Scalable Multipath Routing *(towards)*

71st Meeting of the IFIP WG 10.4 Working Group on Dependability and Security

Ian Welch, School of Engineering and Computer Science; Victoria University of Wellington
email:ian.welch@vuw.ac.nz www:https://ecs.victoria.ac.nz/Main/IanWelch
Huu Trung Truong, Bryan Ng

Friday 27th January 2017

The Talk

BGP protocol and architecture from 1990s.

Facing scalability problems.

Made worse if we **improve** BGP!

Apply standard distributed systems approaches (separation of concerns) to solve it.

Implement using SDN to extend functionality and make more flexible.

Network Architecture

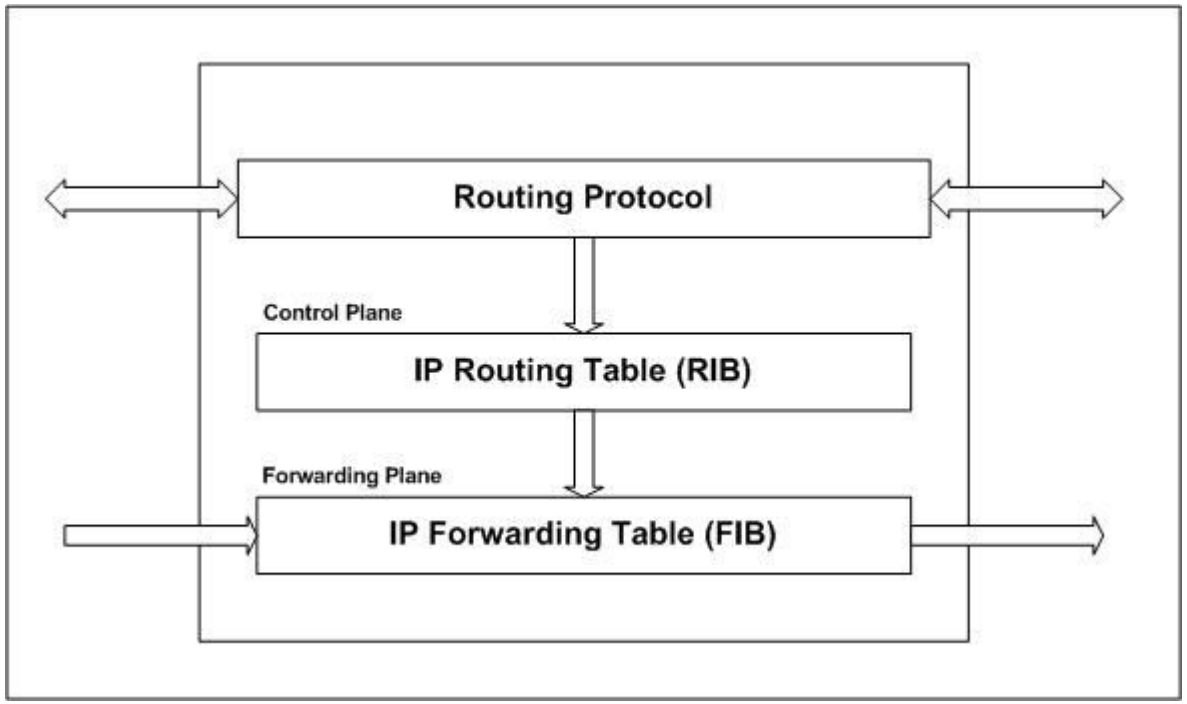
Hierarchical tiers of ethernet switches connect hosts to form local area networks (layer 2)

Local area networks connected by routers (layer 3)

Organisations use edge router (layer 3) to connect to their Internet providers.

Autonomous systems (ASes) assigned range of IPv4 or IPv6 addresses, identifiable using a prefix and subnet

Router Architecture

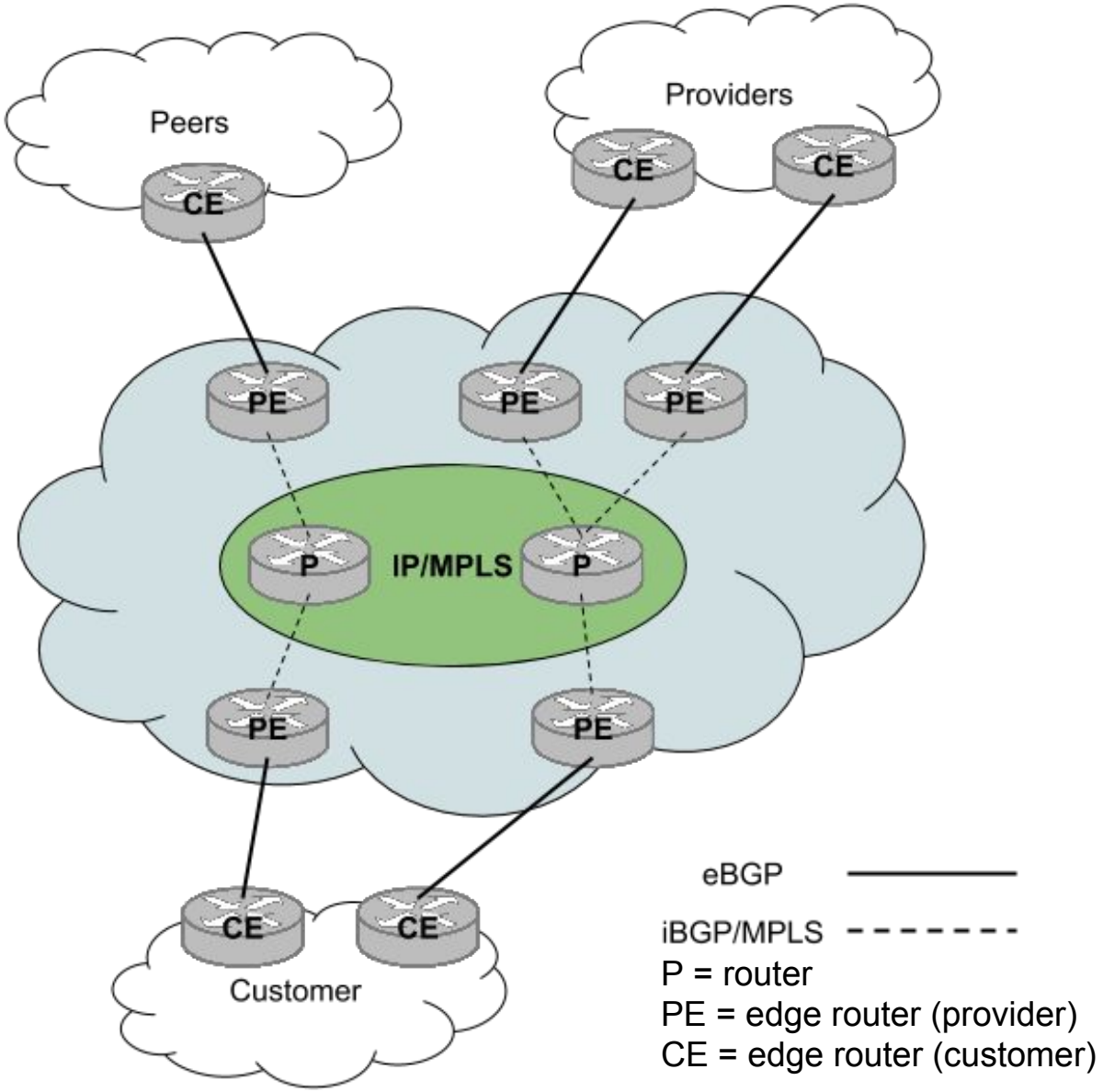


Routing protocol = distributed algorithm (exchange reachability information).

Control plane computes network paths, writes next hop information info fib.

Forwarding (data) plane uses subnet mask to work out where to send packet.

Internet Architecture



Internet service provider = AS

Transit providers ASes

Customer ASes and individuals

Edge/core routing

Border Gateway Protocol (BGP)

IETF RFC 1163 (1990): A Border Gateway Protocol (BGP)
… *used for interdomain routing*

BGP Peering exchanges routing information.

- Destination: IP prefix and subnet mask
- Paths to destination
- Metadata (communities, MED etc.)

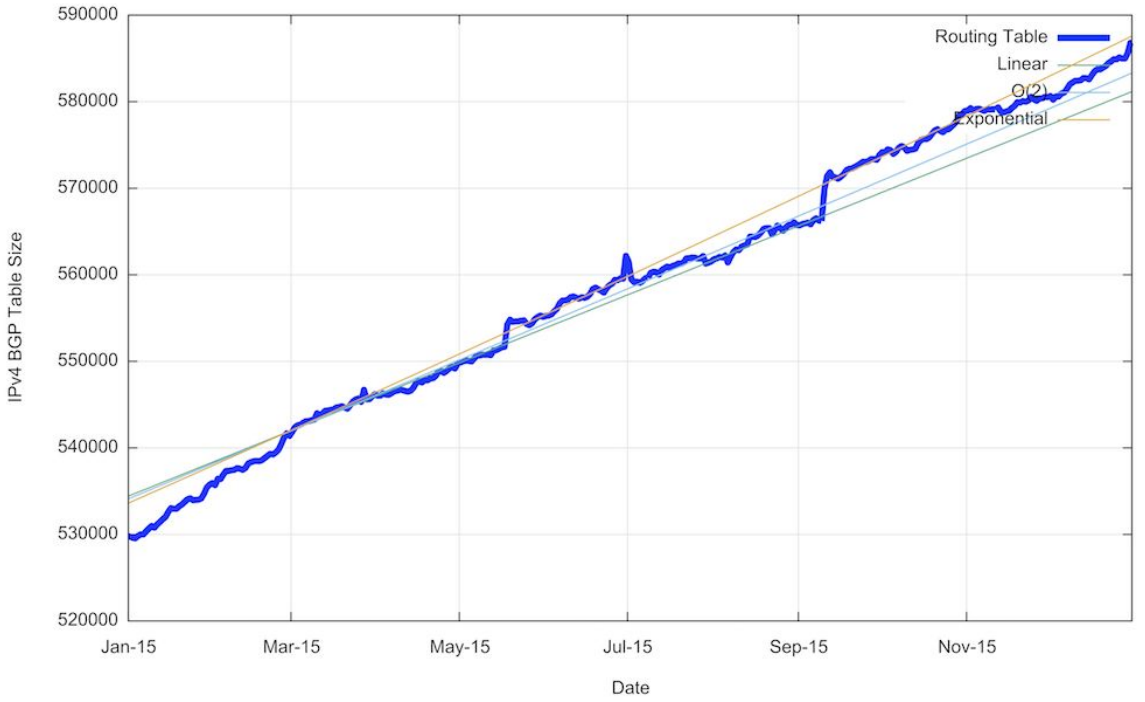
Router applies local policy to construct the FIB.

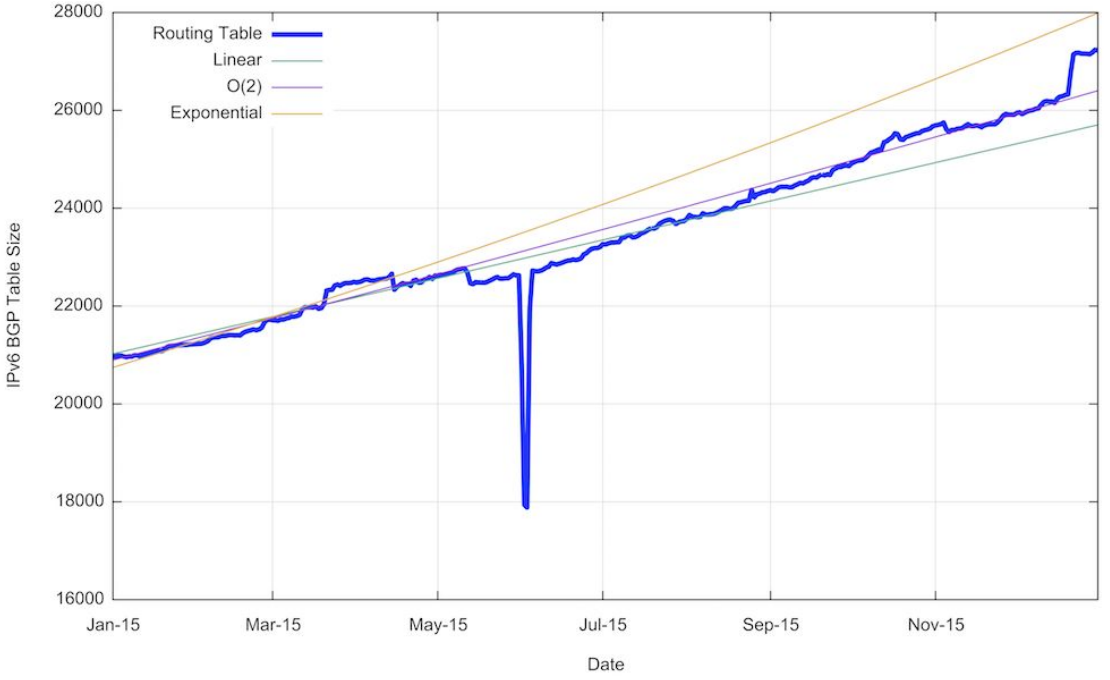
A single “best path” per destination is distributed across the Internet (# hops, $$$$, politics, beer ...)

Scalability

Asia Pacific Network Information Centre analysis (Geoff Houston, 2015)

Table size growing linearly for IPv4 and O(n^2) for IPv6





Why is happening?

IPv4 space exhaustion
Smaller ranges of hosts => more entries

Traffic engineering requirements
multihoming ASes => split subranges

+20% more IPv6 hosts
IoT should make this worse

Same IP range, located in different geographical areas
worse with rise of mobile devices

Why care about scalability?

FIB implemented using content addressable memory
Full FIBs => unreliable routers.
Vertical scaling isn't possible.
Horizontal scaling is expensive/manually intensive.

Control plane processing, increasing rate and number of announcements
Overstressed protocol layer => unreliable routers.
Same scaling issues as FIB above.

Unreliable routers => failed routing => unrelaible routes.

ALSO … having a single path is bad for reliability and quality of service

Different applications = different “best path”
Latency, bandwidth, jitter, political reasons …

Empirical evidence “best path” = suboptimal [Savage et. al. 1999]

Single path

- limits recovery after a failure
- recompute FIB if next hop failure
- wait or upstream routing information if upstream failure

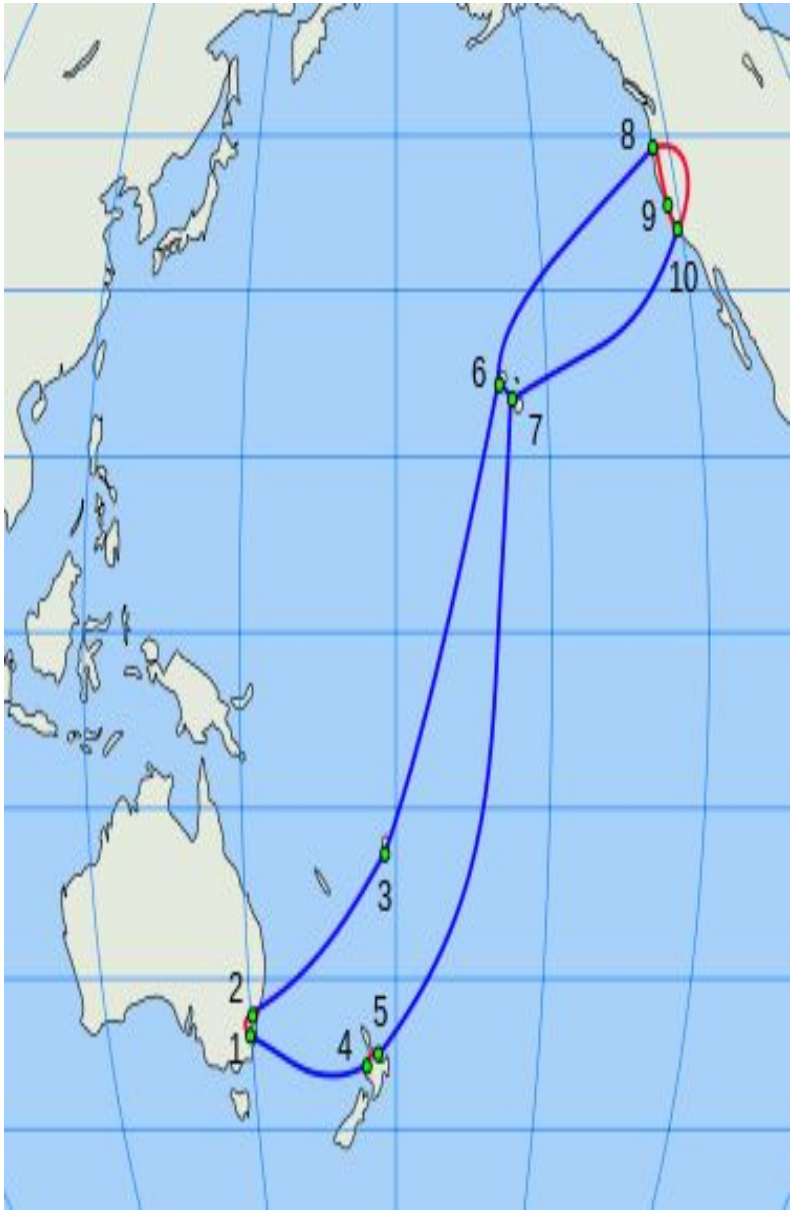
Provide choice - Multipath routing

Multiple paths allow:

- fast failover
- traffic engineering (split bandwidth)
- per-application qos

BUT makes scalability worse:

- control plane complexity
- FIB exhaustion

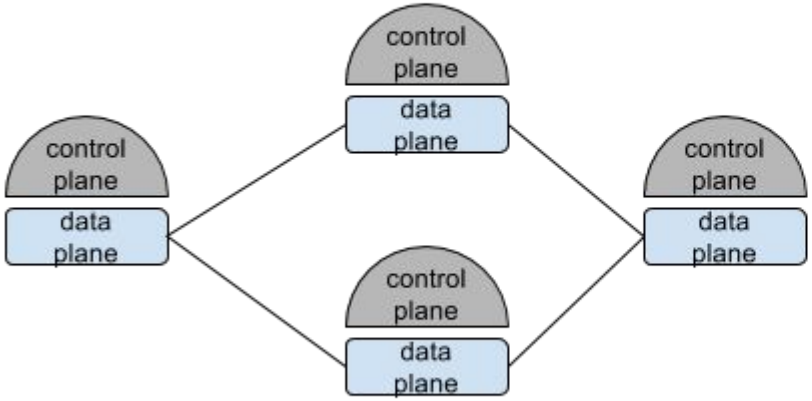


SDN-based architectural solution

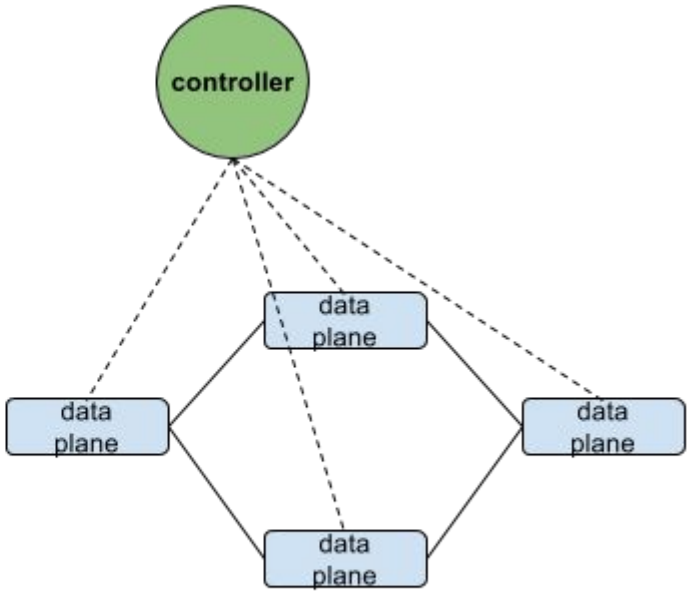
Separate control plane & forwarding plane.

General purpose hardware for control plane.

Multiple switches for forwarding plane (simpler = more reliable?).

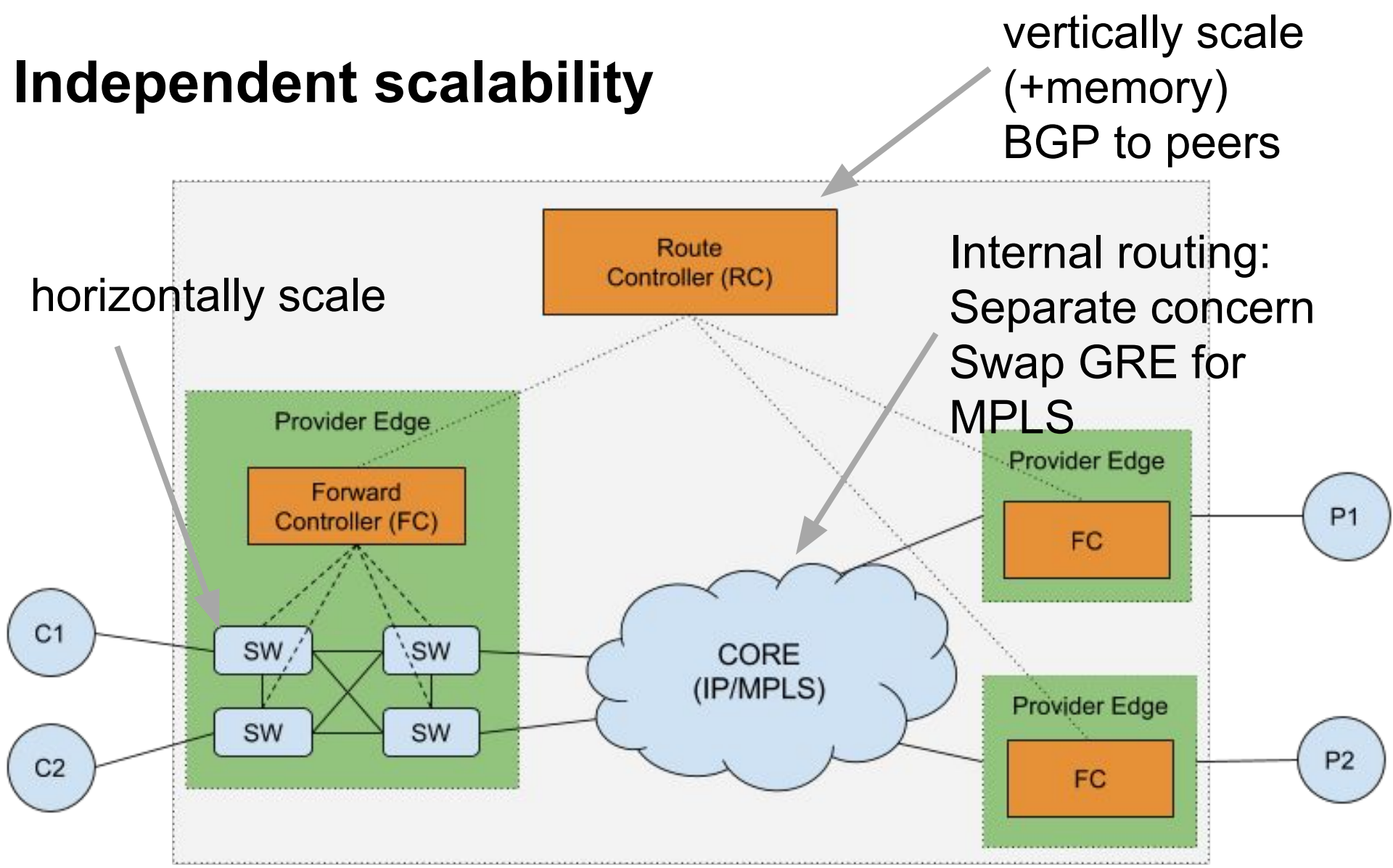


Traditional networking



SDN networking

Independent scalability



Minimising size of FIB on switches

FIB is stored in switch.

Could store prefix+next hop in single table (inefficient).

Split into pipeline of three to minimise space usage.

- Classifier: map L2&L3 metadata -> traffic type
- Multipath: map traffic type + dest -> next hop
- Forwarder: forwarding actions (MAC address manipulation, use of MPLS for internal paths)

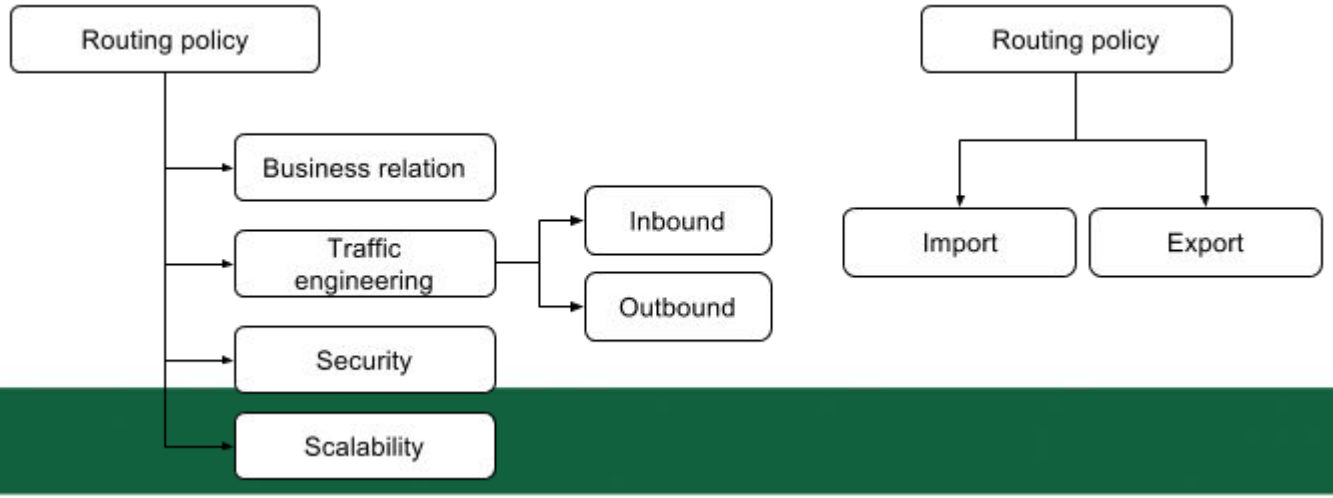
Choosing between multipaths

BGP is limited in expressibility.

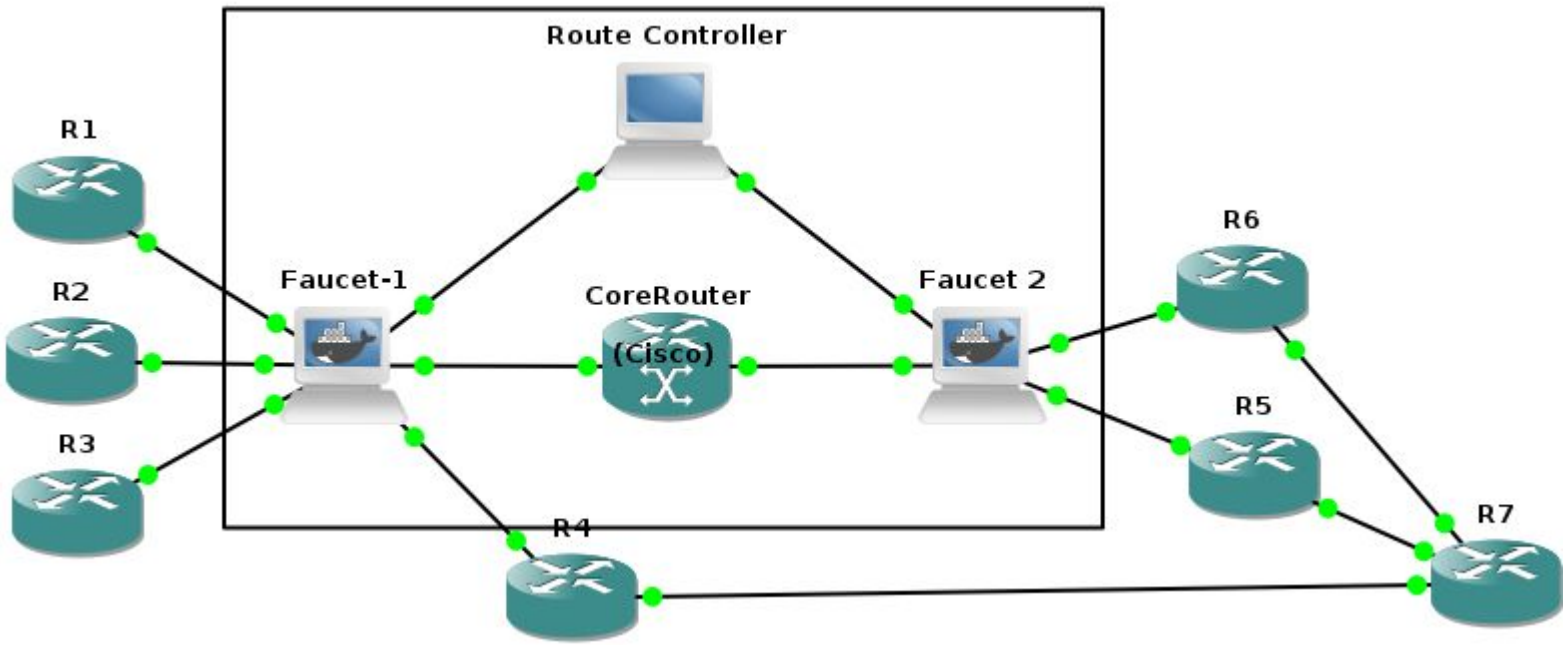
Routing Policy Specification Language (RPSL)

- Not part of BGP
- Used by widely in practice (route reflectors)

Extensible, we implemented simple preference scheme (fast failover between two routes).



Prototype implementation



ONF Openflow flavour of software defined networking

Faucet layer 2 switch (opensource)
Controls switches via open flow protocol
Updates rules used for forwarding at line speed

Current work

Extend classifiers:

- Support per-application paths
- What about dynamic classification?

More sophisticated policy:

- Extend meta data
- Tension between customer requirements and provider

Some existing work on unipath selection requiring extension and tradeoff of customer/provider needs (Morpheus project -- Rexford and co.).

Future directions

Going beyond single AS or ISP.

- gain by having a global view.

Dependability.

- controller replication.
- policy management.