Path Computation Element (PCE)

Adrian Farrel
Old Dog Consulting
adrian@olddog.co.uk

www.mpls2008.com

MPLS 2008

Old Dog Consulting

Agenda

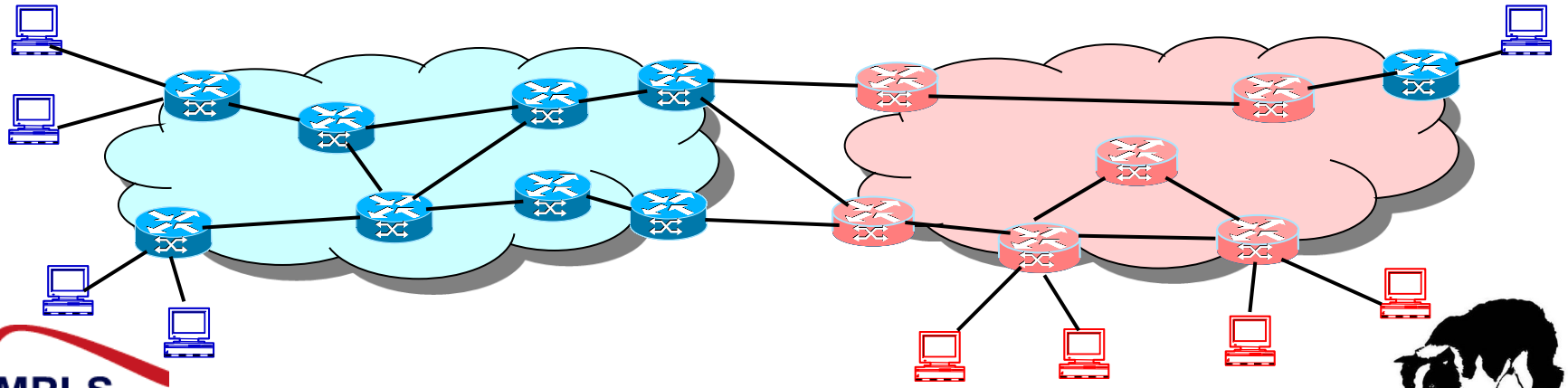
- Historic Drivers
- Generic Requirements
- Architectural Overview
- Discovering PCEs
- PCEP - The Basics of the PCE Protocol
- Usage Scenarios
- Core Protocol Extensions
- Advanced Uses and The Future

Background – MPLS Traffic Engineering

- Objectives are to improve network efficiency, increase traffic performance, reduce costs, and increase profitability
- Adaptive to network changes
- Increasingly achieved through MPLS
- As easy to get wrong as to get right!
- Requires
  - Knowledge of available network resources
  - Understanding of service requirements
  - Planning (computation) of LSP placement
  - Control of provisioning and resource reservation

Historic Drivers

- Virtual PoP
  - Need an MPLS tunnel across a foreign network
  - Guaranteed QoS etc.
- Source domain must decide the correct peering point
  - Should ideally be able to request the LSP "on-demand"



Definition – The Domain

- A domain is defined as
  Any collection of network elements within a common sphere of address management or path computational responsibility
  (RFC 4726 and RFC 4655)
- Classic examples…
  - IGP Areas
  - Autonomous Systems
- More complex examples…
  - Network technology layers
  - Client/server networks
  - Protection domains
  - ITU-T sub-networks
- For us, the problem is the path computational responsibility
  - We need to plan (compute) an end-to-end path
  - But we can only see our domain

Historic Operation – Path Computation

- Path computation limited to within a domain
  - Responsibility of a management/planning station
    - Provisioning based on pre-computed paths
    - Provisioning through management plane or control plane
  - Delegated to an "intelligent control plane"
    - Computation on the head-end LSR
- Domain interconnects by prior arrangement
  - Good for policy and administrative control
  - Bad for responsiveness and dynamic use of resources
  - Not flexible to changes in the network
  - High operational overhead

The Problem of Multi-Domain Path Computation

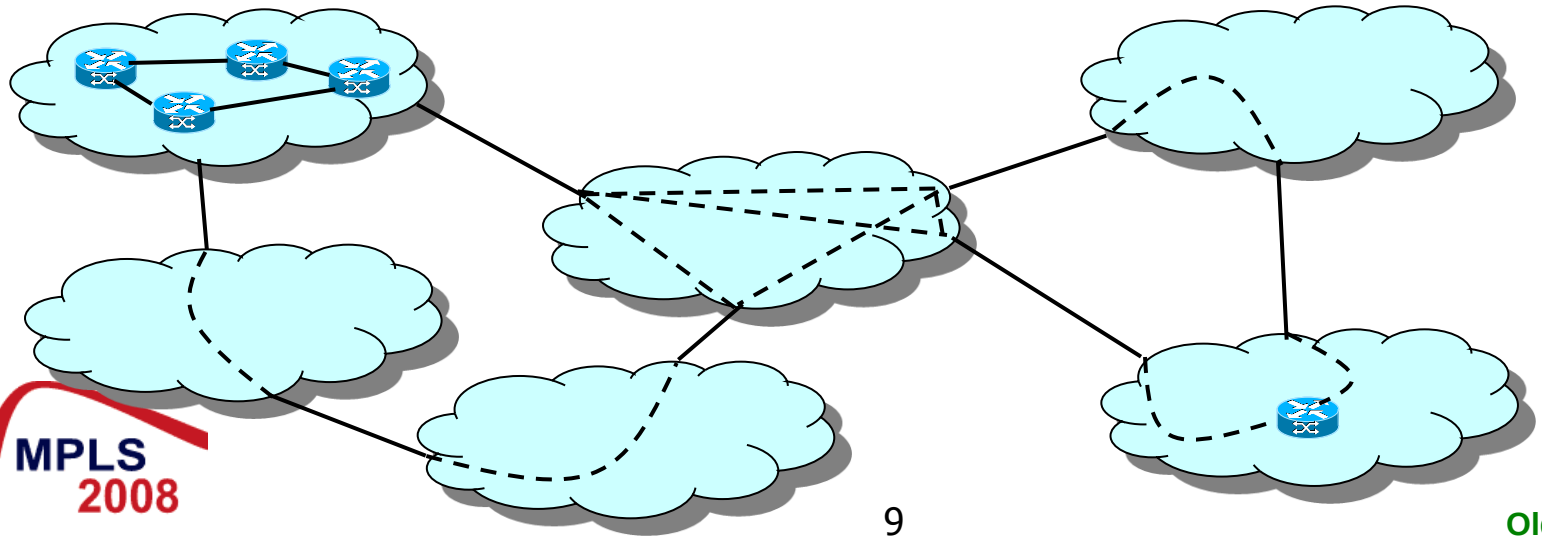
- The Internet is built from administrative domains
  - Scaling reasons
  - Administrative and commercial reasons
- These are IGP areas and Autonomous Systems
- Routing information is not distributed between domains
  - To do so would break
    - Scaling
    - Commercial confidentiality
- Distribution of TE information follows the same rules
  - See RFC 4105 Requirements for Support of Inter-Area and Inter-AS MPLS-TE
  - See RFC 4216 MPLS Inter-AS Traffic Engineering Requirements"
- But, to compute a path we need to be able to see the available links along the whole path

Issues for Routing in Multi-Domain Networks

- The lack of full topology and TE information
- No single node has the full visibility to determine an optimal or even feasible end-to-end path
- How to select the exit point and next domain boundary from a domain
- How can a head-end determine which domains should be used for the end-to-end path?
- Information exchange across multiple domains is limited due to the lack of trust relationship, security issues, or scalability issues even if there is trust relationship between domains

TE Abstraction/Aggregation - A Potential Solution

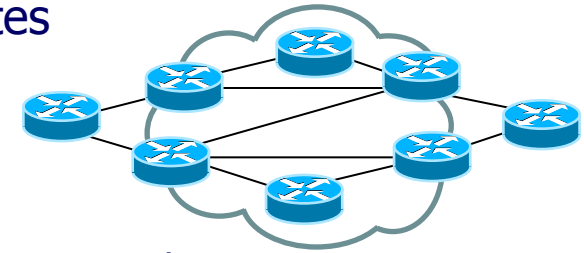
- All we need to know is
  - Details of local domain
  - The connectivity between domains
  - The destination domain to reach
- TE aggregation looks very promising
  - Provide enough information to compute, but still scale
  - But aggregation reduces available information so optimality is in doubt



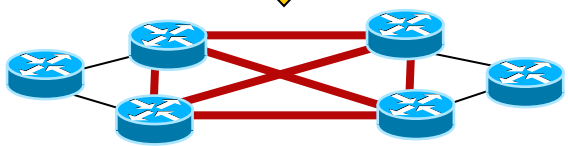
Approaches to TE Aggregation

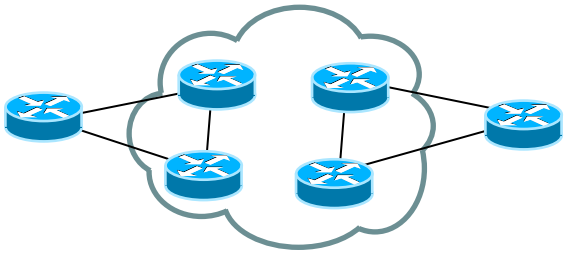
Virtual Link
- "You can reach this destination across this domain with these characteristics"
- BGP-TE model
- Requires large amount of information
- Needs compromises and frequent updates

Virtual Node
- Hierarchical abstraction
- Presents subnetwork as a virtual switch
- Can be very deceptive
  - No easy way to advertise "limited cross-connect capabilities"

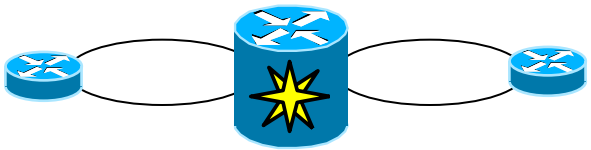


Virtual Link aggregation needs compromises and frequent updates





Virtual Node aggregation hides internal connectivity issues

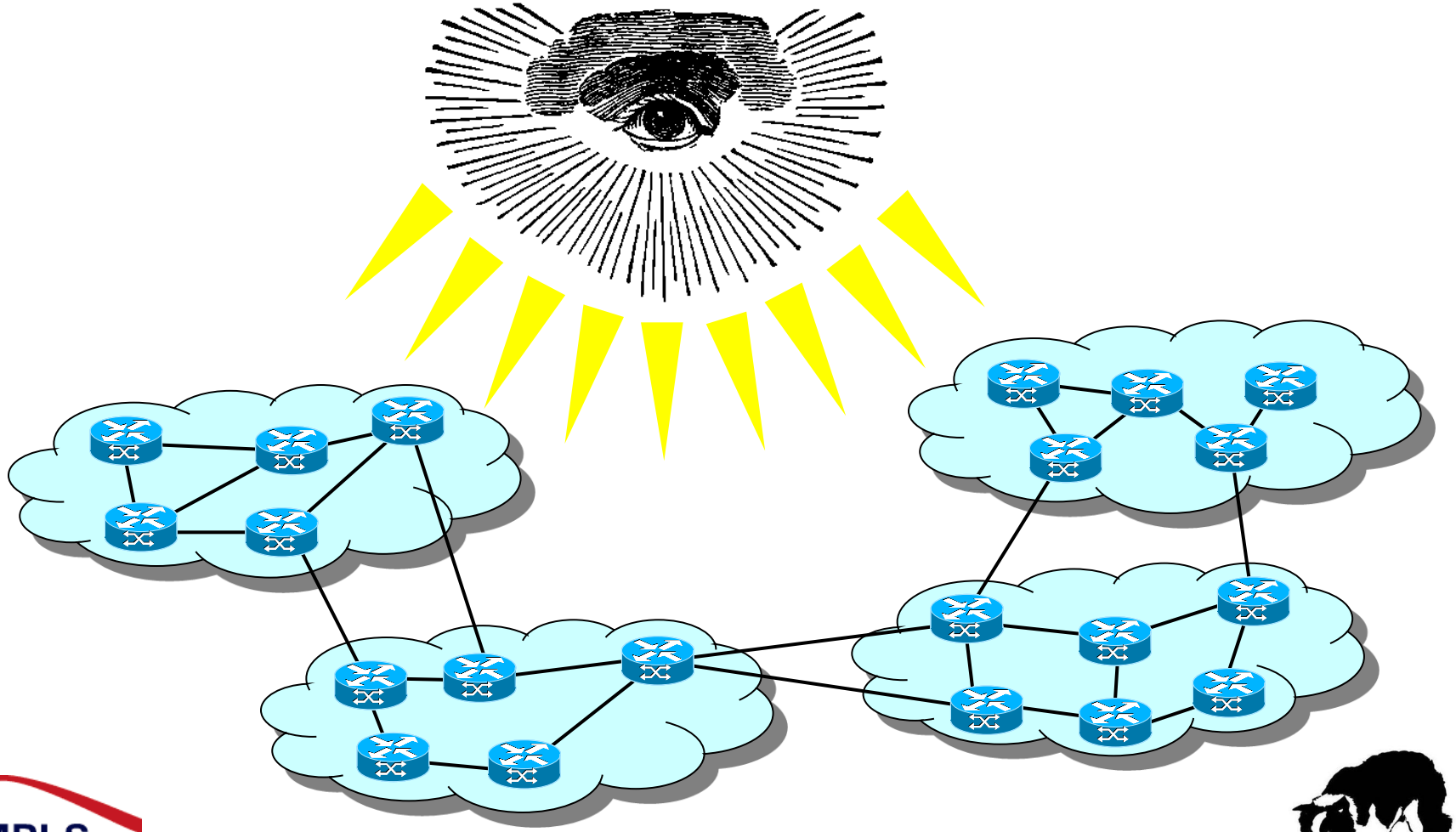


Both rely on crankback signaling and high CPU aggregation

Architectural Concept

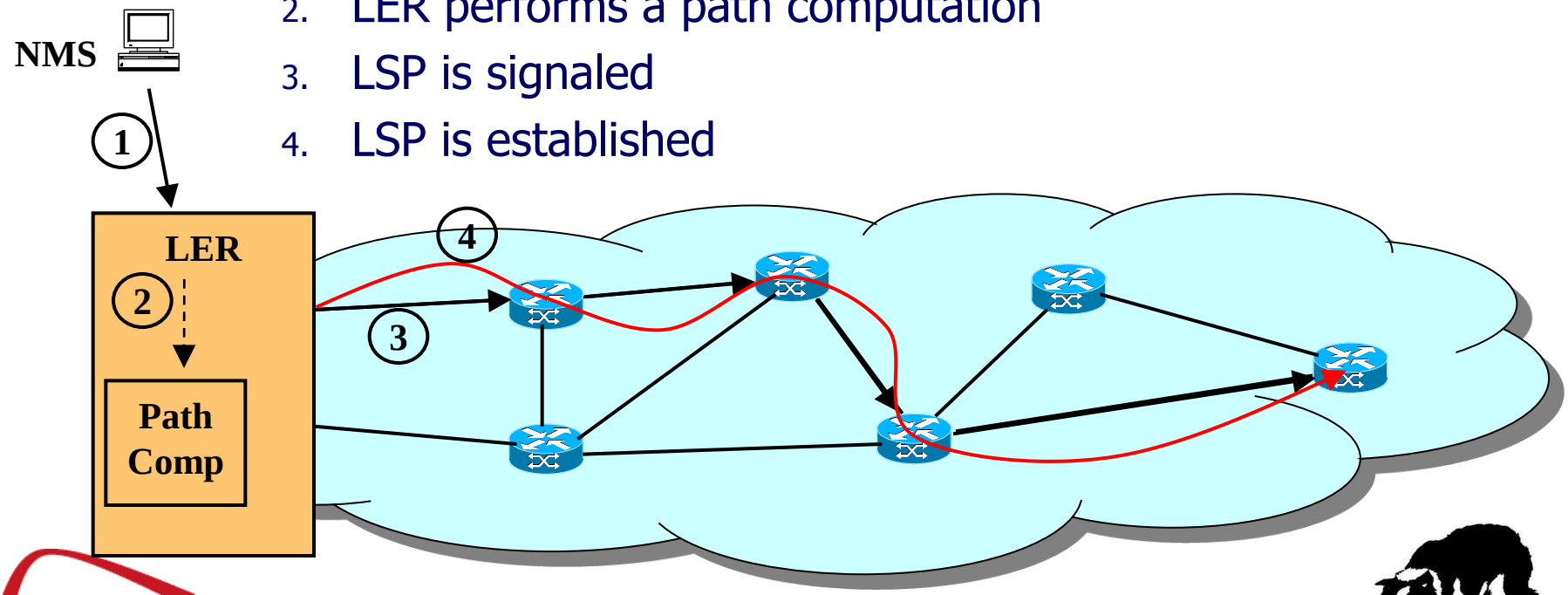
- We need some abstract mechanism to compute paths
  - *"An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints"* (RFC4655)
- PCE is a path computation element (e.g., server) that specializes in complex path computation on behalf of its path computation client (PCC)
- PCEs collect TE information
  - They can "see" within the domain

The All-Seeing Eye – A Myth



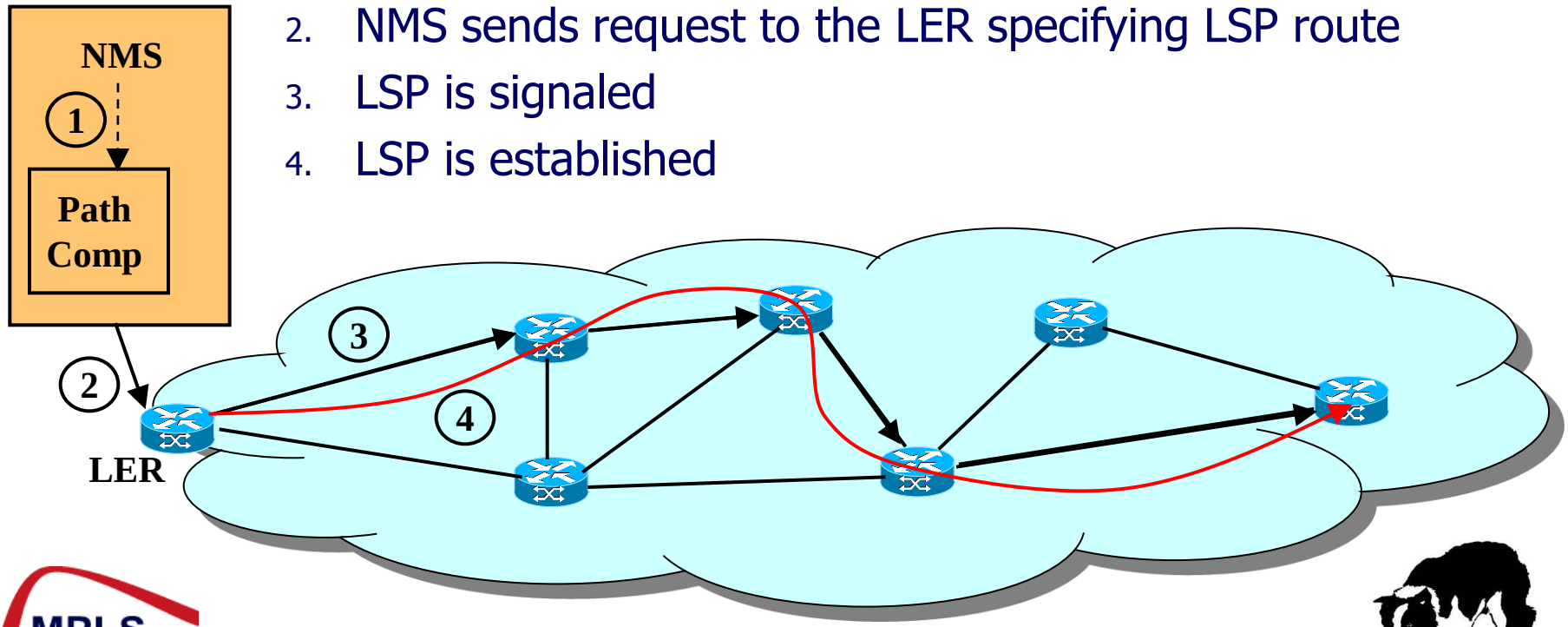
Path Computation – An LER Function

- Path computation is a logical functional component of LERs in existing MPLS-TE deployments
  1. NMS sends request to the LER asking for an LSP
  2. LER performs a path computation
  3. LSP is signaled
  4. LSP is established



Path Computation as an NMS Feature

- Path computation is a logical functional component in many management systems
  1. NMS performs a path computation
  2. NMS sends request to the LER specifying LSP route
  3. LSP is signaled
  4. LSP is established



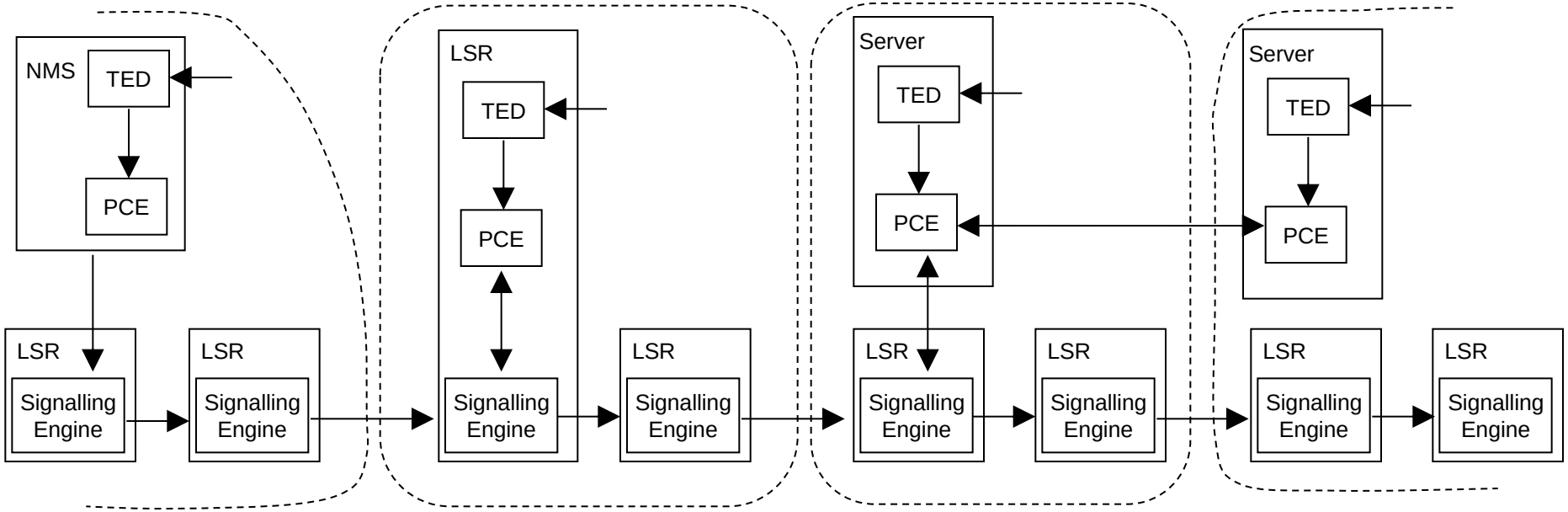
The Traffic Engineering Database

- Path computation requires knowledge of the available network resources
  - Nodes and links
  - Constraints
    - Connectivity
    - Available bandwidth
    - Link costs
- This is the Traffic Engineering Database (TED)
- TED may be built from
  - Information distributed by a routing protocol
    - OSPF-TE and ISIS-TE
  - Information gathered from an inventory management system
  - Information configured directly

The PCE Server and the PCC

- Embedded path computation capabilities
  - Part of the functional model
  - Not very exciting for building networks!
- Path Computation Element (PCE)
  - The remote component that provides path computation
  - May be located in an LSR, NMS, or dedicated server
- Path Computation Client (PCC)
  - The network element that requests computation services
    - Typically an LSR
    - Any network element including NMS

Abstracting The Path Computation Function



"An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints" - RFC 4655

- What's new?
  - Nothing!
  - A formalisation of the functional architecture
  - The ability to perform path computation as a (remote) service

PCC-PCE Communications

- Fundamental to a remote PCE is PCC-PCE communication
- PCC requests a computation
  - From where to where?
  - What type of path? (Constraints)
    - Bandwidth requirement
    - Cost limits, etc.
    - Diversity requirements
- PCE responds with a path (or failure)
  - Details of route of path
  - Details of parameters of path
    - Actual cost, bandwidth, etc.

Multi-Domain PCE

- A single PCE cannot compute a multi-domain path
  - By definition, a PCE can only see inside its domain
- Computation of a multi-domain path *may* use cooperating PCEs
  - PCEs may need to communicate
    - One PCE may send a path computation request to another PCE
    - The first PCE acts as a PCC and the communication is exactly as already described
- Recall: multi-domain path computation is what we are doing this for

Discovering PCEs

- Each PCC needs to know about a PCE
- Maybe more than one PCE
  - Load sharing
  - Different capabilities
    - Support for different constraints
    - Different algorithms
    - Path diversity
- Configuration is an option
  - Management overhead
  - Not flexible to change
- Discovery is the best mechanism
  - Achieved with extensions to the IGP routing protocols

Protocol Extensions

- PCE is probably already participating in the IGP
  - The PCE may be a router (for example, ABR or ASBR)
  - The PCE needs to build the TED
- Advertisement of "optional router capabilities"
  - RFC 4970 for OSPF
    - The Router Information LSA
  - RFC 4971 for IS-IS
    - The Capability TLV
- Define TLVs to carry PCE capabilities
  - RFC 5088 for OSPF
  - RFC 5089 for IS-IS
- TLVs defined for:
  - The IP address of the PCE
  - The domain scope that the PCE can act on
  - The domain(s) in which the PCE can compute paths
  - Neighboring domains toward which the PEC can compute paths
  - Capability flags

Future Discovery Protocol Extensions

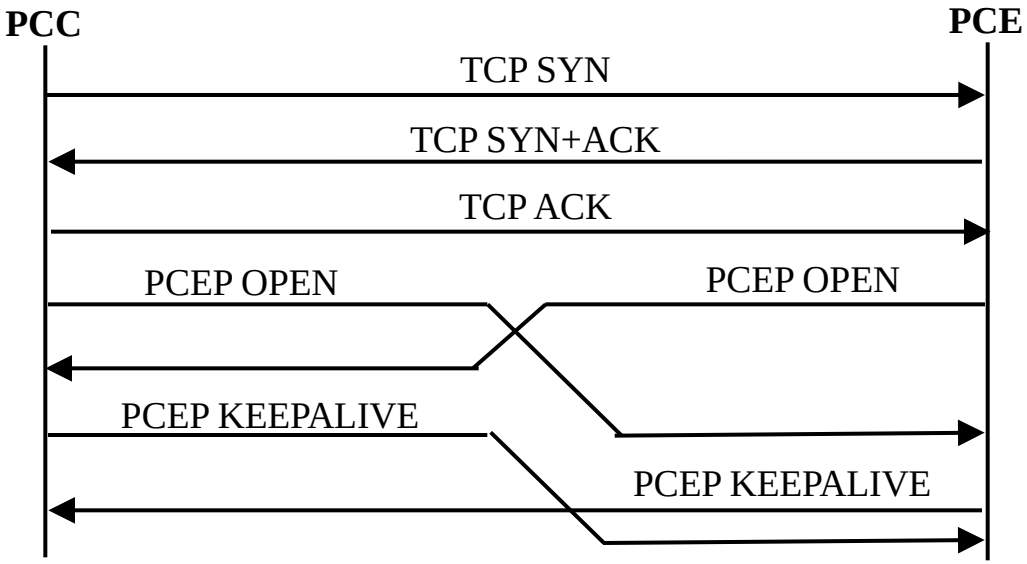
- The Router Information LSA and Capabilities TLV are overloaded
  - They are used for different applications
- *Future* PCE discovery information must be carried in some other way
  - Define a PCE LSA and a PCE TLV
    - Will cause some migration issues
- Exception is capabilities flags that an continue to be used up

PCEP - The Basics of the PCE Protocol

- A request/response protocols
- Operates over TCP
  - Reliability and in-order delivery
  - Security delegated to TCP security issues
- Session-based protocol
  - PCE and PCC open a session
    - Negotiate parameters and learn capabilities
  - All message exchanges within the scope of the session
- Seven messages
  - Open
  - Keepalive
  - Request
  - Response
  - Notify
  - Error
  - Close

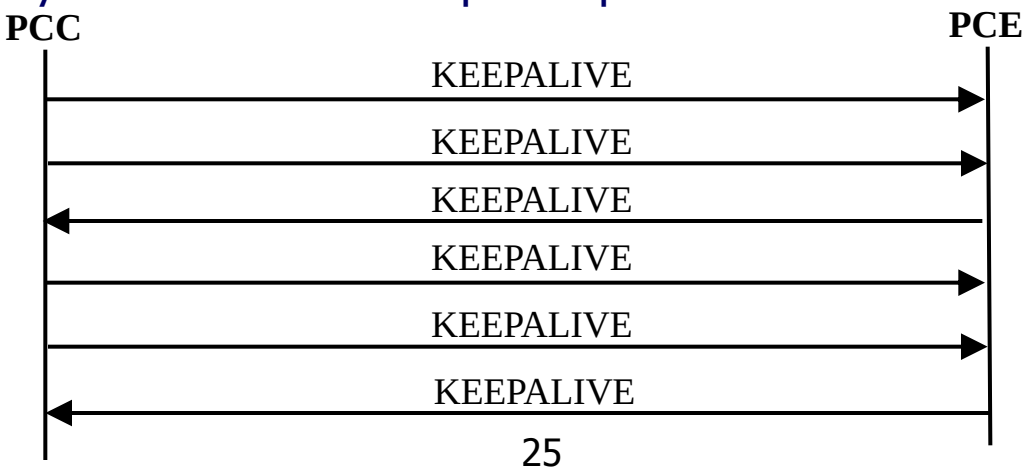
Session Creation

- TCP registered port
  - One connection between any pair of addresses
- Independent two-way exchange of PCEP Open messages
  - Negotiate session capabilities and parameters
  - Accepted with Keepalive message
  - Rejected (for negotiation) with Error message



Session Maintenance

- TCP is not so good at detecting connection failures
  - Connection failure breaks the PCEP session
  - Means that outstanding requests will not get responses
- Many protocols run their own keepalive mechanisms
- The PCEP keepalive process is asymmetrical
  - The Keepalive message is a beacon
  - It is not responded
  - The frequency is set by the receiver on the Open message
  - The session has failed if no Keepalive is received in the Dead Timer period
    - Usually four times the keepalive period



Request / Response Information

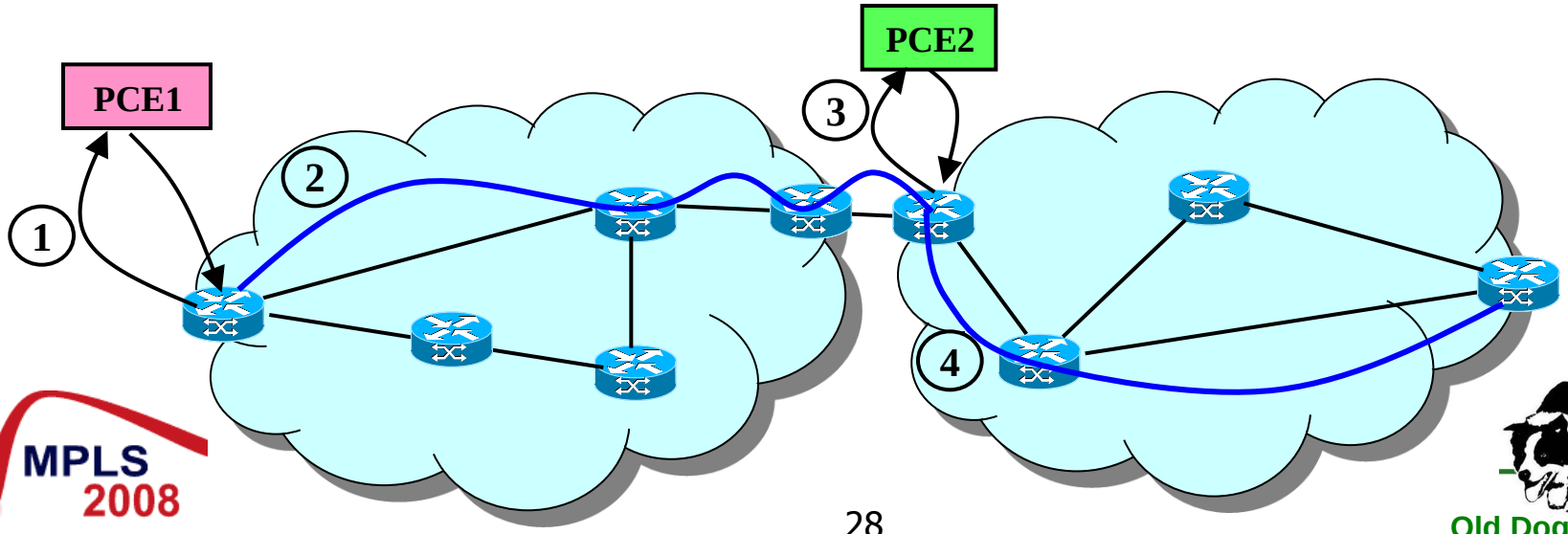
- PCReq message asks for a path computation
  - Start and end points
  - Basic constraints
    - Bandwidth
    - LSP attributes
    - Setup/holding priorities
    - Path inclusions
  - Metric to optimise
    - IGP metric
    - TE metric
    - Hop count
  - Associated paths
- PCRep reports the computed path
  - Explicit route
  - Actual path metrics
  - (Or the failure to find a path)

Multi-Domain Usage Scenarios

- The main purpose of PCE is to solve the multi-domain problem
  - Compute paths across multiple domains
- Three main methods have already been defined
  - Per-domain path computation
  - Simple cooperating PCEs
  - Backward Recursive Path Computation

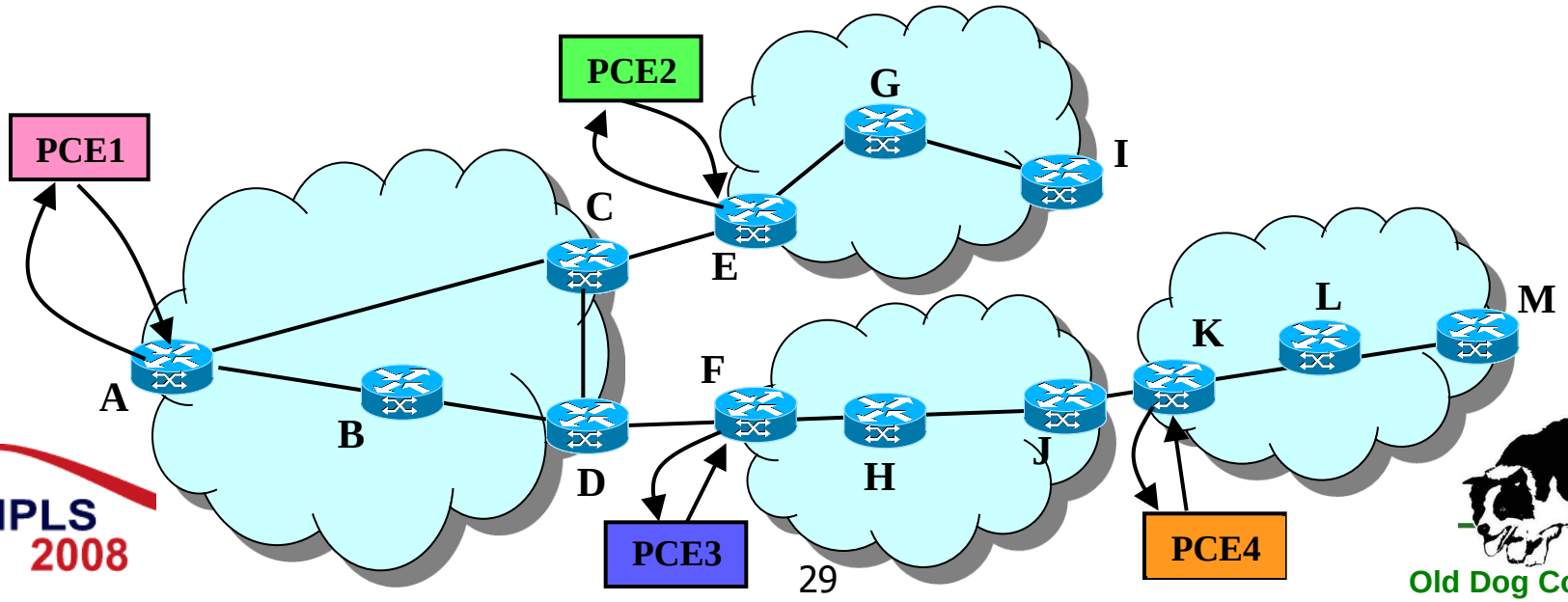
Per-Domain Path Computation

- Computational responsibility rests with domain entry point
- Path is computed across domain (or to destination)
- Simple mechanism works well for basic problems or for "good-enough" paths
- Which domain exit to choose for connectivity?
  - Follow IP routing? First approximation in IP/MPLS networks
  - Sequence of domains may be "known"
- Which domain exit to choose for optimality?



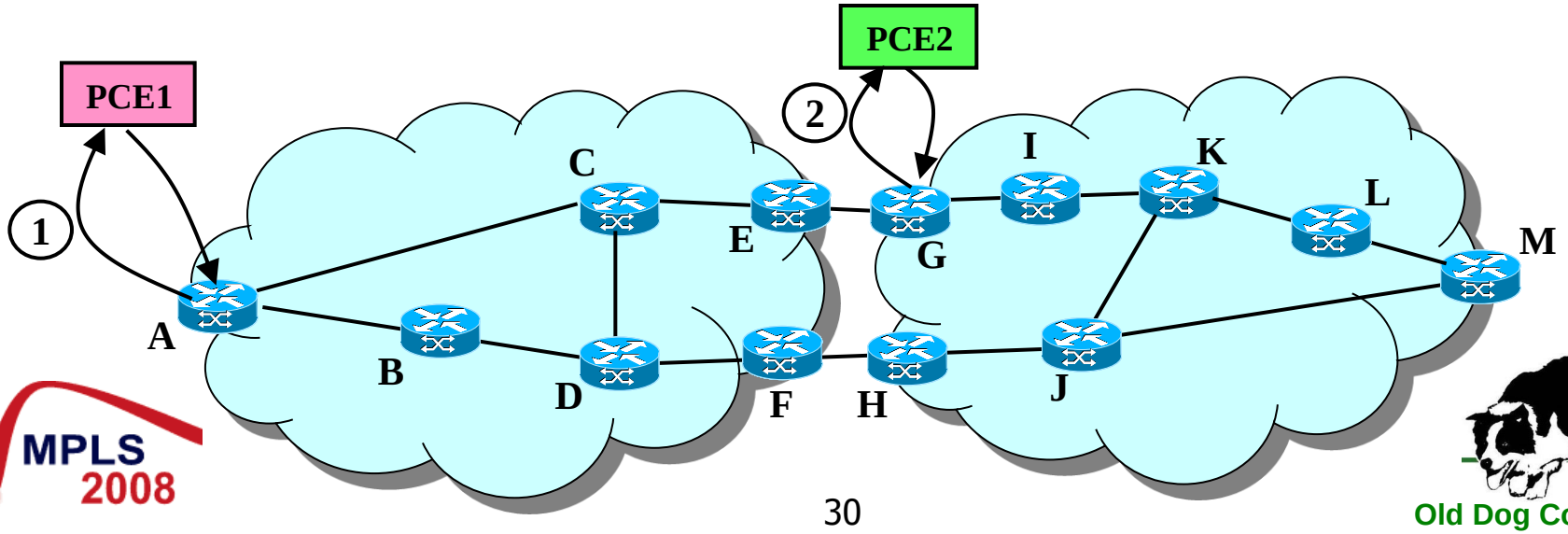
Issues with Per-Domain Computation

- Choice of successive domains
  - PCE1 does not know where the destination is
  - Does it choose the path ACE or the path ABDF?
- There are some signaling solutions that can help
  - For example, crankback
  - Can be very slow and complicated



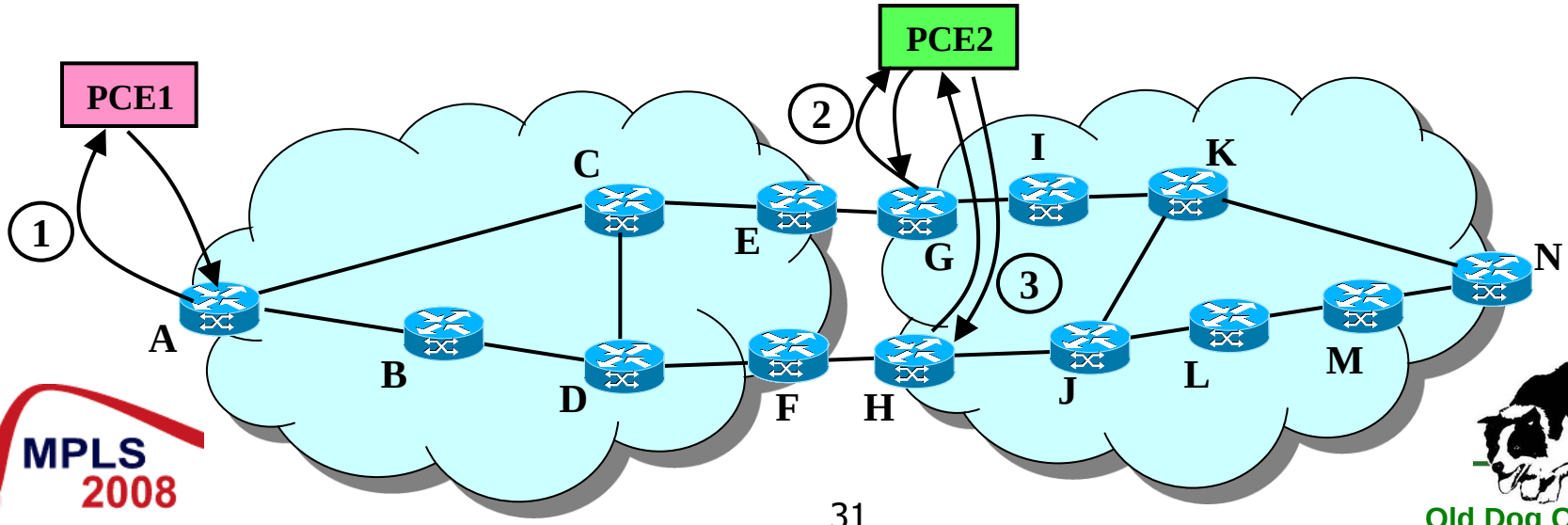
Issues with Per-Domain Computation

- Multiple connections between domains
- PCE1 will select the path ACEG toward the destination
  - Results in the path ACEGIKLM (path length 7)
- A better path would be ABDFHJM (path length 6)
- PCE1 cannot know this

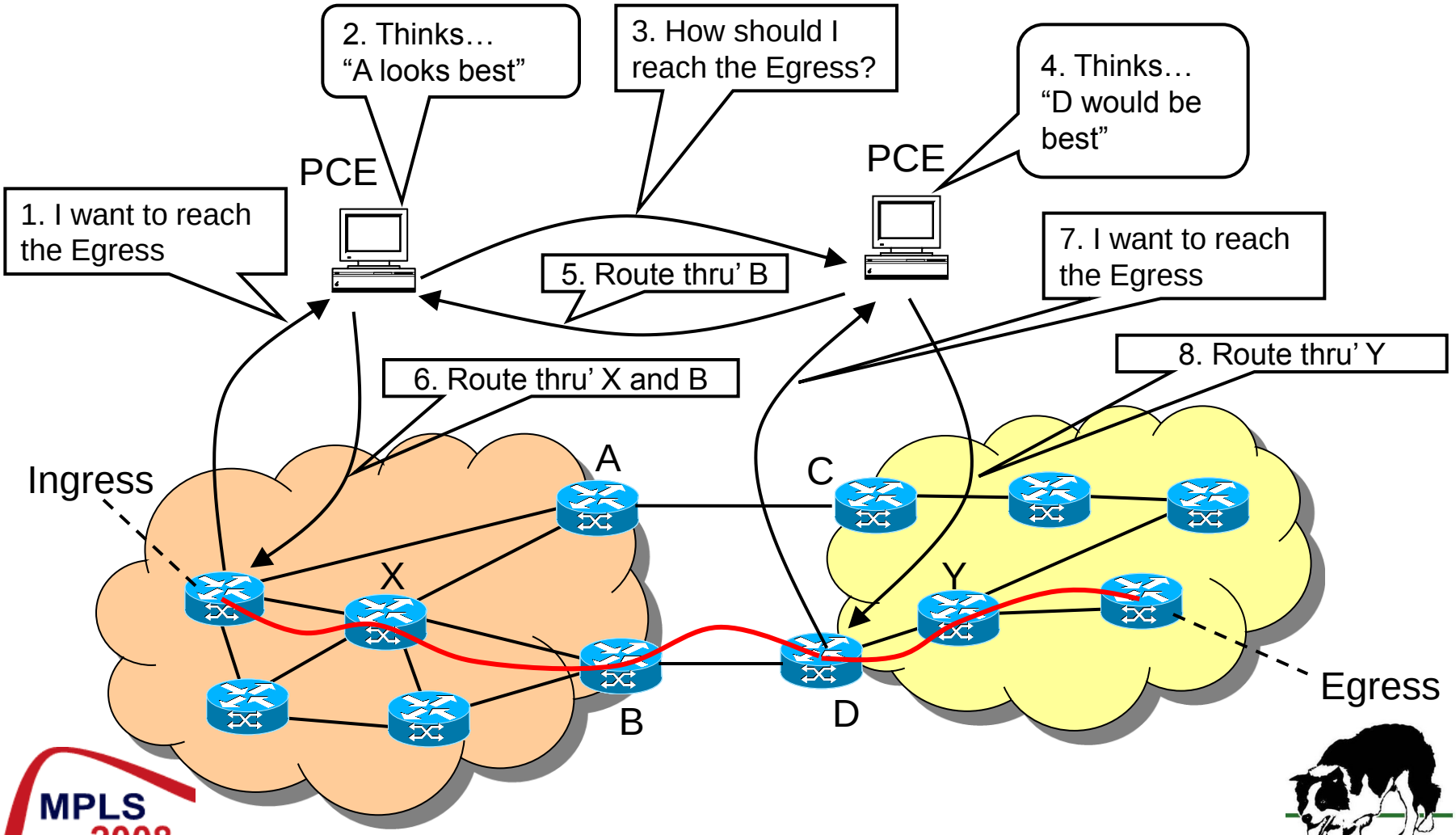


Issues with Per-Domain Computation

- Disjoint paths (for example, for protection)
- PCE1 supplies {ACEG and ABDFH}
  - Disjoint in first network
- Separate requests are made to PCE2 from G and H
  - Results in shortest paths in second network {GIKN and HJKN}
- Resulting paths ACEGIKN and ABDFHJN are not disjoint
  - Link KN is shared
- A possible solution exists {ACEGIKN and ABDFHJLMN}
  - There may be some signaling solutions to this problem in some scenarios

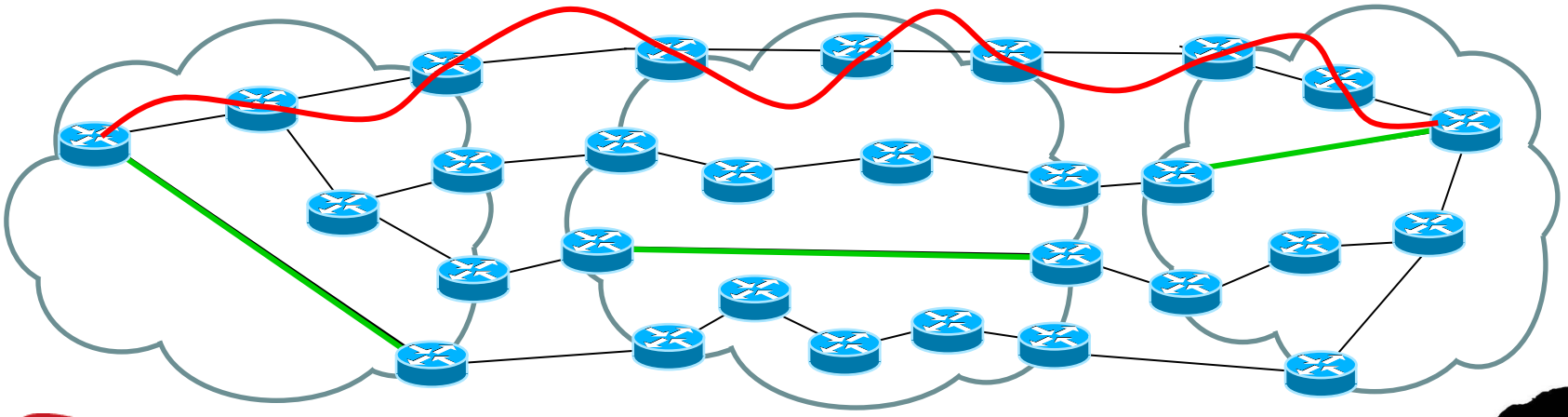


A Simple Example – Cooperating PCEs



Issues with Simple Cooperating PCEs

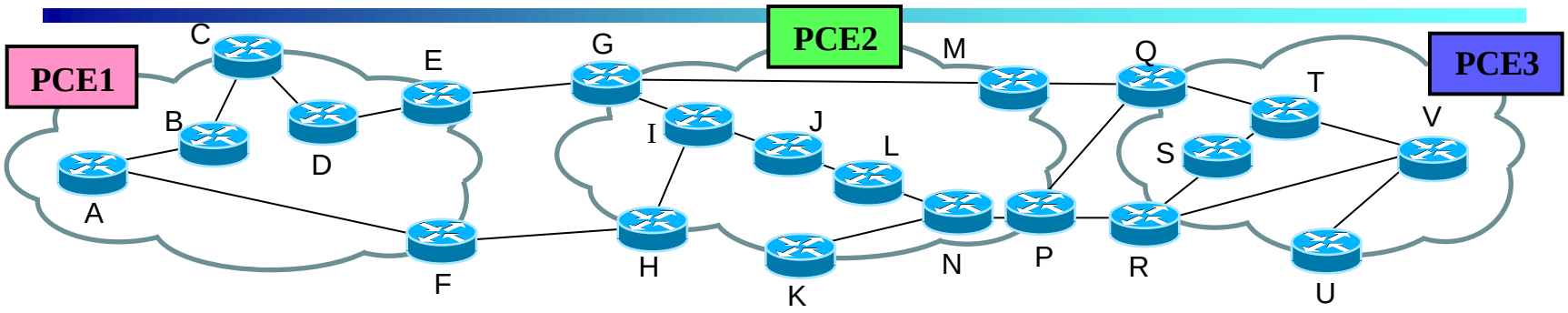
- More than two domains in sequence gets complicated
- Not enough to supply the best path in one domain
  - Hard to achieve optimality
    - The best end-to-end path may use none of the bests paths from each domain



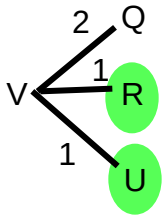
Backward Recursive Path Computation

- PCE cooperation
  - Can achieve optimality without full visibility
  - "Crankback at computation time"
- Backward Recursive Path Computation is one mechanism
  - Assumes each PCE can compute any path across a domain
  - Assumes each PCE knows a PCE for the neighbouring domains
  - Assumes destination domain is known
- Start at the destination domain
  - Compute optimal path from each entry point
  - Pass the set of paths to the neighbouring PCEs
- At each PCE in turn
  - Compute the optimal paths from each entry point to each exit point
  - Build a tree of potential paths rooted at the destination
  - Prune out branches where there is no/inadequate reachability
- If the sequence of domains is "known" the procedure is neater

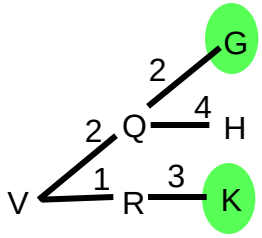
BRPC Example



- PCE3 considers:
  - **QTV cost 2**; QTSRV cost 4
  - RSTV cost 3; **RV cost 1**
  - **UV cost 1**
- PCE3 supplies PCE2 with a path tree



- PCE2 considers
  - **GMQ..V cost 4**; GIJLNPR..V cost 7; GIJLNPQ..V cost 8
  - HIJLNPR..V cost 7; **HIGMQ..V cost 6**; HIJLNPQ..V cost 8
  - **KNPR..V cost 4**; KNPQ..V cost 5; KNLJIGMQ..V cost 9
- PCE2 supplies PCE 1 with a path tree



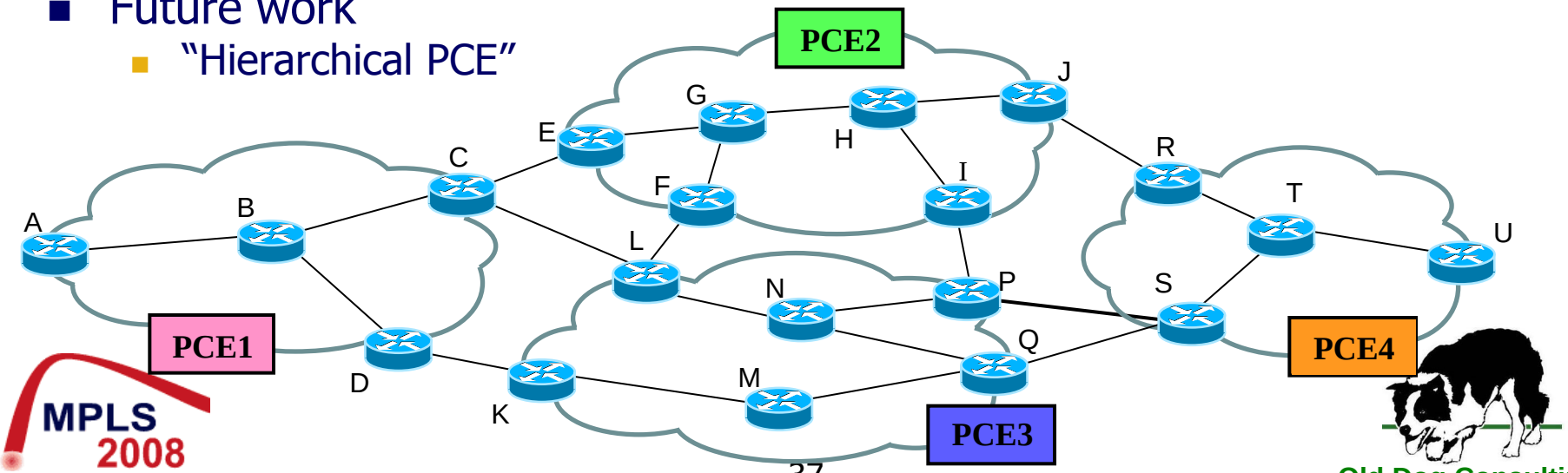
- PCE1 considers
  - ABCDEG..V cost 9
  - **AFH..V cost 8**
- PCE1 selects AFHIGMQTV cost 8

Problems with BRPC

- Destination domain must be known
  - Maybe not unreasonable
    - Destination is known, so destination domain may be known
    - Some mechanisms (like BGP) can distribute location
- Otherwise, need a mechanism to find the destination
  - BGP may suggest a sequence of domains for reachability
    - Works in IP networks
    - Might not be optimal in TE cases
    - IP might not be present (e.g., optical networks)
- Future work
  - "Forward Recursive Path Computation"
    - What is special about backward recursion?
  - "Hierarchical PCE"
    - Discussed later

Problems with BRPC

- Navigating a mesh of domains may be complex
  - Even in a relatively simple example
- PCE4 supplies path trees to PCE2 and PCE3
- PCE2 supplies a tree to PCE3 and PCE3 supplies a tree to PCE2
- PCE1 receives trees from PCE2 and PCE3
  - Maybe several times
- Problem eased by knowing sequence of domains in advance
  - Still some issues with multiple connections
- Future work
  - "Hierarchical PCE"



Core Protocol Extensions

- Explicit route exclusions
  - Identify resources to exclude from the computed path
- Path confidentiality
  - Compute full paths but hide the details of the results
- Objective functions
  - Control of how the PCE interprets the metrics
- DiffServ support
  - Simple additions to specify the DiffServ Class Type

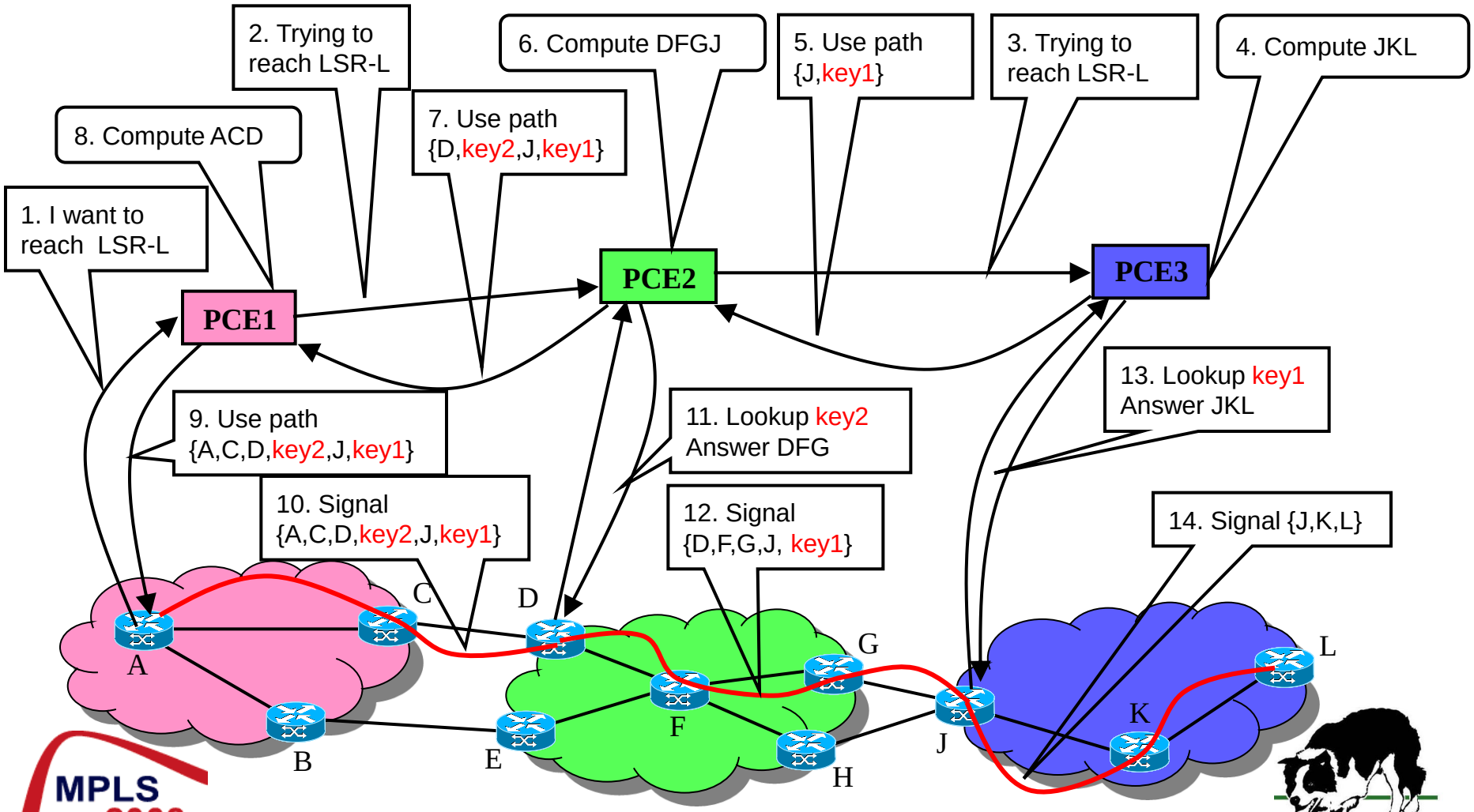
Explicit Route Exclusions

- Operational requirements
  - Find a path that avoids a specific node or link
    - Known issues or reliability or maintenance
  - Find a path that avoids another path
    - Protection function
- Route exclusion allows specification of resources to avoid
  - labels, links, nodes, domains, and SRLGs
- Just another object in the PCReq

Path Confidentiality

- Cooperative PCEs exchange path information
  - This is transferred to signaling to set up the LSP
- But a path fragment reveals information about a domain
  - Some ASes will not want to share this information
    - Confidentiality
    - Security
- Could use loose hops or domain identifiers
  - This hides information efficiently
  - Forces a second computation to be performed during signaling
    - Might lose diversity
- A PCE can replace a path segment with a token
  - We call this a *path key*
- Could be anything
  - No semantic outside the context of the PCE
- De-referenced on entry to a domain

Path Keys


2. Trying to reach LSR-L
6. Compute DFGJ
5. Use path {J,key1}
3. Trying to reach LSR-L
4. Compute JKL
8. Compute ACD
7. Use path {D,key2,J,key1}
1. I want to reach LSR-L
PCE1
PCE2
PCE3
13. Lookup key1 Answer JKL
9. Use path {A,C,D,key2,J,key1}
11. Lookup key2 Answer DFG
10. Signal {A,C,D,key2,J,key1}
12. Signal {D,F,G,J, key1}
14. Signal {J,K,L}
A
B
C
D
E
F
G
H
J
K
L

Objective Functions

- PCEP allows us to convey
  - Path end points
  - Desired path constraints (e.g. bandwidth)
  - Computed path
  - Aggregate path constraints (e.g. path cost)
- But how do we control the way the PCE computes the path?
- An objective function specifies the desired outcome of the computation (not the algorithm to use)
- These can be communicated in a new object
  - Minimum cost path
  - Minimum load path
  - Maximum residual bandwidth path
  - Minimize aggregate bandwidth consumption
  - Minimize the load of the most loaded link
  - Minimize the cumulative cost of a set of paths

Advanced Uses

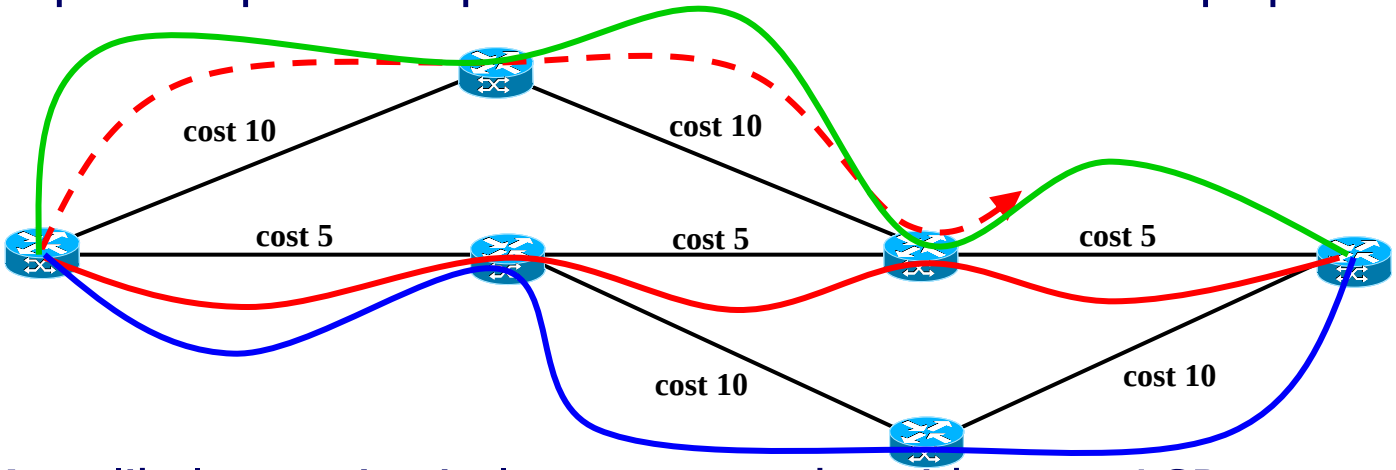
- PCE has become a very powerful concept
- It is being actively examined for use in a wide range of MPLS and GMPLS computation problems
  - Point-to-multipoint LSPs
  - Global concurrent optimization
  - Optical networks
  - VPN management
  - Inter-layer path computation
  - Service and policy management
  - New PCE cooperation techniques
  - Operation of ASON routing
  - Routing multi-segment pseudowires

Point-to-Multipoint Computation Requirements

- Support of complex services
  - High levels of QoS demand multiple constraints
    - Minimal cost, minimal delay, high bandwidth, etc.
    - Computing a minimum-cost tree (Steiner tree) is NP-hard
    - Constraints may conflict with each other
  - Many multiple 'parallel' connections to support one service
- Path diversity or congruence
  - End-to-end protection with link, node, or SRLG diversity
  - Mesh (m:n) service protection
  - Congruent paths for fate-sharing (e.g. virtual concatenation)
- Control of branching points
- Global concurrent network optimisation
  - Compute multiple trees and consider moving existing trees to accommodate new trees
  - Consider multiple complex constraints, including lower (optical) constraints

Global Concurrent Optimization (GCO)

- Sequential path computation can lead to classic "trap" problems



- More likely to arise in larger networks with more LSPs
- Standard PCEP allows a PCC to submit related requests for simultaneous computation
- Trap problems can also arise from multiple head-ends
- GCO allows the coordination of computation of multiple paths
  - Particularly useful for re-optimization of busy networks
  - May require consideration of migration paths

Optical Networks

- Optical network path computation can be split
  - Impairment-free networks
    - The main problem is selecting paths with a continuous wavelength end-to-end
    - The Routing and Wavelength Assignment problem (RWA)
    - Somewhat more complicated than normal CSPF
  - Networks with Optical Impairments
    - Power levels, OSNR, PMD, etc.
    - Very complex path computations
- Large amounts of information required
- Considerable processing requirements
- Optical devices have limited CPU and memory
- Makes sense to devolve path computation to a dedicated server
- A lot of path planning in these networks is off-line

VPN Management

- VPNs provide several routing problems
- Network resources may be partitioned for VPNs
  - There may be policies about how resources are used
  - There may be policies about which VPNs can share
- Network resources may be shared between VPNs
  - PEs will not know how the network is currently used
- CEs may be multi-homed and need to select a PE
  - The PEs may have different connectivity
- Addresses may be scoped per VPN
- Multi-cast VPNs are becoming important

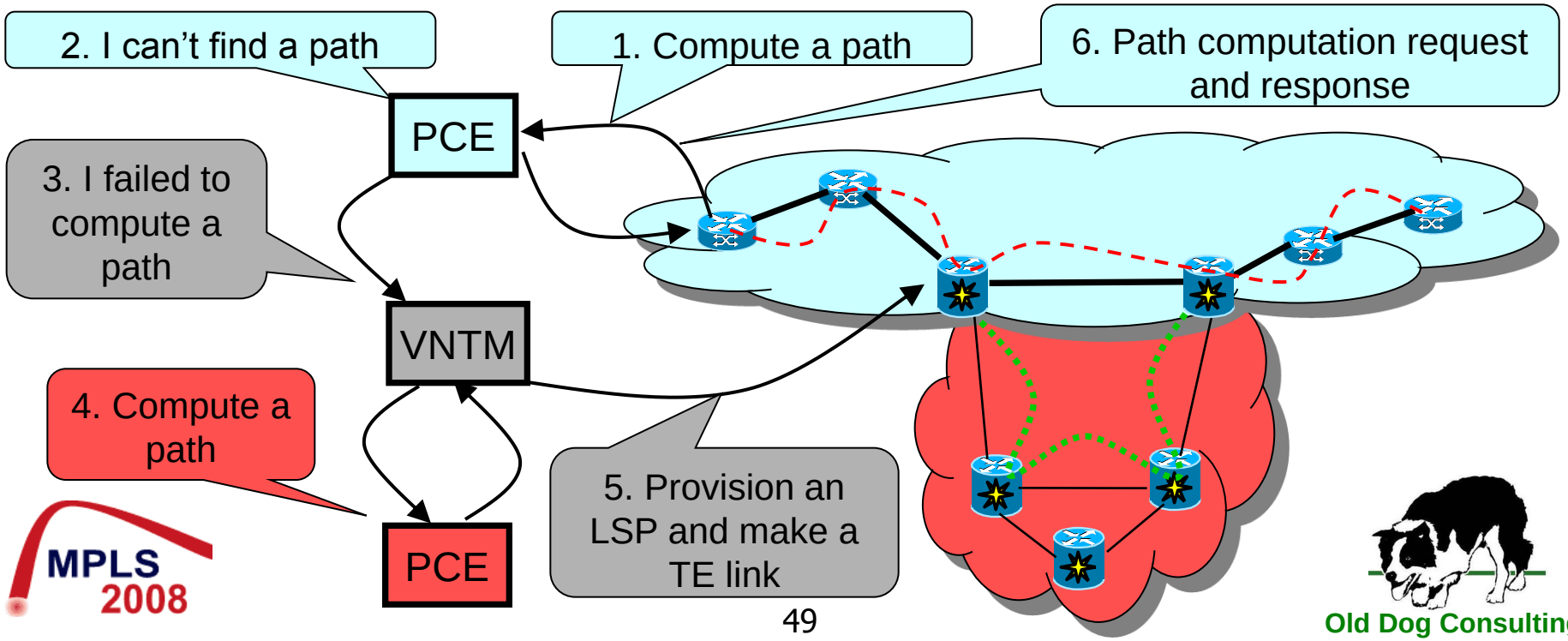
Inter-Layer Path Computation

- Client/server networks

- Several PCE models
  - Single PCE with multi-layer visibility
    - Two TE domains, but one PCE can see both of them
  - Two PCEs without cooperation
    - Per-domain path computation is used
  - Two PCEs with cooperation
    - Some mechanism such as BRPC is used
  - Separate PCEs with management coordination
    - Allows the server network to retain control of expensive transport resources

Virtual Network Topology Manager Interactions with PCE

- VNT Manager is a policy/management component
- Acts on triggers (operator request for a client TE link, client network traffic demand info, client TE link usage info, client path computation failure notification)
- Uses PCE to determine paths in lower layer
- Uses management systems to provision LSPs and cause them to be advertised as TE links in the client layer



Service Management

- ITU-T's Resource and Admission Control Function (RACF)
  - Plans and operates network connectivity in support of services
- Policy Decision Functional Entity
  - Examines how to meet the service requirements using the available resources
- Transport Resource Controller Functional Entity
  - Provisions connectivity in the network (may use control plane)

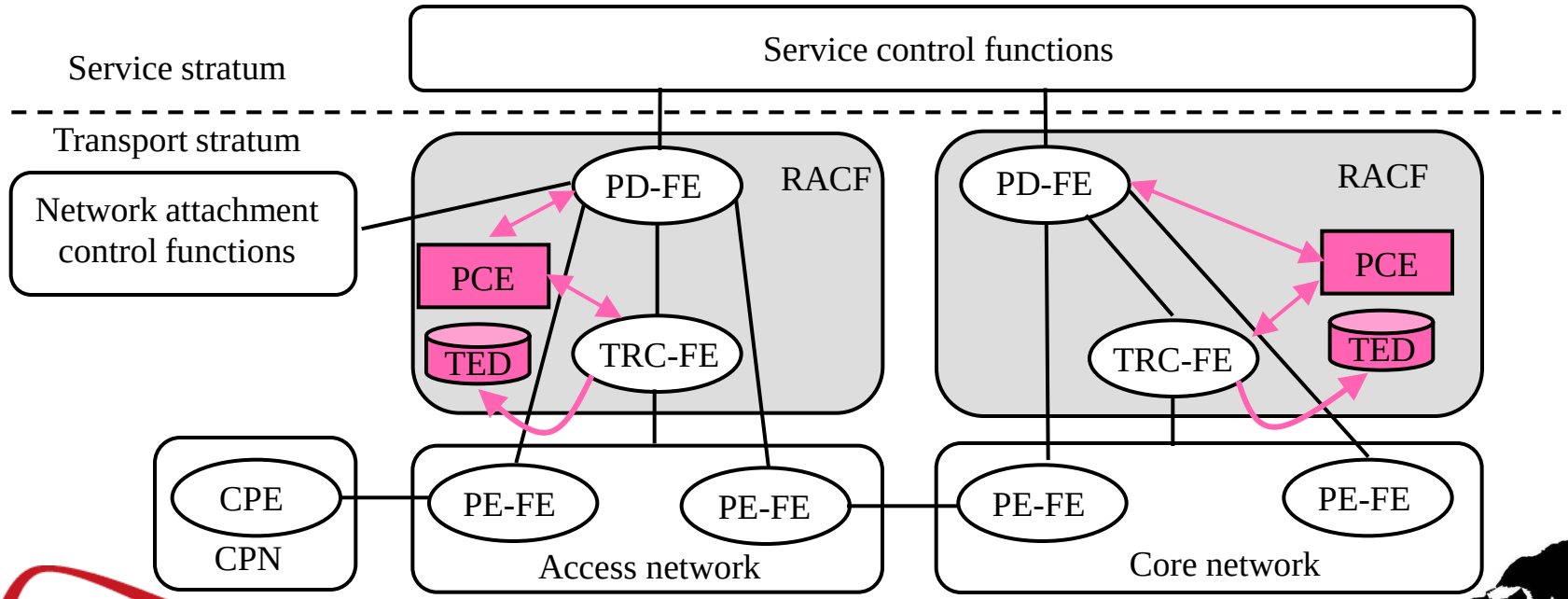


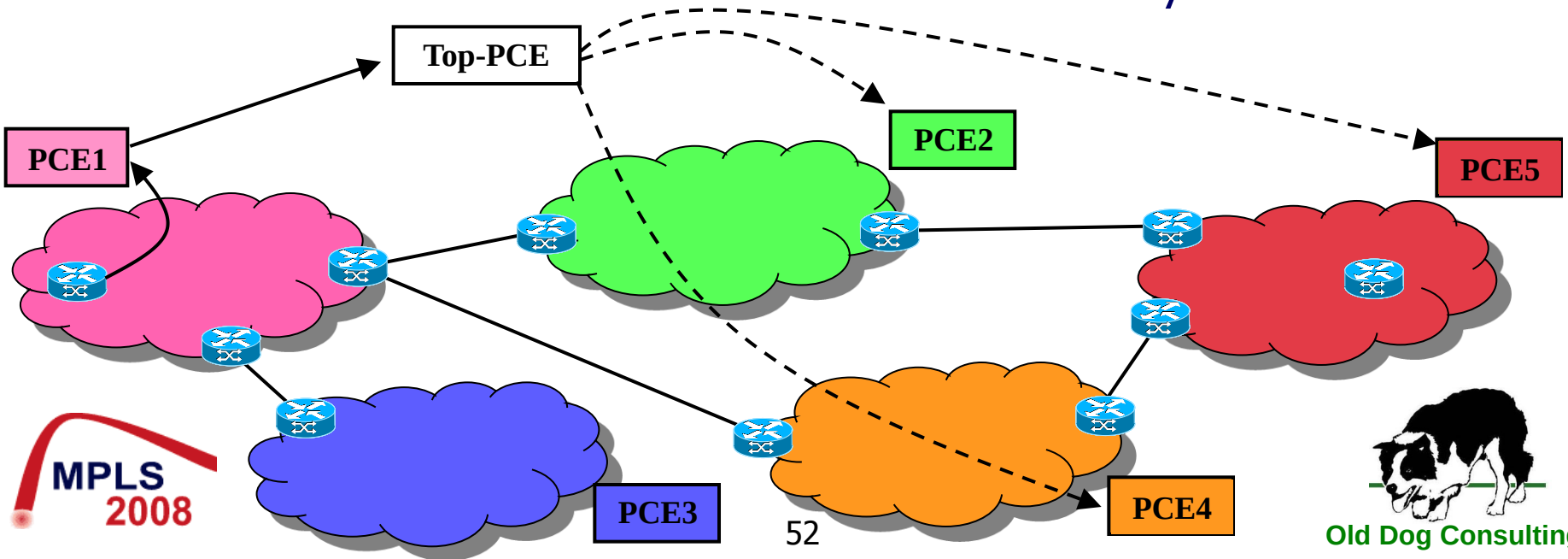
Figure based on ITU-T Y.2111

Integration with Policy

- Policy is fundamental to PCE
  - What should a PCC do when it needs a path?
  - What should a PCE do when it gets a computation request?
  - Which algorithms should a PCE use?
  - How should PCEs cooperate?
- RACF PD-FE is a policy component that could use PCE
- Inter-domain paths are subject to Business Policy
  - IPsphere Forum is working on business boundaries
    - Business policy may guide PCE in its operation
    - Selection of domains based on business parameters is a path computation that PCE could help with

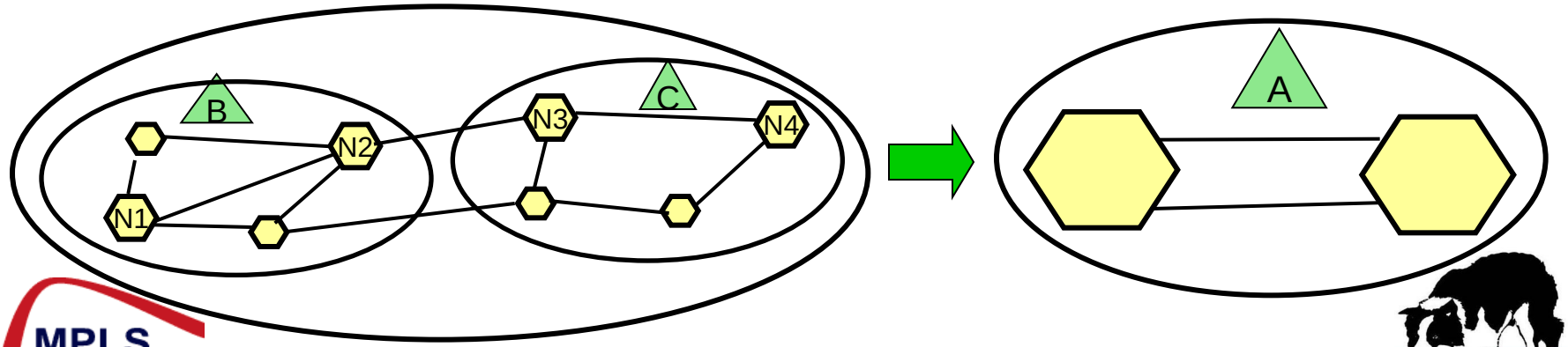
Hierarchical PCE

- A solution to inter-domain TE routing may be hierarchical PCEs
  - Recall that BRPC does not scale well with complex inter-connection of domains
- Hierarchical PCE is ***not*** an all-seeing eye!
  - It knows connectedness of domains
  - It provides *consultative* coordination of subsidiary PCEs
- Per-domain PCEs can be invoked simultaneously



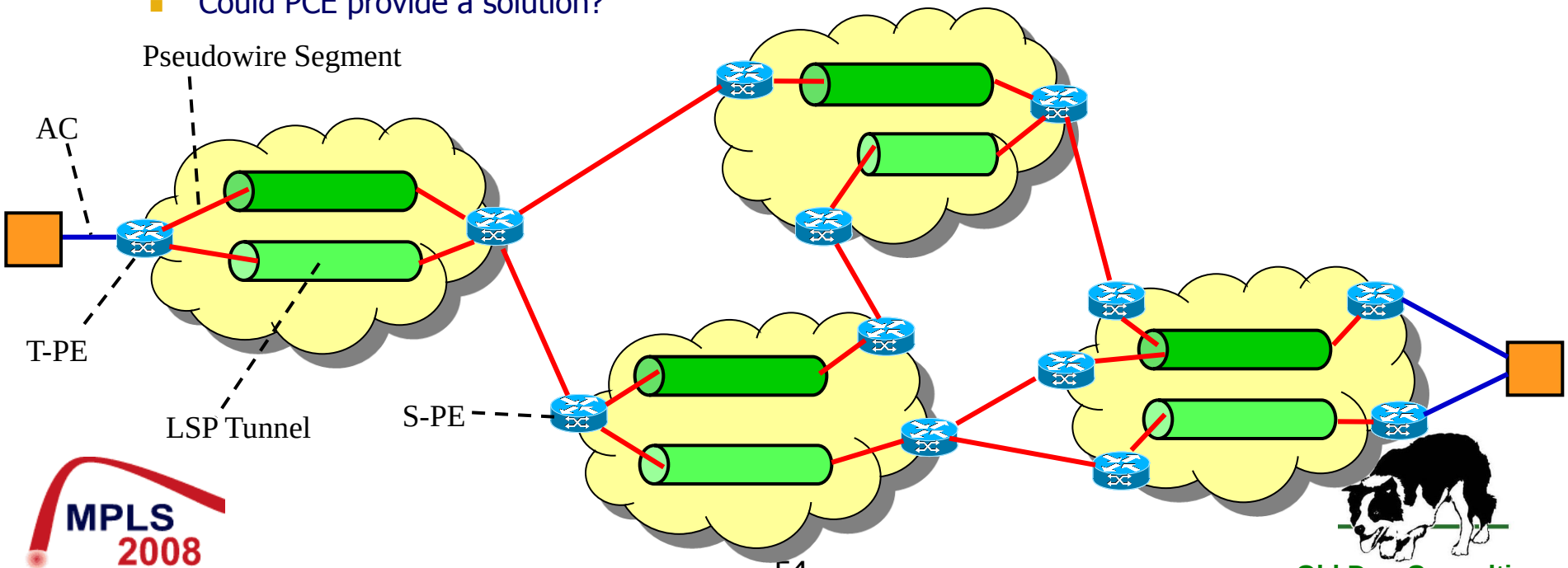
PCE in ASON

- ITU's Automatically Switched Optical Network uses hierarchical routing
  - Networks are constructed from sub-networks
    - Administrative domains
    - Clusters of single-vendor equipment
    - Topological entities (rings, protection domains, etc.)
  - Routing Areas have containment relationships
    - Routing controllers share information between peers
    - There is a parent-child relationship between routing controllers
- Fits particularly well with the hierarchical PCE model



Pseudowire Routing

- Pseudowire networks create a multi-layer routing problem
  - Establishment and routing of LSP tunnels
  - Choice of LSP tunnels to carry pseudowires
  - Choice of "parallel" pseudowires
  - Choice of switching PEs
  - Choice of terminating PEs
- Problem extends to point-to-multipoint pseudowires
- These problems is not properly addressed at the moment
  - Could PCE provide a solution?



Summary

- PCE is a logical functional component
  - It may be centralized within a domain or distributed
  - It is ***not*** an all-seeing oracle
- PCEs may cooperate to determine end-to-end multi-domain paths
- The PCEP protocol is quite simple
  - It can carry lot of information
- The PCE concept is already implemented for MPLS-TE
- PCE is drawing a lot of interest in a wide variety of environments

References

- The Internet Engineering Task Force (IETF) is the main originating body for PCE
  - See the PCE working group home page http://www.ietf.org/html.charters/pce-charter.html
  - The key documents are
    - RFC 4655 *A Path Computation Element (PCE)-Based Architecture*
    - RFC 5088 *OSPF Protocol Extensions for Path Computation Element (PCE) Discovery*
    - draft-ietf-pce-pcep-15.txt *Path Computation Element (PCE) Communication Protocol (PCEP)*
- The IPsphere Forum can be found at http://www.ipsphereforum.org
- The ITU-T has worked on several relevant documents
  - Access documents via http://www.itu.int/publications/sector.aspx?sector=2
    - G.7715.2 *ASON routing architecture and requirements for remote route query*
    - Y.2111 *Resource and admission control functions in Next Generation Networks*

Questions

adrian@olddog.co.uk

PCE Working Group

http://www.ietf.org/html.charters/pce-charter.html