

Fragestunde zur doppelten Anbindung

52. DFN-Betriebstagung Oktober
2009

Thomas Schmid
schmid@dfn.de

- Papier zur doppelten Anbindung
 - enthält eine Beispielkonfiguration eine art ‚Kleinsten gemeinsamer Nenner‘
 - daneben gibt es noch die 749 anderen Möglichkeiten zum Ziel zu kommen
 - welche der 749 anderen Möglichkeiten man wählt, hängt davon ab, wie das eigene Netz aussieht und was man erreichen will
 - 289 Möglichkeiten besprechen wir jetzt 😊

- BGP ist distance-vector Protokoll
 - Routinginformation ist Kombination aus Netz und Netzausgang
 - enthält keine Topologie-Information (vgl. link-state, z.B. OSPF Database)
 - Netzausgang wird über IGP geroutet, sonst terminiert Routingalgorithmus nicht
 - es wird immer so lange rekursiv in Routingprotokollen weitergesucht, bis der next-hop ein lokales Routerinterface ist
- Insbesondere niemals die Adresse des BGP neighbors über BGP annoncieren! -> Henne-Ei-Problem

BGP best path algorithm (Cisco)

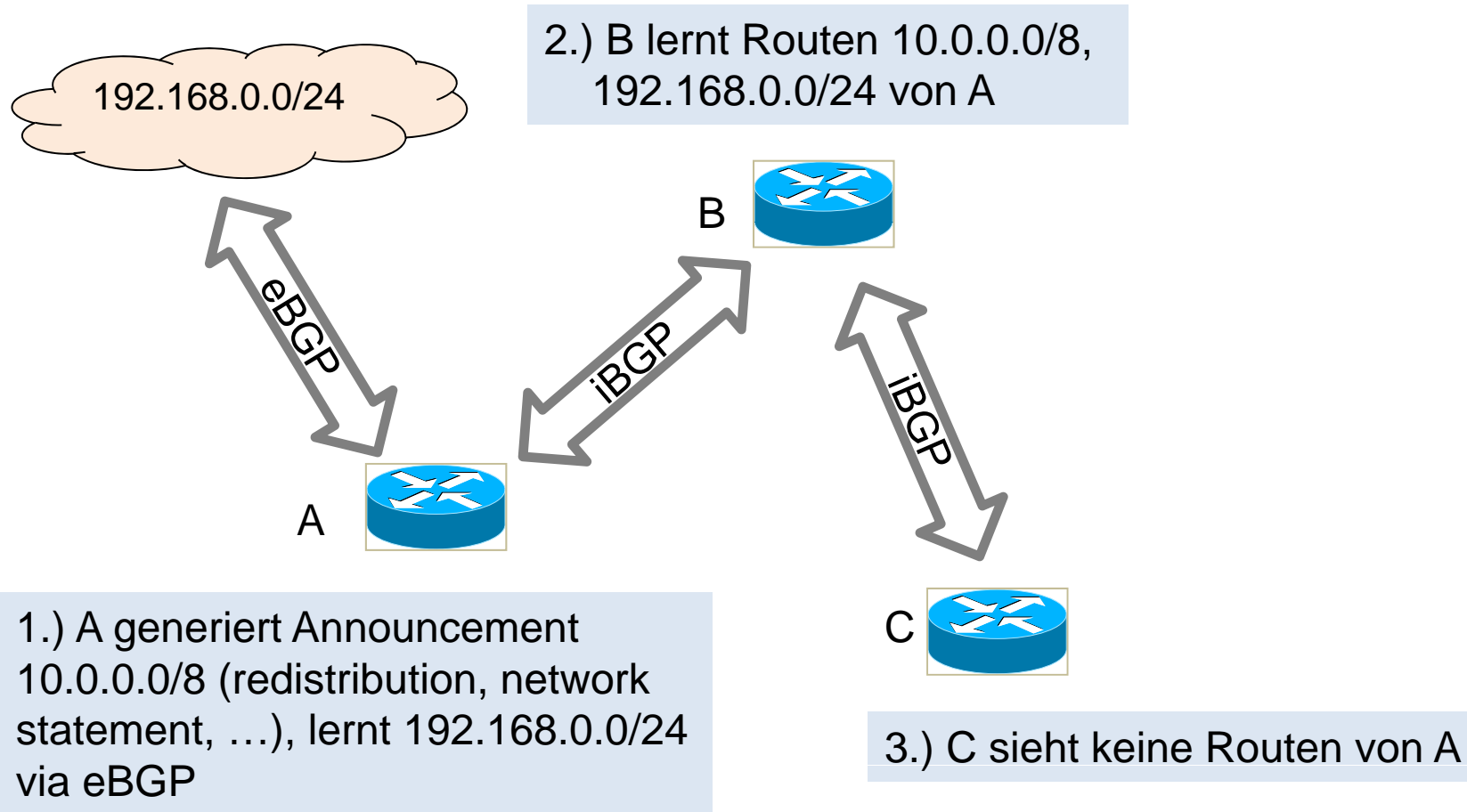
1. highest BGP weight (nur lokal auf dem Router)
2. highest LOCAL_PREF (Teil d. BGP Announcements)
3. locally originated
4. shortest AS_PATH
5. lowest origin type
6. lowest MED
7. prefer eBGP over iBGP
 - wichtig bei 2 KRs und eigener Routingpolicy: wenn nur ein Ausgang verwendet werden soll, MED setzen!
8. lowest IGP metric of next hop
9. multipath?
10. ...
11. prefer route from Router with lowest router ID
12. ...
13. prefer path from the lowest neighbor address

- Einsatz von mehreren Routingprotokollen:
 - ‚longest match‘ schlägt alles
 - /30 **immer** (Ausnahme DVMRP) besser als /28 ...
 - ‚administrative distance‘ als nächstes

Route Source	default distance value
connected	0
static	1
EIGRP summary	5
eBGP	20
EIGRP internal	90
OSPF	110
IS-IS	115
EIGRP external	170
iBGP	200

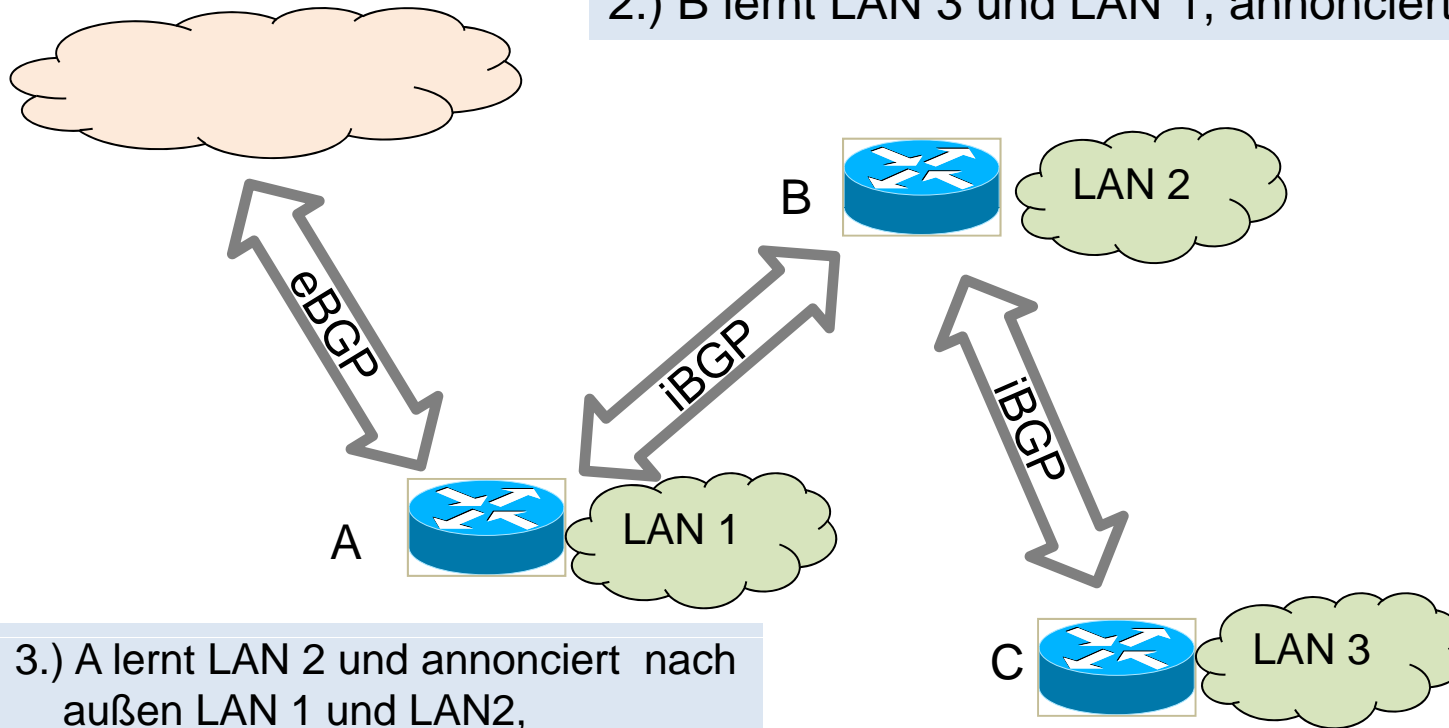
- BGP kann parallel aufgesetzt werden
- Routing im X-WiN folgt der statischen Konfiguration
 - kürzerer AS-Pfad
 - es sei denn, es werden Subnetze über BGP annonciert!
- Überprüfung der Announcements im laufenden Betrieb möglich
- löschen der statischen Routen führt dann zu keinem Ausfall

wer erzählt wem was?



umgekehrt genauso

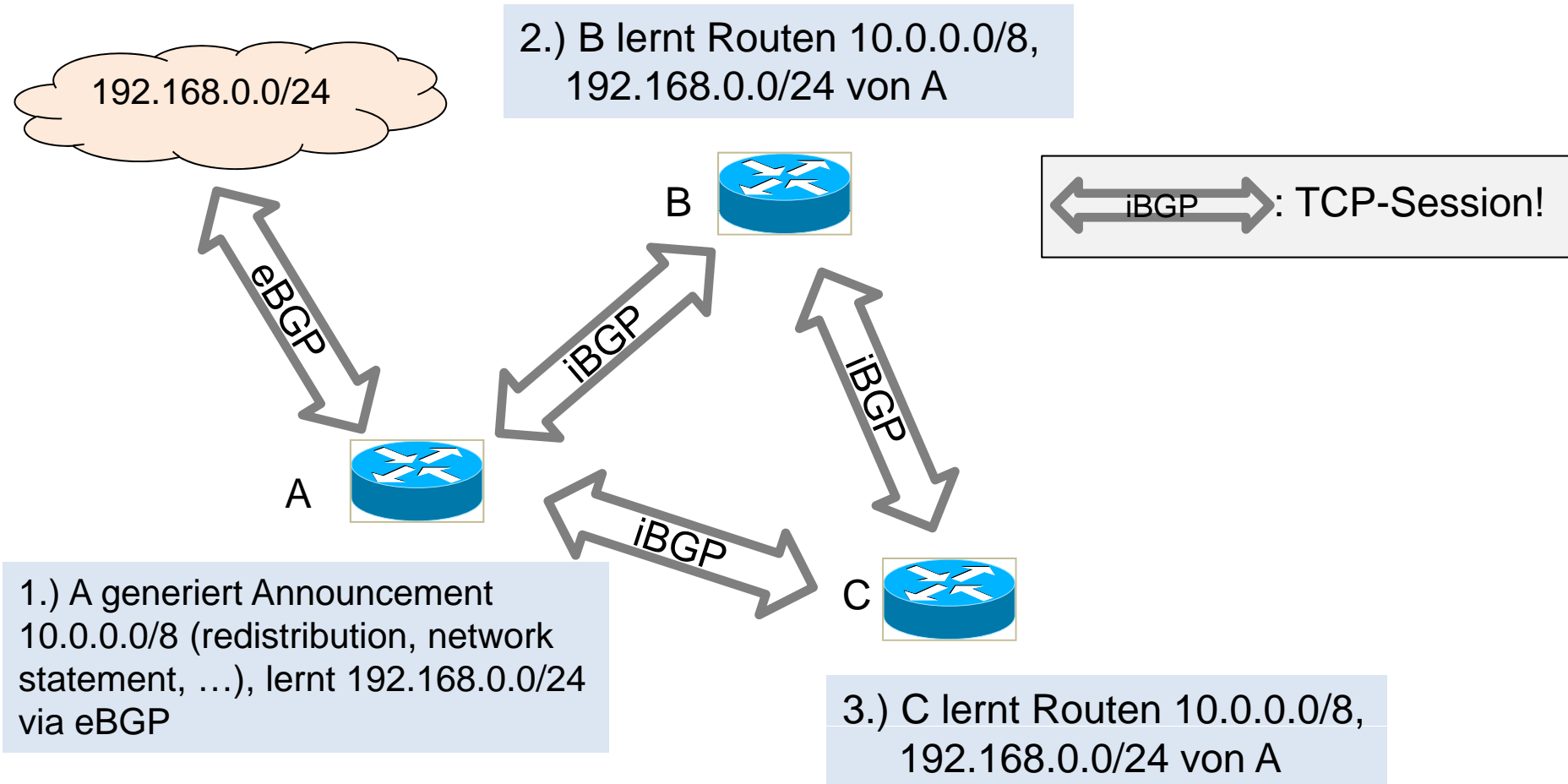
2.) B lernt LAN 3 und LAN 1, annonciert LAN 2



3.) A lernt LAN 2 und annonciert nach außen LAN 1 und LAN 2, aber nicht LAN 3!

1.) C annonciert LAN 3

Vollvermaschung/Reflection



ergo: Vollvermaschung oder Route-Reflection
notwendig, damit Routinginformation überall bekannt ist.

BGP Routing

BGP routing table entry for 194.95.241.0/24

Netz

Versions:

Process	bRIB/RIB	SendTblVer
Speaker	5203915	5203915

BGP

Paths: (2 available, best #1)

Advertised to update-groups (with more than one peer):
0.6 0.11

Path #1: Received by speaker 0

65023, (Received from a RR-client)

188.1.200.36 (metric 1150) from 188.1.200.36 (188.1.200.36)

Origin IGP, metric 50, localpref 100, valid, internal, best

Community: 680:6

Netzausgang

Next Hop

Ausgangsinterface

Routing entry for 188.1.200.36/32

Known via "ospf 680", distance 110, metric 1150, type extern 1

Installed Sep 10 06:11:35.788

Routing Descriptor Blocks

188.1.145.50, from 188.1.200.36, via TenGigE0/7/0/7

Route metric is 1150

No advertising protos.

IGP (link-state)

188.1.145.48/30, version 0, attached, connected, ...

[...]

via TenGigE0/7/0/7 (Next hop table - 0xe0000000), ...

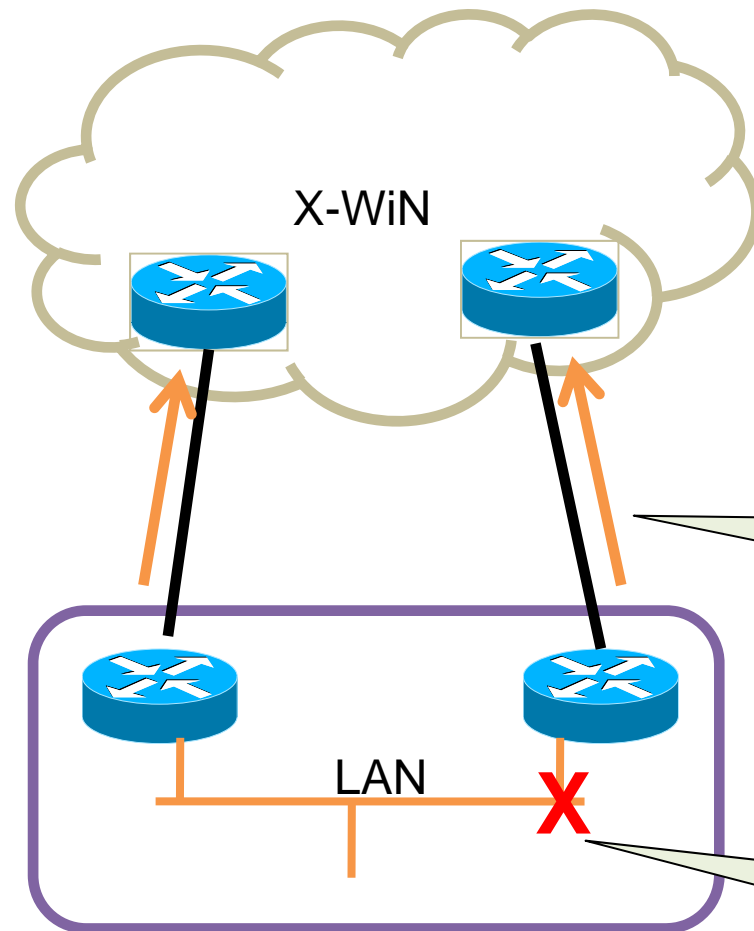
CEF (FIB)

- BGP annonciert ein Netz nur dann, wenn er selbst eine Route für dieses Netz hat
 - ‚hat eine Route‘: d.h. das Netz hat einen Netzausgang zugewiesen und für den Netzausgang existiert ein gültiger next hop
 - diese Route kann statisch, connected, OSPF, IS-IS, RIP, etc. sein
 - BGP-Dokument:

```
! der Router muss eine Route haben, die exakt dem network-Statement  
! im BGP Teil entspricht, damit die Route nach aussen annonciert wird  
ip route 103.10.10.0 255.255.255.0 Null 0 220  
ipv6 route 2001:638:1234::/48 Null 0 220
```

implizite Annahme, dass der Router selbst eine weitere Route für das Netz hat

vorsicht Falle

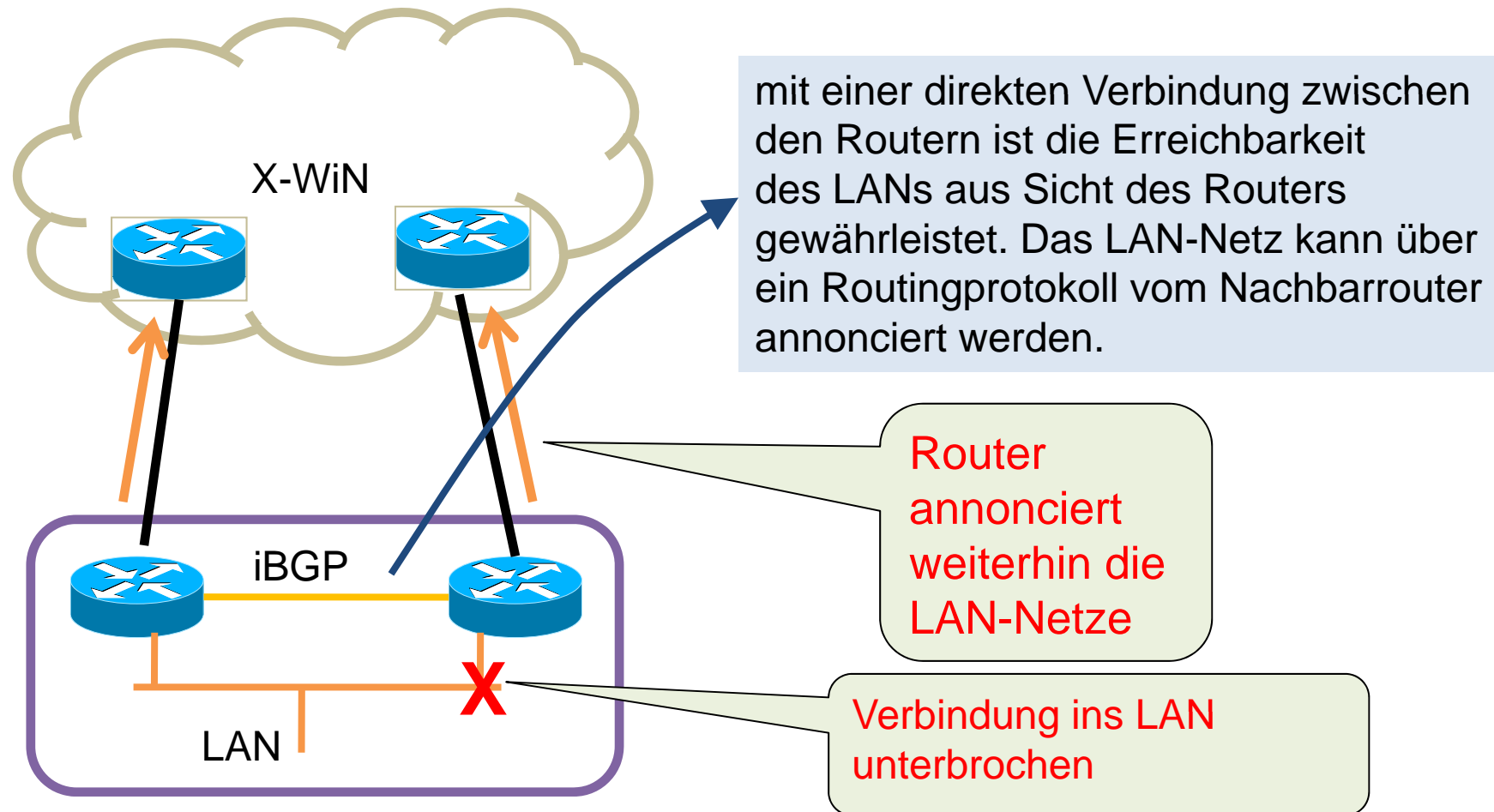


statische Nullroute + network-Statement für das LAN-Netz bedeutet, dass der Router das Netz immer annonciert, auch wenn er selbst das Netz nicht erreichen kann

Router annonciert weiterhin die LAN-Netze

Verbindung ins LAN unterbrochen

- trotzdem ist es sinnvoll, eine Null-Route mit hoher ‚administrative distance‘ für die eigenen Netze zu haben
 - somit wird verhindert, dass Routing-Loops auf der Zugangsleitung entstehen z.B. wenn Pakete an nicht vergebene Adressbereiche der Einrichtung verschickt werden oder ankommen
 - aber: Nullrouten nur für Netzbereiche, für die der Router selbst der einzige Ausgang ist
 - sonst Verlust der Redundanz (s.u.)



gut, aber noch nicht perfekt

- Alternativen zur Kombination **„network“ Statement + statische Nullroute.**
- **Vorsicht:** „network“ Statement muss identisch (Prefix/Maske) mit der weiteren Route sein
 - **„network statement“ + connected Route**
 - der Interfacestatus muss „down“ gehen, wenn die „connected“ Routen nicht mehr erreichbar sind
 - **„network statement“ + IGP-Route**
 - Announcement über OSPF, IS-IS, RIP, EIGRP
- oder lernen der Route über BGP

- Zauberwort ‚Redistribution‘
 - alternative zu ‚network‘-Statement
 - Routinginformation aus anderen Routingprotokollen wird ins BGP übernommen

```
router bgp 65666  
address-family ipv4 unicast  
redistribute ospf route-map filter
```

```
router bgp 65666  
address-family ipv4 unicast  
redistribute connected route-map filter
```

```
router bgp 65666  
address-family ipv4 unicast  
redistribute is-is route-map filter
```

```
router bgp 65666  
address-family ipv4 unicast  
redistribute eigrp route-map filter
```

- Vorteile der Redistribution:
 - IGP-> BGP sinnvoll in komplexeren Topologien
 - die IGP-Information muss natürlich verlässlich sein
 - Router annonciert über BGP nur Routen, die er selber kennt
 - Router erreicht das Netz nicht -> Router annonciert das Netz nicht

- auch sowas geht ...

```
ip route 10.10.10.0 255.255.255.0 10.5.5.1 floating static
```

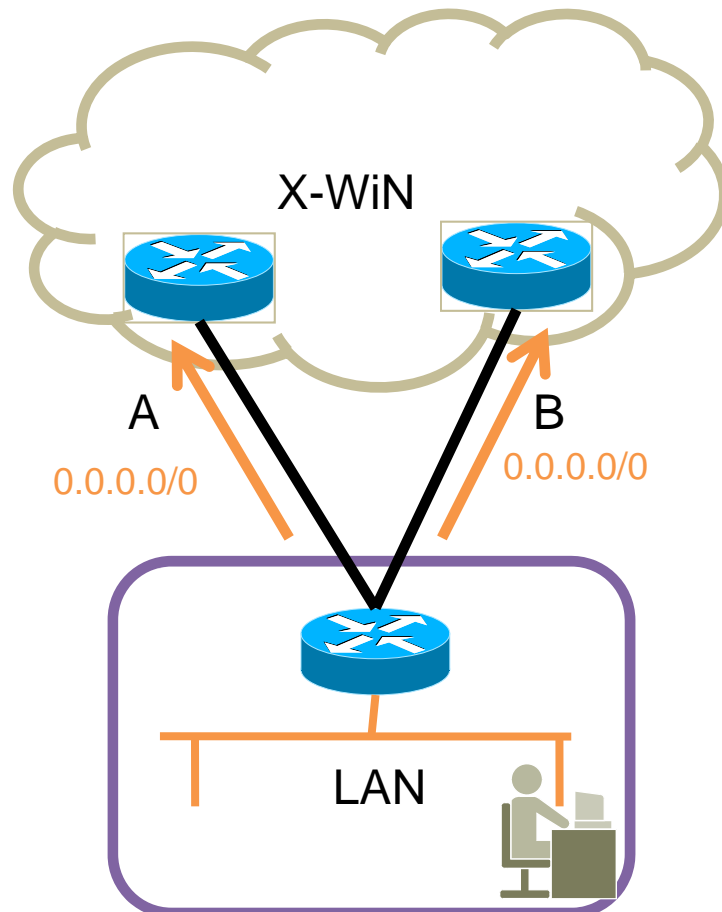
```
router bgp 65666  
  address-family ipv4 unicast  
    redistribute static route-map lan
```

BGP Route 10.10.10.0/24 wird nur dann generiert, wenn 10.5.5.1 in der Routingtabelle

```
>router show ip route 10.5.5.1  
Routing entry for 10.5.5.1/32  
  Known via "ospf 1", distance 110, metric 20, type extern 2, forward metric 1  
  Last update from 195.37.191.38 on GigabitEthernet1/11, 3w5d ago  
  Routing Descriptor Blocks:  
    * 195.37.191.38, from 195.37.191.34, 3w5d ago, via GigabitEthernet1/11  
      Route metric is 20, traffic share count is 1
```

- Wie bekomme ich ein Netz ins BGP?
 1. network-statement
 - in Kombination mit statischer Route, connected Route, IGP-Route
 2. Redistribution
 - RIP, OSPF, IS-IS, statisch, connected, EIGRP, ...
 3. vom BGP-Nachbarn gelernt
 - wenn schon vom Nachbarn, dann am besten über BGP ohne Redistribution

- wieso in Richtung X-WiN nicht beide Links gleich nutzen?



2 Default-Routen, gleich gewichtet.

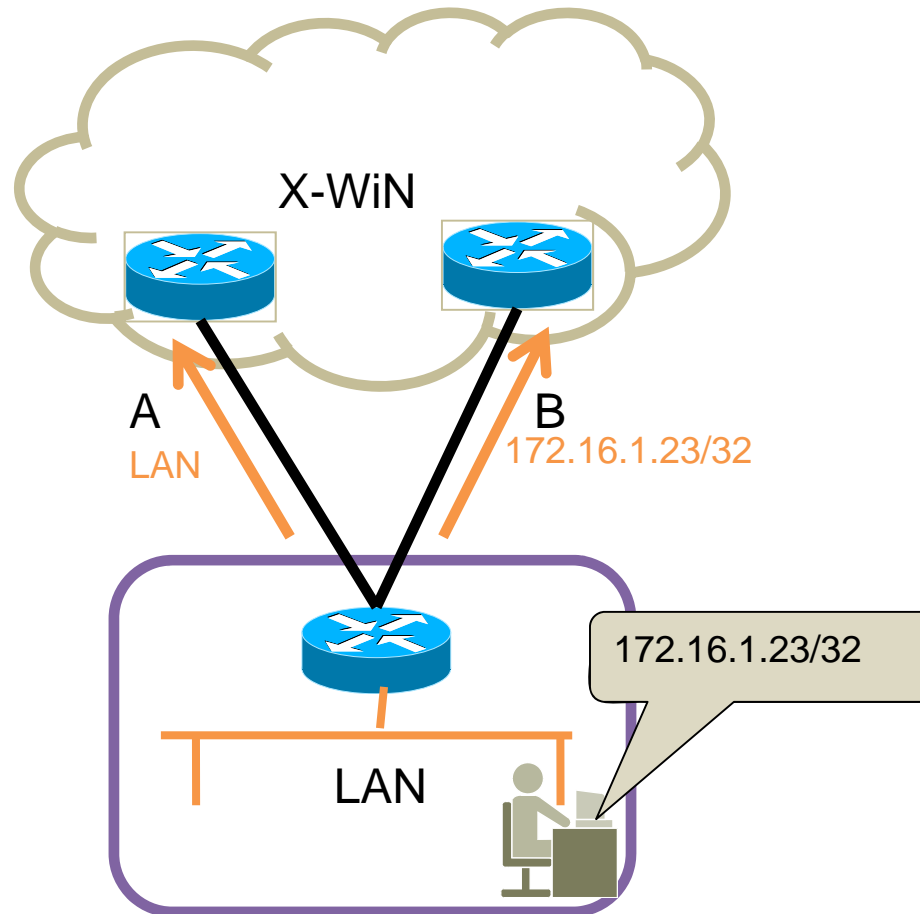
Link A: Hauptleitung

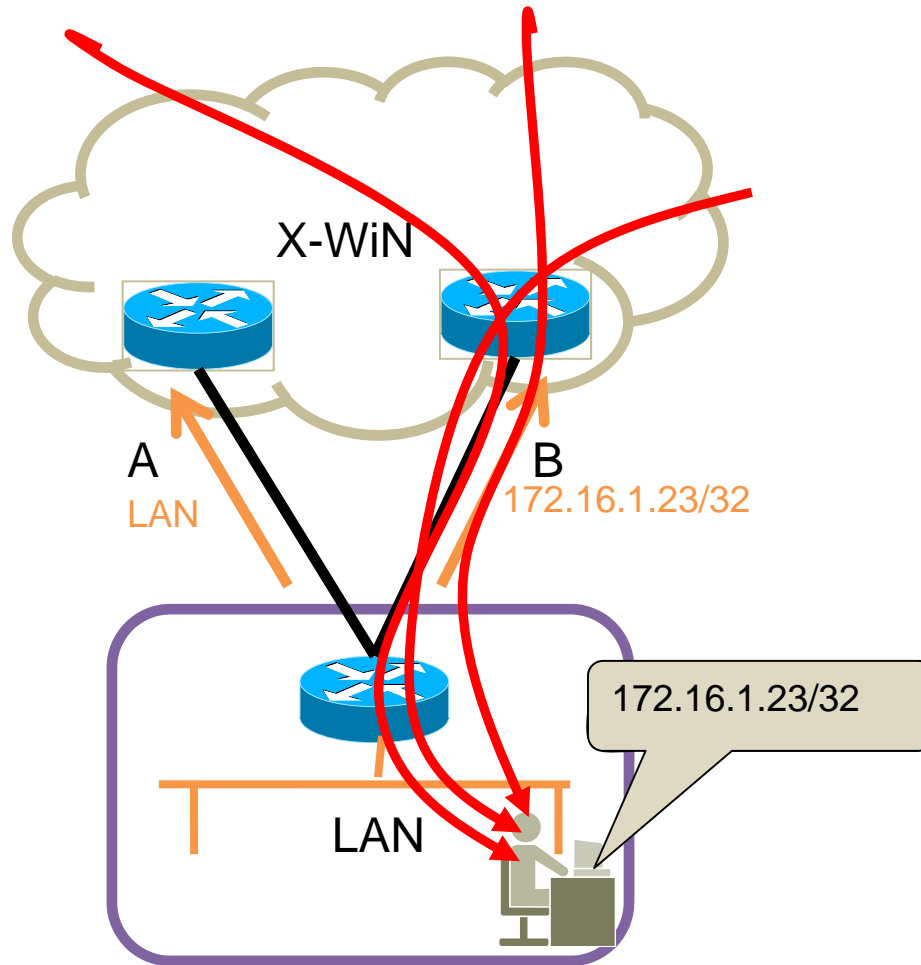
Link B: Nebenleitung, 1/3 Bandbreite von Link A

Flows aus dem LAN ins X-WiN werden round-robin auf die Links verteilt. Somit erhält Anwender mal einen guten Durchsatz, mal nicht.

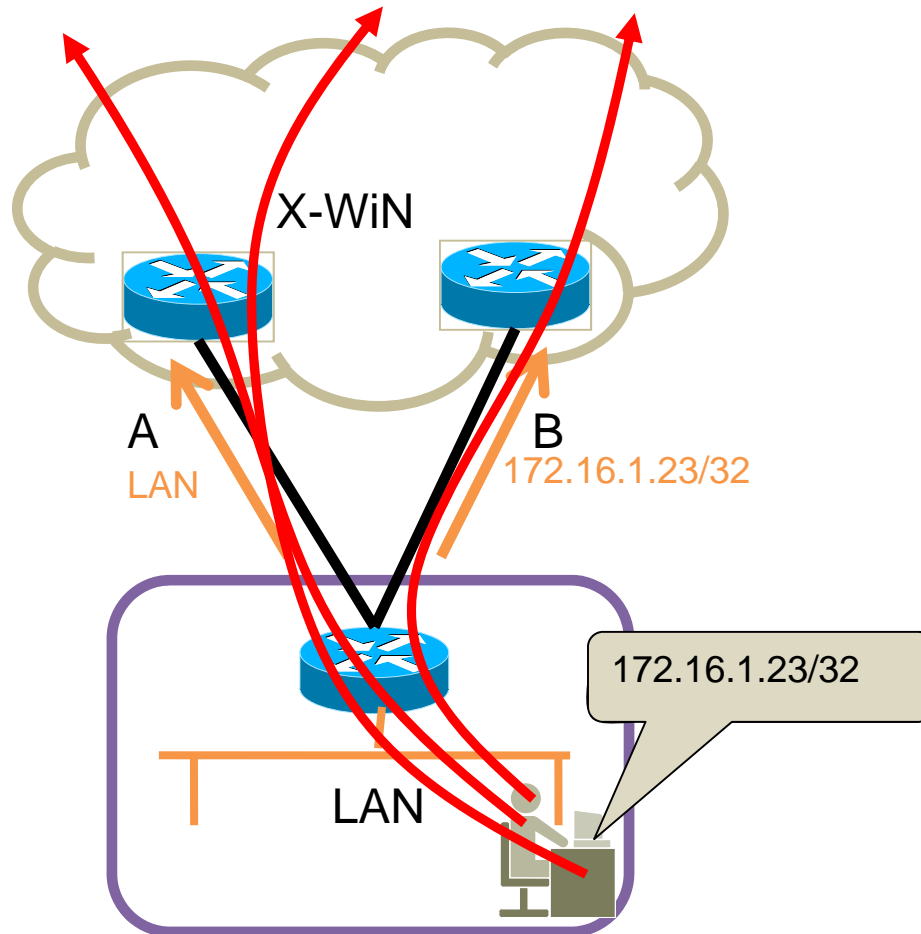
Wenn z.B. Link B voll ist, aber Auf Link A noch Platz ist, werden trotzdem weitere Flows dem Link B zugeordnet -> Überlast h

- Verkehr zu Nutzer mit IP 172.16.1.23/32 soll über Link B geroutet werden





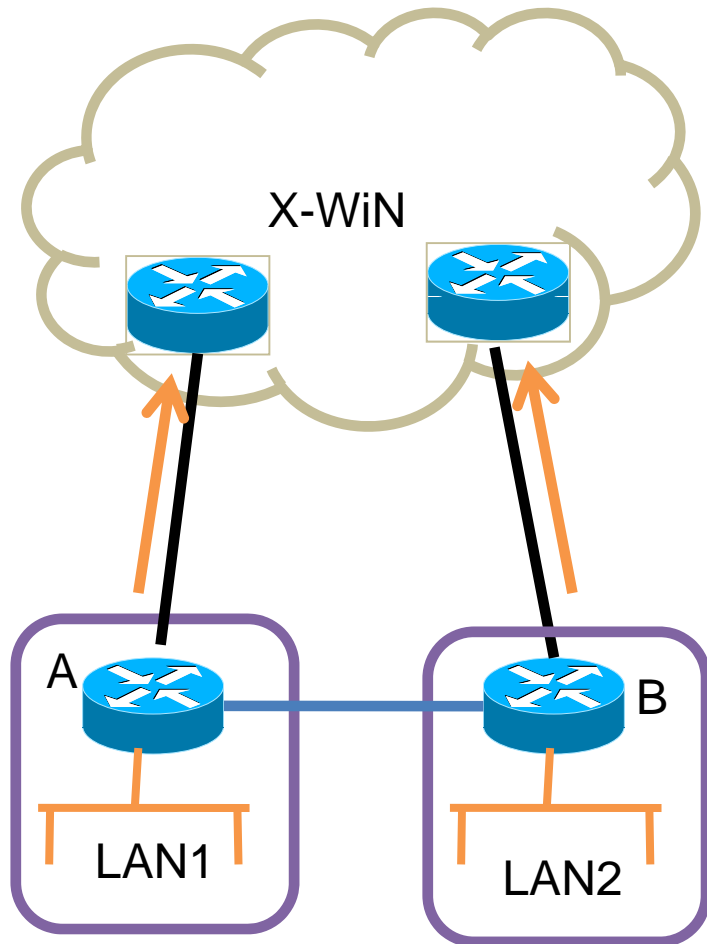
- Subnetz 172.16.1.23/32 wird präferiert (MED) über Link B annonciert
- Aller Verkehr zu dem Host kommt über Link B zu Einrichtung



- Verkehr nach außen läuft entsprechend der lokalen Lastverteilungs-Policy auf Router
- (es sei denn, es wird source-routing gemacht)

- BGP Link Bandwidth
 - The Border Gateway Protocol (BGP) Link Bandwidth feature is used to advertise the bandwidth of an autonomous system exit link as an extended community. This feature is configured for links between directly connected external BGP (eBGP) neighbors. The link bandwidth extended community attribute is propagated to iBGP peers when extended community exchange is enabled. This feature is used with BGP multipath features to configure load balancing over links with unequal bandwidth. [Cisco Doku]

- BGP Link Bandwidth
 - ordnet die Flows nicht round-robin den Links zu, sondern gewichtet entsprechend der zugeordneten Bandbreite (auch bei mehreren Routern!)
 - der Bandbreite, nicht der Auslastung!
 - a priori keine Information über Bandbreitenbedarf des neu hinzukommenden Flows
 - somit kann die Zuordnung auf einen ohnehin schon vollen Link gemacht werden
 - funktioniert nur in Rausrichtung!
 - die Richtung XWiN/Internet -> Anwender wird nicht gewichtet
 - kann trotzdem unter Umständen sinnvoll sein
 - aber nur, wenn man mehrere Router im Netz hat
- Für besonders ambitionierte: Cisco ‚optimized edge routing‘ (OER)
 - berücksichtigt Auslastung / andere Performancemetriken



- dyn. Routing zwischen A und B
 - Router A generiert Announcement für LAN1, Router B für LAN2
- Wunsch: symmetrisches Routing zu/von den Standorten
- Standort -> X-WiN: lokal gelernte Default-Route gewinnt sowieso
 - eBGP vs iBGP, Metrik
- XWiN -> Standort: interne Routingkosten als MEDs an das X-WiN weitergeben
 - alternativ manuell geht immer

```
route-map insXWiN permit 50  
  match ip address <LAN>  
  set metric-type internal
```

so wird am Standort A LAN1 mit kleinem MED, LAN2 mit grösserem MED annonciert.

```
ip prefix-list LAN permit 10.10.10.0/24 ! Standort A
ip prefix-list LAN permit 10.10.11.0/24 ! Standort B
route-map insXWiN permit 10
  match ip address prefix-list LAN
  set metric-type internal
router bgp 65666
[...]
neighbor <X-WiN> route-map insXWiN out
neighbor <X-WiN> distribute-list default in !
```

- die ‚internen Kosten‘ sind die Kosten (z.B. OSPF-Metrik) für das Erreichen des Netzausgangs aus Sicht von BGP! Also die Summe über die Kosten der next-hops bis zum Netzausgang

BGP vs IGP Metrik

BGP routing table entry for 194.95.241.0/24

Versions:

Process	bRIB/RIB	SendTblVer
Speaker	5203915	5203915

BGP

Paths: (2 available, best #1)

Advertised to update-groups (with more than one peer):
0.6 0.11

Path #1: Received by speaker 0
65023, (Received from a RR-client)

188.1.200.36 (metric 1150) from 188.1.200.36 (188.1.200.36)
Origin IGP, metric 50, localpref 100, valid, internal, best
Community: 680:6

IGP Metrik

BGP Metrik

Routing entry for 188.1.200.36/32

Known via "ospf 680", distance 110, metric 1150, type extern 1
Installed Sep 10 06:11:35.788

Routing Descriptor Blocks

188.1.145.50, from 188.1.200.36, via TenGigE0/7/0/7

Route metric is 1150

No advertising protos.

IGP (link-state)

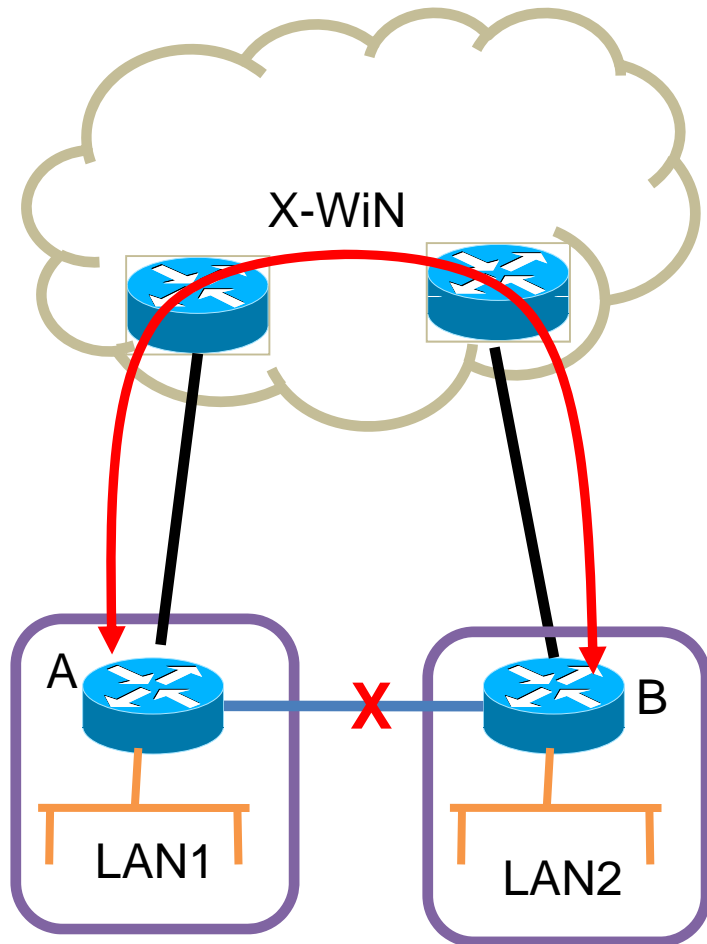
188.1.145.48/30, version 0, attached, connected, ...

[...]

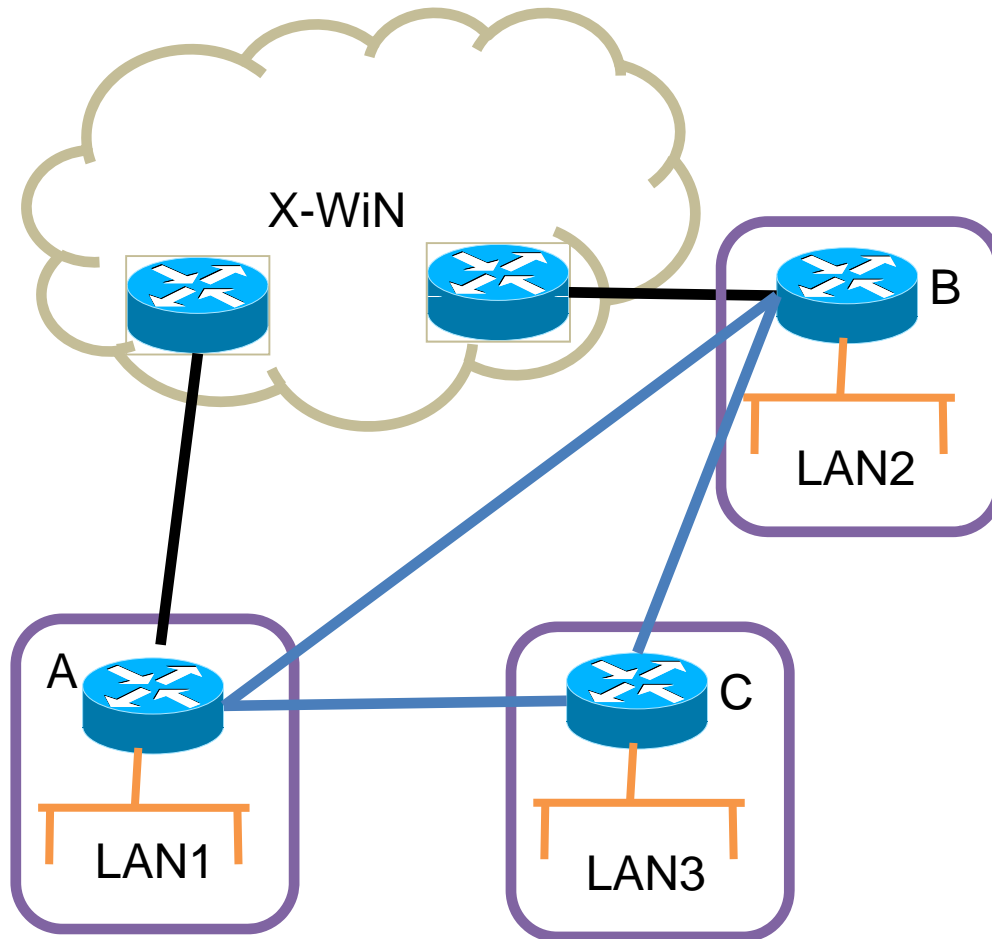
via TenGigE0/7/0/7 (Next hop table - 0xe0000000), ...

CEF (FIB)

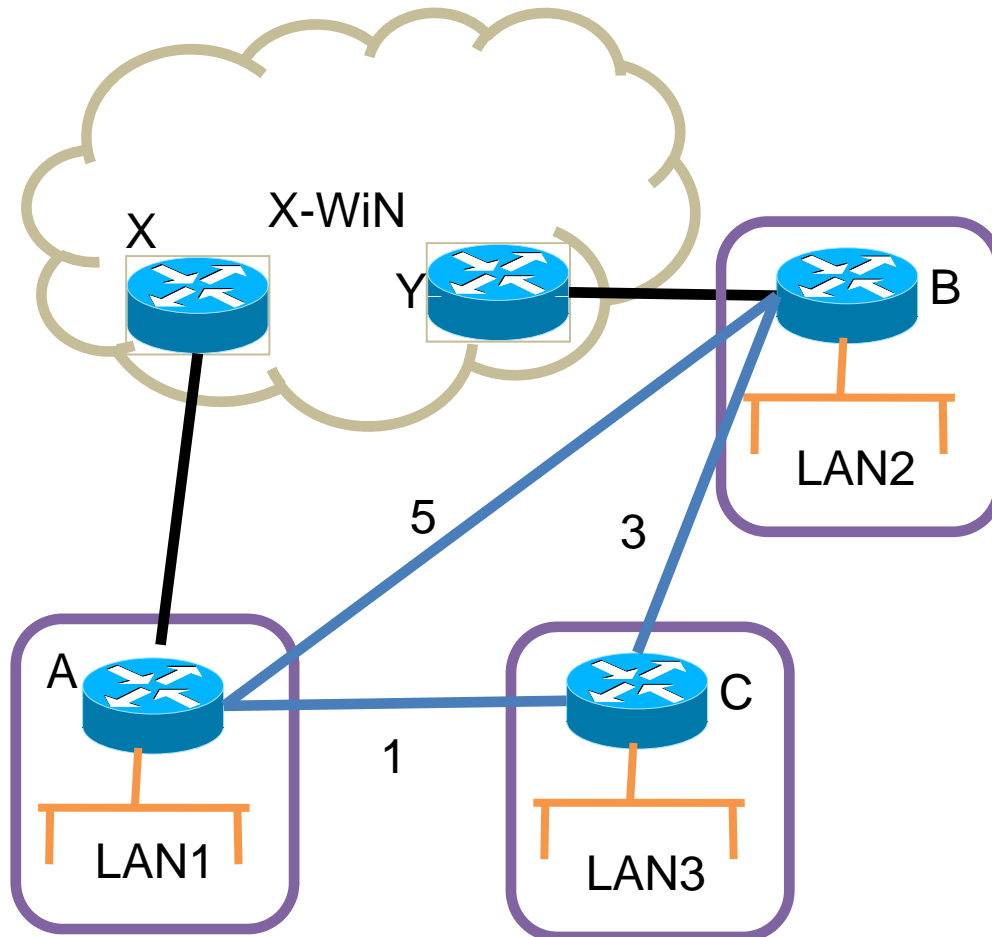
verteilte Standorte, Ausfall



- Verbindung A-B unterbrochen
- A kennt keine Route mehr zu LAN2
 - weil nur B eine Route generiert
- somit Routing über X-WiN



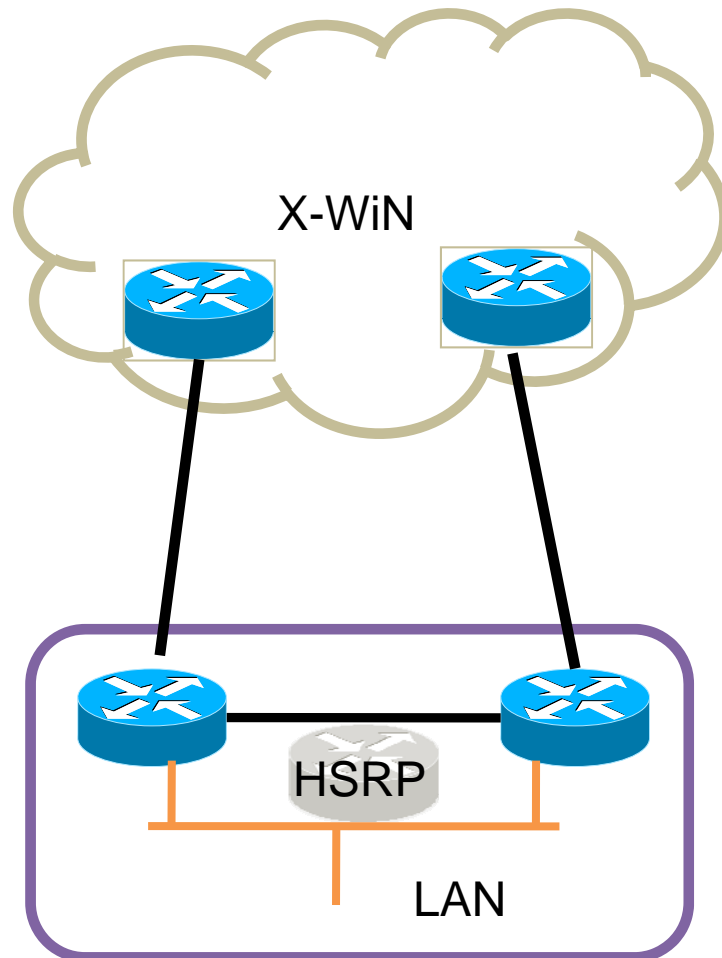
- jeder Router generiert BGP Announcement für lokales LAN
 - und nur dafür!
 - andere Routen werden gelernt
 - geht genauso gut auch mit anderen internen Routingprotokollen + Redistribution



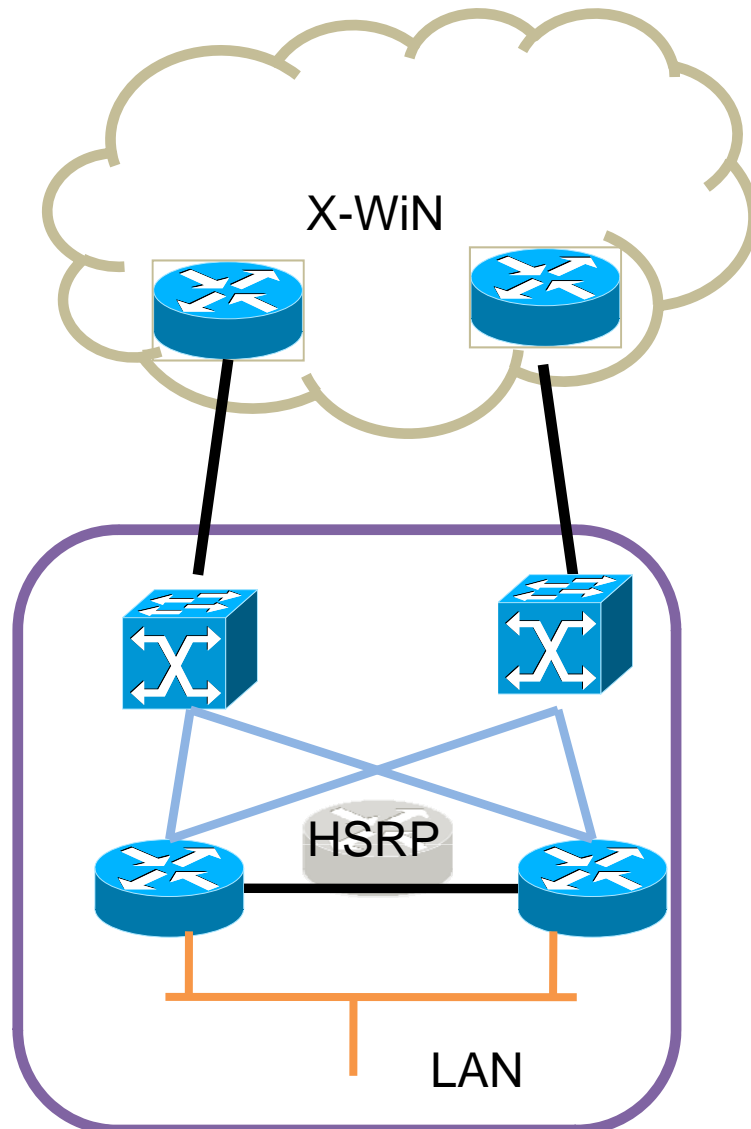
- BGP Volvermaschung
- IGP für next-hops
- set metric-type internal

A annonciert an X
LAN1: MED 0
LAN2: MED 5
LAN3: MED 1

B annonciert an Y
LAN1: MED 5
LAN2: MED 0
LAN3: MED 3



- HSRP aka VRRP
- beide Router erscheinen aus Sicht des LANs wie 1 virtueller Router
 - virtuelle IP Adresse mit eigenem ARP Eintrag
- fällt ein Router aus, übernimmt der andere Router die Rolle als Gateway
- Routing von/nach aussen wie gewohnt
 - Router müssen miteinander physikalisch verbunden sein



- beide Router erscheinen aus Sicht des WiNs wie ein virtueller Router
 - virtuelle IP Adresse mit eigenem ARP Eintrag
- fällt ein Router aus, übernimmt der andere Router die Rolle als KR
- Erweiterung des Adressbereiches auf Zugangsleitung notwendig
- Kosten/Nutzen?

- BGP Szenario 1,2 ohne globale Routingtabelle
 - zusätzliche Anforderungen an Hardware (CPU, Speicher) gering
- BGP Szenario 2 mit globaler Routingtabelle
 - rel. hohe Anforderung an CPU, Speicher
- Ansonsten, Wahl der Hardware primär von anderen Faktoren anhängig
 - Durchsatz pps
 - Architektur
 - Redundanz
 - Portdichte
 - verfügbare Interfacekarten
 - Firewall Features
 - Accounting Features
 - Reporting Features
 - Policy Features
 - Forwarding Features
 - Management Features
 - etc.

Fragen?

