NetFlow & BGP multi-path: quo vadis?

Paolo Lucente <paolo@pmacct.net>
Elisa Jasinska <elisa@netflix.com>

NANOG61, Bellevue

Agenda

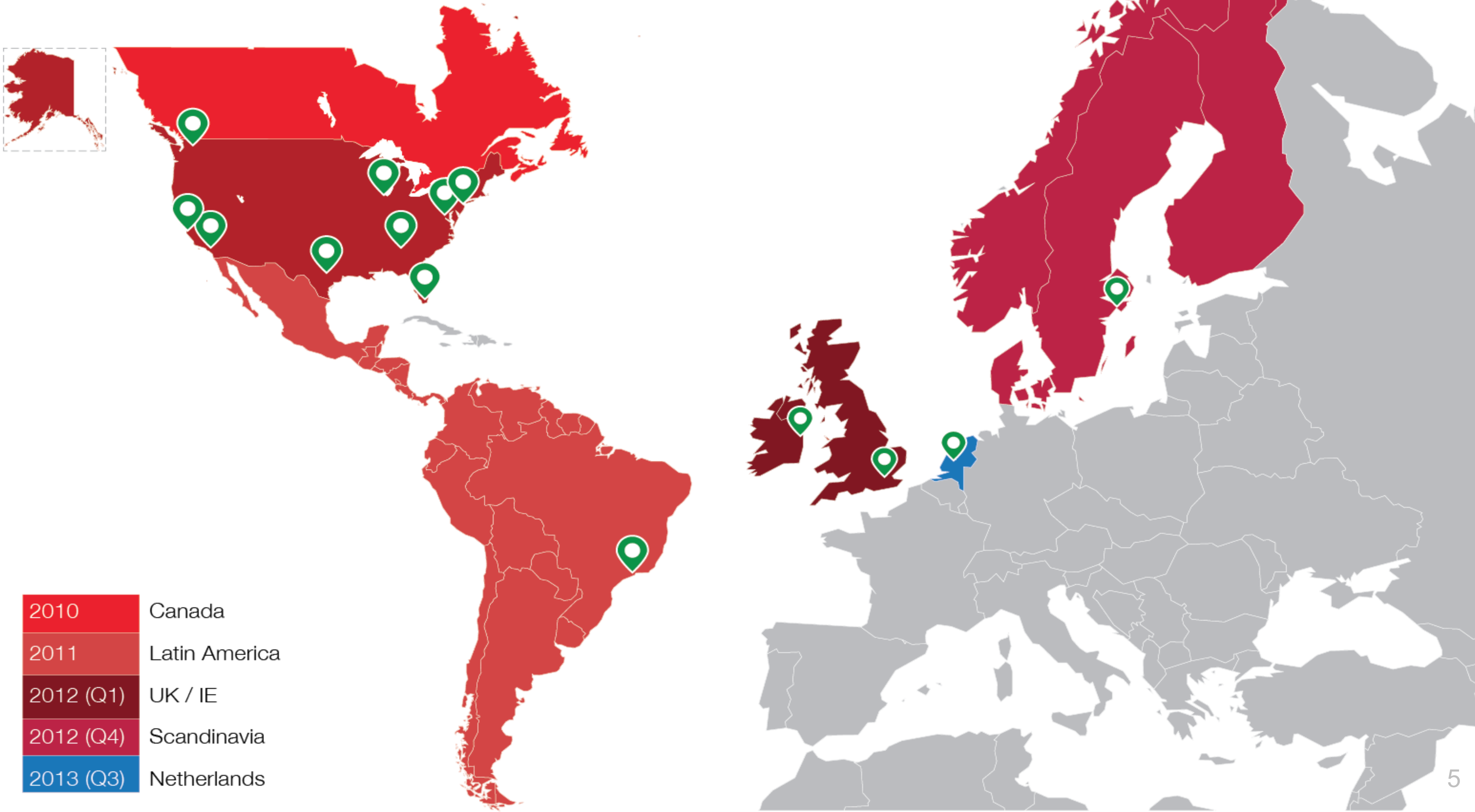
- About Netflix
- About pmacct
- Brief digression on BGP ADD-PATHS
- Putting all the pieces together

About Netflix

Netflix

- Available in over 40 countries
  - North America, including Canada & Mexico
  - Europe: UK, IE, NL, SE, DK, FI, NO
  - Latin America and the Caribbean
- 35 operational POPs
  - 24 in the USA
  - Brazil, London, Dublin, Amsterdam, Stockholm
- Over 48 million subscribers

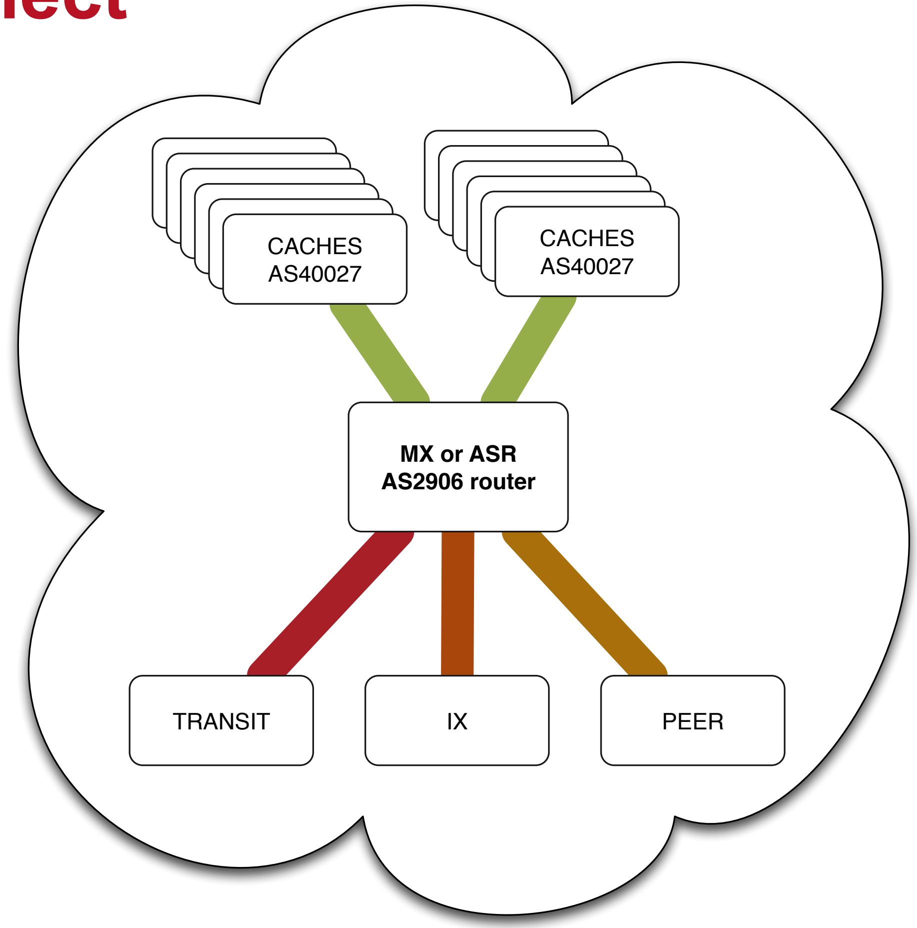
Netflix Service



2010	Canada
2011	Latin America
2012 (Q1)	UK / IE
2012 (Q4)	Scandinavia
2013 (Q3)	Netherlands

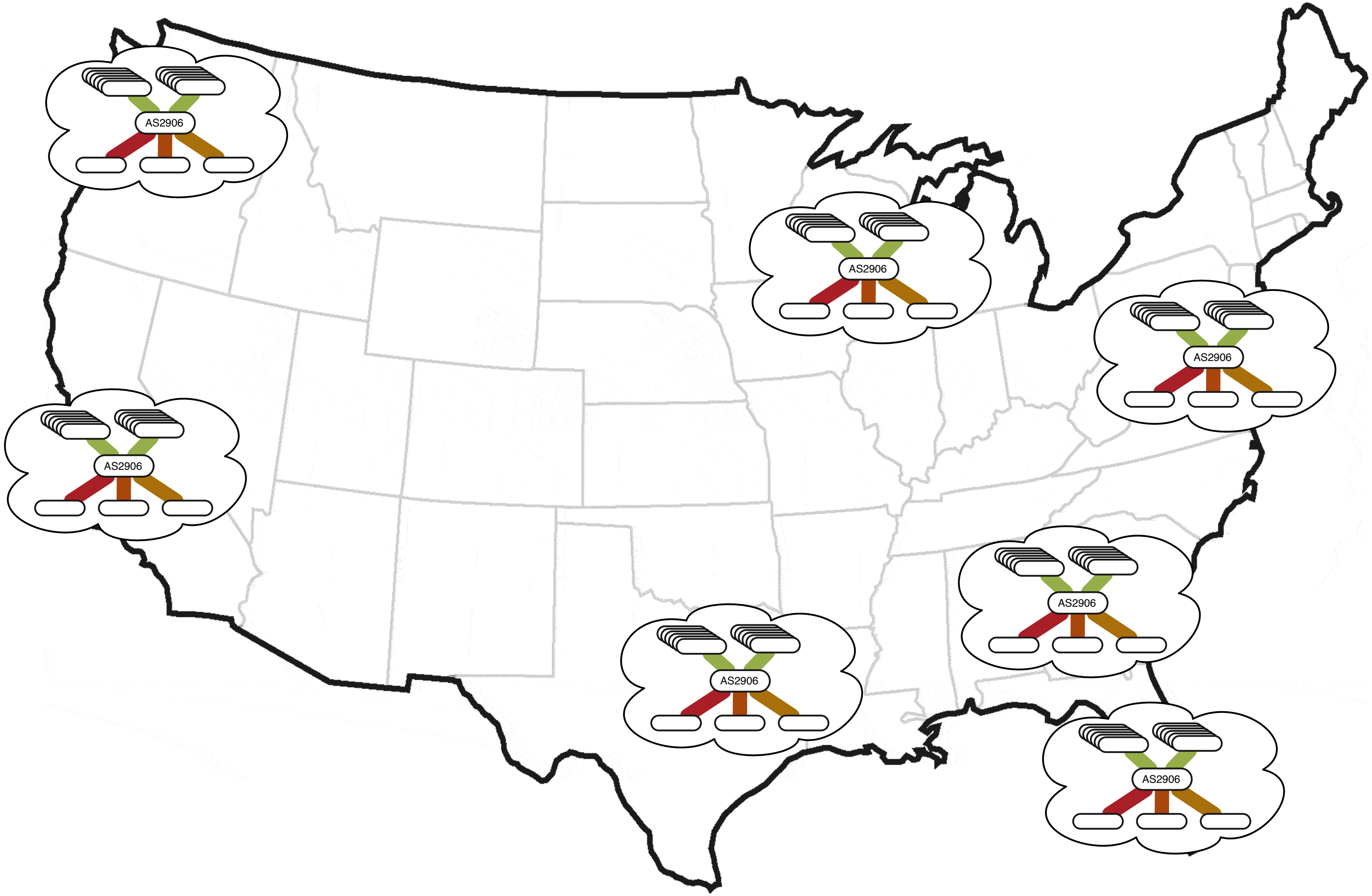
Netflix CDN: Open Connect

- In house CDN
- Designed for efficient video delivery
  - Many POPs
  - No backbone
- Hardware: ASR, MX and some Arista 7500e
- Delivery via:
  - Servers embedded in access network
  - Peering
  - Transit



https://www.netflix.com/openconnect

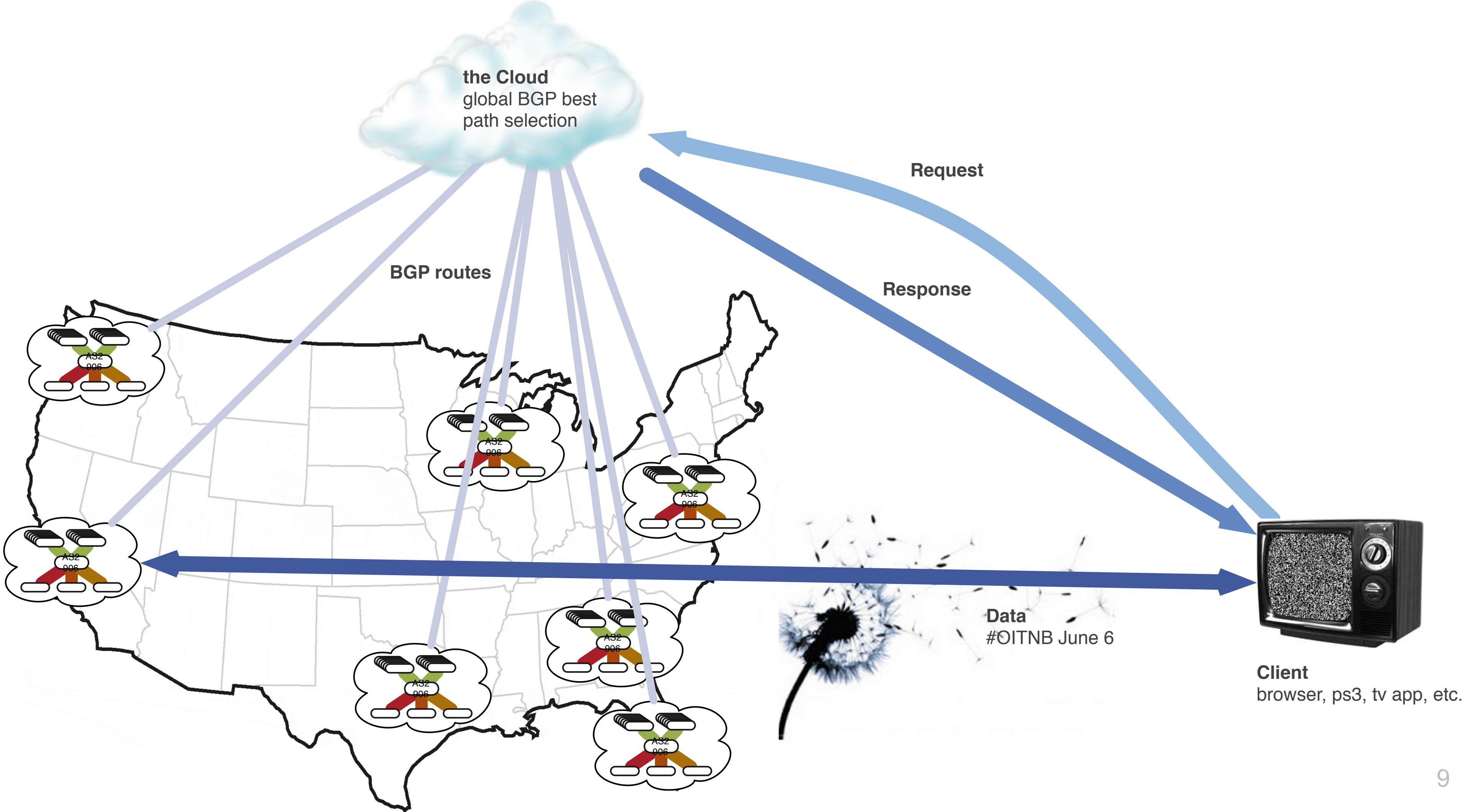
Network Design at Netflix



A Global Network in the Sky

- Routes flow into the cloud and re-aggregate
- BGP path selection algorithm re-implemented with support for massive ECMP/UCMP across distributed devices/pops (as if they were connected)
- Geography, policy, cost, and health used to route viewing sessions to “the best device in the best place”

A Global Network in the Sky



the Cloud
global BGP best
path selection

Request

Response

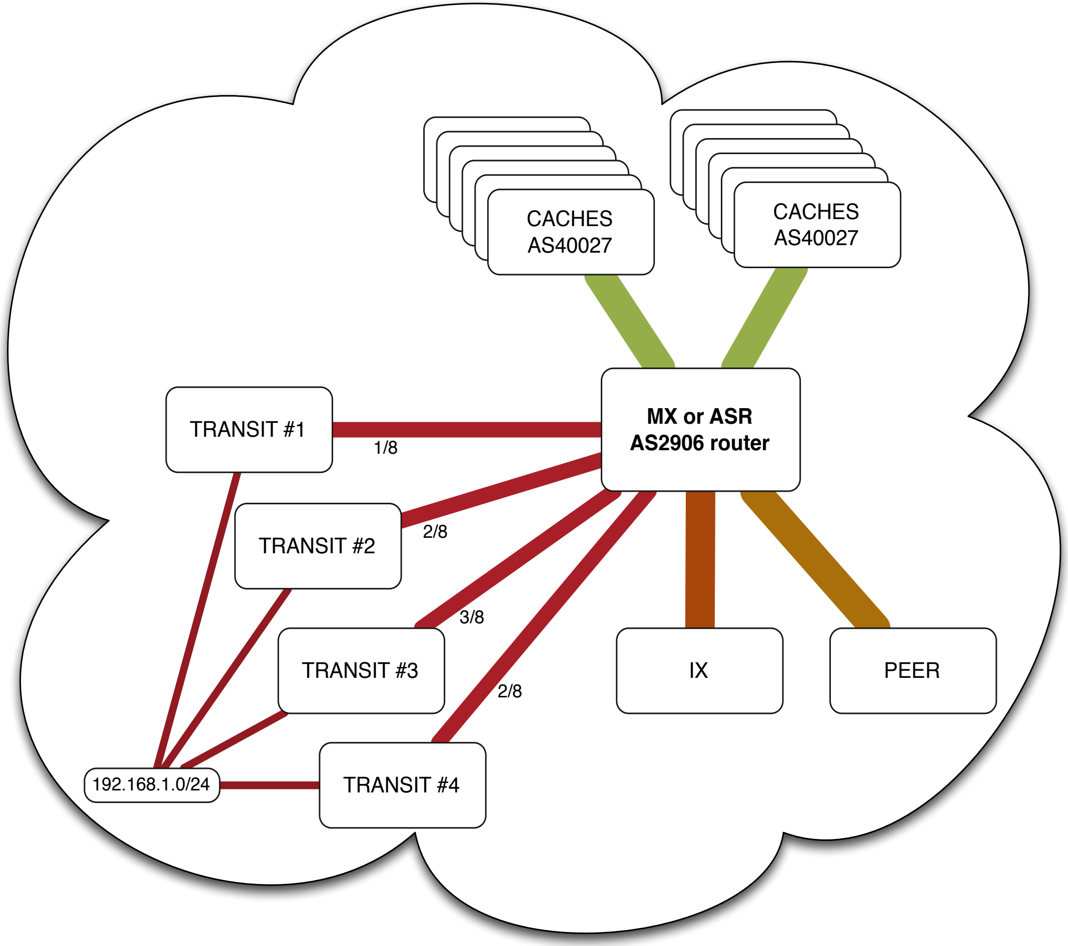
BGP routes

Data
#OITNB June 6

Client
browser, ps3, tv app, etc.

Egress BGP Hacks

- In many cases, too much traffic for 1,2 or even 4 egress partners to handle
- Use of multi-path via different ASN’s

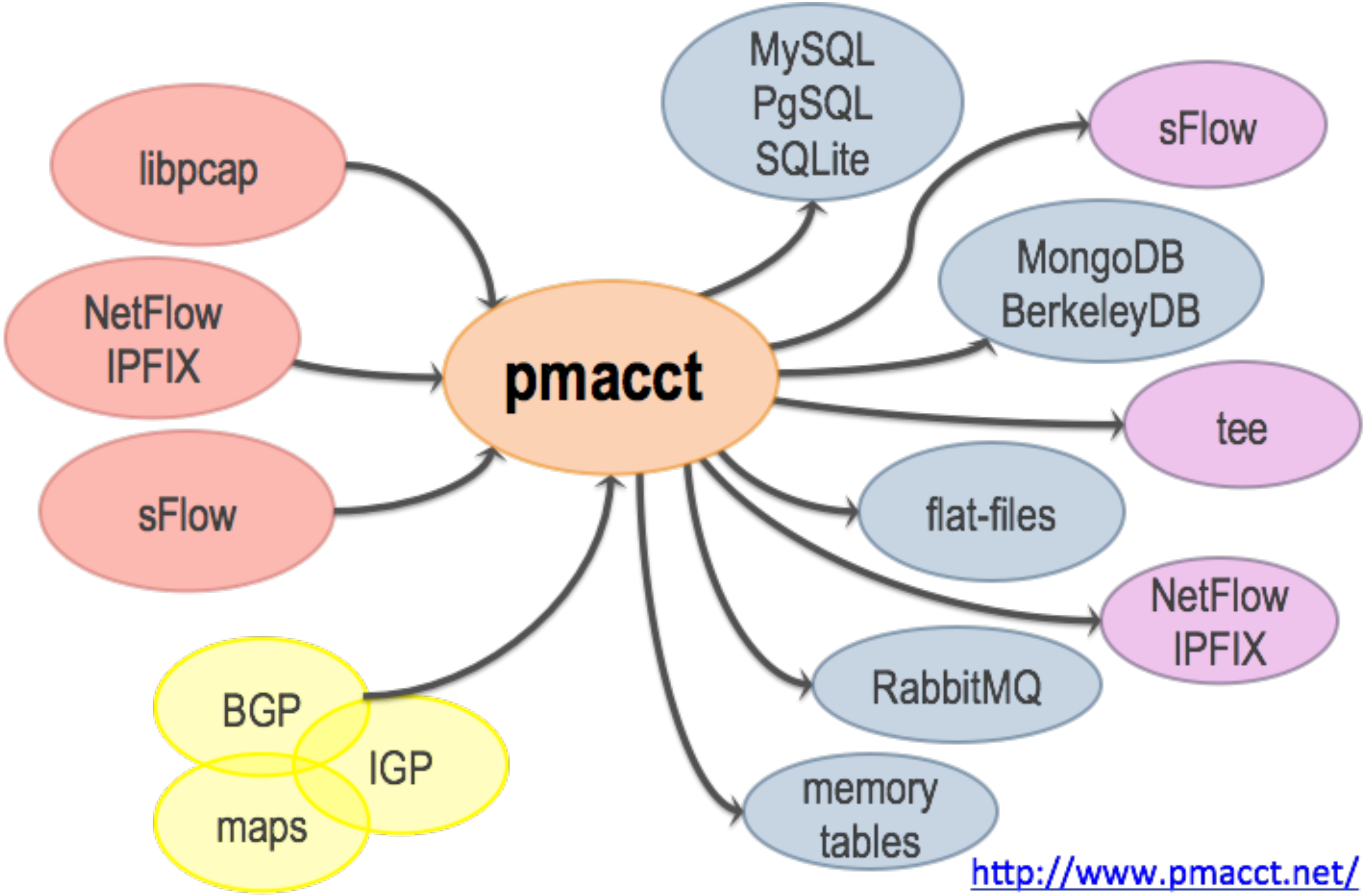


Flow Accounting at Netflix

- Primary goal: peering analysis
  - How much traffic is being exchanged with which ASN?
  - How do they perform?
- Software: pmacct
  - NetFlow/IPFIX augmented by BGP using pmacct
- Problem: multi-path, not only one single best path

About pmacct

pmacct is open source, free, GPL’ed software



http://www.pmacct.net/

pmacct a couple of non-technical facts

- 10+ years old project
- Can’t spell the name after the second drink
- Free, open-source, independent
- Under active development
- Innovation being introduced
- Well deployed around, also large SPs
- Aims to be the traffic accounting tool closer to the SP community needs

pmacct a couple technical facts

- Pervasive data-reduction techniques, ie.:
  - Data aggregation
  - Tagging and filtering
  - Sampling
- Ability to build multiple views out of the very same collected network traffic dataset , ie.:
  - Unaggregated to flat-files for security and forensic purposes
  - Aggregated as [<ingress router>, <ingress interface>, <BGP next-hop>, <peer destination ASN>] to build an internal traffic matrix for capacity planning purposes

pmacct and BGP

- BGP at the collector?
  - Telemetry reports on forwarding-plane, and a bit more
  - Extended visibility into control-plane information
- pmacct introduced a Quagga-based BGP daemon
  - Implemented as a parallel thread within the collector
  - Doesn’t send UPDATEs; passive neighbor
  - Maintains per-peer BGP RIBs
  - Supports 32-bit ASNs; IPv4, IPv6 and VPN families
- ~~Caveats:~~
  - ~~BGP multi-path is not supported~~ Outdated!

Brief digression on BGP ADD-PATHS

On BGP ADD-PATHS

- A BGP extension that allows the advertisement of multiple paths for the same address prefix without the new paths implicitly replacing any previous ones
- Draft at IETF: draft-ietf-idr-add-paths-09

On BGP ADD-PATHS

- New BGP capability, new NLRI encoding:

```
+--------------------------------+
| Path Identifier (4 octets)     |
+--------------------------------+
| Length (1 octet)               |
+--------------------------------+
| Prefix (variable)              |
+--------------------------------+
```

- Capability number: 69

On BGP ADD-PATHS

- BGP ADD-PATHS covers several use cases:
  - Mostly revolving around actual routing
  - Extra path flooding questioned in such context (*)
- Our use-case for BGP ADD-PATHS is around monitoring applications:
  - Not much talk yet in such context
  - Proposal to mark best-paths to benefit monitoring applications: draft-bgp-path-marking (Cardona et al.)

(*) http://www.nanog.org/meetings/nanog48/presentations/Tuesday/Raszuk_To_AddPaths_N48.pdf

Putting all the pieces together: NetFlow and BGP ADD-PATHS with pmacct at Netflix

Wait, so what’s the problem?

- BGP multi-path, traffic not only sent to a single best path
- pmacct is only aware of the best from its BGP feed

BGP Multi-path

```
192.168.1.0/24      [BGP/170] 3w0d 01:19:58, MED 100, localpref 200
                      AS path: 789 I, validation-state: unverified
                    > to 10.0.0.1 via ae12.0
                    [BGP/170] 3w0d 01:15:44, MED 100, localpref 100
                      AS path: 123 456 789 I, validation-state: unverified
                    > to 10.0.0.2 via ae8.0
                    [BGP/170] 3w0d 01:13:48, MED 100, localpref 100
                      AS path: 321 654 789 I, validation-state: unverified
                    > to 10.0.0.3 via ae10.0
                    [BGP/170] 3w0d 01:18:24, MED 100, localpref 100
                      AS path: 213 546 789 I, validation-state: unverified
                    > to 10.0.0.4 via ae1.0
```

Traditional BGP to pmacct

```
* 192.168.1.0/24              10.0.0.1       100 200     789 I
```

BGP ADD-PATHS FTW!

- ADD-PATHS provides visibility into the N best-paths

BGP Multi-path

```
192.168.1.0/24      [BGP/170] 3w0d 01:19:58, MED 100, localpref 200
                      AS path: 789 I, validation-state: unverified
                    > to 10.0.0.1 via ae12.0
                    [BGP/170] 3w0d 01:15:44, MED 100, localpref 100
                      AS path: 123 456 789 I, validation-state: unverified
                    > to 10.0.0.2 via ae8.0
                    [BGP/170] 3w0d 01:13:48, MED 100, localpref 100
                      AS path: 321 654 789 I, validation-state: unverified
                    > to 10.0.0.3 via ae10.0
                    [BGP/170] 3w0d 01:18:24, MED 100, localpref 100
                      AS path: 213 546 789 I, validation-state: unverified
                    > to 10.0.0.4 via ae1.0
```

BGP ADD-PATH to pmacct

```
* 192.168.1.0/24          10.0.0.1       100 200     789 I
                          10.0.0.2       100 100     123 456 789 I
                          10.0.0.3       100 100     321 654 789 I
                          10.0.0.4       100 100     213 546 789 I
```

pmacct and BGP ADD-PATHS

- In early Jan 2014 pmacct BGP integration got support for BGP ADD-PATHS
  - GA as part of 1.5.0rc3 version (Apr 2014)
- Why BGP ADD-PATHS?
  - Selected over BMP since it allows to not enter the exercise of parsing BGP policies
  - True, post-policies BMP exists but it’s much less implemented around and hence not felt the way to go

NetFlow/IPFIX and BGP ADD-PATHS

- OK, so we have visibility in the N best-paths ..
- .. but how to map NetFlow traffic onto them?
  - We don’t want to get in the exercise of hashing traffic onto paths ourselves as much as possible
  - NetFlow will tell! BGP next-hop in NetFlow is used as selector to tie the right BGP information to traffic data
  - Initially concerned if the BGP NextHop in NetFlow would be of any use to determine the actual path
    - We verified it accurate and consistent across vendors

NetFlow/IPFIX and BGP ADD-PATHS

NetFlow

```
SrcAddr:         10.0.1.71
DstAddr:         192.168.1.148
NextHop:         10.0.0.3
InputInt:        662
OutputInt:       953
Packets:         2
Octets:          2908
Duration:        5.112000000 sec
SrcPort:         80
DstPort:         33738
TCP Flags:       0x10
Protocol:        6
IP ToS:          0x00
SrcAS:           2906
DstAS:           789
SrcMask:         26 (prefix: 10.0.1.64/26)
DstMask:         24 (prefix: 192.168.1.0/24)
```

BGP ADD-PATH to pmacct

```
* 192.168.1.0/24        10.0.0.1       100 200      789 I
                        10.0.0.2       100 100      123 456 789 I
                        10.0.0.3       100 100      321 654 789 I
                        10.0.0.4       100 100      213 546 789 I
```

Netflix + NetFlow/IPFIX + pmacct + ADD-PATHS

- Multiple pmacct servers in various locations
- NetFlow is being exported to the pmacct servers:
  - Mix of NetFlow v5, v9 and IPFIX
- BGP ADD-PATHS is being set up between routers and the pmacct servers
  - Sessions configured as iBGP, RR-client
  - Juniper ADD-7 (maximum)
  - Cisco ADD-ALL

Thanks!! Questions?

Paolo Lucente <paolo@pmacct.net>

Elisa Jasinska <elisa@netflix.com>