Multicast Within SR-MPLS
A Comparative Review

IETF 103 PIM WG

Ian Duncan - Presenter

iduncan@ciena.com

David Allan – Collaborator

david.i.allan@ericsson.com

On Multicast

Some Axiomatic Truisms

1. We have operational network scenarios making good use of multicast
2. Further, efficient multicast has value — efficiency defined roughly as "least cost" traffic replication
3. Combined sets of unicast & multicast flows often used to provide a single unified service function
4. Multicast can be hard

Established Multicast Protocols

- Fine assortment too numerous to enumerate all
  - BIER has obvious benefits, except its data plane
  - mLDP is cool but requires LDP & all that that implies
  - RSVP-TE PtMP even weightier implications
  - PIM is PIM, and independent of SR & MPLS
  - mOSPF provides helpful historical precedent
  - And so does 802.1aq – SPB/M
- draft-zzhang-pim-sr-multicast "Summary" states
  - "BIER is the best choice" (but for that data plane)
  - if "efficient multicast replication is required, then run mLDP/RSVP-TE/PIM"
  - else "use static configuration or controller signaling"
  - as answers, these seem inadequate to requirements

On Multicast & SR-MPLS

We Like BIER … Except

- No upgrade potential on existing SR-MPLS capable systems means "forklift" of legacy
- Merchant ASICs have no BIER support
  - Nearly none of current available switches
  - "BIER capable" is 3-5.5x cost & 4x Watts (per 10GE)
- MPLS data plane motivation is simple economics

For MPLS Segment Routing Multicast

- neither mLDP or RSVP-TE PtMP seems ideal
  - operational utility limited to coordination only
  - requiring parallel protocols with independent DBs, diagnostics, conceptual models & more
  - however both provide template for how best to integrate inscribing paths for multicast
- Tree-SID is fine as static replication anchor, so well suited to specific SDN models (only)
- Spray is edge replication replicated once again; i.e., not network multicast

Dynamic Efficient SR-MPLS Multicast

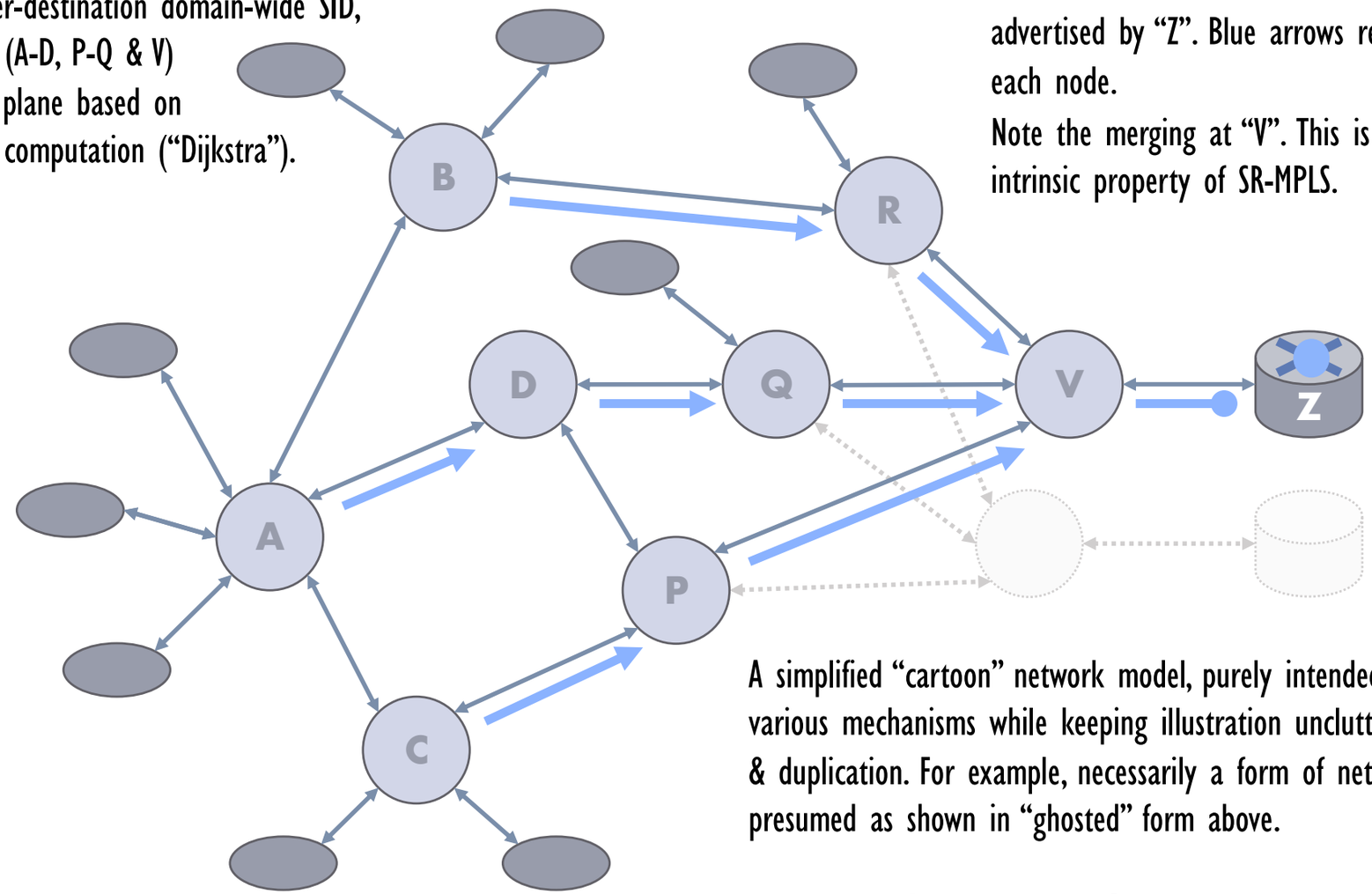
- per-tree data plane state is an agreeable cost
- synchronized flooded multicast state is also a desirable & viable burden
- Unified control plane providing congruence with unicast SR forwarding & control
- Operational simplicity & potential alignment for proactive OAM
- LS IGP providing "computed trees"
  - group membership state stable & constant during topology changes
  - all join/leave processing well-ordered

Substantial Architectural Re-use

- A merging of two complementary IETF protocol designs
  - stock RFC 3032 data plane, as used today in LDP/mLDP, TE PtP/PtMP & SR
  - RFC 6329 "IS-IS Support of IEEE 802.1aq"

A Unicast “SPF” Baseline

Unicast SR-MPLS relies on per-destination domain-wide SID, flooded via IGP. Each node (A-D, P-Q & V) locally installs label in data plane based on conventional “shortest path” computation (“Dijkstra”).



Traffic toward “Z" follows LSP based on the “blue” SID advertised by “Z”. Blue arrows represent link “next hop” on each node.

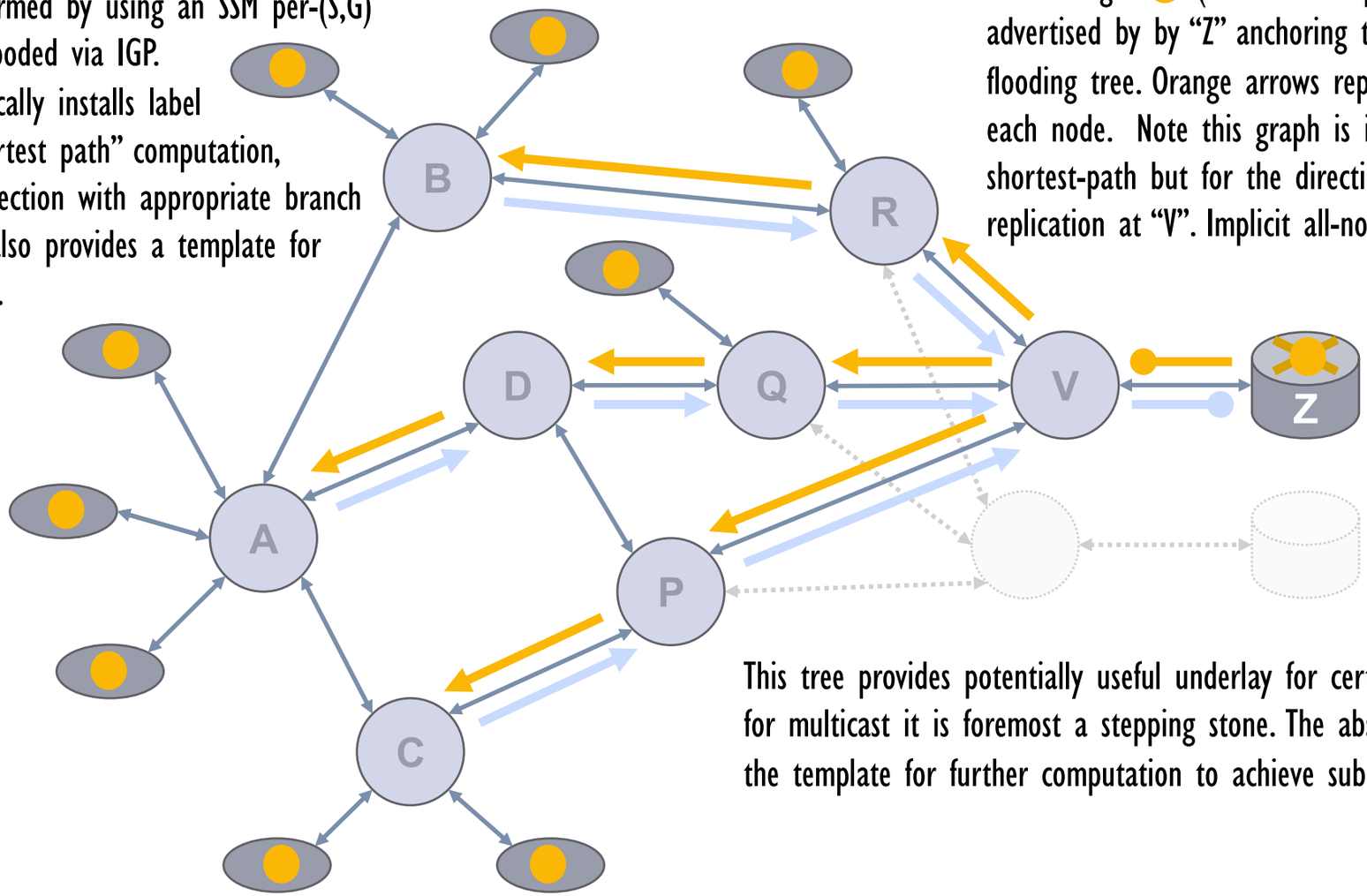
Note the merging at “V”. This is inherent to sink-trees, an intrinsic property of SR-MPLS.

A simplified “cartoon” network model, purely intended to show detail of the various mechanisms while keeping illustration uncluttered by complications & duplication. For example, necessarily a form of network resilience is presumed as shown in “ghosted” form above.

This is simple SR-MPLS

Broadcast Tree & "Template" Tree

Broadcast SR-MPLS tree is formed by using an SSM per-(S,G) domain-wide SID, similarly flooded via IGP.
Each node (A-D, P-Q & V) locally installs label in data plane based on "shortest path" computation, but in reverse-forwarding direction with appropriate branch replication. In abstract, this also provides a template for the next levels of illustration.

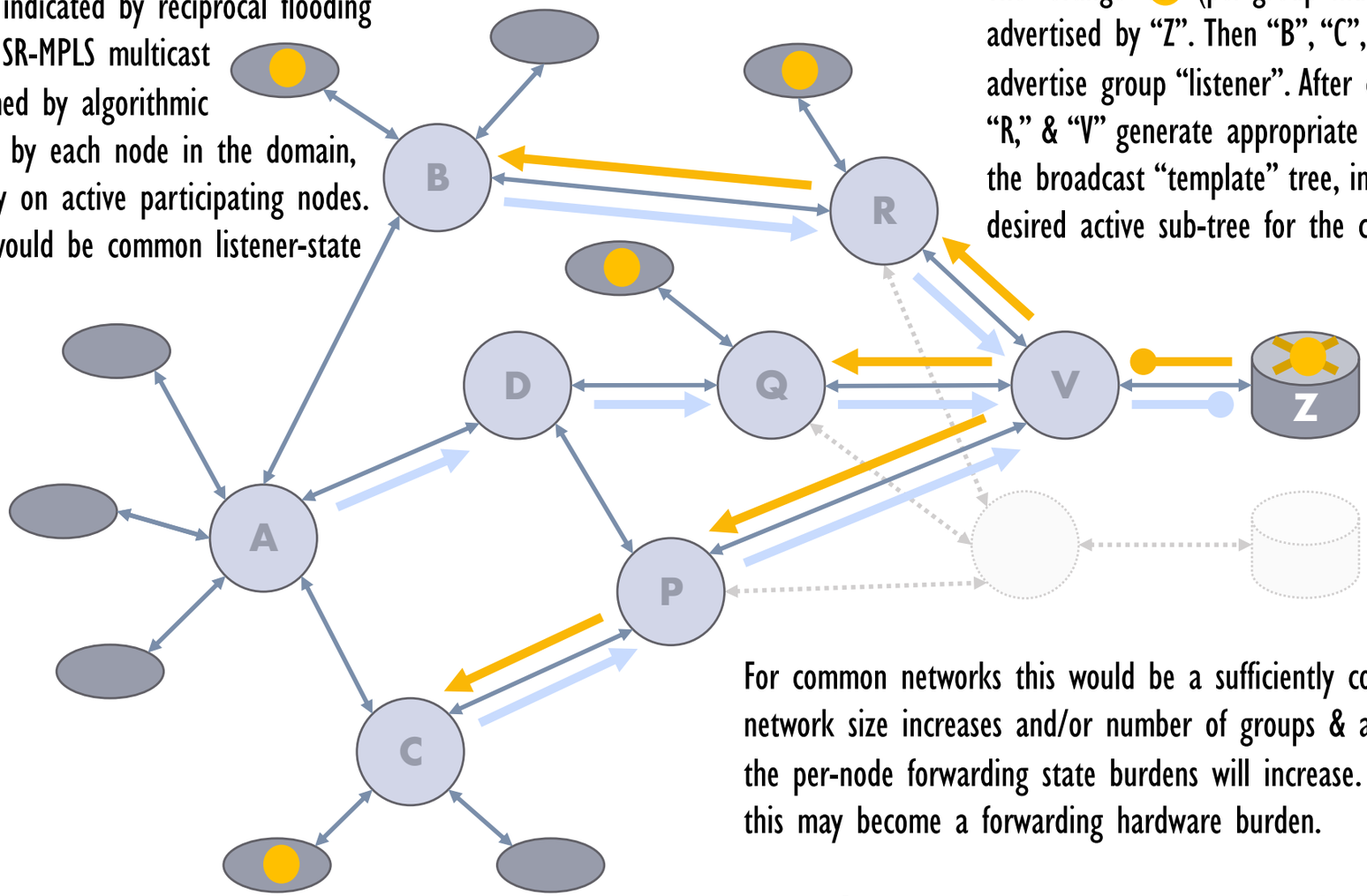


The "orange" SID (a "multicast-specific" from SRGB) is advertised by by "Z" anchoring the source for a broadcast flooding tree. Orange arrows represent link "next hop" via each node. Note this graph is isomorphic to unicast shortest-path but for the direction of arrows, with a branch replication at "V". Implicit all-nodes flooding membership.

This tree provides potentially useful underlay for certain services. However, for multicast it is foremost a stepping stone. The abstract graph is key as the template for further computation to achieve sub-graphs for multicast.

Simple Model Multicast

The "listener" membership is indicated by reciprocal flooding of the group SID. The active SR-MPLS multicast group sub-graph tree is formed by algorithmic pruning of the template tree by each node in the domain, with label state installed only on active participating nodes. The MLD & IGMP protocols would be common listener-state inputs.



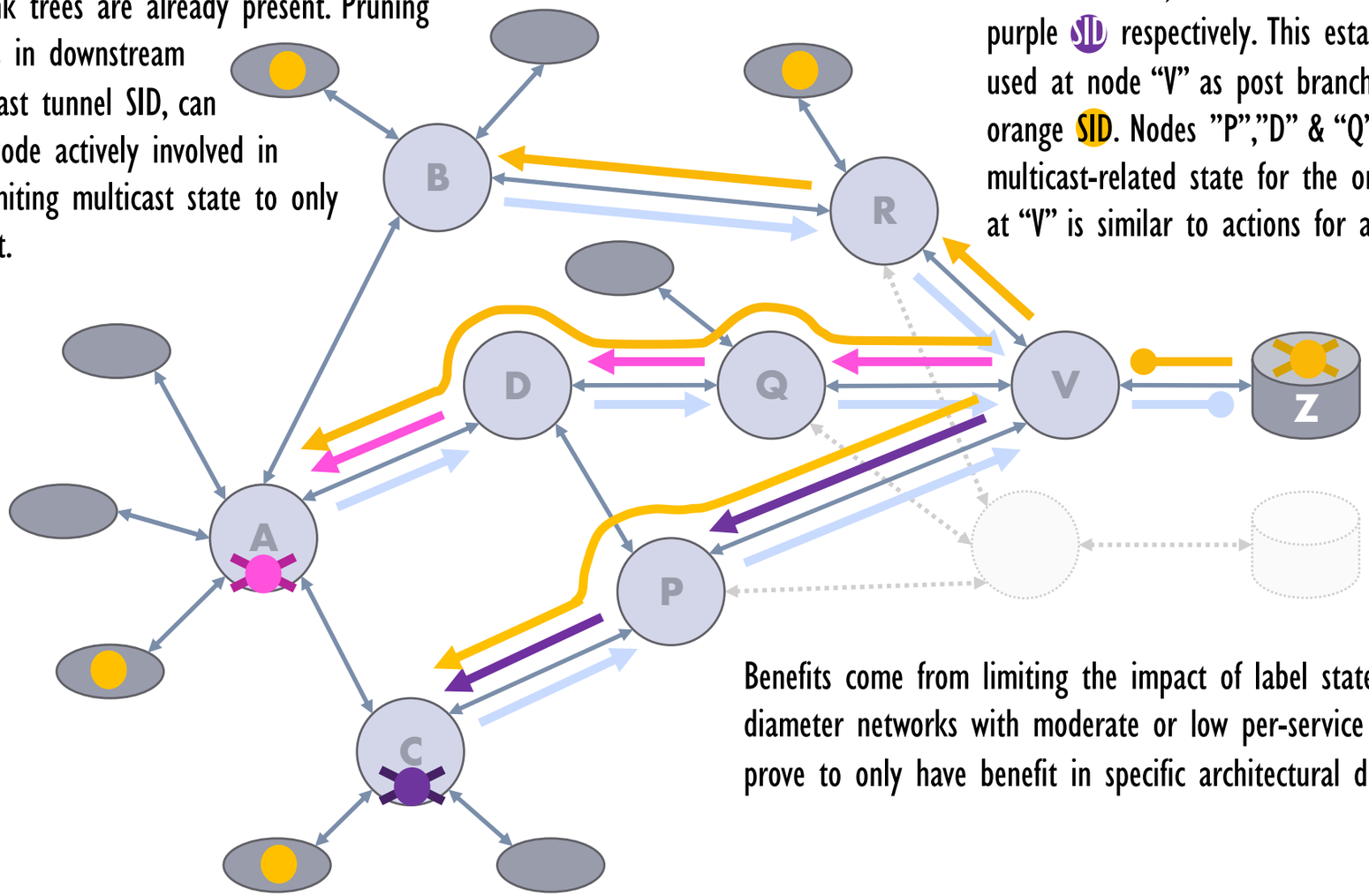
The "orange" SID (per-group multicast-specific from SRGB) is advertised by "Z". Then "B", "C", "Q", & "R" advertise advertise group "listener". After computing, "B", "C", "P", "Q", "R," & "V" generate appropriate forwarding state by pruning the broadcast "template" tree, installing & maintaining the desired active sub-tree for the current listeners.

For common networks this would be a sufficiently complete solution. As the network size increases and/or number of groups & active leaves increases, the per-node forwarding state burdens will increase. At some levels of scale this may become a forwarding hardware burden.

SR Multicast Intuition Extended

Multicast Tree With Unicast Tunnels

Unicast SR-MPLS per-node sink trees are already present. Pruning will identify transparent hops in downstream branches, and using the unicast tunnel SID, can provide bypass to the next node actively involved in multicast forwarding. Thus limiting multicast state to only those nodes truly requiring it.



In base unicast, nodes “A” & “C” advertise a pink SID & purple SID respectively. This established infrastructure gets used at node “V” as post branch-replication “next hop” for orange SID. Nodes ”P”,”D” & “Q” are unburdened by any multicast-related state for the orange tree. The forwarding at “V” is similar to actions for a binding SID.

Benefits come from limiting the impact of label state scale in larger diameter networks with moderate or low per-service leaf count. This may prove to only have benefit in specific architectural deployment scenarios.

SR Multicast Per Current IETF I-D

RFC 1584 (c.1994)
Multicast Open Shortest Path First

NIRVANA

An Era When

- A software upgrade was actually hardware
- A Casio watch was almost cool
- A pager was crucial for those operationally responsible
- A 33 MHz 32 bit CPU & 128KB of DRAM really was cool

Past Mythology Around Scale Long Obsolescent

- 1.4 GHz i64 x86 Quad Core & 16GB DDR4 is commonplace
- > 30Gb of combined control CPU I/O (2x 10GE + PCIe gen3)
- Generations of Moore's law later O(logN) bit less worrying
- Larger IETF IGP process 'footprint' readily accommodated

Summary & Conclusion

- No interest in "boiling the ocean"
  - This actually seems relatively easy (for a multicast project)
  - We love BIER — on most existing/deployed routers & current merchant ASICs using SR-MPLS is our only viable option
  - Remember, there is multi-topology should degrees of isolation be a concern
- Request WG help formulating how to proceed on suitable capture in I-Ds
  - previously shared <draft-allan-pim-sr-mpls-multicast-framework> should be considered a useful starting point
  - open to starting afresh with separate & distinct I-Ds
    - one to cover requirements & general architecture
    - another (perhaps few) to capture architectural specifics such as flex-algo & BIER interworking

And thanks to all for your ongoing feedback & advice