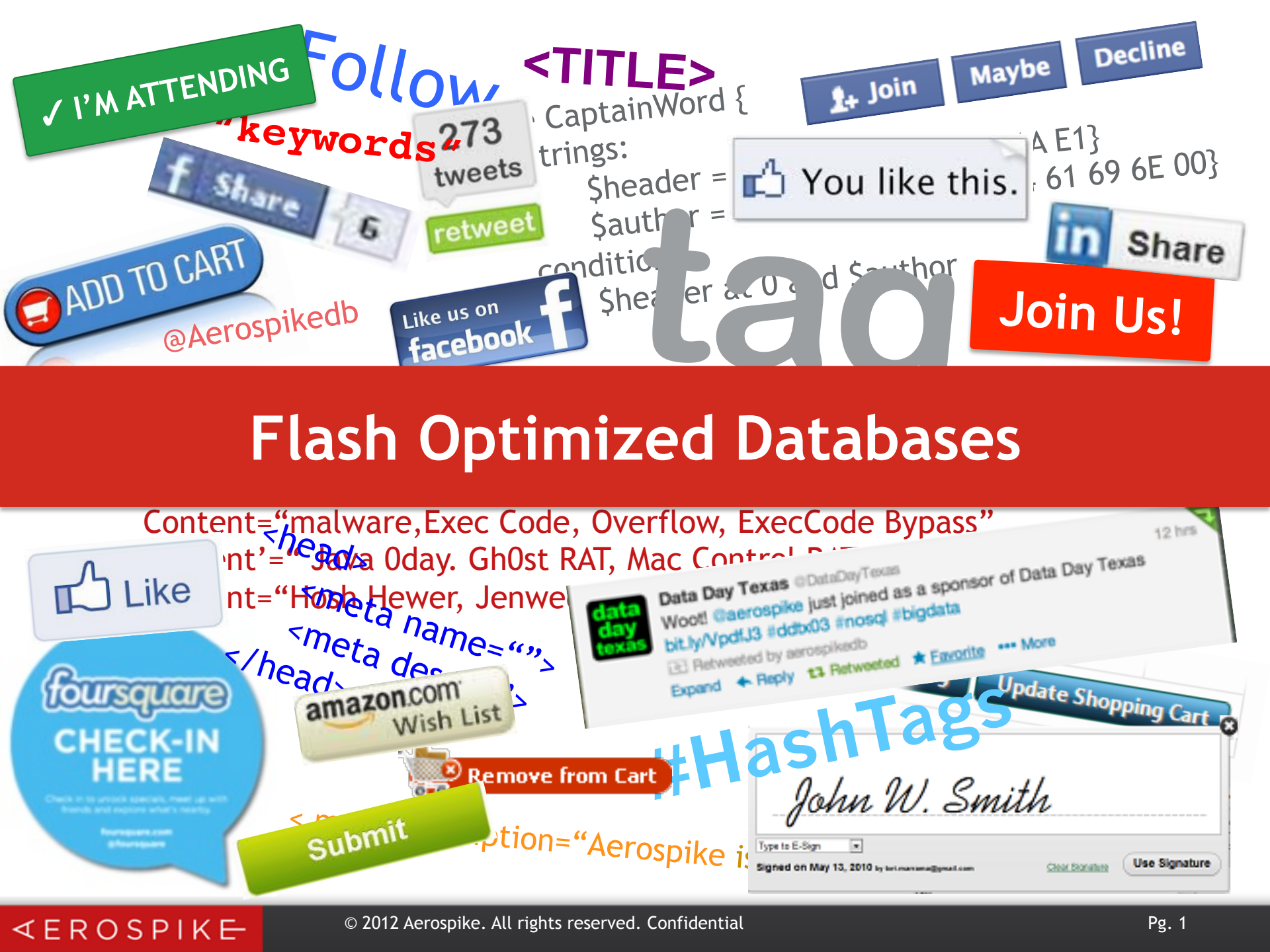
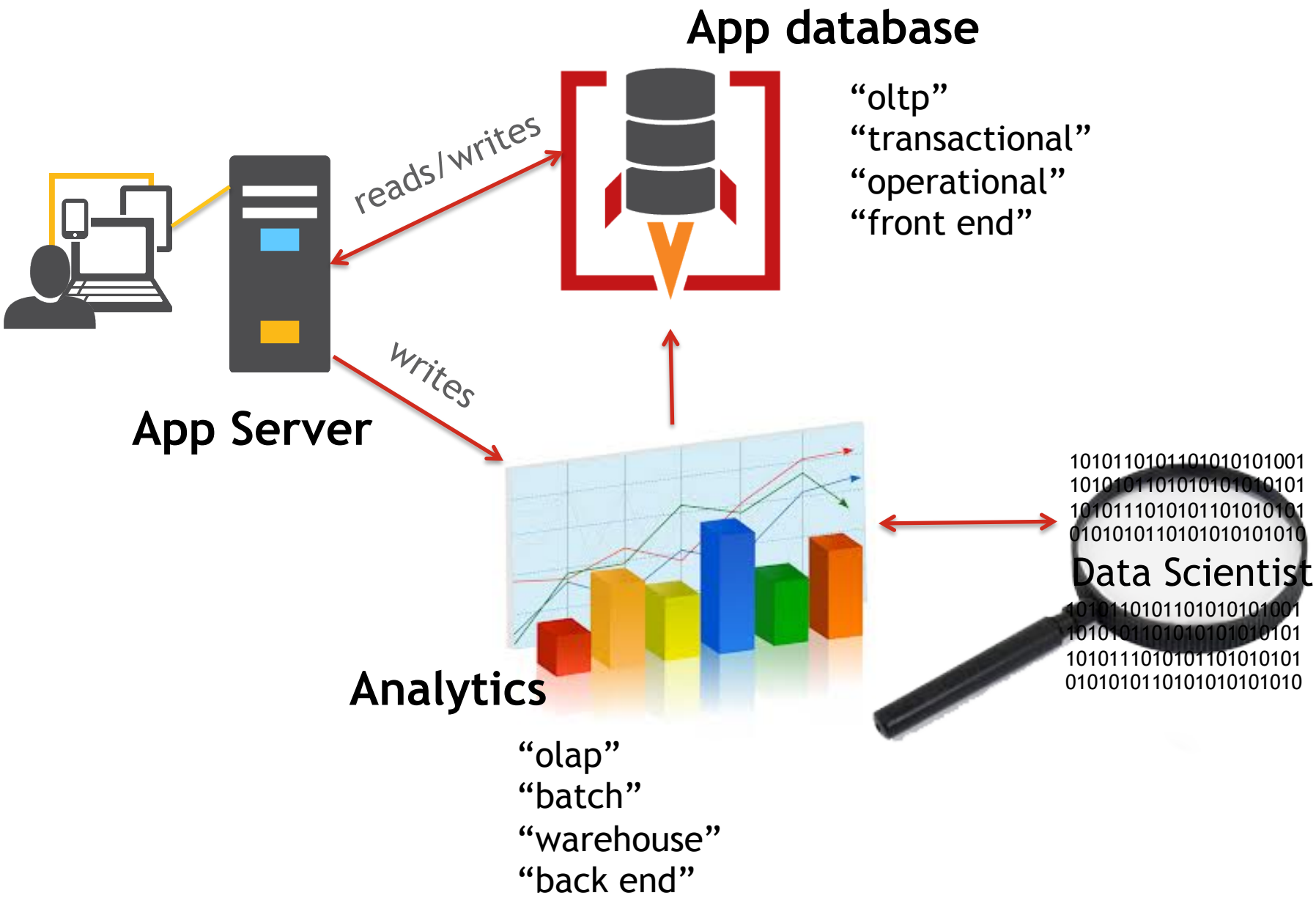
Flash Optimized Databases



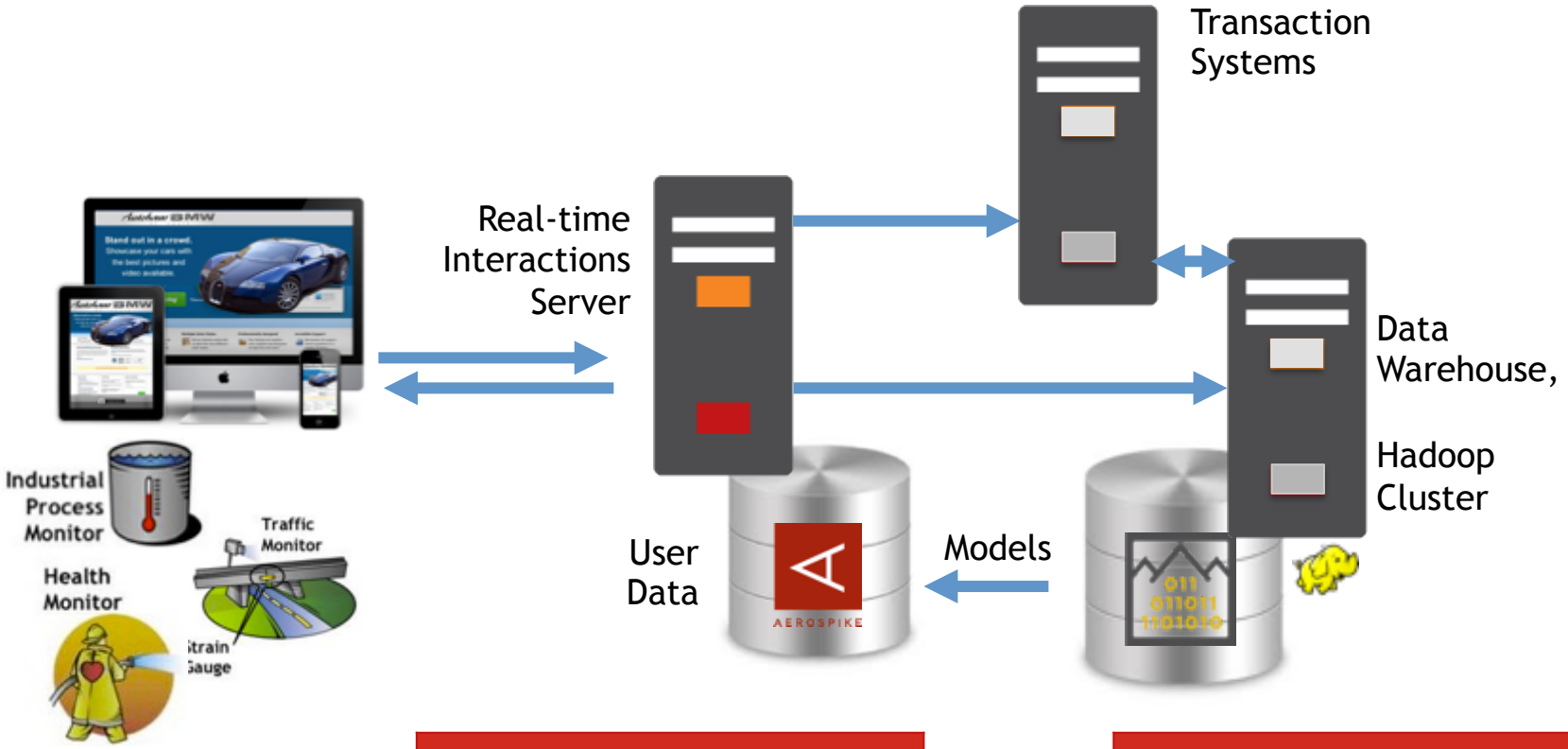
AEROSPIKE

ROCKET FUEL FOR BIG DATA APPS™

Aerospike aer . o . spike [air-oh- spahyk]
noun, 1. tip of a rocket that enhances speed and stability



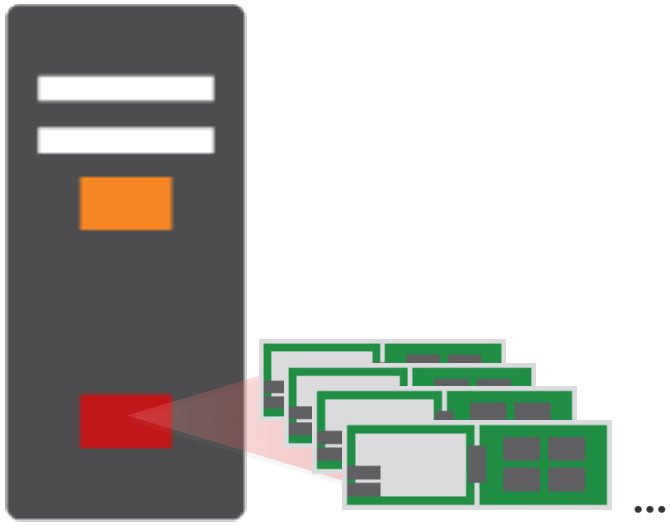
Typical Deployment



Real-time Interactions	Batch Analytics
• Frequency caps • Recent ads served • Recent search terms	• User segmentation • Location patterns • Similar audience

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



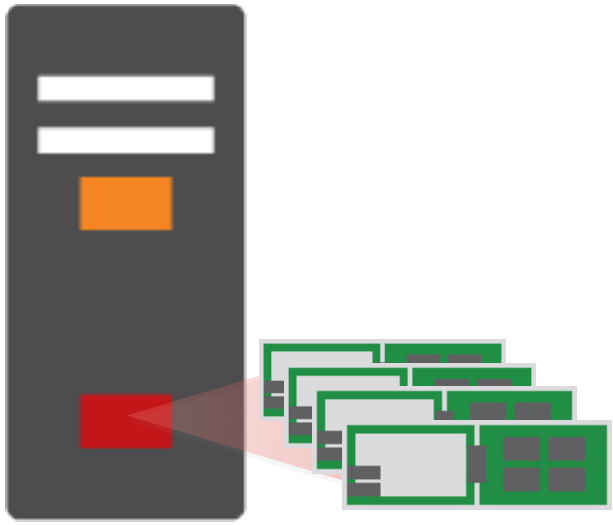
- This Year
  - 12 core Xeon
  - 128G RAM
  - 2T~4T SATA / PCIe (12 s3700 / 4 P320h)
  - 100,000 read TPS, 50,000 write TPS
  - 3K object size / 1B objects
  - 4 to 10 node cluster

In-memory Big Data – a contradiction?

How do customer use in-memory big data?

- Advertising optimization
- Fraud detection

  (fraud is everywhere)
- Retail “deals” calculation
- Financial positions
- “T0+” financial analysis
- Streaming machine learning with Esper, Storm?

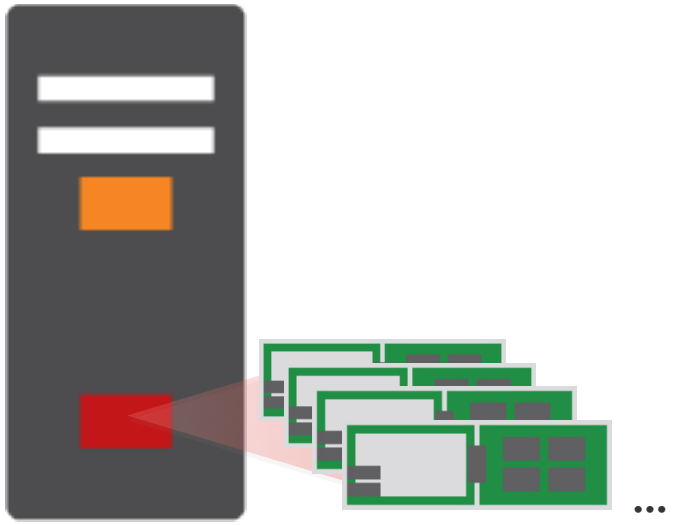


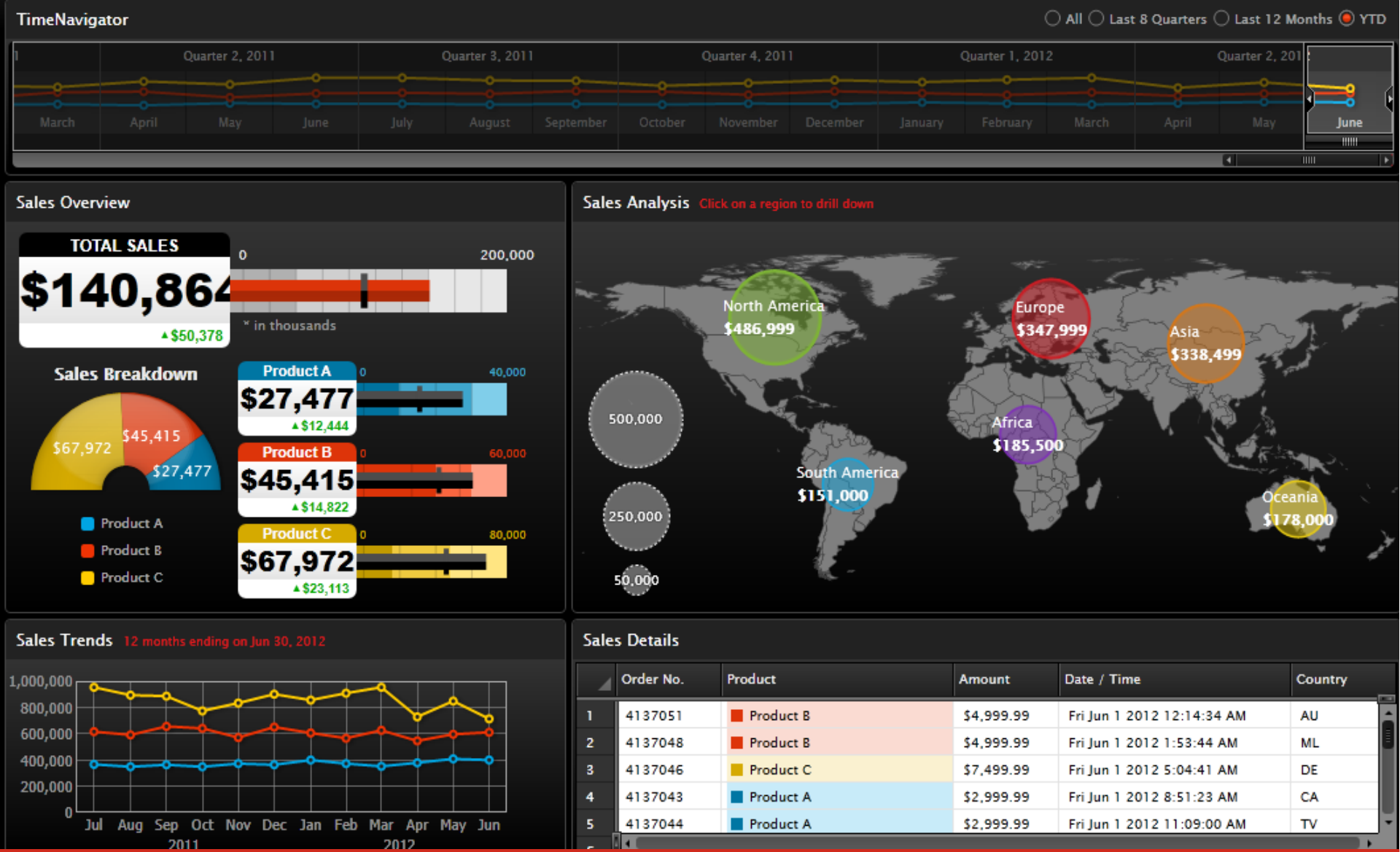
…

How do customer use in-memory big data?

- MapReduce over …
  - Data subsets
  - Time ranges
  - In _milliseconds_
- Track every ….

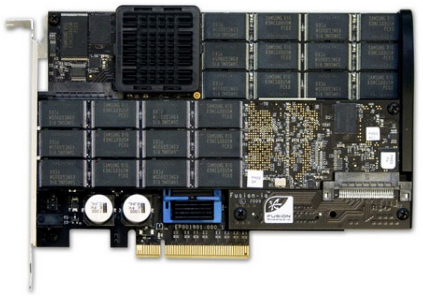
  IP address, cookie, search term ?
- Load new data sets in minutes ?





Advertising: recent activity + predictions

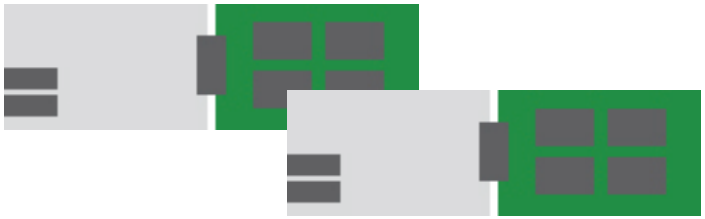
facebook + = $200+M



Facebook and Apple bought *at least* $200+M in FusionIO cards in 2012

(55% of $440M revenue estimate, reported in quarterly FusionIO earnings)

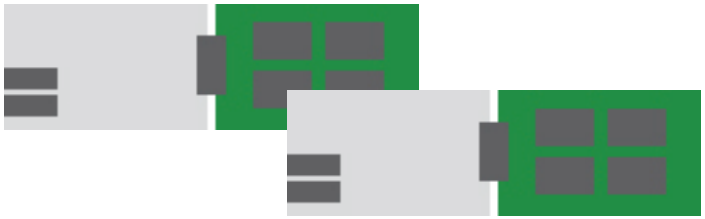
Everyone wants that “facebook architecture”



What about SSDs?

All databases go faster with SSD, right?

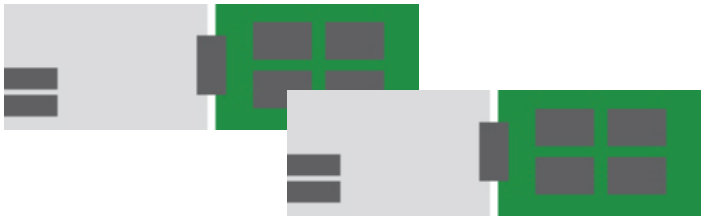
(Most DBs go 3 ~ 4x faster with SSD)



SSDs are “different”

Read locality doesn’t matter (like main memory)

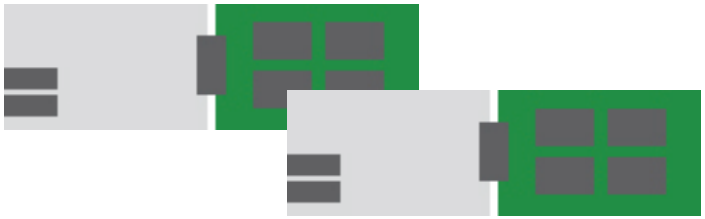
Streaming speeds are 3 ~ 4 times faster



SSDs are “different”

Write on large blocks

(helps disk’s internal defragmentation)



SSDs are “different”

Gain parallelism

(use OS routines that queue)

10T example
(a reasonable project budget)

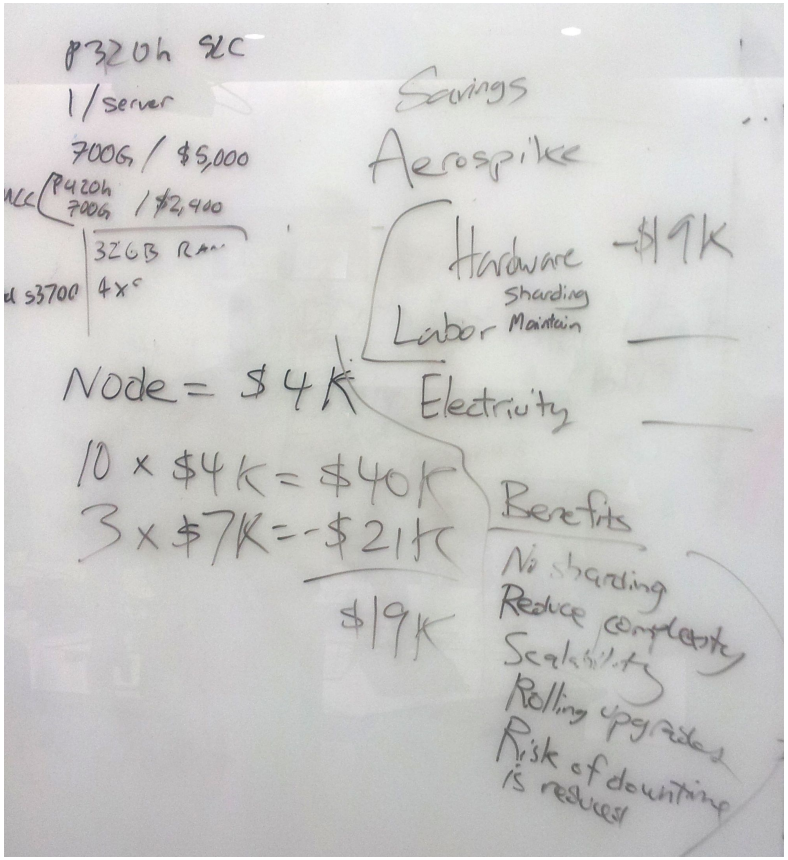
Storage type	SSD	DRAM
Storage per server	2.4 TB (4 x 700 GB)	180 GB (on 196 GB server)
TPS per server	500K	500K
Cost per server	23000	30000
# Servers for 10 TB (2x Replication)	10	110
Server costs	230,000	3,300,000
power/Server (kWatts)	1.1	0.9
Cost kWh ($)	0.12	0.12
Power costs for 2years	46,253	416,275
Maintenance costs for 2 years		$$$
Total	**$276,253**	**$3,716,275**

"…data-in-DRAM implementations like SAP HANA.. should be bypassed…
..current leading data-in-flash database for transactional analytic apps is Aerospike."

- David Floyer, CTO, Wikibon

http://wikibon.org/wiki/vData_in_DRAM_is_a_Flash_in_the_Pan

Real world calculation



200G

Redis with DRAM:
10 servers @ $4K = $40K

Aerospike with Flash:
3 servers @ $7K = $21K

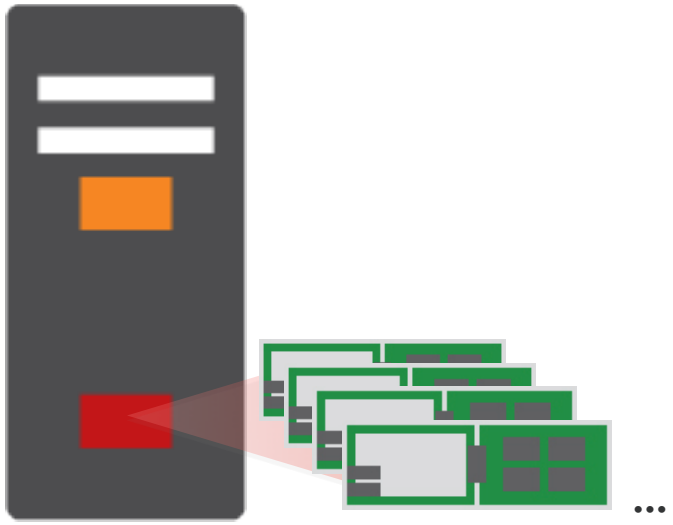
saves $19K

(and more as you scale up)

How do you optimize for SSD?

Flash-optimized Storage Layer

- Log structured file system, “copy on write”
- Data written in flash optimal large block patterns
- All indexes in RAM for low wear

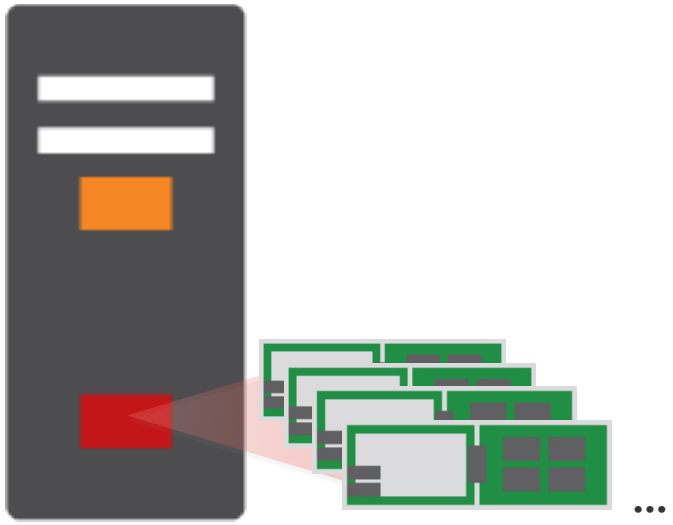


SSD performance varies widely

- Aerospike has a certified hardware list
- Free SSD certification tool, CIO, is also available

Flash-optimized Storage Layer

- Constant background defragmentation
- Random distribution using hash does not require RAID hardware
- Fast restart through shared memory



SSD performance varies widely

- Aerospike has a certified hardware list
- Free SSD certification tool, CIO, is also available

Flash-optimized Storage Layer

- Direct attach storage optimized

  (nothing else is fast)

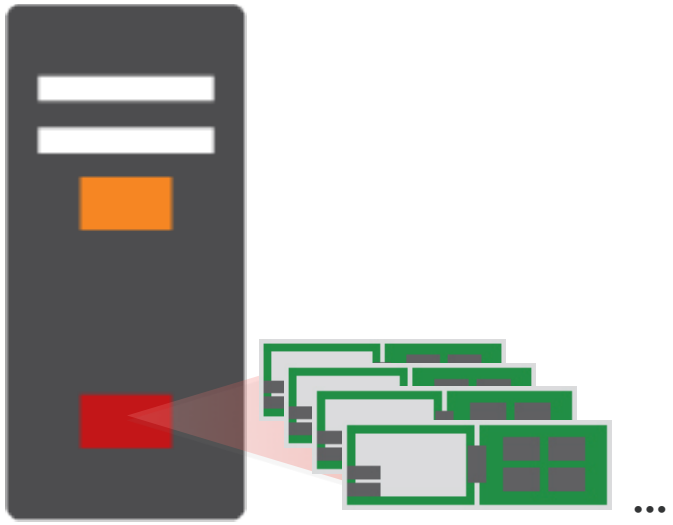
- Don't use TRIM

  (Tends to block the device)

- Multiple servers copies for ultimate HA

  (We all know servers fail)

  (no one trusts Flash storage **yet**)



SSD performance varies widely

- Aerospike has a certified hardware list
- Free SSD certification tool, CIO, is also available

Next generation interfaces

- OpenNVM KV

  Aerospike – first implementation

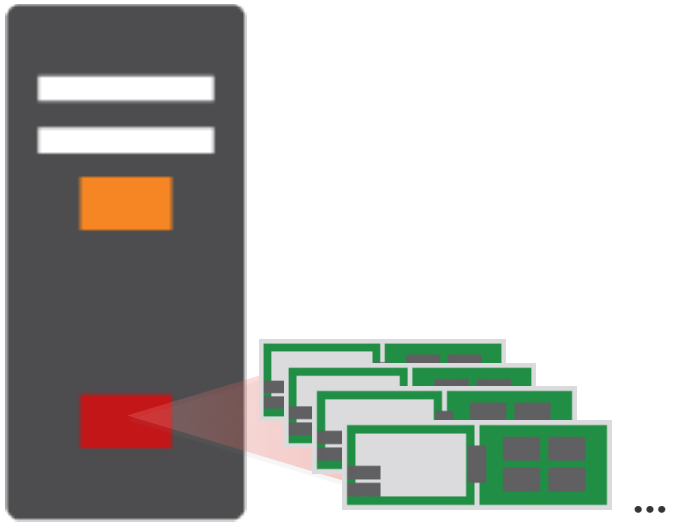
- PCIe optimization

  Non Transparent Bridge Mode

- RDMA

- In device computation ?

- Others – see me



SSD performance varies widely

- Aerospike has a certified hardware list
- Free SSD certification tool, CIO, is also available

Which SSDs really work?

Measure your drives!

Aerospike Certification Tool (**ACT**)
http://github.com/aerospike/act

Transactional database workload

Reads: 1.5KB
(can't batch / cache reads, random)

Writes: 128K blocks
(log based layout)
(plus defragmentation)

Turn up the load until latency is over required SLA

"Quit feeding him so many bananas! He's our biggest customer, what if he falls?"

Micron P320h – ACT results

[root@144.bm-general.dev.nym2 act]#
latency_calc/act_latency.py -l
actconfig_micron_75x_1d_rssdb_20130503232823.out
trans device %>(ms) %>(ms)

hour	1	8	64	1	8	64
1	0.17	0.00	0.00	0.03	0.00	0.00
2	0.17	0.00	0.00	0.03	0.00	0.00
3	0.18	0.00	0.00	0.03	0.00	0.00
4	0.18	0.00	0.00	0.03	0.00	0.00
5	0.18	0.00	0.00	0.03	0.00	0.00
6	0.19	0.00	0.00	0.04	0.00	0.00

150K read IOPS @ 1.5K
225MB writes @ 128K
225MB reads @ 128K

$8/GB

Test data – the next generation

6K reads per second, 9MB/sec write load

	> 1 ms	> 8 ms	> 64 ms
Intel s3700, 20% OP - 6k iops	1.6	0	0 ($3/GB)
Intel s3700, 20% OP - 12k iops	5.4	0	0
Intel s3700, 20% OP - 24k iops	12.29	0	0
Intel s3700, NO OP - 24k iops	15.33	0	0
FusionIO Iodrive 2 – 6k iops	2.63	0.01	0 ($8/GB)
FusionIO iodrive 2 – 12k iops	7.32	0.1	0

Test data – the previous generation

2K reads per second, 3MB/sec write load

	> 1 ms	> 8 ms	> 64 ms
Intel X25-M + w/No OP (160G):	17.9%	0.6%	0.4%
Intel X25-M + OP (126G):	3.4%	0.1%	0.08%
OCZ Deneva 2 SLC + OP (95G):	0.9%	0.08%	0%
Samsung SS805 (100G):	2.0%	0.09%	0%
Intel 710 + OP (158G):	4.0%	0.01%	0%
Intel 320 + OP (126G):	5.6%	0%	0%
OCZ Vertex 2 + OP (190G):	6.3%	0.5%	0.01%
SMART XceedIOPS + OP (158G):	5.4%	0.4%	0%
Intel 510 + OP (95G):	6.2%	4.0%	0.03%
Micron P300 + OP (79GB):	1.3%	1.0%	0.7%

Test data – the previous generation

6K reads per second, 18MB/sec write load

	> 1 ms	> 8 ms	> 64 ms
OCZ Deneva 2 SLC + OP (95G):	3.2%	0.4%	0%
Samsung SS805 (100G):	10.1%	0.8%	0.02%
Intel 320 + OP (126G):	22.0%	0.3%	0.03%
OCZ Deneva 2 MLC (Sync)	8.8%	0.6%	0.06%
OCZ Vertex 2 + OP (190G):	27.6%	4.6%	0.4%
SMART XceedIOPS + OP (158G):	24.5%	5.4%	1.0%

Aerospike for HA and scale

Proven in Production

> *"Aerospike has operated without interruptions and easily scaled to meet our performance demands."*
> – Mike Nolet, CTO, AppNexus

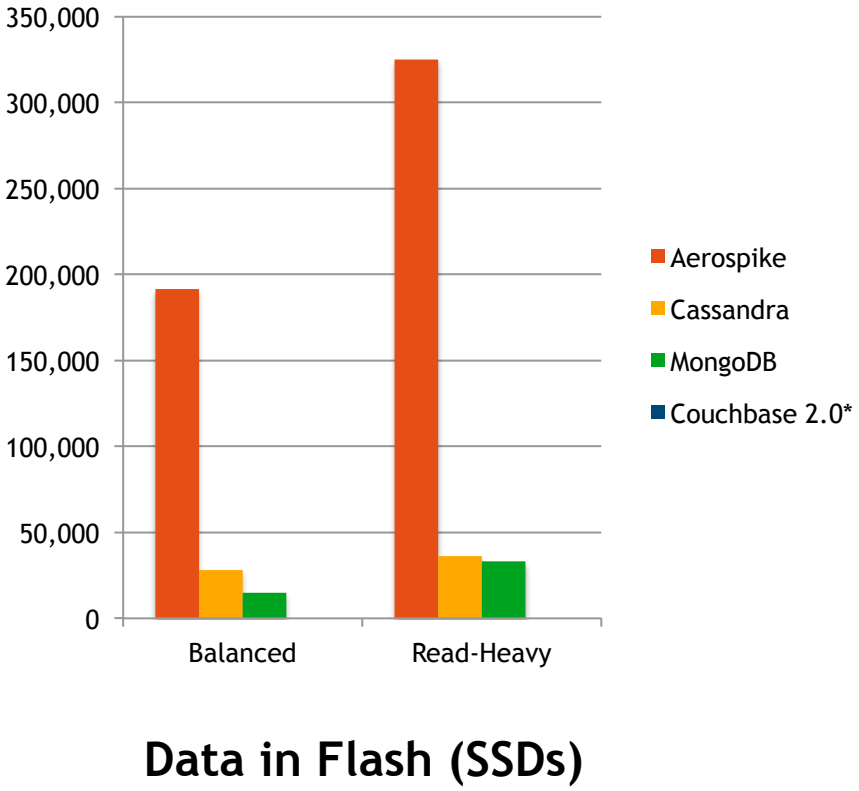
- AppNexus - #2 RTB after Google
  - 27 Billion auctions per day
  - 600+ QPS
  - Aerospike servers in 6 clusters in 3 data centers
- Chango – #2 Search after Google
  - Sees more Searches than Yahoo! + bing
  - Data on 300 Million users
- TradeDesk – first Ad Exchange
  - Facebook Exchange partner
  - FBX serves 25% of Ads on the Internet
  - 1200% growth in 2012

Speed at Scale

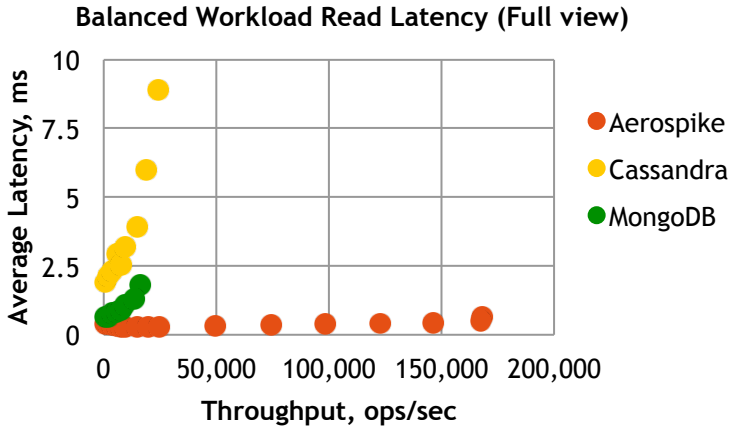
"Aerospike was the dominant performer, showing durable, replicated behavior 5-10 times faster than what others could achieve".

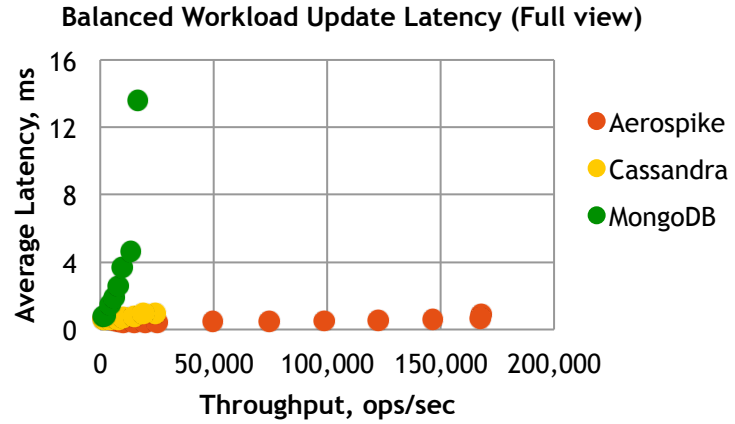
thumbtack TECHNOLOGY

- High Throughput, Low Latency



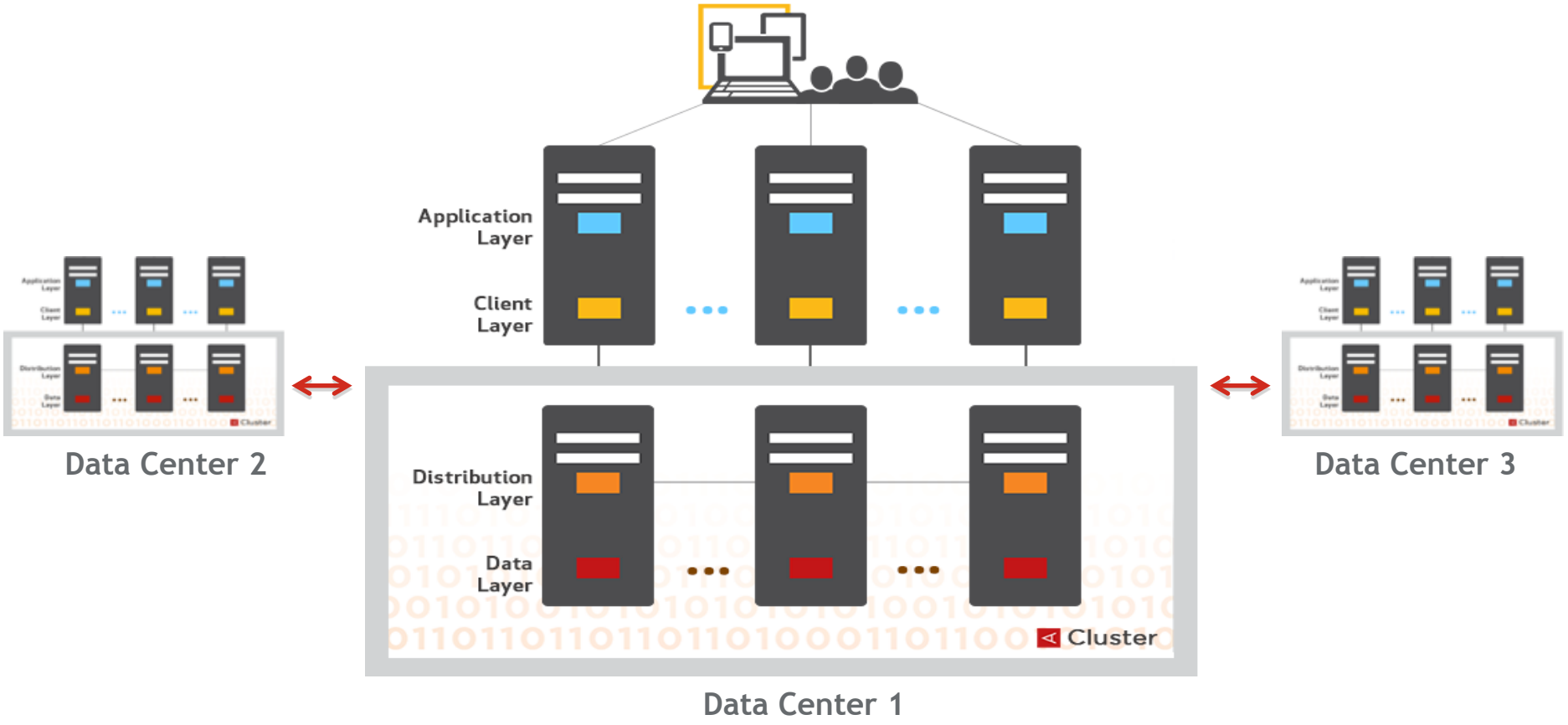
Data in Flash (SSDs)





*We were forced to exclude Couchbase ..since when run with either disk or replica durability on it was unable to complete the test."
- Thumbtack Technology

Shared-Nothing Architecture



Every cluster node is Identical and handles both transactions and long running tasks

Replication supported with immediate consistency

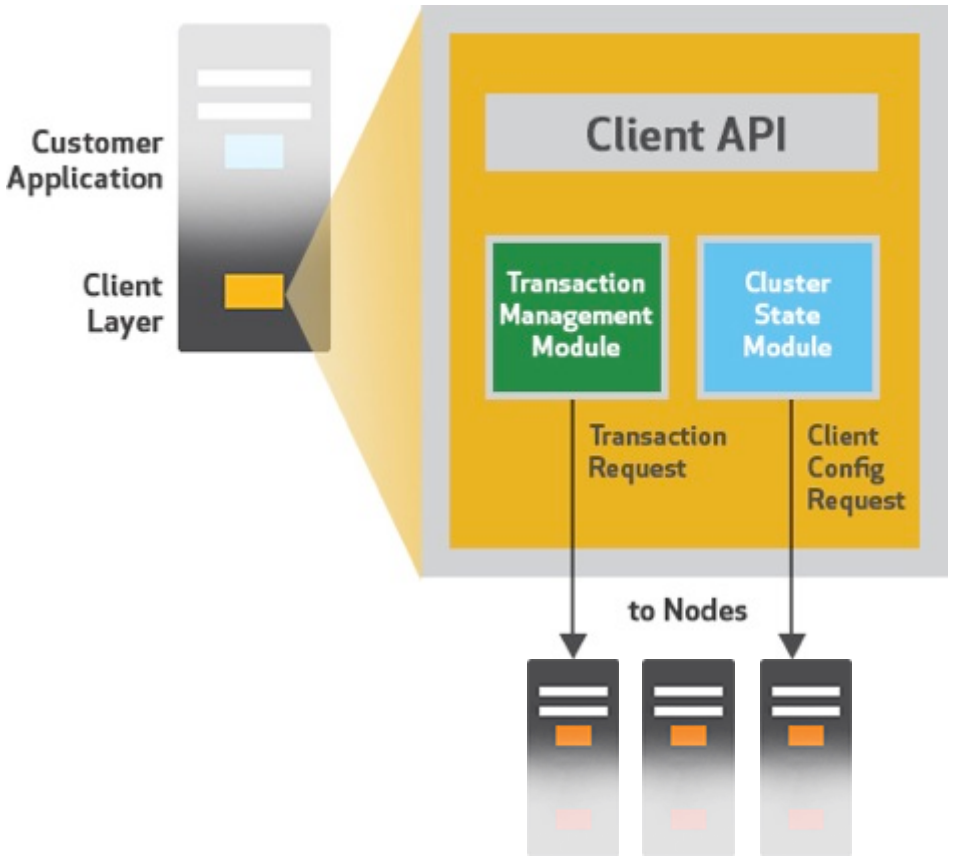
Intelligent Client

Shields Applications from the Complexity of the Cluster

- Implements Aerospike API
- Optimistic row locking
- Optimized binary protocol

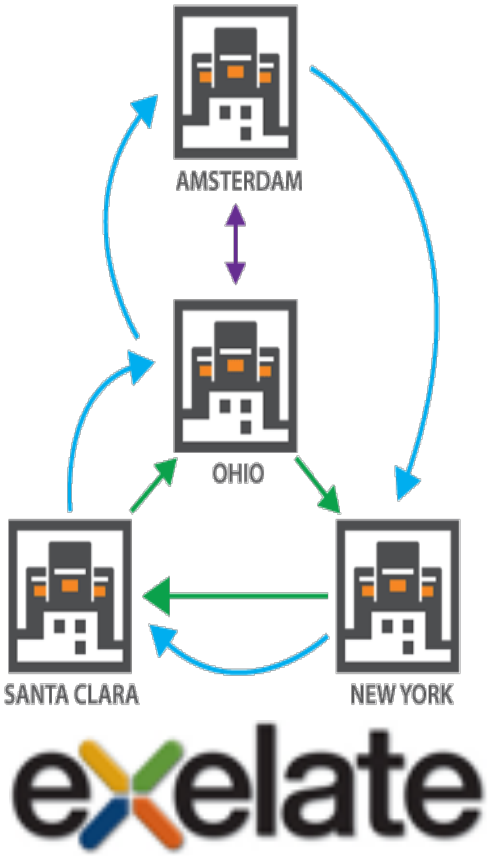
- Cluster tracking
  - Learns about cluster changes, partition map
  - Gossip protocol

- Transaction semantics
  - Global transaction ID
  - Retransmit and timeout



Aerospike Cross Data Center Replication™ (XDR)

- XDR configured per namespace
- Any combination of star (master/slave) and ring (master/master) patterns
  - Conflict resolution is via timestamps or multiple versions
- Asynch Replication
  1. Transaction journal on partition master and replica
  2. XDR process writes batches to destination
  3. Transmission state shared with source replica
  4. Retransmission in case of network fault
  5. When data arrives back at originating cluster, transaction ID matching prevents subsequent application and forwarding
- XDR In Action: Super Storm Sandy
  - NYC Data Center loses power, service continues from other data centers, clusters synchronize when NYC comes back online



Monitoring

- Graphical and text based
- Cluster Health
  1. RAM and DISK usage
  2. Alerts
  3. Node Selection
  4. Throughput
  5. Nodes
  6. Namespaces
  7. XDR

How to get Aerospike?

Free Community Edition

- For developers looking for speed and stability and transparently scale as they grow
  - All features for
    - 2 nodes, 100GB
    - 1 cluster
    - 1 datacenter
  - Community support

Enterprise Edition

- For mission critical apps needing to scale right from the start
  - Unlimited number of nodes, clusters, data centers
  - Cross data center replication
  - Premium 24x7 support
  - Priced by TBs of unique data (not replicas)

AEROSPIKE

Questions