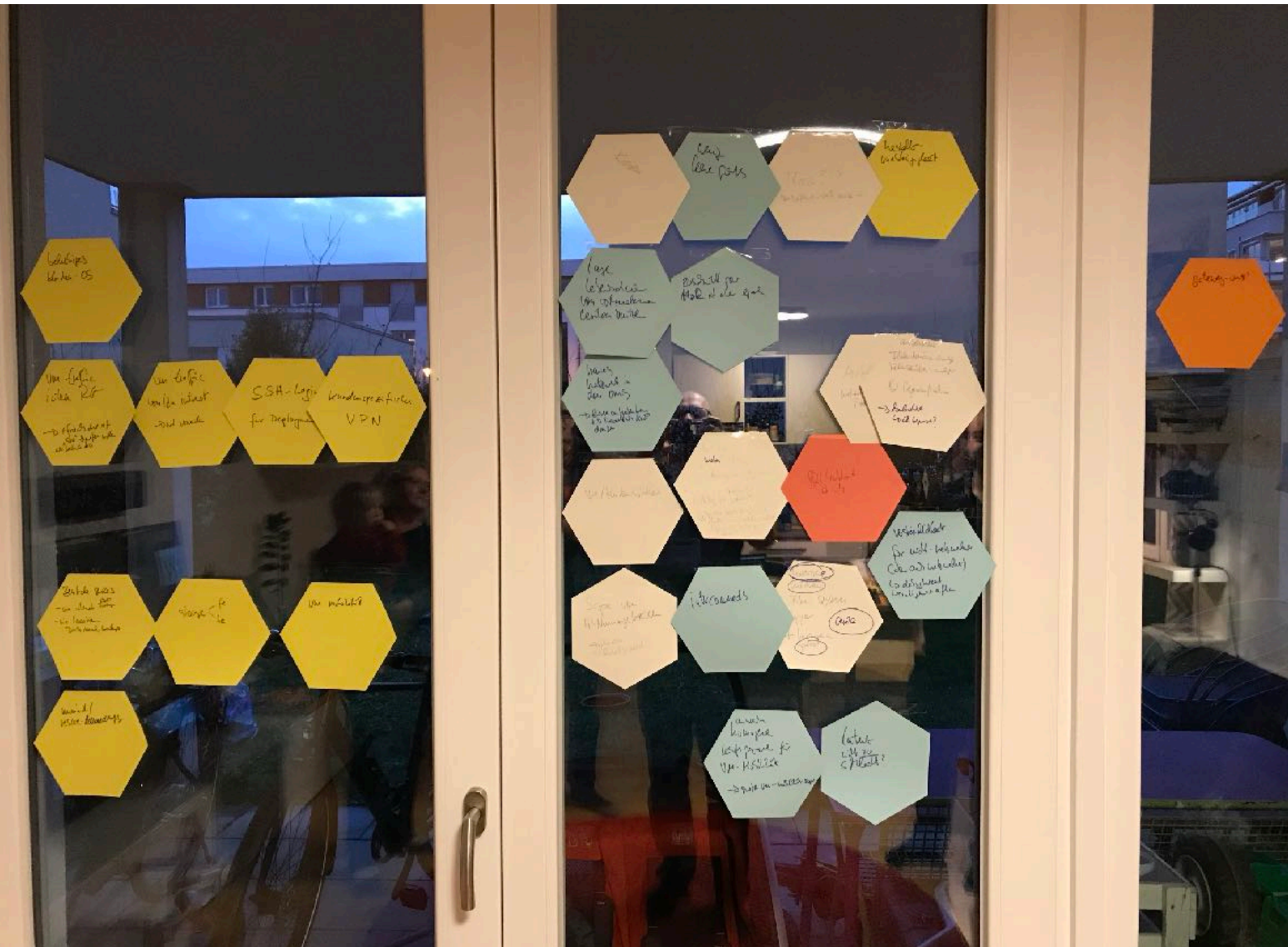
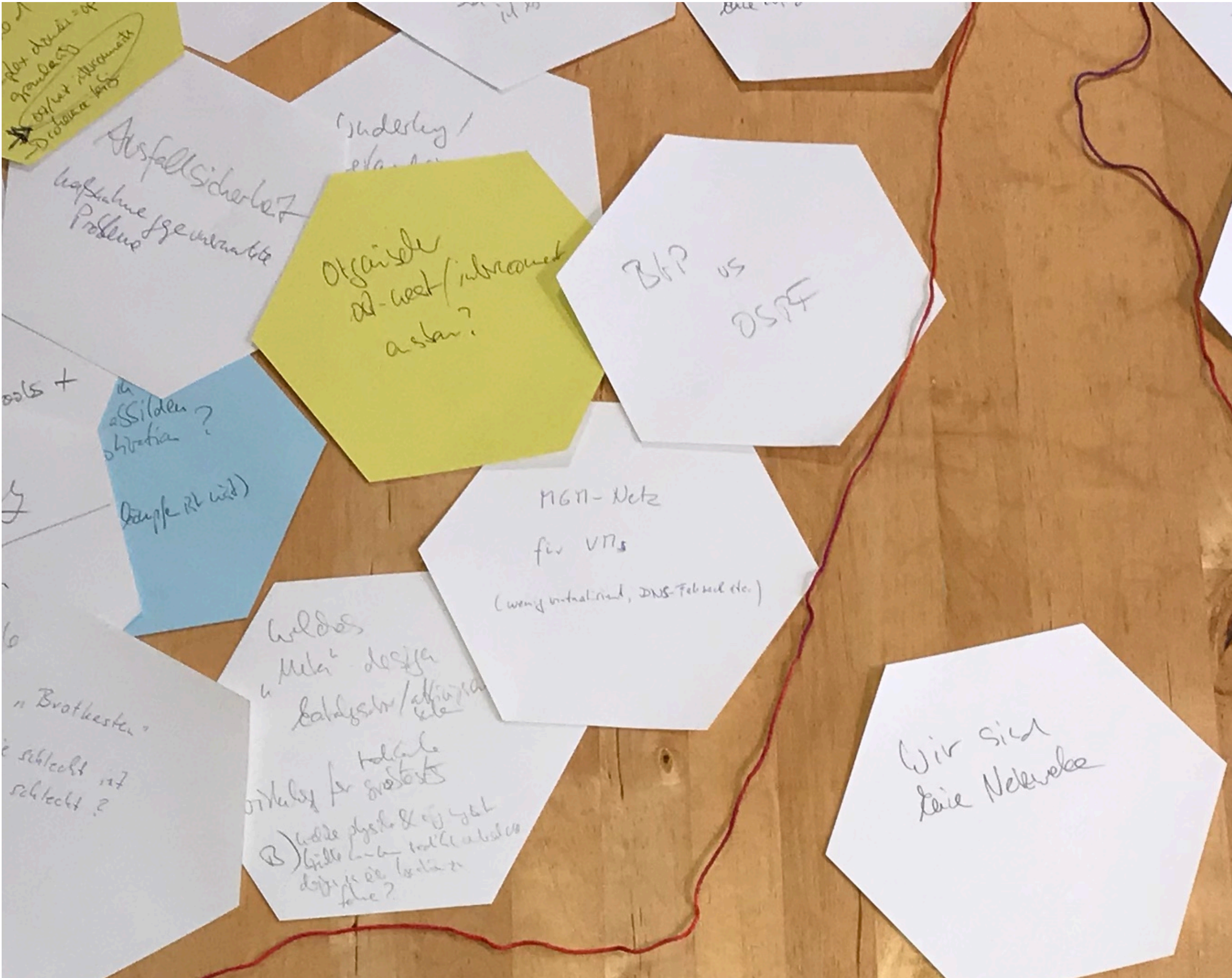


Von L2 auf EVPN-VXLAN in 4 Jahren

Migration eines PaaS-
Netzwerkes mit größtmöglichem
Open-Source-Einsatz

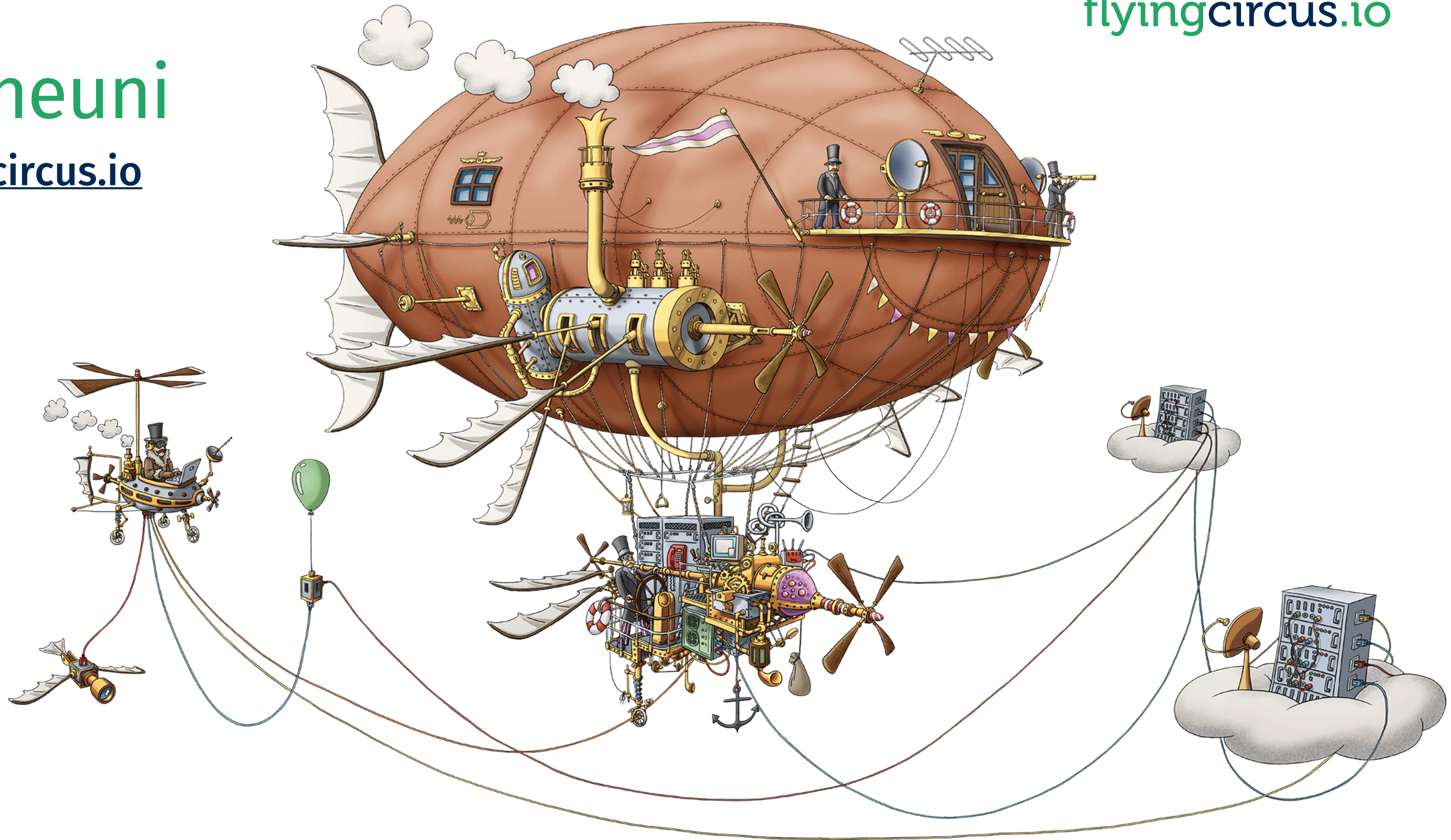
Christian Theune am 17. August 2024 auf der FrOSCon



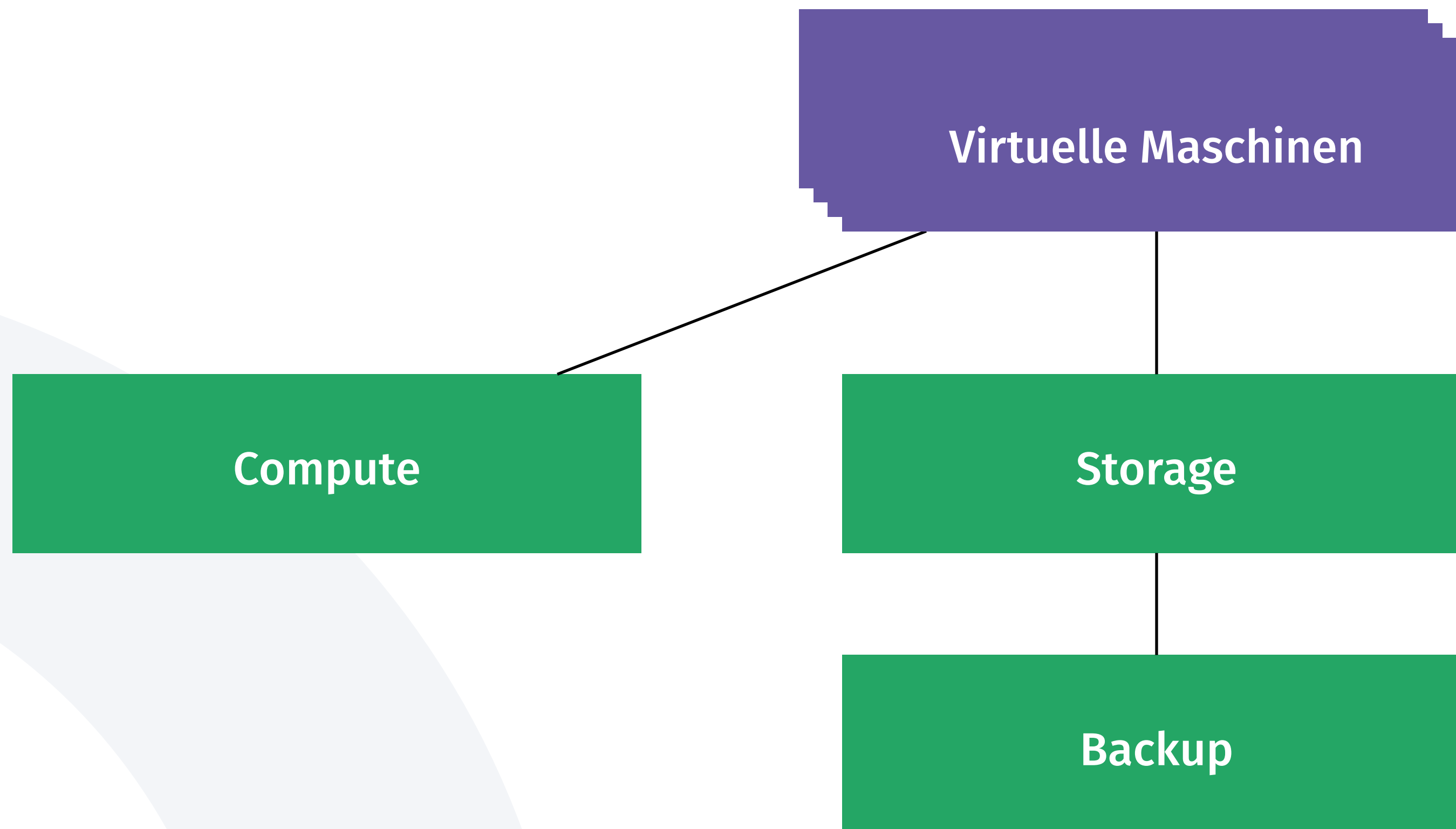


@theuni
flyingcircus.io

flyingcircus.io



Infrastruktur



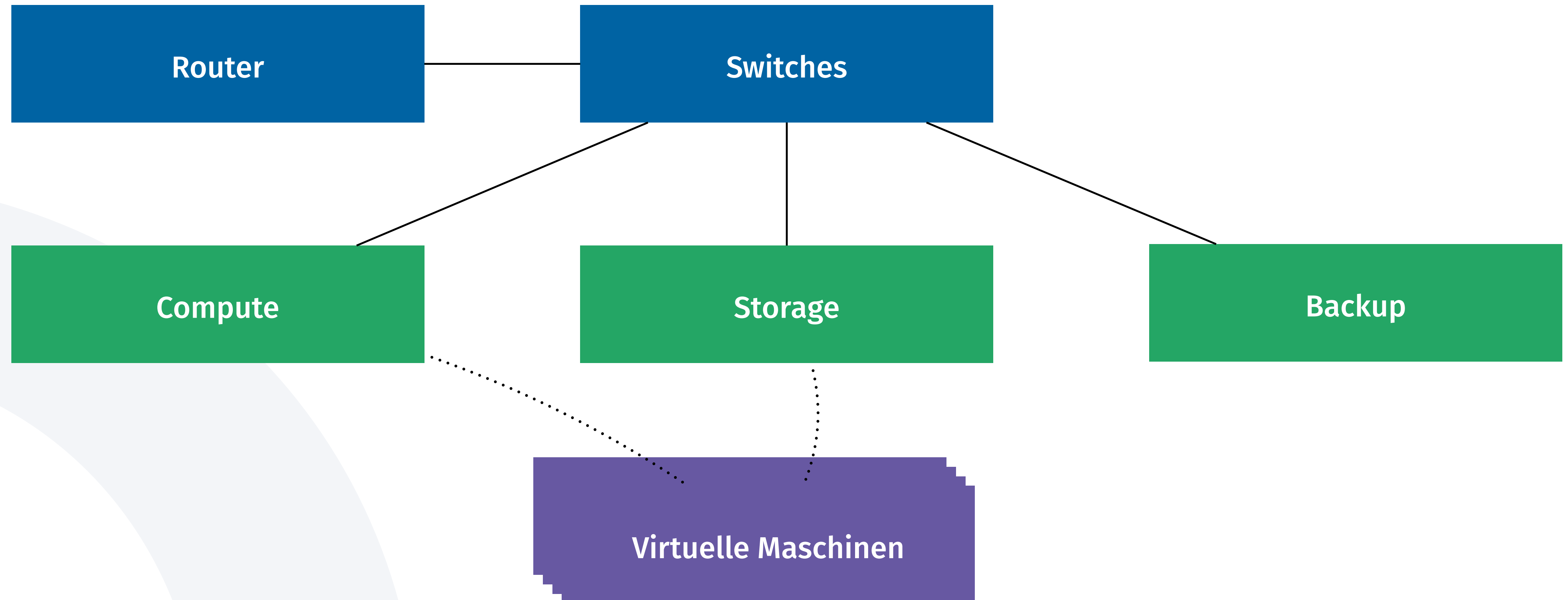
Da fehlt doch was?



> Erwartungen an das Netzwerk

- ▶ Es muss funktionieren.
RFC 1912, Punkt 1
 - ▶ also: Redundanz wäre schon gut
- ▶ Traffic segmentieren
- ▶ Nicht dauernd irgendwas konfigurieren müssen.
- ▶ Einmal Linux, bitte!

Infrastruktur - mit Netzwerk



Netzwerkhersteller die @theuni auf dem Kerbholz hat:

✦ HP ProCurve / ProVision

✦ Brocade

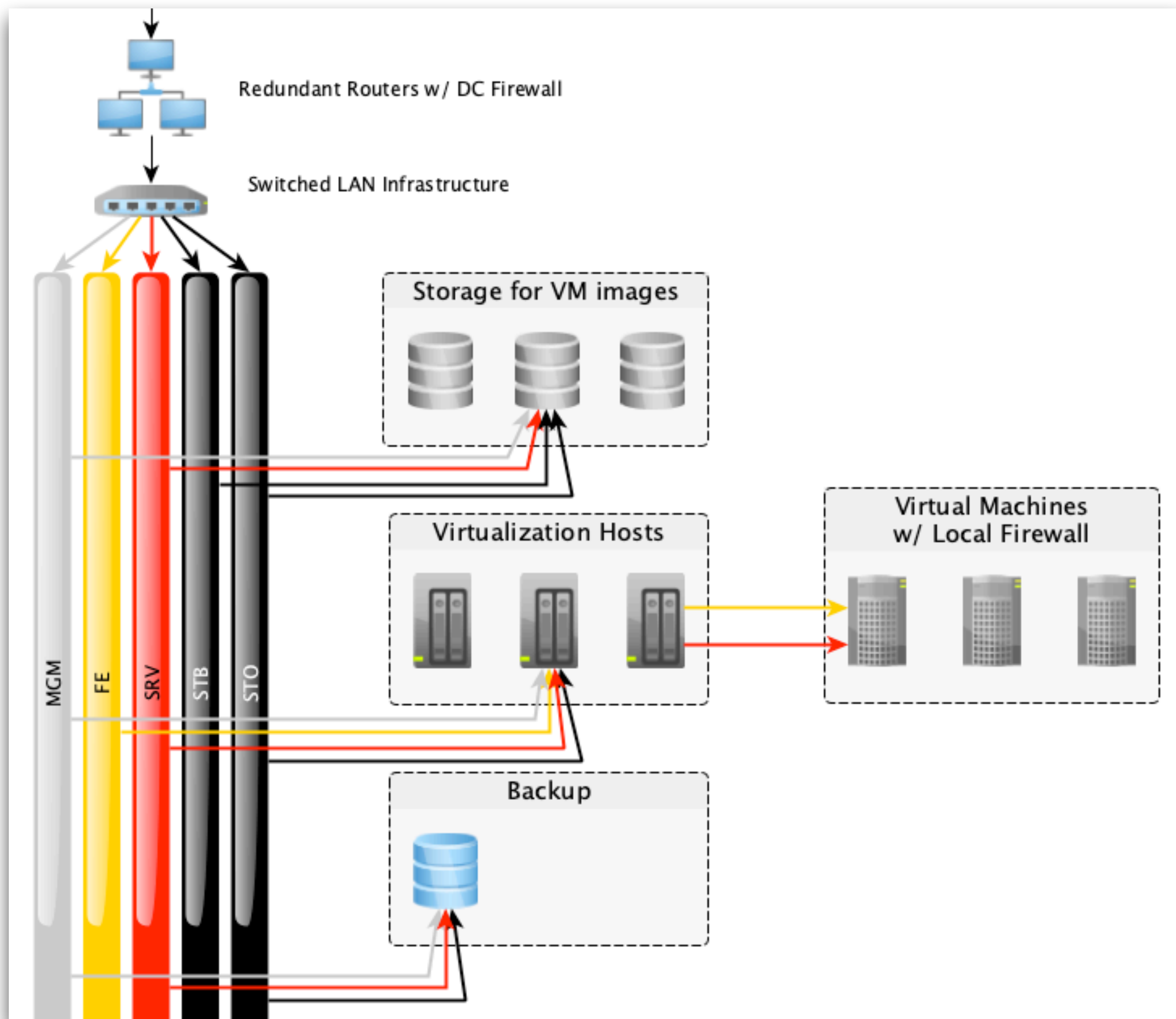
✦ Cumulus

✦ Pica 8

✦ Juniper

(Ich hab' doch nur geguckt!)





Wunschzettel

- ▶ Aktive Redundanz
- ▶ Weniger Ports
- ▶ Migration zu Glasfaser
- ▶ Alle Netze 10G+
- ▶ Herstellerunabhängig / Open Source
- ▶ Kosten im Griff behalten
- ▶ Stärkere Segmentierung



Roadmap anno 2019

- Migration zu neuer Hardware und vereinfachtem L2
- Schrittweise Kupfer durch Glas ersetzen
- BGP Underlay mit ECMP einführen
- EVPN/VXLAN Overlay einführen

The screenshot shows the flyingcircus.io dashboard with two main sections: 'Design-Entwürfe' and 'Projekte'.

Design-Entwürfe:

- Ausfallsicherheit, diverse Switches, Hersteller, Medien (2 cards)
- "Location Net" (1 card)
- Zentrales GW / Firewall zwischen RGs (1 card)
- GW-VM pro RG (1 card)
- GW-VM pro RG (Variante 2) (1 card)
- + Add another card

Projekte:

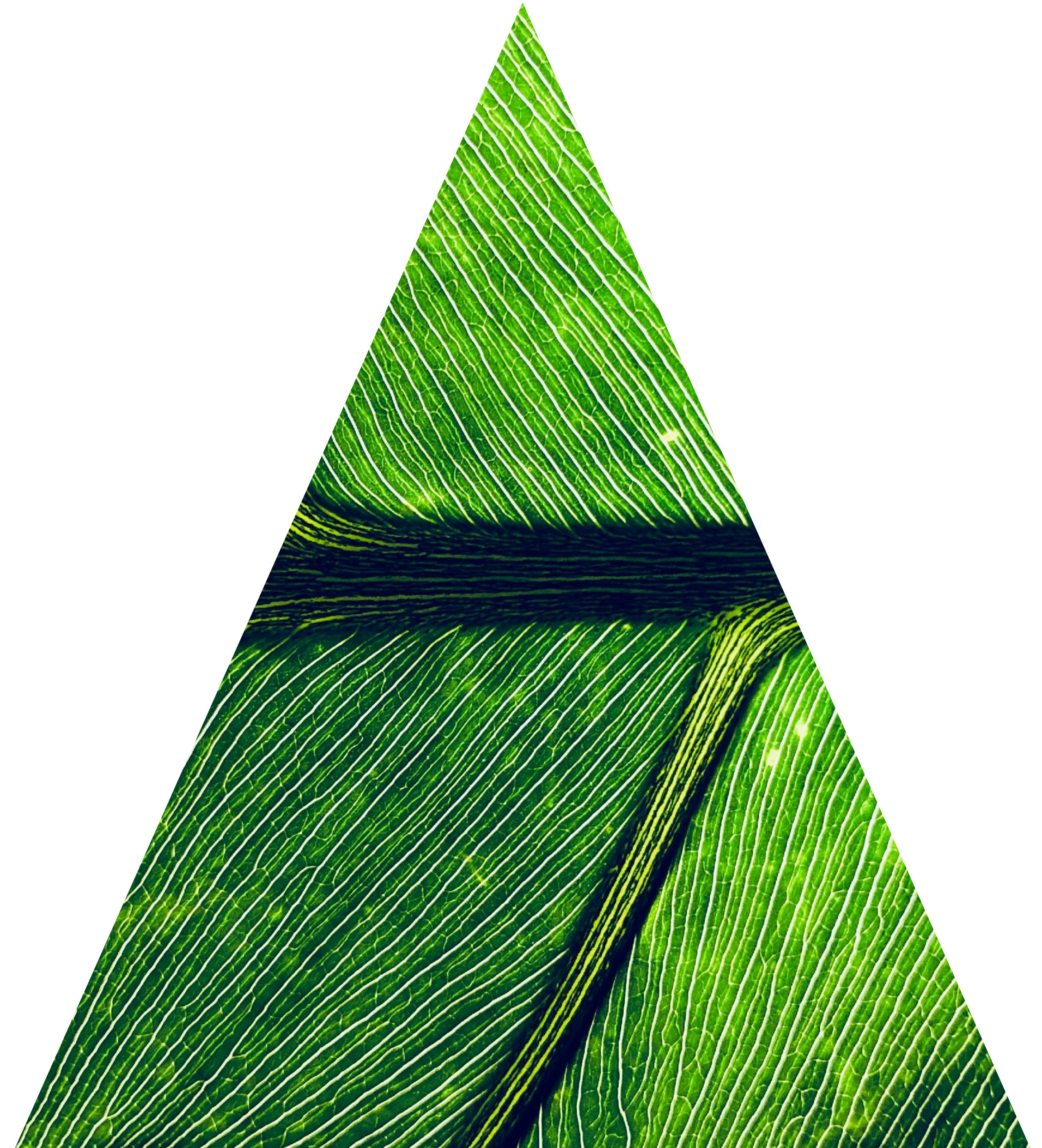
- Gateway-Konzept / Management-RG
- BGP - mit oder ohne EVPN
- Storage BGP - Separation STO/STB?
- NixOS-Update
- Neue Hardware - Anforderungen
- Gebrauchte Hardware & Hersteller mixen (1 card)
- SFP-Kompatibilität recherchieren (Uplinks)
- + Add another card





Details, Details, Details

- ▶ Routing On the Host
- ▶ BGP Unnumbered
- ▶ Config Management
- ▶ Performance
- ▶ Redundanz
- ▶ Hotspares und Labor
- ▶ Access Network
- ▶ Der finale Rollout



> Routing On the Host

Ich dachte das machen doch schon alle coolen Kids?

- ▶ “Das Netzwerk” ist nicht mehr durch “Netzwerkgeräte” definiert.
- ▶ Virtualisierung bedeutet: auch der Compute-Server ist ein Switch.
- ▶ Ein Overlay wie VXLAN bedeutet: alles ist Netzwerk.
- ▶ BGP auf dem Host ist wohl ein gelöstes Problem?

> BGP Unnumbered – RFC 5549

Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop

- ▶ Massiver Reduktion von Konfigurationsaufwand
- ▶ L3-EVPN-Setup fühlt sich operativ wieder wie ein “dummer” Switch an
- ▶ Weniger potentielle Fehler beim Verkabeln - lass die Remotehands einfach irgendwas stecken

> Config-Management: Switches

- ▶ User-Management
- ▶ Lizenz-Management
- ▶ Konfiguration von Interfaces und BGP
- ▶ Sensu- und Prometheus-Integration
- ▶ Hardware-Monitoring

```
[environment]
service_user = root
update_method = rsync-ext
branch = production

;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; VXLAN fabric switch A1
[host:stan33]
components = system, network

data-mgmt-address = 172.22.1.195/24

data-vxlan-underlay-address = 172.23.64.3/20
data-vxlan-underlay-ports = 1-54

data-port-count = 54
data-port-speeds = {
  "default": 10,
  "49-54": 40}

;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
; VXLAN fabric switch A2
[host:stan34]
components = system, network
.
data-mgmt-address = 172.22.1.196/24
```


> Config-Management: Switches

Commit Hash 66ef9da0153f6dd476b7a5f8f2c840a1b9f63149
Tree f0528e3b5d091b0694661b8e0739153d49eaf895
Author Molly Miller <mm@flyingcircus.io>
Date Wed, 15 May, 2024, 15:21
Parent 1f91f386c662a06f88a69ae31f0e5018eb2c5cab
Stats 1 file changed: -2 +2

rzob: more VXLAN ports

▼ environments/rzob/environment.cfg -2 +2

```
30 30
31 31 data-mgmt-address = 172.22.1.195/24
32 32
33 33 data-vlan-stb-ports = 1,3,5,7,17,19,21,39-42,47-54
33 33 data-vlan-stb-ports = 3,5,19,47-54
34 34
35 35 data-vxlan-underlay-address = 172.23.64.3/20
36 36 data-vxlan-underlay-ports = 2,4,6,8-16,18,20,22-38,43-46
36 36 data-vxlan-underlay-ports = 1,2,4,6,7,8-16,17,18,20,21,22-38,39-42,43-46
37 37 data-vxlan-bridge-vlans = stb
38 38
39 39 data-port-count = 54
```


> Config-Management: Switches

```
ctheune@chazzzzz ~/Code/flyingcircus/fc-switches-batou (master*) $ ./batou deploy --predict whq
batou/2.4.1 (cpython 3.7.11-final0, Darwin 23.6.0 arm64)
===== Preparing =====
main: Loading environment `whq`...
main: Verifying repository ...
You are using external rsync. This is a non-verifying repository -- continuing on your own risk!
main: Loading secrets ...
===== Connecting hosts and configuring model ... =====
services13: Connecting via ssh (1/3)
stan30: Connecting via ssh (2/3)
stan52: Connecting via ssh (3/3)
===== Predicting deployment actions =====
stan30: Scheduling component system ...
stan52: Scheduling component system ...
stan30 > System > Telegraf > Extract('telegraf-1.14.3_linux_amd64.tar.gz') > Untar('telegraf-1.14.3_linux_amd64.tar.gz') [total=1.67s, verify=1.67s, update=NaN, sub=0.00s]
stan30: Scheduling component network ...
stan52 > System > Telegraf > Extract('telegraf-1.14.3_linux_amd64.tar.gz') > Untar('telegraf-1.14.3_linux_amd64.tar.gz') [total=1.69s, verify=1.69s, update=NaN, sub=0.00s]
services13: Scheduling component sensuserver ...
stan52: Scheduling component network ...
stan30 > Network > Exec('ifreload -a') > File('/etc/network/interfaces') > Content('interfaces')
interfaces ---
interfaces +++
interfaces @@ -562,7 +562,7 @@
interfaces      mstpctl-portadmedge yes
interfaces      mstpctl-bpduguard yes
interfaces      link-autoneg off
interfaces - link-speed 1000
interfaces + link-speed 10000
interfaces      mtu 9216
stan30 > Network > Exec('ifreload -a')
services13 > SensuServer > File('/etc/local/sensu-client/switch-stan52.json') > Content('switch-stan52.json')
```


Config-Management: Hosts

- ▶ NixOS + eigene Config-Management DB als (öffentlich) geheime Zutat.
- ▶ Netzwerk-Abstraktionen in der CMDB umbauen
- ▶ Vorteil: Unit- und funktionale Tests und strukturierte DB-Migrationen

Network	
ipmi	ipmi
	BMC (ipmi)
2a02:238 <input type="text"/>	/64
172.20.1.182	/24
sto	vxlan
	02:00:00:04:13:83 (vtep)
2a02:238 <input type="text"/>	/64
172.20.4.111	/24
mgm	untagged
	onboard/left (10G)
2a02:238 <input type="text"/>	/64
172.20.1.188	/24

NICs		
fe	02:00:00:02:13:83	vtep
srv	02:00:00:03:13:83	vtep
sto	02:00:00:04:13:83	vtep
BMC	0c:c4:7a:dd:5e:39	ipmi
external/ lower	a0:36:9f:27:c9:06	10G
external/ upper	a0:36:9f:27:c9:04	10G

> Performance

- ▶ Linux auf den Servern will für “UDP in UDP” ein bisschen getuned werden, sonst harte Paketverluste wenn wir Richtung 10G wollen.
- ▶ NIC-Offloading (ethtool):
 - ▶ Interrupt Moderation: rx-usecs 1
 - ▶ Große Ring Buffers: -G rx 4096 tx 4096
 - ▶ Large Receive Offload: lro on
 - ▶ VXLAN offload: udp_tnl
- ▶ Intel:
 - ▶ "ixgbe.InterruptThrottleRate=1"
- ▶ Mellanox (Nvidia) ❤️ Intel 💥

Performance

- ▶ Connection Tracking (im Underlay) vermeiden
- ▶ Speicher/Buffer groß machen
- ▶ IRQ Balancing!

```
boot.kernel.sysctl = {  
  
    "vm.min_free_kbytes" = "513690";  
  
    "net.core.netdev_max_backlog" = "300000";  
    "net.core.optmem" = "40960";  
    "net.core.wmem_default" = "16777216";  
    "net.core.wmem_max" = "16777216";  
    "net.core.rmem_default" = "8388608";  
    "net.core.rmem_max" = "16777216";  
    "net.core.somaxconn" = "1024";  
  
    "net.ipv4.tcp_fin_timeout" = "10";  
    "net.ipv4.tcp_max_syn_backlog" = "30000";  
    "net.ipv4.tcp_slow_start_after_idle" = "0";  
    "net.ipv4.tcp_syncookies" = "0";  
    "net.ipv4.tcp_timestamps" = "0";  
  
                                # 1MiB   8MiB   # 16 MiB  
    "net.ipv4.tcp_wmem" = "1048576 8388608 16777216";  
    "net.ipv4.tcp_rmem" = "1048576 8388608 16777216";  
    "net.ipv4.tcp_mem" = "1048576 8388608 16777216";  
  
    "net.ipv4.tcp_tw_recycle" = "1";  
    "net.ipv4.tcp_tw_reuse" = "1";  
  
    # Supposedly this doesn't do much good anymore, but in one of my tests  
    # (too many, can't prove right now.) this appeared to have been helpful.  
    "net.ipv4.tcp_low_latency" = "1";  
  
    # Optimize multi-path for VXLAN (layer3 in layer3)  
    "net.ipv4.fib_multipath_hash_policy" = "2";  
};  
  
services irqbalance.enable = true;
```


> Redundanz

- ▶ BGP + BFD mit ECMP
- ▶ Fail-Over bei unerwartetem “Kabel weg” und “Switch weg” < 2s
- ▶ Geplantes Umstellen geht nahezu “hitless”
- ▶ Trifft auch immer nur ~50% des Traffics

```
ctheune@kyle20 ~ $ ip r
default via 172.20.3.1 dev brsrv proto static metric 60
default via 172.20.1.1 dev ethmgm proto static metric 90
10.0.0.0/24 dev brsrv proto static scope link
10.102.99.0/24 via 172.30.3.110 dev brsrv onlink
172.16.1.0/24 dev enp3s0f1 proto kernel scope link src 172.16.1.100 linkdown
172.20.1.0/24 dev ethmgm proto kernel scope link src 172.20.1.188
172.20.2.0/25 dev brfe proto static scope link
172.20.3.0/24 dev brsrv proto kernel scope link src 172.20.3.165
172.20.4.0/24 dev brsto proto kernel scope link src 172.20.4.111
unreachable 172.21.64.0/22 metric 335544321
172.21.64.2 nhid 593 via inet6 fe80::1eea:bff:fe6a:39e9 dev ul-extern-upper proto
tric 20
172.21.64.3 nhid 591 via inet6 fe80::669d:99ff:fe3a:f5f9 dev ul-extern-lower proto
etric 20
172.21.64.5 nhid 592 proto bgp src 172.21.64.20 metric 20
    nexthop via inet6 fe80::669d:99ff:fe3a:f5f9 dev ul-extern-lower weight 1
    nexthop via inet6 fe80::1eea:bff:fe6a:39e9 dev ul-extern-upper weight 1
172.21.64.6 nhid 592 proto bgp src 172.21.64.20 metric 20
    nexthop via inet6 fe80::669d:99ff:fe3a:f5f9 dev ul-extern-lower weight 1
    nexthop via inet6 fe80::1eea:bff:fe6a:39e9 dev ul-extern-upper weight 1
172.21.64.7 nhid 592 proto bgp src 172.21.64.20 metric 20
    nexthop via inet6 fe80::669d:99ff:fe3a:f5f9 dev ul-extern-lower weight 1
    nexthop via inet6 fe80::1eea:bff:fe6a:39e9 dev ul-extern-upper weight 1
172.21.64.8 nhid 592 proto bgp src 172.21.64.20 metric 20
    nexthop via inet6 fe80::669d:99ff:fe3a:f5f9 dev ul-extern-lower weight 1
    nexthop via inet6 fe80::1eea:bff:fe6a:39e9 dev ul-extern-upper weight 1
```


> Labor und Hot-Spares

- ▶ “Infrastructure as Code” → also bitte auch mehrstufiges Dev/QA/Staging/...
- ▶ Für max. 96 Server in Produktion braucht es 5 Switches (2x48+2x48+1x48)
- ▶ Im Labor: 4x48
- ▶ Im Backup-RZ: 2x48
- ▶ Kostenpunkt klassische Hersteller: $11 * 25.000 = 275.000 \text{ €}$ - und ich muss einmal alles neu Anschaffen
- ▶ Kostenpunkt Whitebox mit Mischung Cumulus + Sonic: 55.000 € (bzw. 15.000 € weil bestehendes Equipment)

> Access- und Management-Netzwerke

- ▶ Tolles Netzwerk - aber was machen wir jetzt mit den ganzen dummen Geräten, die kein BGP können?
- ▶ Das sind geplant und zum Glück extrem wenige
 - ▶ RIPE Atlas Probe
 - ▶ ein paar USVs und PDUs
 - ▶ Hardware-Monitoring
 - ▶ IPMI-Controller
- ▶ Zwei billige, separates Layer-2-Netzwerke, die hinter der Firewall stecken



> Der Wochenplan

One, few, many ...

- ▶ Montag:
 - ▶ ankommen, 3 Paletten Zeug sichten, Backup-Server umstellen
- ▶ Dienstag:
 - ▶ ~10 Server umstellen, 1. Router austauschen
- ▶ Mittwoch:
 - ▶ ~20 Server umstellen, 2. Router austauschen, Rückbau und aufräumen
- ▶ Donnerstag:
 - ▶ ~30 Server umstellen, mehr Rückbau und aufräumen
- ▶ Freitag:
 - ▶ Nacharbeiten, aufräumen, Abreise

> One-Shot-Config-Replacement auf Linux-Servern

- ▶ Reboots aller Maschinen unbedingt vermeiden - kostet zu viel Zeit
- ▶ Größtes Problem:
 - ▶ alte Configs im laufenden Betrieb sauber abbauen
 - ▶ interface renames
- ▶ SystemD + udev sind keine Freunde von konvergenter Config
- ▶ NixOS scripted networking
 - ▶ leider kein allgemeines Interesse mehr dran
 - ▶ in unserem Fork recht einfach strukturiert mit harten Calls zu “ip link set name” und “geh mir aus der Sonne”

> Ceph Wartungsmodus

- ▶ Ceph ist sehr latenz-sensibel, unnötiges Stoppen/Starten von OSDs vermeiden. Aber Netzwerk darf auch nicht einfach wackeln...
- ▶ Wir haben automatisierte Maintenance-Mechaniken.
- ▶ Bei Maintenance alle OSDs auf einem Host:
 - ▶ “ceph osd set-group noup 1,2,3,4,...”
 - ▶ “ceph osd down 1,2,3,4,...”
- ▶ Und danach:
 - ▶ “ceph osd unset-group noup 1,2,3,4,...”

> Die NASA-Checkliste

	kyle50	cartman30	barbrady07	ingress00	kyle49	barbrady06	cartman36	kyle48	cartman35	kyle46	cartman34	cartman33	cartman32	cartman31		kyle28	barbrady04	kyle27	cartman23	kyle26	cartman17	kyle25	cartman16	kyle33	cartman15	kyle34		cartman21	cartman20	kyle23	kyle45	cartman19	
Start time	13.05.24 19:38	13.05.24 20:24	13.05.24 18:40	14.05.24 14:25	14.05.24 10:49	14.05.24 10:16	14.05.24 10:38	14.05.24 11:49	14.05.24 11:07	14.05.24 13:50	14.05.24 12:17	14.05.24 12:25	14.05.24 12:38	14.05.24 12:46		15.05.24 09:12	15.05.24 09:28	15.05.24 09:56	15.05.24 10:04	15.05.24 10:14	15.05.24 10:25	15.05.24 10:35	15.05.24 10:45	15.05.24 10:53	15.05.24 11:02	15.05.24 11:11		15.05.24 11:40	15.05.24 11:56	15.05.24 12:23	15.05.24 12:35	15.05.24 12:45	
KVM: evacuate (next host)	✓				✓			✓		✓						✓		✓		✓		✓		✓		✓					✓	✓	
Old port ethrv	stan39-29	stan41-41		stan39-27	stan39-6	stan41-39	stan39-48	stan39-2	stan41-46	stan41-13	stan39-44	stan41-44	stan41-3	stan41-1		stan39-28	stan41-4	stan41-35	stan41-17	stan39-33	stan41-7	stan39-35	stan39-12	stan39-16	stan41-5	stan39-11		stan39-46	stan39-42	stan41-36	stan39-17	stan39-40	
Old port ethfe	???	???			???	-		???		???																							
Old port stb	stan32-21	stan32-48	stan32-1	stan32-25	stan32-16	stan32-35	stan32-8	stan32-5	stan32-10	stan32-15	stan32-14	stan32-24	stan32-44	stan32-46		stan32-27	stan32-4	stan32-32	stan32-2	stan32-26	stan32-17	stan32-39	stan32-20	stan32-11	stan32-12	stan32-6		stan32-40	stan32-47	stan32-29	stan32-22	stan32-45	
Old port stb	-	stan33-48			-	-	stan33-47	-	stan33-45	-	stan33-43	stan33-38	stan33-44	stan33-46					stan33-7		stan33-21		stan33-1		stan33-13			stan33-41	stan33-39				stan33-40
New Switch ports:	4	6	2 + 2	28	12	8	10	16	14	26	18	20	22	24		34	32	23	25														
Prepare: gather cables and modules, prewire PB + new ports	✓	✓	✓	■	✓	✓	✓	■	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	■		✓	✓	✓	✓	■	
Ceph: clean?		✓					✓		✓		✓	✓	✓	✓					✓		✓		✓		✓			✓	✓			✓	
Ceph: set noout		✓					✓		✓		✓	✓	✓	✓					✓		✓		✓		✓			✓	✓			✓	
Directory: set out of service + nix config snippet	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
fc-manage -ebv	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Router: turn on stop file																																	
Directory: Convert network config / clean up policies	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
KVM: wait for clear	✓				✓			✓		✓						✓		✓		✓		✓		✓							✓	✓	
Cleared for maintenance																																	
Theuni: Rewire / Set modules / Remove old cables	✓	✓	✓	■	✓	✓	✓	■	✓	✓	✓	✓	✓	■		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	■		✓	✓	✓	✓	✓	
Molly: Switch: reconfigure	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Molly: Disable interfaces except mgm (stop network-addresses-srv/fe)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Molly: Apply config: fc-manage -ebv	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Validate config	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Directory: set in service, remove config snippet	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
fc-manage -ebv	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
KVM: fc-maintenance run	✓				✓			✓		✓						✓		✓		✓		✓		✓						✓	✓		
Back in service																																	
KVM: evaluate non-prod VM	✓				✓			✓		✓						✓		✓		✓		✓		✓		✓					✓	✓	
Ceph: unset noout		✓					✓		✓		✓	✓	✓	✓					✓		✓		✓		✓			✓	✓			✓	
Router: turn off stopfile																																	
Router: force failover to validate config																																	
End time	13.05.24 20:18	13.05.24 20:38	13.05.24 19:31	14.05.24 14:42	14.05.24 11:04	14.05.24 10:30	14.05.24 10:48	14.05.24 12:13	14.05.24 11:17	14.05.24 14:02	14.05.24 12:24	14.05.24 12:33	14.05.24 12:45	14.05.24 12:53		15.05.24 09:28	15.05.24 09:46	15.05.24 10:04	15.05.24 10:13	15.05.24 10:24	15.05.24 10:33	15.05.24 10:43	15.05.24 10:52	15.05.24 11:00	15.05.24 11:10	15.05.24 11:20		15.05.24 11:54	15.05.24 12:22	15.05.24 12:33	15.05.24 12:43	15.05.24 12:51	
Duration	40m 14s	13m 52s	50m 39s	17m 7s	14m 52s	14m 10s	10m 26s	23m 54s	10m 9s	12m 9s	6m 51s	8m 1s	7m 47s	7m 31s	0	15m 45s	17m 30s	7m 53s	8m 9s	10m 1s	8m 36s	8m 17s	7m 36s	7m 40s	8m 53s	8m 48s	0	14m 14s	25m 57s	10m 8s	8m 50s	6m 26s	
	evacuate a kvm host while performing a migration of a ceph host; stuck sfp module		bad cable (wrong polarity on patchbox cable), confused routing and tracing the cables → trace from switch side not server side	took a little longer due to updating interface labels				technical difficulties getting the facetime call set up																									

> Die NASA-Checkliste

“kyle43 evacuated, ready for critical path.” – “kyle43, ack. links going down” - “Ack. Links are down on kyle43.”

- ▶ 25 Schritte pro Server: Evacuation, Config, Rewiring, Test, Reintegration
- ▶ Zu zweit mit Headsets: einer macht Physik, einer macht Config
- ▶ Zeiten:
 - ▶ Minimum: 6 Minuten
 - ▶ Durchschnitt: 13 Minuten
 - ▶ Maximum: 50 Minuten
 - ▶ Gesamt: 544 Minuten (9h) für 43 Server



> 100%? Fast.

- ▶ Slow-Fail auf Glasfaser
 - ▶ Sieht aus wie MTU-Problem
 - ▶ Hat aber jmd. ins Kabel gebissen
 - ▶ Link Up → ECMP verteilt fröhlich weiter
 - ▶ trifft nur große Probleme
- ▶ Fehlkonfigurierter Switchport
 - ▶ egal, schnell umstecken
- ▶ Weirdes FRR Flooding-Problem

> Frr - das Sorgenkind

- ▶ Frr und komisches Flooding
- ▶ Interaktion mit keepalived ist fishy
- ▶ Netlink-Messages die verloren gehen ...
- ▶ Traffic-Sniffing vs. Netlink
- ▶ Müssen wir in der Community noch besser ankommen

> Ist noch Zeit für zwei kleine Goodies?

```
8512 /nix/store/...-bird-2.0.10/bin/bird -c /etc/bird/bird2.conf -u bird2 -g bird2
692862 /nix/store/...-frr-8.5.5/libexec/frr/bgpd -f /etc/frr/bgpd.conf -A localhost -p 0
4135619 /nix/store/...-frr-8.5.5/libexec/frr/zebra -f /etc/frr/zebra.conf -A localhost
4135620 /nix/store/...-frr-8.5.5/libexec/frr/bfdd -f /etc/frr/bfd.conf -A localhost
```

```
3: ethtr: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc
    link/ether ac:1f:6b:1e:3e:29 brd ff:ff:ff:ff:ff:ff
    altnam enp5s0
    altnam onboard-right
    altnam connected-switch-stan17-port-45
4: ul-extern-upper: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu
    link/ether 0c:42:a1:0b:c2:ae brd ff:ff:ff:ff:ff:ff
    altnam external-upper
    altnam connected-switch-stan31-port-swp39
5: ul-extern-lower: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu
    link/ether 0c:42:a1:0b:c2:af brd ff:ff:ff:ff:ff:ff
    altnam external-lower
    altnam connected-switch-stan51-port-swp39
```


> Und jetzt?

- ▶ Happy: Redundanz, weniger Geräte, weniger Kabel, mehr Performance
- ▶ Etwas weniger “keine Netzwerker”
- ▶ Open Source heißt für uns immer wieder: eher Make und Cooperate als nur Buy – viel gelernt, viel selbst gestaltet - digitale Souveränität!
- ▶ Open-Networking: Cumulus ist un-tot, es lebe SONiC? – Community ist mir da noch unklar
- ▶ Hardware-Weiterverwendung durch Multi-Vendor-Mischbetrieb
- ▶ Netzwerk-Segmentierung für Kunden jetzt evolutionär ein “reines Software-Problem”