

Strong Consistency with High Performance in a Primary Key database

Brian Bulkowski, CTO and Founder

HPTS GONG SHOW, October, 2017

This discussion of new product features is *not* intended to describe a system which Aerospike will make available.

It represents current research.

■

Practical Multi-Million TPS temporal database

With Strong Consistency

Based on existing real, deployed technology

Give up *Availability*, not Performance (P in CAP is *not* performance)

Who am I ?

I'm Brian Bulkowski, and I'm addicted to building infrastructure

I founded Aerospike in 2008 to bring modern systems approaches to databases

We have a scheme for high performance transactions

Which requires a little background....



History of Aerospike

- **Solve “internet scale” data**
 - Multi-core, multi-server, clustering & HA
 - Allows competition with Google, Amazon, etc
 - Use *Flash*, software only, “cloud-enabled”
- **2008 ~ 2009 - Prototype**
 - Proved 100K TPS+ per server, distribution mechanism
 - DRAM only, no persistence (cache)
- **2010 ~ 2012 - Flash, Adtech, and Lies**
 - Claimed ACID prematurely
 - Funded by boutique VC firms (Alsop/Louie, Tim Draper)
 - Key-value stores working at 100k+++ read / writes over Flash
 - Avg 0.3 ms, 95% < 1ms, 99% < 5ms
 - 80-ish paying customers



History of Aerospike

- **2013 ~ 2015 – Open source, queries, data structures**

- Funding by major VC firms (NEA)
- Added query, in-database compute, lists / maps
- telco (data-oriented routing)
- fraud detection in real time
(behavioral analytics involving money)
- One retail brokerage

- **2016 ~ 2017 – The road to strong constancy**

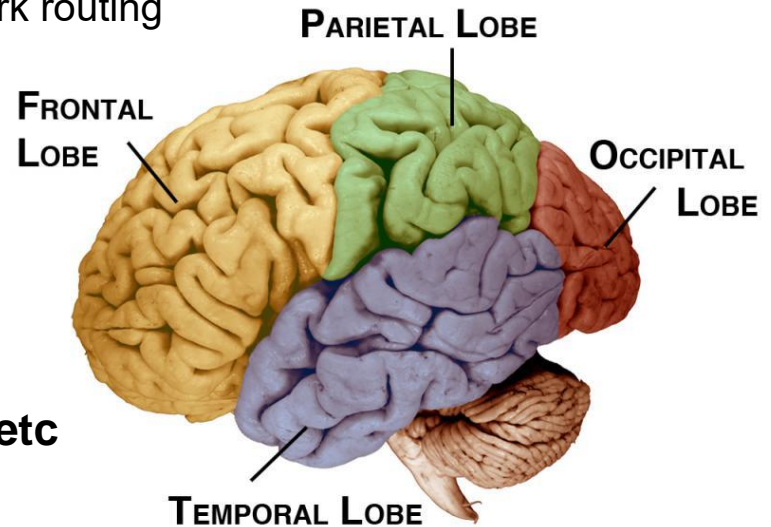
- Promised to stop lying (HPTS 2016, thanks Kyle)
- Tired of hearing “what about split brain”



AEROSPIKE

Why do people use Aerospike?

- **Temporal “Edge” uses**
 - Internally immutable, “update” to an application
 - “Analytics” like ad pricing, fraud detection, network routing
 - Behavioral suggest Availability over Consistent
- **Extreme Insert & Ingest**
 - 50% write rate common
- **Lower latency than DynamoDB, BigTable, etc**
- **Works great with Flash / NAND**
 - Beyond “In Memory Databases”



From AdTech Outward

ADTECH



ECOMMERCE



GAMING / BETTING



FINANCIAL



TELECOM



TECHNOLOGY



Extreme Data Integrity Required

- **“Enterprises” demand strong consistency**
 - “System of Record” needs higher performance
- **“System of Record” with cache is broken**
 - “System of Engagement” is a fast SoR
- **“Data services” in enterprise need speed**
 - Architects understand they are fighting Amazon
 - “Stateless” apps need *more* database
- **Aerospike provides speed**



Thus, Strong Consistency

- **Performance**
 - 1M++ TPS *per server* over Flash ; 10M++ DRAM / Batch
- **Mixed read / write workload**
 - Extreme insert rates
- **“Fairly” high availability**
 - Allow “n-1” replica failure with “no” availability hit
- **“Local”**
 - < 10ms network latency
- **Durability**
 - Disk-buffered currently, adding based on customer demand (Flash)



Scheme – combine known techniques

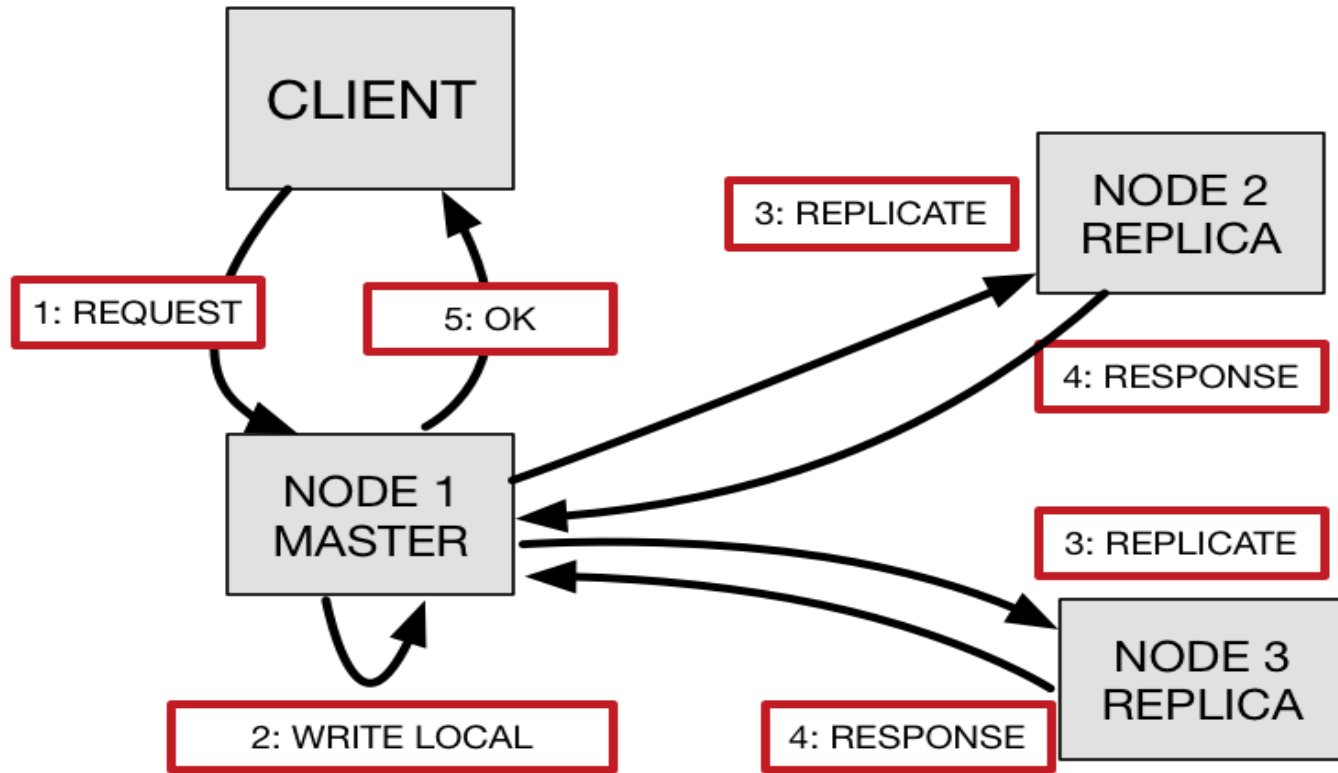
- **Single Master**
 - Requires high quality cluster management (*)
- **Writes: synchronous**
 - *Aerospike already does it*
- **Reads: Single-master (“session”) vs All (“linearize”)**
 - Linearize only needs to check other servers
 - Per-read transaction choice
 - Stale reads in case of client accessing server about to leave cluster
- **Lamport Clock**

Only One Master

- **Rule 1**
 - If a sub-cluster has all of the known masters and replicas, CONTINUE PROCESSING
- **Rule 2**
 - If a *majority sub-cluster* has data (master or replica), CONTINUE PROCESSING (promote, replicate)
- **Rule 3 (tie-breaker)**
 - If an equal sub-cluster, and had a master of the data, CONTINUE PROCESSING (replicate)
- **If none of the above, NO reads and writes (retain data for fast resync)**



Writes



Thank You
Questions?

Talk to me about
Storage Class Memory,
The Future of Infrastructure