THE NEXT SPRING FORWARD.

HOW DO WE SUCCESSFULLY EVOLVE TODAY'S IP/MPLS NETWORKS?

ROB SHAKIR
IP ARCHITECT
GOOGLE, INC.

PARIS, FRANCE – MARCH 2018.

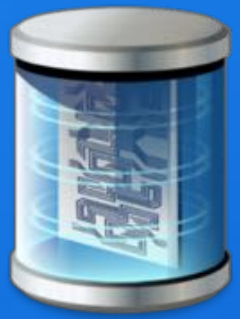
HI!

ROB SHAKIR

IP ARCHITECT
GOOGLE

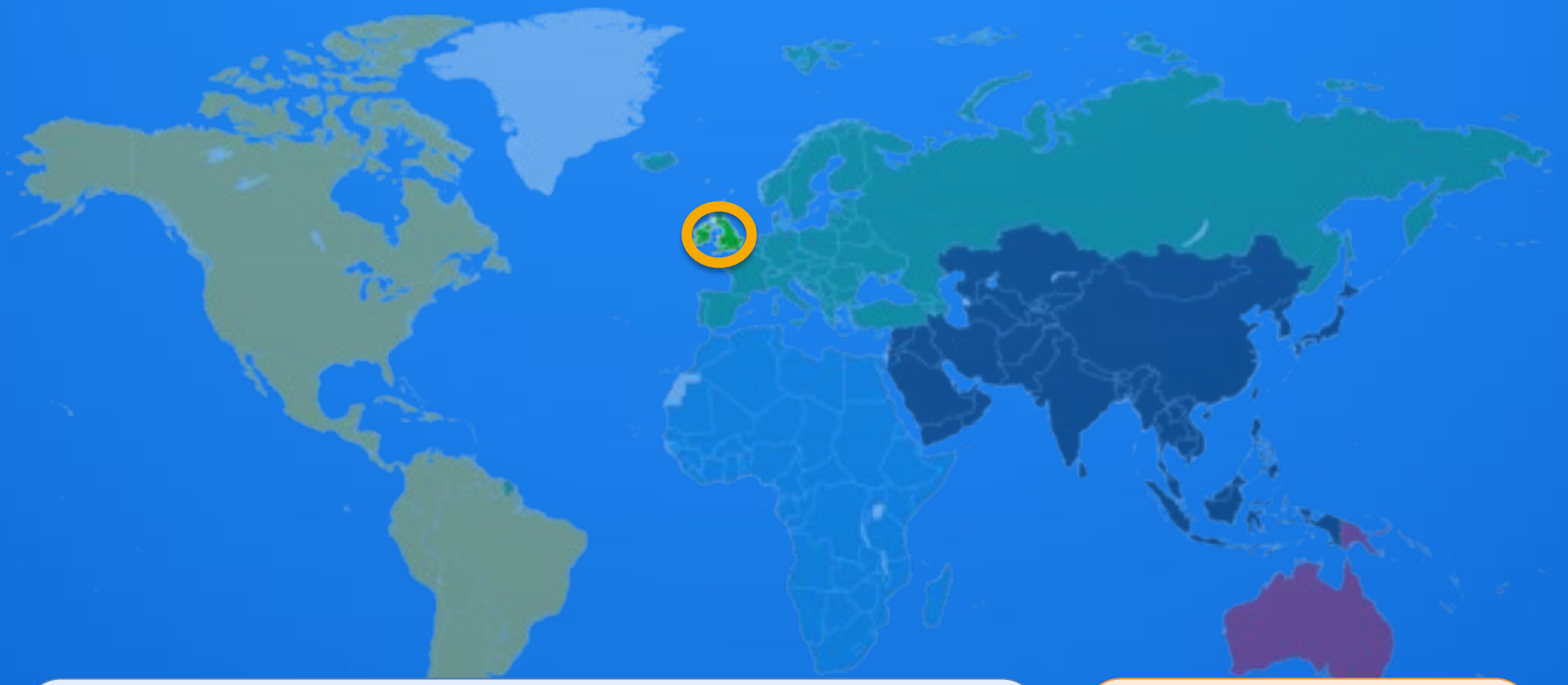
Network analysis and planning

Automated network operations



Software-driven infrastructure

IN THE BEGINNING.



catalyst2

Small peering network offering content hosting and managed services.

3 UK PoPs,
3 staff,
<50Mbps traffic!

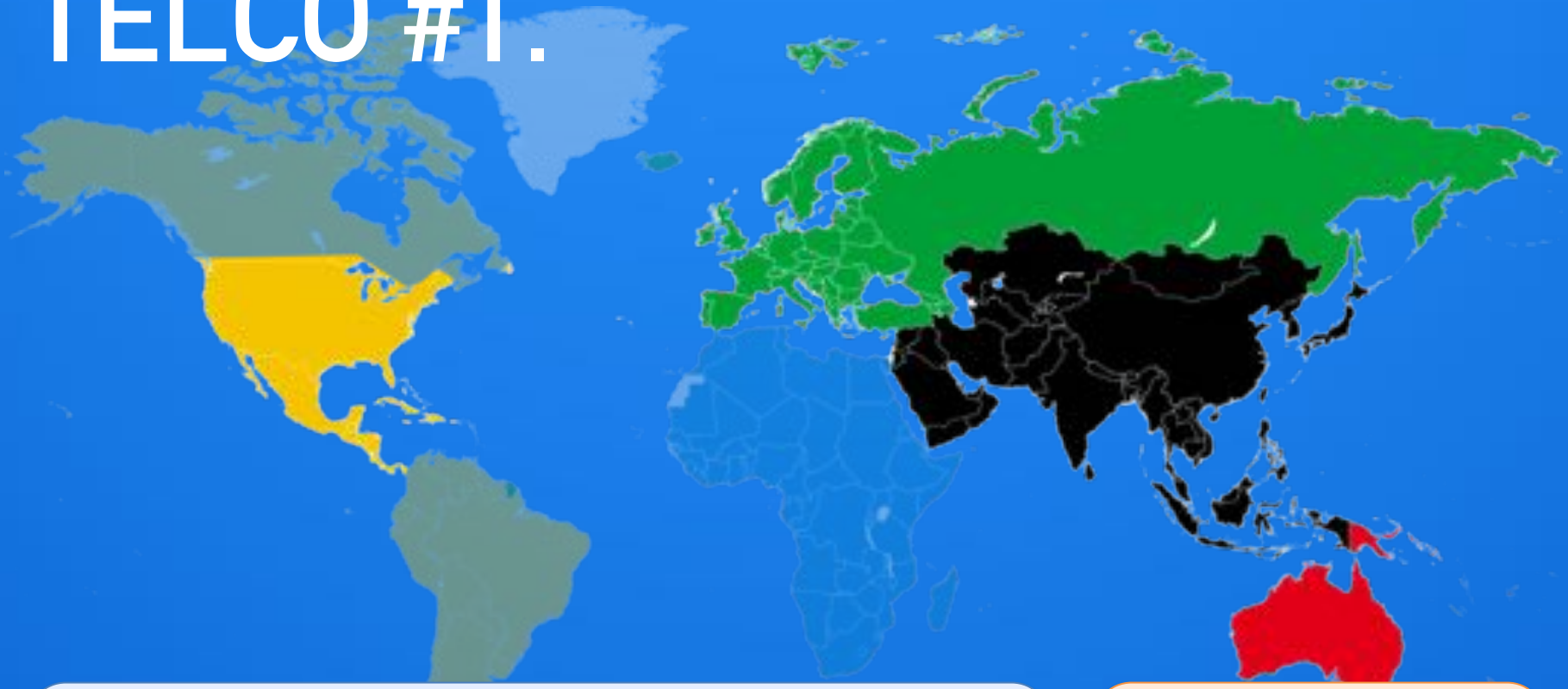
DISCOVERING LEGACY.

pipex

NSP offering Internet, data, voice and hosting services – lots of acquisitions and legacy!

20+ EU PoPs,
250 staff,
10s of Gbps traffic.

GOING GLOBAL – TELCO #1.



Cable&Wireless Worldwide

Global telco offering transport, IP/MPLS, voice services – focus on multi-service & IP convergence.

200+ PoPs,
~10,000 staff,
100s of Gbps traffic.

THE INCUMBENT…

Many strategic and legacy networks with O(10,000s) elements – spread over ~6,000 PoPs in 180 countries.

Delivering private IP and public Internet services – edge capacity of terabits/second peak.

BT

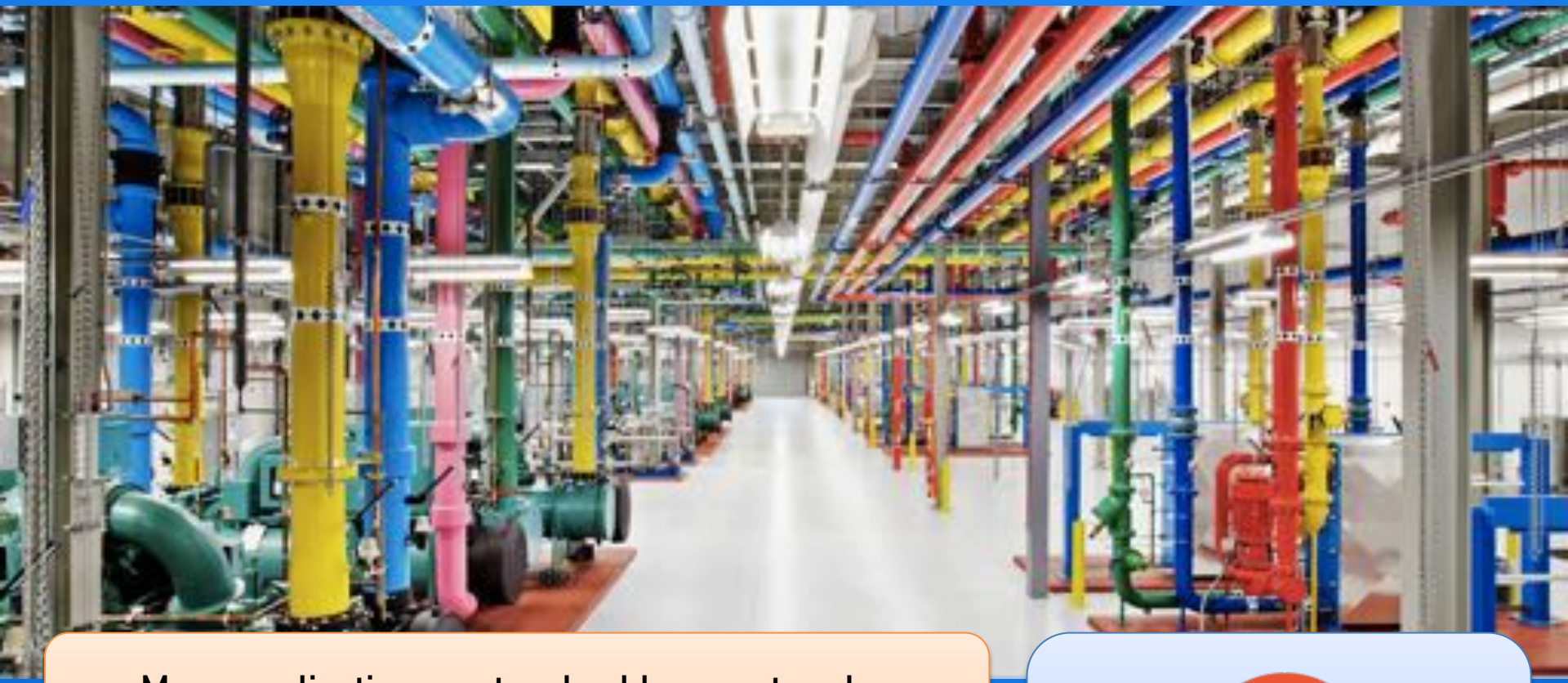
MOVING UP THE STACK!

Distributed platform – backbone is overlay on telco infrastructure and the Internet.

Automated application-aware infrastructure.

JIVE

TODAY…



Many applications on two backbone networks.

Large scale SDN and automation.

“Web Scale” infrastructure.

GOOGLE'S BACKBONES.

B2 – USER FACING BACKBONE
HYBRID CONTROL PLANE

B4 – INTER-DC BACKBONE
CENTRALISED "SDN" NETWORK

INTERNAL TRAFFIC DOMINATES VOLUME.
GOOGLE (B2) ESTIMATED TO BE 25-30% OF TOTAL INTERNET TRAFFIC.

NETWORKS AT SCALE.

>30,000 OPERATIONAL CIRCUITS

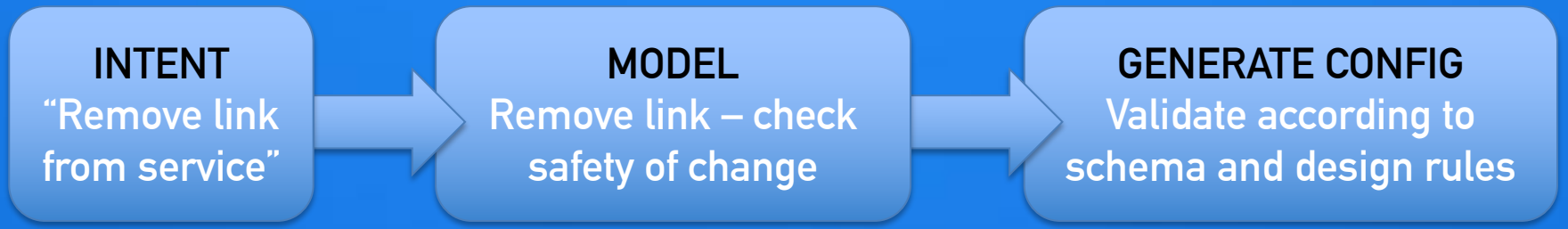
LINES OF CONFIGURATION >4M

>1000 CHANGES PER DAY

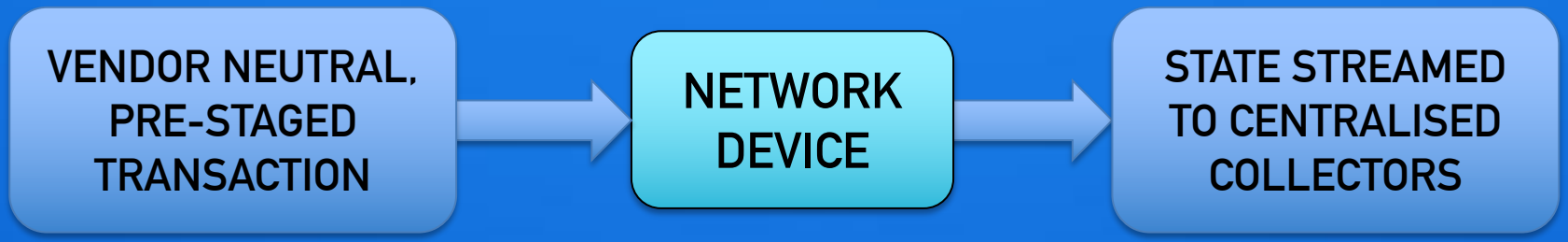
TELEMETRY: >300K UPDATES/S

WE CANNOT MANAGE NETWORKS WITH HUMANS.
OUR CURRENT TECHNOLOGIES DO NOT SCALE TO MEET MODERN NETWORKS.

MODERN MANAGEMENT PLANE.

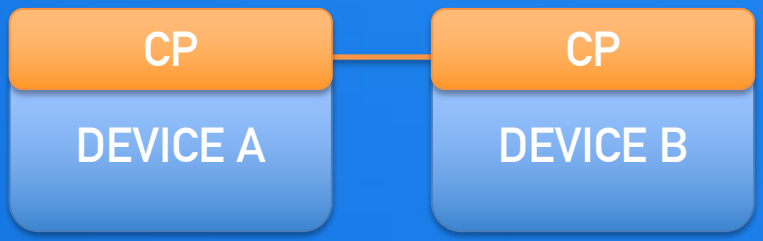


TAKING HIGH-LEVEL NETWORK PLAN AND AUTOMATICALLY REALISING THIS IN A SAFE MANNER IN THE NETWORK.

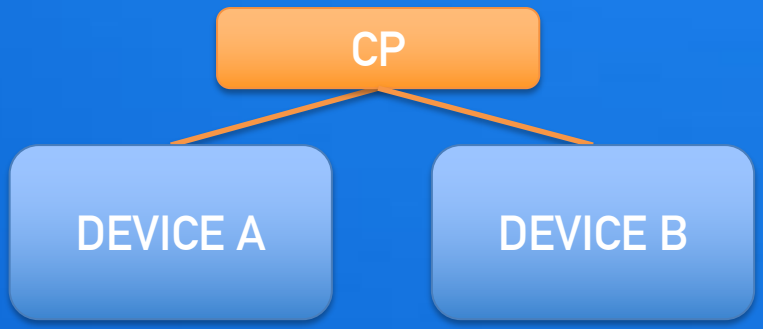


A REAL API TO THE NETWORK DEVICE – NOT HUMANS TYPING.
TELEMETRY SCALING TO ENABLE ON/OFF BOX CONTROL.

WHERE DOES THE CONTROL PLANE LIVE?



COMPLETELY ON BOX.
DISTRIBUTED COMPUTATION.
HIGHLY SURVIVABLE.
COMPLEX API REQUIREMENTS.



COMPLETELY OFF BOX.
CENTRALISED COMPUTATION.
HIGHLY FLEXIBLE.
SIMPLE API/DUMB DEVICES.

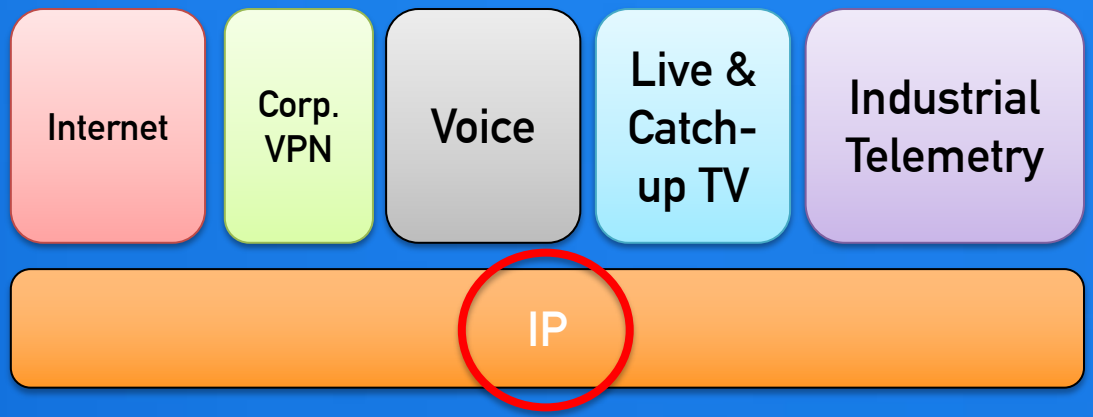
WE REALLY REQUIRE A HYBRID – ON-BOX SURVIVABILITY WITH SIMPLE API.
HOW SHOULD THIS BE DEFINED?

WHY TRAFFIC ENGINEERING?

COMPLEX, GLOBAL SYSTEM – WITH DIFFERENT APPLICATIONS (SEARCH, CLOUD, ADS, ETC.)

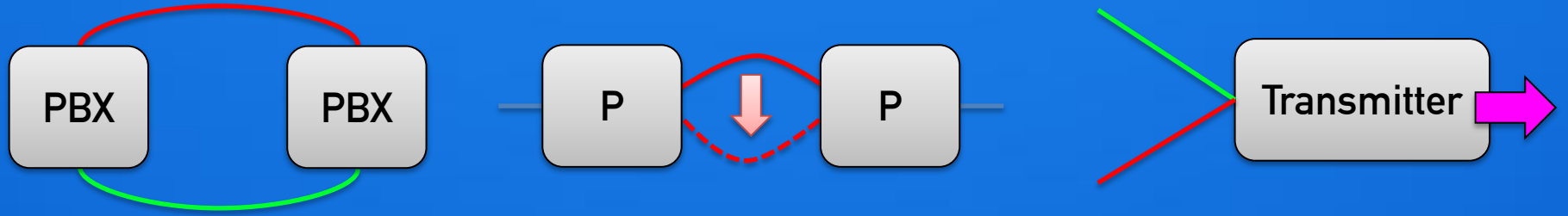
GROWING AT A HUGE RATE – EFFICIENCY IS EXCEPTIONALLY IMPORTANT.

COMMON THEME: EVERYTHING IS IP.



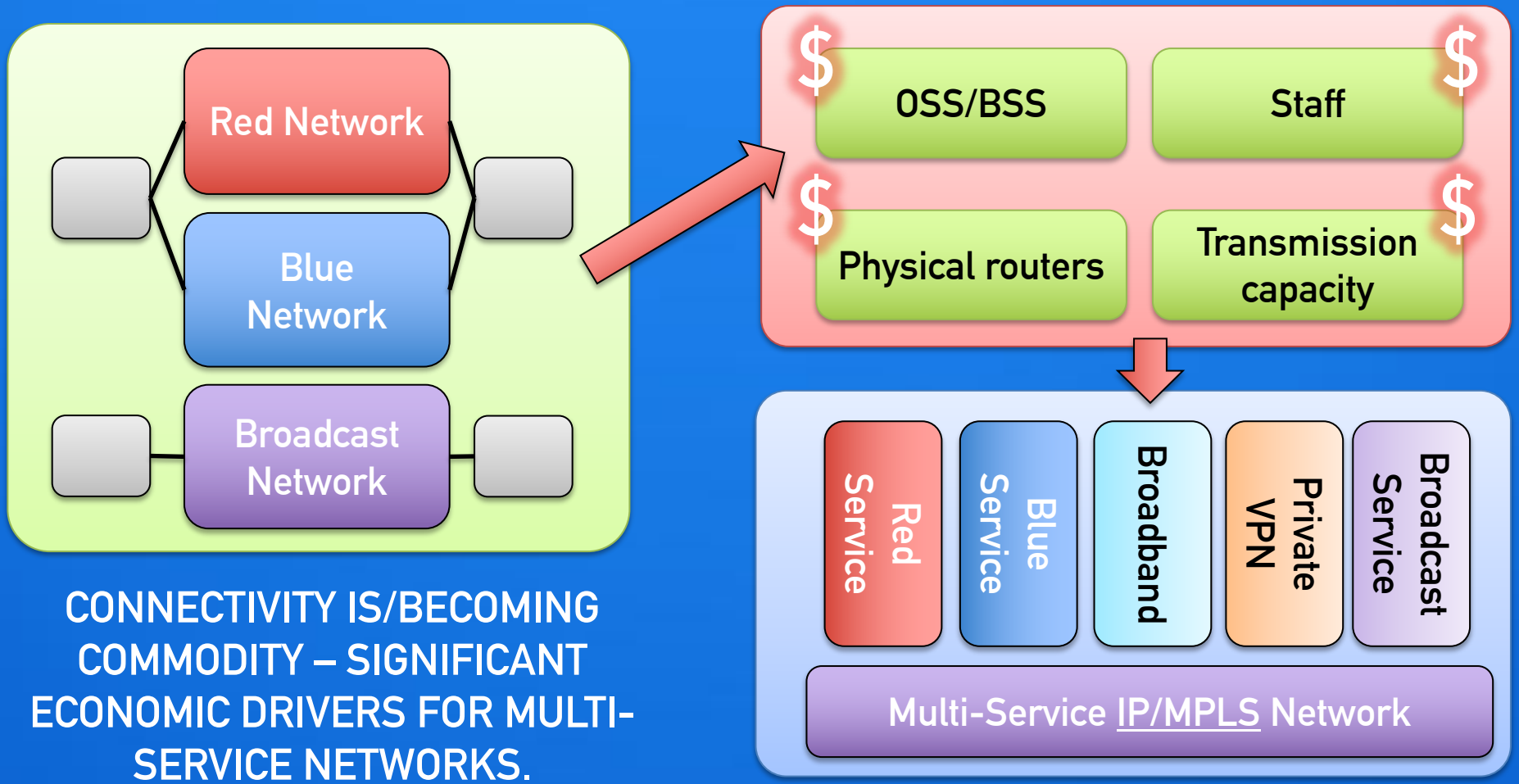
MANY APPLICATIONS WITH DIFFERENT TRANSPORT REQUIREMENTS – ALL ON TOP OF IP NETWORKS.

How does this impact engineering of the IP layer?



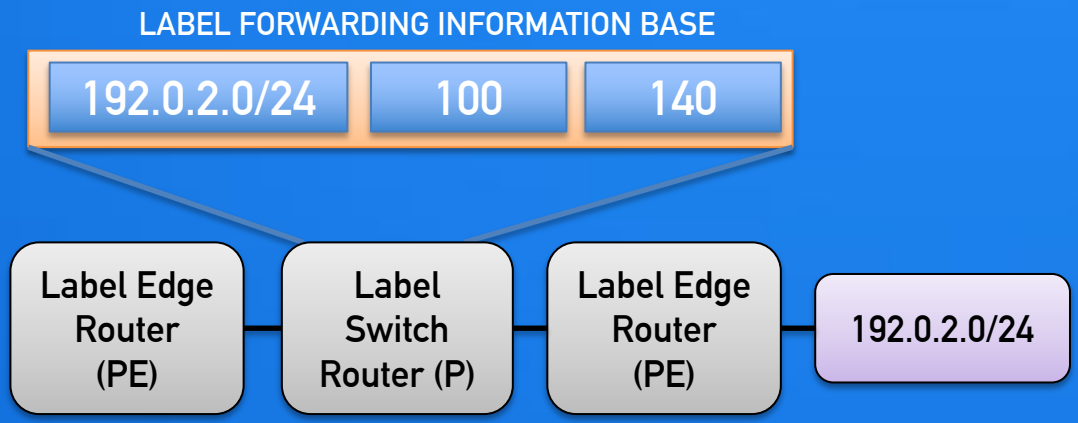
APPLICATIONS EXPECT LEGACY NETWORKING BEHAVIOUR – INTRODUCING NEW PATH ROUTING AND PERFORMANCE REQUIREMENTS TO IP.

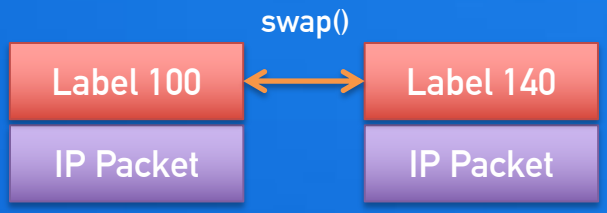
COMMON THEME: MULTI-SERVICE NETWORKS.

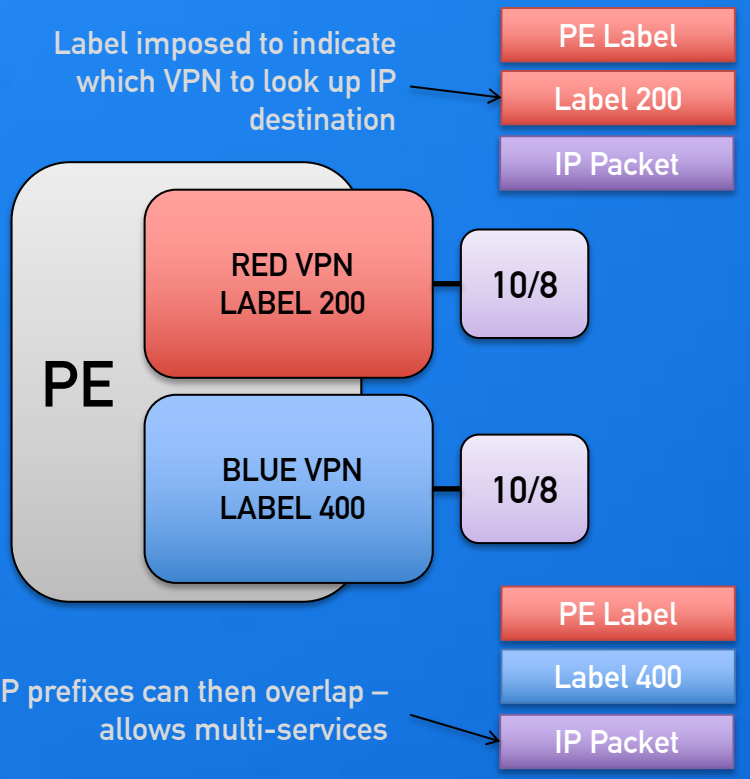


CONNECTIVITY IS/BECOMING COMMODITY – SIGNIFICANT ECONOMIC DRIVERS FOR MULTI-SERVICE NETWORKS.

MULTI-PROTOCOL LABEL SWITCHING?



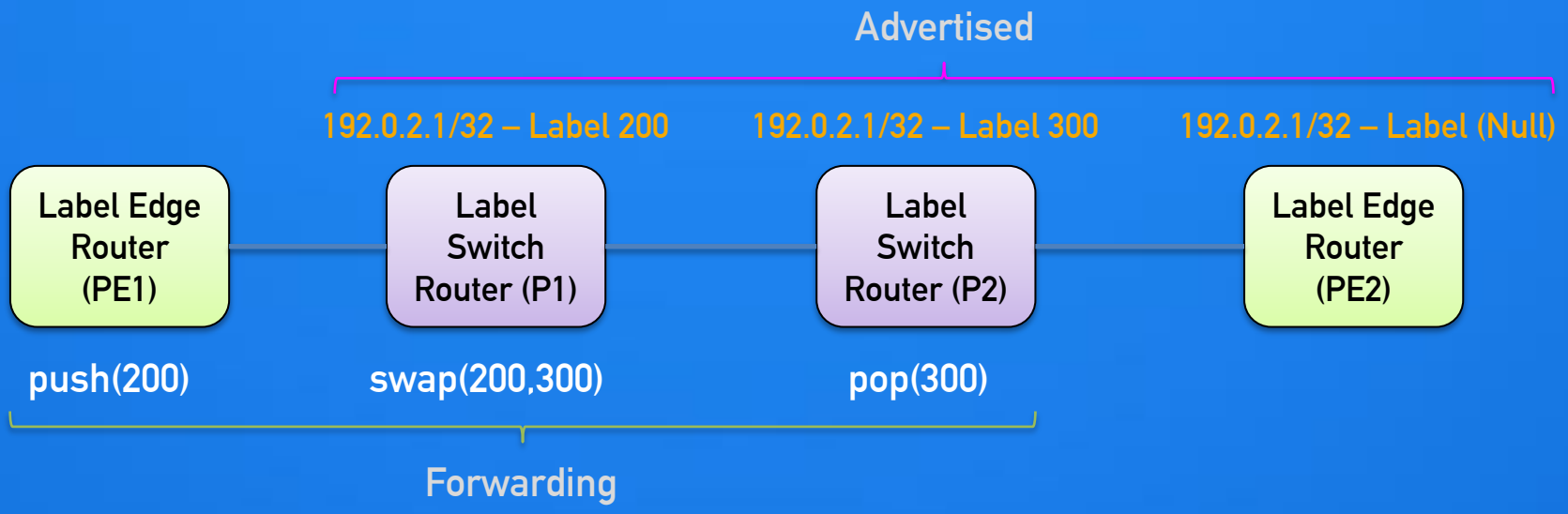


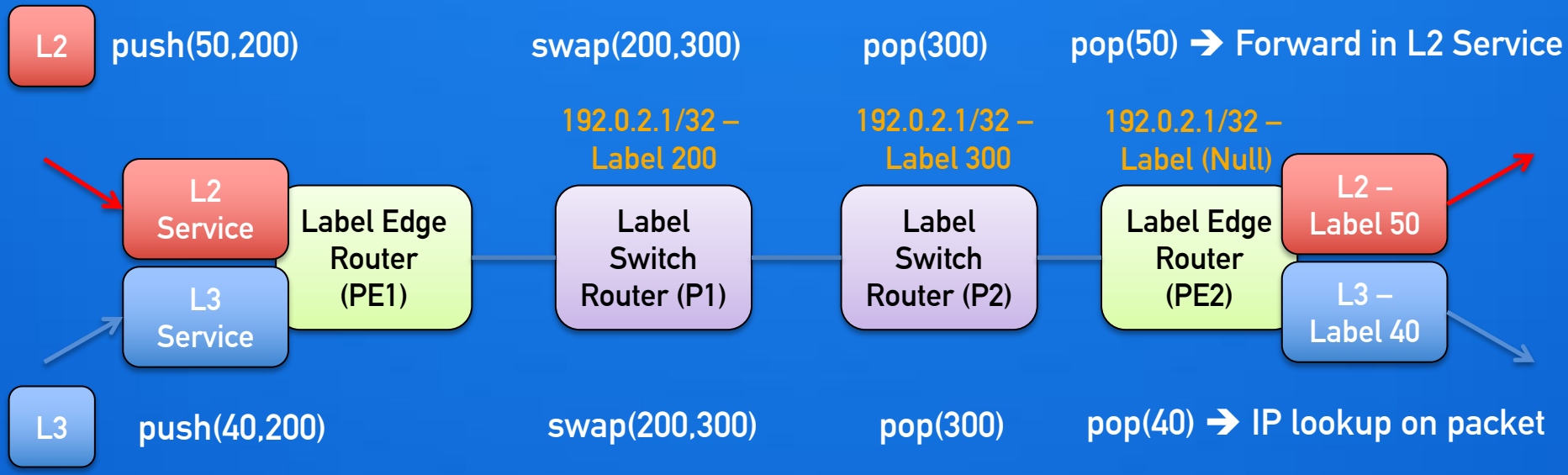


FORWARDING BASED ON LABEL INFORMATION RATHER THAN IP DESTINATION – IP TUNNELLED INSIDE MPLS PATH (FEC).

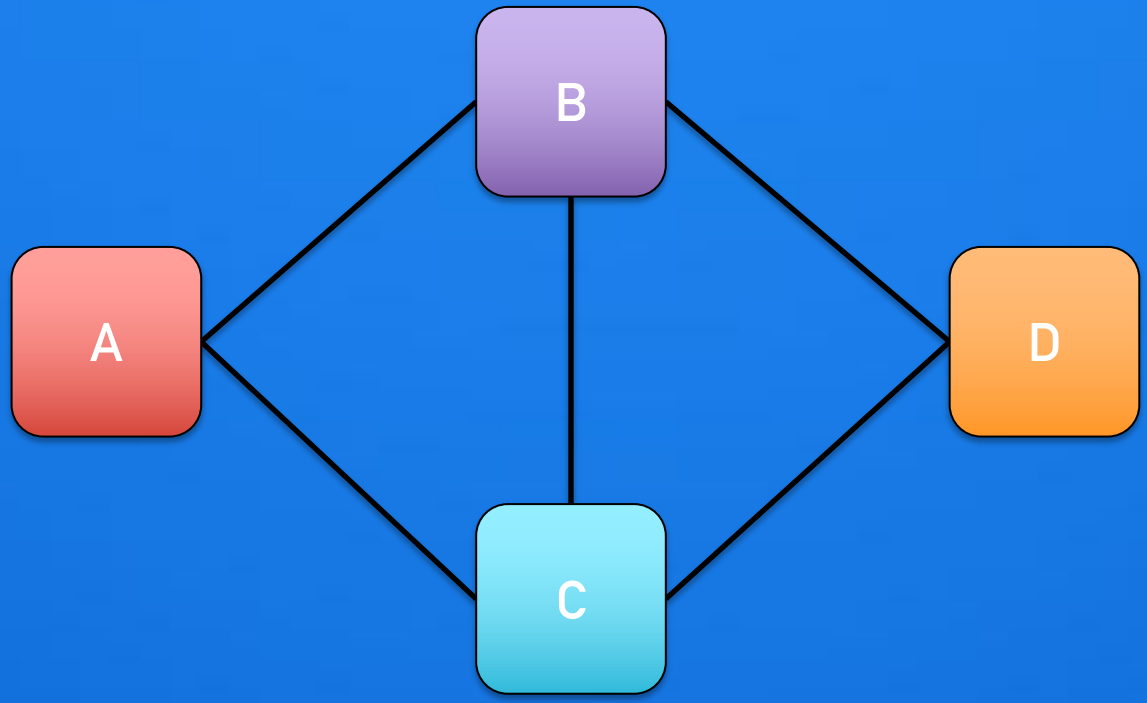
LABEL USED TO INDICATE PACKET'S CONTEXT – ALLOWS MULTIPLE SERVICES TO BE USED – BOTH L3 + L2.

BASIC MPLS FORWARDING.



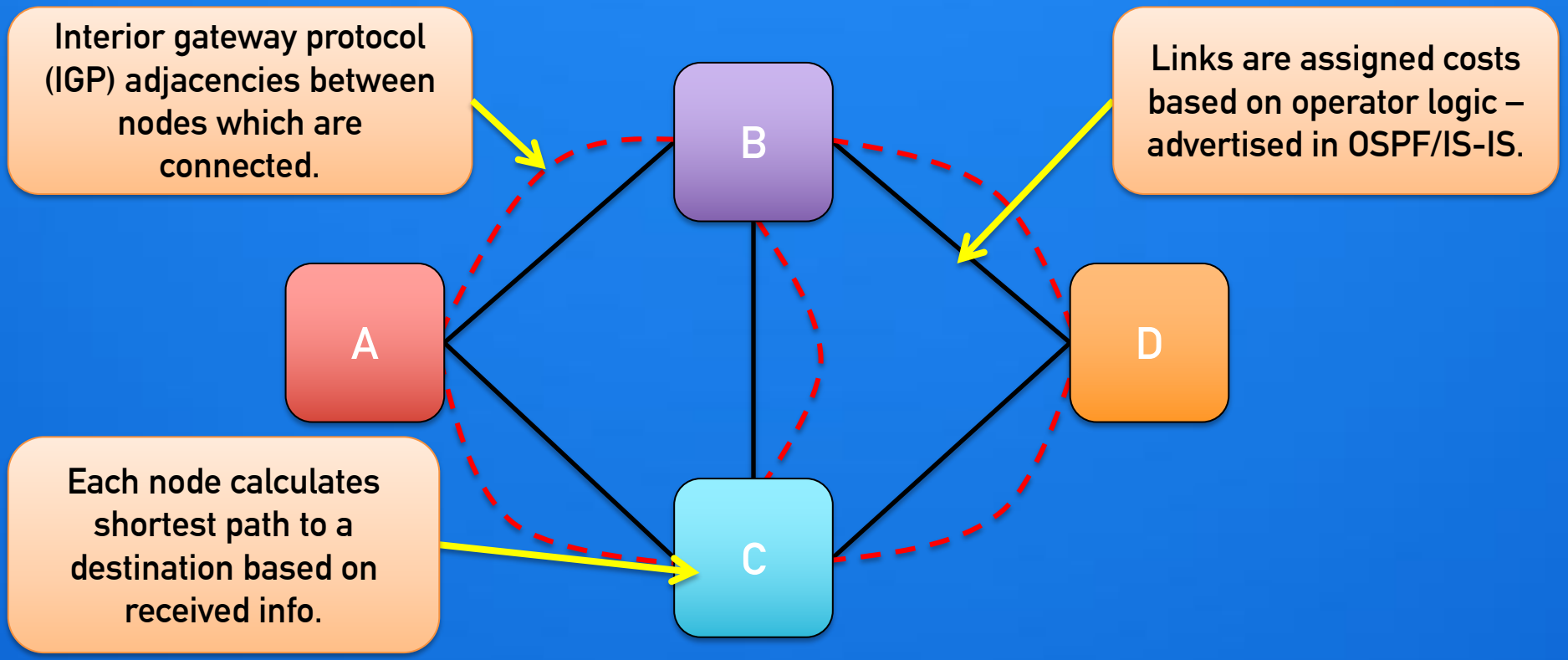


TODAY'S BASIC IP/MPLS NETWORK ANATOMY.



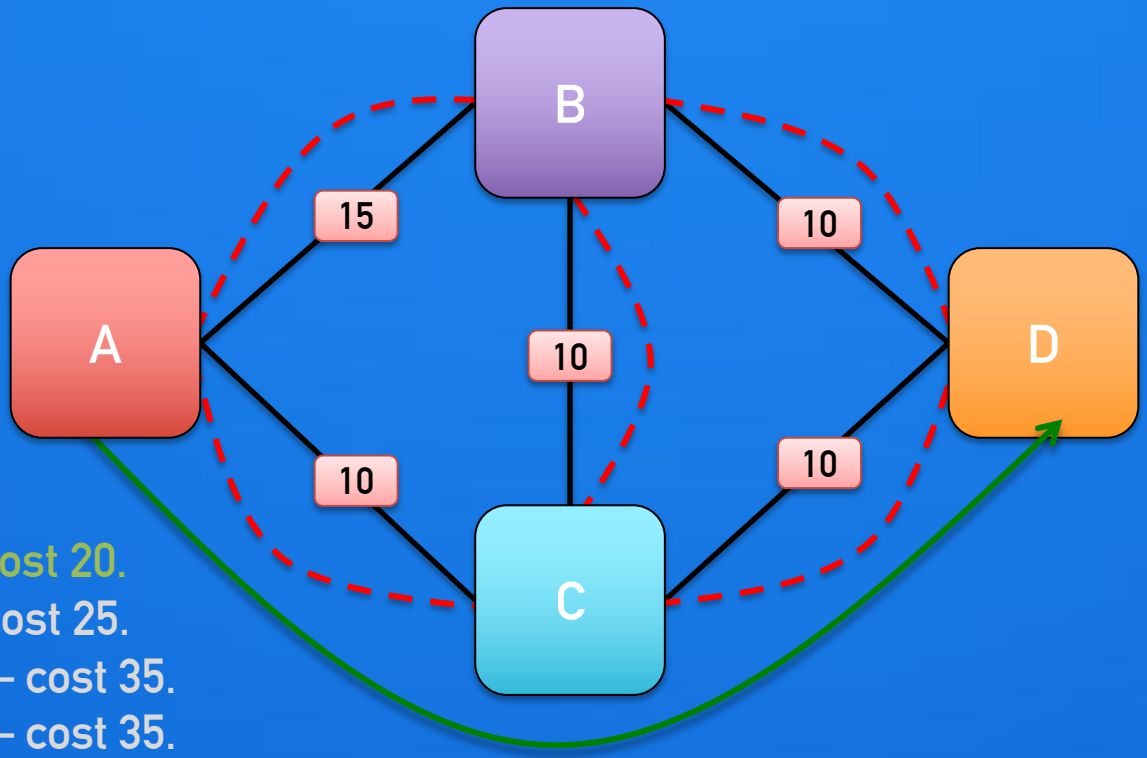
BASIC NETWORK – HOW DO WE DISCOVER WHICH LINKS ARE UP/DOWN AND WHICH DESTINATIONS ARE REACHABLE THROUGH ADJACENT NODES?

TODAY'S BASIC IP/MPLS NETWORK ANATOMY – IGP.



INTERIOR GATEWAY PROTOCOL (IGP) – TYPICALLY OSPF OR IS-IS – PROVIDES INFORMATION ABOUT REACHABILITY BETWEEN NODES, ALLOWING SHORTEST-PATH CALCULATIONS BASED ON ADVERTISED COST.

TODAY'S BASIC IP/MPLS NETWORK ANATOMY – IGP.

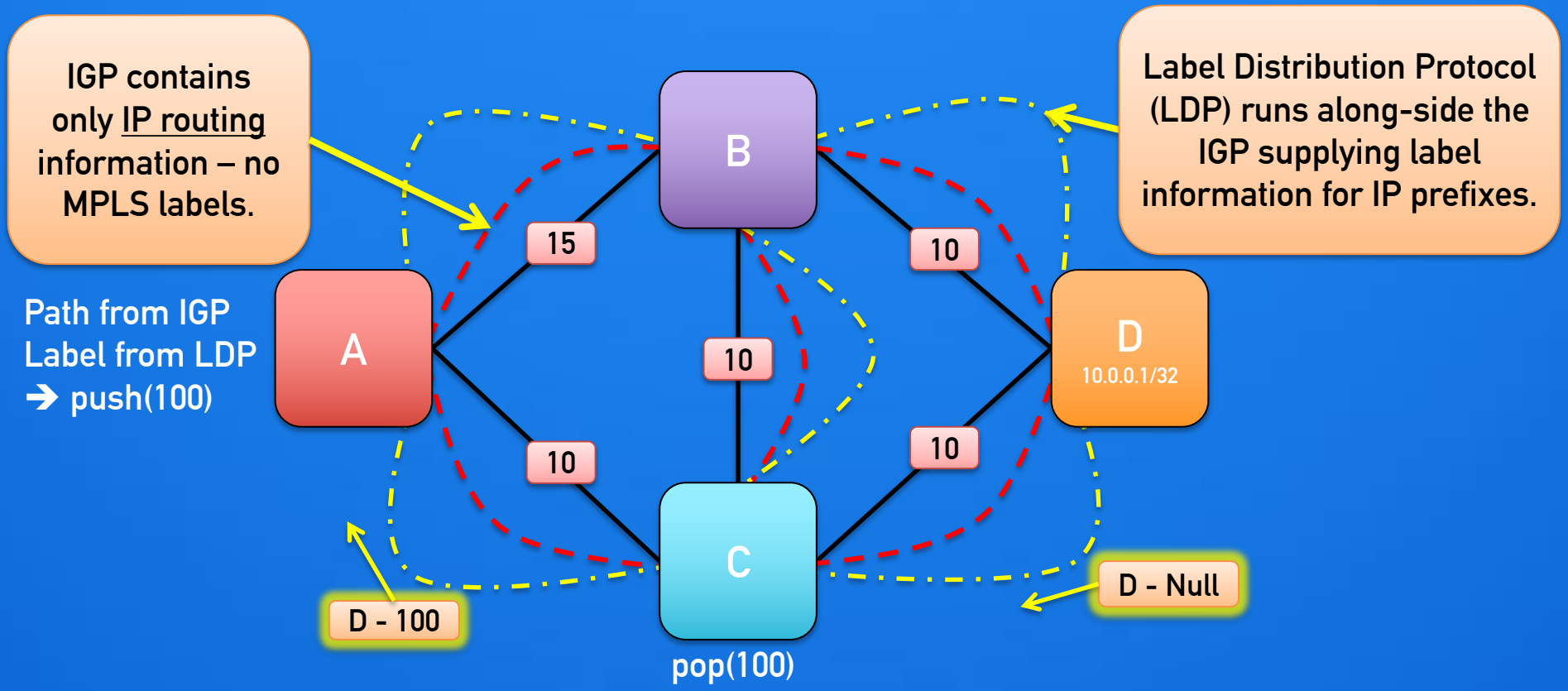


Dijkstra→D:

- A-C-D – cost 20.
- A-B-D – cost 25.
- A-B-C-D – cost 35.
- A-C-B-D – cost 35.

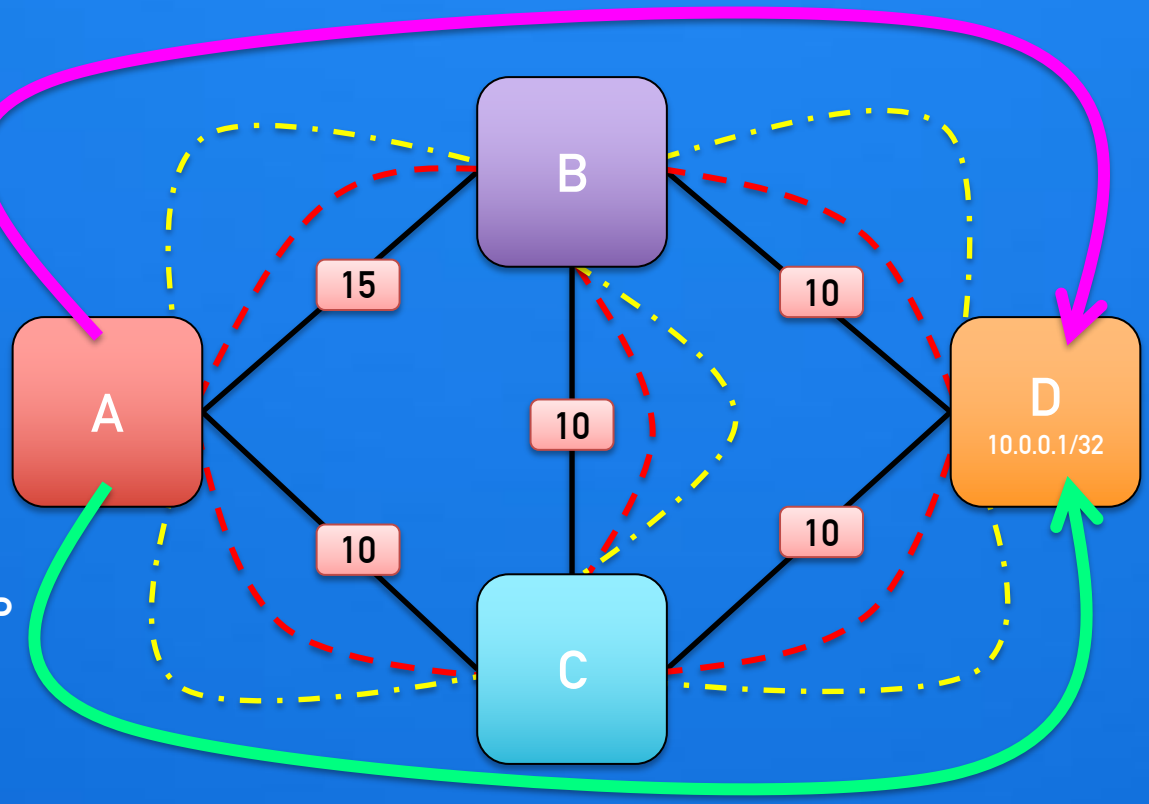
FORWARDING THROUGH THE NETWORK IS THEN BASED ON SHORTEST PATH INFORMATION ONLY – SINGLE SET OF METRICS FOR THE NETWORK.

TODAY'S BASIC IP/MPLS NETWORK ANATOMY – LDP.



ALLOWS MPLS FORWARDING ALONG THE IGP SHORTEST PATH – BUT INCREASES THE NUMBER OF PROTOCOLS DEPLOYED AND OPERATIONAL COMPLEXITY.

SELECTING A NON-SHORTEST PATH.

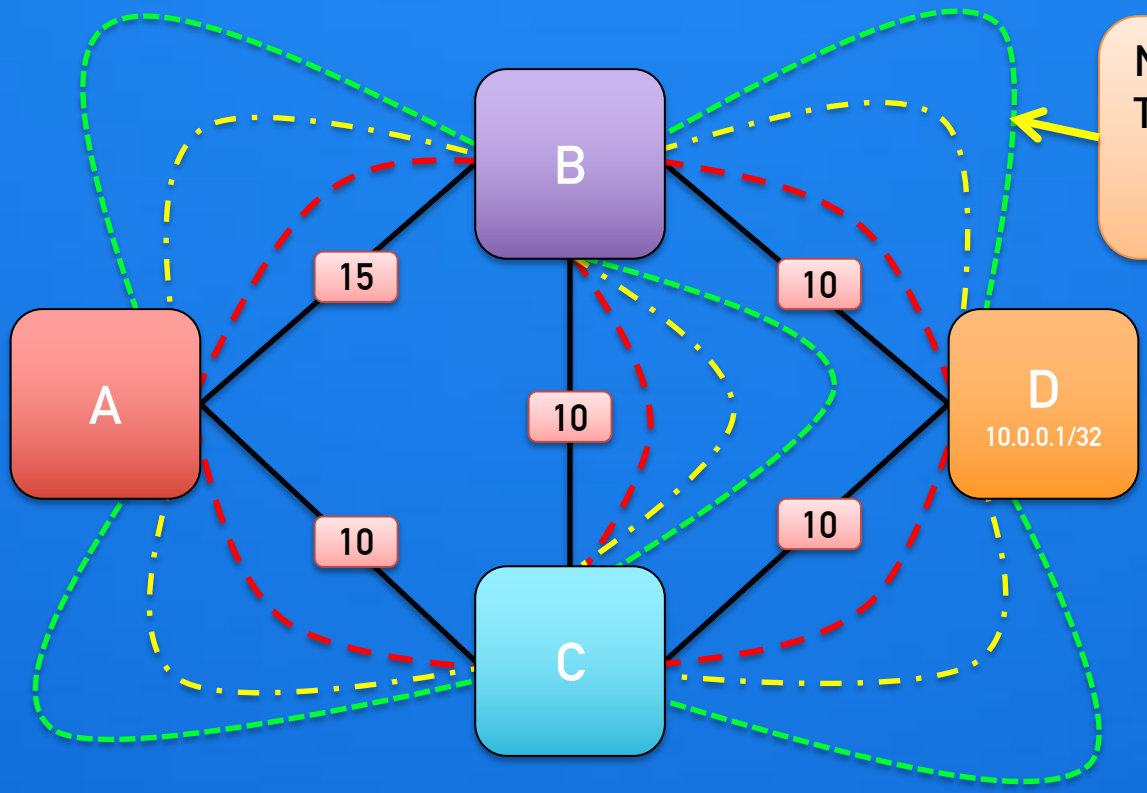


Pink path is divergent from lowest cost path (cost = 25, rather than 20)

Green path is IGP shortest path.

FOR SOME MULTI-SERVICE APPLICATIONS – E.G., PATH DISJOINTNESS – WE NEED TO SELECT A PATH WHICH IS NOT THE IGP SHORTEST PATH.

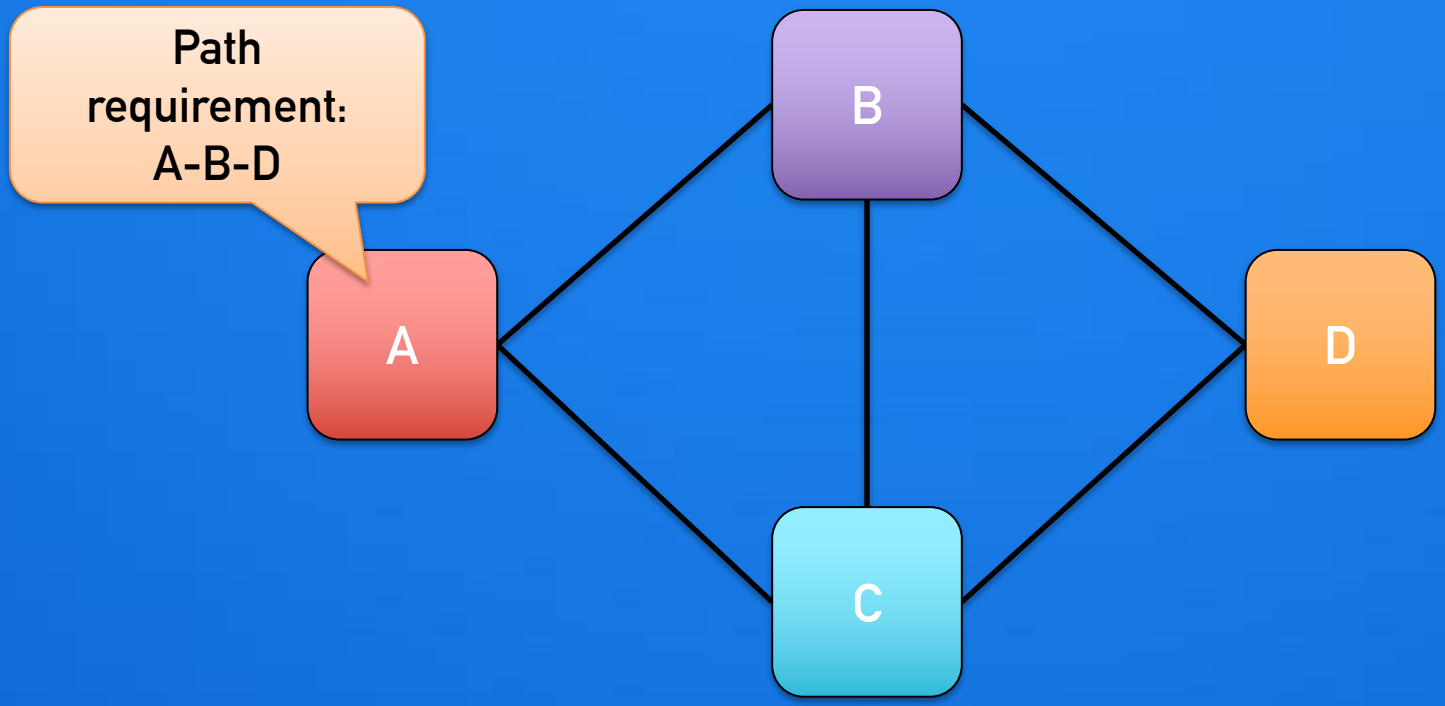
SELECTING A NON-SHORTEST PATH – RSVP-TE.



New protocol – RSVP-TE introduced to allow explicit paths to be signalled.

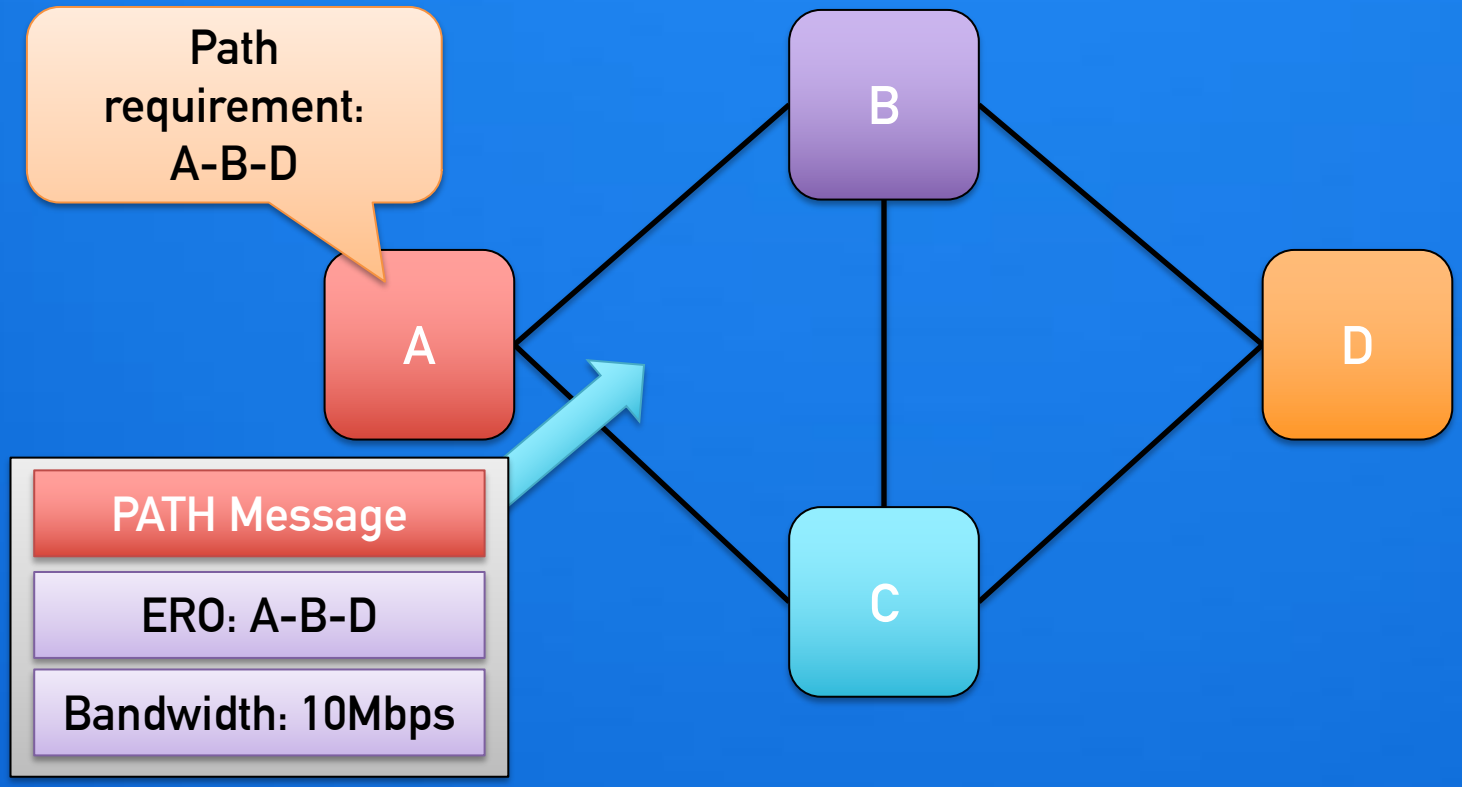
FURTHER PROTOCOL INTRODUCED TO ALLOW FOR EXPLICIT PATHS – WHICH MAY BE A SMALL SUBSET OF THE TRAFFIC CARRIED ON THE NETWORK.

OPERATION OF RSVP-TE WITHIN A NETWORK.



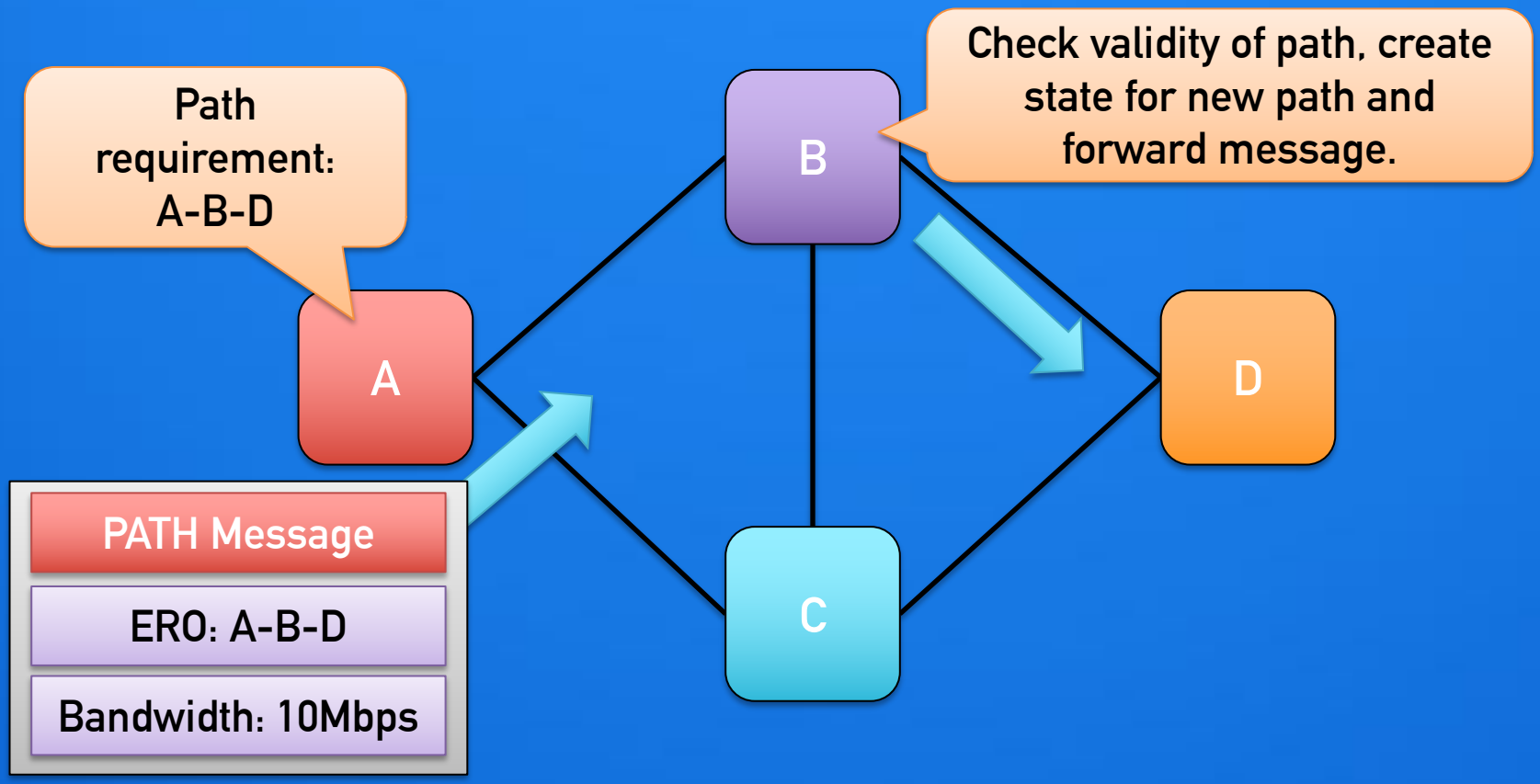
HEAD-END (INITIATOR) OF LABEL SWITCHED PATH (LSP) CALCULATES PATH THROUGH THE NETWORK THAT IS REQUIRED - LDP/IGP DOES NOT GUARANTEE THIS PATH.

OPERATION OF RSVP-TE WITHIN A NETWORK.



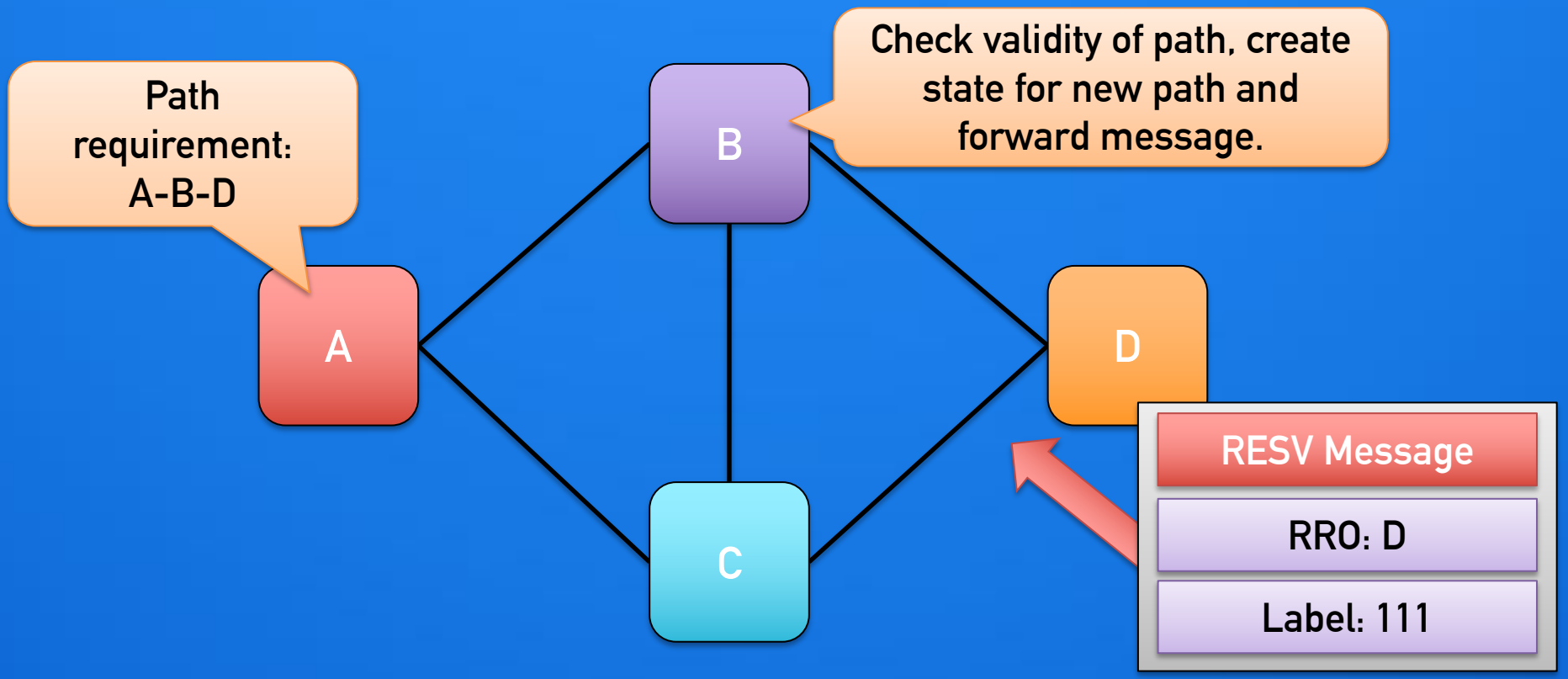
HEAD-END DEVICE BUILDS A MESSAGE TO SIGNAL THIS LSP – INCLUDING ANY CONSTRAINTS REQUIRED FOR THE NEW TUNNEL.

OPERATION OF RSVP-TE WITHIN A NETWORK.



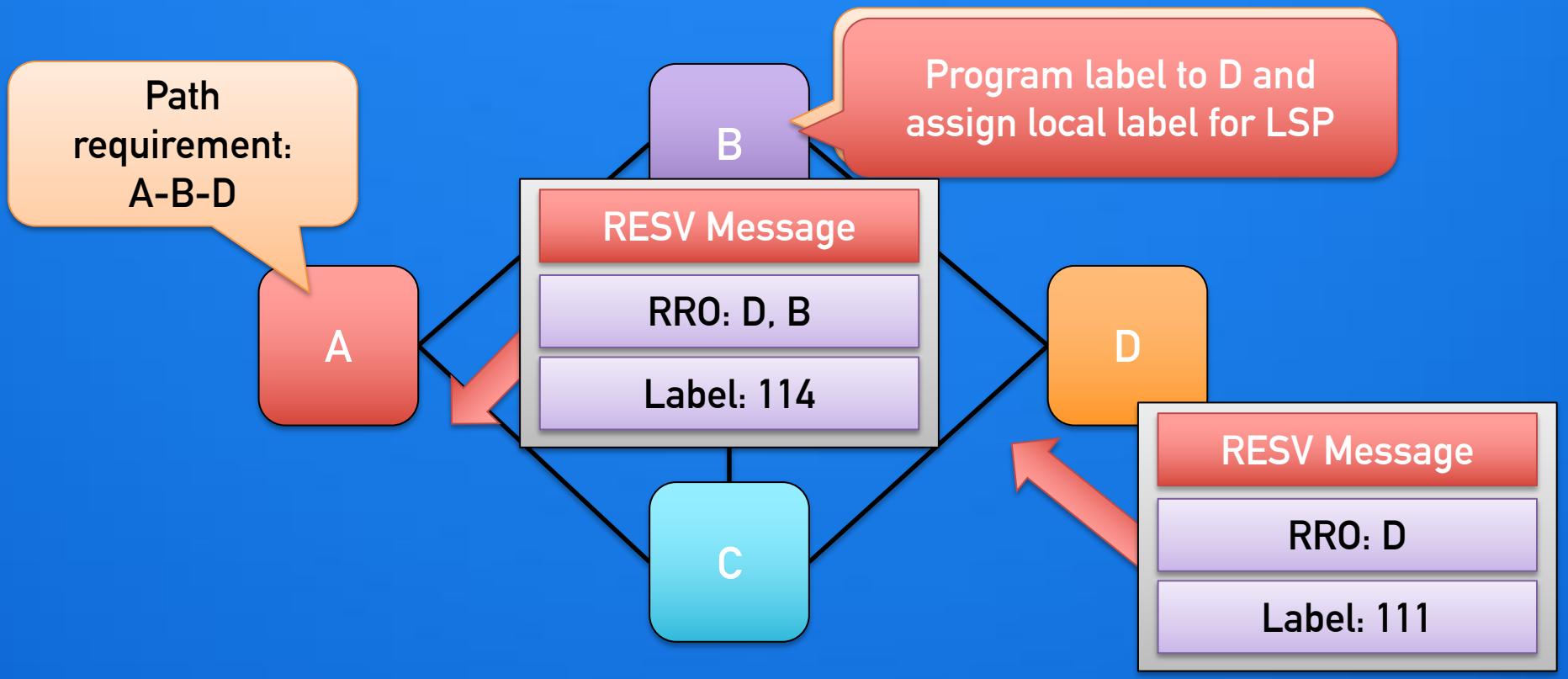
HEAD-END DEVICE BUILDS A MESSAGE TO SIGNAL THIS LSP – INCLUDING ANY CONSTRAINTS REQUIRED FOR THE NEW TUNNEL.

OPERATION OF RSVP-TE WITHIN A NETWORK.



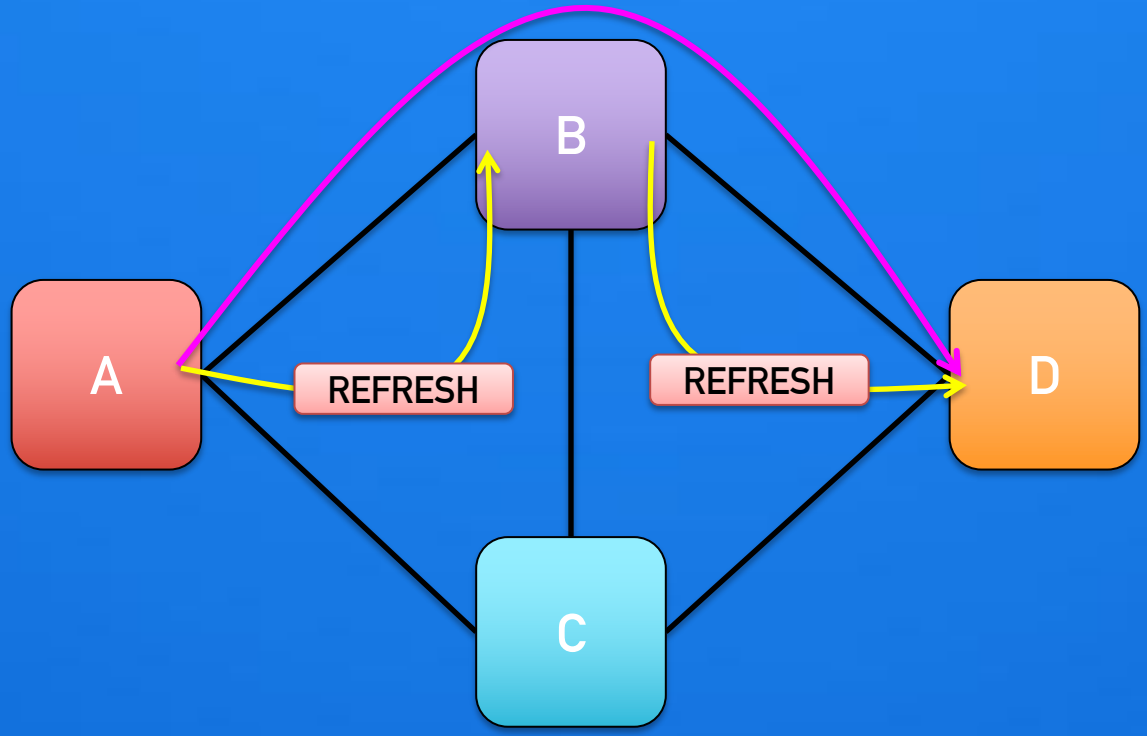
WHEN DESTINATION RECEIVES PATH MESSAGE, IT CREATES A RESERVATION MESSAGE PROVIDING LABEL INFORMATION FOR THE LSP.

OPERATION OF RSVP-TE WITHIN A NETWORK.



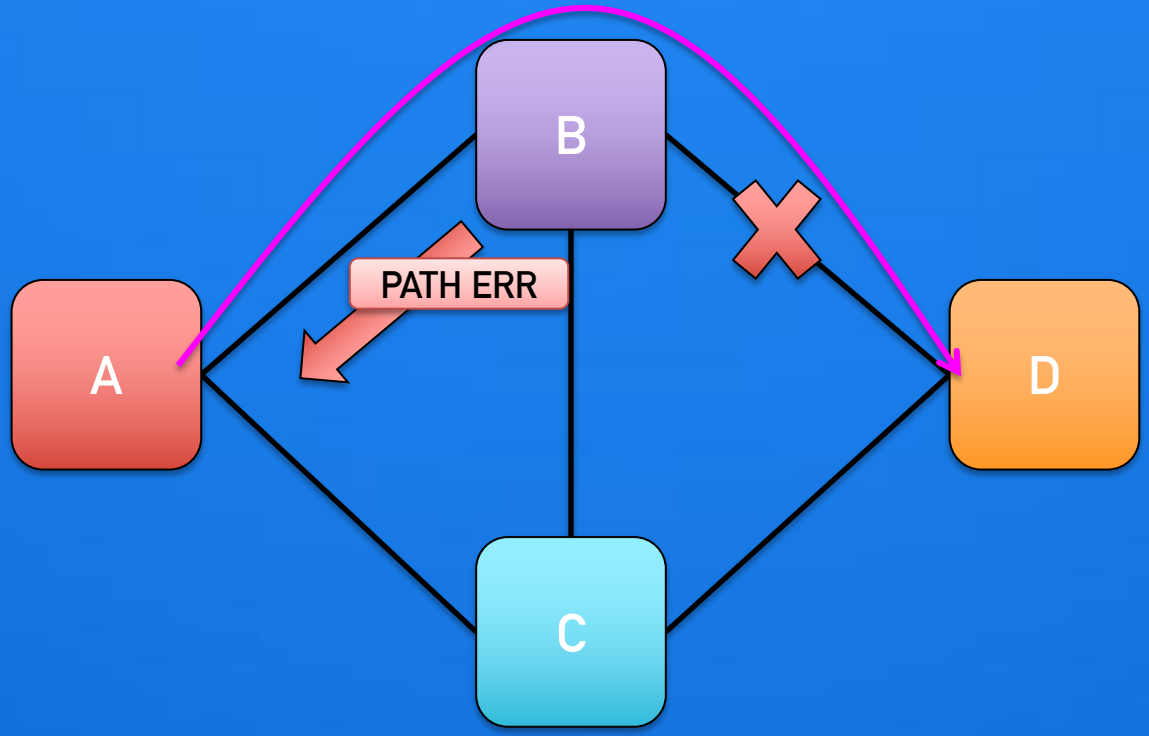
RESERVATION MESSAGE CREATES STATE ON EACH NODE ALONG THE PATH – AND REPORTS LABEL INFORMATION BACK TO THE HEAD-END FOR FORWARDING.

OPERATION OF RSVP-TE WITHIN A NETWORK.



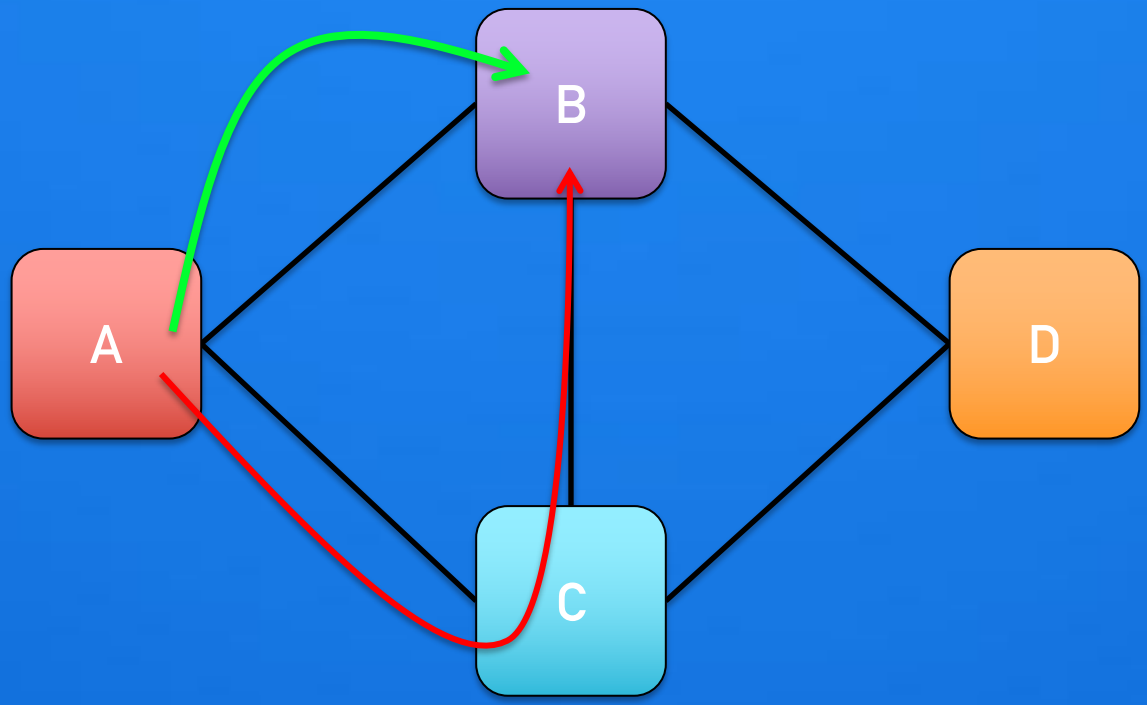
AS WELL AS DIFFERING SET UP PROCEDURES, THE RESERVATIONS MUST BE REFRESHED PERIODICALLY (SOFT-STATE PROTOCOL REQUIREMENT).

OPERATION OF RSVP-TE WITHIN A NETWORK.



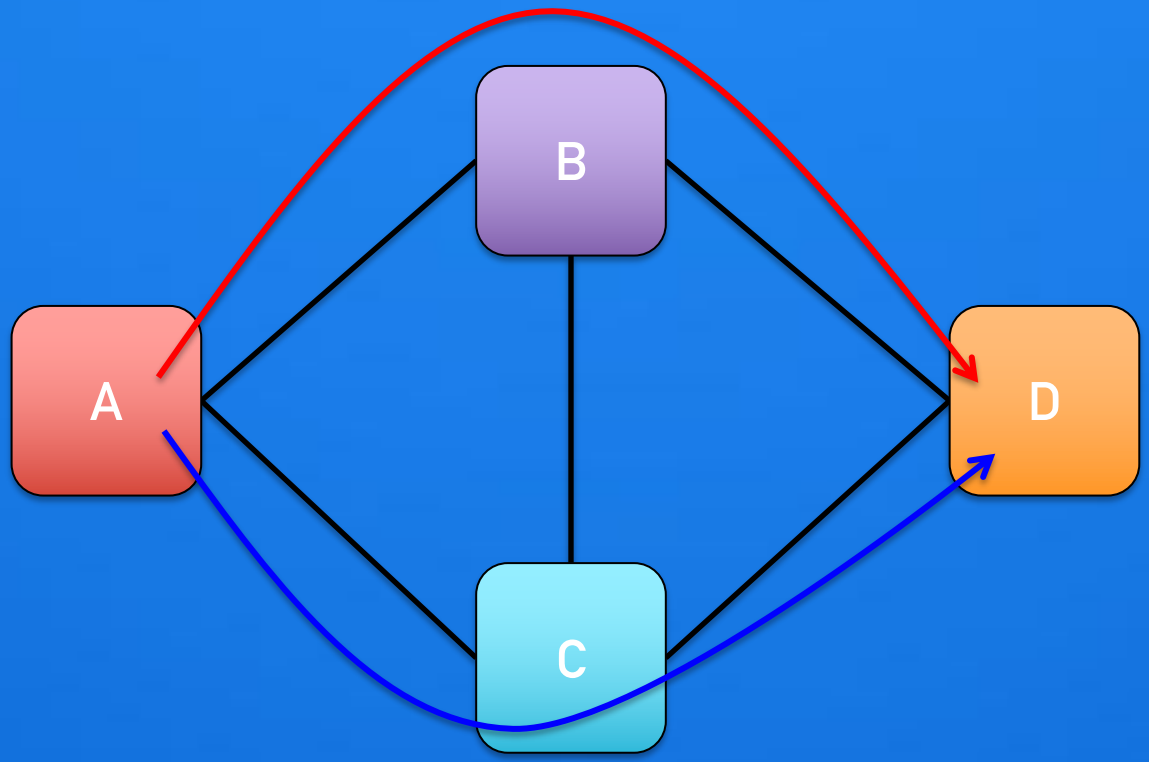
DURING LINK FAILURES – HEAD-END IS NOT REQUIRED TO RE-CONVERGE PATHS – BUT RATHER NODES ADJACENT TO THE FAILURE MUST REPORT PATH FAILURE – RESULTING IN RE-SIGNALLING PROCESS RE-STARTING.

APPLICATIONS OF RSVP-TE LSPS.



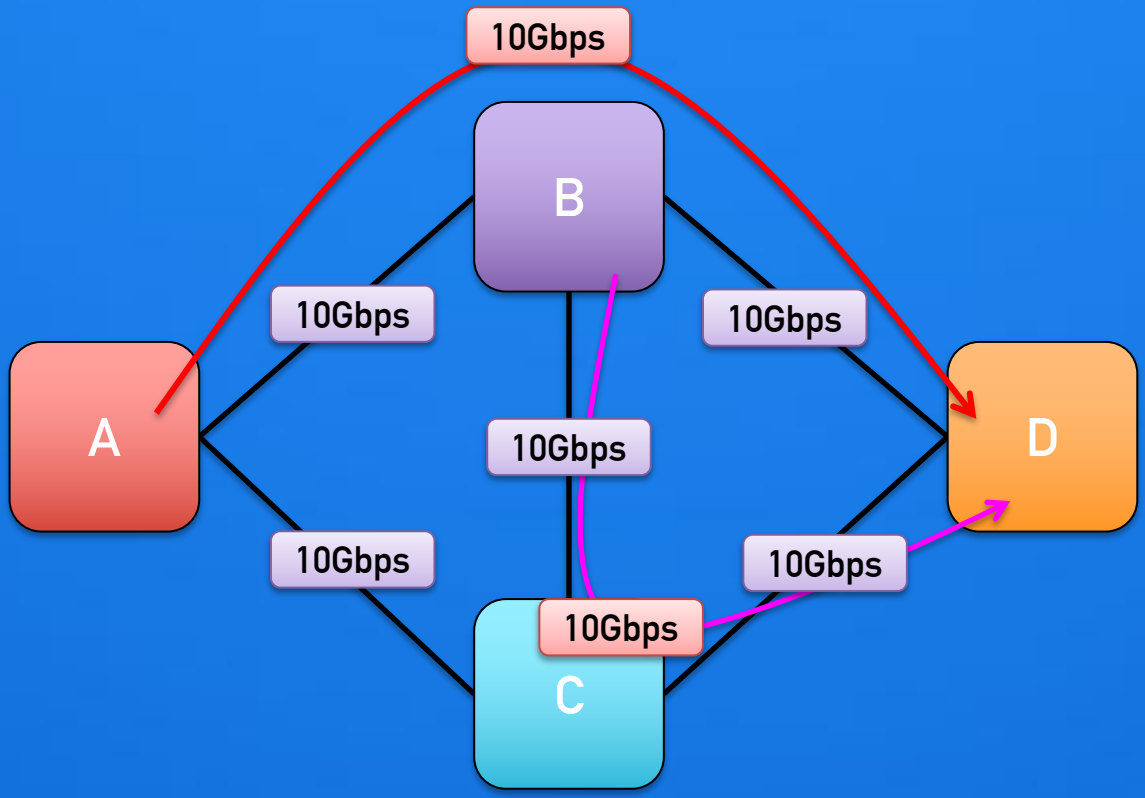
FAST RE-ROUTE:
EXPLICIT PATH AVOIDING A LINK OR NODE TO BE PROTECTED WHICH CAN BE USED TO PROVIDE FAST RESTORATION OF A PATH.

APPLICATIONS OF RSVP-TE LSPS.



DISJOINT PATHS:
RED/BLUE TOPOLOGIES CAN BE SIGNALLED BASED ON LSPS WHICH TRAVERSE PARTICULAR SETS OF LINKS.

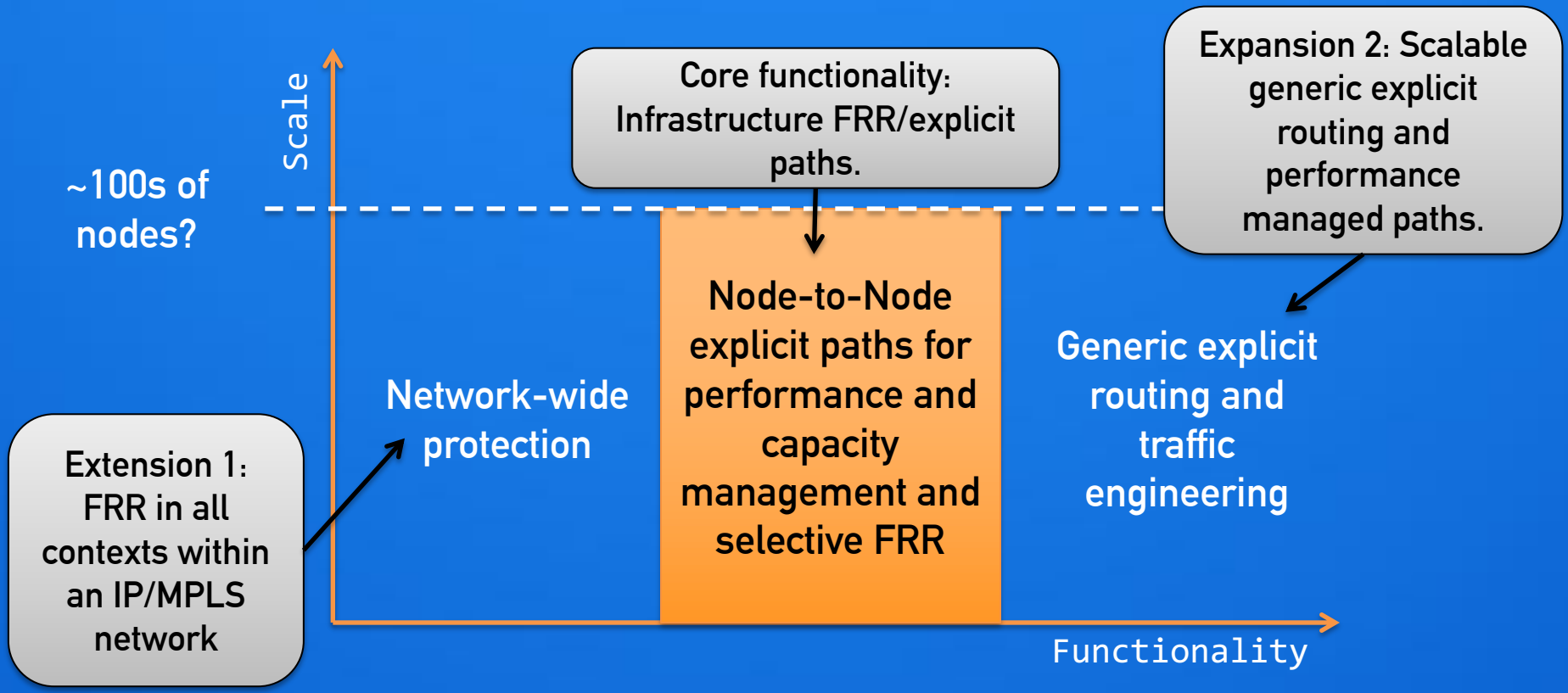
APPLICATIONS OF RSVP-TE LSPS.



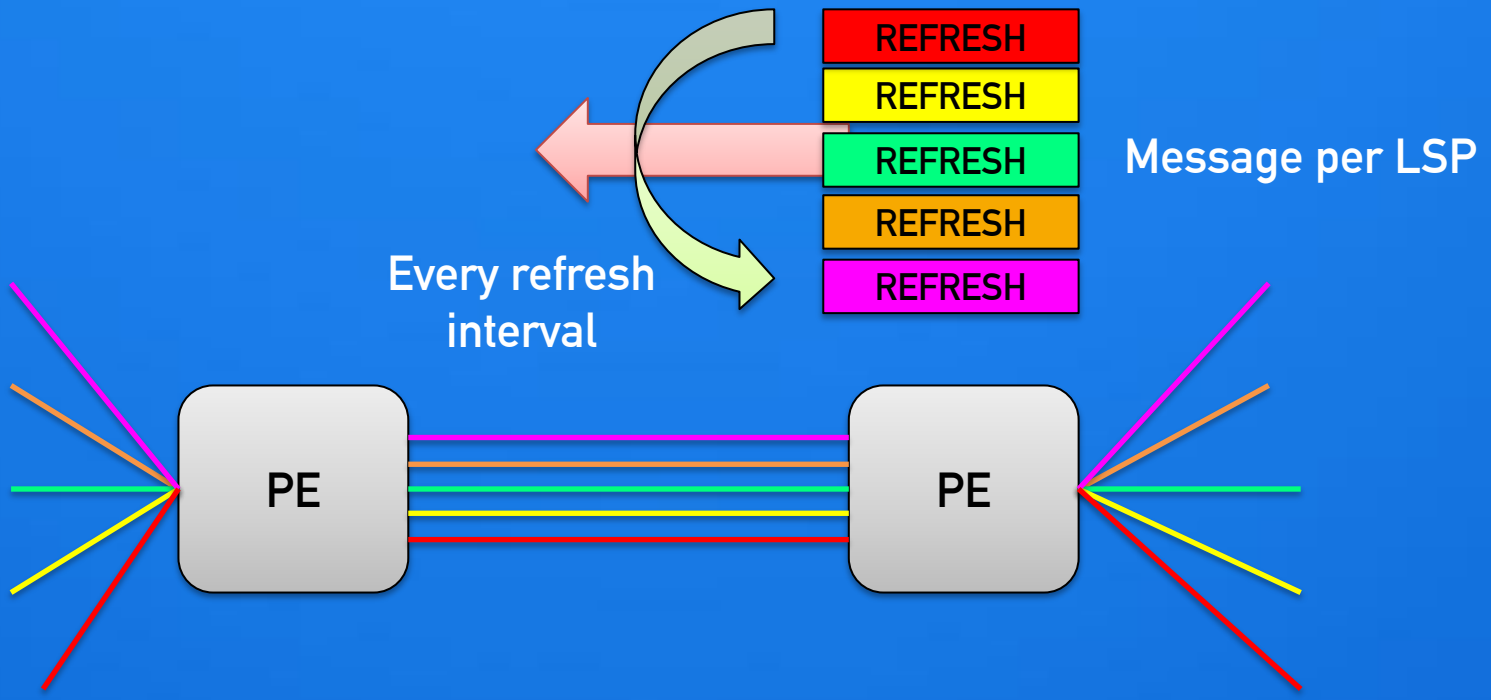
TRAFFIC ENGINEERING:
PLACEMENT OF PATHS ACCORDING TO AVAILABLE RESOURCES SUCH AS BANDWIDTH OR LATENCY.

RSVP-TE IN THE CONTEXT OF RFC5218.

Looking at RSVP-TE in the context of 5218 – we can see where these solutions were intended to fit – and the possible expansions out to the generic multi-service IP/MPLS network use-cases discussed originally.

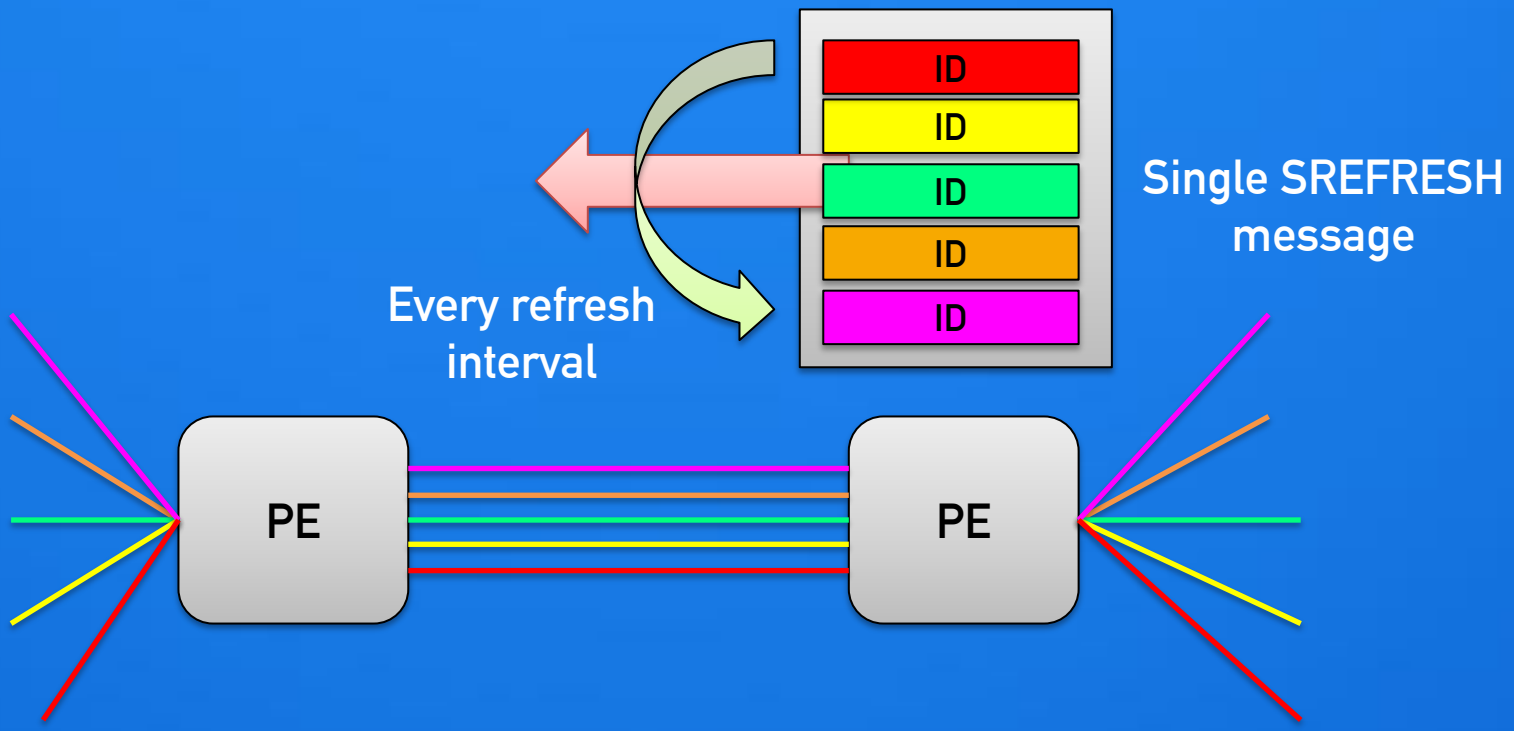


CHALLENGE 1: EXPLICIT PATH ROUTING.



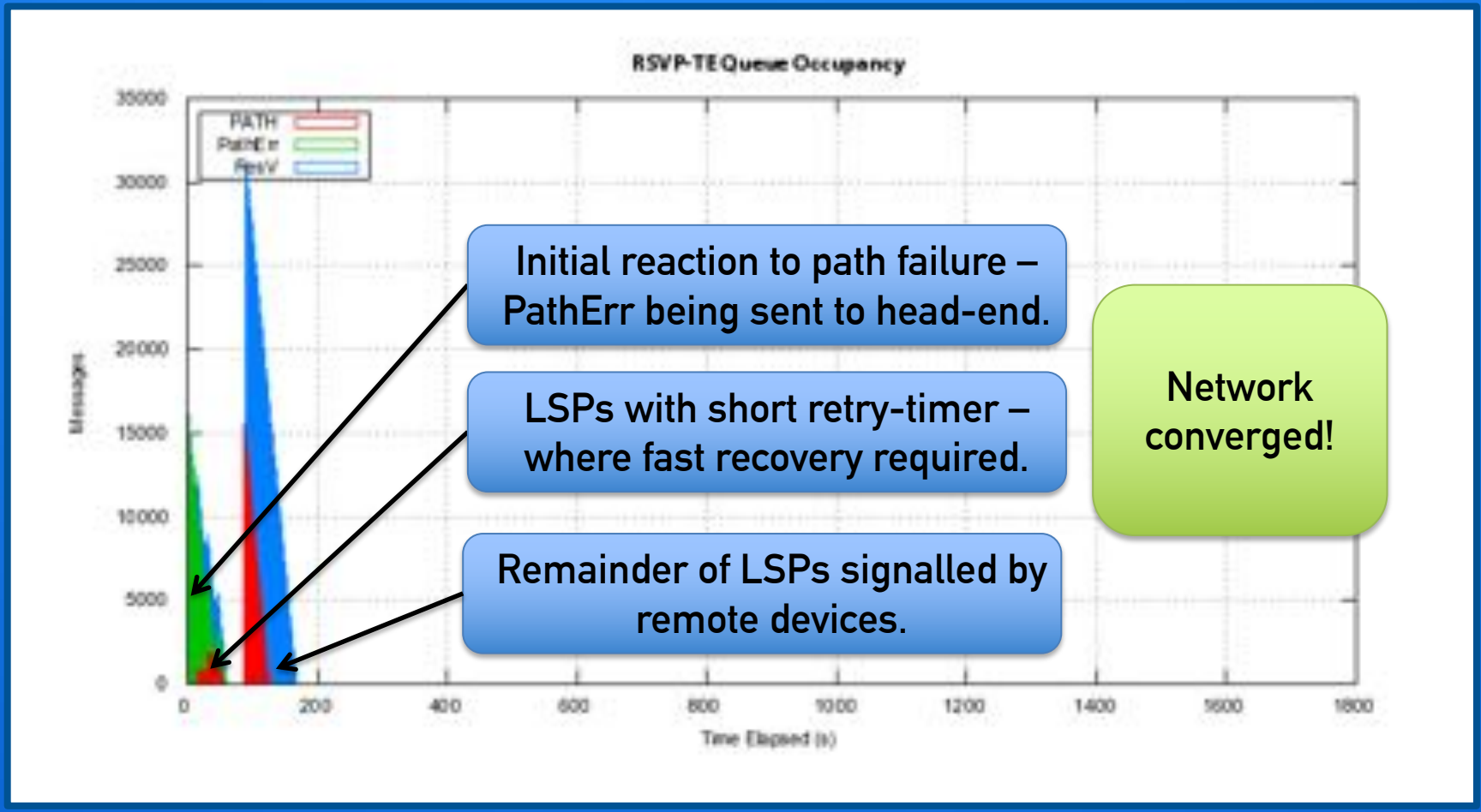
SCALING SOFT-STATE PROTOCOLS:
ALTHOUGH RSVP-TE CAN SIGNAL THESE PATHS – SCALE IS LIMITED BY THE NUMBER OF PATHS THAT CAN BE REFRESHED WITHIN THE INTERVAL (OR RATHER, IN THE SCHEDULED CPU CYCLE TIME).

CHALLENGE 1: EXPLICIT PATH ROUTING.



RELATIVELY EASY TO SOLVE – BUT REQUIRED ADDITIONS TO PROTOCOL. REFRESH ALL LSPS WITHIN A SINGLE MESSAGE – REDUCES NUMBER OF MESSAGES THAT MUST BE GENERATED.

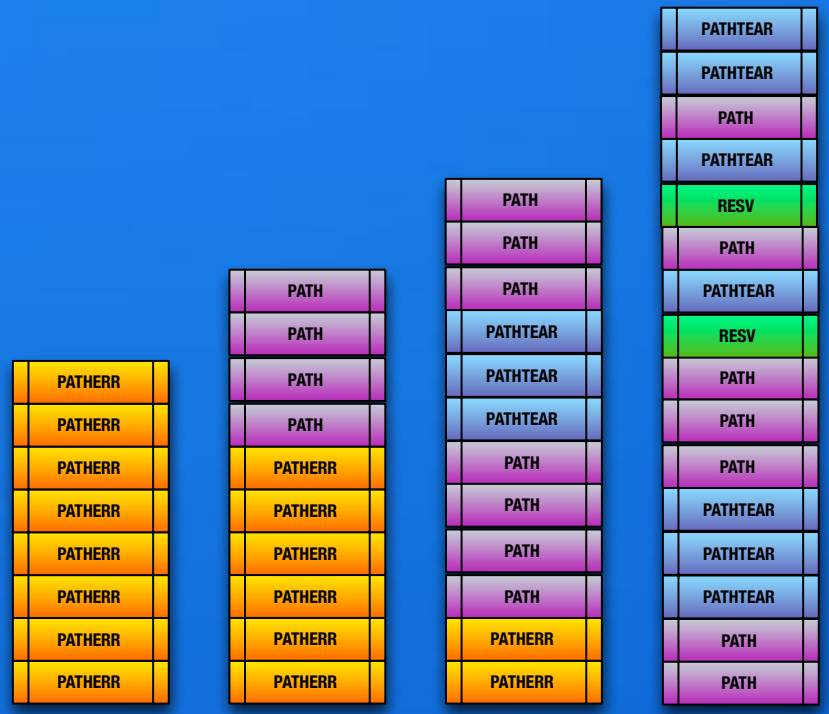
SIGNALLING FOLLOWING MID-POINT FAILURE:



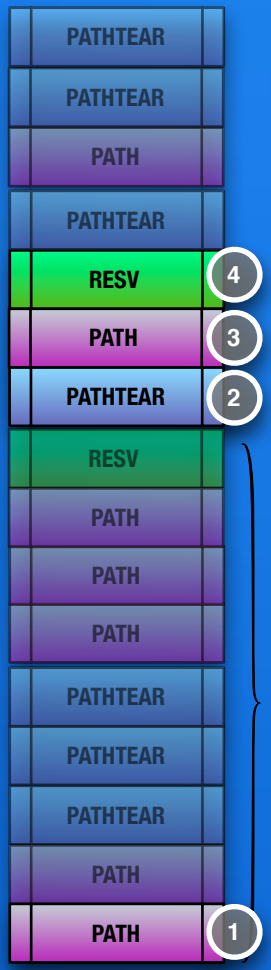
CHALLENGE 1: EXPLICIT PATH ROUTING.

But the soft state problem is not wholly solved outside of steady-state operation – for example, look at a large failure on a mid-point carrying many LSPs (quite common due to sub-sea cable connectivity)!

1. **INITIAL PATHERRS.**
   Indicating FRR to head-end LSRs.
2. **GLOBAL REVERT PATH.**
   HE LSRs begin to signal new LSPs.
3. **NEW LSP PATHTEAR.**
   Slow response results in new LSP teardown.
4. **RESV FROM UPSTREAM.**
   Torn down LSP is setup.
5. **PATH FROM HE RETRY.**
   Subsequent attempts to setup LSP.



CHALLENGE 1: EXPLICIT PATH ROUTING.



1. GLOBAL REVERT PATH.

Head-end transmits a Path message for LSP Tunnel ID = M, LSP ID = N.

2. PATHTEAR SENT AFTER RETRY INTERVAL.

Head-end tears down Tunnel ID = M, LSP ID = N– retry period expired.

3. HEAD-END RE-SENDS PATH.

HE tries to signal new LSP – Tunnel ID = M, LSP ID = N+1.

4. MID-POINT RECEIVES RESV BACK.

Resv received for Tunnel ID = M, LSP ID = N – out of date!

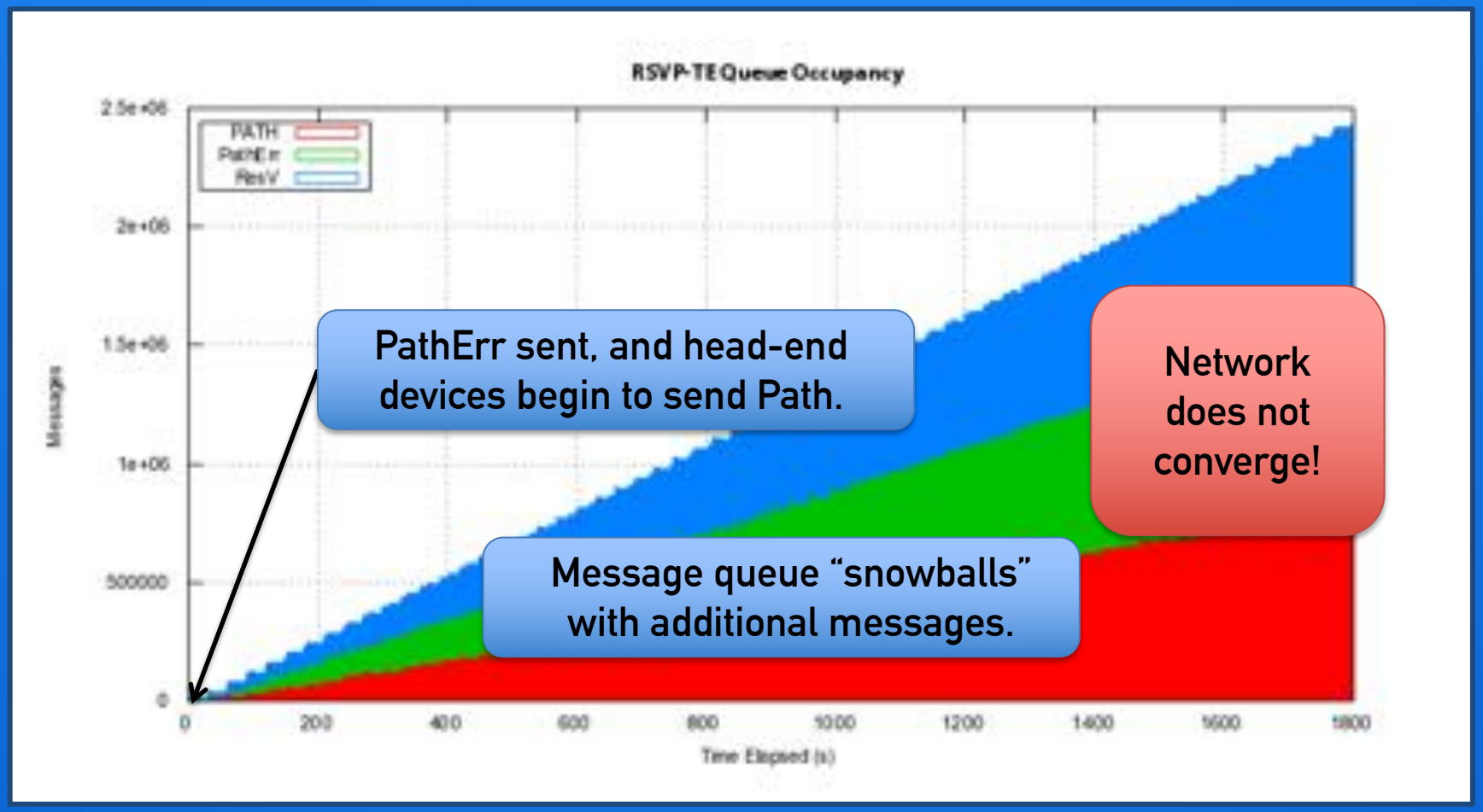
CONTINUAL LOOP!

Midpoint never processes the 'current' LSP ID.

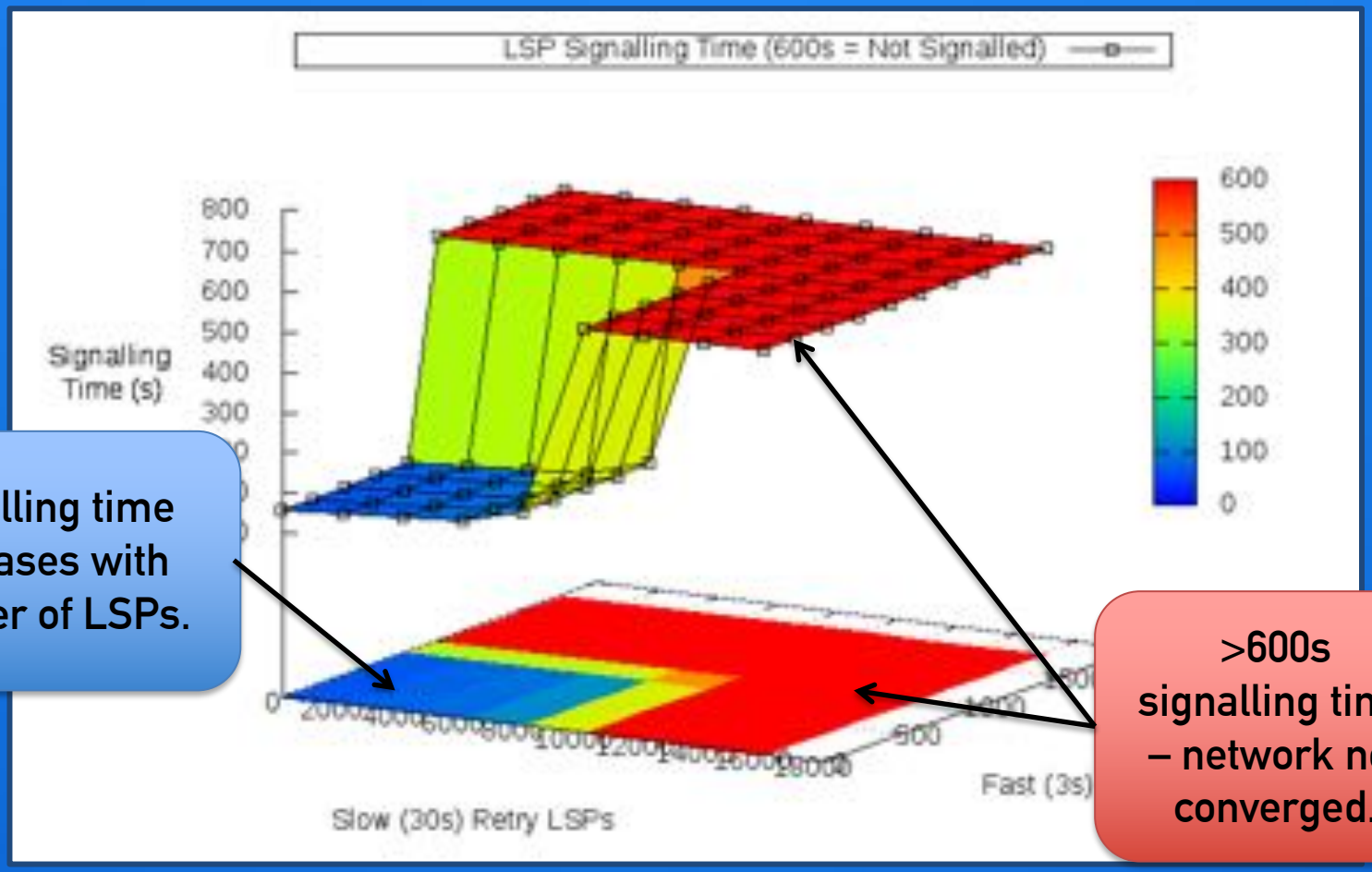
Results in at least two further messages in the queue.

Results in a "Snowball" effect.

RSVP-TE BEYOND SCALE LIMIT.



RSVP-TE BEYOND SCALE LIMIT – LSP RECOVERY TIME.



Signalling time increases with number of LSPs.

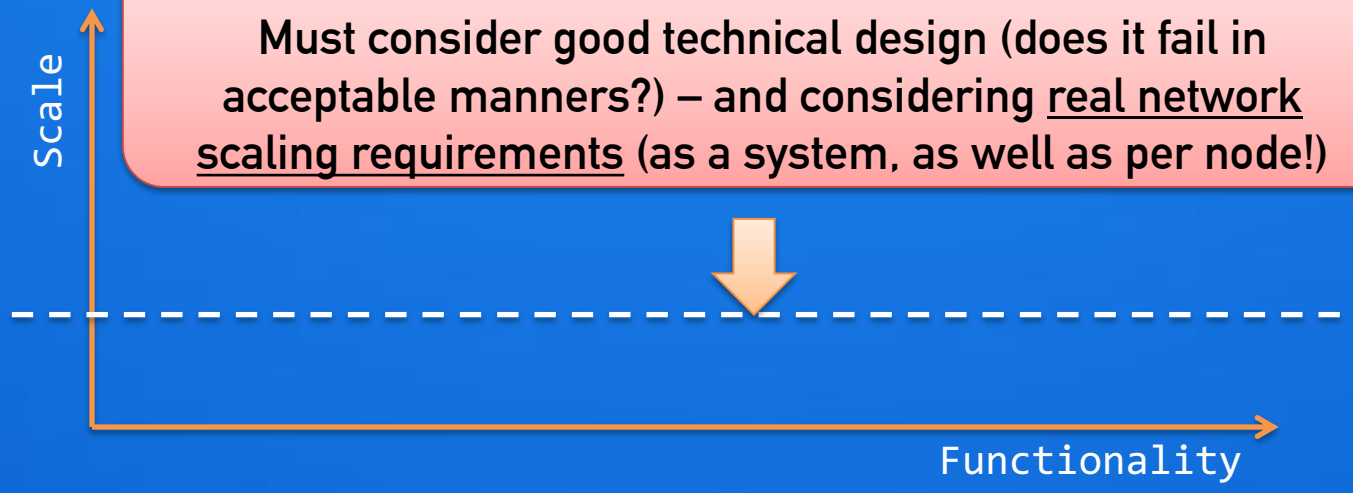
>600s signalling time – network not converged.

CHALLENGE 1: EXPLICIT PATH ROUTING.

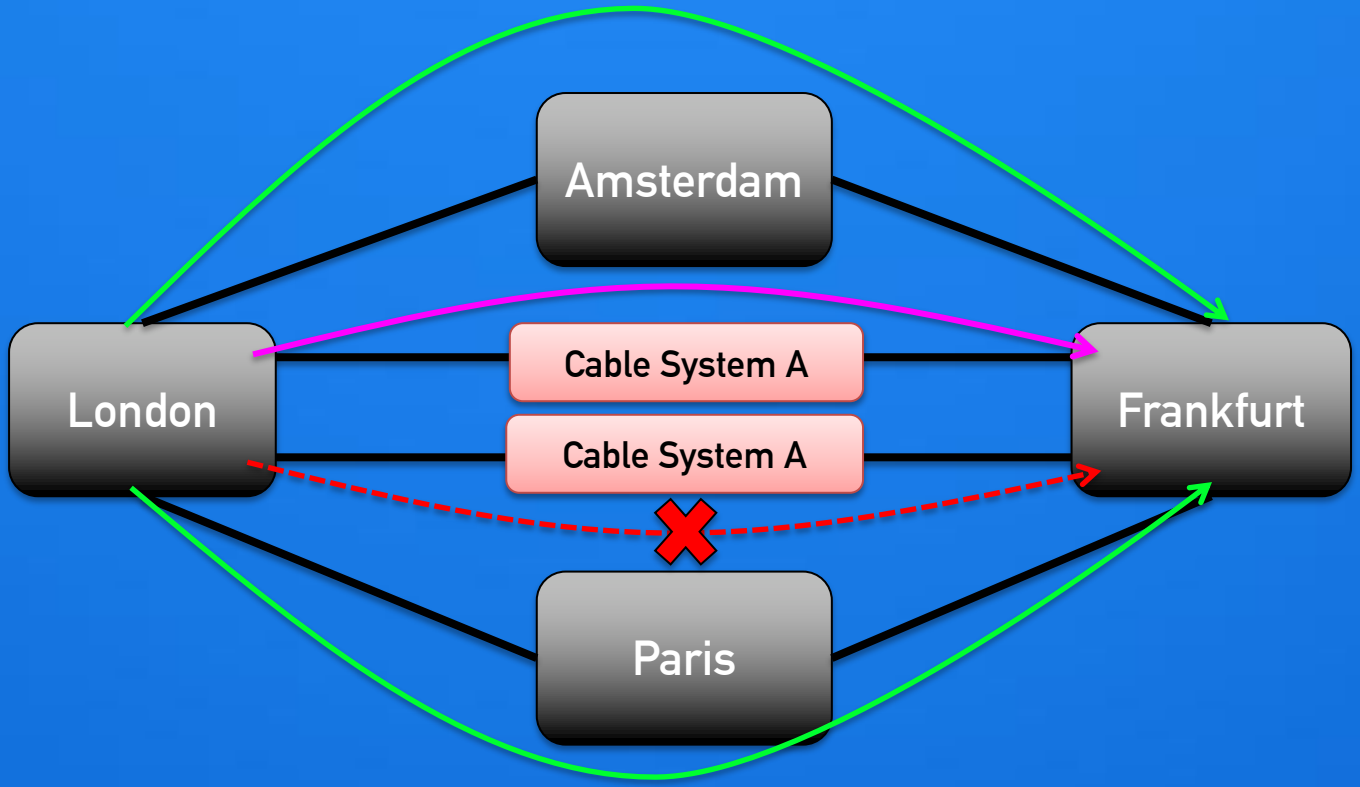
RSVP-TE had an incremental deployment advantage and solves real operator challenges…

But…comes with scalability limits which can result in significant fragility being introduced into real networks – this is definitely negative net value!

FOR PROTOCOL DESIGNERS:
Must consider good technical design (does it fail in acceptable manners?) – and considering real network scaling requirements (as a system, as well as per node!)

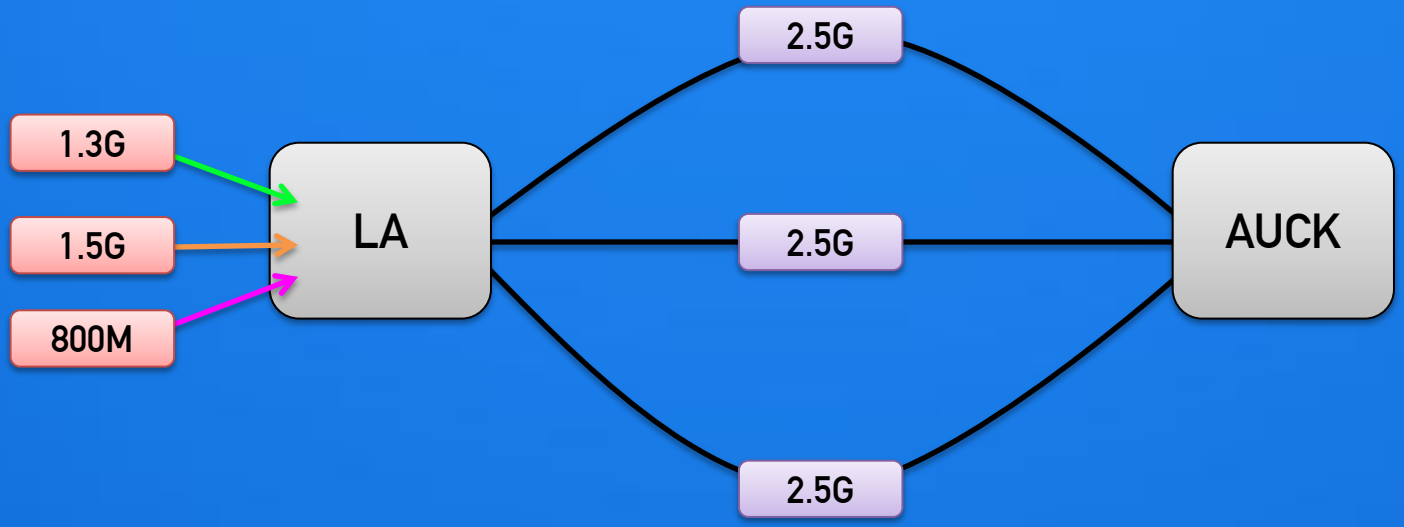


CHALLENGE 2: OPERATIONAL COMPLEXITY.



PATH PLACEMENT:
WHERE SHOULD A PARTICULAR LSP BE PLACED WITHIN THE NETWORK TO ENSURE NO SHARED RISKS, AND SUCH THAT WE MAXIMISE ROBUSTNESS?

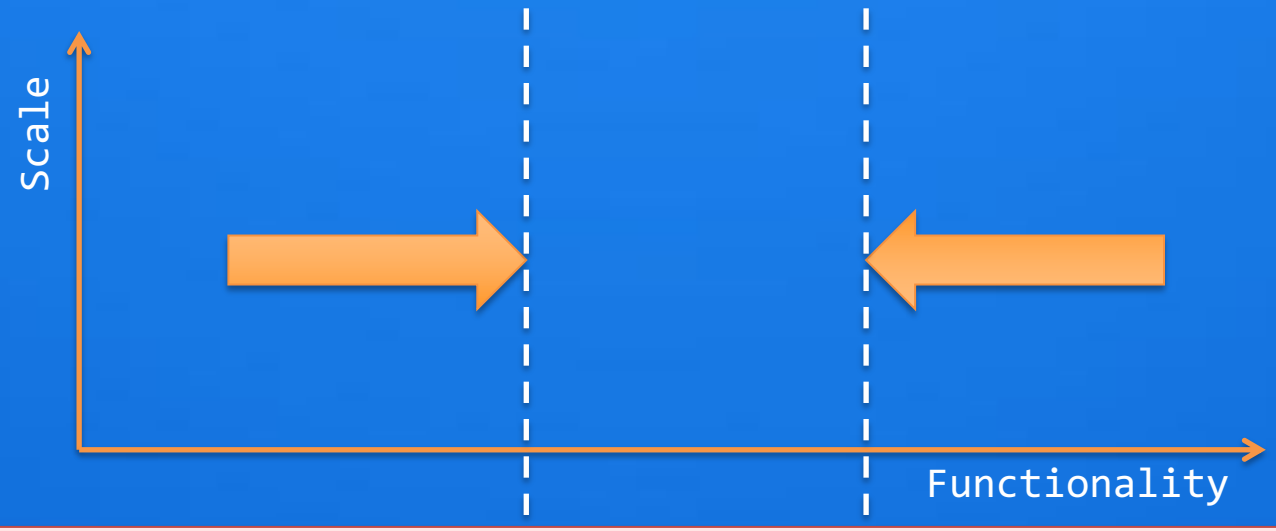
CHALLENGE 2: OPERATIONAL COMPLEXITY.



PATH PLACEMENT FOR TRAFFIC ENGINEERING:
WHERE WAS A PARTICULAR PATH PLACED (NOT JUST RELATED TO THE AVAILABLE LINKS) AT A PARTICULAR TIME?
WHICH SERVICES WERE IMPACTED BY A CERTAIN FAILURE?

CHALLENGE 2: OPERATIONAL COMPLEXITY.

The overall cost of deployment must be considered – RSVP-TE introduces a requirement for additional systems surrounding the network for both path placement and path monitoring. Additionally, costs are incurred based on operational training of staff where characteristics (e.g., reversion) differ to other protocol operation.



FOR PROTOCOL DESIGNERS:

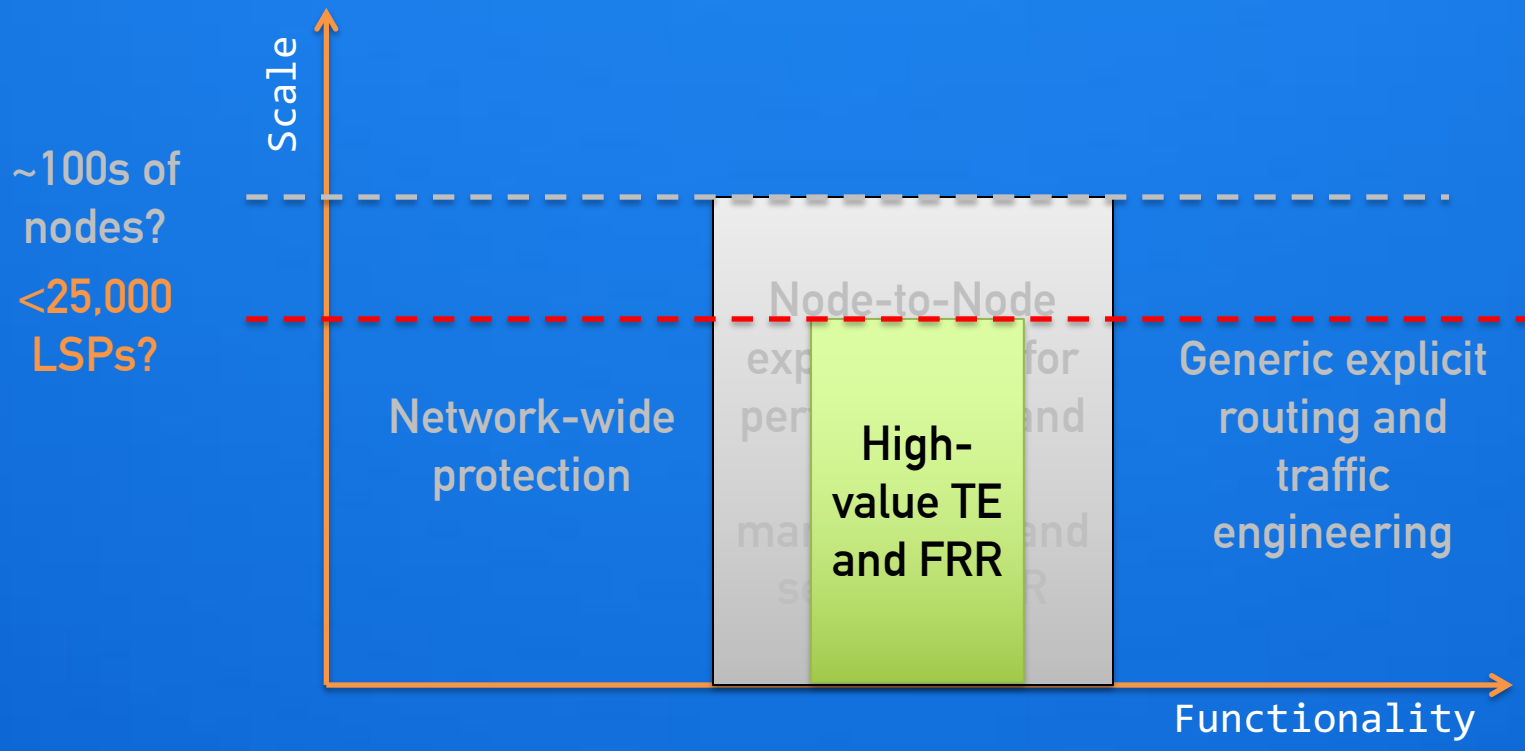
Minimising the number of additional systems required only for the protocol minimises cost and deployment barriers – and makes for simplified roll-out.

LOOKING AT 5218: WAS RSVP-TE A SUCCESS?

Real problem?	AS DEMONSTRATED – EXPLICIT PATHS SEEM TO BE A REAL REQUIREMENT IN MULTI-SERVICE IP/MPLS NETWORKS.
No new hardware?	ONLY NEW CODE REQUIRED (IN GENERAL) – NEW HARDWARE IS A SIGNIFICANT CHALLENGE (LEGACY REMAINS!)
Existing ops/processes?	NO – SIGNIFICANT DIFFERENCE FOR REVERSION AND MONITORING – REQUIRES SPECIFIC TRAINING OF OPS STAFF.
Relieving operational pain?	NOT PARTICULARLY – BUT WE BALANCE OPERATIONAL COMPLEXITY AGAINST REDUCED PARALLEL DEPLOYMENTS.
Incrementally deployable?	CAN BE DONE FOR SUBSETS OF TRAFFIC – GOOD IN THIS RESPECT.
Good technical design?	INHERENT RELIANCE ON SOFT-STATE HAS SIGNIFICANT CHALLENGES – WAS THIS THE RIGHT CHOICE?

PARTIAL SUCCESS: WHEN THE BARRIER'S TOO HIGH.

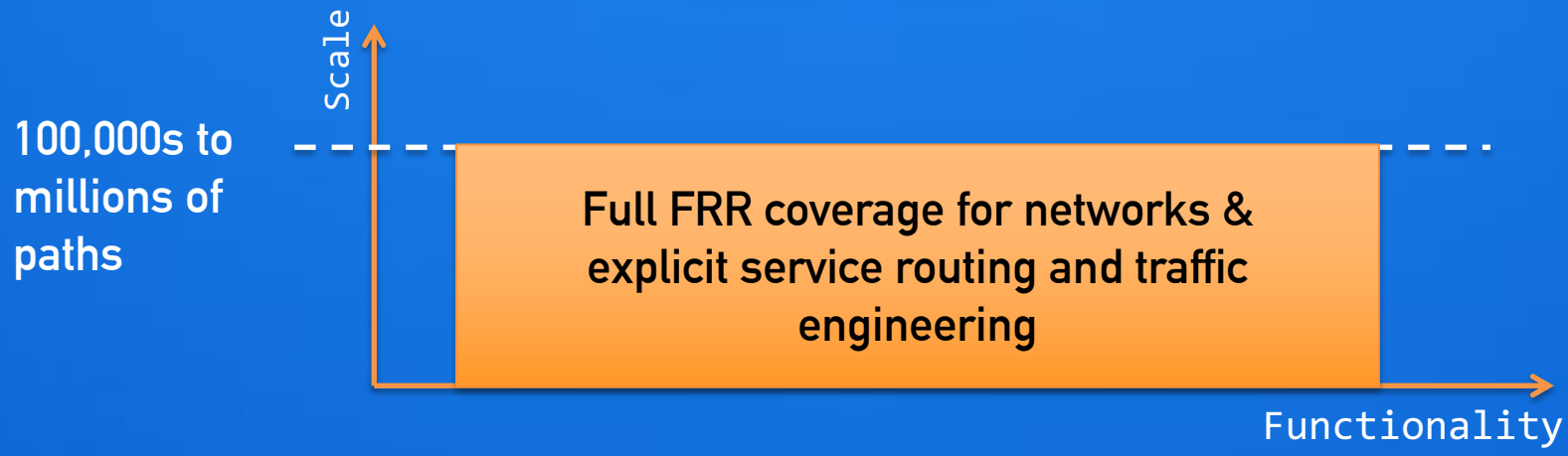
RSVP-TE is deployed – but in more limited scenarios, with lower scale than envisaged – a partial success, limited by scalability and operational complexity.



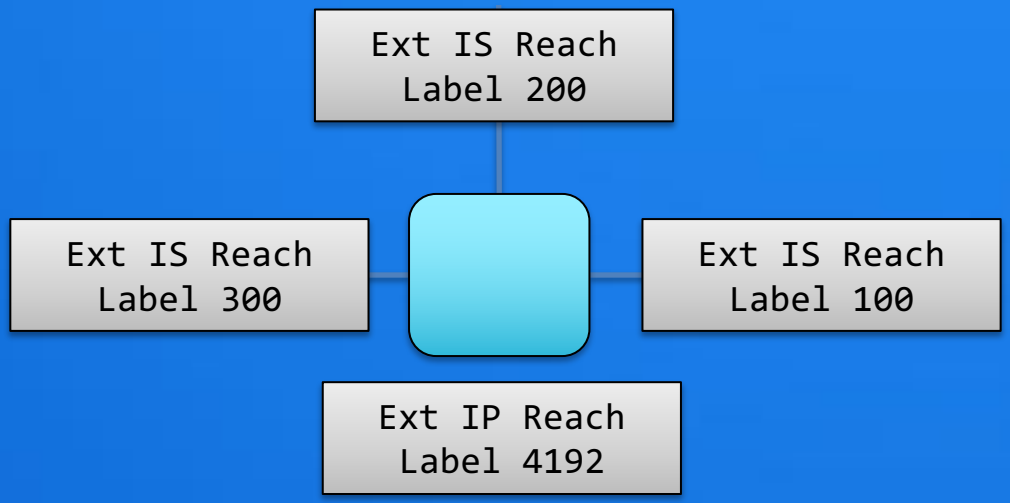
SEGMENT ROUTING/SPRING: CAN WE DO ANY BETTER?

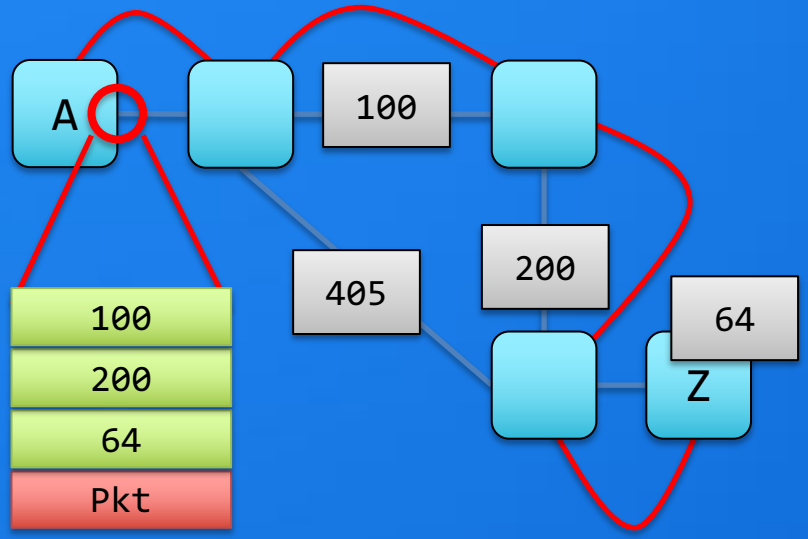
Real world problems still exist – e.g., B4 [SIGCOMM '13] – systems implementing traffic engineering over and above those identified in this discussion.

CAN WE IMPLEMENT EXPLICIT PATHS WHICH CAN BE PERFORMANCE AWARE IN A MANNER WHICH SCALES TO TODAY'S REQUIREMENTS, AND LOWERS THE COMPLEXITY TO SUPPORT THEM?



WHAT IS SEGMENT ROUTING?

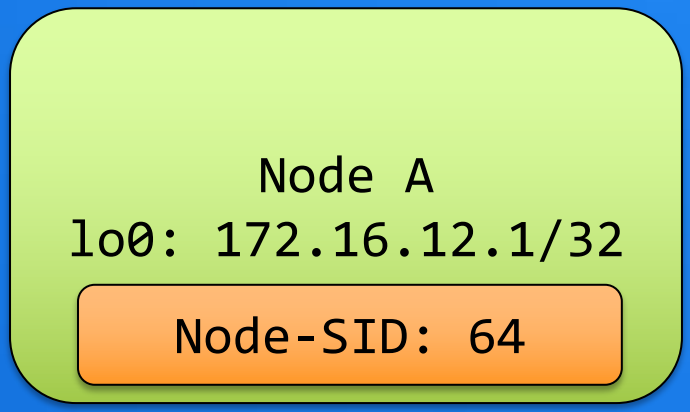




LABEL ADVERTISEMENT IN THE IGP.

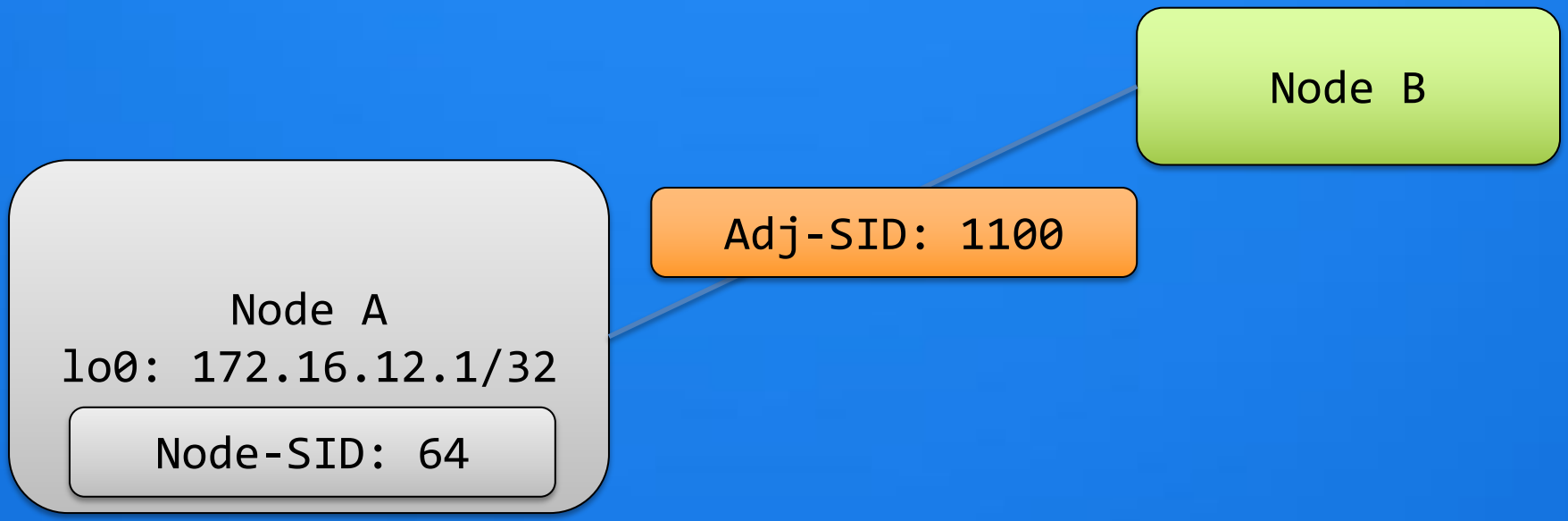
FORWARDING BASED ON STACKED LABELS.

SEGMENT IDENTIFIERS.



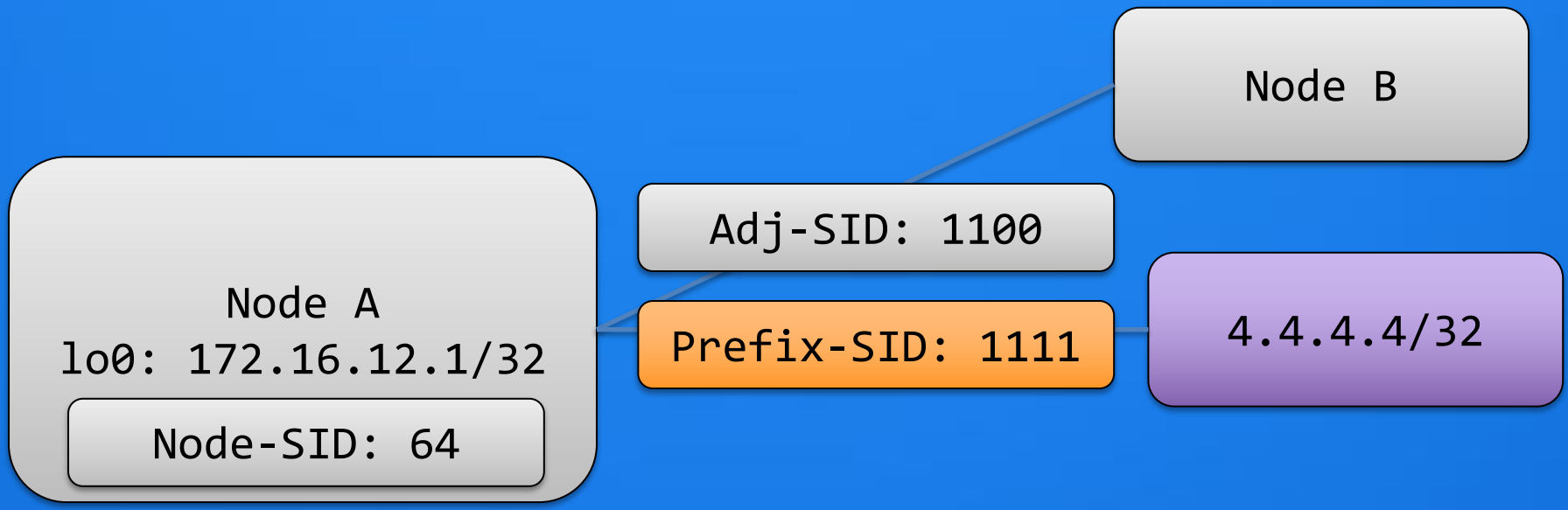
NODE SID:
GLOBAL (INDEXED) LABEL ALLOCATION INDICATING SPT TO ADVERTISING NODE (SPECIAL PREFIX SID).

SEGMENT IDENTIFIERS.



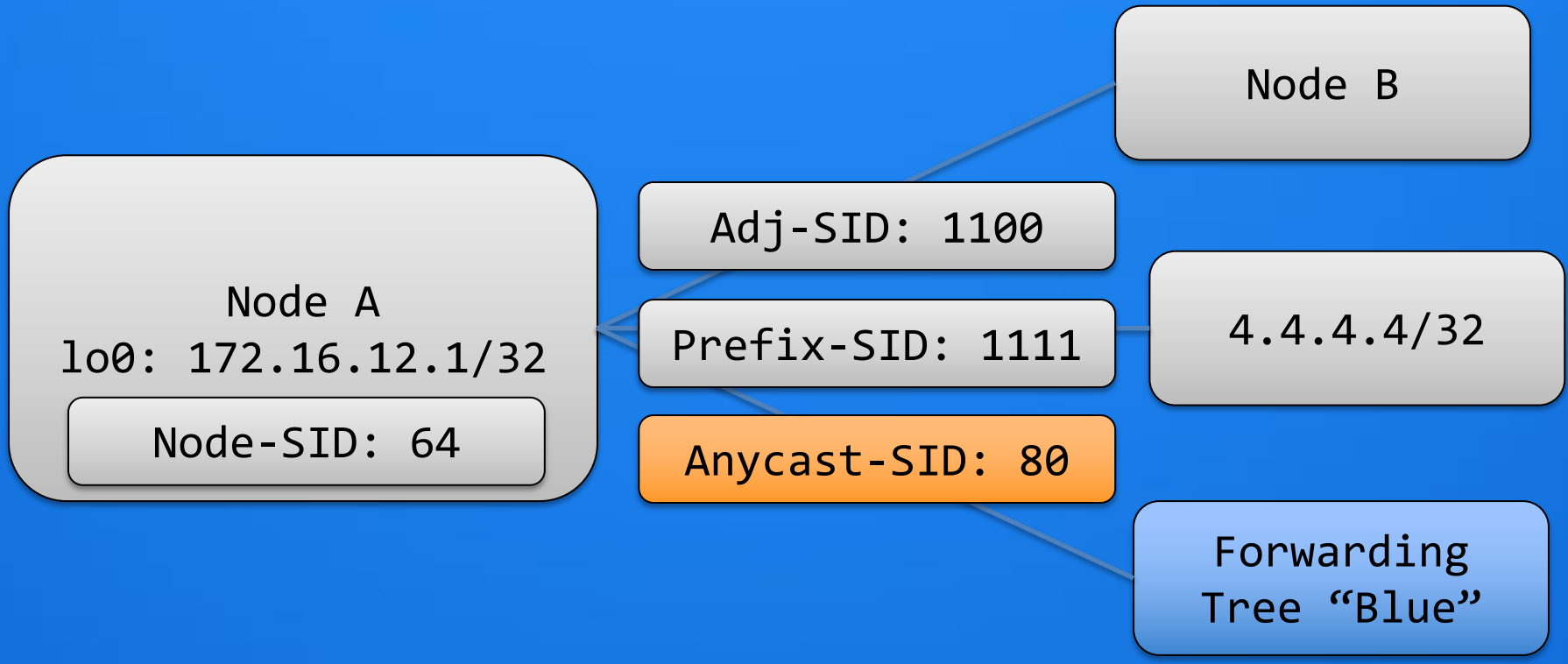
ADJACENCY SID:
LOCAL LABEL ALLOCATION INDICATING A LINK (OR SET OF LINKS) WITHIN THE IGP TOPOLOGY.

SEGMENT IDENTIFIERS.



PREFIX SID:
LOCAL LABEL ALLOCATION INDICATING AN IGP "LEAF" IP PREFIX (E.G. ATTACHED NODE).

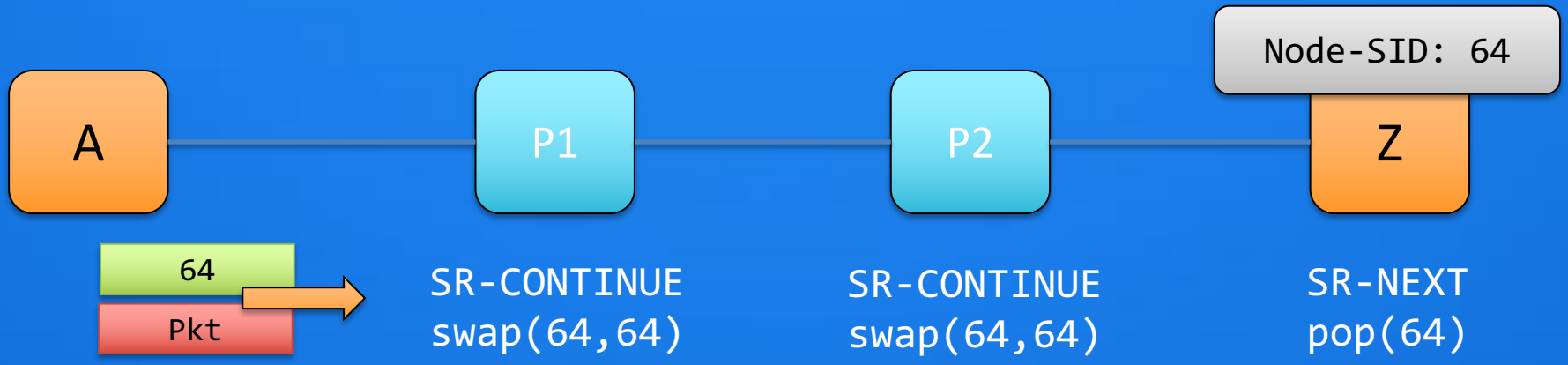
SEGMENT IDENTIFIERS.



IGP-ANYCAST SID:
GLOBAL LABEL ALLOCATION INDICATING REACHABILITY TO A CERTAIN RESOURCE OR FORWARDING PATH.

SPRING FORWARDING.

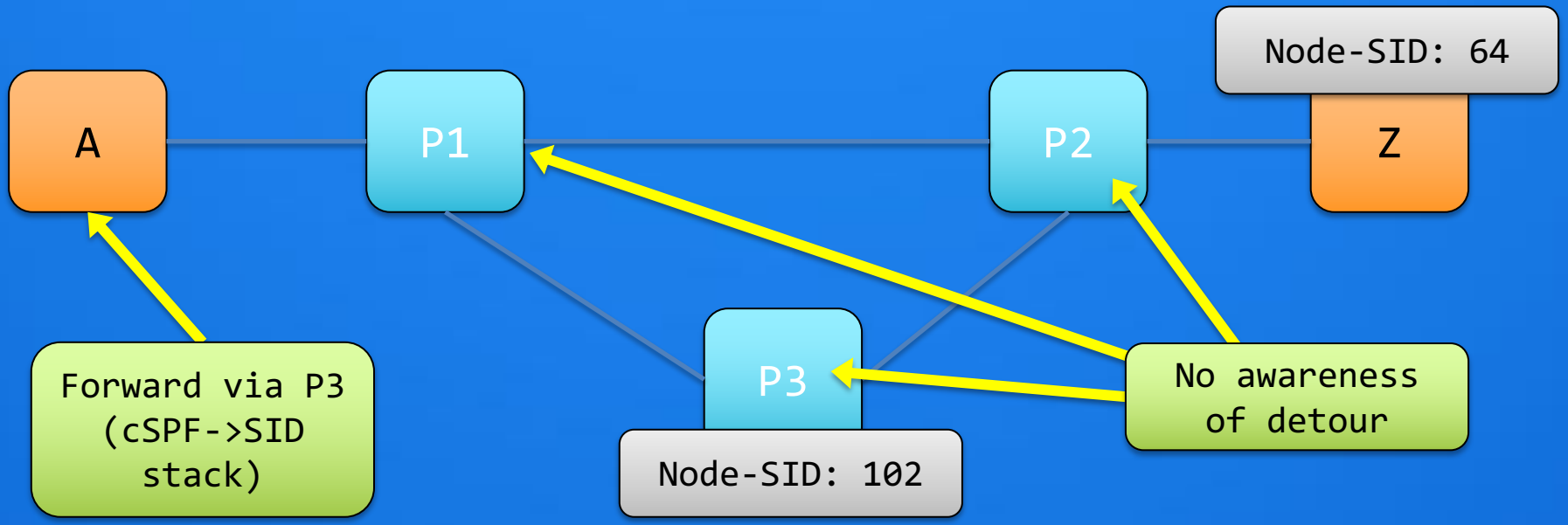
NODE-TO-NODE ALONG SPT:

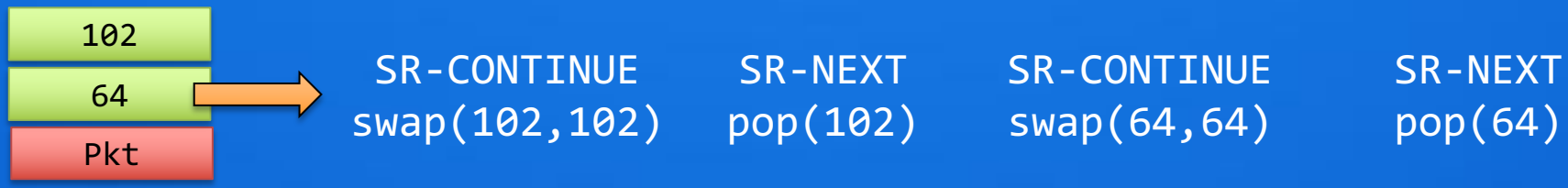


NO NEED FOR LDP FOR FORWARDING TO NODES WITH NODE-SID, OR IP ROUTES WITH IGP-PREFIX-SID – CAN ELIMINATE LDP AND LDP-IGP SYNC.

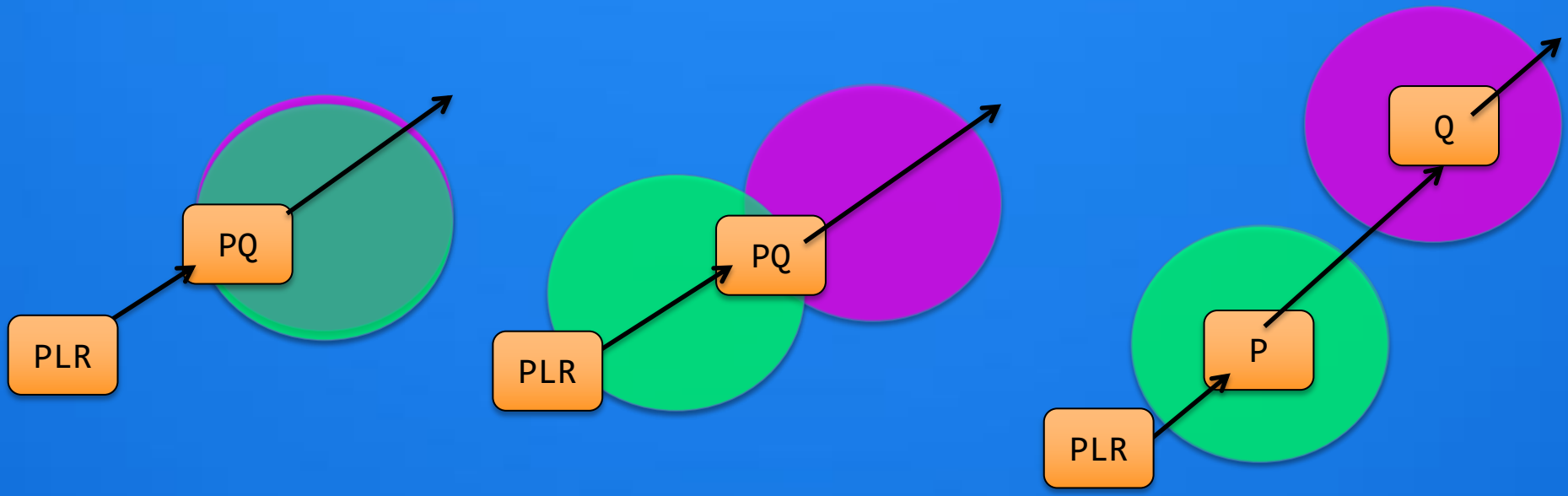
SPRING TACTICAL TE.

NODE-TO-NODE – TACTICAL TE:





IP FRR WITH SPRING.



"VANILLA LFA":

PUSH()

"REMOTE LFA":

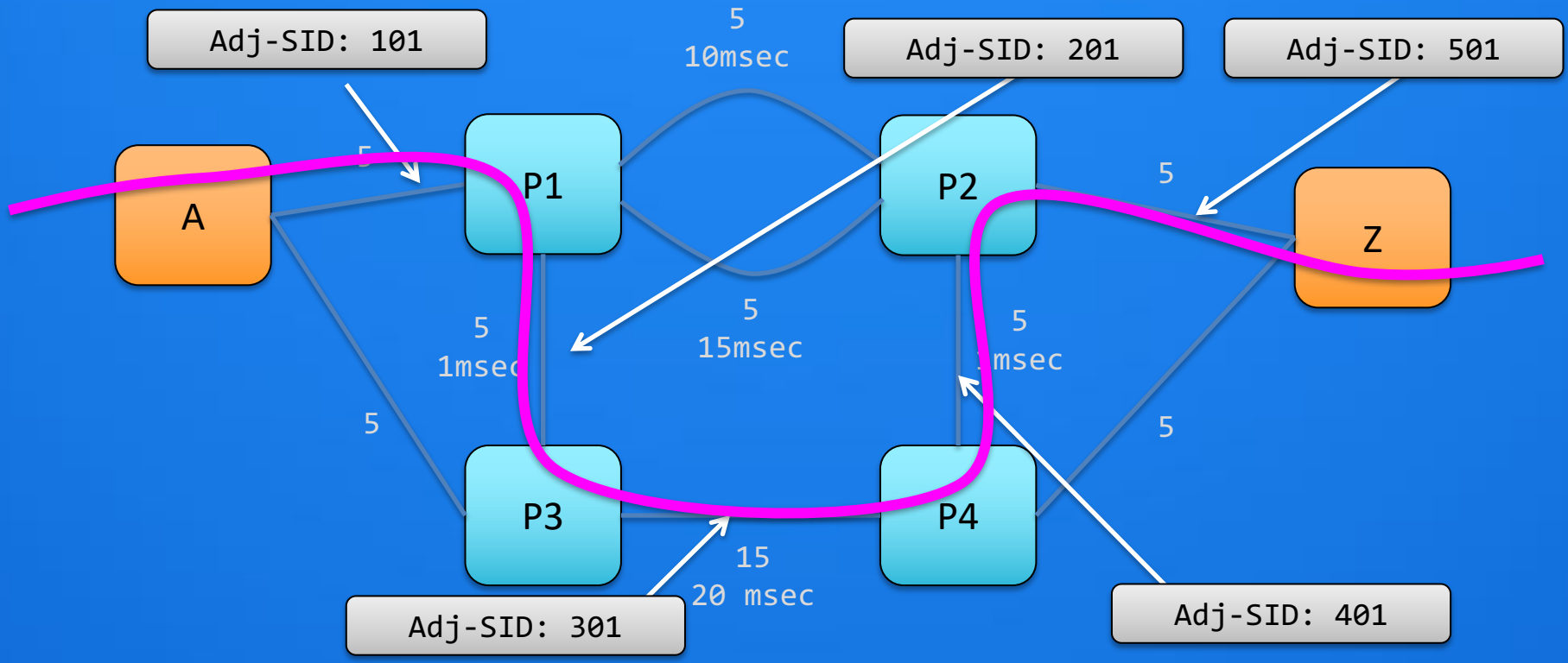
PUSH(PQ)

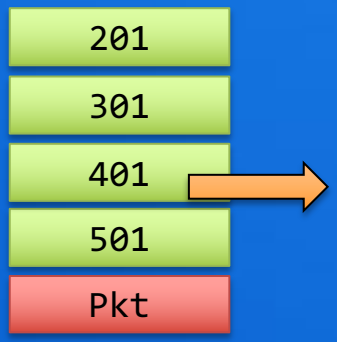
"DIRECTED LFA":

PUSH(P,P-Q)

SINGLE FORWARDING APPROACH FOR ALL FRR TYPES – NO ADDITIONAL CONTROL-PLANE REQUIRED.

EXPLICIT FORWARDING.





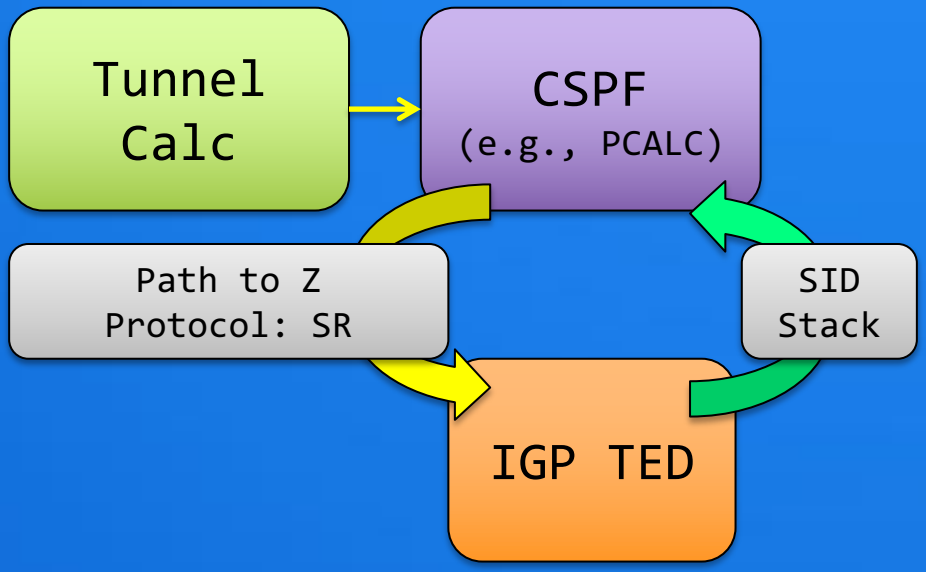
SR-NEXT
pop(201)
NHOP: P3

SR-NEXT
pop(301)
NHOP: P4

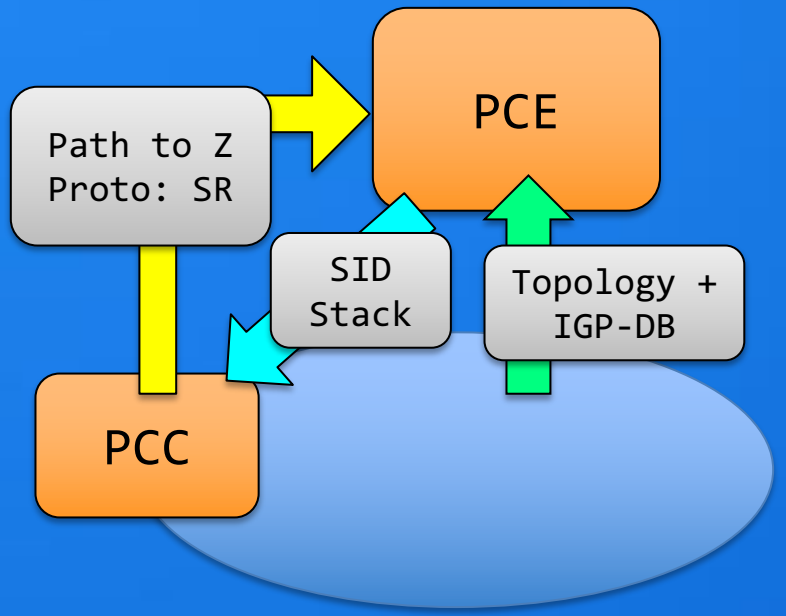
SR-NEXT
pop(401)
NHOP: P2

SR-NEXT
pop(501)

BASE COMPLEXITY: PATH CALCULATION.

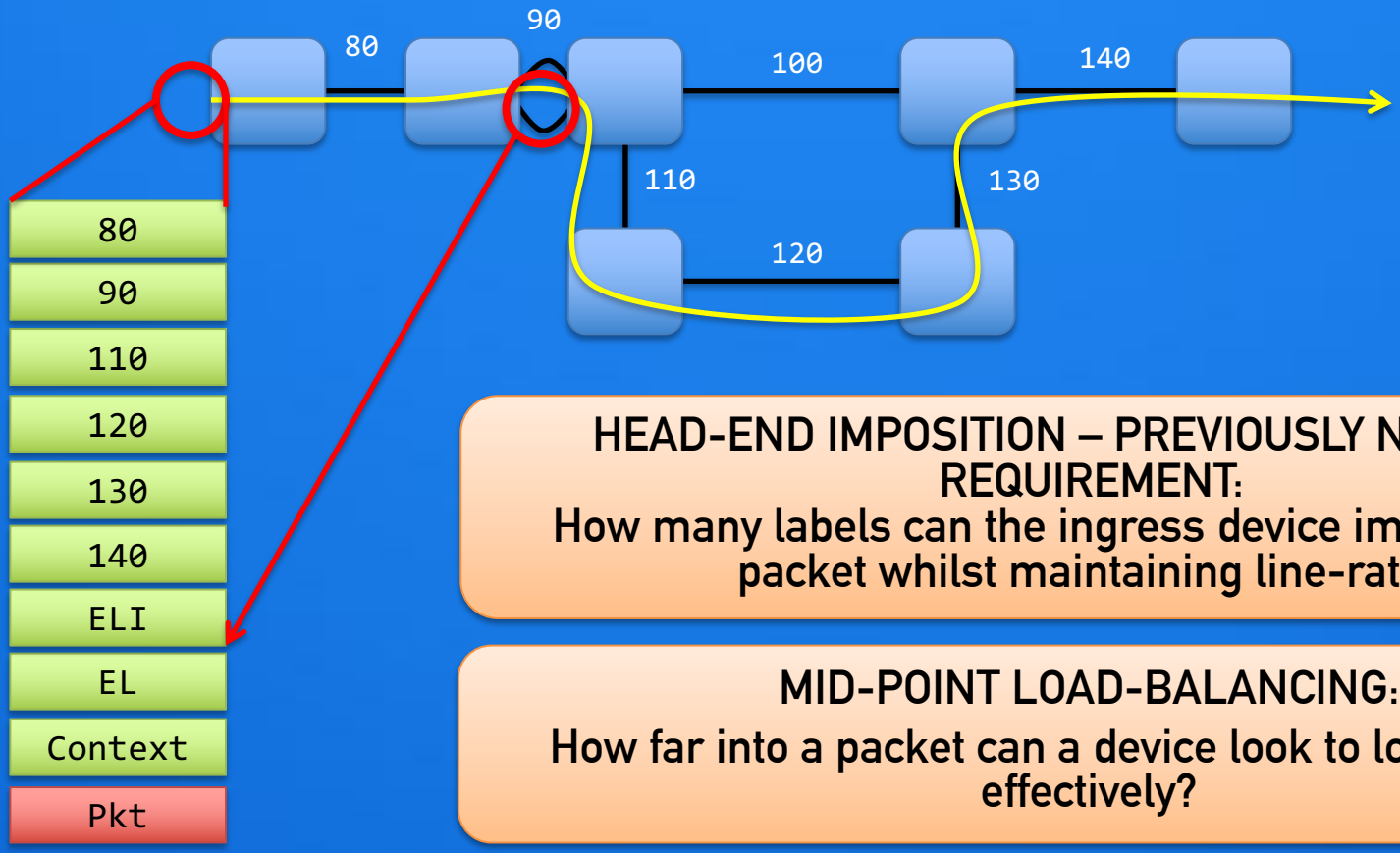


RE-USE OF EXISTING CALCULATION MACHINERY: WHERE HEAD-END HAS VISIBILITY OF ALL REQUIRED ROUTING INFO.



USE OF PATH COMPUTATION ELEMENT: GIVING HEAD-END ADDITIONAL VISIBILITY OR EXTERNAL INFO.

CHALLENGE: STACKED LABEL DEPTH.

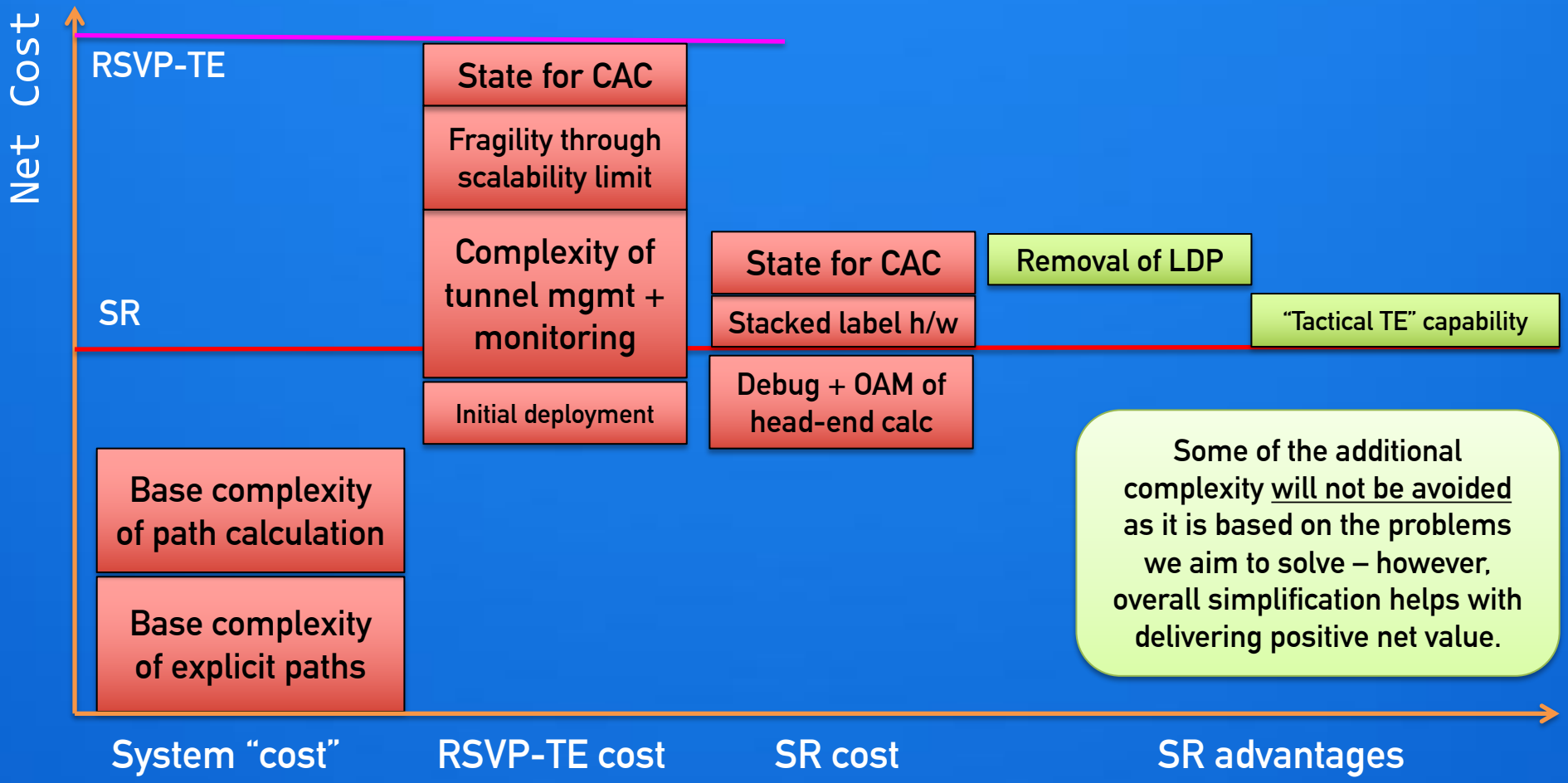


HEAD-END IMPOSITION – PREVIOUSLY NOT A KEY REQUIREMENT:
How many labels can the ingress device impose onto a packet whilst maintaining line-rate?

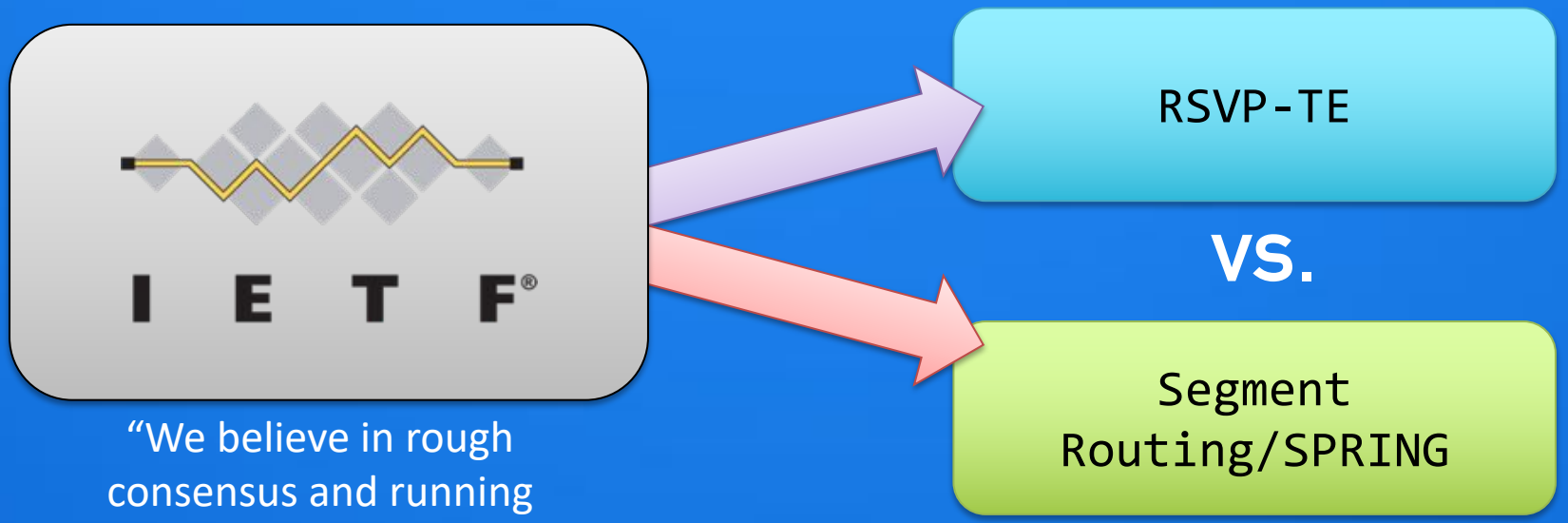
MID-POINT LOAD-BALANCING:
How far into a packet can a device look to load-balance effectively?

DIFFICULT CHALLENGE – DEPENDENT ON OPERATOR TOPOLOGY AND VENDOR HARDWARE OPTIMISATIONS

COMPARISON TO RSVP-TE IN TERMS OF COMPLEXITY.



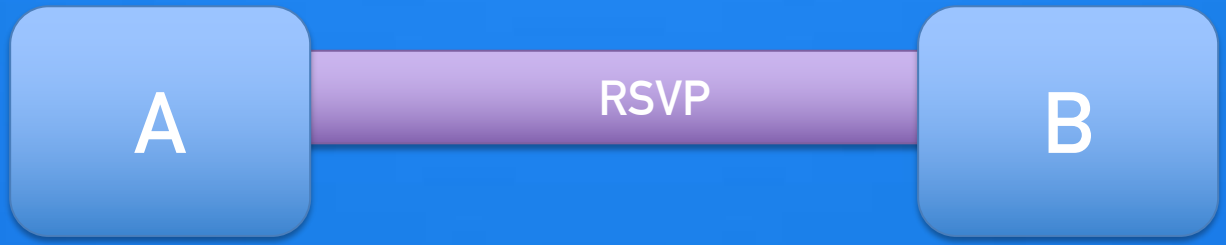
TECHNICALLY SUPERIOR VS. POLITICALLY ACCEPTABLE



"We believe in rough consensus and running code" – David Clark.

ADOPTION OF A TECHNOLOGY CAN BE VERY DEPENDENT UPON HOW IT IS PERCEIVED – OFTEN VERY RELATED TO THE INDIVIDUALS INVOLVED. A SUCCESSFUL PROTOCOL NEEDS SOME LUCK!

OOPS – HOW DO WE MIGRATE TO SR?



CURRENT RSVP NETWORK – 100% OF BANDWIDTH IS RSVP RESERVED.

DOUBLE UP CAPACITY – CAPITAL INTENSIVE AND SLOW TO MIGRATE.

NEW TECHNOLOGY NEEDS TO CONSIDER HOW IT CAN BE INTRODUCED INTO EXISTING DEPLOYMENTS – NO GREEN FIELDS!

“DARK BANDWIDTH”

MEASURE SR UTILISATION – LIE TO RSVP-TE ABOUT LINK SIZES CHANGING BASED ON SR USAGE.

CONSTRAINED COMPLEXITY AND EFFICIENT RESOURCE UTILISATION

UNTESTED MODE OF EXISTING SYSTEM. CHANGING STABLE TO CANARY UNSTABLE.

DUAL RUNNING/COEXISTENCE IS ONE OF THE HARDEST ELEMENTS OF PROTOCOL DESIGN IN <L4!

SOME CONCLUDING THOUGHTS…

OPERATOR'S ISSUES ARE WIDE AND VARIED…	Protocol design that looks to solve multiple sets of problems, for multiple operators are those that maximise their probability of success.
IT'S EASY TO SEE PROBLEMS IN THE REAR-VIEW MIRROR…	The collected thoughts in this presentation are based on issues in live operational networks – thinking about "what-ifs" in real networks is a good plan (but won't capture every issue).
CONSIDER THE COST OF CHANGE…	Protocols that are already deployed have significant advantages over wholly new approaches – and may win out based on this – however, this does not prevent new deployments where there is sufficient value.

THANKS – QUESTIONS?

ROB SHAKIR ROBJS@GOOGLE.COM

SOUND INTERESTING?

Interested in software development around networks?

robjs@google.com

REFERENCES/LINKS

Some insight on use of TE in a large global network:

B4: Experiences with a Globally-Deployed Software Defined WAN, S.Jain, et al. (b4-sigcomm@google.com), SIGCOMM '13, https://people.eecs.berkeley.edu/~sylvia/cs268-2014/papers/b4-sigcomm13.pdf

"Patches" to RSVP-TE:

https://tools.ietf.org/html/draft-ietf-teas-rsvp-te-scaling-rec-03
https://tools.ietf.org/html/draft-ravisingh-teas-rsvp-setup-retry-01

Segment Routing/SPRING:

Requirements - https://tools.ietf.org/html/rfc7855
Architecture - https://tools.ietf.org/html/draft-ietf-spring-segment-routing-10
MPLS Instantiation of SR –
https://tools.ietf.org/html/draft-ietf-spring-segment-routing-mpls-06
OSPF SR Extensions –
https://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-10
IS-IS SR Extensions -
https://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-09