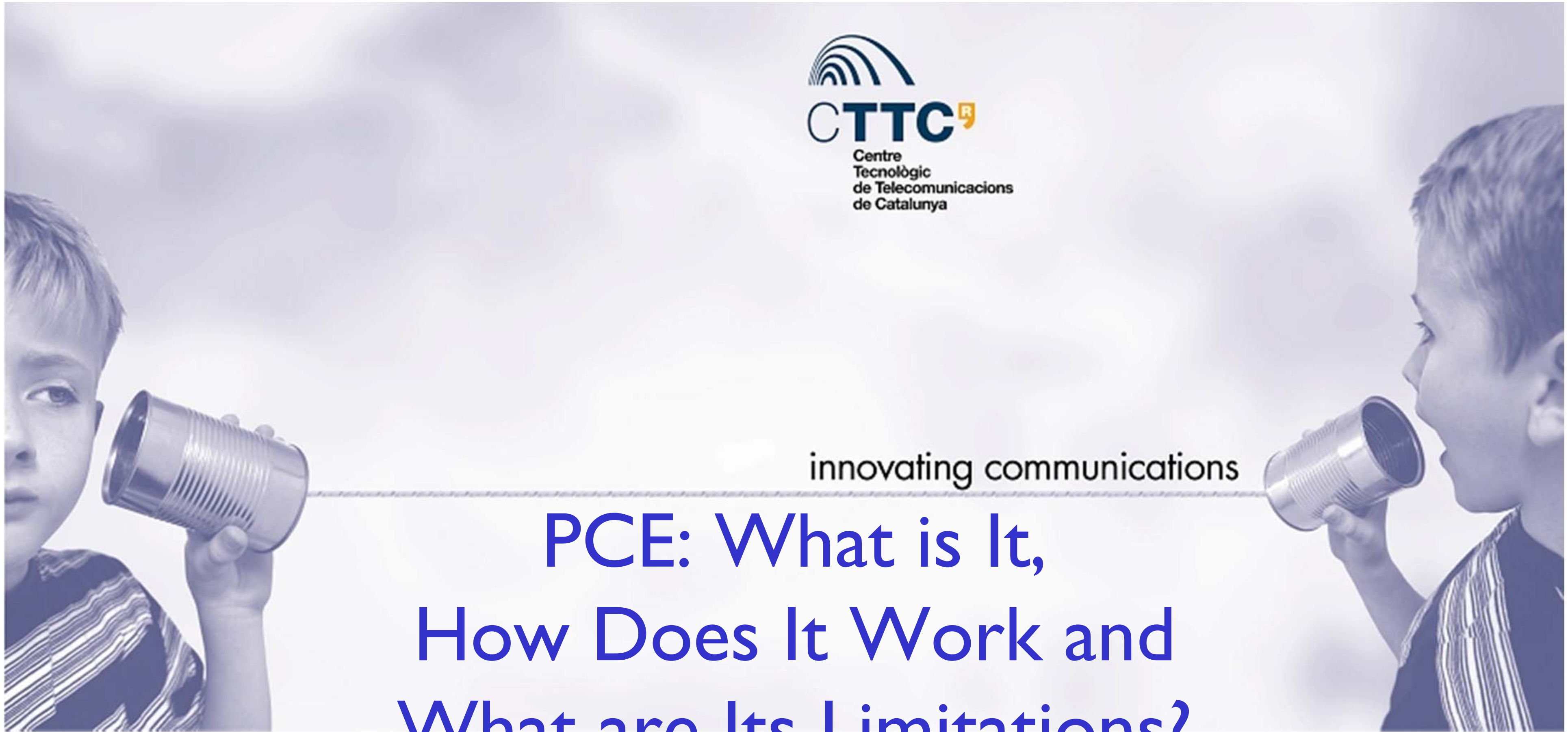


PCE: What is It, How Does It Work and What are Its Limitations?

Raul Muñoz, Ramon Casellas, Ricardo Martínez.

Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)

Castelldefels (Barcelona) - Spain

OFC/NFOEC 2013, Anaheim, CA, USA, March 20th, 2013

Outline

- GMPLS-controlled optical networks and PCE architecture:
  - Routing, signaling and path computation.
- Limitations of GMPLS-controlled optical networks and PCE-based solutions.
  - Impairment-aware path computation.
  - Multi-domain path computation.
  - Multi-layer path computation.
- Limitations of PCE.
  - Synchronization of the Traffic Engineering Database.
  - Increase of the path computation blocking.
  - Suboptimal path computation algorithms.
- Stateful PCE.
  - Applicability to SDN and its limitations.
- Conclusions.

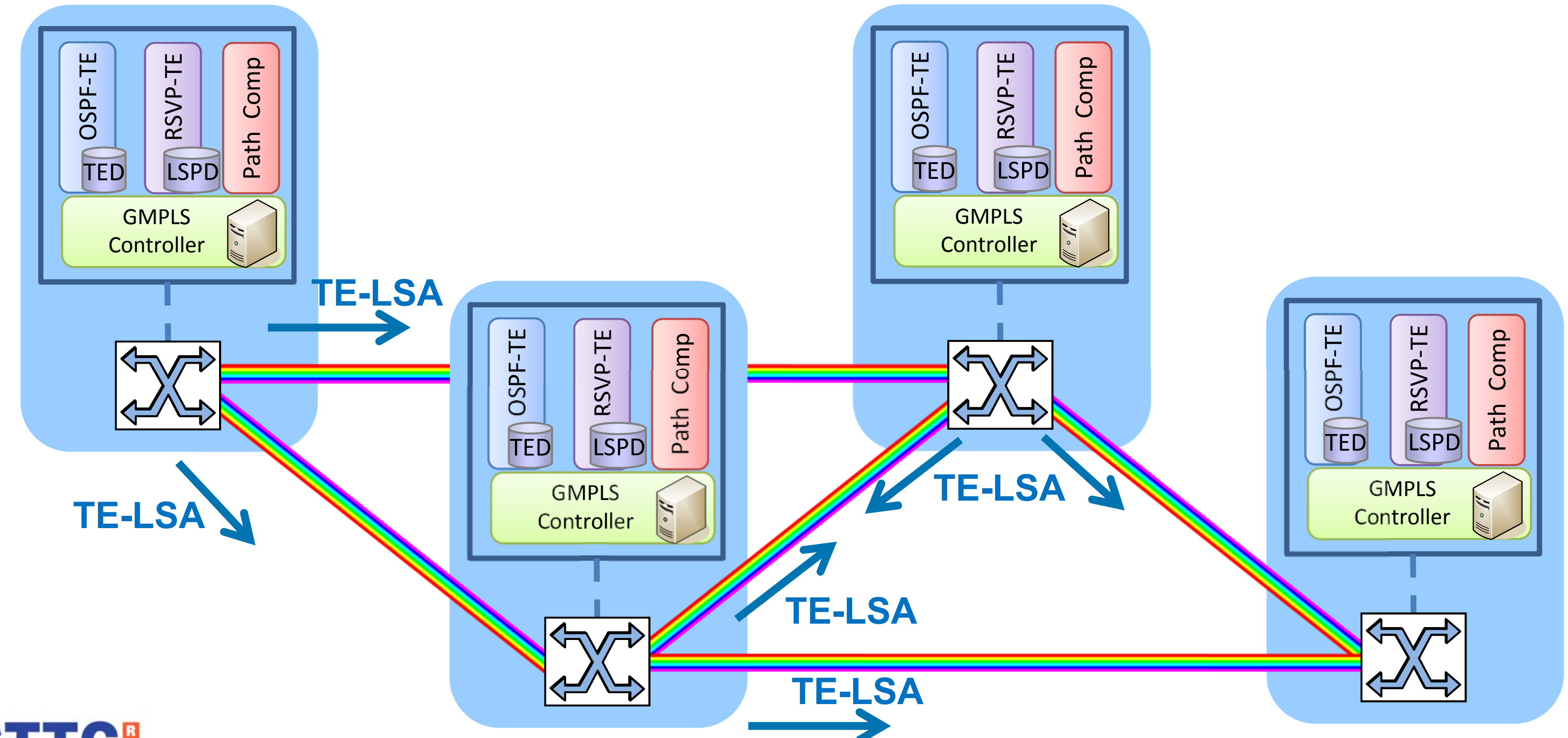
GMPLS-controlled optical networks and PCE architecture

GMPLS-controlled optical transport networks

- A GMPLS control plane allows the automation of transport networks through a common set of functions (e.g., path computation) and interconnection mechanisms (e.g., signaling and routing):
  - Connection provisioning and recovery, traffic engineering and QoS.
- IETF defines a set of standard GMPLS protocols, mainly:
  - OSPF-TE / IS-IS-TE routing protocol used for topology and network resource dissemination.
  - RSVP-TE signaling protocol used for setting up the end-to-end connections.
  - LMP Link management protocol for the creation of the control channel infrastructure and fault localization.
- A GMPLS control plane is a distributed entity composed of:
  - GMPLS Connection Controllers (one per node) which execute several collaborative processes (RSVP-TE, OSPF-TE, path computation, etc.).
  - A Data Communication Network based on IP control channels (IPCC) allows the exchange of control messages between GMPLS controllers.

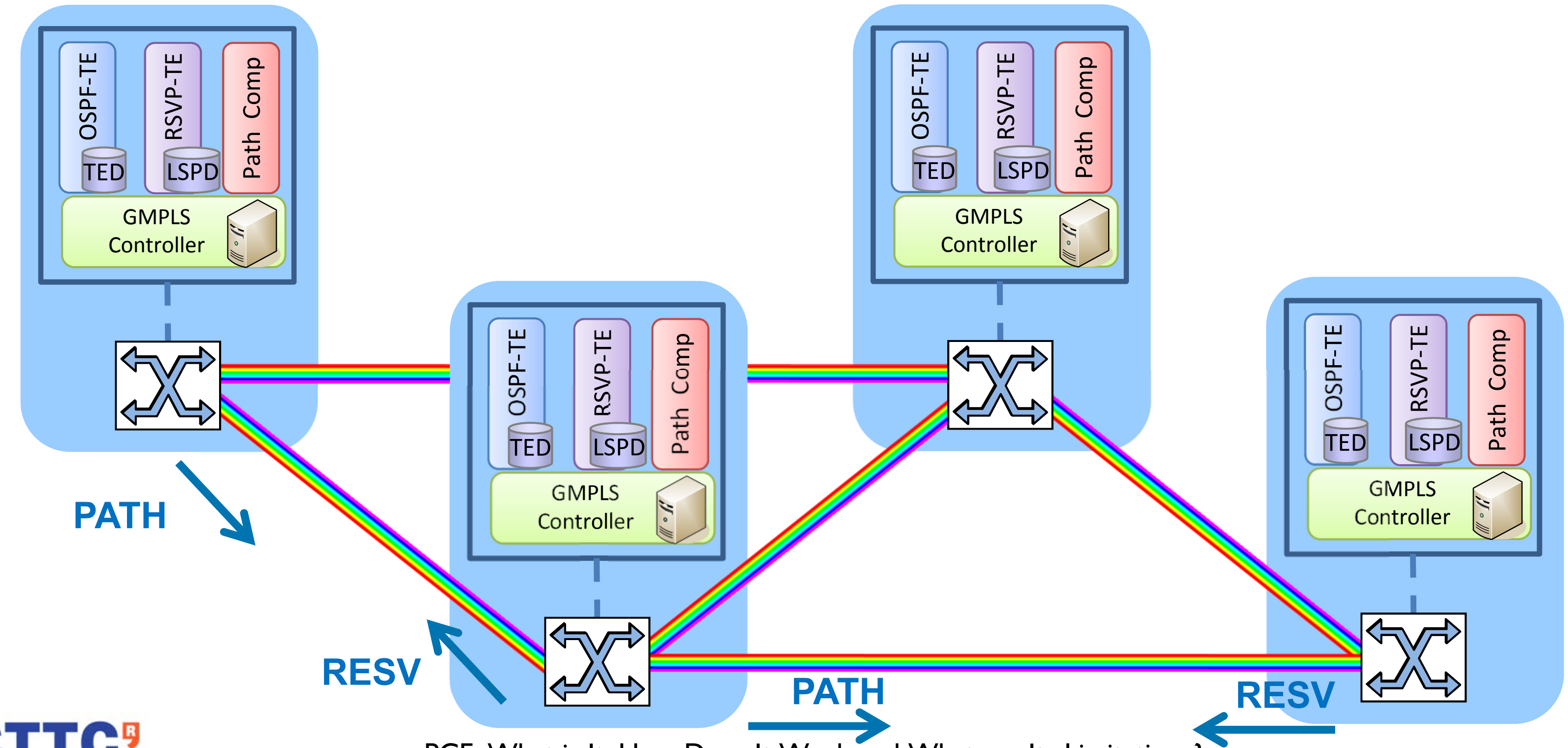
GMPLS OSPF-TE/IS-IS-TE routing

- GMPLS routing protocol disseminates changes in the network state (topology and resources) through the exchange of TE LSAs being then collected in the Traffic Engineering Database (TED).
- It allows GMPLS controllers to update local TEDs and maintain a global picture of current network topology and resource availability.



GMPLS RSVP-TE signaling

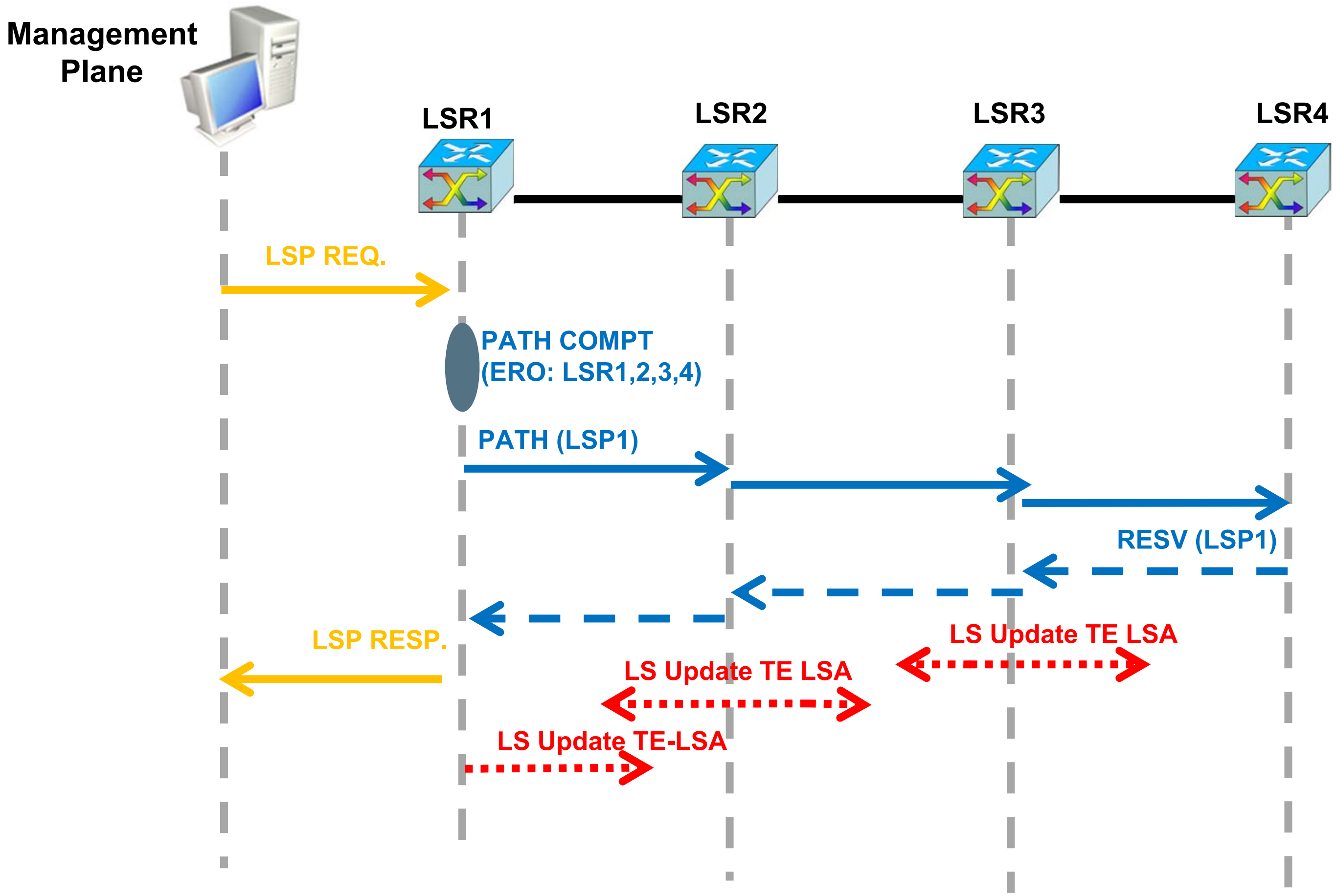
- GMPLS RSVP-TE signaling is responsible for the provisioning and control of connections (i.e., LSPs).
- Each RSVP-TE connection controller manages the state of all the connections originated, terminated or passing-through a node, stored in the LSP Database (LSPD).
- Signaling involves the exchange of (mainly) PATH and RESV messages, hop-by-hop.



GMPLS Path Computation – source-based

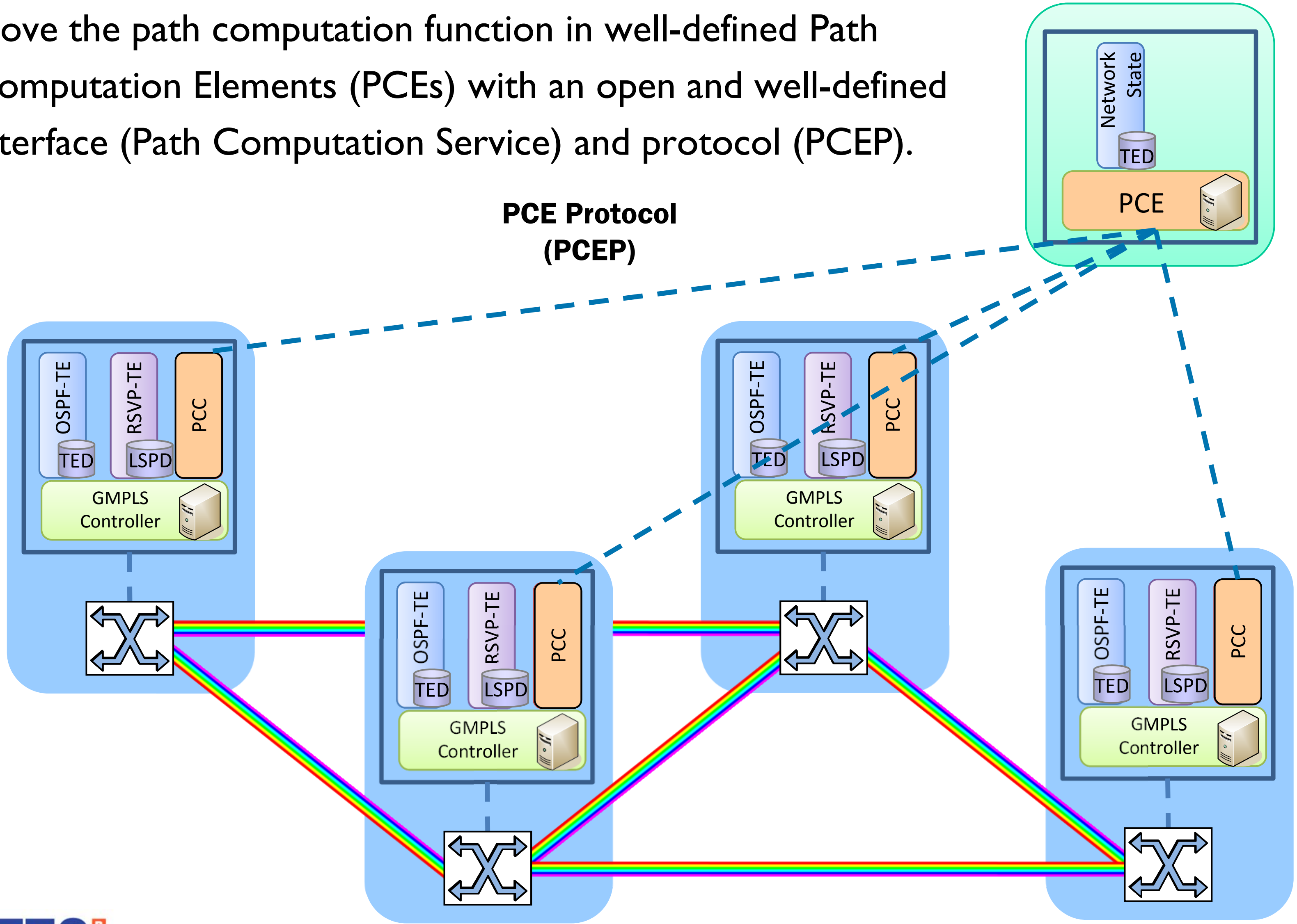
- Combined routing and resource assignment:
  - The source node computes both the spatial path (nodes and links) and assigns the resources (wavelengths, 3R, spectrum) using as input the TED information.
  - This approach This approach requires that the routing protocol disseminates detailed network information (e.g., wavelength/nominal central frequency availability, 3R/WC, etc.).
- Routing and distributed resource assignment:
  - The source node only computes the spatial path using TED information (e.g., link aggregated unreserved bandwidth)
  - Resource assignment is performed at the destination / intermediate nodes during signaling in the backwards direction.
  - This approach requires collecting resource state information (e.g., wavelength/nominal central frequency availability, 3R/WC, etc.) during the signaling.

Example of GMPLS source-based path computation and provisioning



Introduction to Path Computation Element (PCE)

- Move the path computation function in well-defined Path Computation Elements (PCEs) with an open and well-defined interface (Path Computation Service) and protocol (PCEP).



Analogy of centralized Path Computation: Google maps

- Definition of PCE: an entity (component, application or network node) that is capable of computing a network path or route based on a network graph (TED) and applying computational constraints.

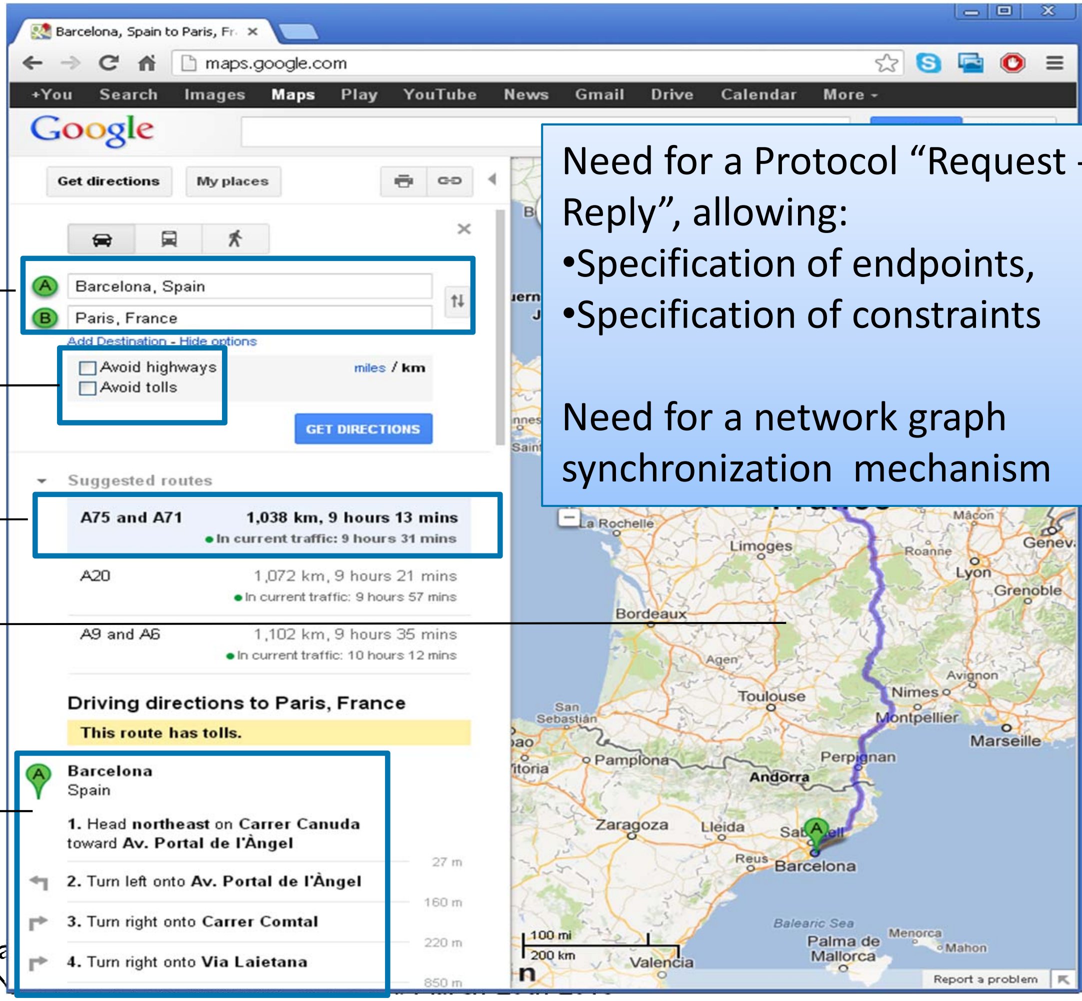
Request Endpoints

Constraints

Total Path Cost / Metric

Topology /
Traffic Engineering Database

Explicit Route (path)



Need for a Protocol “Request -> Reply”, allowing:
- Specification of endpoints,
- Specification of constraints

Need for a network graph synchronization mechanism

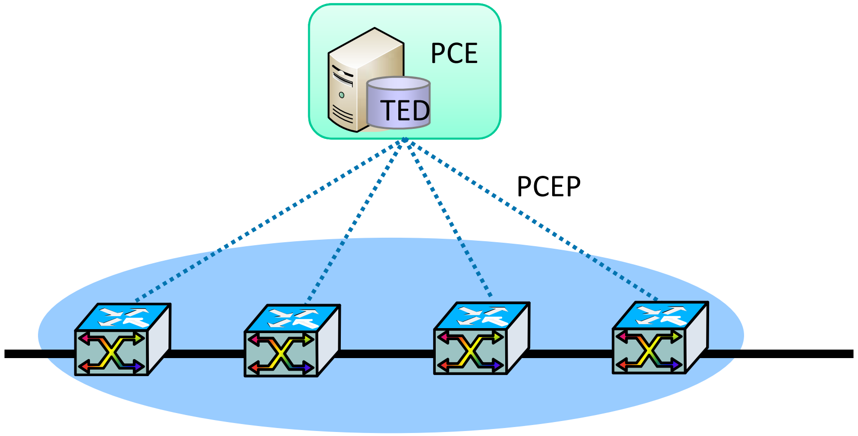
IETF PCE

- The initial driver for the deployment of PCEs was the increasing complexity of path computation justifying dedicated computational resources
- The decoupling of the path computation function from the GMPLS-enabled control plane in one or more dedicated entities provides:
  - More flexibility for Network operators in the control of their networks
  - The ability to apply their own policies in the development of the path computation algorithms, not bound to software updates within the controllers (closed and vendor-dependent)
  - Third party customized developments and upgrades of path computation algorithms
- IETF defines an standard and functional formalization of:
  - PCE global architecture
  - Communication interface and protocol (PCEP).
- Defined in 2006 for MPLS path computation (RFC4655) and eventually extended for GMPLS.

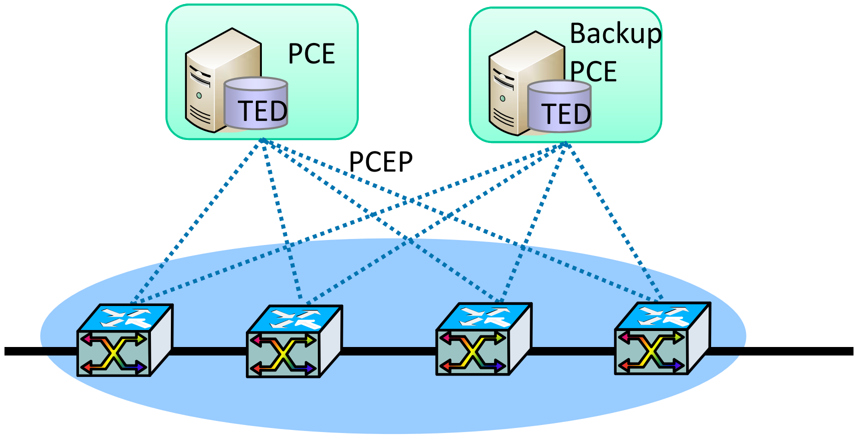
Basics of PCE Architecture

- Two main components:
  - Path Computation Client (PCC): any client application requesting a path computation to a Path Computation Element (PCE).
  - Path Computation Element (PCE): an entity computing a network path based on the network graph (TED) and applying computational constraints.
- A PCC can be located within a network node or in the NMS.
- A PCC may use different PCEs for path computation (e.g., to distribute the set of requests for load balancing purposes).
- The PCE entity is an application that can be located within a network node (composite PCE node) or in a dedicated server (external PCE node).
- PCE-based path computation models:
  - Centralized Path computation: single and centralized PCE.
  - Distributed Path computation: several PCEs are deployed in different switching layers or domains, each one serving requests from a subset of GMPLS controllers.

PCE-based Centralized Path Computation

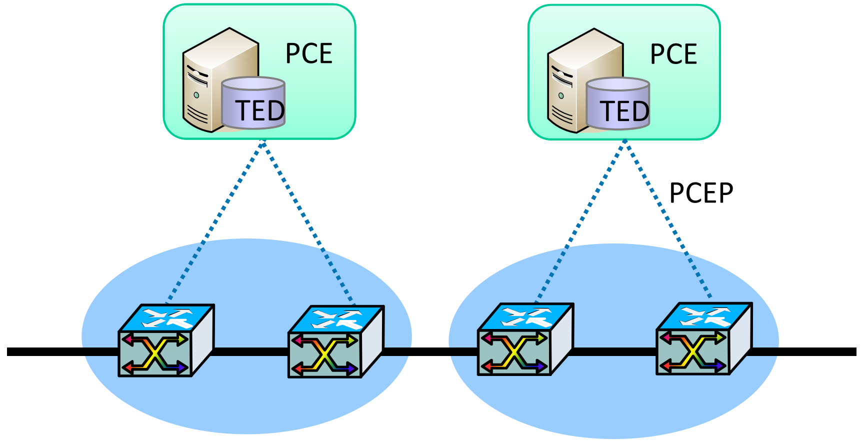


Centralized Path Computation

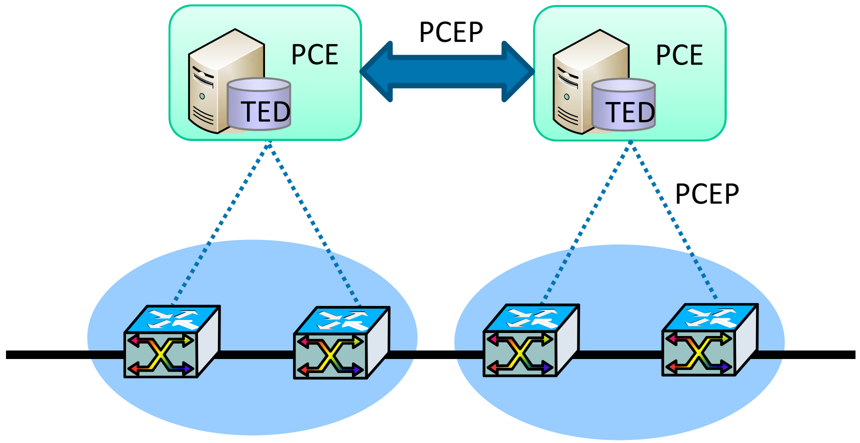


Centralized Path Computation with backup

PCE-based Distributed Path Computation



Distributed path computation without collaboration



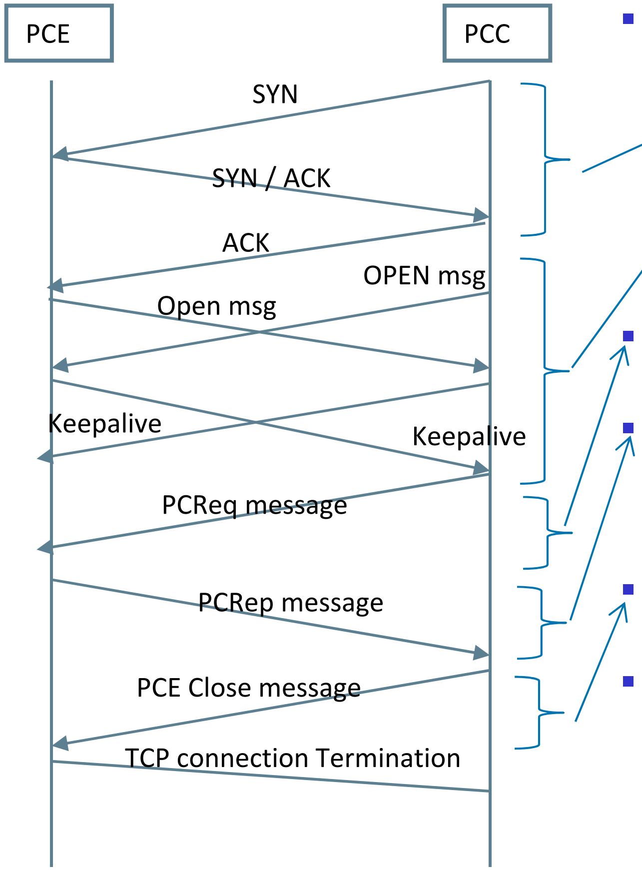
Distributed path computation with collaboration

TED synchronization

- Information on network topology and resource status to build the TED may be provided by:
  - Participation in IGP distribution of TE information :
    - A PCE may collect TE information by maintaining a routing adjacency with a GMPLS-enabled node in the domain (PCE can act as a IGP passive listener)
    - Local PCE TED can be constructed by sniffing e.g., OSPF-TE TE LSA exchange
  - Out-of-band synchronization:
    - Some mechanism (e.g. a topology server) is used by the PCE to retrieve the TED
    - Such a mechanism could be either incremental (like IGP) or involving a bulk transfer of the complete TED -> may lead to TED synchronization problems
- Enhanced TED may include additional info obtained from other means than IGP
- In anycase, the TED can be updated by the PCE after the path computation.

Basics of PCEP Session

- PCEP is used for communicating both PCC and PCE, as well as between PCEs.
- It is a standard, flexible and extensible interface and protocol, defined in RFC 5440.

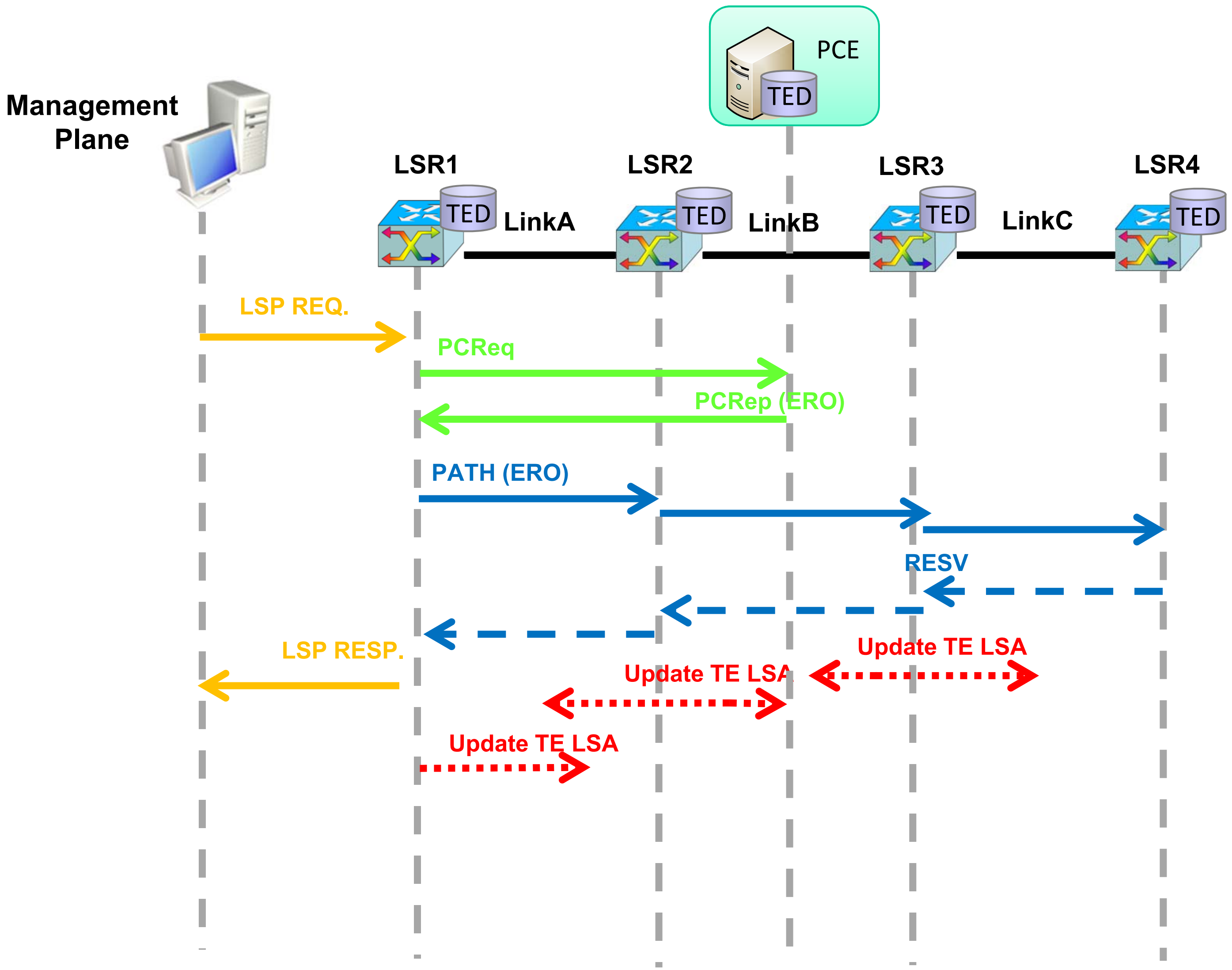


- Initialization phase:
  - Establishment of a TCP connection.
  - Establishment of a PCEP session (negotiation: Keepalive and Dead Timers, supported OFs, capabilities)
- Path computation Request sent by a PCC to a PCE
- Path computation Reply sent by a PCE to a PCC: (positive and negative reply)
- Termination of the PCE session and TCP connection
- A PCEP session can be either *transient* (closed after the request is served) or *permanent* (monitored by means of keep-alive messages).

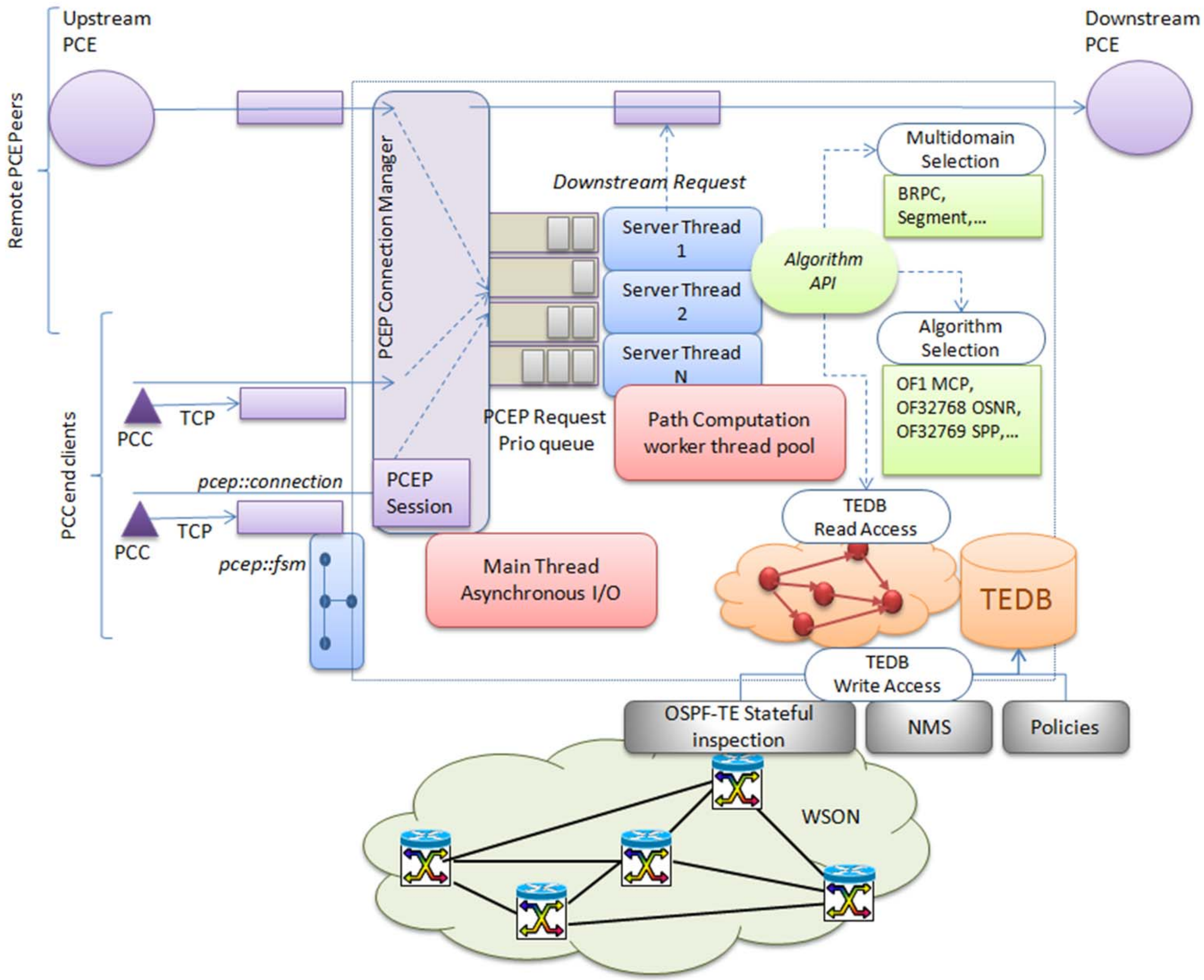
Path Computation Request / Reply

- A PCC sends a path computation request to the PCE (PCReq message) with a variety of objects that specifies the set of constraints and attributes:
  - Endpoints (source and destination node addresses).
  - Objective Function: Requested algorithm (optimization criteria).
  - Traffic parameter (e.g., Requested bandwidth).
  - Requested Switching Capability and Encoding Type.
  - Exclusion of network nodes, links, labels or whole domains (Exclude Route Object).
  - Inclusion of network nodes, links, labels or whole domains (Inclusion Route Object).
  - Re-optimization of an existing path avoiding double-booking (Reported Route Object).
  - Non-synchronized computation of a set of paths.
  - Synchronized computation of a set of paths (SVEC Object).
  - Global Objective Function and constraints applied for global concurrent optimization.
- If path computation succeeds, PCE replies (PCRep message) with the computed path specified by means of Explicit Route Objects (EROs).

Example of GMPLS PCE-based path computation



PCE design with PCE peers developed in the CTTC ADRENALINE testbed



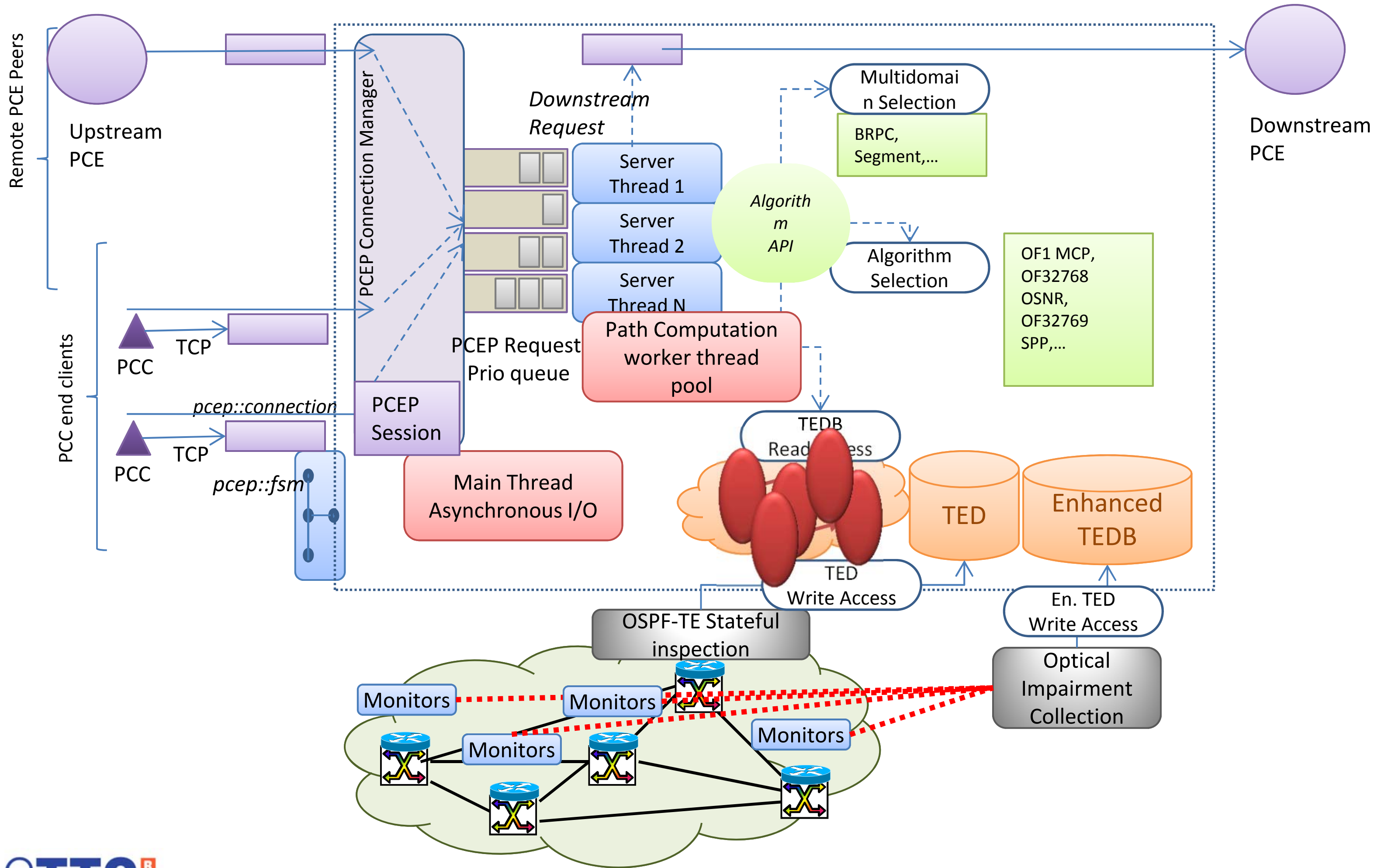
Limitations of GMPLS-controlled optical networks and PCE-based solutions

Impairment-aware path computation problem

- In WSON/SSON, the optical signal degrades due to the accumulation of physical impairments whilst the signal travels from the source towards the destination:
  - ASE noise, Chromatic Dispersion, PMD and non-linear effects.
- Optical physical impairments must be taken into account during the RWA/RSA.
- The dissemination of optical impairments may cause a significant control plane overhead problem requiring at each controller to maintain a large amount of data.
  - No standard routing protocol extensions have been defined so far.
- Some impairment information may not be directly mapped to link/node TE attributes disseminated by the GMPLS routing protocol.
- Collection of physical impairments through RSVP-TE is sub-optimal:
  - The information is not available at the path computation time which may lead to compute unfeasible paths-> the path is only validated.

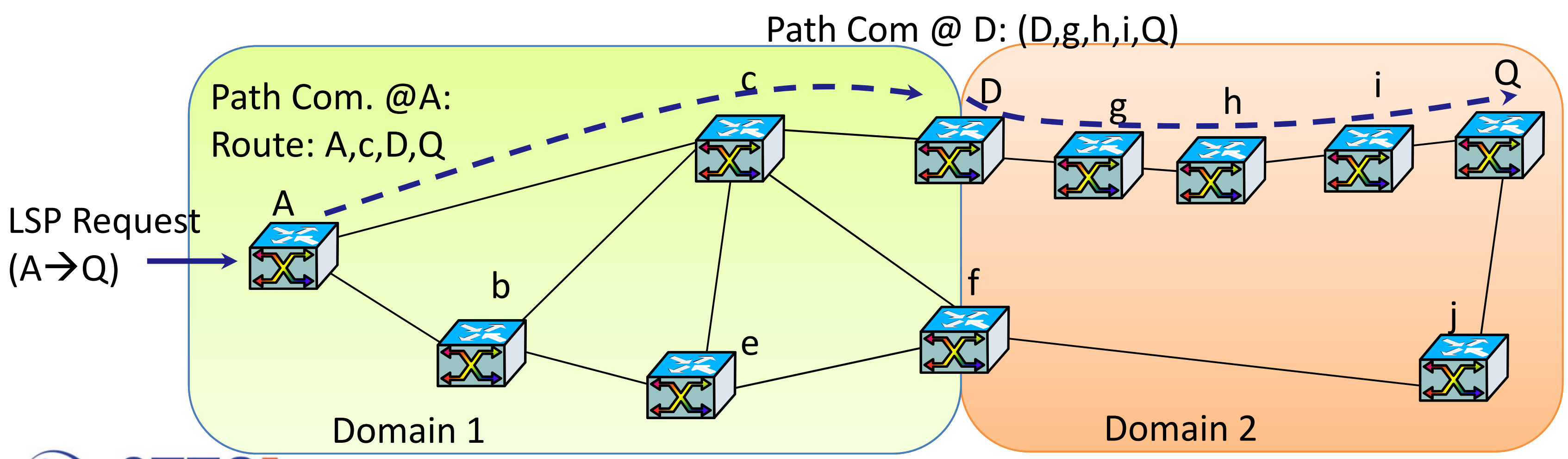
PCE-based solution: Enhanced TED

- Physical impairment information is gathered by dedicated monitors and stored in the Enhanced TED at the PCE (i.e., not flooded by the GMPLS controllers)



Multi-domain path computation problem

- Operators require to segment their network infrastructure into several domains (as IGP areas or AS) to enhance the scalability and/or for confidentiality reasons.
- The exchange of TE information between domains is limited to the dissemination of reachability:
  - GMPLS controllers have a complete view of the network topology and resources within their domain boundaries, but a limited visibility of other domains.
  - A source node is not able to compute an end-to-end multi-domain path with an strict list of nodes and links -> only a distributed per-domain path computation can be used.
  - The source node determines the egress domain node (not optimal).



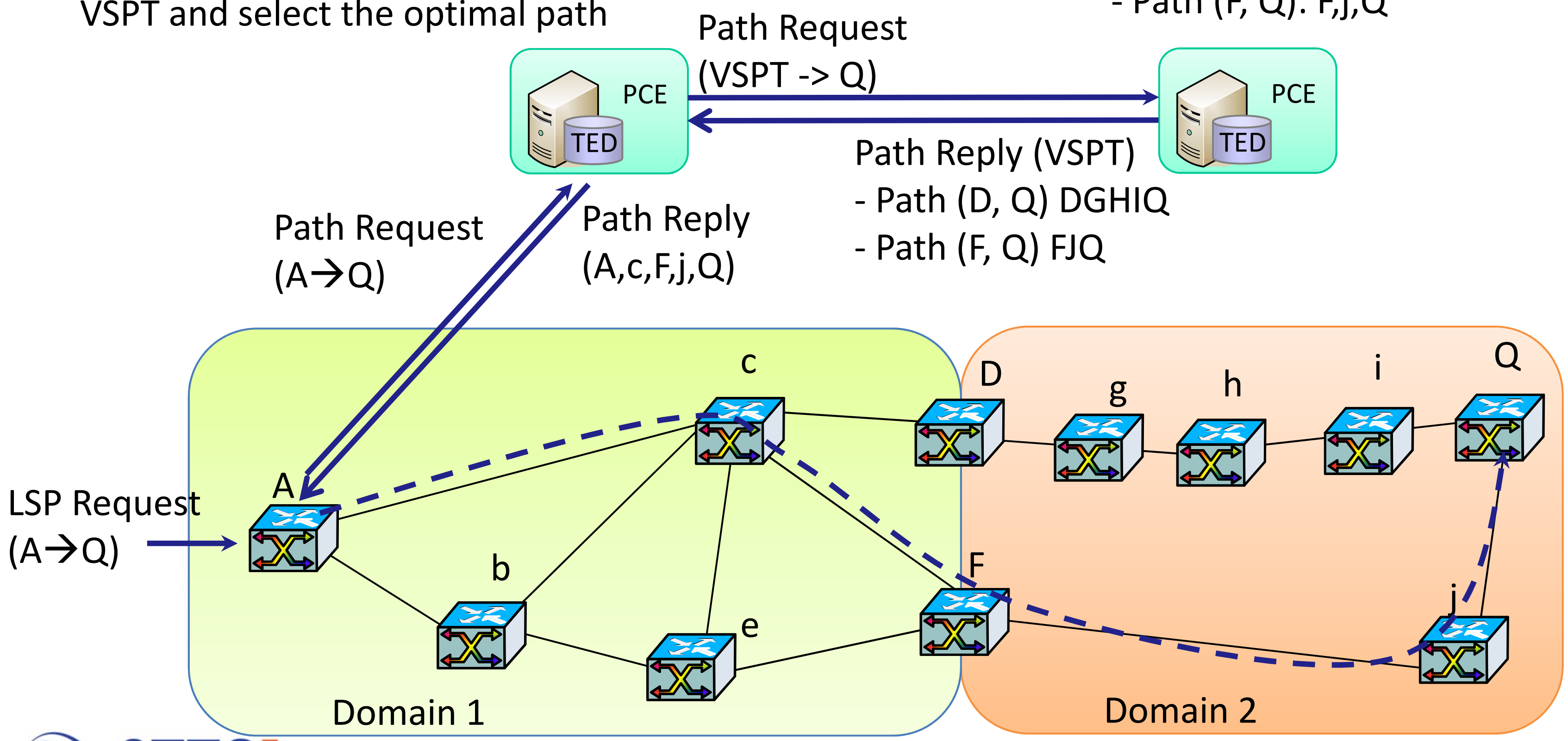
PCE-based solutions: Backward Recursive Path Computation (BRPC)

- BRPC computes the path in a reverse way, starting from the destination domain:
  - The destination domain PCE computes a virtual shortest path tree (VSPT) from the domain ingress nodes to the destination node.
- The destination domain PCE sends the computed VSPT to the upstream PCE
- Upstream PCEs compute their own VSPT, by:
  - Computing the optimal path from each domain ingress node that are adjacent to the upstream domain to each domain egress node adjacent to the downstream domain.
  - Building a tree from each of the ingress nodes to the destination node, using the computed optimal paths and the received paths in the VSPT.
  - Pruning the sub-optimal paths from the VSPT, before sending it to the next upstream domain PCE.
- The upstream domains apply this procedure up to the source domain.
- BRPC attains optimal path computation if the sequence of domains is known.
- Applied to meshed domains may be complex.

Example of BRPC

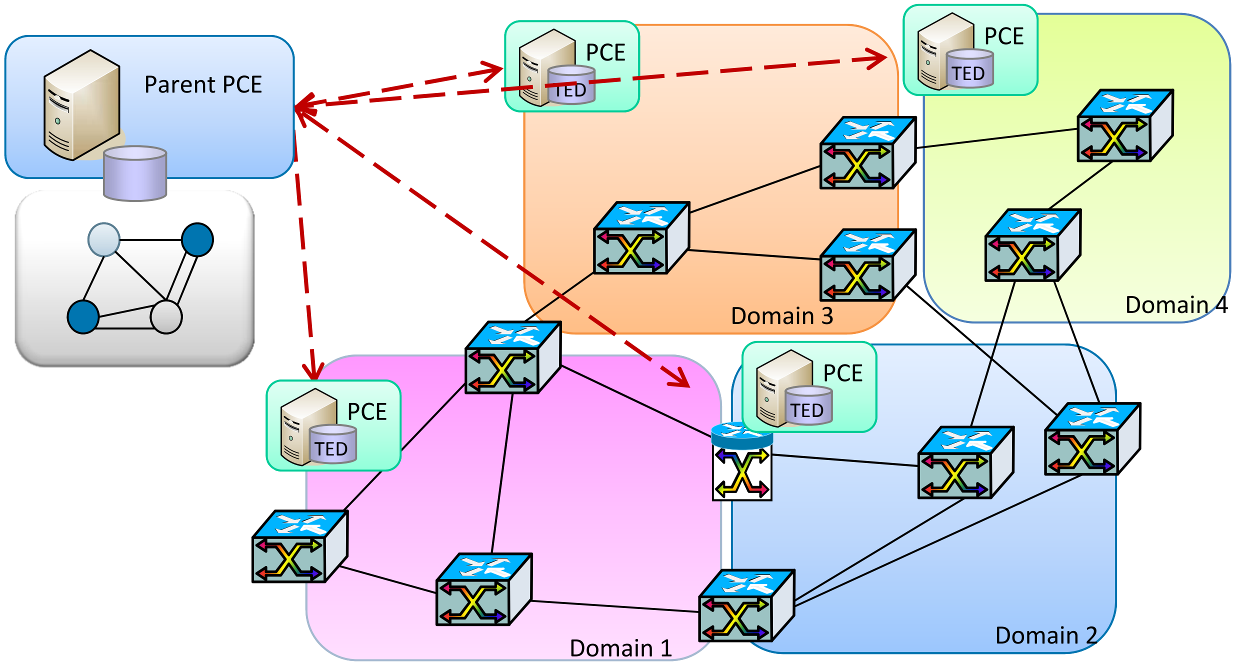
Compute optimal paths from source node to egress domain nodes:
- Path (A,D): A,c,D
- Path (A,F): A,c,F
Build a tree from the source to the destination node using the received VSPT and select the optimal path

Compute optimal path from ingress domain nodes to destination node:
- Path (D, Q): D,g,h,I,Q
- Path (F, Q): F,j,Q



PCE-based solutions: Hierarchical PCE (H-PCE)

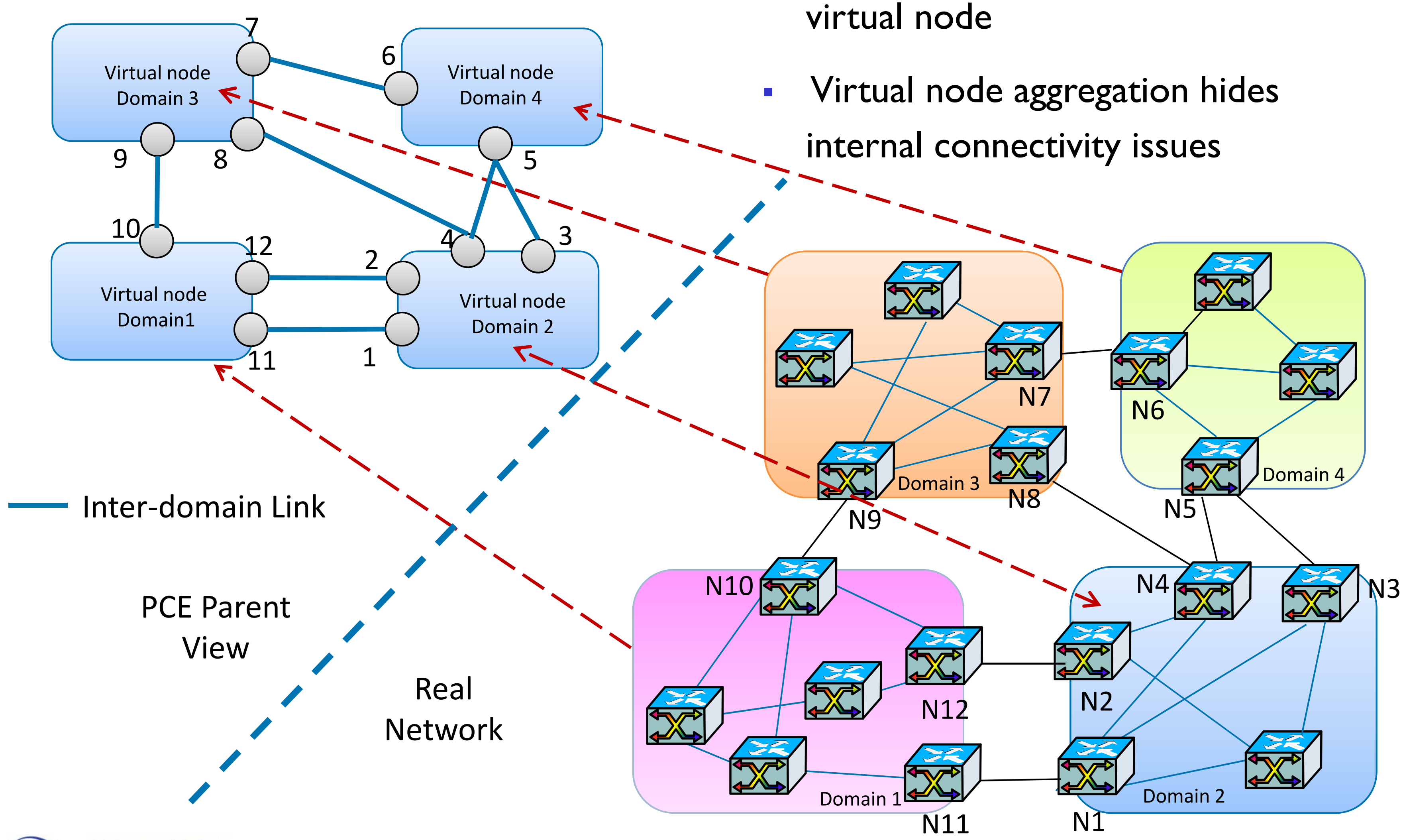
- 2 level H-PCE:
  - A single parent PCE maintains a domain topology map .
  - Each domain has at least one PCE capable of computing paths within the domain.



PCE-based solutions: Hierarchical PCE (H-PCE)

- The main issue in H-PCE is the selection of the optimal Domain Sequence in order to compute the optimal path, but the selection of such a sequence depends on the sub-paths in the domains.
- Topology
  - Parent PCE builds a simplified topology (how simplified?), examples:
    - Mode A: a domain is a node for the parent
    - Mode B: A child PCE computes paths that appear as virtual links
  - Child PCEs use regular TED
- Computation relies on a 2-step process
  - Domain sequence selection (parent)
    - may / may not account for inter-domain links
  - Domain segment computation (delegated to child PCEs) + composition
    - Vertical communication only → No siblings.

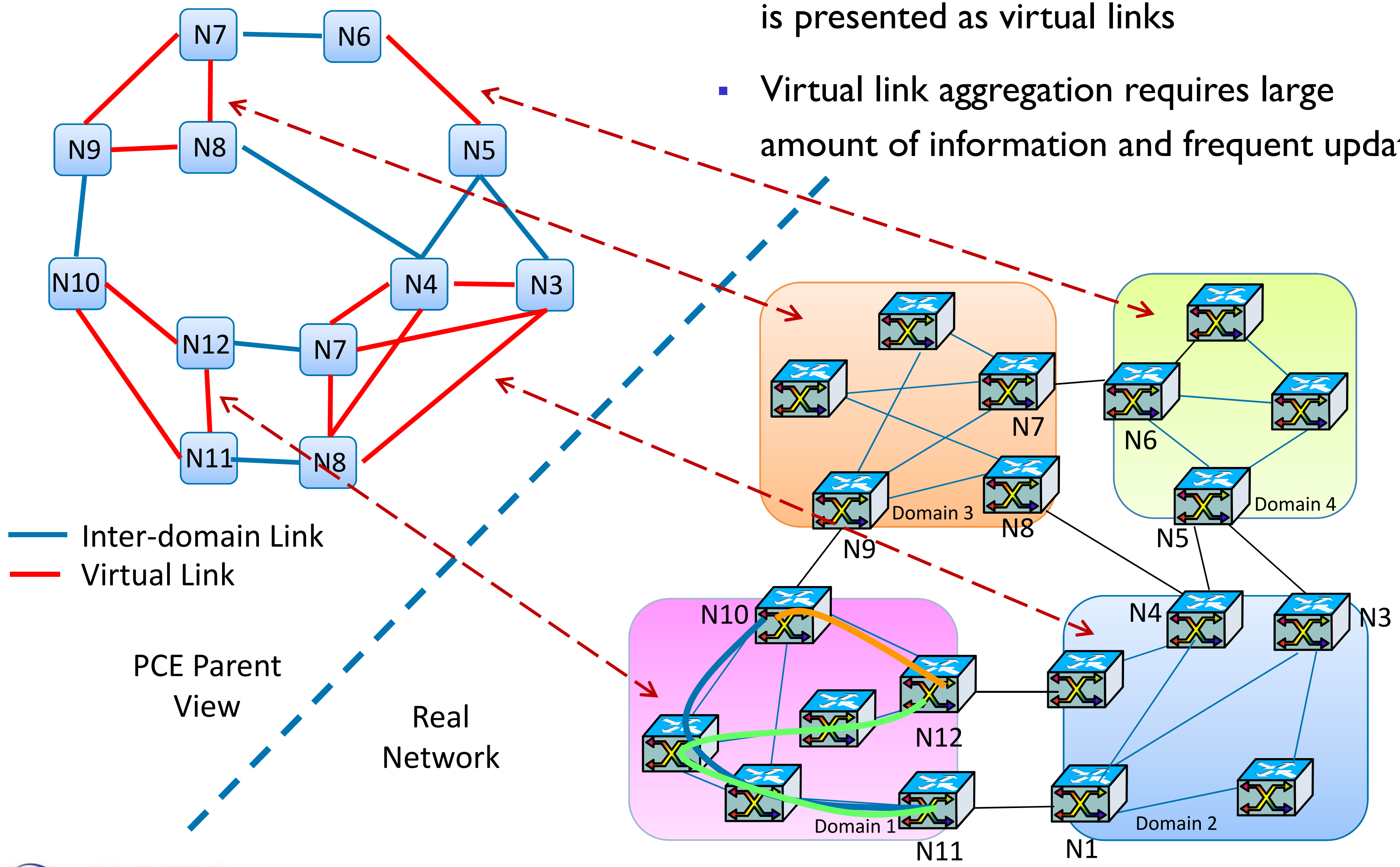
Virtual node aggregation



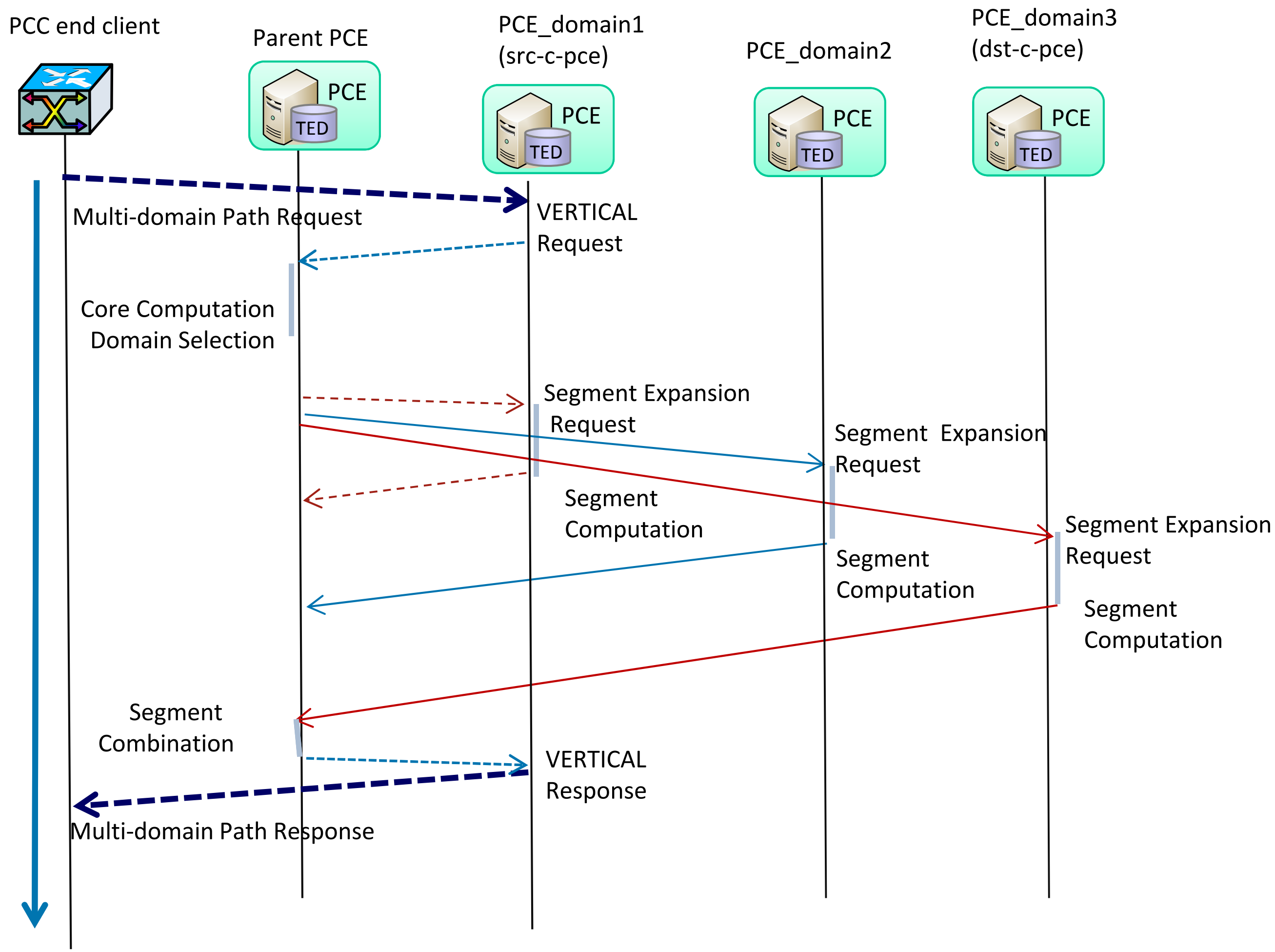
- Each domain is presented as a virtual node
- Virtual node aggregation hides internal connectivity issues

Virtual link aggregation

- Internal connectivity between border nodes is presented as virtual links
- Virtual link aggregation requires large amount of information and frequent updates

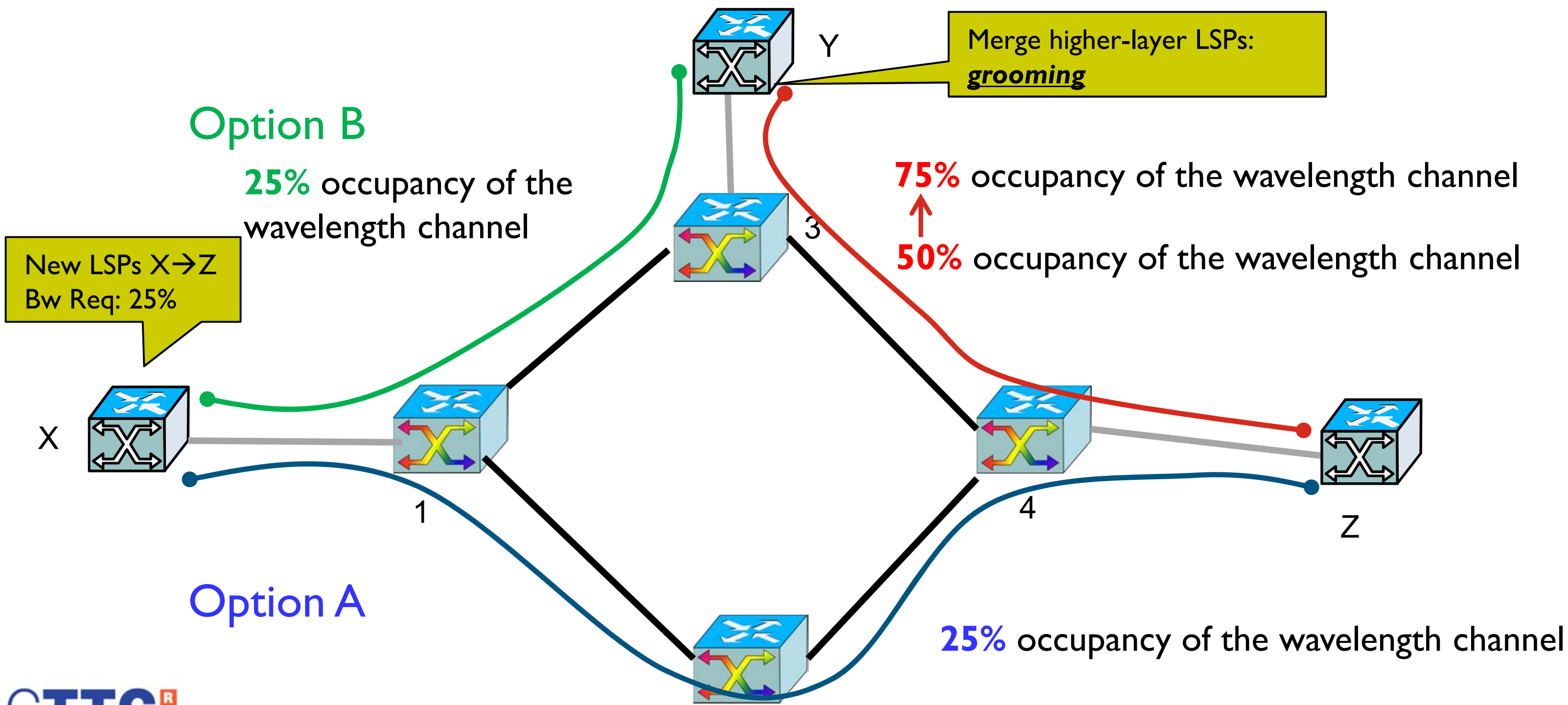


Segment expansion: Parallelized requests



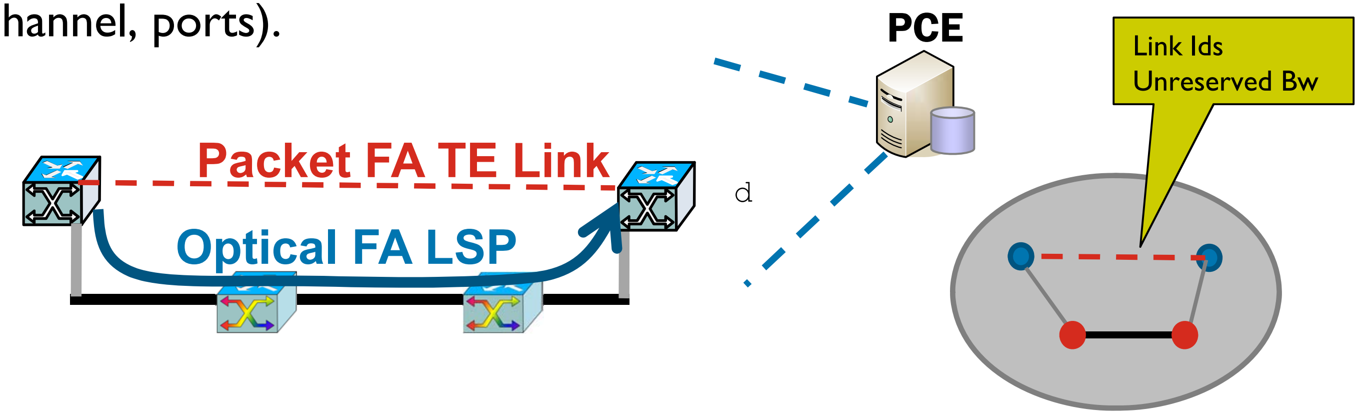
Peer control model with unified control plane

- Peer model: a single and unified control plane instance governs all switching layers
  - The routing protocol disseminates information relative to any switching layer -> a single TED with a global view of network resources and topology.
  - Path computation uses this full network visibility to attain global network resource optimization through applying *Multi-Layer TE strategies (grooming)* .

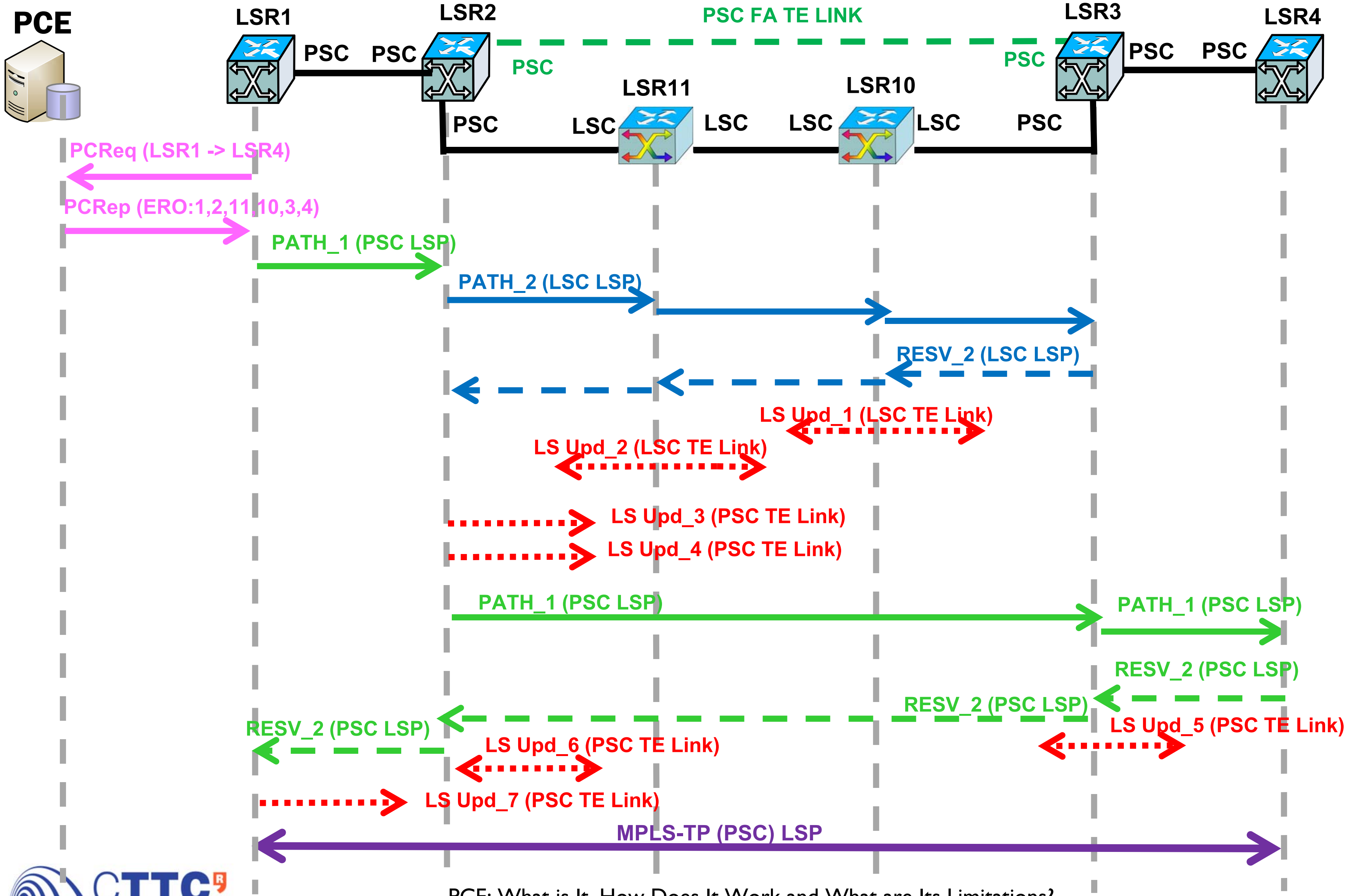


Unified Control Plane: Forwarding Adjacency

- LSPs between layer border nodes can be used as data link in the upper layers -> GMPLS FA TE link concept.
- General use of GMPLS Forwarding Adjacencies (FA):
  1. A lower-layer (e.g., optical) LSP is set up.
  2. OSPF-TE advertises such a lower layer LSP as (*virtual / FA*) TE link in the higher-layer topology -> Virtual Network Topology (VTN).
  3. Subsequent higher-layer (e.g., packet) LSP requests may reuse the remaining available bandwidth in preference to occupy new resources (e.g., WDM channel, ports).

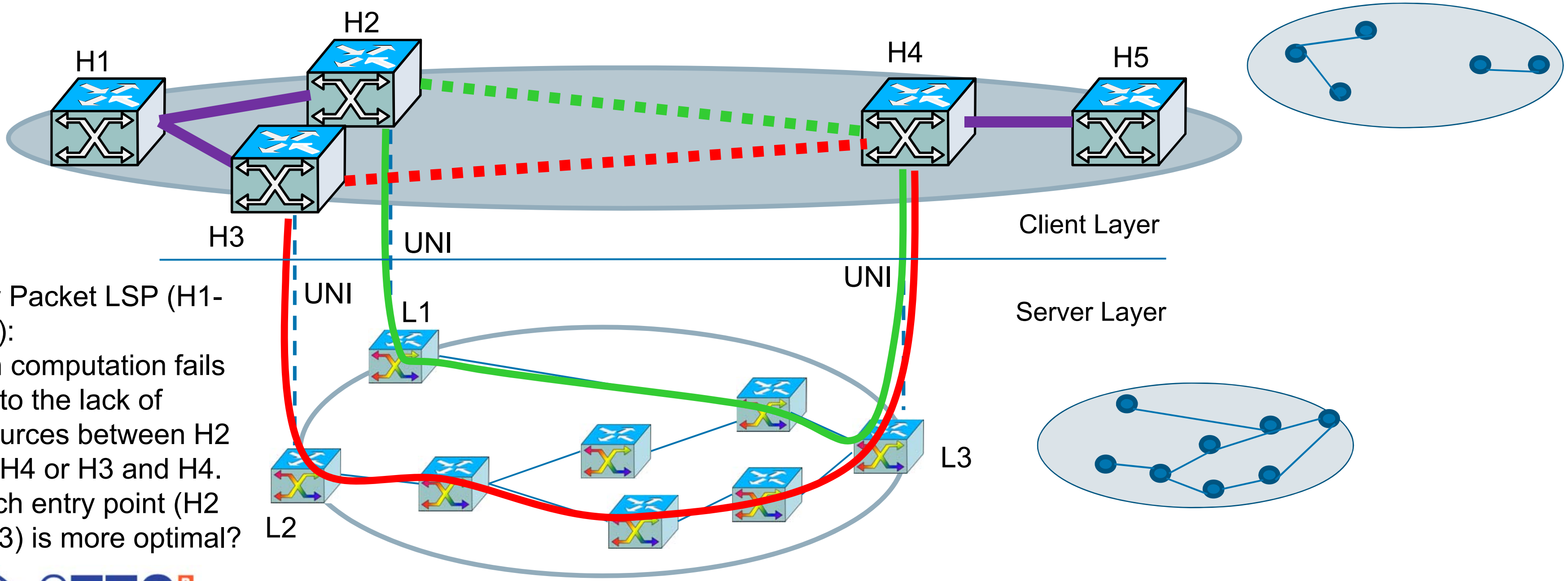


Example of multi-layer path computation and hierarchical signaling



Overlay model with separated control instances

- Overlay model: each switching layer has its own control plane instance
  - No exchange of network topology and resource information between layers
  - Exchange of signaling through the UNI is allowed for provisioning
- The lack of a multi-layer network topology info prevents to globally optimize network resource utilization (end-to-end path computation across multiple layers)
  - Network resource optimization is performed at each layer independently.

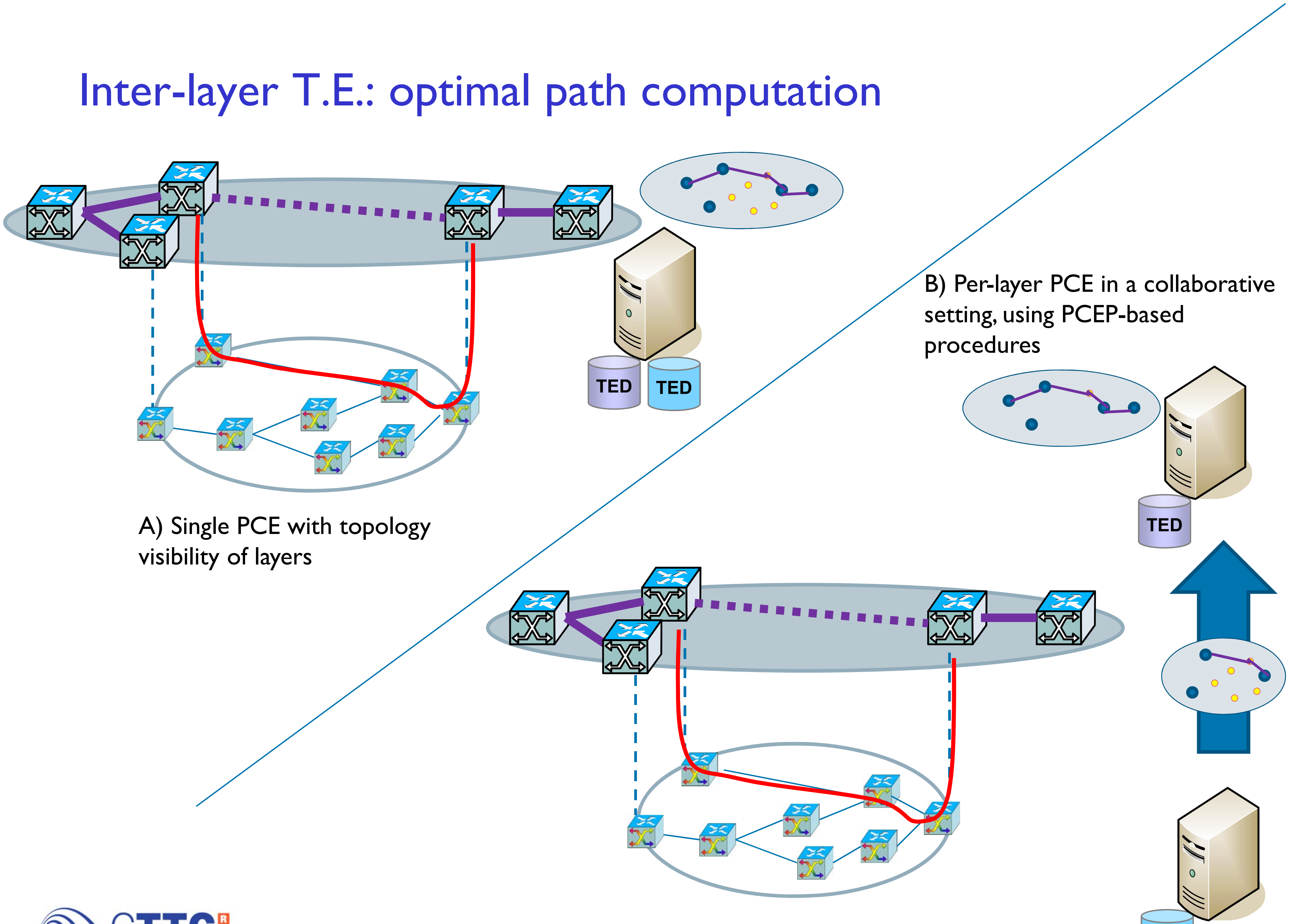


- New Packet LSP (H1->H5):
- Path computation fails due to the lack of resources between H2 and H4 or H3 and H4.
- Which entry point (H2 or H3) is more optimal?

PCE-based multi-layer path computation in an overlay model

- Optimality requires coordinated PCE-based path computation or full topology visibility, by means of:
  - A) Single PCE with topology visibility of layers
    - Either in a single combined TED or in a joint processing of each layer TED
    - Only the PCE knows the full topology, not the routing agents in each layer LSR
  - B) Per-layer PCE in a collaborative setting, using PCEP-based procedures to ensure optimality
    - Each layer PCE knows its layer topology, but needs to ensure that the optimal region boundary nodes are selected
    - Use methods conceptually similar to BRPC in a multi-domain context
    - Use H-PCE based solutions
- Path Optimality → path computation output in the form of a strict ERO object including each layer sub-objects.

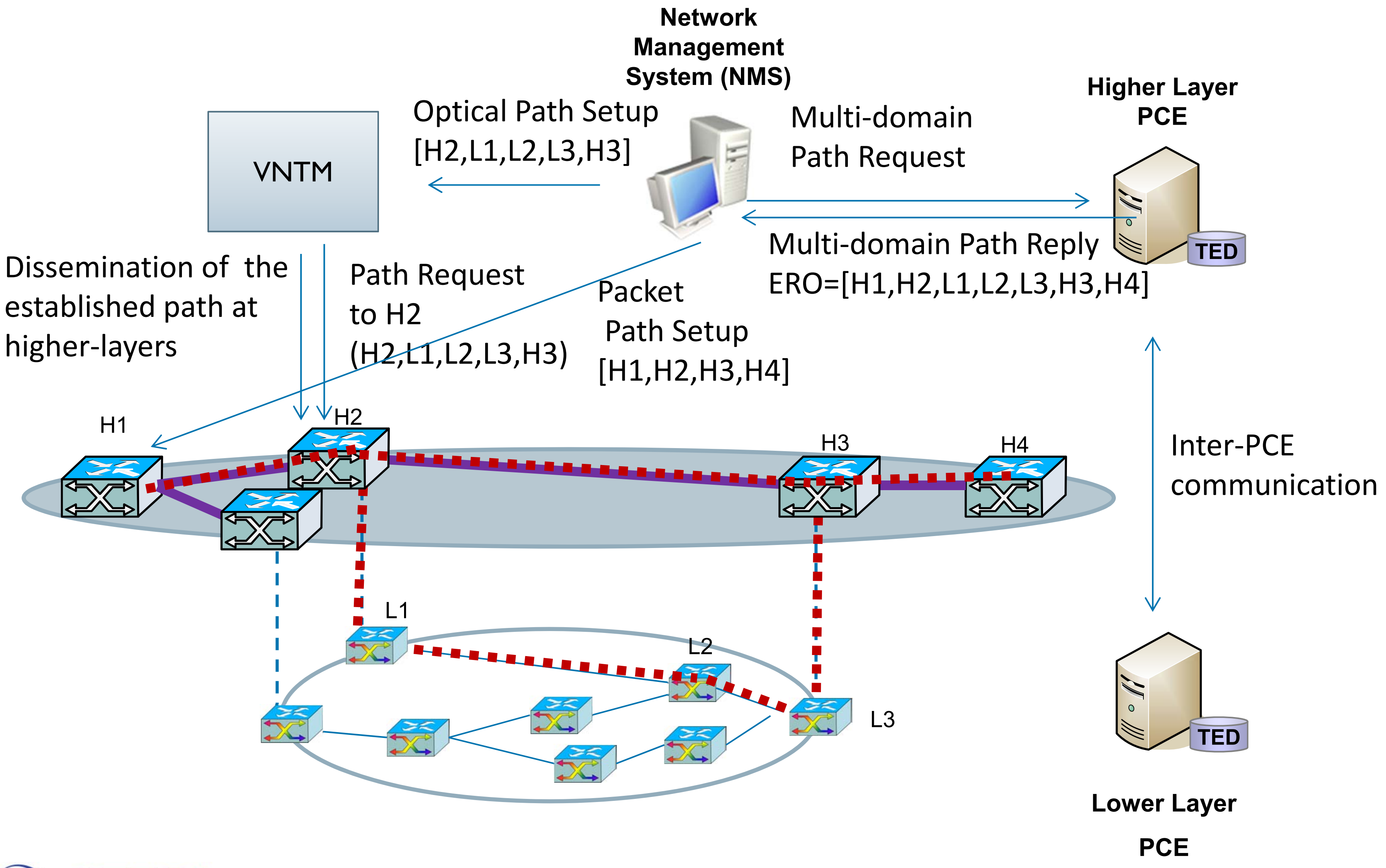
Inter-layer T.E.: optimal path computation



Inter-layer TE: dynamic (automated) provisioning

- Inter-layer automated provisioning depends on the ability to provision all (server/client) layers:
  - Hierarchical signaling
    - The establishment of a client LSP triggers the establishment of a server layer connection at region boundaries
  - Layered provisioning
    - An entity that is able to coordinate the layered establishment of server segments and finally the client layer (VNTM)
- Both approaches are based on the promotion of a server connection as a client layer (logical) TE link:
  - Forwarding Adjacency concept as considered in a unified control plane
- Disseminated by:
  - Server layer connection head-end LSR
  - The VNTM via the server layer connection head-end LSR

Inter-Layer provisioning: Layered Provisioning



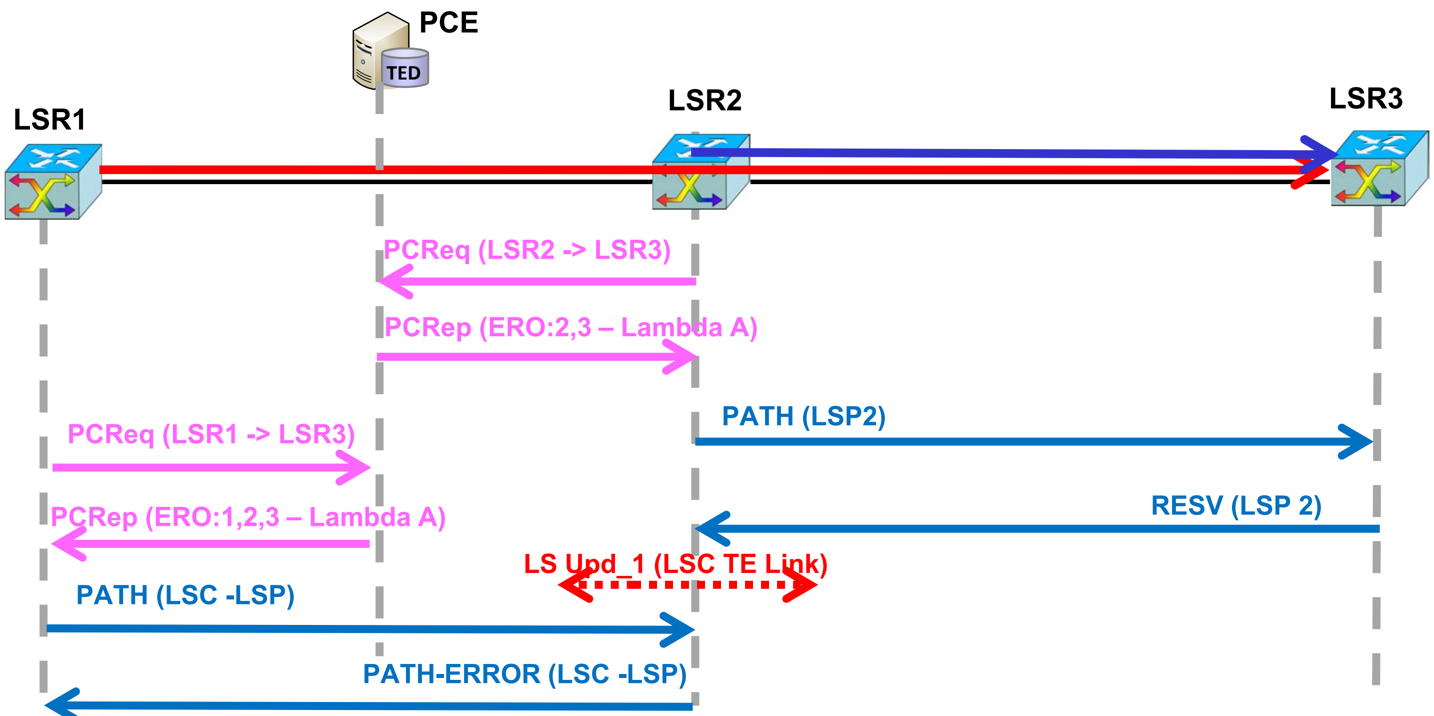
Limitations of PCE

TED synchronization

- A stateless PCE operates with network state information (topology and resource) collected in the TED provided by a synchronization mechanism:
  - Initial synchronization mechanism based on IGP passive listener for intra-domain TED.
  - … but extended methods are needed to discover neighboring domains, border nodes, inter-domain links or peering PCE addresses.
  - New synchronization mechanisms: embedded PCEP notifications, dedicated topology servers or new protocols such as BGP-LS to obtain the TED by BGP peering.
- Not necessary to "remember" computed paths and a request / set of requests is processed independently of each other.
  - Stateless PCE computes paths based on TED information which may not be synchronized with the actual network state, e.g. due to recent PCE-computed paths changes -> Increase the path computation blocking.

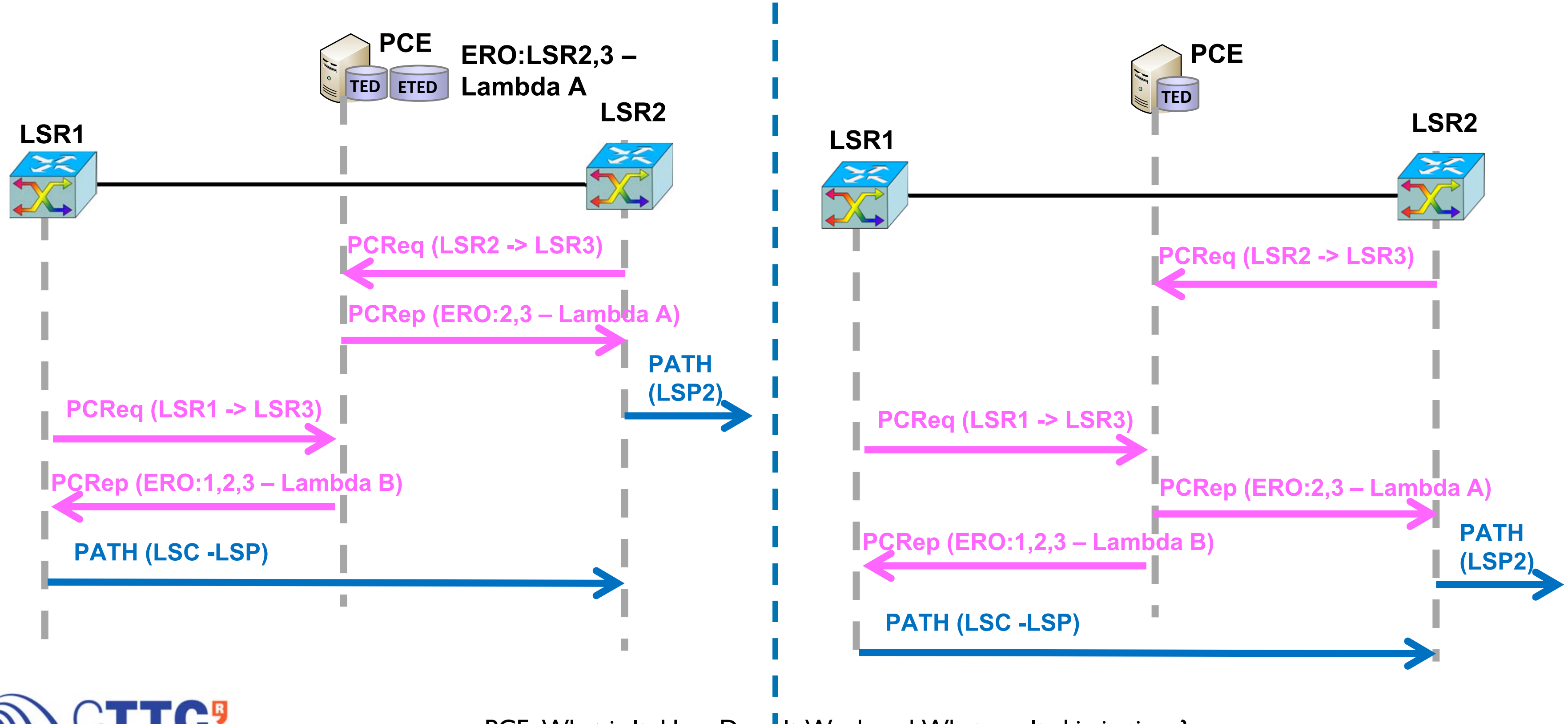
Increase of the path computation blocking

- IGP distribution mechanisms take some time (routing convergence time) to update the TED network state.
- Two requests must be separated more than the OSPF-TE to ensure that the PCE operates with a fully updated TED:
  - The same resources may be assigned to different LSPs -> resource contention.



Minimization of the out-of-sync TED

- The PCE may retain for a limited period of time some information from recently computed paths so that it avoids the use of the same resources for other LSPs.
- The PCE may store for a limited period of time some LSP request, and process concurrently all the received requests when the timer expires.



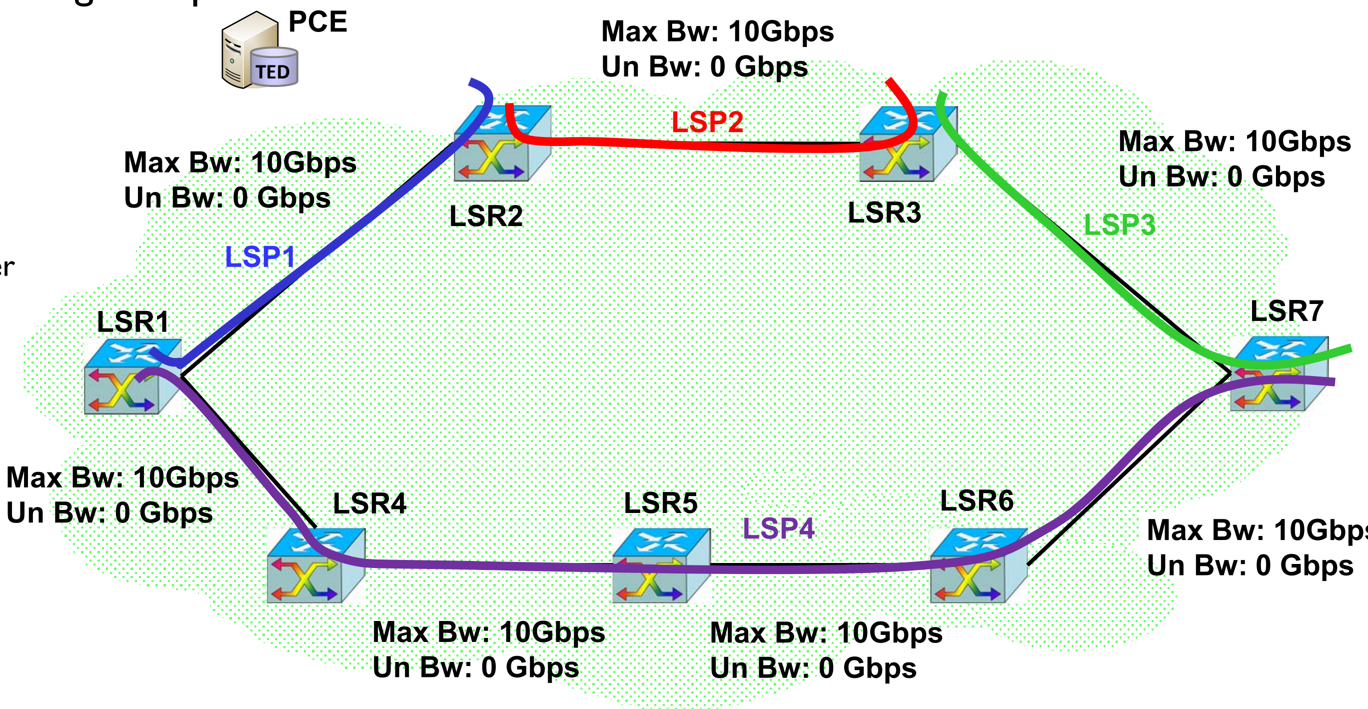
Sub-optimal path computation: lack of global LSP state

- The lack of global LSP state information (e.g., LSP route and reserved resources) may result in sub-optimal PCE algorithms:
  - Minimal perturbation problem → route a demand along the path that requires the lowest number of preemptions. Without knowledge of LSPs, preempting low-priority LSP based on the minimum number of links may not result in the smallest number of LSPs being disrupted.

New High-Priority LSP
LSR1 -> LSR7

- Computed Path without Global LSP state: LSR1,2,3,7
- Minimization of number of links
- 3 LSP must be preempted

- Computed Path with Global LSP state: LSR1,4,5,6,7
- Minimization of number of LSP
- Only 1 LSP must be preempted.



Sub-optimal path computation: PCE control of path reservations

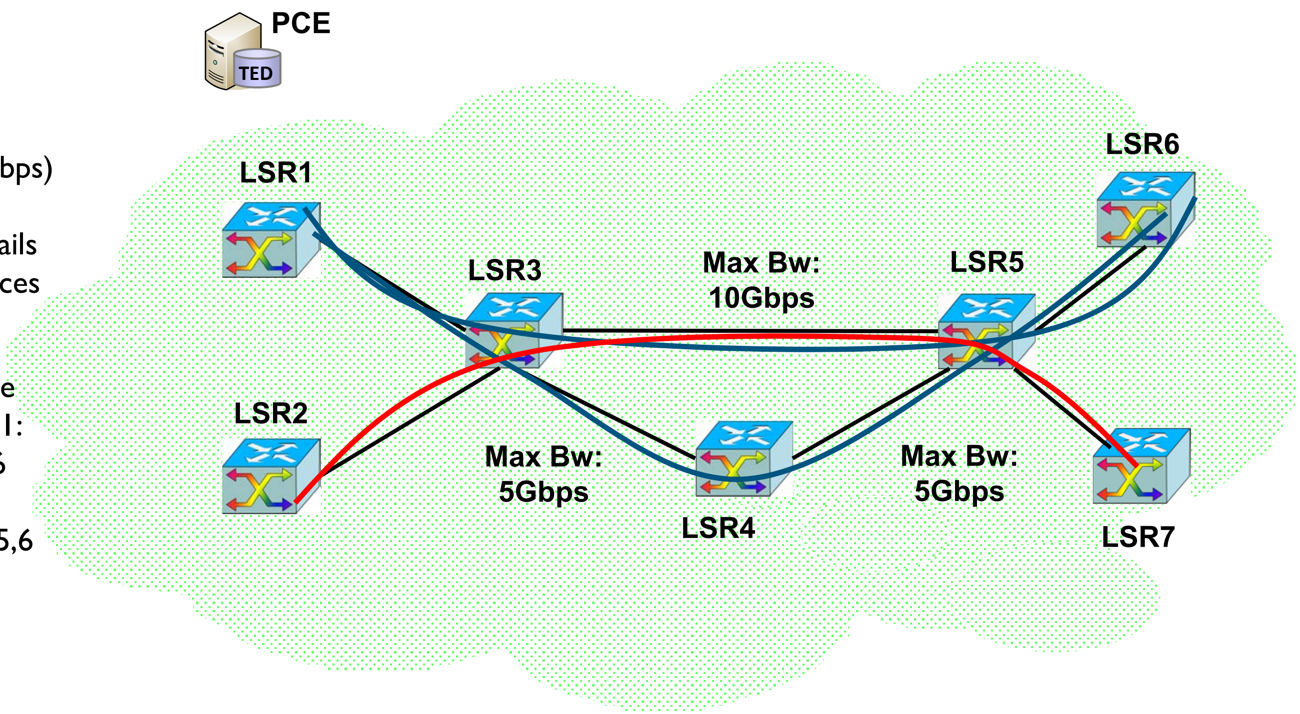
- Also the lack of PCE control of path reservations (sequence and timing of re-optimization, provisioning and release of LSPs) may result in sub-optimal PCE algorithms:

New LSP 1 request (5Gbps)
LSR1 -> LSR6
- ERO: LSR1,3,5,6.

New LSP 2 request (7Gbps)
LSR2 -> LSR7
- Path computation Fails due to lack of resources

LSP2 would not fail if the PCE could re-route LSP1:
- New ERO: LSR1,4,5,6

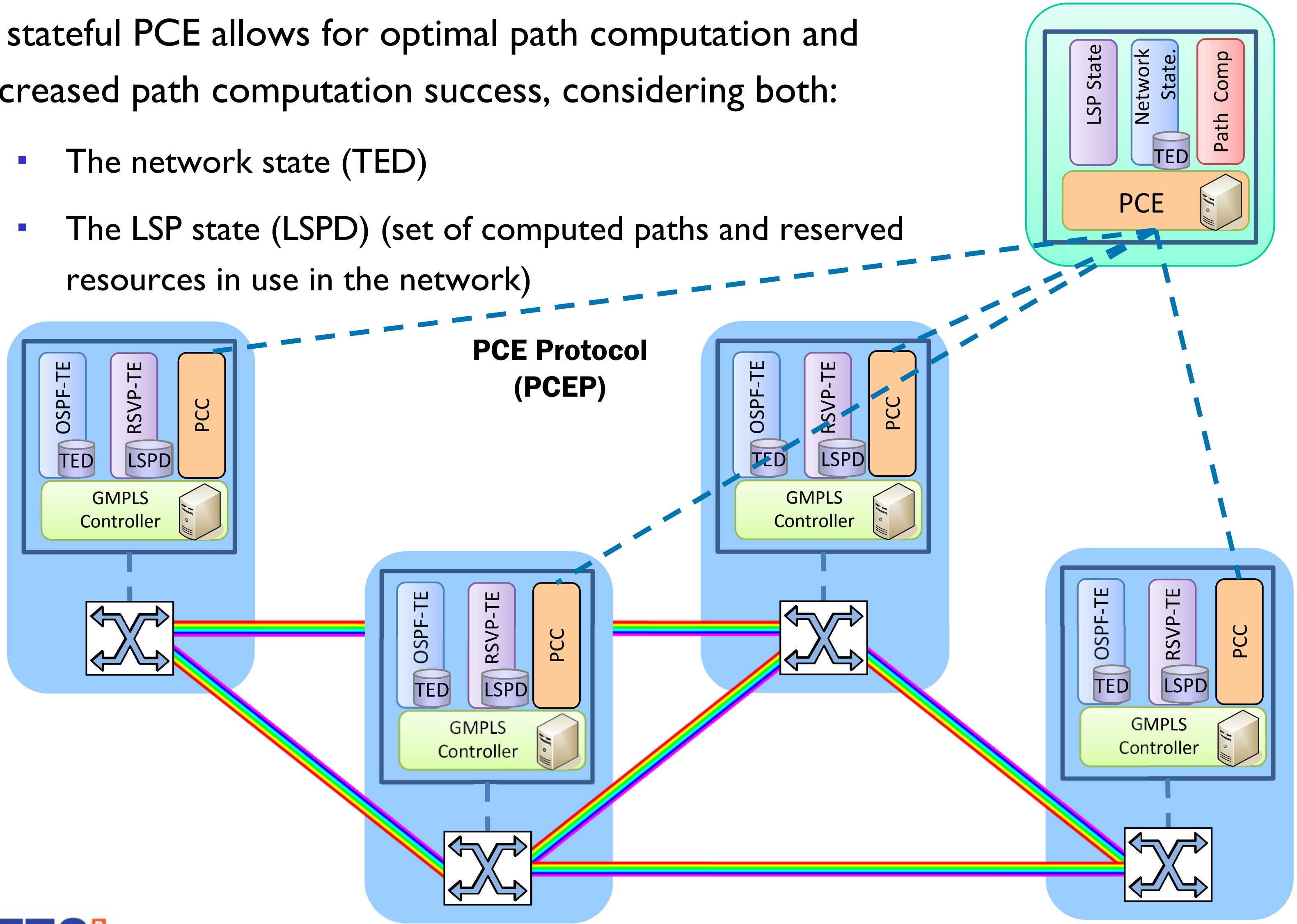
Then, LSP2 could be routed through LSR1,3,5,6



Stateful PCE, applicability to SDN and its limitations

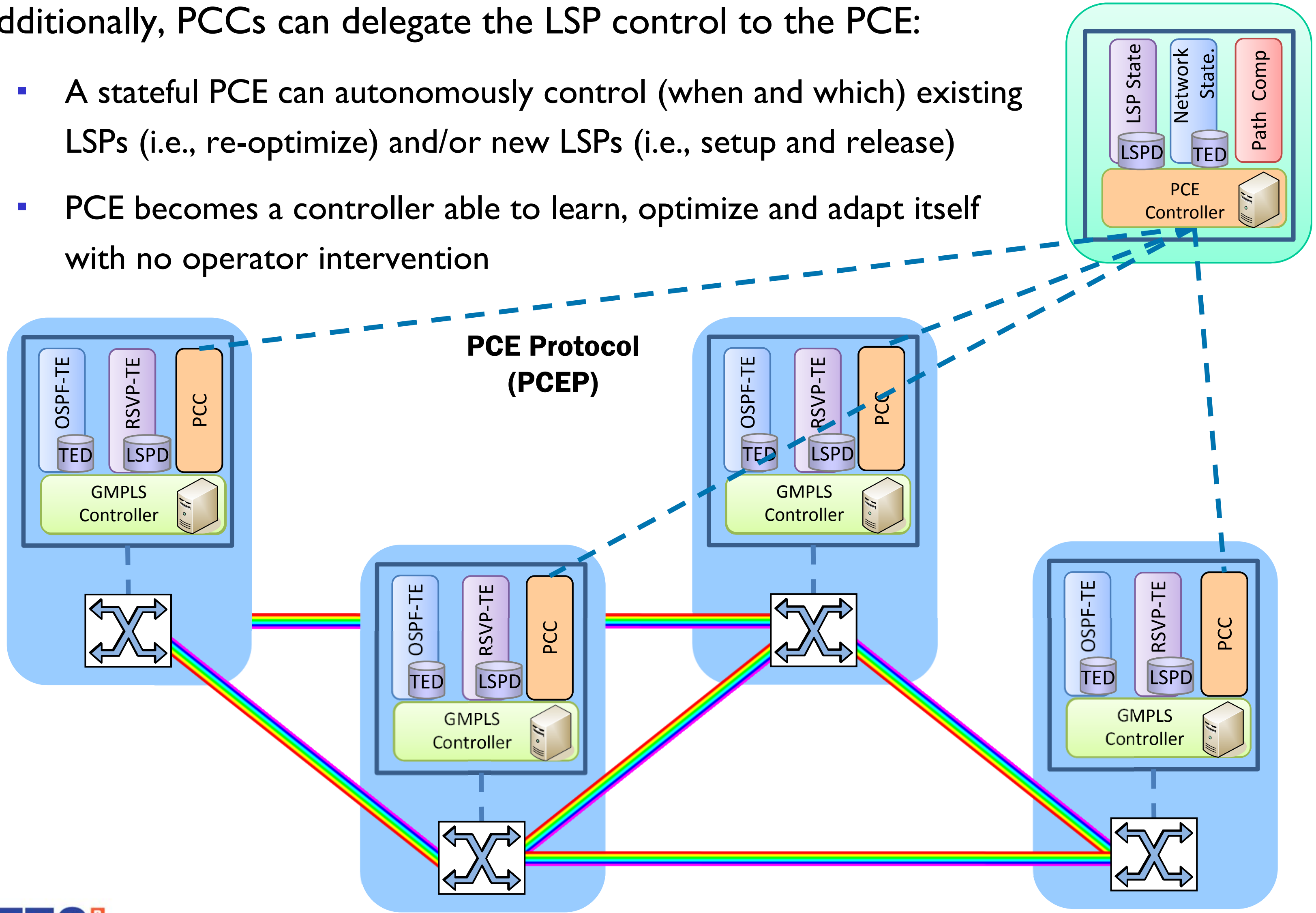
Introduction to Stateful PCE (Passive)

- A stateful PCE allows for optimal path computation and increased path computation success, considering both:
  - The network state (TED)
  - The LSP state (LSPD) (set of computed paths and reserved resources in use in the network)



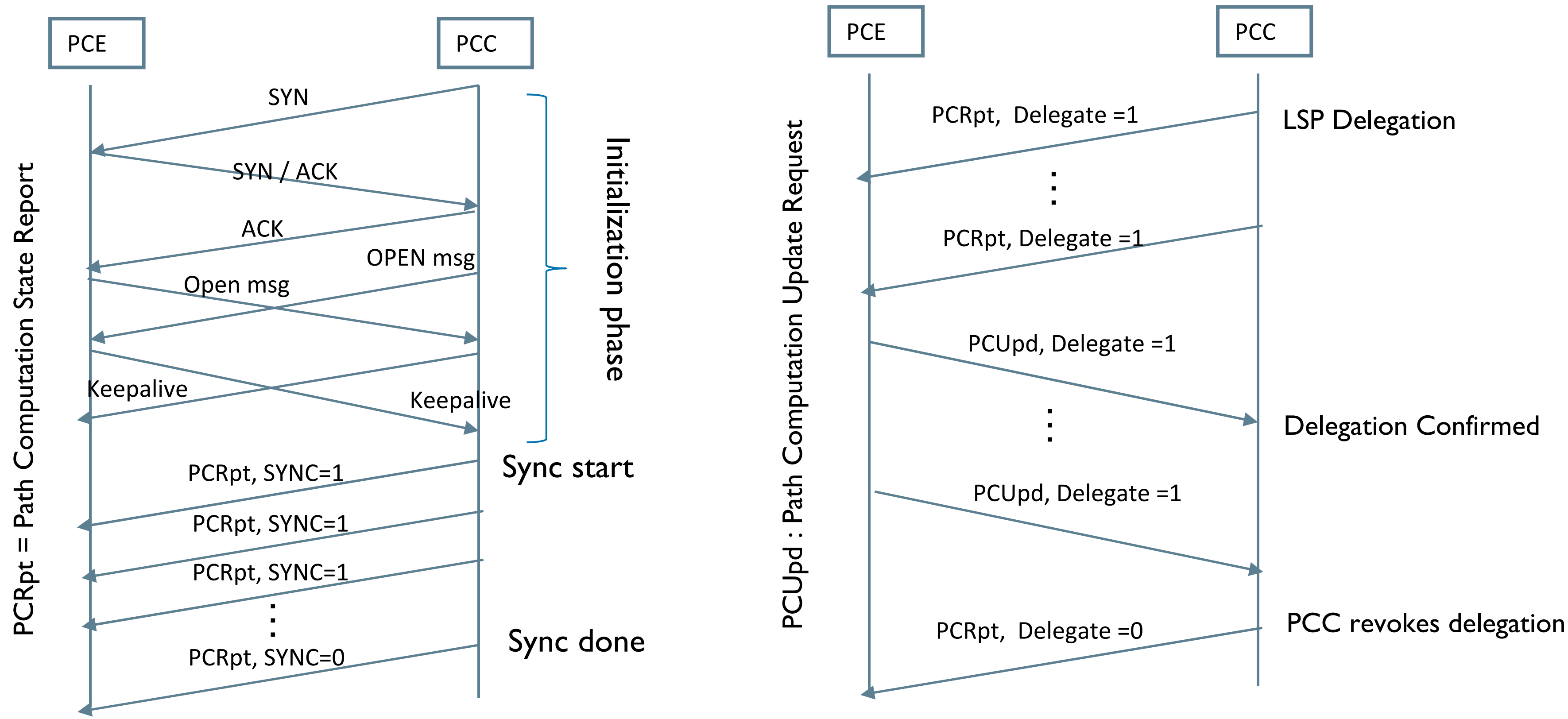
Introduction to Stateful PCE (Active)

- Additionally, PCCs can delegate the LSP control to the PCE:
  - A stateful PCE can autonomously control (when and which) existing LSPs (i.e., re-optimize) and/or new LSPs (i.e., setup and release)
  - PCE becomes a controller able to learn, optimize and adapt itself with no operator intervention



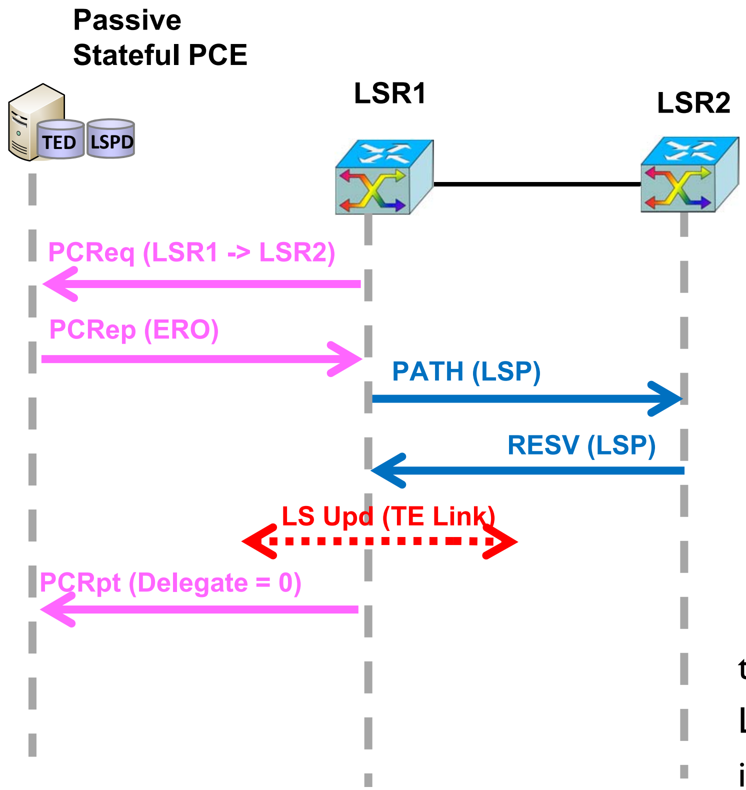
LSP State Synchronization and Delegation

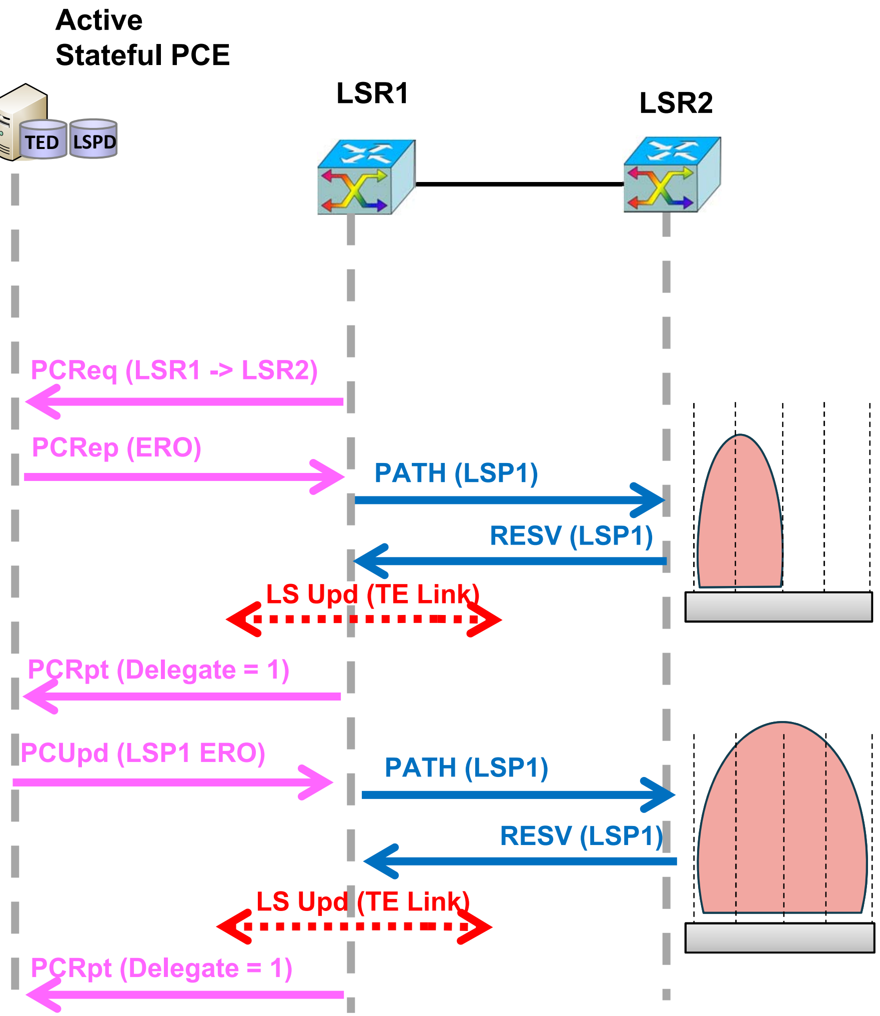
- PCCs are always the owner of LSP state, but:
  - The PCE maintains strict synchronization with PCCs to learn the LSP state (PCEP)
  - PCCs can delegate the control of a set of LSPs to an active stateful PCE



E. Crabbe, PCEP Extensions for Stateful PCE, draft-ietf-pce-stateful-pce-02

Passive vs Active Stateful PCE Path computation request/response





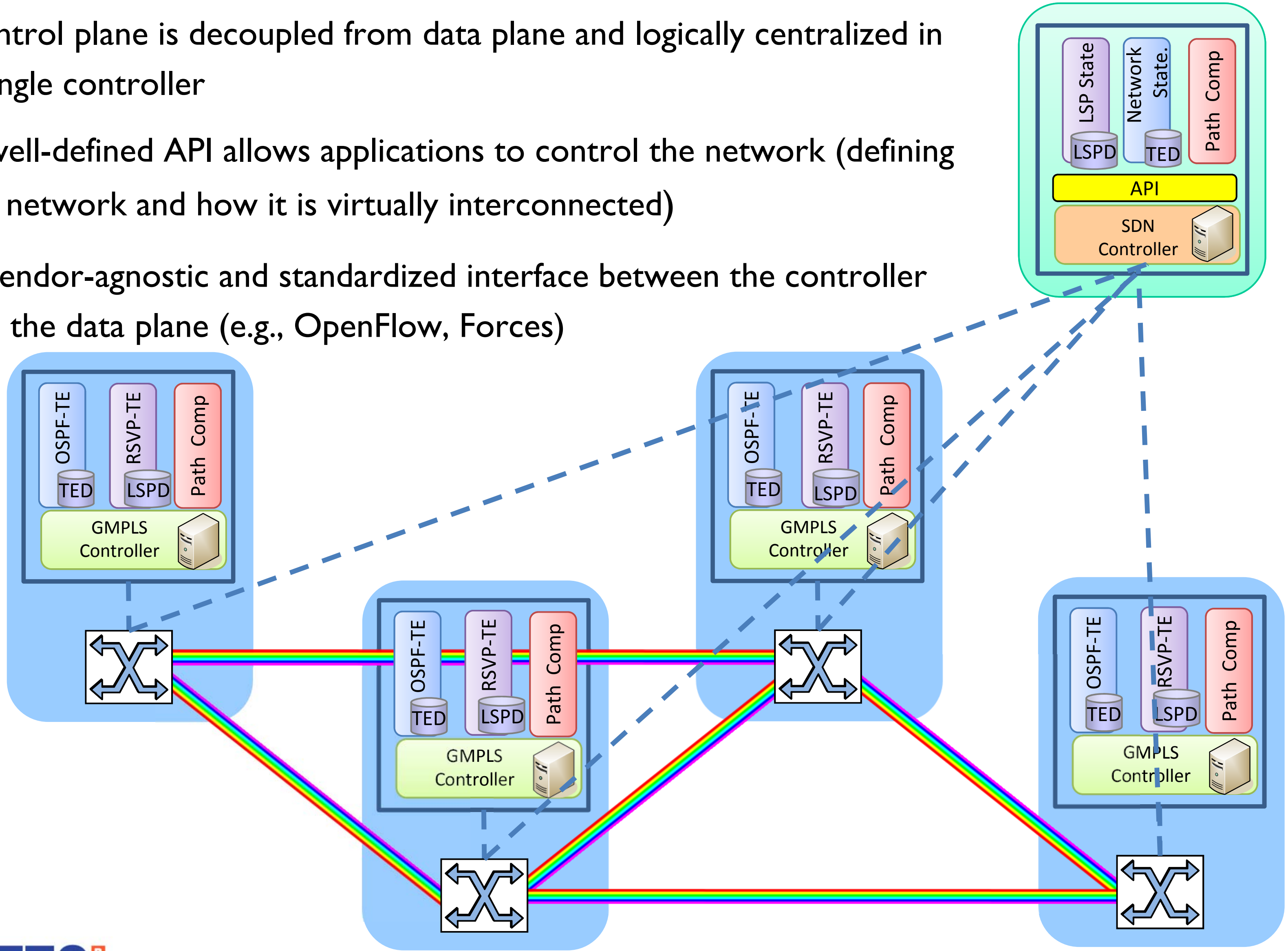
PCE decides to update LSP1 (e.g. to increase the BW)

- PCC decides to create, remove or optimize LSPs

- PCC decides to create LSPs and the PCE can optimize them.
- The PCE could also decide to create and remove LSPs

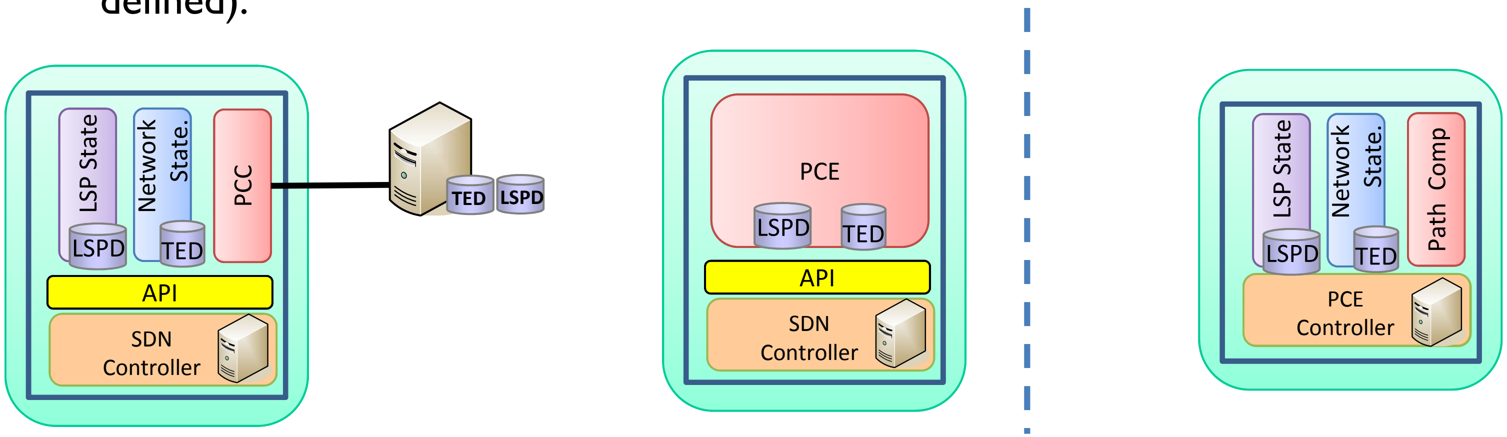
Introduction to Software Defined Networks

- Control plane is decoupled from data plane and logically centralized in a single controller
- A well-defined API allows applications to control the network (defining the network and how it is virtually interconnected)
- A vendor-agnostic and standardized interface between the controller and the data plane (e.g., OpenFlow, Forces)



Integration with SDN/Openflow

- As an application of the SDN controller:
  - A) The PCE is formally separated from the SDN controller, a PCC is an application on top of the controller that requests path computation from the PCE.
    - The TED may be obtained from a topology server, or requested back to the SDN controller.
  - B) The PCE is directly an application on top of the SDN/Openflow controller
    - Better integration with TED and LSPD. Requires fully-featured / complete API.
- As a SDN controller: active-stateful PCE.
  - However, a stateful PCE does not interface with applications and services (not defined).



Limitations of stateful PCE

- In large networks the maintenance of a LSP database can be non-trivial and may require substantial control plane resources:
  - Reliable synchronization mechanism
- If there is a single PCE per domain, all path computations are done by the PCE; the PCE remains synchronized, but:
  - In multi-domain networks with several PCEs, the path computation and LSP state information are distributed among PCEs
    - PCEs would require to synchronize the LSP database by communicating with each other, as done with the TED.
    - PCEs would require to coordinate the path computation by communicating with each other
- Path computations considering both TED and LSP databases would be highly complex

Conclusions

Conclusions

- We overviewed the PCE architecture and how it can mitigate some weaknesses of GMPLS-controlled WSON/SSON
- We have identified some of its own limitations and the way they are being addressed, along with the advanced deployments in SDN/Openflow
- Summary of main trends:
  - PCEs are being integrated within other control paradigms outside their original scope (MPLS/GMPLS):
    - Coupled with NMS or SDN
  - PCE architecture is moving from stateless to stateful
  - Active stateful PCEs are able to autonomously decide where and when to setup, re-optimize and release data connections:
    - Behave similarly to a SDN controller, i.e., capable to learn, optimize and adapt themselves in cognitive networks.
  - Stateful PCE is a very recent topic requiring significant research effort

http://wikiona.cttc.es

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement n° 317999 (IDEALIST project) and MINECO (Spanish Ministry of Economy and Competitiveness) through the project FARO (TEC2012-38119)

CTTC
Centre Tecnològic de Telecomunicacions de Catalunya

AENOR
Gestión I+D+i
UNE 166002
IDI-0024/2011