Flash-Optimized, High-Performance NoSQL Database for All

Khosrow Afrozeh
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)
...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

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

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

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

K. Kruglov - CTO

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

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”

... I don’t have that much data!

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

- 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

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

**AGE OF CUSTOMER = READ/WRITE PATTERN**
BIG DATA: EMERGING ARCHITECTURE
Add-a-Layer Architecture

Request routing and sharding

LOAD BALANCER
APP SERVERS
CACHE
DATABASE
STORAGE

CONTENT DELIVERY NETWORK
Minimalism Makes a Comeback

APP SERVERS

CACHE

DATABASE

STORAGE
Minimalism Makes a Comeback

APP SERVERS
CACHE
DATABASE
STORAGE

APP SERVERS
Hybrid NoSQL
BIG DATA: EDGE DATABASES
Financial Services – Intraday Positions

LEGACY DATABASE (MAINFRAME)

Start of Day
Data Loading

End of Day
Reconciliation

REAL-TIME DATA FEED

ACCOUNT POSITIONS

Read/Write

Query

XDR

RT Reporting App

Records App

Finance App

ACCOUNT POSITIONS

10M+ user records
Primary key access
1M+ TPS planned

Finance App

Records App

RT Reporting App

ACCOUNT POSITIONS

REAL-TIME DATA FEED

ACCOUNT POSITIONS

Start of Day
Data Loading

End of Day
Reconciliation

LEGACY DATABASE (MAINFRAME)

Finance App

Records App

RT Reporting App

ACCOUNT POSITIONS

REAL-TIME DATA FEED

ACCOUNT POSITIONS

Start of Day
Data Loading

End of Day
Reconciliation

LEGACY DATABASE (MAINFRAME)

© 2014 Aerospike. All rights reserved.
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

Java application tier

Data abstraction and partitioning

Recent user generated content

Content and Historical data

MODIFIED REDIS (SSD ENABLED)

MYSQL or POSTGRES (ROTATIONAL DISK)
Real-time bidding

Real-time Bidding Online Advertising Ecosystem

Ad Verification

Media Buying Desks (MBDs)*

Demand-side Platforms (DSPs)

Supply-side Platforms (SSPs)

Ad Networks

Ad Exchanges (Open & Private)

© 2012 Parks Associates

* Also referred to as agency trading desks

© 2014 Aerospike. All rights reserved.
Real-time bidding
Modern Scale-Out Architecture

LOAD BALANCER
Simple stateless

APP SERVERS
Fast stateless

HYBRID NoSQL

CONTENT DELIVERY NETWORK

RESEARCH WAREHOUSE
Long term cold storage
Modern Scale-Out Architecture

- CONTENT DELIVERY NETWORK
- LOAD BALANCER
  - Simple stateless
- HYBRID NoSQL
- Fast stateless
- RESEARCH WAREHOUSE
  - Long term cold storage
- APP SERVERS
- HDFS BASED
  - node
  - NGINX
  - jetty
  - Aerospike
  - MongoDB
  - Cassandra
  - Hadoop
  - Apache HBase
The Power of Flash Storage
Flash Storage Proven and Growing

Worldwide Solid State Drive Revenue Forecast (Billions of US Dollars)

Source: IHS iSuppli Research, January 2013
Facebook and Apple bought \textit{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
  - 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 and I’ll tell you the answer.

Ask me. I’ll look up the answer and then tell it to you.
## Flash Big Data Economics

10x FASTER
10x FEWER
SERVERS REQUIRED

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

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

© 2014 Aerospike. All rights reserved.
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

© 2014 Aerospike. All rights reserved.
ROBUST DHT TO ELIMINATE HOT SPOTS
How Data Is Distributed (Replication Factor 2)

- 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.

<table>
<thead>
<tr>
<th>Partition ID</th>
<th>Master node</th>
<th>Replica node</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1820</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1821</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4096</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
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

![TPS/Instance Chart](chart_image.png)
Amazon EC2 results

Number of Nodes vs TPS

- r3.large
- r3.2xlarge
- m3.xlarge
- m1.large

© 2014 Aerospike. All rights reserved.
Amazon EC2 results
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)
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
Large and capable, but not fast.

http://www.aerospike.com/community/labs/
Complex Event Processing (CEP)
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 Scenario – 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.

UN-PREDICTABLE LATENCY

- Cassandra: 70 – 760 ms
- MongoDB: 128 – 300 ms
- Aerospike: 7 – 10 ms

QPS
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 Free
Questions and Answers

@parshua
khosrow@aerospike.com