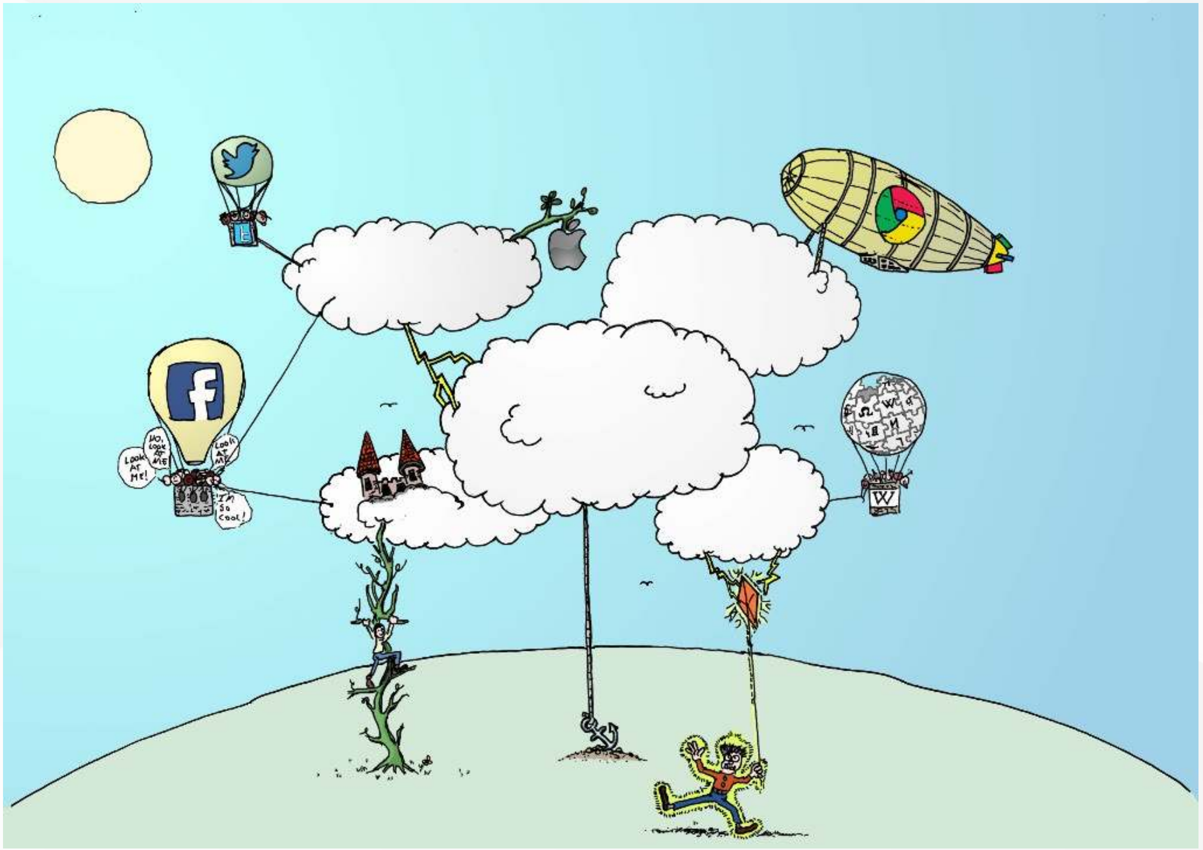
amsix
amsterdam
internet exchange

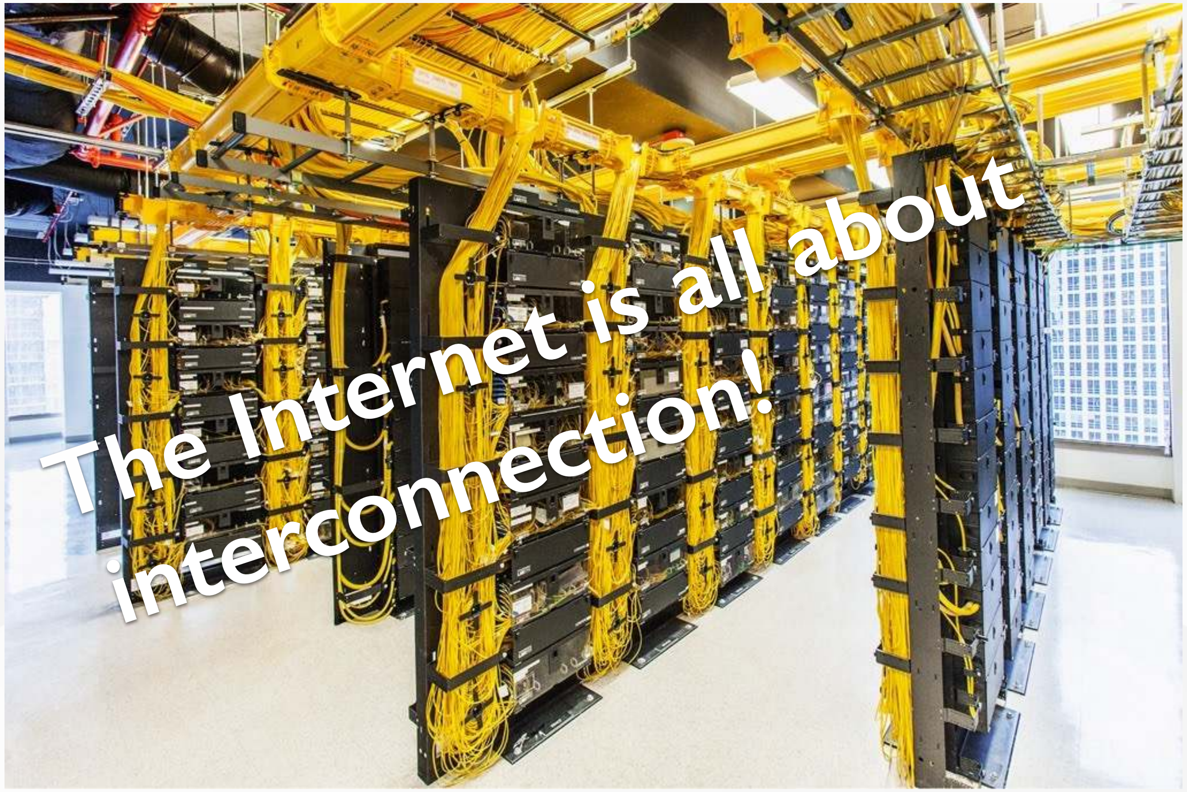
Interconnection, Peering IXPs

What and How

Interconnection



Interconnection



Interconnection

Typically Interconnection between networks in the Internet is implemented in two ways

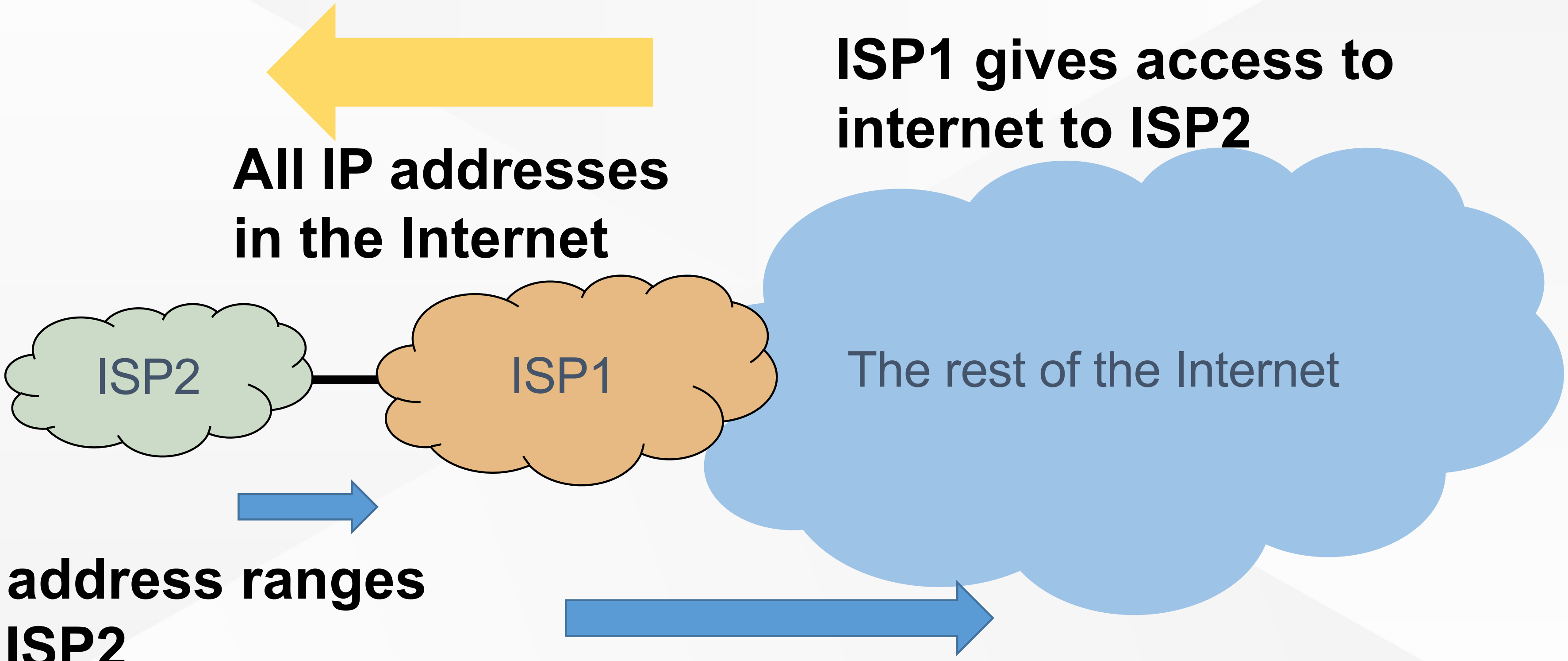
- Transit
  - Buy interconnection to the rest of the internet from a service provider
- Peering or direct interconnection
  - Interconnect directly to other networks

Interconnection

- Interconnection is implemented physically by creating a connection between two routers
  - Physical media Fiber or sometimes still copper
  - Datalink layer almost always Ethernet (IEEE 802.3)
  - Physical layer: 802.3.. (1, 10, 100GE etc)
    - We see the first customer requests for 400GE

- Logical interconnection is implemented using eBGP
  - Advertise reachability information between Autonomous Systems (AS)
    - AS is an identifier for a network
    - The reachability information in eBGP consists of the IP (v4 or v6) address ranges that are part of the AS to be announced
  - Each router calculates shortest path (in AS hops) to destination

Interconnectievormen: Transit



ISP1 gives access to internet to ISP2

All IP addresses in the Internet

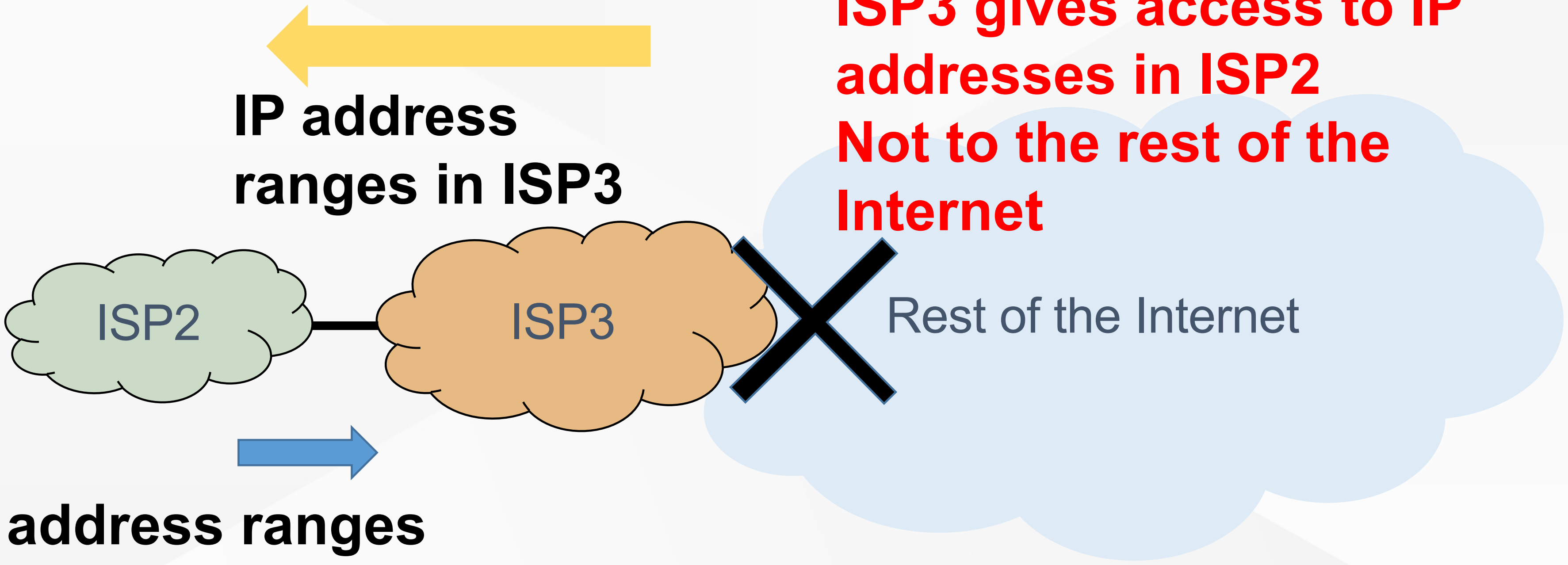
IP address ranges in ISP2

IP addresses
Ranges in ISP1 plus ISP2

Peering

- The Exchange of traffic between parties where only each others' customers are advertised is called peering
- "Peer" stands for “equal party”
  - Large carriers peer with large carriers and small ISPs with small ISPs
  - Providers peer where there is equal gain
- Peering typically happens without financial settlements but not necessarily
  - Specifically not one party is much larger or has more negotiating power than the other
- Benefits of peering:
  - Reduced need for upstream connectivity, thus lower costs for exchanging IP traffic
  - Shorter paths between networks, thus faster data flows (lower latency, less jitter)

Interconnectievormen: Peering



IP address ranges in ISP3

IP address ranges in ISP2

ISP3 gives access to IP addresses in ISP2
Not to the rest of the Internet

Why Peering?

- Transit is easy, but ….
  - By definition you add always at least one AS hop to your destination
    - Unless the destination is the transit provider itself
  - Quality of traffic flows are dependend on quality of networks between you and destination
    - Transit provider can give quality assurances on its own network but not on other networks in the path to destination
  - Although transit pricing is still declining it can still be costly
    - Depending on location in the world
    - Depending on who buys

Peering Implementation

- Direct connection (private interconnect, most common)
  - Two routers co-located (in same datacenter) interconnected by means of a direct fiber connection.
  - Can become cumbersome if you have hundreds of peers in one location

- Multiple routers (more than 2) connected to a shared infrastructure
  - Internet Exchange Point (IXP)
  - Single physical connection but allows for multiple logical connections
    - For example on AMS-IX with this one connection you can peer > 800 other networks
  - If IXP extends to multiple datacenters no need for routers to be co-located

Peering

- Peering needs to be arranged
  - Transit you can “just” buy
  - Peering needs to be managed
    - Especially since Peering always goes together with transit as you never can peer with all the networks in the internet
      - Exception being the few “Tier 1” transit free operators
    - Traffic engineering
      - Do I set up peering to reach a network or do I use transit
      - Is it worth to go to another IXP instead of transit
    - On a large IXP as AMS-IX you have the possibility to peer with over 800 networks

Peering

- Need to contact the other network (peering coordinator) and agree on peering, i.e. agree on a common interest and roughly equal gain
    - Often just an e-mail is enough, many networks on an IXP advertise they have an open peering policy and peer with anyone
    - At gatherings of peering coordinators
    - Global or Regional peering events
    - RIPE/Nanog/Apricot/Sanog, etc.

Changes in Peering on AMS-IX

- Originnally mostly (eyeball) ISPs with some content in their own networks
- Later a mix of ISPs and content providers
  - This evolved in AMS-IX becoming a distribution point for content.
    - Big traffic streams from content providers to ISPs
- Big traffic streams moved from AMS-IX to private interconnects
  - AMS-IX used for the “long tail” of peering
- Large ISPs moved away from AMS-IX to better control interconnection

Col-Location: Equinix AM5

Equinix AM5 Amsterdam ZO

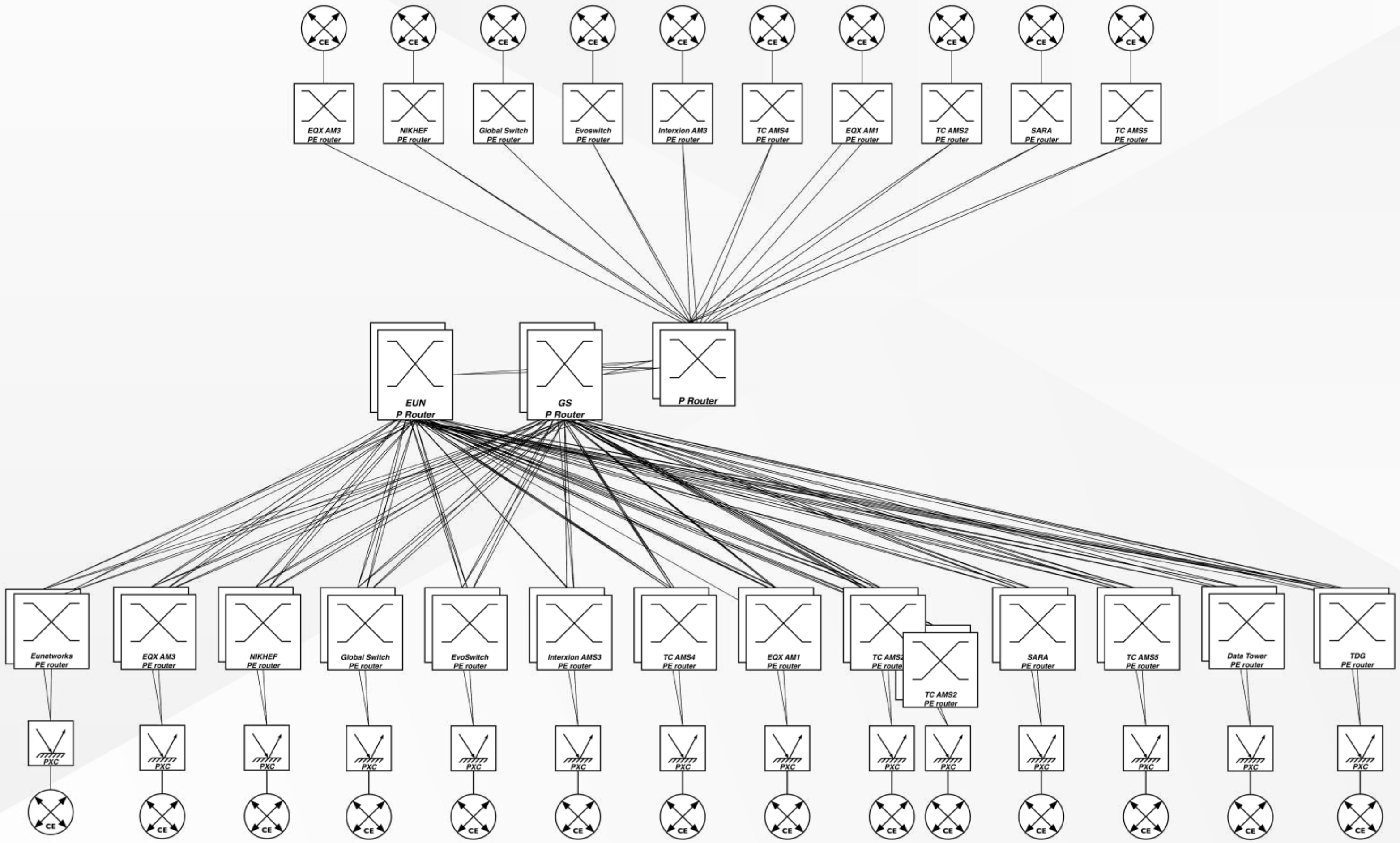
Meet Me Room: MMR

amsix
amsterdam
internet exchange

AMS-IX Platform and Infrastructure

Typical AMS-IX Cage

AMS-IX Amsterdam Platform



Customer router

Low Speed access

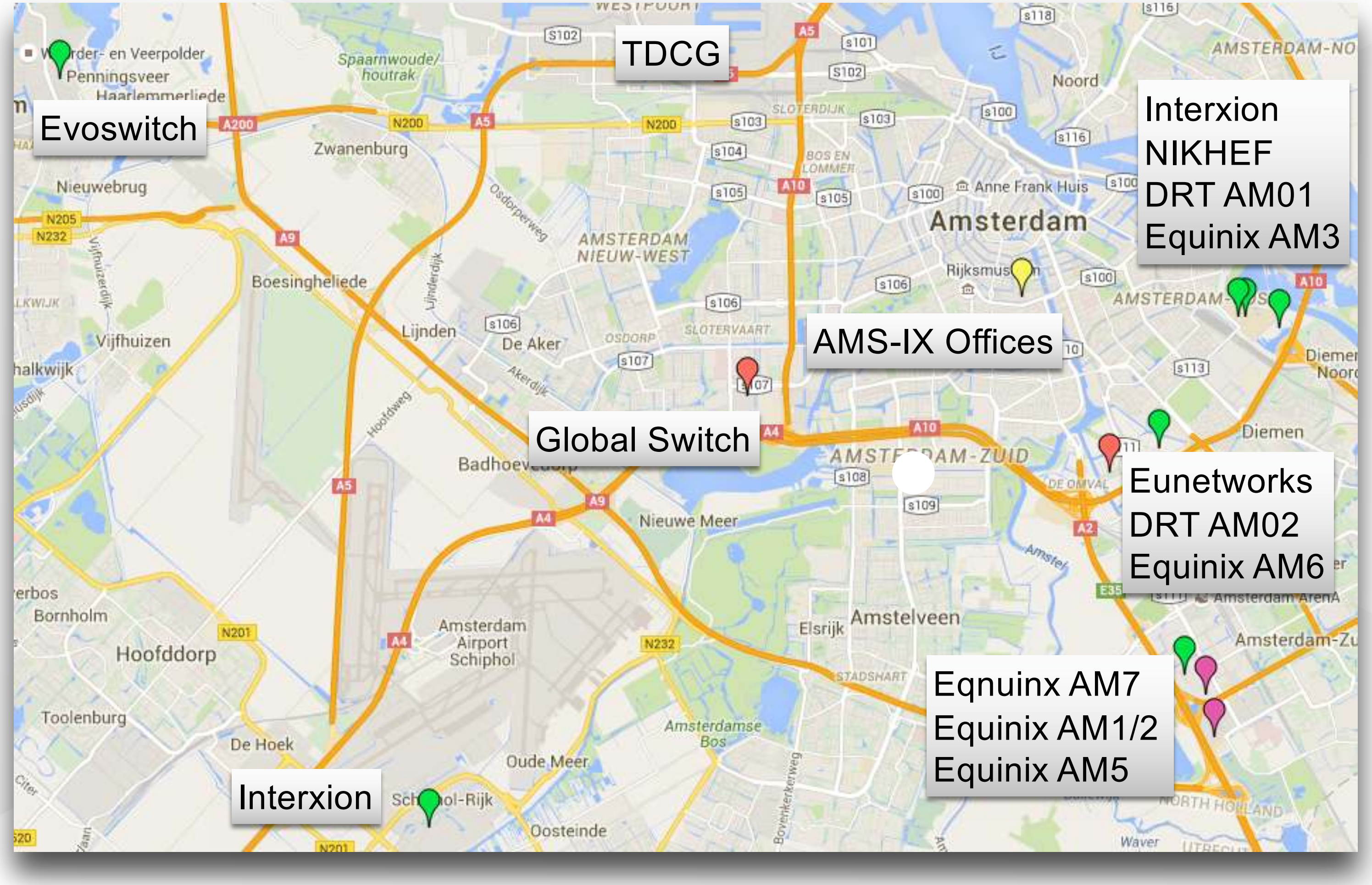
Core or Spine

High Speed access

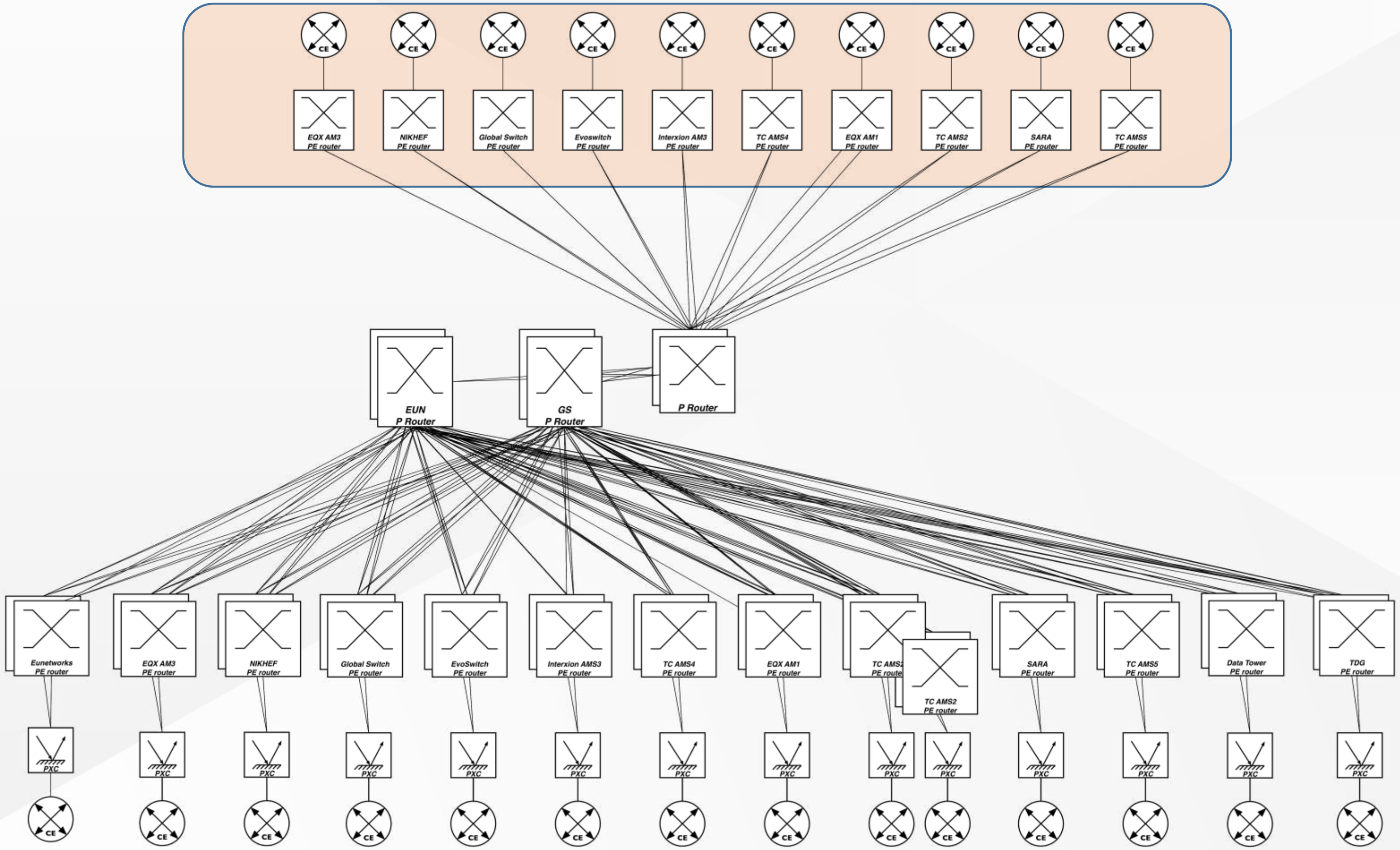
Optical access

Customer router

AMS-IX in Amsterdam



AMS-IX Amsterdam Platform



Customer router

Low Speed access

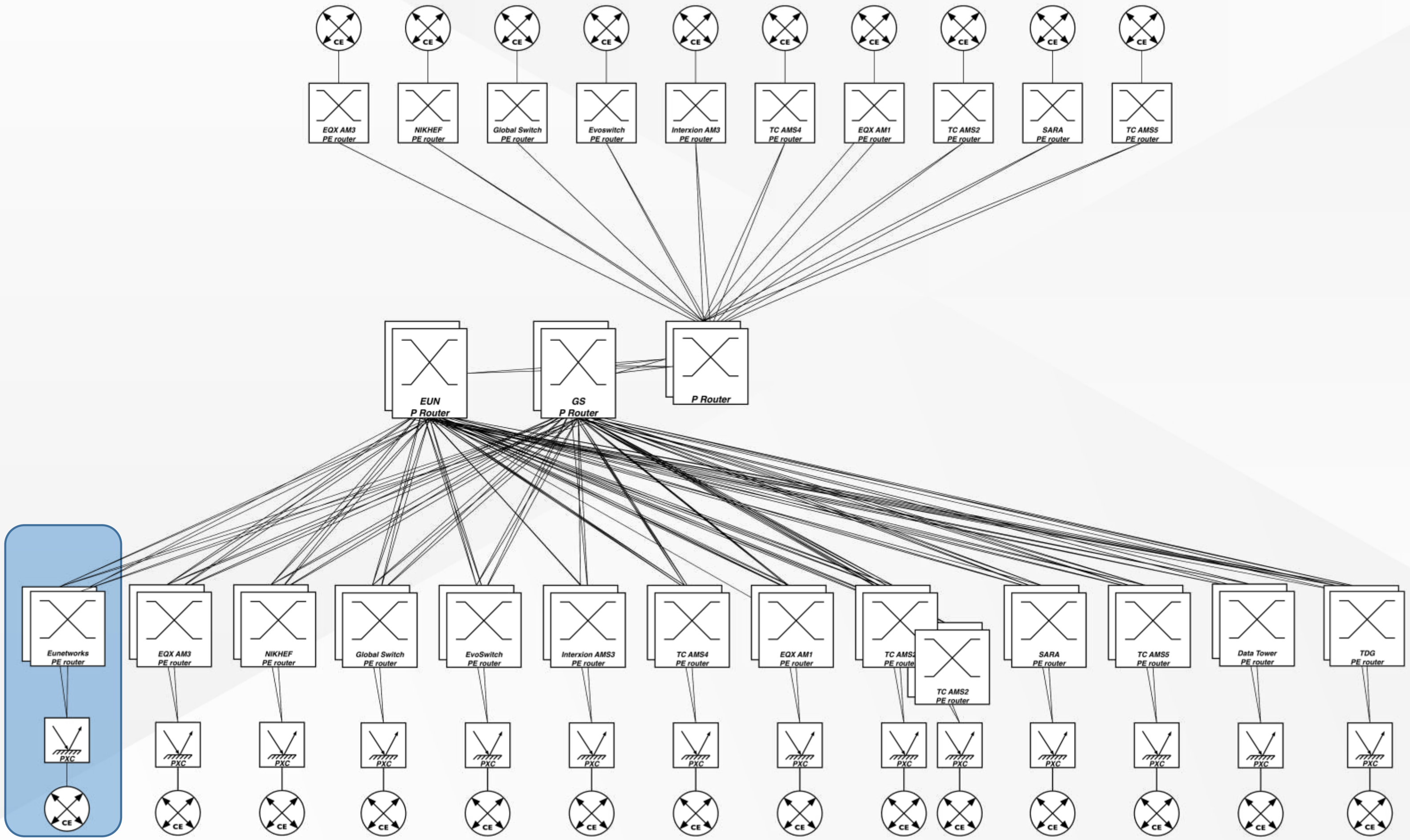
Core or Spine

High Speed access

Optical access

Customer router

AMS-IX Amsterdam Platform



Customer router

Low Speed access

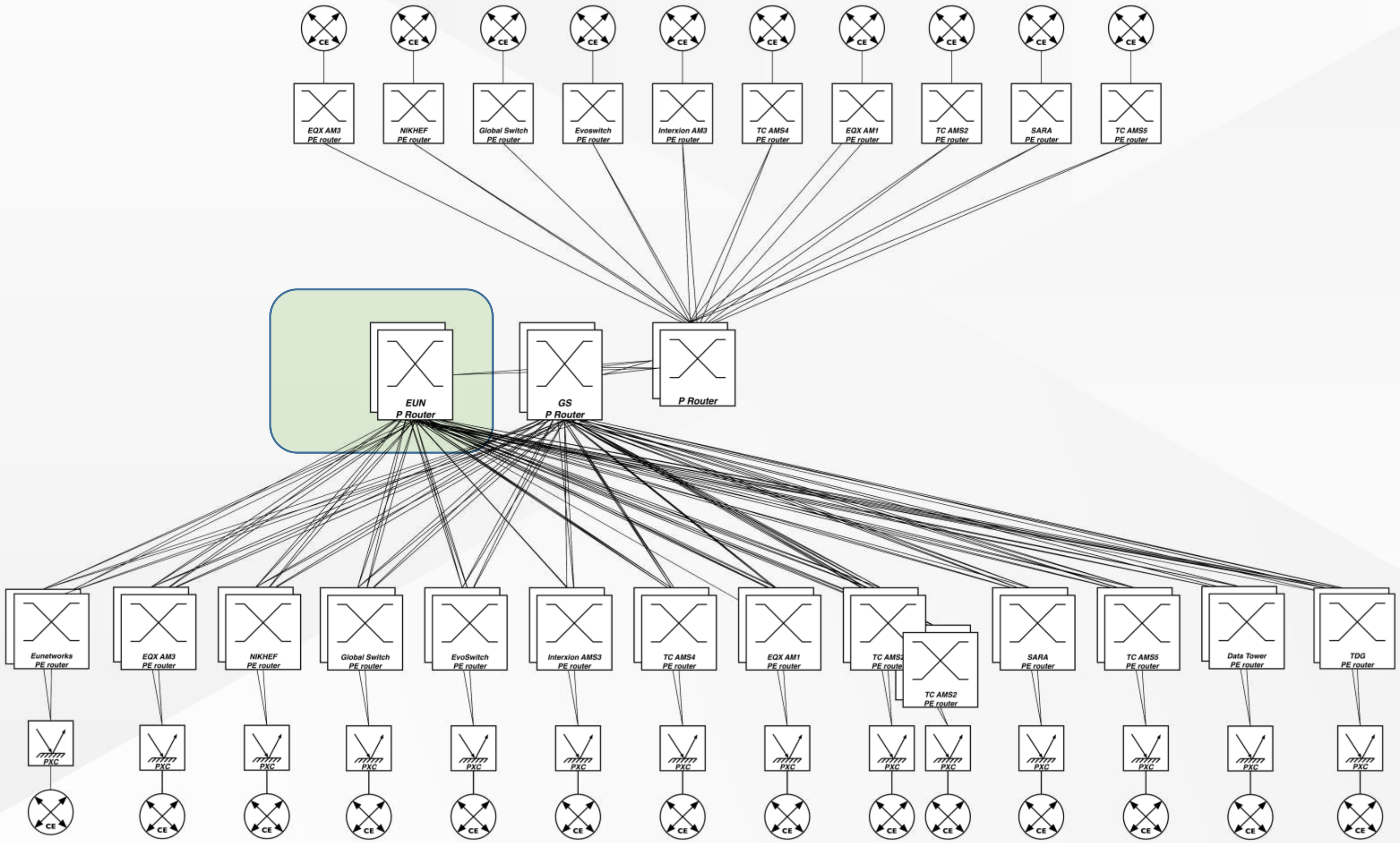
Core or Spine

High Speed access

Optical access

Customer router

AMS-IX Amsterdam Platform



Customer router

Low Speed access

Core or Spine

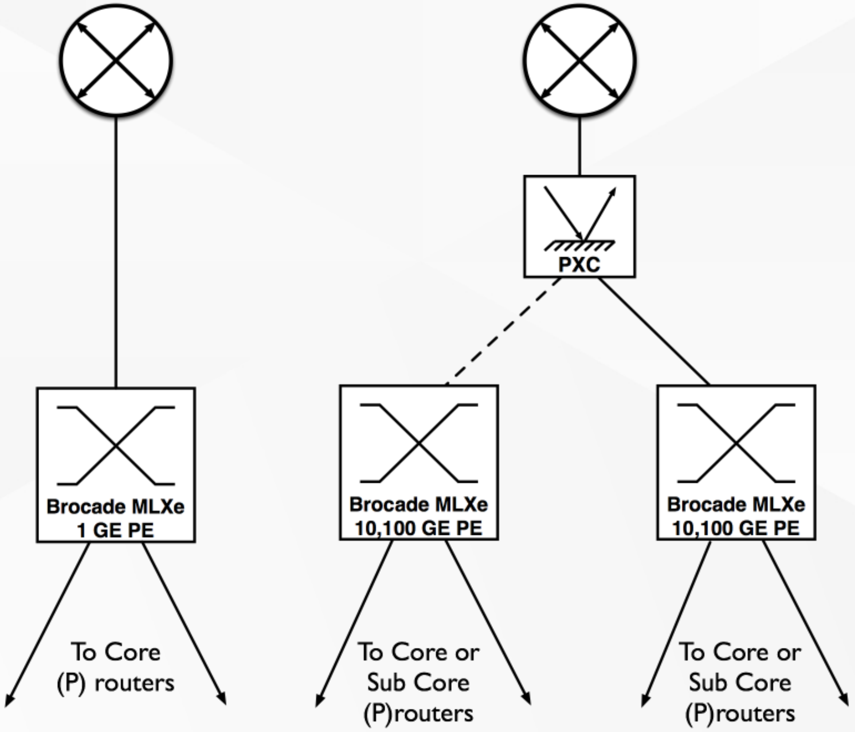
High Speed access

Optical access

Customer router

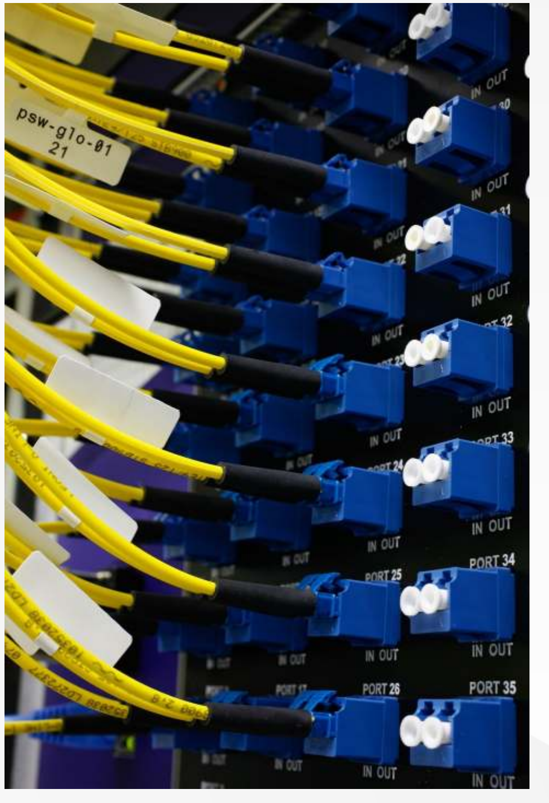
Access Connections

High Speed Access connection protected

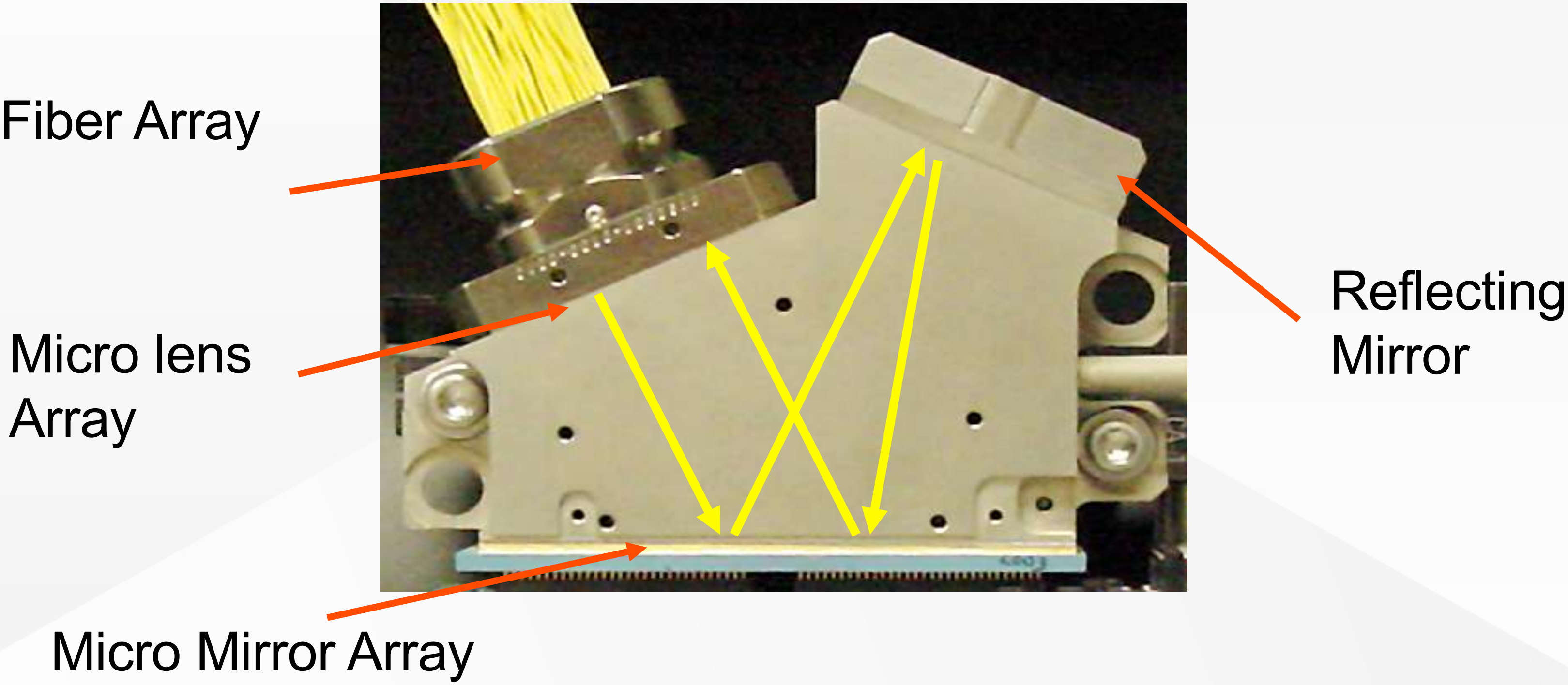


Photonic Switching

- Glimmerglass Networks switch
- 64 to 192 port MEMS based switch
- Connect any port to any other port



Glimmerglass PXC: Switching engine



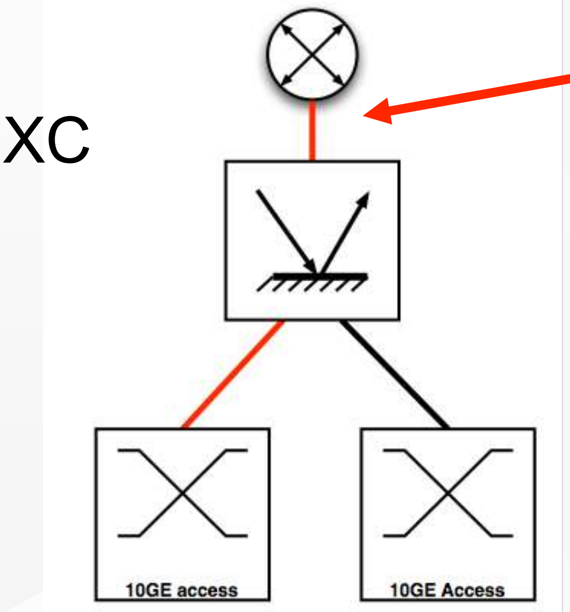
PXC Application

- PXC used for protection of CE to PE
  - Swap connection between identical pair of PEs
  - Hard and software failures on PEs manageable
  - Helps in troubleshooting
  - Allows for non service interrupting maintenance

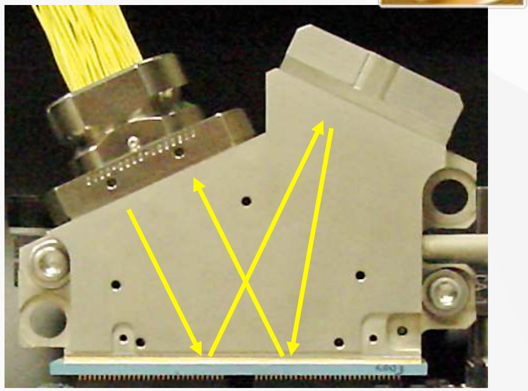
The Platform

X * 10GE, X >= 1

PXC

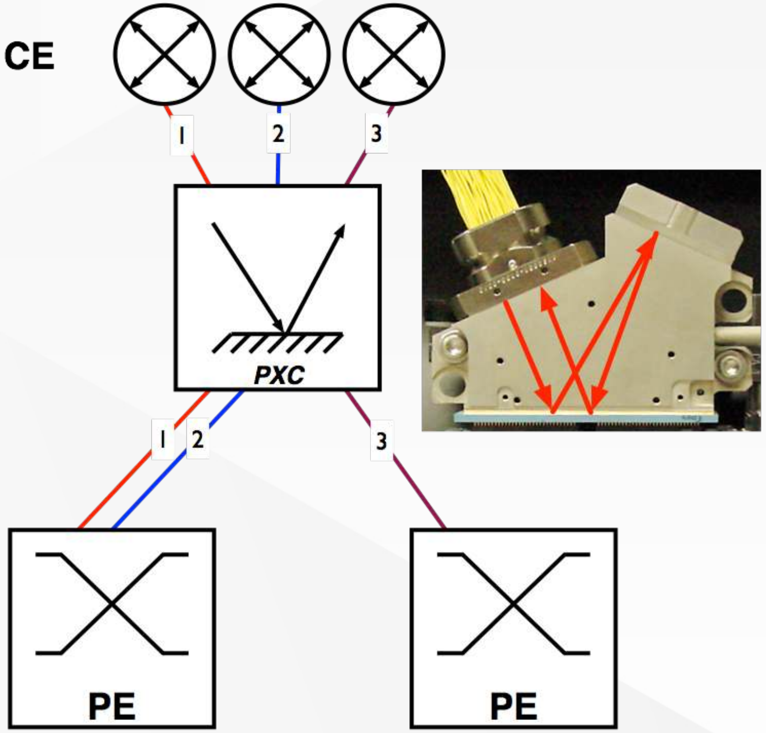


10 and 100 GE PE



PXCD

- PXCD
  - Manages Photonic Cross Connects
  - Directs failover of customer connections beween pair of PEs
  - Triggers are manual or events in the platform
    - LSP up/down

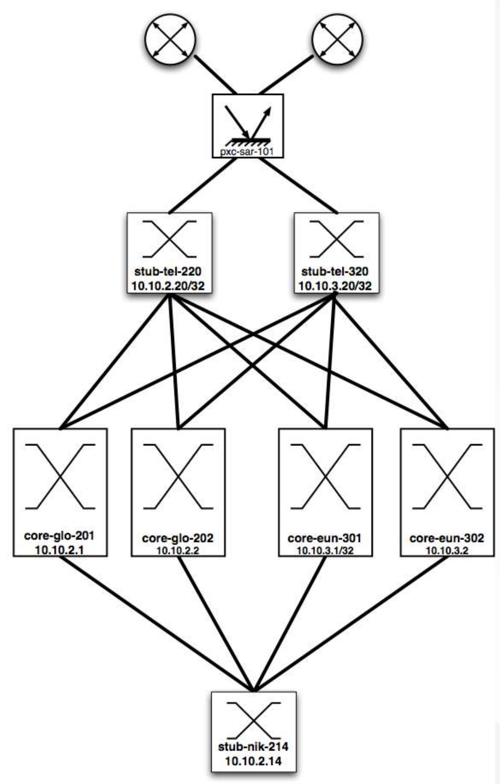


AMS-IX Technical Infrastructure

The MPLS setup

AMS-IX Platform

- MPLS/VPLS-based peering platform
  - X LSPs between each pair of access switches
  - over one or more core (P) routers
  - Load balancing of traffic over multiple LSPs
- 10/100GE access switch resilience
  - 10/100GE customer connection on PXC
  - Protection of access connection



AMS-IX Platform

- OSPF
  - BFD for fast detection of link failures
- RSVP-TE signaled LSPs over predefined paths
  - primary and secondary (backup) paths defined
- VPLS instance per VLAN
  - Static defined VPLS peers (LDP signalled)
  - Load balanced over parallel LSPs over all core routers
- Layer 2 ACLs to protect customer port

AMS-IX Platform

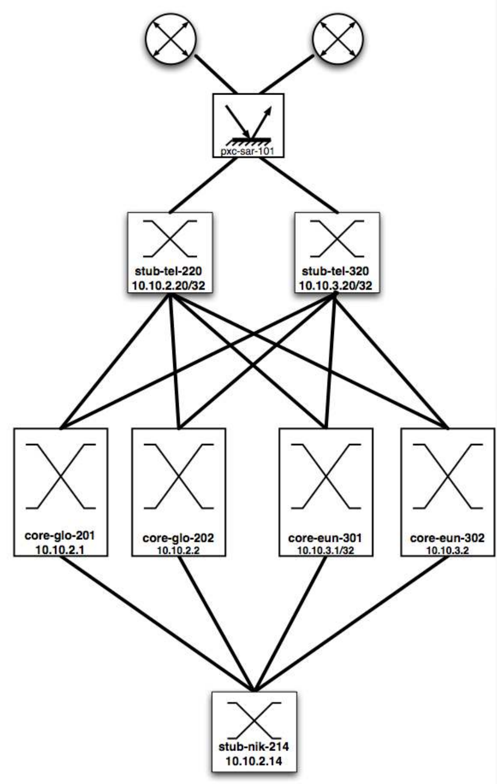
- Single OSPF area
  - Loopback addresses and backbone links in OSPF
  - Choice for OSPF (instead of ISIS) arbitrary based on available expertise

- BFD for rapid detection of failure in forwarding path
  - Bi-directional Forwarding detection
    - Detect faults in bi-directional path between two forwarding engines
    - Allows for very fast convergence of OSPF in case of link failure
  - *bfd interval 50 min-rx 50 multiplier 10*

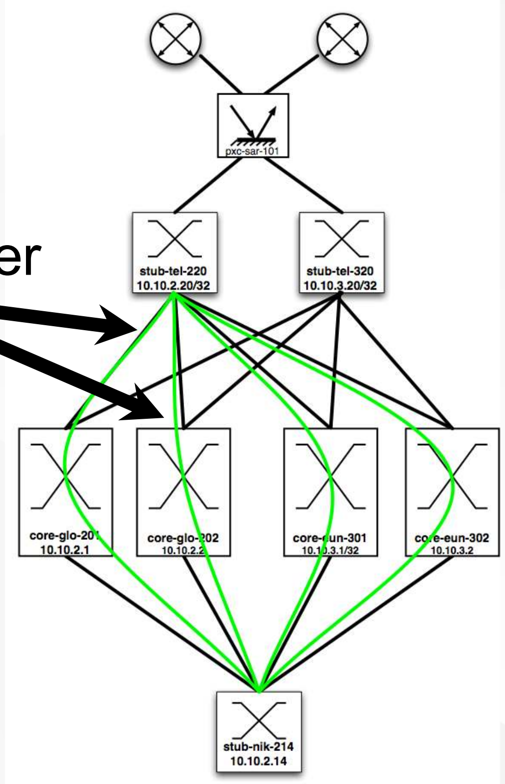
AMS-IX Platform

- Access switches (PE) act as Label Edge Router
- Core (P) act as transit Label Switch Router
  - Penultimate, label is popped on core instead of egress LER
- LSPs follow pre-defined paths through the network
- RSVP-TE for LSP signaling

MPLS/VPLS setup: LSP Definitions

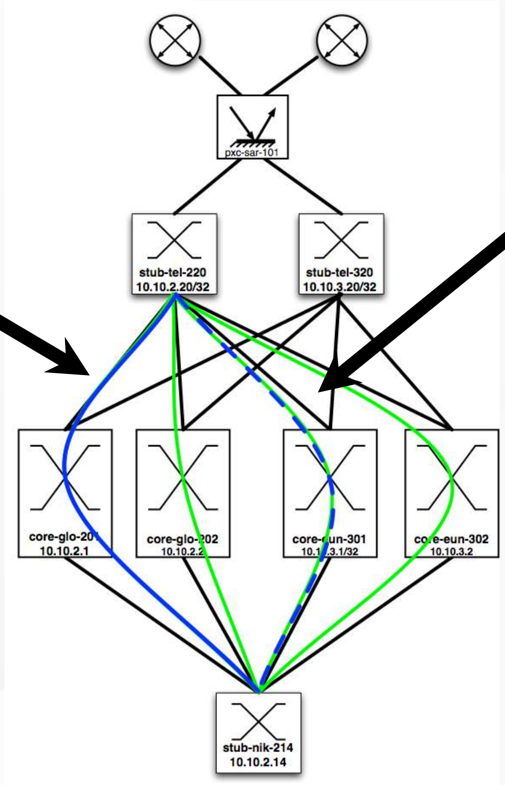


Pre-defined paths
between PEs
over each core router

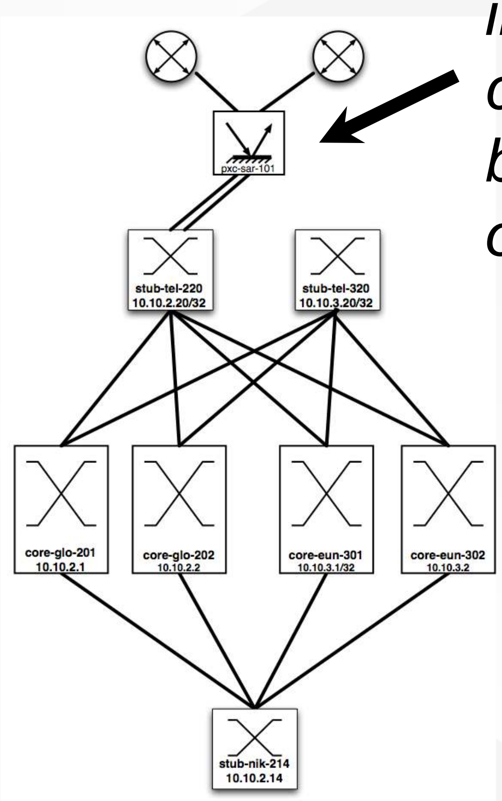


MPLS/VPLS setup: Resilience

LSP over primary Path



LSP over backup Path



Resilience in access connection by means of PXC

AMS-IX Platform
VPLS: Multipoint to Multipoint VPN

- VPLS to emulate the shared L2 infrastructure
  - LDP used in control plane.
    - Distribution of VPLS labels and MAC addresses
  - PEs pre-defined
  - Full mesh of LSP (virtual circuits) between each PE (access) device
    - Actually X LSPs (one over each core) between each pair
    - Manually configured
  - Traffic between pair of PEs load balanced over these X LSPs
  - Association of customer interface (L2) to VPLS instance
  - One VPLS instance per VLAN
  - Loop free as by default no packets arrived over an LSP is forwarded on another LSP

ROUTE SERVER

Basic About BGP Routing & The Internet

Key Concepts – Autonomous System

Regional Internet Registry (RIR)	Government Independent Body who manage and assign internet resource (IP/AS). There are 5 RIR for each region of the world APNIC - Asia Pacific AfriNIC - Africa ARIN - North America LACNIC - South America RIPE - Europe and Middle East
Autonomous System (AS)	Represent the network of a company or an organization on the Global Internet
Autonomous System (AS) Number	Unique Number given to an AS by the RIR (Regional Internet Registry). A company/organization can have more than one AS numbers
AS Path	Path from one AS to another AS which can consist multiple AS. I.E. **AS_PATH: 6939 4826 38803 56203**

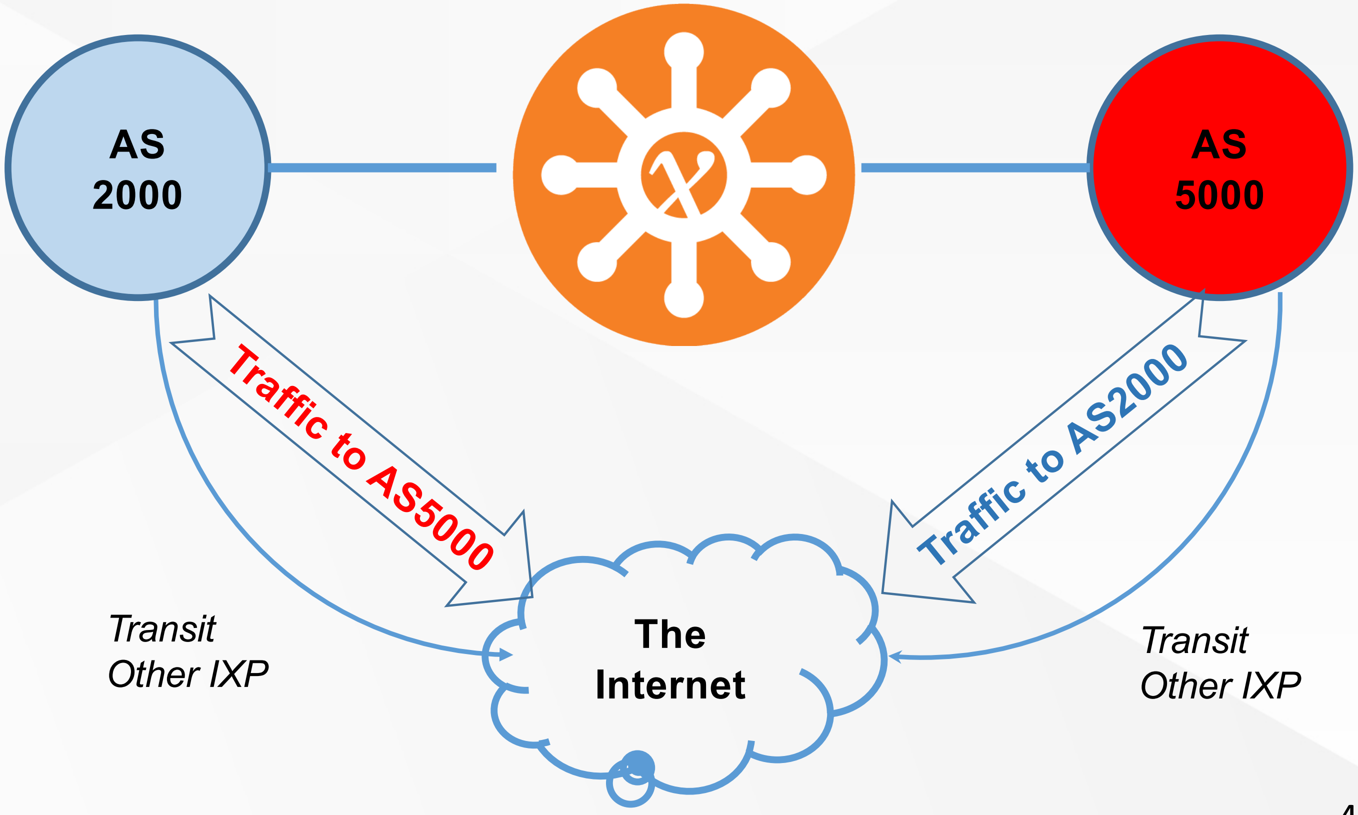
Basic About BGP Routing & The Internet

Key Concepts – IP/Router/Border Gateway Protocol

IP address	**Internet Protocol Address**, address given to device connect to the internet. There are two IP versions; IPv4 and IPv6, which is not inter-operable
IP prefixes	A group of IP address in the same range
NLRI	Network Layer Reachability Information; use by router to decide which path to forward internet traffic. Also known as **BGP prefixes**
Router	Device use within network to forward internet traffic base on IP
Border Gateway Protocol (BGP)	Routing Protocol use to exchange NLRI between routers, current on version 4 (BGP-4)
Global Routing Table	Table consist of EVERY known IP prefixes on the internet
BGP Transit	Provide gateway to Internet for a network via BGP Global Routing Table
BGP Peering	The process of exchanging **NLRI** information between two routers via BGP
BGP Peering Session	The application level session between 2 routers to exchange **NLRI**, setup using TCP/IP

BGP Peering on Internet Exchange Platform
Why BGP peering ?

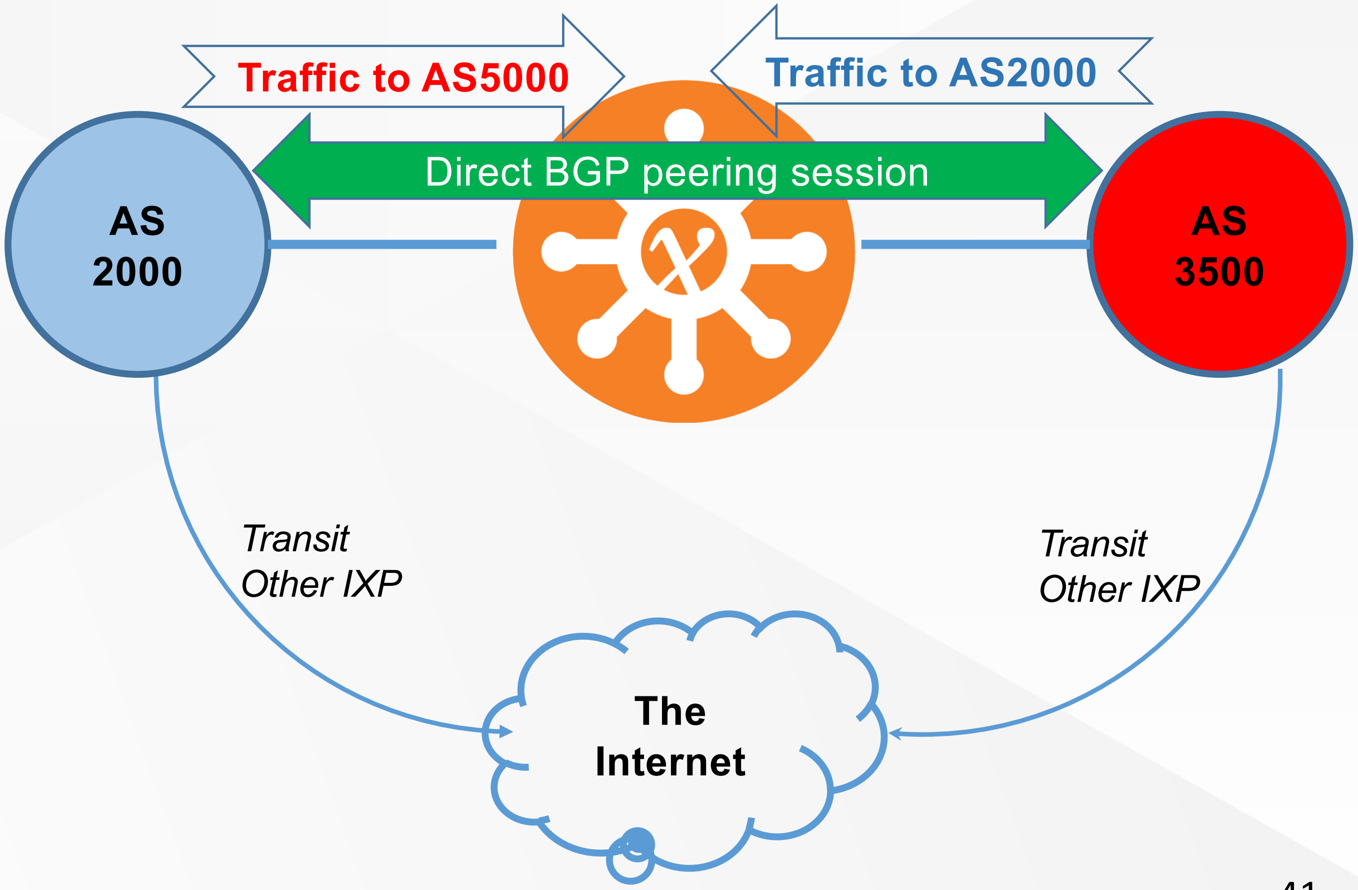
- Having AMS-IX connections **does not** mean 2 AS start exchanging traffic immediately
- Their routers do not know about the available path via AMS-IX



BGP Peering on Internet Exchange Platform

Why BGP peering ? - Direct Peering

- As the 2 AS set up direct BGP peering session they start exchanging **NLRI** (or BGP prefixes) information
- After that they can start exchanging traffic



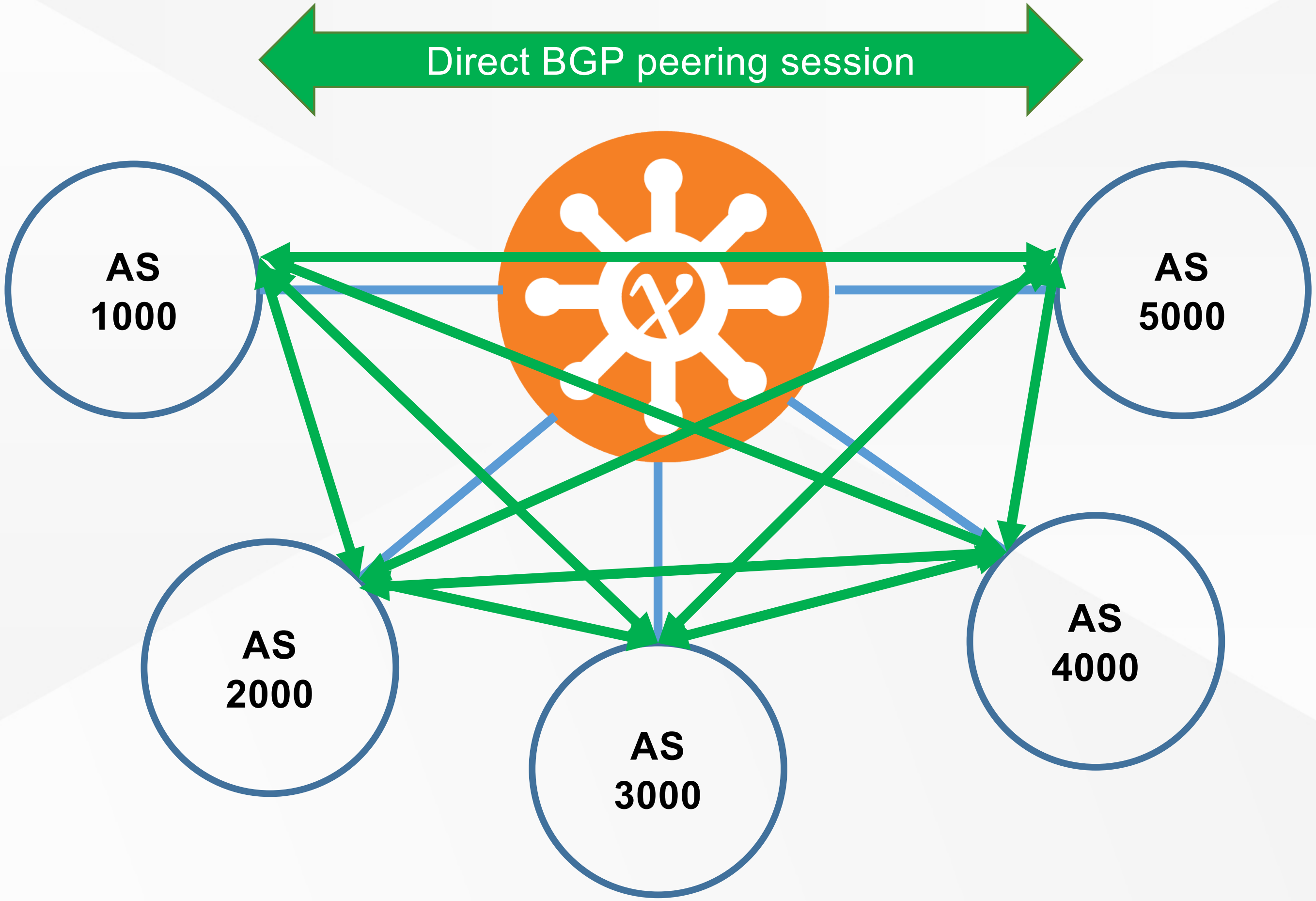
BGP Peering on Internet Exchange Platform
What is a BGP peering session ?

BGP peering is the process of exchanging **NLRI (Network Layer Reachability Information** between **routers** via **BGP (Border Gateway Protocol)**

BGP peering session the application level session between **two routers** to exchange **NLRI**, setup using **TCP/IP**

BGP Peering on Internet Exchange Platform

If there are only direct peering



BGP Peering on Internet Exchange Platform
Route Server

- The Network Administration Question

  “ BGP peering is setup only to exchange NLRI between AS

  So what if I have central place where I can advertise my NLRI and receive other NLRI ? Which will reduce the number of BGP sessions I have to manage a lot ”

- The answer is **Route Server**

BGP Peering on Internet Exchange Platform
Route Server

- The goals of the route server are
  - **to facilitate** the implementation of peering arrangements
  - **to lower the barrier** of entry for new participants on the peering platform

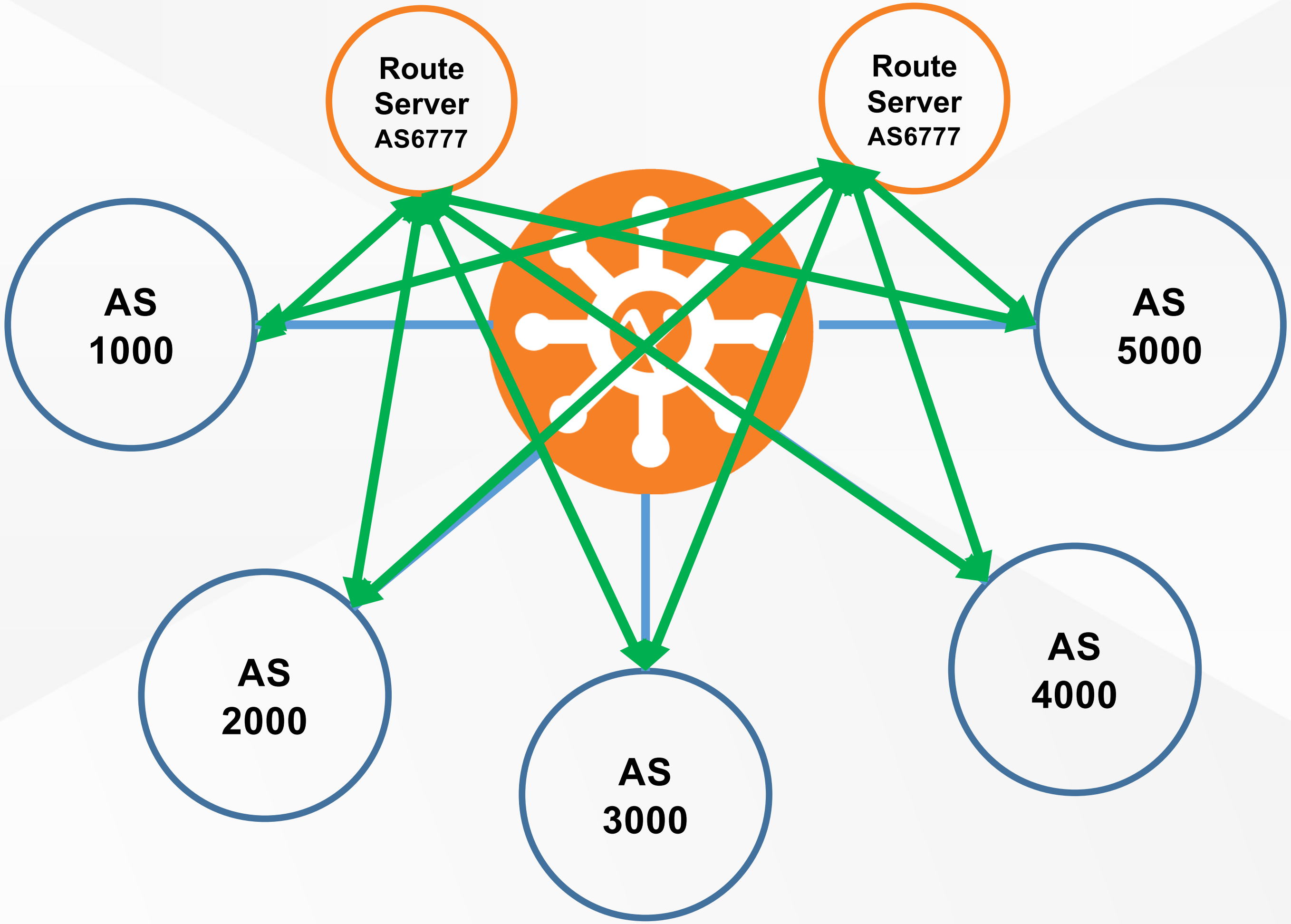
- The route servers **DO NOT Participate** in the forwarding path, so they do not forward any traffic.

- The route servers AS number is not added to the forwarding path

- Peering with a route server does not mean that you must accept routes from all other route server participants.

BGP Peering on Internet Exchange Platform

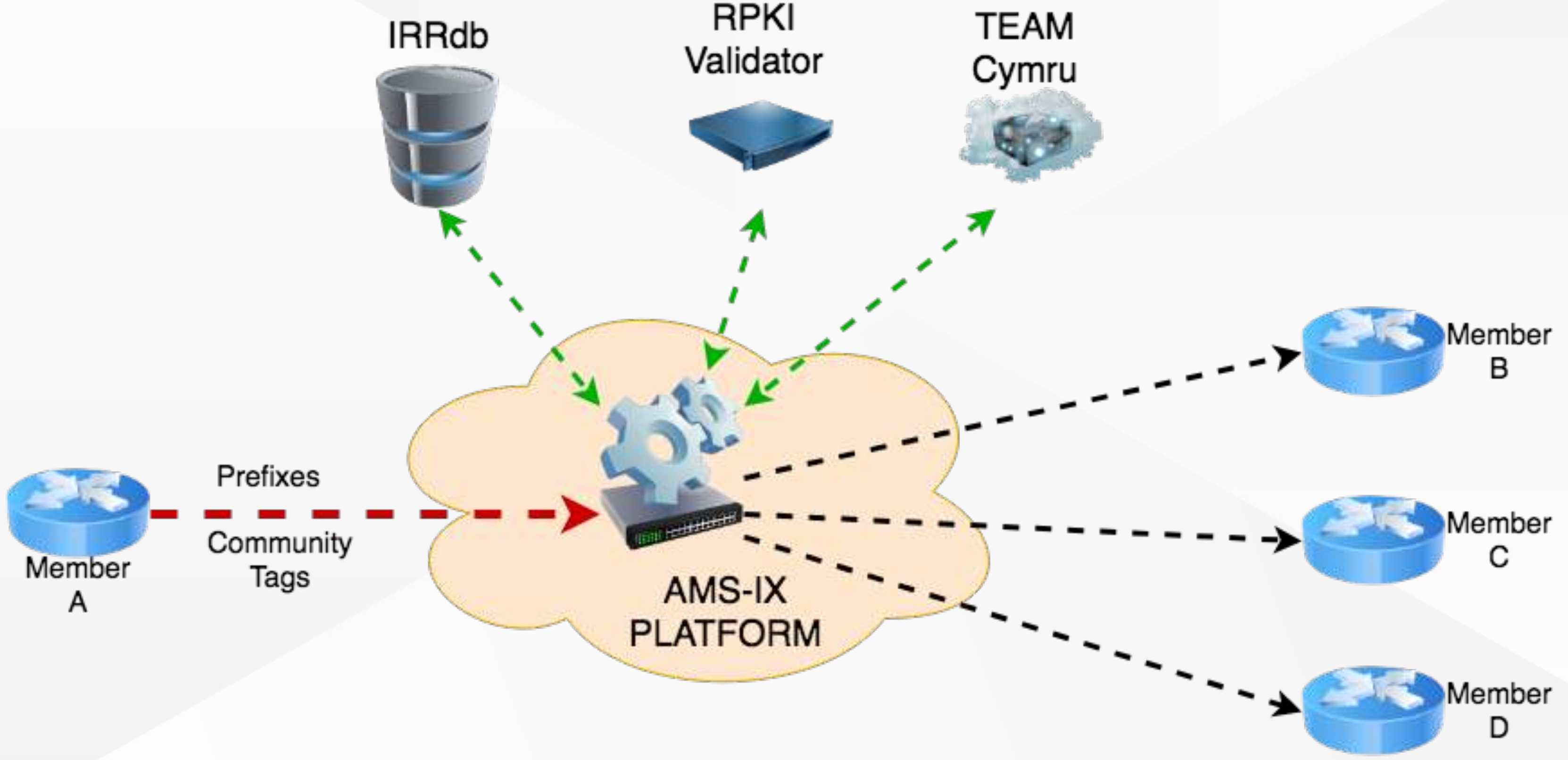
Route Server



Route Servers Deployment
Criteria for choosing route-server

- Route Server is **NOT a Route Reflector** !
- Route Server **DOES NOT** require high network bandwidth (1x1GE normally is sufficient)
- Route Server does need adequate CPU & Memory to calculate BGP routing information, base on the scale the exchange

AMS-IX RS architecture



AMS-IX RS features

- Receive Prefixes / Propagate best paths
- Ensure peering rules are satisfied
- Perform IRR and RPKI based filtering
  - The 4 filtering modes
- Perform community-based filtering
- Expose info to looking glass and notification system

Peering rules (ingress)

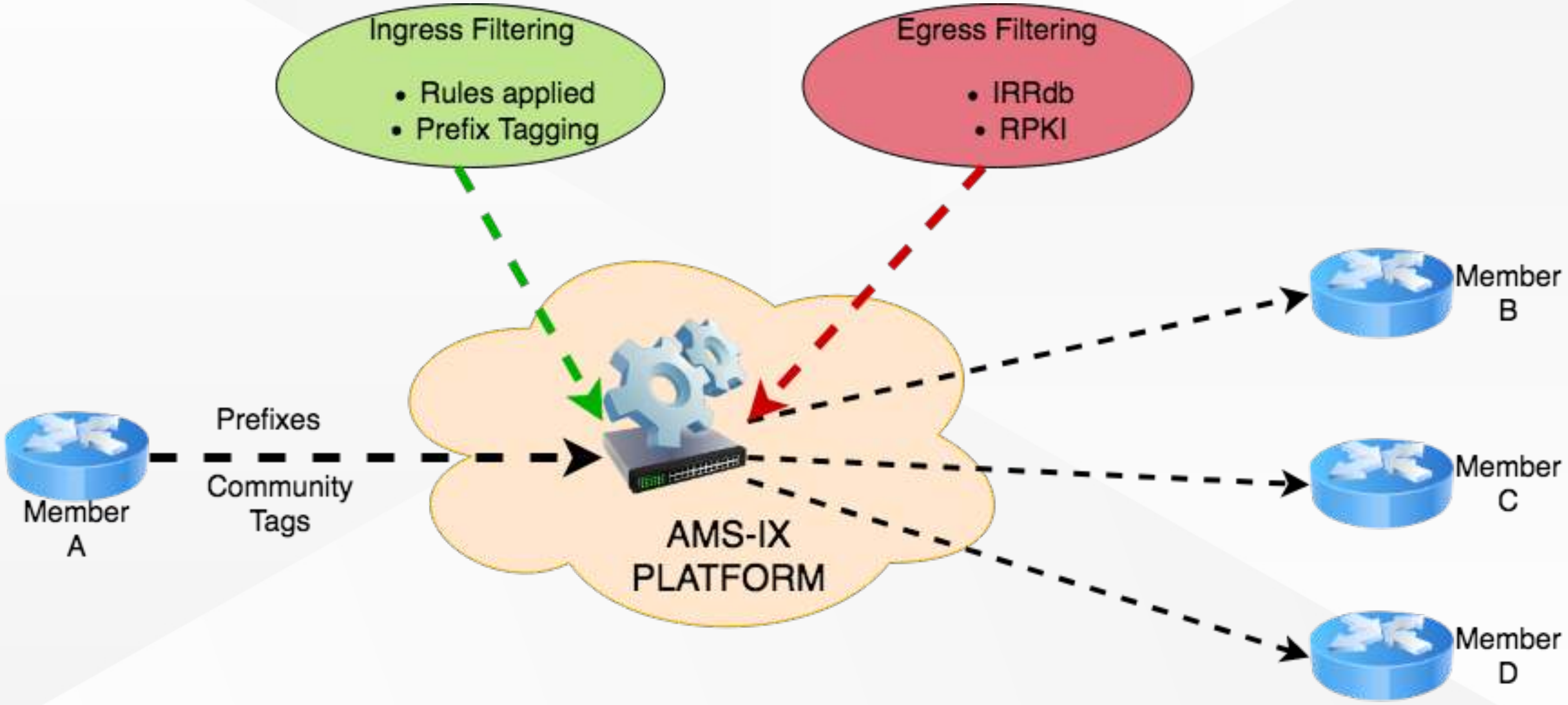
- Not accepted prefixes:
  - Bogons & Martians
    - Invalid networks on the Internet
      - Such as Private address space, link local, loopback
  - AMS-IX prefixes
  - Prefixes with AS path length > 64
  - The first AS in AS path is **not** the customer one
  - BGP next hop not belonging to the router advertising the prefix

ACCESS DENIED

The 4 filtering modes (egress)

- "*Filtering based on both IRRdb and RPKI data”* (**default**)
- *“Filtering based on IRRdb data”*
- *“Filtering based on RPKI data”*
- *“Just tagging”*

Where is filtering applied



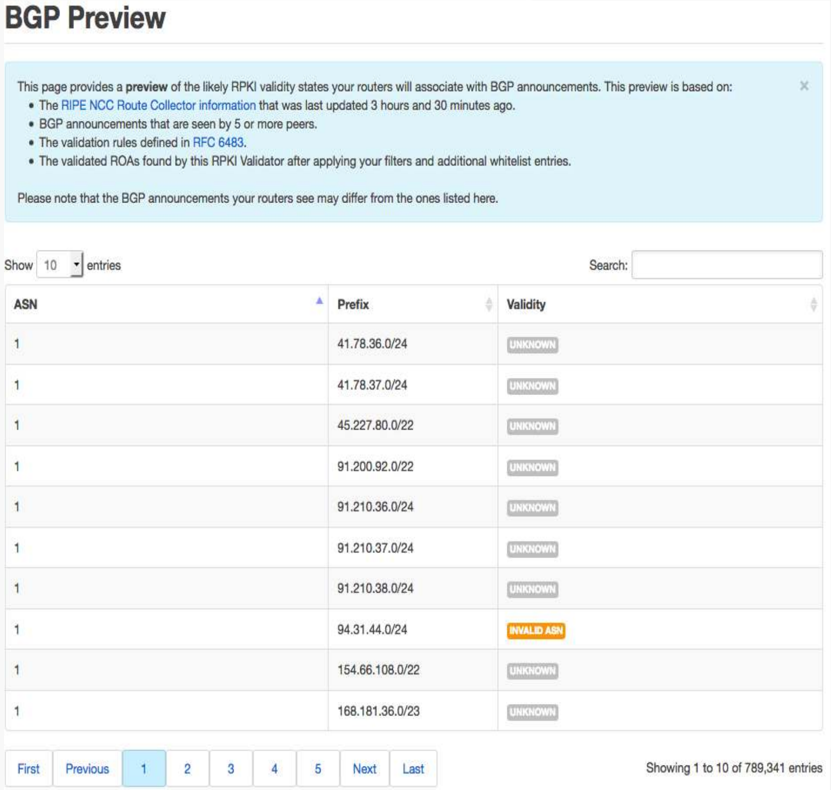
IRRdb Filtering

- RS config is generated automatically based on IRRdb parser scripts
  - Info gathered from all major IRR DBs
  - We detect policy changes every hour
- Import-via/export-via are supported
- Outgoing filtering based on IRR policies
  - You define your policy -> you instruct the RS
- Keep IRR objects up-to-date

```
aut-num: AS1200
as-name: AMS-IX1
org: ORG-AIEB2-RIPE
import: from AS-AMS-IX-PEERS action pref=100;
accept ANY AND NOT {0.0.0.0/0}
export: to AS-AMS-IX-PEERS announce AS1200
import: from AS6777 accept ANY
```

RPKI Filtering

- BGP announcements are validated with RIPE's RPKI validator
  - Only for prefixes that have a "route origin authorization" regsitered
- The prefixes that are being blocked are the ones with ROA status "INVALID"



BGP Preview

This page provides a **preview** of the likely RPKI validity states your routers will associate with BGP announcements. This preview is based on:

- The RIPE NCC Route Collector information that was last updated 3 hours and 30 minutes ago.
- BGP announcements that are seen by 5 or more peers.
- The validation rules defined in RFC 6483.
- The validated ROAs found by this RPKI Validator after applying your filters and additional whitelist entries.

Please note that the BGP announcements your routers see may differ from the ones listed here.

Show 10 entries

Search:

ASN	Prefix	Validity
1	41.78.36.0/24	UNKNOWN
1	41.78.37.0/24	UNKNOWN
1	45.227.80.0/22	UNKNOWN
1	91.200.92.0/22	UNKNOWN
1	91.210.36.0/24	UNKNOWN
1	91.210.37.0/24	UNKNOWN
1	91.210.38.0/24	UNKNOWN
1	94.31.44.0/24	INVALID ASN
1	154.66.108.0/22	UNKNOWN
1	168.181.36.0/23	UNKNOWN

Showing 1 to 10 of 789,341 entries

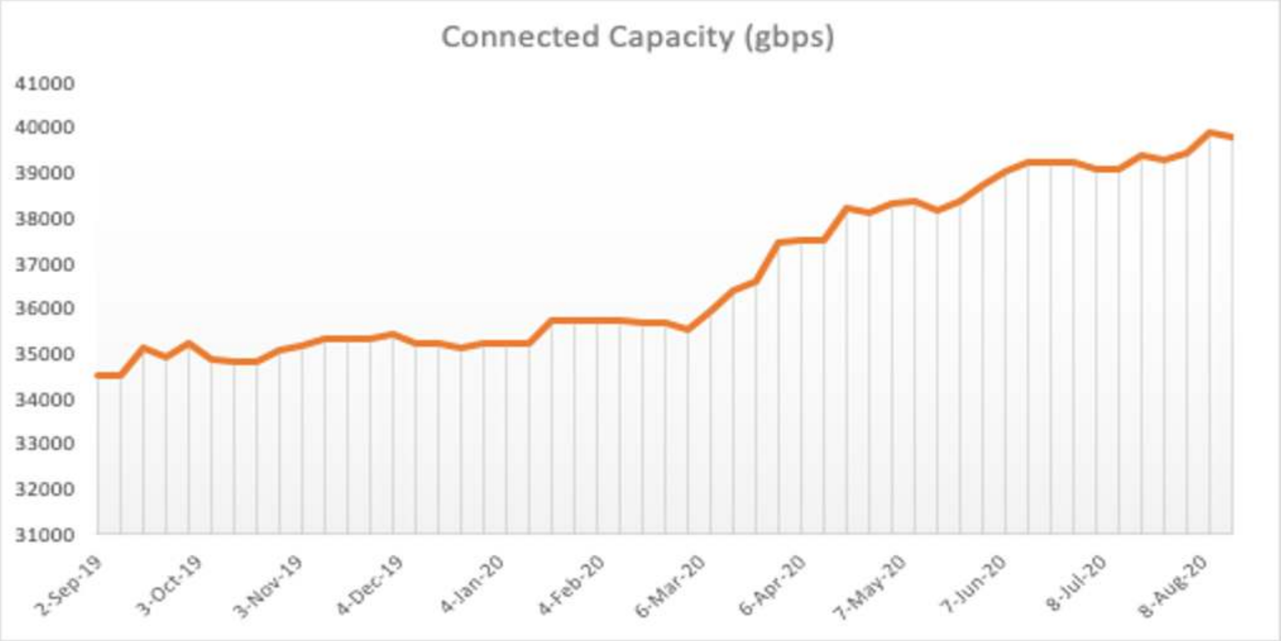
BGP communities

- Manipulate prefix announcement via BGP community attributes:
  - Do not announce a prefix to a certain peer (**0:peer-as)**
  - Announce a prefix to a certain peer (**6777:peer-as)**
  - Do not announce a prefix to any peer (**0:6777)**
  - Announce a prefix to all peers (**6777:6777)**

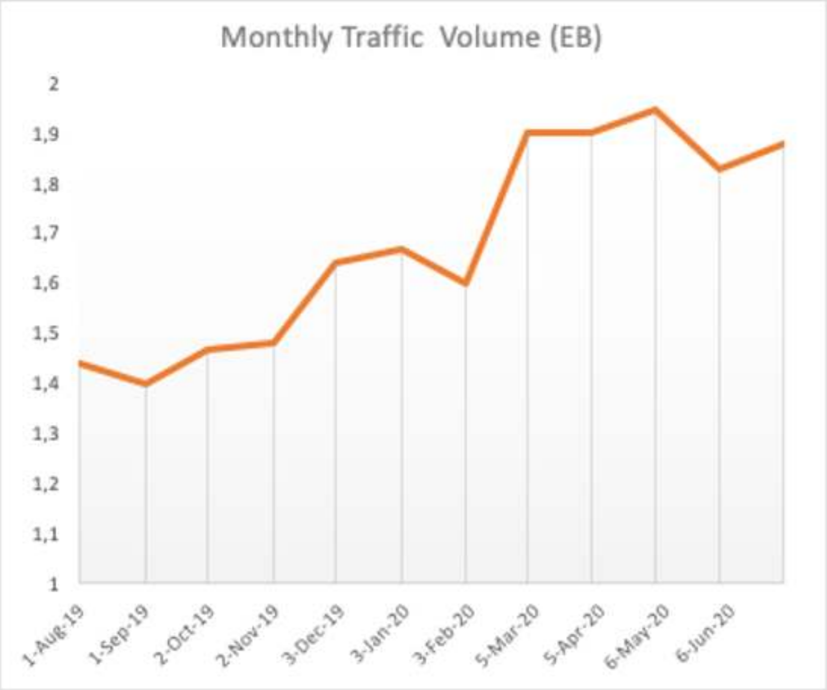
amsix
amsterdam
internet exchange

AMS-IX: Some statistics

Some statistics

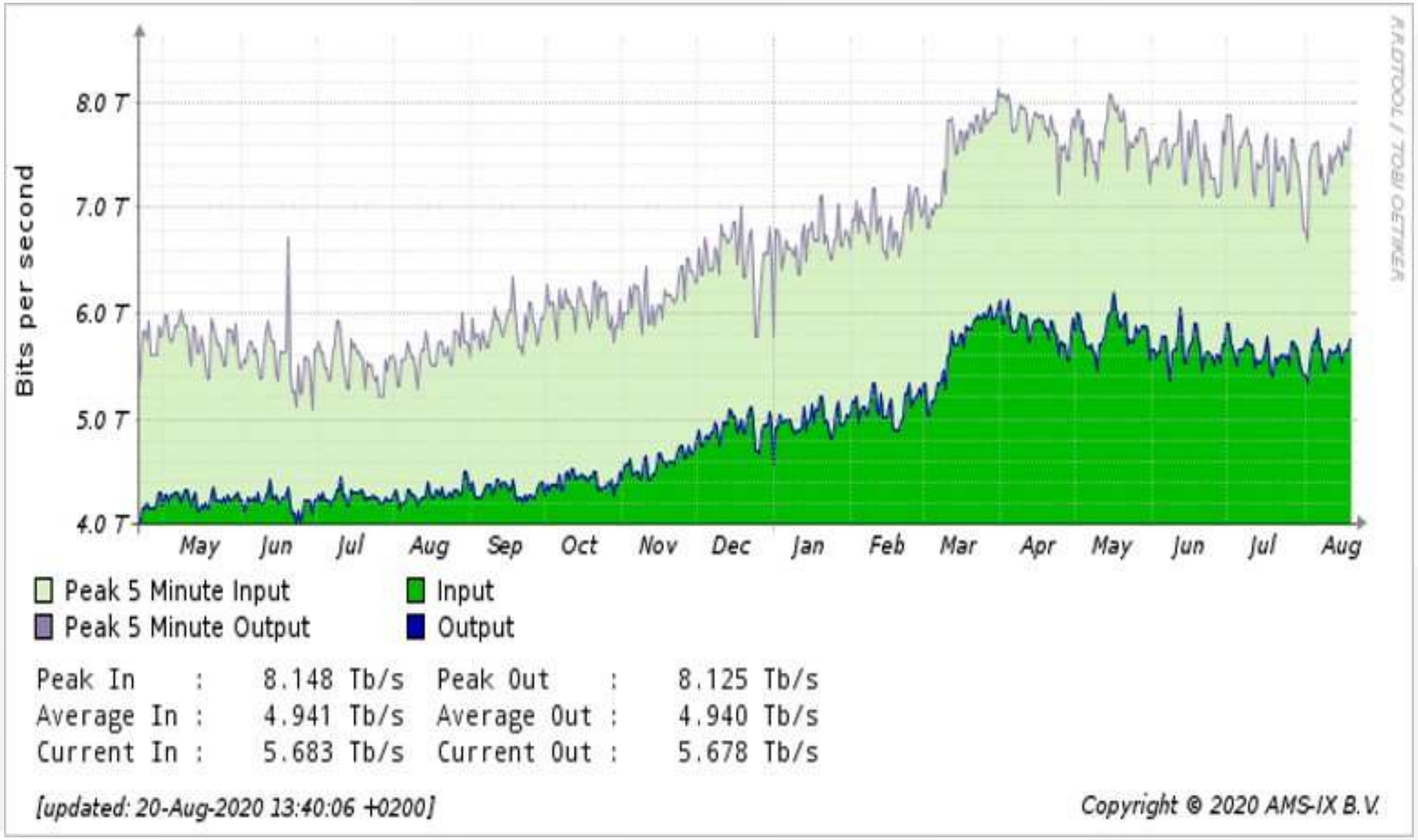


Connected Capacity (gbps)



Monthly Traffic Volume (EB)

Traffic rate



```
Peak In      :   8.148 Tb/s  Peak Out     :    8.125 Tb/s
Average In  :   4.941 Tb/s  Average Out :    4.940 Tb/s
Current In  :   5.683 Tb/s  Current Out :    5.678 Tb/s
```

[updated: 20-Aug-2020 13:40:06 +0200]

Copyright © 2020 AMS-IX B.V.

Challenges

- Staff ☺
  - Hard to get good network engineers end or software developpers
    - Extremely hard to get software developers that know of networks

- Automation
  - It is our aim to automate as much as possible
  - Ultimate goal no touch service offering
    - Certainly “no touch” provisioning

amsix
amsterdam
internet exchange

Questions ?