PEAK WEB CONSULTING

ENGINEERED TO SOLVE PROBLEMS

Extending the Life of Layer 3 Switches in a 256k+ Route World

a.k.a.

"Lose Routes Now, Ask Me How!"

NANOG44
October 13, 2008

Dani Roisman
droisman~at~peakwebconsulting.com

What's the Problem

- A few popular Layer 3 switches have forwarding table size limitations:
  - Cisco Sup2/MSFC2: 256,000 routes
  - Cisco Sup720-3B: 192,000 or 239,000 routes (after "mls cef maximum-routes" & reload)
- The public Internet v4 routing table is currently > 263,000 routes (Oct, 2008)
- Oops, that just won't work at all!

Who May Be Interested

- Enterprise or Datacenter network operators with older equipment
  - not for you if you provide BGP to customers
- Networks that are multihomed to diverse ISPs and taking a full BGP feeds:
  - to achieve better traffic engineering
  - for metrics gathering (e.g. bandwidth per-AS)
- Constrained by funds, power, rack space, time, human resources

What is the Ideal Solution

- So many routes, but so few next hops
- Why do we need so many routes in the forwarding table?
  - We’re already having this discussion to try and avoid issues with *new* hardware and the fear of an ever-expanding IPv4 + growing IPv6 forwarding tables
  - Why can’t router software perform aggregation where possible?

What is the Current Solution

- Purchase new hardware
  - May just mean upgrading your management modules
  - For some vendors may mean upgrading every linecard in the chassis as well (distributed forwarding)
  - May include provisioning additional power, changing fan blades, and even juggling linecard positions in a chassis
    - e.g. Cisco Sup2 fits in slots 1 & 2 of a 6509, but Sup720-3BXL goes into slots 5 & 6

Is There an Acceptable Workaround

- What are the goals:
  - Quick, easy, inexpensive
  - Retain as much relevant routing information as possible
  - Maintain "*routing accuracy*" when making ISP next-hop decisions
- This comes down to abbreviating the IP forwarding table ("route pruning")

Routing Accuracy

- Personal definition: amount of traffic that can be forwarded by a default-free routing table
- Traffic that cannot follow a shorter match in the forwarding table will take the default path, this is your “inaccuracy”
- For Enterprise and Content networks this is not severe or detrimental – your upstream ISP will know how to deliver this traffic

How is This Done

- In order to prevent a service disruption, you must first receive default route 0.0.0.0 from *EACH* of your ISPs in addition to the full routing table
- Then simply throw away some long prefixes and check results:
  - forwarding table size reduction
  - routing accuracy: how much of your network's traffic is forwarded without following default (ask me how)

Filter Parameters We've Used

- For IP blocks in 91.0.0.0/8:
  - allow up to and including /24
- For IP blocks in "Class A" or "Class B" 0.0.0.0 – 191.255.255.255:
  - allow up to and including /23
- For everything else (historical "Class C" is left) 192.0.0.0 – 223.0.0.0:
  - allow up to and including /24

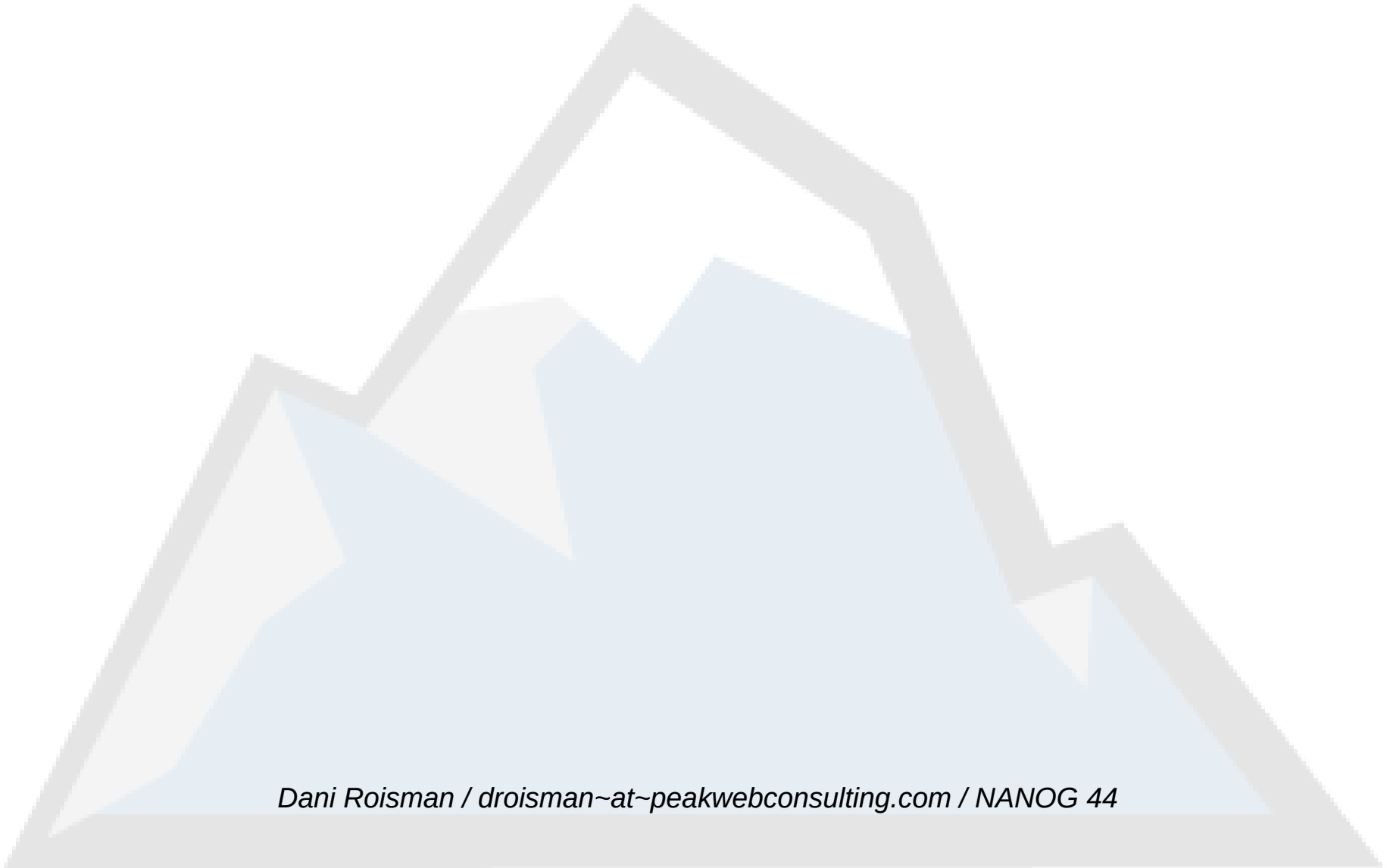
Real World Results

- May / June 2008:
  - Forwarding table size decreased from 253,000 routes to 199,000 routes
- September 2008:
  - Forwarding table at 205,000 routes
- Routing accuracy > 99%
  - Total: 1.5Gbps, following default: < 9Mbps
  - Total: 300Mbps, following default: < 450Kbps
  - Total: 5Gbps, following default: < 30Mbps

Tips

- Apply a few broad strokes, and get back to other work
- Don’t dwell on minimum allocations
  - Many examples found get too specific
- Diminishing margin of return:
  - Reducing forwarding table size further results in reduced “forwarding accuracy”

Any Questions?



Reference: Actual Configuration

- In Cisco IOS:

```
ip prefix-list REJECT-DEAGGREGATES seq 5 permit 0.0.0.0/2 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 10 permit 64.0.0.0/4 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 15 permit 80.0.0.0/5 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 20 permit 88.0.0.0/7 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 25 permit 90.0.0.0/8 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 30 permit 92.0.0.0/6 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 35 permit 96.0.0.0/3 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 40 permit 128.0.0.0/2 ge 24
ip prefix-list REJECT-DEAGGREGATES seq 45 permit 0.0.0.0/0 ge 25

route-map ISP-IN deny 10
 match ip address prefix-list REJECT-DEAGGREGATES
route-map ISP-IN permit 20...
 (your standard route-map begins here...)
```

Reference: Additional Results

- Restrictions applied to “A” and “B” blocks:

```
Allow up to /22 from everywhere else = 186k routes
Allow up to /21 from everywhere else = 174k routes
Allow up to /20 from everywhere else = 164k routes
Allow up to /19 from everywhere else = 153k routes
```

- Accuracy drop to < 90% with table drop to 153k

Reference: Minimum Allocations

- ARIN:

  http://www.arin.net/reference/ip_blocks.html#ipv4
- RIPE:

  (here you'll find detail about 91.0.0.0/8)
  https://www.ripe.net/ripe/docs/ripe-ncc-managed-address-space.html
- APNIC:

  http://www.apnic.net/db/min-alloc.html
- AFRINIC:

  http://www.afrinic.net/docs/policies/afpol-v4200407-000.htm