AEROSPIKE

SUMMIT '19

AEROSPIKE

Systems of Record and Edge-based Systems

Tim Faulkes
Director of Solutions Architecture
Aerospike

What is a System of Record vs Edge System?

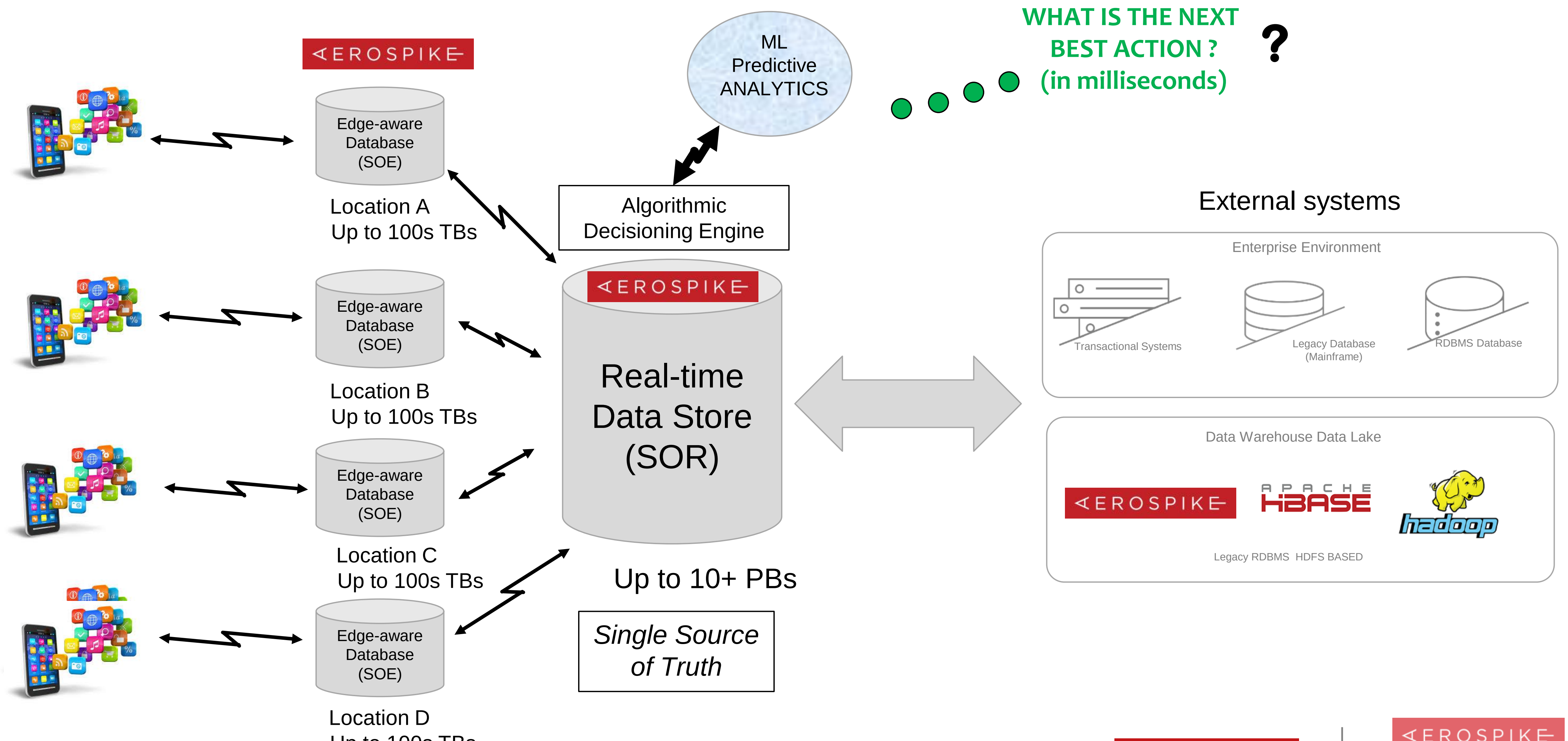
Edge based System

- User facing data, typically fast moving.
- Measured up to 100’s of TB
- Pulls source data from SOR and can push its data to the SOR
- Normally strongly consistent, can prefer availability in some use cases
- Objects tend to be smaller, targeted information for the use case

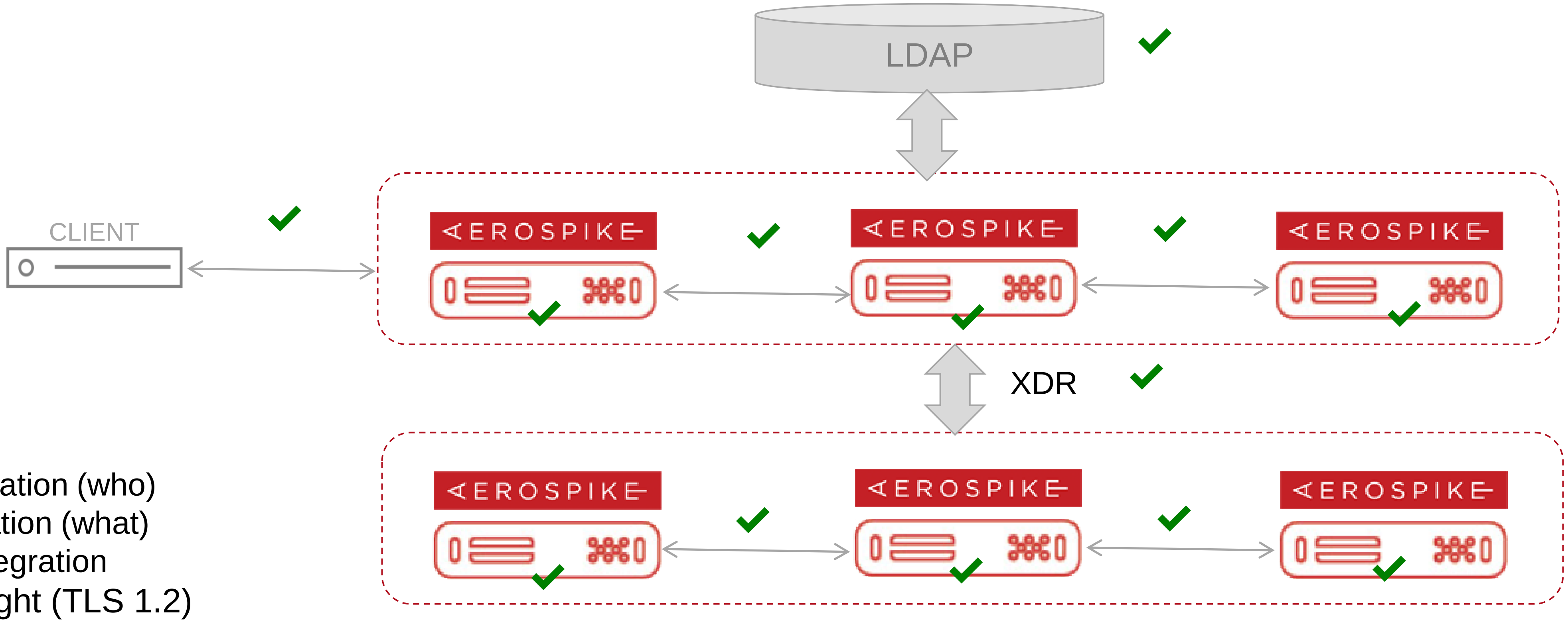
System of Record

- Slower moving data
- Measured in TB, PB or larger
- Must be able to ingest and egest data from multiple sources
- Must be strongly consistent
- Objects tend to be larger, a more compete picture

The Aerospike Difference – Performance and Scale



Full security is a must for a SOR

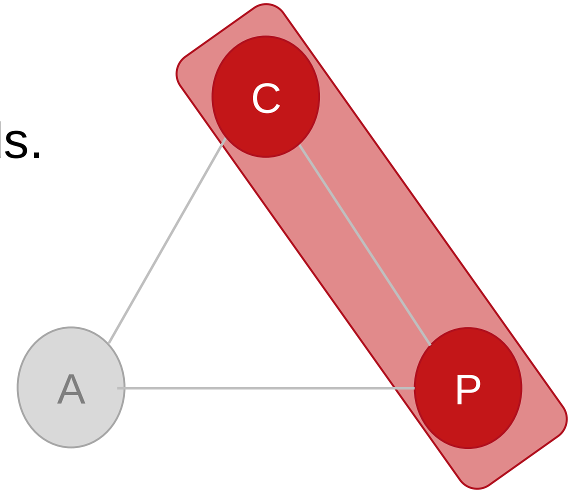


- Security
  - Authentication (who)
  - Authorization (what)
  - LDAP integration
- Encryption in flight (TLS 1.2)
  - Client to Cluster
  - Cluster to Cluster
  - Server to Server
- Encryption at Rest (AES128 / AES256)

Strong Consistency

Version 4 of Aerospike introduces Strong Consistency

- No acknowledged write can be lost
- Fully support for linearized, session consistency or relaxed consistency reads.
- Minimal changes to client code
- Uniquely: high performance
  - Almost zero performance impact under common configurations
- Configurable at the namespace level
  - SC and AP in the same cluster!



"Aerospike does appear to provide linearizability through network partitions and process crashes"

--- Kyle Kingsbury, Jepsen.io

Aerospike 4.0: High Performance with Strong Consistency

Aerospike internal benchmark of Strong Consistency versus Availability

	Linearizable Consistency	Sequential Consistency	Availability Mode
OPS	1.87 million	5.95 million	6 million
Read Latency	548 µs	225 µs	220 µs
Update Latency	630 µs	640 µs	640 µs

In-memory configuration with persistence enabled

5 node cluster
500M keys

Replication factor 2
Objects were 8 byte integers

Strong Consistency use cases

Social media:
- Maintain friends lists, posts from friends

Credit card processing:
- Monetary amount “at risk” when processing transactions for fraud.

Fraud detection:
- Heuristics, so ok with wrong data in error situations
- But need to be able to measure when the data is wrong -> Need strong consistency.

Inter-person money transfers
- Any loss of data or inaccuracies unacceptable

Trading Systems
- Intra-day system of record, mainframe offloading

Loyalty programs
- Keep track of customer’s reward points, points usage

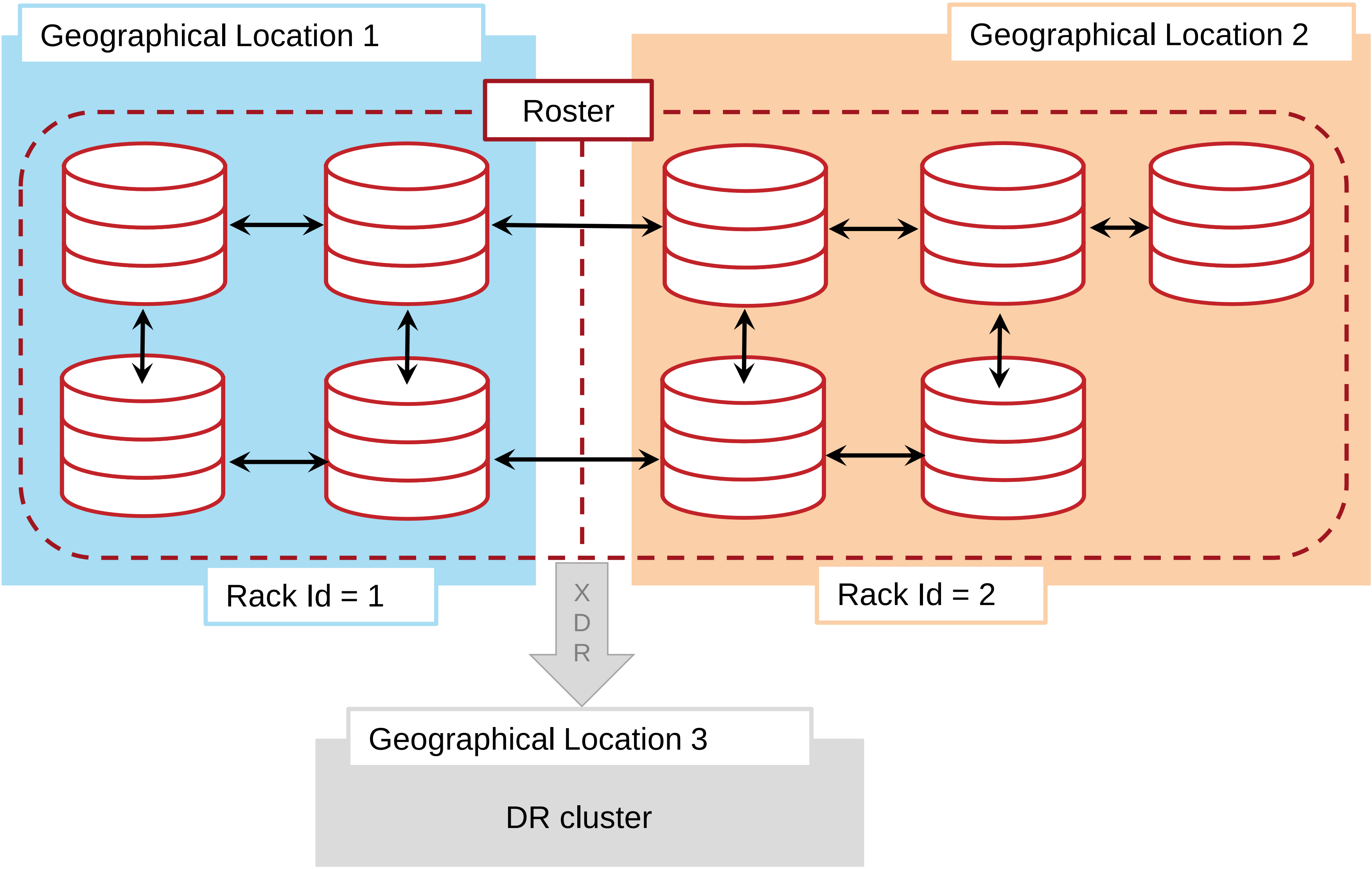
Shopping carts

Strong Consistency and Rack Awareness

 | Proprietary & Confidential |

Geographical redundancy with Strong Consistency

- 2 Physical DCs with 1 “stretch” cluster
- Rack awareness implemented so master and replicas are in different clusters
- Clusters should be geographically “close” to minimize latency costs
- Writes will always incur 1 or 2 inter-zone hops
- Separate, async cluster for pure DR
- Reads: Consistency selectable on a per-read basis: *linearizable, session consistency, sequence, prefer rack.*
- Non-linearizable reads (minorly!) trade the chance of stale reads for increased availability / lower inter-AZ costs.

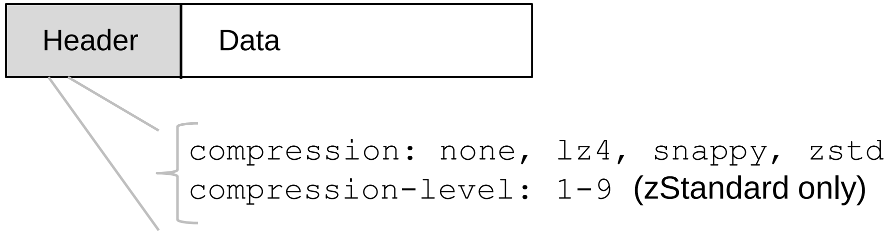


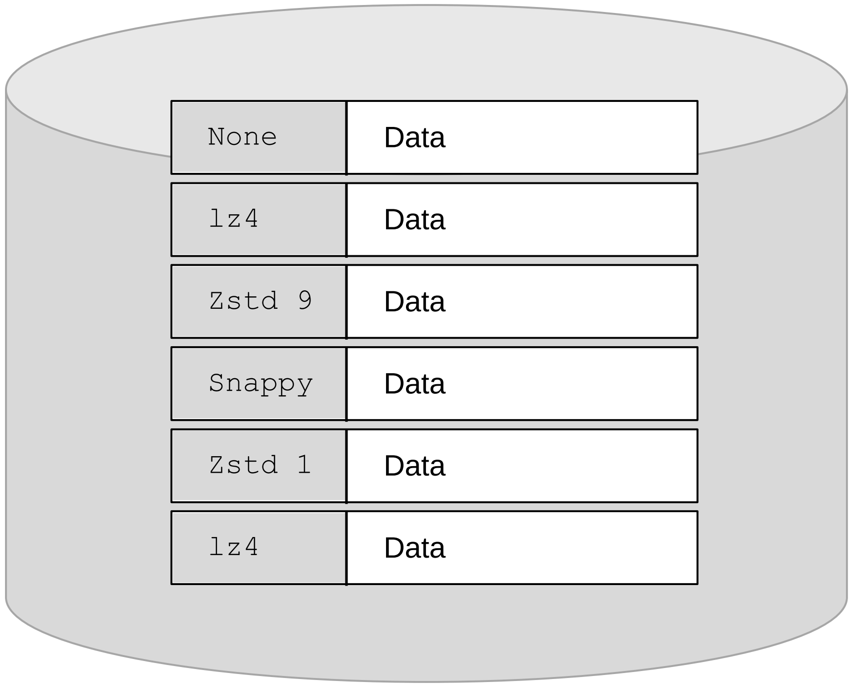
AEROSPIKE | AEROSPIKE SUMMIT ‘19

Compression at rest

Data Compression

- Available in EE v 4.5.0+, requires separate license key
- Compression exchanges CPU for storage space
- Transparent to application: secondary indexes, CDT operations, etc continue to work.
- If a compressed record is larger than the uncompressed one, uncompressed one is written.
- Different records on the same drive can have different compression schemes, whatever is active when the record is written.
- Both `compression` and `compression-level` are dynamic and specified in the `storage-engine` section





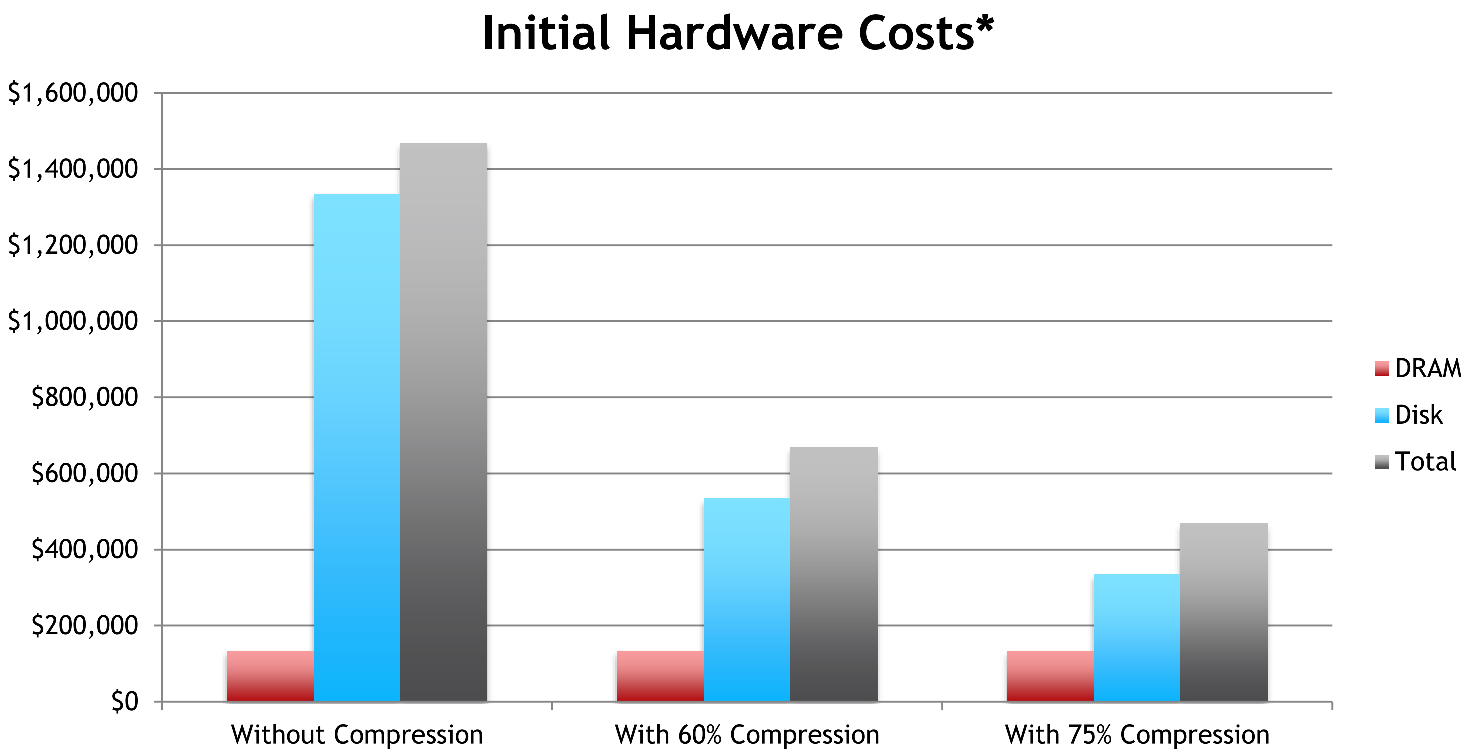
Data Compression Benefits

- Amount of DRAM stays the same for the indexes of data
- Storage for data reduces by the compression ratio.
- Example: 100B records, each 10k in size, 2 copies of the data (1PB business data)
- Assume DRAM is $9/GB, SSD $0.40/GB

Cost savings: Up to 68% of hardware costs

The compression ratio depends on the use-case and must be tested to determine the actual compression ratio

Note that CPU should be monitored too as some compression algorithms are CPU intensive.



* Note: Costs do not include servers, power, cooling, maintenance, operations personnel, etc

All Flash

ALL FLASH Configuration

Aerospike Server Version 4.3.0.2+ introduces ALL FLASH storage option.

- Allows user to store the PRIMARY INDEX (PI) on device (NVMe SSD).

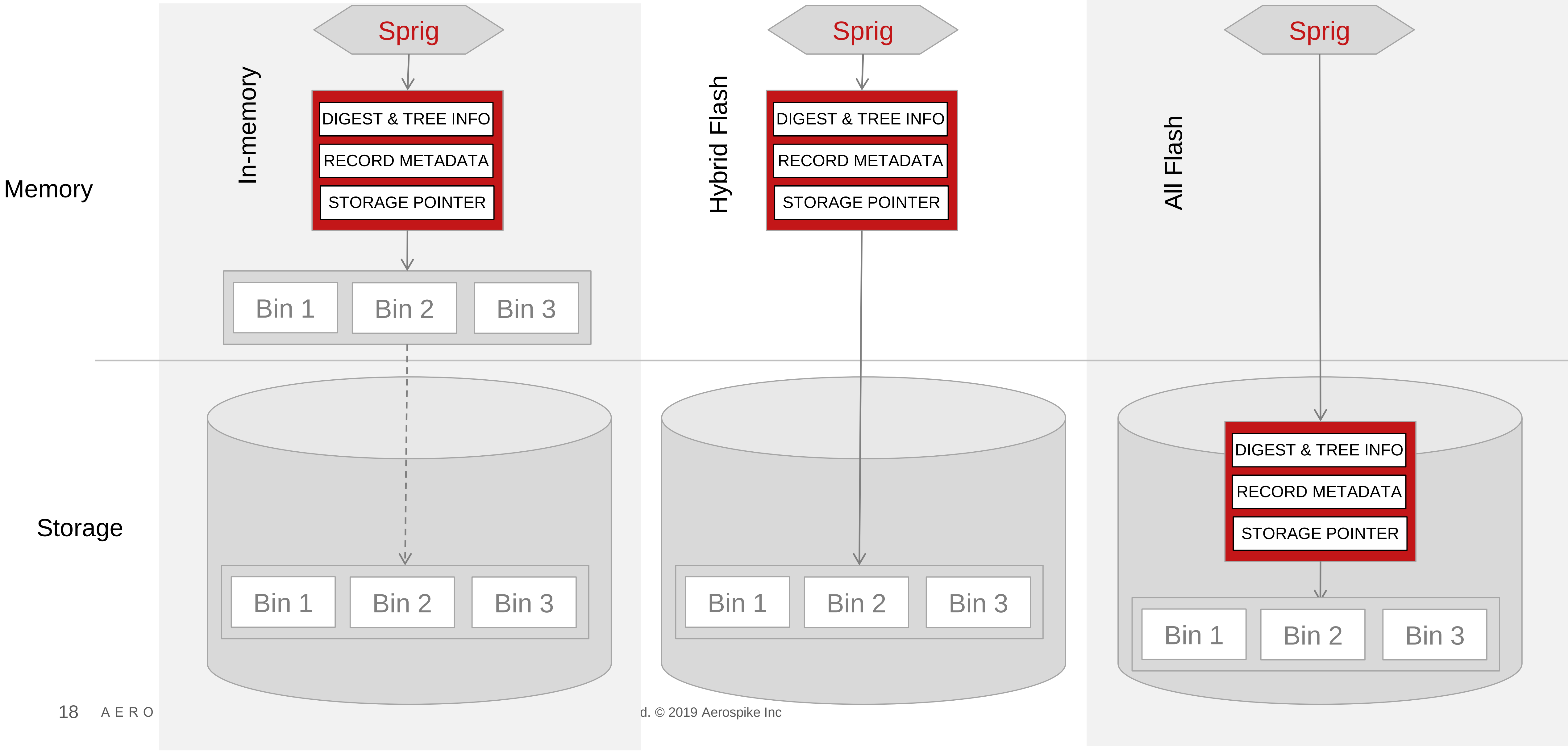
Edge Systems

- For large number of very small size records with relaxed latency needs.
- RAM vs SSD storage space ratio approaches 1:1 causing server sprawl.
- Significant cost savings by using ALL FLASH storage.

System of Record

- Cost savings with very large data stores. (> 100 TB)

All Flash



Memory
Storage
In-memory
Sprig
DIGEST & TREE INFO
RECORD METADATA
STORAGE POINTER
Bin 1
Bin 2
Bin 3
Bin 1
Bin 2
Bin 3
Hybrid Flash
Sprig
DIGEST & TREE INFO
RECORD METADATA
STORAGE POINTER
Bin 1
Bin 2
Bin 3
All Flash
Sprig
DIGEST & TREE INFO
RECORD METADATA
STORAGE POINTER
Bin 1
Bin 2
Bin 3

25 X

In-Memory

- Index and Data in DRAM
- Latency < 0.2 ms
- Hot Data – High TCO

< 1 TB

5 X

Hybrid Flash

- Index in DRAM, Data in Flash
- Latency < 1 ms
- Hot Data – Medium TCO

< 250 TB

1 X

All Flash

- Index and Data in Flash
- Latency < 5 ms
- Cold data – Low TCO

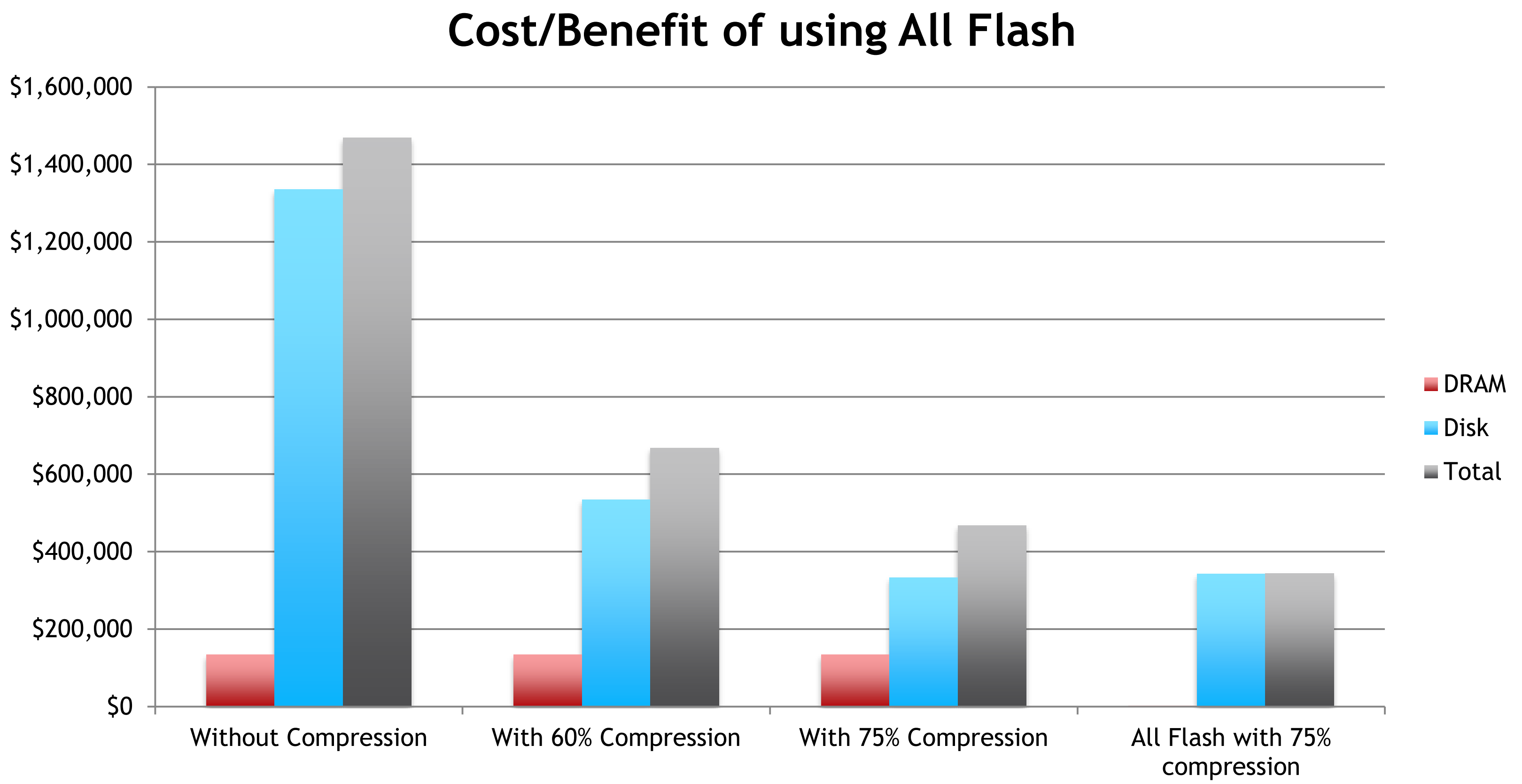
> 250 TB

SPEED & INFRA COST

SCALE

Cost benefit of using All Flash for 100B records

- **DRAM is around $9/GB, NVMe SSDs around $0.35/GB**
- **Moving 12.8TB DRAM to Disk saves significant money – at the cost of a few milliseconds of latency**



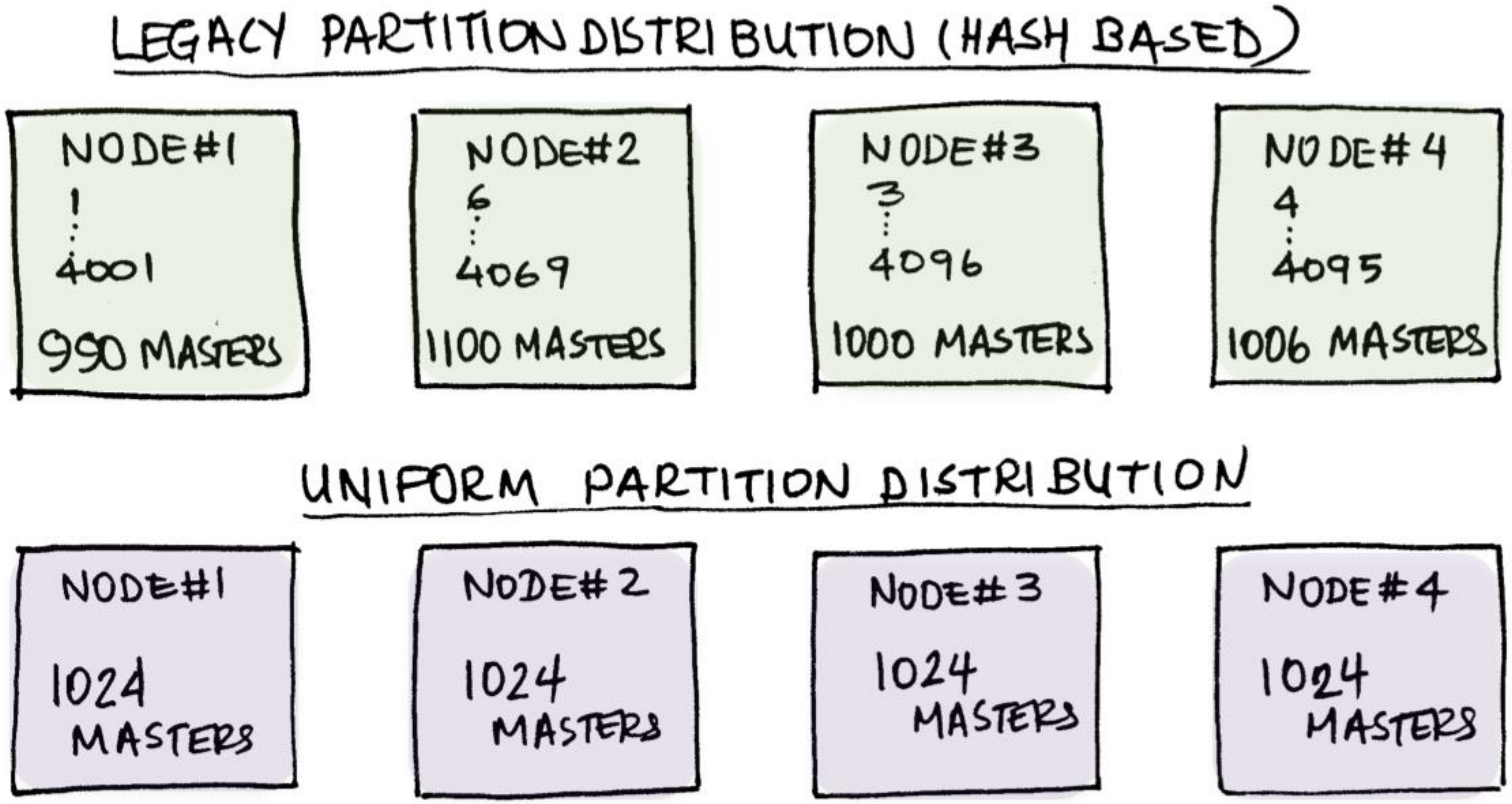
Total cost savings using compression and All Flash: 77%

* Note: Your use case may vary. Please see an Aerospike Solutions Architect to discuss your particular use case.

Uniform Balance

Uniform Distribution of Partitions across Cluster Nodes

Aerospike Server Version 4.3.0.10 introduced option to uniformly balance partition distribution across the nodes of a namespace.



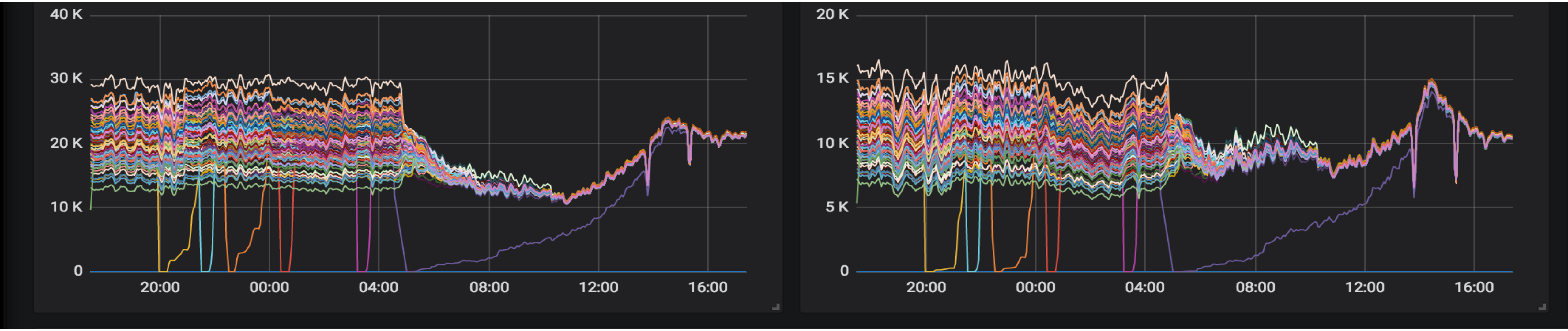
 | Proprietary & Confidential |

AEROSPIKE

AEROSPIKE SUMMIT '19

Prefer uniform balance results – real customer

Large cluster data for transactions per second (TPS).



Spread was 50% of Nominal TPS and data, tightened to < 5% after Uniform Balance.

- For large clusters, the removal of data skew saves costs
- Consider a cluster with optimal number of disks/node is 8. With 50% skew, some nodes will need 6 drives, others 10.
- As the usage is unpredictable and can shift with migrations, all nodes need 10 drives.
- With uniform balance, all nodes need 8 drives, saving 20% on drive costs.

Total Cost of Ownership

 | Proprietary & Confidential |

Cost of Ownership

- **Business problem:**
  - 1PB Unique data over 100B objects (average object size 10kB)
  - 2+ copies of the data
  - 500k writes/s, 20k reads/s

- **Cost comparison between**
  - In-memory solution
  - Cassandra + cache
  - Aerospike

Let's pretend they have strong consistency!

Cost of Ownership – In memory

- **Quorum based, so 3 copies of data: 3PB replicated data**
  - Assume 0% fragmentation (unrealistic!)

System 1: Commodity hardware
- 256GB DRAM, 200GB usable for data
- 3PB replicated data => 3,000,000GB
- Need 3,000,000 / 200 = 15,000 servers
- Cost per server:
- DRAM @ $9/GB: $1,800
- Server cost: $1,000
- Total cost: $2,800
- Total Cost: $42M

System 2: Heavy DRAM servers
- 1TB DRAM, ~0.95TB usable for data
- 3PB replicated data => 3,000,000GB
- Need 3,000,000 / 950 = 3,158 servers
- Cost per server
- DRAM: 18x64GB DIMMS, $500/ea
- = $9,000 for DRAM
- Server cost: $1,200
- Total cost: $10,200
- Total Cost: $32M

System 3: Heavy DRAM + PMEM servers
- 4TB DRAM, ~3.95TB usable for data
- 3PB replicated data => 3,000,000GB
- Need 3,000,000 / 3,960= 760 servers
- Cost per server
- PMEM not yet commercially available!

BEST CASE: $32M to purchase servers, but would need > 3,000 servers!
NOT PRACTICAL!

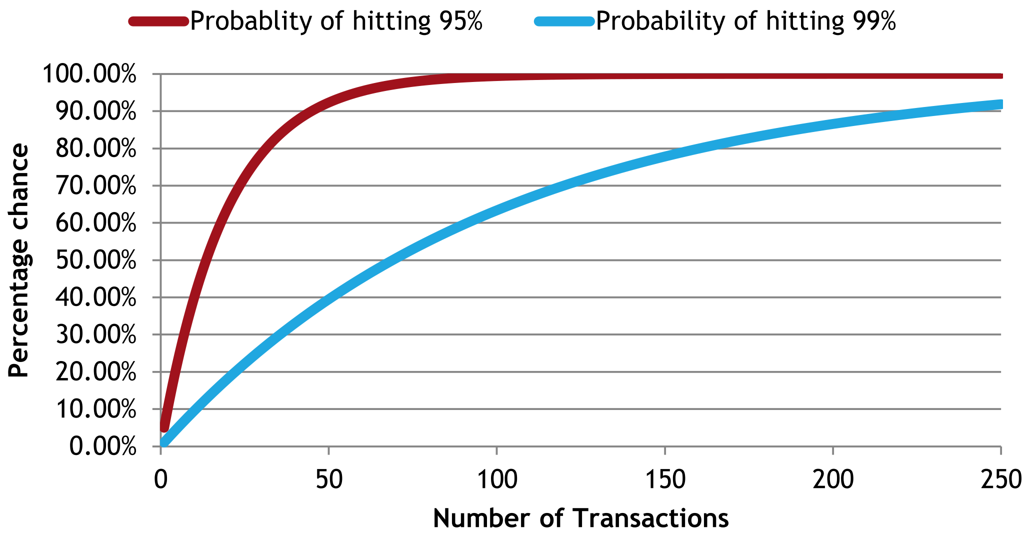
Cost of Ownership – Cassandra + Cache

- **Quorum based, so 3 copies of data: 3PB replicated data**
- **Recommended maximum data per node: 1 TB*.**
- **3PB => 3,000 nodes.**
- **SSTable fragmentation buffers of 50% => 6PB data**
- **Cost:**
  - 200TB DRAM at $9/GB: $1.8M
  - 3,000 servers at $1,200/server: $3.6M
  - 6PB rotational drive at $0.10/GB: $0.6M
  - Total cost of hardware: $6.0M

*** https://docs.datastax.com/en/dse-planning/doc/planning/planningHardware.html**

Cassandra + Cache: Latency

- **250 database lookups / transactions:**
  - Probability of hitting a 95%: 99.999%
  - Probability of hitting a 99%: 91.894%



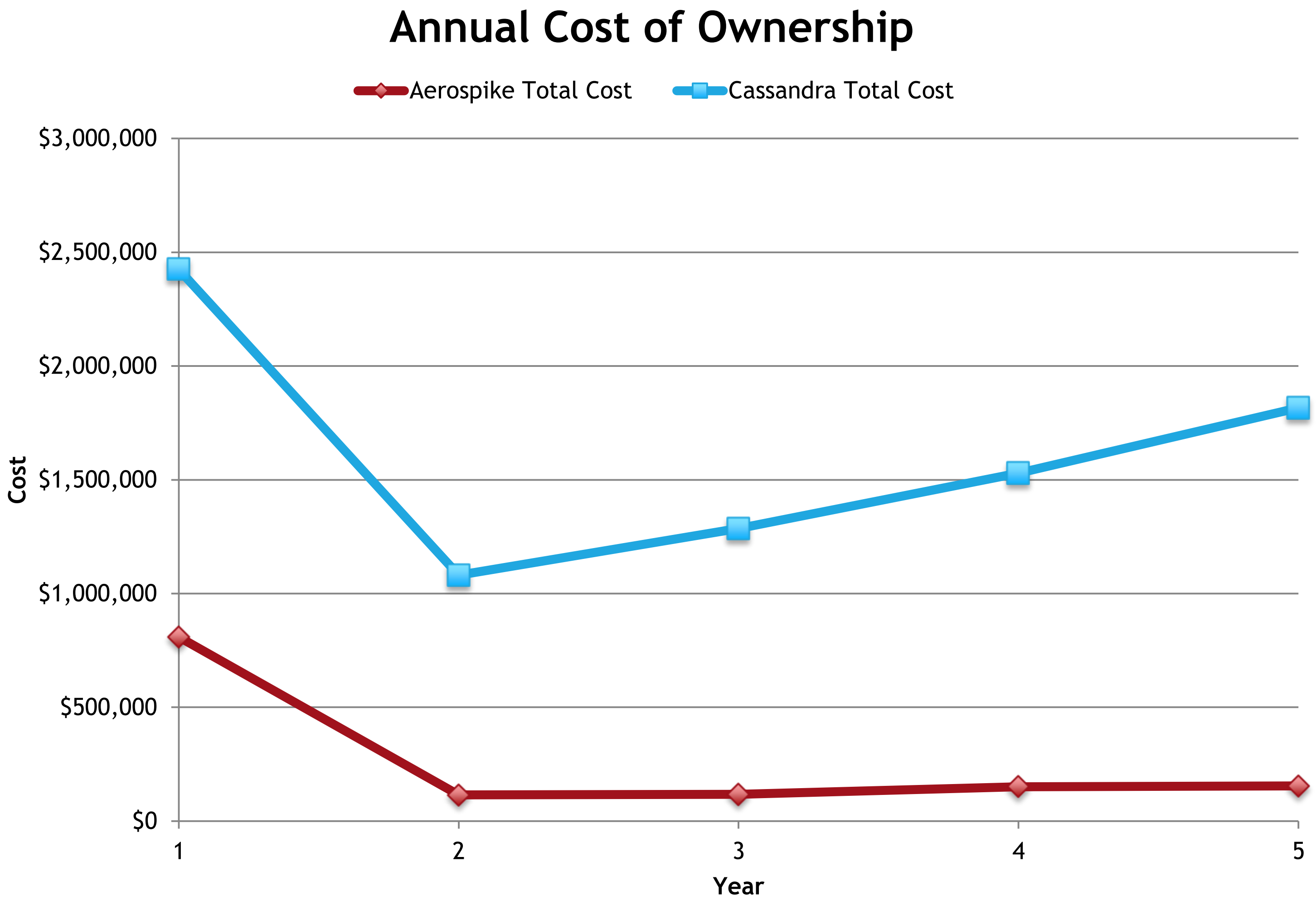
AEROSPIKE | AEROSPIKE SUMMIT '19

Cost of Ownership – Aerospike

- **Non-quorum based, so 2 copies of data: 2PB replicated data**
- **Allow for fragmentation of 50% => 4PB total data**
- **Memory: 100B objects x 2 x 64B => 12.8TB DRAM**
- **SSDs: ~$0.35/GB for 100x drives, 10 x 6.4TB/node**
- **Nodes: 4PB / (10x6.4TB) ~= 63 nodes**
- **Cost:**
  - 12.8TB DRAM at $9/GB: $115k
  - 63 servers at $1,200/server: $75k
  - SSD cost: 4PB @ $0.35/GB: $1,400k
  - Total cost of hardware: $1,590k

Cost of Ownership – Assumptions

- Cassandra compression: 75%
- Aerospike compression: 60%
- Growth: 10% / year
- Bare metal servers
- Cassandra servers: $2,000 each
- Aerospike servers: $25,000 each
- 1 system admin can manage 180 servers full time and cost $150k/years
- License costs not included
- Power, cooling, DC space costs not included



 | Proprietary & Confidential |

AEROSPIKE

AEROSPIKE SUMMIT '19

Other Considerations

 | Proprietary & Confidential |

Other considerations

- **64TB / node on Aerospike => Cold restarts will be *slow*.**
- **Consider using Intel Optane PMEM to store indexes persistently for several orders of magnitude performance improvement.**

- **The numbers presented here are representative only. Aerospike Solution Architects can work with you on your use case to get TCO numbers applicable to you.**

- **Some of the features mentioned in this presentation require separate license agreements. Please contact Aerospike for more information.**

Questions?

