



Fast Online Access to Massive Offline Data

Software Engineering Conference in Russia

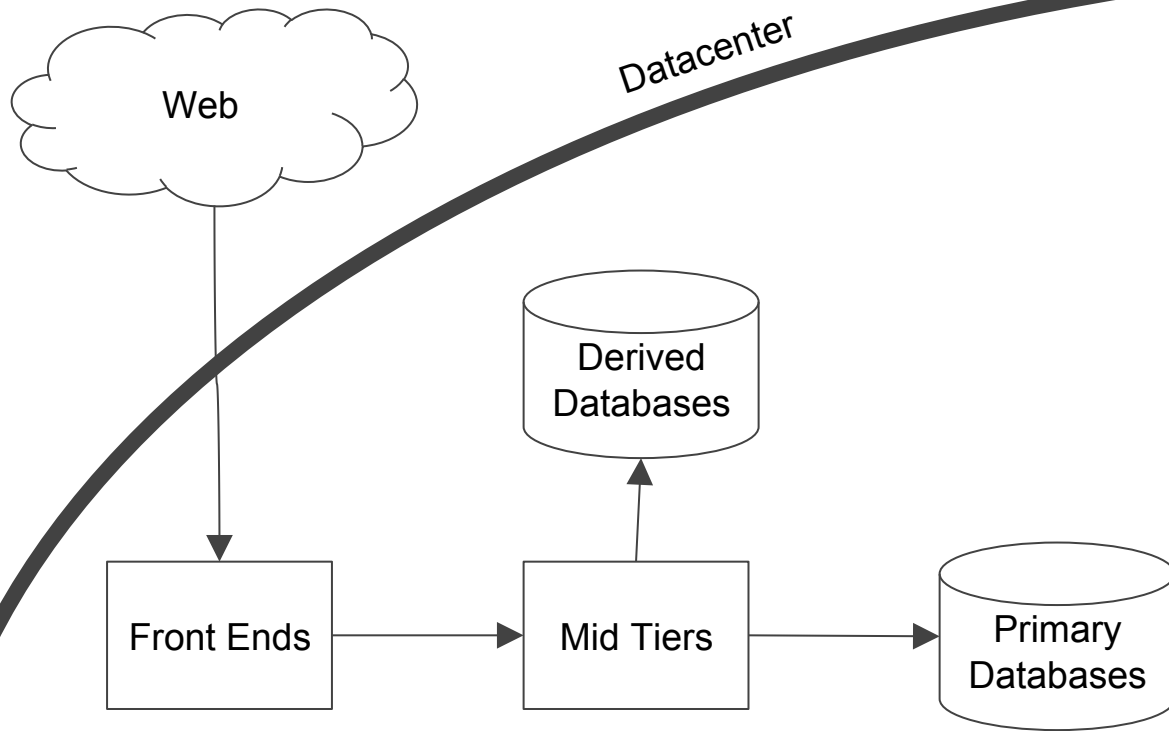
28 October 2016

By Felix GV

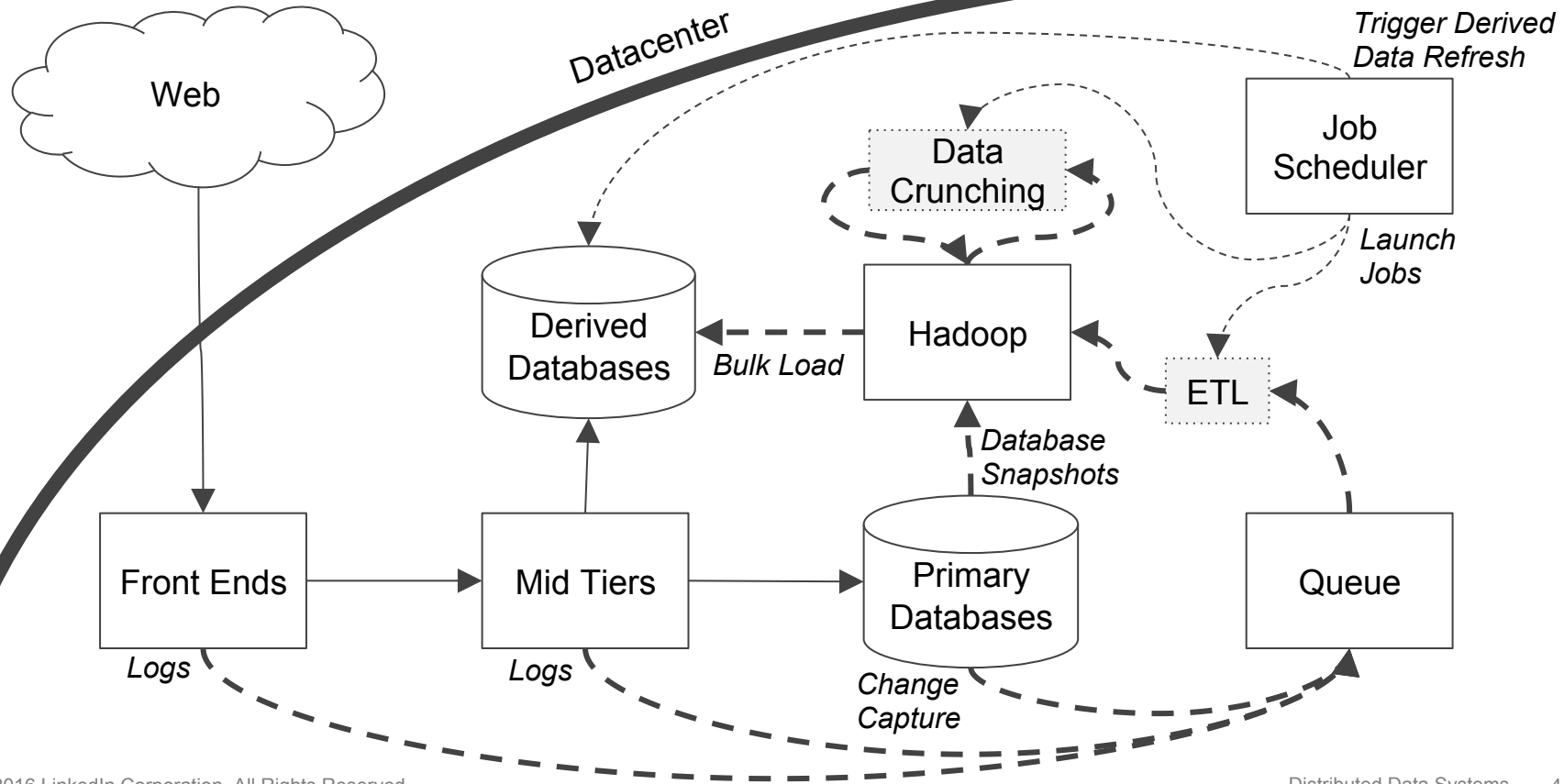
Agenda

- Introduction
 - High Level Web Architecture
 - What is Primary Data and Derived Data?
 - What is Voldemort?
- Recent Improvements to Voldemort RO
 - Cross-DC Bandwidth
 - Multi-tenancy
 - Performance
- How to Get Started?

High Level Web Architecture



High Level Web Architecture



What is Primary Data and Derived Data?

Primary Data is “Source of Truth” data:

- Users’ profiles, private messages, etc.
- Typically requires strong consistency & read-your-write semantics

Derived Data comes from crunching primary data:

- “People You May Know”, “Jobs You May Be Interested In”, etc.
- Aggregation, Joins, Machine learning
- Typically generated offline, in Hadoop

How can we serve derived data back to online apps?

What is Voldemort?

Voldemort is a Distributed Key Value Store with two modes:

- Read-Write
 - Random access writes, tunable consistency
- Read-Only
 - Bulk-loaded from Hadoop

Can be single DC, or globally distributed.

Pluggable:

- Storage Engine (BDB-JE, RocksDB, Read-Only format, etc.)
- Serialization (Avro, JSON, Protobuf, Thrift, raw bytes, etc.)

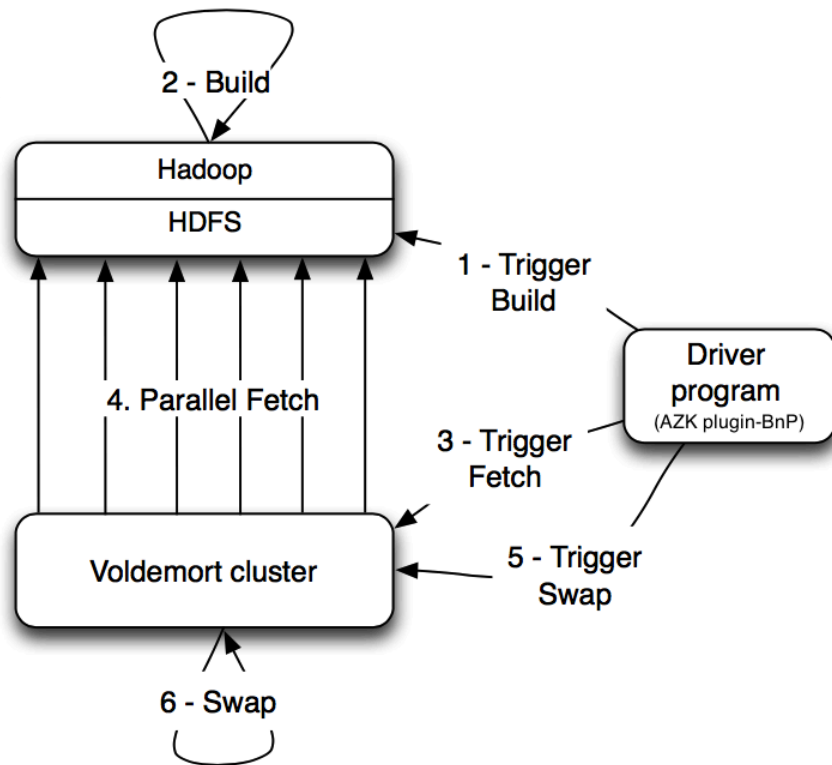
What is Voldemort Read-Only?

Voldemort RO Servers:

- Fetch entire datasets from HDFS
- Serve key-value read requests

Build and Push job:

- Validates store & schema
- Triggers MR job
 - MR job partitions data
- Triggers server fetches
- Swaps new dataset



A close-up shot of a pale, bald man with wide, light-colored eyes and an open mouth, suggesting a state of shock or fear. The lighting is dim and blue-toned, creating a somber and intense atmosphere. The man's skin appears wrinkled and aged.

For Voldemort to die, all pieces must die.

Recent Improvements to Voldemort RO

- **Cross-DC Bandwidth**
 - **Block-level Compression**
 - **Throttling**
- **Multi-tenancy**
 - Nuage Integration
 - Storage Space Quotas
- **Performance**
 - Build and Push Performance
 - Client Latency

Voldemort Cross-DC Bandwidth

Voldemort RO among largest users of cross-DC bandwidth at LinkedIn

- >100 TB / day ingested across the WAN, every day

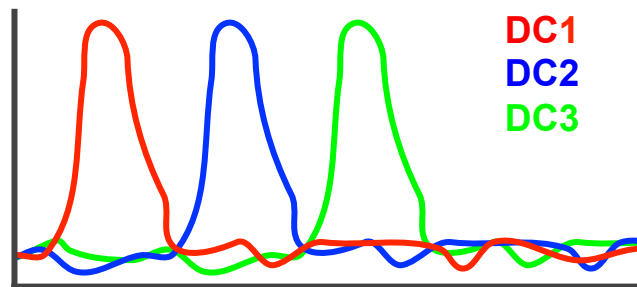
We added block-level compression:

- Output of BnP's MR job is compressed (GZIP)
- Voldemort servers decompress on the fly
 - CPU cost of decompression deemed negligible
- Completely transparent to store owners
- ~18% reduction in dataset size

Voldemort Cross-DC Bandwidth

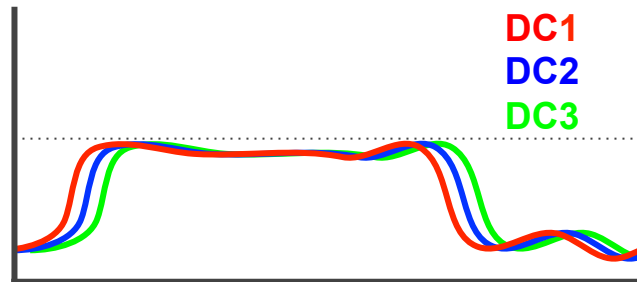
Inbound bandwidth used to look like this:

- Very spiky
- Each DC was fetching sequentially
- Love/hate relationship with Net Ops



Now, with parallel fetching and throttling:

- No more spikes
- DCs are fetching in parallel
- Best friends with Net Ops



Recent Improvements to Voldemort RO

- Cross-DC Bandwidth
 - Block-level Compression
 - Throttling
- **Multi-tenancy**
 - **Nuage Integration**
 - **Storage Space Quotas**
- Performance
 - Build and Push Performance
 - Client Latency

Voldemort Multi-tenancy

Nuage is LinkedIn's internal (AWS-like) self-service provisioning infra

All new stores at LinkedIn are now created via Nuage

- Prevents accidental pushes to the wrong cluster
- Store creation in production is now self-service
- ~3 stores / week created in production!

Voldemort Multi-tenancy

Storage Space Quotas

- Precise storage requirement measured during build phase
- Voldemort server validates quota before starting to fetch
- The goal is to prevent unexpected growth from killing clusters

LOTS of improvements on admin commands

- Stability
- Performance

Recent Improvements to Voldemort RO

- **Cross-DC Bandwidth**
 - Block-level Compression
 - Throttling
- **Multi-tenancy**
 - Nuage Integration
 - Storage Space Quotas
- **Performance**
 - **Build and Push Performance**
 - **Client Latency**

Build and Push Performance

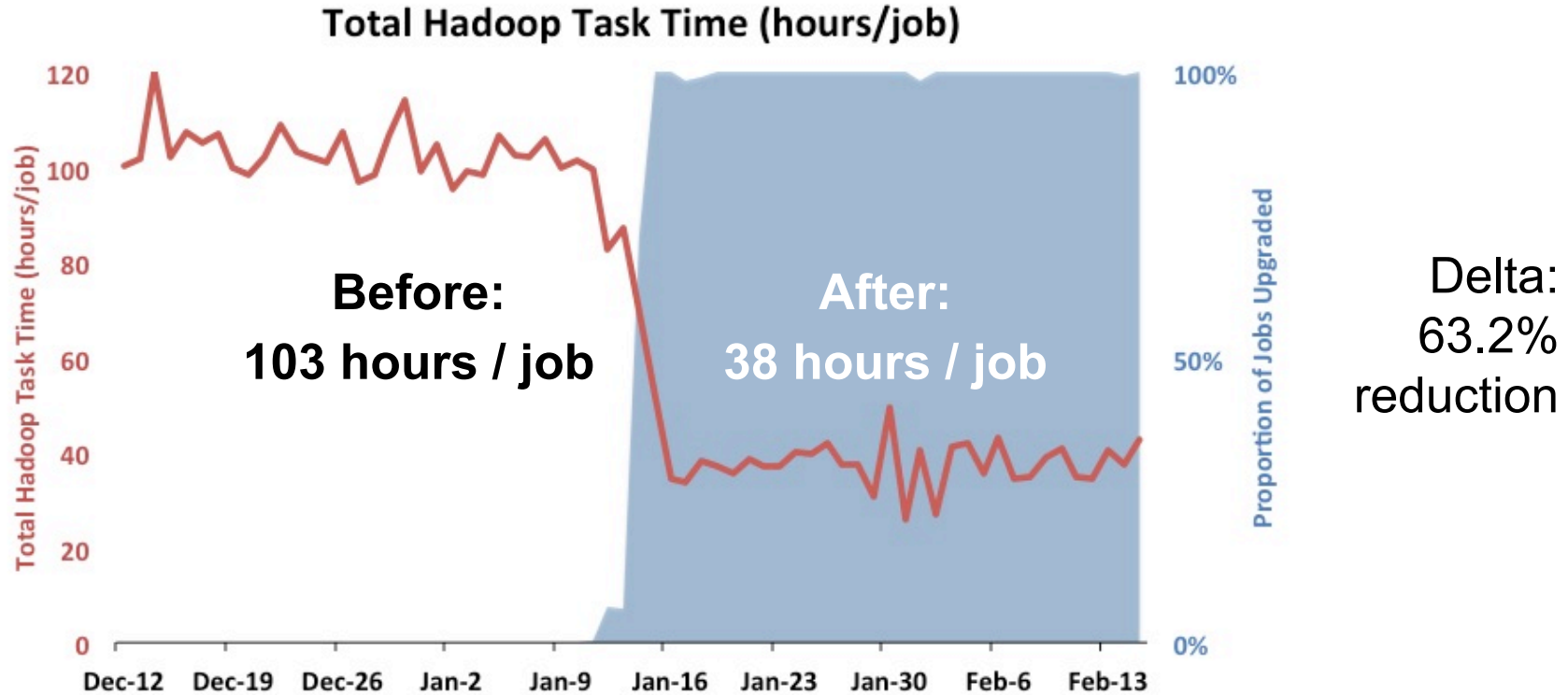
New datacenter with denser clusters required major BnP refactoring

As a side effect of that rewrite, BnP is also a lot more efficient

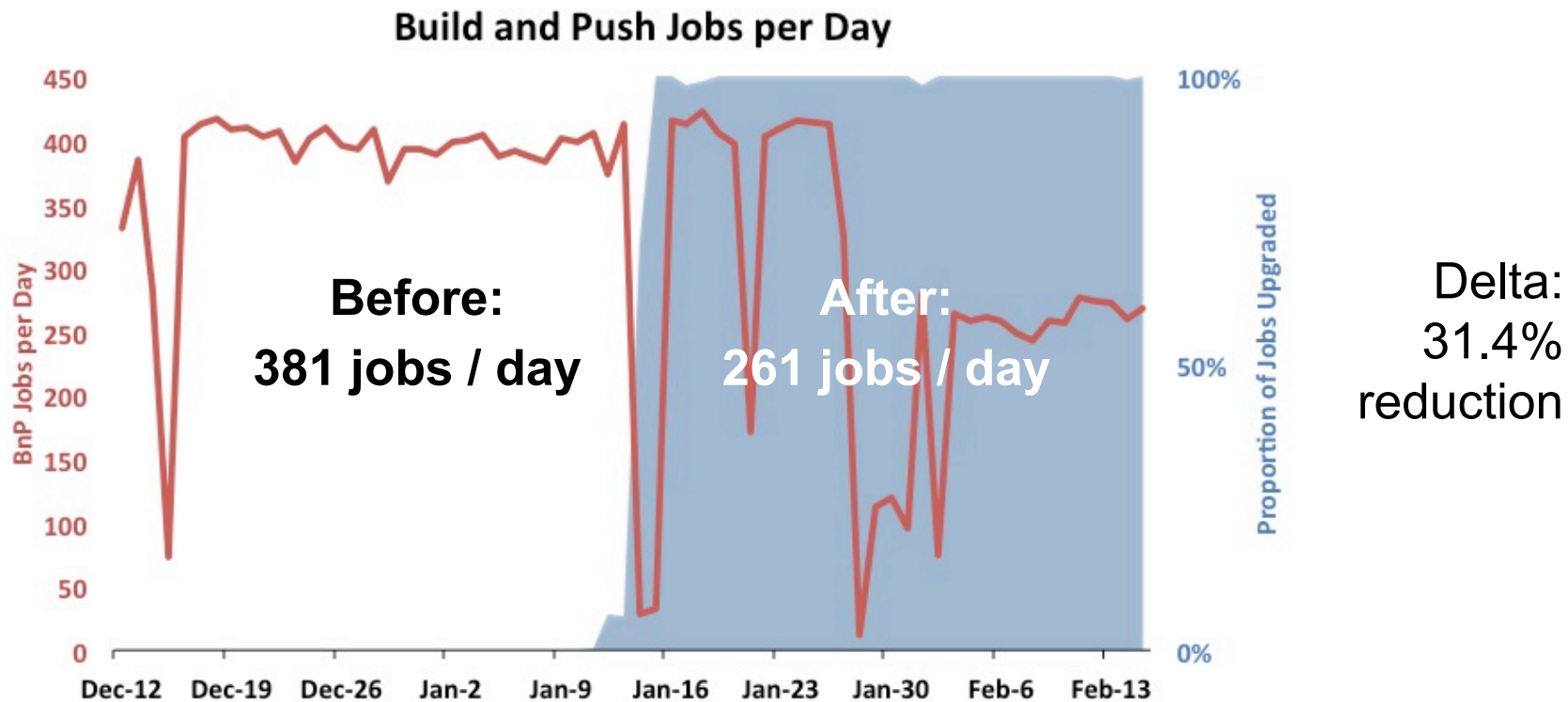
- Each partition replica is now built only once
- The following metrics are reduced (roughly by half) :
 - Number of reduce tasks
 - Amount of shuffle bandwidth
 - Amount of data written to HDFS

Leveraged new datacenter buildout to get rid of duplicate BnP jobs

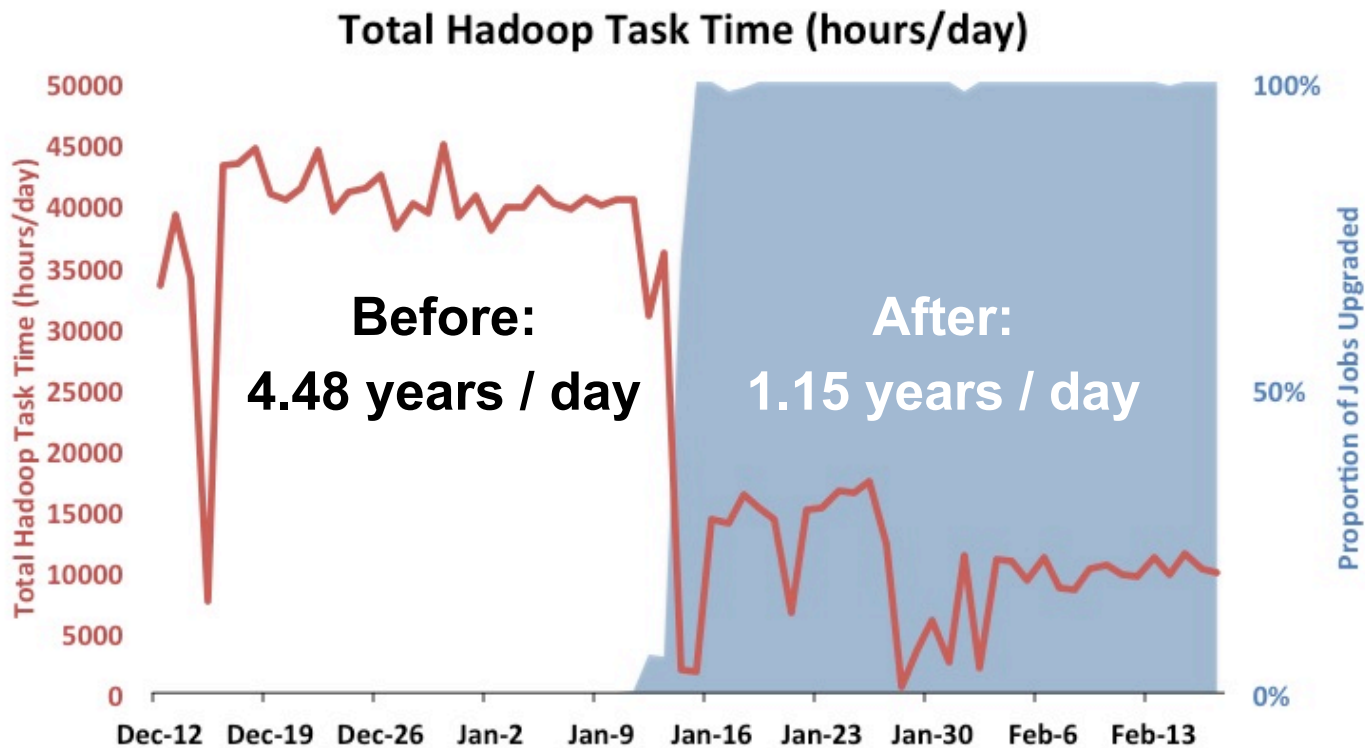
Build and Push Performance



Build and Push Performance



Build and Push Performance



Delta:
74.4%
reduction

Client Latency

Major client/server communication rewrite:

- Client and server's memory footprint reduced by half
- Client garbage collection reduced by 80%
- Failure detection is more accurate
- Solved the stability issues of our highest throughput clients

Major effort to upgrade all clients to the latest version.

Client Latency

Set up:

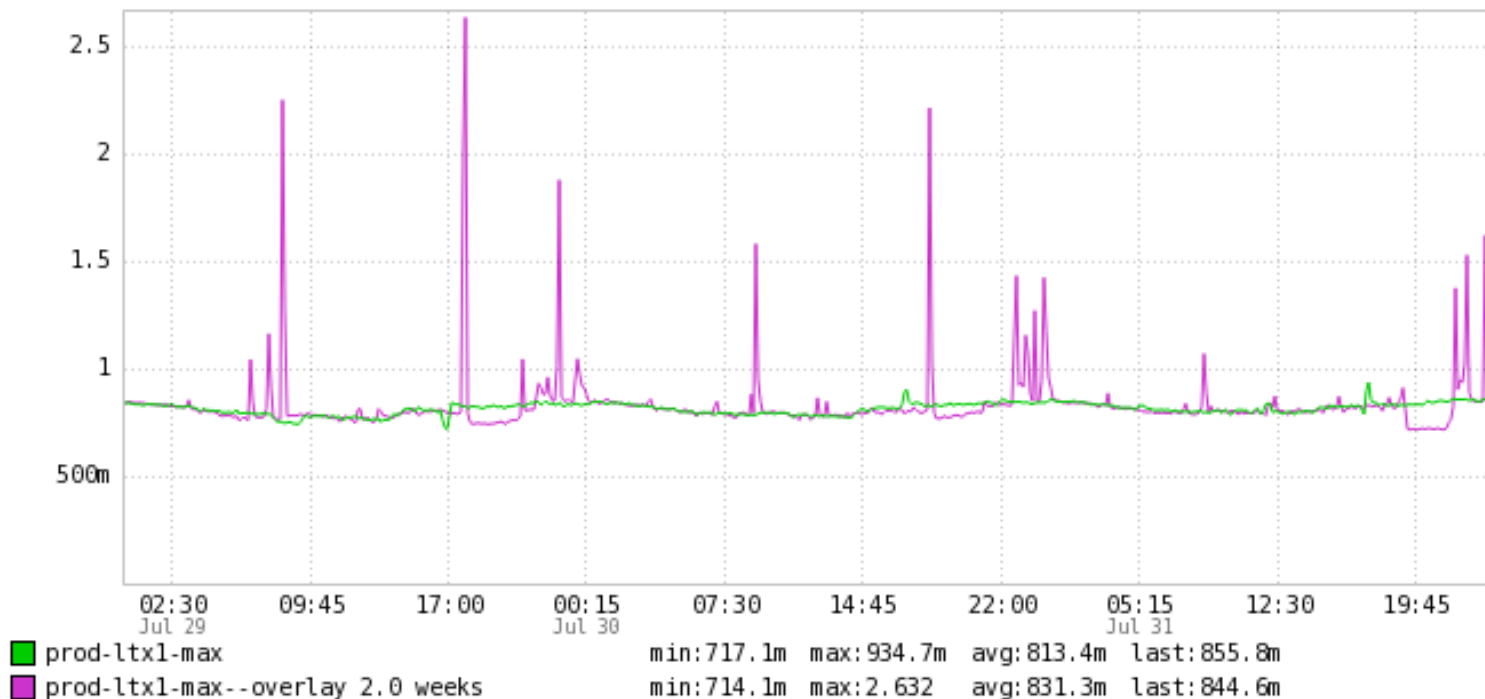
- 20 client instances
- 40 nodes multi-tenant Voldemort cluster
 - >90 stores
 - 56K QPS at peak

Metric:

- Worst p95/p99 latency from all clients
- Units are in milliseconds (“m” mean microseconds)
- Old code in pink
- New code in green

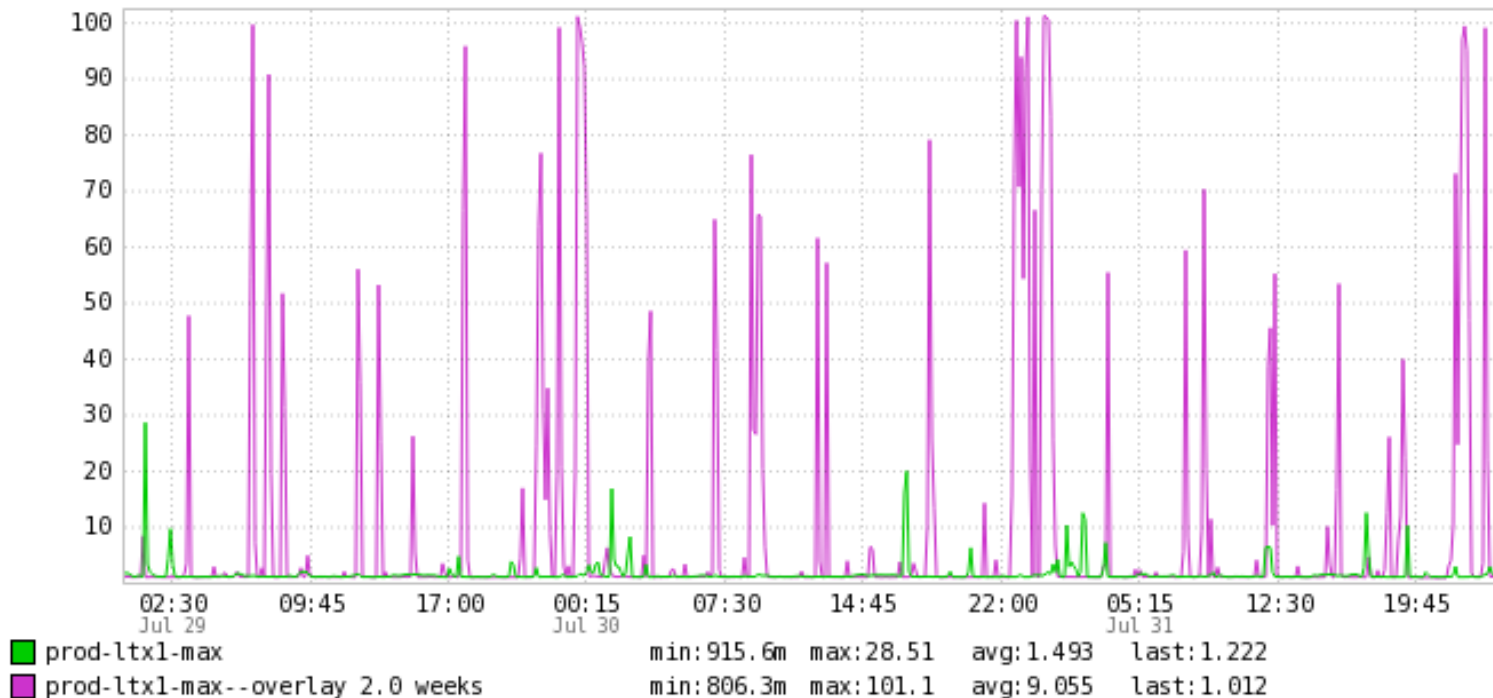
Client Latency

95th percentile



Client Latency

99th percentile



B



The DAILY PROPHET

★ THE WIZARD WORLD'S BEQUILING BROADCAST OF CHOICE ★

10,000 GALLONS ON BLACK'S HEAD
SEE INSIDE FOR FULL DETAILS PG. 3

National Weather
south - cloudy & rain ice
north - cloudy & rain ice
central - sunny period ice
London - cloudy & rain ice

Zodiac * Aspects
ta - my moon ☾ steps ♀♀
♊ - ♋ series - sun ☉ ♋ ♋
♈ ♀ ♀♈ - ♀♈ ♀♈ ♀♈ ♀♈

FIRST-SECOND EDITION
985890 - London - UK

TODAY ☽ is Aquarius
Letters or photos to the Editor should be sent only "by post" and with a clear stamp to The Daily Prophet - UK



EXCLUSIVE

MAYHEM AT HIGH SECURITY PRISON

MASS BREAKOUT FROM AZKABAN

by r. amorm - security editor

In a mass breakout - which all apprehensions for a period of time were made by international top - brass of "Azkaban" in early October, several hundred Azkaban prisoners have escaped from the high security prison. The escape was a complete surprise to the authorities and has caused a major crisis in the prison system. The escapees were taken to the mountains and hidden in a series of caves. The prison authorities are now searching for the escapees and are expected to capture them within a few days. The escape is believed to have been planned for some time and was the result of a long and careful preparation. The escapees are now being held in a series of temporary detention centres and are being questioned about their escape. The prison authorities are expected to launch a major operation to recapture the escapees and to prevent any further breakouts. The escape is a major blow to the prison system and is expected to have a major impact on the way in which the prison is run.

SEVERAL

ATCH BACKP

How to Get Started?

Get code

```
git clone https://github.com/voldemort/voldemort  
cd voldemort
```

Launch Voldemort Servers

```
./gradlew jar  
./bin/voldemort-shell.sh config/readonly_two_nodes_cluster/node0/config &  
./bin/voldemort-shell.sh config/readonly_two_nodes_cluster/node1/config &
```

Launch Build and Push job

```
./gradlew bnpJar  
./bin/run-bnp.sh <config_file>
```

Questions?

(We're hiring!)

