Spanning the World with Ethernet

The Five Rules that Ensure Interoperability among Ethernet Service Providers on a Global Scale

Norman Finn, Cisco Systems

Notes

- This presentation represents the opinions of the author on certain aspects of the “Ethernet Service Providers” problem.
- It is the author’s intention to present some variant and/or subset of this presentation to most of the relevant standards bodies, in the interest of generating a consensus on a model for interoperability among Ethernet Service Providers.
- This presentation will change, as new ideas are advanced.

Part 1: Avoiding Global Broadcast Storms

What are “Ethernet Service Providers” trying to accomplish?

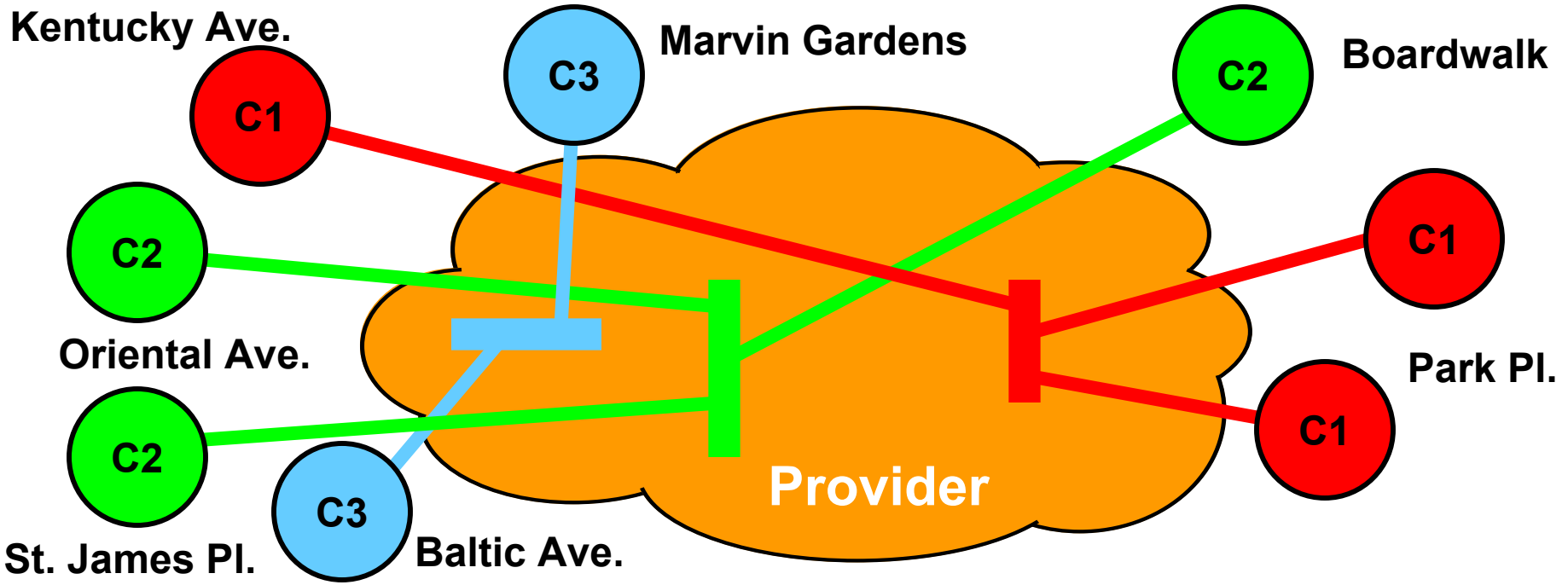
- An Ethernet Provider wants to sell, to many different Customers, what seem to be Ethernet connections.

  Each “Customer Service Instance” may span a city or a continent.

  Some Providers want to use bridges and routers to implement the services.

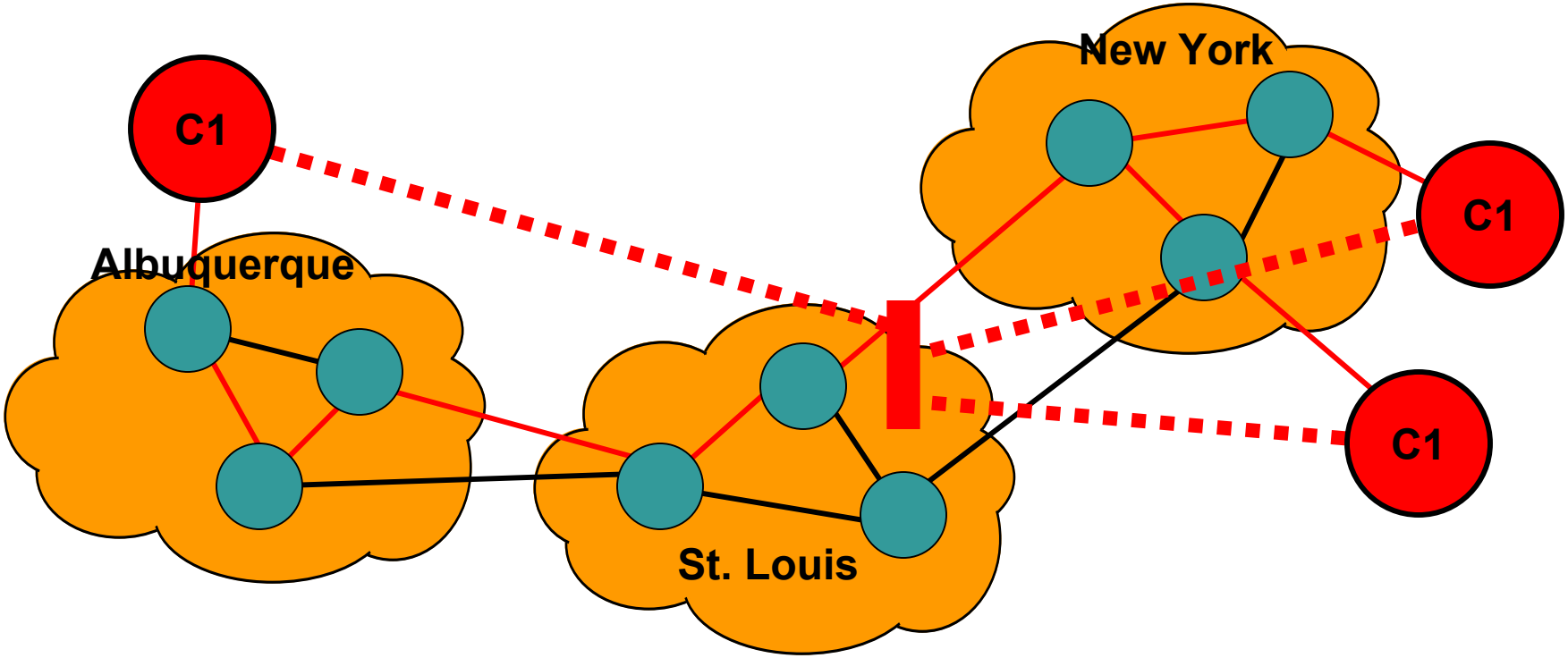
  These Providers want to interconnect with other, similar, Ethernet Providers.

What are “Ethernet Service Providers” trying to accomplish?



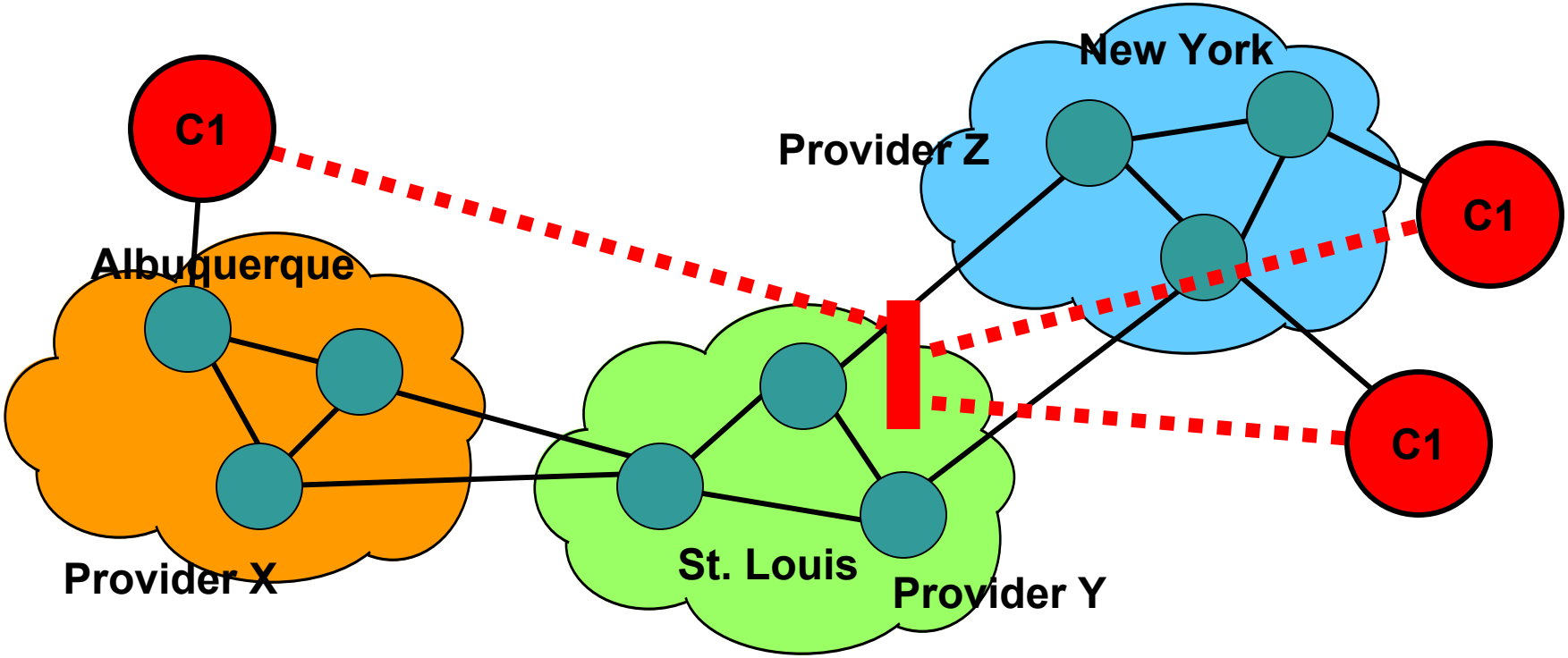
- Each Customer, Red, Green, and Blue, purchases a separate instance of what appears, to each, to be a point-to-point or shared-medium Ethernet.

What are “Ethernet Service Providers” trying to accomplish?



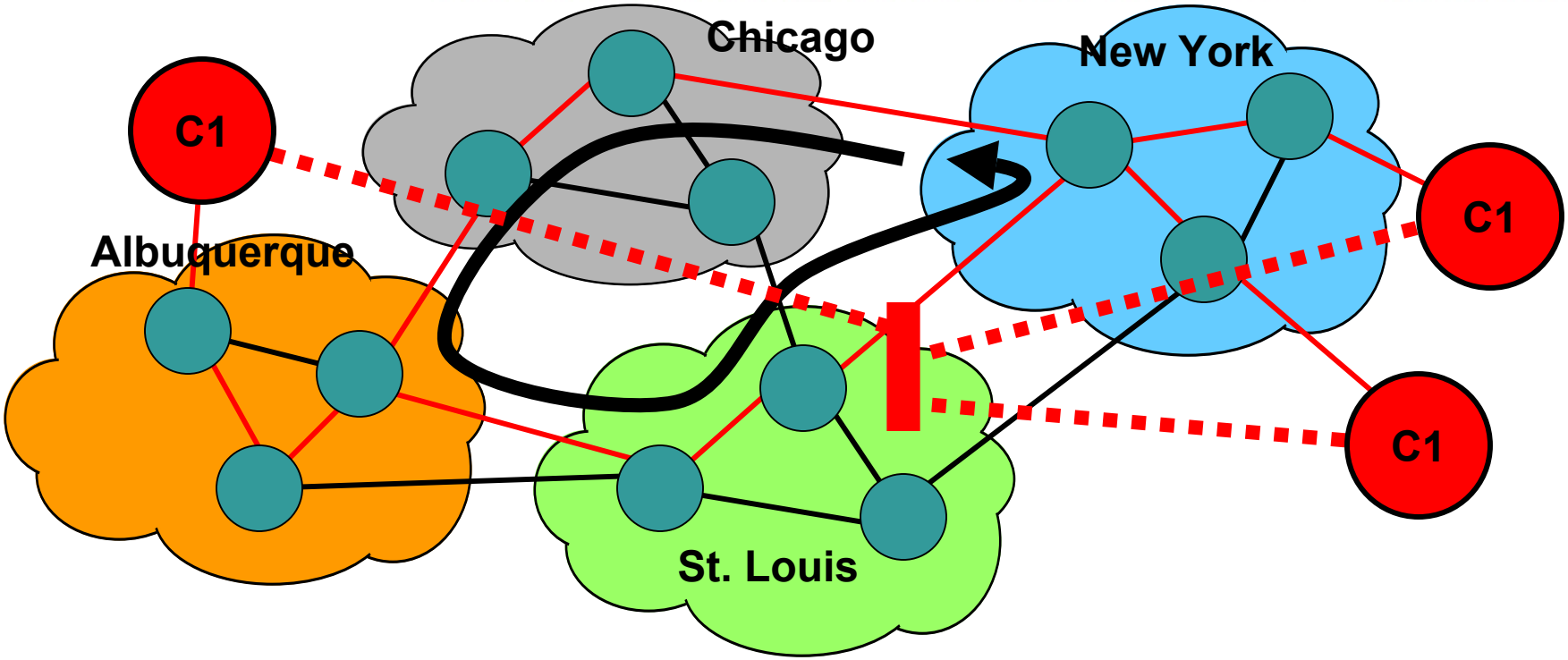
- Provider’s network is composed of bridges and routers, perhaps spanning continents.

What are “Ethernet Service Providers” trying to accomplish?



- Providers may be interconnected

What are “Ethernet Service Providers” trying to accomplish?



- We must prevent Layer 2 forwarding loops!

How might one implement a Customer Service Instance?

- **If all were point-to-point, it wouldn't matter.**

  Any technology could carry Ethernet frames transparently end-to-end without even knowing that it was carrying Ethernet frames.

- **But, many Customers want multipoint-to-multipoint services!**

  If they wanted every frame output on every port in the service, it *still* would be easy to use any underlying technology capable of broadcast/multicast, in ignorance of Ethernet.

How might one implement a Customer Service Instance?

- But, many Customers want, and are willing to pay for, an “intelligent” service.

  Service is multipoint-to-multipoint. (It swims.)

  Service must carry frames which are not IP packets. (It flies.)

  Service should not deliver frames to Customer ports where they are not wanted or needed. (It quacks.)

- In short, many Customers want a service which closely resembles a Bridged LAN. (It’s a duck!)

So, why don't you just …

- Run a spanning tree algorithm to prevent loops?
- But, bridges cannot span the world; they relay on Spanning Tree Protocols!
  1. The Spanning Tree algorithms do not scale to cover the world.
  2. The Spanning Tree algorithms assume a single operational authority.

So, why don't you just …

- **Run the spanning tree on one VPLS instance at a time?**

  No one VPLS is too large for a spanning tree!

- **But, each Provider Bridge may handle hundreds, or even thousands, of VPLSs.**

  1. **Will all technologies, and all standards bodies, and all Providers, agree to use a common spanning tree?**

So, why don't you just …

- Use routers and routing protocols on Ethernet MAC addresses?
- But, Ethernet frames cannot be forwarded using the existing routing protocols!

1. All the routing protocols can temporarily forward packets in a circle while responding to a change in the network topology.
2. Routed protocols (e.g. IP) have a Time-To-Live field that kills looping packets.
3. Ethernet frames have no TTL field.

Not to mention …

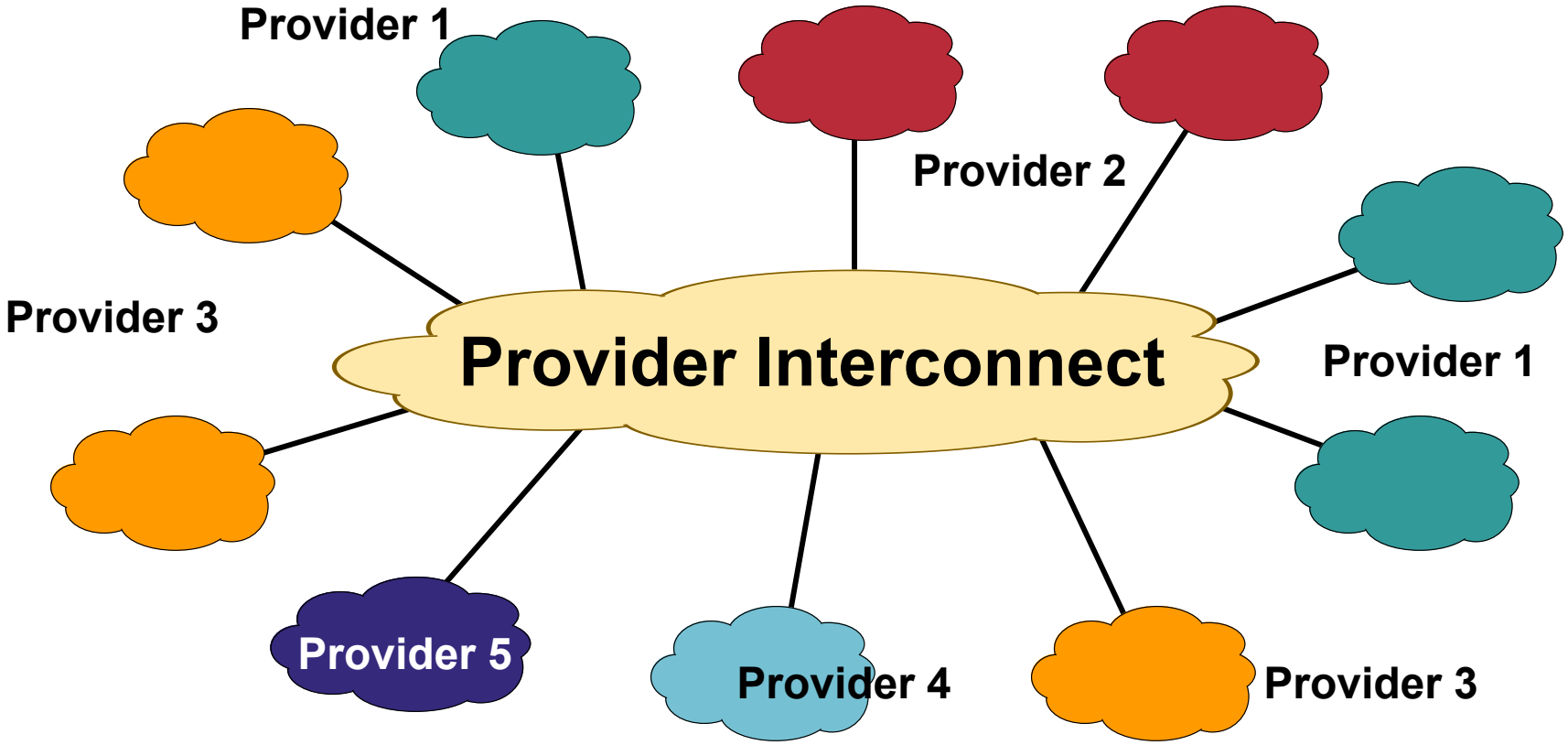
- The routing protocols allow out-or-order packet delivery, whereas users' native Ethernet protocols may fail when frames are delivered out of order.
- The Layer 2 address space is perfectly flat, with no provision for the summarization of geographically related addresses.

Spanning the World at L2 is hard!

- So, if we cannot bridge, because the spanning tree protocols cannot expand to global scales;
- And, we cannot route, because Ethernet frames have no TTL; then

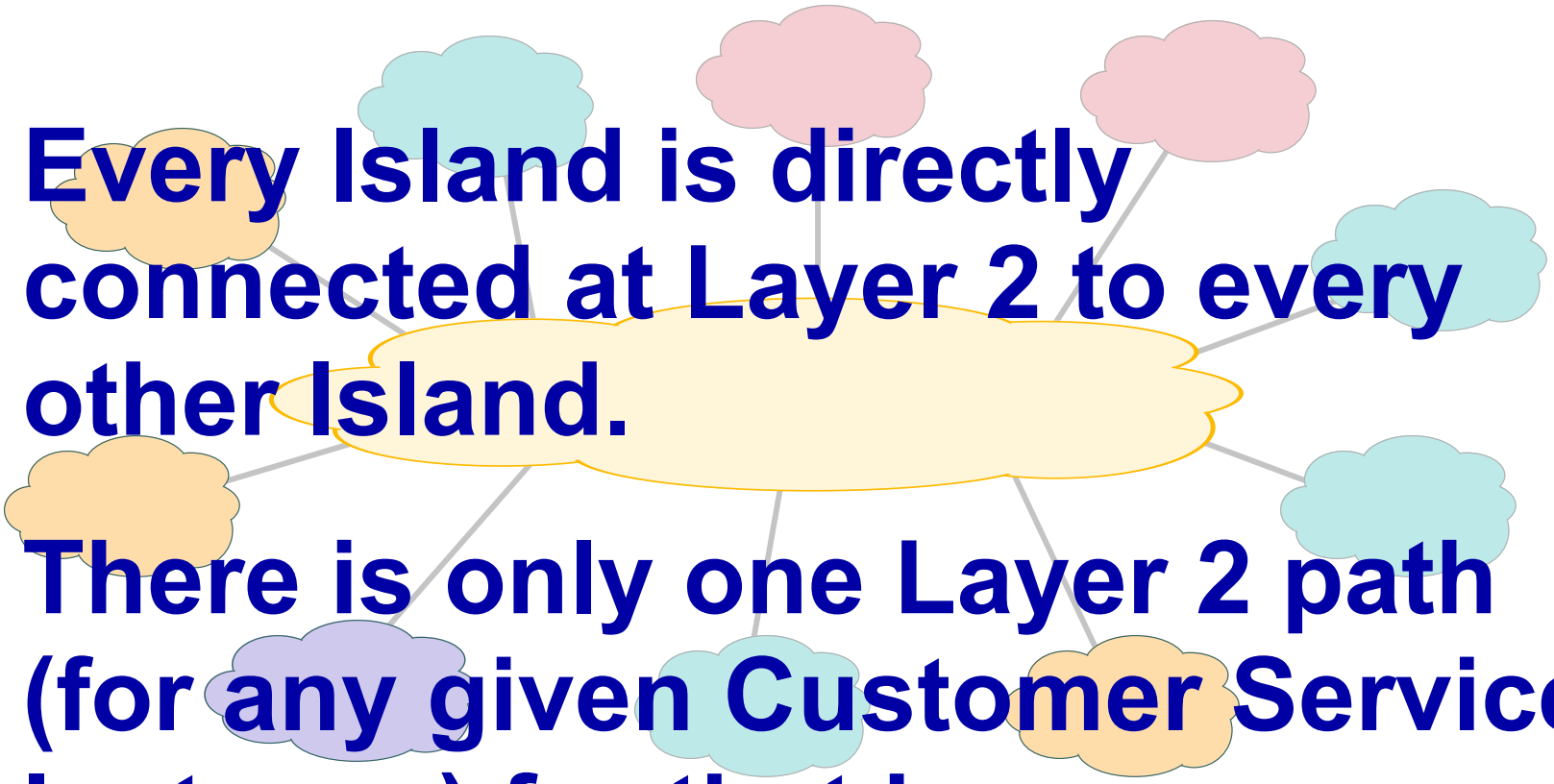
How do we avoid "Spanning the World" with forwarding loops?!

Answer: Eliminate loops using a topology constraint:



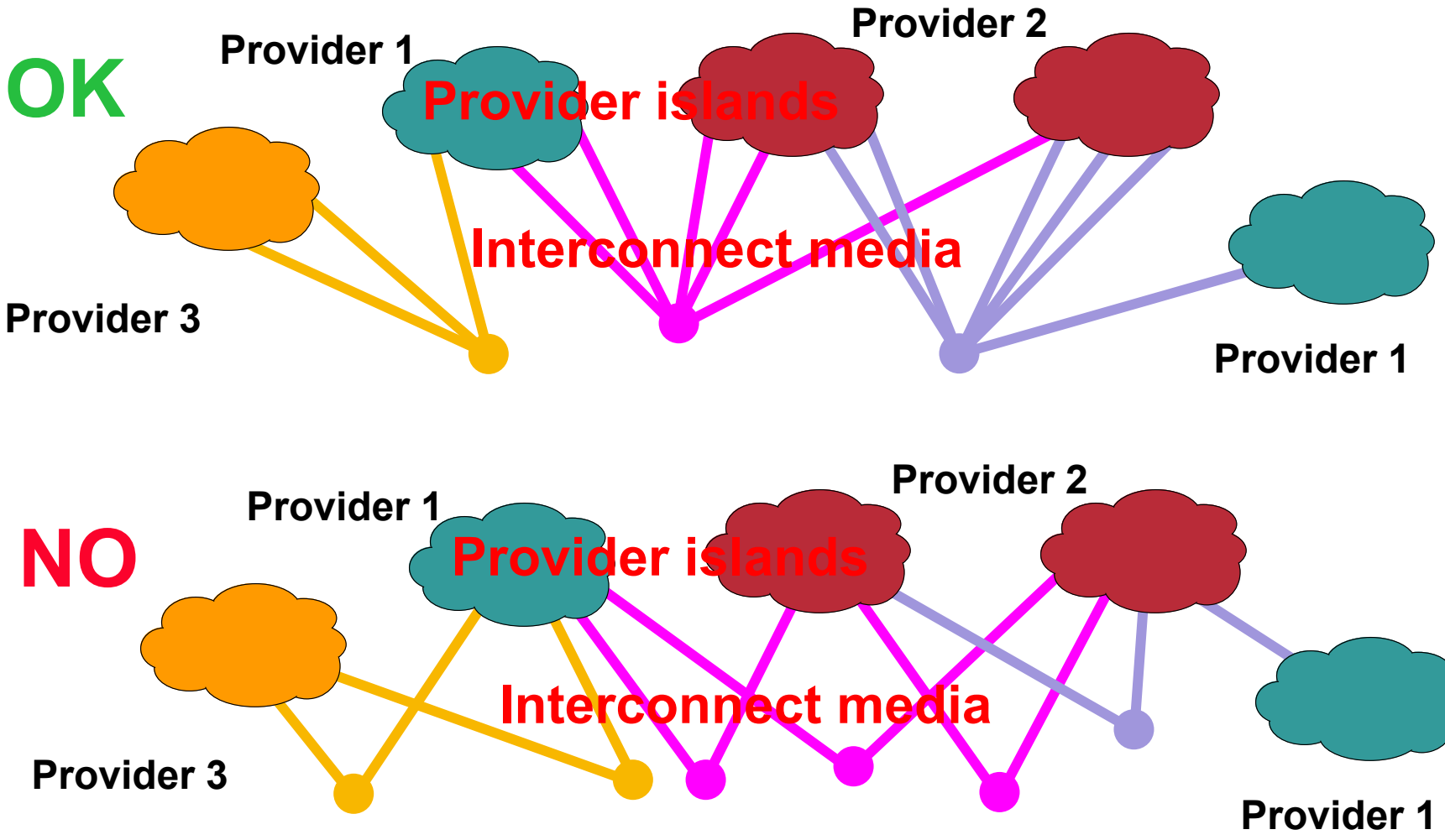
"Islands" of Provider Networks are all (and only!) connected via a single interconnect cloud!

In other words,



- Every Island is directly connected at Layer 2 to every other Island.
- There is only one Layer 2 path (for any given Customer Service Instance) for that hop.

Zooming in on the Topology Constraint: We define “Interconnect Media”



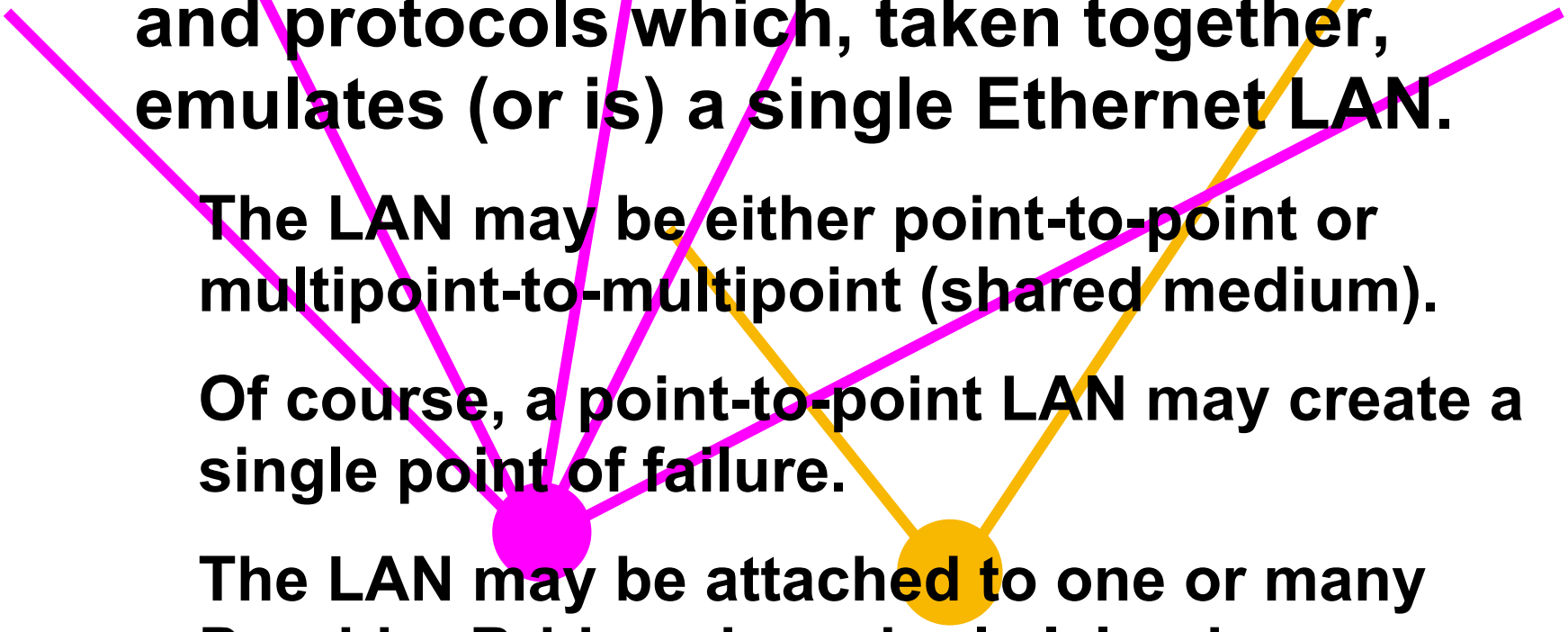
What is an Interconnect Medium?

- A system of real and/or virtual data paths and protocols which, taken together, emulates (or is) a single Ethernet LAN.

  The LAN may be either point-to-point or multipoint-to-multipoint (shared medium).

  Of course, a point-to-point LAN may create a single point of failure.

  The LAN may be attached to one or many Provider Bridges in a single Island.



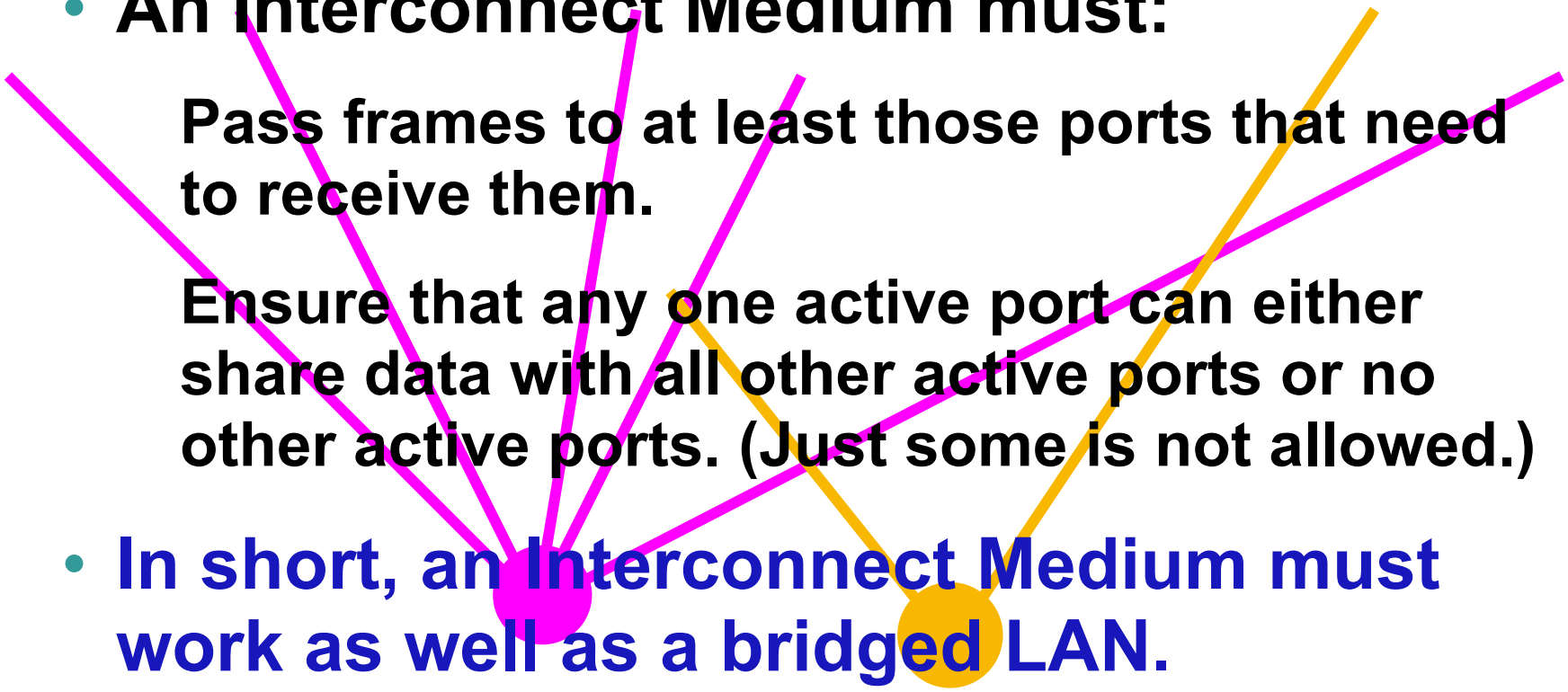
What is an Interconnect Medium?

- **An Interconnect Medium must:**

  Pass frames to at least those ports that need to receive them.

  Ensure that any one active port can either share data with all other active ports or no other active ports. (Just some is not allowed.)

- **In short, an Interconnect Medium must work as well as a bridged LAN.**



What is an Interconnect Medium?

- **Examples of Interconnect Media:**

  An 802.3 10G Ethernet connection.

  An 802.17 Resilient Packet Ring.

  An emulated LAN consisting of a full mesh of Pseudowires implemented over MPLS.

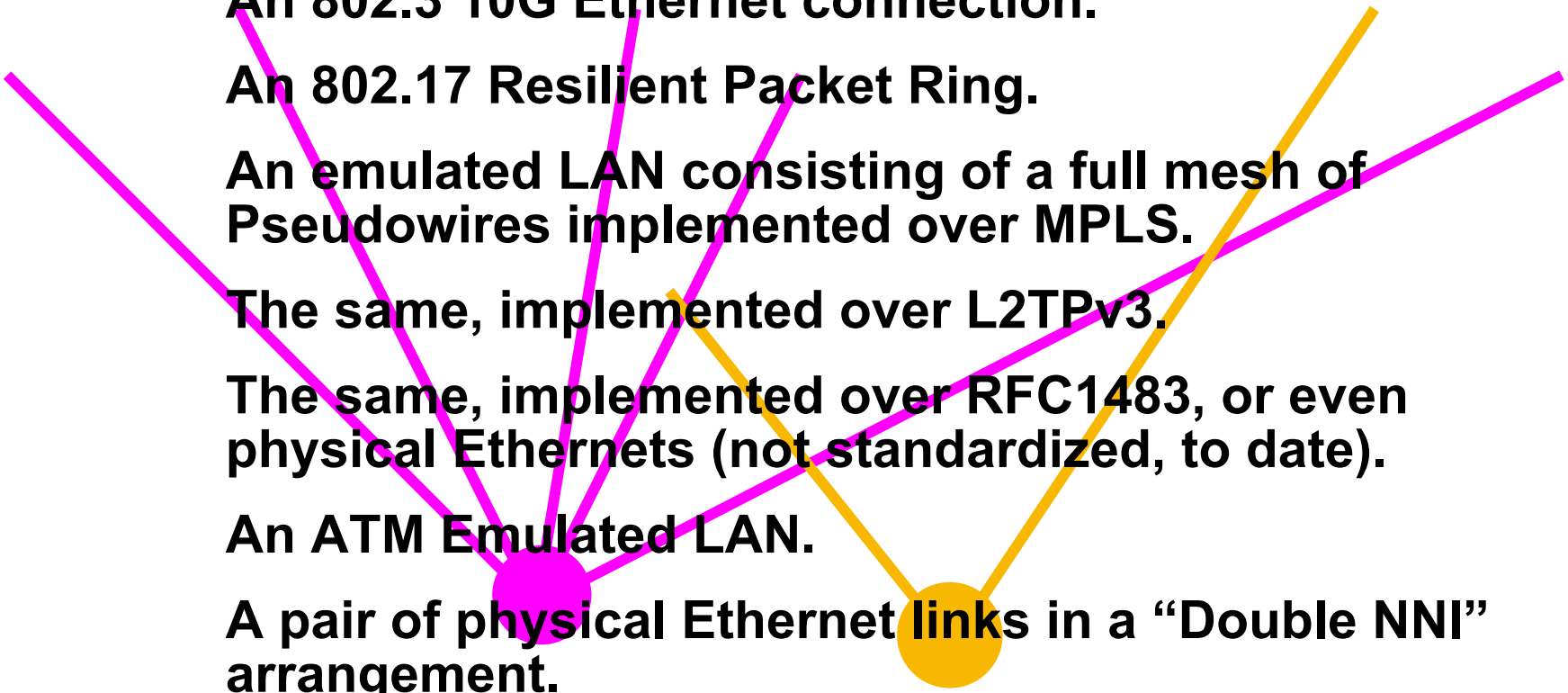
  The same, implemented over L2TPv3.

  The same, implemented over RFC1483, or even physical Ethernets (not standardized, to date).

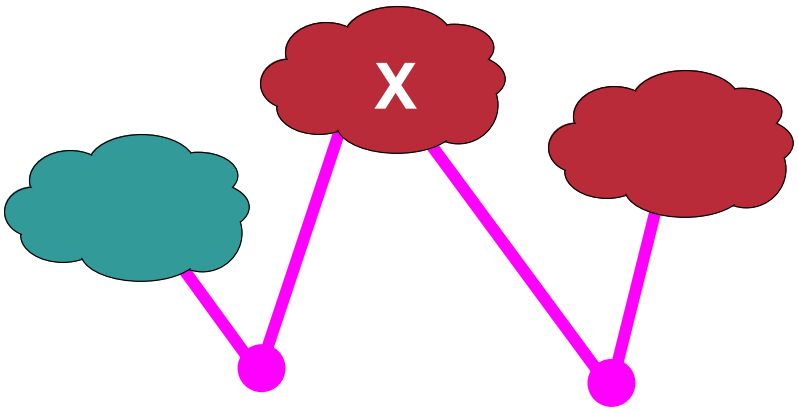
  An ATM Emulated LAN.

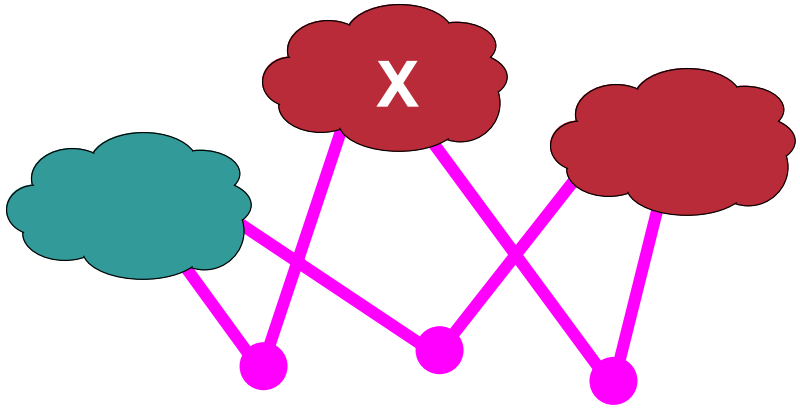
  A pair of physical Ethernet links in a "Double NNI" arrangement.

  An H-VPLS Hierarchical mesh.



One more element is required

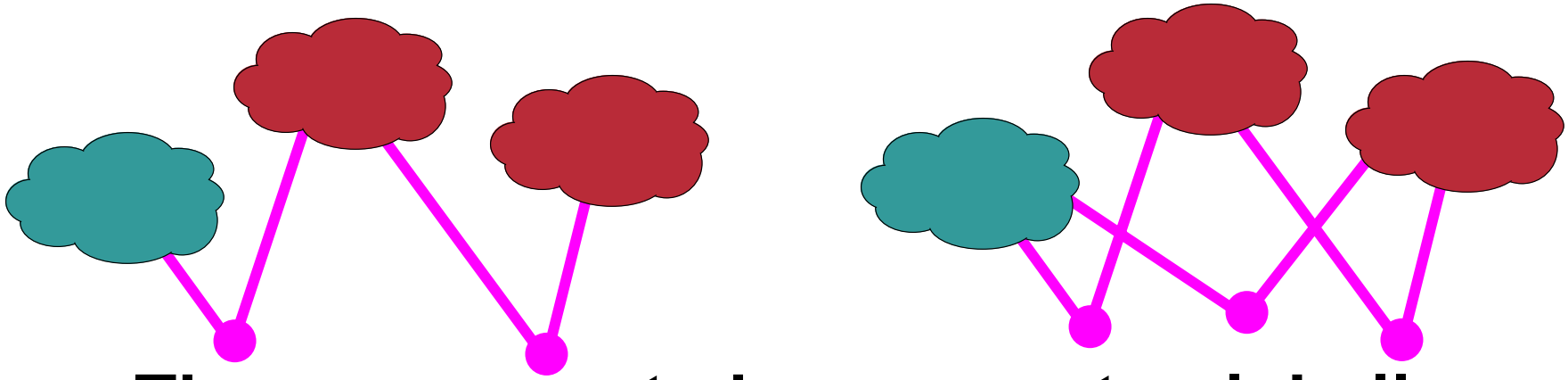




- This interconnect would be OK.
- This interconnect would loop.

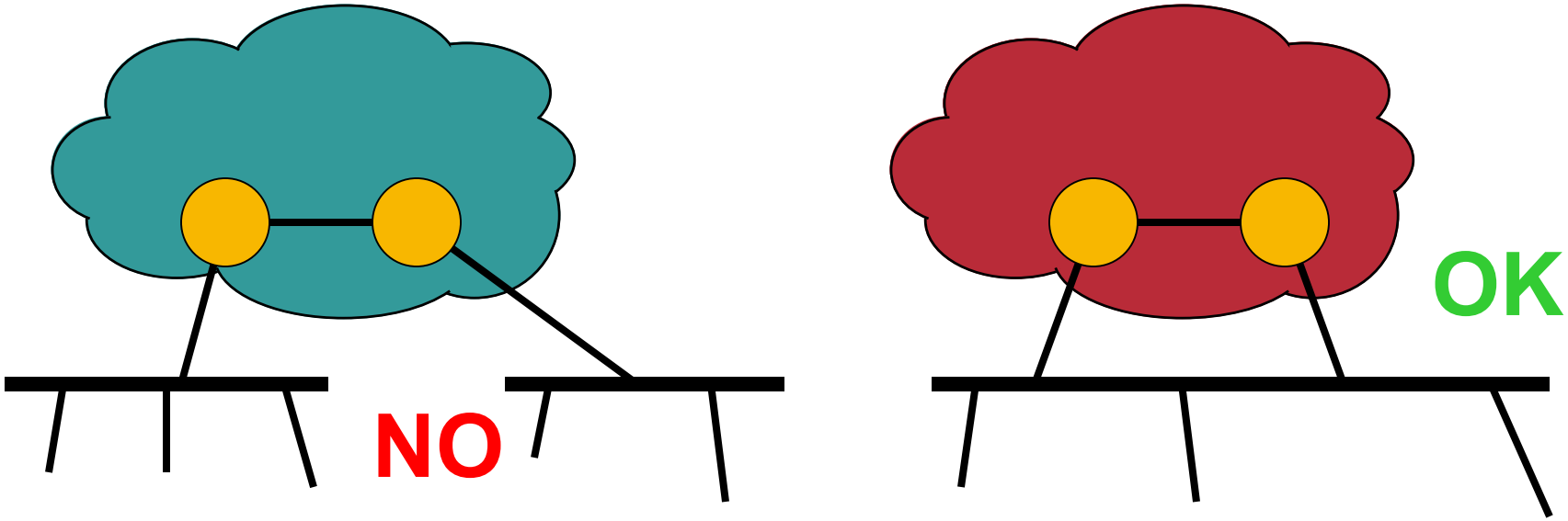
- But how does X tell which configuration it has, unless it runs a protocol that spans all Islands, i.e., spans the world?

One more element is required



- There appears to be no way to globally enforce a “no loops” rule without a global spanning tree.
- Therefore, each Island enforces a stricter rule: “Only one connection to a separate IM per Customer Service Instance”. Both configurations are outlawed!

In other words,



- **No Island can relay data from one Interconnect Medium to another.**

This does NOT prevent redundant connections to *one* IM for reliability!

What is an Island?

No man?

- **Within a given Island, certain matters are strictly local:**

  Forwarding data among UNI ports and between UNI ports and Interconnect Media.

  Avoiding forwarding loops.

  Attaching at most one Interconnect Medium to a given Customer Service Instance.

  Making sure that only one port is available to a Customer frame passing between the Interconnect Medium and the Island.

What is an Island? Examples:

A network of IEEE 802.1AD Provider Bridges (Q-in-Q), including the case of a single bridge.

A MAC-in-MAC network.

A gateway to a Frame Relay cloud.

A Lasserre-VKompella PE-rs/MTU cloud.

A Router, File Server, or other L2 endstation ***directly attached*** to the provided L2 service.

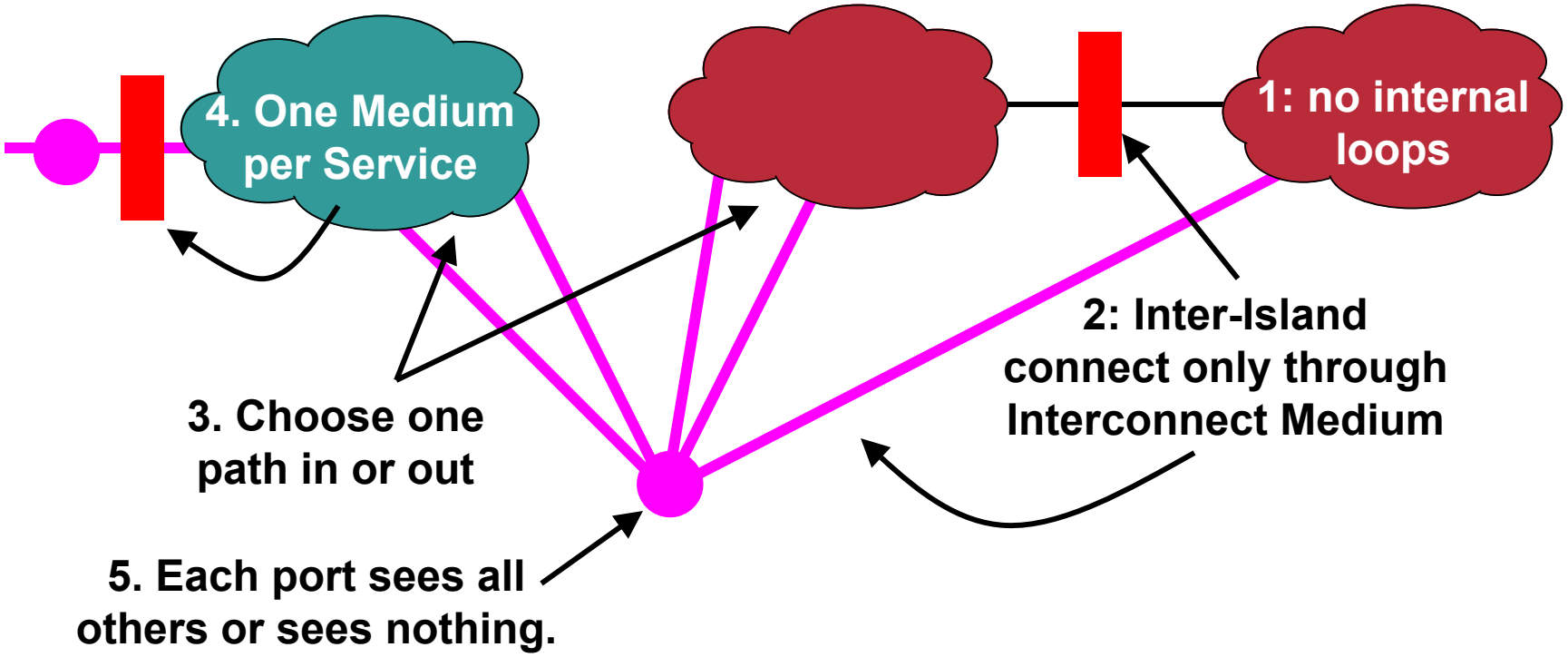
Similar Customer Equipment that is trusted by, and integrated into, the Provider's network.

How do you know there are no global forwarding loops? The Five Rules:

1. Each Island is responsible for preventing internal forwarding loops.
2. Islands connect to other Islands only through Interconnect Media.
3. Each Island ensures that no customer data frame passes through more than one Interconnect Medium attachment into or out of the Island.
4. Each Island ensures that it attaches any given Customer Service Instance to no more than one Interconnect Medium.
5. An Interconnect Medium ensures that if an attached port can talk to ***any*** other attached port, it can talk to ***all*** of the ports attached to that Medium.

How do you know there are no global forwarding loops?

- Thus, the global connectivity plan for one Customer Service Instance must look something like:



Oh, yes! The Customers!

- By definition, Customer Ports (UNIs) are present only in Islands.
- Trusted Customer Equipment, e.g. a Router, might be attached directly to an Interconnect Medium, but it would then be Provider Equipment, not a normal Island UNI.
- Yes, a “Customer” could own and operate an Island. But, to the protocols, it is still a Provider Island.

Just in case you missed it!

- Those Five Rules were the whole point of this presentation!
- Islands can be built with many different technologies.
- Interconnect media can be built with many different technologies.
- Using Islands, Interconnect Media, and the Five Rules, Ethernet can span the world!
- And, no global routing protocol is required!

So, what's the big deal?

- **Trivially obvious to the most casual observer?**

  If there is no forwarding, there are no loops!

  If there is only one path, there are no loops!

- **Duh!**
- **The subtlety:**

  Although there are no loops, redundant paths are still provided; reliability is maintained.

  The Islands are totally independent with regard to loop avoidance protocols.

Part 2: Building Islands

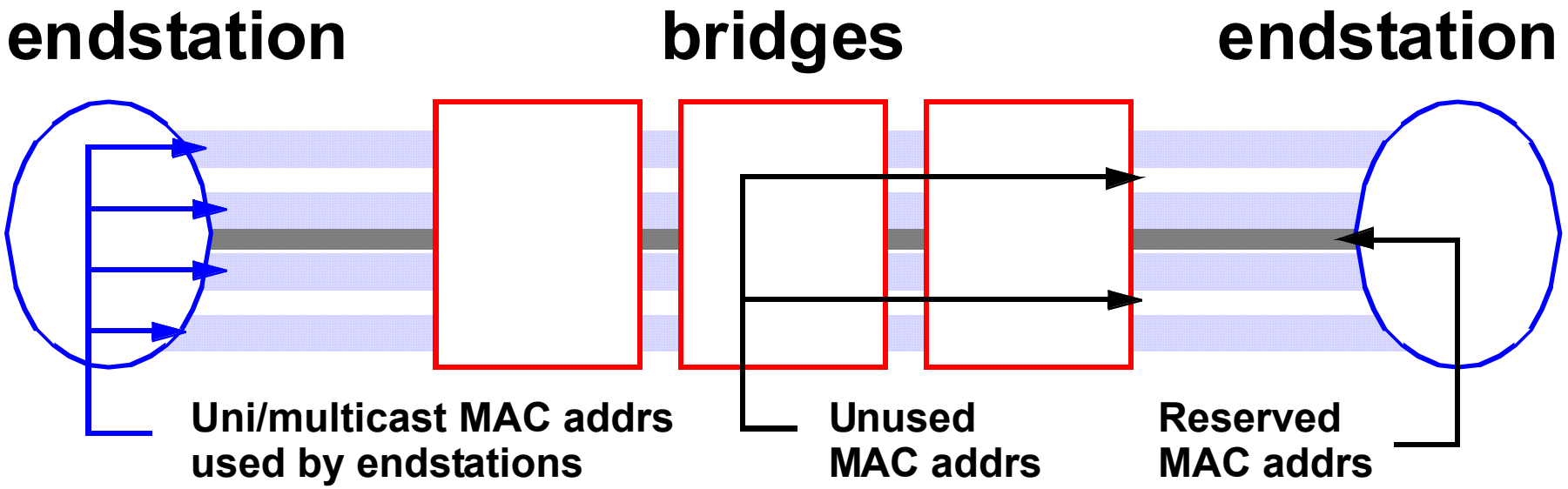
One simple way to build an Island

- **A network of IEEE 802.1AD Provider Bridges.**

  A PB is much like a normal 802.1Q VLAN bridge, but uses a different set of BPDU/Control MAC addresses, and a different EtherType for its VLAN tags.

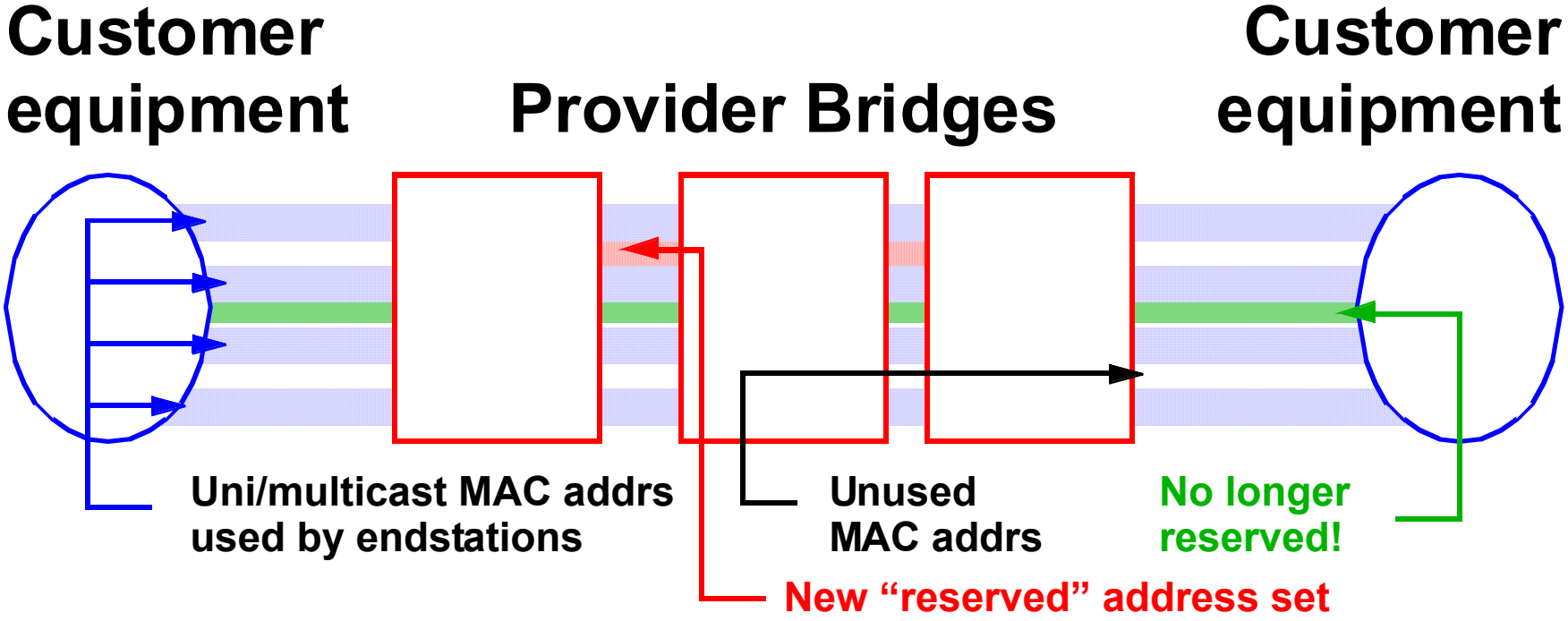
  Each Provider VLAN (P-VLAN) tag corresponds to one Customer Service Instance.

IEEE 802.1D/Q Bridges



- 802.1Q bridges are ***transparent*** to endstations' traffic, ***except*** for 802.1Q tag addition/removal, and ***except*** for the band of 33 reserved multicast MAC addresses.

IEEE 802.1AD Provider Bridges



- **If the Provider Bridges use a new set of VLAN tags and MAC addresses, then they are transparent to Customer Equipment *even* if CE is a bridge!**

A less simple way to build an Island

- **A MAC-in-MAC Network.**

  A “MAC-in-MAC” Edge Bridge encapsulates each Customer frame in a wrapper whose outer source and destination MAC addresses refer to the edge bridges, or even the specific UNI ports.

  Wrapper identifies the Customer Service Instance and other useful information.

  A MAC-in-MAC Island can probably be larger than a single Q-in-Q Island.

More ways to build an Island

- **Anything that passes Customer data frames through an attachment to an Interconnect Medium, and follows the rules for avoiding large-scale loops.**

Connections within Islands

- One may connect Provider Bridges (or MAC-in-MAC, for that matter) within an Island using point-to-point Ethernets.

  We would expect that to be the normal case.

- But a Pseudowire is equivalent to, and can be used just like, a physical Ethernet.
- Even an Emulated LAN consisting of a full mesh of Pseudowires can serve within an Island, rather than as an IM.

Restricting multicast distribution

- **Imagine an Island with 4000 Customer Service Instances = 4000 P-VLANs. How do you prevent each broadcast, multicast, or unknown unicast from reaching every bridge in the Island?**
- **Three standard answers:**

  Configure ports to restrict P-VLAN distribution.

  Use GMRP (or IGMP snooping) to restrict multicast distribution.

  Use GVRP to restrict P-VLAN distribution.

Restricting multicast distribution

- **Configuring P-VLAN distribution**

  Time consuming and error-prone. What happens when a bridge or link fails?

- **GMRP/IGMP snooping for multicasts**

  Takes effort to snoop, especially if Customer C-VLANs are taken into account.

- **GVRP for P-VLAN distribution**

  GVRP is much heavier than RSTP.

- **Which will scale up to 4K VLANs? TBD**

Scaling GVRP to 4K VLANs

- In GVRP, it takes 4 bytes per P-VLAN to transmit joins or leaves.
- It takes 12 GVRP PDUs to join 4094 VLANs.

  Perhaps a bit vector structure in the PDUs could reduce this to 1 or 2 PDUs.

- GVRP leaves require a timeout, even on point-to-point links.

  On a point-to-point link, GVRP should be configurable to believe a single leave.

Part 3: Building Interconnect Media

Emulated LAN: Full-mesh split-horizon Pseudowires

- A “Pseudowire” is a point-to-point, bidirectional tunnel for carrying Ethernet frames.

  A Pseudowire is a Layer 3 tunnel; the Ethernet frames are packed inside an IP Packet.

  To the best of its ability, a Pseudowire carries every frame to the other end.

  A Pseudowire does not care about MAC addresses or VLAN tags. It’s just a wire.

  Pseudowires are defined for carrying Ethernet over MPLS and over L2TPv3.

Emulated LAN: Full-mesh split-horizon Pseudowires

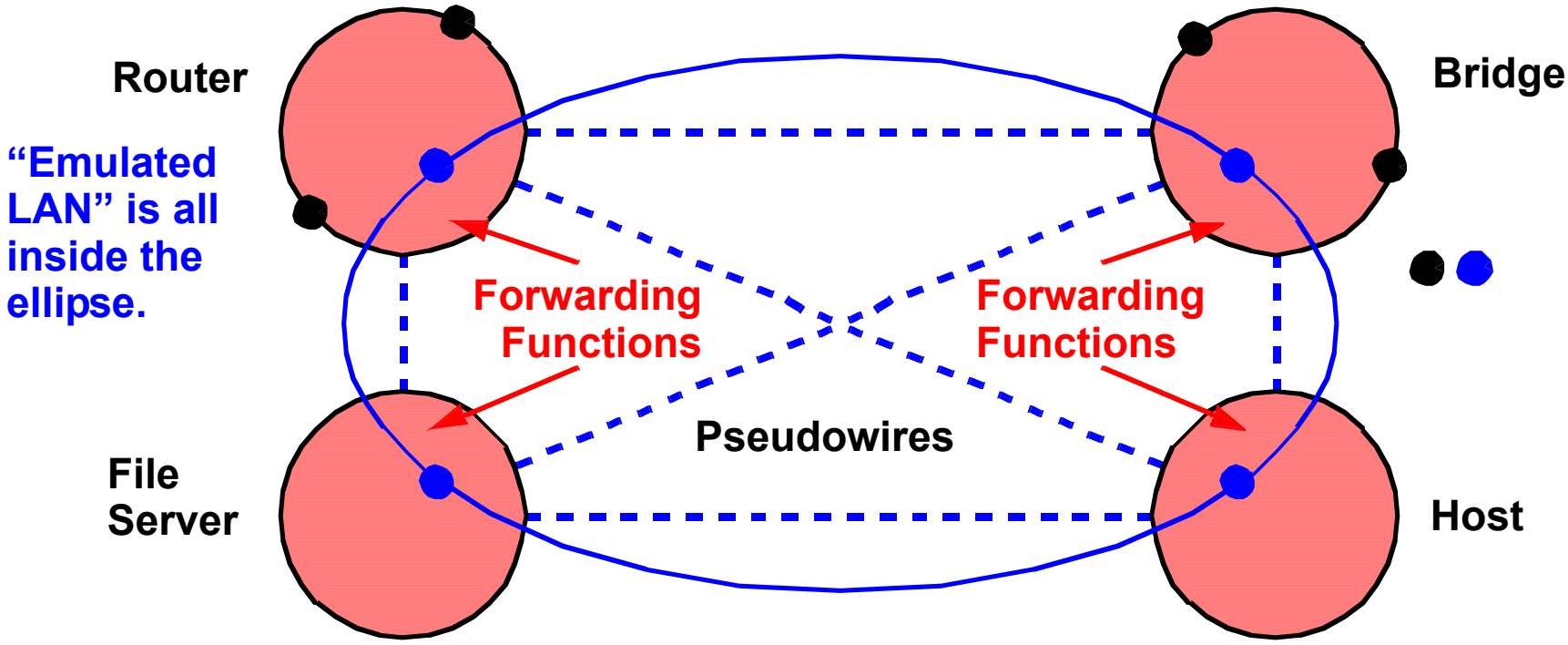
- Each device connects to the Emulated LAN through a “Forwarding Function” (See also L2VPN Requirements).

  FF creates a full mesh of Pseudowires among all of the participants in the Emulated LAN.

  FF learns associations between MAC addresses and Pseudowires.

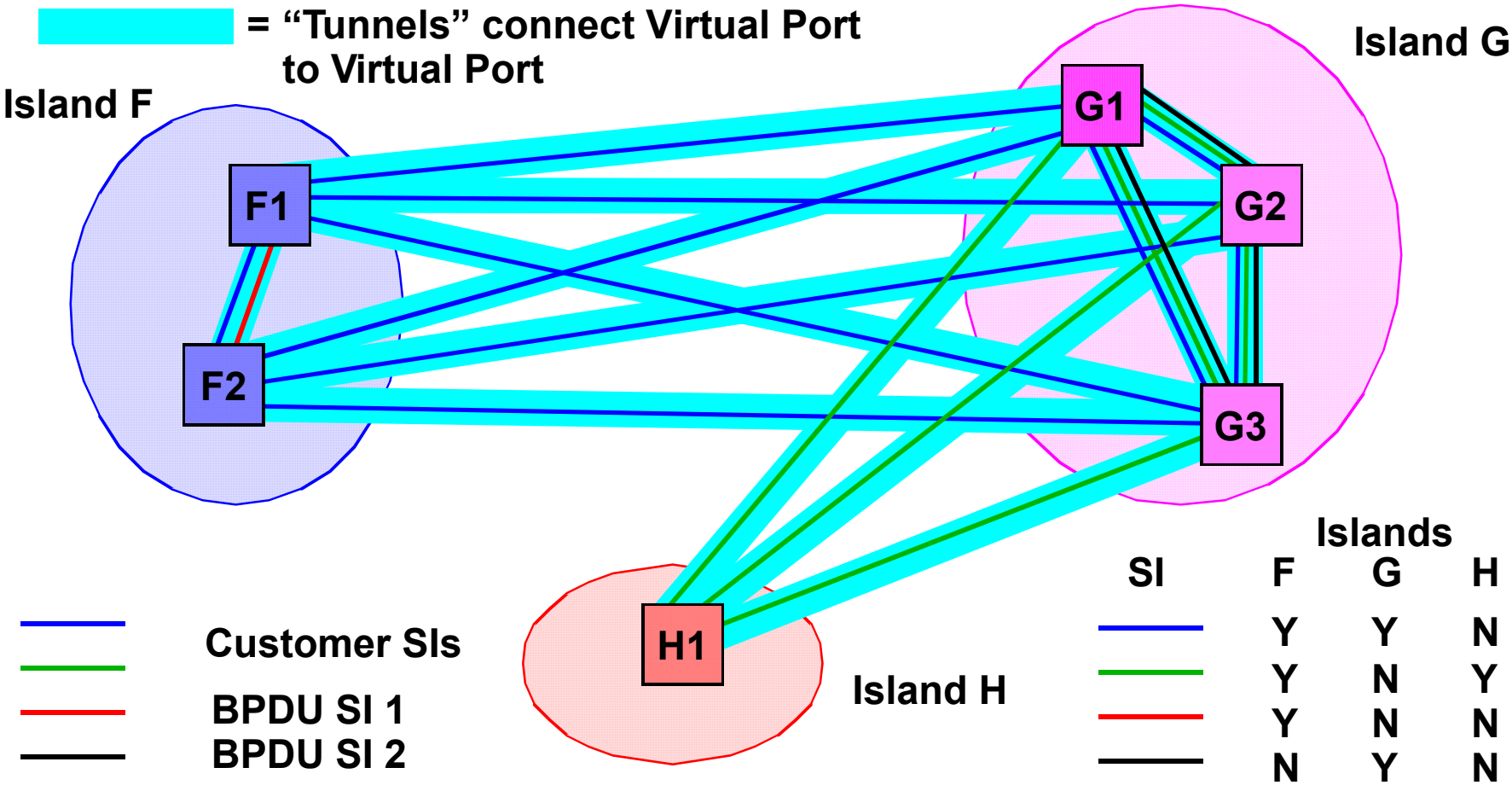
  FF forwards multicasts and unicast floods to all Pseudowires, known unicasts to just to the one it has learned is the right one.

Emulated LAN: Full-mesh split-horizon Pseudowires



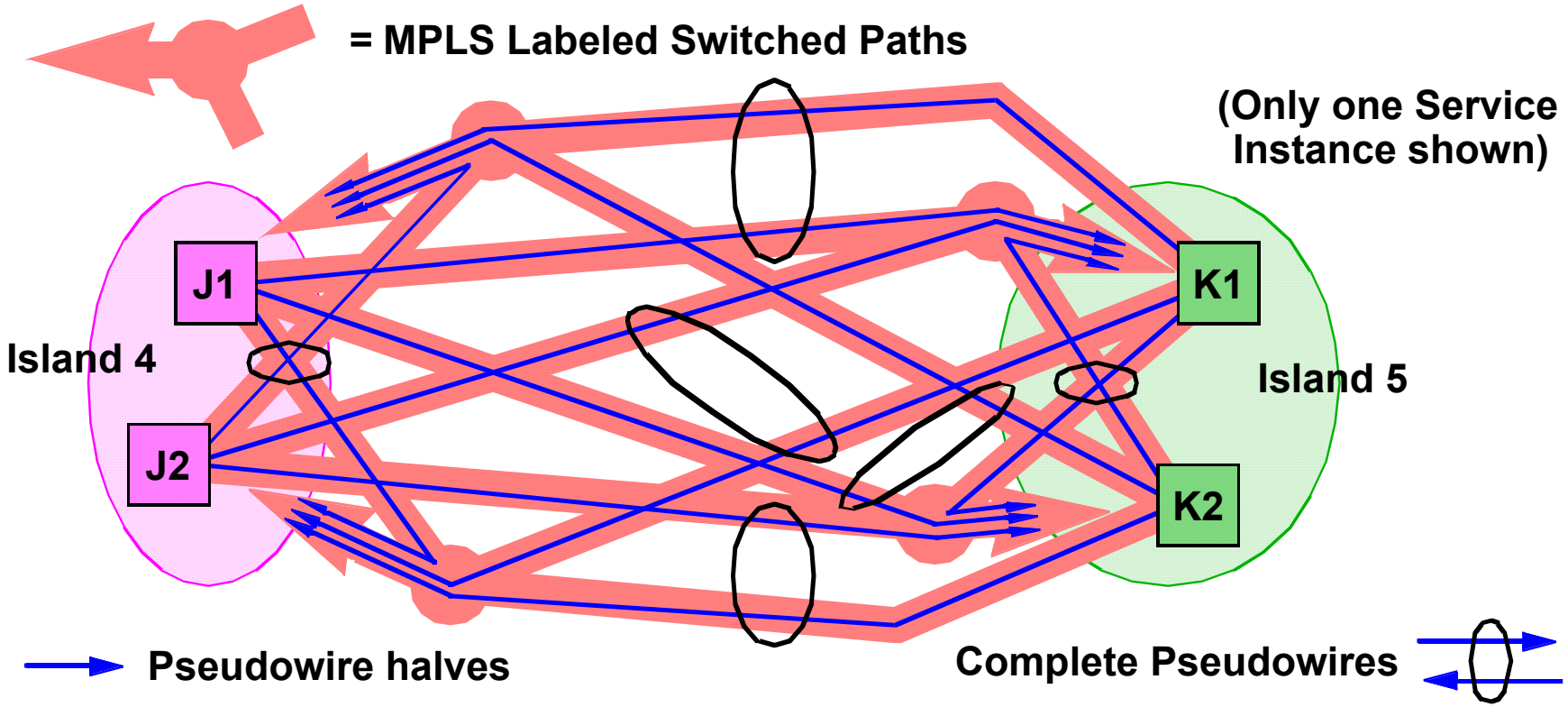
- Emulated LAN looks like a shared medium Ethernet (through •)to the upper layers.

Emulated LAN: Full-mesh split-horizon Pseudowires



- **Many Pseudowires can share each tunnel.**

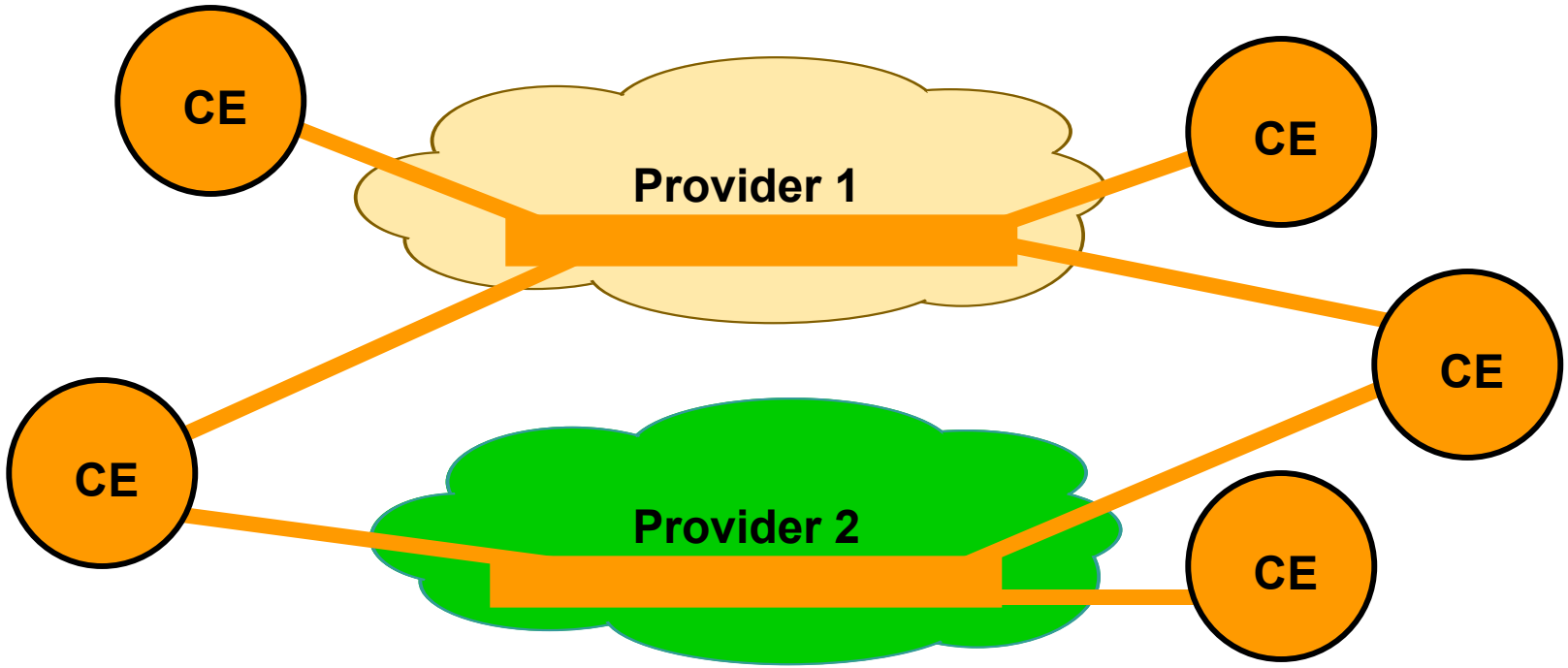
Emulated LAN: Full-mesh split-horizon Pseudowires



- **Accurate view of one Customer Service Instance's Pseudowires over MPLS**

(Assuming no Pseudowire-based load sharing)

Requirements for any LAN Emulation Technology



- **Emulated LANs will be carrying BPDUs for at least some Customers' spanning trees.**

Requirements for any LAN Emulation Technology

- **To carry Customers' BPDUs, an Emulated LAN must guarantee, to a high degree of probability, that any BPDU transmitted into one attachment port is received on all other attachment ports.**

  This guarantee can be violated for a "short time", if the Customer Equipment's timing constants are suitably adjusted.

  Therefore, the violation time must be bounded.

Requirements for any LAN Emulation Technology

- We know, through experience, that one bridged LAN can meet these requirements, and can carry another bridged LAN's data and Spanning Tree BPDUs, as long as the timing parameters in the networks are compatible.
- However, if a physical shared medium in the inner network is replaced by an Emulated LAN, the inner network's requirements are imposed upon the embedded Emulated LAN.

What connectivity scenarios must be avoided?

- **Full mesh failure is defined as:**

  Two or more Forwarding Functions' ifOperStatus == Up.

  Those FFs do not have a full mesh of operational, bidirectional Pseudowires.

  This condition persists for more than *X* milliseconds, where *X* is (roughly) equal to the Hello time of the Spanning Tree Protocol using the Emulated LAN.

Why is a full-mesh failure bad?

- Some connectivity is lost, of course.
- If a Pseudowire is unidirectional:

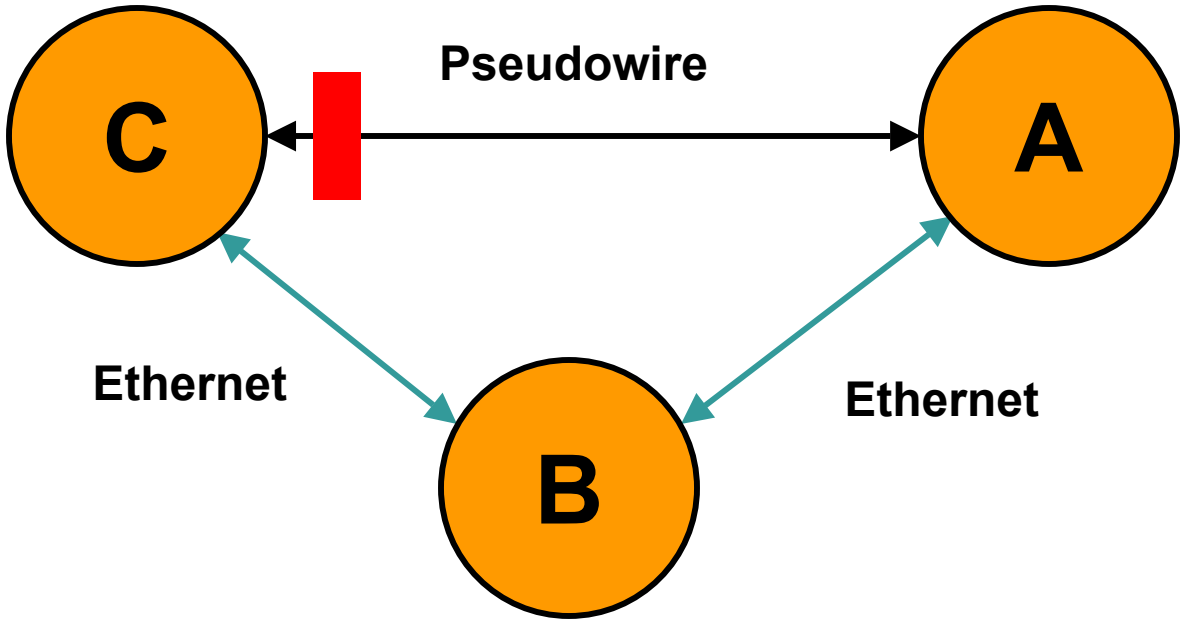
  Bridges at both ends may forward data.

  This causes a loop, leading to a broadcast storm.
- If a Pseudowire is missing:

  Two Bridges may both forward data, though not to each other.

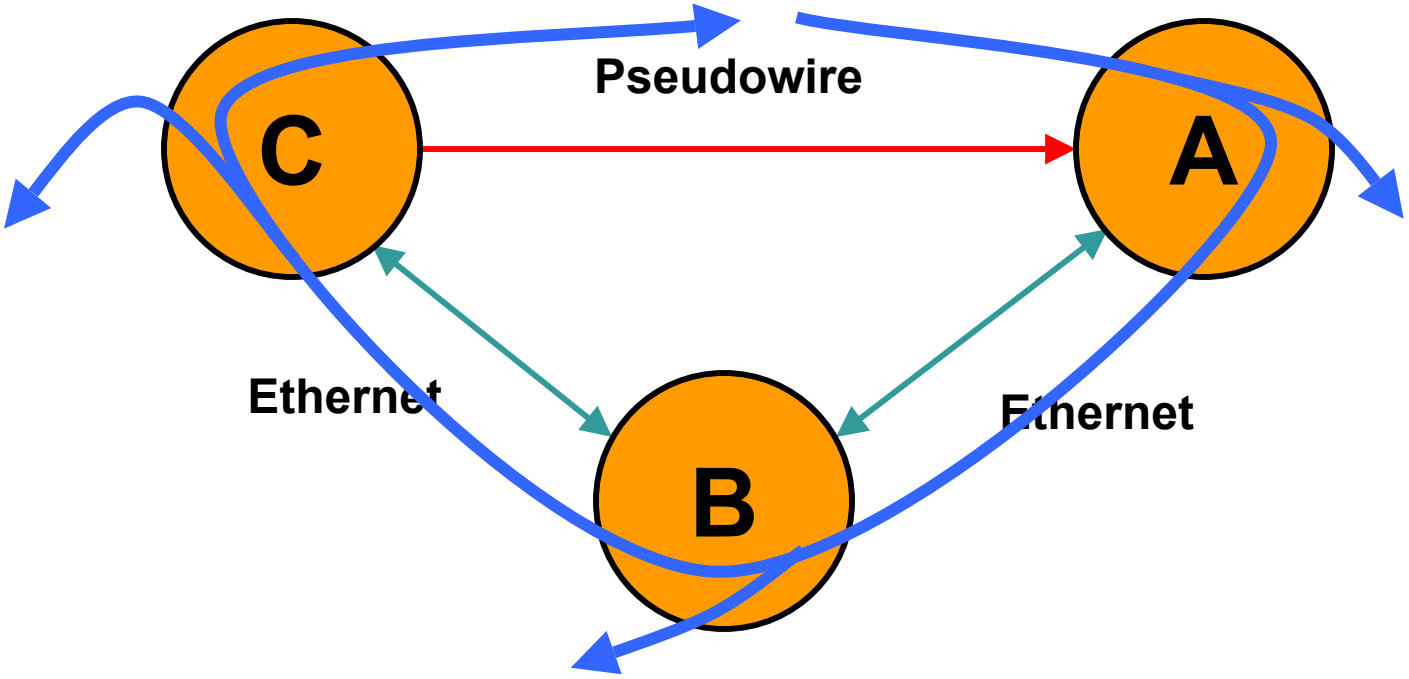
  So, frames may be delivered multiple times.

Unidirectional Pseudowire



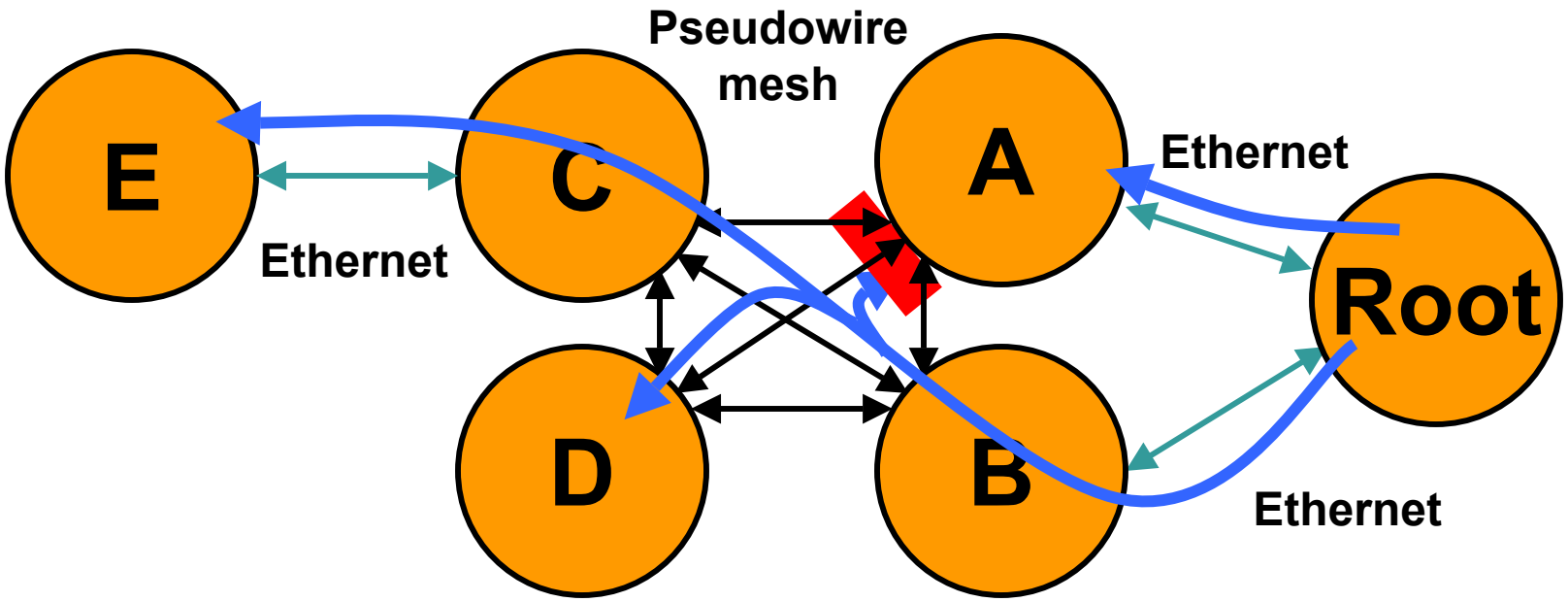
- **Normally, Bridge C blocks the Pseudowire port. No loops.**

Unidirectional Pseudowire



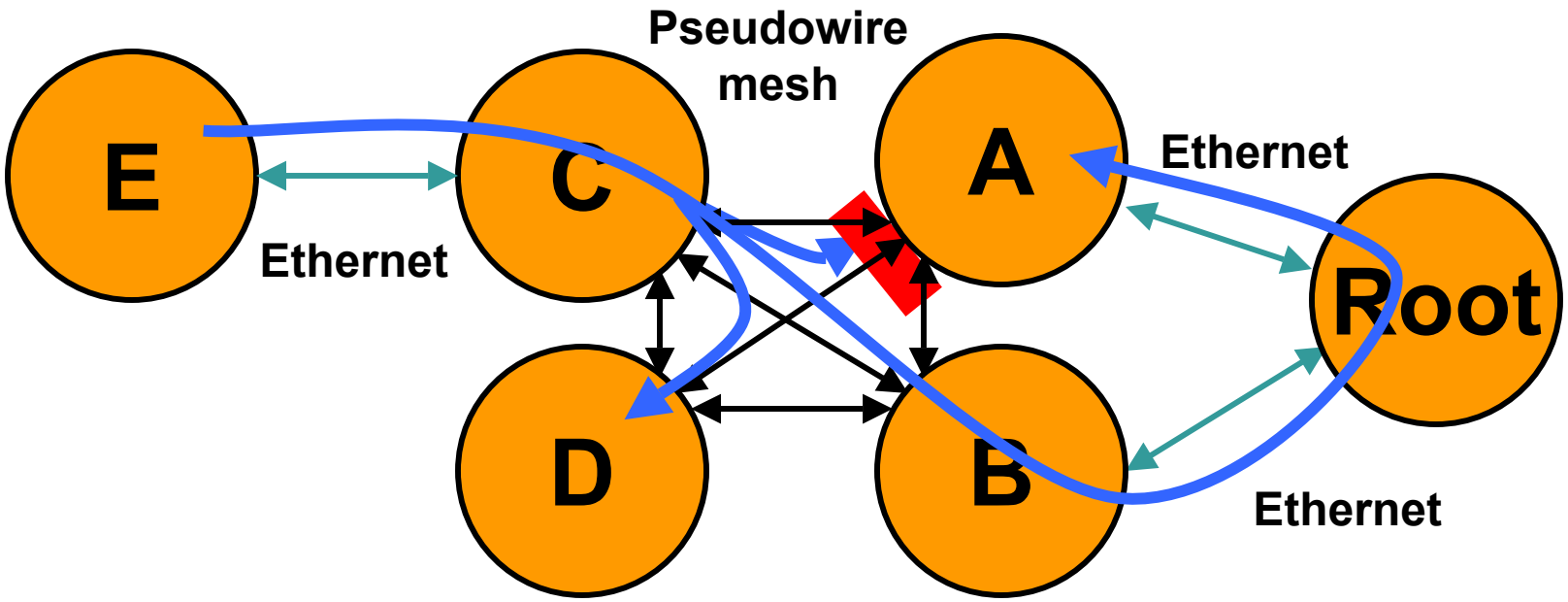
- **Unidirectional Pseudowire prevents C from seeing A's BPDUs, so C forwards data to Pseudowire. We have a storm!**

Missing Pseudowire from mesh



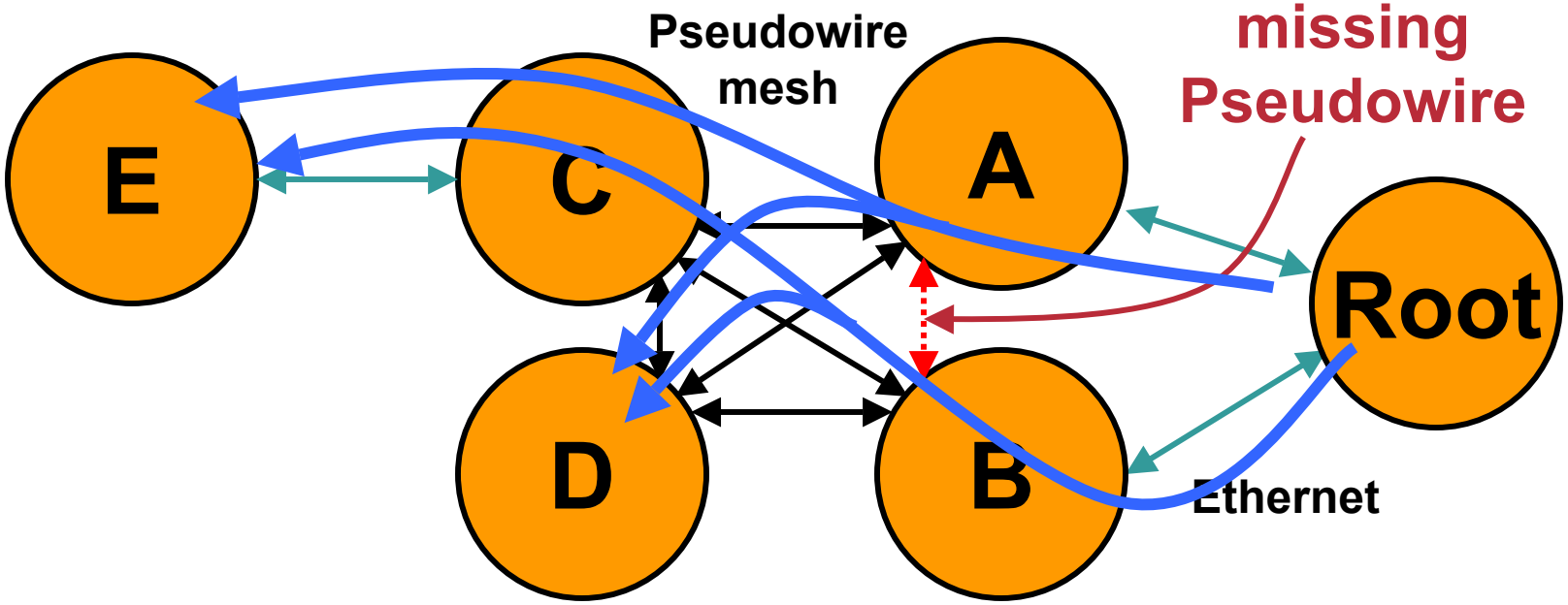
- **Normally, Bridge A blocks its Pseudowire port. No loops on frames sent from the Root.**

Missing Pseudowire from mesh



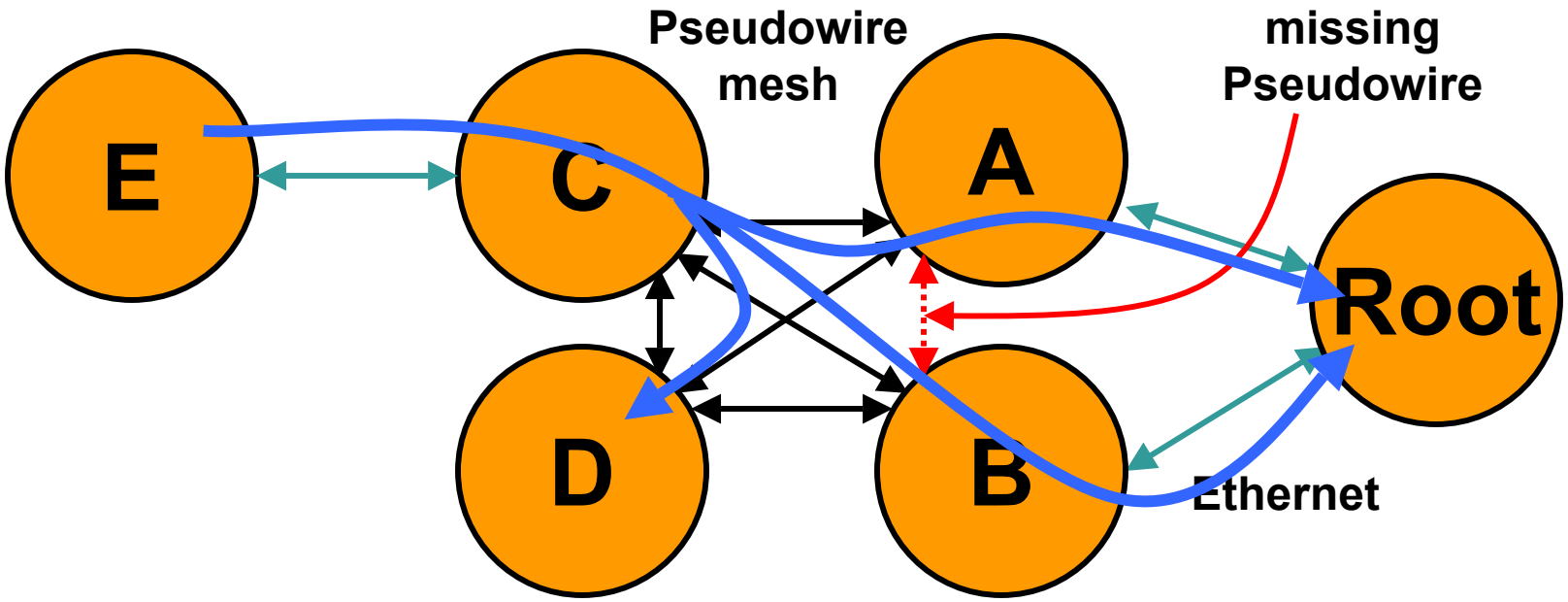
- And, no loops on frames sent from E.

Missing Pseudowire from mesh



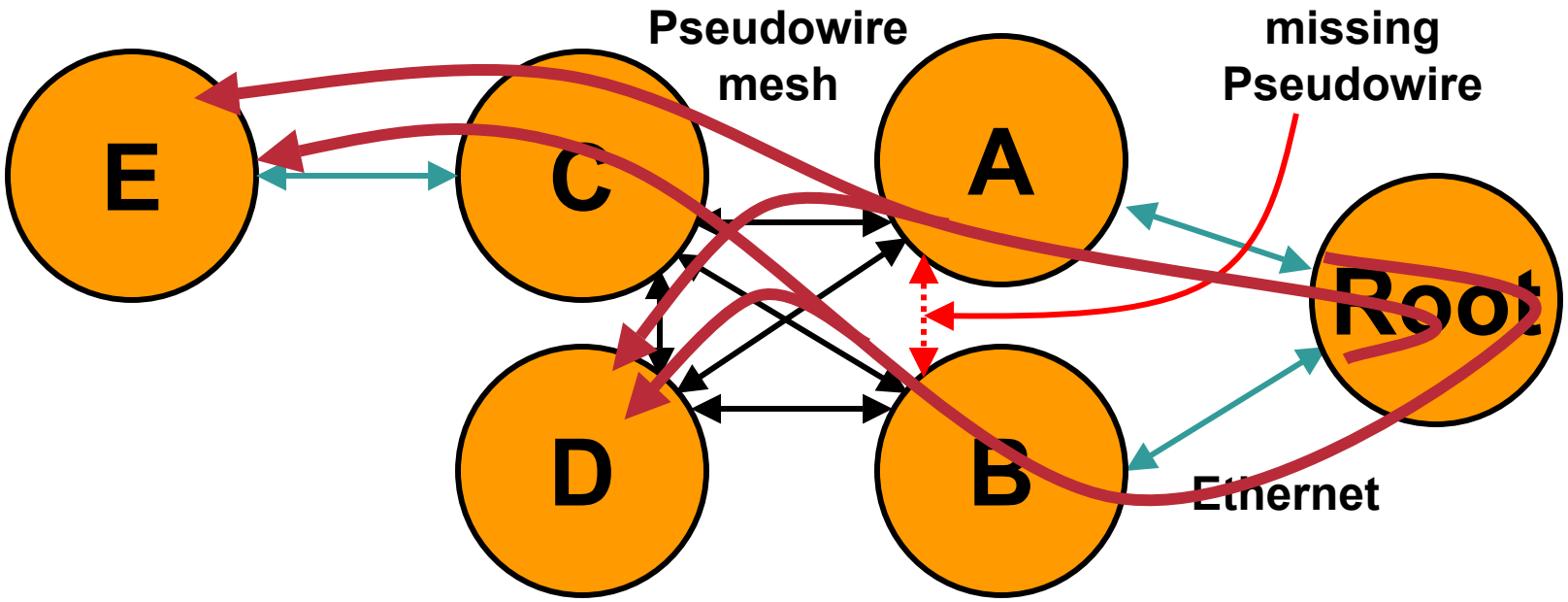
- If A and B cannot see each others' BPDUs, both forward to the full mesh. C, D, and E see two copies of every multicast or unknown flood from the Root.

Missing Pseudowire from mesh



- Even worse is any frame, even a normal unicast, sent from E. The frame is delivered twice to the Root.

Missing Pseudowire from mesh



- Then, the Root reflects both copies back!
- Bridges E, A, B, and the Root each see 2 copies of the frame. Bridges C and D each see 3 copies!
- Also, the bridges learn the source address on the wrong ports, because of the reflection!

Avoiding Full-Mesh Failures

- **Fast Pings can detect unidirectional links, the worst failure.**

  A unidirectional link can and must be quickly converted to a fully-failed link.

  BUT: LSP Pings *may* not follow same path as Pseudowire Pings; routers may load-share on a per-Pseudowire basis.

  This would require per-Pseudowire pings.

- **Therefore, one may trade load-sharing flexibility against Ping overhead.**

Avoiding Full-Mesh Failures

- **Fully failed links will usually be caused by:**

  The failure of a node containing a Forwarding Function, which is perfectly acceptable.

  The failure of an intermediate router, which means that the link will be reestablished in short order.

- **A fully failed link does not initiate a broadcast storm; brief outages to restore routing connectivity *may* be acceptable to some Providers and Customers.**

Avoiding Full-Mesh Failures

- **Theoretically, a missing Pseudowire (a fully failed link) is not possible, except during membership transitions.**

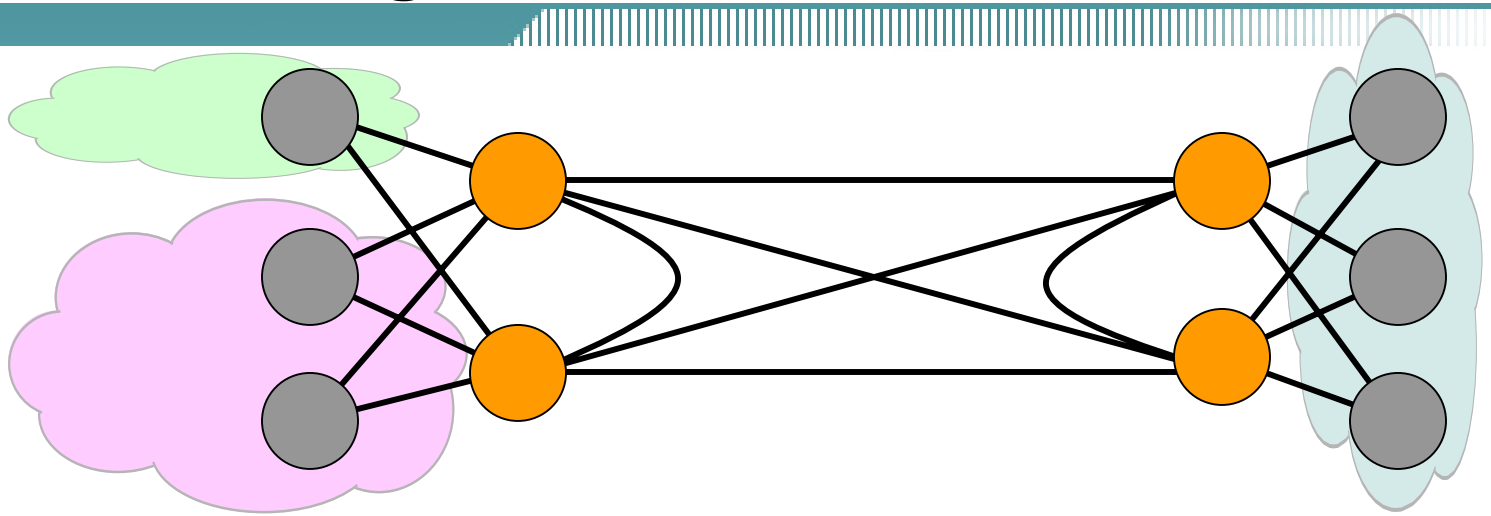
  If A can reach B and B can reach C, then A must be able to reach C (through B!).

  However, an erroneous Access Control List (ACL) in a router could prevent the operation of one Pseudowire.

- **It would be very nice to have a mechanism to (at least) recognize and report such a failure.**

  Ideally, it would have the optional capability to disconnect a node(s) to prevent duplicate deliveries.

What other Interconnect Media technologies are available?



- **H-VPLS Meshes**

  Although designed for edge-to-edge services, an H-VPLS Mesh could equally serve as an Interconnect Medium.

  H-VPLS offers a means for scaling an Interconnect Medium to sizes larger than a full mesh can support.

What other Interconnect Media technologies are available?

- **ATM LAN Emulation almost works,**

  But, its timing parameters are out of date; they would have to be updated for ATM LANE to carry Customers' RSTP BPDUs.

- **A MAC-in-MAC backbone would work.**

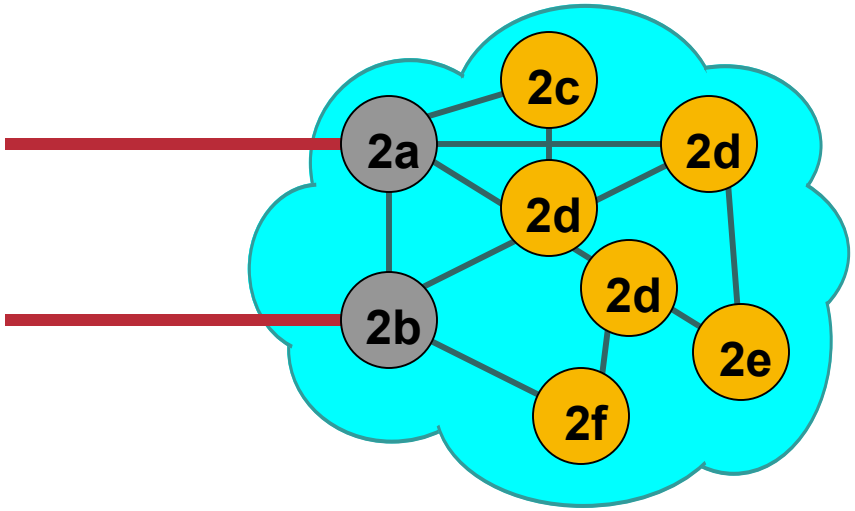
  But, it cannot scale to Layer 3 sizes.

  Further definition of the required spanning tree interactions is needed.

- **Multiple different Interconnect Media technologies can interoperate, as long as everyone obeys The Five Rules.**

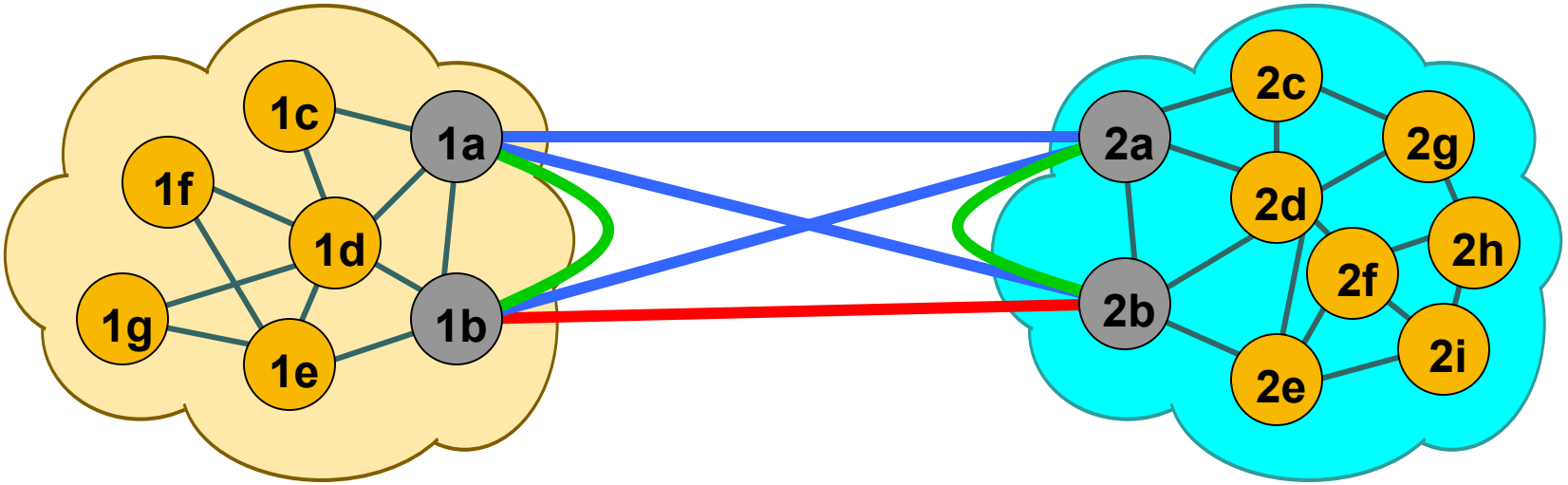
Part 4: Attaching Islands to Interconnect Media

Rule 3: No Customer data frame goes in or out through two IM attachments.



- How do Provider Bridges 2a and 2b cooperate in their use of the physical links to other Islands (shown in red)?

Why bother with a full mesh?

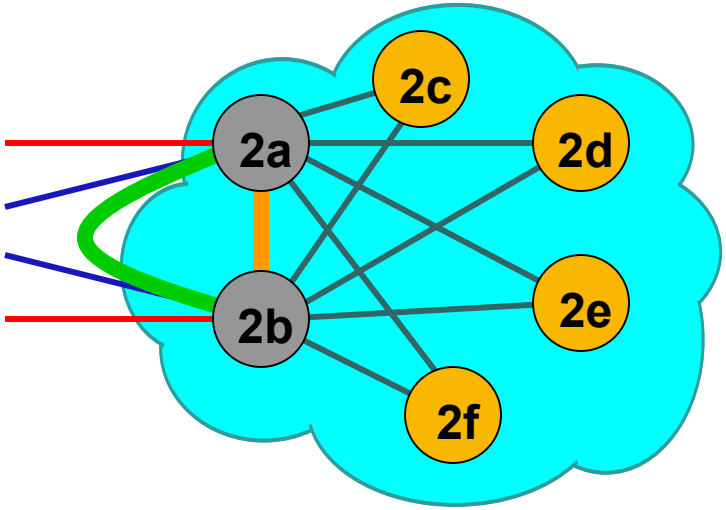


Minimal connectivity

Efficient unicast reachability

Flexibility and emergency

Why a full mesh?
Two POPs, many two-link edge bridges.



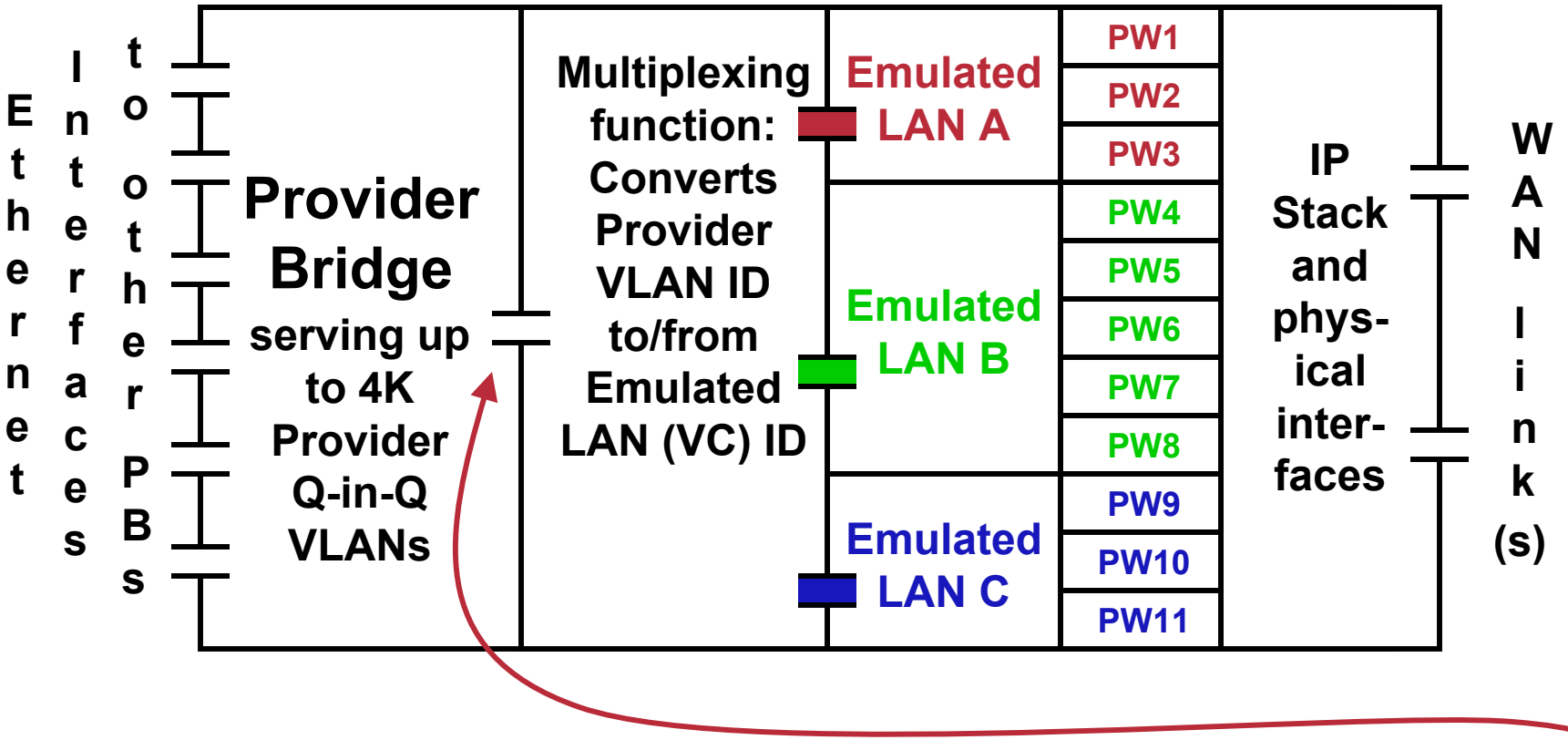
- What happens if the orange Ethernet fails?

  UNI on 2c no longer can reach UNI on 2f?

  Edge bridge 2d transports all 2a-2b traffic?

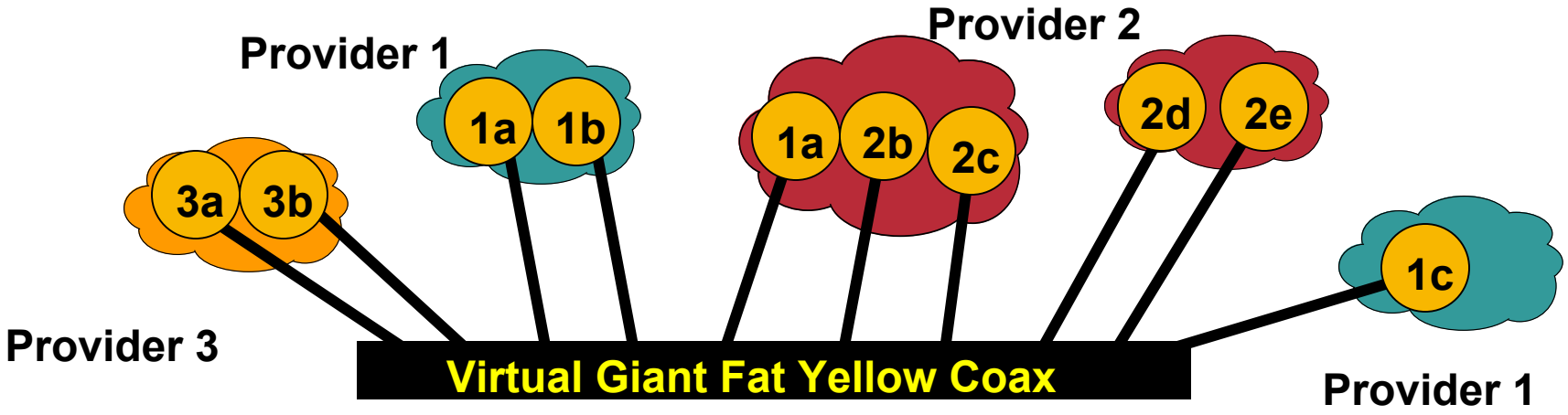
- Perhaps better: the green Pseudowire replaces the orange Ethernet.

Interior of a Provider Bridge



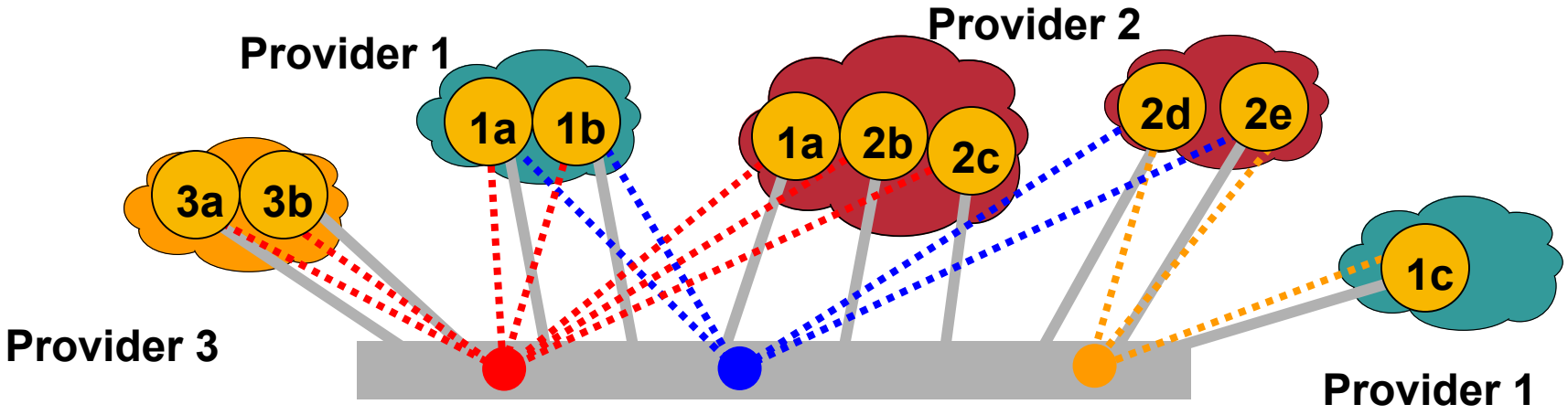
- Inside a Provider Bridge, there is ***one*** Bridge Port to the entire Layer 3 world!

Provider Bridges' view of the World



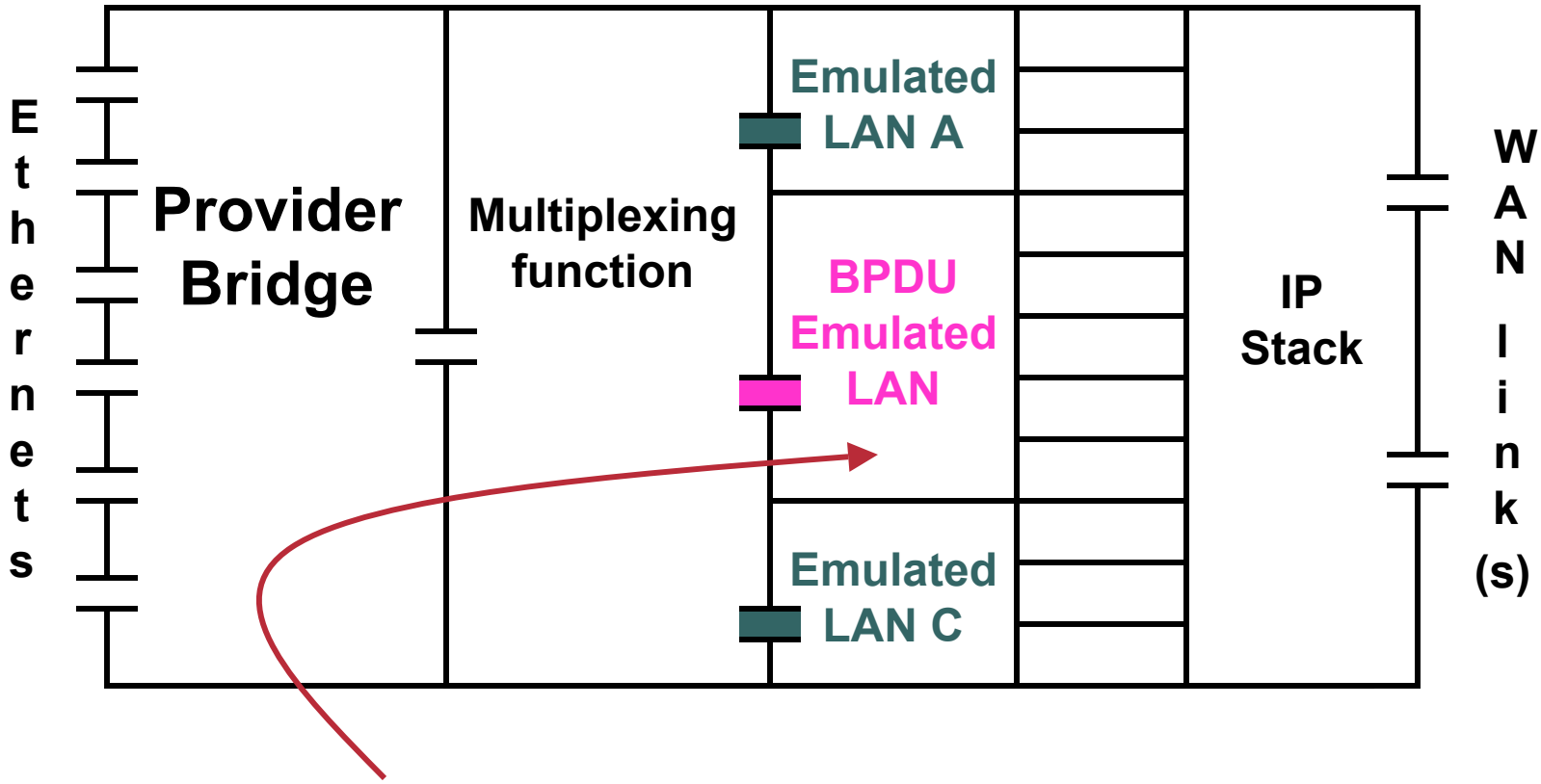
- So, Provider Bridges think that the world looks like this.
- One giant shared-medium Ethernet carrying all VLANs (Customer Service Instances).

Provider Bridges' view of the World



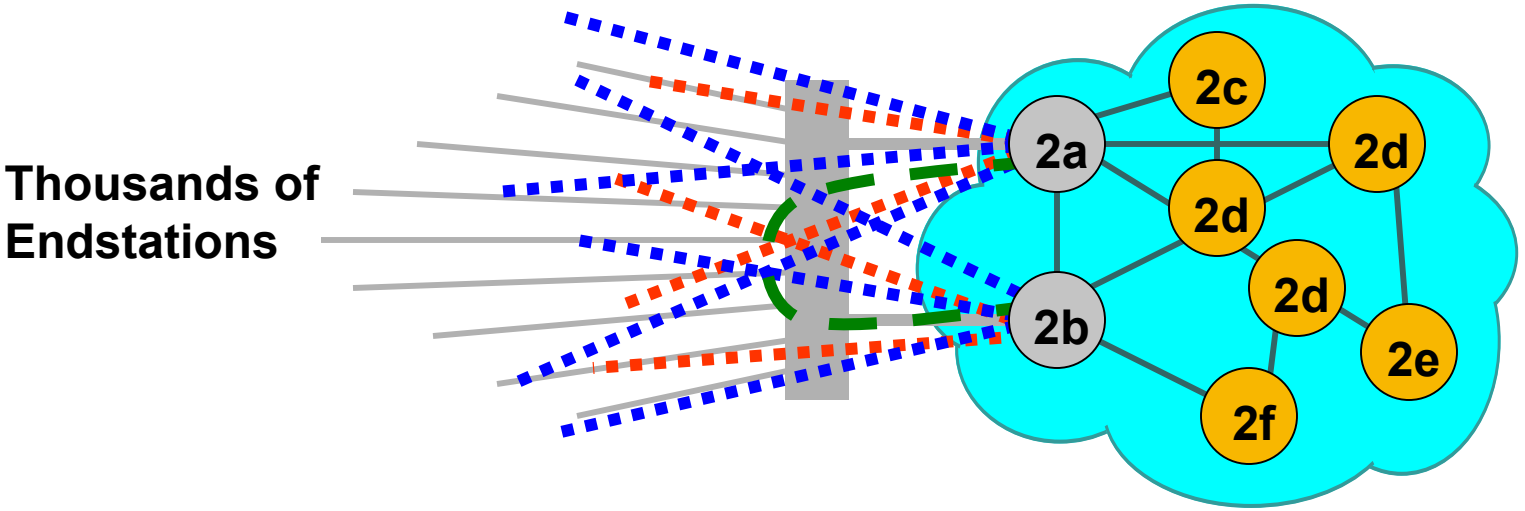
- **Service Instances (Interconnect Media) are overlaid on the single Virtual Fat Yellow Coax, just like VLANs are overlaid on a single physical shared medium.**

Interior of a Provider Bridge



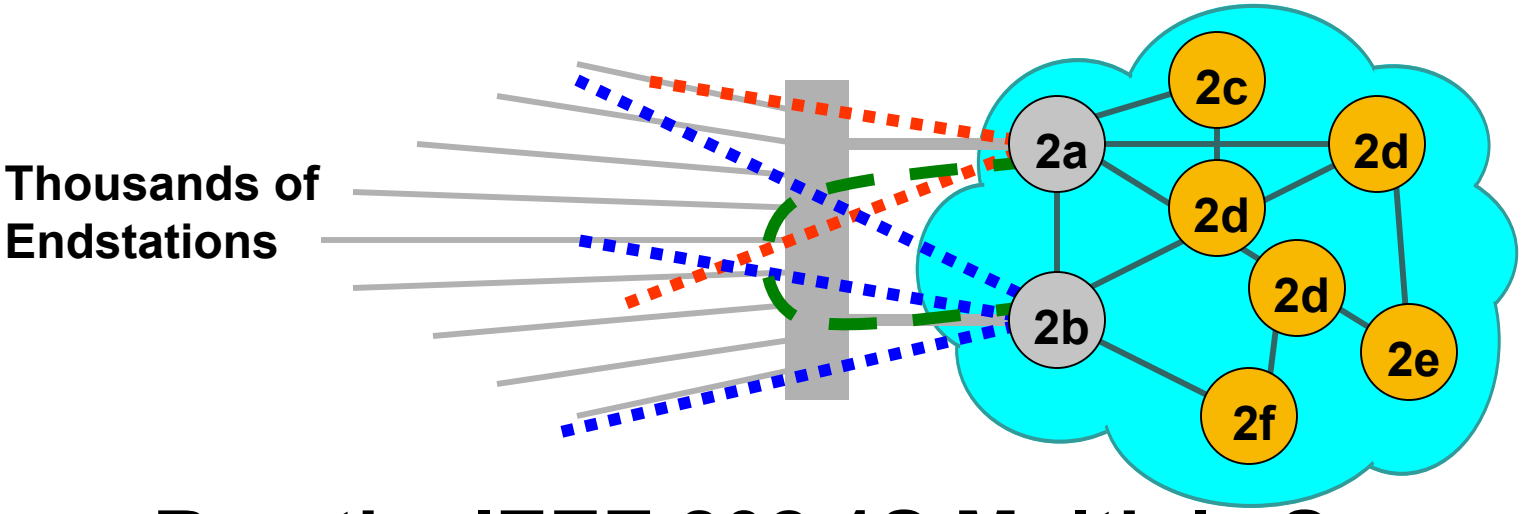
- Now, if we reserve ***one*** Emulated LAN for MSTP BPDUs for just this Island…

Rule 3: Preventing multiple entrances or exits.



- The other Islands disappear, because 2a and 2b see only each others' BPDUs.
- The endstations in the other Islands appear to be attached directly to the Giant Fat Yellow Coax.

Rule 3: Preventing multiple entrances or exits.



- Run the IEEE 802.1S Multiple Spanning Tree Protocol in the Island, including the Giant Fat Yellow Coax (on the BPDU IM).
- 2a and 2b now share the load over their attachments to the Emulated LAN on a per-Customer Service Instance basis.

Why use this "VLAN for this Island's Bridges only" scheme?

- **Because it fulfills Rule 3, requiring the controlling and sharing of the Layer 3 links, for absolutely any configuration, and without inventing any new protocols.**

Danger, Will Robinson!

- **But, this would mean that the BPDUs are not taking the same path as the data. This is dangerous!**

  As discussed for LSP Pings, load sharing by a router may cause Pseudowires in the same LSP to take different routes.

  If such load sharing is avoided, then the BPDU Pseudowire will take the same path as the data Pseudowires, within this Island.

- **That is sufficient to guarantee no loops within this Island, and therefore, no loops anywhere.**

Improvement to Provider Bridges: Suppressing needless transmissions

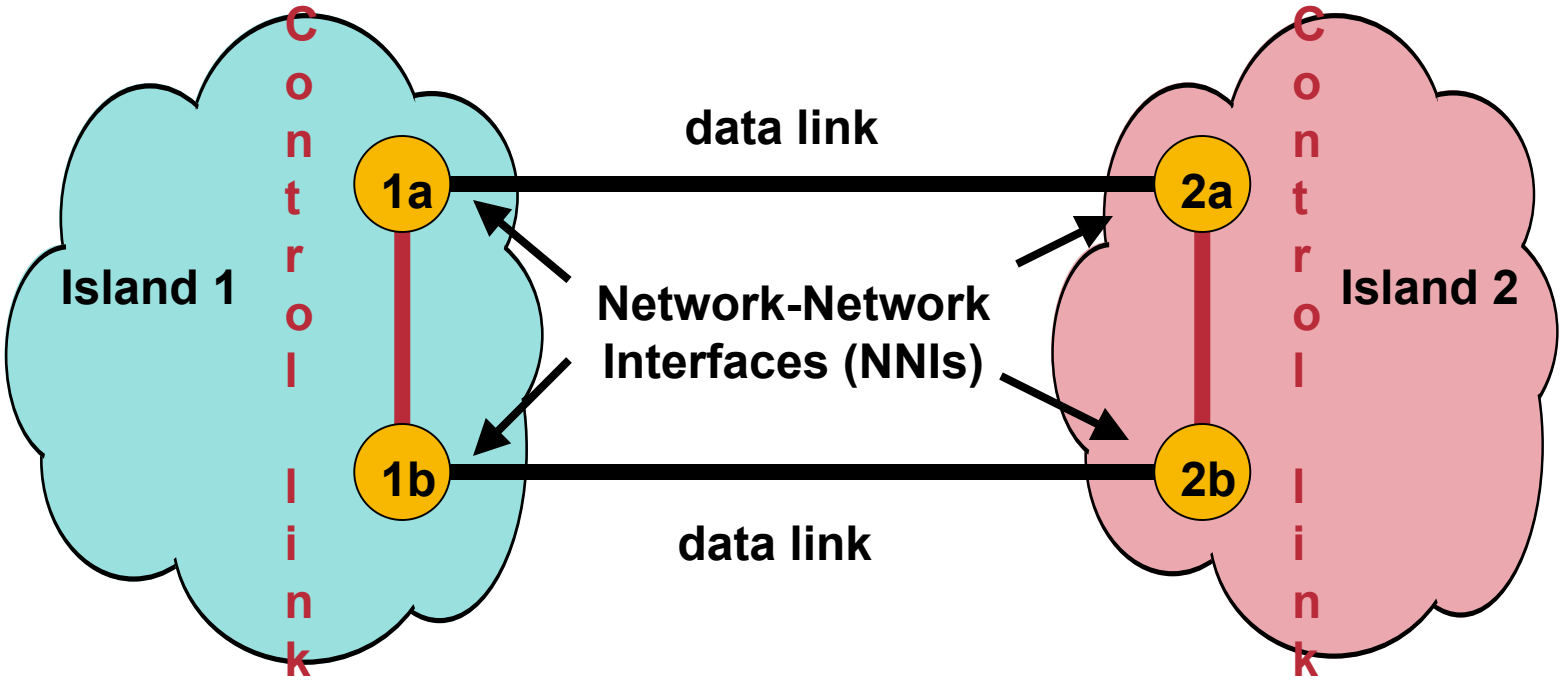
- In normal bridged Ethernet, there is no penalty for sending data down a point-to-point LAN to a blocked port.

  However, in a network of bridges connected by Pseudowires, each Pseudowire may share a physical link with other data.

  There is, then, a penalty for sending data towards a blocked port.

- A bridge should be configurable to not transmit data (from certain ports) towards a port it knows is blocked.

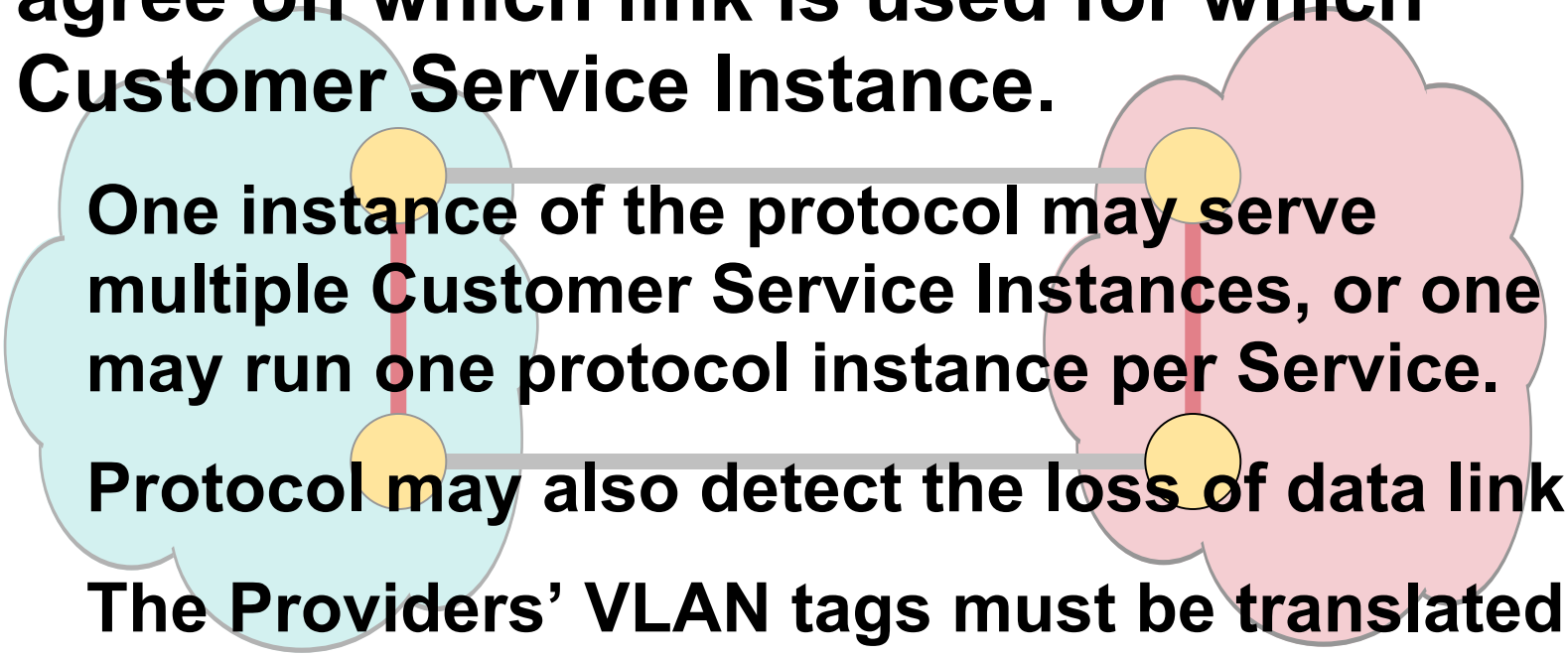
The Double NNI: One type of Interconnect Medium?



- A protocol runs among the 4 Provider Bridges to ensure that only one data link is used for any given Customer Service Instance.

The Double NNI: One type of Interconnect Medium?

- The protocol ensures that the bridges agree on which link is used for which Customer Service Instance.



One instance of the protocol may serve multiple Customer Service Instances, or one may run one protocol instance per Service.

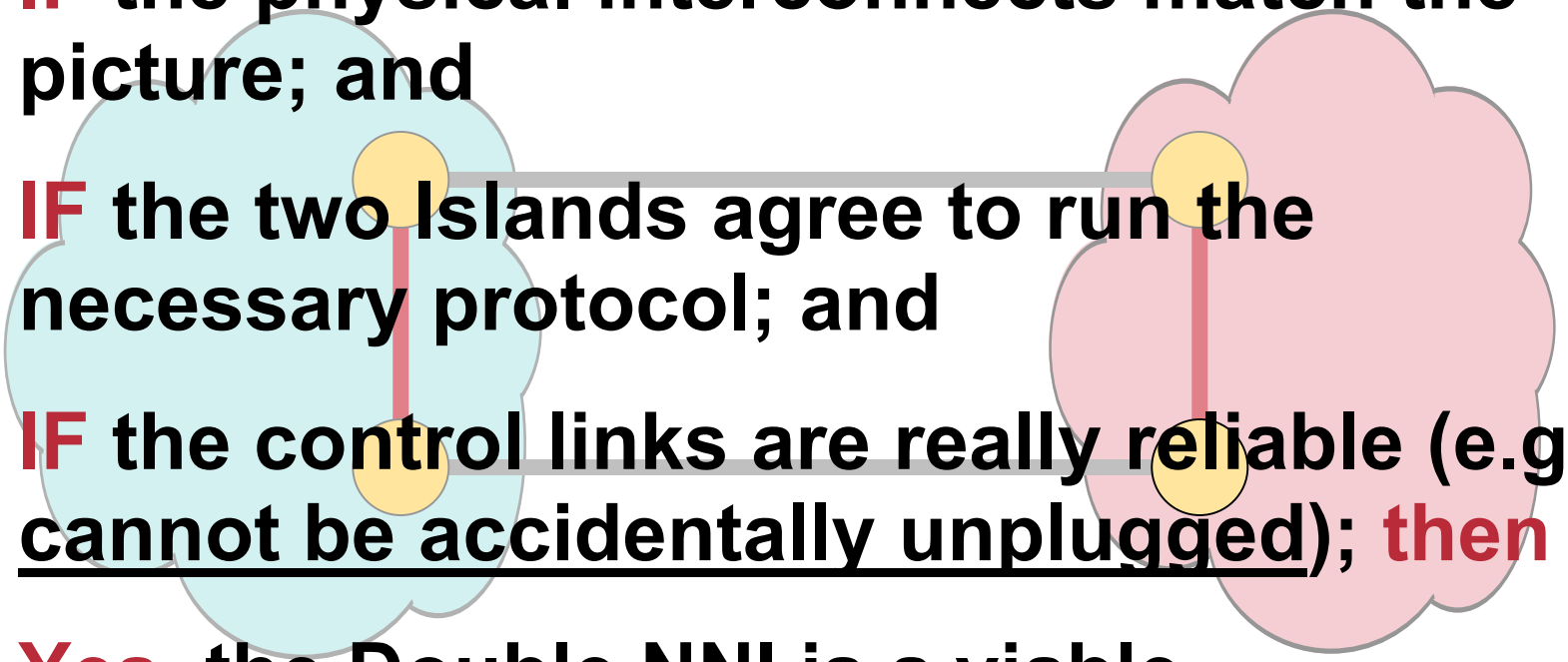
Protocol may also detect the loss of data link.

The Providers' VLAN tags must be translated.

- The loss of a control link may prevent the use of the data links – or worse!

The Double NNI: One type of Interconnect Medium?

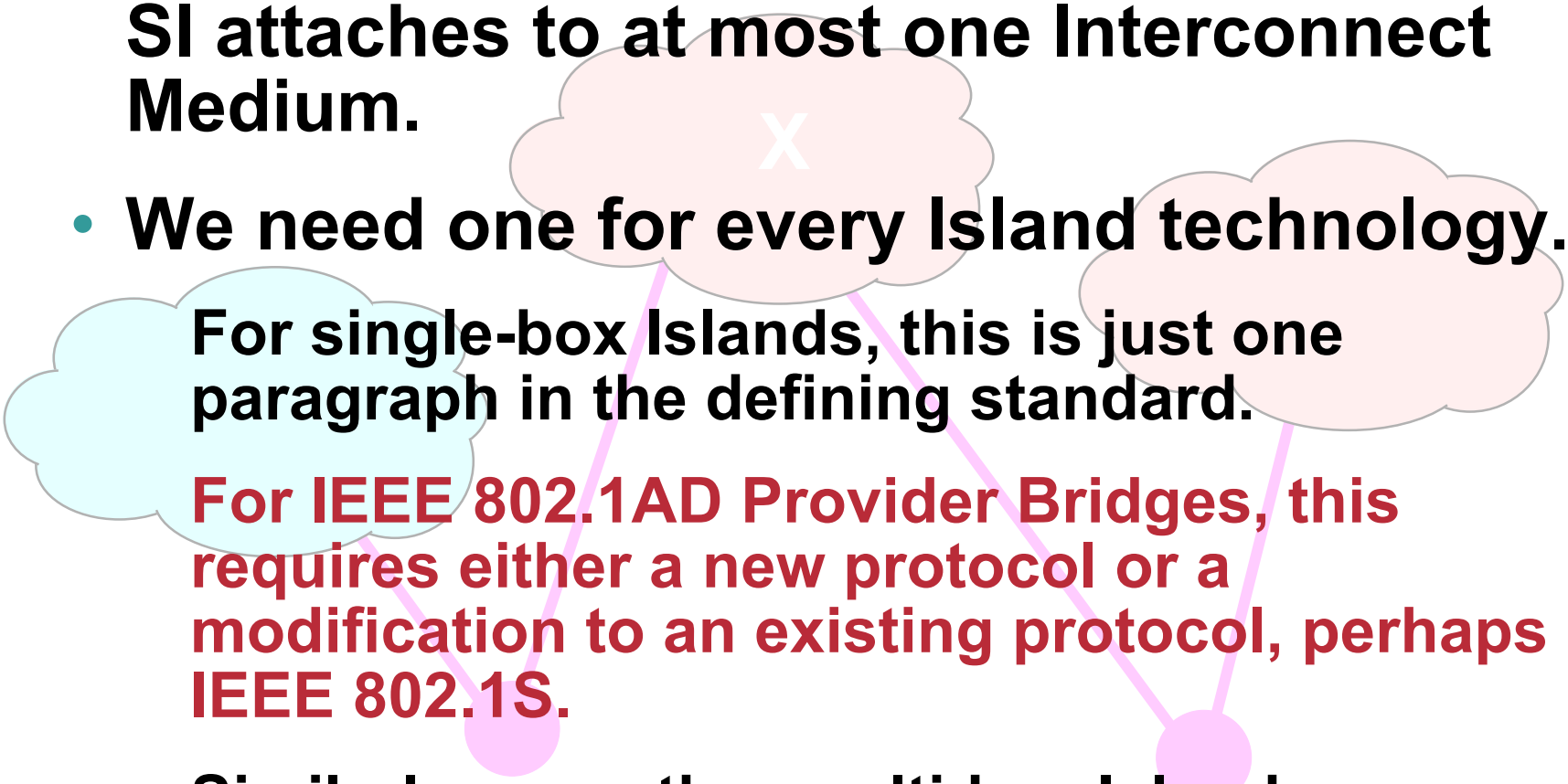
- IF the physical interconnects match the picture; and
- IF the two Islands agree to run the necessary protocol; and
- IF the control links are really reliable (e.g. cannot be accidentally unplugged); then
- Yes, the Double NNI is a viable Interconnect Medium.



Part 5: Ensuring No Multiple Attachments

Standardizing Rule 4.

- There is no standard for for ensuring one SI attaches to at most one Interconnect Medium.
- We need one for every Island technology.



For single-box Islands, this is just one paragraph in the defining standard.

For IEEE 802.1AD Provider Bridges, this requires either a new protocol or a modification to an existing protocol, perhaps IEEE 802.1S.

Similarly, any other multi-box Island technology will require a protocol.

Standardizing Rule 4.

- One idea for a Rule 4 Protocol for IEEE 802.1AD Provider Bridges:

  Each Provider Bridge attached to an Interconnect Medium (PB+IM Bridge) transmits a multicast list of {P-VLAN, Interconnect Medium ID} pairs.

  In case of Rule 4 violation,only the PB+IM with the best Bridge ID operates its IM attachment.

  This would also allow PB+IM Bridges to verify that all are on the Island's BPDU IM.

  This also catches inconsistent {P-VLAN, IM-ID} mappings in different PB+IM Bridges.

Standardizing Rule 2.

- One way to guarantee that an Island has no “back doors” to other Islands outside the Interconnect Media.

  Carry an “Island Name” in the MSTP BPDU.

  Receipt of a different Island Name in an MSTP BPDU blocks all VLANs on the port.

  Island Name could be a delimited field in the Configuration ID, or could be a new field.

  But, this makes changing Island Names difficult. (: I don’t have all the answers…Yet. :)

Part 6: End-to-End Protocols

What “End-to-End” Protocols are needed?

- **“Forget these MAC addresses” protocol(s).**

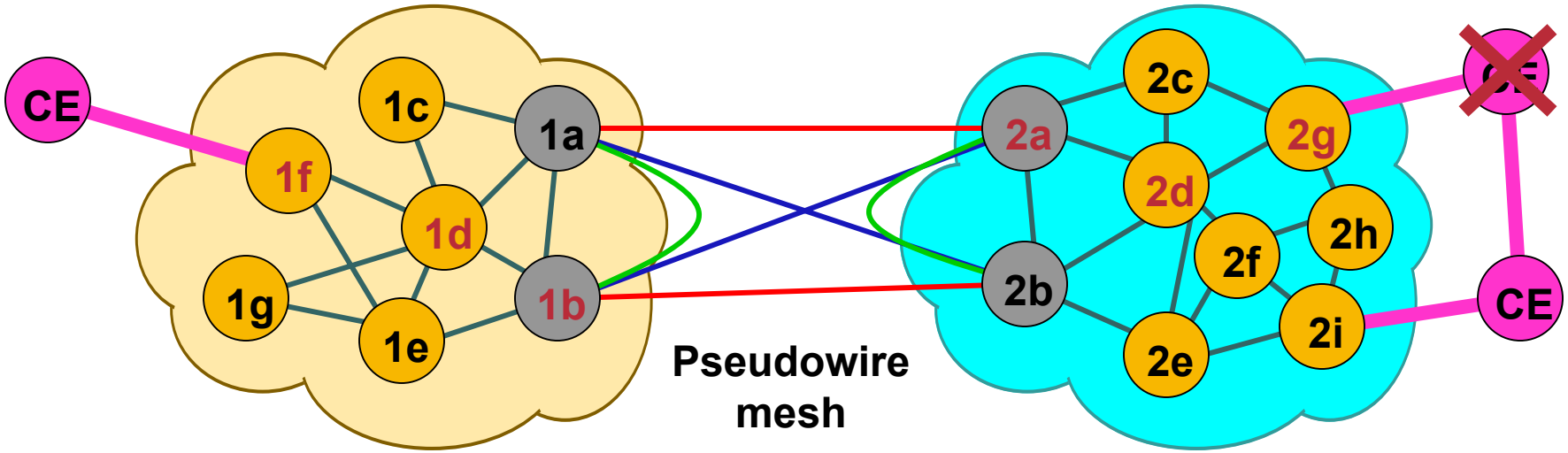
  Events in one Island may require bridges and/or LAN Emulation modules in other Islands to forget some or all of the MAC addresses in a given Service Instance.

  Events in one Island may require a Forwarding Function to forget learned MAC addresses.

- **OAM protocol(s).**

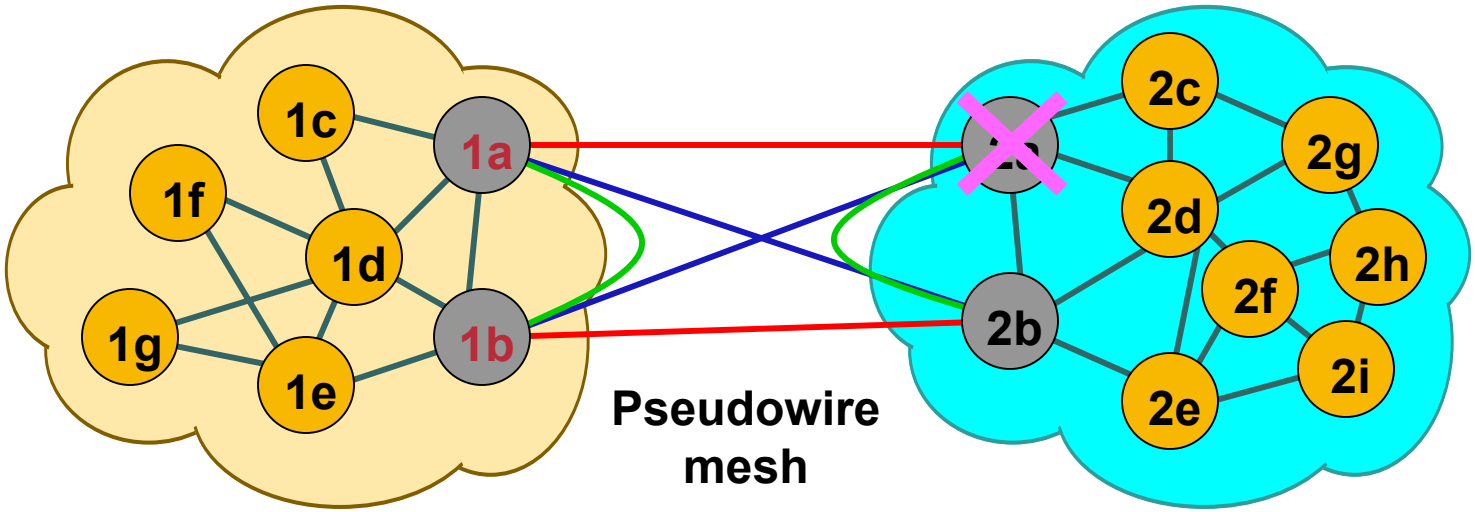
  Debugging connectivity problems in this network may be greatly facilitated by a few new protocol features.

“Forget these MAC addresses 1” protocol



- A spanning tree topology change in the Customer’s network may require other Islands’ bridges and LAN Emulation Forwarding Functions to forget some of that Customer’s MAC addresses.

“Forget these MAC addresses 2” protocol



- A spanning tree topology change in the Provider’s network may require other Islands’ LAN Emulation Forwarding Functions only to forget a number of Customers’ MAC addresses.

“Forget these MAC addresses” protocols

- **The two Islands may use very different technologies. Therefore:**

  The “End-to-End Forget” protocol must be an in-band Ethernet frame.

  The “Forwarding Function Forget” protocol may be an Ethernet frame or a technology-dependent control frame, but must not pass through a Forwarding Function.

End-to-End OAM Protocols

- Proposals exist in IETF and MEF for “end-to-end OAM protocols”.
- The protocols suggested by MEF utilize only Ethernet frames, and are therefore compatible with all Island and all Interconnect Medium technologies.

End-to-End OAM Protocols

- **OAM packet types suggested, so far:**

  Which Provider edge bridge "owns" this Customer MAC address?

  What is the path to this Customer MAC address?

  I'm a Provider edge bridge on this Service Instance, and I'm alive [or about to die].

  Ping. (To a Provider edge bridge, perhaps to collect statistics.)

The “Traceroute” Problem

- **Three questions a Provider Network administrator might well ask:**

  What is the path for a customer’s frame from Customer source MAC A to destination MAC B, through this Provider’s network?

  What is that path through the other Provider’s network?

  What was that path before it stopped working?

The “Traceroute” Problem

- **What is the path for a customer’s frame from Customer source MAC A to destination MAC B, through this Provider’s network?**

  This question can be answered by an application running in a management station, using the currently defined standard MIBs.

  It might be answered more easily and quickly by a hop-by-hop, in-band “Traceroute” Ethernet frame.

The “Traceroute” Problem

- **Through another Provider’s network?**

  This question could be answered by an application running in a management station, using the currently defined standard MIBs.

  However, Providers may not trust each other to that degree.

  It might be answered more easily and quickly by a hop-by-hop, in-band “Traceroute” Ethernet frame.

  Of course, Providers may not trust each other to that degree, either, but this is more likely.

The “Traceroute” Problem

- **What was that path before it stopped working?**

  A “Traceroute” function will likely not help, here, as the MAC address(es) have likely been forgotten.

- **But, maybe …**

  If the information has not been forgotten, a “Traceroute” function might generate useful information about a lost path.

Part 7: Alternatives to Fully Independent Islands

Preventing Global Loops Directly

There is a very different model for preventing large-scale loops which merits attention:

- **Run one instance of the spanning tree, globally, for each Customer Service Instance.**

  This has the obvious and very significant advantage of guaranteeing no loops, anywhere.

One Spanning Tree Instance for each Customer Service Instance

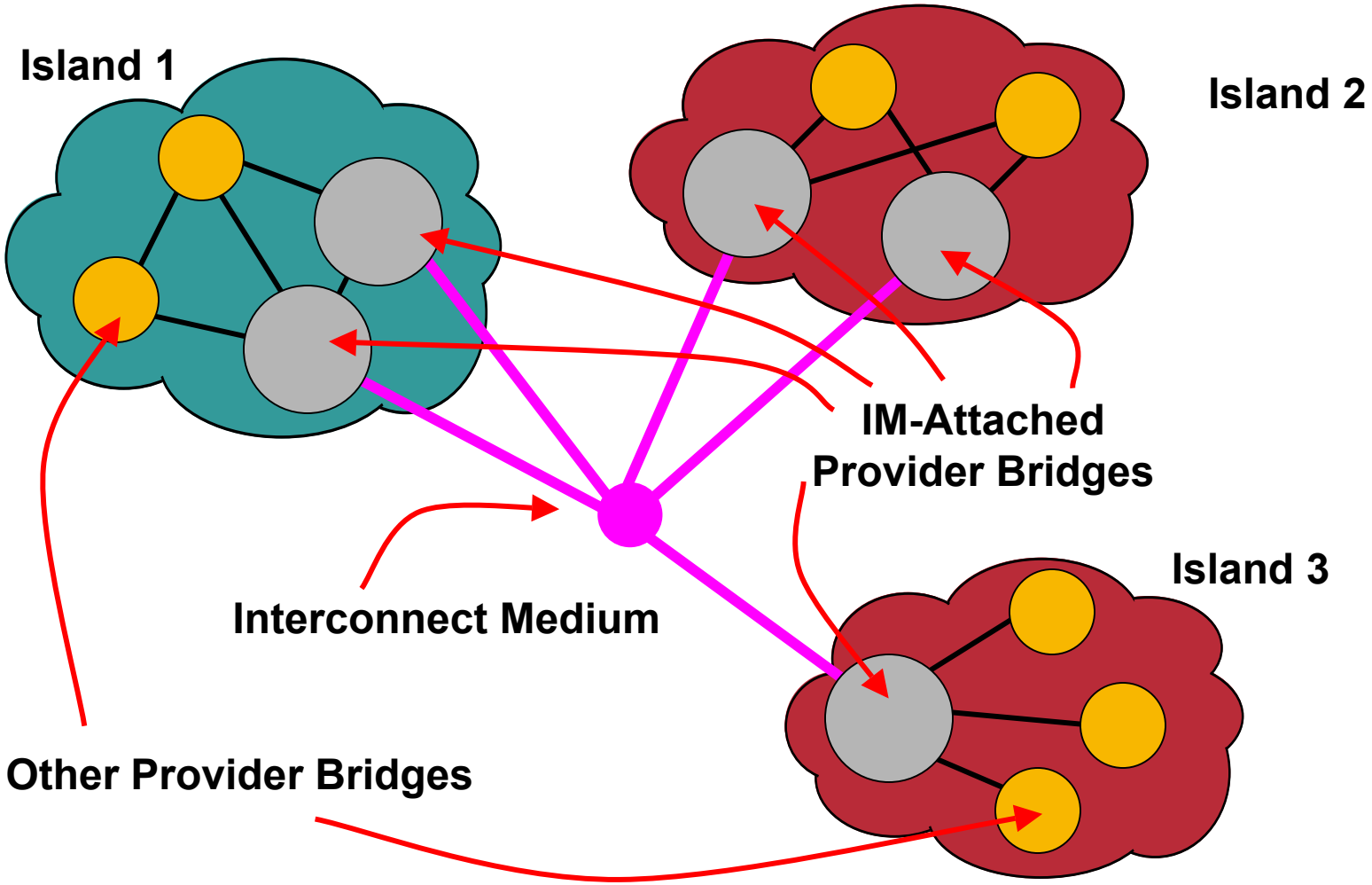
- Within one island, MSTP can carry up to 64 spanning tree instances in each BPDU.
- By sending up to 64 BPDUs per transmission event, an Island could support a separate spanning tree instance for each of the 4094 P-VLANs.

  Typically, this trick costs the Provider Bridge's CPU little more than it costs to run 64 spanning trees in separate BPDUs.

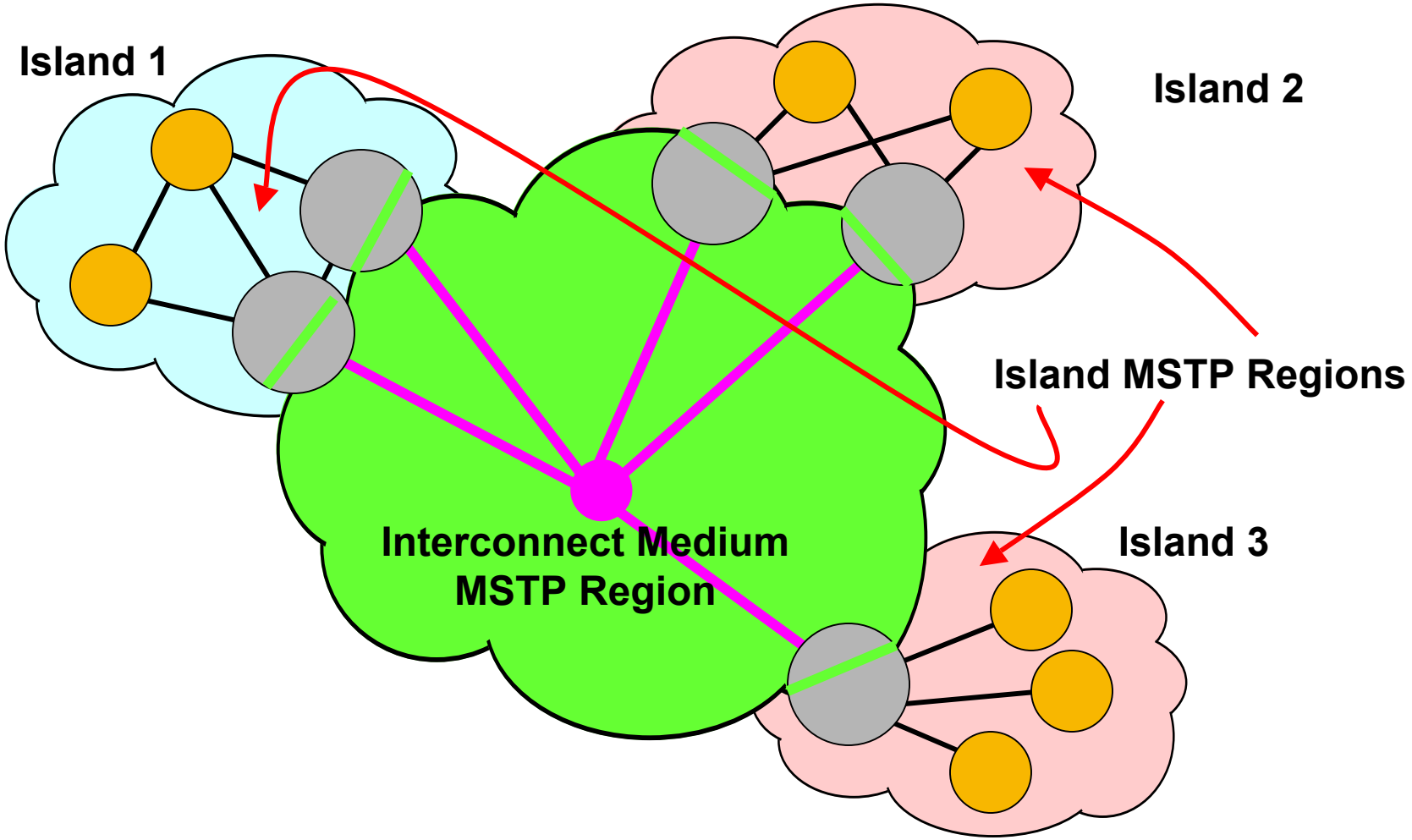
Millions of Spanning Trees (MoST)

- **Each Island is a single MSTP Region, and runs one spanning tree instance per P-VLAN.**
- **The twist: An Interconnect Medium may carry up to 4k Customer Service Instances, distinguished by an IM-VLAN ID (an outer 802.1Q tag) in each frame.**

Millions of Spanning Trees (MoST)



Millions of Spanning Trees (MoST)



Millions of Spanning Trees (MoST)

- At each attachment to an Interconnect Medium, the Provider Bridge must perform two translations:

  Between Interconnect Medium VLAN IDs and the Island's P-VLAN IDs; and

  Between the IM's MSTP BPDUs and the Island's MSTP BPDUs.

- Thus, each Island remains separate, but there is one spanning tree instance for each Island that spans the world.

Millions of Spanning Trees (MoST)

- **The easy parts:**

  Every Provider Bridge's MST Configuration Table is set for one spanning tree instance per VLAN, with MST IDs matching VLAN IDs, so the Configuration Digests always match.

  The Interconnect Medium, in essence, is an MST Region which interfaces with each Island.

  The Configuration Name and Revision Level fields, and/or similar, new fields, may be used to discover and/or verify that all Provider Bridges attached to an IM agree to the IM's identity.

Millions of Spanning Trees (MoST)

- **The easy parts:**

  In order to enable Islands to pass data directly among each other over the IM, without needing relays from Island to Island, the root of each spanning tree instance should be in an IM-Attached Provider Bridge.

  There can be any number of Interconnect Media. An IM attached Provider Bridge has separate MSTP and VLAN ID translation functions for each IM.

  The IM is an MSTP Region. However, because it is simply an Ethernet emulation, MSTP is not needed to select paths within the IM.

Millions of Spanning Trees (MoST)

- **Now, one could extend this idea to make the IM equivalent to an Island, and construct arbitrarily large clouds of clouds!**
- **However, this makes Customer Service Instances arbitrarily large.**

  Such a large network, covering multiple administration authorities, is unlikely to work.

Fully Independent Islands (FII) vs. Millions of Spanning Trees (MoST)

Cisco.com

Fully Independent Islands are better:

- Does Provider A trust Provider B enough to share a spanning tree instance?
- Must a low-cost edge bridge carrying only 8 Customer Service Instances support 4K spanning trees and 64 BPDUs per Send time?
- FIIs can interoperate with any Island technology: even one that does not utilize spanning trees at all.
- One full mesh of Pseudowires per Customer Service Instance maximizes delivery efficiency.
- MoST invites the creation of networks that are just plain too big to work.

Fully Independent Islands (FII) vs. Millions of Spanning Trees (MoST)

No, Millions of Spanning Trees is better:

End-to-end spanning trees are attractive because they **guarantee** loop-free forwarding.

The need for "Forget" messages is eliminated, the "Double NNI" connection is not special, and Rule 4 (Island-to-Island forwarding) can be relaxed.

Sharing multiple Customer Service Instances over a single IM ensures that the BPDUs traverse the same paths as the data, **even** if Pseudowire-based load sharing is present.

MoST typically creates fewer Pseudowires than FIIs. Fewer to signal, fewer to manage.

Yet another possibility: FII + Pseudowire Multiplexing

- **Borrowing an idea from MoST:**

  The FII model can work using Customer Services Instances multiplexed using VLAN tags over one Pseudowire for data, and using untagged BPDUs over the same Pseudowire.

  In this case, Islands must differentiate between own-Island BPDUs and other-Island BPDUs, and ignore the latter.

  Again, additional fields in the MSTP BPDU, and/or new uses of the existing Configuration Name field, may provide a means for Island identification in the BPDUs.

FII + Pseudowire Multiplexing

- Islands' spanning trees are still independent; an Island is not required to run one spanning tree instance per P-VLAN.
- No spanning tree covers the world.
- BPDUs follow the same paths as data.
- A Provider Bridge could work with both untagged Interconnect Media and Pseudowire Multiplexed Media.

FII + Pseudowire Multiplexing

- One may trade off between many VLANs per Emulated LAN (few BPDUs, but VLANs reach unwanted ports) and few VLANs per ELAN (many BPDUs, but efficient delivery).
- Islands must cooperate to the extent they must be able to differentiate own-Island from other-Island BPDUs.

What Shall We Do?

- Clearly, further discussion is needed.
- In any case, separation of the problem into Islands and Interconnect Media greatly clarifies the discussion.
- If the one-hop-to-everywhere model is not followed, then either:

  A global spanning tree is required; or

  We risk global forwarding loops.

Part 7: Standardization

Several standards bodies are at work

- IETF's PPVPN Working Group is defining Layer 2 Virtual Private Networks (L2VPNs).
- IEEE 802.1 is defining Provider Bridges and Link Security.
- Metro Ethernet Forum is defining requirements and working on some protocols for Metro Ethernet Service Providers.
- ITU is considering work in both SG13 and SG15 for OAM and Multiple Services, including Ethernet.

What's needed from IEEE 802?

- Complete the IEEE 802.1AD Provider Bridges standard as so-far envisioned.
- Provide for not sending data towards blocked ports.
- Control multiple attachments from one Island for one Service Instance.
- Define OAM and End-to-End "Forget".
- Protocol or mods to prevent "back doors".
- Scale GVRP up to 4K VLANs.
- Define when "Forwarding Function Forget" should be generated?

What's needed from MEF?

- Definitions of services expected by Customers.
- Ownership of the overall multi-technology requirements and architecture.
- One-stop-shopping documents that pull together the various other standards groups' documents into a coherent whole.
- Other standards that are required, but not felt by other standards bodies to be within their scopes. (Line Management Interface?)

What's needed from IETF?

- **A method for constructing and ensuring the correct operation of Emulated LANs over suitable L3 substrates, e.g. MPLS or L2TPv3, including:**
  - **A means for Island devices to discover their peers in the Emulated LAN (in progress).**
  - **A means for more efficient broadcast / multicast transmission than, "Send a copy on each Pseudowire." (slow progress)**
  - **A means for telling intermediate routers that per-Pseudowire load-sharing is not allowed on this MPLS tunnel. (new requirement)**

What's needed from IETF?

- **Still more for proper LAN Emulation:**

  Perhaps, the ability to scale to more nodes than a simple full mesh can reliably support (H-VPLS? BUS?). (slow progress)

  A way to meet the connectivity requirements of Rule 5. (in progress)

  Forwarding Functions and non-bridges must generate and obey "End-to-end Forget" messages. (new requirement)

  A "Forwarding Function Forget" message is needed. (new requirement)

What's needed from Somebody? (Everybody?)

- Select among the contenders for how Interconnect Media may be constructed: FII? MoST? FII+Muxing? All?
- Create a globally unique universal ID (a string and/or a number) for each Customer Service Instance.

  Each IM or Island technology can map its local identifier, e.g. P-VLAN ID or Pseudowire VC ID, to the universal ID.

  This ID is needed to ensure that ID remapping functions in IM-attached Provider Bridges do not cause a global forwarding loop.

What's needed from ITU?

- **Models for interaction and interconnection between circuit-based and frame-based approaches to providing Ethernet services.**

What's needed from Vendors?

- **Value-add features within the framework for interoperability provided by the standards.**

  But, be really, really, *REALLY* careful not to cause global loops!

Again, The Five Rules:

1. Each Island is responsible for preventing internal forwarding loops.
2. Islands connect to other Islands only through Interconnect Media.
3. Each Island ensures that no customer data frame passes through more than one Interconnect Medium attachment into or out of the Island.
4. Each Island ensures that it attaches any given Customer Service Instance to no more than one Interconnect Medium.
5. An Interconnect Medium ensures that if an attached port can talk to ***any*** other attached ports, it can talk to ***all*** of the ports attached to that Medium.

That's too difficult (restrictive, …)

- **But, *my* variant of Spanning Tree *can* span the whole world!**
  - **Only if *every* Provider participates in it.**
  - ***At best*, you then have a global L2 network operating in parallel with the existing global L3 network.**

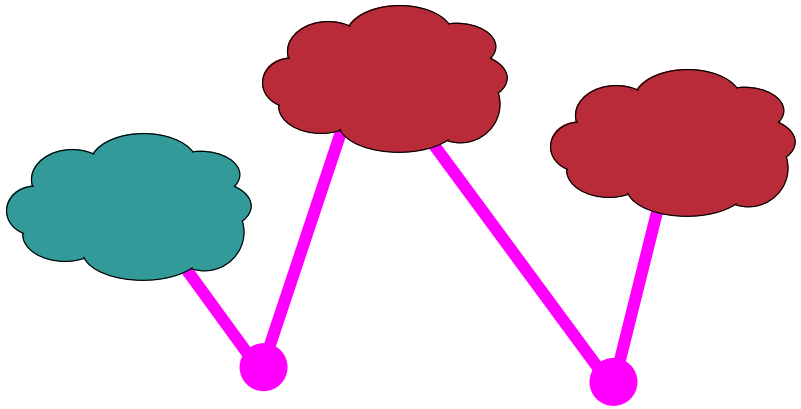
That's too difficult (restrictive, …)

- **But, we can just configure Inter-Provider connections using ad-hoc "NNI" ports.**
  - **How do you know that there is only one ad-hoc connection for a given Customer between Providers A and B?**
  - **How do you know that there is not a path for a given Customer from Provider A to Provider B to Provider C and back to Provider A?**
  - **What happens when (not if!) you're wrong?**

That's too difficult (restrictive, …)

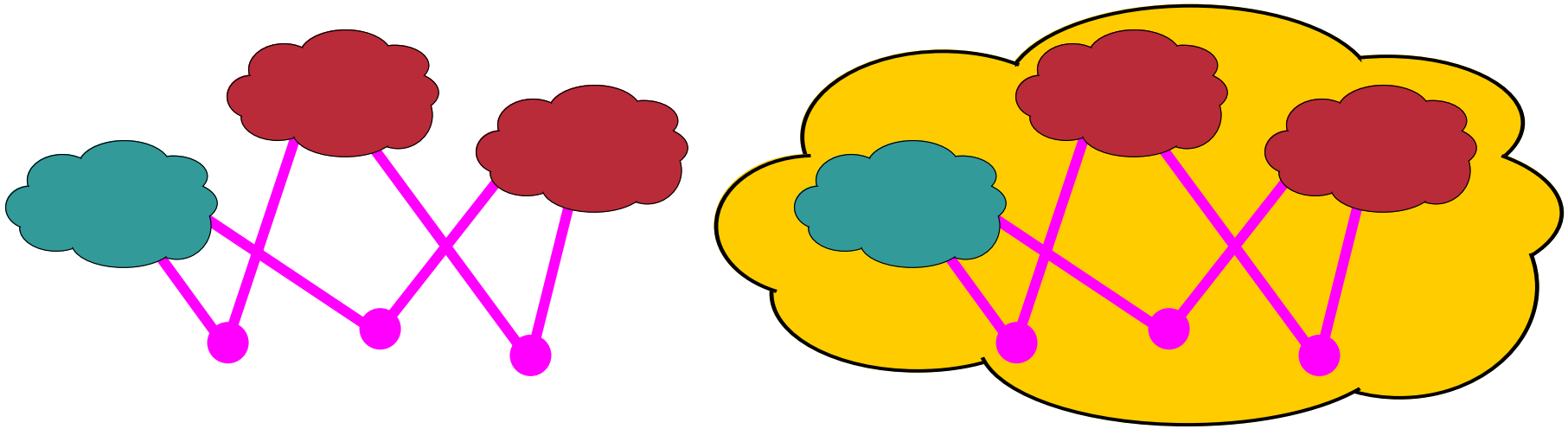
- **But, Rule 4 disallows some non-looping configurations.**
  - **Again, how do you know that there is only one ad-hoc connection for a given Customer between Providers A and B?**
  - **Again, how do you know that there is not a path for a given Customer from Provider A to Provider B to Provider C and back to Provider A?**
  - **Again, what happens when (not if!) you're wrong?**

That's too difficult (restrictive, …)



- **I don't care! I'm going to setup multiple links, anyway.**
  - **Well, we can't stop you.**
  - **But, please, be careful!**

That's too difficult (restrictive, …)



- **Why can't you run a spanning tree among the Islands?**
  - **You can! But, that makes the group one Island.**
  - **There is no rule against remapping P-VLAN IDs within an Island.**

That's too difficult (restrictive, …)

- **But, I don't want my L2 service to traverse the Big-I Internet!**
  - **Just as now, for L3 services, there are many ways to isolate one virtual network from another over the Big-I Internet.**
  - **Just as now, for L3 services, one may construct one's own "Intranet".**
  - **However, any given customer's service instance must reside in at most one of these virtual L3 Interconnect Media!**

That's too difficult (restrictive, …)

- **But, L2 traffic with a Service Level Agreement may be inelastic, and thus incompatible with the Big-I Internet.**
  - **So is phone traffic. The Internet must deal with that, too.**

That's too difficult (restrictive, …)

- **But, global Ethernet services are a bad idea. You should route, instead.**
  - **You'll get no argument from this author on that point!!**
  - **In fairness, however, not every protocol can be routed.**
  - **And other people (some of whom have money to spend!) have different "religions" with regard to Layer 2 vs. Layer 3.**

That's too difficult (restrictive, …)

- **But, you don't need Bridges. You can do this all with Layer 3 tunnels end-to-end.**
  - **Not if you offer multipoint-to-multipoint non-IP services that do not deliver every transmitted frame to every UNI. (Flap, flap, paddle, paddle, quack, quack – remember?)**
  - **If you don't want to supply "smart" mp-2-mp services, then go for it! The Layer 2 and Layer 3 solutions can coexist; there is no requirement to interoperate.**