Aerospike @ PayPal

Athreya Gopalakrishna
NoSQL Engineering Lead

Aerospike NoSQL DB

Key/Values

Hardware Evolution

NoSQL Database Evolution

Case Study – Fraud Detection System

NoSQL DB Storage Architectures

Designing a 50TB NoSQL DB

System Efficiency, Performance etc.

Q & A

Key Values

Key Value, Computer Languages

KV is a storage paradigm to store and retrieve data

key	value
k1	1
k2	Aa,b1,1c
k3	1,a,5/17/2018

KV Store in primitive forms

Map Container in C++ STL [`map <int, int> kvstore;`]

HashMap in Java [`HashMap<Integer, Integer> kvstore = new HashMap(..)`]

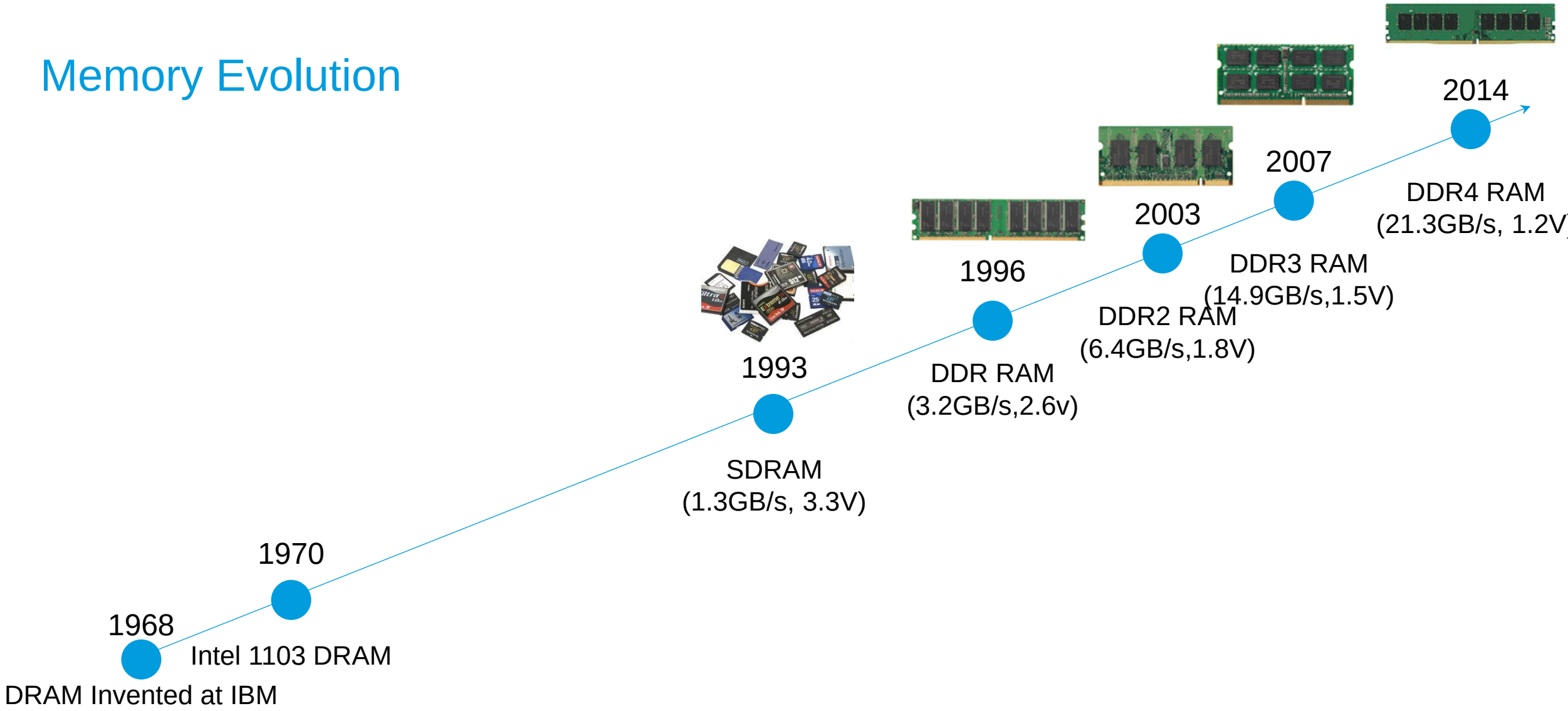
Dictionary in Python [`kvstore = {}`]

Simple interface

`put(key, value), put(key, value, TTL), get(key), delete(key)`

Hardware Evolution

Memory Evolution



Disk Evolution

1986
SCSI

First 2.5” HDD

1988
Texas Memory System
(16KB)

1989
SandDisk SSD
(20MB)

1991
M-Systems
SSD with high MTBF

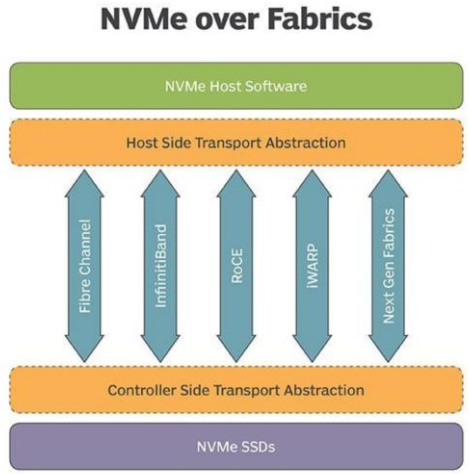
1995
SATA
6GBit/s,
32 queues

SAS
12GBit/s,
256 queues

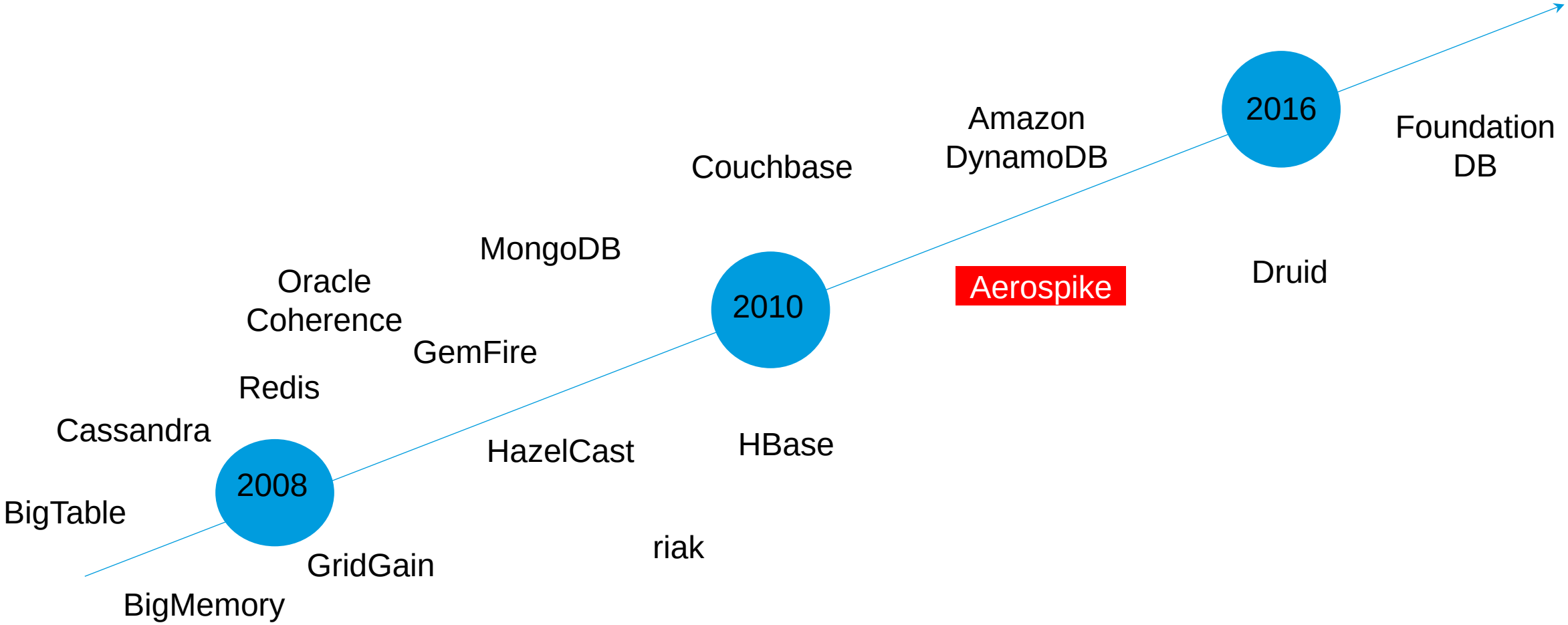
PCIe/NVMe
32GBit/s,
65K queues

NVMf

3D-NAND



NoSQL Database Evolution



NoSQL Database Evolution

Memory Evolution

Disk Evolution

Foundation DB

Druid

2016

Amazon DynamoDB

Couchbase

MongoDB

Oracle Coherence

2010

GemFire

Redis

Cassandra

2008

gTable

HazelCast

HBase

GridGain

BigMemory

riak

1993

SDRAM GB/s, 3.3V)

1995

1970

1968

Intel 1103 DRAM

DRAM Invented at IBM

1991

M-Systems SSD with high MTBF

1989

SandDisk SSD (20MB)

1988

Texas Memory System (16KB)

1986

Aerospike

2014

2007

2003

199

DDR4 RAM
21.3GB/s, 1.2V)

3D-NAND

PCIe/NVMe

NVMf

SAS

SATA

32GBit/s,
65K queues

12GBit/s,
256 queues

6GBit/s,
32 queues

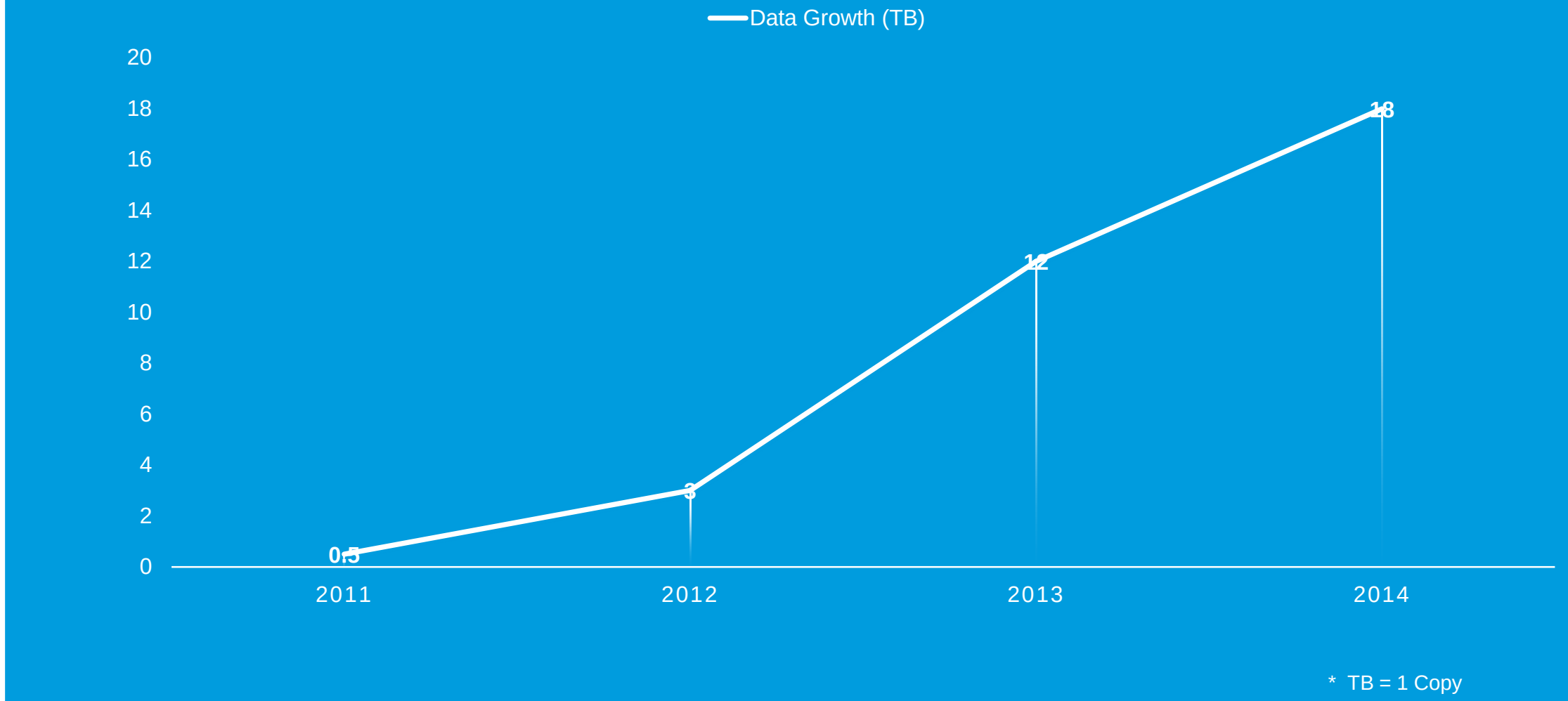
©2015 PayPal Inc. Confidential and proprietary.

Case Study

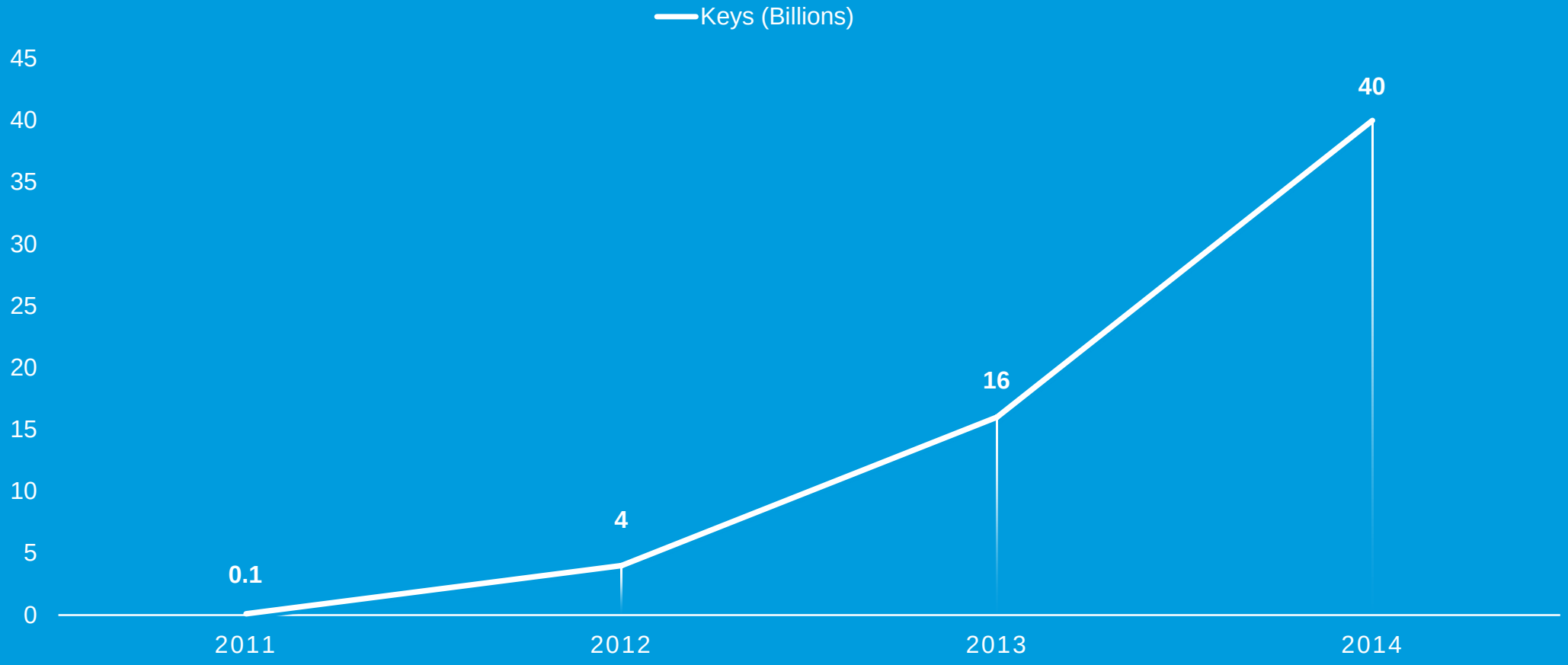
Fraud Detection System

- Analytical system
- Built on Relational and KV databases
- Requires Low Latency and High Throughput
- 1-200KB avg. object size
- **1ms@99** Percentile
- **4-5 Millions** of **transactions/second**
- **Trillions** of keys
- 100+ of Terabytes of Storage

KV STORAGE GROWTH TREND



KEY SPACE GROWTH TREND



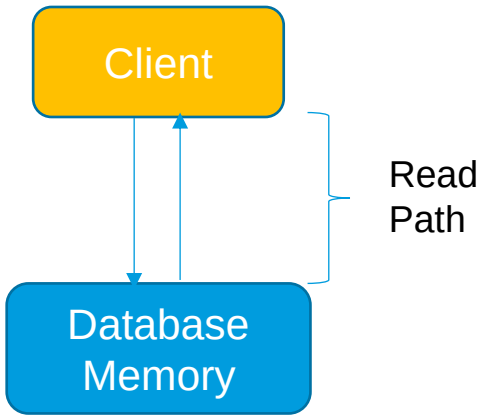
Storage Architecture

High Performance NoSQL

Cache	In-Memory	Memory-First	Hybrid-Memory
• Index, Data Stored and Served from Memory • No Data on Disk	• Index, Data Stored and Served from Memory • Data is persisted on Disk	• Index, Data Stored in Memory and Disk OR • Index in Memory and Data on Disk • Data is persisted on Disk	• Index in Memory and Data on Disk • Data is persisted on Disk

Read Path – Latency and Consistency

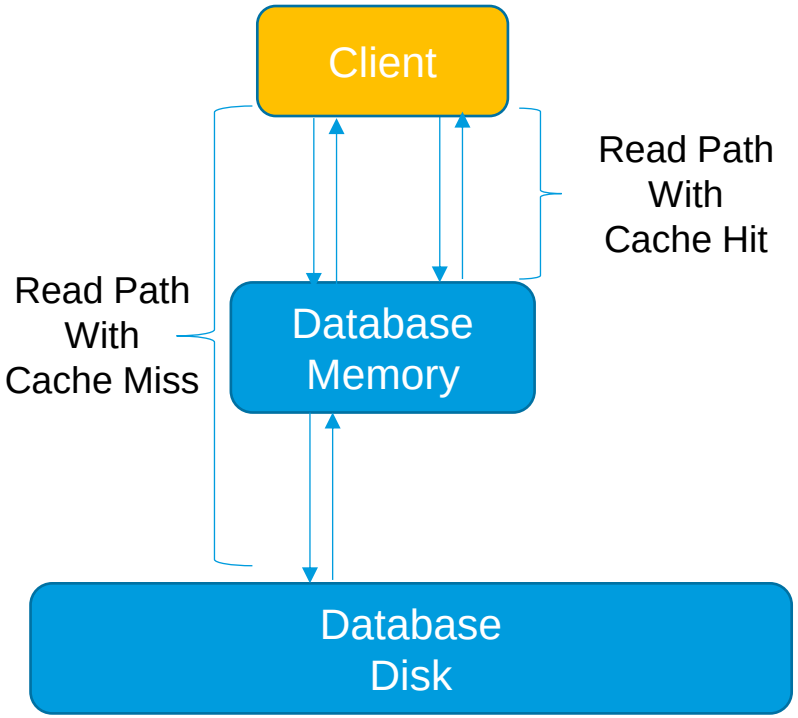
In-Memory/Cache



Predictable

Latency - Low
Throughput - Consistent

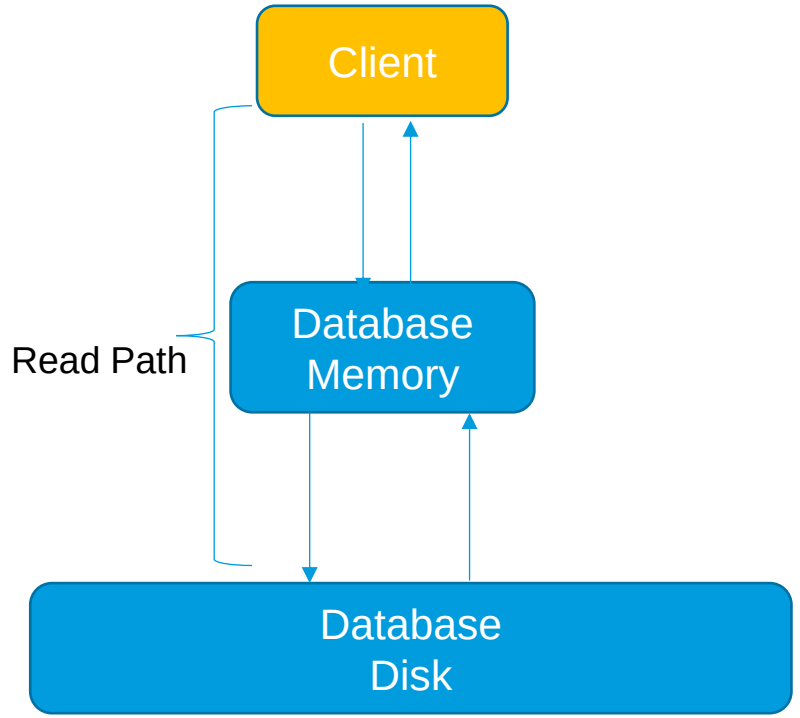
Memory-First



Latency – Varying
Throughput - Varying

Varying

Hybrid-Memory

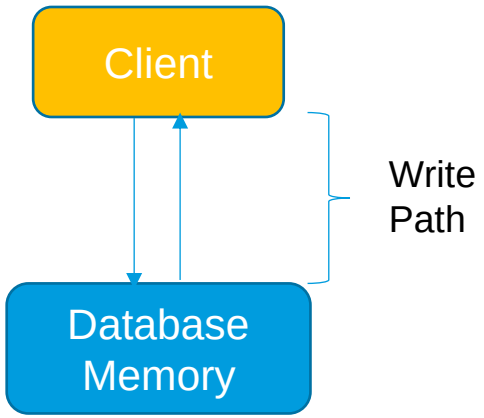


Latency – Low or High
Throughput - Consistent

Predictable

Write Path – Latency and Consistency

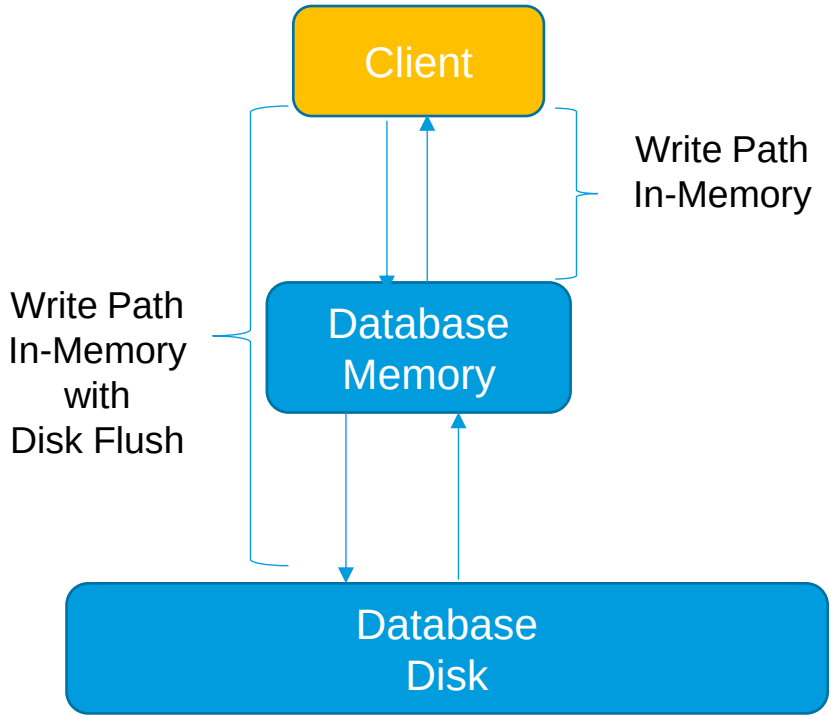
In-Memory



Predictable

Latency - Low
Throughput - Consistent

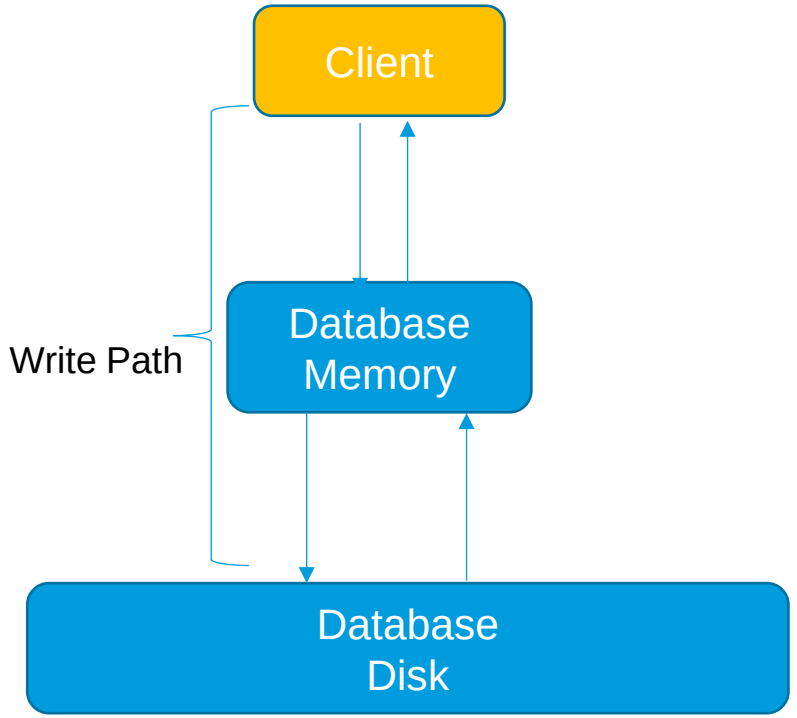
Memory-First



Latency – Varying
Throughput - Varying

Varying

Hybrid-Memory

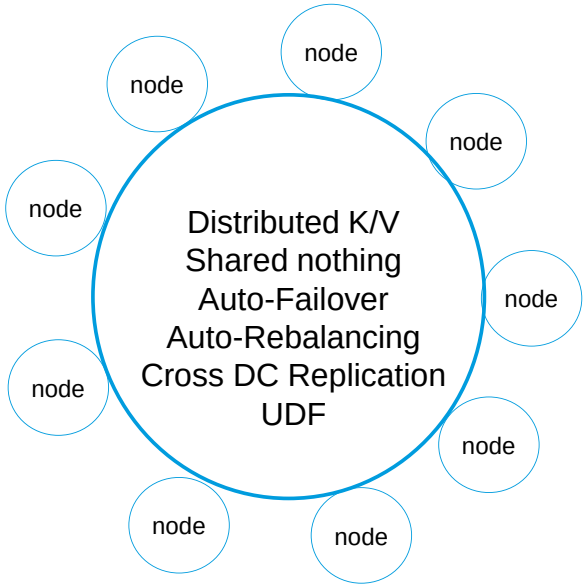


Latency – Low or High
Throughput - Consistent

Predictable

Why Aerospike ?

Aerospike Architecture

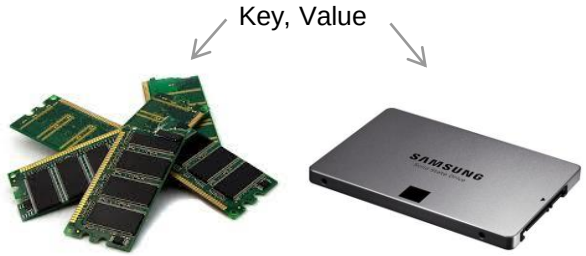


Key Differentiators in NoSQL space

Aerospike

Proprietary file system
(Achieves – Consistent Device Latency,
Follows Device throughput)

Ground up, Designed for SSDs.
(Achieves – Even wear and tear on Device)



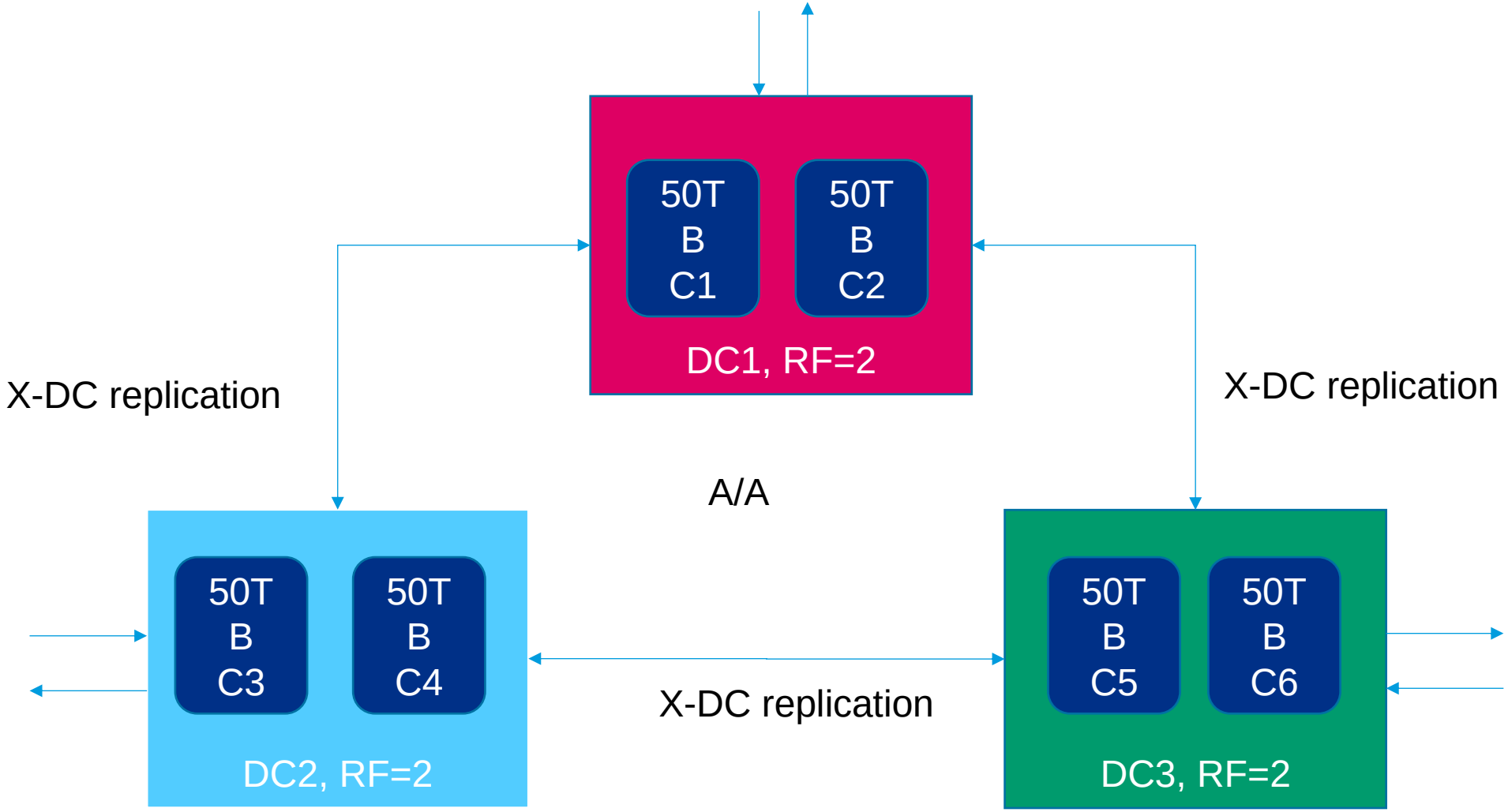
Hybrid Storage – Predictable capacity.
(Achieves – Enables huge storage on SSD)

Aerospike as NoSQL Database

Aerospike

- Written in C
- Simple KV database
- Distributed shared nothing architecture
- **AP and CP** (Strong consistency) modes
- Operates **In-Memory** or **Hybrid-Memory Modes**
- Low write amplification
- **SSD optimized for consistent performance**
- High storage density
- Low CPU utilization
- **UDF** for server side computations

Designing a 50TB A/A Database



PayPal

In-Memory Database for 50TB

(**Predictable performance**)

RAM – 384GB/Server

Servers ~ 1024

DC1

Racks~18

DC2

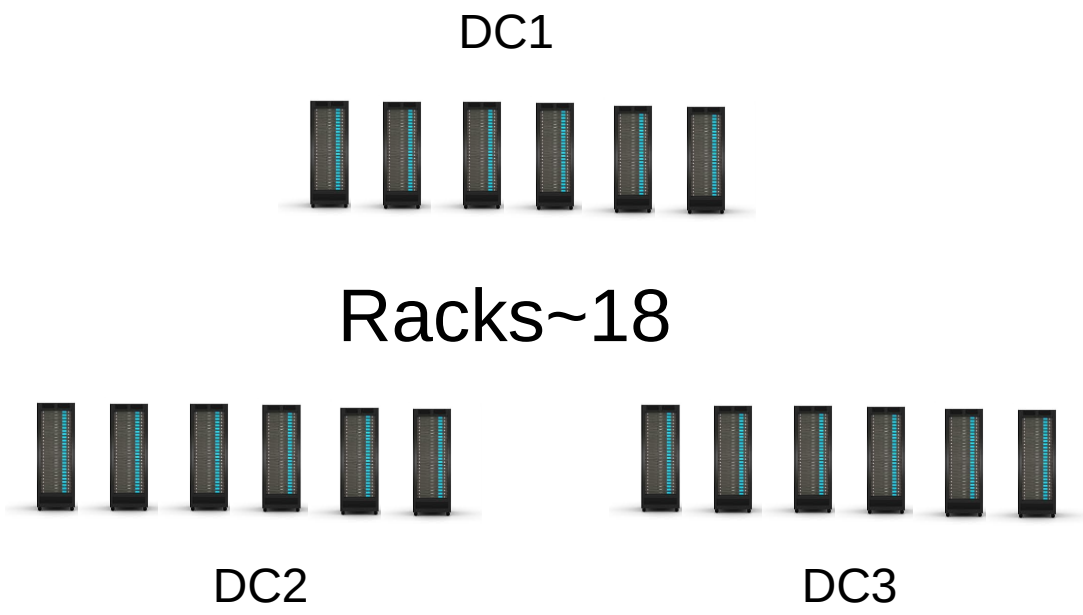
DC3

Price ~$12M

PayPal

Memory-First Database for 50TB

(**Predictable performance**)



Servers ~ 1024

Price ~$15M

PayPal

Memory-First Database for 50TB

(Unpredictable performance)

Servers ~ 120

DC1

Racks~3

DC2

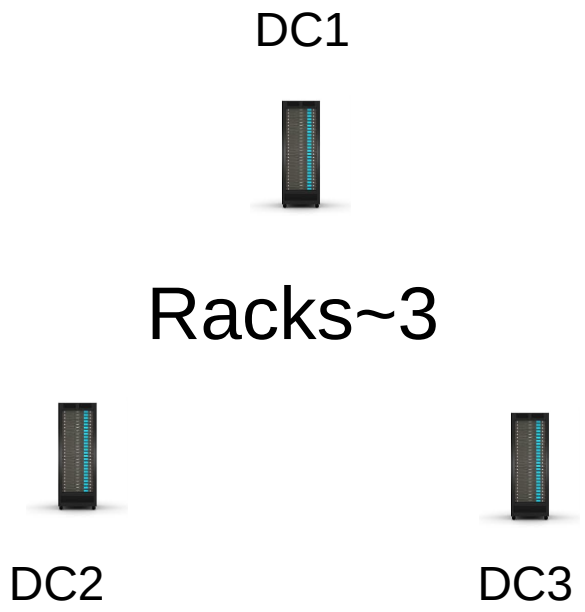
DC3

Price ~$1.8M

PayPal

Hybrid Memory Database for 50TB

(**Predictable performance**)



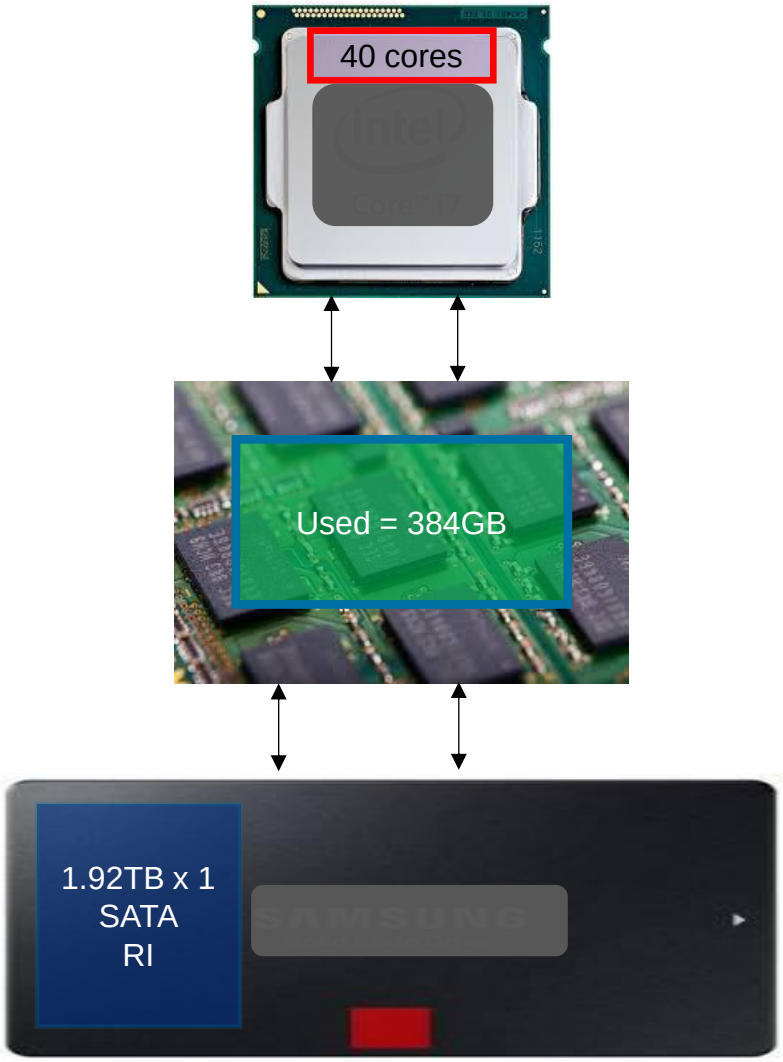
Servers ~ 120

Price ~$1.8M

PayPal

System Efficiency

System configuration

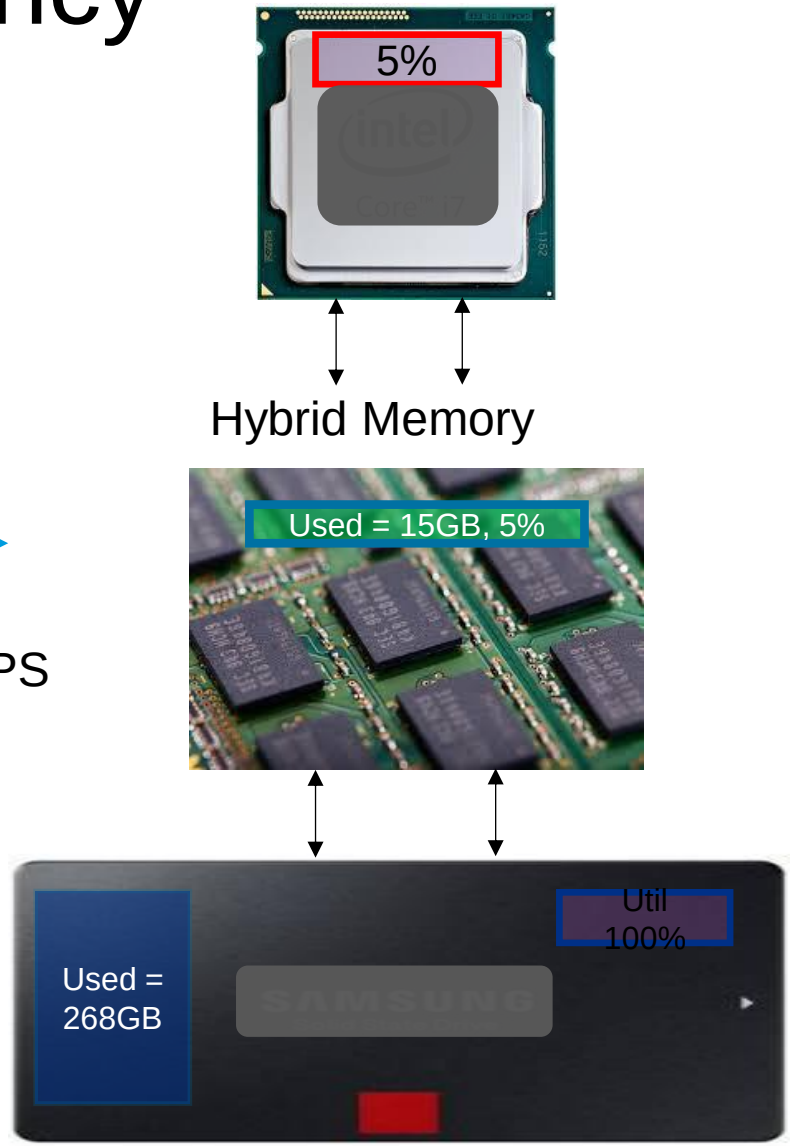


Load 250M
1KB Value Size
Raw Data = 250GB

Max Write = 100K TPS
CPU = 5%
Mem = 15GB
Disk = 268GB

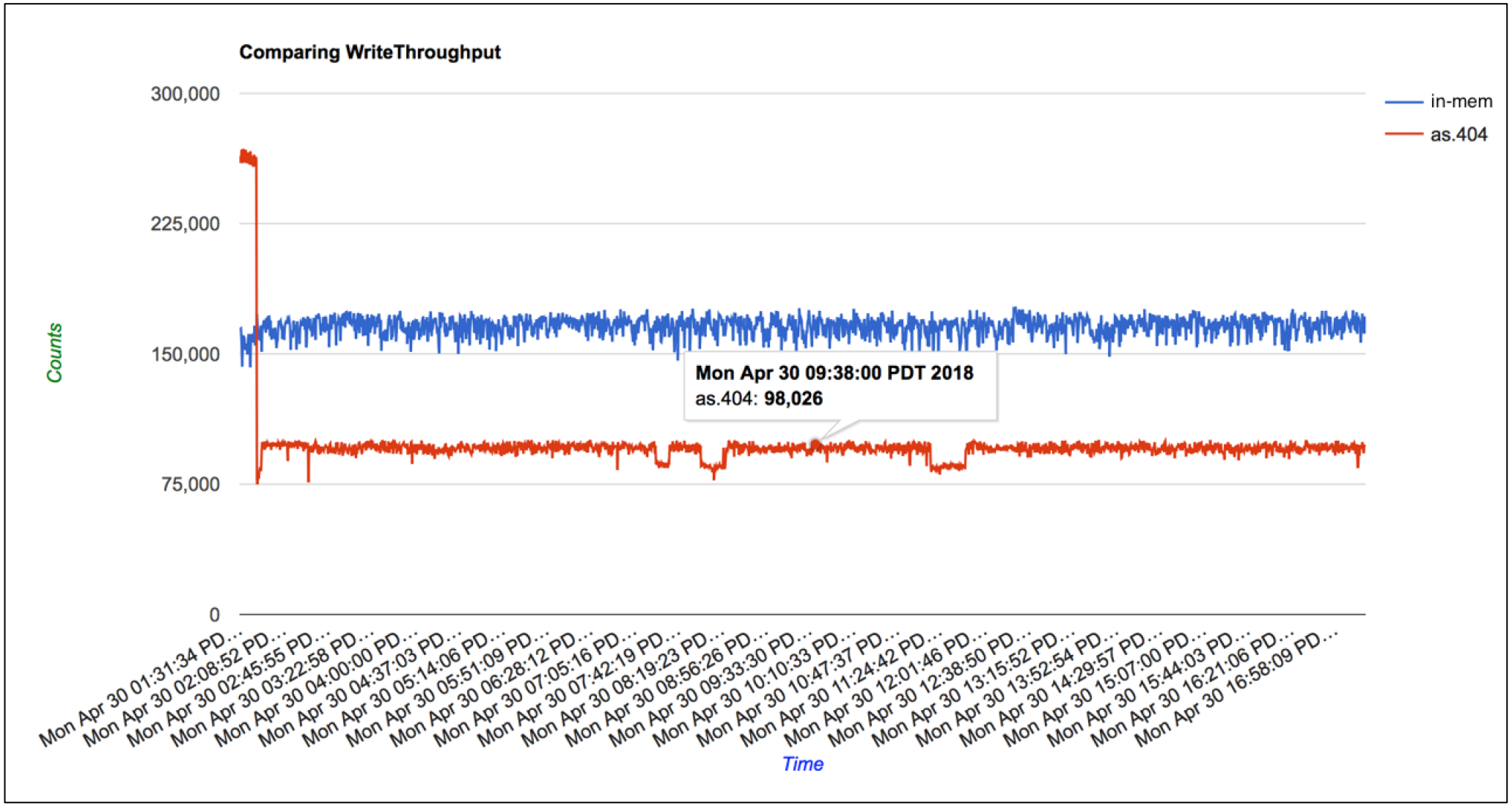
Aerospike

Hybrid Memory



PayPal

Write Throughput (in-mem vs hma)



PayPal

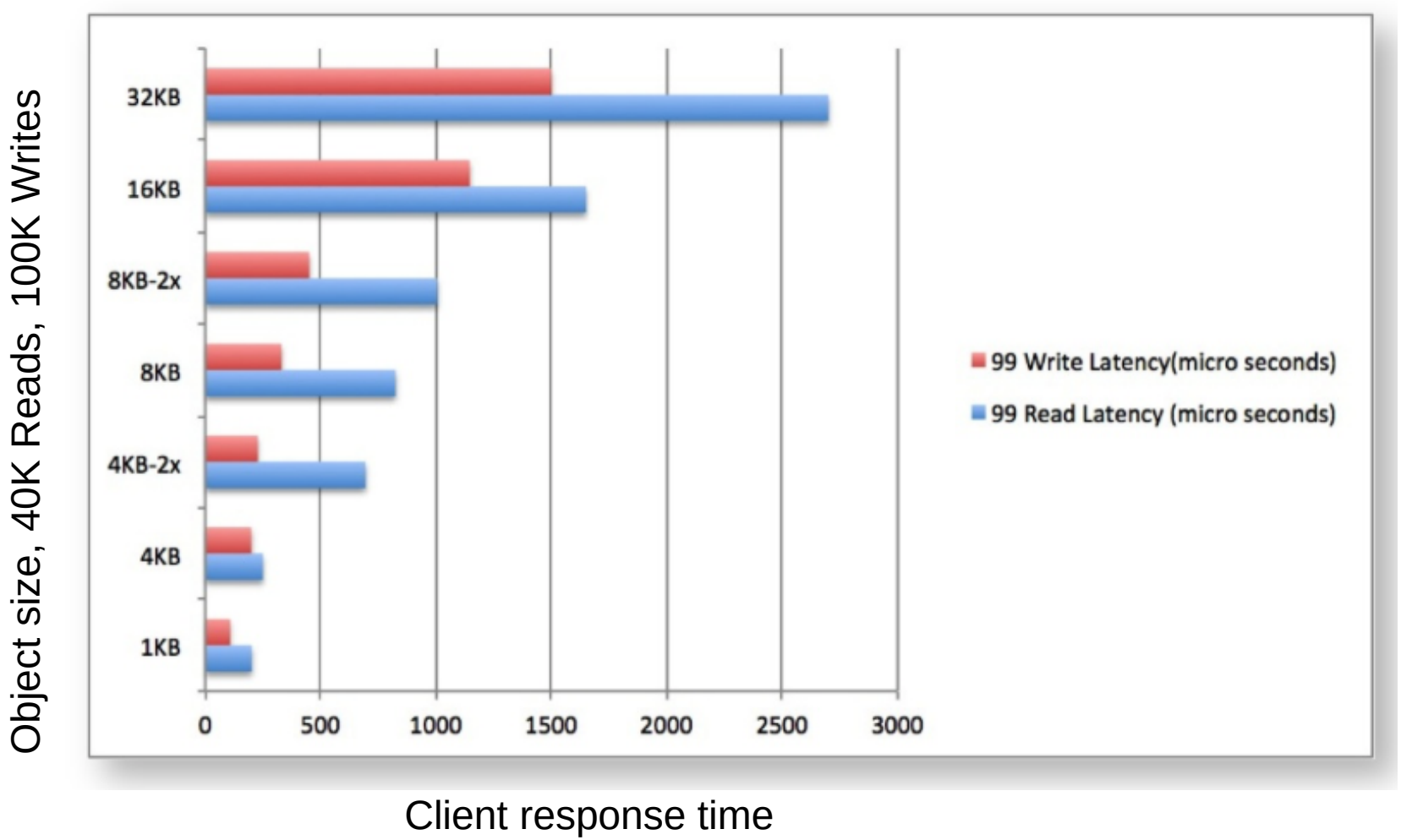
HW Profiles

Package	M(Samsung)	XL(Samsung)	L(Intel)	XL(HP-OEM)
Server	 **Dell Power Edge R730XD Haswell**	 **Dell Power Edge R730XD Haswell**	 **Dell Power Edge R730XD Haswell**	 **HP DL360P Gen 9 Haswell**
Cost/Unit	28,000	$45,382	35,000	$23,518
Cost/GB				
RAM	56GB	512GB	512GB	384GB
Storage Capacity	.4TB	12.8TB	8TB	12TB
SSD Model	 **Samsung SM1715**	 **Samsung SM1715**	 **Intel P3700**	 **Intel S3700**
SSD Form Factor	PCIe NVMe	PCIe NVMe	PCIe NVMe	SAS
SSD Storage Density	3.2TB	3.2TB	2TB	1.92TB
SSD #	2x	4x	4x	6x
Random Read IOPS (4KB)	750,000	750,000	450,000	75,000
Random Write IOPS (4KB)	130,000	130,000	175,000	36,000
Sequential Read (MB/s)	3000	3000	2800	500
Sequential Write (MB/s)	2200	2200	2000	460
Aerospike Max Read throughput (1KB)	380,000	760,000	780,000	200,000
Aerospike Max Write throughput (1KB)	420,000	840,000	800,000	542,000
Aerospike Read Latencies(Min/Avg/95th/99th/Max) 80K-R/200K-W @1KB	Percentile / Latency(µs) Min 113 Avg 115 95th 108 99th 108 Max 108	Percentile / Latency(µs) Min 113 Avg 115 95th 108 99th 108 Max 108	Percentile / Latency(µs) Min 112 Avg 113 95th 114 99th 114 Max 114	Percentile / Latency(µs) Min 271 Avg 402 95th 407 99th 645 Max 856
Aerospike Write Latencies(Min/Avg/95th/99th/Max) 80K-R/200K-W @1KB	215/264/288/288/288(µs)	215/264/288/288/288(µs)	233/236/240/240/240(µs)	189/196/202/221/277(µs)
Cold restart (Minutes)	23m	1hr 45m 52s	2hr 10m 16s	1h 11m 53s
Warm restart(Minutes)	1m 57s	3m 14s	7m 5s	3m 10s

PayPal

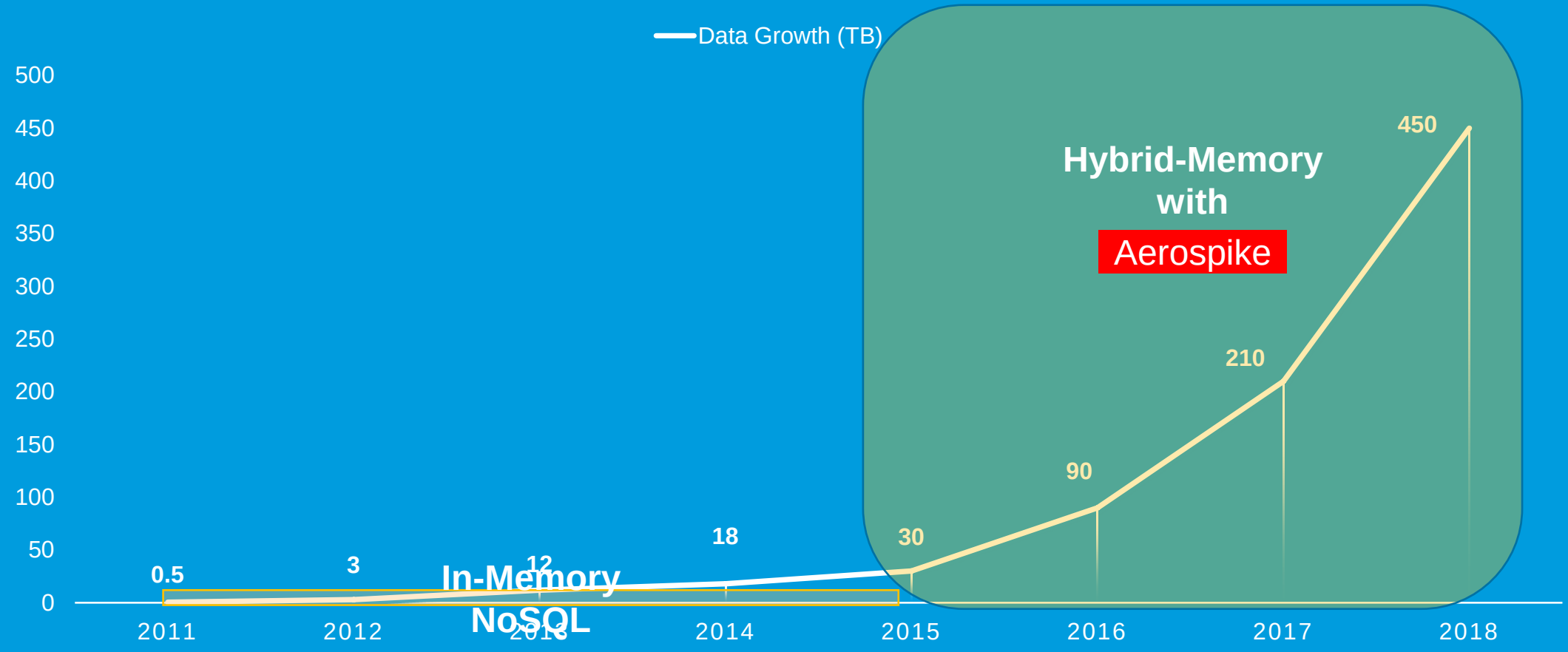
Aerospike Object size vs. Latency

2x = 80K read + 200K writes



PayPal

FRAUD DETECTION SYSTEMS WITH AEROSPIKE



* TB = 1 Copy

PayPal

In-Memory NoSQL (50TB)

HP DL360P – 1U

RAM – 384GB/Server

Metric	Cost
Read Latency	~1-2ms
Write Latency	~1ms
Read Throughput	200K
Write Throughput	200K

Total Cost = $12.5m

Aerospike (Hybrid Memory NoSQL) (50TB)

Dell 630/730– 1U/2U

RAM – 512GB/Server

NVMe/PCIe SSD
(2x3.2TB = 6.4TB)

Metric	Cost
Read Latency (2R:5W)	200 microseconds
Write Latency (2R:5W)	150 microseconds
Peak Read Throughput	840K
Peak Write Throughput	800K

M Size	256 GB RAM/ 6.4 TB SSD
L Size	512 GB RAM/ 12.8 TB SSD

Total Cost = $1.8m

Performance Cost

3x

Space/Power

8x

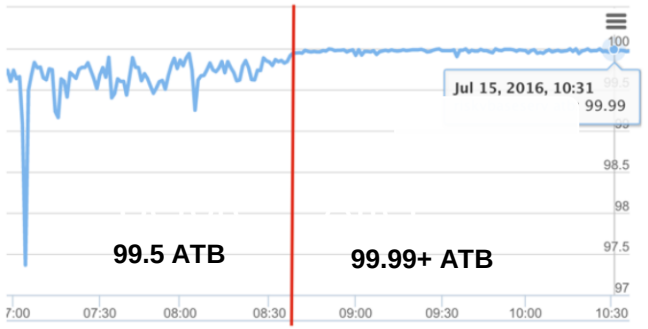
18 Racks
# of server 1024

3 Racks
# of servers = 120

Availability

10x

- **Inconsistent Performance**
- **Unstable**
- **Average Throughput – 200K TPS**
- **Low latency (~1ms)**



- **Consistent Performance**
- **Stable**
- **Avg Throughput – 1M TPS**
- **Ultra low latency (~200us)**

Thank you

Q & A