UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Building your own CGN with Linux

Maximilian Wilhelm

DENOG12

IMT: Zentrum für Informations- und Medientechnologien

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Agenda

- Background
- Why?
- How?
- The struggles
- The solution
- Questions

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Background

- Mid-size German university
  - 20k students
  - 2.5k employees
  - Some guests
  - Main campus
    + 3.5 remote sites
  - Part of eduroam

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

The WIFI challenge

- 10k users (peak, pre-covid)
- 7 IPv4 (public) prefixes
- VLAN mapping magic in radius
  - Based on first character of username
  - Has to be updated each semester
- We‘re out of prefixes to add
- 4Gb/s WIFI traffic

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

The Plan

- Add IPv6
- Use one large private IPv4 prefix
- Get rid of VLAN-mapping magic
- NAT external IPv4 traffic

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

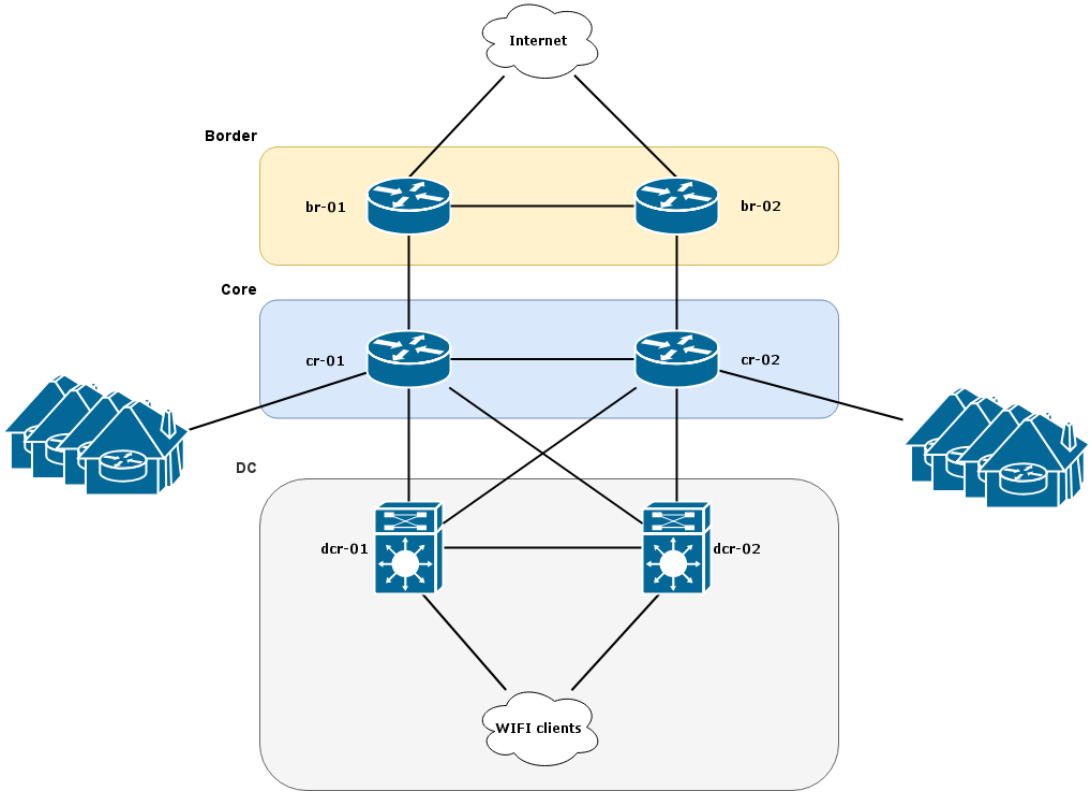
Constraints

- Two groups of users
  - Internal (student, employee, local guest)
  - External (roamed user from remote org)
- Distinction required for some services
  - Intranet
  - Licensed stuff (library, etc.)
  - Has to be represented in IP prefixes

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Network topology

- Classic campus design
  - Border
  - Layer 3 core
  - Distribution
  - Access
  - WIFI with CAPWAP to DC



Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Plan A – use existing boxes

- Simple, border routers should NAT
  - It‘s shiny ASR 9ks, they can do it!
- Turns out, they can‘t
  - ASR 9k need Virtualized Services Module for NAT
  - ✗ Doesn‘t fit into our ASR 9001

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Plan B – buy

- 2x ASR 1001-HX
  - 2RU
  - IOS-XE
  - 4x 10G
  - 4 million NAT sessions
  - 2x 750W PSU
- ~150 k€

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Plan C – build

- Get two commodity servers
- Shove decent NICs into it
- Install (Debian) Linux
- Configure bird
- Configure nftables
- Profit

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Sizing – CPU

- One or two CPUs?
  - Avoid NUMA*
- How many cores?
  - Distribute packet processing load, utilise NIC queues
  - 1 Core routes 3Gb/s
- Intel oder AMD?
  - It‘s gonna be EPYC

➔ Single socket EPYC system

* https://www.youtube.com/watch?v=8NSzkYSX5nY

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Sizing – RAM

- Last peak: 10k users
  - Assume 200 sessions per users
  - ➔ Design for 5m sessions
- One session roughly consists of two 5-tuple + x
  - 5-tuple <= 16B
  - Let‘s asume 48B / session
- 5m * 48B / 10^20 = 228MB

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Sizing – Network

- NIC should have
  - Large Buffers
  - Decent number of queues
  - 2x 10/25G ports
- Not Broadcom & Intel
- ➔ Mellanox ConnectX 4

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Hardware configuration

- 2x Dell R6515
  - 1RU
  - 1x EPYC 7542, 32C
  - 4x 10/25G Mellanox ConnectX 4
  - 32GB RAM
    - >1 billion sessions*
  - 2x 550W PSU
- ~10 k€

*calculated

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

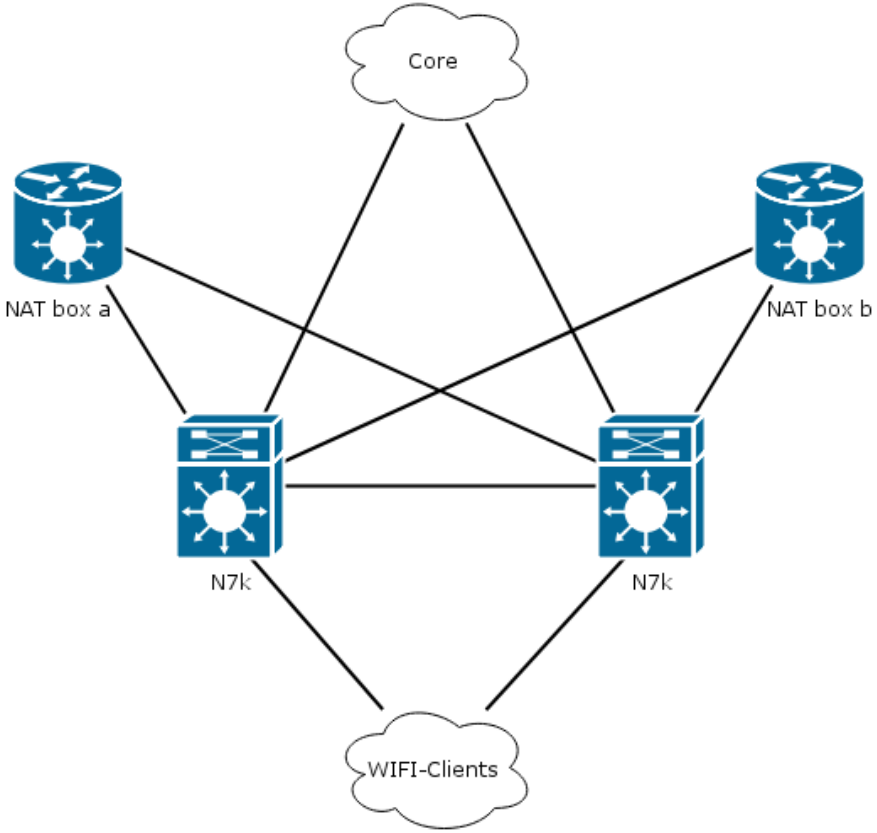
How to integrate CGN into network?

- Active-Active setup would be nice
  - BGP + Anycast
  - Requires ECMP
- Policy
  - Internal traffic should be routed regulary
  - Traffic to external destinations should be NATed
  - ➔ Will need policy-based routing (PBR)

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

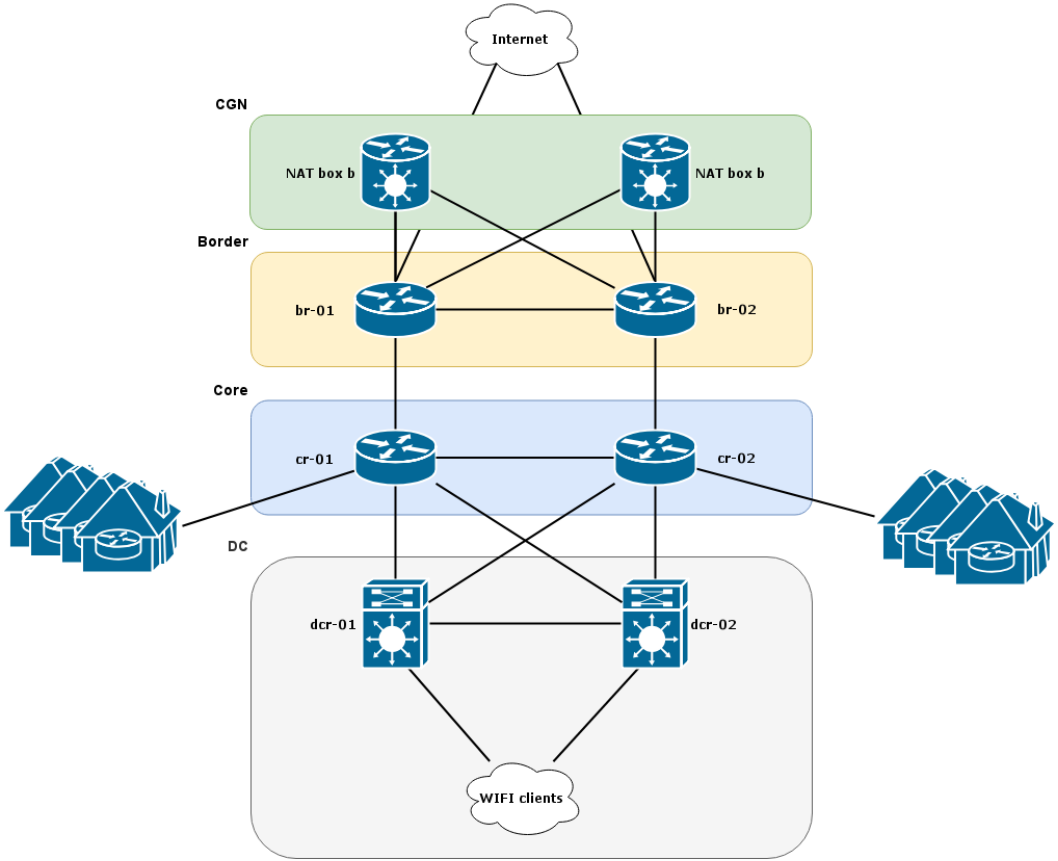
Option 1 – DC

- Nexus 7010
  - NX-OS
  - PBR via route-map
- PBR on SVIs to WIFI clients
  - set next-hop to CGN-IP
- ✗ Only useful for DC networks



Option B – Border

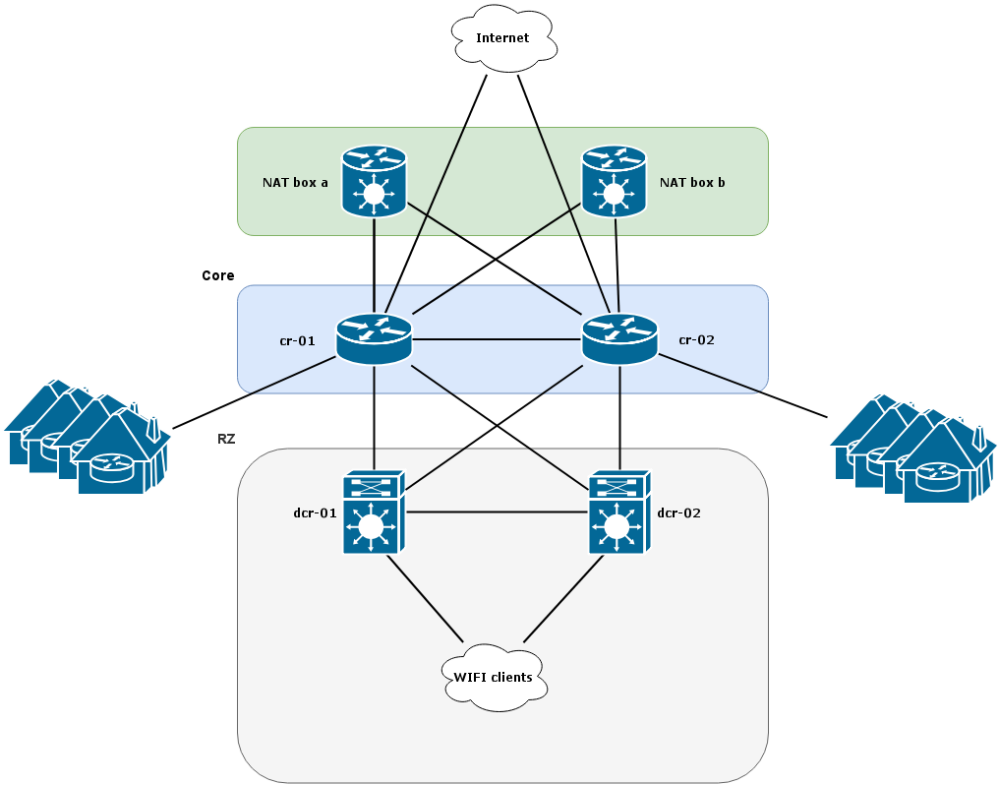
- ASR 9001
  - 4x SFP+ 8x XFP
  - IOS-XR
  - Supports PBR via VRF
- ✗ Router ports expensive
- ✗ Config complicated



Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Option C – Core

- Catalyst 9500-48Y4C
  - IOS-XE
  - PBR via route-map
- PBR on interfaces do DC
  - Easily extendible
- ✔ Ports available, cheap(er) and 25G possible



Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Core it is

- Straight forward setup
  - eBGP Core ↔ CGN nodes
  - Anycast for CGN-srv-IP
  - Route-map with
    - ACL to catch traffic
    - **`set next-hop recursive`**

✗ So you would think...

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

PBR on IOS-XE on Catalyst 9500

- IOS-XE 16.9.4
  - Recommended release when I started
  - Configurable in route-map
  - ✗ Route-map not applied to interface
  - ✗ Log entry that something isn‘t supported
- ➔ Upgrade to IOS-XE 16.12.3e

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

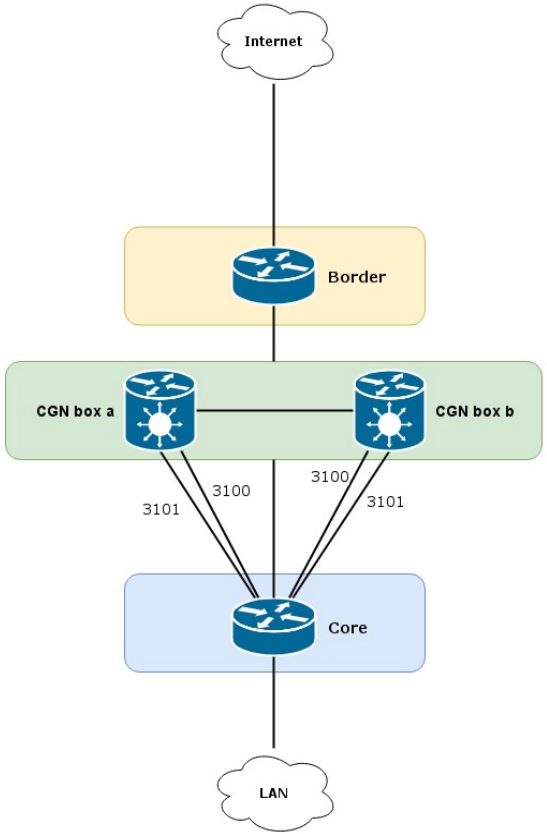
PBR on IOS-XE on Catalyst 9500

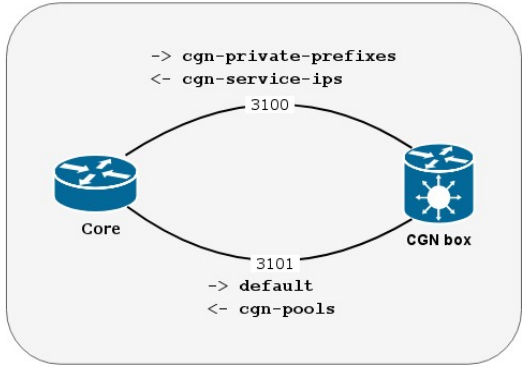
- IOS-XE 16.12.3e
  - Configurable in route-map
  - Applied to interface
  - Seems to work (with one NH)
- 2nd box added, 1st box drained
  - ✗ 1st drained box get‘s ALL traffic
- TAC says
  - ✗ „Not supported on Cat9500“
  - ✗ „Not on the BU roadmap either“

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

The solution

- Two sub-interfaces
  - Internal / external
- BGP in GRT
- VRF cgn
  - Static default route to CGN-srv-IP in GRT
- PBR
  - **`set vrf cgn`**





Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

NAT – Nftables

```
# NATs are good
table ip nat {
  chain postrouting {
    type nat hook postrouting priority 100; policy accept;
    ip saddr 100.64.0.0/12 snat to 192.0.2.0-192.0.2.15
    persistent
    ip saddr 100.127.0.0/16 snat to 203.0.113.240-
    203.0.113.247 persistent
  }
}
```

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Conntrackd

- User-space daemon for conntrack table sync
- Allows sync via unicast or multicast
- Setup
  - Unicast
  - NOTRACK mode
  - Internal/external cache disabled
  - TCPWindowTracking Off
  - ExpectationSync On

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

The CGN box

- Debian stable
- **`bird`** for BGP
  - Plus drain switch
- **`nftables`** for NAT (three rules!)
- **`conntrackd`** for stateful failover
- Bloody details: BLOG

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Bird

```
root@cgn-o2c-01[~]# birdc show route

0.0.0.0/0          via 198.51.100.28 on twe1-2.3101 [cr_cua_01_e 2020-09-15] * (100) [AS65049i]
                   via 198.51.100.40 on twe2-1.3101 [cr_n2a_01_e 2020-09-15] (100) [AS65049i]
                   via 198.51.100.24 on twe1-1.3101 [cr_o2g_01_e 2020-09-15] (100) [AS65049i]

100.64.0.0/10      via 198.51.100.17 on twe1-2 [cr_cua_01_i 2020-09-15] * (100) [AS65049i]
                   via 198.51.100.33 on twe2-1 [cr_n2a_01_i 2020-09-15] (100) [AS65049i]
                   via 198.51.100.9  on twe1-1 [cr_o2g_01_i 2020-09-15] (100) [AS65049i]

# NAT Pools für interne User
192.0.2.0/28       unreachable [nat_pools 2020-09-15] * (200)
192.0.2.32/28      unreachable [nat_pools 2020-09-15] * (200)

# NAT Pools für externe User
203.0.113.240/29   unreachable [nat_pools 2020-09-15] * (200)
203.0.113.248/29   unreachable [nat_pools 2020-09-15] * (200)

# Host-IP
198.51.100.252/32  dev lo [nat_srv_ip 2020-09-15] * (240)

# Anycast CGN Service-IP
198.51.100.254/32  dev anycast_srv [nat_srv_ip 2020-09-15] * (240)
```

Maximilian Wilhelm
Zentrum für Informations- und Medientechnologien

Thank you

- Questions?