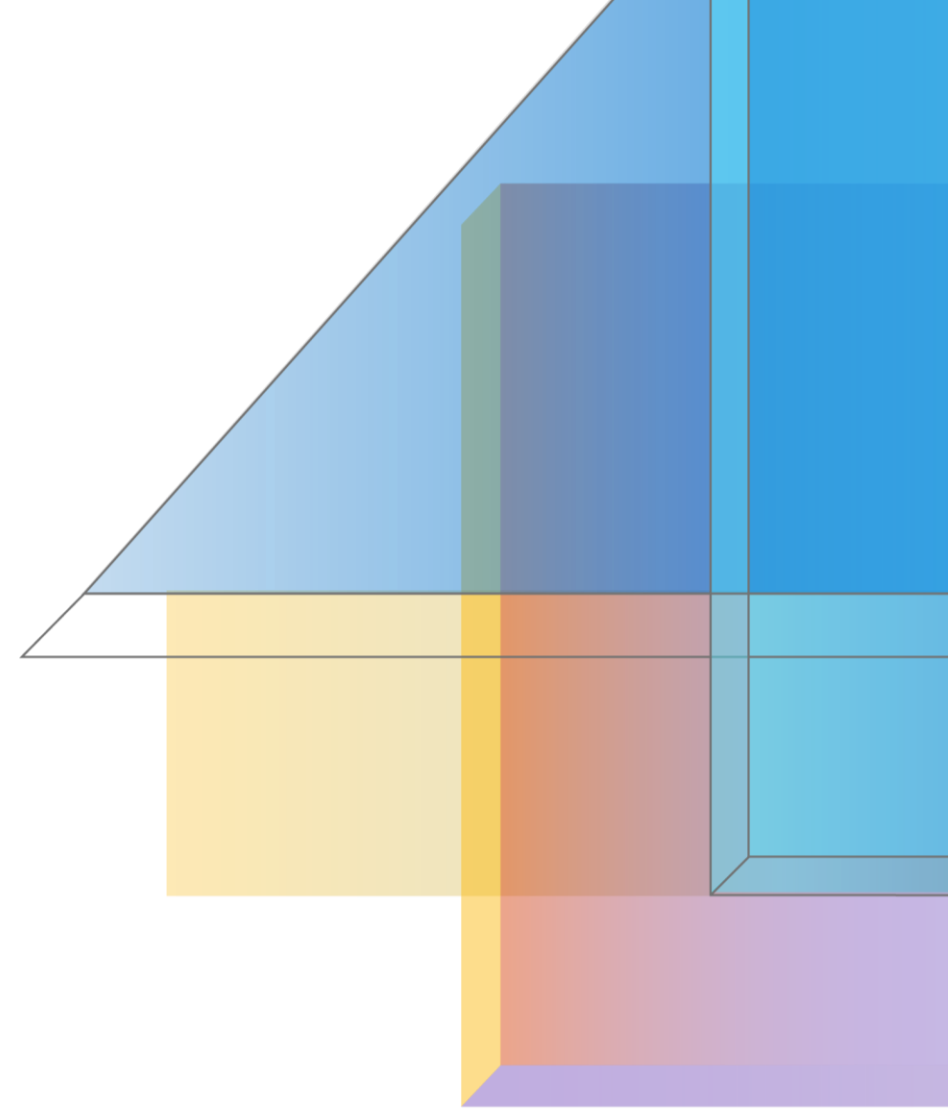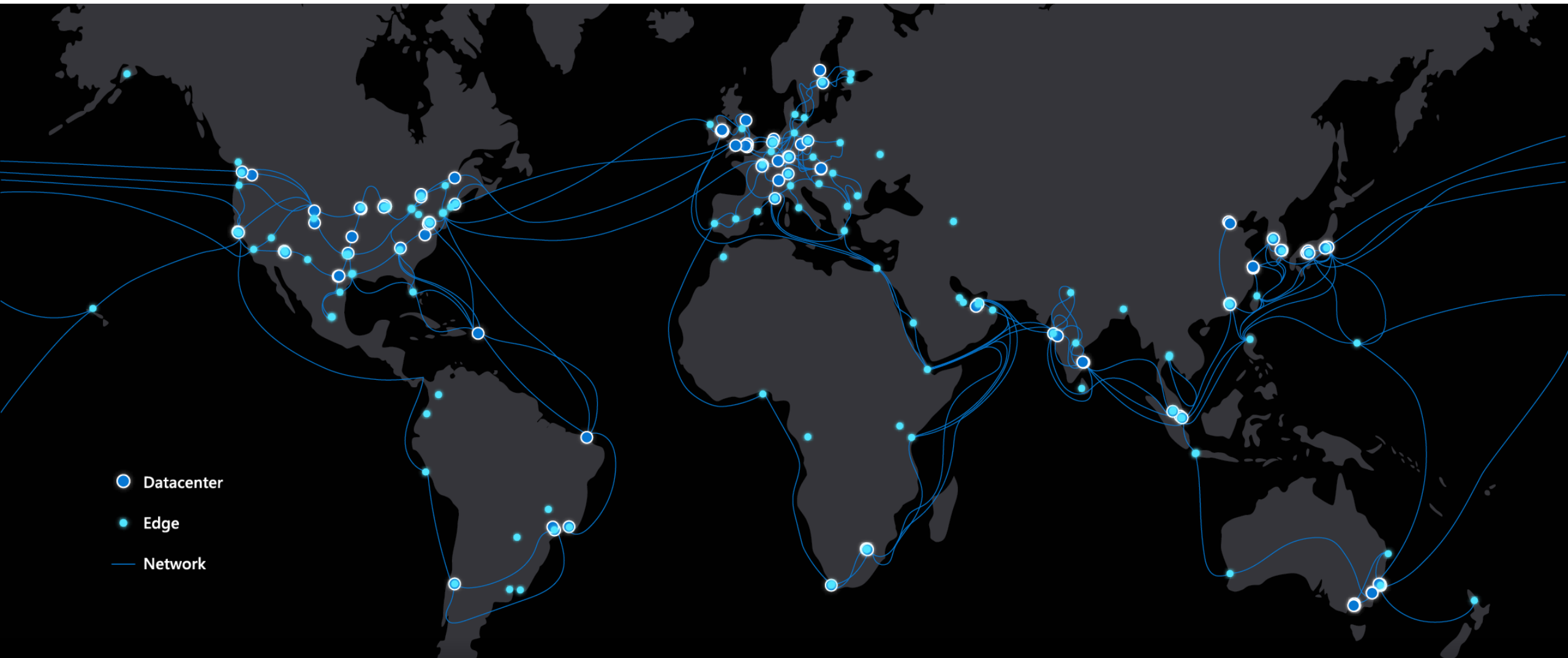# OneWAN is better than two: Unifying a split WAN architecture

Umesh Krishnaswamy*, Rachee Singh, Paul Mattes, Paul-Andre Bissonnette, Nikolaj Bjørner, Zahira Nasrin, Sonal Kothari, Prabhakar Reddy, John Abeln, Srikanth Kandula, Himanshu Raj, Luis Irun-Briz, Jamie Gaudette, Erica Lan

Microsoft global network

Legend:
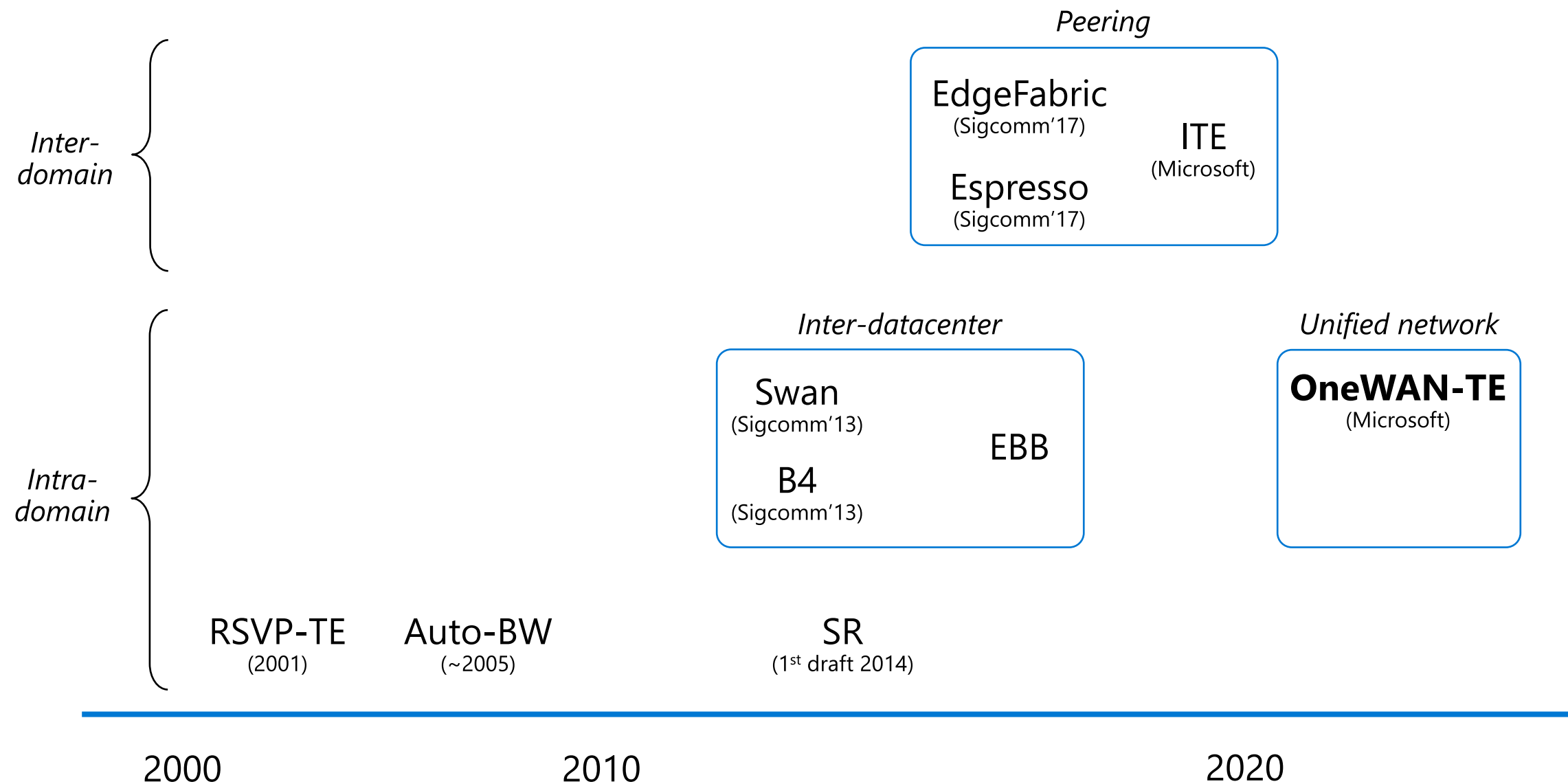- Datacenter
- Edge
- Network

60+ Azure regions

175k+ miles of fiber + subsea cables

190+ Network Edge PoPs

200+ Express Route partners

20k+ peering connections

Traffic Engineering

Inter-domain

Peering

EdgeFabric
(Sigcomm'17)

ITE
(Microsoft)

Espresso
(Sigcomm'17)

Intra-domain

Inter-datacenter

Unified network

Swan
(Sigcomm'13)

EBB

OneWAN-TE
(Microsoft)

B4
(Sigcomm'13)

RSVP-TE
(2001)

Auto-BW
(~2005)

SR
(1st draft 2014)

2000   2010   2020

Two-WAN architecture

Regional gateway

Swan (SDN)

Core (RSVP-TE)

Edge

Core

Inter-datacenter
flow

DC

Internet
flow

Peers
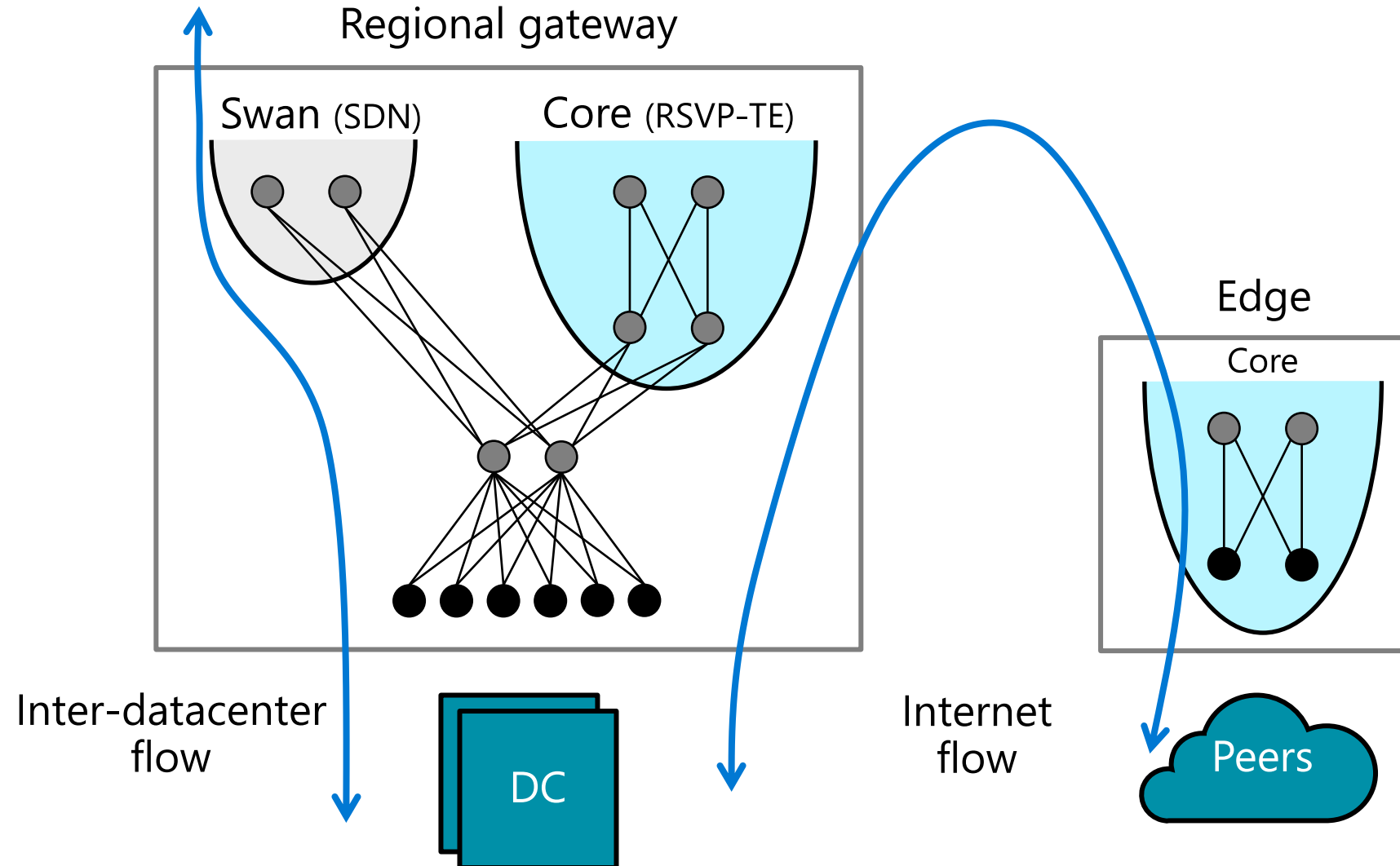
Two-WAN positives

Global TE, operate Swan at higher utilization, and time-shifting discretionary traffic lowered capex.

Core carried Internet flows with higher SLO. Swan evolved to serve critical customer workloads.

BlastShield (NSDI'22) slices enabled global TE without global blast radius.

| Traffic type	SLO
Customer (e.g. Azure, Office, Teams)	99.999%
Discretionary (e.g replication, backup)	99.9%

BlastShield slices

Two-WAN headwinds

Internet traffic growing faster since 2020.

RSVP-TE reaching scale limits in Core. Full mesh involves network-wide change.

Higher device and power costs inside regional gateway due to two WANs.

Capacity planning harder with two WANs. One network can be over-utilized while other is under-utilized.

Traffic growth

Internet

Inter-DC

2020   2021

Extra aggregation
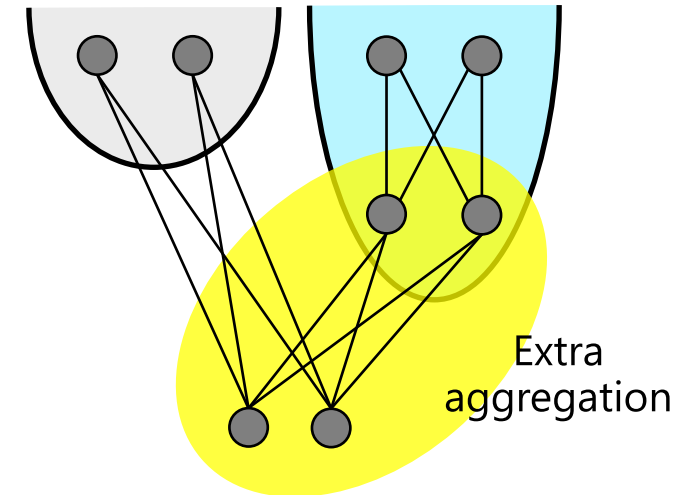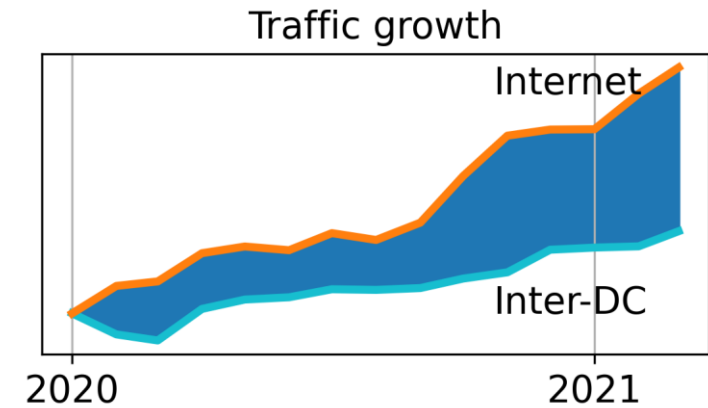
New challenges with unified WAN

10x increase in route scale: $O(10^5) \rightarrow O(10^6)$.

10x increase in number of TE devices: $O(100) \rightarrow O(1000)$.

Match RSVP-TE fast reroute repair times.

Use existing hardware.

Hitless transition in live network.

1. **Traffic steering**

2. Local repair

3. TE optimization
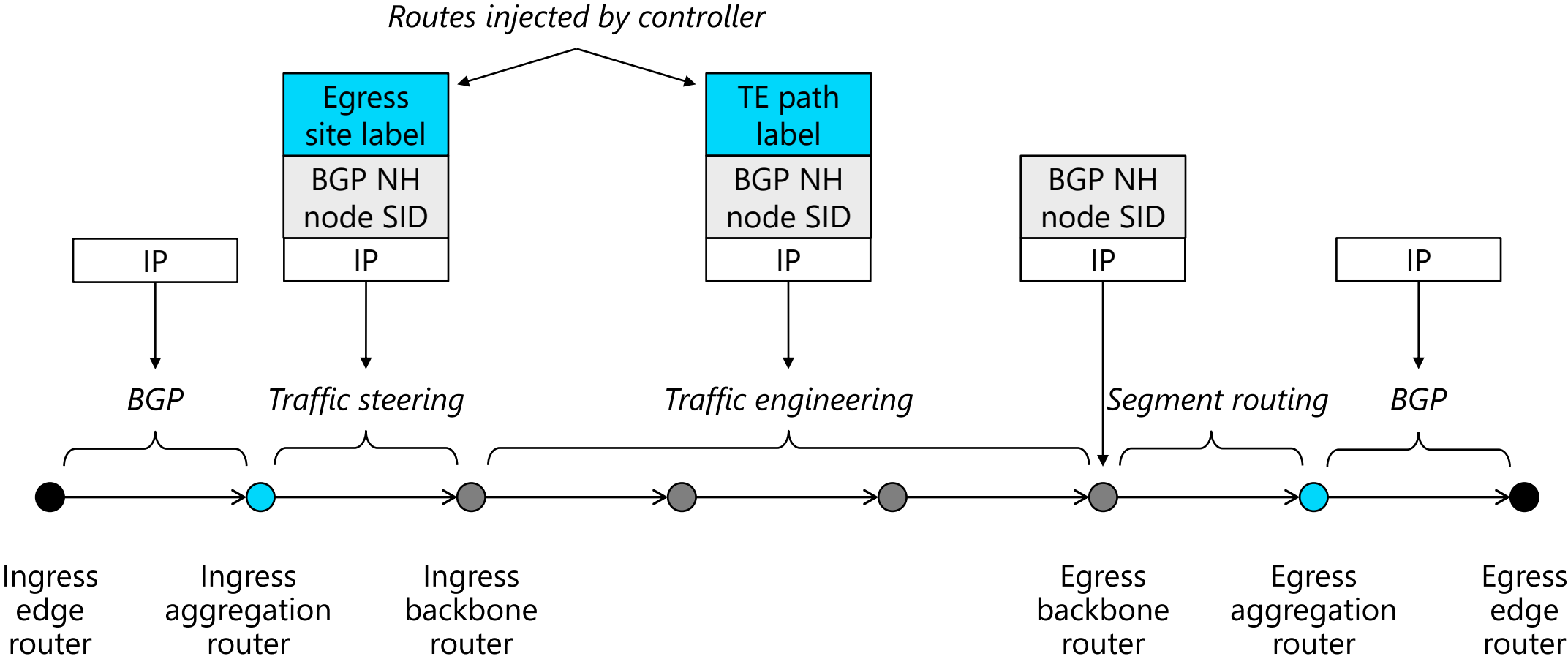
4. Traffic matrices

5. Traffic migration

Why traffic steering?

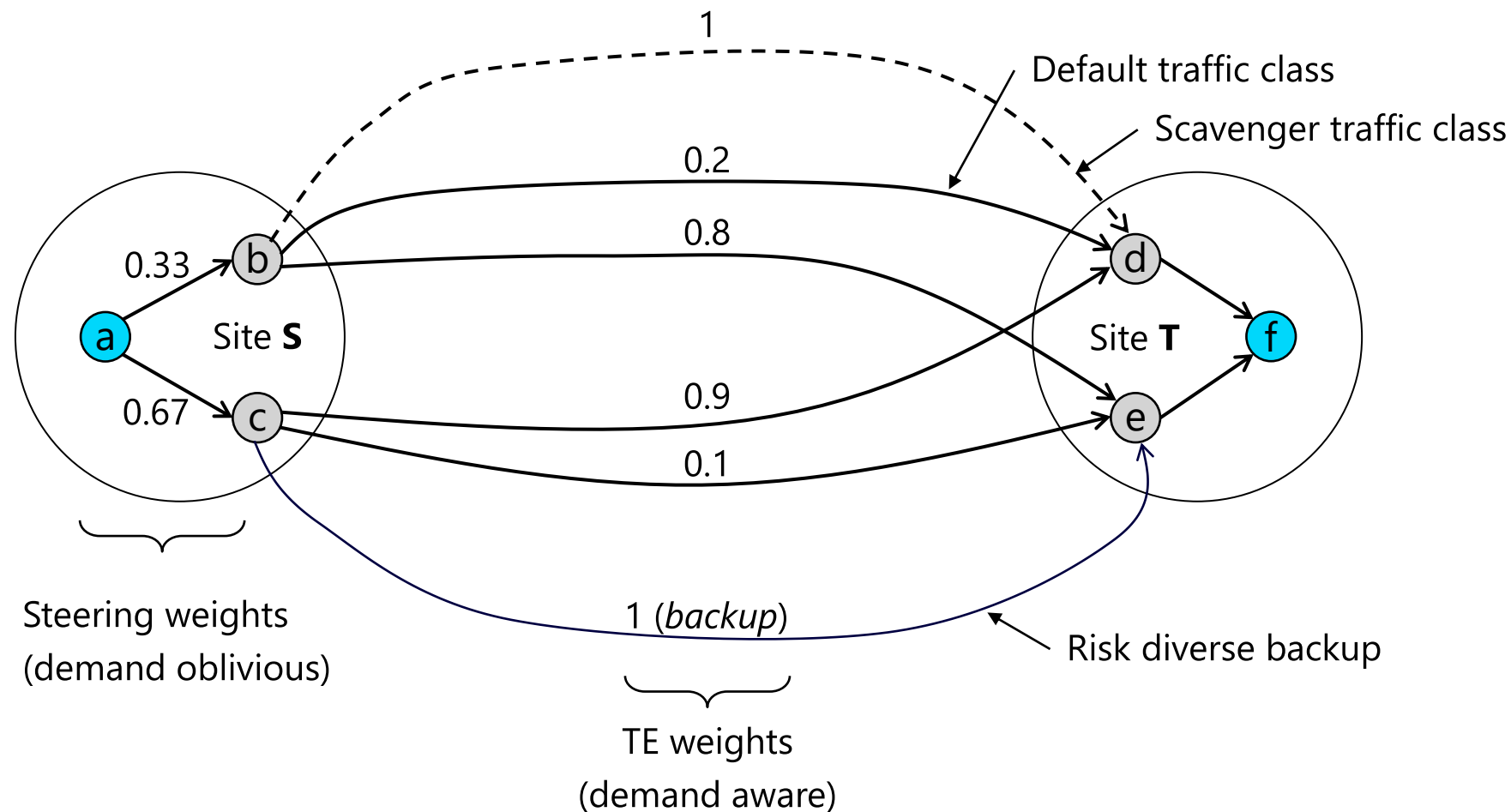The controller does not need to deal with route scale.

Use standard BGP to resolve prefixes to controller routes.

Only aggregation routers hold full Internet routes.

OneWAN routing in three parts

Routes injected by controller

Egress site label
BGP NH node SID
IP

TE path label
BGP NH node SID
IP

BGP NH node SID
IP

IP

IP

BGP *Traffic steering* *Traffic engineering* *Segment routing* *BGP*

Ingress edge router

Ingress aggregation router

Ingress backbone router

Egress backbone router

Egress aggregation router

Egress edge router

OneWAN-TE traffic splitting

1. Traffic steering

2. **Local repair**

3. TE optimization

4. Traffic matrices

5. Traffic migration

Why local repair?

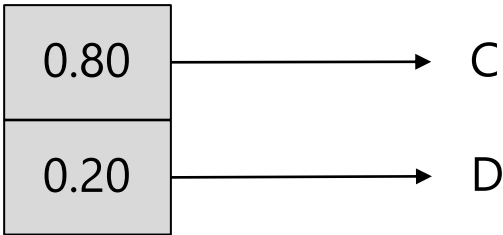Interactive applications need fast route convergence.

Agent can make pre-programmed route changes faster than a controller.
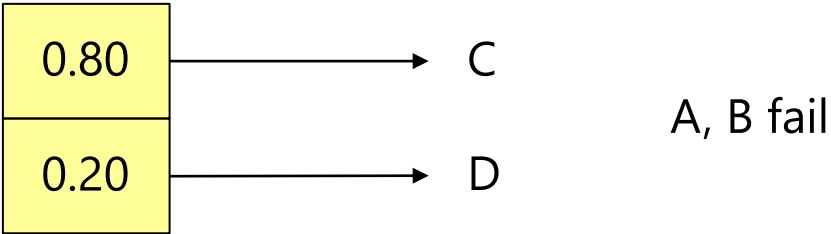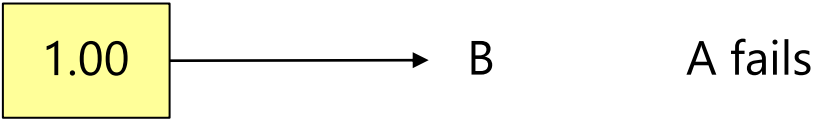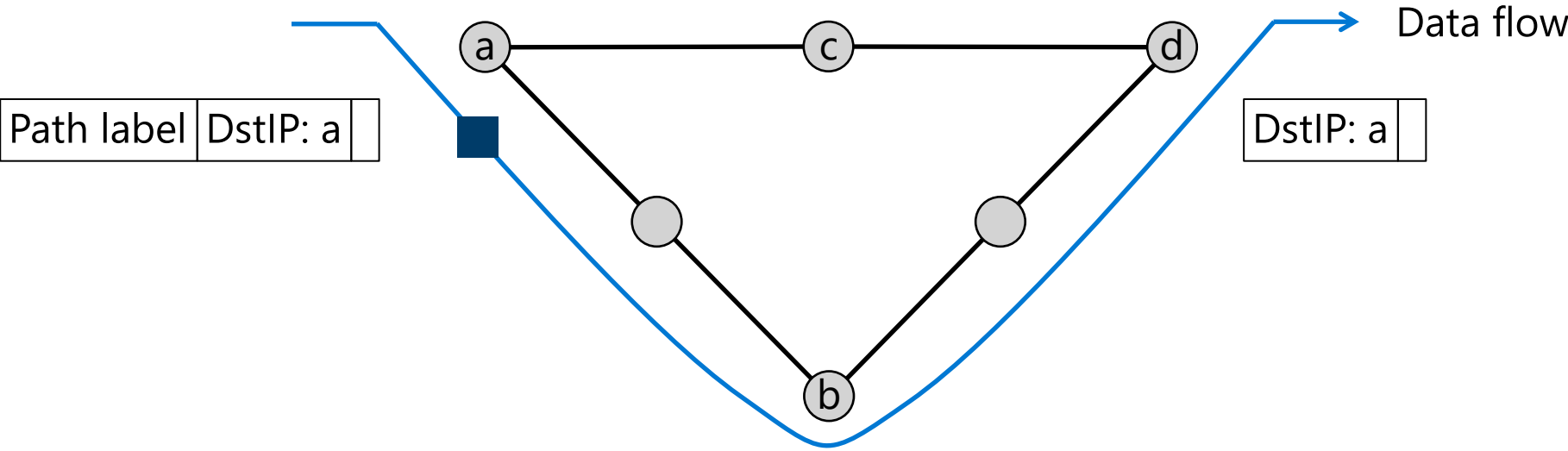
Local repair

Primary default

| 0.75 | → A |
| 0.25 | → B |
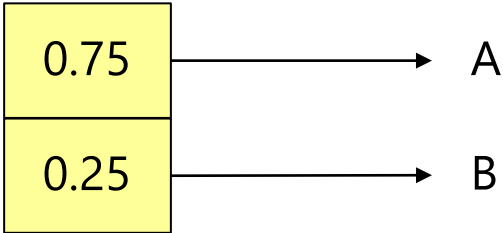
Primary default after local repair

| 1.00 | → B | A fails |

Backup default

| 0.80 | → C |
| 0.20 | → D |

| 0.80 | → C |
| 0.20 | → D |

A, B fail

Pop site label

All fail

Unidirectional tunnel probing in Swan

Data flow

| Path label | DstIP: a | |

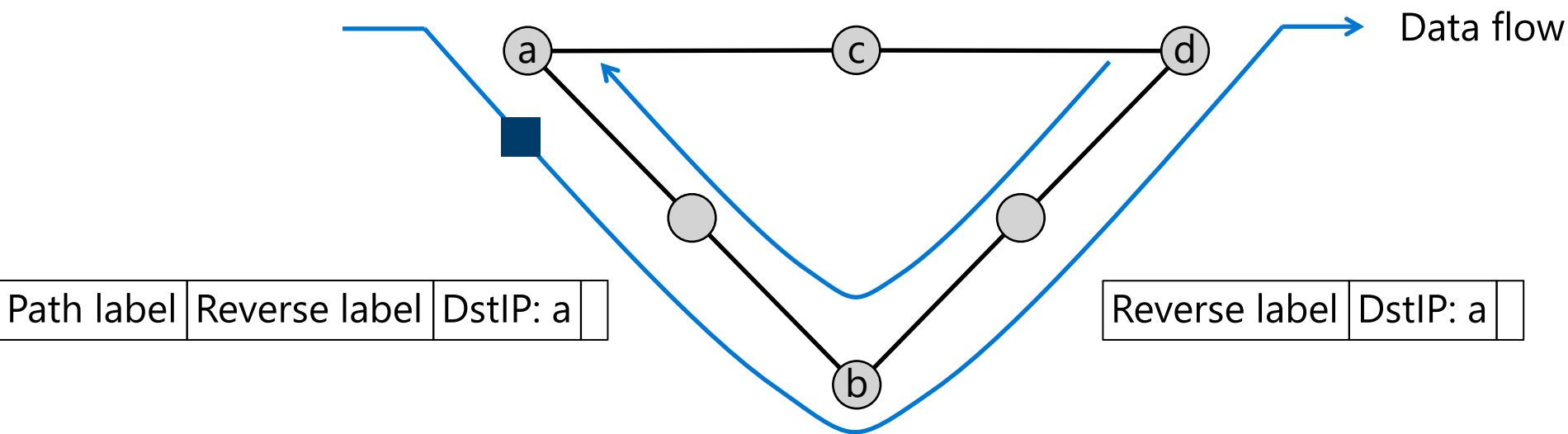| DstIP: a | |

Probes return to tunnel ingress using IS-IS routes.

Route convergence impact

Primary tunnel

Backup tunnel
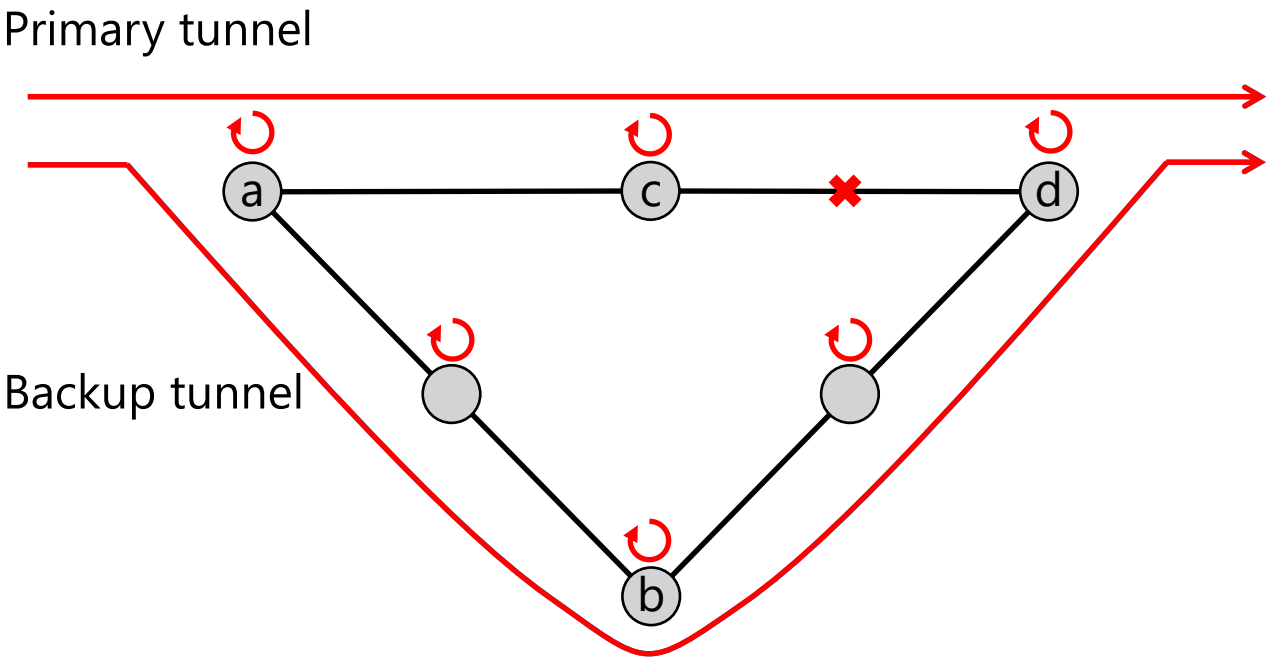
Bidirectional tunnel probing in OneWAN

Data flow

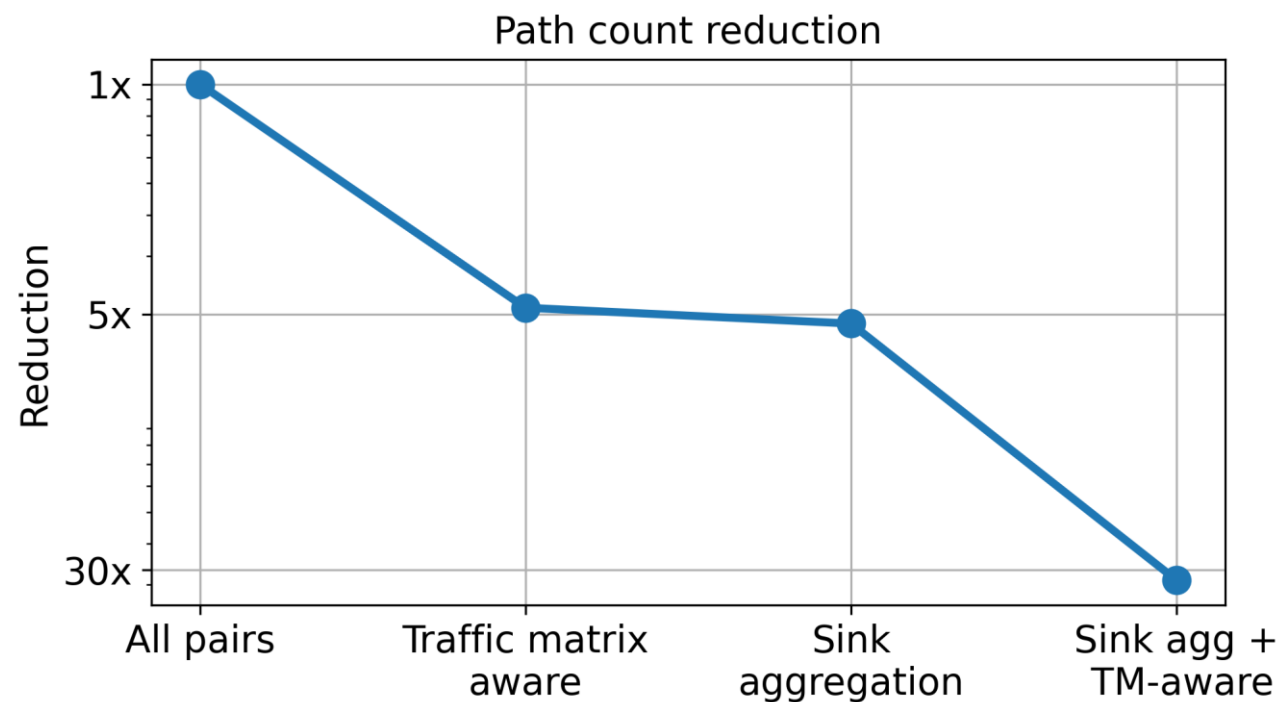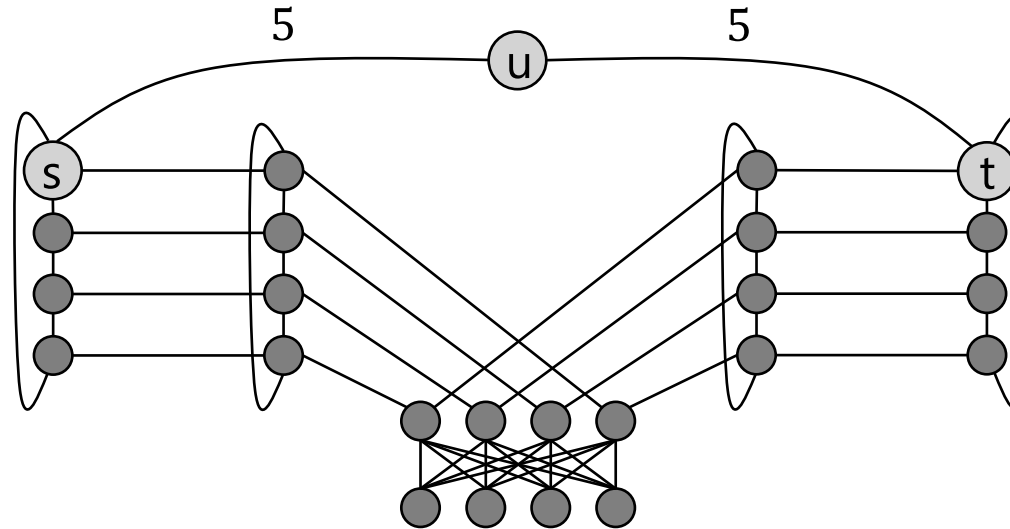| Path label	Reverse label	DstIP: a	

Reverse label	DstIP: a	

1. Traffic steering

2. Local repair

3. **TE optimization**

4. Traffic matrices

5. Traffic migration

Online path computation
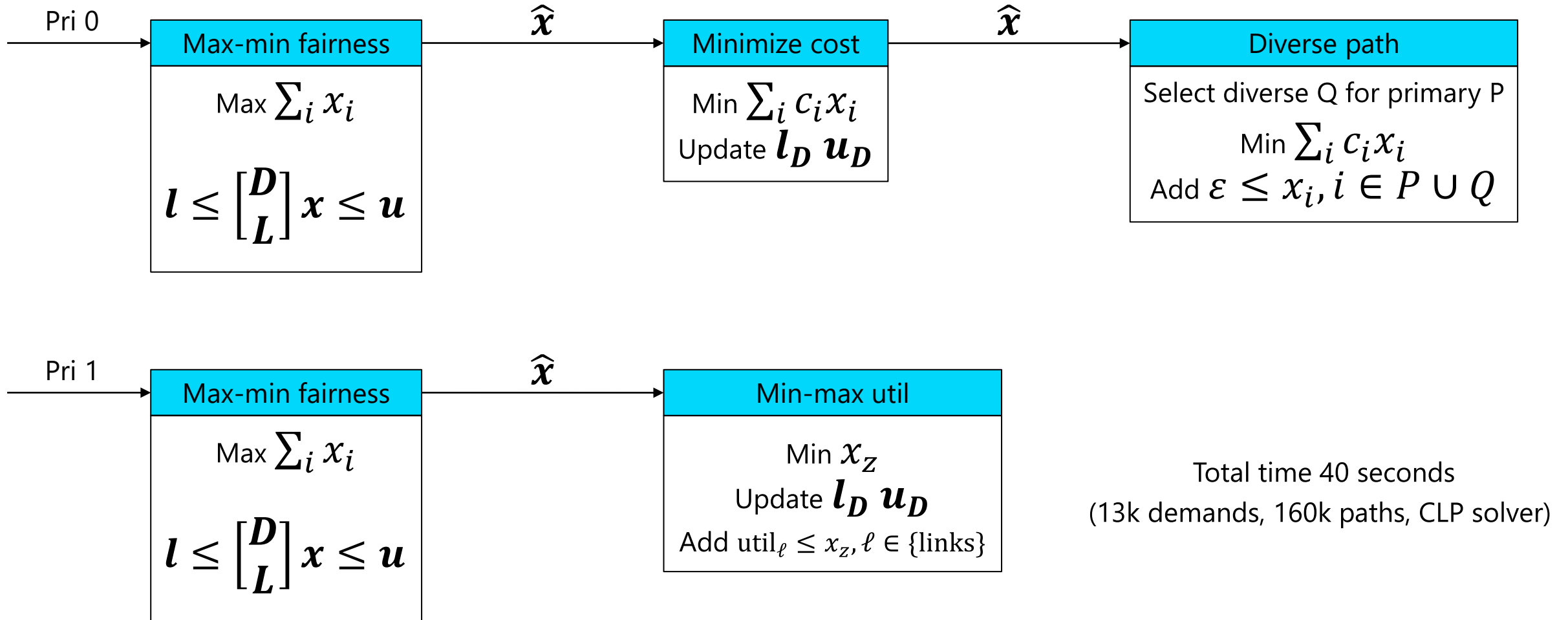

Path count reduction
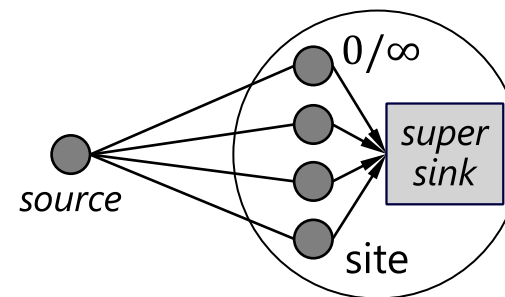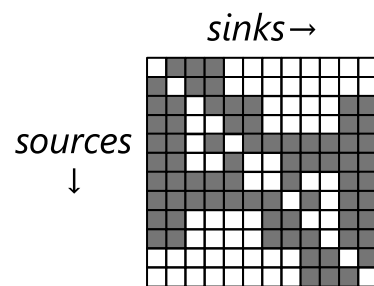
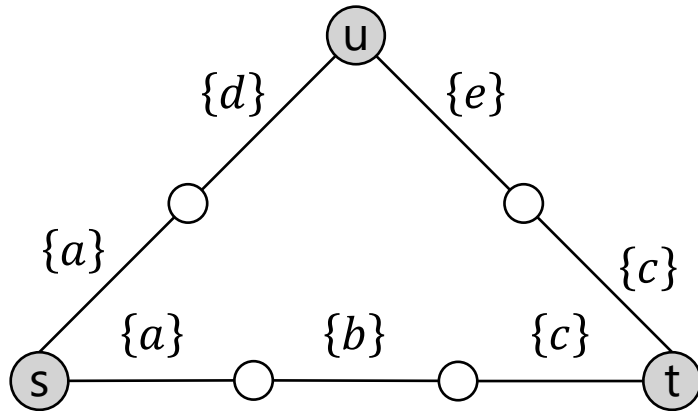# Why we do not use k-shortest paths

k-shortest path finds $s - u - t$ after 668 shorter paths.
Single commodity max-flow path finder is more efficient.

Priority fairness solver chaining

Pri 0 →

Max-min fairness

$$\text{Max} \sum_i x_i$$

$$l \le \begin{bmatrix} D \\ L \end{bmatrix} x \le u$$

$\widehat{x}$ →

Minimize cost

$$\text{Min} \sum_i c_i x_i$$

Update $l_D \; u_D$

$\widehat{x}$ →

Diverse path

Select diverse Q for primary P

$$\text{Min} \sum_i c_i x_i$$

Add $\varepsilon \le x_i, i \in P \cup Q$

Pri 1 →

Max-min fairness

$$\text{Max} \sum_i x_i$$

$$l \le \begin{bmatrix} D \\ L \end{bmatrix} x \le u$$

$\widehat{x}$ →

Min-max util

$$\text{Min } x_z$$

Update $l_D \; u_D$

Add $\text{util}_\ell \le x_z, \ell \in \{\text{links}\}$

Total time 40 seconds
(13k demands, 160k paths, CLP solver)
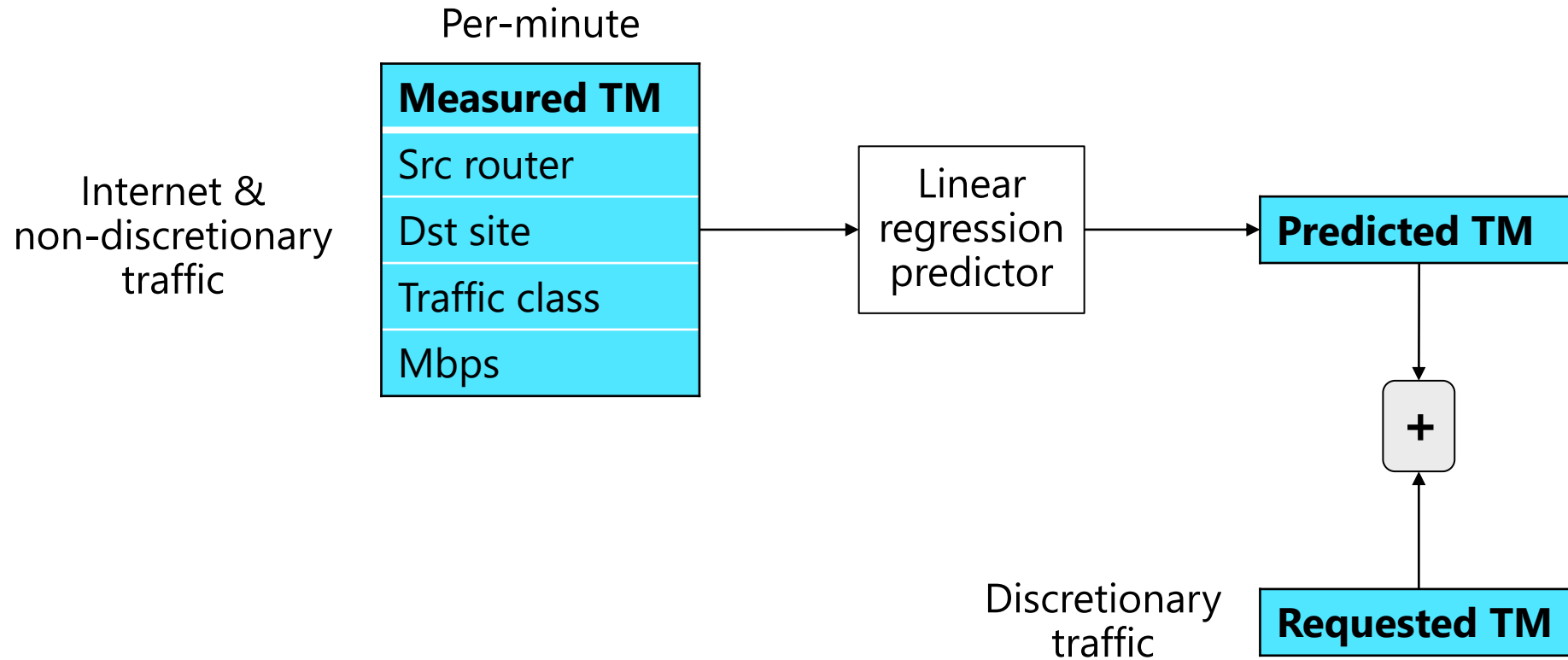
Diverse path solver

Path $s - u - t$ only protects risk $\{b\}$.
Penalizing path finder explores risk diverse paths.
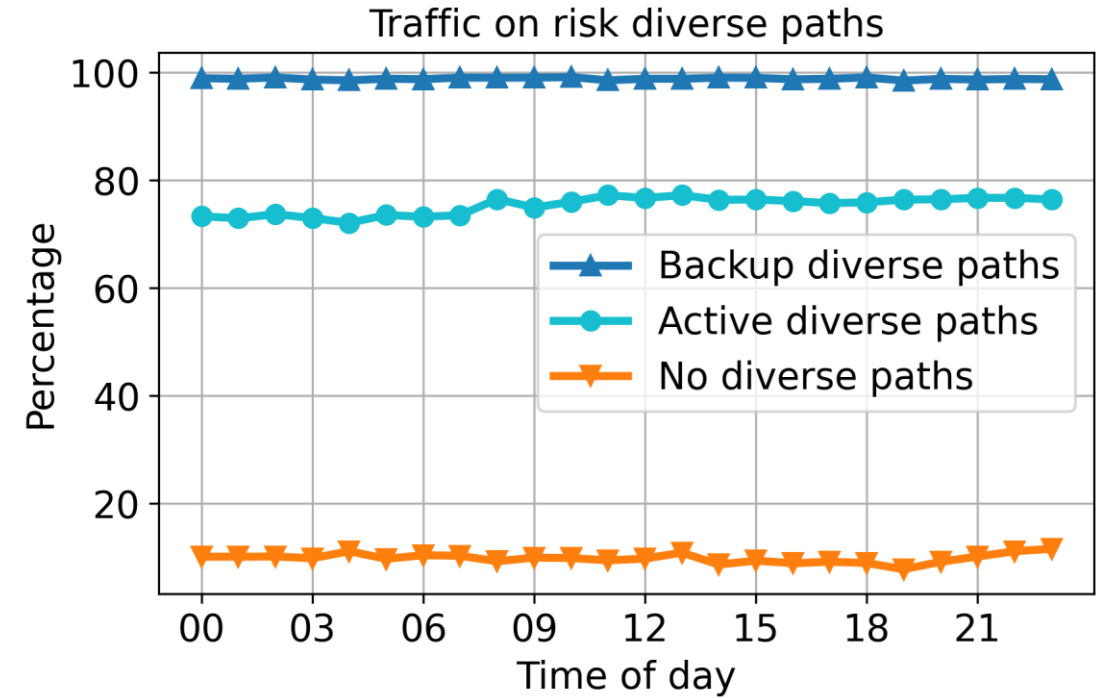
Backup diverse paths do not carry traffic in non-fault cases.
Greater freedom to select diverse paths when backup.

1. Traffic steering

2. Local repair

3. TE optimization

4. **Traffic matrices**

5. Traffic migration

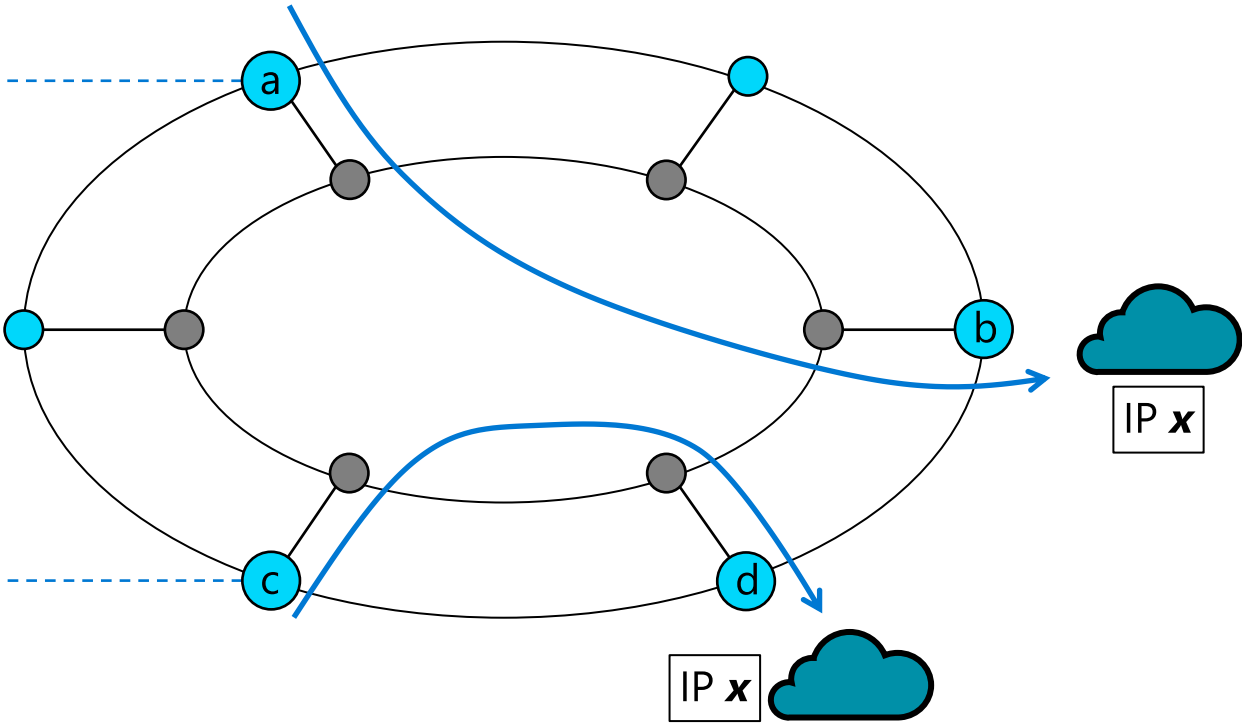# Traffic matrix measured by flow sampling

Per-minute

Measured TM
Src router
Dst site
Traffic class
Mbps

Internet & non-discretionary traffic

Linear regression predictor

Predicted TM

+

Discretionary traffic

Requested TM

Measuring anycast traffic

Switch : **a**

Egress interface

BGP nexthop : **b**

...

Flow records

Switch : **c**

Egress interface

BGP nexthop : **d**

...

IP **x**

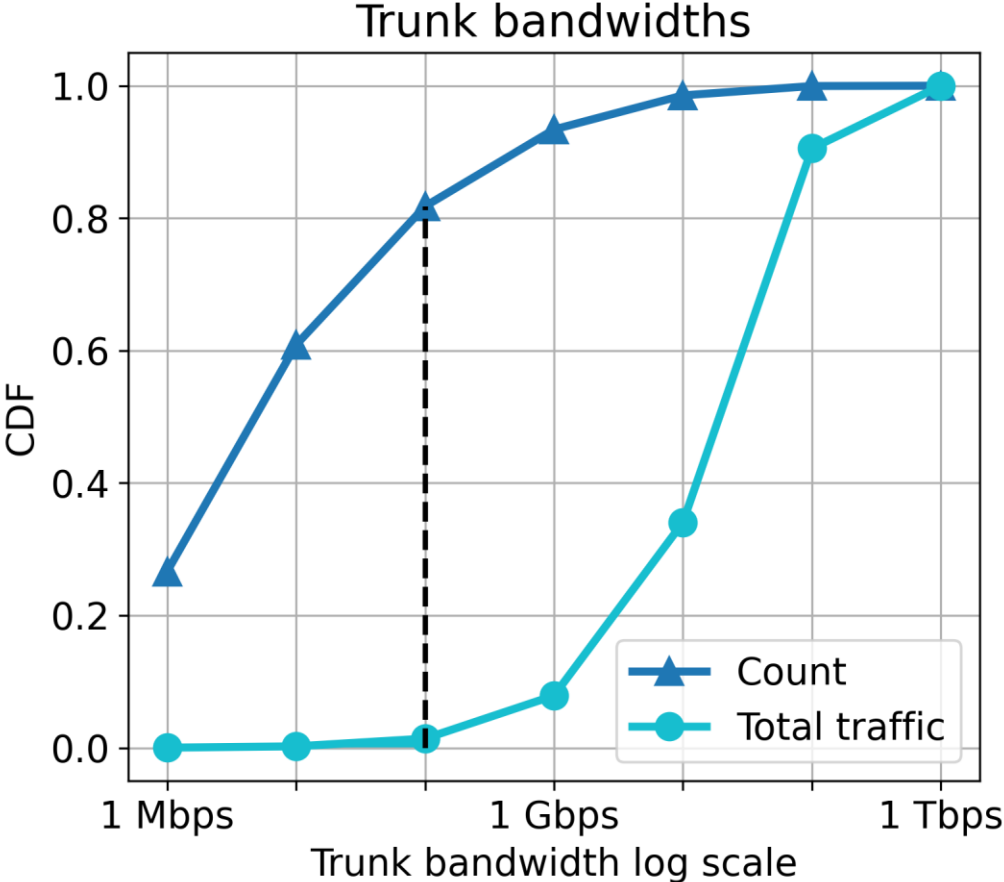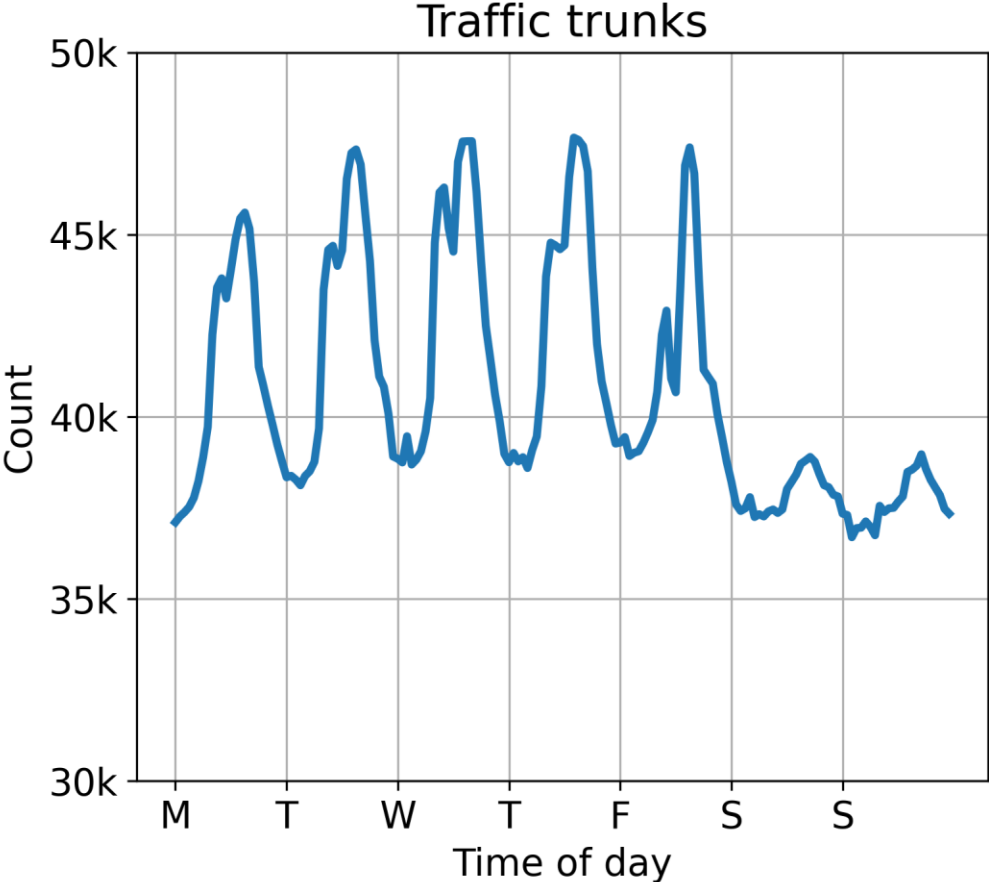IP **x**

Traffic matrix error correction
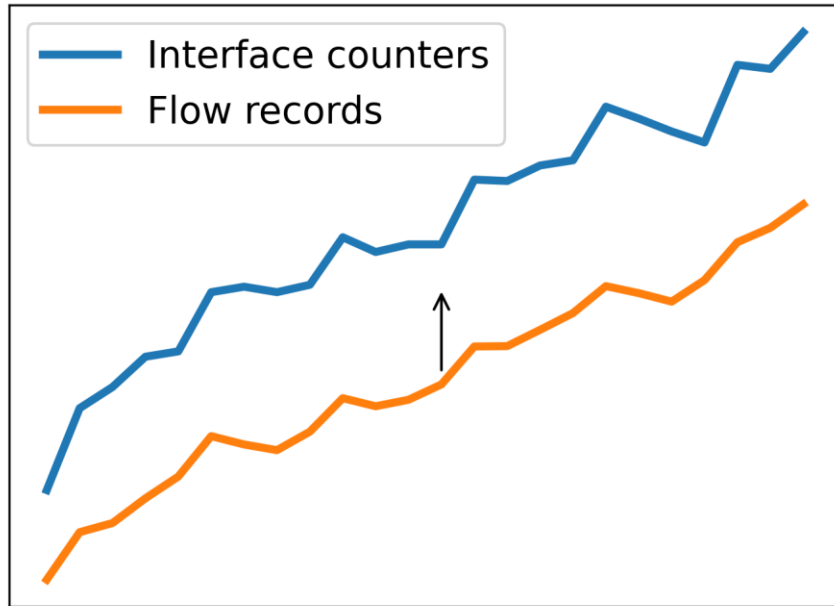
Interface counters correct for underprediction from data loss

$f(TM)$
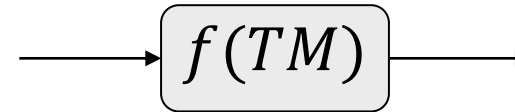
TM operations correct for firmware limitations
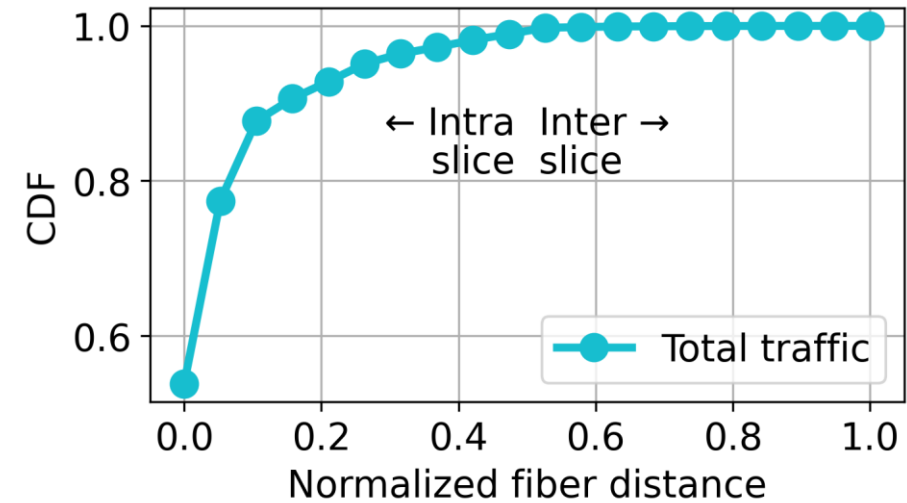
OneWAN traffic matrix characteristics

Traffic trunks

Trunk bandwidths

1. Traffic steering

2. Local repair

3. TE optimization

4. Traffic matrices

5. **Traffic migration**

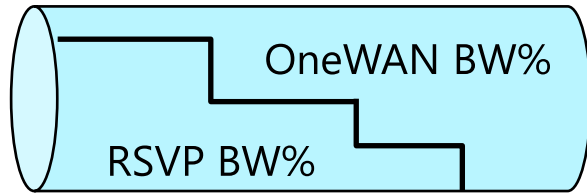# Traffic migration

BlastShield slices

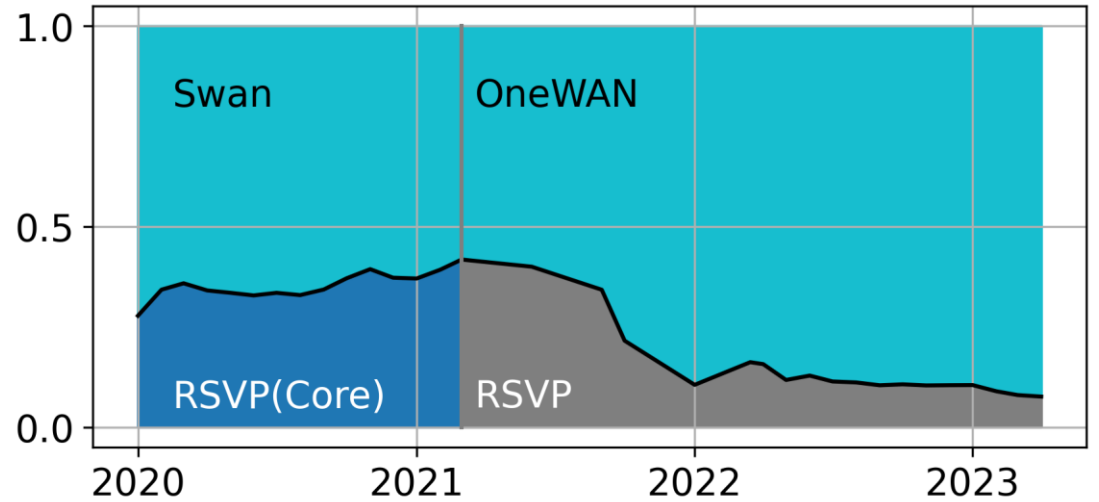New BlastShield slices for OneWAN controllers.

Migrate intra-slice before inter-slice.

Traffic migration

Step down RSVP bandwidth reservation till zero.

Traffic migration from RSVP to OneWAN.

Summary

Inter-DC flow

Internet flow

OneWAN

Routes injected by controller

Egress site label		TE path label		BGP NH node SID		IP
BGP NH node SID		BGP NH node SID		IP		
IP		IP				

BGP   *Traffic steering*   *Traffic engineering*   *Segment routing*   *BGP*

Swan   OneWAN

RSVP(Core)   RSVP