deutsches forschungsnetz

Bericht aus dem NOC

81. DFN-Betriebstagung | 8.10.2024

Thomas Schmid, Robert Stoy

Agenda

I. Core-Router Ersatz

II. IPv6

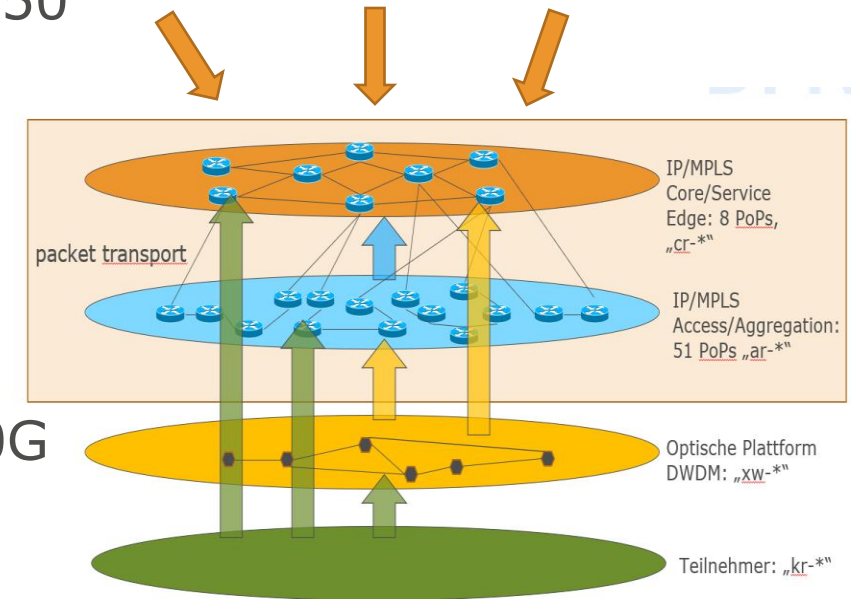
III. Migration 800G SuperCore : Eine Monitoring Perspektive

IV. Update Monitoring inkl. Performance / Alarmierung

V. Ende-zu-Ende Performance - einige Hinweise

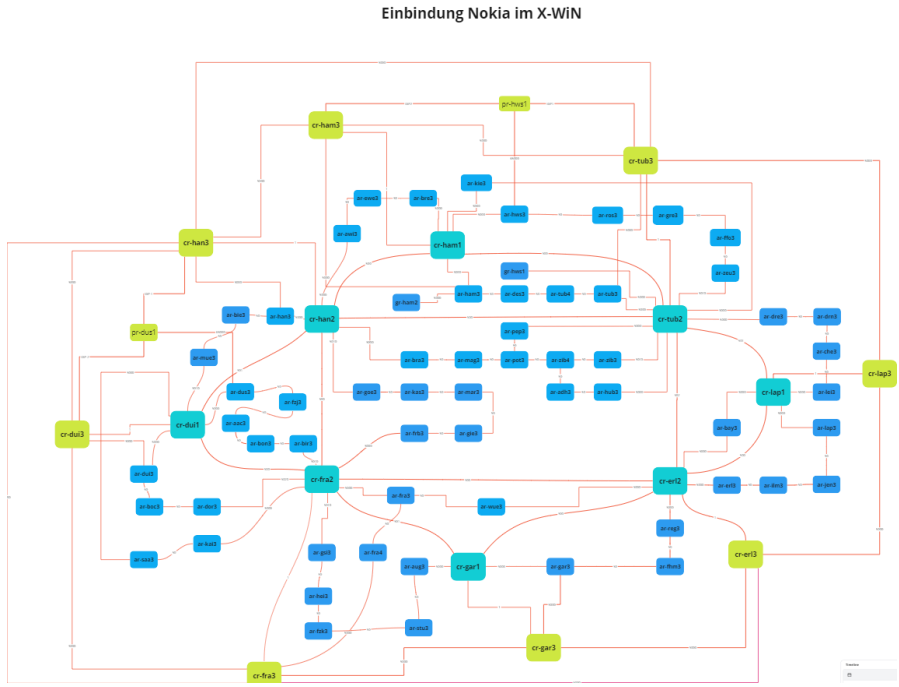
Core-Router: was, warum, wann?

- Was?
  - Ersatz der Cisco ASR 9010 und 9912 durch Nokia 7750 SR-7s und SR-2s
    - Neu: 2 Peering-Router an den Standorten HWS (Hamburg Global Connect) und DUS (Digital Realty Düsseldorf)
- Warum?
  - Upgrade der Infrastruktur von 100G auf 400G + 800G unter Beibehaltung aktuell eingesetzter Features
- Wann?
  - Plan, die Migration bis Ende des Jahres abgeschlossen zu haben



Was?: was bisher geschah …

- Alle Router-Einbauten abgeschlossen
  - 1 DoA: pr-hws3
- Geräte sind im Monitoring, Alarmierung, DDoS-Mitigation integriert
- Zwischentopologie in Betrieb
  - 2 x 400G Backbone fertig
    - Aber noch nicht für Produktionsverkehr genutzt
  - Verbindung Cisco-Nokia aktuell mit 100G – 300G
  - Finale Topologie mehr oder weniger identisch zur alten Topologie



400/800G: erste Erfahrungen

- /ˈbliː.dɪŋ ɛd͡ʒ/
  - Interoperabilitätsproblem: Unterbrechung im DWDM-Pfad triggert einen RF-Alarm zum Client und somit ein LOS auf dem Nokia 400G-Transceiver. Problem: kein clear des LOS, wenn RF-Alarm weg ist.
    - R&D von Ribbon und Nokia im regen Austausch
  - Workaround nach Firmwareupgrade auf DWDM-Karten inzwischen implementiert
    - DWDM-Platform schickt LOL (Laser aus) an Nokia -> Transceiver funktioniert wieder
  - Kein showstopper für die Inbetriebnahme
- Erster echter 800G-Kanal auf der DWDM-Platform FRA-ERL in Betrieb
  - Nutzung zum Transport von 2 x 400G Client-Signalen

Wie? (I)

- Umzug der „einfachen“ Fälle zuerst
  - Standard IP-Anschlüsse über PWHE (aka „spoke-sdp“ bei Nokia). Beginn seit September
    - Ca. 1050 Anschlüsse, aktuell ca. 300 bereits umgezogen
  - Umzug nur in der Konfiguration ohne Umstecken
  - i.d.R. innerhalb von 10 min erledigt

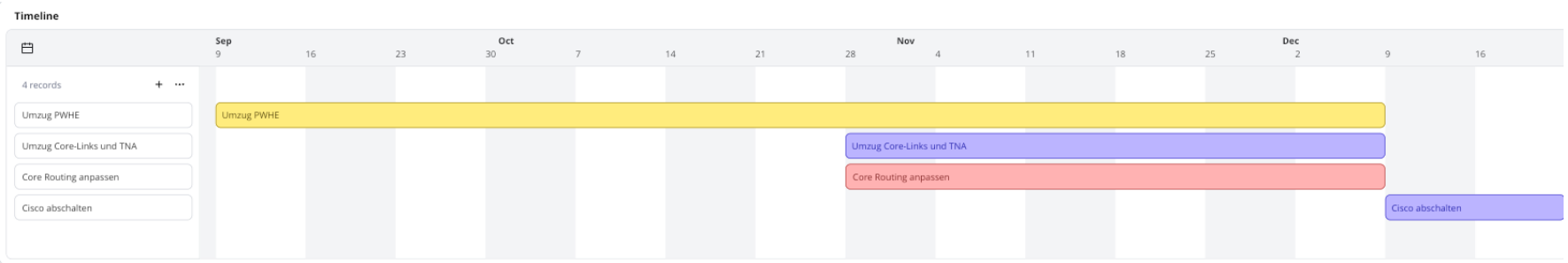
Umzug


Umzug in Datenbank
BGP shutdown
Konfigurations erzeugung
Aufspielen der Konfig
Testen und aufräumen
ONE DOES NOT SIMPLY
WASH, RINSE, REPEAT

Wie? (II)

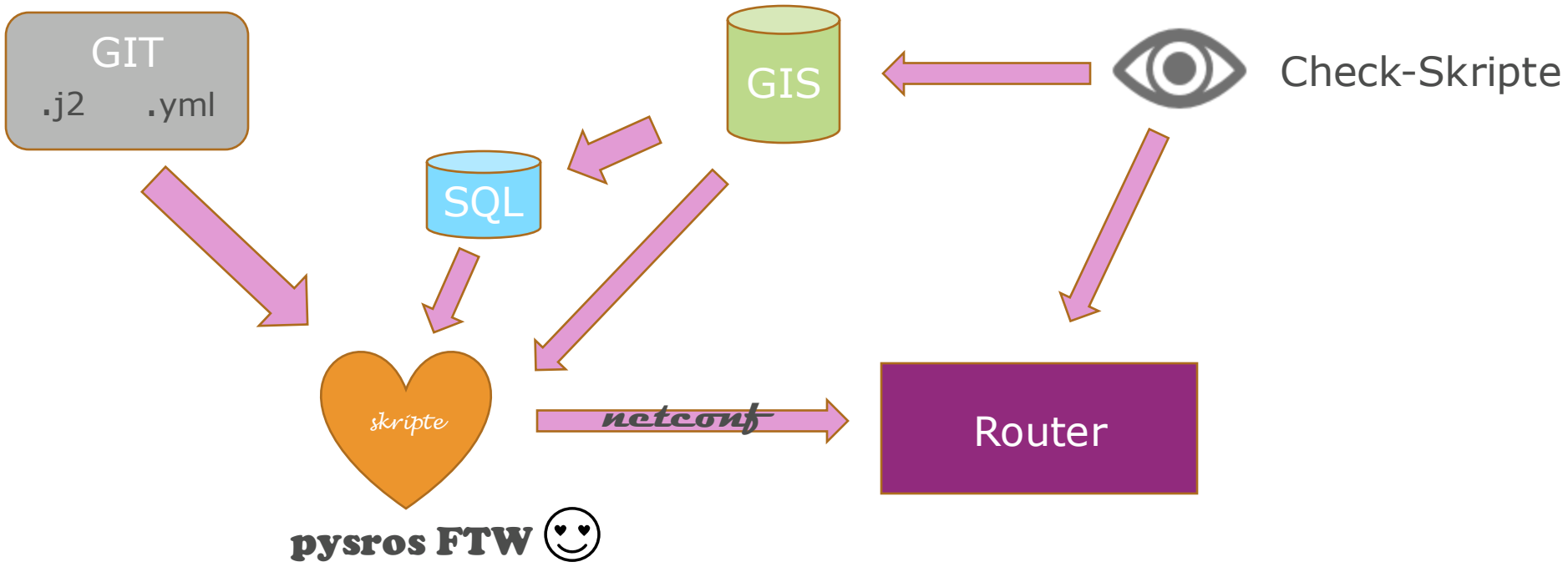
- Nächste Schritte
  - „Schwierige Fälle“ umziehen (ab Ende Oktober)
  - Standorteinsätze
    - Verbindungen AR-CR auf Nokia umziehen
    - Direkt am Cisco angeschlossene Teilnehmer umstecken
    - Ca. 260 Anschlüsse
  - Routing auf 400G-Infrastruktur legen
    - 2 parallele Backbones -> Bottleneck vermeiden!
  - Peering-Router in Betrieb nehmen
  - Cisco abschalten (Dezember)

Ende Oktober - Dezember



Wie?: Konfigurationserzeugung

- GIS (globales Informationssystem): unser OSS/BSS/Inventory Management/Datenbank/single-source-of-truth für alles
- GIT: Jinja2 Templates, YML und Konfiguration



Nokia-Konfigs

- Nokia-Konfiguration relativ komplex
  - Ein einfacher PWHE Regelanschluss auf Nokia erzeugt z.B. 700 Zeilen Konfiguration vs. ca. 150 Zeilen auf Cisco
- Anschlussvarianten entsprechend der Entgeltordnung
  - Regelanschluss
  - Versorgeranschluss
  - Clusteranschluss
    - VpD, (VfA: Auslaufmodell, nur noch Bestrandsschutz)
  - VPNs
    - Layer3 oder Layer 2 P2P oder MP2MP

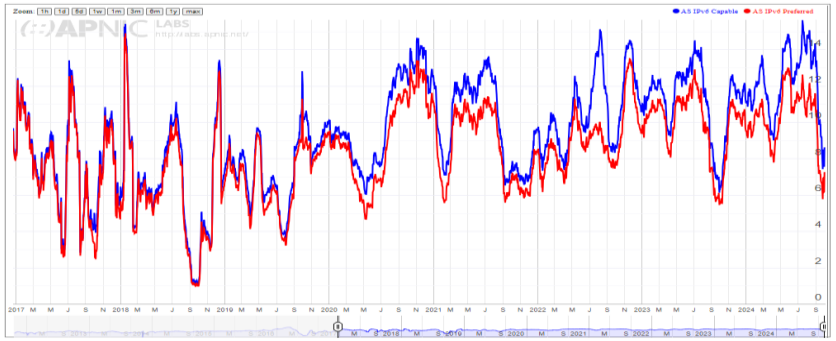
IPv6

IPv6 Wasserstand

- 33% der BGP-Sessions zu Teilnehmern mit IPv6 ☹
- Verkehrsanteil IPv6:

cr-fra2

Google



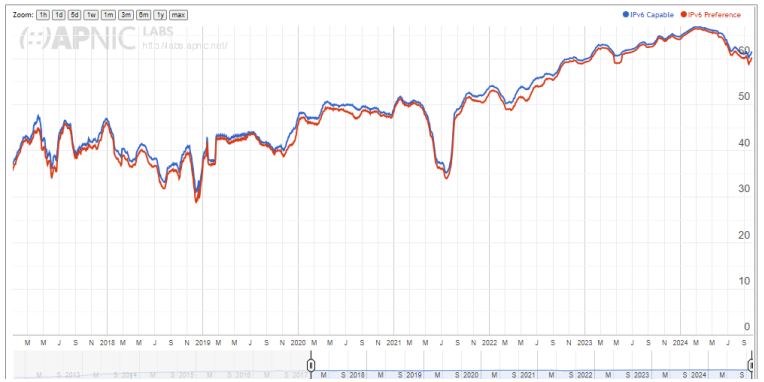
Verkehrsanteil IPv6

https://stats.labs.apnic.net/ipv6/

<- DFN: max 15%

(Zahlen seit 2018)

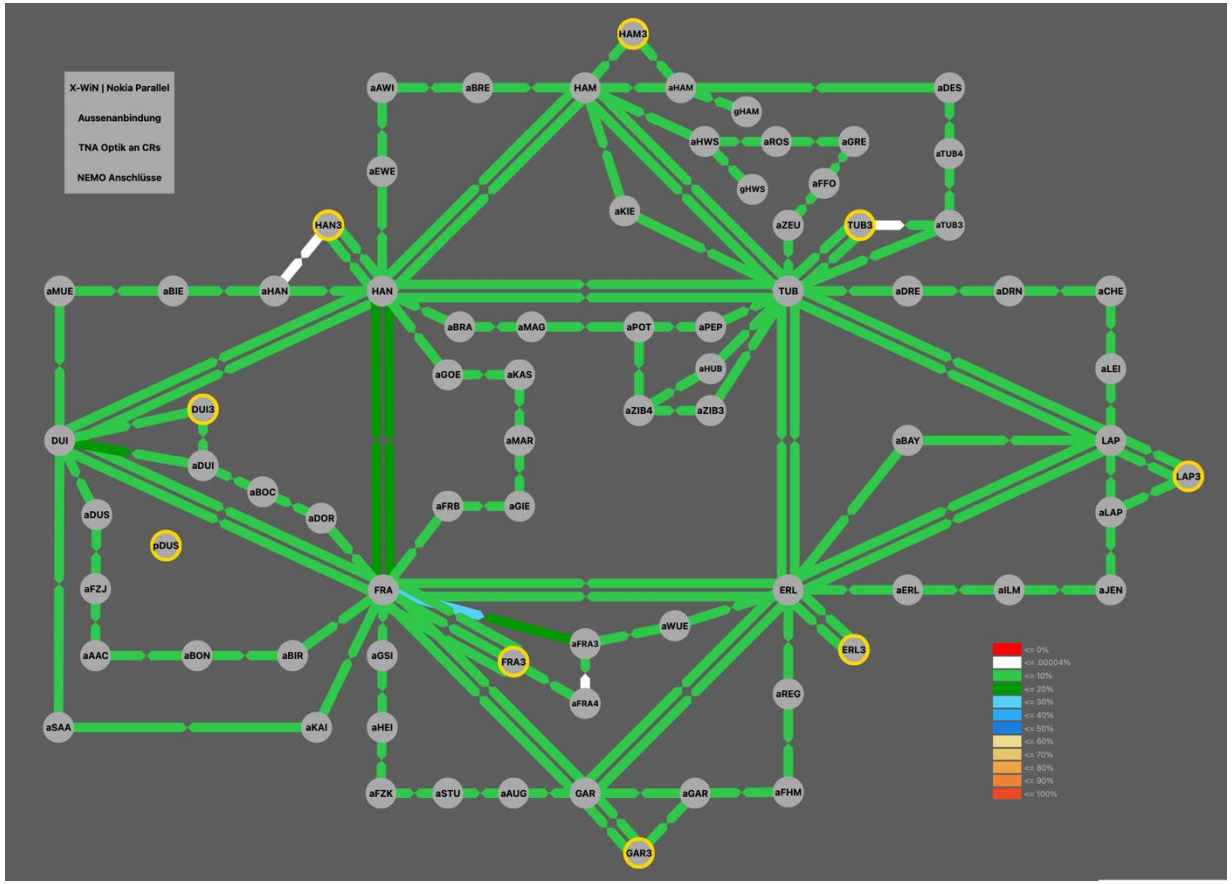
Deutschland: max 70 % ->



Migration zum 800G SuperCore

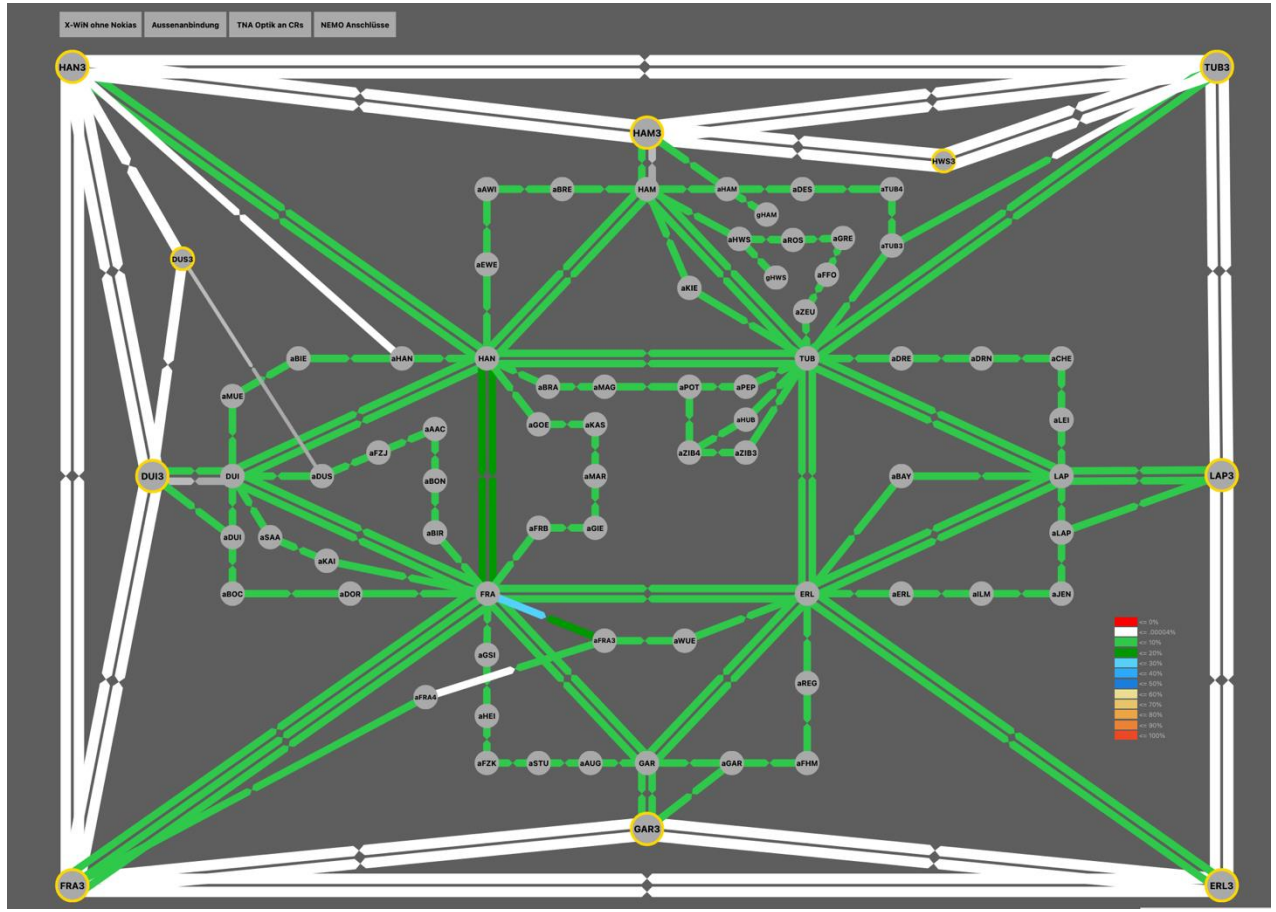
Die Monitoring Perspektive

X-WiN Kernnetzmigration Phase 1+2

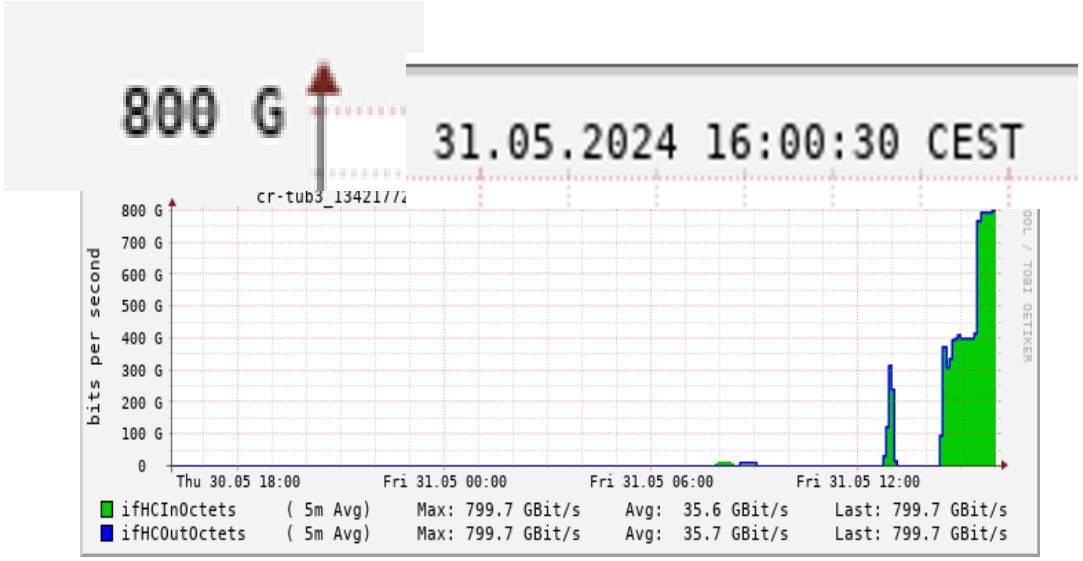


- DWDM Netz Erweiterung bis 07/2024
- Einbau Nokia Router 05-08/2024

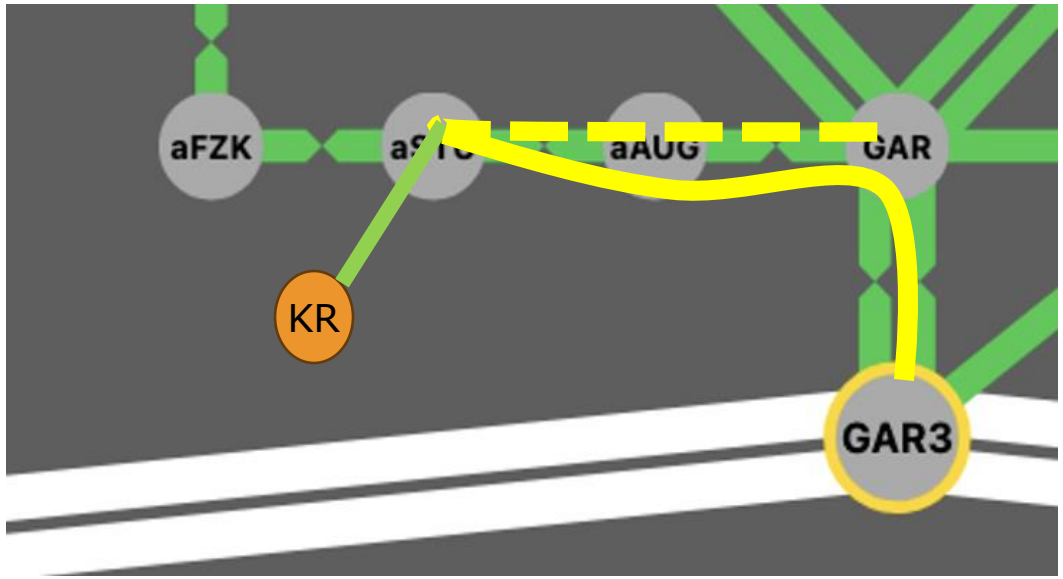
X-WiN Kernnetzmigration : Phase 3



- Tests und Herstellung Betriebsbereitschaft 2*400 G fertig 09/2024



X-WiN Kernnetzmigration : Phase 4
Umzüge Pseudowires



- Umzug von schmalbandigen (<10G) Teilnehmeranschlüssen 1081 Stück
  - Beginn in 09/2024 , ca ¼ erledigt
  - Reiner config-change der MPLS Kanäle, gestützt auf automatische Generierung der Routerkonfiguration
  - Routing über neue Core Router
  - Verifikation

X-WiN Kernnetzmigration
kommende Phase(n)

- Umzug physikalischer Anschlüsse an alten Core Routern
  (Techniker Einsatz am Kernnetz Knoten)
  - breitbandige (10G,100G) Teilnehmeranschlüsse von alten auf neue CRs
  - Umzug Aggregationsketten
- Sukzessive Umstellung des Routings von alten 200G auf neue 800G Links
  Änderung der Wegekosten nach ausgearbeitetem Plan
- Abschaltung alte Core Router

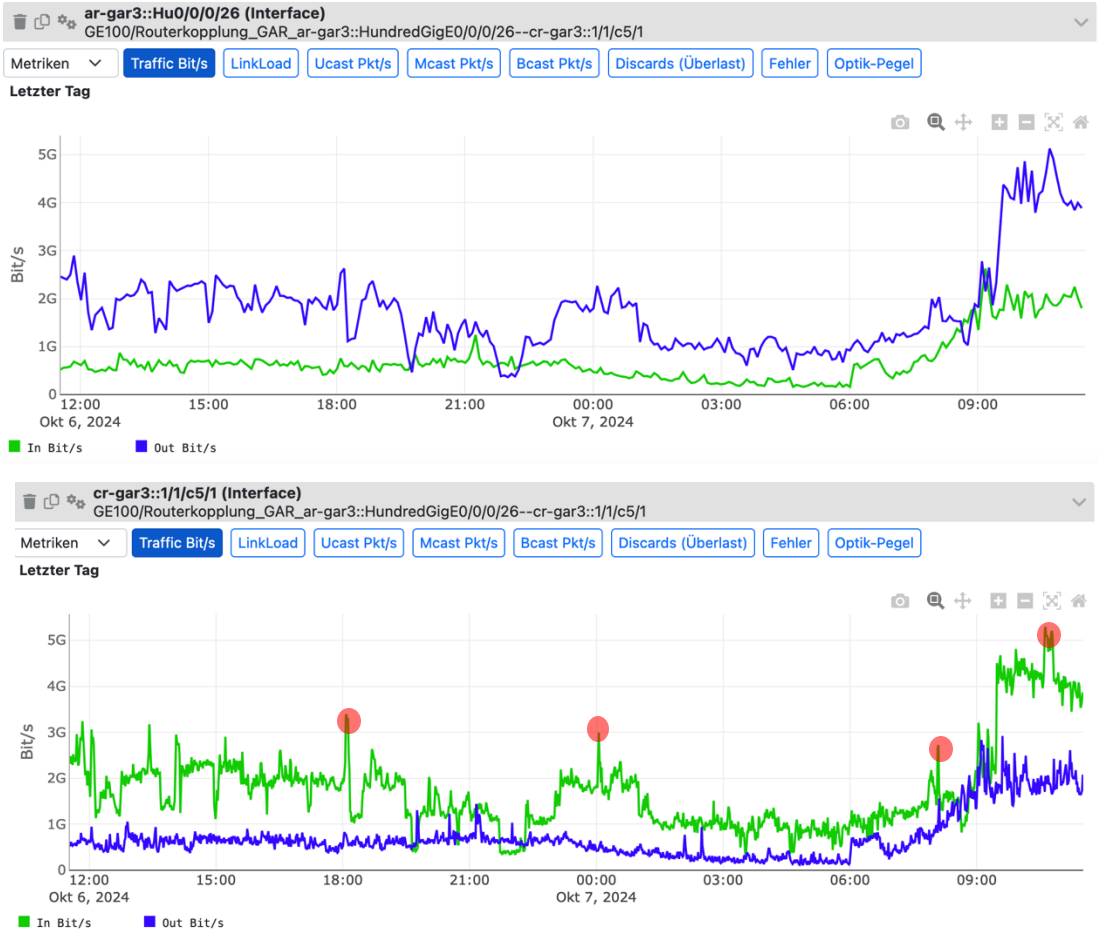
Monitoring : Neue SW Technologie

Streaming Telemetry in DMon

Router Monitoring
Weiterentwicklung : Streaming Telemetry (I)

- Software Technologie als Ersatz für SNMP
  für höherfrequentes, zeitnahes und weniger Router belastendes Monitoring
- IETF standardisierte oder Hersteller spezifischen YANG Datenmodelle als Grundlage
  -> (Sensor-Pfade)
- NMS abonniert einen Sensor-Pfad im Router
  mit gewünschten update Intervallen oder für einmalige Information
- Router sendet permanenten Strom von Messdaten zum Kollektor in NMS
- Tests zeigen mögliche minimale update Intervalle bei Zählern auf Routerports
  im bis 3s und auf logischen Interfaces bis 10s
  Einsetzbar temporär in Laborumgebungen oder während Troubleshooting

Router Monitoring (DMon)
Weiterentwicklung : Streaming Telemetry(II)

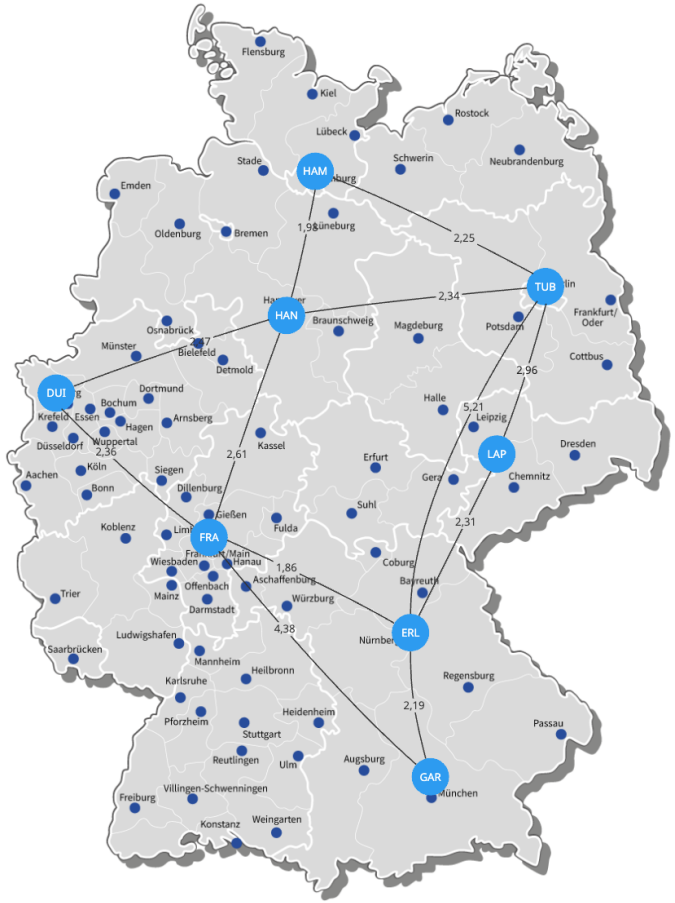


- Geringere Belastung der Router Control Plane
- Erhöhung der zeitlichen Messpunkte um Faktor 5
  (Intervall bisher 5min , neu: 1min)
- -> genauere Erfassung von Spitzenwerten (Höhe und Dauer)
  Im Beispiel bis zu ca. 30% höher
- -> schnellere Alarmierung
  Verzögerung max. 1 Minute

Performance Monitoring und Überwachung

Aktive Messungen

Überwachung der Netzperformance (I)



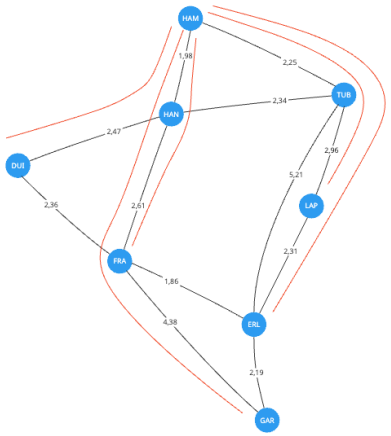
- Setup
  - Spezielle Messrechner in X-WiN Supercore Knoten , GPS zeitsynchronisiert
  - perfSONAR Software
  - Vollvermaschte unidirektionalen Messpfade 8*7=56 Stück
  - Permanenter Messstrom auf jedem Messpfad
    10 Pkt/s , 15 kbit/s
  - Wegeführung der Messpfade entsprechend allgemeinem X-WiN Routing

- Aufgezeichnete Messergebnisse
  - Kurzzeit Unterbrechungen im 100ms Bereich. (Verbesserung auf 50ms im Plan)
  - Einweglaufzeiten mit Messgenauigkeit +/- 0,05 ms

- Alarmierung bei
  - Kurzzeitunterbrechung eines Messpfades
  - Abweichungen vom Soll-Routing
  - Einweglaufzeit Schwankungen

Überwachung der Netzperformance (II)

Routen von HAM



Hop	Router	IP	Delay	MTU
1	cr-ham1-te0-2-0-3-9-200.x-win.dfn.de	188.1.222.9	0.5ms	
2	cr-han2-be3.x-win.dfn.de	188.1.144.38	4.4ms	
3	cr-dui1-be17.x-win.dfn.de	188.1.144.189	9.2ms	
4	owd-sc-ms-dui.x-win.dfn.de	188.1.222.30	8.6ms	

Hop	Router	IP	Delay	MTU
1	cr-ham1-te0-2-0-3-9-200.x-win.dfn.de	188.1.222.9	0.8ms	
2	cr-tub2-be13.x-win.dfn.de	188.1.144.58	4.9ms	
3	cr-erl2-be7.x-win.dfn.de	188.1.146.209	15.3ms	
4	owd-sc-ms-erl.x-win.dfn.de	188.1.222.18	14.8ms	

Hop	Router	IP	Delay	MTU
1	cr-ham1-te0-2-0-3-9-200.x-win.dfn.de	188.1.222.9	0.7ms	
2	cr-han2-be3.x-win.dfn.de	188.1.144.38	4.5ms	
3	cr-fra2-be12.x-win.dfn.de	188.1.144.133	9.5ms	
4	owd-sc-ms-fra.x-win.dfn.de	188.1.222.2	9.1ms	

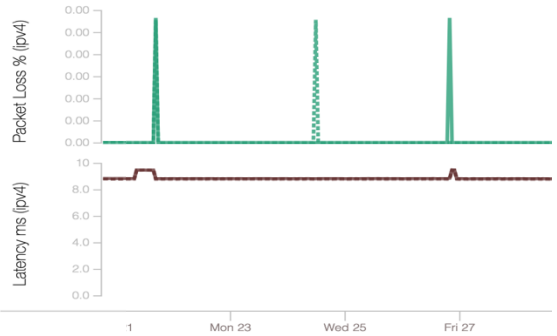
Hop	Router	IP	Delay	MTU
1	cr-ham1-te0-2-0-3-9-200.x-win.dfn.de	188.1.222.9	0.7ms	
2	cr-han2-be3.x-win.dfn.de	188.1.144.38	4.4ms	
3	cr-fra2-be12.x-win.dfn.de	188.1.144.133	9.4ms	
4	cr-gar1-be6.x-win.dfn.de	188.1.145.230	20.2ms	
5	owd-sc-ms-gar.x-win.dfn.de	188.1.222.26	17.7ms	

Hop	Router	IP	Delay	MTU
1	cr-ham1-te0-2-0-3-9-200.x-win.dfn.de	188.1.222.9	0.6ms	
2	cr-tub2-be13.x-win.dfn.de	188.1.144.58	5ms	
3	cr-lap1-be8.x-win.dfn.de	188.1.144.25	10.9ms	
4	owd-sc-ms-lap.x-win.dfn.de	188.1.222.34	10.3ms	

Mesergebnis vom Knoten HAM visualisiert

```
Status 06.10.2024 17:30
Lese letzte OWD Messwerte fuer jede Messtrecke aus perfSONAR
Measurement Archive...
```

/	DUI	ERL	FRA	GAR	HAM	HAN	LAP	TUB
DUI	--	4.12	2.36	6.66	4.30	2.39	6.35	4.69
ERL	4.12	--	1.82	2.20	7.37	4.41	2.31	5.21
FRA	2.35	1.86	--	4.37	4.51	2.61	4.12	4.89
GAR	6.62	2.19	4.38	--	8.80	6.90	4.43	7.31
HAM	4.32	7.34	4.55	8.85	--	2.01	5.13	2.24
HAN	2.42	4.40	2.62	6.92	1.98	--	5.23	2.34
LAP	6.36	2.32	4.10	4.43	5.13	5.25	--	2.99
TUB	4.68	5.20	4.88	7.30	2.25	2.34	2.96	--



- Routing folgt kürzesten Wegen
- Im Sollzustand des Netzes
- Längster Weg
  HAM <-> GAR:
  Einweglaufzeit: 8,9ms
- Kürzester Weg
  FRA <-> ERL:
  Einweglaufzeit 1,9ms

- Graphen über der Zeit
  - Kurzzeitunterbrechungen
  - Laufzeiten
  - Alarm Trigger

Ende-zu-Ende Performance Messungen

Einige Hinweise

Laufzeiten, Bandbreite, erreichbarer Ende-zu-Ende Durchsatz

- Der wesentliche Unterschied zwischen WAN (X-WiN und darüber hinaus) und LAN sind die Paketlaufzeiten.
  Im X-WiN: Ende-zu-Ende: im Bereich ca. 3 bis 30 ms

- TCP Durchsatz wird begrenzt durch Kombination von:
  - Paketlaufzeiten : Durch Pfade vorgegeben.
  - Paketverluste : Im X-WiN Kernnetz = 0
  - Pfad-MTU: für > 10Gbit/s empfohlen: Erhöhung auf 9000
  - TCP congestion window (cwnd) : Obergrenze wird definiert im Endsystem, evlt. Tuning nötig

- Unerwartet geringe TCP Durchsätze, typische Ursachen:
  - Paketverluste im Pfad, meist im LAN oder durch überlastete Zugangsleitungen
  - cwnd Obergrenzen Einstellung im Endsystem, zu gering für WAN
  - Performance Engpässe in Endsystemen durch virtualisierte Betriebsysteme
    Ältere Hardware unterhalb heutiger „Mittel-Klasse Servern“

Iperf3 Messungen zur Problemeingrenzung(I)

- Messrechner an X-WiN Supercore Knoten , aktuell angeschlossen mit 2 * 10 Gbit/s (upgrade auf 2*100 Gbit/s geplant)
- Einsatz von perfSONAR (mit iperf3) innerhalb X-WiN
- Freischaltung iperf3 Server bei Bedarf für Ende-zu-Ende Messungen für Teilnehmer
- Aktuelle iperf3 Version in Linux Distributionen: v3.12 (siehe >**`$ iperf3 -v`**)
- perfSONAR Zugang für Teilnehmer zunächst in kleinem Pilotprojekt angedacht.

Iperf3 Messungen zur Problemeingrenzung (II)

```
@ms:~$ iperf3 -c bw10g-od-ms-xxx -i 1
Connecting to host bw10g-od-ms-fra, port 5201
[  5] local 188.1.x.x port 55184 connected to 188.1.x.x port 5201
[ ID] Interval           Transfer     Bitrate         Retr  Cwnd
[  5]   0.00-1.00   sec  1.11 GBytes  9.50 Gbits/sec    0   31.8 MBytes
[  5]   1.00-2.00   sec  1.15 GBytes  9.88 Gbits/sec    0   31.8 MBytes
[  5]   2.00-3.00   sec  1.15 GBytes  9.89 Gbits/sec    0   31.8 MBytes
[  5]   3.00-4.00   sec  1.15 GBytes  9.90 Gbits/sec    0   31.8 MBytes
[  5]   4.00-5.00   sec  1.14 GBytes  9.75 Gbits/sec    0   31.8 MBytes
[  5]   5.00-6.00   sec  1.15 GBytes  9.89 Gbits/sec    0   31.8 MBytes
[  5]   6.00-7.00   sec  1.15 GBytes  9.90 Gbits/sec    0   31.8 MBytes
[  5]   7.00-8.00   sec  1.15 GBytes  9.88 Gbits/sec    0   31.8 MBytes
[  5]   8.00-9.00   sec  1.15 GBytes  9.90 Gbits/sec    0   31.8 MBytes
[  5]   9.00-10.00  sec  1.15 GBytes  9.89 Gbits/sec    0   31.8 MBytes
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bitrate         Retr
[  5]   0.00-10.00  sec  11.5 GBytes  9.84 Gbits/sec    0
sender
[  5]   0.00-10.01  sec  11.4 GBytes  9.77 Gbits/sec
receiver
```

Neben der Bitrate wichtige Messwerte

- **Retr : TCP Retransmissions**
  - Jede Retransmission ist die Folge von **Paketverlusten**, Senderate wird zurückgesetzt
- **Cwnd : Congestion Window**
  - Aktuelle im Sender gepufferte Datenmenge bis zum Eintreffen der Empfangsbestätigung vom Empfänger
  - Wird bei Paketverlusten zurückgesetzt.
  - Obergrenze wird in Endsystemen eingestellt.

Retransmissions = 0, -> keine Paketverluste ✓
Cwnd stabil bei 32 Mbyte (keine Limit im Endsystem) ✓
Erwarteter TCP Durchsatz ca. 9.8 Gbit/s erreicht ✓

Danke für Ihre Aufmerksamkeit

- DFN-NOC Team
  - Nils Beyrle
  - Peter Heiligers
  - Valentin Kirchner
  - Maximilian Müller
  - Thomas Schmid
  - Thilo Scholpp
  - Aljoscha Schulte
  - Frank Schröder
  - Robert Stoy
  - Hubert Waibel

E-Mail: noc@dfn.de
Telefon +49 71163314-112