ARISTA
Fast Server Failover for Hadoop & HBase

Benoît “tsuna” Sigoure
Member of the Yak Shaving Staff
tsuna@aristanetworks.com
What is SDN?

**Purist View**
a strict separation of control plane and data plane

**Pragmatic View**
a network architecture designed to be programmed by high-level languages and APIs

**A Common View**
SDN = Network Virtualization
SDN = OpenFlow
For Today SDN is: Making the Network Work Better With HBASE & hadoop
Host Failures in Hadoop Clusters

- Hardware failures
- Kernel panics
- Operator errors
- NIC driver bugs
Host Failures for HBase & Hadoop

- **RegionServer**: wait for ZooKeeper lease timeout
  Typical: ~30s

- **DataNode**: wait for heartbeats to timeout enough for NameNode to declare dead
  Typical: ~10min (or ~30s with `dfs.namenode.check.stale.datanode` see HDFS-3703)
Host Failures for HBase & Hadoop

- **RegionServer**: wait for ZooKeeper lease timeout
  Typical: ~30s

- **DataNode**: wait for heartbeats to timeout enough for NameNode to declare dead
  Typical: ~10min (or ~30s with dfs.namenode.check.stale.datanode see HDFS-3703)
Host Failures for HBase & Hadoop

- **RegionServer**: wait for ZooKeeper lease timeout
  Typical: ~30s

- **DataNode**: wait for heartbeats to timeout enough for NameNode to declare dead
  Typical: ~10min (or ~30s with dfs.namenode.check.stale.datanode see HDFS-3703)
Manually Mitigating Host Failures

Assumptions:
- Modern ToR (Top-of-Rack)
- L3 network design
- Hosts are singly homed
Manually Mitigating Host Failures

Process:
• Log into ToR
• Add IP address of the failed host as a secondary IP on the SVI used as the default gateway
• Remove IP when the host comes back
Manually Mitigating Host Failures

Process is brittle and not bullet proof
How Does This Work?

- Redirect traffic to the switch
- Have the switch’s kernel send TCP resets
- Immediately kills all existing and new flows
How Does This Work?

- Redirect traffic to the switch
- Have the switch’s kernel send TCP resets
- Immediately kills all existing and new flows
Arista invited a few customers to talk Hadoop

How can we help you make Hadoop run smoother?

Well there is this manual process I use to work around machine failures...

It sucks. Can't we just get the network to do it for us?

And that’s how **Fast Server Failover** was born.
Fast Server Failover

- Switch learns & tracks IP ↔ port mapping
- Port down → take over IP and MAC addresses
- Kicks in as soon as hardware notifies software of the port going down, within a few milliseconds
- Port back up → resume normal operation
- Can also run custom shell script on each event

EOS = Linux
Under the Hood

- Custom callback for ARP & MAC table changes
- Custom MAC table entry to tell hardware to send packets to Linux
- Rule in iptables to reject traffic (TCP RST, ICMP Destination Unreachable, etc.)
- Devil is in the details:
  - Server moving to another interface
  - Aggregated links (LAG)
  - Multi-chassis Link Aggregation (MLAG)
  - Handle IPs / routes changing on the fly
  - Static MAC entries
  - Rate-limiting traffic to not overwhelm Linux
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