What Enterprises Can Learn from Real-time Bidding
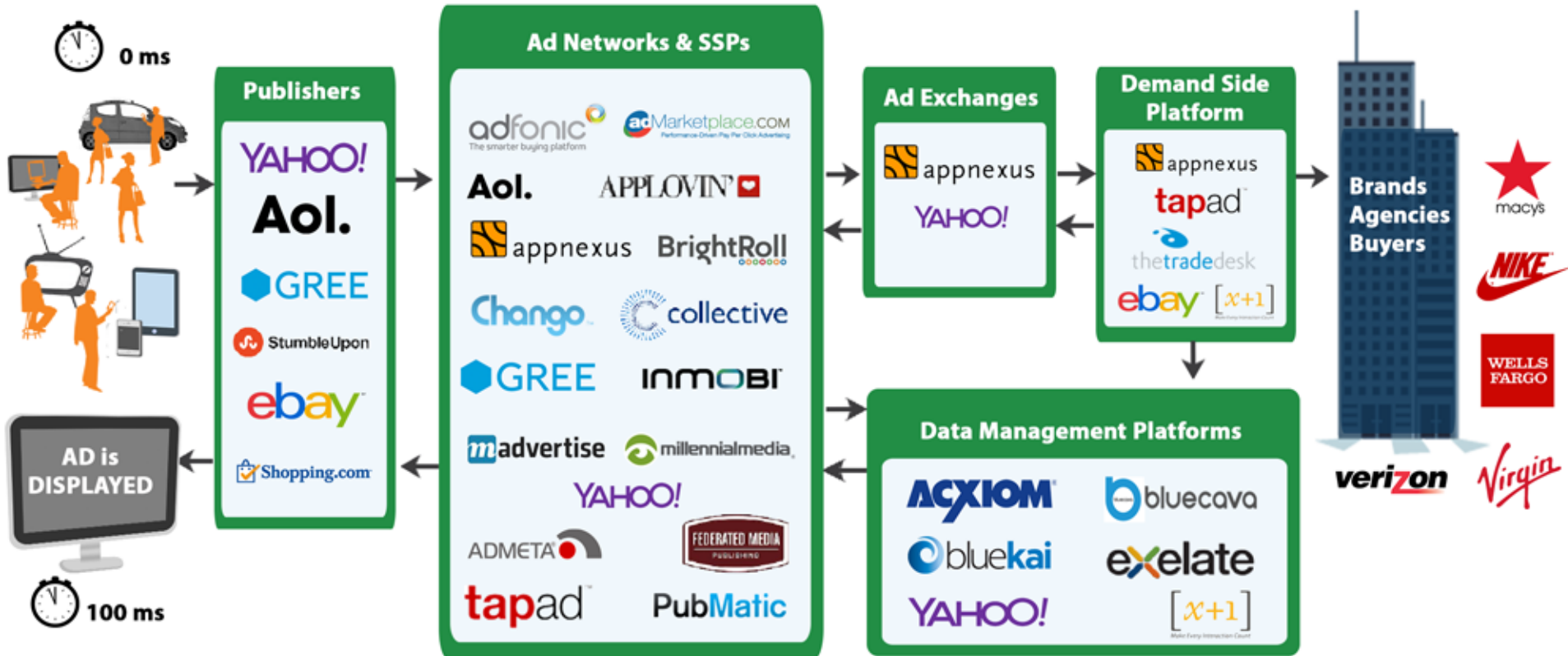
How, and why, to achieve Operational Big Data

Brian Bulkowski
CTO and co-founder
Aerospike

REQUIREMENTS FOR INTERNET ENTERPRISES

time pressure

North American RTB speeds & feeds

- 1 to 6 billion cookies tracked
  - Some companies track 200M, some track 20B

- Each bidder has their own data pool
  - Data is your weapon
  - Recent searches, behavior, IP addresses
  - Audience clusters (K-cluster, K-means) from offline Hadoop

- "Remnant" from Google, Yahoo is about 0.6 million / sec
- Facebook exchange: about 0.6 million / sec
- "other" is 0.5 million / sec

Currently about 3.0M / sec in North American

Advertising requirements

- 100 millisecond or 150 millisecond ad delivery
  - De-facto standard set in 2004 by Washington Post and others

- North America is 70 to 90 milliseconds wide
  - Two or three data centers

- Auction is limited to 30 milliseconds
  - Typically closes in 5 milliseconds

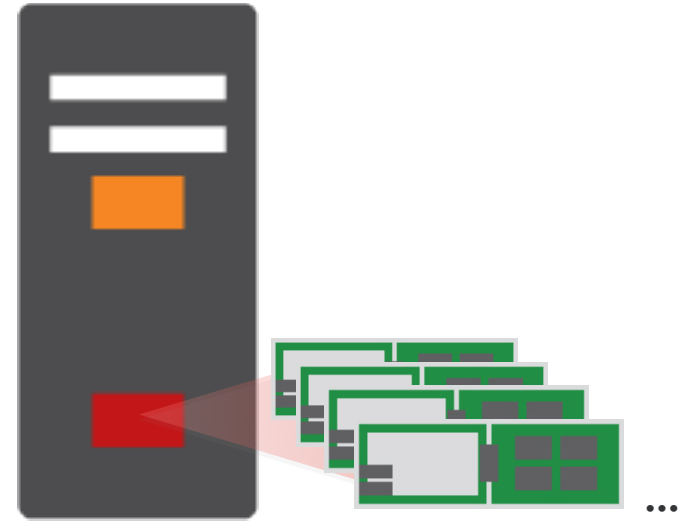- Winners have more data, better models – in 5 milliseconds
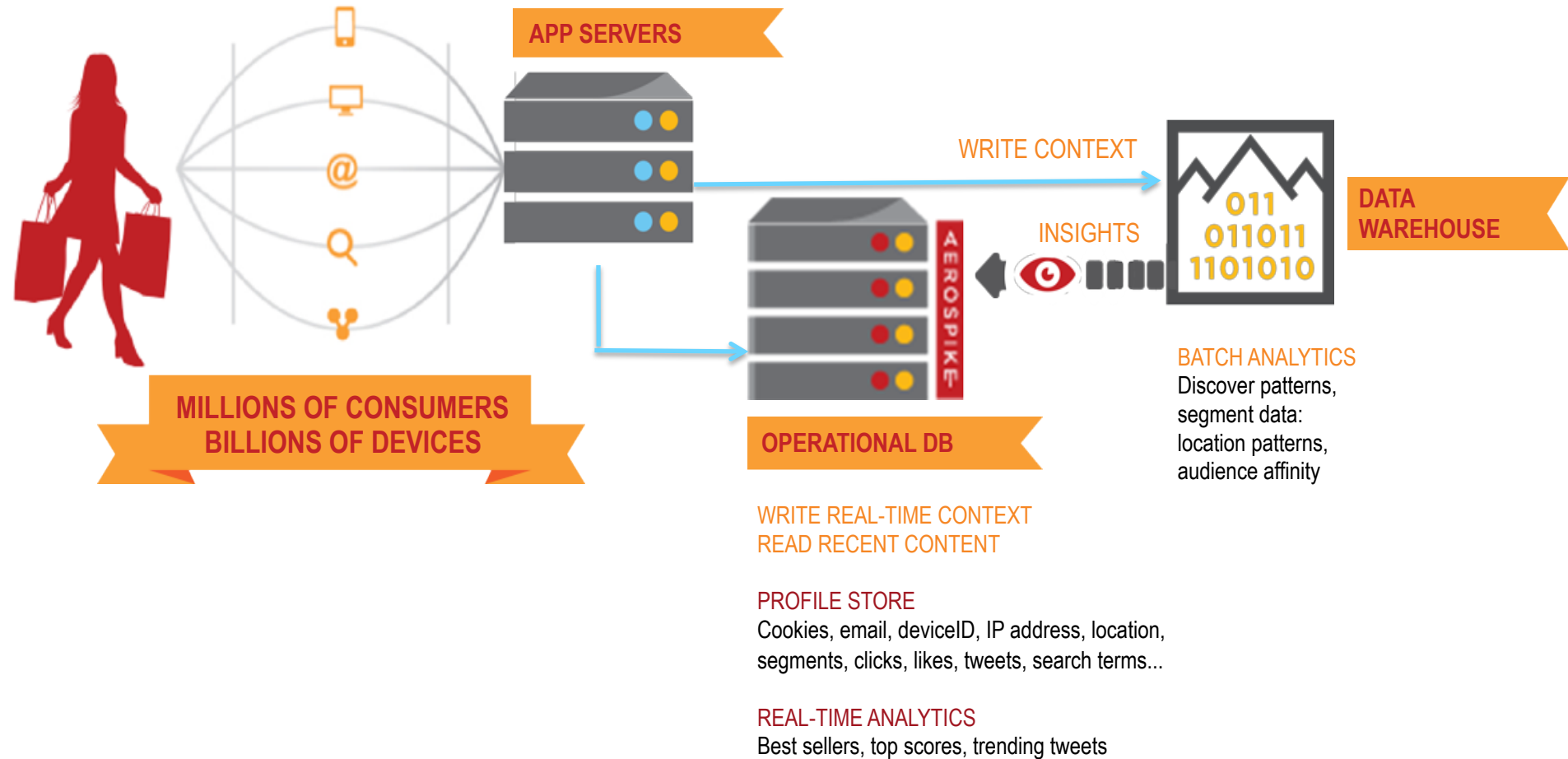
Typical Deployment

➢ Last Year
- ➢ 8 core Xeon
- ➢ 24G RAM
- ➢ 400G SSD (SATA)
- ➢ 30,000 read TPS, 20,000 write TPS
- ➢ 1.5K object size / 200M objects
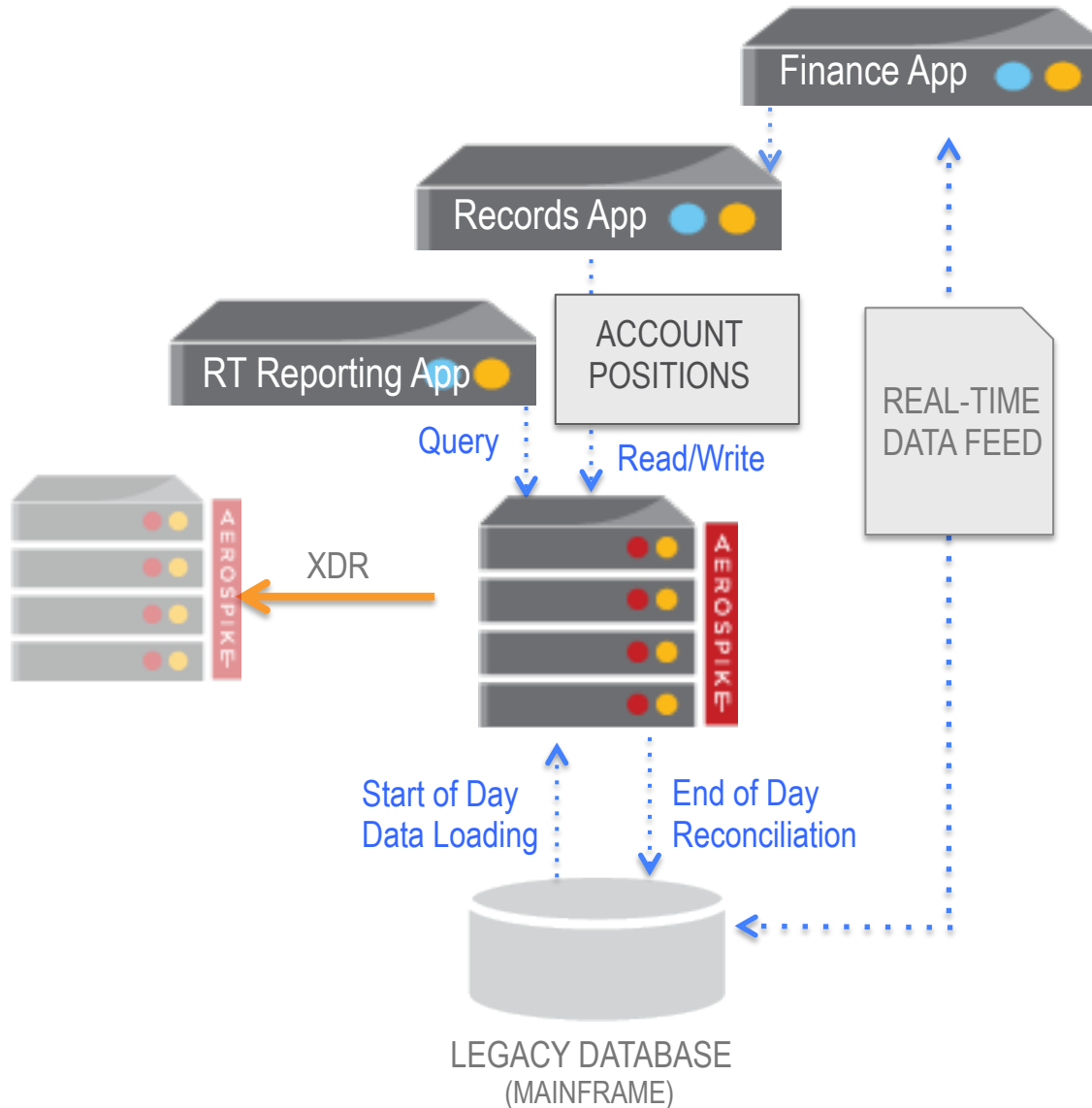- ➢ 4 to 40 node clusters

➢ This Year
- ➢ 16 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
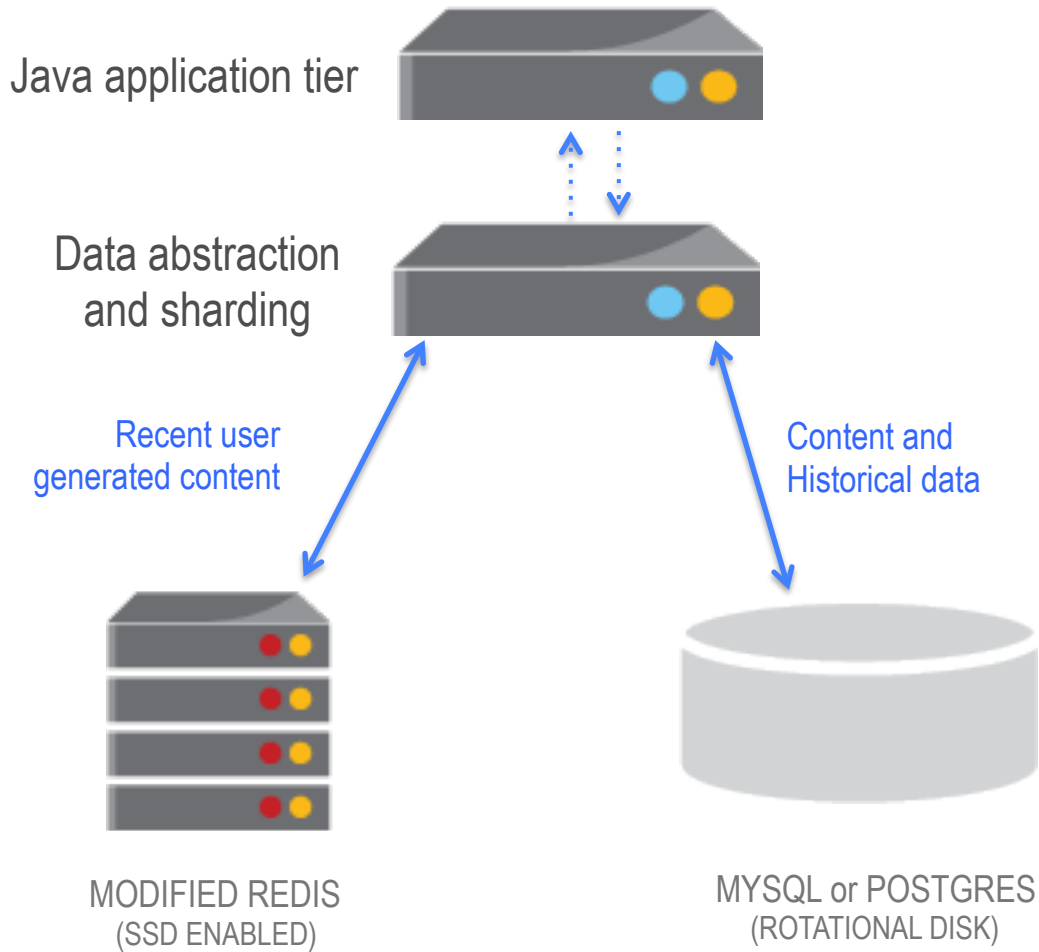- ➢ 4 to 20 node cluster

...

Advertising Technology Stack

APP SERVERS

WRITE CONTEXT

INSIGHTS

DATA WAREHOUSE

MILLIONS OF CONSUMERS
BILLIONS OF DEVICES

OPERATIONAL DB

BATCH ANALYTICS
Discover patterns,
segment data:
location patterns,
audience affinity

WRITE REAL-TIME CONTEXT
READ RECENT CONTENT

PROFILE STORE
Cookies, email, deviceID, IP address, location,
segments, clicks, likes, tweets, search terms...

REAL-TIME ANALYTICS
Best sellers, top scores, trending tweets

Finance App

Records App

ACCOUNT POSITIONS

RT Reporting App

Query

Read/Write

REAL-TIME DATA FEED

XDR

Start of Day Data Loading

End of Day Reconciliation

LEGACY DATABASE (MAINFRAME)

10M+ user records

Primary key access

1M+ TPS planned

Social Media

Java application tier

Data abstraction
and sharding

Recent user
generated content

Content and
Historical data

MODIFIED REDIS
(SSD ENABLED)

MYSQL or POSTGRES
(ROTATIONAL DISK)

新浪微博
weibo.com

Alibaba.com®

Tencent 腾讯

AEROSPIKE

Travel Portal

Travel App

SESSION MANAGEMENT

PRICING DATA

Session Data

Read Price

Store Latest Price

XDR

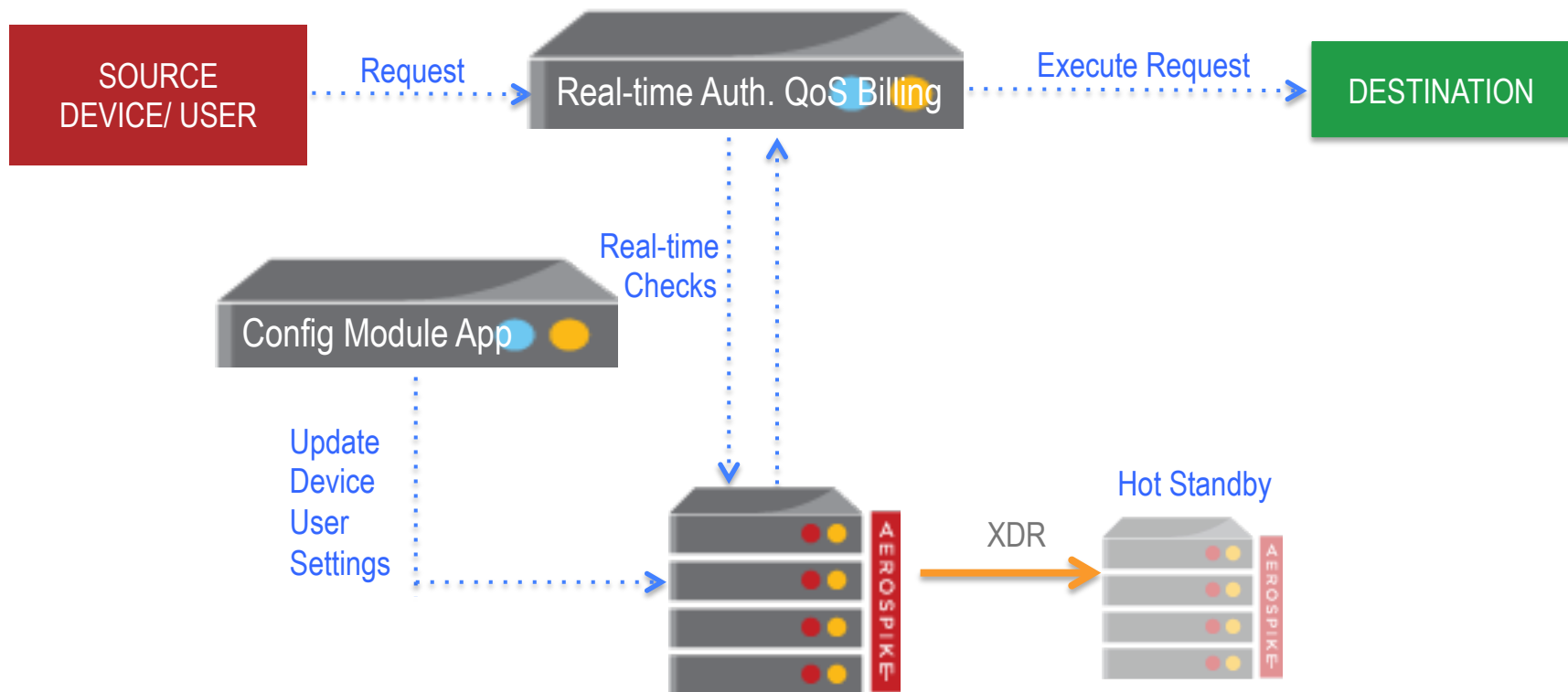Poll for Pricing Changes

PRICING DATABASE
(RATE LIMITED)
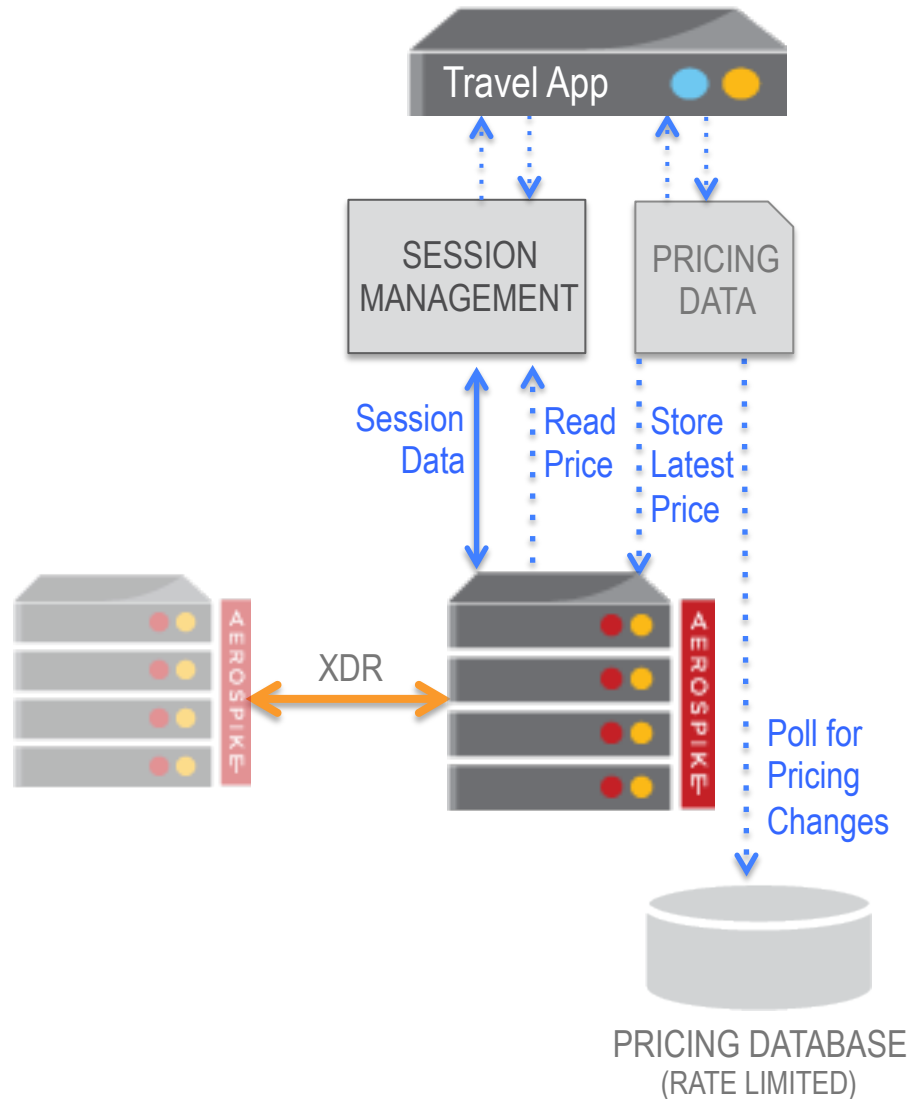
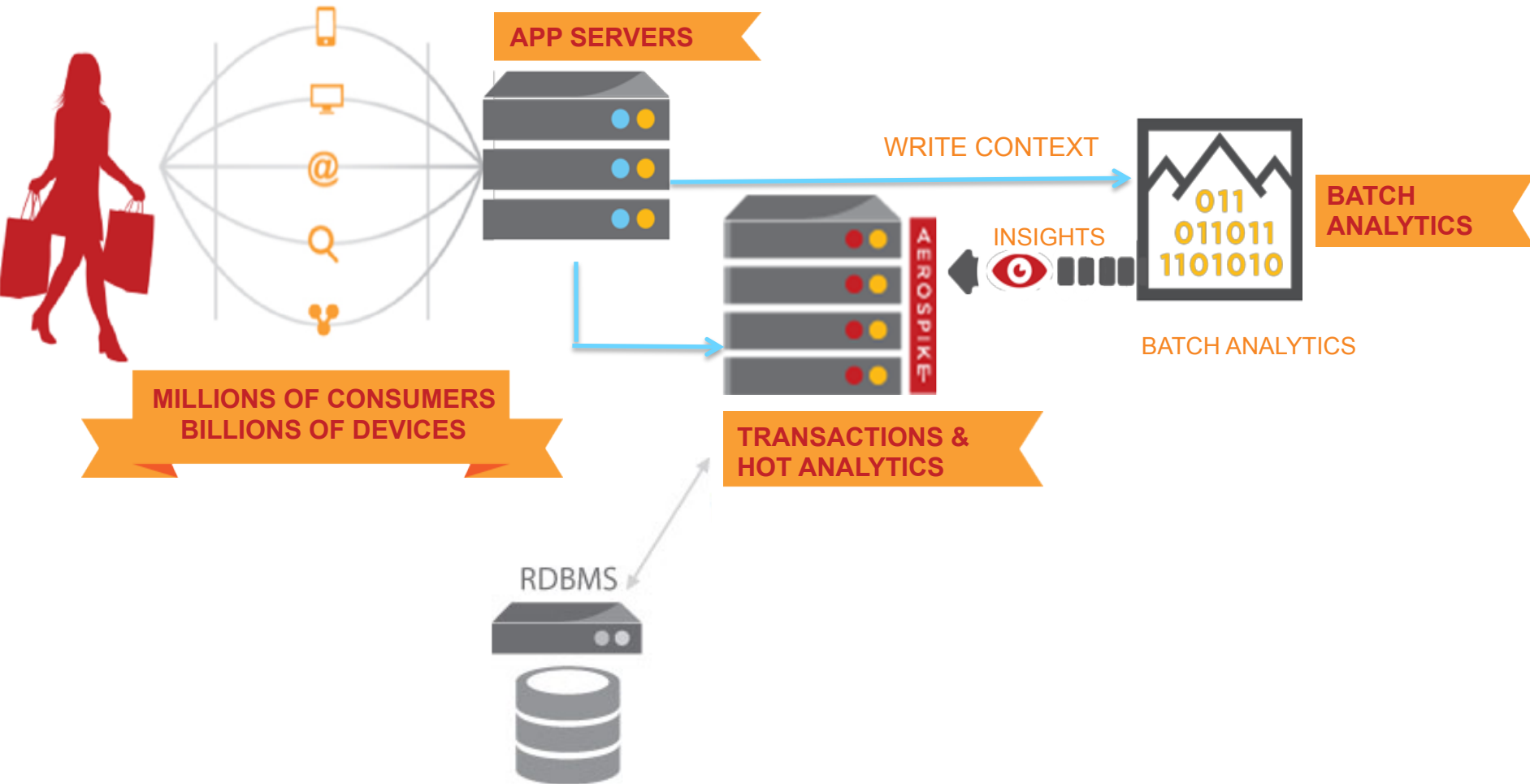Airlines forced interstate banking

Legacy mainframe technology

Multi-company reservation and pricing
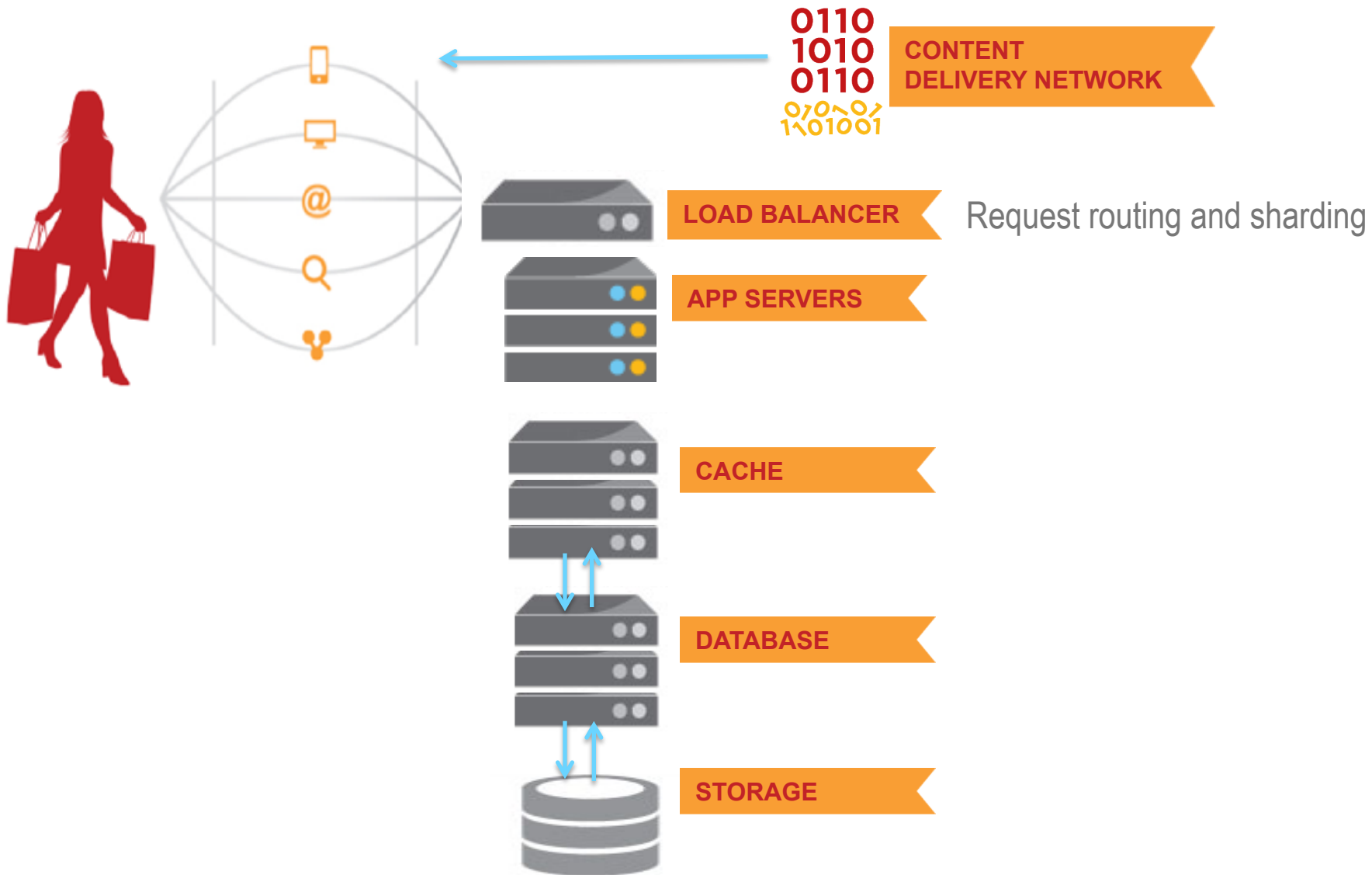
Requirement: 1M TPS allowing overhead

- In-switch Per HTTP request Billing
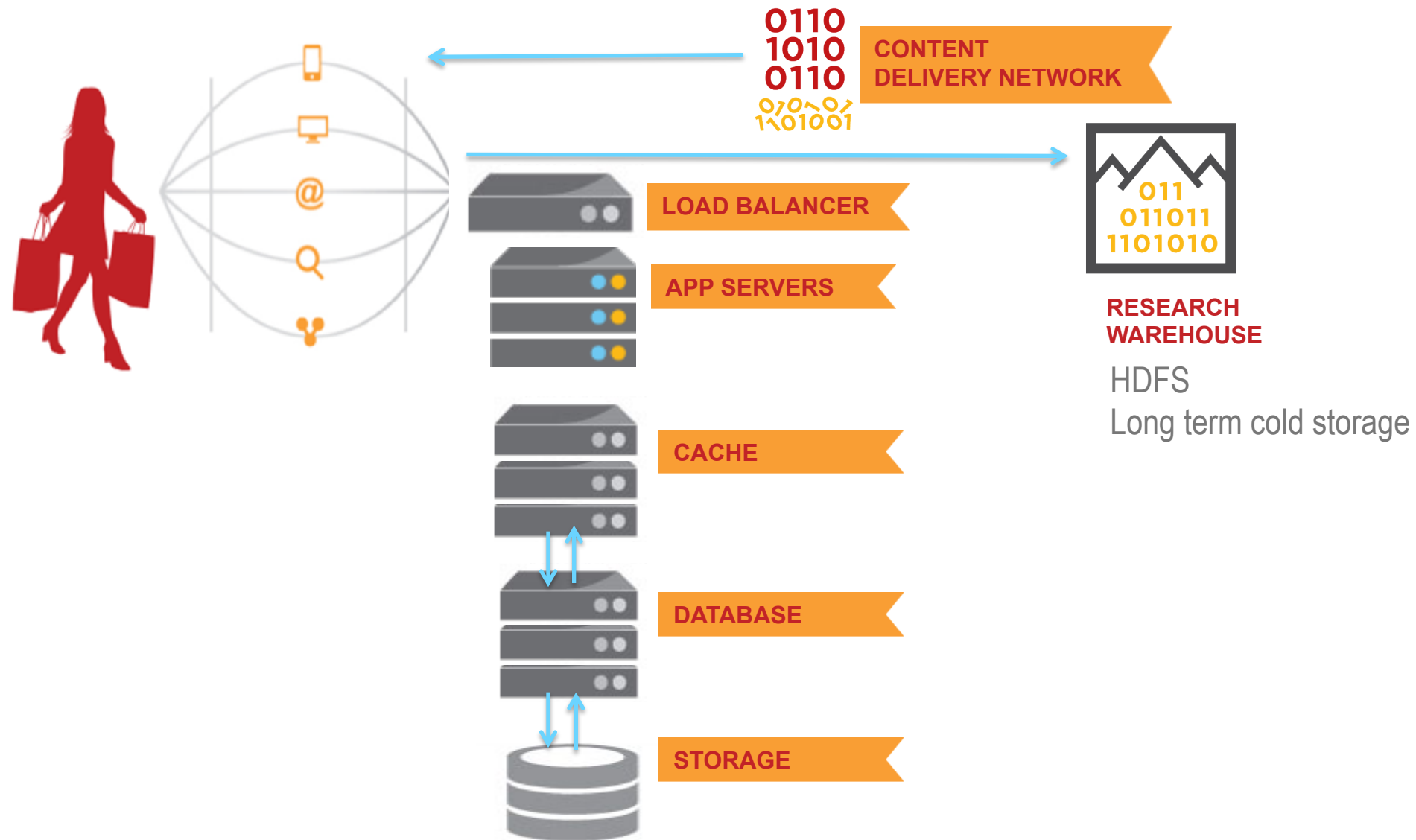 - US Telcos: 200M subscribers, 50 metros
- In-memory use case

SOURCE DEVICE/ USER

Request

Real-time Auth. QoS Billing

Execute Request

DESTINATION

Real-time Checks

Config Module App

Update Device User Settings

XDR

Hot Standby

ΛEROSPIKE

The New Architecture

APP SERVERS

WRITE CONTEXT

BATCH ANALYTICS

INSIGHTS

BATCH ANALYTICS

011
011011
1101010

MILLIONS OF CONSUMERS
BILLIONS OF DEVICES

TRANSACTIONS &
HOT ANALYTICS

RDBMS

AEROSPIKE

CONTENT DELIVERY NETWORK

LOAD BALANCER — Request routing and sharding

APP SERVERS

CACHE

DATABASE

STORAGE

Early Big Data Architecture

CONTENT DELIVERY NETWORK

LOAD BALANCER

APP SERVERS

CACHE

DATABASE

STORAGE

RESEARCH WAREHOUSE

HDFS
Long term cold storage

AEROSPIKE

Modern Scale Out Architecture

0110
1010
0110
0ı0110ı
ıı01001

CONTENT DELIVERY NETWORK

LOAD BALANCER

Simple stateless

APP SERVERS

Fast stateless

IN-MEMORY NoSQL

011
011011
1101010

RESEARCH WAREHOUSE

Long term cold storage

Direct attach Flash DB

- Reliability
- Durability
- High Availability
- Volume Management

AEROSPIKE

Modern Scale Out Architecture

CONTENT DELIVERY NETWORK

LOAD BALANCER

Simple stateless

APP SERVERS

Fast stateless

IN-MEMORY NoSQL

RESEARCH WAREHOUSE

Long term cold storage

HDFS BASED

Build a data layer

Focus on Key Value

RECOMMENDATIONS

Use open source
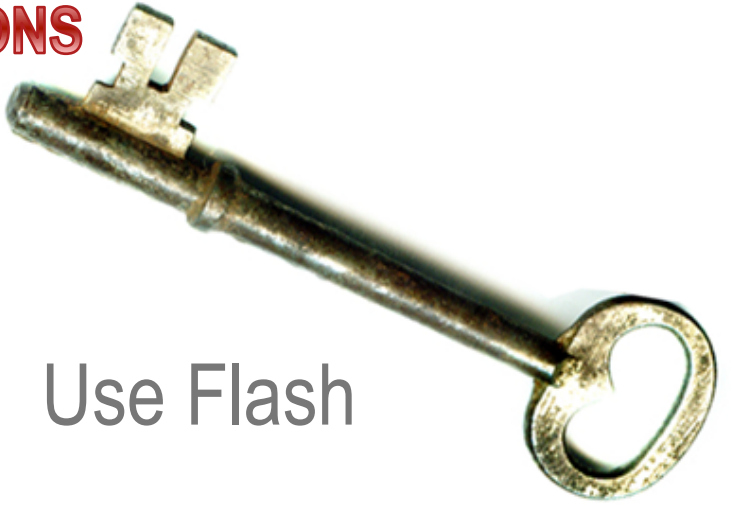
Use Flash

Use In-memory NoSQL

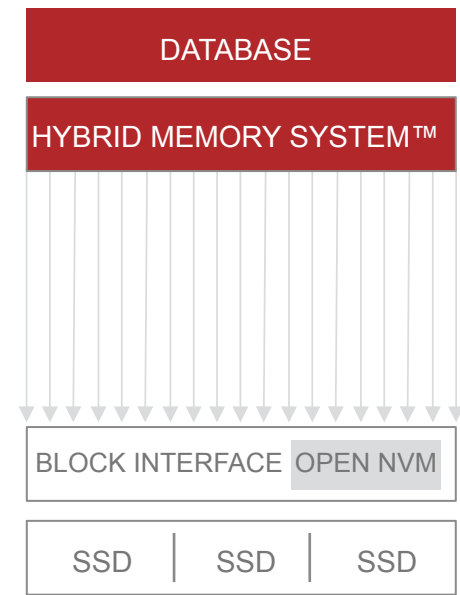The Power of Flash Storage

FLASH OPTIMIZED HIGH PERFORMANCE

- Direct device access
- Large Block Writes
- Indexes in DRAM
- Highly Parallelized
- Log-structured FS "copy-on-write"
- Fast restart with shared memory

Ask me. I'll look up the answer and then tell it to you.
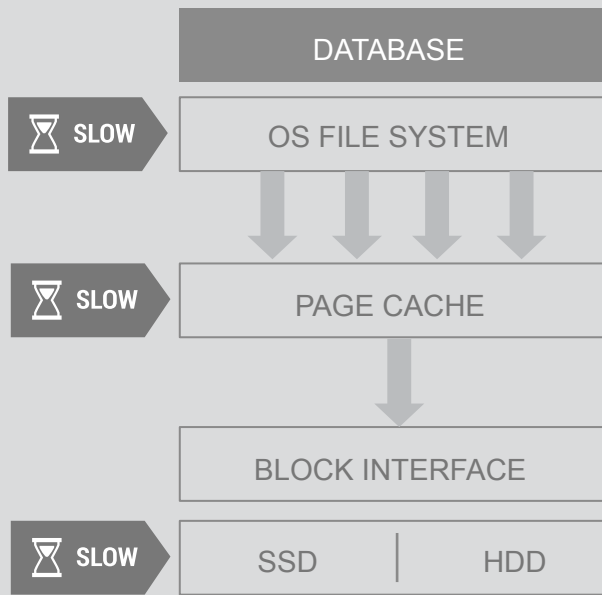
Ask me and I'll tell you the answer.

DATABASE
OS FILE SYSTEM

SLOW

PAGE CACHE

SLOW

BLOCK INTERFACE

SSD	HDD

SLOW

DATABASE
HYBRID MEMORY SYSTEM™

BLOCK INTERFACE	OPEN NVM

SSD	SSD	SSD

AEROSPIKE

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?"

AEROSPIKE

- Know your Flash
  - ACT benchmark http://github.com/aerospike/act
  - Read-write benchmark results back to 2011

- All clouds support flash now
  - New EC2 instances
  - Google Compute
  - Internap, Softlayer, GoGrid…

- Write durability usually not a problem with modern flash
  - Durability is high (5 "drive writes per day" for 5 years, etc)
  - Read performance suffers under write load anyway
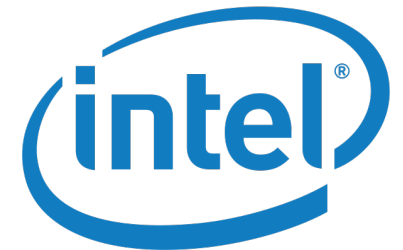
Aerospike's Flash Experience

- Densities increasing
    - 100G 2 years ago → 800G today
    - SATA vs PCI-E
    - Appliances: 50T per 1U this year

- Prices still dropping: perhaps $1/G next year

- Intel P3700 results
    - 250K per device @ $2.5 / G
    - Old standard: Micron P320h 500K @ $8 / G

- "Wide SATA"
    - 20 SATA drives
    - LSI "pass through mode"
    - 250K+ per server
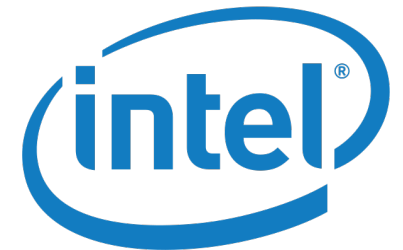
10T example
(a reasonable project budget)

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**

Wikibon

"…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

Aerospike

In-memory reliable Database

Flash and DRAM

AEROSPIKE
SHARED-NOTHING SYSTEM
100% DATA AVAILABILITY

- Every node in a cluster is **identical**, handles both transactions and long running tasks

- Data is replicated **synchronously** with immediate consistency within the cluster

- Data is replicated **asynchronously** across data centers

OHIO Data Center
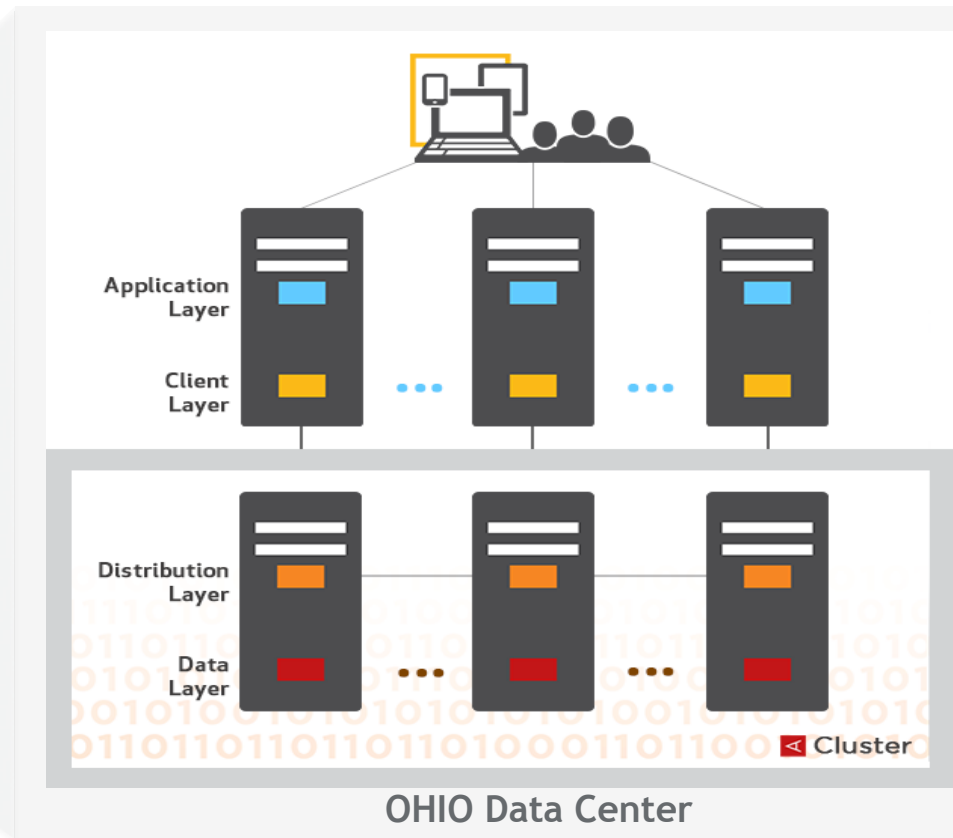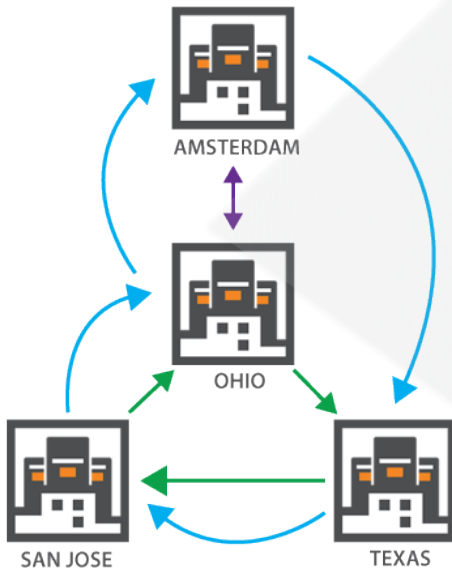
Aerospike: the trusted In-Memory NoSQL

Performance
- Over ten trillion transactions per month
- 99% of transactions < 2 ms
- 150K TPS per server

Scalability
- Billions of Internet users
- Clustered Software
- Maintenance without downtime
- Scale up & scale out

Reliability
- 50 customers; zero down-time
- Immediate Consistency
- Rapid Failover; Data Center Replication

Price/Performance
- Makes impossible projects affordable
- Flash-optimized
- 1/10 the servers required