IONOS

Connecting 500 000 hosts

Building the European cloud backbone

What is IONOS?

- > 6 million customers
- > 40 locations
- > 30 years of history
- Provides backbone connectivity for 15 sister companies
- Publicly listed since February 2023

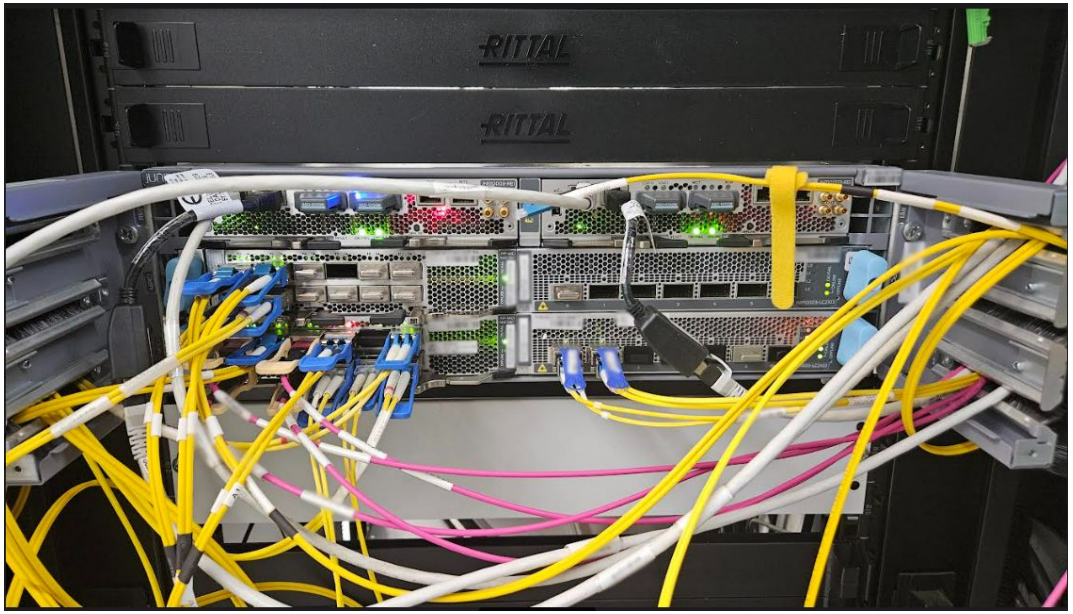
Fundamentals

- Dual vendor strategy
- MACSEC everywhere
- Lots of history
- Five setups
  - Small / big Juniper Backbone PoP
  - Small / big Cisco Backbone PoP
  - Data center handover

Hardware selection

Routers

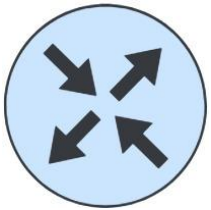
Juniper
MX 204
MX 960
MX 10003
PTX 10004

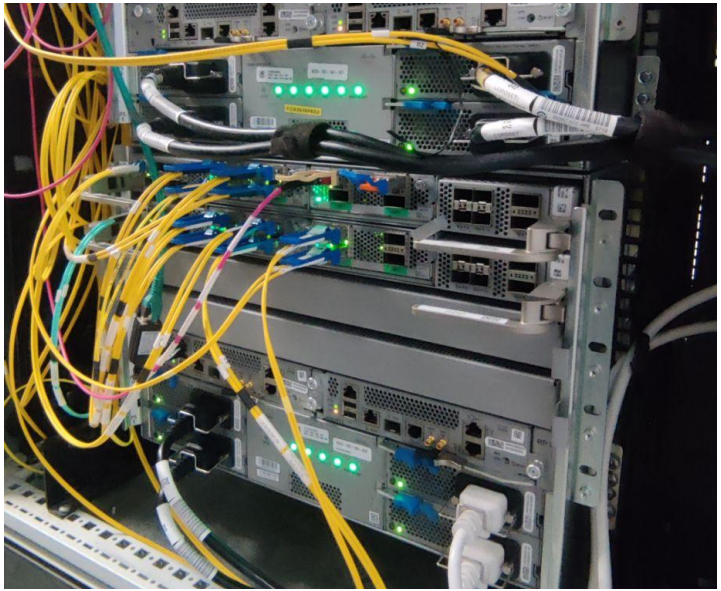


Hardware selection

Routers

Cisco
NCS-57C3
NCS-5504
ASR-9910





Hardware selection

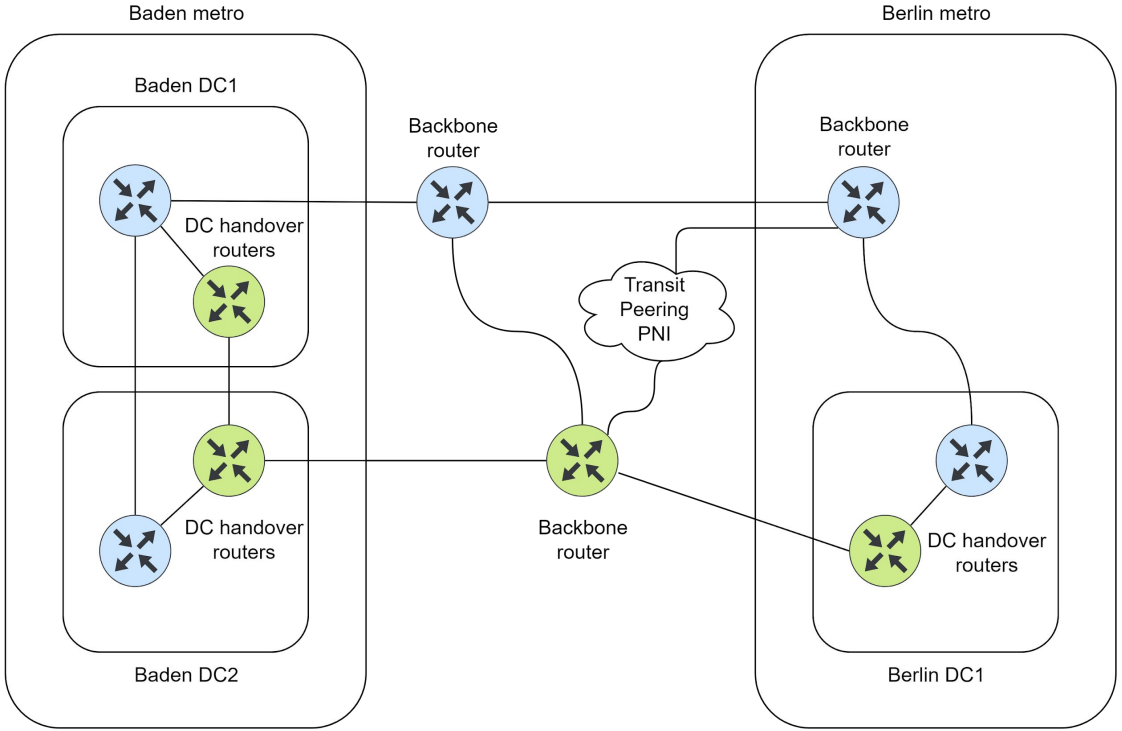
WDM

Infinera Groove G30 and G31
ADVA FSP 3000 being phased out

Routing architecture

Basic overview

- SR-MPLS on IS-IS everywhere
- Latency determines metrics
- Unnumbered interfaces everywhere[1]
- Dual stack everywhere

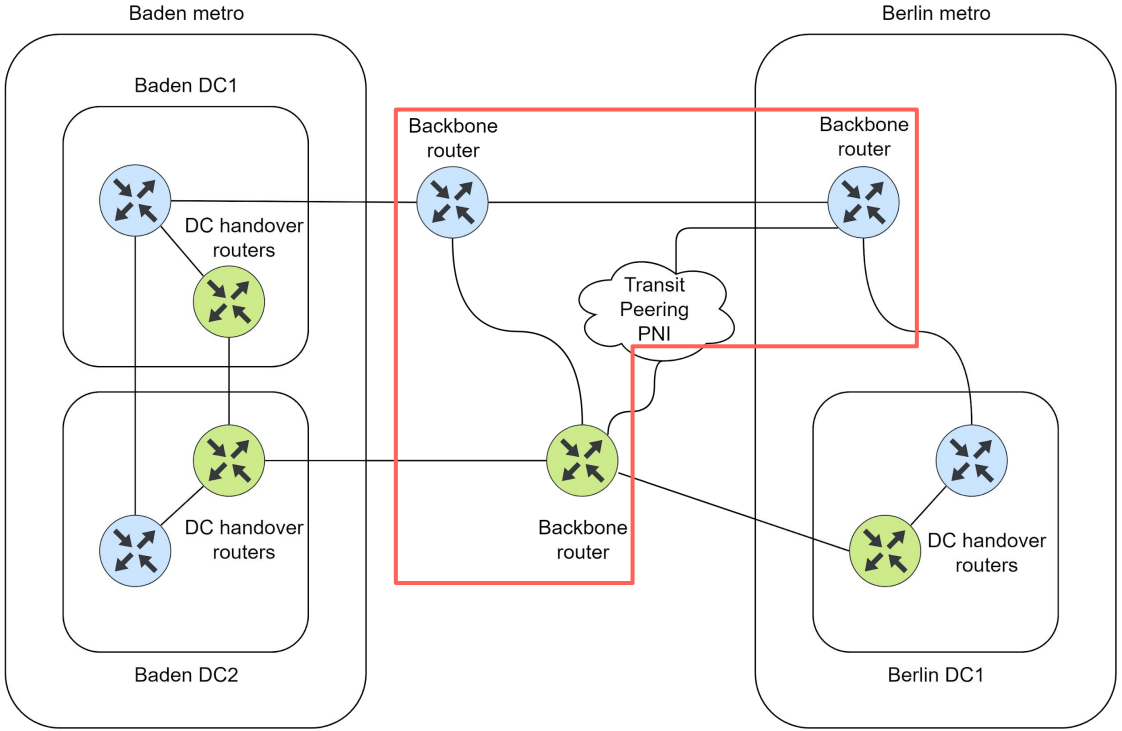


[1]There is a limitation to the usability of unnumbered interfaces that will be discussed later on

Routing architecture

Basic overview

- IBGP full mesh in the backbone for GRT
- Route reflector sessions for VPN services
- Next-hop filtering policies applied[1]

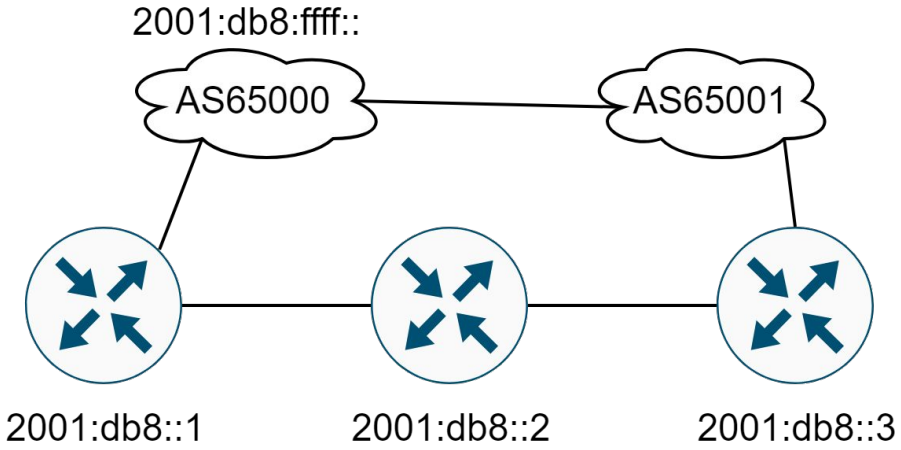


[1]IOS-XR: router bgp 64511 address-family ipv6 unicast nexthop route-policy mypolicy
[1]Junos: set routing-options resolution rib inet6.0 import mypolicy

Routing architecture

next-hop-filtering

- Aggregate route to 2001:db8::/48
- Path from ::2 to ::3 is worse due to AS-PATH

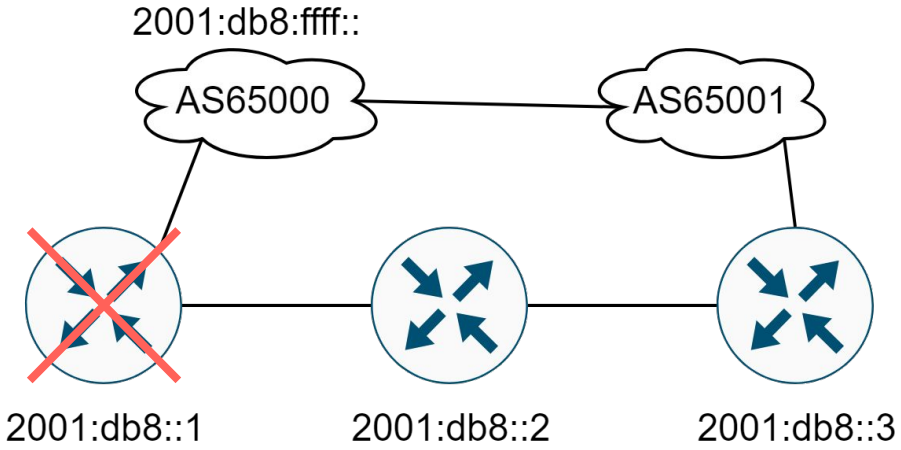



```
2001:db8:ffff::/48 *[BGP/170] 00:01:00, localpref 100, from 2001:db8::1
                     AS path: 65000 I, validation-state: unverified
                   > to fe80::e00:2aff:fe09:b901 via ge-0/0/0.0
                   [BGP/170] 00:01:00, localpref 100, from 2001:db8::3
                     AS path: 65001 65000 I, validation-state: unverified
                   > to fe80::e00:edff:febb:c102 via ge-0/0/1.0
```

Routing architecture

next-hop-filtering

- Best-path router ::1 has failed
- Route to ::1 has been removed from IGP
- BGP has not timed out yet
- BGP resolves NH ::1 through aggregate

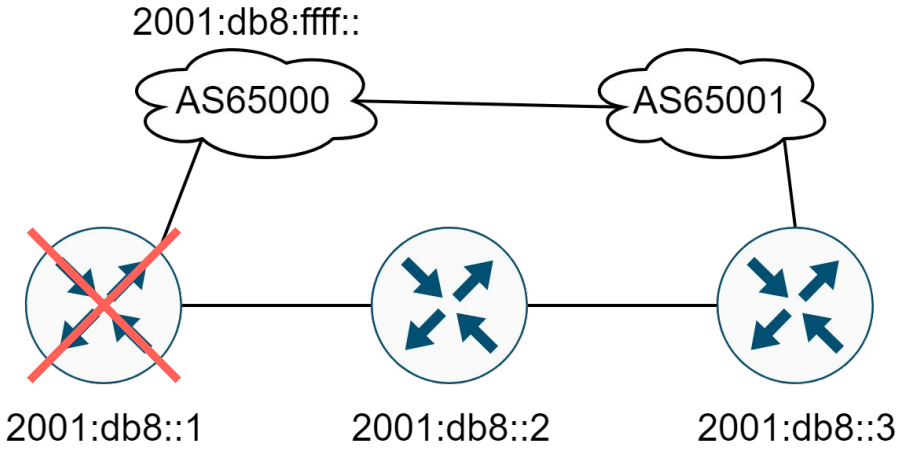


2001:db8:ffff::/48 *[BGP/170] 00:00:07, localpref 100, from 2001:db8::1
AS path: 65000 I, validation-state: unverified
to Discard
[BGP/170] 00:08:26, localpref 100, from 2001:db8::3
AS path: 65001 65000 I, validation-state: unverified
> to fe80::e00:edff:febb:c102 via ge-0/0/1.0

Routing architecture

next-hop-filtering

- Best-path router ::1 has failed
- Route to ::1 has been removed from IGP
- Next-hop hop policy does not allow aggregate next-hop

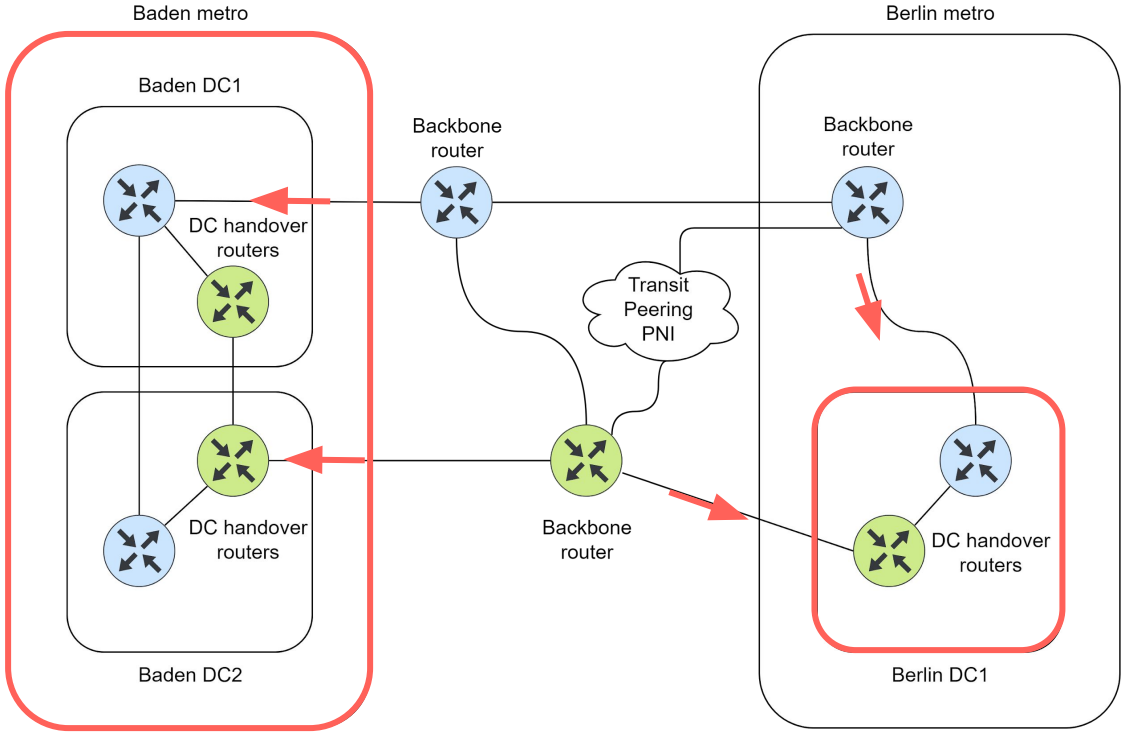


2001:db8::ffff/128 *[BGP/170] 00:10:32, localpref 100, from 2001:db8::3
AS path: 65001 65000 I, validation-state: unverified
> to fe80::e00:edff:febb:c102 via ge-0/0/1.0
[BGP/170] 00:00:21, localpref 100, from **2001:db8::1**
AS path: 65000 I, validation-state: unverified
Unusable

Routing architecture

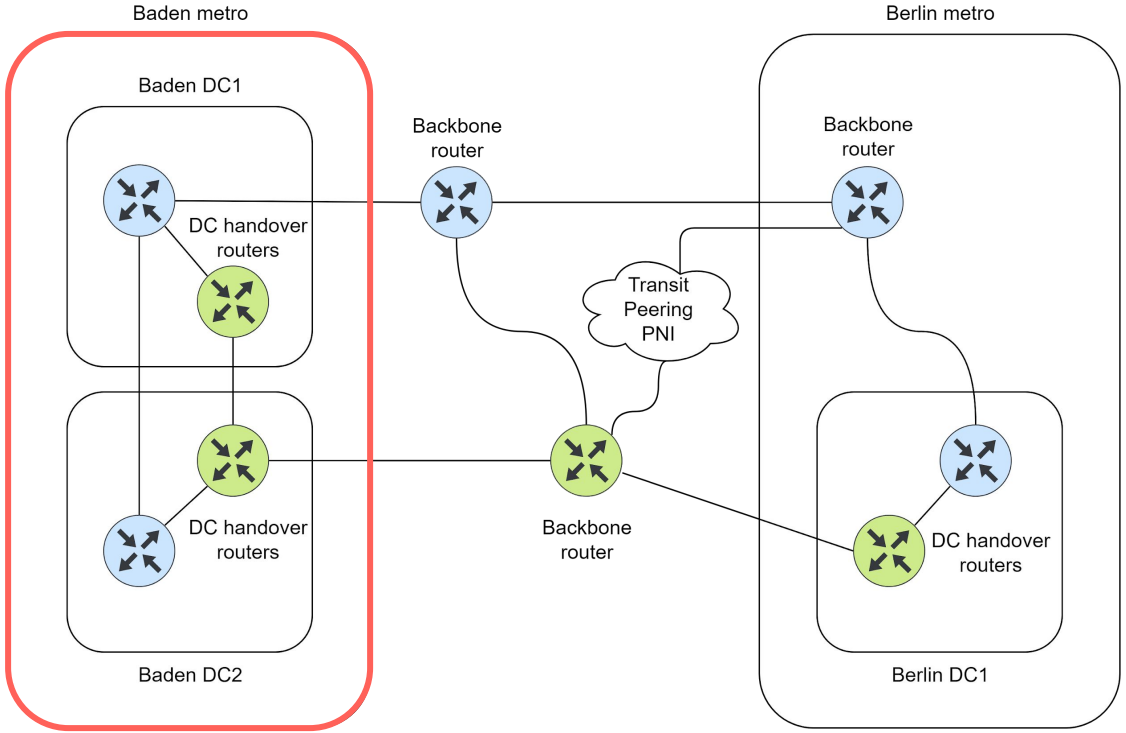
Basic overview

- DC handover routers have an IBGP metro mesh
- Heavily reduced routing table reflected to DC handover routers
- DC handover routers used to connect different regions through MPLS



Routing architecture

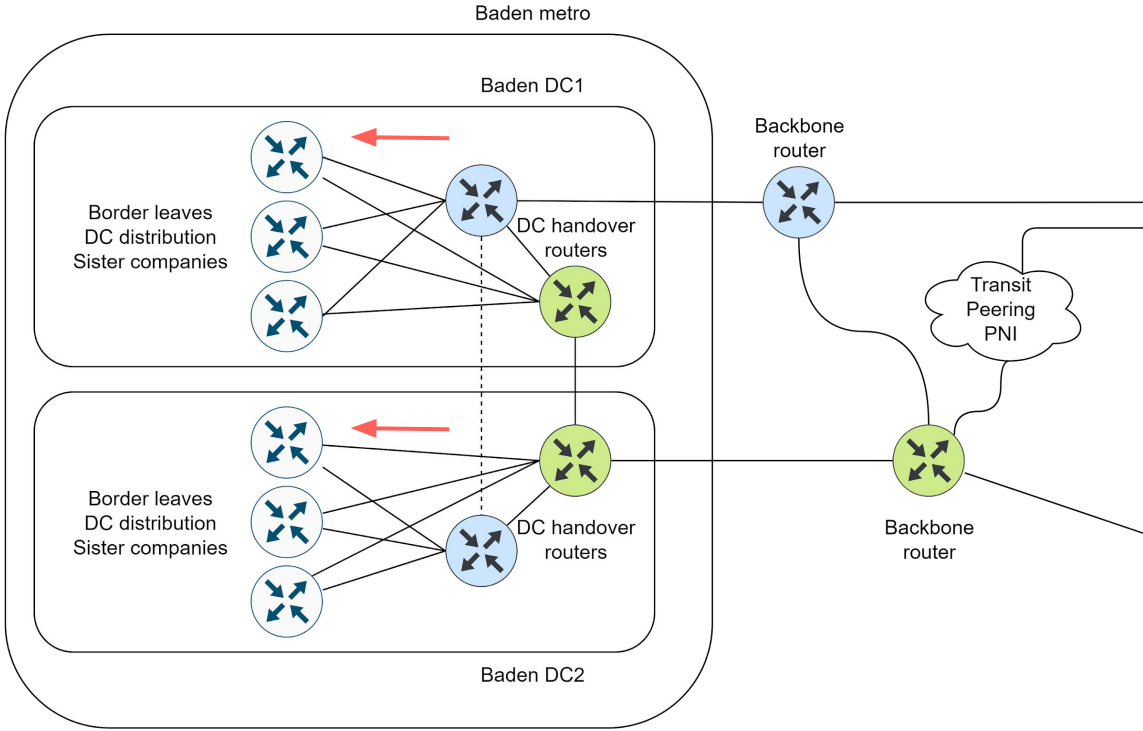
Basic overview



Routing architecture

Basic overview

- DC handover routers announce default routes + necessary routes
- Connecting downstreams
  - eBGP strongly preferred
  - Some setups require iBGP
  - Inter-department IS-IS can lead to problems



Routing architecture

Lessons learned

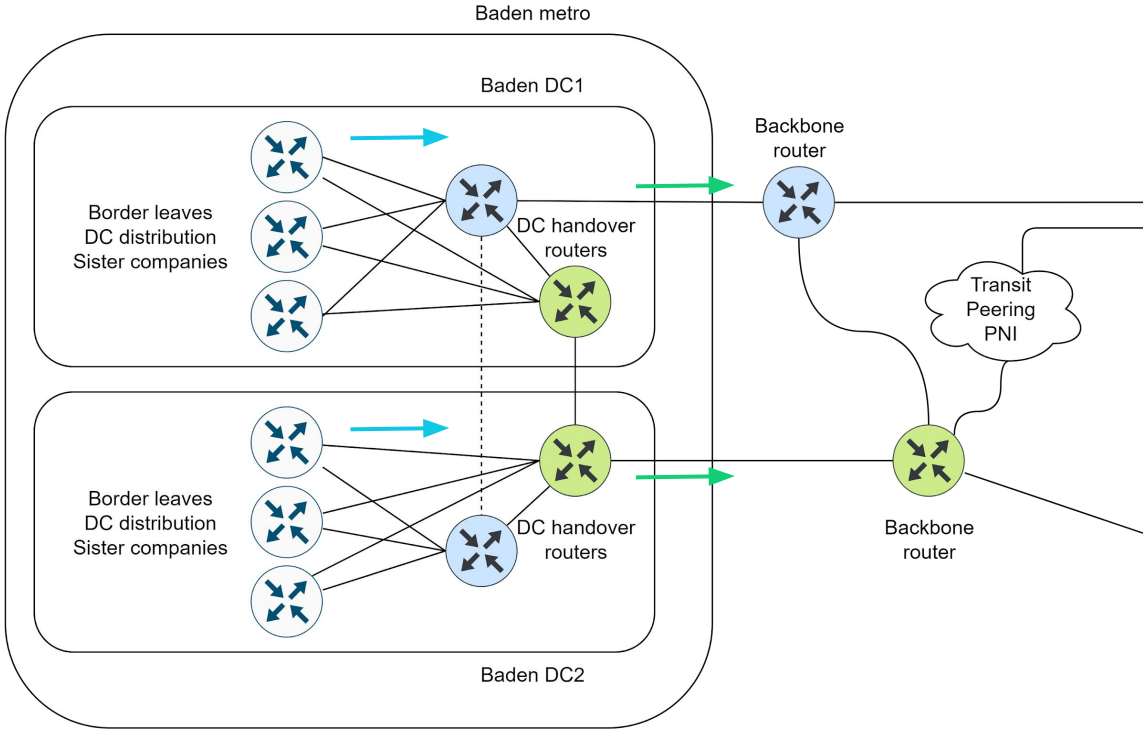
- Avoid IGP adjacencies between departments
  - Very strict policy required
  - High risk of miscommunication
  - People do not necessarily know the impact of their actions

- Accidentally setting the wrong metrics
  - Potentially drawing MPLS traffic through non-MPLS interfaces
- A downstream device leaking link local prefixes into L2 LSDB
  *RP/0/RP0/CPU0[…]: %ROUTING-ISIS-4-MARTIAN : **Level 2 LS** […] **contains** an IPv6 Unicast prefix **advertisement to the invalid prefix fe80::6664:9b04:cd64:4bf0/128***
- An engineer trying to address the problem accidentally redistributes 200k prefixes into L2 LSDB

The scenarios on the right represent examples of what could go wrong in the described scenario ;)

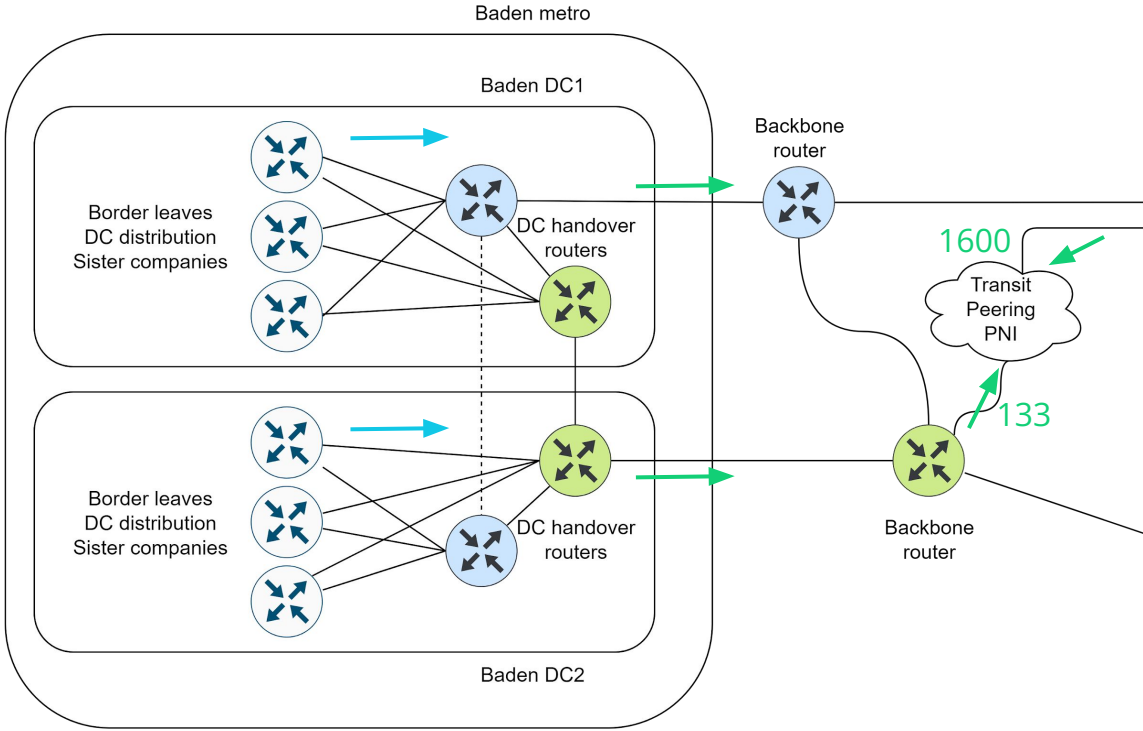
Aggregation

- Downstreams announce more-specifics
- DC handovers originate Metro-aggregates
- More-specifics spread as little as possible



Announcements

- Public announcements with latency-based MED
- Community-based filters



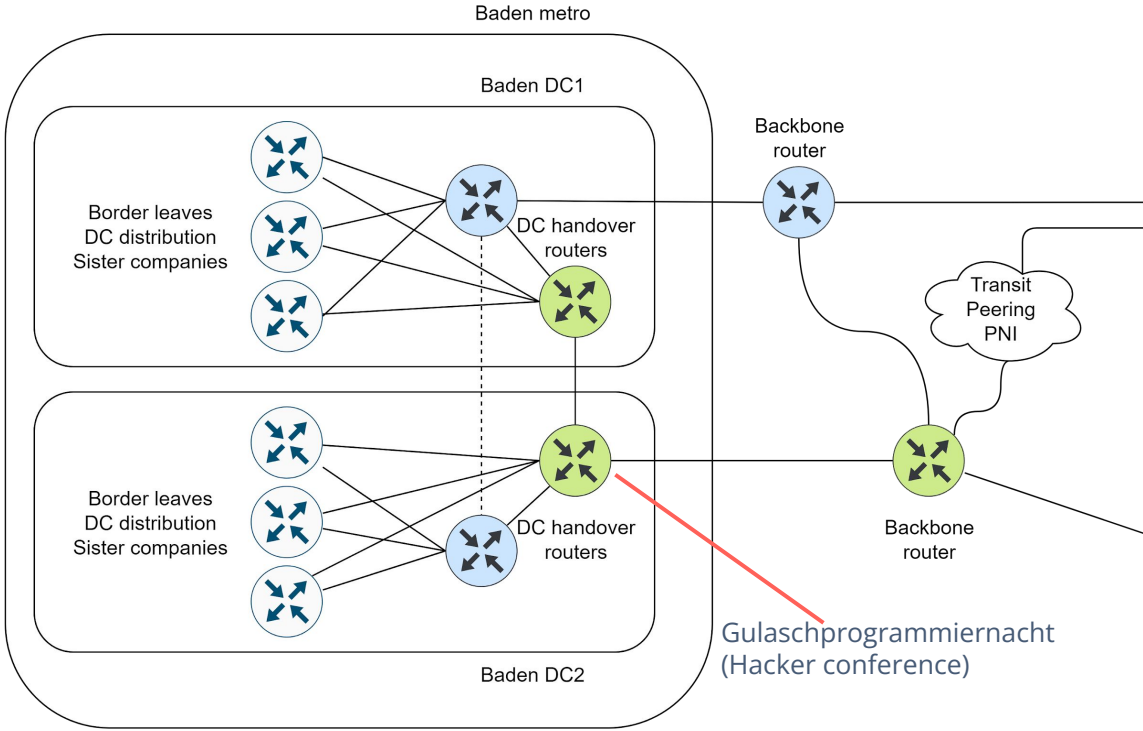
Communities

- Community set at the origin
- Communities for
  - Internal
  - Downstreams
  - Peers
  - Upstreams
  - Origin region / DC
  - Route type (aggregate / more-specific)
  - Each peer / IXP / transit

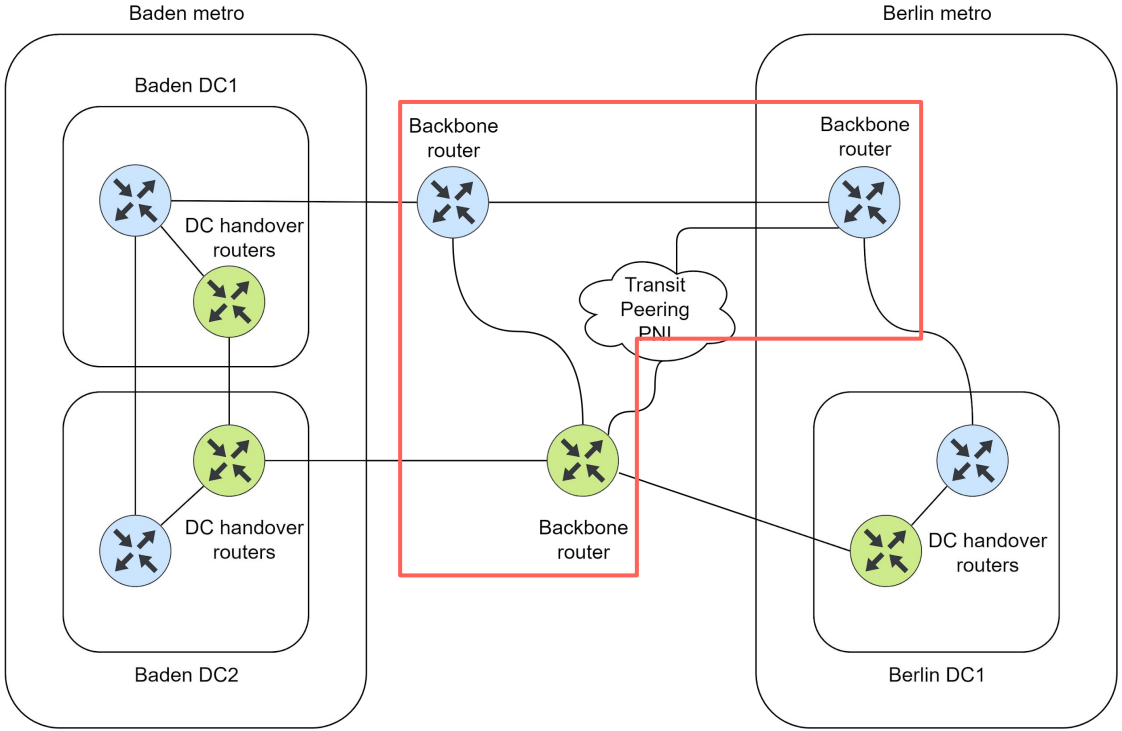
- Example routes
  - 1.1.1.0/24 has
    - Is peer
    - From DECIX
    - From Europe
    - From InterXion Fra8
  - 2001:8d8::/32 has
    - Is our own
    - Is aggregate
    - From Europe
    - From Germany

Getting a fulltable

- Sometimes we need a fulltable in the DC
  - DC handover routers don't have one
- Downstream peers
  - Get a session to the DC handover + multihop from two backbone routers
- Events (transit only)
  - EVPN-VXLAN service from DC handover to the closest backbone router



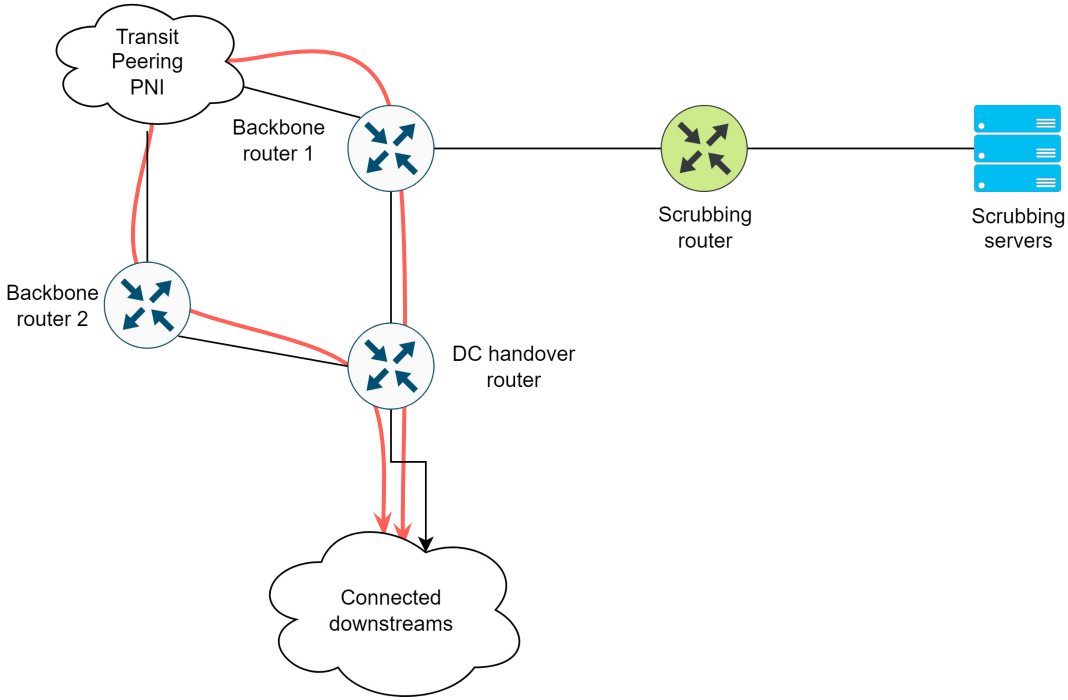
DDoS mitigation



DDoS mitigation

Regular traffic flow

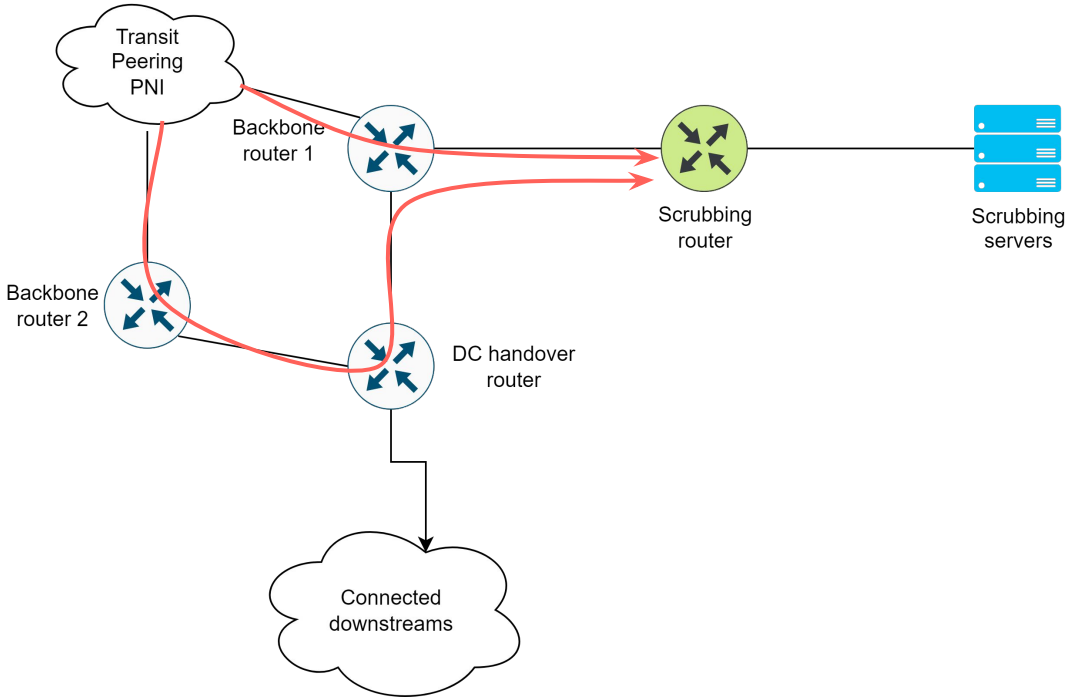
- DC handover originates aggregate
- Backbone routers route traffic to DC handover



DDoS mitigation

Mitigation flow

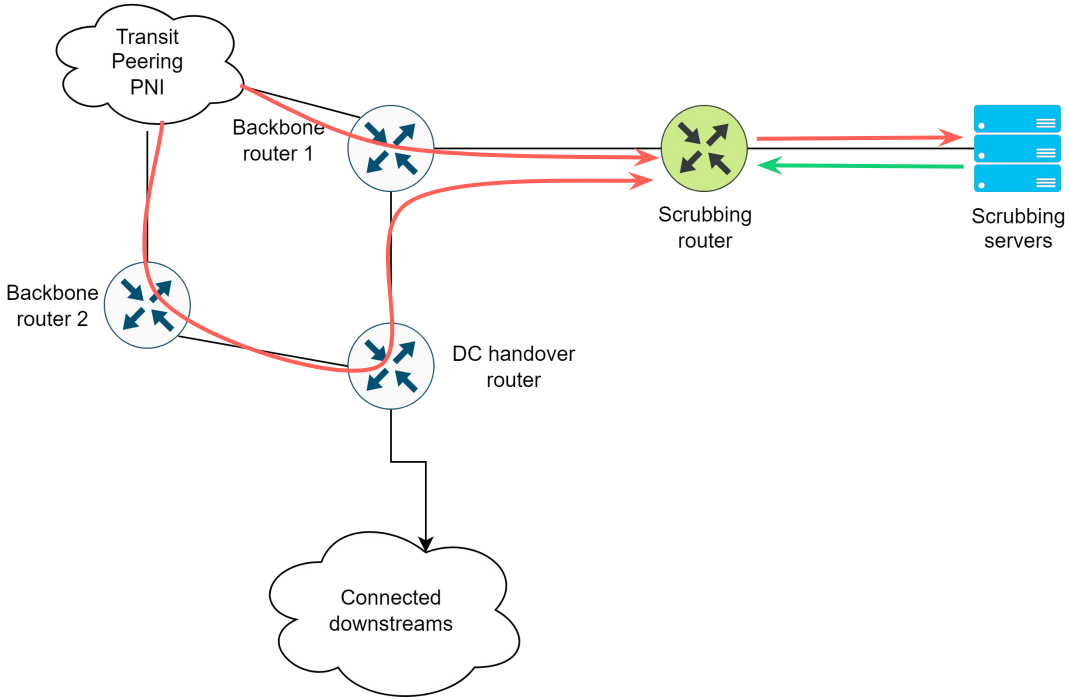
- Scrubbing router announces a most-specific
- DC handover keep their aggregate
- Traffic is routed to the closest scrubbing router



DDoS mitigation

Cleaned traffic

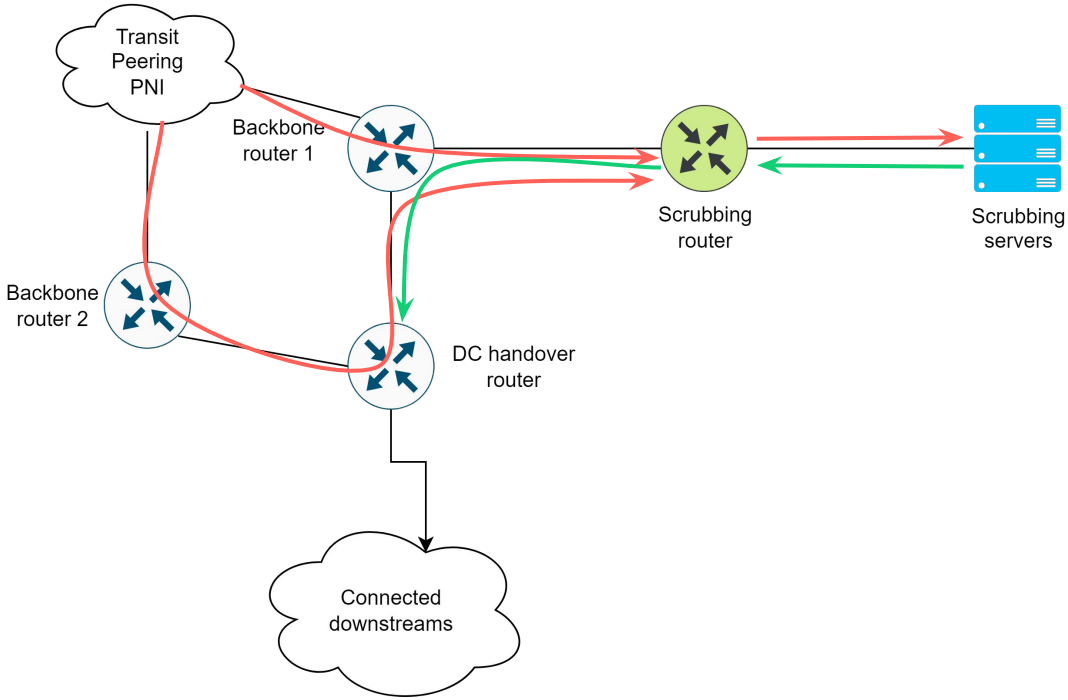
- Scrubbing router leaks the traffic into a dirty VRF
- Scrubbing servers route clean traffic back
- VRFs are only required on the scrubbing router, nowhere else



DDoS mitigation

Cleaned traffic

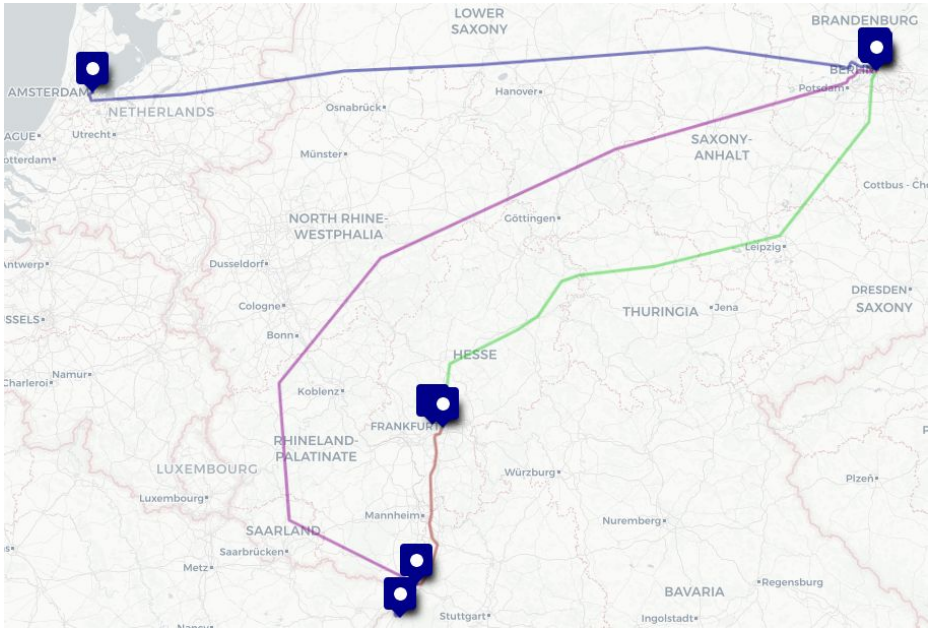
- Scrubbing router stacks SIDs to switch the traffic to the DC handover
- The DC handover routes the clean traffic as if nothing happened



Going dark

Splitting AS8560

- Berlin has three routes
  - Amsterdam
  - Frankfurt
  - Karlsruhe
  - Two new bundles to Frankfurt and Karlsruhe are **ordered**
- Aggregates are originated on the backbone routers

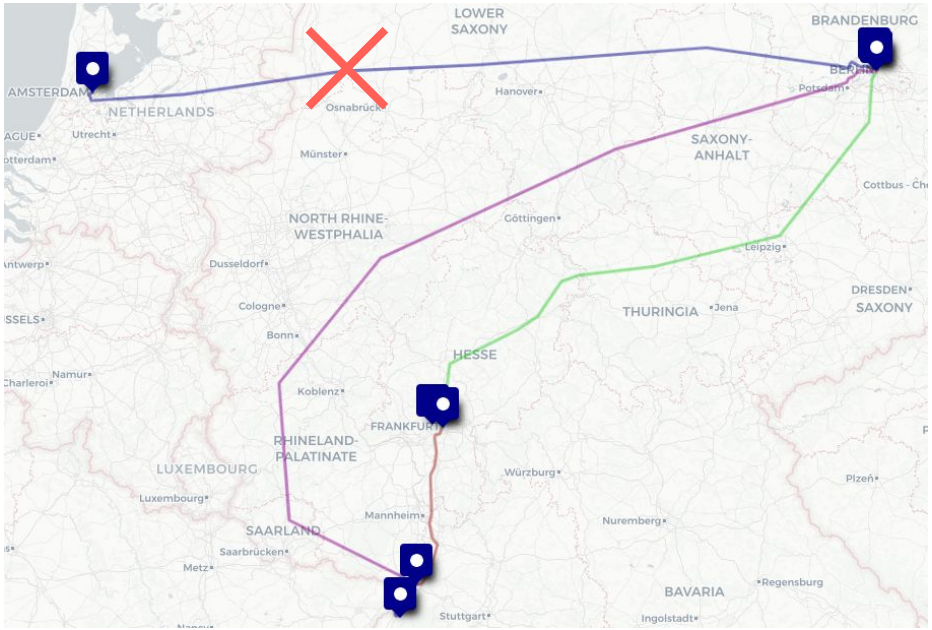


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Going dark

FF 6 months

- Berlin has ~~three~~ **two** routes
  - ~~Amsterdam~~
    - PoP offline
  - Frankfurt
  - Karlsruhe
  - Two new bundles to Frankfurt and Karlsruhe are **delayed**
- Aggregates are originated on the backbone routers

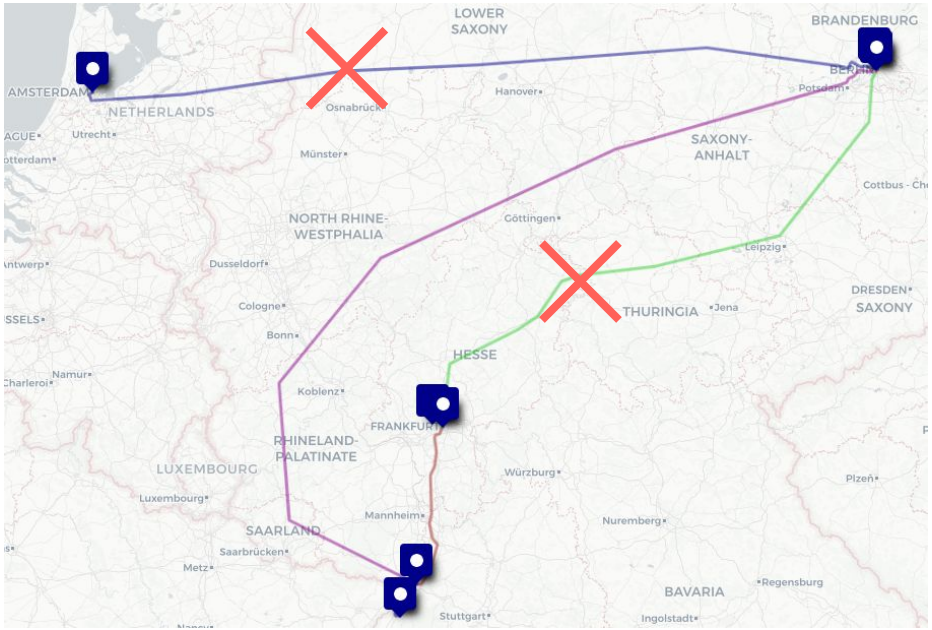


Fictional lines and fictional data center locations.

Going dark

Fiber cuts

- Berlin has ~~three~~ **one** route
  - ~~Amsterdam~~
  - ~~Frankfurt~~
    - Fiber was cut; technicians are on their way
  - Karlsruhe
  - Two new bundles to Frankfurt and Karlsruhe are **delayed**
- Aggregates are originated on the backbone routers

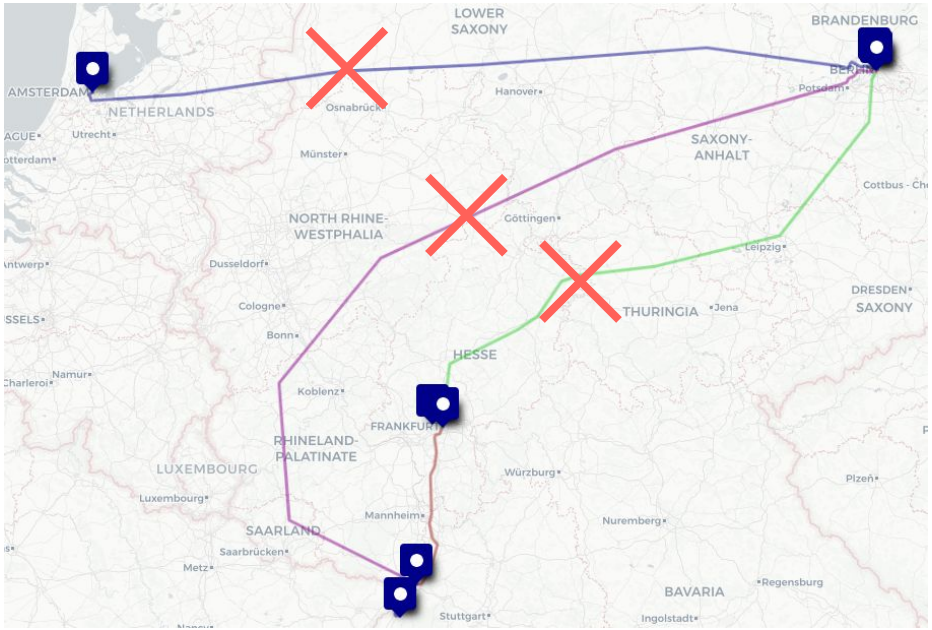


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Going dark

Fiber cuts

- Berlin has ~~three~~ **no** route
  - ~~Amsterdam~~
  - ~~Frankfurt~~
  - ~~Karlsruhe~~
    - Floods took out the fiber
  - Two new bundles to Frankfurt and Karlsruhe are **delayed**
- Aggregates are originated on the backbone routers

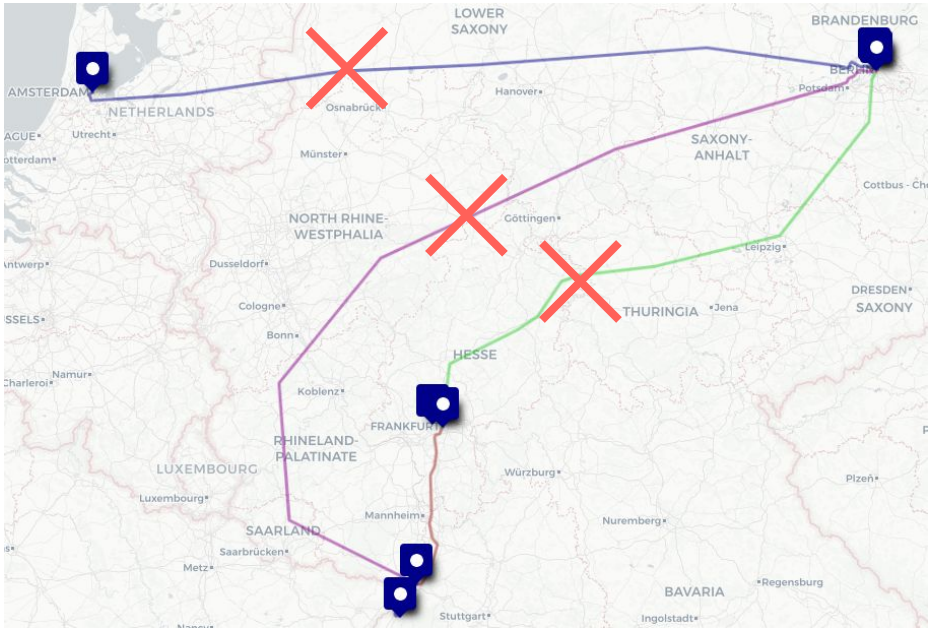


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Going dark

DC offline

- Berlin has ~~three~~ **no** route
  - Local transit / peering, but is no longer connected to the rest of the network
  - Parts of the net unreachable depending on the ingress PoP
- Aggregates are originated on the backbone routers
  - Some routers announce aggregates for networks they can’t reach

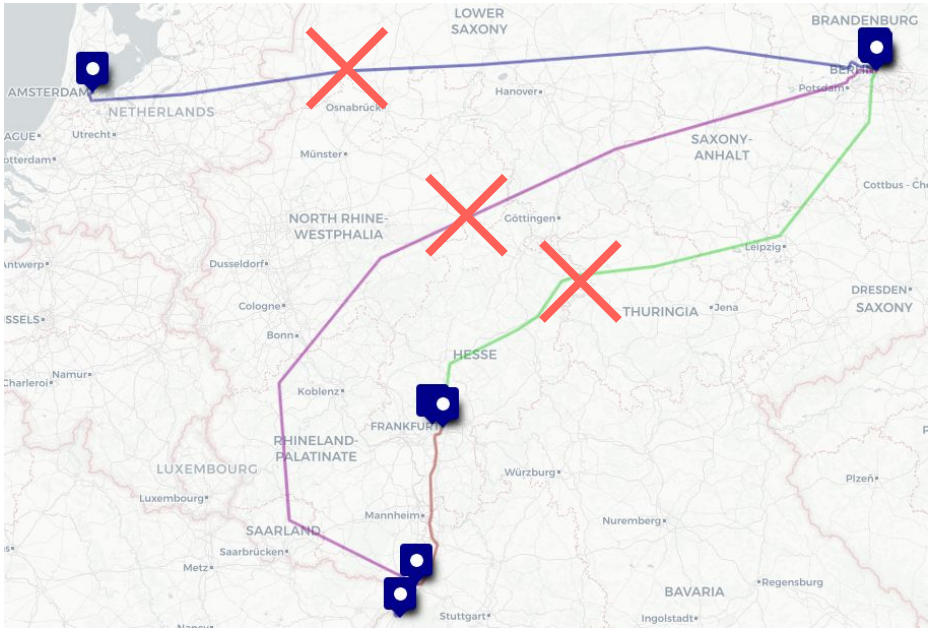


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Going dark

Splitting AS8560

- Berlin has ~~three~~ **no** route
  - Internal services unreachable
    - Amplified impact
- Aggregates are originated on the backbone routers
  - Some routers announce aggregates for networks they can’t reach

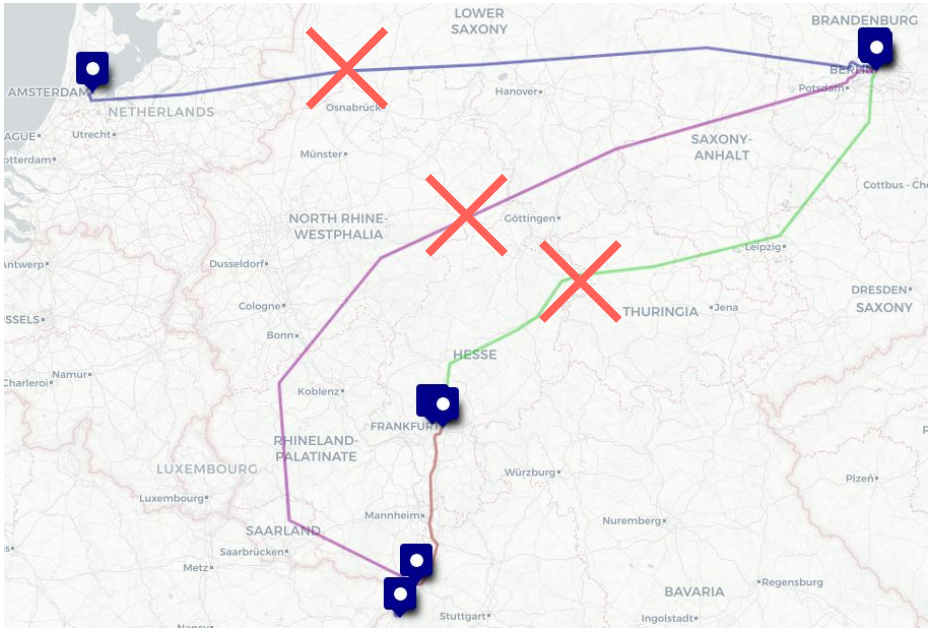


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ **no** route
  - DF carriers do not have enough capacity
    - Fibers cannot be swapped to protected paths
  - Technicians are on site in Frankfurt
  - GRE over transit evaluated

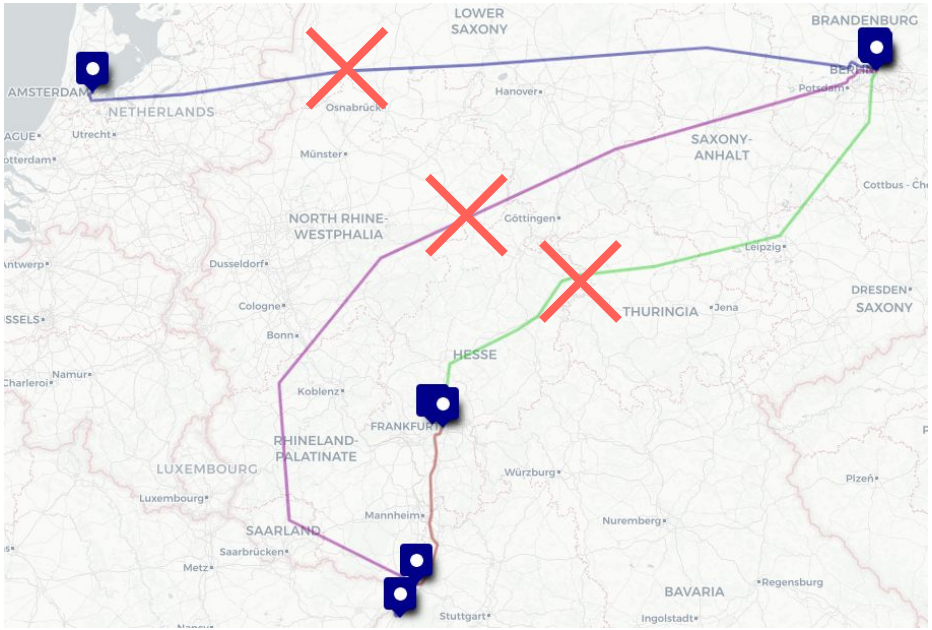


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ **no** route
  - DF carriers do not have enough capacity
  - Technicians are on site in Frankfurt
    - Fiber cut is next to a leaking gas pipeline; repair not possible
  - GRE over transit evaluated

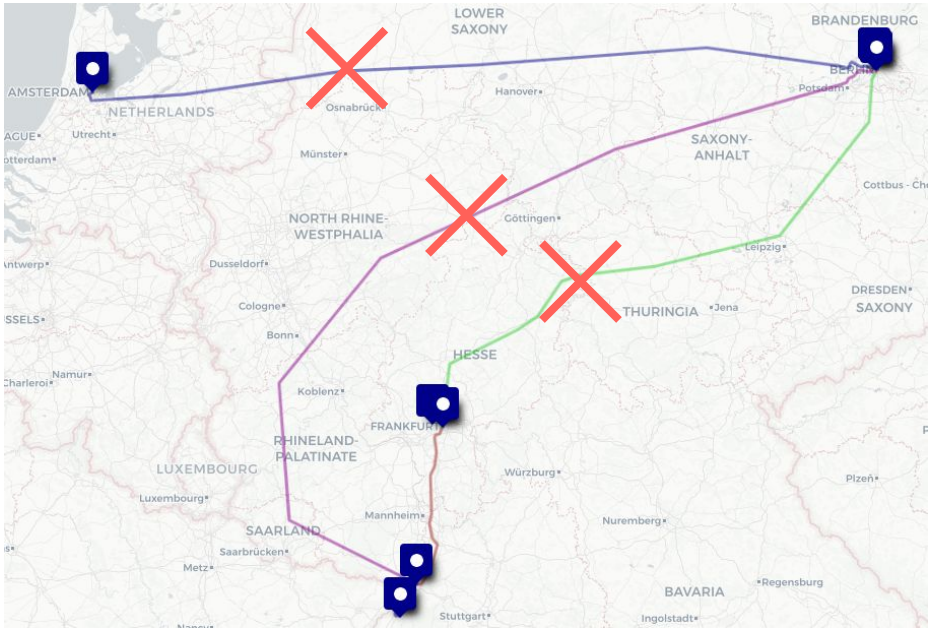


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ **no** route
  - DF carriers do not have enough capacity
  - Technicians are on site in Frankfurt
  - Tunnels over transit evaluated
    - Enough unencrypted capacity available
    - Engineers are identifying what traffic needs to be dropped

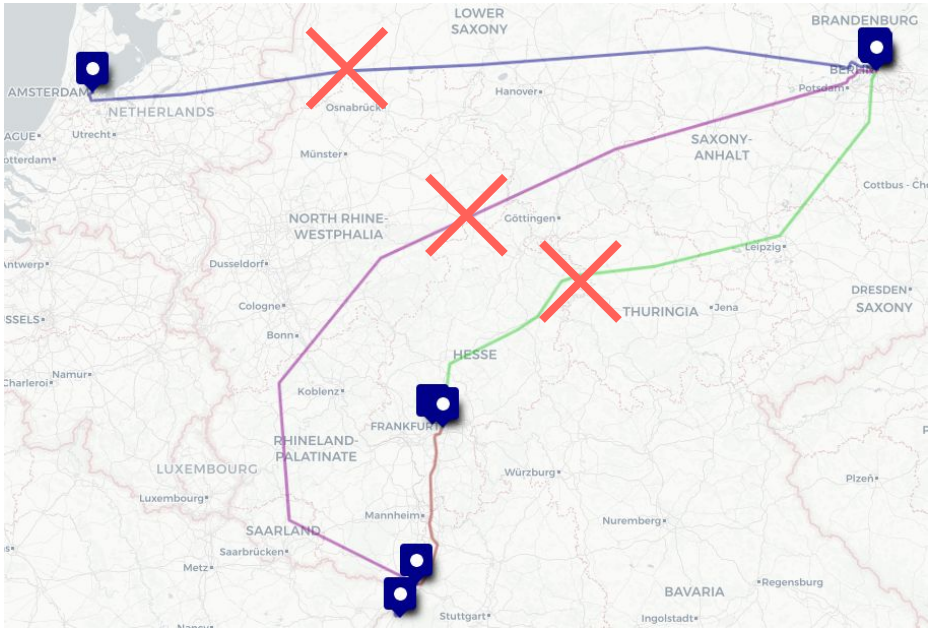


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ **no** route
  - Identified black holes
    - Withdrew discard announcements
  - Identified critical traffic
    - Load-shed less important traffic
    - Identify traffic that cannot go over unencrypted lines

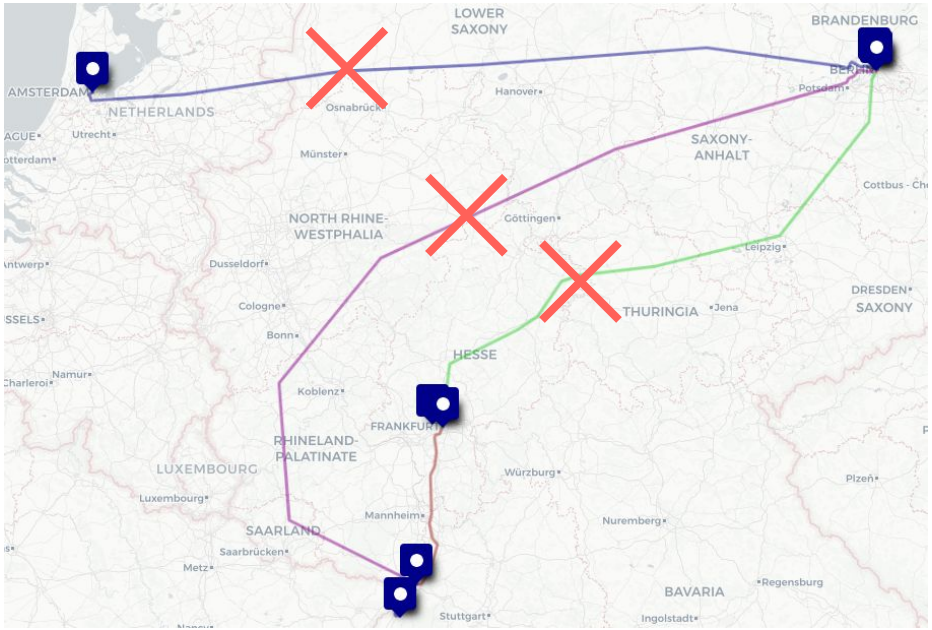


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ ½ route
  - Tunnel-Glue applied
    - Interop bug found
    - Sensitive traffic filtered
    - Lines near capacity

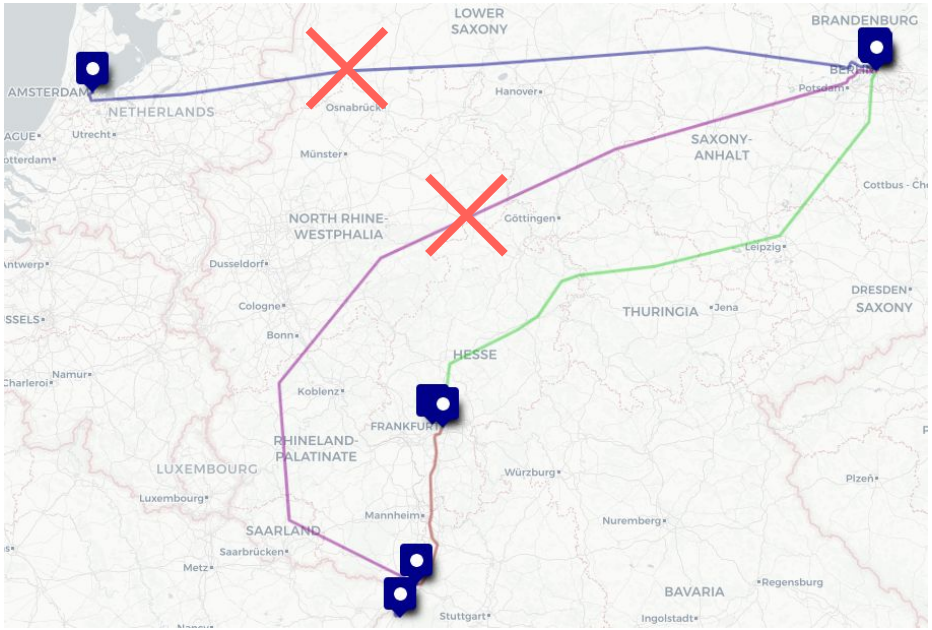


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ **one** route
  - Tunnel-Glue applied
  - DF carrier spliced a new line in Frankfurt

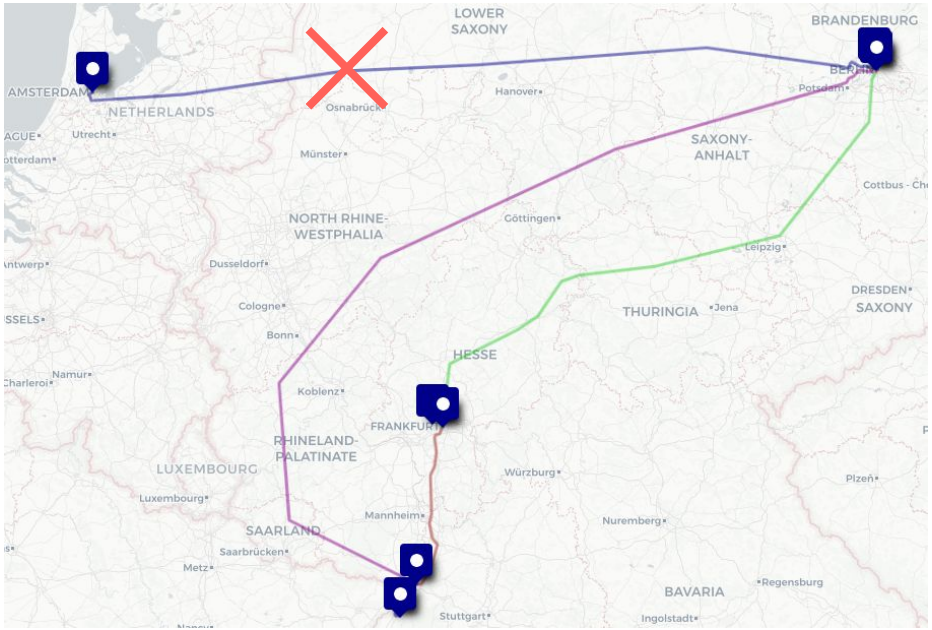


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Gluing AS8560 together

- Berlin has ~~three~~ **two** routes
  - Tunnel-Glue applied
  - DF carrier spliced a new line in Frankfurt
  - DF carrier spliced another line to Karlsruhe

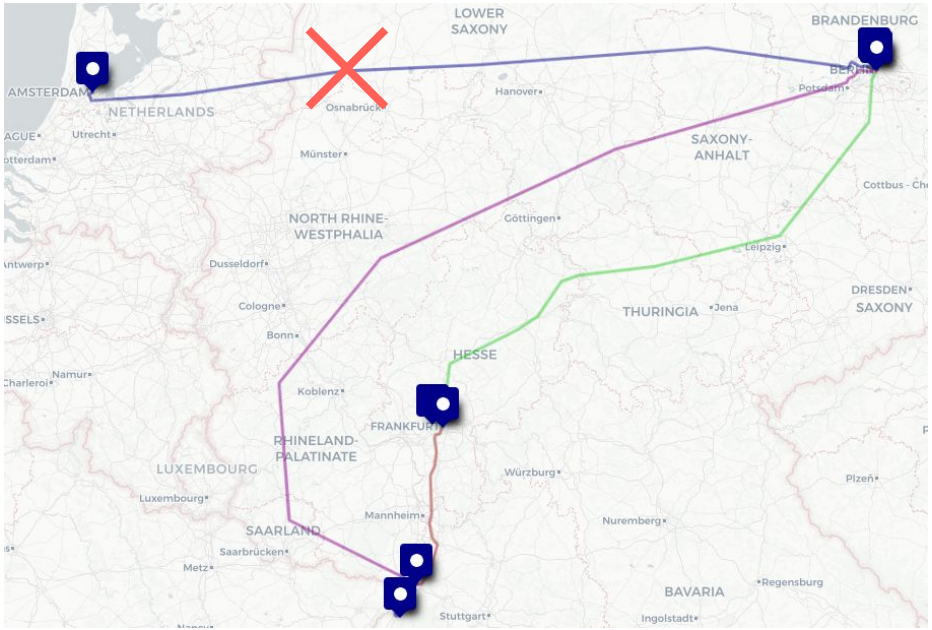


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Life savers

- Communication channels independent of IONOS servers
- Engineers knowing more than they need to
- Lowering bureaucracy with increasing stress levels
- Working on the problem, not worrying about the problem
- The help of dozens of engineers and managers

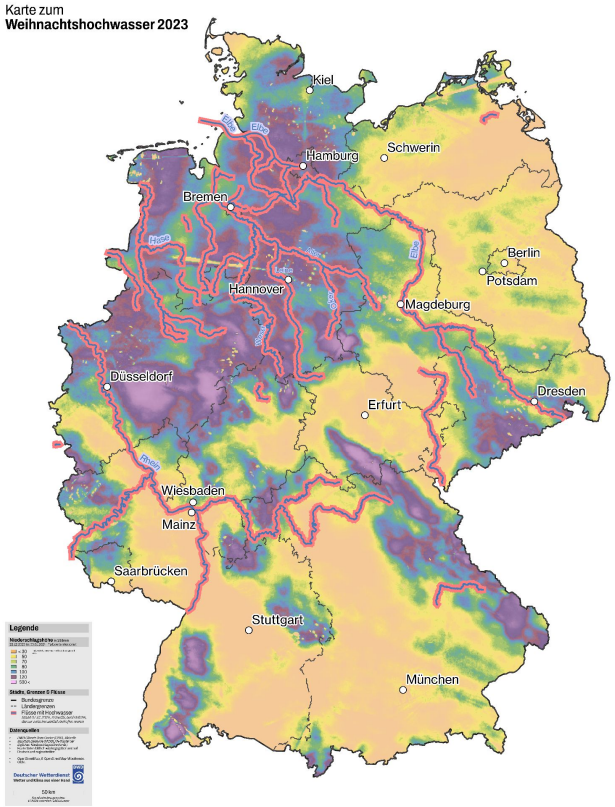


Fictional lines and fictional data center locations.
map data © OpenStreetMap contributors under ODbL , Map tiles by CartoDB, under CC BY 3.0. map data © OpenStreetMap contributors under ODbL

Coming back

Lessons learned

- Non-crossing might not be enough
  - Natural disasters in w. Germany took out two out of three paths
  - Every region needs 2N+1 through at least three exits leaving in different cardinal directions

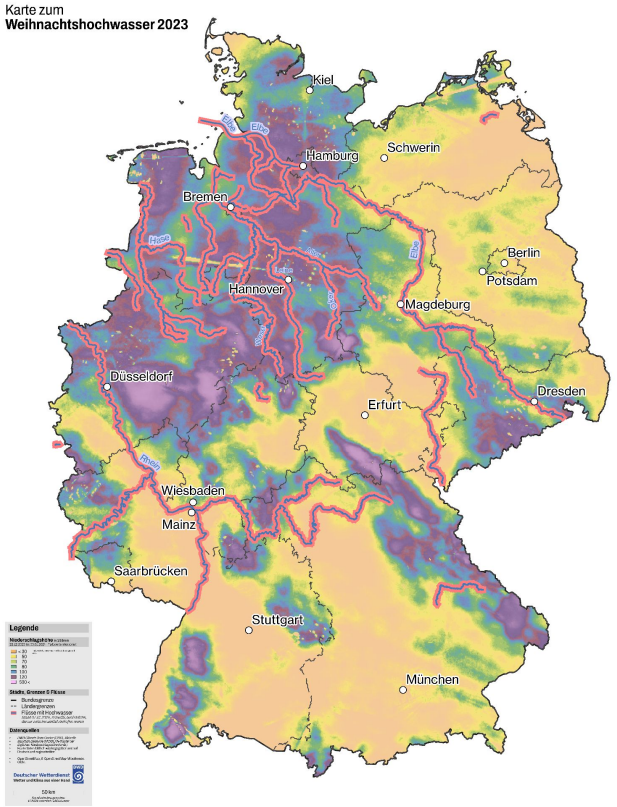


Fictional lines and fictional data center locations.
Weather data: German weather service https://opendata.dwd.de/
Image by user Ovinator using OpenStreetMap

Coming back

Lessons learned

- You cannot count on the AS staying physically connected
  - Move the aggregates
  - Establish last-resort backup paths
- Unexpected issues will occur at the worst time. Have a backup plan for your backup plan
  - Murphy is watching

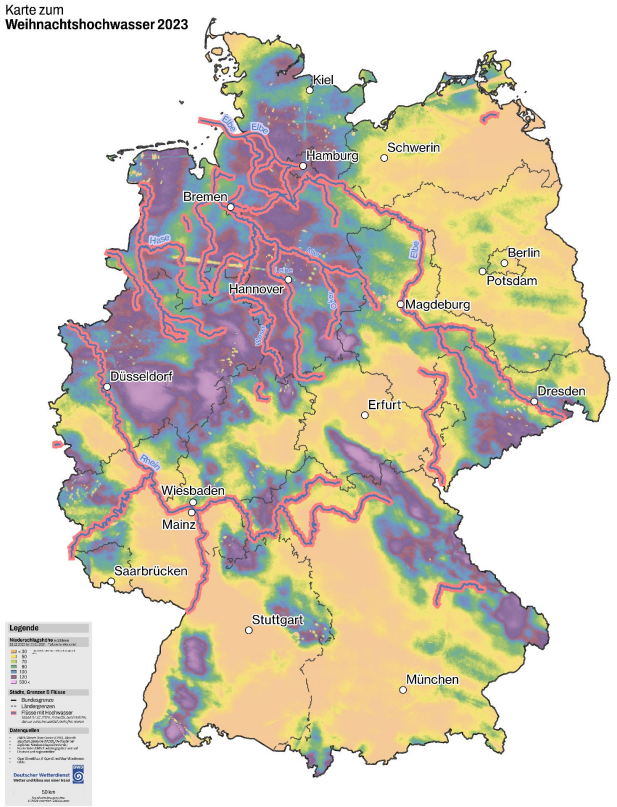


Fictional lines and fictional data center locations.
Weather data: German weather service https://opendata.dwd.de/
Image by user Ovinator using OpenStreetMap

Coming back

Lessons learned

- Know how your network is used
  - It’s not enough to connect A to B
    - Know the most important services
    - Know how they interact
    - People might depend on “add-on” features. Identify who uses the features to prevent further issues

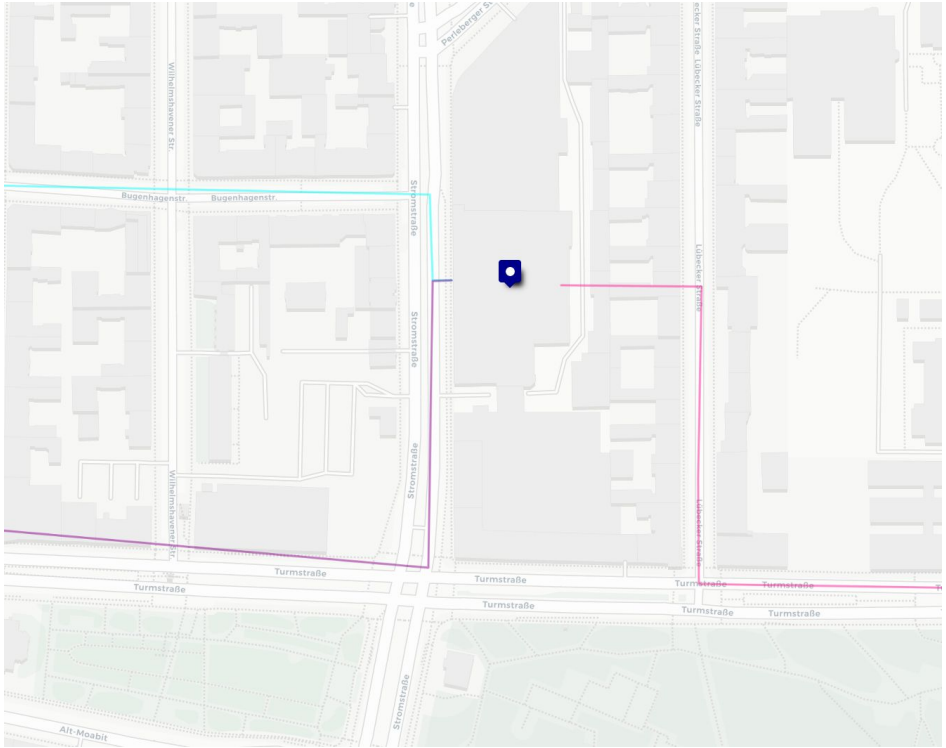


Fictional lines and fictional data center locations.
Weather data: German weather service https://opendata.dwd.de/
Image by user Ovinator using OpenStreetMap

Data center interconnect

Metro regions

- The lines diverge ASAP, should never cross
- More than two house entries often unavailable
- At least two paths leave any DC
- All lines are encrypted

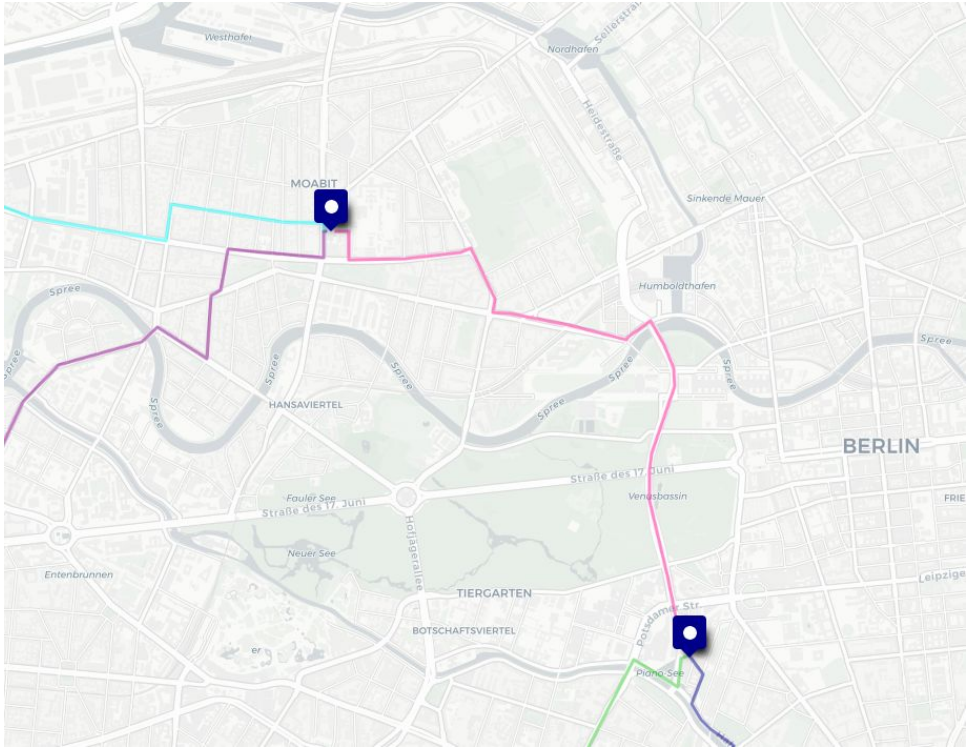


Fictional lines and fictional data center locations.

Data center interconnect

Metro regions

- At least three paths leave the region
- Lots of metro DF illuminated with OTUC2 and OTUC4
  - Transports 100/400GBASE-R
- Encrypted on the WDM or router

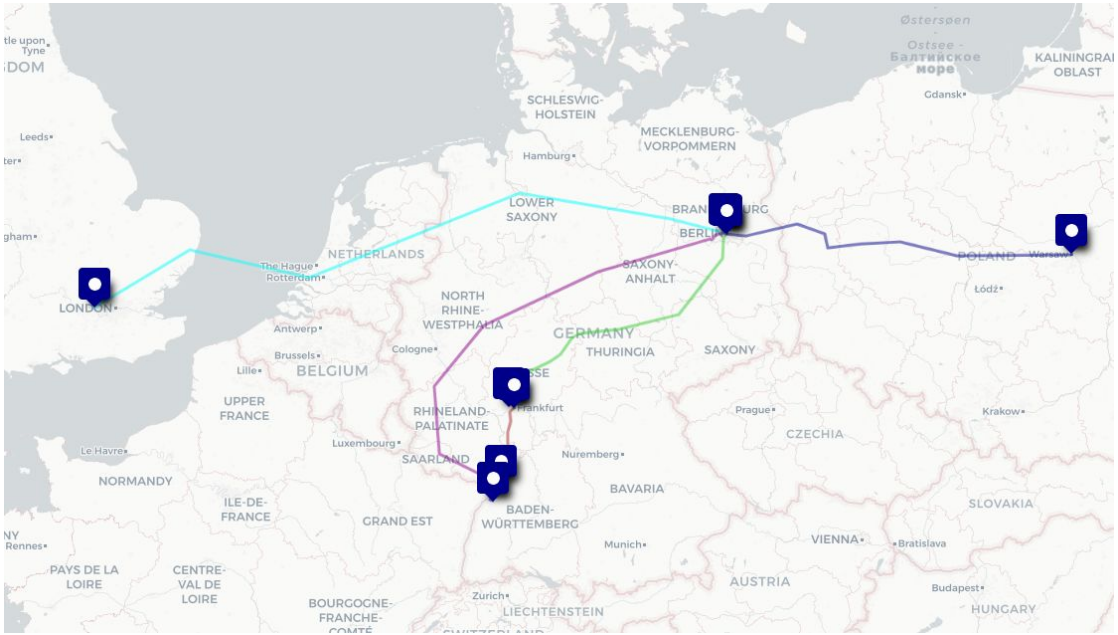


Fictional lines and fictional data center locations.

Data center interconnect

Berlin’s connection

- Longer lines as leased waves
- Traffic is encrypted on the router
- Lines are routed in different cardinal directions
- Partial mesh of tunnels with high metrics prepared

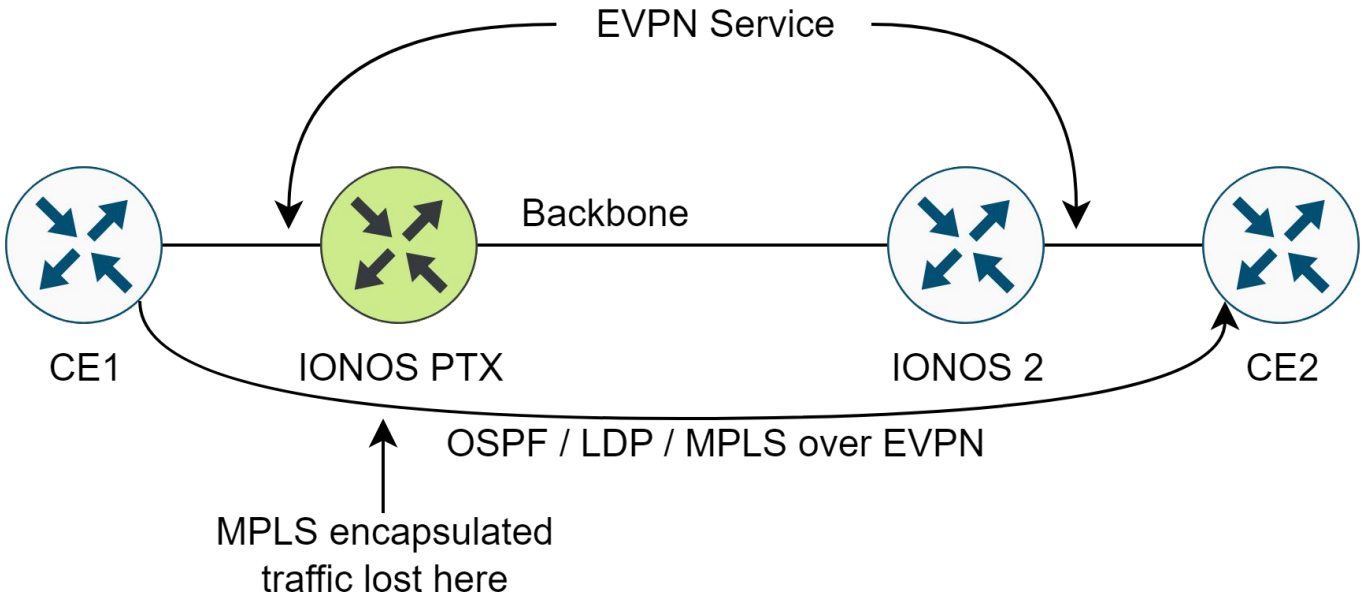


Fictional lines and fictional data center locations.

Interoperability

EVPN-MPLS

- “instance-type evpn” no longer supported on JunosEVO
- MPLS service within EVPN causes traffic loss



Interoperability
MACSEC

- Juniper does not include the SCI per default

  [edit security macsec connectivity-association narina-fantasyland-ca]
  cipher-suite gcm-aes-xpn-256;
  security-mode static-cak;
  include-sci;
  pre-shared-key-chain narina-fantasyland-kc;

- Unnumbered IPv4 breaks sometimes