SEAMLESS MPLS

Rafał Szarecki

slide deck by Kireeti Kompella

MPLS AND PACKET TRANSPORT

Packet transport is fundamental to the notion of a “Next Generation Network”

- The “next generation” aspect is the shift away from transporting bits over TDM links to moving packets over non-TDM links, usually Ethernet

There is still some confusion over what the underlying infrastructure should be

- PBT, PBB-TE, PBB, T-MPLS, MPLS-TP, …
- Why not just plain MPLS?

RECAP OF THE PURPLE LINE ARGUMENT

Last year, I suggested the use of MPLS for this infrastructure, incorporating the notions of:

- A “Transport Router” optimized for MPLS switching
- Optical integration
- Organizational integration

The reasons included:

- Reduction/simplification of layers for packet transport
- Reduction/simplification of features for transport
- Reduction of control planes (no duplication)
- Reduction/simplification of the number of components in transport (devices, transponders)

"COST-PER-BIT" AND "VALUE-PER-BIT"

The previous arguments show how to reduce CapEx and OpEx, which is a Good Thing

- This addresses the "cost-per-bit" aspect of NGNs

Others in Juniper talked about "value-per-bit"

- the idea of a *service* infrastructure that allows the deployment of new value-added services

Question: is this completely orthogonal to the choice of packet transport infrastructure?

- This talk will attempt to answer this

EXPLORING THE ENTIRE NETWORK ARCHITECTURE

Let's look at the whole network, which can be partitioned into …

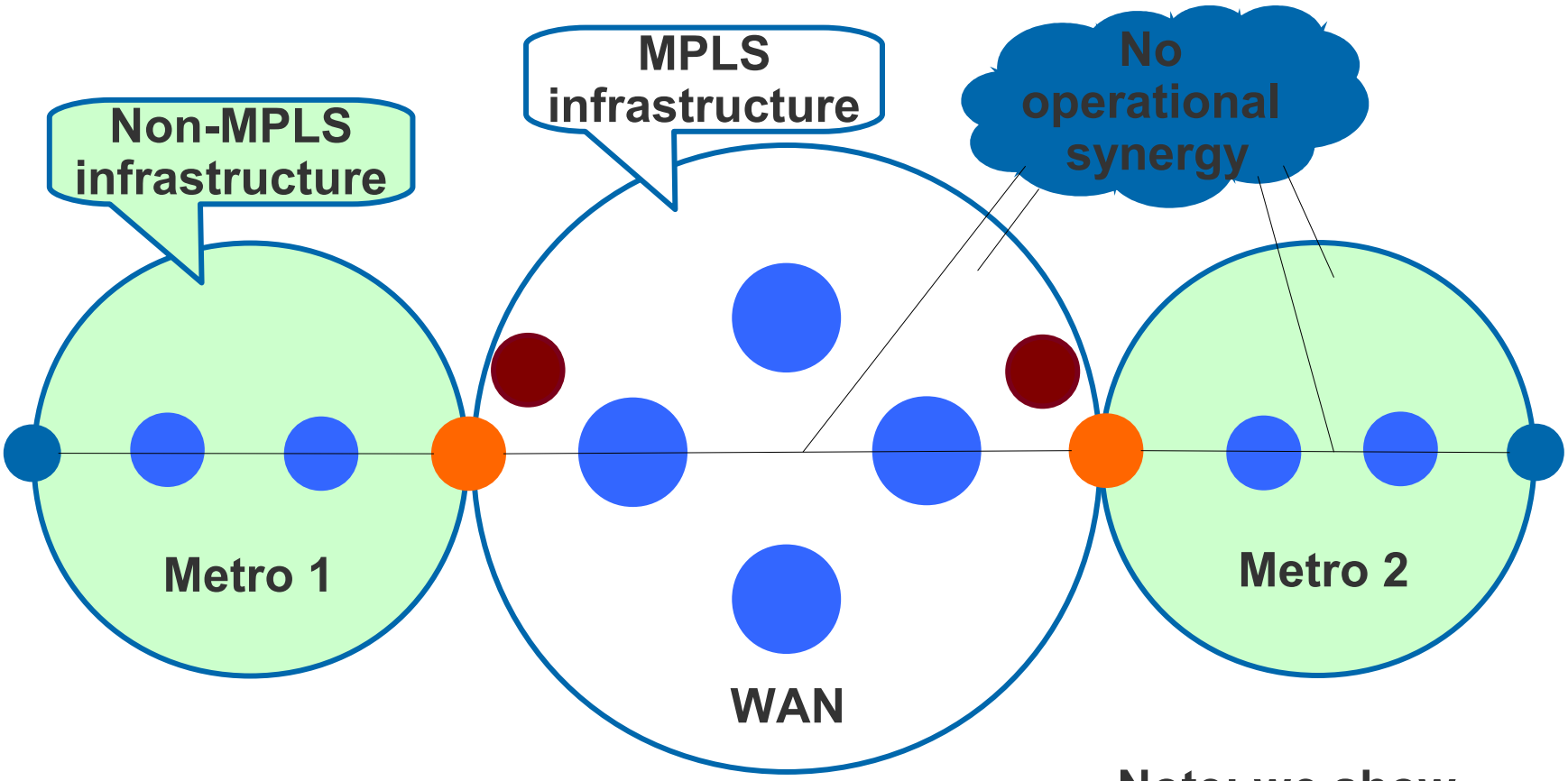
- Metro regions (including access); and
- WAN regions which connect up all the metros

… and component devices:

- Access nodes – where packets enter/leave the network
- Transport nodes – basically, packet movers
- Service nodes – where services are delivered
- Service helpers – these enable or enhance services

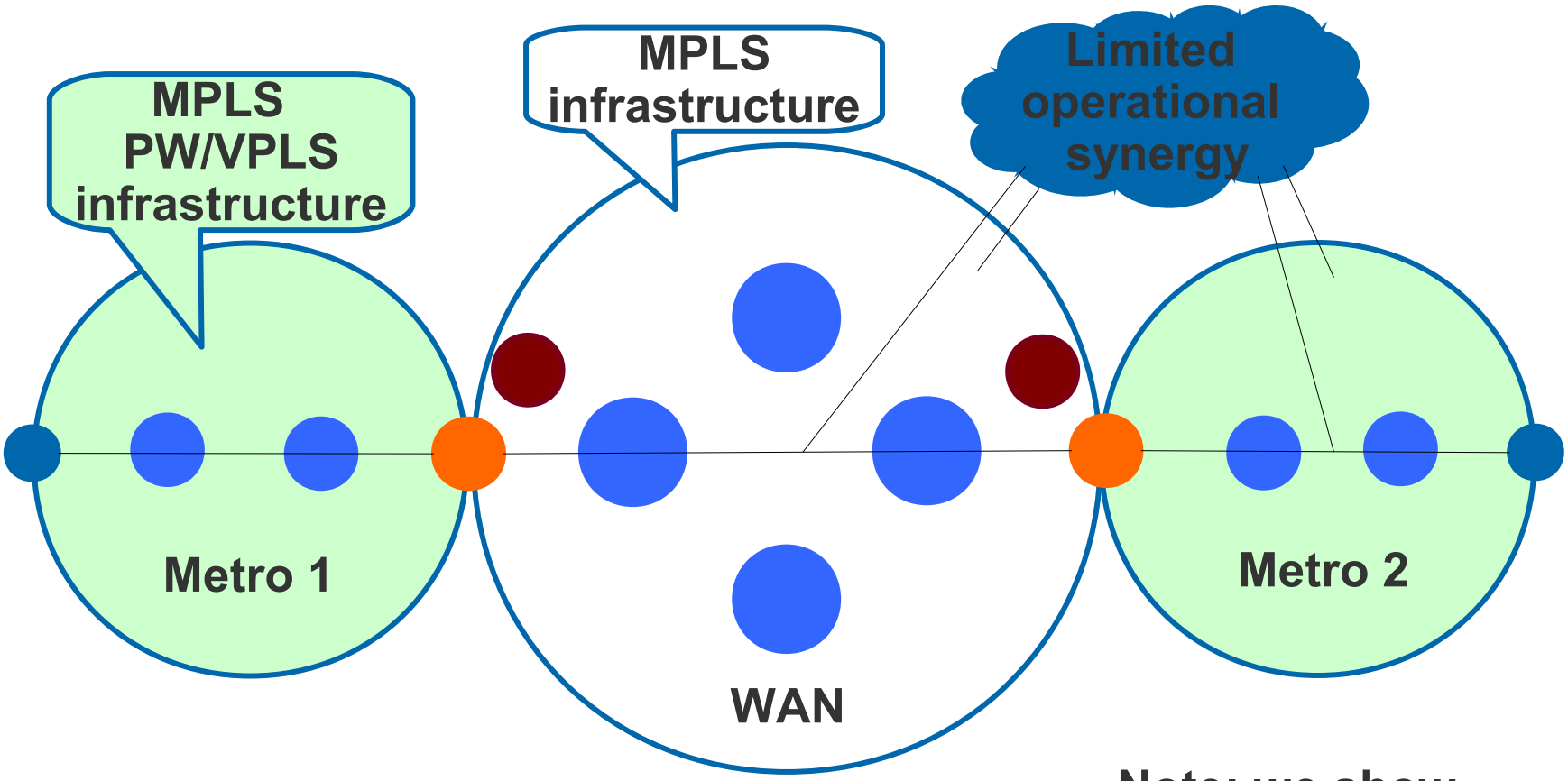
This high-level view encompasses residential and business as well as fixed and mobile subscribers

NETWORK ARCHITECTURE, FIRST TRY



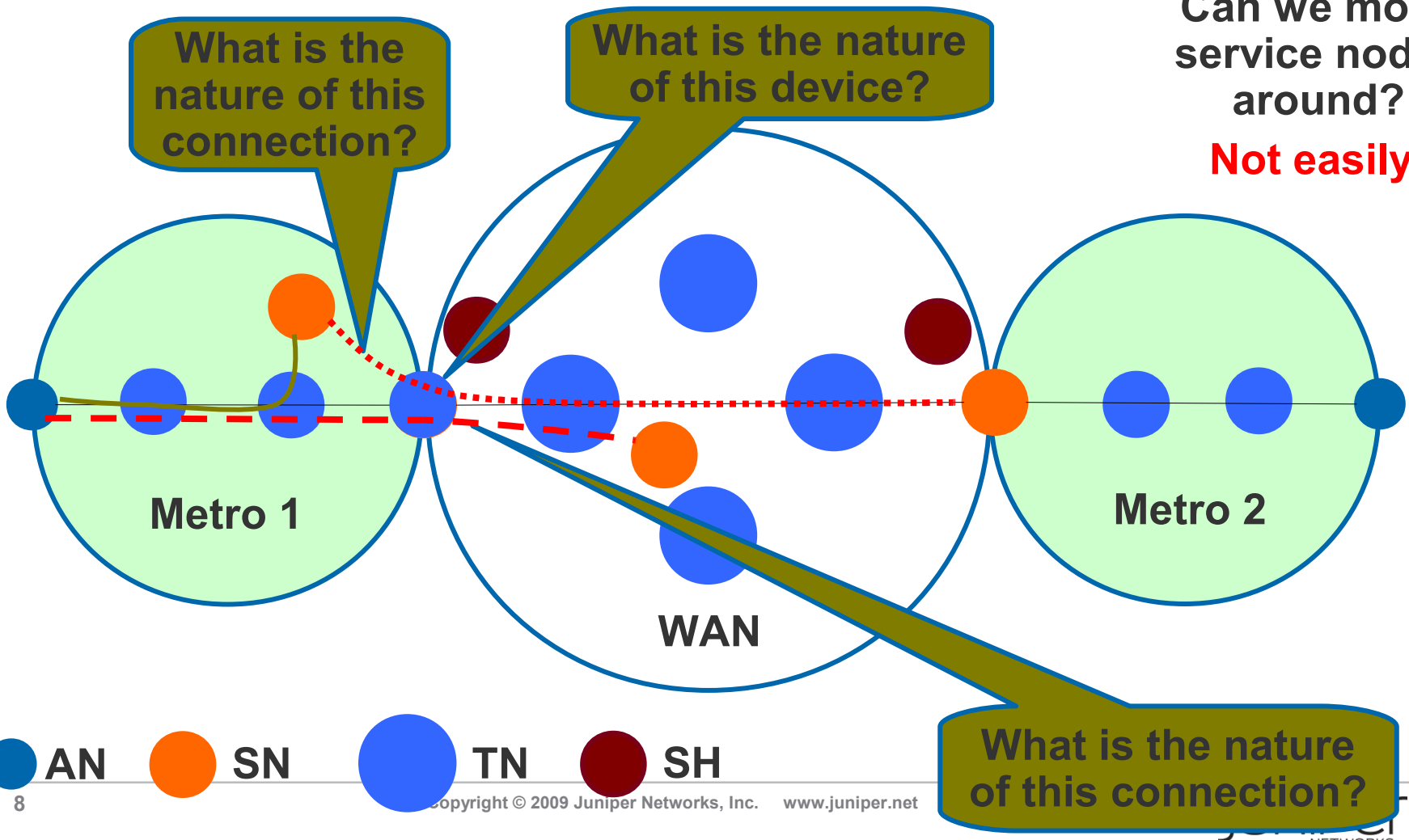
Note: we show *connections* rather than *links*

NETWORK ARCHITECTURE, FIRST AND HALF TRY



Note: we show *connections* rather than *links*

NETWORK ARCHITECTURE, FIRST TRY



WHAT IS SEAMLESS MPLS?

seam·less *adjective* *1.having no seams.* *2. smoothly continuous or uniform in quality; combined in an inconspicuous way*

End-to-end MPLS: from the moment a customer packet enters the network till it exits the network

- no breaks, no discontinuities; uniform, “converged”
- whether the customer is residential or business, fixed or mobile, commodity bit-pipe or deeply service- oriented, Layer 2 or Layer 3 or even Layer 7

BENEFITS

Convergence – since the network is uniform

True service freedom

- Deploy services when you want, where you want
- Bring up new services quickly and easily, and move them around as their requirements evolve

The network exists to enable services

- Too often, network architecture dictates what can be offered (and how), rather than services dictating
- Services should determine connectivity paradigms, quality of experience and resilience requirements
- Services are why your customers will stay

DECOUPLING ARCHITECTURES

Ultimately, a Service Provider needs to provide services (even if it is just basic connectivity)

- The network is a means to an end, not the end itself

Service architecture defines where and how a service is delivered, and the interaction of service nodes and service helpers to enable the service

Network architecture provides the underlying connectivity functions (QoS, CAC, FRR, …) to make each service as effective as possible

These architectures need to be as decoupled and independently managed as feasible

PLACEMENT OF SERVICE DELIVERY POINTS

Today, services are most often delivered at the boundary of metro and core: at the PoP

- Mobile services are even more centralized

However, SPs are seeing increasing value of deploying services in a more distributed fashion

- Location-based services
- Local ad generation and insertion
- Caching of high-bandwidth services: video, P2P, …

But some services are best kept centralized

As services evolve, the most efficient placement of their delivery points can change

MOVEMENT OF NETWORK ELEMENTS

As the network itself evolves, network elements may have to move or be replaced

- Layers may collapse or expand
- Region boundaries may change

The requirements of geographic presence, scale, resilience, and/or new connectivity paradigms may also drive such change

- Such change should be possible without affecting services and service delivery

GOALS FOR SEAMLESS MPLS

Very large scale: from < 1000 nodes today to 10 to 100 thousand nodes in a single MPLS network

All-encompassing: access, metro, core

Robust: protocols, devices, OAM

Resilient: 50 msec service restoration

Service flexibility

- The network architecture to achieve the above requirements **must not** constrain services in any way

NETWORK ARCHITECTURE, SECOND TRY

Divide the network into “regions” (all with MPLS)

Establish connectivity within regions

- IP connectivity for control plane, service helpers, management
- MPLS connectivity for all customer packets

Establish inter-region connectivity

- Via “Border Nodes” – variant of TNs

Build “transport” PWs to get packets from ANs to SNs and vice versa

Build connectivity between SNs of various flavors, depending on the service

NETWORK ARCHITECTURE

A region can be an IGP instance or area or an AS

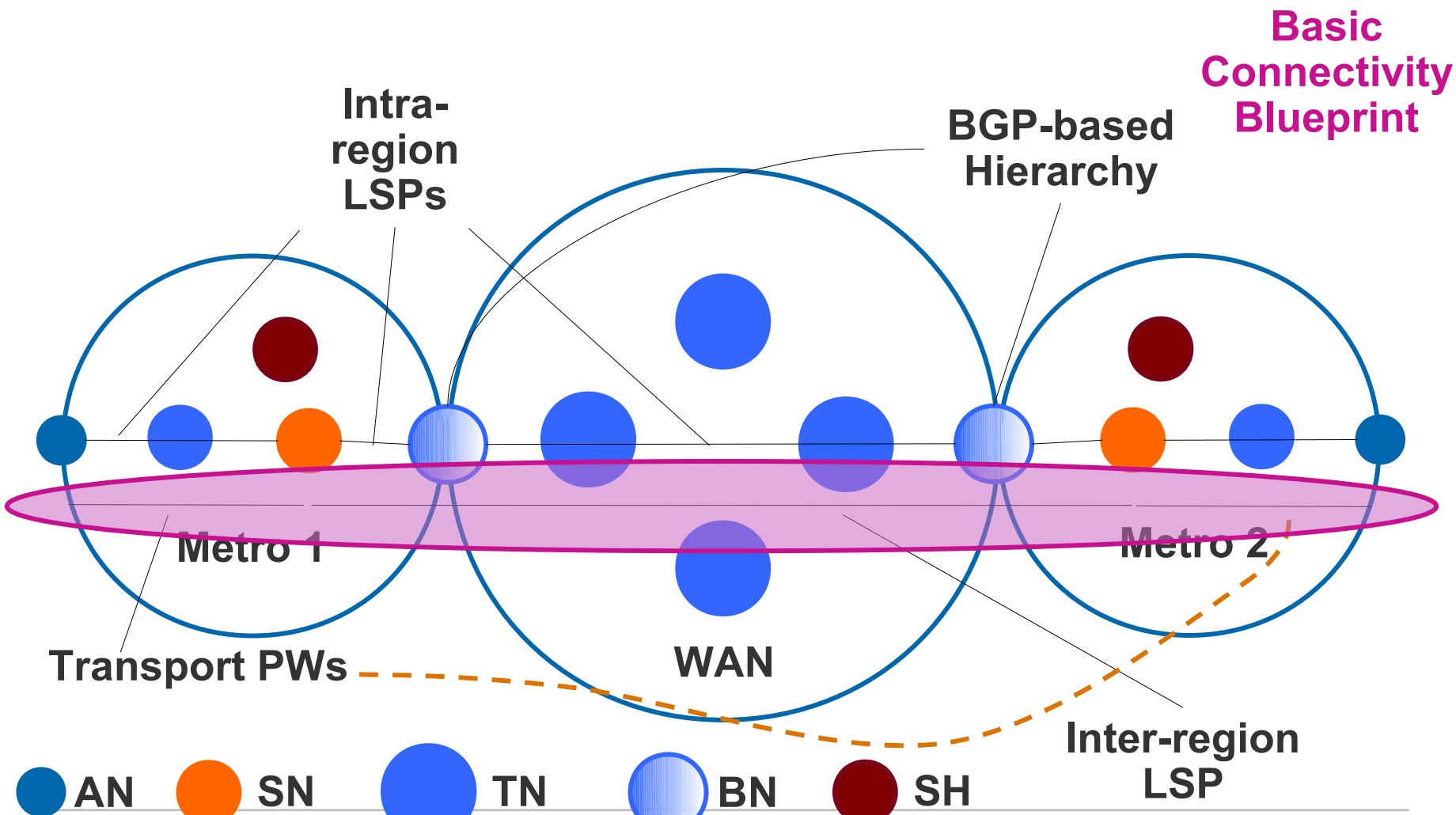
Each region is responsible for connectivity (both IP and MPLS) within the region

Each region can independently decide whether to run LDP or RSVP-TE or even LDP-over-RSVP

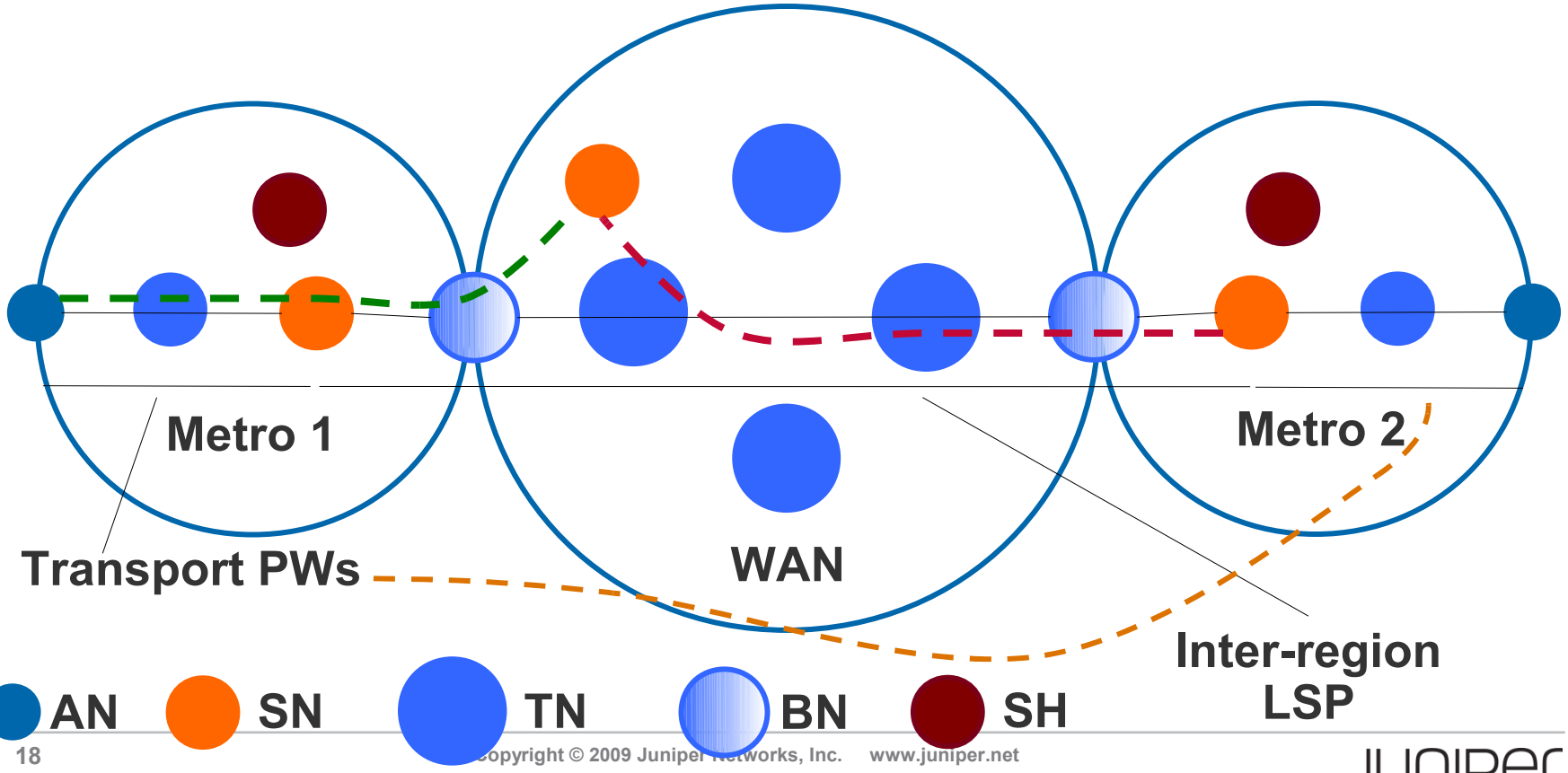
Region border nodes are responsible for inter-region connectivity

- For IP, this is done by prefix aggregation
- For MPLS, this is done by LSP hierarchy based on “labeled BGP” (RFC 3107)

NETWORK ARCHITECTURE



NETWORK ARCHITECTURE



ENABLING “VALUE-PER-BIT”

This network architecture, besides reducing cost via network and operational convergence, facilitates value-added services

- Incorporating “service helpers” into the architecture allows fine-grained control of services (UAC, policy, …)
- Allowing services nodes to be deployed as needed and easily moved as a service evolves encourages free thinking in the domain of services
- Using the “transport pseudowire” concept presents new paradigms of service provisioning, and even self-provisioning

SEAMLESS MPLS – MULTIPLE COMPONENTS

How to establish and maintain LSP in the network at scale

How to use MPLS as customer transport between AN and SN

How to provide Multicast service

How to ensure service HA

We will touch all of above – next PLNOGs

In this Look closer for basic connectivity of unicast LSP in single provider network.

SCALING

If we do this in a systematic way, we can have 10-100 thousand nodes in the network

- 10-20 thousand mobile access nodes (CSGs)
- 20-50 thousand DSLAMs and OLTs
- Several thousand transport and service nodes

Can this possibly scale???

Wait: we already have this many nodes!

- The current network scales, because we don't attempt to connect every node to every other node
- We implicitly have "connectivity blueprints" today, only the components are built out of different technologies

SCALING MPLS TRANSPORT LSPS

Using BGP as the inter-AS routing and label distribution protocol

- Fairly well established and deployed
- RFC 3107

Also using BGP as the inter-Area routing and label distribution protocol

- Seemingly obvious, but took a number of iterations before arriving at this solution

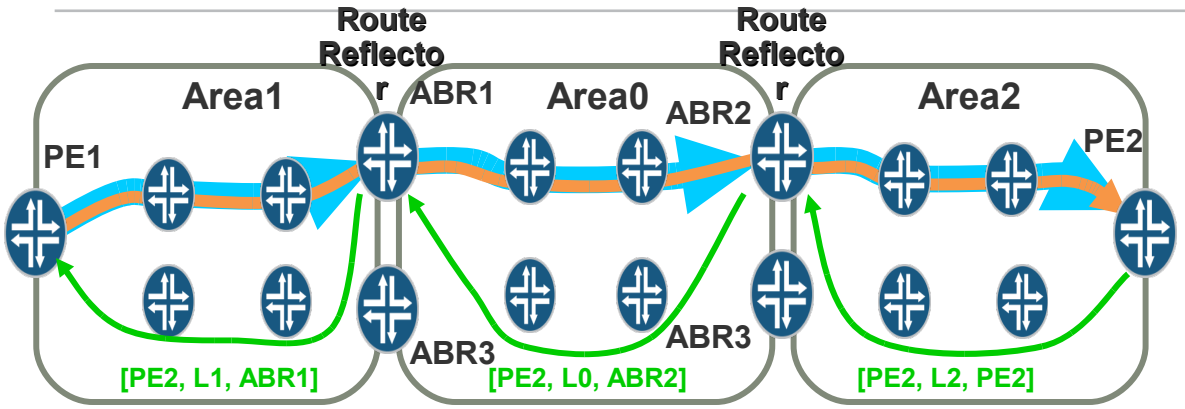
BGP as the inter-Area routing and label distribution protocol for unicast

- draft-leymann-mpls-seamless-mpls-00.txt, “Seamless MPLS Architecture”, N. Leymann, Editor, DT

BGP as the inter-Area protocol for inter-Area P2MP segmented LSPs for multicast

- Draft-raggarwa-mpls-seamless-mpls-multicast-00.txt

LABELED BGP (RFC3107) AS INTER-AREA ROUTING AND LABEL DISTRIBUTION PROTOCOL - EXAMPLE



Inter-area LSP

Intra-area LSP

Labeled BGP

We have three intra-area LSPs:

In Area 1, from PE1 to ABR1: use label L1'

In Area 0, from ABR1 to ABR2: use label L0'

In Area 2, from ABR2 to PE2: use label L2'

We will denote a BGP route to D with label L and Next-Hop N by [D, L, N].

(1) PE2 announces (via IBGP) to ABR2 in Area 2 a route [PE2, L2, PE2].

(2) When ABR2 receives this route, it does the following:

ABR2 resolves the Next-Hop PE2 to the Area 2 LSP to PE2, i.e., label L2'. Now ABR2 can reach PE2 via label stack <L2', L2>

ABR2 announces (via IBGP) to ABR1 in Area 0 a route [PE2, L0, ABR2].

ABR2 also installs LFIB state to swap <L0> to <L2', L2>.

(3) When ABR1 receives this route, it does the following:

ABR1 resolves the Next-Hop ABR2 to an Area 0 LSP to ABR2, i.e., label L0'. Now ABR1 can reach PE2 via label stack <L0', L0>.

ABR1 announces (via IBGP) to PE1 in Area 1 a route [PE2, L1, ABR1].

ABR1 also instantiates LFIB state to swap <L1> to <L0', L0>.

(4) When PE1 receives this route, it resolves the Next-Hop ABR1 to an Area 1 LSP to ABR1, i.e., label L1'.

(5) Now PE1 can reach PE2 via label stack <L1', L1>.

LABELED BGP (RFC3107) AS INTER-AREA ROUTING AND LABEL DISTRIBUTION PROTOCOL

Use BGP to distribute (routes + labels) across IGP area boundaries for LSPs that span multiple IGP areas within an AS

- Provides both routing and label information to establish inter-area LSPs

Area Border Routers (ABRs) act as BGP Route Reflectors, BUT also set BGP Next-Hop to “self”, AND create an MPLS forwarding state for inter-area LSPs

- Additional hierarchy of Route Reflectors may be used to further facilitate scaling
  - Although these would be “conventional” Route Reflectors
    - Unmodified propagation of Next-Hop; no creation of an MPLS forwarding state

LSPs that span multiple IGP areas within the same AS are carried within each such area over intra-area LSPs within that area

- Using LSP hierarchy
- Intra-area LSPs could be established via LDP or RSVP-TE

Benefits of labeled BGP (RFC3107) as inter-area routing and label distribution protocol

The scope of the label distribution protocol for establishing intra-area LSPs is confined to a single IGP area

- Different IGP areas within the same AS may use different (intra-area) label distribution protocols
  - E.g., RSVP-TE in Area 0, LDP in all other areas

No “leaking” of /32 routes for PEs across IGP area boundaries

Ps within each area do not maintain any state for inter-area LSP

- Ps within each area maintain state only for intra-area LSPs within that area
  - Applies to both control and data planes on Ps
- Only ABRs and PEs maintain state for inter-area LSPs

Support for LSP end-point (ABR) resiliency (more on this later…)

Common mechanism for both inter-area and inter-AS LSPs

- BGP (RFC3107) as a common routing and label distribution protocol

SCALING

There is a deep dive into how the network can be made to scale

- for fixed and mobile backhaul
- for IP connectivity (service helpers, management)
- for MPLS connectivity
- for unicast and for multicast services

… in the “MPLS in the access” tutorial

SERVICE RESTORATION

"Fast Re-route" (FRR) is a means to an end

- The end is fast service restoration

FRR has inherent limitations for the egress node, and sometimes for the last link

- However, where FRR is applicable, it is a good tool

Fast end-to-end service restoration *is* possible, at scale, and within 50-100ms

- Was presented in past PLNOG
- Key requirement: mechanisms for service state replication for all kinds of services
- Another key: local failure detection and repair

“CHANGE IS THE ONLY CONSTANT”

If you buy the idea of a “purple line” that separates the services and transport layers, then

- service requirements dictate network functions (the service layer is a client of the network/transport layer)
- each layer should be designed independently
- each layer should be managed independently
- each should have its own OAM, serving different purposes and with different outcomes

Between the layers is a “contract” for each service; this can be provisioned statically or dynamically, and changed as needed

- This is part of the role of the service helpers