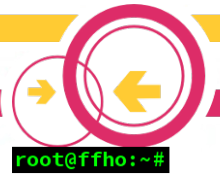


Salt-Orchtestrated Software Defined (Freifunk) Networks

Service-Provider-Netze aus OpenSource-Komponenten bauen

Maximilian Wilhelm
Freifunk Hochstift

Frühjahrsfachgespräch 2017 #ffg2017

Wer bin ich?

- Maximilian Wilhelm
  - @BarbarossaTM
  - @ffho_noc
- Senior Infrastructure Architect, Uni Paderborn
- Vorstand Freifunk Hochstift e.V.
- Linuxer
- Netzwerker
- OpenSource Hacker

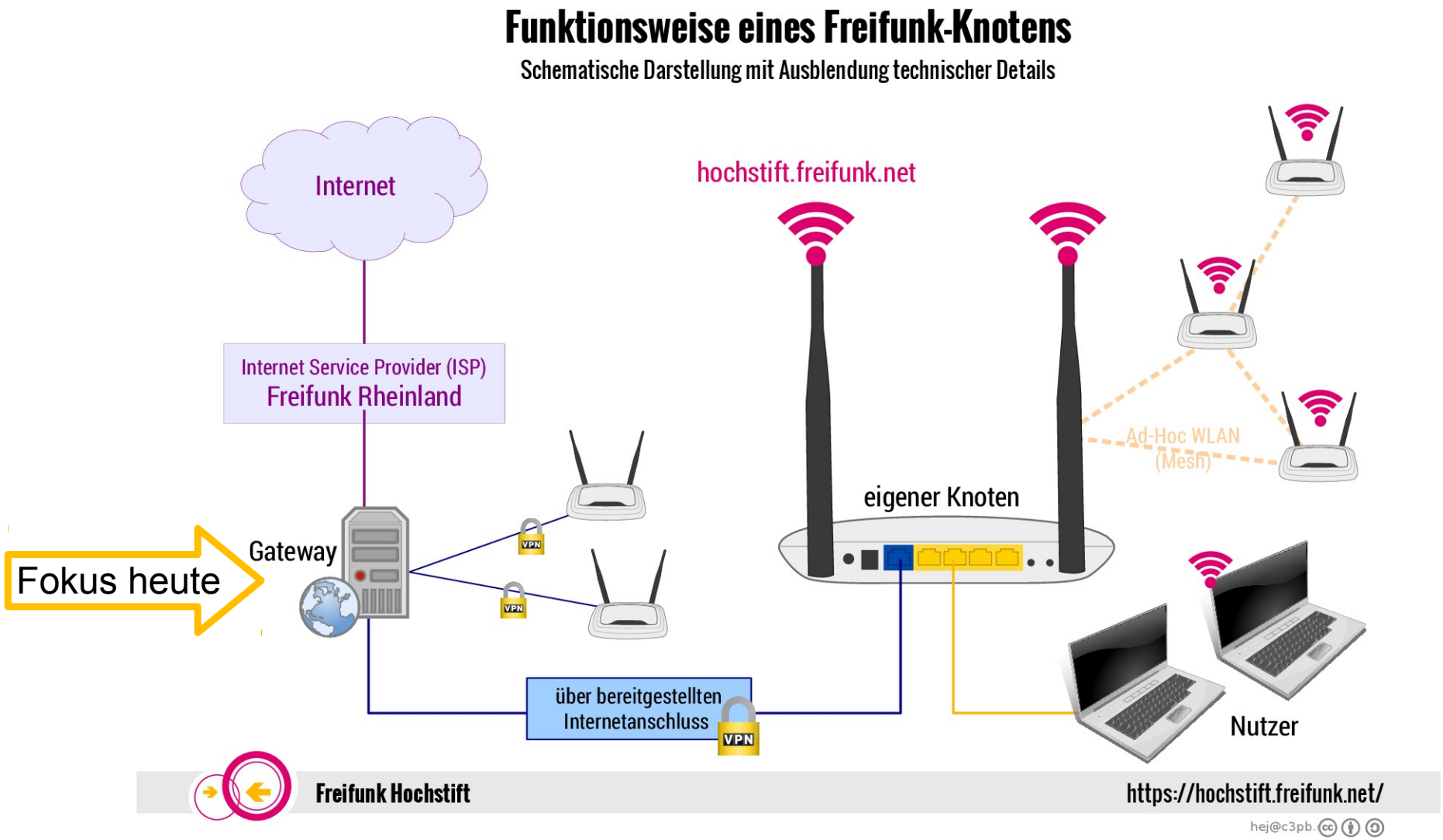
Agenda

- Hintergrund
- Problem Statement
- Automatisierung
  - Ausfallsichere Services
  - Elegantes und sicheres Routing (VRFs)
- Layer2-Overlay (VXLAN)

Background: Freifunk Hochstift

- Gegründet als Freifunk Paderborn (12/2013)
- Schnell stark gewachsen dank viel Unterstützung
  - Bäcker
  - Werbegemeinschaft Paderborn
  - Land NRW
  - Seit 2016 auch Stadt + Kreis Paderborn
- Expandiert zum Freifunk Hochstift (09/2015)
- ~900 Knoten, ≥1.500 Clients

Exkurs: Freifunk-Technik



Exkurs: B.A.T.M.A.N. adv.

- Better Approach To Mobile Adhoc Networking
- Routingprotokoll für Mesh-Netze
- Layer2-Overlay über Layer2-Netz
  - Enkapsuliert Ethernet-Frames in Ethernet
- Klassischerweise: BATMAN über
  - Fastd (L2-VPN)
  - Richtfunk
  - Ad-Hoc WLAN
  - LAN

B.A.T.M.A.N. im Freifunk-Netz

- Historisch: eine große L2-Domain
  - Skaliert nicht
  - ≥100kb/s Grundrauschen (ARP, ND, DHCP, RA, ...)
- ➔ Mittlerweile 10 kleinere L2-Domains
- Pro L2-Domain
  - ein IPv4 + IPv6 Prefix
  - Full-Mesh zwischen allen Gateways nötig
- Jedes B.A.T.M.A.N. Gateway ist DHCP-Server
  - Pro Gateway ein “Sub-Prefix” als DHCP-Pool

Netzwerk-Setup

- Alle Systeme per DualStack verbunden
  - Dynamisches Routing mit Bird
  - OSPF
    - Loopback-Reachability
    - Propagation von Mgmt-Netzen der POPs
  - iBGP mit 3 Core-Routern (RRs)
    - Default-Route (kommt per eBGP)
    - Announcement von Client-Netzen
    - Announcement von (Anycasted) Service-IPs
    - Traffic Engineering

`https://github.com/FreifunkHochstift/ffho-salt-public/tree/master/bird`

Recap

- ✔ Erledigt
  - ✔ Geroutete Infrastruktur
- Next up
  - Automatisierung
  - Ausfallsichere Services
  - Elegantes und sicheres Routing
  - B.A.T.M.A.N.-Overlays

SDN für Linux (Debian)?

- Klassisches ifupdown nur mäßig automatisierbar
- /etc/network/interfaces generieren einfach
- Aber wie neu laden?
  - »service networking restart« disruptive
  - Kein Tool für “neu laden” vorhanden
  - Untrivial zu bauen
- CumulusNetworks Ifupdown2
  - Rewrite von ifupdown in Python
  - https://github.com/CumulusNetworks/ifupdown2

ifupdown2

- Leider keine Featureparität mit ifupdown
- Kann von Hause aus
  - Dependency Resolution
  - `ifreload`
  - VLAN-Aware-Bridges
  - VRFs
  - VXLAN
- Kann (noch) nicht:
  - ppp

Ifupdown2 Patches

- Dank Python leicht erweiterbar
- Upstream kommunikativ
- Kann mittlerweile
  - B.A.T.M.A.N. Interfaces
  - Tunnel (GRE, SIT, IPIP)
- Offene Pull-Requests für
  - Filter für Bridge-Interfaces
  - Phys-dev für VXLAN
  - Pointopoint-Bugfix

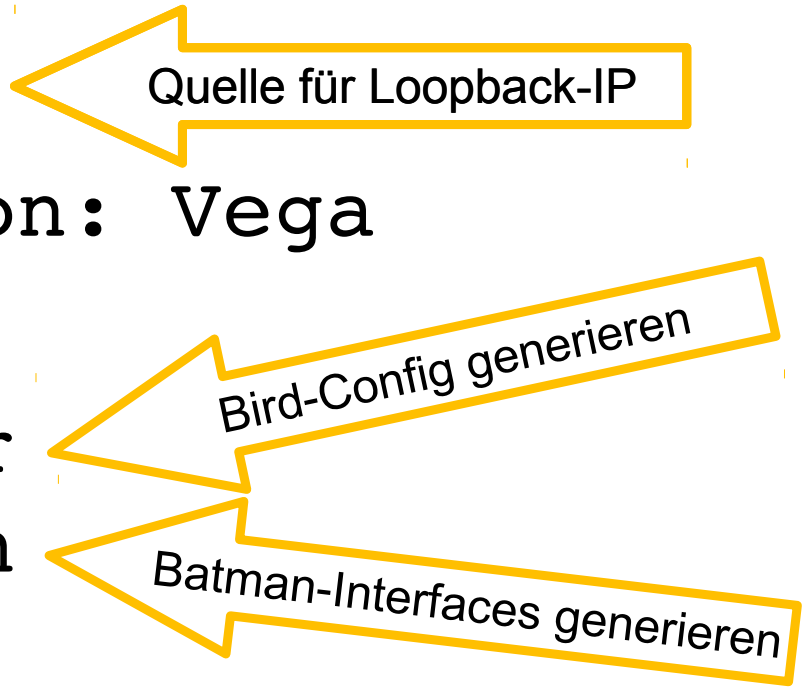
Automatisierung mit SaltStack

- Node-Informationen in »pillar«
  - Strukturierter Key-Value-Store in YAML-Syntax
- Eigene Python Module für “SDN” und mehr
- Daraus generiert:
  - /etc/network/interfaces
  - Bird-Config (OSPF, iBGP, eBGP)
  - OpenVPN
  - DHCPd

`https://github.com/FreifunkHochstift/ffho-salt-public`

Pillar Example

```
bbr-vega.in.ffho.net:
  id: 198
  sysLocation: Vega
  roles:
    - router
    - batman
    - bbr
  sites:
    - pad-cty
```



Pillar Example contd.

```
ifaces:
    bond0:
      bond-slaves: "eth0 eth1 eth2"

    vlan1002:
      desc: "<-> gw04"
      vlan-raw-device: bond0
      prefixes:
        - 10.132.253.58/31
        - 2a03:2260:2342:fe1c::1/126
      batman_connect_sites: pad-cty
[...]                                                    #
```

Recap

- ✔ Erledigt
  - ✔ Geroutete Infrastruktur
  - ✔ Automatisierung
- Next up
  - Ausfallsichere Services
  - Elegantes und sicheres Routing
  - B.A.T.M.A.N.-Overlays

Exkurs: Anycast

- Idee: Service-Prefix mehrfach announcen
- Client verbindet immer zu nahem Server
  - Nähe aus Sicht des Routings!
- Realisierung: anycast-healthchecker + bird
  - Service-Prefix nur announcen, wenn Dienst OK
  - Obacht: Flow-Based ECMP nutzen (ab Kernel 4.4)
- Fällt ein Server aus, “Umleitung” zum nächsten

```
https://github.com/unixsurfer/anycast_healthchecker
https://github.com/FreifunkHochstift/ffho-salt-
public/blob/master/bird/ff-policy.conf
```

Exkurs: Traffic Engineering

- Steuerung von Traffic-Flows
  - Ingress-Traffic steuern
  - Kürzeste Weg zu Batman-Gateway-Prefixen
  - ➔ Ost-West-Traffic vermeiden
- Lösung:
  - More-Specific Routen von/zu gewünschtem Ziel
  - Mit spezieller BGP-Community gekennzeichnet

VRFs

- Virtual Routing and Forwarding
- Unabhängige Routinginstanzen
  - Layer3 Separation
  - Strikte Trennung von Netzen
  - Überlappende Prefixe möglich
- L3-VPNs
  - Üblicherweise im Kombination mit MPLS
- “VRF ohne MPLS” → VRF-lite
  - Hier: VRF-lite

VRFs vs. Policy Routing

- Alt bekannt: Policy-Routing (seit Kernel 2.2)
- Fussschusspotential
  - Rules für v4 und v6 vorhanden?
  - Rules für alle Interfaces vorhanden?
  - Rules für alle Source-Prefixe vorhanden?
  - Pipe-Protokoll in Bird
  - Management-Katastrophe

VRFs vs. Network Namespaces

- Seit Kernel 2.6.24++
- NetNS haben eigene
  - Routingtabellen
  - Routing Policies
  - Netfilter Regeln
- Device Layer separation
- Prozesse müssen in NetNS laufen
- Uncharmant im Freifunk Umfeld
  - “Zuviel des Guten”

VRFs unter Linux

- Separierung für Layer3 Kommunikation
- VRF-Interface als Master für “echte” Interfaces
  - Legt Routing-Table für VRF fest
- Ab Kernel 4.[345] (nehmt >= 4.9)

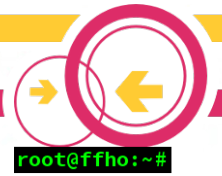
`https://git.kernel.org/cgit/linux/kernel/git/torvalds/linux.git/tree/Documentation/networking/vrf.txt`

`https://cumulusnetworks.com/blog/vrf-for-linux/`

`https://de.slideshare.net/CumulusNetworks/operationalizing-vrf-in-the-data-center`

VRFs unter Linux

```
ip link add VRF_DEVICE type vrf
            table ID

ip link set dev DEVICE
            master VRF_DEVICE
```



VRFs mit ifupdown2

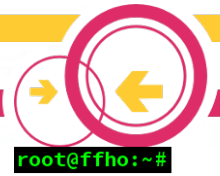
```
auto eth0
iface eth0
        address 185.46.137.163/25
        address 2a00:13c8:1000:2::163/64
        gateway 185.46.137.129
        gateway 2a00:13c8:1000:2::1
        vrf vrf_external


auto vrf_external
iface vrf_external
       vrf-table 1023
```

VRF-Konzept

- Haupt-VRF mit internem Freifunk-Netz
  - OSPF + iBGP
  - Routen zu allen internen Hosts und Diensten
  - Debugging mit Standardtools
- Ggf. “external” VRF für Interfaces mit public IPs
  - GRE-Tunnel zu AS201701
  - OpenVPN-Verbindungen
  - Fastd-Einwahl
  - Eigene public facing services

Inter-VRF-Kommunikation

- Nur in Ausnahmefällen erforderlich
- Erfordert vEth-Paar
  - Quasi virtuelles Netzwerkkabel
- Ein Ende in Haupt-VRF, eins in VRF “external”
- Bird spricht BGP mit sich selbst
  - Exportiert aggregierte Prefix(e)
  - Importiert Public IP(s)
- Public IP(s)s intern redistributiert



Veth unter Linux

```
ip link add VETH_END1 type veth
            peer name VETH_END2
```

```
# ip l
[...]
24: VETH_END2@VETH_END1:   [...]
25: VETH_END1@VETH_END2:   [...]
```

OpenVPN vs. VRFs

- Viele OpenVPN Tunnel im Einsatz
- OpenVPN muss über VRF “external” reden
- Dafür brauchts einen kleinen Patch

```
setsockopt (sd, SOL_SOCKET,
SO_BINDTODEVICE, dev, strlen(dev)
+1);
```

- https://github.com/OpenVPN/openvpn/pull/65

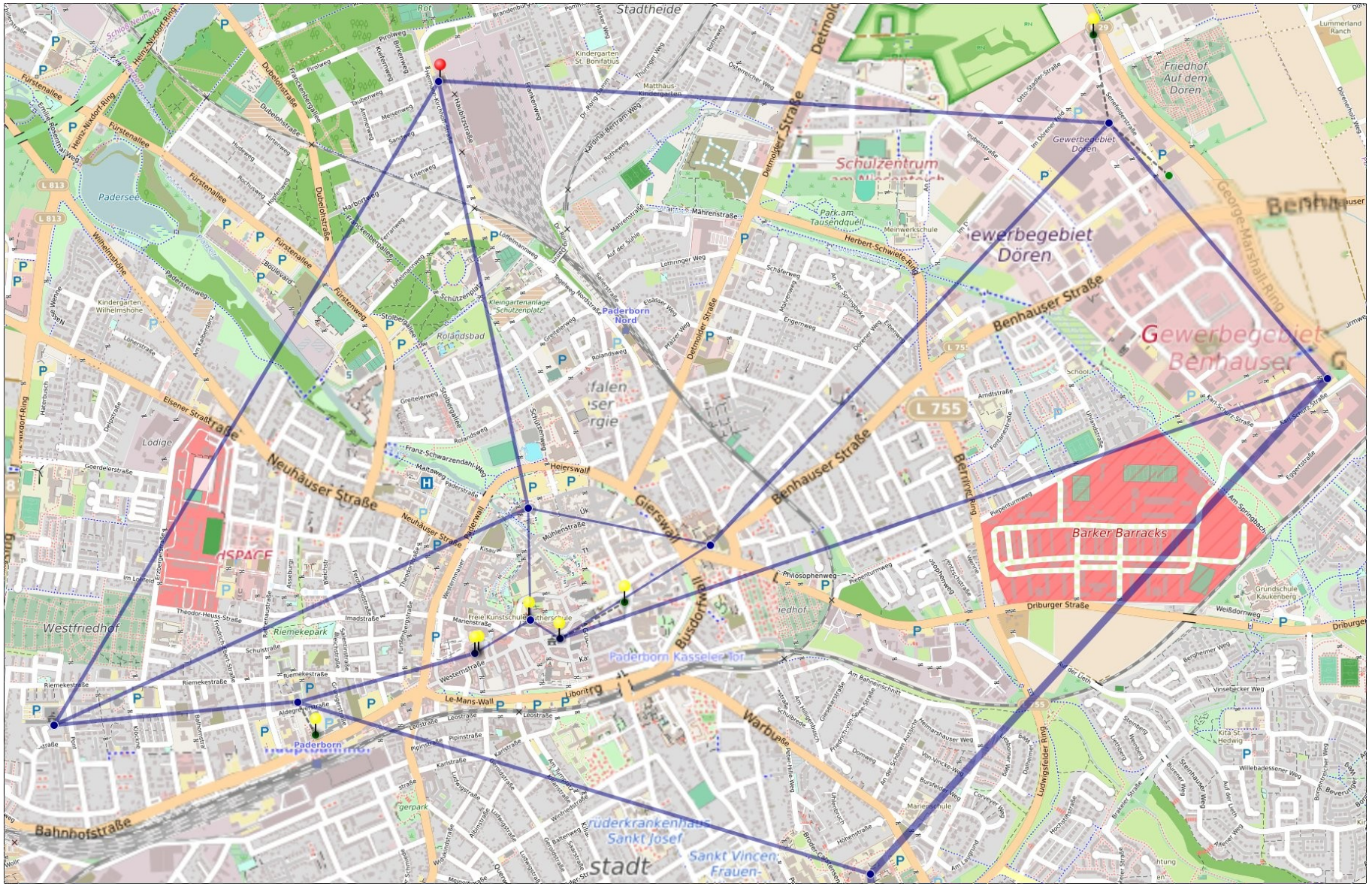
Recap

- ✔ Erledigt
  - ✔ Geroutete Infrastruktur
  - ✔ Automatisierung
  - ✔ Ausfallsichere Services
  - ✔ Elegantes und sicheres Routing
- Next up
  - B.A.T.M.A.N.-Overlays

Recap Freifunk / B.A.T.M.A.N.

- Freifunk basiert auf Batman-Meshes
  - Eine oder mehr Layer2-Domains
  - Ermöglicht netterweise Roaming
    - Das ist in der City total cool
- B.A.T.M.A.N Adv.
  - Layer2-in-Layer2 Mesh-Protokoll
  - Keine Interface-Kosten, nur Hop-Penalty
  - Gateway-Knoten == DHCP-Server
  - Jede Instanz braucht eigene Layer2-Verbindung zu Peers

Wireless Backbone (Planung)

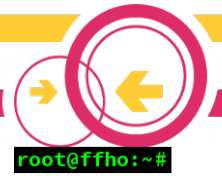


Wege aus der VLAN-Hölle

- B.A.T.M.A.N. braucht Layer2-Verbindung
- IP-Backbone vorhanden
- Layer2-Overlay wäre praktisch!
  - MPLS unter Linux bisher nur tw. Vorhanden
  - ➔ VXLAN

VXLAN

- “Ethernet over UDP”
  - Oder: “Poor mans approach to MPLS”
- Designed als Layer2-Overlay in Datacentern
  - Multi-tenant Overlay über IP-Fabric
  - 24Bit VNI => 16M Instanzen möglich
  - Unicast/Multicast Kommunikation
  - Endpunkt = VTEP (VXLAN Tunnel End Point)
- RFC7348



VXLAN unter Linux

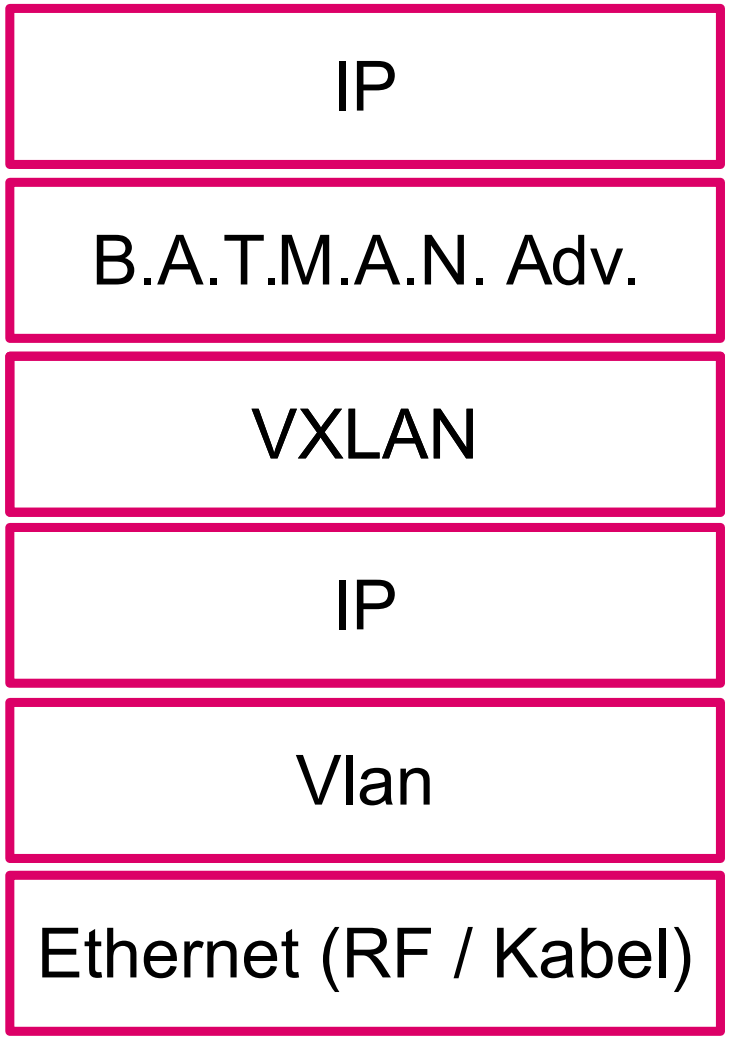
```
ip link add DEVICE type vxlan id ID
[ dev PHYS_DEV ]
[ { group | remote } IPADDR ]
[ local { IPADDR | any } ]
[ … ]
```

VXLAN mit ifupdown2

```
auto vx_v1002_padcty
iface vx_v1002_padcty
        vxlan-id          655617
        vxlan-physdev     vlan1002
        vxlan-svcnodeip 225.10.2.1
        #
        hwaddress f2:00:c1:01:10:02
        mtu 1560
```

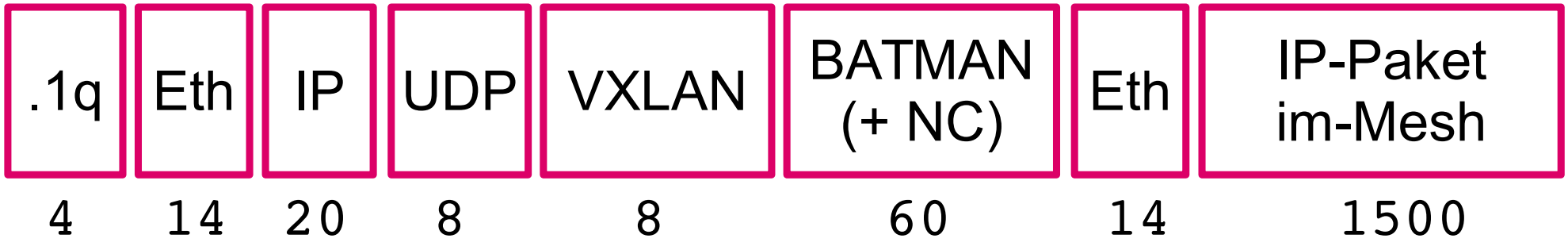
IPoBATMANoVXLANoVLANoRF

- Wait, what?



MTU

- »Increasing MTUs for fun and profit«
  - Mesh-Netz (BATMAN): 1500
  - BATMAN-Underlay (VXLAN): 1560
  - VXLAN-Underlay (VLAN): 1610
  - On-Wire: 1628



https://github.com/FreifunkHochstift/ffho-salt-public/blob/master/_modules/ffho_net.py#L77

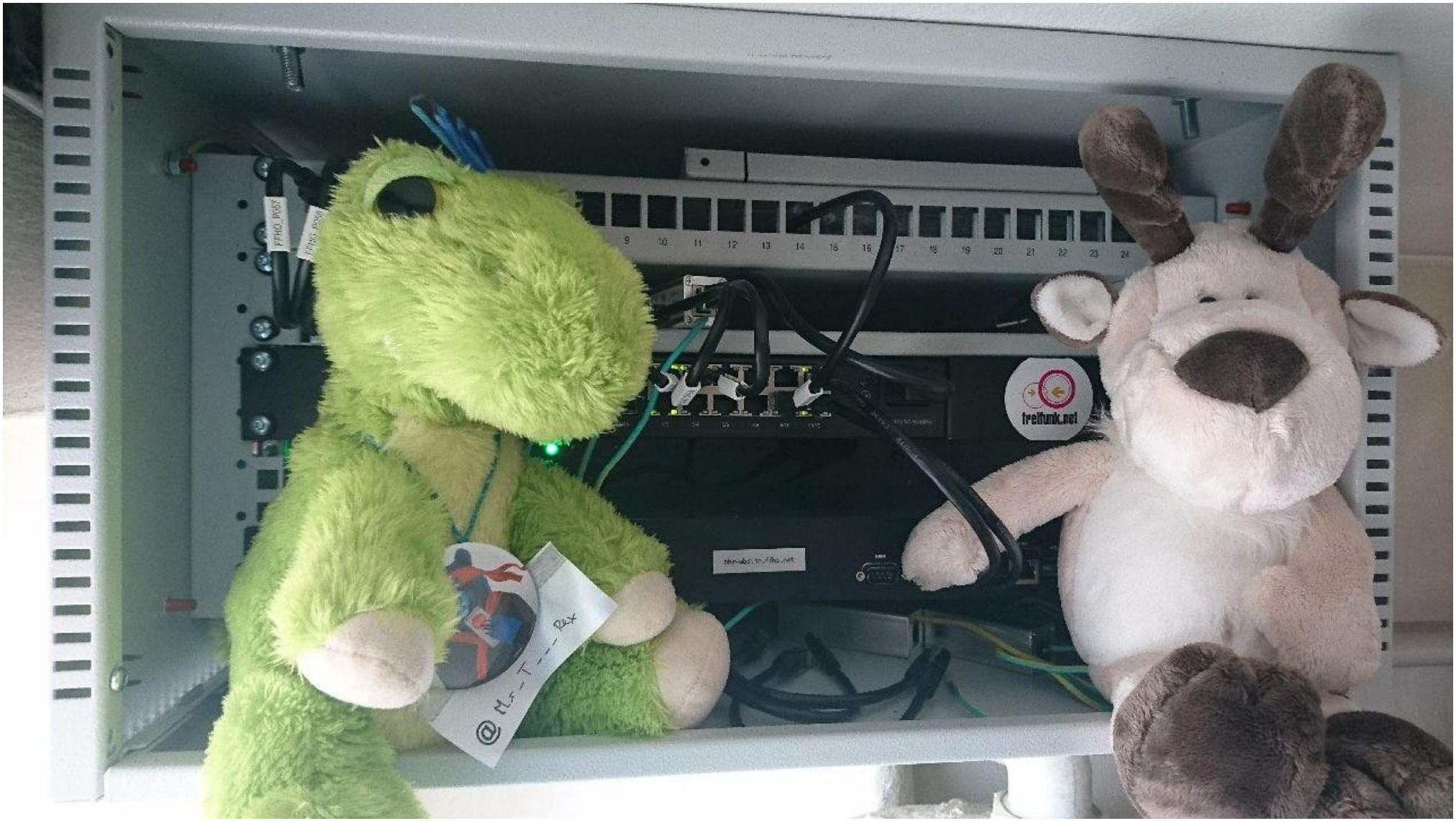
Recap

- Erledigt
  - Geroutete Infrastruktur
  - Automatisierung
  - Ausfallsichere Services
  - Elegantes und sicheres Routing
  - B.A.T.M.A.N.-Overlays

Hardware

- Zoo aus z.T. gesponsorter Hardware
  - Server, Switche, Richtfunk, ..
- Versuch der Homogenisierung
  - PCengines APU2
  - Netonix WISP Switches
  - Ubiquiti Networks
    - PowerBeam
    - LiteBeam
    - AC Mesh Pro

Hardware



Lessons Learned

- Offload ist ein Thema voller Missverständnisse
  - Abschalten! (GS, GRO, GSO, TSO)
  - 4KB/s vs. 40MB/s
- BCP38 ausrollen
- Kernel >= 4.9 nehmen
  - Mit 4.6 / 4.7 IPv6-Routing in VRF subtil kaputt
  - Mit 4.8
    - Problem mit Bridges und B.A.T.M.A.N.
    - IPv6 und Fragmentation

Systemd + OpenVPN vs. ifup

- Einige OpenVPN Instanzen konfiguriert
- “up /etc/openvpn/ifup”
  - ifup “$1”
- Dank systemd starten Instanzen parallel
  - Einige ifup-Aufrufe parallel
  - Nahezu keine IPs mehr konfiguriert

➔ `flock –exclusive –wait 30`

Fragen?

Maximilian Wilhelm @BarbarossaTM

Freifunk Hochstift @ffhochstift + @ffho_noc

Upcoming: VRF support für pppd

- An einem Standort DSL-Uplink vorhanden
- ppp0 sollte in VRF “external”
- Mit post-up scripts geht's leider nicht
- Brauchts wohl auch einen Patch :-)

Backend-Infrastruktur

- Mittlerweile über 60 Systeme
  - Debian Linux auf Blech oder in VMs
  - An immer mehr POPs im Hochstift
  - Server + VMs verteilt in .de
  - Angebunden per RiFu-Backbone oder VPNs
- Konfiguration vollständig per SaltStack verwaltet
- Monitoring per LibreNMS + Icinga2 (upcoming)

Lessons Learned