare Metal Commodity Infrastructure

70% of Aerospike deployments

- Extremely low-latency requirements outweighs the cost of infrastructure
- Very high-performance required, cannot even afford latencies due to virtualized infra
- High data sensitivity
- Examples : Fraud Prevention, High-Freq Trading, Real-Time charging, Payment Processing etc.,

AWS Google Cloud Platform Azure

25% of Aerospike deployments

- Ready to trade-off latencies induced to cloud infra
- Ease of operation more critical than performance
- Able to live to noisy networks occasionally
- Examples : Session store, cache, profile store, e-commerce catalog store etc.,

Pivotal Cloud Foundry

http://network.pivotal.io

Newly Introducing for Enterprises

Only Pivotal Partner who has all 3 different type of TILEs - ServiceBroker, Managed Service, OnDemand Service

- Ready to trade-off latencies induced to cloud infra
- Ease of operation more critical than performance
- Data sensitivity critical
- Ability for leveraging cloud yet control infra is critical
- Examples : Session store, cache, profile store etc.,

What’s it good for?

Case Studies: HMA - Lower TCO & better SLA

Customer	Situation	Problem	Hybrid Memory System
Trading Account Account Status, Trades, Risk	DB2+Gemfire cache	150 Servers growing to 1000	Single cluster – 12 servers
Fraud Detection	2 ORCL RAC clusters + Terracotta cache	System Stability & missing SLA's	3 Clusters – 20 Servers each
User Integrity Checking for Internet Transactions	DataStax/Cassandra	168 DataStax Servers growing to 450+	30 Servers – 2 clusters
Customer 360 and Rich Consumer Application	Green Field / Oracle / X.500	Largest Telco needs "MyService" application, integrated customer DB	15 Servers – 2 clusters
Telco Device and User Access	ORCL Coherence / DataStax Cassandra	Existing SOE solutions unstable & Costly	5 successful POC's
Telco Revenue Assurance	DataStax/Cassandra PostgreSQL + cache	Hundreds of cache & Cassandra Servers Scalability challenges	Significant reduction of server footprint – global deployment

Retail Banking Positions – Trading and Risk

Business Challenge

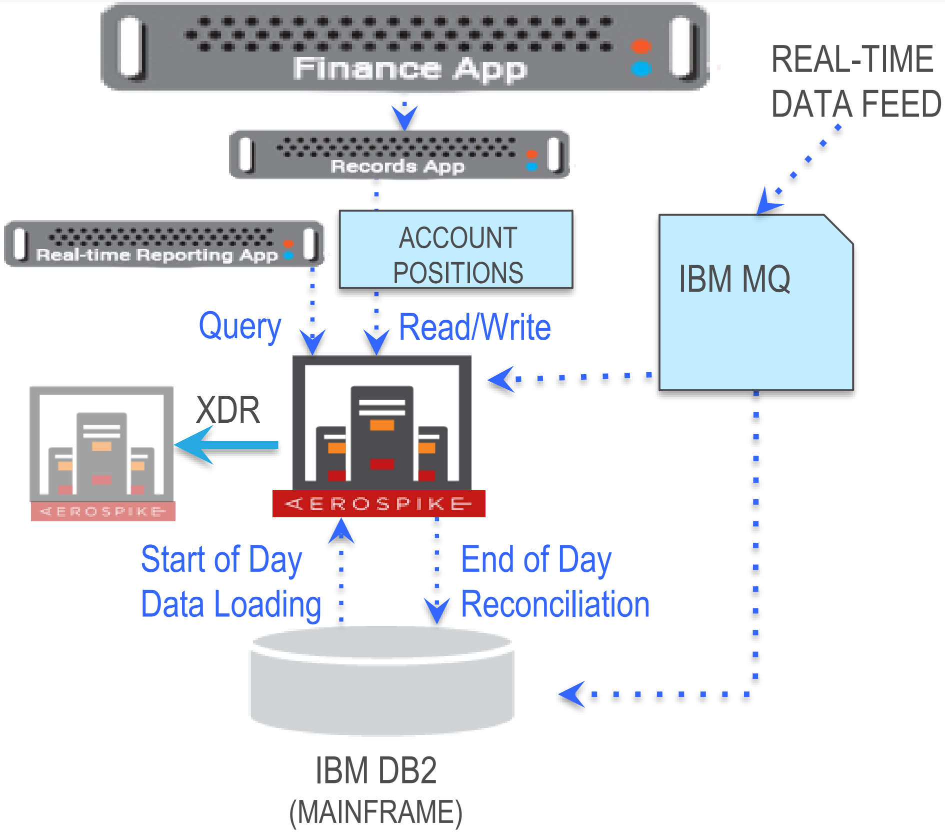
- Must update stock prices, show balances on 300 positions, process 250M transactions, 2 M updates/day
- High access from mobile was killing the DB2 under normal transaction load
- Calculate risk metrics on portfolios on a continuous basis

Caching solution failed

- Running out of memory, data inconsistencies, restarts at 1 hr
- 3 → 13 TB, 100 → 400 Million objects, 200k → I Million TPS

Hybrid Memory Advantage

- Built for persistent Flash – eliminated inconsistencies
- Predictable Low latency at High Throughput – handled mobile access easily for enhanced transaction load
- 10-12 Server Cluster – reduced from 150 in-memory cache servers
- Growth from 150 to 1000 cache servers triggered db change



Telco – Customer 360

Business Challenge

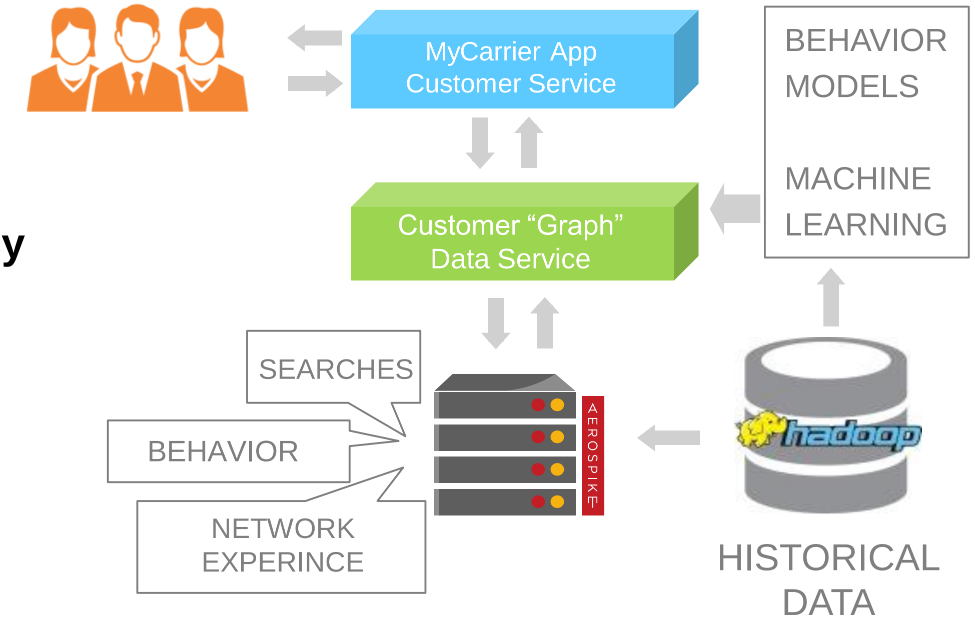
- 1 Billion potential customers
- Data Sources regarding past customer history, behavior, satisfaction, offer responses, advertising
- Integration with flow & network monitoring
- Existing solutions (X.500) were failing at scale

High Availability, Reliability, Low latency

- > TBs of data
- 1B objects
- 10-200K TPS

Selected Aerospike

- Rich application programming model
- Scale-up and Scale-out
- Strong Consistency
- Support of Cache + Operational uses



Fraud Prevention for Interactive Payments

Business Challenge

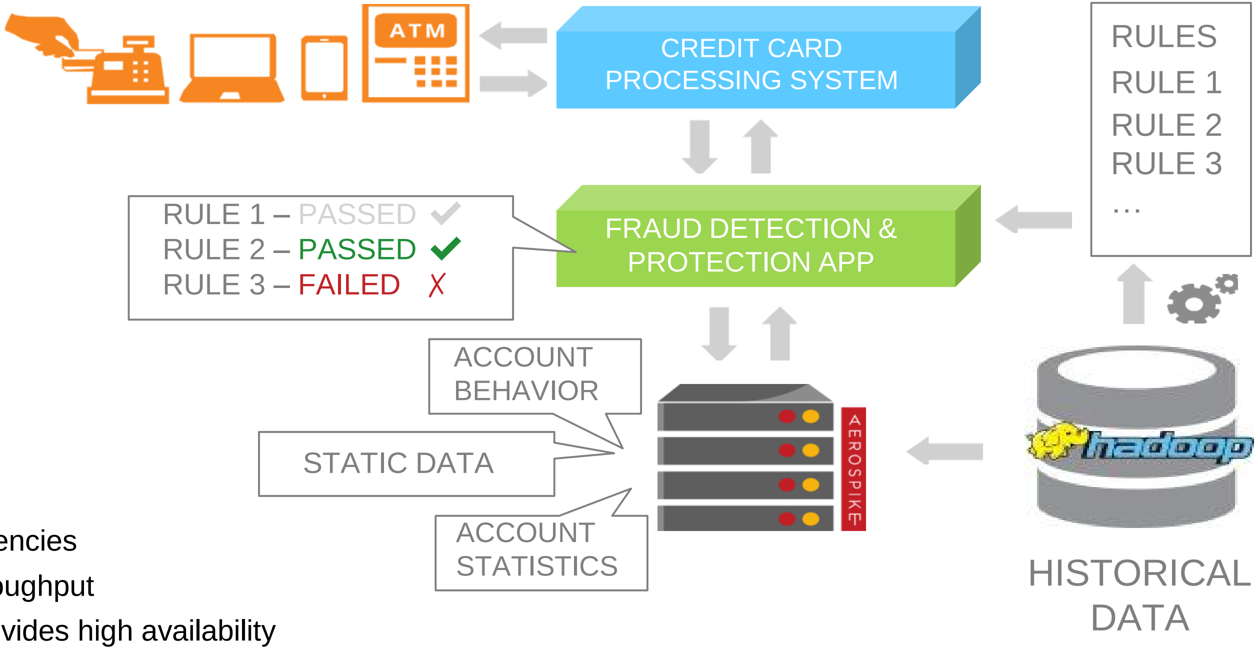
- Every payment transaction requires hundreds of DB reads/writes
- Missed latency SLA lost business
- Caching solution too expensive

Need to scale up

- 10 → 100 TB
- 10B → 100 B objects
- 200k → I Million+ TPS

Selected Aerospike

- Built for Flash – eliminated inconsistencies
- Predictable Low latency at High Throughput
- Cross data center (XDR) support provides high availability
- 20 Server Cluster reduced from 150 in-memory cache servers
- Used latest technology to reduce cost – Dell 730xd w/ 4NVMe SSDs



NAND Flash Performance Matters

Come to

Riverside Beer Chat
(Battersea Barge)

6pm tonight

Thanks!

brian@aerospike.com
@bbulkow

Come to our event!