ARISTA

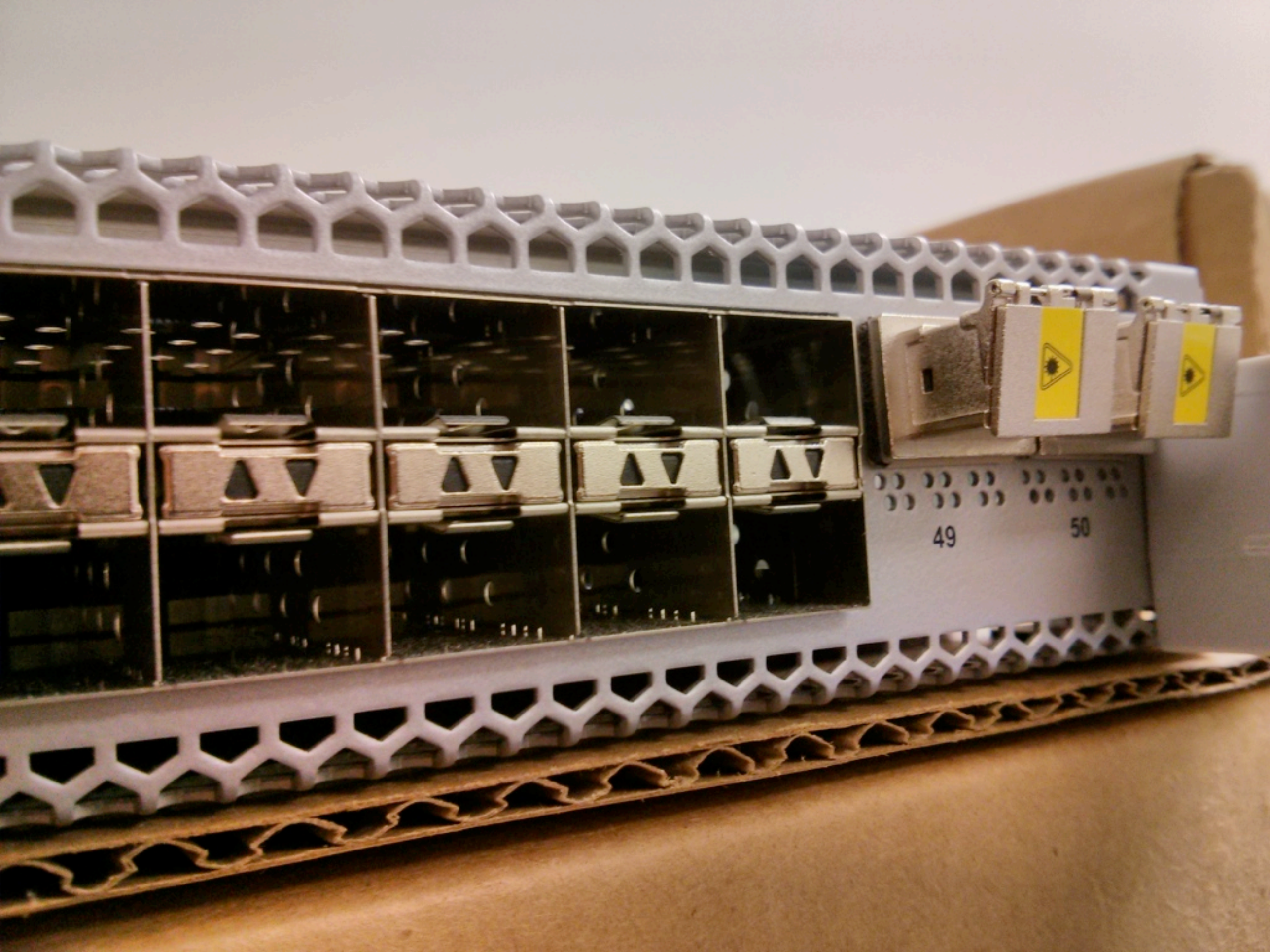
Modern & Scalable Data Center Networks

Benoît “tsuna” Sigoure
Member of the Yak Shaving Staff
tsuna@aristanetworks.com

@tsunanet

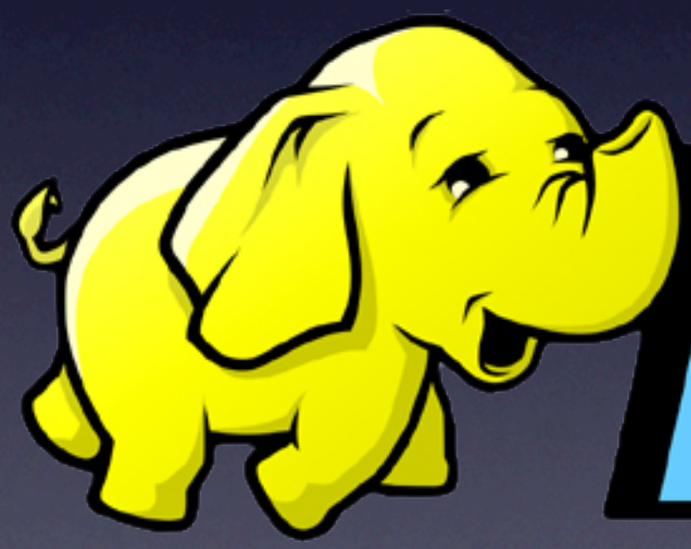
Agenda

1. What are we doing with Hadoop/HBase at Arista?
2. Update on AsyncHBase 1.5 and 0.95+ compatibility
3. Update on OpenTSDB 2.0 and the future

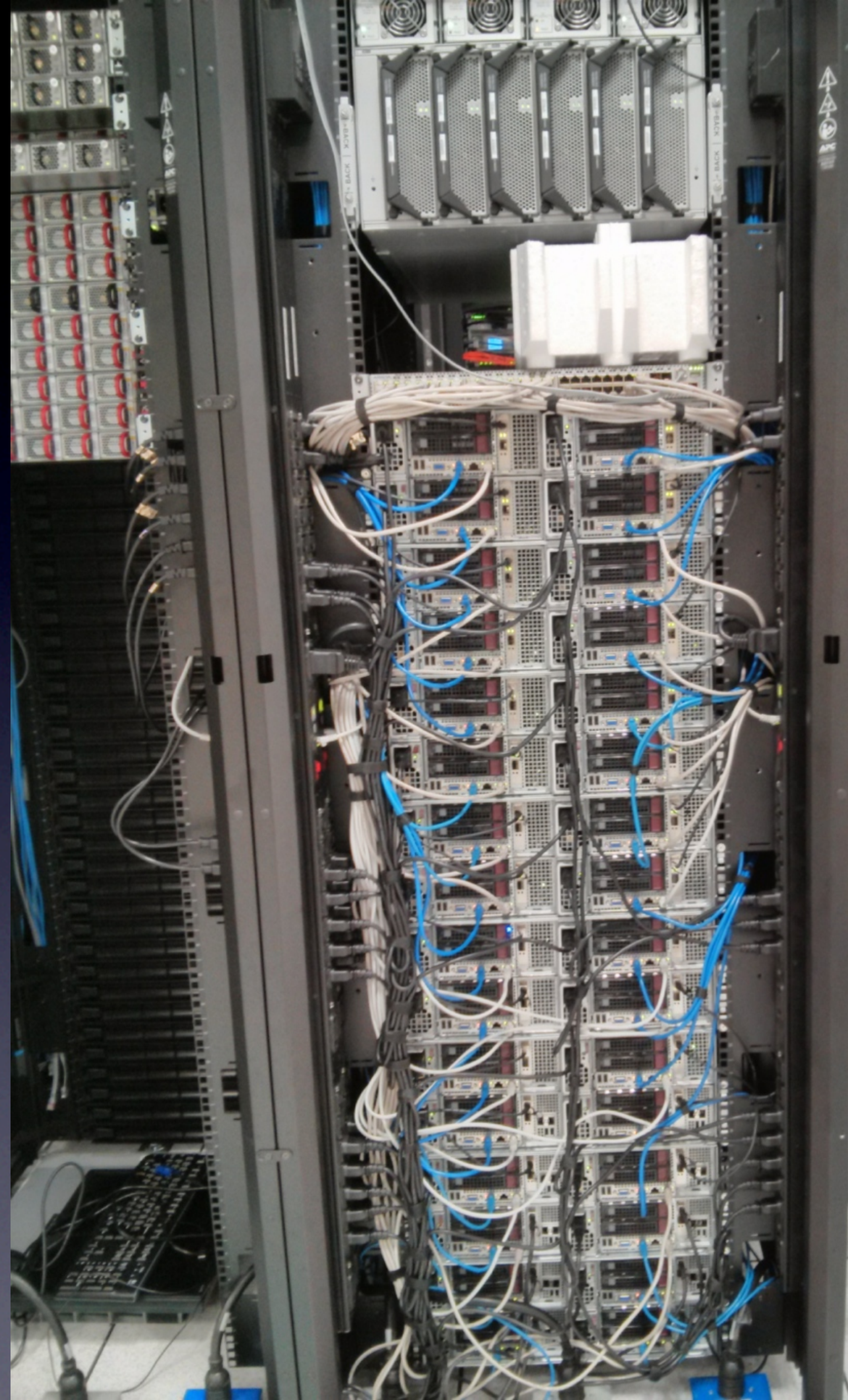

49
50

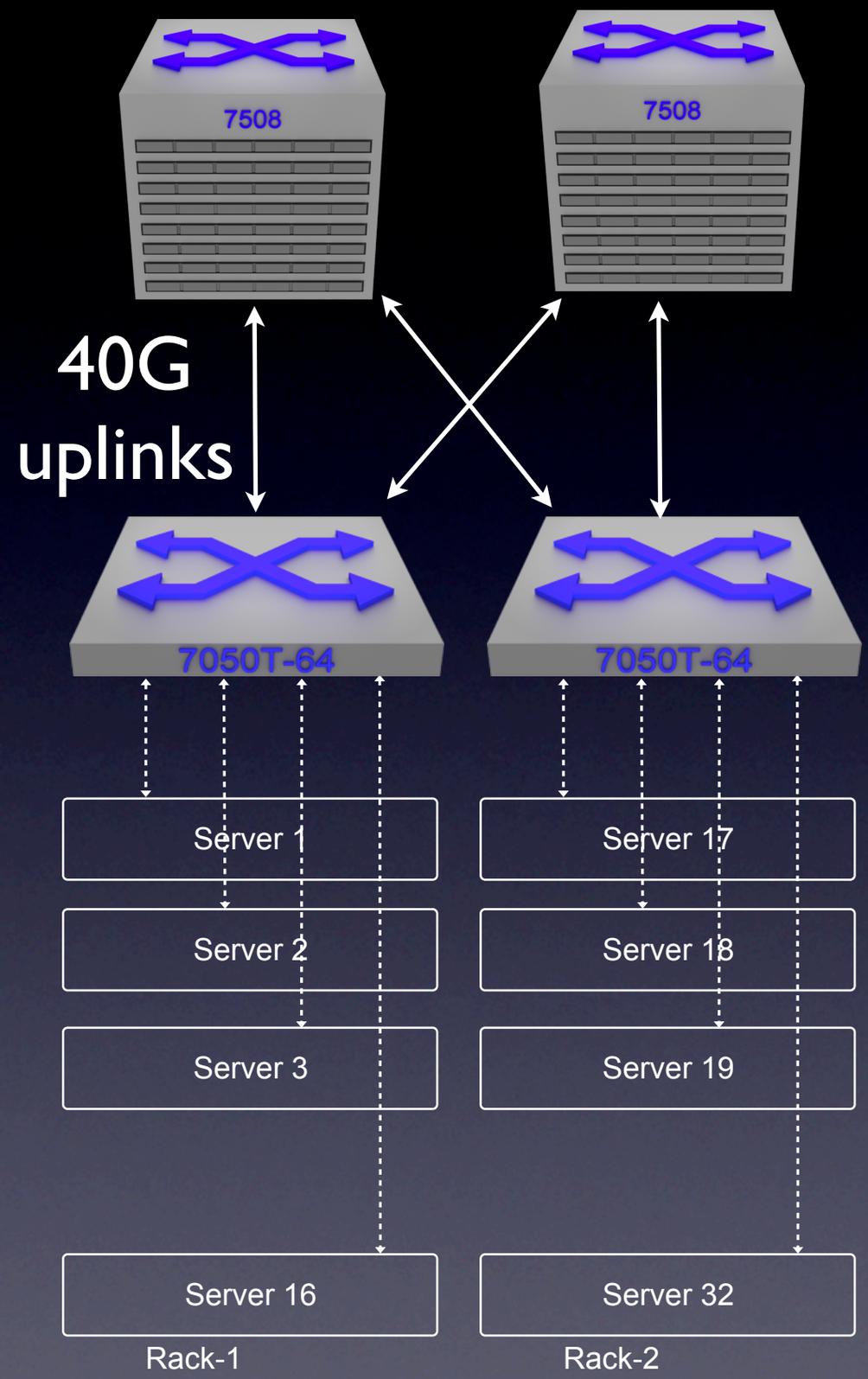
MapReduce Tracer

For



hadoop

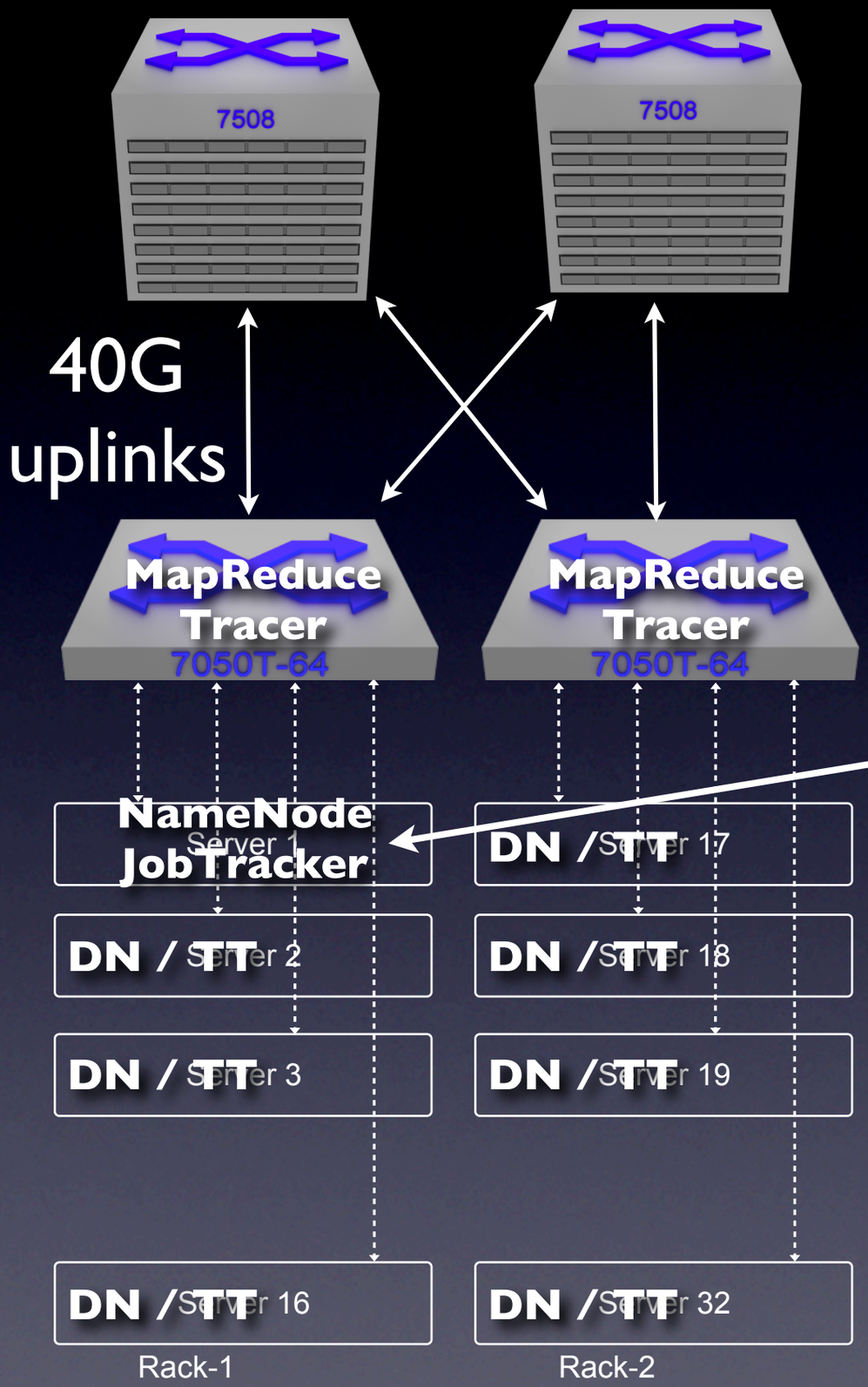




- 2 racks, 16 hosts each
- L3 network (/27 per rack, ECMP)
- Both TOR use the following additional config:

```
monitor hadoop
    cluster batch
        jobtracker host 172.19.33.1
        jobtracker user tsuna
        tasktracker http-port 10060
        no shutdown
```

- No change required to Hadoop
- No plugin, config change, or additional daemon to run



- 2 racks, 16 hosts each
- L3 network (/27 per rack, ECMP)
- Both TOR use the following additional config:

```
monitor hadoop
    cluster batch
        jobtracker host 172.19.33.1
        jobtracker user tsuna
        tasktracker http-port 10060
        no shutdown
```

- No change required to Hadoop
- No plugin, config change, or additional daemon to run

Hadoop Tracer
Phase 1: Visibility

- Keep track of nodes in the Hadoop cluster
- Poll statistics about Hadoop activity:
  - What jobs are running, when, where
  - How “big” each job is, network-wise
  - Historical data of MapReduce activity
  - MapReduce vs HDFS traffic accounting
- CLI integration for real-time visibility
- Multi-cluster support
- Current target: EoY release

Hadoop Versions Supported

- Apache Hadoop:
  0.20.203 to 1.2.1
- Cloudera's distribution:
  CDH3 and CDH4
- HortonWorks' distribution:
  HDP 1.1, 1.2, 1.3
- Microsoft HDInsight 0.10

- Incompatible:
  Apache 0.20.x and earlier
  Apache 0.21.x, 0.22.x, 0.23.x
  Apache 2.0.x

HBASE

TRACER

HE SE

TR ER

The idea

- Connect to ZooKeeper
- Find all nodes in the cluster
- Scan .META. to locate all regions
- Figure our which regions are local to the switch
- Collect statistics and annotate network events
- Feed information back into HBase?

AsyncHBase 1.5

Same API, same speed
Now with protobufs

What's new?

- 3 new contributors, 54 files changed, 4394 insertions(+), 443 deletions(-)
- ScanFilters (thanks Viral Bajaria from GrepData)
- RPC fast-fail
- Various bug fixes
- Compatibility with 0.95+

➡ As usual compatibility with previous versions of HBase is retained.

AsyncHBase 1.6 and beyond

- Secure RPC (Kerberos) – patch from Francis Christopher Liu at Yahoo!
- Streaming RPC – patches from Arthur van Hoff, Jonathan Payne, and Lei Zheng at Flipboard
- Atomic append
- Anybody would like to contribute coprocessor support code?

(Need logo)

OpenTSDB 2.0

Brought to you from the East coast
by Chris Larsen (ManOLamancha)

What's new?

- 11 contributors, 151 files changed, 40283 insertions(+), 1874 deletions(-)
- RESTful API, CORS
- Annotations
- Meta data
- Trees
- Plugins: search, data streaming, ingest
- Millisecond resolution

Backward compatible

Current Status

- First release candidate out
- Second release candidate: next week?
- Please help test it out!

- We're likely to have a 1.2 release too

Annotations

- Can be global or on a specific time series
- Start time, end time, description, notes
- Store custom key-values in annotations

- Annotations are stored in-line with the data
- Representation is JSON
- Qualifier must length must be 3 bytes and the first byte must be 0x01

Trees

- dal
  - web1.dal.mysite.com
    - app
      - connections (tsuid=010101)
      - errors (tsuid=0101010306)
    - cpu
      - system (tsuid=0102040101)
      - user (tsuid=0202040101)
    - web2.dal.mysite.com
      - cpu
        - system (tsuid=0102040102)
        - user (tsuid=0202040102)
    - web3.dal.mysite.com
      - cpu
        - system (tsuid=0102040103)
- lax
  - web1.lax.mysite.com
    - cpu
      - system (tsuid=0102050101)
  - web2.lax.mysite.com
    - cpu
      - system (tsuid=0102050102)

- Organize your time series in a tree-like fashion
- You can have 65535 trees
- Each tree has a rule set that expresses how to match time series at each level of the tree
- Helps bridge the gap with Graphite

Meta Data

- Track what time series (combination of metric name and tags) exist
- In a separate table: tsdb-meta
- Tracked with atomic increments
- If the value returned by increment is 1:
    - Create meta data objects for the new time series
    - Notify the search plugin if there is one
    - Process the tree rules

Millisecond Resolution

Challenge: introduce this feature without changing data representation and without adding extra costs for second-level resolution data points

Design doc @ http://goo.gl/Zbb35k

Millisecond Resolution

Bit Pattern (first 12 bits)	Value (delta in seconds)
0b000000000000	0
0b000000000001	1
0b000000000010	2
...	
0b111000001101	3597
0b111000001110	3598
0b111000001111	3599
0b111000010000	3600 – *unused*
... all unused ...	
0b111100000000	3840 – *unused*
0b1111...	*... all unused*
0b111111111111	4095 – *unused*

Millisecond Resolution: Column qualifier format

```
[ 0b11111101, 0b10111011, 0b10011111, 0b11000111 ]
    <--><------------------------------------>^^^<->
```

msec precision

delta (3599999 milliseconds)

type value length

2 unused bits

```java
final byte[] qualifier = …
final int delta;  // in ms
if (qualifier[0] & 0xF0) {  // 4 MSB set => delta in milliseconds
  final int qual = Bytes.getInt(qualifier, 0);
  delta = (qual & 0x0FFFFFC0) >> 6;
} else {  // Traditional delta in seconds
  final short qual = Bytes.getShort(qualifier, 0);
  delta = ((qual & 0xFFFF) >>> 4) * 1000;
}
final long timestamp = baseTime() * 1000 + delta;
```

Thank You

We're hiring in **SF**, Santa Clara, Vancouver, and Bangalore

ARISTA

Benoît "tsuna" Sigoure
Member of the Yak Shaving Staff
tsuna@aristanetworks.com

@tsunanet


#Hi
#Haters