

SANDOG

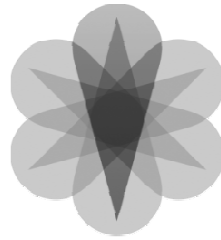
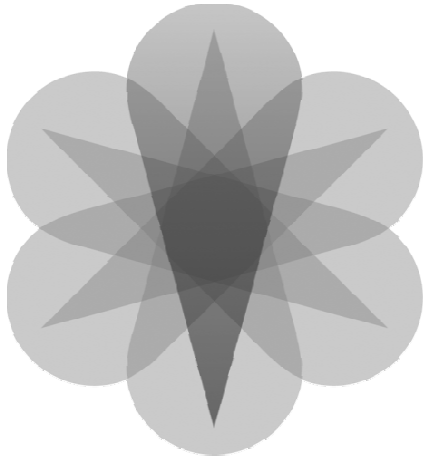
MOBILITY

SECURITY

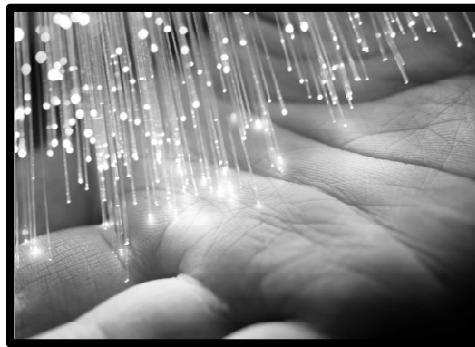
LTE

VIDEO

4D



SEAMLESS MPLS : FLEXIBLE SERVICE DELIVERY

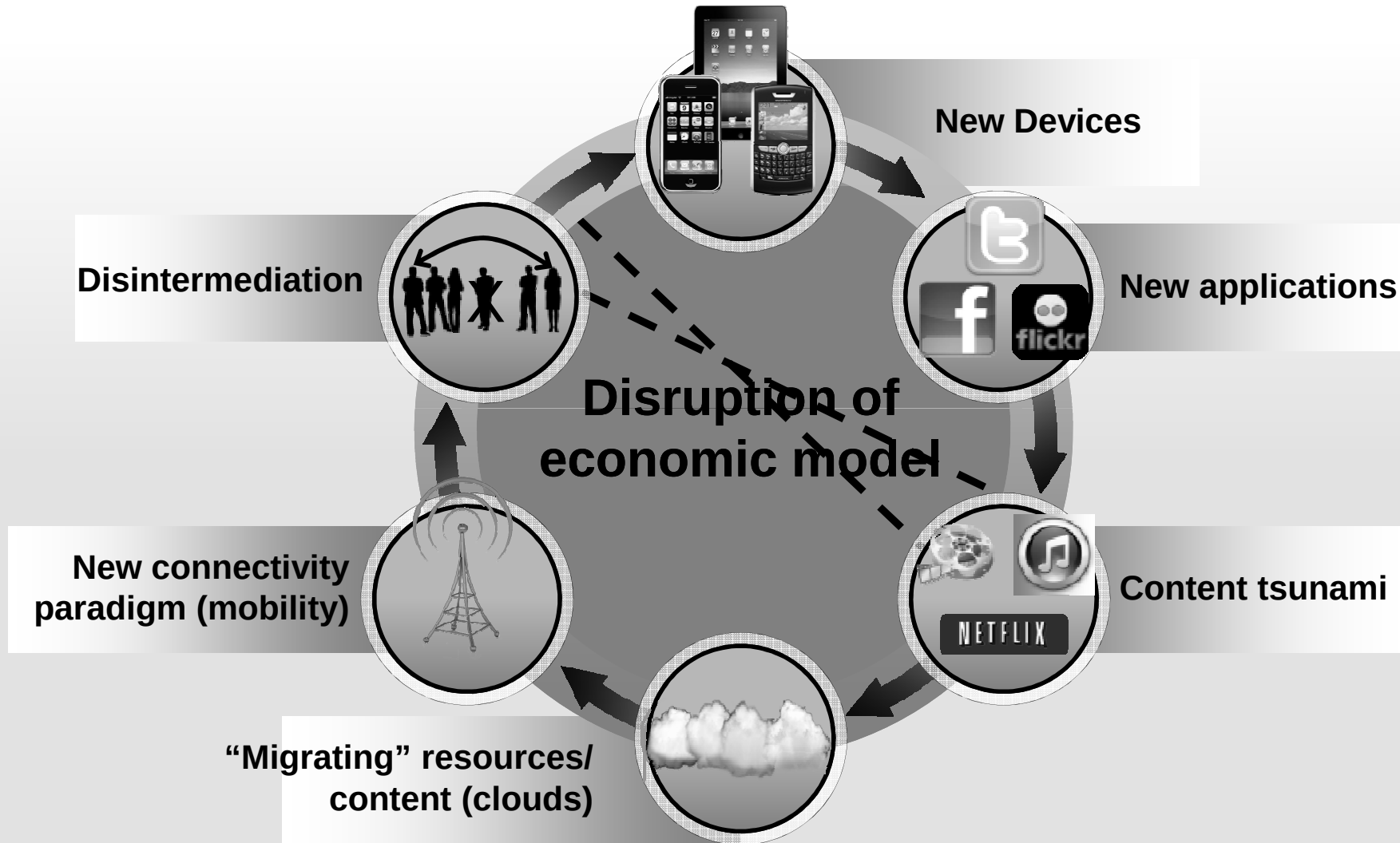


Agenda

1. The Market Dynamics
2. The Principle
3. The Technology - Seamless MPLS
4. The Evolved Backhaul
5. The Practical Implementation
6. Conclusion

Market Dynamics

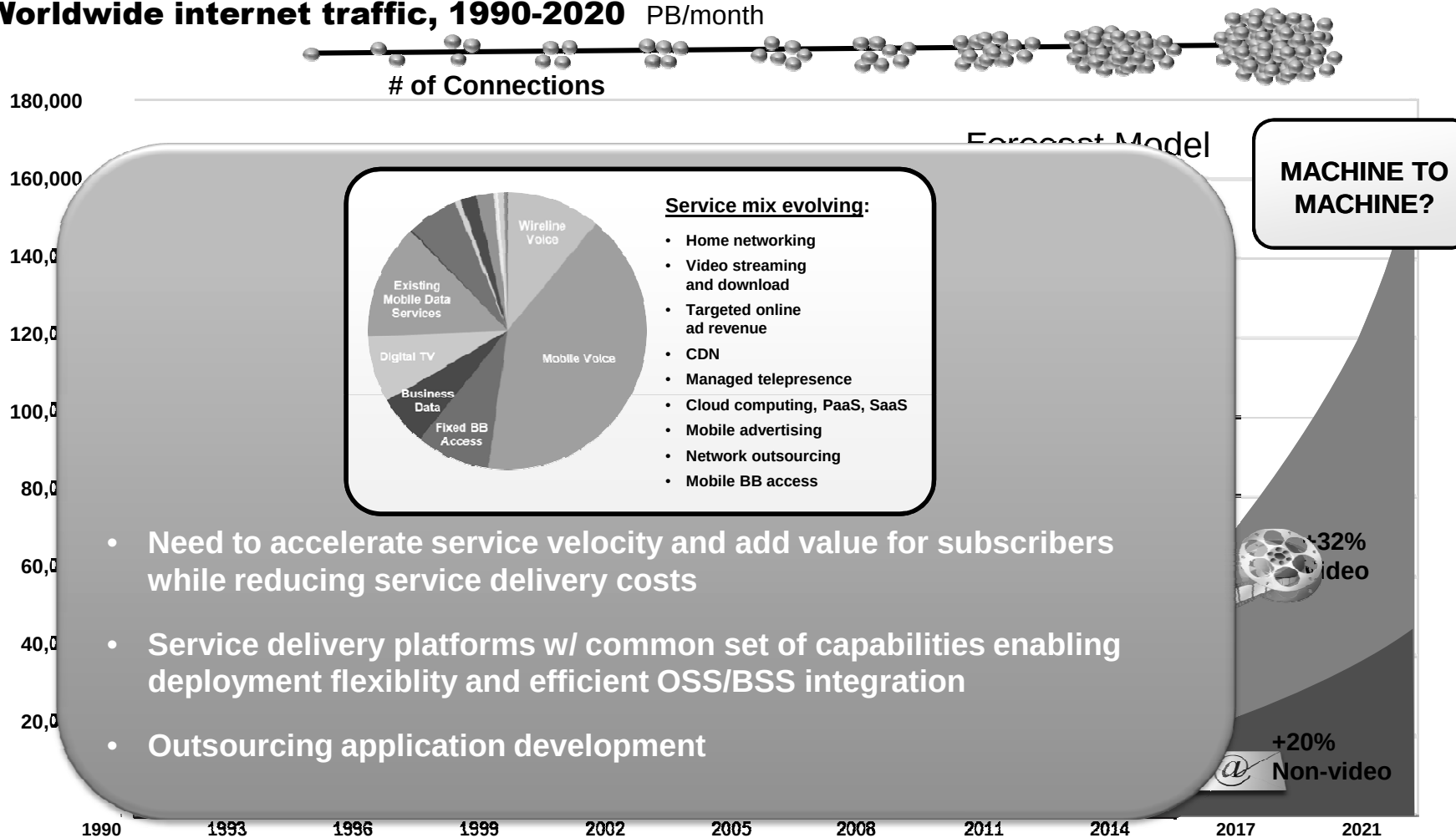
Unpredictability



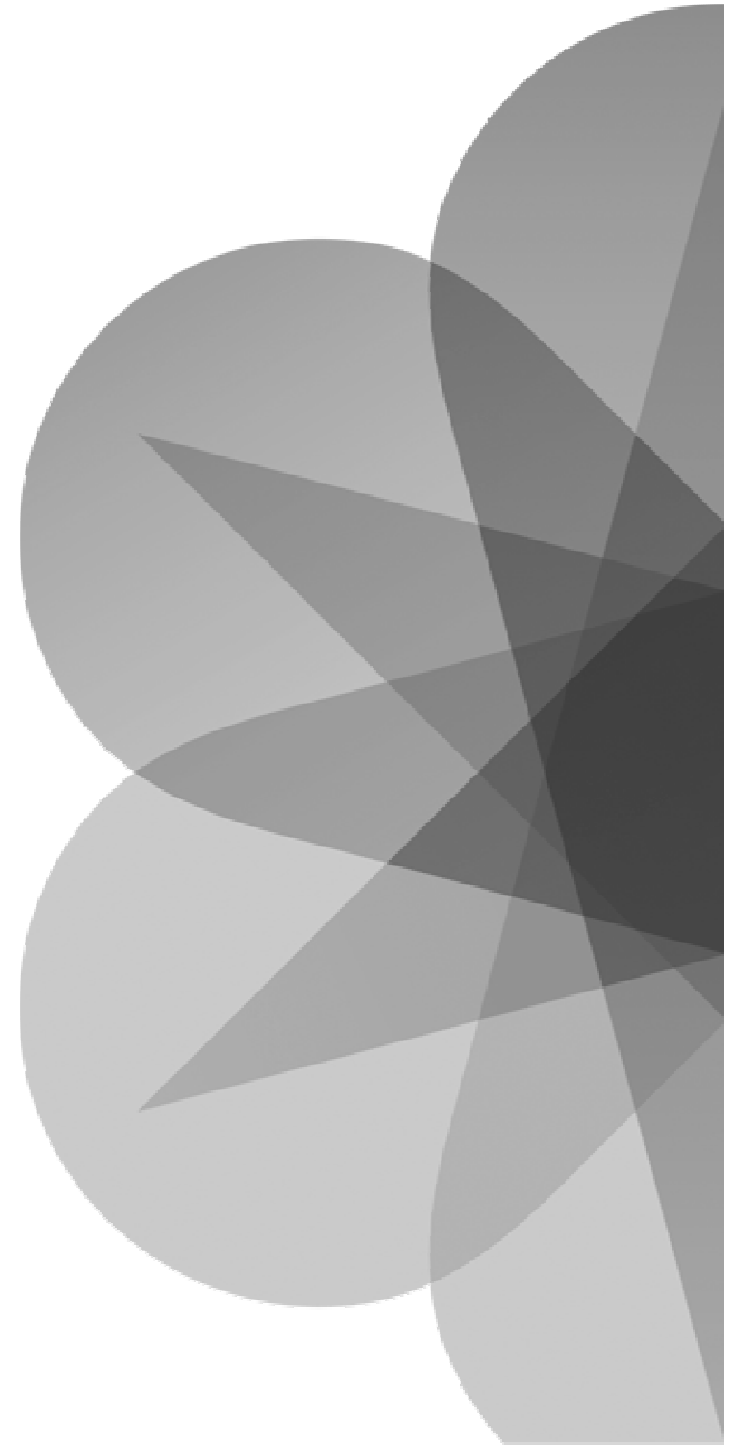
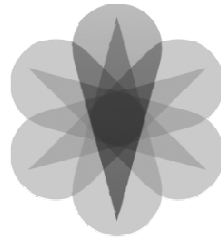
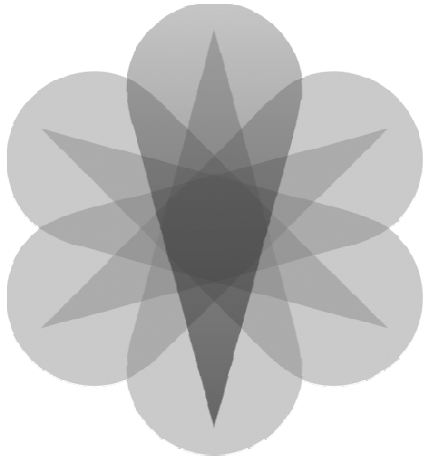
MARKET DYNAMICS

Unprecedented traffic growth

Worldwide internet traffic, 1990-2020 PB/month



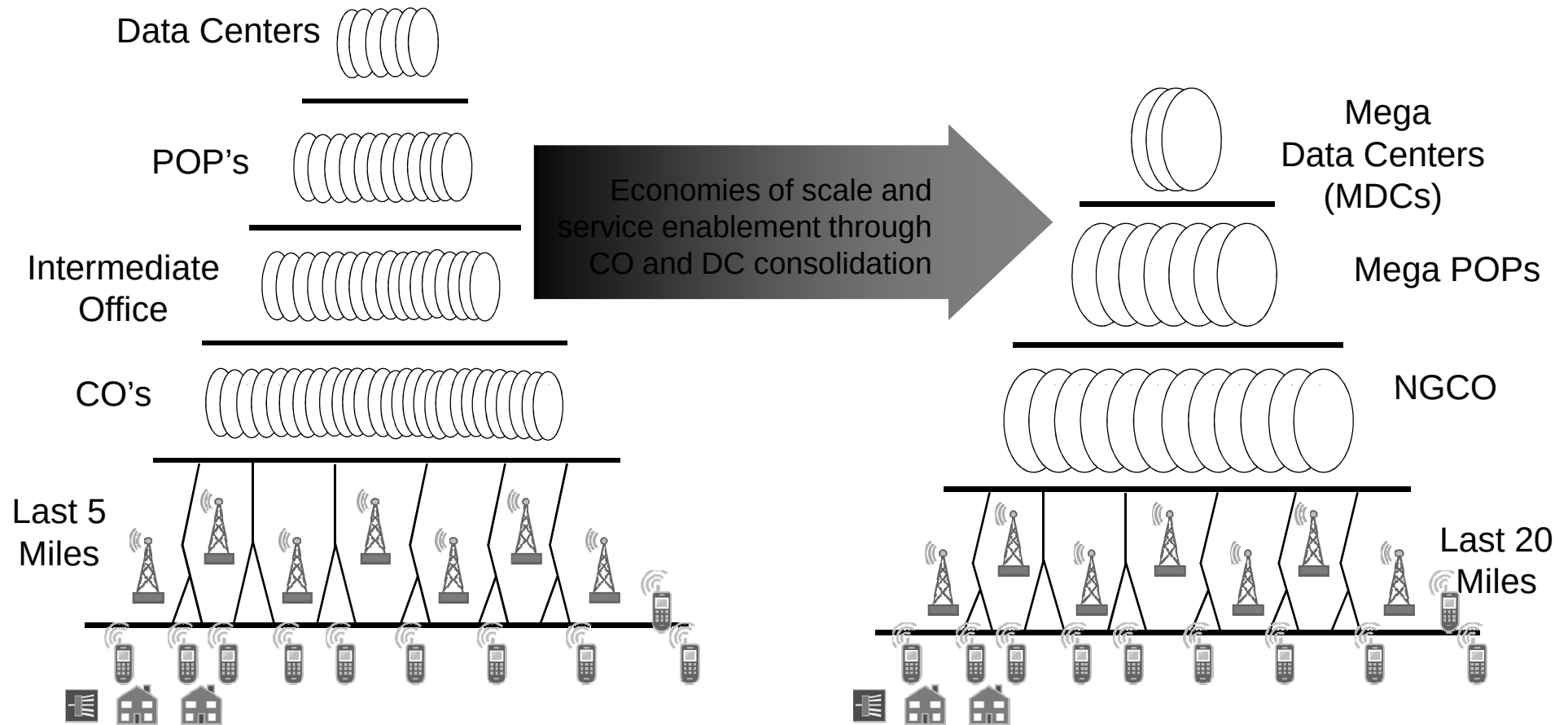
Source: Juniper, Cisco, MINTS



THE PRINCIPLE

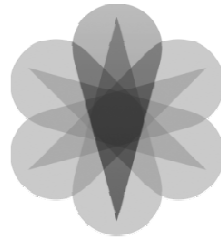
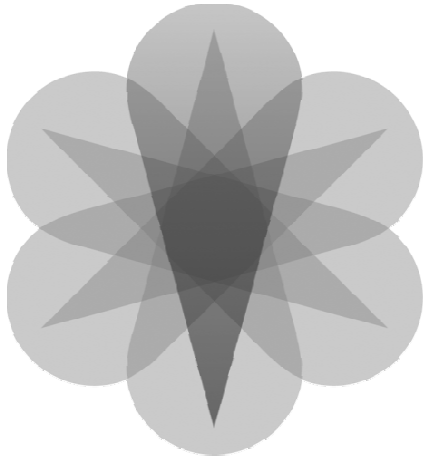
THE PRINCIPLE

Centralize what you can and distribute what you must

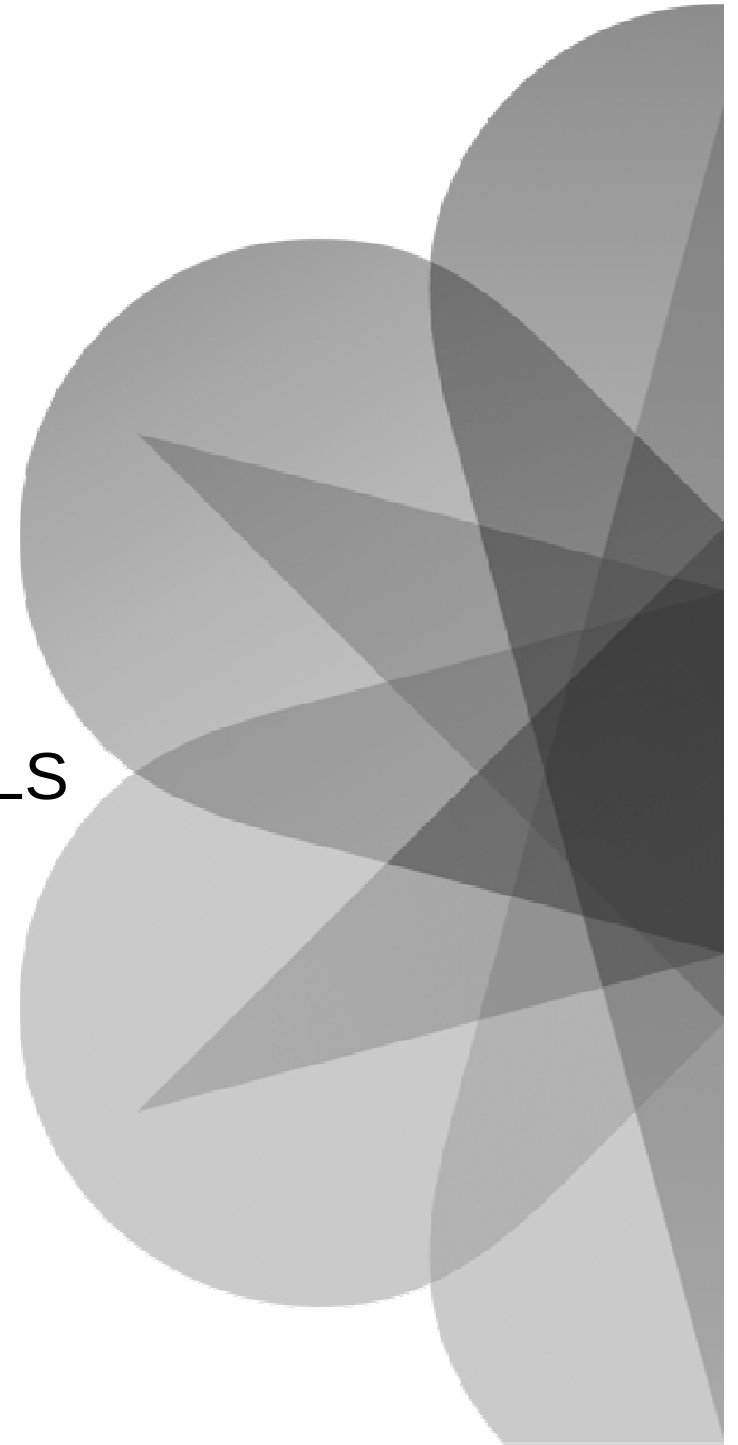


- COs and datacenters currently the biggest OPEX drains for all Service Providers
- Advanced CO and DC infrastructures required to enable Service Ecosystem

- Removing and consolidating resources realizes economies of scale
- Fabric links layers together to drive service innovation



THE TECHNOLOGY - SEAMLESS MPLS



Seamless MPLS Service and Network Architecture

Requirements addressed across the three main architectural dimensions

(1) Scale – enables 100,000s of devices in ONE PSN network

- Large network scale via MPLS LSP hierarchy and robust network protocol stack (IGP, BGP)
- No service dependency whatsoever – all packet services supported
- Low-cost/low-end access devices accommodated natively without adding complexity (MPLS labels on demand)

(2) E2E service restoration – enables sub-50ms recovery from any event

- Service restoration made independent of scale, services and failure types
- Achieved with full coverage of local-repair mechanisms for sub-50ms restoration
- Deterministic for any failure domain size / radius

(3) Decoupled network and service architectures

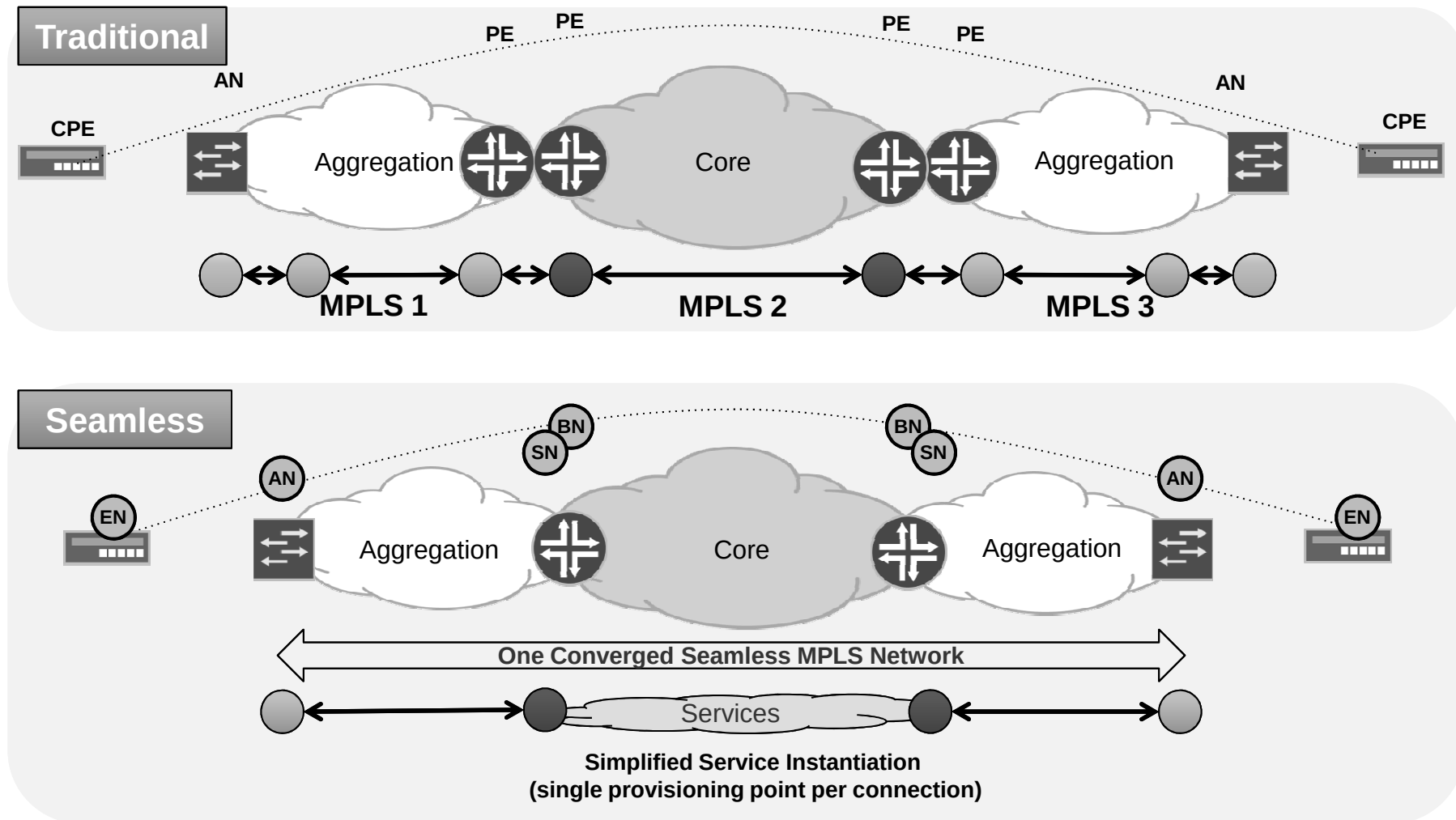
- Flexible topological placement of services enabled via MPLS Pseudowire Termination into Services
- E2E virtualization of network service delivery with tight integration of Ethernet, IP and MPLS
- Minimized number of provisioning points, simplifying service delivery and IT systems(!)

Node definitions

- **Access Node (AN).** For example, DSLAM or Cell-Site Router.
- **Service Node (SN).** For example, PE for VPN services or Broadband Network Gateway (BNG).
- **Border Node (BN).** Resides at the boundary between different domains (e.g. IGP areas)
- **Transport Node (TN).** A P-router that is not performing any of the other functions listed above.

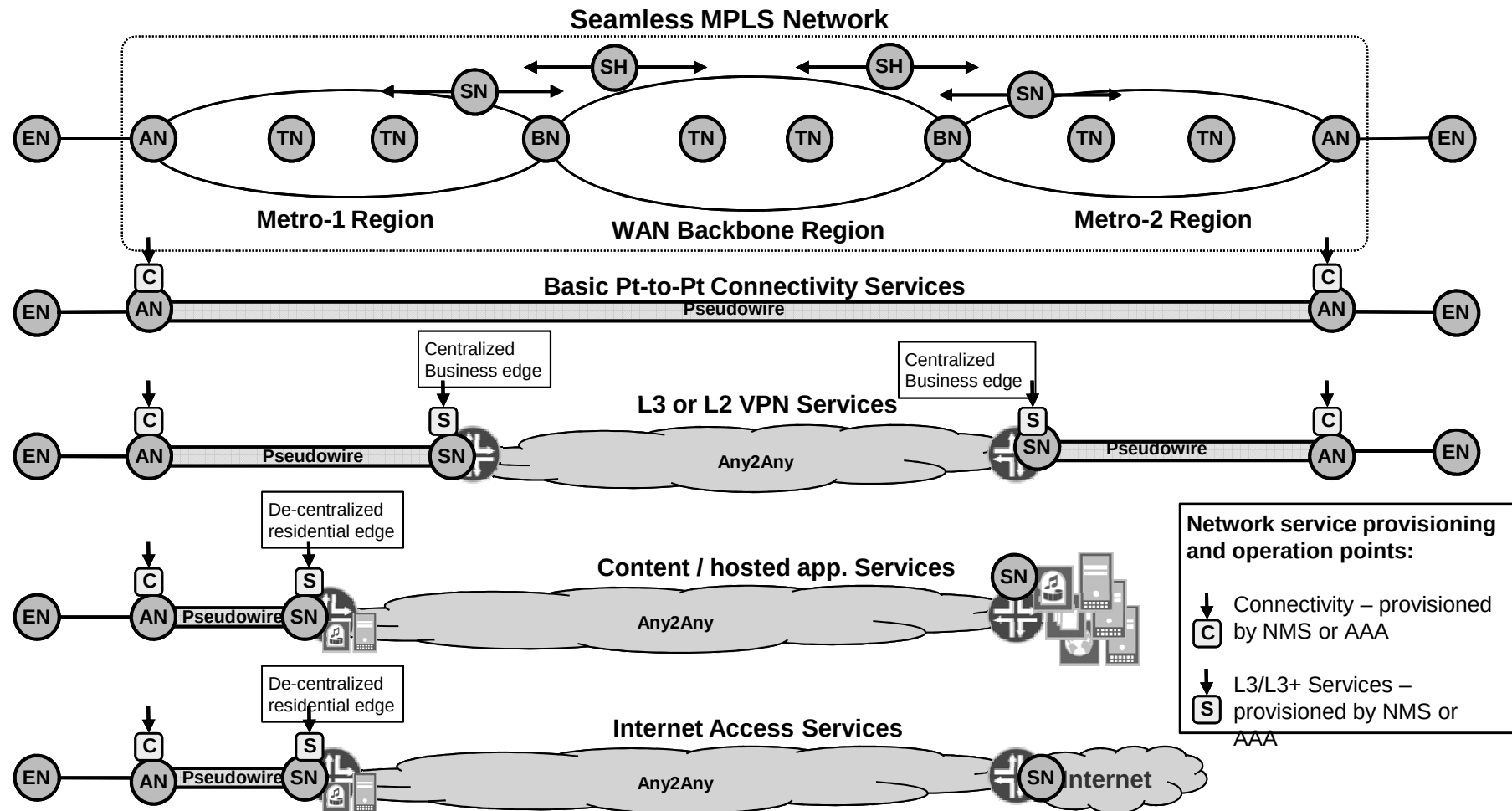
Seamless MPLS

Simplified Service Delivery



Seamless MPLS Architecture

Connectivity and Services Blueprint



MPLS scaling?

How do we scale to a network containing ~100 000 ANs and thousands of SNs??

If using RSVP, does that mean we have millions of LSPs? ($\sim N^2$)

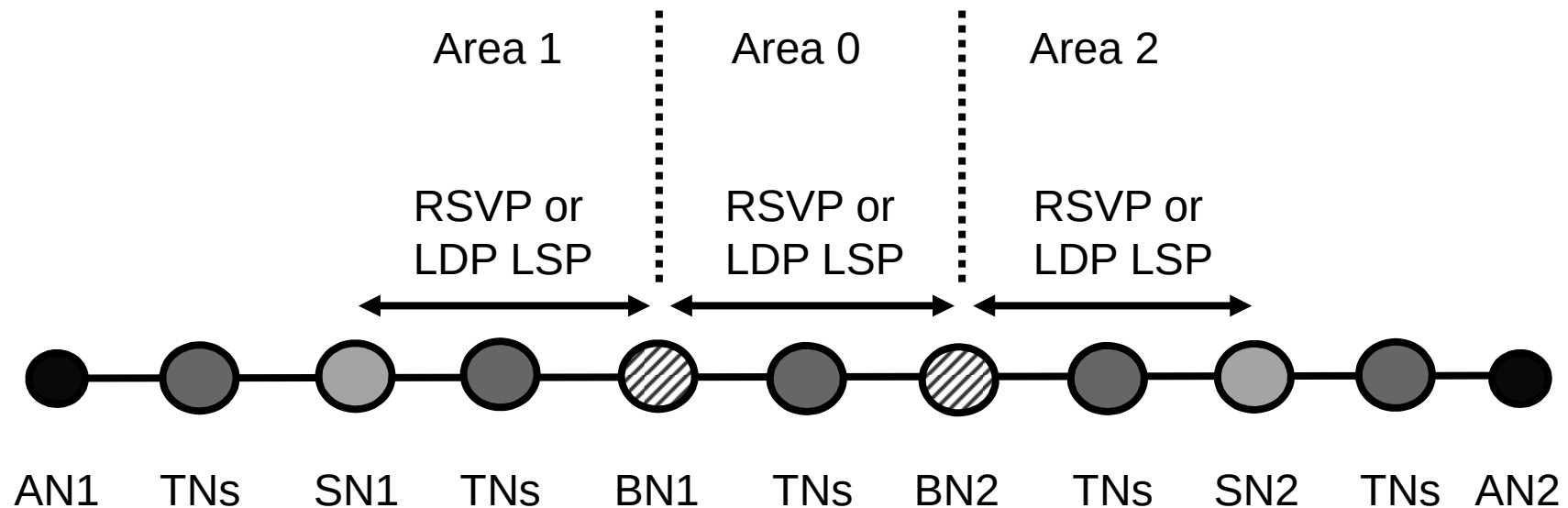
- Use RSVP LSP hierarchy to solve problem?

If using LDP, does that mean we have ~100 000 FECs, and ~100 000 /32's in the IGP?

- Relax LDP's "IGP exact match" rule?
- LDP hierarchy?

It was then realised that there was a much simpler way to address the scaling...

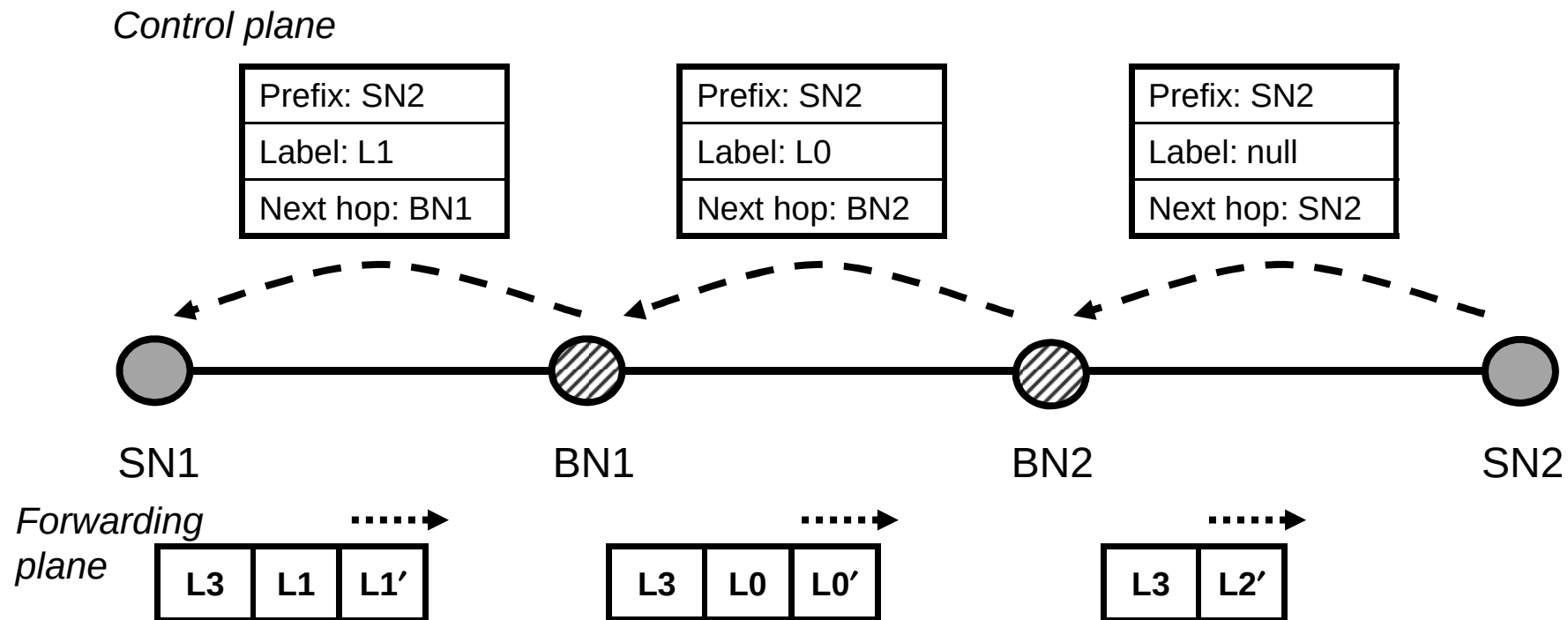
Scaling MPLS



Key point: LDP, RSVP and IGP do *not* cross area boundaries, only *labelled-BGP*.

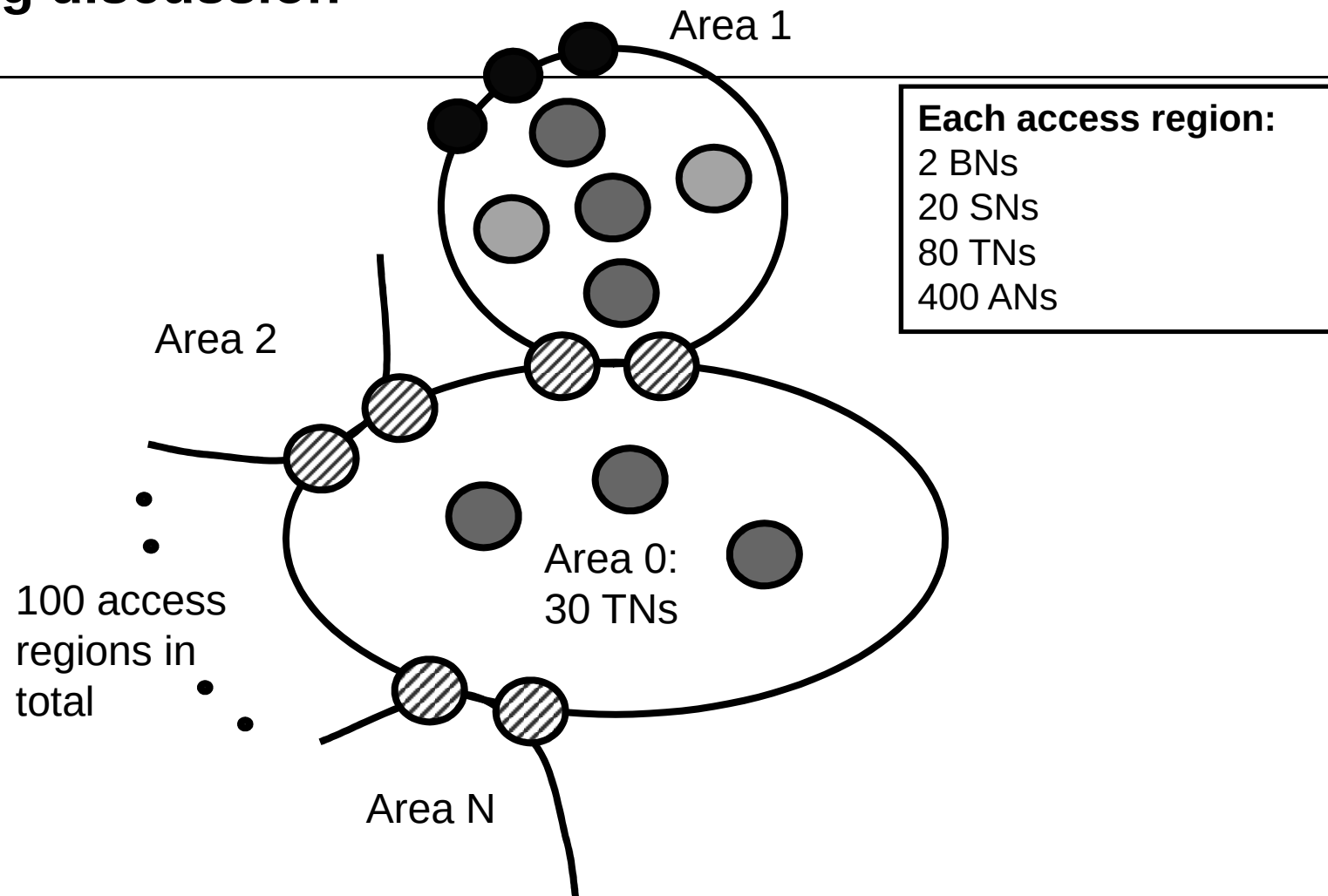
Labelled BGP (RFC 3107) allows BGP to advertise MPLS labels. It has been around for several years but here we are using it in a new context. (In the past, labelled BGP has mostly been used in the context of “Interprovider VPN Option C”)

BGP operations



- BNs are route-reflectors for labelled-BGP routes (SNs are the RR-clients)
- Top-level BGP mesh between BNs
- BN sets next-hop to self when reflecting labelled-BGP routes
- Label stack sent by SN1: Bottom label (L3) is the service label e.g. VPN label. L1 is the label corresponding to SN2. Top label (L1') is the label to reach BN1.

Scaling discussion



Network totals:

200 BNs
2000 SNs
8030 TNs
40000 ANs

Case 1: LDP in each domain, labelled BGP between domains

- Each **Access Node (AN)** has **~500 LDP FECs**, as there are ~500 nodes in each access region
- Each **Service Node (SN)** has **~500 LDP FECs**. Also has **~1980 labelled-BGP entries** (corresponding to the other SNs in the other regions).
- Each **Border Node (BN)** has **~500 LDP FECs** from the attached access region plus **~230 LDP FECs** from nodes in the core region. Also has **~2000 labelled-BGP entries** corresponding to the SNs in the network.
- **Transport Node (TN)**. Has **~500 LDP FECs** (if in access region) or **~230 LDP FECs** (if in core region)

Case 2: RSVP in each domain, labelled BGP between domains

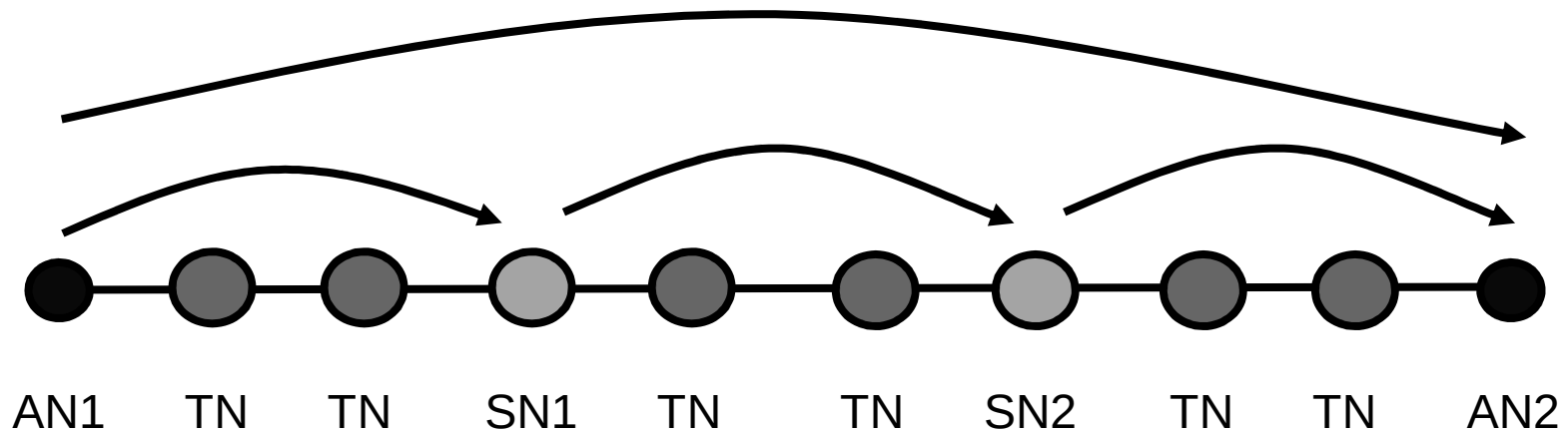
(Assuming each AN is homed to 2 SNs. Bear in mind that RSVP LSPs are unidirectional)

- Each **Access Node (AN)** has **4 RSVP LSPs** (one LSP in each direction with the two parent SNs)
- Each **Service Node (SN)** has **80 RSVP LSPs** to/from ANs in the region **plus 4 RSVP LSPs** to/from the BNs Also has **~1980 labelled-BGP entries** (corresponding to the other SNs in the other regions).
- Each **Border Node (BN)** has **40 RSVP LSPs to/from SNs** in the attached access region plus **~400 RSVP LSPs** to/from the other BNs. from nodes in the core region. Also has **~2000 labelled-BGP entries** corresponding to the SNs in the network.
- **Transport Node (TN)**. In access region, is transit for **proportion of the RSVP LSPs** mentioned above. In core region, is transit for a **proportion of the ~200² BN-BN RSVP LSPs**.

Seamless MPLS variations

What if we want ANs to act as SNs for one or more service types?

E.g. If the *service* is a point-to-point pseudowire, why not just run it directly between ANs, rather than stitching PWs at the SNs?



Running PW services directly between ANs

There could be tens of thousands of ANs in the network.

A particular AN would only have a few customer facing ports, so would only have a few PWs on it, to a handful of other ANs.

But when the network is built, we don't know in advance which AN pairs will have PWs between them.

How do we ensure that an AN knows the MPLS transport labels for the ANs that it needs to send traffic to?

Running PW services directly between ANs (cont'd)

Option 1: Labelled BGP on the AN

- A high-end AN could support Labelled BGP.
- If we put all of the AN loopbacks into labelled BGP, that implies tens of thousands of prefixes in labelled BGP. That is fine from the *RIB* point of view.
- The AN only installs a corresponding entry in the *FIB* if there is a “route” (e.g. a PW) that needs to resolve over it –there will only be a handful of these

Option 2: LDP Downstream on Demand (DoD) on the AN

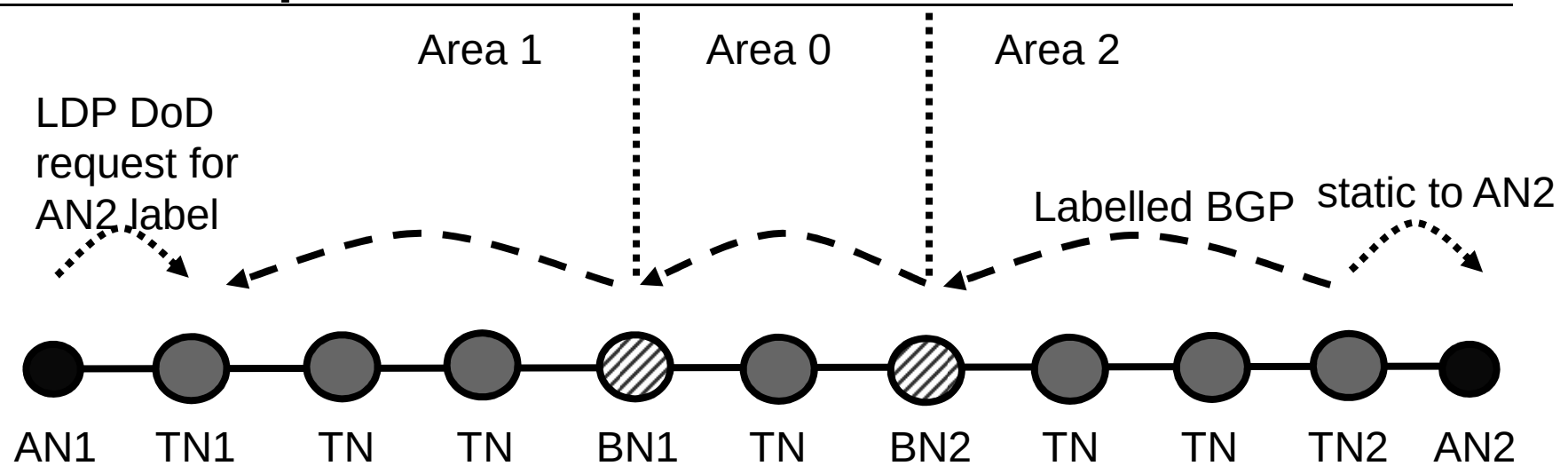
- An AN sends explicit requests only for the labels it needs from its upstream LDP neighbour
- Good scheme for low-end AN –can just have LDP DoD plus static routes

LDP Downstream on Demand (DoD)

LDP DoD has been around for many years

- Discussed in RFC 5036 (LDP Specification)
- Originally intended for ATM switches, so not traditionally used/supported by routers
- Seamless MPLS is a new use-case for LDP DoD

LDP DoD operations

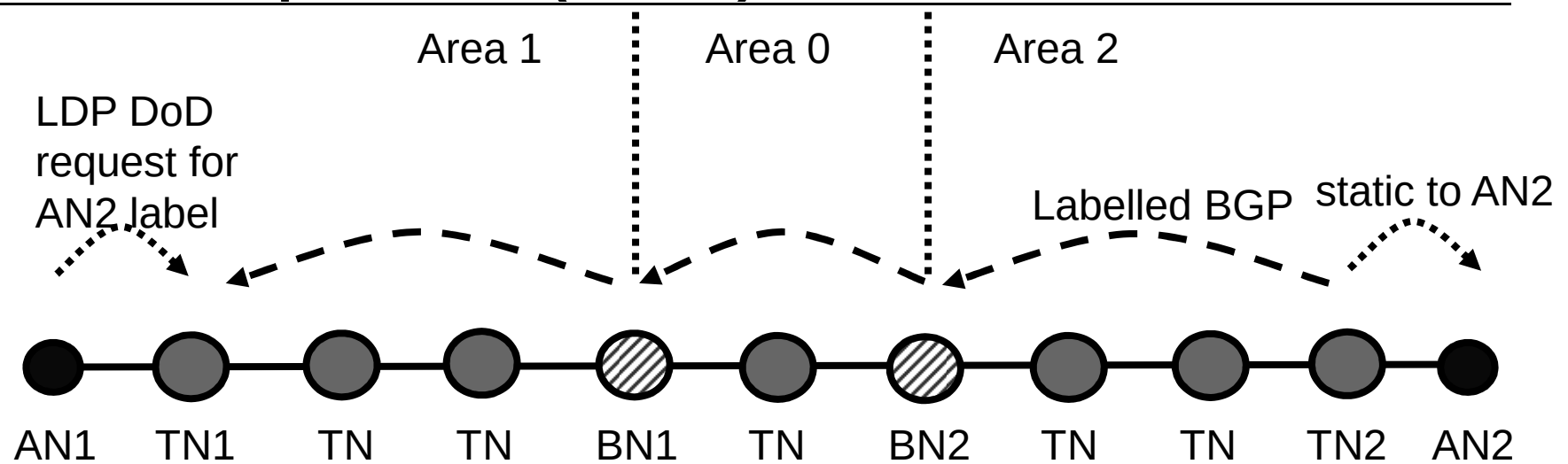


Note that the LDP DoD is providing an extension to the Labelled-BGP layer. The diagram above only shows that layer.

In addition, there are LDP or RSVP LSPs to carry traffic from TN1 to BN1, from BN1 to BN2 and BN2 to TN2. (Not shown on diagram).

Note: there is also signalling at the PW layer between AN1 and AN2. (Not shown on diagram)

LDP DoD operations (cont'd)

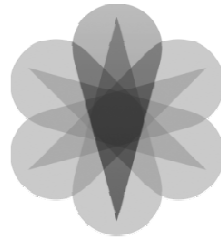
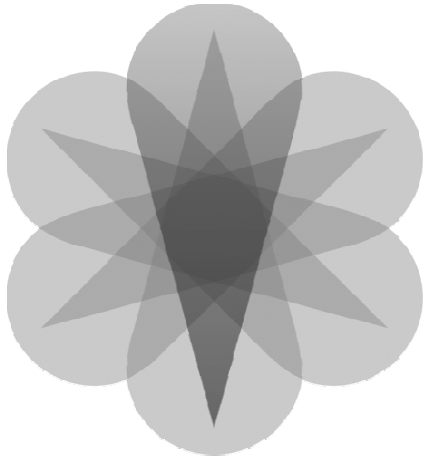


If a TN is attached to an AN, it has a static route to that AN, which it redistributes into labelled-BGP. These TNs are RR clients of the BNs, for labelled BGP.

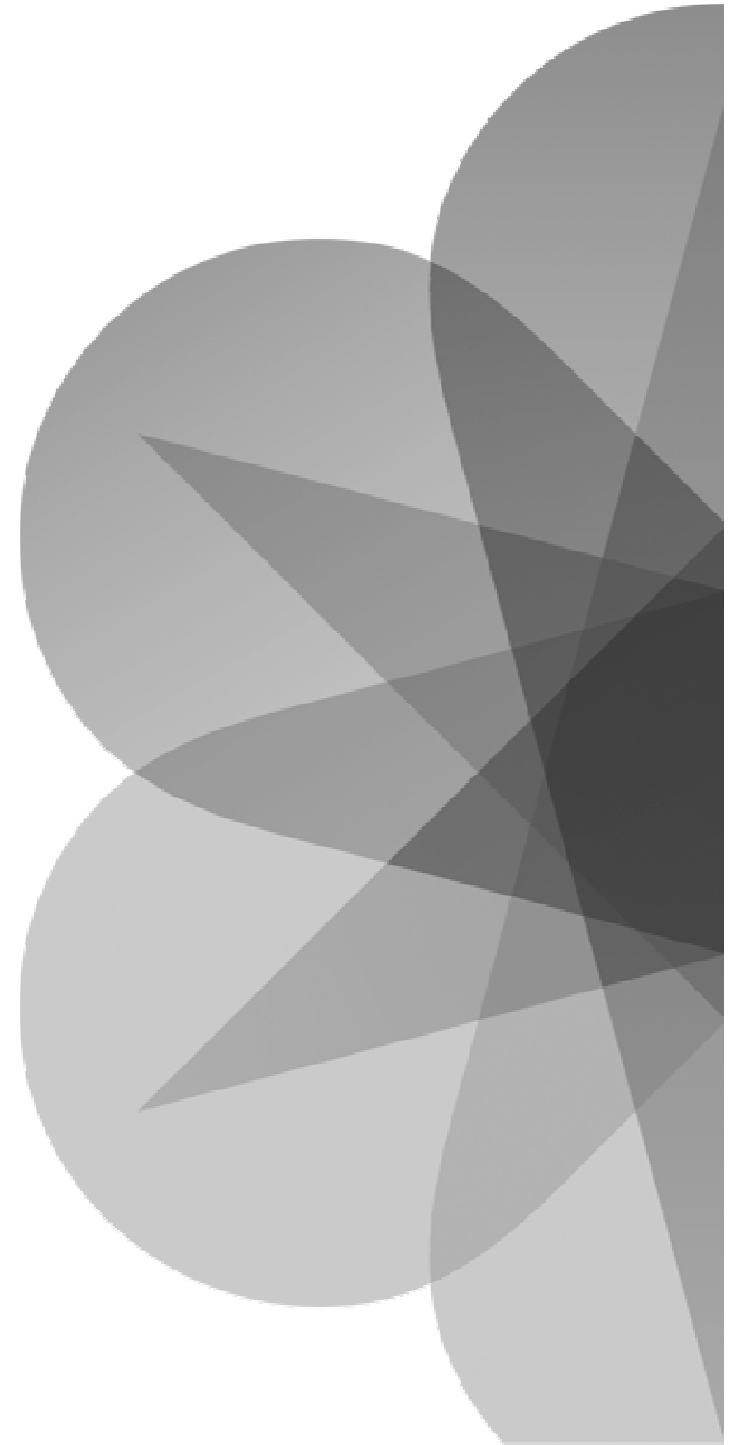
LDP DoD runs between AN and attached TN.

AN1 knows it needs a label for AN2, because it needs to build a PW to it (identity of AN2 is manually configured or auto-discovered via BGP-L2VPN). So AN1 uses LDP DoD to request AN2's label from TN1.

Note: there is also signalling at the PW layer between AN1 and AN2 (not shown on diagram)



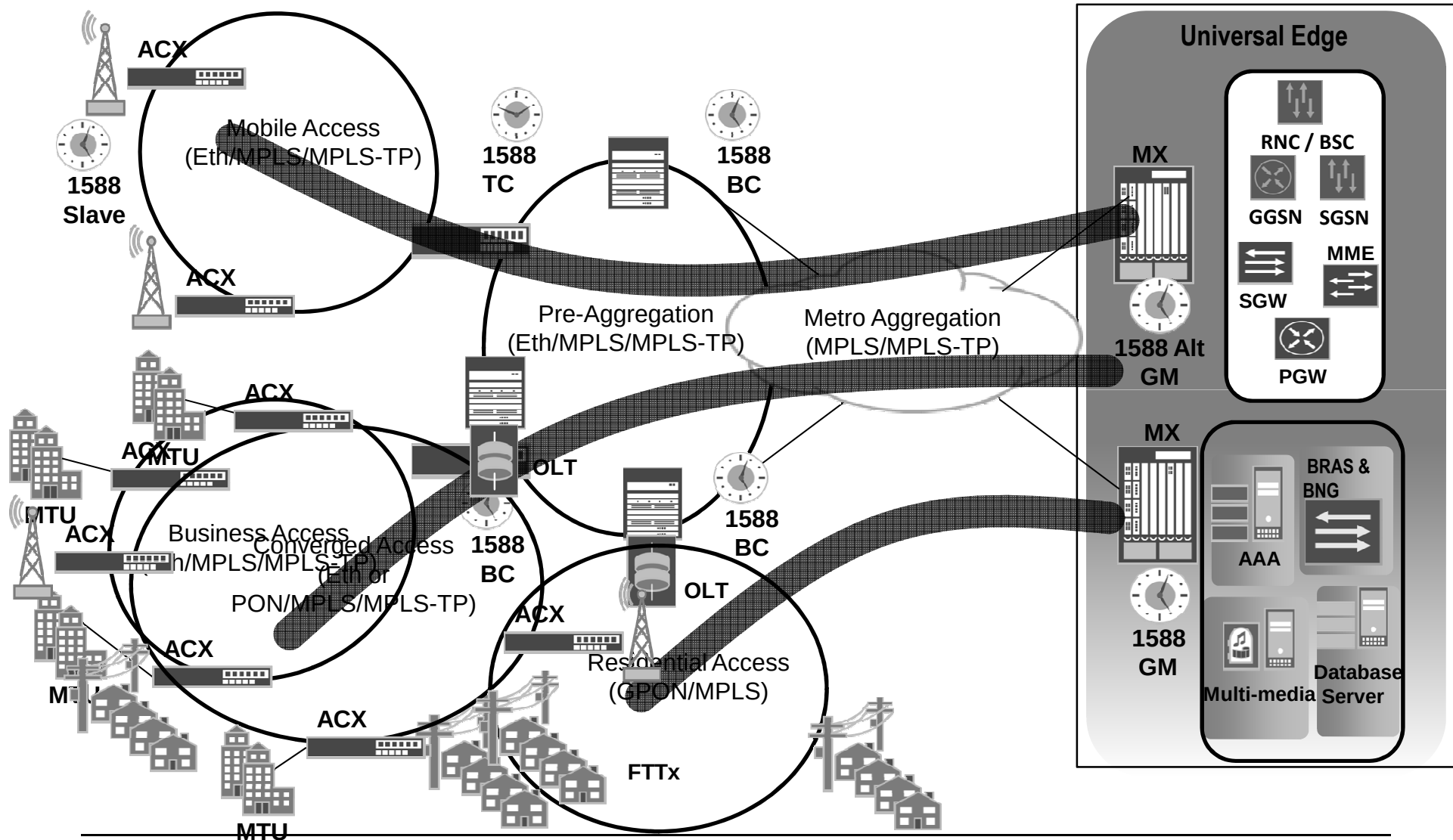
THE EVOLVED BACKHAUL



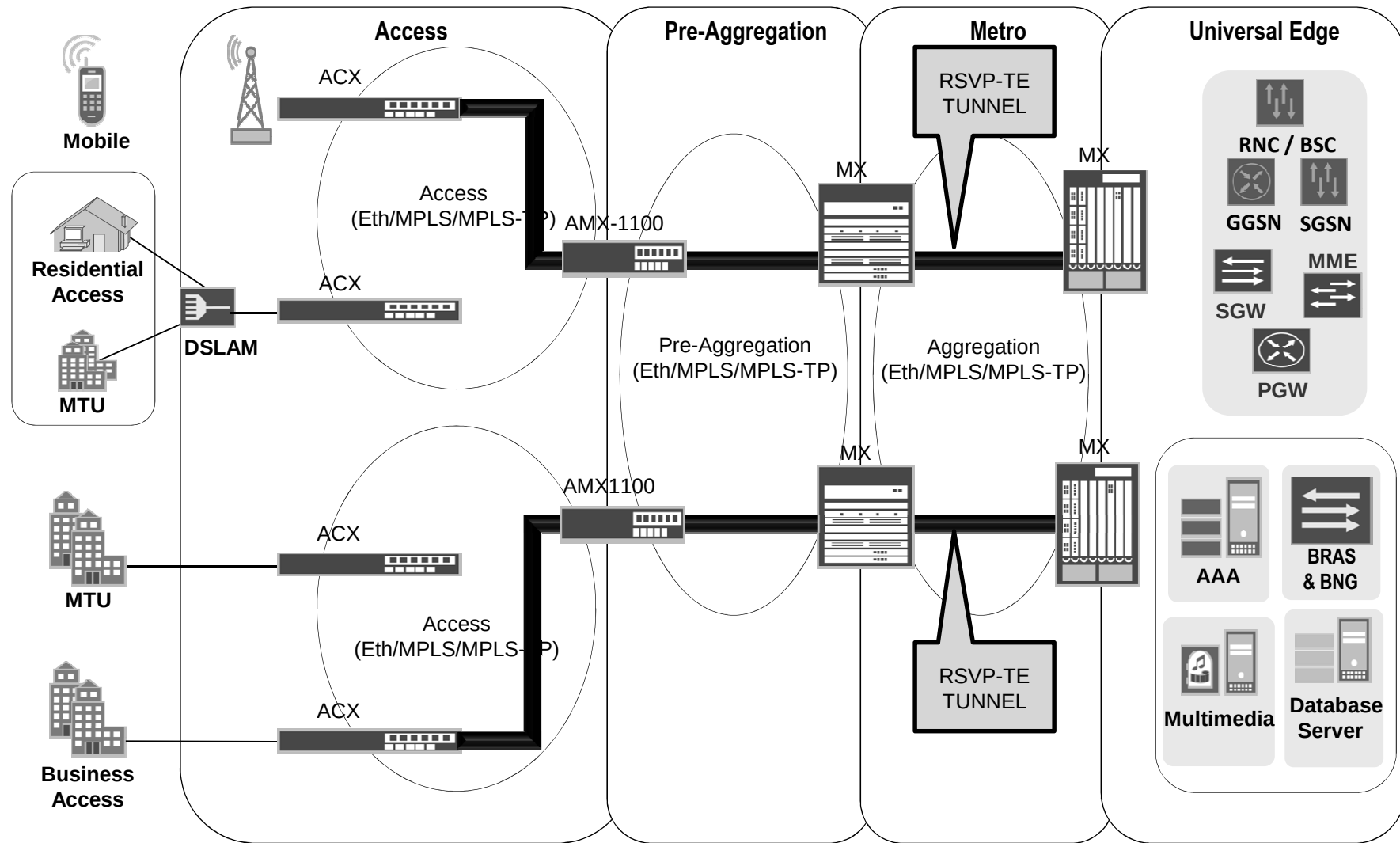
The diagram illustrates the Evolved Transport Network architecture. It shows two main service clouds at the top: **L2/L3 Services** and **L3 Any-to-Any Services**. These are connected to a central core of network elements. The core includes **AAA**, **BRAS & BNG**, **Database Server**, **Multimedia**, **RNC / BSC**, **GGSN**, **SGSN**, **SGW**, **MME**, and **PGW**. These elements are interconnected via **LSP** (Label Switched Path) and **PW** (Pseudowire) links. The network is connected to an **Evolved Transport Network** backbone, which consists of **Access & Aggregation** and **Backbone** components. The backbone is shown as a large cloud with a central cross symbol. The entire network is labeled **Evolved Transport Network** at the bottom.

- 26

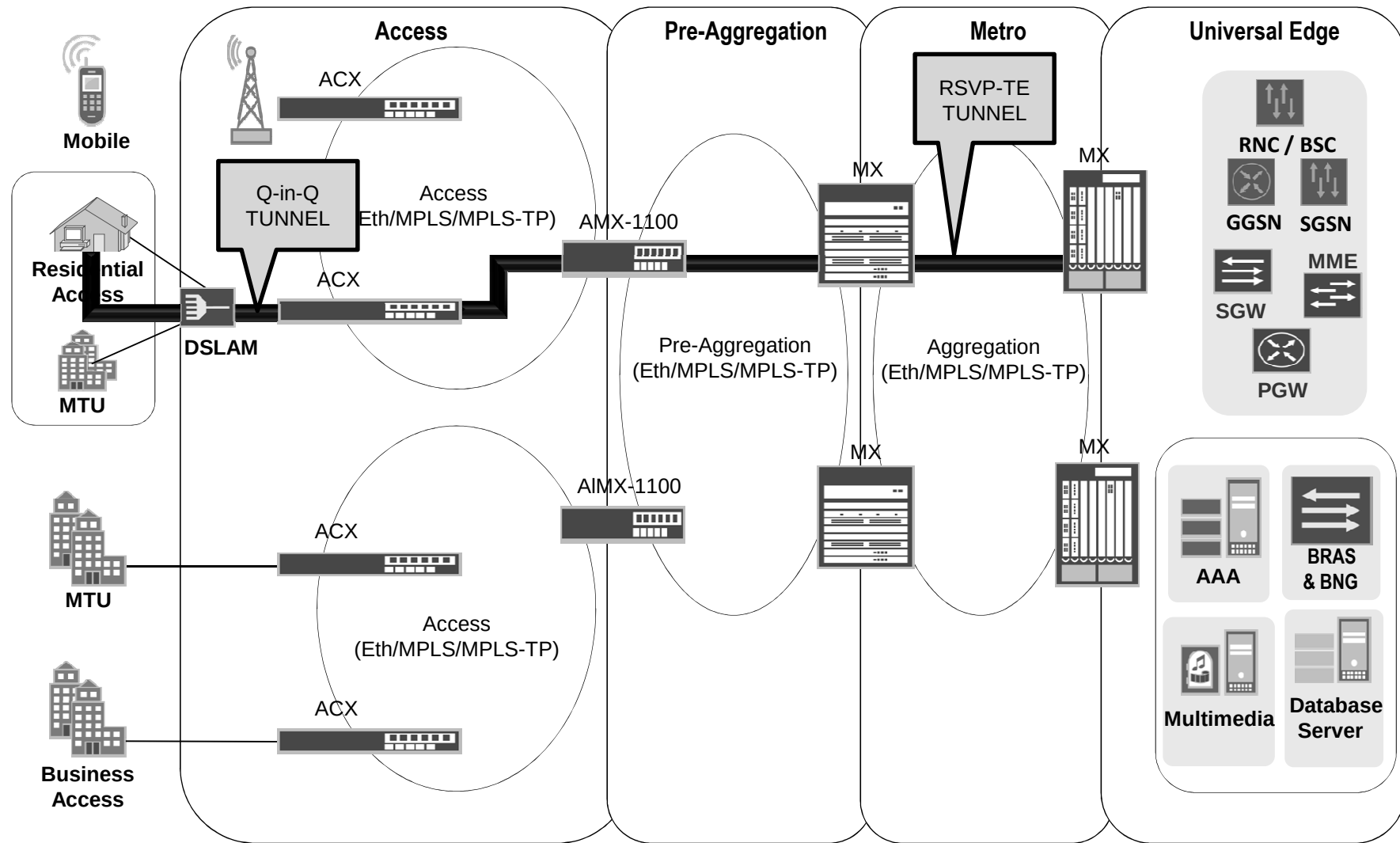
CONVERGED ACCESS AND AGGREGATION: SYNCHRONIZATION AND SEAMLESS MPLS



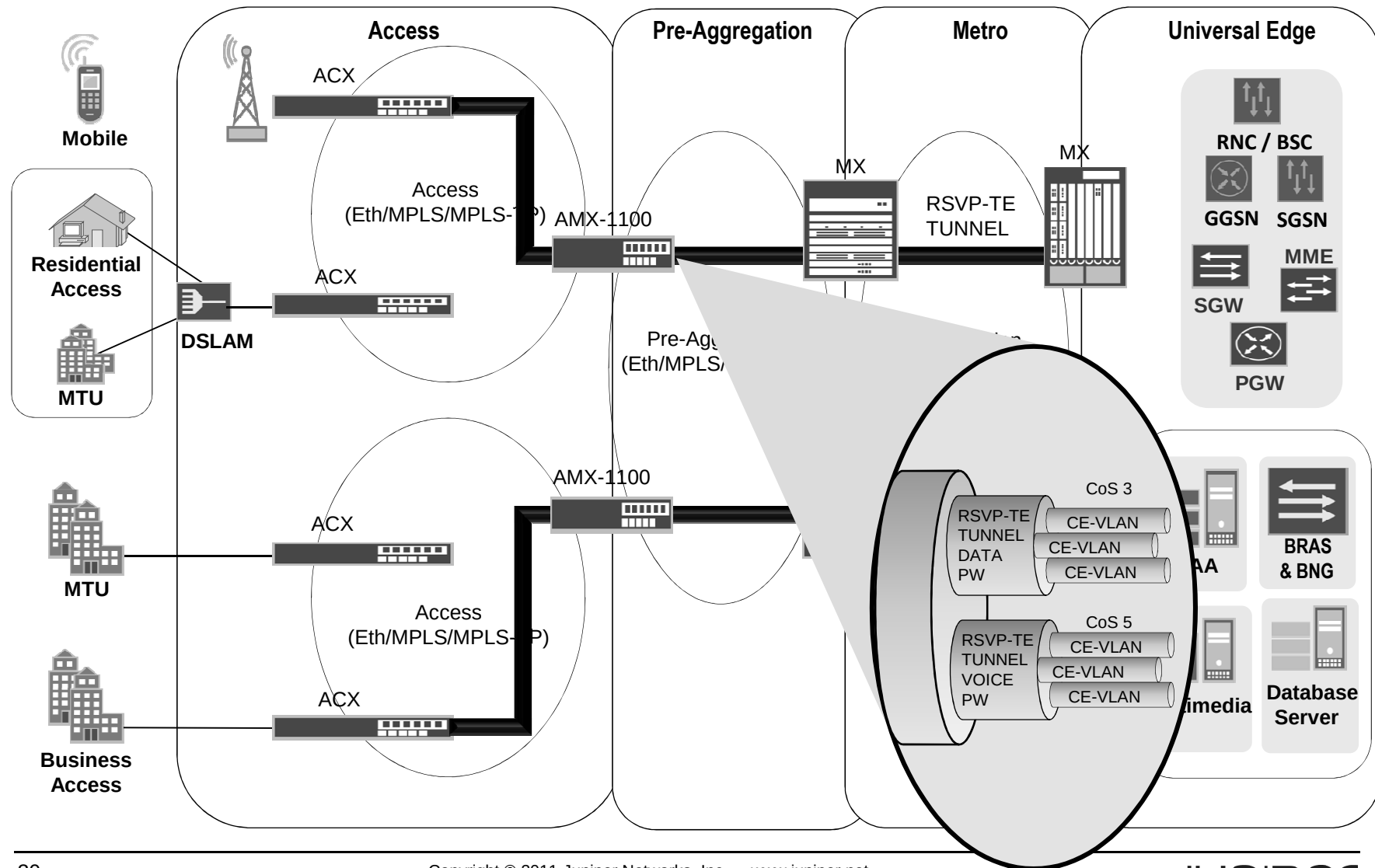
Seamless MPLS: Transport Plane



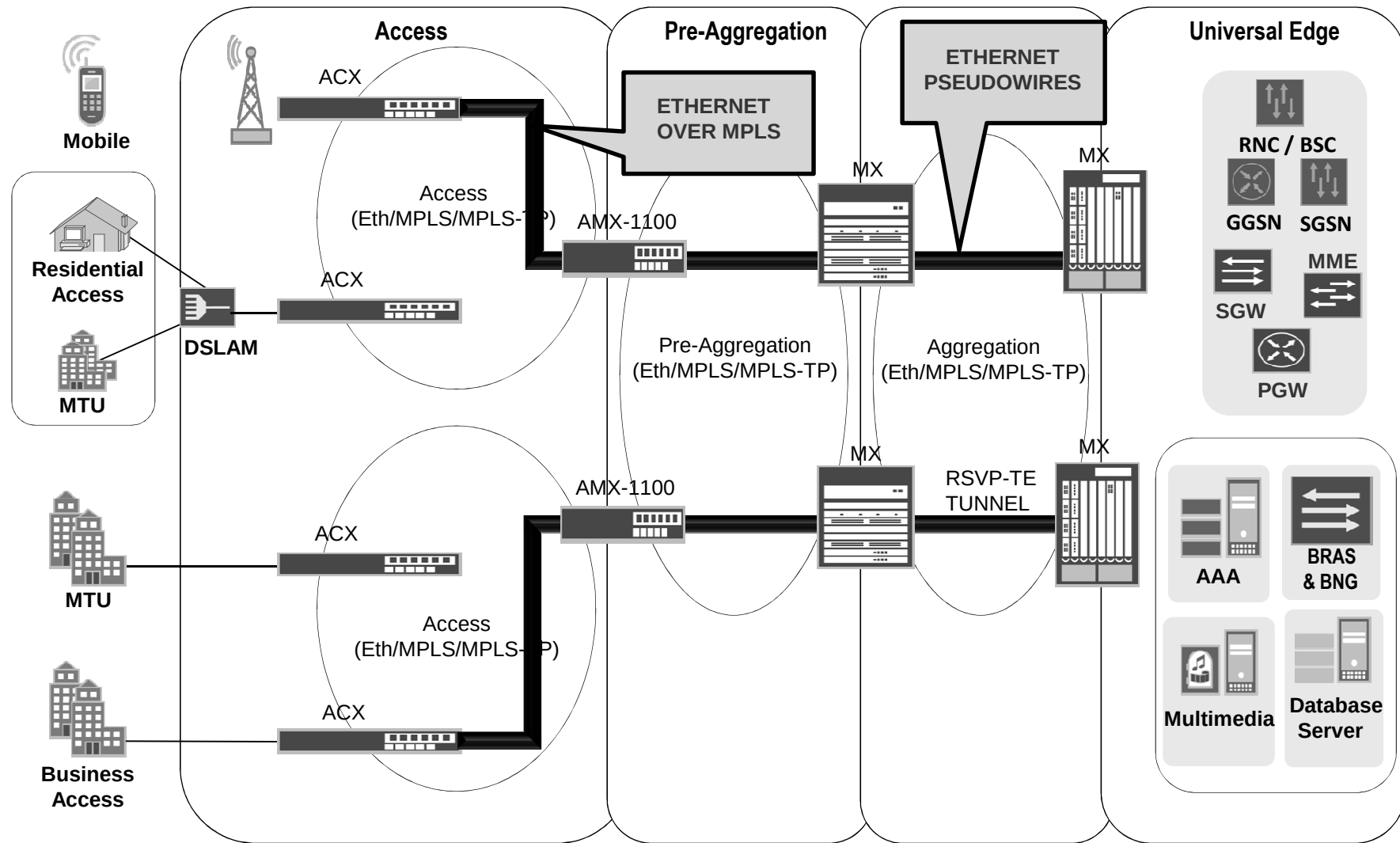
Seamless MPLS: Transport Plane, Eth & MPLS



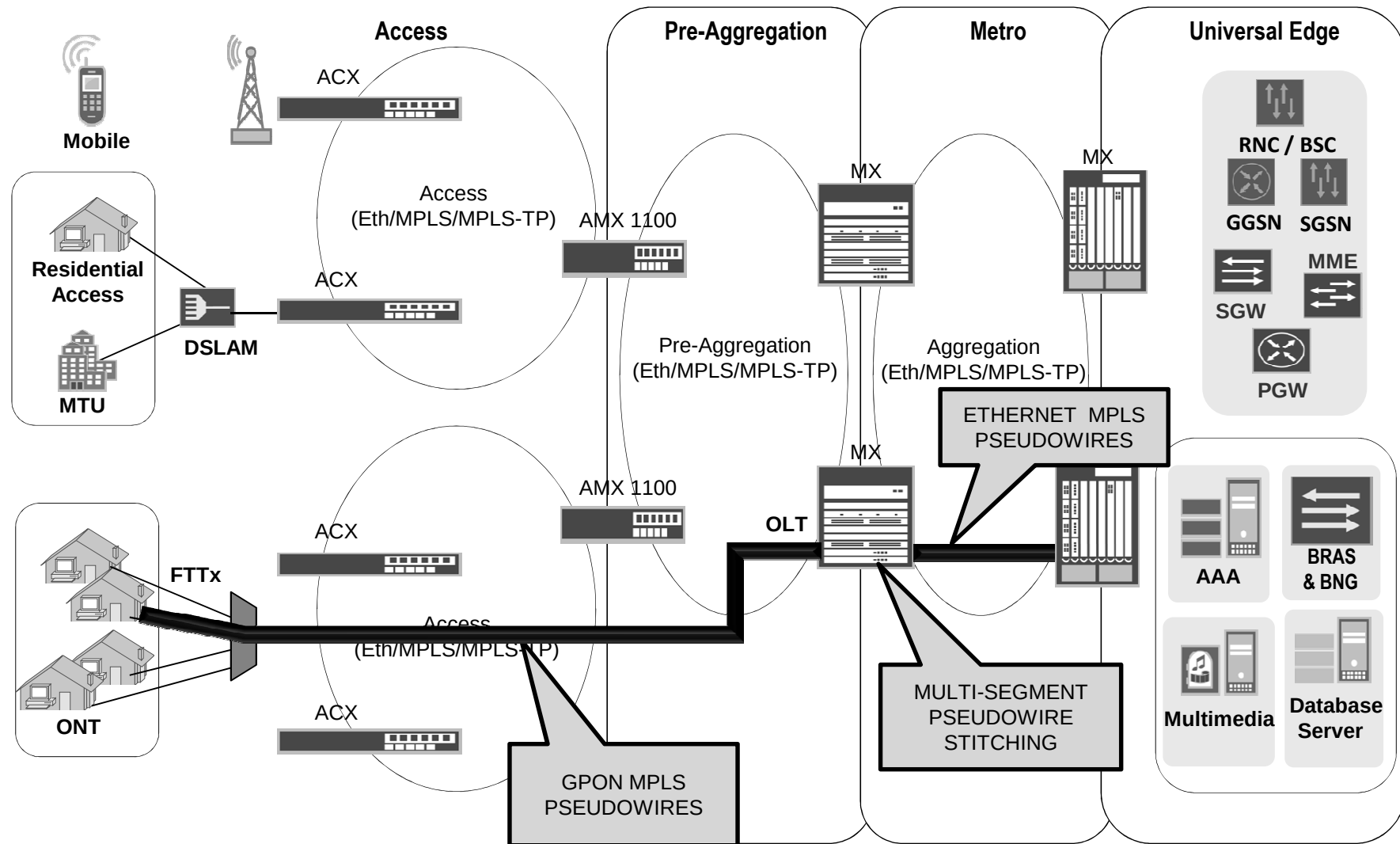
SEAMLESS MPLS: SERVICES Plane



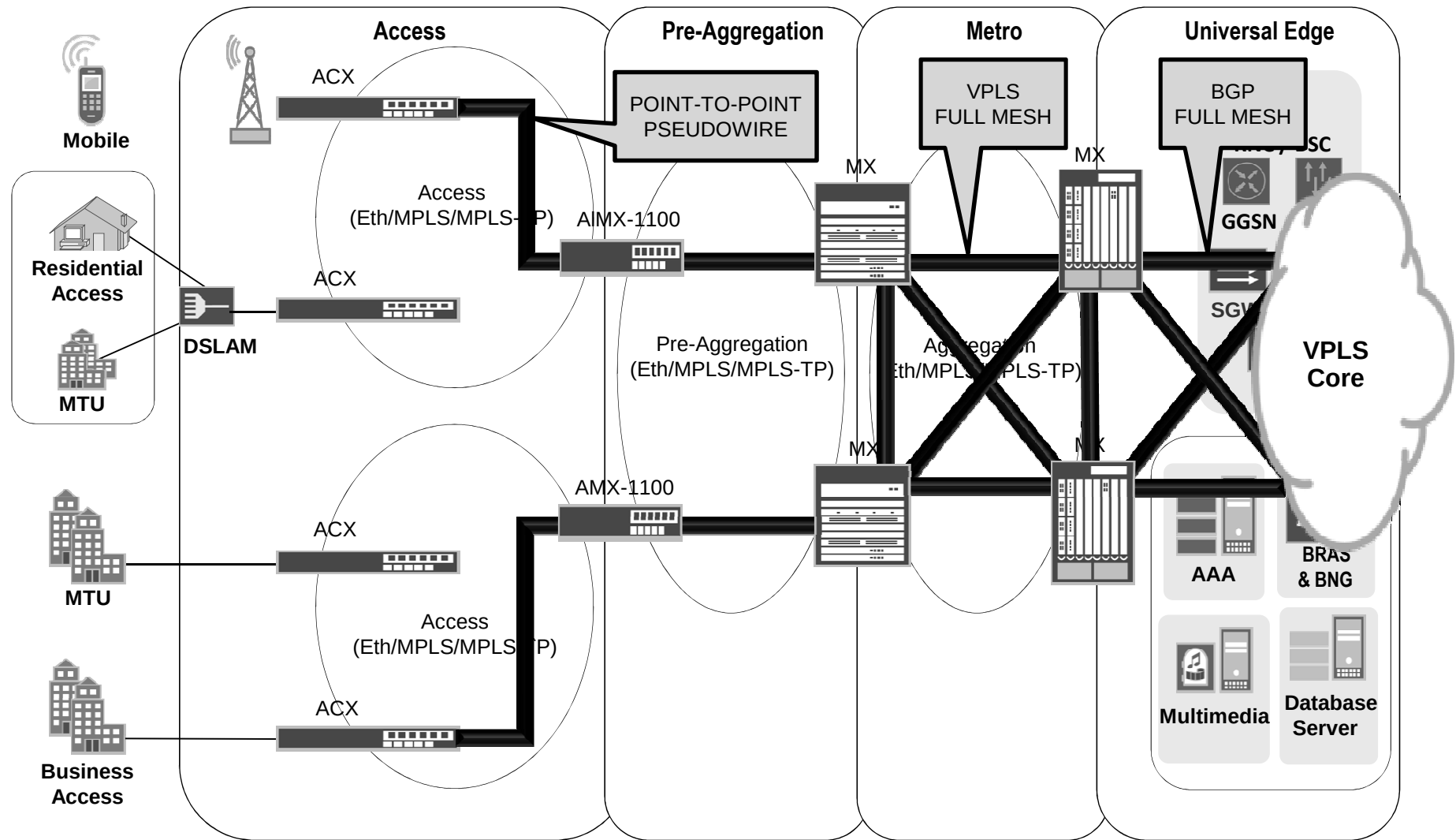
SEAMLESS MPLS: Service Plane TDM to Packet Migration



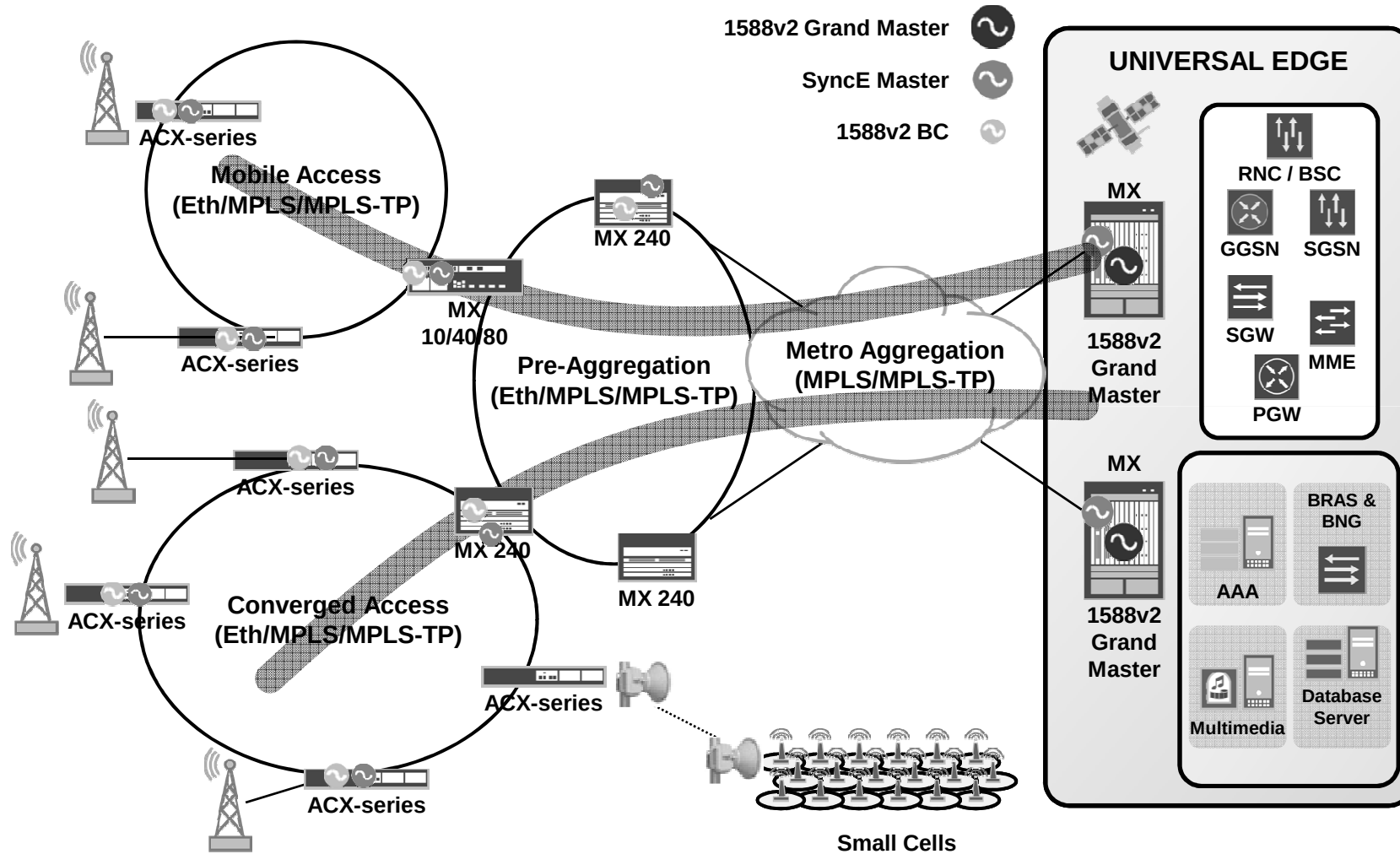
SEAMLESS MPLS: Service Plane

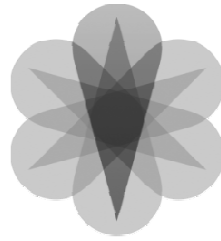
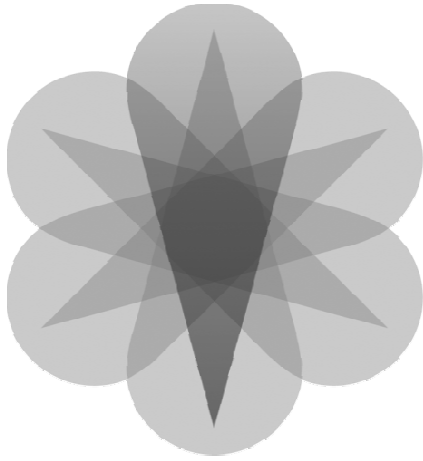


SEAMLESS MPLS: Service Plane, L2VPN & L3VPN

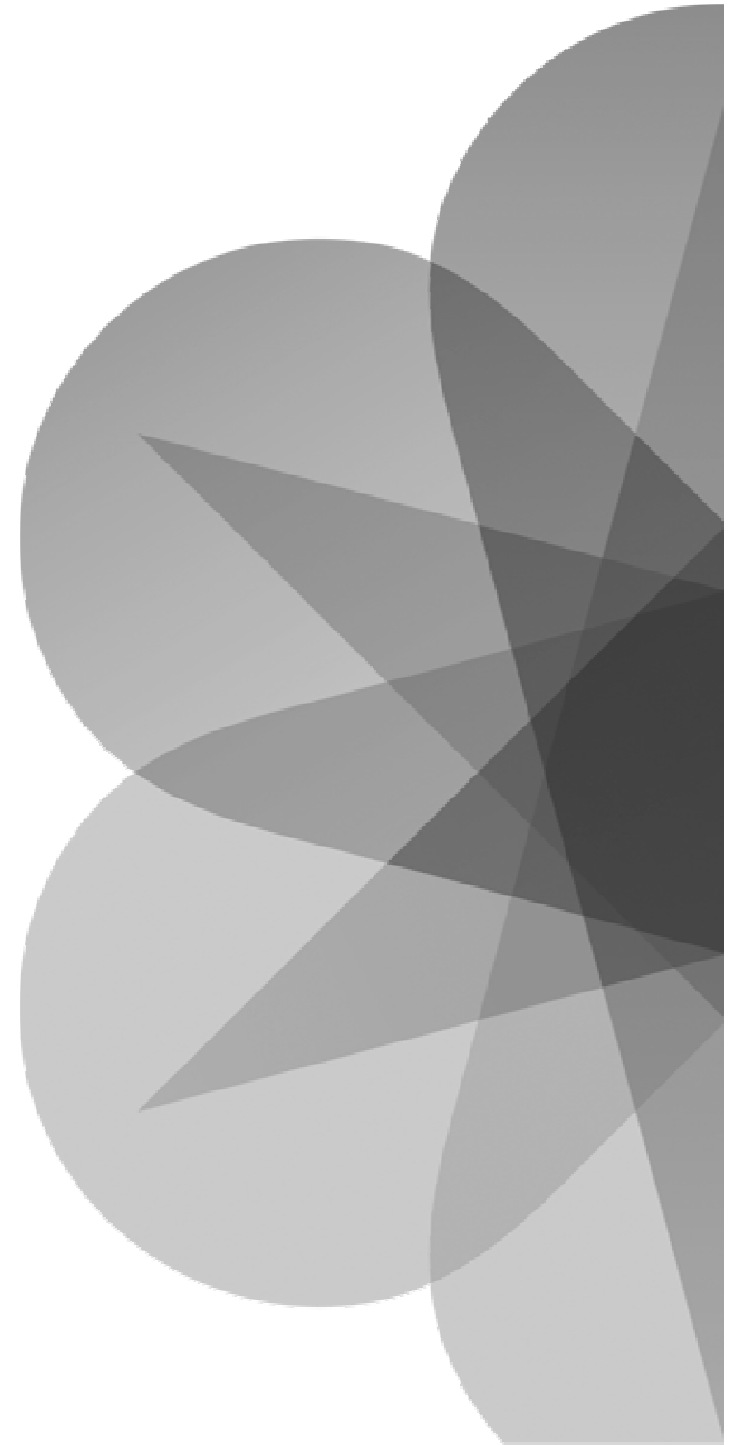


MPLS in Access : The Next Generation MBH solution





Practical Implementation



Seamless MPLS Service and Network Architecture

RECAP

Requirements addressed across the three main architectural dimensions

(1) Scale – enables 100,000s of devices in ONE PSN network

- Large network scale via MPLS LSP hierarchy and robust network protocol stack (IGP, BGP)
- No service dependency whatsoever – all packet services supported
- Low-cost/low-end access devices accommodated natively without adding complexity (MPLS labels on demand)

(2) E2E service restoration – enables sub-50ms recovery from any event

- Service restoration made independent of scale, services and failure types
- Achieved with full coverage of local-repair mechanisms for sub-50ms restoration
- Deterministic for any failure domain size / radius

(3) Decoupled network and service architectures

- Flexible topological placement of services enabled via MPLS Pseudowire Termination into Services
- E2E virtualization of network service delivery with tight integration of Ethernet, IP and MPLS
- Minimized number of provisioning points, simplifying service delivery and IT systems(!)

Seamless MPLS – Use Case Network Scale

Design

- Split the network into regions: access, metro/aggregation, edge, core
- Single IGP with areas per metro/edge and core regions
- Hierarchical LSPs to enable e2e LSP signaling across all regions
- IGP + LDP for intra-domain transport LSP signaling
 - RSVP-TE as alternative
- BGP labeled unicast for cross-domain hierarchical LSP signaling
- LDP Downstream-on-Demand for LSP signaling to/from access devices
- Static routing on access devices

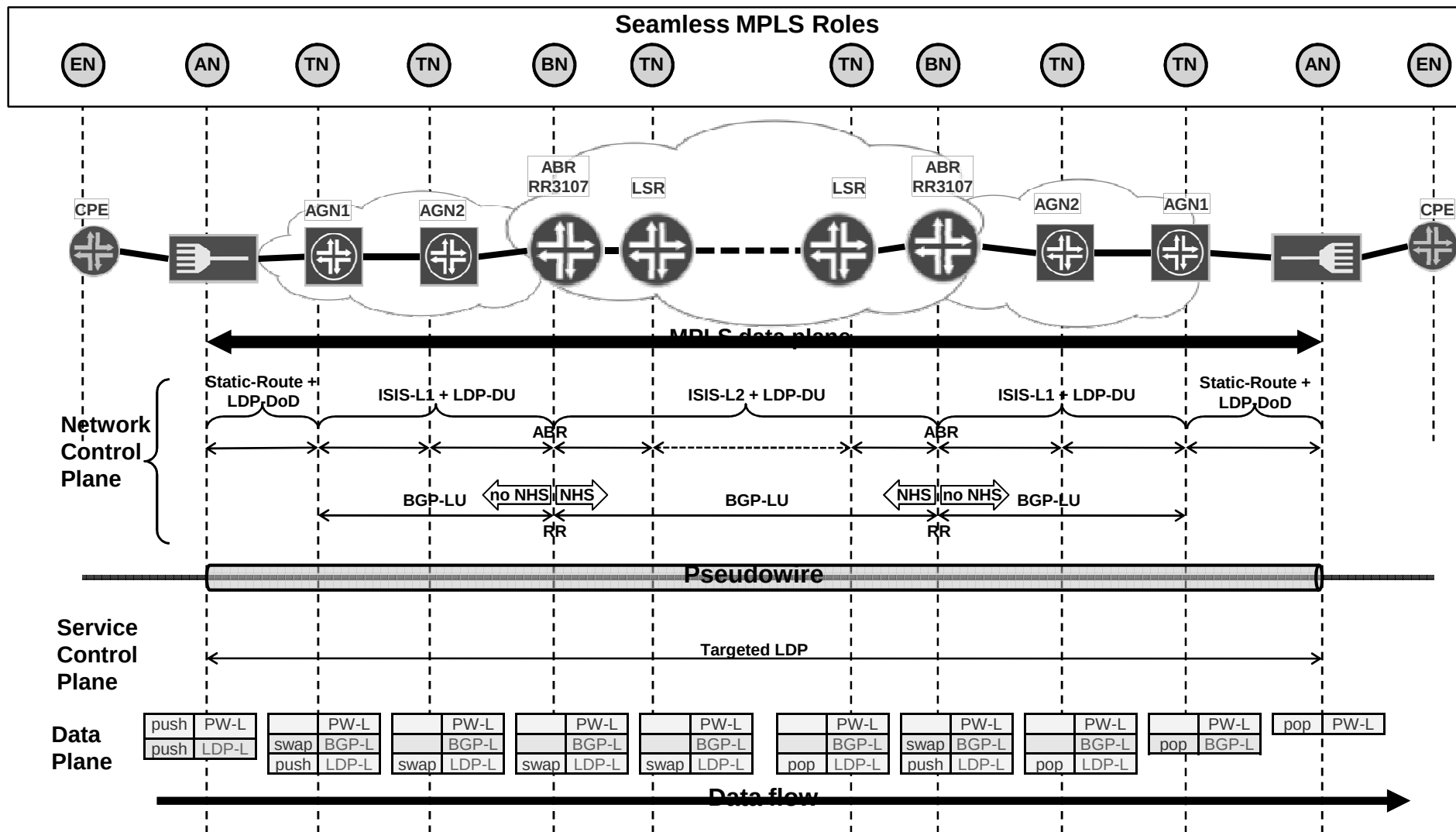
Properties

- Large scale achieved with hierarchical design
- BGP labeled unicast enables any-to-any connectivity between >100k devices – no service dependencies (e.g. no need for PW stitching for base VPWS service)
- A simple MPLS stack on access devices (static routes, LDP DoD)

Seamless MPLS – Use Case

IP/MPLS Network Infrastructure

LDP DoD – LDP Downstream on Demand, RFC5036
 LDP DU – LDP Downstream Unsolicited, RFC5036
 BGP LU – BGP Label Unicast, RFC3107
 NHS – BGP next-hop-self



Scale Enablers

LDP Downstream-on-Demand (LDP DoD)

IP/MPLS routers implement LDP Downstream Unsolicited (LDP DU) label distribution

- Advertising MPLS labels for all routes in their RIB
- This is very insufficient for Access Nodes
 - Mostly stub nodes, can rely on static routing and need reachability to a small subset of total routes (labels)

AN requirement addressed with LDP DoD

- LDP DoD enables on-request label distribution ensuring that only required labels are requested, provided and installed

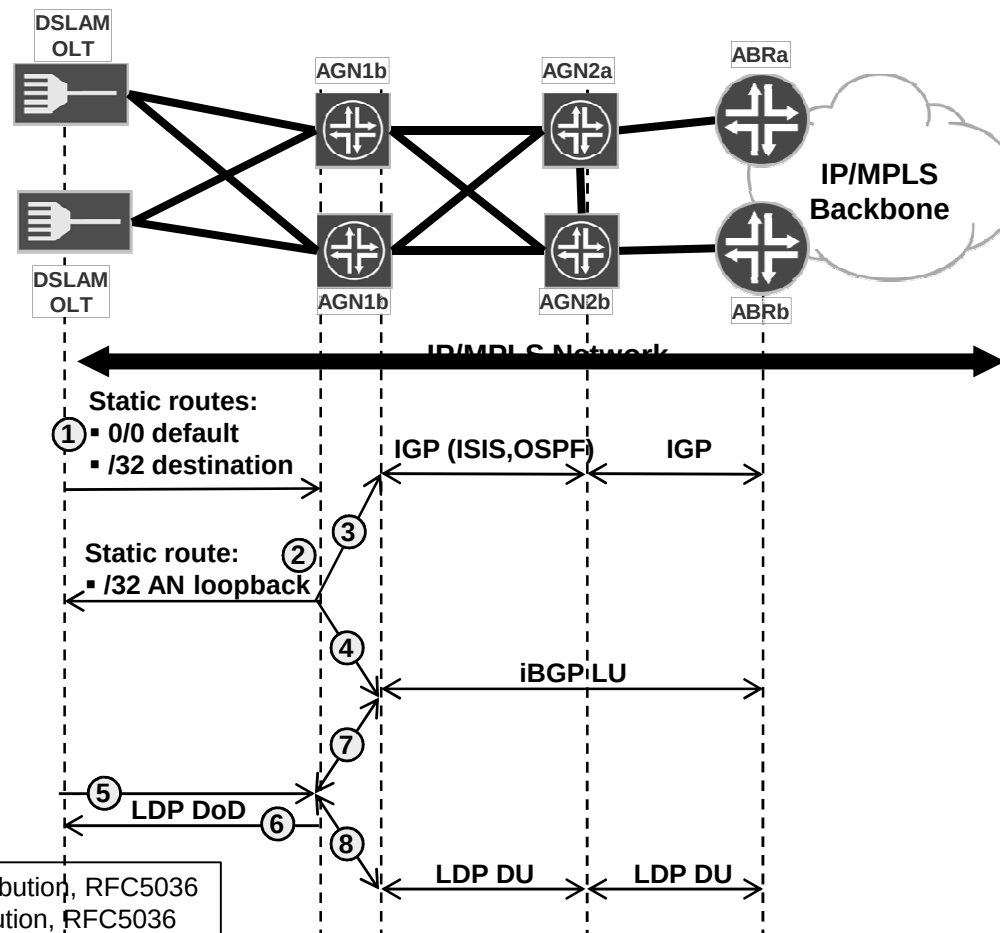
LDP DoD is described in RFC5036

- But not widely available in IP/MPLS routers apart from MPLS over ATM/FR
- This is being fixed now ☺

LDP DoD – Seamless MPLS Use Case Configuration and Operation

- ① **AN:** provisioned static routes
- ② **AGN1:** provisioned static routes
- ③ **AGN1:** statics redistributed into IGP (optional)
- ④ **AGN1:** statics redistributed into BGP-LU
- ⑤ **AN:** LDP DoD lbl mapping requests for FECs associated with /32 static routes and configured services using /32 routes matching default route(*)
- ⑥ **AGN1:** LDP DoD lbl mapping requests for static route /32 FECs
- ⑦ **AGN1:** AN loopbacks advertised in iBGP LU
- ⑧ **AGN1:** if (3) AN loopbacks advertised in LDP DU

LDP DoD – Label Distribution Protocol, Downstream on Demand distribution, RFC5036
 LDP DU – Label Distribution Protocol, Downstream Unsolicited distribution, RFC5036
 BGP LU – Border Gateway Protocol, Label Unicast extensions, RFC3107



(*) Requires LDP support for longest match prefix in RIB (in addition to the exact match) as per RFC5283.

Scale Enablers

BGP Labeled Unicast (RFC3107)

BGP-LU enables distribution of /32 router loopback MPLS FECs

- Used between Seamless MPLS regions for any2any MPLS reachability
- Enables large scale MPLS network with hierarchical LSPs

Not all MPLS FECs have to be installed in the data plane

- Separation of BGP-LU control plane and LFIB
- Only required MPLS FECs are placed in LFIB
 - E.g. on RR BGP-LU FECs with next-hop-self
 - E.g. FECs requested by LDP-DoD by upstream
- Enables scalability with minimum impact on data plane resources – use what you need approach

End-to-End Restoration

Local vs. Global Repair

Local-repair

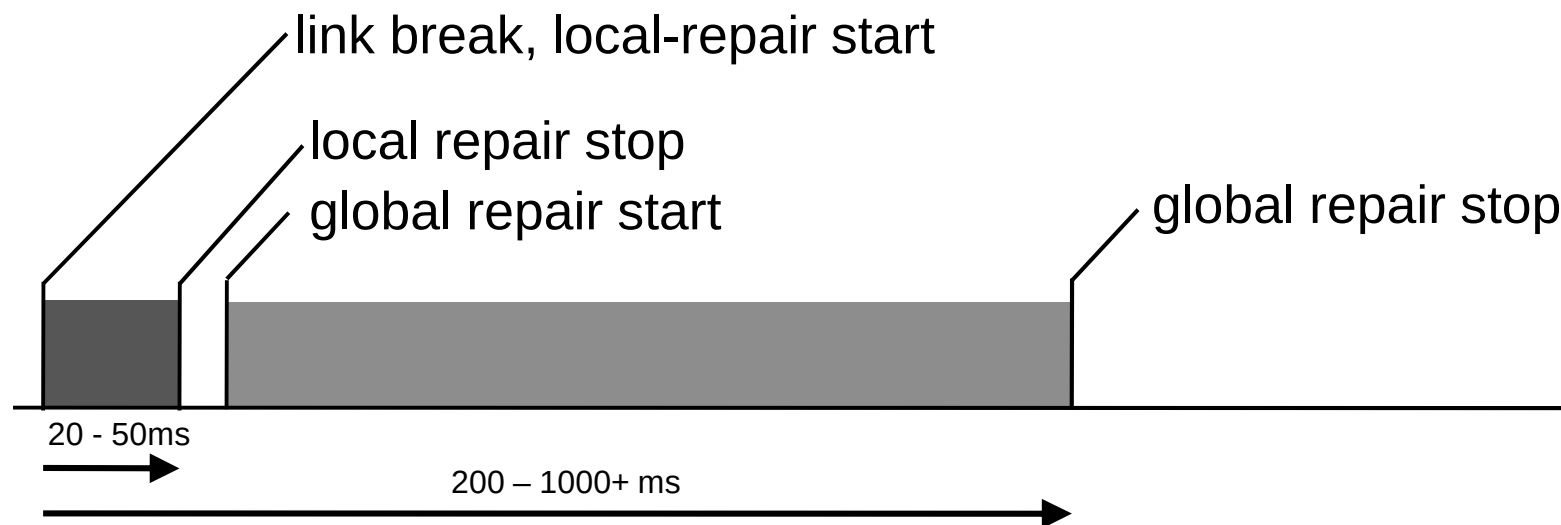
- Based on the pre-computed local backup forwarding state - provides sub-50msec restoration

Global-repair

- Requires signaling to take place after failure detection - can provide sub-1sec or longer restoration times

Local-repair *complements* Global-repair

- Local-repair keeps traffic flowing while
- Global-repair gets things right
- Variation of “Make before break”



End-to-End Restoration

IP/MPLS Local-Repair Coverage – 100% Achieved!

Ingress: CE-PE link, PE node failure

- ECMP, LFA

Transit: PE-P, P-P link, P node failure

- LFA based on IGP/LDP; if no 100% LFA coverage, delta with RSVP-TE
- RSVP-TE FRR

Egress: PE-CE link failure

- BGP PE-CE link local protection

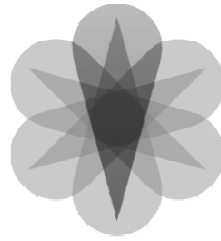
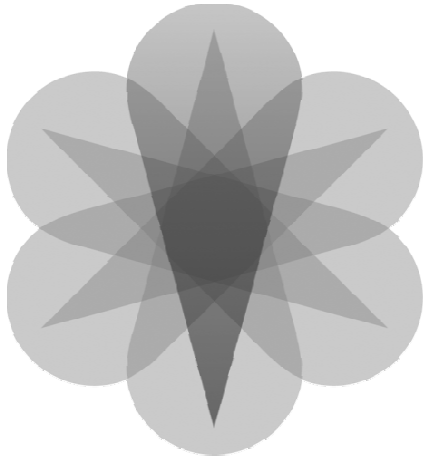
Egress: PE node failure (new)(*)

- LSP tailend protection with context label lookup on the backup PE
- Failure repaired locally by adjacent P router using LFA (or TE-FRR)

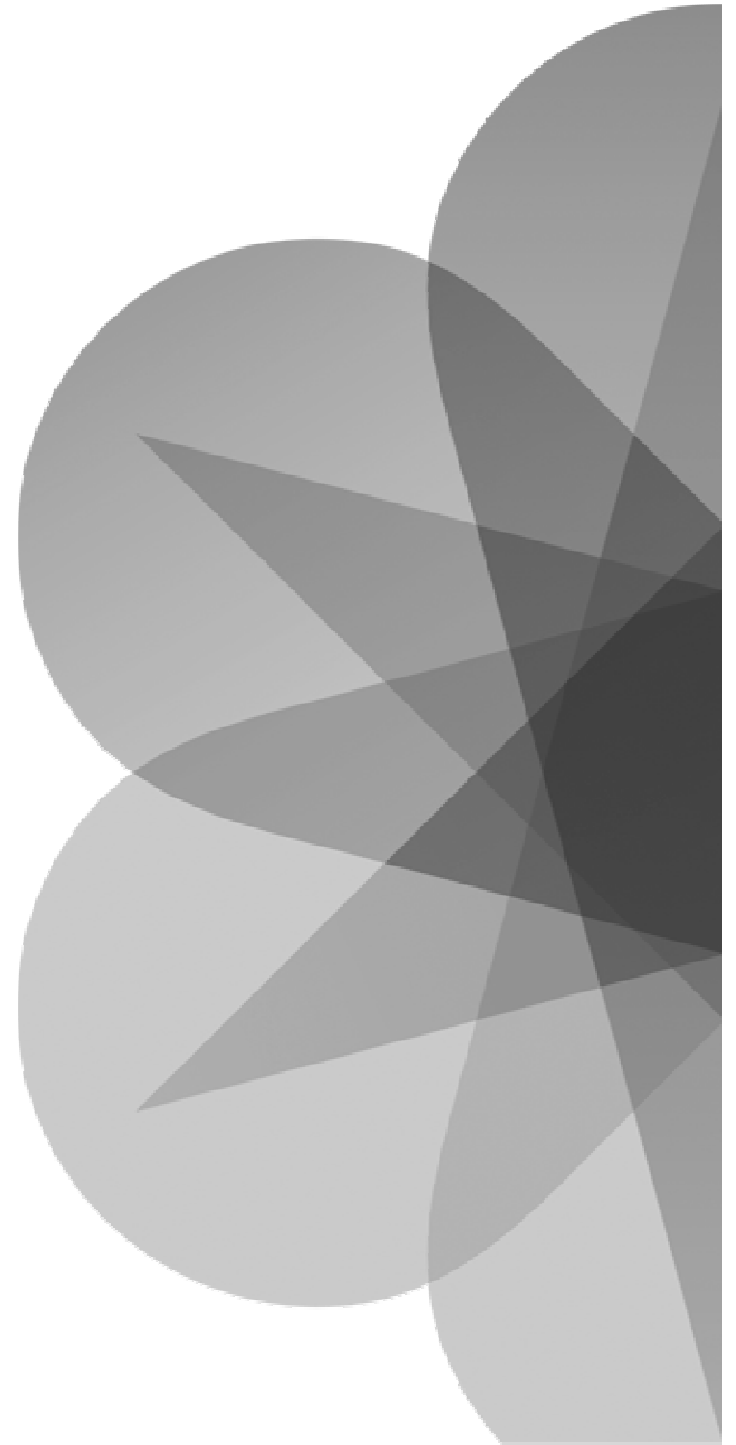
Packet based networks finally can provide E2E service protection similar to SDH 1:1 protection, regardless of network size and service scale

This provides **network layer failure transparency to service layers**, becoming a major enabler for network consolidation

(*) "High Availability for 2547 VPN Service", Y.Rekhter, MPLS&Ethernet World Congress, Paris 2011.



Conclusion



In Conclusion...

Seamless MPLS approach addresses all key requirements for converged packet network design

- Support for all packet services across fixed, mobile, business, residential, wholesale
- Support for large scale incl. high number of low end access devices
- E2E fast restoration sub-50msec for all network failures
- Simplified service delivery with flexible topological placement

Useful References

Seamless MPLS Architecture

- <http://tools.ietf.org/html/draft-ietf-mpls-seamless-mpls-00>

RFC 3107: Carrying Label Information in BGP-4

- <http://www.ietf.org/rfc/rfc3107.txt>

LDP Downstream-on-Demand in Seamless MPLS

- <http://tools.ietf.org/html/draft-ietf-mpls-ldp-dod-00>

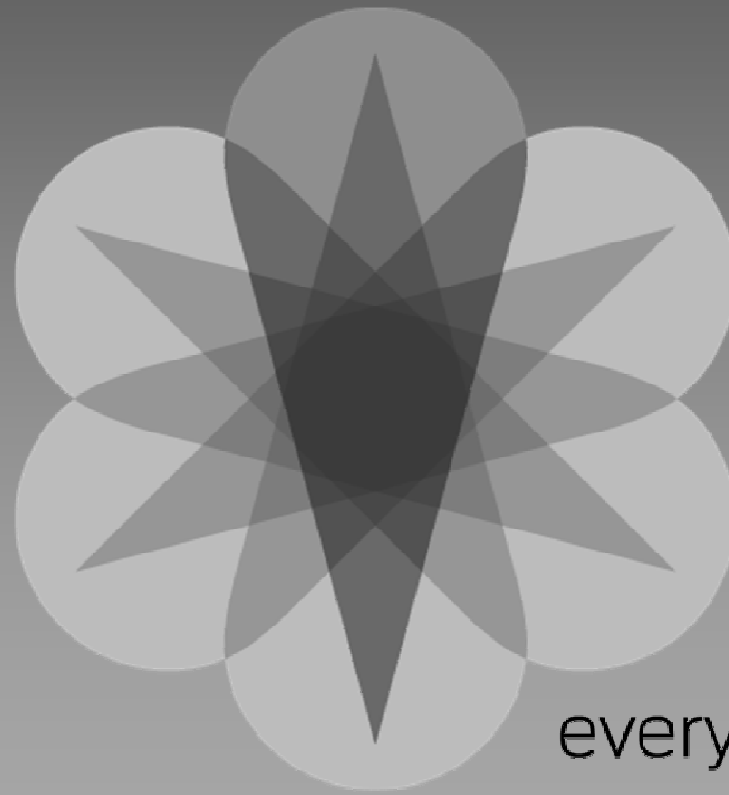
Inter-Area P2MP Segmented LSPs

- <http://tools.ietf.org/html/draft-raggarwa-mpls-seamless-mcast-03>

Seamless MPLS Whitepapers

- <http://www.juniper.net/us/en/local/pdf/whitepapers/2000316-en.pdf>
- <http://www.juniper.net/us/en/local/pdf/design-guides/8020013-en.pdf>

MPLS Enabled Applications, Third Edition, Chapter 16: MPLS in Access Networks and Seamless MPLS



everywhere