

Linux Router

mit Alpine Linux



Thomas Liske <liske@ibh.de>

Leiter NOC

Chemnitzer Linux-Tage 2023

whoami

📍 Dresden, Germany

🏢 IBH IT-Service GmbH

👤 @liske

✉ @liske@ibh.social

Beruf

► Leiter NOC

► RZ- & ISP-Infrastruktur

Linux

► seit der 8. Klasse – ist schon länger her

► Distributionen

► Debian

► Alpine Linux



Software-Projekte

► apt-dater

► ifstate

► imvirt

► needrestart

► python-apds9960

...

IBH IT-Service GmbH



RECHENZENTRUM

- Housing / Rackspace
- Dedizierte Hardware
- Cloud-Dienste
- Managed Services



IT-INFRASTRUKTUR

- Netzwerke
- Server
- Storage
- Datensicherung
- Virtualisierung
- IP-Telefonie
- USV-Anlagen



IT-SECURITY

- PEN-Analyse
- Firewall
- Sicherheitszertifikate
- Antivirus
- 2-Faktor-Authentifizierung



PLANUNG UND BAU VON RECHENZENTREN

- Bau oder Ertüchtigung von Rechenzentren
- Planung und Realisierung von Gebäuden, Innenausbau und TGA
- Lieferung und Inbetriebnahme der technischen Innenausstattung
- Betrieb von Rechenzentren für Kunden



PHYSISCHE INFRASTRUKTUR

- Strukturierte Verkabelung
- Projektierung, Installation und Inbetriebnahme von USV-Anlagen
- Elektroanlagen
- Brandmeldeanlagen
- DGUV V3 Überprüfung



IT-MANAGEMENT

- Service- und Wartungsverträge
- Monitoring-Services
- Update-Service
- Managementverträge



INTERNET

- Breitbandversorgung
- Direktverbindungen und SD-WAN
- Domains
- WLAN-Dienst aeroLAN



CONSULTING

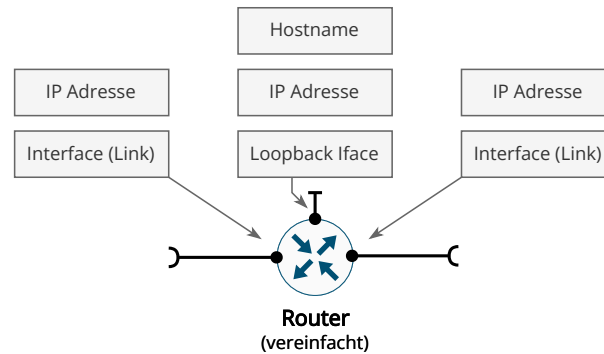
- IT-Infrastruktur-Konzepte
- Workshops und Schulungen

 www.ibh.de

 [@news@ibh.social](https://twitter.com/news@ibh.social)

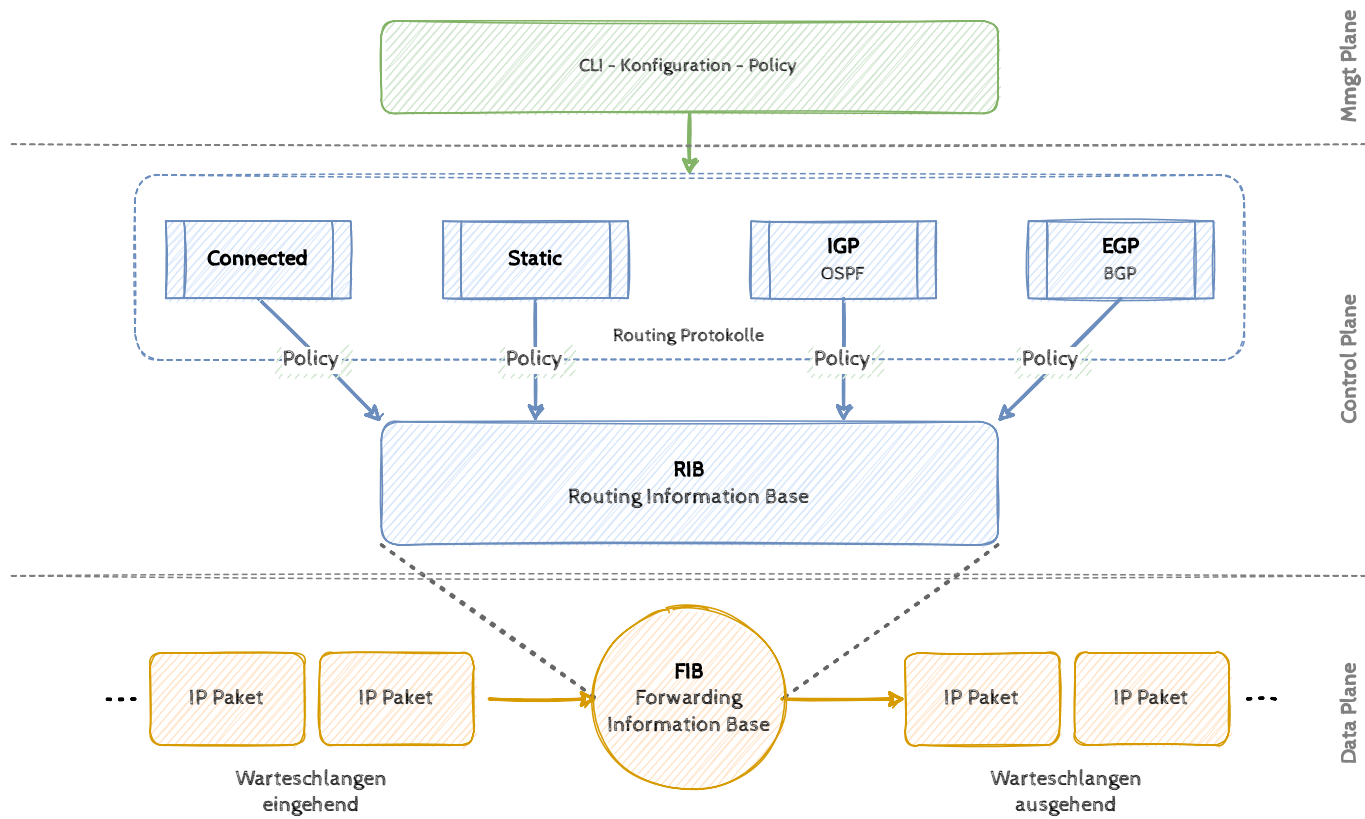
Router

- ▶ arbeitet auf dem Network Layer
Schicht 3 des OSI-Referenzmodells
- ▶ vermittelt Pakete



- ▶ benötigen Wegbeschreibungen:
 - ▶ statische Routen
lokal konfiguriert
 - ▶ dynamische Routen
mittels eines Routing Protokolls gelernt
- ▶ Weiterleitung zum Ziel anhand des aktuell "*besten*" verfügbaren Weges

Aufbau (Prinzip)



Also was von ... ?

Wunschzettel

- ▶ geeignet für Netzwerkschränke (2 Pfosten)
- ▶ serielle Konsole
- ▶ kein Dateisystem was "einfach" kaputt geht (Power Cycle)
- ▶ gespeicherte vs. aktive Konfiguration
- ▶ dynamische Routing Protokolle
OSPF, OSPFv3, BGP
- ▶ Abbilden komplexer Netzstrukturen
VLAN Tags 802.1Q; Q-in-Q 802.1ad; LACP 802.3ad; VRF
- ▶ Abbilden komplexer Routing Policies
- ▶ Full Table IPv4 & IPv6 der DFZ
inklusive Wachstum; RIB & FIB
- ▶ Nachhaltiger Betrieb
 - ▶ offenes System
 - ▶ Nutzung bis Hardware defekt
 - ▶ oder Leistungsgrenze erreicht

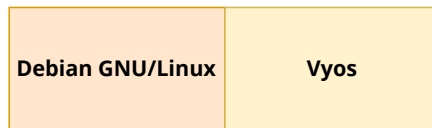
**Wie wir es
~~nicht~~ gelöst
haben...**

Anti-Pattern



Routing Daemon

- ▶ Cisco-like Bedienung & Konfiguration ☯
- ▶ nicht DevOps tauglich ([frr-reload.py](#))
- ▶ konfiguriert IP-Adressen, aber keine Links
unklare Abgrenzung zur Netzwerkkonfiguration des OS



Betriebssystem & Software Stack

- ▶ klassisches Dateisystem
kein read-only Betrieb
- ▶ viel unnötige Dienste
- ▶ Paketverwaltung nicht deklarativ
- ▶ Routing Daemon & Policy (FRR)
- ▶ basiert auf Debian 8 
siehe [Vynos FAQ](#)
- ▶ funktionale Beschränkungen



Hardware

- ▶ hohe Portdichte auf 1HE schwierig
z.B. 32x SFP+
- ▶ im Feld selten Serverschränke verfügbar
- ▶ unnötig aufwendig
Disk Cages, VGA Konsole, ...

Linux Router

mit

Alpine Linux

Hardware



Lanner Network Computing Appliance

4x NIC Slot



Intel Fortville XL710

8x 10GbE



Intel E810-CAM1

4x 25GbE



Intel E810-CAM2

2x 100GbE

Alpine Linux

Überblick

- ▶ hat eine der kürzesten Bootzeiten aller Betriebssysteme
- ▶ als libc wird **musl** verwendet
- ▶ **BusyBox** als coreutils
- ▶ **apk** als deklarativer Paketmanager!
- ▶ **OpenRC** als Init System
- ▶ Anwendungen: Position Independent Executables (PIE) mit Stack Smashing Protection

Einsatz

- ▶ Container
- ▶ Embedded Geräte
- ▶ Software Router

Alpine Linux

Diskless Mode

- ▶ System läuft in einer RAM-Disk
- ▶ Änderungen nicht direkt persistent
- ▶ Konfiguration bzw. geänderte Dateien werden in apkovl.tar.gz Datei gespeichert

Primary Image - SYSLINUX - kernel & initrd - packages - alpine configs 8 GB - sda1	Secondary Image - SYSLINUX - kernel & initrd - packages - alpine configs 8 GB - sda2	TBD 16 GB - sda5
---	---	---------------------

```
sparrow [~]# ls -l /media/sda1/  
apks  
boot  
cache  
ldlinux.c32  
ldlinux.sys  
nca-demo.20171102233000.tar.gz  
nca-demo.20171104231052.tar.gz  
nca-demo.20171105203828.tar.gz  
sparrow.20171119125350.tar.gz  
sparrow.20171119135306.tar.gz  
sparrow.20171120045037.tar.gz  
sparrow.apkovl.tar.gz
```

```
sparrow [~]# lbu status  
U etc/ifstate/config.yml
```

```
sparrow [~]# lbu diff  
--- a/etc/ifstate/config.yml  
+++ b/etc/ifstate/config.yml  
@@ -1,7 +1,8 @@  
  interfaces:  
-- name: eth0  
- addresses: []  
+- name: outside  
+ addresses:  
+   - 192.0.2.1/24  
  link:  
- state: down  
+ state: up  
  kind: physical  
- address: '8c:16:45:dc:b1:ad'  
+ businfo: '0000:00:1f.6'
```

Alpine Linux

Installation

► ISO booten

► Laufwerk partitionieren (BusyBox 😭)

apk add util-linux

fdisk

mkfs.vfat

► Laufwerk bootfähig machen

setup-bootable

► System konfigurieren

setup-alpine

```
Device      Boot Start      End Sectors Size Id Type
/dev/vda1   *      2048 8388607 8386560   4G  c W95 FAT32 (LBA)
```

```
Command (m for help): w
```

```
The partition table has been altered.
```

```
Calling ioctl() to re-read partition table.
```

```
Syncing disks.
```

```
localhost:~# mkfs.vfat -v /dev/vda1
```

```
Device '/dev/vda1':
```

```
heads:16, sectors/track:63, bytes/sector:512
```

```
media descriptor:f8
```

```
total sectors:8386560, clusters:1046275, sectors/cluster:8
```

```
FATs:2, sectors/FAT:8175
```

```
volumeID:6404a0cf, label:''
```

```
localhost:~# setup-bootable -v /dev/sr0 /dev/vda1
```

```
Using /dev/vda1 as target (mounted on /media/vda1)
```

```
Extracting /dev/sr0 to /media/vda1/.new/
```

```
Installing /dev/vda1 to alpine-router-3.17 230214
```

```
Found boot/syslinux/syslinux.cfg
```

```
Flushing cache...
```

```
Replacing existing files...
```

```
Making /dev/vda1 bootable...
```

Alpine Linux

Installationsmedien

► Standard

Basisystem, analog Debian Netinst

► Extended

einige Routing Pakete enthalten

► ...

► Passt nicht?

eigene Images

► Skripte aus **aports**-Repository

<https://gitlab.alpinelinux.org/alpine/aports/>

► eigenes Profil erstellen

► bauen mit mkimage.sh

```
./mkimage.sh --tag 3.17 --arch x86_64 --hostkeys ↵  
--outdir ~/iso ↵  
--repository http://dl-cdn.../alpine/v3.17/main ↵  
--repository http://dl-cdn.../alpine/v3.17/community ↵  
--profile router
```

```
profile_router() {  
    profile_extended  
    profile_abbrev="router"  
    title="Router"  
    desc="Most common used packages included.  
        Suitable for routers.  
        Runs from RAM."  
    arch="x86_64"  
    kernel_addons="xtables-addons zfs"  
    kernel_cmdline="console=tty1 console=ttyS0,115200"  
    boot_addons="intel-ucode"  
    initrd_ucode="/boot/intel-ucode.img"  
    syslinux_serial="0 115200"  
    apks="$apks  
        e2fsprogs run-parts figlet bc shadow coreutils  
        bash procs busybox-extras utmps keepalived  
        bird lldpd dpns wireguard-tools-wg ifstate  
        py3-pygments hostapd libbbpf bpftool bpfmon atop  
        xdp-tools mc fping irtt prometheus-node-exporter  
        smartmontools perl bmon docker docker-compose  
        nftables apt-dater-host doas minicom  
        usb-modeswitch-udev dnsmasq-dnssec  
        open-vm-tools xe-guest-utilities"  
    local k_a  
    for k in $kernel_flavors; do  
        apks="$apks linux-$k"  
        for a in $kernel_addons; do  
            apks="$apks $_-$k"  
        done  
    done  
    apks="$apks linux-firmware"  
}
```

Alpine Linux

eigenes Image

- ▶ alle Pakete, die wir brauchen
- ▶ Air Gap tauglich
- ▶ Upgrade mit **setup-bootable -u**
- ▶ reproduzierbare Systeme
- ▶ weitere Anpassungen

```
profile_router() {  
    # ...  
  
    apkovl="genapkovl-dhcp.sh"  
}
```

```
# ...  
  
mkdir -p "$tmp"/etc  
makefile root:root 0644 "$tmp"/etc/hostname <<EOF  
$HOSTNAME  
EOF  
  
mkdir -p "$tmp"/etc/network  
makefile root:root 0644 "$tmp"/etc/network/interfaces <<EOF  
auto lo  
iface lo inet loopback  
  
auto eth0  
iface eth0 inet dhcp  
EOF  
  
mkdir -p "$tmp"/etc/apk  
makefile root:root 0644 "$tmp"/etc/apk/world <<EOF  
alpine-base  
EOF  
  
rc_add devfs sysinit  
rc_add dmesg sysinit  
rc_add mdev sysinit  
rc_add hwdrivers sysinit  
rc_add modloop sysinit  
  
rc_add hwclock boot  
rc_add modules boot  
rc_add sysctl boot  
rc_add hostname boot  
rc_add bootmisc boot  
rc_add syslog boot  
  
rc_add mount-ro shutdown  
rc_add killprocs shutdown  
rc_add savecache shutdown  
  
tar -c -C "$tmp" etc | gzip -9n > $HOSTNAME.apkovl.tar.gz
```

Netzwerkkonfiguration

ifupdown

bis Alpine 3.13

► es gab ja nichts besseres

IfState

für statische Konfiguration

- logische und physische Links
- Link Parameter (Speed, Duplex)
- Sysctl (Reverse Path Filter)
- Traffic Shaping
- IP-Adressen

Routing

The BIRD Internet Routing Daemon

Sponsored by CZ.NIC

*BIRD is used in several Internet exchanges, such as the London Internet Exchange (**LINX**), **LONAP**, **DE-CIX** and **MSK-IX**) as a route server, where it replaced Quagga because of its scalability issues. According to the 2012 Euro-IX survey BIRD is the most used route server amongst European Internet exchanges.*

Features:

- ▶ IPv4 & IPv6
- ▶ IPv6 RA
- ▶ BGP
- ▶ OSPF
- ▶ Powerful language for route filtering
- ▶ ...

Accounting

pmacct - promiscuous mode traffic accountant

Features:

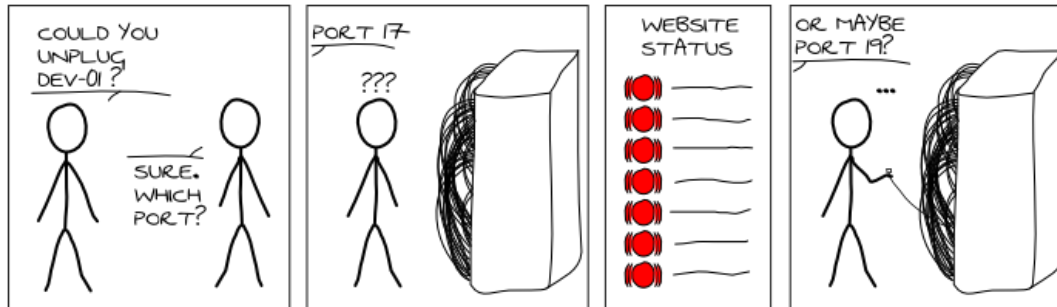
- ▶ entwickelt für ISPs, IXPs
- ▶ IPv4 & IPv6 Support
- ▶ statistische Aggregation des Traffics
- ▶ diverse (export) Protokolle
- ▶ ...

Nutzer:

- ▶ DE-CIX
- ▶ Netflix
- ▶ Spotify

Link Layer Discovery Protocol (802.1AB)

lldpd



Features:

► diverse Protokolle

LLDP, CDP, EDP, SONMP, FDP

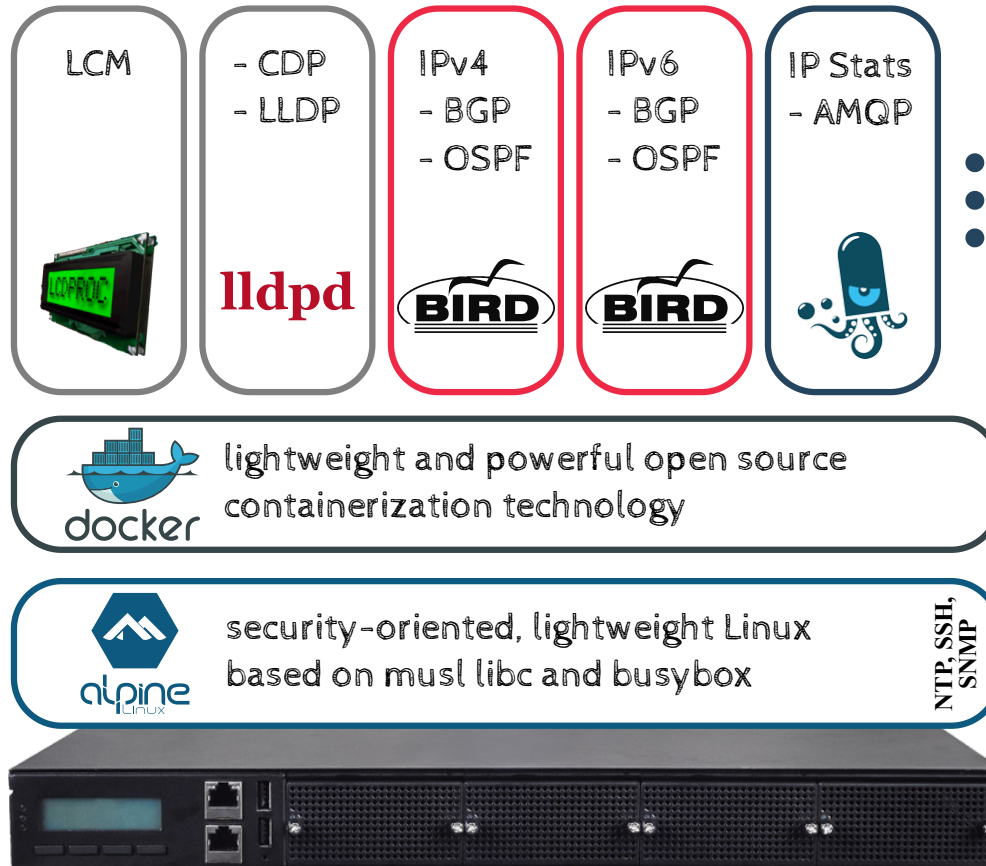
► SNMP Subagent

```
sparrow [~]# lldpctl
```

```
-----  
LLDP neighbors:  
-----
```

```
Interface:      eth0, via: LLDP, RID: 10, T  
Chassis:  
  ChassisID:      mac 00:c8:8b:9b:81:c0  
  SysName:         dc-cl-cs1.ibh.net  
  SysDescr:        Cisco IOS Software, IOS  
                   Technical Support: http  
  TTL:             120  
  MgmtIP:          212.111.232.41  
  MgmtIP:          2a01:7700:0:1f::30  
  Capability:      Bridge, on  
  Capability:      Router, on  
Port:  
  PortID:          ifname Gi2/5  
  PortDescr:       L2T 1G sparrow.ibh.net  
  PMD autoneg:     supported: yes, enabled  
    Adv:            10Base-T, HD: yes, FD  
    Adv:            100Base-TX, HD: yes,  
    Adv:            1000Base-T, HD: yes,  
    MAU oper type: 1000BaseT-FD - Four-p  
  VLAN:            1, pvid: yes  
-----
```

Software Stack 2017



auf Basis von Alpine Linux 3.7

Performance

Messgrößen

- ▶ minimale Latenz
- ▶ maximale Paketrage

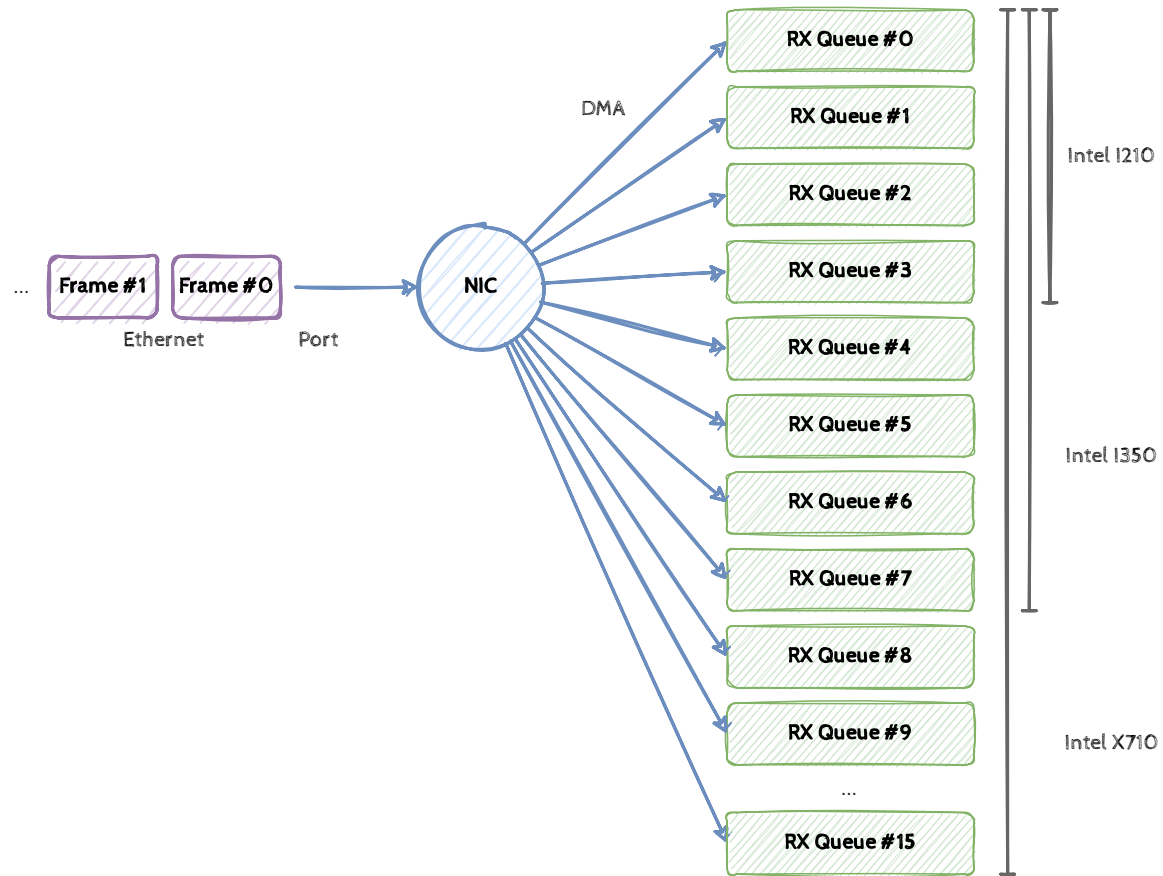
Einflüsse

- ▶ Übertragung
 - ▶ CPU-Cores & -Takt
 - ▶ PCIe-Performance (GT/s)
 - ▶ RAM-Zugriff
- ▶ Paketverarbeitung

Interface	100MbE	1GbE	10GbE	25GbE	100GbE
PPS	150 Kpps	1,5 Mpps	15 Mpps	37 Mpps	149 Mpps

Quelle: Vincent Bernat

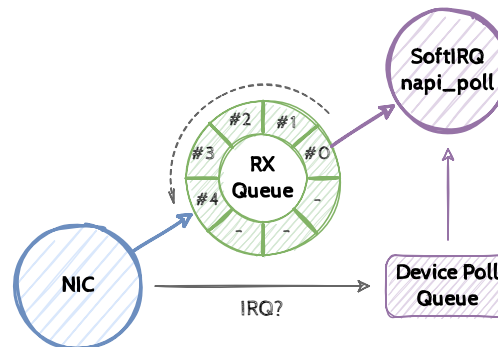
Netzwerkkarten



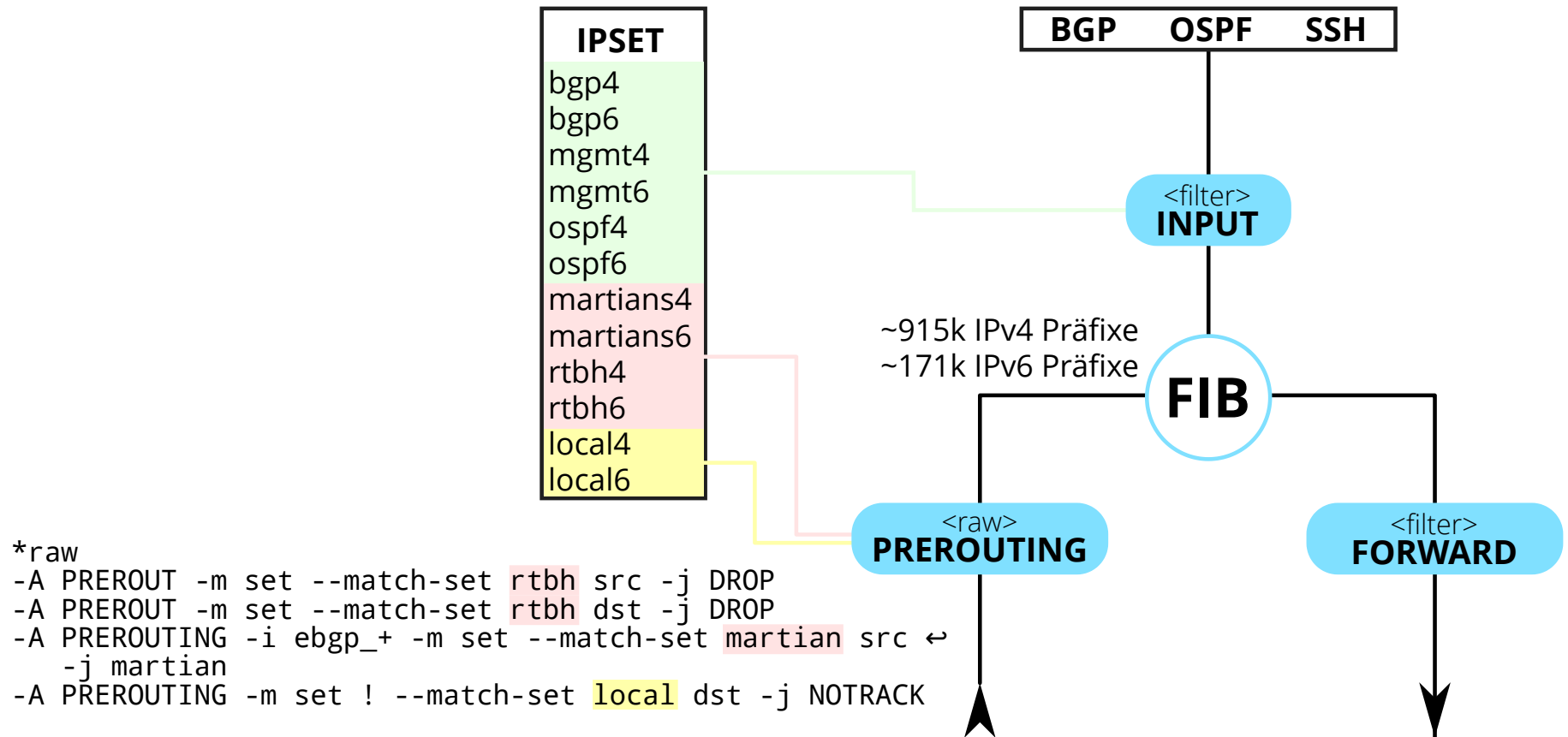
Pakete im Linux Kernel

New API (NAPI)

- Problem: ein Paket auf NIC empfangen, dann IRQ auslösen
→ IRQs sind teuer, skalieren nicht
- Lösung: Polling
 - Pakete via DMA in Ringpuffer (*ohne* CPU)
 - Pakete im regelmäßigen Polling verarbeiten (SoftIRQ)
 - zu viele Pakete → Ringpuffer schreibt weiter (kein Trashing)
 - nicht effizient bei geringem Traffic
- Treiber wechselt lastabhängig zwischen Interrupt und Polling Modus



Paketfluss* Netfilter



*) stark vereinfacht

Durchsatz

TRex - Realistic Traffic Generator

- Software zur Lastmessung
- basiert auf dem Data Plane Development Kit (DPDK)

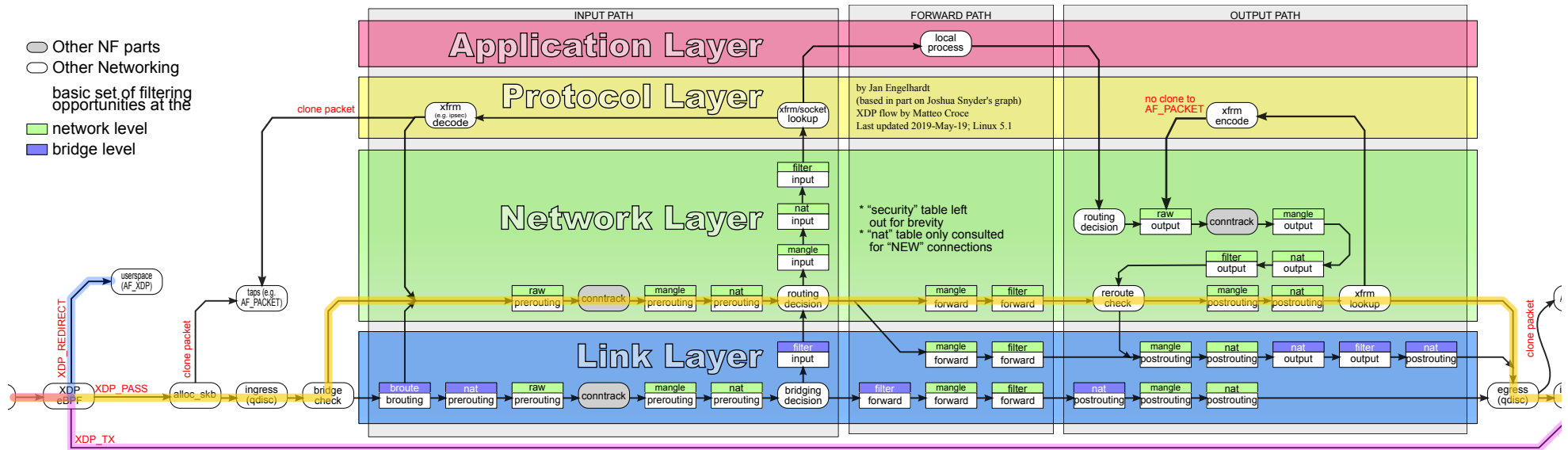
"Testszenario"



- Intel E-2236 6C/12T @ 3,4 GHz | 25GbE: AMD EPYC 7313 16C/32T @ 3,0 GHz
- Device Under Test

	Intel I210 1 GbE	Intel I350 1 GbE	Intel X710 10 GbE	Intel E810 25 GbE
64B	408 kpps 208 Mbps	420 kpps 214 Mbps	5.000 kpps 2,57 Gbps	6.940 kpps 3,55 Gbps
1518B	81,3 Kpps 984 Mbps	81,3 Kpps 984 Mbps	813 kpps 9,86 Gbps	2.031 kpps 24,66 Gbps

Paketfluss



Optionen

- klassisches Forwarding
- eXpress Data Path
- XDP mit User Space Forwarding
- DPDK basiertes Forwarding (nicht im Bild)

Bild: Jan Engelhardt

Durchsatz (2)

	Intel I210 1 GbE	Intel I350 1 GbE	Intel X710 10 GbE	Intel E810 25 GbE
64B Kernel	408 kpps	420 kpps	5.000 kpps	6.940 kpps
64B XDP	1.340 kpps	1.355 kpps	12.000 kpps	16.200 kpps
	328 %	323 %	240 %	233 %

xdp-router

<https://github.com/Nat-Lab/xdp-router/>

- minimaler XDP Unicast Router mit VLAN Support
- eBPF Programm
 - ~90 Zeilen C
- Routing Entscheidung durch Kernel FIB
 - bpf_fib_lookup

Zustand eines Routers (Firewall)

Statisch

- Pakete → Alpine Router ISO
- Konfiguration → apkovl.tar.gz

Orchestrierung
reproduzierbar

Dynamisch

- dynamische Routen (RIB) → Routing Daemon
- Neighbour Tables → Linux Kernel
- Connection Tracking → Linux Kernel

Flüchtig
kurze Konvergenzzeit

Orchestrierung

Ansible Deployments

- grundsätzlich in git Repositories
 - internes GitLab
- eigene Repositories für verschiedene Aufgaben
 - z.B. Edge-, PoP- oder CPE-Router
- ~25 standardisierte Rollen
- keine automatische Aktivierung von Konfigurationsänderungen (Data Plane)

Rolle	Inhalt
base	Grundkonfiguration (\$EDITOR, SSH-Keys, ...)
bgp-communities	Variablen für BGP (Large) Communities
bird	BIRD Routing Daemon
chassis	Variablen für Hardware Spezifika
ifstate	Netzwerkkonfiguration
ipset	Konfiguration Kernel ipsets
lm-sensors	Konfiguration Hardware Sensoren
...	...

Router Recovery

- Konfiguration
 - reproduzierbar aus VCS
- Alpine Router ISO
 - internes Artefakt
- Hardware-Konfiguration
 - Spare Pool

Contribution

- wir (alle) nutzen FOSS
- wir (alle) leben von FOSS

Unserer Beitrag. Aus unserem Alltag.

- Bug Reports & MR
- Software-Projekte
- Paketbetreuung

accel-ppp

bird_exporter

bird-lg-go

dpns

ifstate

irtt

lcdproc

py3-pyroute2

py3-wgnlpy

xdp-tools

Und euer Beitrag? Aus eurem Alltag?

- Sharing is Caring

Zusammenfassung

Status Quo

- ▶ Plattform hat sich bewährt
- ▶ fehlende Pakete für Alpine paketiert
- ▶ Daily Business

liske@ibh.de 
@liske@ibh.social 

Weiterentwicklung

- ▶ harte Mandantenfähigkeit (NetNS statt VRF)
- ▶ Evaluation
 - ▶ Linux *eXpress Data Path* (XDP)
 - ▶ FD.io *Vector Packet Processing* (VPP)

Die Präsi enthält...

- `animate.css`
- `highlight.js`
- `fork-awesome`
- `impress.js`
- `impress.js-progress`
- `impress-console`