CHI-NOG: Multipath Traffic Engineering (MPTE)

Colby Barth, Senior Distinguished Technologist, HPE

May 2026

Agenda

- **Why MPTE and the MPTE Concept**
- **MPTE Architecture: DAG Tunnels and Junctions**
- **MPTE Signaling options: RSVP (and BGP)**
- **Protection: Facility or Detours?**
- **Closing: MPTE Differentiation**

Why MultiPath Traffic Engineering (MPTE)

Networks are Increasingly Built with Inherit Multipathing Attributes

- Moving from single plane to multiplane via ‘N’ redundant chassis and/or CLOS topologies
- ECMP, while present, does not address all networking challenges e.g. Bandwidth Management

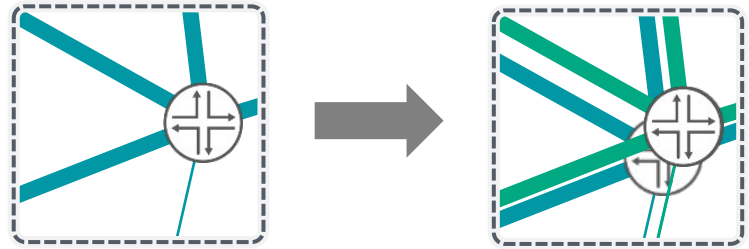
How do we add TE to multi-path (WANs)?

How do we add multi-path to TE (DC fabrics)?

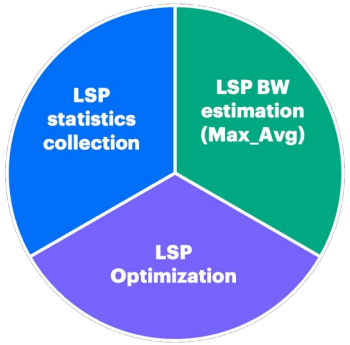
Next Evolution in Bandwidth Engineering

- Bandwidth engineering (e.g. Auto-bandwidth) has arguably been one of the best MPLS tools we have in our solution set
- Steady investment/evolution of auto-bandwidth for more than a decade

MPTE is the next [BIG] step in that evolution



From single plane to multiple planes



LSP Statistics Collection — Improvements in real-time TE tunnel stats collection

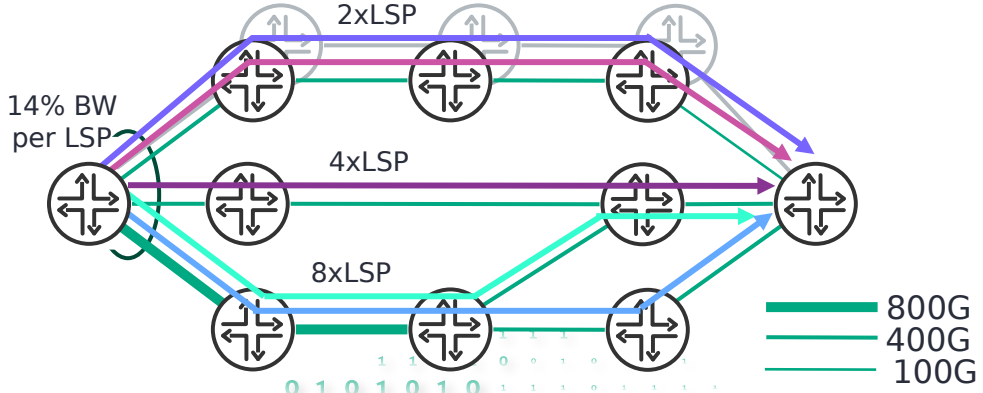
LSP BW Estimation — Introduction of novel, on-box and off-box, BW prediction techniques

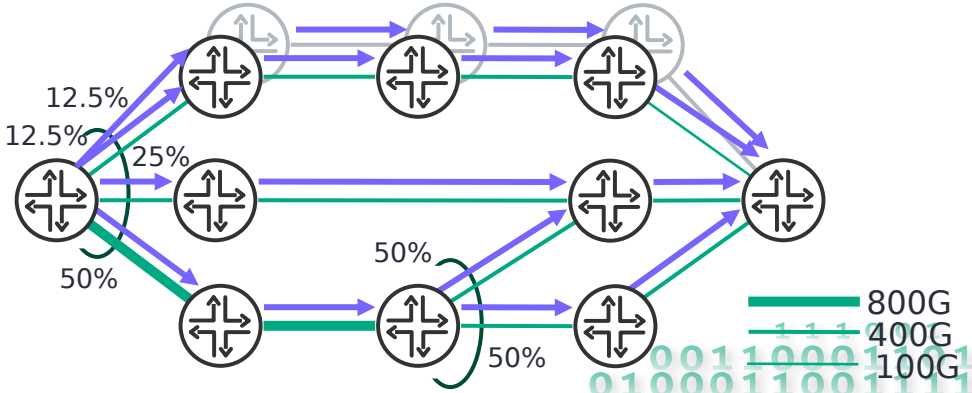
LSP Optimization — In-place BW updates, Resource notifications, **DAG based computation**

Evolution of Bandwidth Engineering

The Concept: Conventional TE vs. DAG based MPTE

A **Directed Acyclic Graph (DAG)** is a graph made of nodes connected by one-way links, where following the direction of the links can never lead back to the same node.





Single Path Computation

Spawn & Maintain 14 individual TE Tunnels

Equal cost bandwidth splitting
ECMP across all ingress tunnels

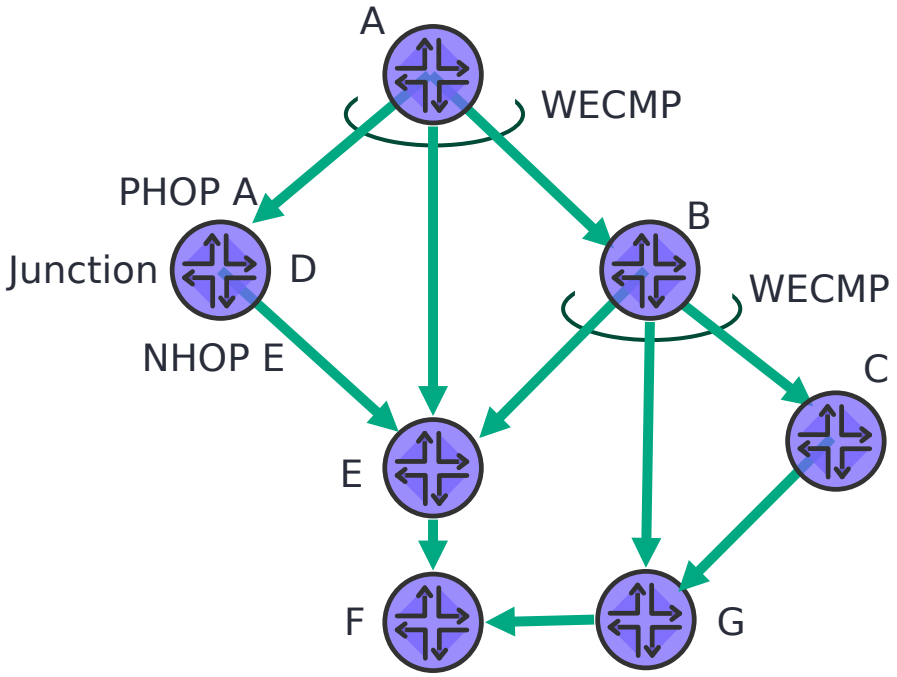
DAG based Computation

Spawn & Maintain 1 MPTE DAG Tunnel

DAG may split/combine at each junction
Optimal LB ratios maintained

MPTE Architecture

- MPTE combines classical TE with Unequal Cost Load Balancing.
- MPTE changes the computation result from a **single constrained path to a constrained multipath DAG** between ingress (multiple) and egress (multiple)
- The **DAG is defined** by an *ingress set, an egress set, a metric objective, optional **slack**, and a set of path constraints*.
  - ✓ The slack parameter is important because it allows near-optimal paths to be admitted even if they are not strictly equal cost.
- Each participating node is modeled as a **Junction** that maintains a set of previous hops (**PHOPs**), next hops (**NHOPs**), bandwidth information, and load-sharing weights.
  - ✓ This turns the TE object into a per-node forwarding policy across a constrained multipath structure.
- **MPTE is transport agnostic** and can be applied to MPLS, Segment Routing, and native IP forwarding models.



MPTED Tunnels and Junctions

MPTED tunnel

— TE construct that contains a constrained set of paths representing an optimized Directed Acyclic Graph (DAG) from one or more ingresses to one or more egresses

— The paths that make up an MPTED tunnel traverse a set of junction nodes

MPTED junction

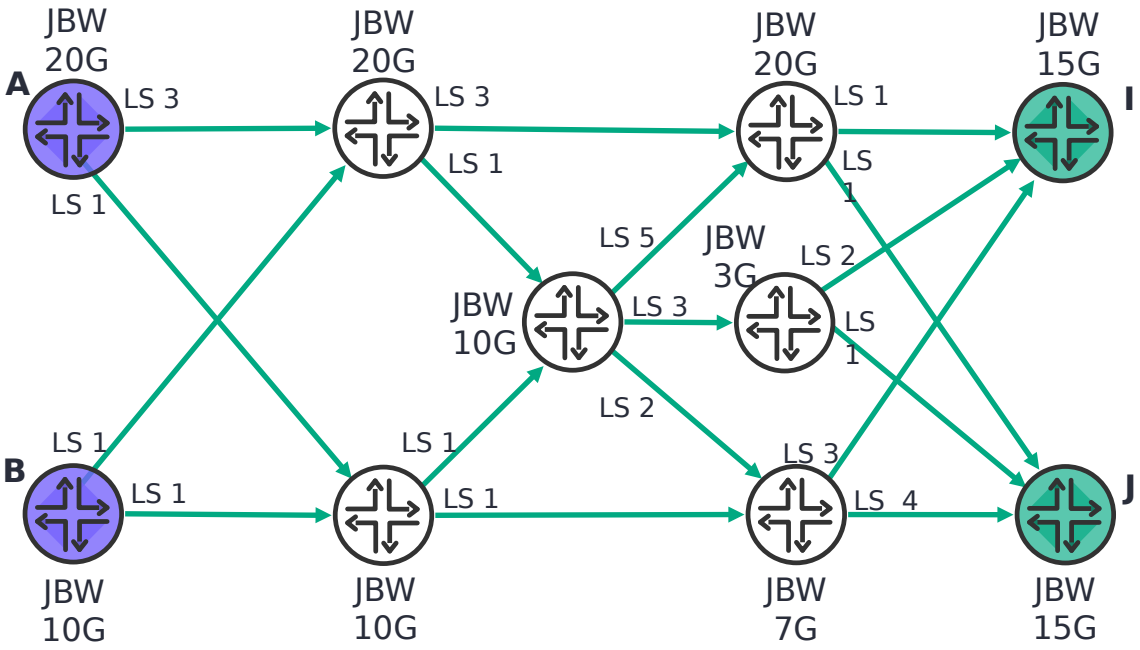
— TE construct associated with the MPTED tunnel at each junction node and constitutes a set of previous-hops (JCT-PHOPs) and a set of next-hops (JCT-NHOPs) over which traffic is load-balanced in a weighted fashion

Provisioning

— Provisioning an MPTED tunnel involves provisioning the control and forwarding plane state associated with the MPTED junction at each junction node

MPTED Tunnel: Tun-West-To-East [30G]

- ✓ Ingresses – {A [20G], B [10G]}; Egresses – {I, J}
- ✓ Constraint: Include Green (Resource Affinity)
- ✓ Optimization Objective: TE metric
- ✓ Node 1 - tunnel originator, DAG computer, and signaling source



JBW: Junction Bandwidth
LS: Load Share

RSVP Signaling: Messages for Junction Management

Optimized Signaling Procedures

- 💡 Minimize “Refresh” message processing
- 💡 Avoid unnecessary signaling adj failures
- 💡 Minimize “Trigger” message processing
- 💡 Minimize the number of signaling notifications triggered when a link fails/degrades
- 💡 Maintain Ordered Control/Programming

(Signaling) Source to Junction (S2J) Messages

- JunctionCreate - RSVP MPTED Path
- JunctionUpdate - RSVP MPTED Path
- JunctionDelete - RSVP MPTED PathTear (with or without CONDITIONS object)

Junction to Source (J2S) Messages

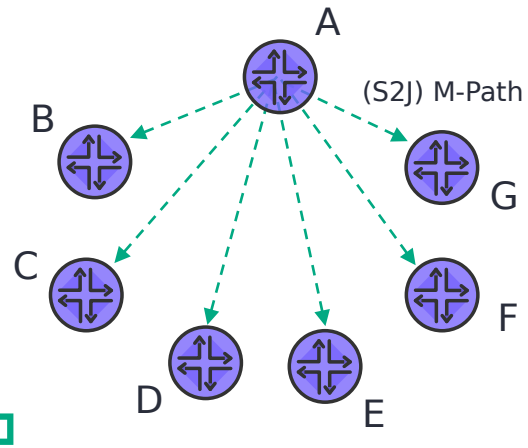
- JunctionNotify - RSVP MPTED Notify
- ResourceNotify - RSVP Rsrc Notify

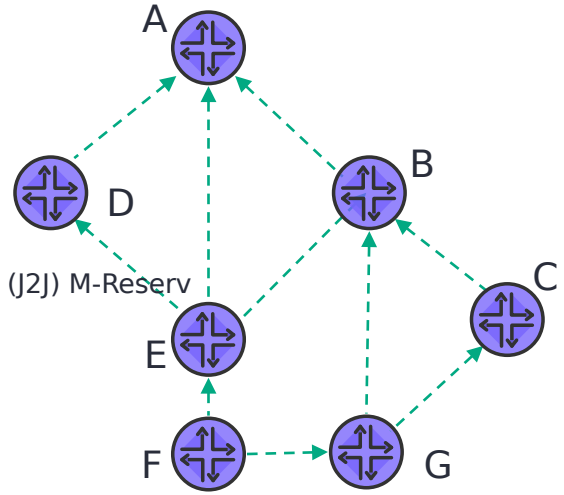
Junction to Junction (J2J) Messages

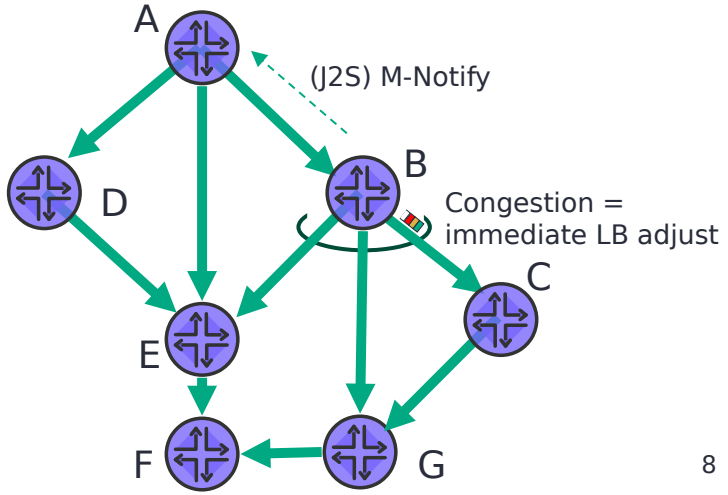
- Upstream (J2JU) Messages
  - JunctionNextHopReservation - RSVP MPTED Resv
  - JunctionDown - RSVP MPTED Notify
- Downstream (J2JD) Messages
  - JunctionDelete (Conditional) - RSVP MPTED PathTear (with CONDITIONS object)
  - JunctionNotReady - RSVP MPTED ResvErr

RSVP Signaling: Set-up and Maintenance

- Transport problems are inherently one of constrained multipath across unequal paths
  - ✓ The RSVP-based MPTE model provisions the DAG as a single MPTED tunnel rather than as many separately signaled point-to-point member LSPs.
  - ✓ The signaling source distributes junction descriptors that define local PHOP, NHOP, bandwidth, and load-share state at each node.
  - ✓ **M-Path** messages create or update junction state, while **M-Resv** messages establish upstream reservation and ordered readiness.
- If the DAG shape does not change, updates can be localized to junction bandwidth and NHOP weights, which aligns well with auto-bandwidth behavior.
  - ✓ RSVP notification mechanisms such as **ResourceNotify** can report degraded or unavailable branches and drive local and/or upstream reaction.





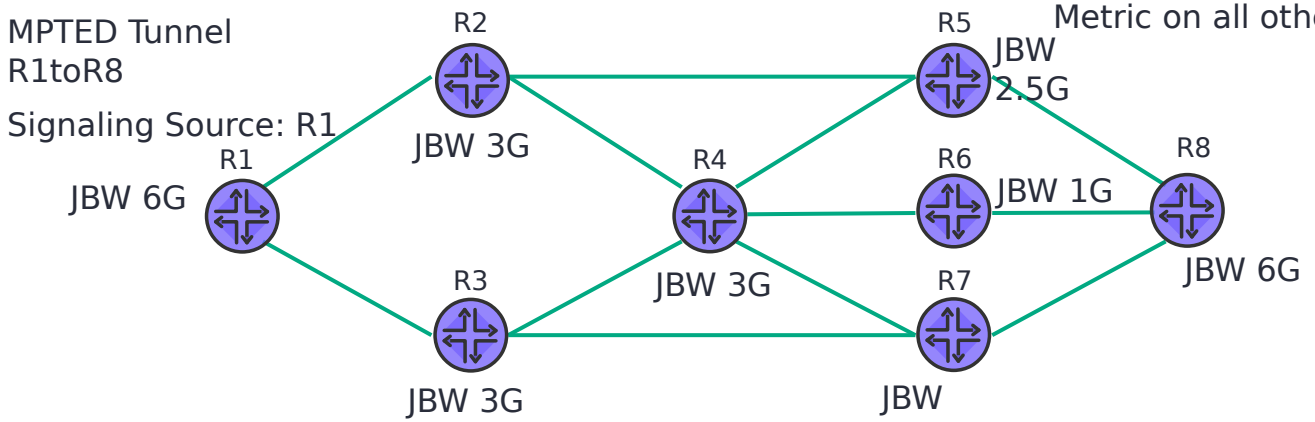


Example: Initial RSVP Signaling Setup Sequence

Metric on links 2-4, 4-5,3-4 and 4-7 = 10

Metric on 4-6 and 6-8 = 15

Metric on all other links = 20



1 **Initiation of setup sequence on MPTED tunnel signaling source, R1:**

- R1 sends an M-Path message to each junction node (R2, R3, R4, R5, ... R8)
- R1 processes the ingress JUNCTION, constructs a JSB, and waits for an M-Resv message to arrive from each JCT-NHOP (R2 and R3).

2 **M-Path message processing on transit junction nodes (R2, R3, R4, R5, R6, R7):**

- Each transit junction node processes the JUNCTION, constructs a JSB, and waits for an M-Resv message to arrive from each JCT-NHOP.

3 **M-Path message processing on egress junction node, R8:**

- R8 processes the JUNCTION and constructs a JSB.
- R8 sends an M-Resv message to each JCT-PHOP (R5, R6, and R7)
- R8 sends an M-Notify message to R1, indicating that the junction processing is complete at R8.

4 **M-Resv message processing on transit junction nodes (R2, R3, R4, R5, R6, R7):**

- Each transit junction node waits until M-Resv messages are received from all available JCT-NHOPs and then:
  - Updates BW reservation on TE-links.
  - Allocates a label for each JCT-PHOP and programs the corresponding labeled route.
  - Sends an M-Resv message to each JCT-PHOP with the corresponding allocated label.
  - Sends an M-Notify message to R1, indicating that the junction processing is complete on the node.

5 **M-Resv message processing on ingress junction node, R1:**

- R1 waits until M-Resv messages are received from all JCT-NHOPs (R2 and R3) and then:
  - Updates BW reservation on TE-links.
  - Programs a route for the MPTED tunnel.
  - Notifies the signaling source (itself) that the junction processing is complete on the ingress node.

6 **M-Notify message processing on the signaling source:**

- The signaling source (R1) considers the setup sequence complete when confirmation of junction provisioning is received from all junctions.

Protection

Classical TE provides mechanisms to build protection paths for an existing path to protect against link or node failures

- **Detour protection**: 1-to-1 LSP protection
- **Facility protection**: Link and node protection
- **Constraint and optimization objective inheritance**: Bringing additional granularity to facility protection

MPTE reduces the need to build dedicated protection paths

- Load-balancing weights need only local adjustment upon a failure
- There may be cases where a node has only a single next hop, or all next hops share a common failure mode
- **MPTE will follow a ‘detour’ methodology thus providing per DAG protection**

MPTE is Real

```
root@R1_re# set protocols mpted tunnel 1to9 ?
Possible completions:
> admin-group           Administrative group policy
> admin-group-extended  Extended administrative group policy
  compute-engine        Compute engine address
  compute-only          Compute DAG only, do not signal tunnel
  description           Text description of tunnel
  disable               Disable tunnel
+ egress                Tunnel egress address
> hop                   Hop constraints configuration
> ingress               Tunnel ingress configuration
  max-jcts              Maximum number of junctions (1..255)
  max-nhops             Maximum number of nexthops per junction (1..255)
  min-nhops             Minimum number of nexthops per junction (1..255)
> optimize              Optimization configuration
  originator            Tunnel originator address
  priority              Preemption priorities
  signaling-source      Signaling source address
  signaling-type        Signaling protocol type
  type                  Tunnel type
```

Per hop min-bandwidth

```
root@R1_re# set protocols mpted tunnel 1to9 hop ?
Possible completions:
  minimum-bandwidth    Minimum bandwidth per hop (bps)
```

Control the span of the DAG

Multiple optimization objectives

```
root@R1_re# set protocols mpted tunnel 1to9 optimize ?
Possible completions:
  delay                Delay metric optimization
  igp                  IGP metric optimization
  metric-margin        Metric margin for path selection (0..4294967295)
  te                   Traffic engineering metric optimization
```

MPTE References – IETF Drafts

Workgroup: TEAS WG
Internet-Draft: draft-kompella-teas-mpte-01
Published: 7 July 2025
Intended Status: Standards Track
Expires: 8 January 2026

K. Kompella
Juniper Networks
L. Jalil
Verizon
M. Khaddam
Cox Communications
A. Smith
Oracle Cloud Infrastructure

Multipath Traffic Engineering

Workgroup: LSR WG
Internet-Draft: draft-kompella-lsr-mptecap-00
Updates: 5073 (if approved)
Published: 7 July 2025
Intended Status: Standards Track
Expires: 8 January 2026

K. Kompella
Juniper Networks

Multipath Traffic Engineering Capabilities

TEAS WG
Internet-Draft
Intended status: Standards Track
Expires: 8 January 2026

K. Kompella
V. P. Beeram
C. Ramachandran
Juniper Networks
7 July 2025

RSVP-TE Extensions for Multipath Traffic Engineered Directed Acyclic Graph Tunnels
draft-kbr-teas-mptersvp-01

TEAS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 January 2026

V. P. Beeram
K. Kompella
Juniper Networks
7 July 2025

A YANG Data Model for Multipath Traffic Engineering Directed Acyclic Graph (MPTED) Tunnels and Junctions
draft-beeram-teas-yang-mpted-00

PCEP Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 January 2026

V. P. Beeram
K. Kompella
Juniper Networks
7 July 2025

Path Computation Element Communication Protocol (PCEP) Extensions for Multipath Traffic Engineered Directed Acyclic Graph (MPTED) Tunnels
draft-beeram-pce-pcep-mpted-00

Closing

- **MPTE is a TE construct**, not a specific control-plane or data-plane
- MPTE **decouples multipath forwarding from the equal-cost assumption**, this is useful for AI transport paths that are often near-optimal rather than identical
- **Supports bandwidth-aware load sharing** across NHOPs instead of uniform per-flow hashing
- **Preserves TE-style constraints and optimization goals**
  - ✓ A generalized mechanism for supporting various deployment models (multi-plane, 3-5-7-stage Fat Tree, Dragon Fly and meshed DCIs)
- Enables **graceful degradation**: a partially impaired branch can be down-weighted instead of forcing immediate full reroute or tunnel teardown
  - ✓ This is better aligned with AI workloads, where transient degradation is common and binary failure handling is often too coarse
- In practical terms, **MPTE combines the control of traffic engineering with the resiliency of engineered multipath**

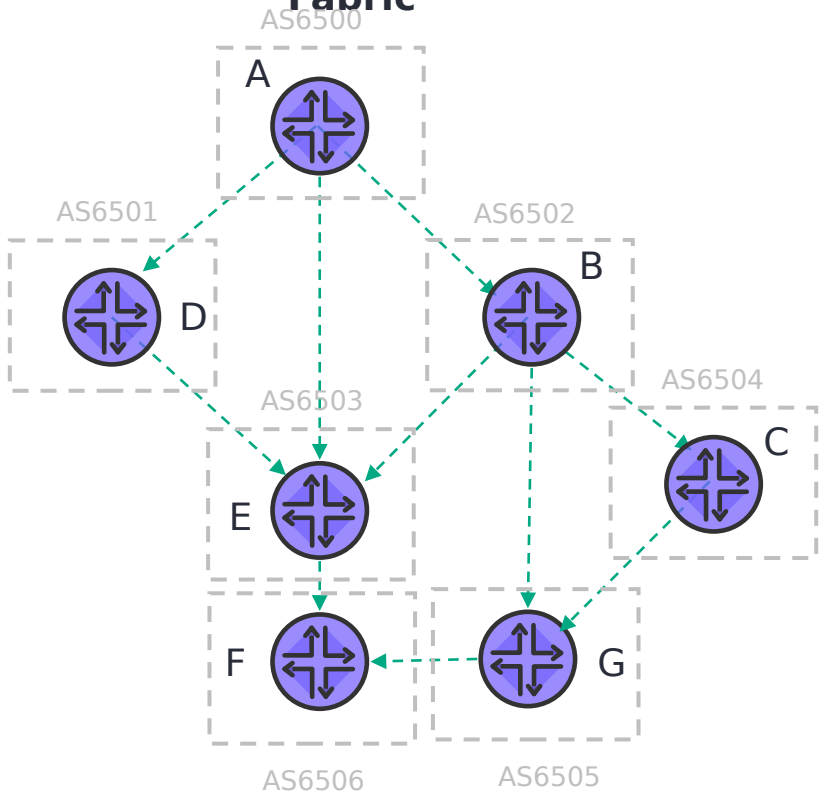
Thank You

Colby Barth, Jonathan.barth@hpe.com

Signaling MPTE with BGP

- MPTE can be signaled using Junction State routes that describe the DAG state to participating nodes.
- A dedicated **MPTE SAFI** identifies these routes as MPTE signaling information.
- The signaling state carries the MPTED computing entity address, MPTED identifier, version, tunnel type, junction node address, originator address, and junction bandwidth.
- PHOP and NHOP information is encoded using the BGP **Tunnel Encapsulation Attribute** and associated sub-TLVs.
- Each node imports the junction state relevant to it and installs the corresponding forwarding behavior locally.
- **Ordered control** is important because it delays full forwarding installation until downstream readiness is confirmed, reducing the risk of blackholing traffic.
- This model is operationally attractive because it aligns with the BGP-only control infrastructure already common in DC and AI DC/DCI fabrics.

Inline BGP MPTE signaling on BGP-Only Fabric*

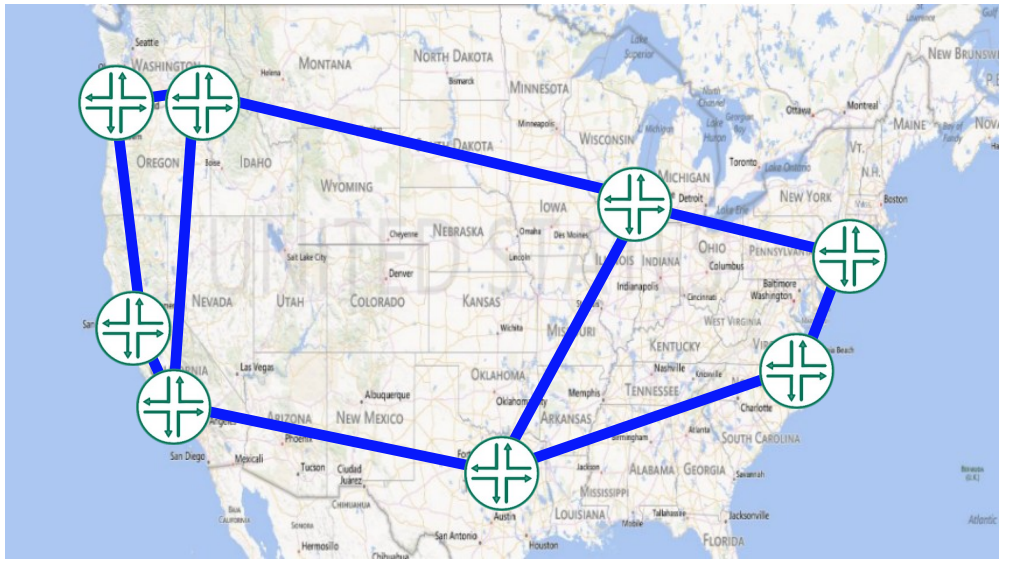


*** Route Reflector models also relevant**

Applicability of MPTE

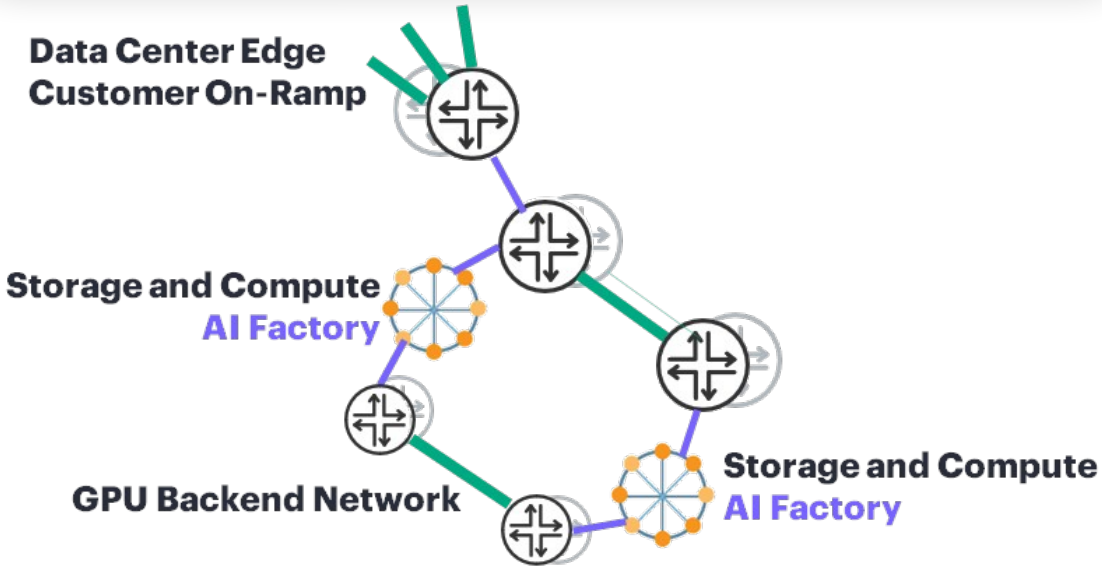
Wide Area Networks

- ✓ RSVP/MPLS incremental evolution
- ✓ Optimize to new/evolving topologies
- ✓ Improved scale and performance
- ✓ Incremental deployment



Data Center InterConnect (DCI)

- ✓ Adapt to changing traffic patterns of AI/ML workloads
- ✓ Transition to end-to-end IP data-planes
- ✓ Either RSVP or BGP-only control-planes



Data Center (DC) BGP/IP-only fabrics

- ✓ Improve AI/ML workload performance
- ✓ BGP-only control-plane and IP-only data-planes