"NVMe, Storage Class Memory and Operational Databases: Real-World Results"

Brian Bulkowski, CTO and Founder

August, 2016

What is Aerospike ?

Large-scale DHT Database (10B ++ objects, 100T++, **O(1) get / put**)
… with queries, data structures, UDF, fast clients ...
... On Linux ...

High availability **clustering & rebalancing** (proven 5 9's, no load balancer)

Very **high performance C code** – reads and writes
(2M++ TPS from Flash, 4M++ TPS from DRAM ***PER SERVER***)

KVS++ provides **query, UDF, table/columns, aggregations, SQL**

Direct attach storage; persistence through replication and **Flash**

Cloud-savvy – runs with **EC2, GCE others; Docker**, more …

Dual License: Open Source for devs, Enterprise for deployment

Enterprise Requirement: 2-Speed IT

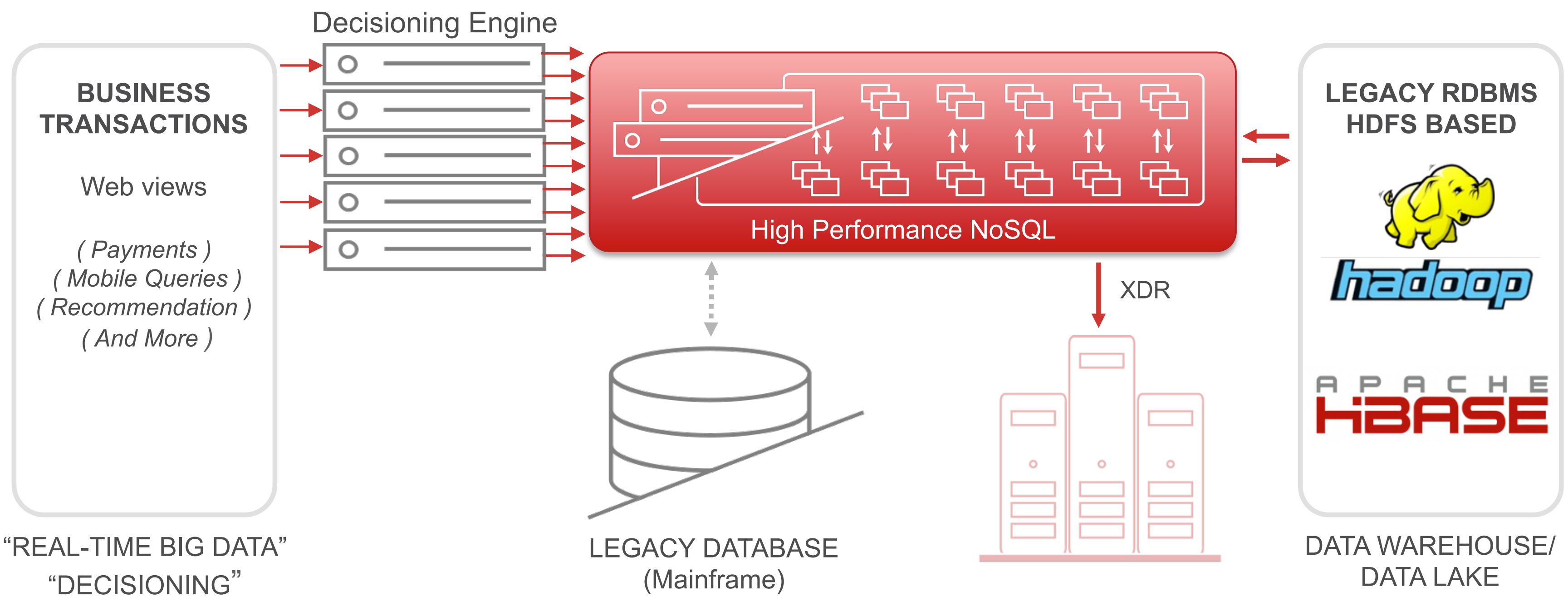
The only way for traditional enterprises to easily build Digital business is adapting to 2-Speed IT decoupling Systems of Record and Systems of Engagement

Front office Consumer Scale Digital Applications that move at a Faster pace and act as Systems of Engagement

Back office legacy Enterprise Scale Applications that move at a slower pace and act as Systems of Record

AEROSPIKE

Architecture Overview – Flash based system of engagement



Decisioning Engine

BUSINESS TRANSACTIONS

Web views

(Payments)
(Mobile Queries)
(Recommendation)
(And More)

High Performance NoSQL

XDR

LEGACY RDBMS HDFS BASED

“REAL-TIME BIG DATA”
“DECISIONING”

LEGACY DATABASE
(Mainframe)

DATA WAREHOUSE/
DATA LAKE

500 Business Trans per sec	X	**5000** Calculations per sec	=	**2.5 M** Database Transactions per sec

Real-world engagements

AdTech – Real-Time Bidding

Challenge

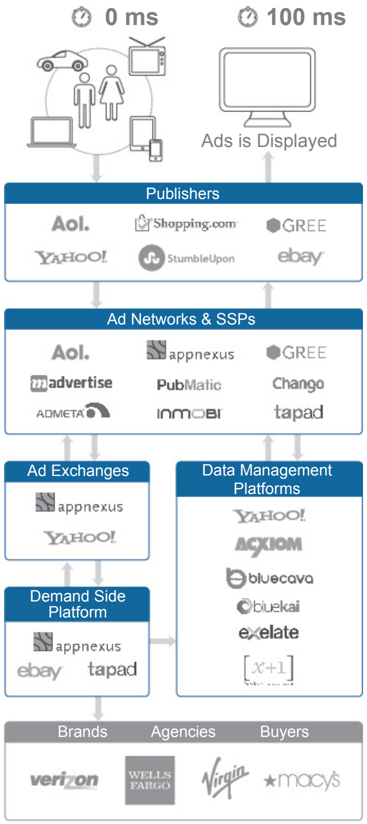
- Low read latency (milliseconds)
- 100K to 5M operations/second
- Ensure 100% uptime
- Provide global data replication

Performance achieved

- 1 to 6 billion cookies tracked
- 5.0M auctions per second
- 100ms ad rendering, 50ms real-time bidding, 1ms database access
- 1.5KB median object size

Selected NoSQL

- 10X fewer nodes
- 10X better TCO
- 20X better read latency
- High throughput at low latency



Fraud Prevention

Challenge

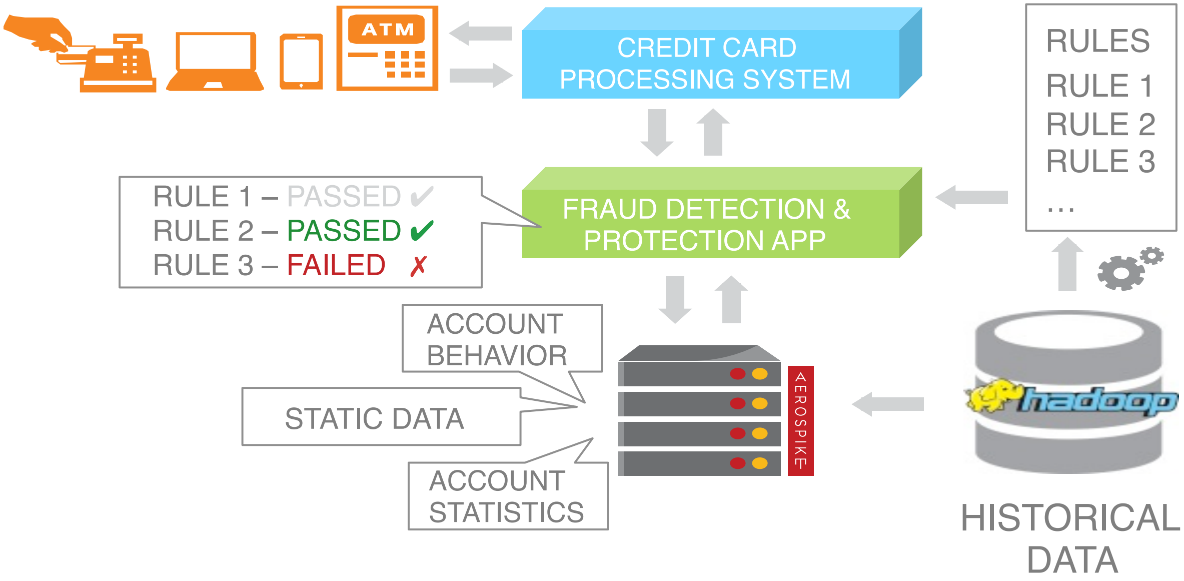
- Overall SLA 750 ms
- Loss of Business due to latency
- Every Credit Card transaction requires hundreds of DB reads/writes

Need to scale reliably

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

Selected NoSQL

- Built for Flash
- Predictable Low latency at High Throughput
- Immediate consistency, no data loss
- Cross data center (XDR) support
- 20 Server Cluster
- Dell 730xd w/ 4NVMe SSDs



AEROSPIKE

Fin Serv – Positions System of Record

Challenge

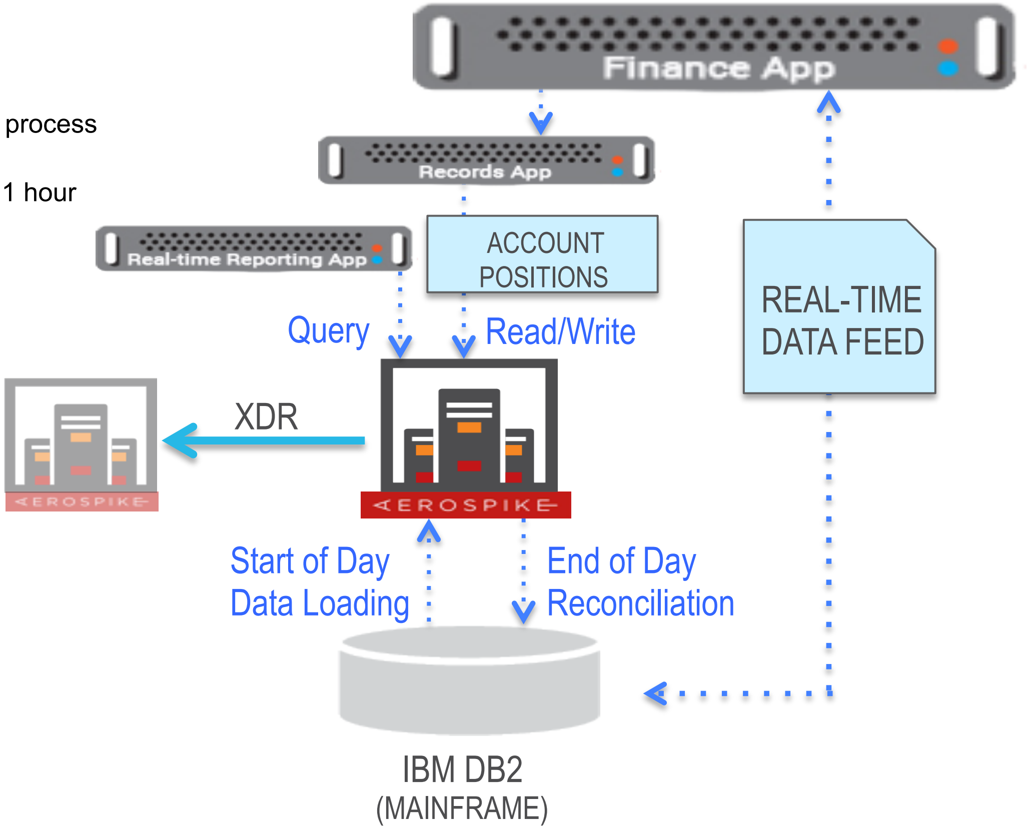
- DB2 stores positions for 10 Million customers
- Must update stock prices, show balances on 300 positions, process 250M transactions, 2 M updates/day
- Running out of memory, data inconsistencies, restarts take 1 hour
- 150 Servers -> Growing to 1000

Need to scale reliably

- 3 → 13 TB
- 100 → 400 Million objects
- 200k → I Million TPS

Selected NoSQL

- Built for Flash
- Predictable Low latency at High Throughput
- Immediate consistency, , no data loss
- Cross data center (XDR) support
- 10 Server Cluster



AEROSPIKE

Telco – Real-Time Billing and Charging Systems

Challenge

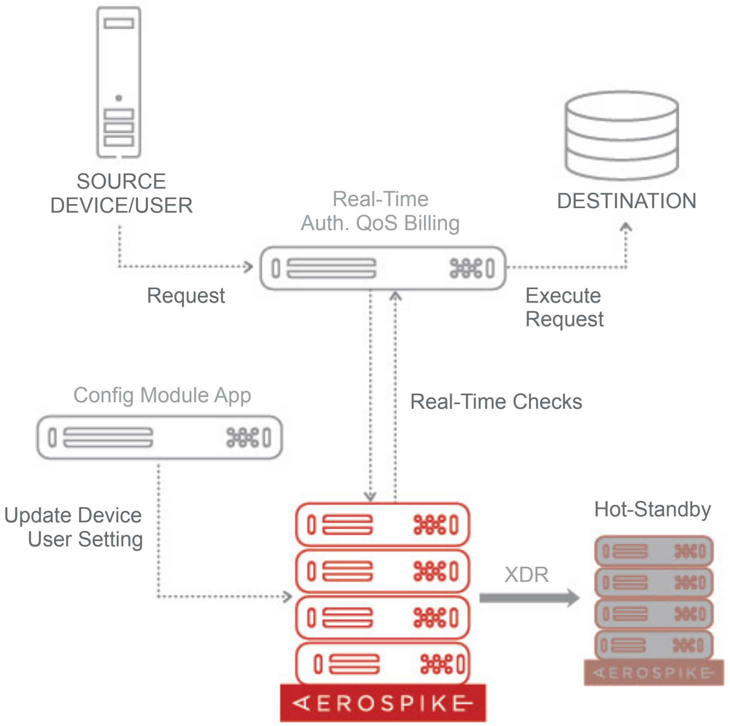
- Edge access to regulate traffic
- Accessible using provisioning applications (self-serve and through support personnel)

Need for Extremely High Availability, Reliably, Low latency

- > TBs of data
- 10-100M objects
- 10-200K TPS

Selected NoSQL

- Clustered system
- Predictable low latency at high throughput
- Highly-available and reliable on failure
- Cross data center (XDR) support



NAND Status
Wide SATA, PCIe, NVMe

Historical Perspective – Early Days

- **Early SATA (2009, 2010)**
  - Intel X25M
  - Samsung SS805
  - Devices provide 95% < 1ms for about 2,000 IOPS
  - $3 / GB

- **FusionIO - 2010**
  - PCIe, but custom driver
  - CPU, bus load
  - $8 / GB

AEROSPIKE

Historical Perspective – 2013, 2014

- **Micron proves fast PCIe possible**
  - P320 (SLC) with low bus overhead, excellent driver
  - Over 200,000 IOPS with 99.7% < 1ms
  - (SFF-8639 hot-swap 2.5” pci-e drives)
- **“Wide SATA” generally used**
  - 12 to 20 2.5” SATA drives per 2U chassis
  - Intel S3700, S3500 ; Samsung 843 favored
  - 8 drives per controller (many issues)
  - 150,000 IOPs per chassis achievable
- **Violin, FusionIO troubled, DSSD sold to EMC**
- **NVMe available but not practical**

Twentieth Century Fox

NVMe Arrives – 2015 to present

- **Linux, Windows drivers achieve performance**
- **U.2 and M.2 form factors available**

- **Intel P3700, P3600, P3500 available –**
  - 250k IOPs per card
- **Samsung PM1735 available –**
  - 120k IOPs per card
- **Micron 9100**
- **HGST, Toshiba – 30k to 50k per card**

- **SAS / SATA lingers**
  - Samsung SM1635, PM1633; Intel S3700; Micron S600 still shipping

Flash in the Public Cloud

- **Every public cloud provider has Flash**
  - AWS / EC2 has sophisticated offerings
  - Google Compute is NVMe - high performance
  - Softlayer allows own-hardware
- **Private clouds manage Flash**
  - Docker offers storage metadata
  - Pivotal manages Flash and traditional storage

1 Million TPS on 1 Server - NVMe

Intel Reaches 1 Million TPS on 6.4TB Flash Using a Single Aerospike Server

Options for storage on a database before Aerospike:

- **RAM**, which was fast, but allowed very limited storage
- **Disk**, which allowed for a lot of storage, but was limited in speed

Intel achieved 1M TPS using 4 Intel P3700 SDs with 1.6 TB capacity on a single Aerospike server. The cost per GB is a fraction of the cost of RAM, while still having very high performance.

Storage Class Memory
(and trends)

Trends in NAND

- **Diverging “Drive Writes Per Day”**
- **Low-write devices**
  - 1~2 DWPD
  - Sandisk Inifiniflash, Micron
  - Increase density, lower cost
  - Hadoop / Datalake “all flash” use
- **High-write devices**
  - 10~15 DWPD
  - P3700, Hitachi, Samsung
  - Optane may disrupt everything

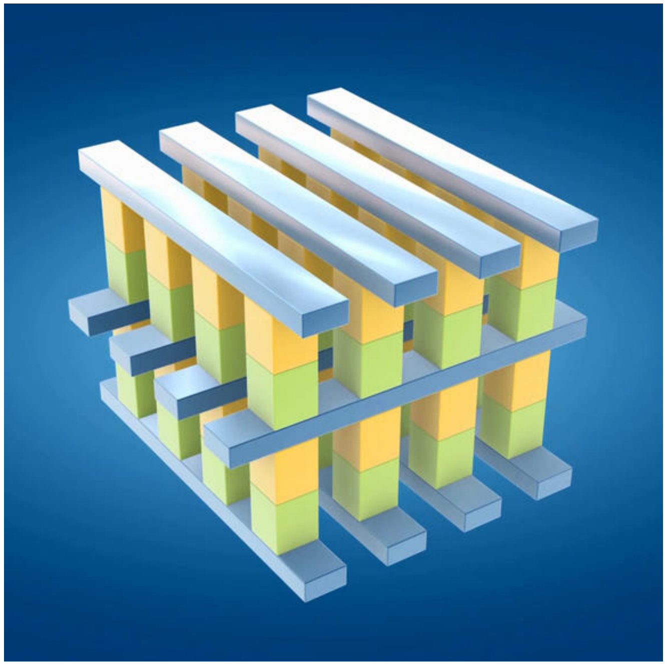
“All Flash arrays”

- **Database knows best, not storage**
  - Database should manage
    consistency vs availability
  - Database should manage views and snapshots
- **Array vendors have started making “databases”**
  - “Object stores”
- **High Density “flash aggregation”**
  - Sandisk Infiniflash – SATA
    - High-read, or write-once, applications
  - Apeiron ASD1000 – NVMe
    - Read and write applications
  - Vexata, others

AEROSPIKE

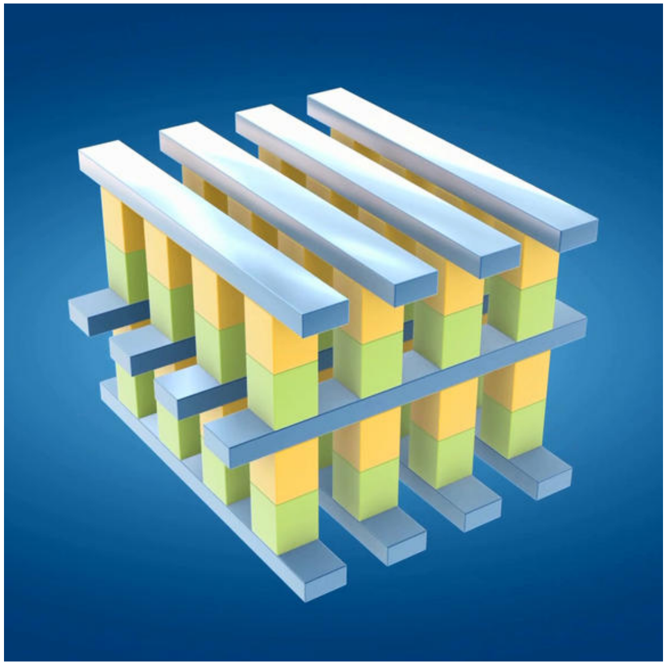
What is 3D Xpoint?

- **Persistent storage using chips**
  - No power while idle
- ***Does not use transistors***
  - Resistor / phase change “but different”
- **Chips almost as fast as DRAM**
  - DRAM – 10 ns ; Nand 10 micro
    Xpoint 1 micro to 100ns
- **“Infinite” write durability**
- **128B read and write granularity**
  - NAND write granularity --- 16 MB
  - DRAM write granularity --- 64 B



Intel’s 3D Xpoint roadmap (public info)

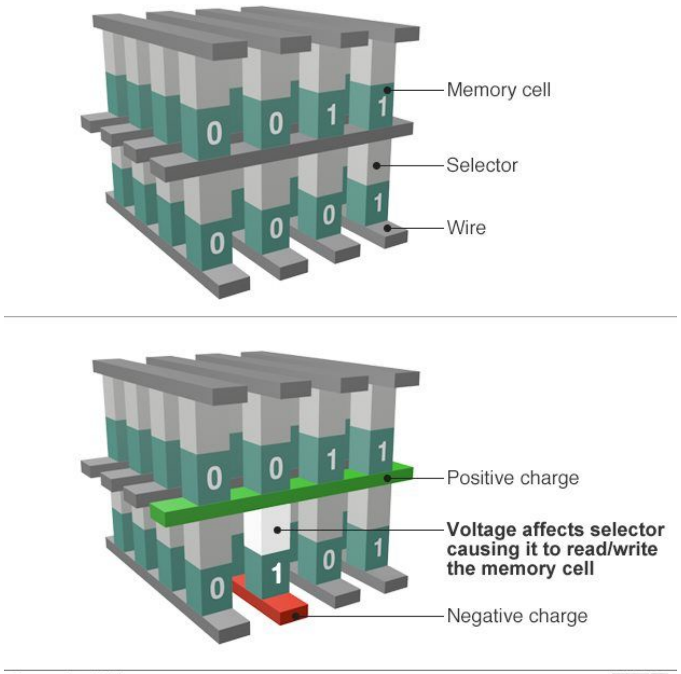
- **Optane *this year***
  - 3D Xpoint in 2.5” NVMe package
  - “7x faster” – limited by NVMe !
  - Very high write durability
  - Replaces SLC for some use cases
  - Unknown pricing
- **NVDIMM (on memory bus)**
  - Removes NVMe limit
  - Intel cagy on delivery – “uncommitted”
  - Competes with DRAM
  - *Hard to program to*
  - *Really* changes the world



How to architect for DDR 3D XPoint

- **It's not exactly like DRAM**
  - *It persists*
  - When a system restarts, need to reset
  - Slower, so different data structures required
- **It's not exactly like storage**
  - It's on the memory bus
  - Small blocks for reads and writes
  - New instructions for persistent control
- **Cache approaches will be inferior**

How 3D XPoint memory works


Memory cell
Selector
Wire
Positive charge
Voltage affects selector causing it to read/write the memory cell
Negative charge
BBC

Source: Intel, Micron

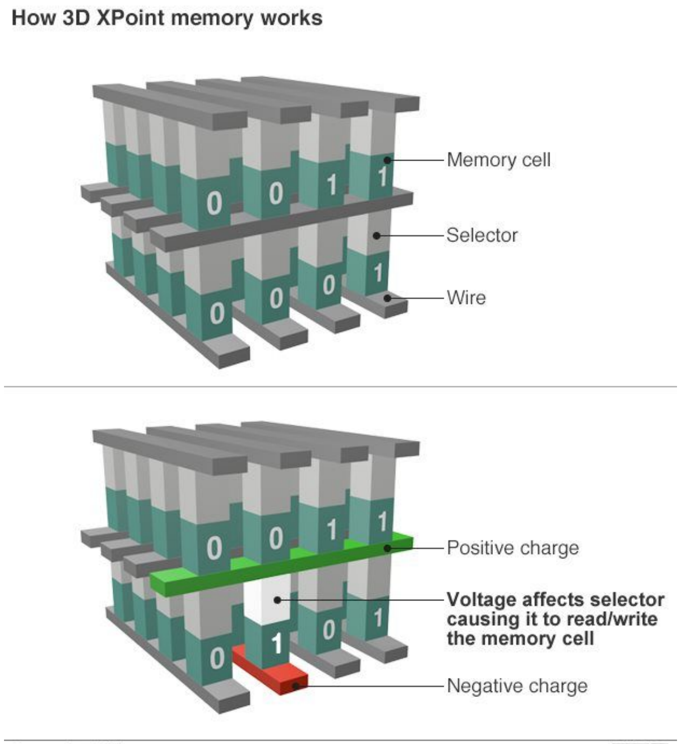
Architecting for DDR 3D XPoint

- **Think of it like DRAM**
  - Lower power consumption
  - Much higher density (1T++)

- ***4x efficiency gain***
  - No defrag required
  - No overprovisioning

- Aerospike thesis:

  Indexes in 3D Xpoint


How 3D XPoint memory works
0
0
1
1
0
0
0
1
Memory cell
Selector
Wire
0
0
1
1
0
1
0
1
Positive charge
Voltage affects selector causing it to read/write the memory cell
Negative charge

Source: Intel, Micron

Thank You

Questions?

Like this ? Shout out !
@bbulkow
#aerospike #nosql