Scaling networks through software

March 16th 2015 | João Taveira Araújo
@jta
network systems @ fastly
scalability
constraints
Scaling networks through software constraints

knowledge
constraints
knowledge
technology
constraints
knowledge
technology
complexity
constraints

time
money
people
constraints

time
money
people
Scaling networks through software constraints time money people
Becoming a multi terabit network

Number of PoPs                      ~20
BGP announcements                  ~2000
Requests per second                ~1000000
Becoming a multi terabit network

Number of PoPs ........................................... ~20
BGP announcements ................................. ~2000
Requests per second ............................... ~1000000

Network ops ............................................. 2
Becoming a multi terabit network

Number of PoPs .......................... ~20
BGP announcements ..................... ~2000
Requests per second ..................... ~1000000

Network ops ............................. 2
Network software ........................ me
scalability
Observations on network scalability from a company that used to be a startup
anything you don’t explicitly control is an implicit liability
the internet
Scaling networks through software
How do you:

• load balance traffic
• gracefully failover if a server fails
Scaling networks through software

PoP

router

A
B
C
D

DNS

A
B
C
D

25%
25%
25%
25%
Scaling networks through software
Bad idea:

- gets hard to manage
- do one thing and do it well
- you don’t control TTL
Scaling networks through software

load balancer

PoP

A B C D
Scaling networks through software

PoP

load balancer

A  B  C  D

Cost

State

© COPYRIGHT 2014 FASTLY, INC.
Scaling networks through software

PoP

load balancer

A  B  C  D

Cost

State

Existing demand
Scaling networks through software

PoP

load balancer

A → B → C → D

Cost

State

Existing demand

Projected demand

© COPYRIGHT 2014 FASTLY, INC.
Bad idea:

- you don’t control demand
- you don’t control DDOS
Scaling networks through software

PoP

ECMP

A  B  C  D
Scaling networks through software

Destination network | Next hop
---|---
10.0.0.0/24 | A
10.0.0.0/24 | B
10.0.0.0/24 | C
10.0.0.0/24 | D
Scaling networks through software

ECMP

Destination network | Next hop
--- | ---
10.0.0.0/24 | A
10.0.0.0/24 | B
10.0.0.0/24 | C
10.0.0.0/24 | D

PoP

ECMP

A
B
C
D
Scaling networks through software

ECMP

Destination network | Next hop
--- | ---
10.0.0.0/24 | B
10.0.0.0/24 | C
10.0.0.0/24 | D

Bad idea:

- connection resets
- you don’t control rehashing
- you don’t control vendor roadmaps
don’t resign to fate just because everything sucks
faild
Scaling networks through software
Scaling networks through software

<table>
<thead>
<tr>
<th>Destination network</th>
<th>Next hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.1</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.2</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IP Address</th>
<th>MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.A.1</td>
<td>A:A</td>
</tr>
<tr>
<td>10.1.A.2</td>
<td>A:A</td>
</tr>
<tr>
<td>10.1.A.3</td>
<td>A:A</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Scaling networks through software

**PoP**

```
<table>
<thead>
<tr>
<th>Destination network</th>
<th>Next hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.1</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.2</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IP Address</th>
<th>MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.A.1</td>
<td>A:A</td>
</tr>
<tr>
<td>10.1.A.2</td>
<td>A:A</td>
</tr>
<tr>
<td>10.1.A.3</td>
<td>A:A</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
```

**ECMP**

A B C D
drain a host
Scaling networks through software

<table>
<thead>
<tr>
<th>Destination network</th>
<th>Next hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.1</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.2</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

PoP

ECMP

A  B  C  D
## Scaling networks through software

### Destination network

<table>
<thead>
<tr>
<th>Destination network</th>
<th>Next hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.1</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.2</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### IP Address

<table>
<thead>
<tr>
<th>IP Address</th>
<th>MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.A.1</td>
<td>B:A</td>
</tr>
<tr>
<td>10.1.A.2</td>
<td>C:A</td>
</tr>
<tr>
<td>10.1.A.3</td>
<td>D:A</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Scaling networks through software
cut off to failed state
Scaling networks through software

<table>
<thead>
<tr>
<th>Destination network</th>
<th>Next hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.1</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.2</td>
</tr>
<tr>
<td>10.0.0.0/24</td>
<td>10.1.A.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

PoP

ECMP

A

B

C

D

© COPYRIGHT 2014 FASTLY, INC.
Scaling networks through software
Scaling networks through software

ecmp

Active flows

Time
load balancer
Scaling networks through software

faild

- Active flows decreasing over time

- State increasing over time
faild
Scaling networks through software

Time

Active flows

State

Time

faild
if it’s expensive you probably don’t need it.
F5 BIG-IP 10350v
F5 BIG-IP 10350v

$200,000
F5 BIG-IP 10350v

$200,000

$0
load balancer
load balancer balancing
load balancer balancing

(a load balancer is just an appliance which provides load balancing)
distributed load balancer balancing

(a load balancer is just an appliance which provides load balancing)
How to build a PoP

- buy a router
- get BGP table from each provider
- install routes to FIB
- servers use default gateway
Juniper MX960 Router
Juniper MX960 Router

~$500,000
router
router routing
router routing

(a router is just an appliance which provides routing)
distributed router routing

(a router is just an appliance which provides routing)
Arista DCS-7150S switch family
Arista DCS-7150S switch family

$29,995
How to build a Fastly PoP

- buy switches
- reflect BGP down to servers
- inject multipath routes into FIB
How to build a Fastly PoP

- buy switches
- reflect BGP down to servers
- inject multipath routes into FIB
How packets egress Fastly

- switches emit nexthop IP and MAC
- servers configure p2p link / ARP
- send directly to provider nexthop!
How packets egress Fastly

- switches emit nexthop IP and MAC
- servers configure p2p link / ARP
- send directly to provider nexthop!
joao@cache:~$ sudo birdc show route count
BIRD 1.4.4 ready.
2099355 of 2099355 routes for 524852 networks
Fastly PoPs: engineering perspective
Fastly PoPs: investor perspective
It’s easier to make people less busy than hire people.
YOU WOULDN'T DOWNLOAD A CAR
You wouldn't download a car

Yes I would
Scaling networks through software

software
networking

“you wouldn’t do that to a switch”
“Networking is hard”
“Networking is hard”

resource constraints
“Networking is hard”

resource constraints
protocol standards
“Networking is hard”

- resource constraints
- protocol standards
- security concerns
“Networking is hard”

resource constraints
protocol standards
security concerns
network vendors
where is time spent needlessly?
pinpointing path failures
st-ping: probe all upstreams

<table>
<thead>
<tr>
<th>Upstream</th>
<th>Intf</th>
<th>Nexthop</th>
<th>Sent</th>
<th>Loss</th>
<th>Min</th>
<th>Avg</th>
<th>Max</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>cogent</td>
<td>p5p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.023</td>
<td>1.042</td>
<td>1.056</td>
<td>0.022</td>
</tr>
<tr>
<td>cogent</td>
<td>p3p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.018</td>
<td>1.042</td>
<td>1.079</td>
<td>0.034</td>
</tr>
<tr>
<td>cogent</td>
<td>p3p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.014</td>
<td>1.029</td>
<td>1.059</td>
<td>0.011</td>
</tr>
<tr>
<td>cogent</td>
<td>p5p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.024</td>
<td>1.036</td>
<td>1.063</td>
<td>0.039</td>
</tr>
<tr>
<td>l3</td>
<td>p3p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>0.867</td>
<td>0.878</td>
<td>0.902</td>
<td>0.016</td>
</tr>
<tr>
<td>l3</td>
<td>p5p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.347</td>
<td>1.357</td>
<td>1.383</td>
<td>0.038</td>
</tr>
<tr>
<td>l3</td>
<td>p3p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.3</td>
<td>1.318</td>
<td>1.341</td>
<td>0.021</td>
</tr>
<tr>
<td>l3</td>
<td>p5p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>0.88</td>
<td>0.887</td>
<td>0.902</td>
<td>0.027</td>
</tr>
<tr>
<td>* telia</td>
<td>p3p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>26.485</td>
<td>26.634</td>
<td>27.243</td>
<td>0.32</td>
</tr>
<tr>
<td>* telia</td>
<td>p3p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>27.963</td>
<td>28.587</td>
<td>29.692</td>
<td>0.674</td>
</tr>
<tr>
<td>* telia</td>
<td>p5p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>25.81</td>
<td>26.621</td>
<td>27.24</td>
<td>0.446</td>
</tr>
<tr>
<td>* telia</td>
<td>p5p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>27.953</td>
<td>29.058</td>
<td>29.669</td>
<td>0.634</td>
</tr>
</tbody>
</table>
changing route preferences
switch#conf
switch(config)#l3
switch(config-if-Et3)#show active

interface Ethernet3
  description l3_1 [asia,dns1,dns2,dns3,dns4,http1,http2,http3,http4,site] is up since 2015-02-27
  load-interval 5
  ip access-group inbound in
  ip access-group outbound out
  queue-monitor length thresholds 1024 128
  no lldp receive
upstream alias

```
switch\> (config-if-Et3)#show active
interface Ethernet3
    description l3_1 [asia,dns1,dns2,dns3,dns4,http1,http2,http3,http4,site] is up since 2015-02-27
    load-interval 5
    ip access-group inbound in
    ip access-group outbound out
    queue-monitor length thresholds 1024 128
    no lldp receive
```
announced prefixes

```
switch#conf
switch(config)#l3
switch(config-if-Et3)#show active
interface Ethernet3
  description l3_1 [asia,dns1,dns2,dns3,dns4,http1,http2,http3,http4,site] is up since 2015-02-27
  load-interval 5
  ip access-group inbound in
  ip access-group outbound out
  queue-monitor length thresholds 1024 128
  no lldp receive
```
live BGP info
switch (config-if-Et3)#desc +15169
switch (config-if-Et3)#show active
interface Ethernet3
description l3_1 [asia,dns1,dns2,dns3,dns4,http1,http2,http3,http4,site] {+15169} is up since 2015-02-27
load-interval 5
ip access-group inbound c in
ip access-group outbound out
queue-monitor length thresholds 1024 128
no lldp receive
increase Google localpref

```
switch# (conf-if-Et3)# desc +15169
switch# (conf-if-Et3)# show active
interface Ethernet3
  description l3_1 [asia,dns1,dns2,dns3,dns4,http1,http2,http3,http4,site] {+15169} is up since 2015-02-27
  load-interval 5
  ip access-group inbound in
  ip access-group outbound out
  queue-monitor length thresholds 1024 128
  no lldp receive
```
switch (config-if-Et3)#desc +15169
switch (config-if-Et3)#show active
interface Ethernet3
description l3_1 [asia,dns1,dns2,dns3,dns4,http1,http2,http3,http4,site] {+15169} is up since 2015-02-27
load-interval 5
ip access-group inboundc in
ip access-group outbound out
queue-monitor length thresholds 1024 128
no lldp receive

localpref overrides
```
joao@cache:~$ sudo tt-ping 8.8.8.8
Pinging 8.8.8.8 via 12 upstreams.

<table>
<thead>
<tr>
<th>Upstream</th>
<th>Intf</th>
<th>Nexthop</th>
<th>Sent</th>
<th>Loss</th>
<th>Min</th>
<th>Avg</th>
<th>Max</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>cogent</td>
<td>p3p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.018</td>
<td>1.028</td>
<td>1.037</td>
<td>0.035</td>
</tr>
<tr>
<td>cogent</td>
<td>p5p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.02</td>
<td>1.037</td>
<td>1.052</td>
<td>0.044</td>
</tr>
<tr>
<td>cogent</td>
<td>p3p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.011</td>
<td>1.031</td>
<td>1.06</td>
<td>0.028</td>
</tr>
<tr>
<td>cogent</td>
<td>p5p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.026</td>
<td>1.033</td>
<td>1.049</td>
<td>0.026</td>
</tr>
<tr>
<td>l3</td>
<td>p3p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.3</td>
<td>1.319</td>
<td>1.363</td>
<td>0.035</td>
</tr>
<tr>
<td>l3</td>
<td>p5p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>1.344</td>
<td>1.357</td>
<td>1.383</td>
<td>0.034</td>
</tr>
<tr>
<td>l3</td>
<td>p3p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>0.866</td>
<td>0.879</td>
<td>0.899</td>
<td>0.033</td>
</tr>
<tr>
<td>l3</td>
<td>p5p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>0.869</td>
<td>0.885</td>
<td>0.925</td>
<td>0.038</td>
</tr>
<tr>
<td>telia</td>
<td>p3p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>25.802</td>
<td>26.55</td>
<td>27.202</td>
<td>0.379</td>
</tr>
<tr>
<td>telia</td>
<td>p5p1</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>26.481</td>
<td>26.713</td>
<td>27.231</td>
<td>0.346</td>
</tr>
<tr>
<td>telia</td>
<td>p5p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>27.943</td>
<td>28.803</td>
<td>29.47</td>
<td>0.619</td>
</tr>
<tr>
<td>telia</td>
<td>p3p2</td>
<td></td>
<td>10</td>
<td>0.0%</td>
<td>27.948</td>
<td>28.579</td>
<td>29.669</td>
<td>0.667</td>
</tr>
</tbody>
</table>
```
changing prefix announcements
interface Ethernet3
    description l3_1 ![http1,http2,http3,http4,asia,dns1,dns2,dns3,dns4,site] {+15169} is feed since 2015-02-27
    load-interval 5
    ip access-group inbound c in
    ip access-group outbound out
    queue-monitor length thresholds 1024 128
    no lldp receive
withdraw all HTTP anycast prefixes
Interface Ethernet3
  description l3_1 [!http1,!http2,!http3,!http4,asia,dns1,dns2,dns3,dns4,site] {+15169} is feed since 2015-02-27
  load-interval 5
  ip access-group inbound in
  ip access-group outbound out
  queue-monitor length thresholds 1024 128
  no lldp receive

BGP session status
BGP session status

description l3_1 ![http1, http2, http3, http4, asia, dns1, dns2, dns3, dns4, site] {15169} is fed since 2015-02-27
load-interval 5
ip access-group inbound in
ip access-group outbound out
queue-monitor length thresholds 1024 128
no lldp receive
changing global routing policy
We generate lots of BGP announcements

- changing policy manually is hard
- changing policy per-device takes long
We generate lots of BGP announcements

- changing policy manually is hard
- changing policy per-device takes long
Stage and deploy via Github

- generate diff of routing policy and exported routes
- peer reviewed, endlessly revertible
Staging lists affected switches and prefixes

- human error could withdraw Fastly from the Internet
- hard to automate, so make sure people can get it right first
Seems so simple...

- reduced time spent needlessly
- reduced human error dramatically
- allowed us to train netops out of our datacenter team
- Arista eAPI allows description changes: instant RESTful orchestration
existing best practices won’t save you.
Saving money

- buy bare essentials
- distribute everything
- efficiency matters

Saving time

- correct architecture helps!
- reduce cognitive overhead
- solve ops first, automate later
Be wary of:

- best practices
- cool stuff
- perfect