Choices for Modeling IB-BEB DRNI

It is becoming clearer – I think

Rev. 1

Norman Finn

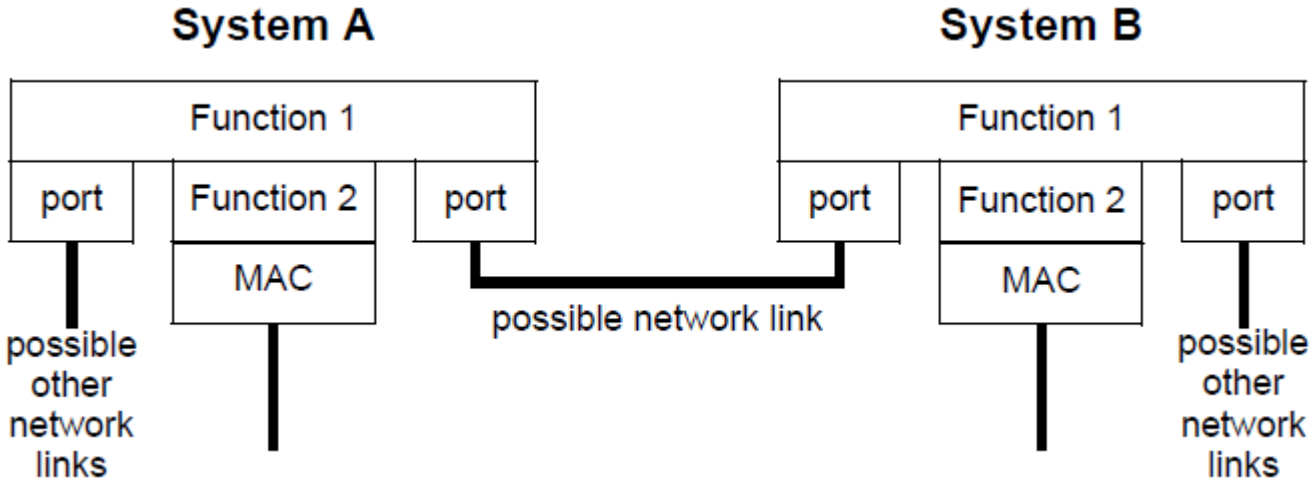
nfinn@cisco.com

SUMMARY

- Over the last six months, some very smart people have been going around in circles over the DRNI model.
- This is because we have not organized the interdependent decisions sufficiently clearly.
- The “Distributed Relay” is fundamental. The DRNI is simply a case of applying the Distributed Relay to Link Aggregation.
- The Distributed Relay can be applied repetitively.
- The importance of which flows are co-resident with which other flows has been underemphasized – this choice actually drives the differences in the models.
- **We cannot pick a model until we have a proposal for what is required to do a MEP that is shared among multiple Systems.**

The Distributed Relay

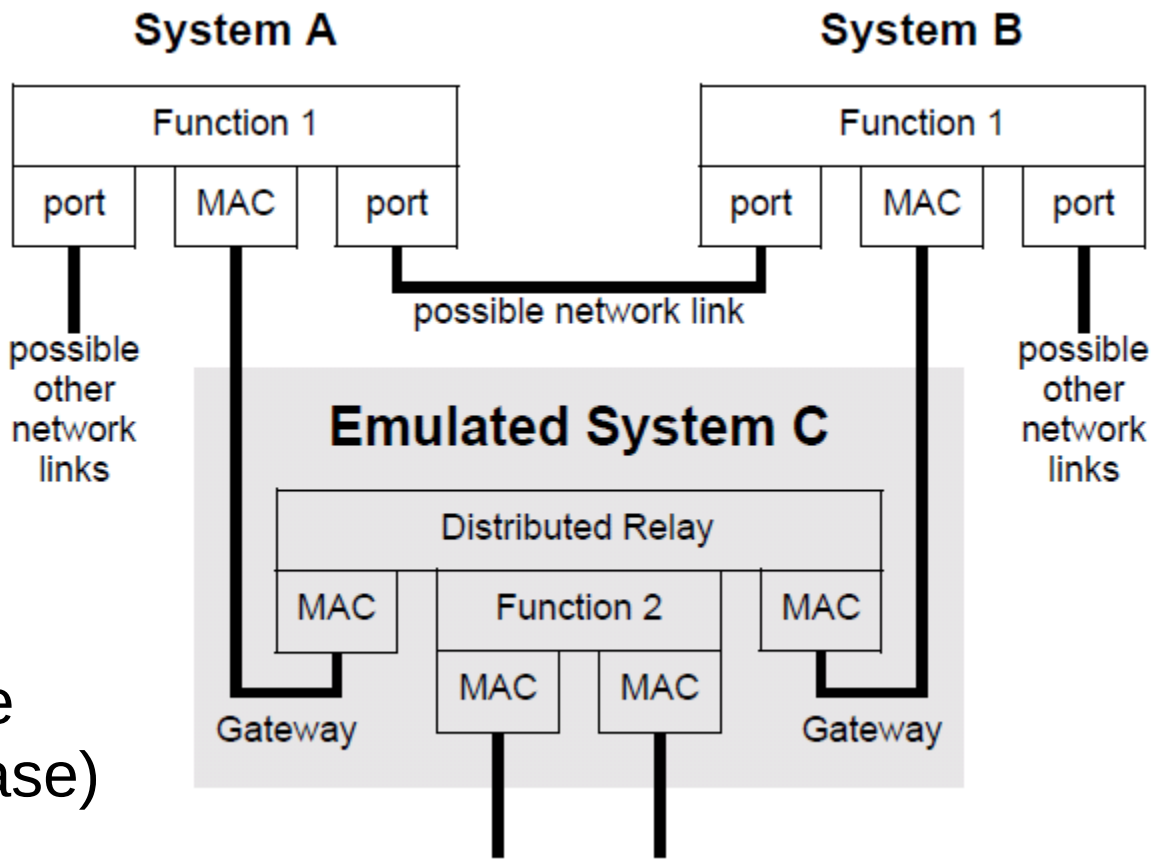
The Distributed Relay



- This is the starting point.
- Two separate Systems that may or may not be connected to each other and/or to other Systems.
- “Function 1” performs the interconnect among the ports in each System. There may be no other ports (for example, in a host computer).
- “Function 2” is the core of the exercise.

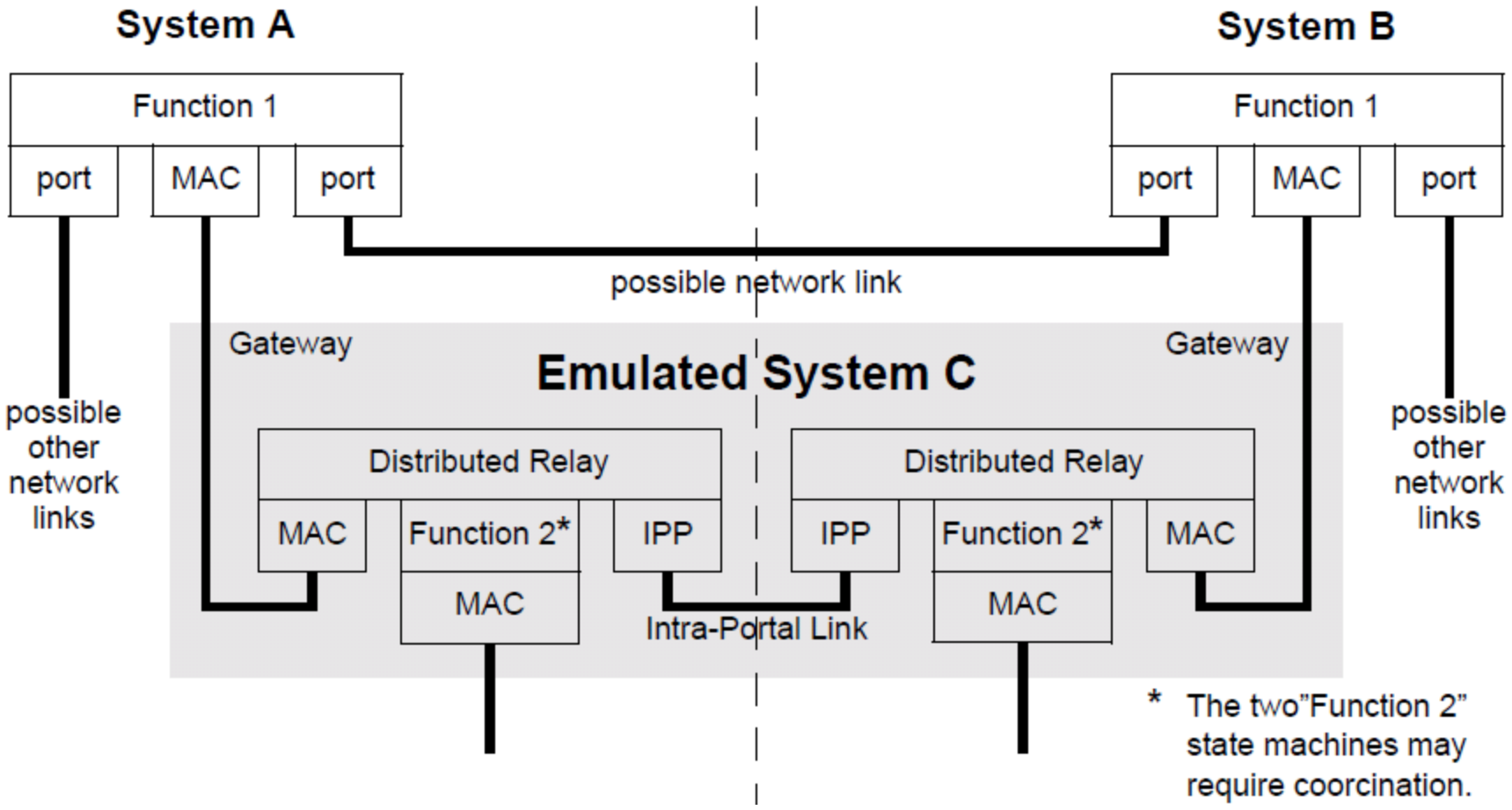
The Distributed Relay

- **This is the object of our efforts**.
- We want the two Systems A and B to emulate a third System C with a single Function 2 with two MACs.
- NOTE: N MACs above System C, (2 in this case) and N MACs below.



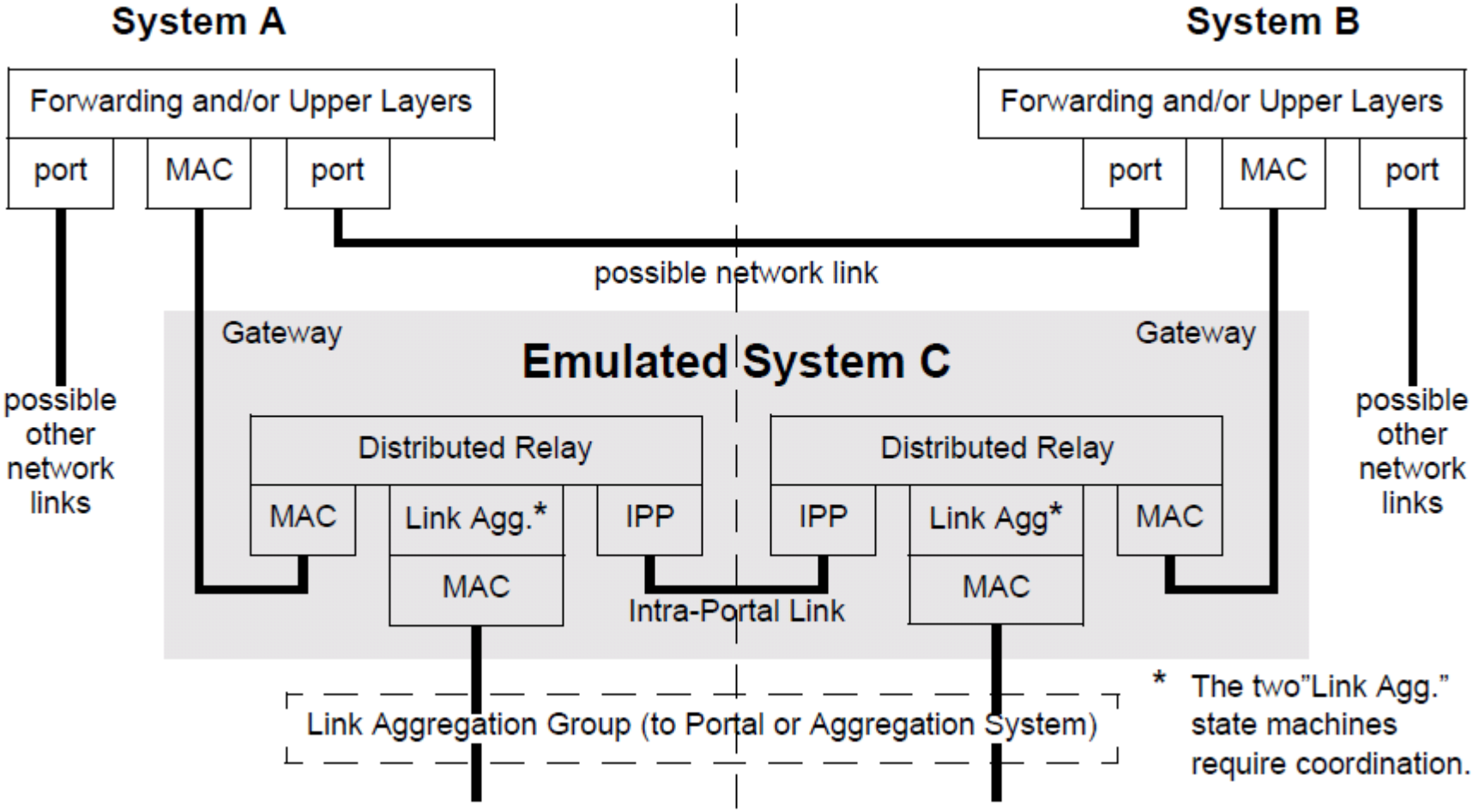
- **Two Systems visible above, with sole access from the network to a third (emulated) system with one or two down links, visible to Systems below.**

The Distributed Relay



- The **Distributed Relay** is how we accomplish this task.
- N MACs above, 1-N MACs below.

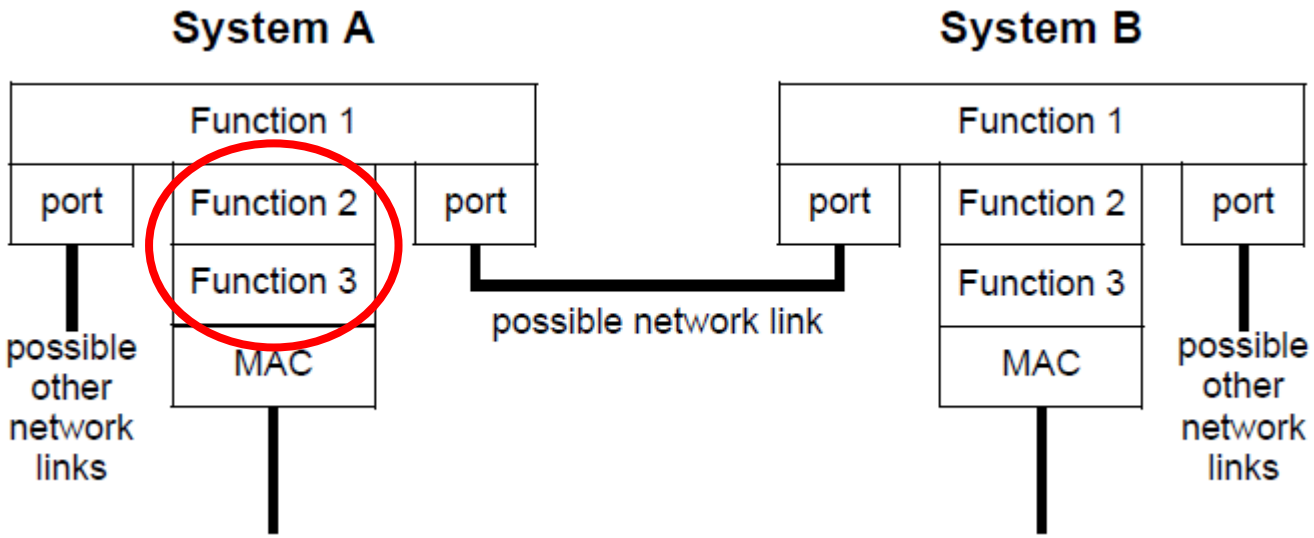
The Distributed Relay: the DRNI



- **DRNI: Function 1 = bridge/router/host.**
- **Function 2 = Link Aggregation.**

The Distributed Relay reapplied

Was just Function 2

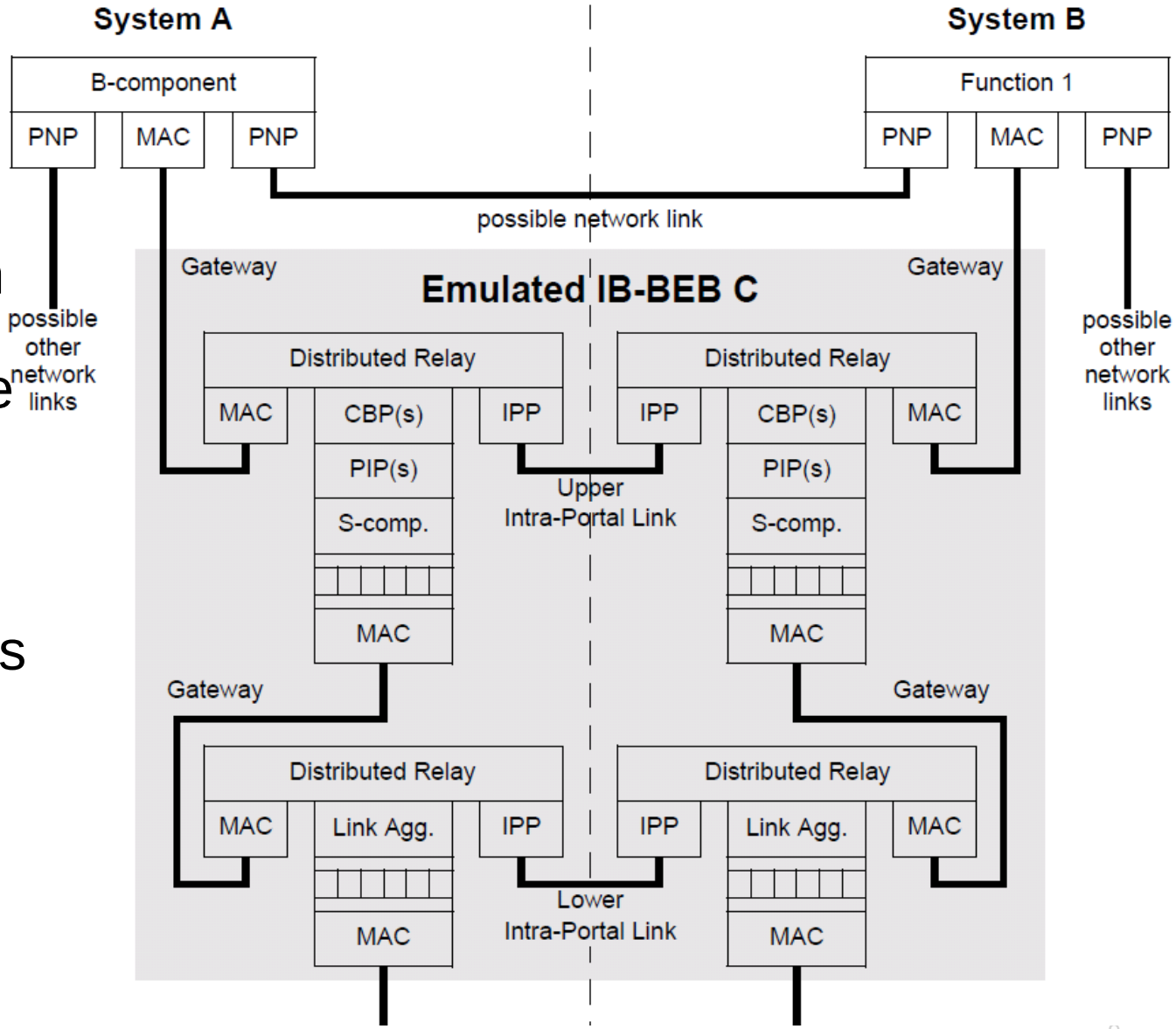


- We can apply this methodology again if we add another function to the stack.
- **Substitute two Functions 2 and 3 for the former Function 2, reapply the idea, and you get ...**

REFERENCE model for IB-BEB

- If we substitute the parts of the IB-BEB for Function 2 and Function 3, then we get the complete picture of an IB-BEB DRNI.

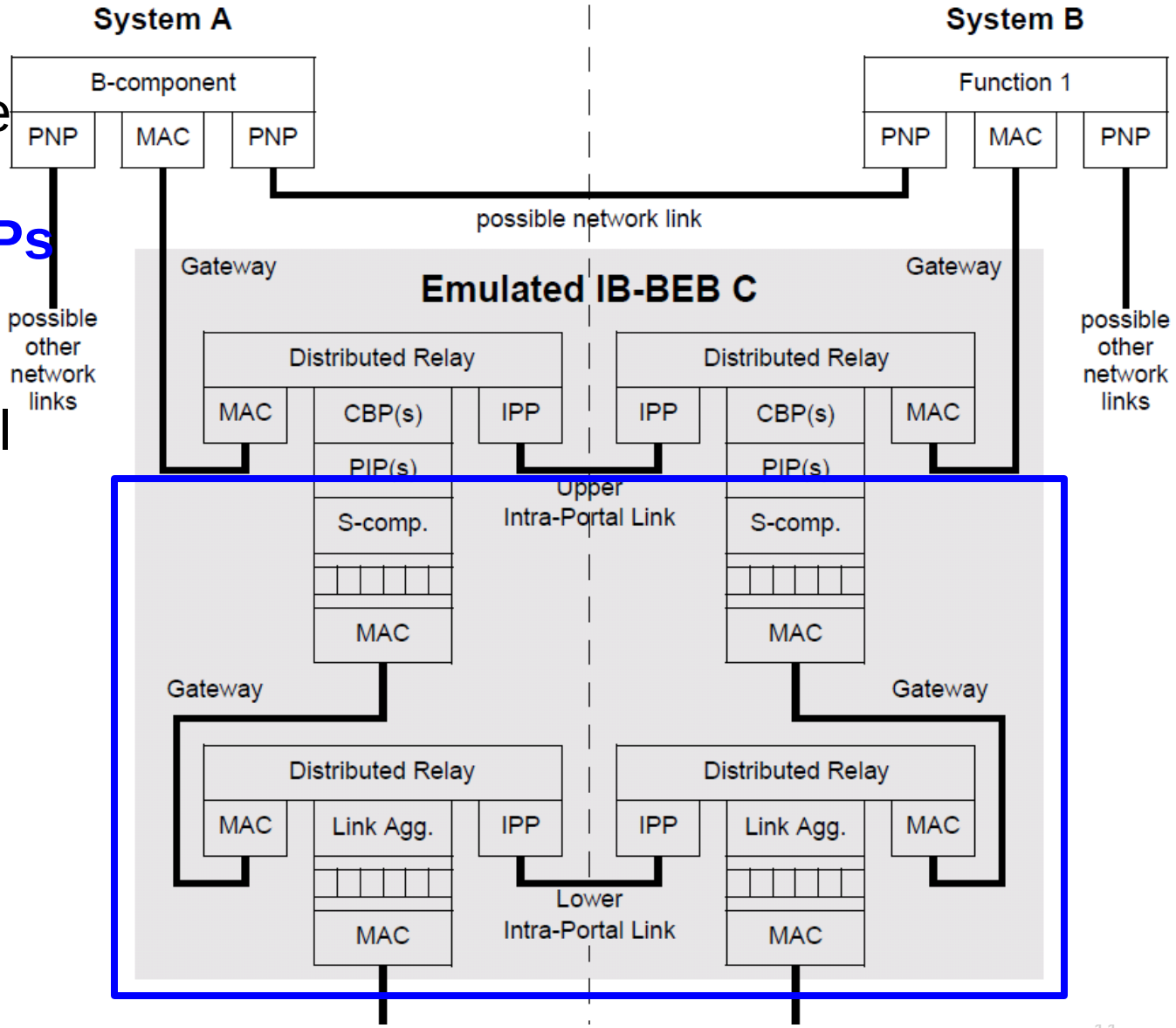
- F1 = B-comp.
  F2 = CBPs+PIPs +S-comp. (+CFM?)
  F3 = Link Agg. (+CFM?)



Aside: Placement of S-VLAN MEPs

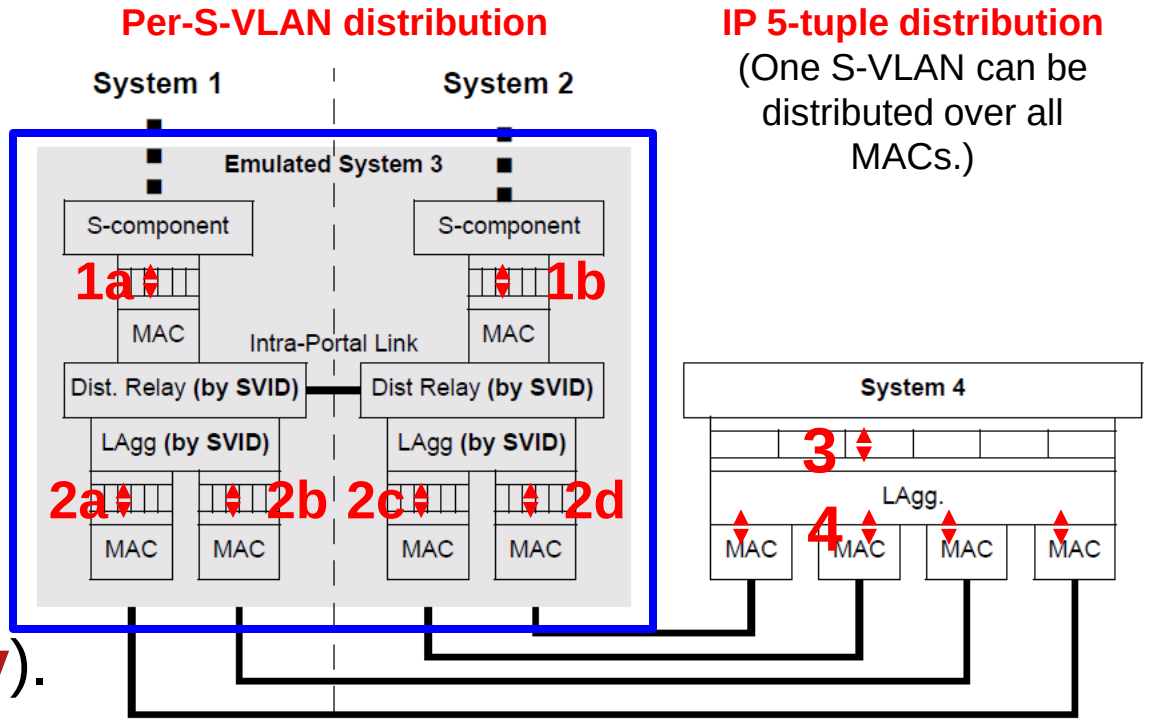
Placement of S-VLAN MEPs

- There is an issue about whether the **S-VLAN MEPs** should go above or below the lower Intra-Portal Link (or both!).
- Let us look at just **this part** of the picture.



Placement of S-VLAN MEPs

- Let us look at just **this part** of the picture.
- And, let's connect it to a legacy Aggregation System that distributes frames by IP 5-tuple (**we want compatibility**).



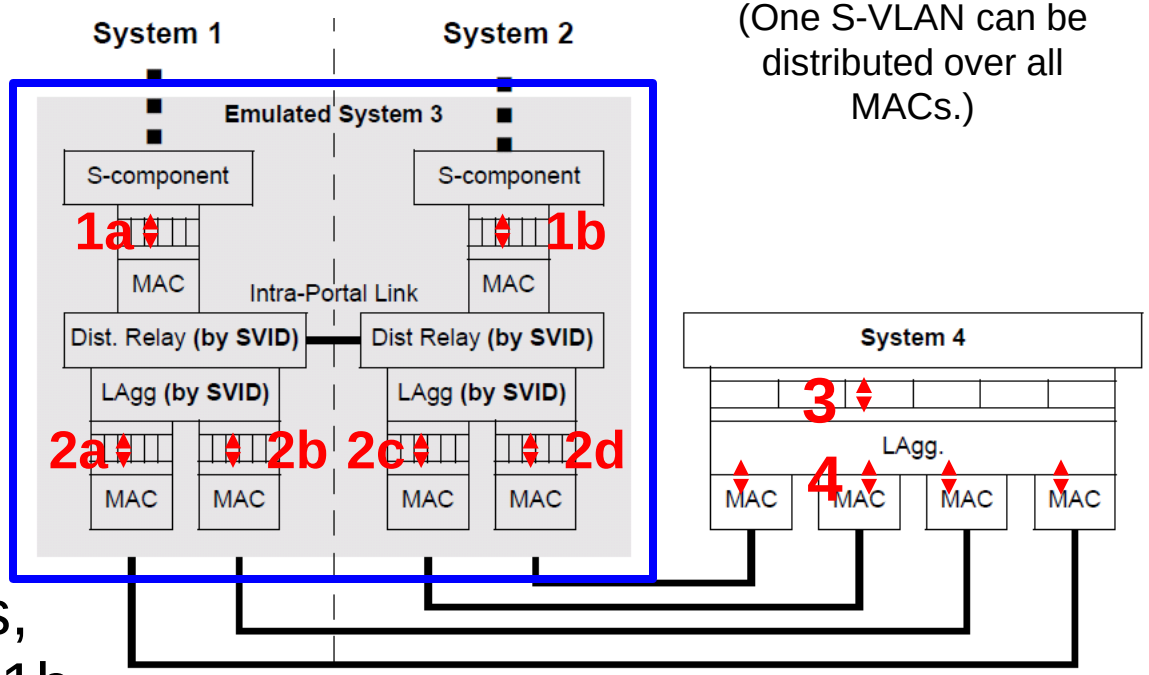
- Where do we put the **Per-S-VLAN MEPs**? In **System 4**, 802.1AX says that they **must** go in **position 3**. An implementation could reasonably place them in all four links (position 4), but the implementer must make sure that four MEPs **emulate a single MEP**.

Placement of S-VLAN MEPs

- Let us look at just **this part** of the picture.
- The only place you can put a single MEP in the emulated System 3 is position 1a or 1b. If the localization plan permits, only one of either 1a or 1b is needed, and Standby MEPs will work; Shared MEPs are not needed.

Per-S-VLAN distribution

IP 5-tuple distribution
(One S-VLAN can be distributed over all MACs.)



Placement of S-VLAN MEPs

- It is reasonable to consider a CFM stream to be an 802.1AX “conversation”. Therefore, all of the CFM from System 4, at least on any given VLAN, will take the same physical link. It is only the data frames of a VLAN that are likely to be distributed over all links.
- Therefore, for the purpose of simple CCM continuity, you could use Standby mode MEPs in each one of the positions 2a – 2d in System 1 or System 2. Only one would be Active, the one that talks to System 4.
- But, measuring frame loss using the CCM counters would require some kind of protocol interaction between System 1 and System 2. That is, a Shared MEP.

Placement of S-VLAN MEPs

- On the other hand, as Vissers has pointed out, these S-VLAN MEPs are part of Inter-Network Maintenance Entities. Putting the S-VLAN MEP above the lower IPL means that a link belonging to the System 1+2 network is included in the Inter-Network ME. This is undesirable from an operations point of view, because it makes it harder to determine which operator needs to take action when a failure occurs.
- **Resolution of this issue will take further discussion.**

Distributing functions

CBP/PIP Localization Plans

- Distributing a function can require a protocol and state machines to coordinate the separated parts; the "localize vs. distribute" choice thus drives the protocol requirements. (For example, distributing Link Aggregation requires protocol changes.)
- Such protocols are, in essence, what must change in **existing** protocols for them to live with the DRNI.
- The "localize vs. distribute" choice also drives whether a B-space Intra-Portal Link, an S-space IPL, or both, are necessary.
- **We maximize the utility of 802.1AX-REV by minimizing protocol interactions among distributed functions.**

CBP/PIP Localization Plans

- We can characterize "**localization plans**" by what services and stateful entities in the CBP/PIP pair(s) in the emulated third IB-BEB are **localized** to the same Portal System, and what are **distributed** among the Systems of a Portal.

  Per CBP: A CBP/PIP is localized, no matter what B-VLANs or services it operates. Different CBPs can be localized in different Portal Systems.

  Per Segment: All services assigned to the same pair of protection segments, no matter what B-VLAN or CBP they belong to, are localized together. Services assigned to different protection segment sets can be distributed.

  Per S-VLAN: Each S-VLAN is localized, but different S-VLANs in the same I-SID may be distributed.

- Other plans are possible (e.g. per B-VLAN) that do not offer fundamentally different tradeoff potentials, as the above do.

Localization Plan Dynamics

- The services localized by the localization plan in use can be **Active** in one Portal System and **Quiet** (or nonexistent) in all other Systems in the Portal. Depending on administrative configuration or events in the network or DRNI, Active conversations or functions can be shifted to another System in the Portal.

CFM Distribution

- Some localization plans result in a need to distribute a MEP or MIP across multiple Portal Systems.

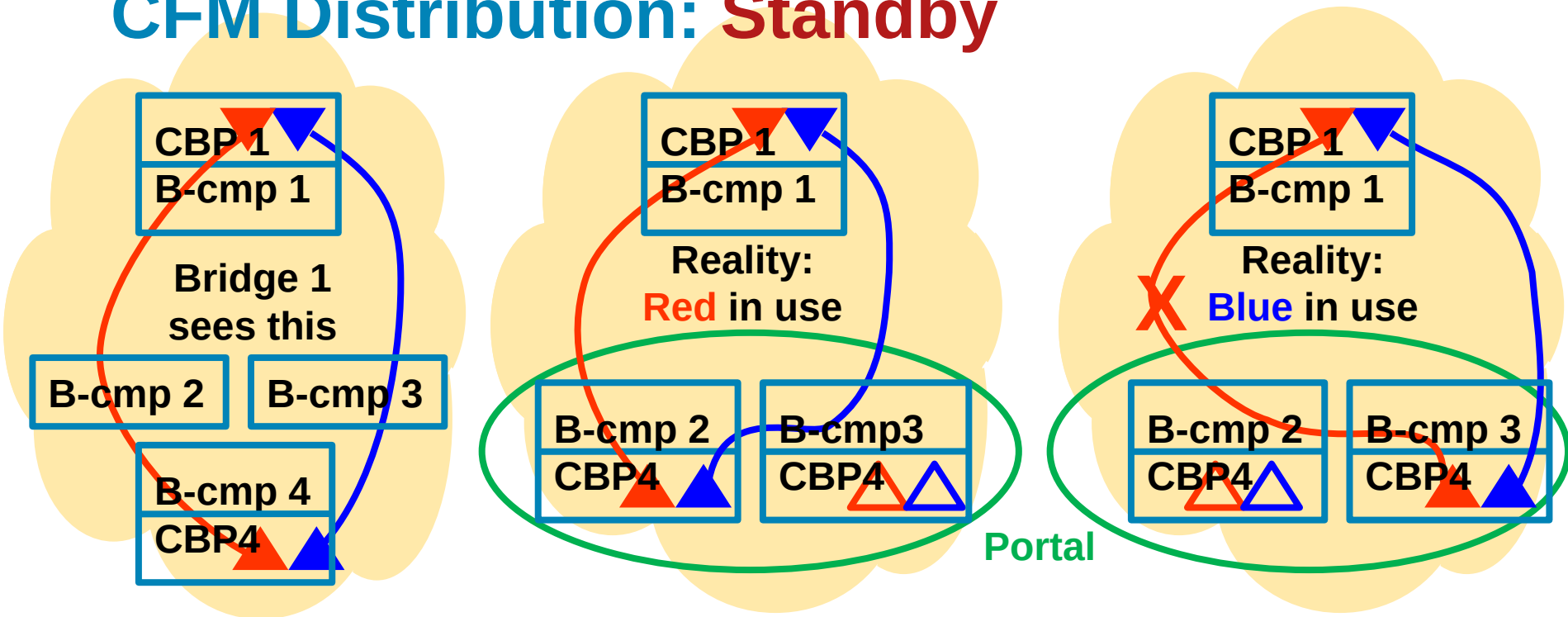
  For example, a CBP can have only one B-VLAN Up MEP (one whose frames carry no I-TAG) per B-VLAN. If the Per I-SID plan is used, and if it assigns two I-SIDs in a single CBP and single B-VLAN to two different Portal Systems, that Up MEP needs to be distributed, in some fashion.

- We will discuss two ways to distribute a MEP or MIP, **Standby** and **Shared**.

Standby CFM Distribution

- Duplicate the MEP/MIP in each Portal System as needed. At any given moment, only one System's MEPs are Active; the others are Quiet.
- All sets have the same MEPID (MEP only) and MAC address. (There may be cases where the MAC address could change, but that is not discussed, here.)
- Both sets of MEPs' MIBs are visible to the administrator.
- Duplication is largely transparent to other MEPs in the network. When a shift occurs, it appears to other MEPs that a very fast reboot happened; CFM sequence numbers and frame counters are reset.

CFM Distribution: Standby



CBP4 has Active MEPs on **B4** and **B6** in Portal System 2. A standby CBP4 with Quiet **B4** and **B6** MEPs is on Portal System 3. Data + CFM on ESP **B4** and only CFM on the ESP **B6**. The failure of ESP **B4 could** cause CBP4 to migrate; its MEPs become Quiet, and the System 3 CBP4 MEPs become Active.

D4	S1	B4	CFM

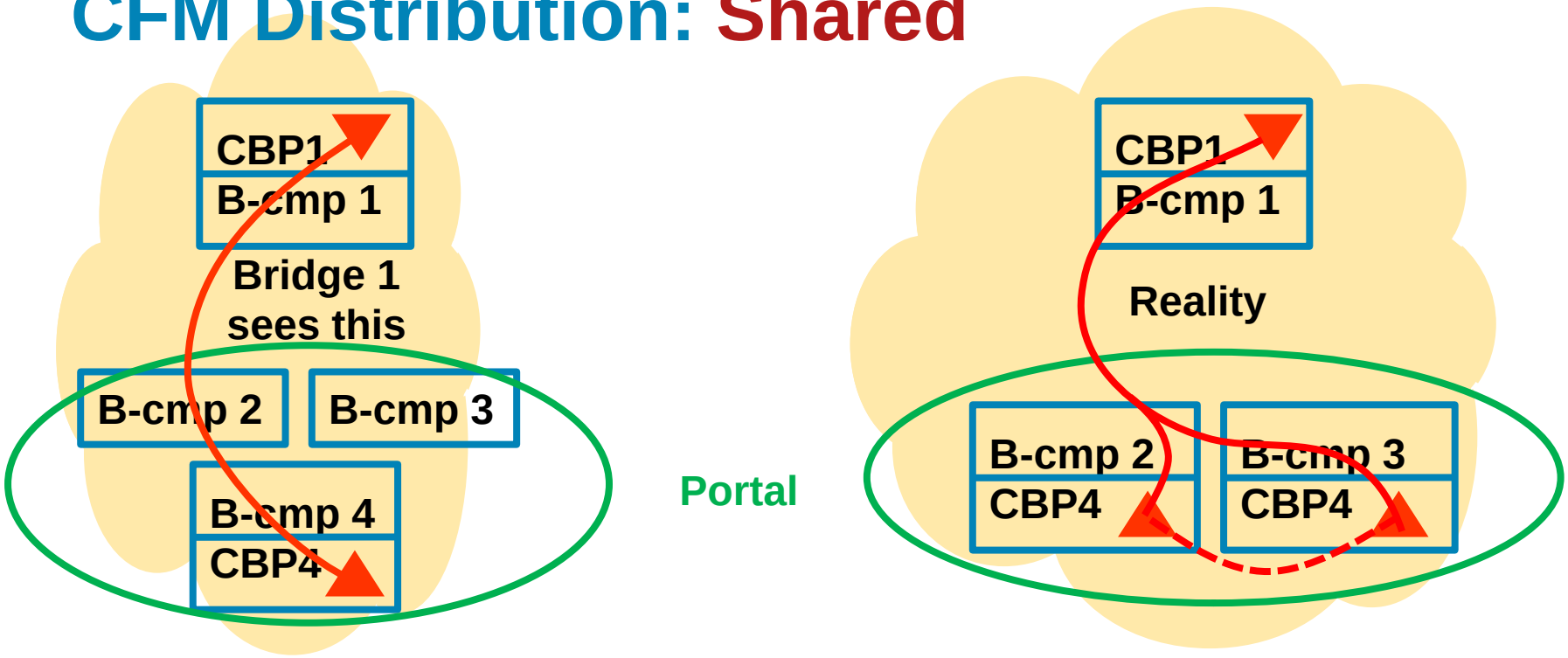
D4	S1	B4	I10	SD data

D4	S1	B6	CFM

Shared CFM Distribution

- There is an instance of the MEP/MIP in more than one System of the Portal.
- All are Active, but appear to the far end of the Maintenance Association to be a **single** MEP/MIP, of course with one MEPID and one MAC address.
- Some functions, e.g. acquiring statistics, are easy; the administrator could collect both sets and add them together.
- Others, e.g. figuring out who sends CCMs, are more difficult.
- Still others, e.g. frame loss measurements, require new or modified protocols.
- **We need specific proposals.**

CFM Distribution: Shared



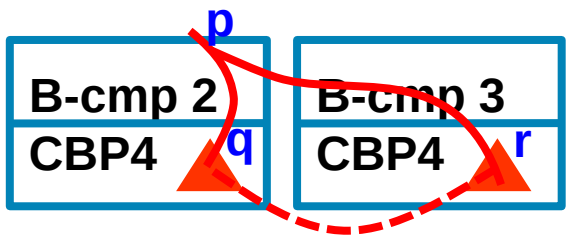
A TESI Up MEP (special I-TAG with no Customer addresses) is placed in a CBP that is distributed across two Portal Systems (some S-VLANs to one and some S-VLANs to the other). The two MEPs must look like one MEP, even for frame loss measurements.

D4	S1	B4	I10	SD data

D4	S1	B4	I10	00 CFM

D4	S1	B4	I10	00 CFM

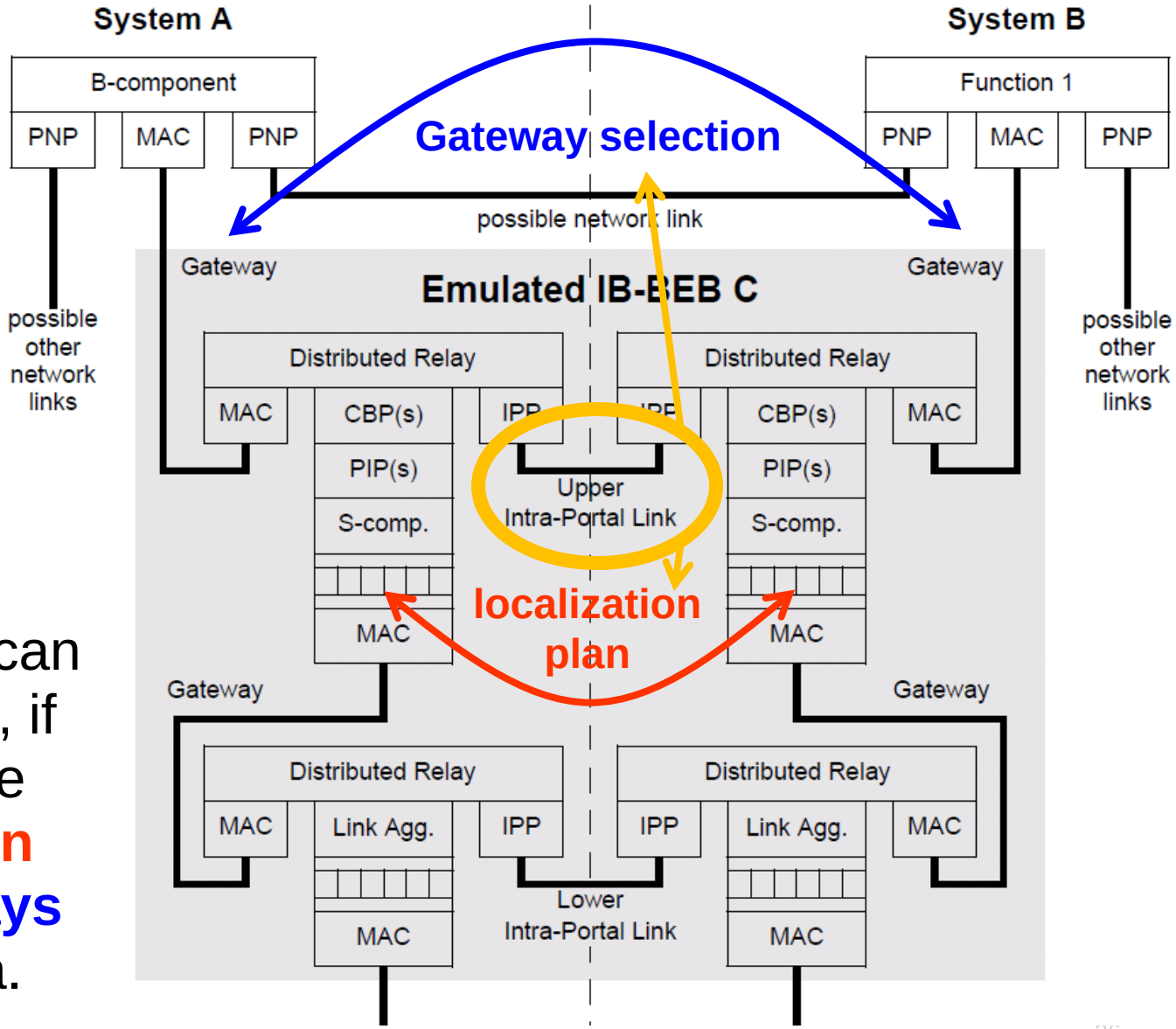
CFM Distribution: Shared



- A flow enters B2, and is split and delivered both down to the CBP and across the IPL to B3 and then down. The problem is that the last point in common to both branches is where the stream enters B2 (point **p**).
- A MEP at point **p** can count all the traffic (e.g. for Frame Loss Measurement), but that leaves the path **p**–**r** unprotected; a loss between **p** and **r** is undetected.
- Network management could combine statistics for MEPs at **q** and **r**, but how would Frame Loss Measurement OAM work across these points? What would transmission delay measure? (There may be answers to these questions. But, I don't have them.)

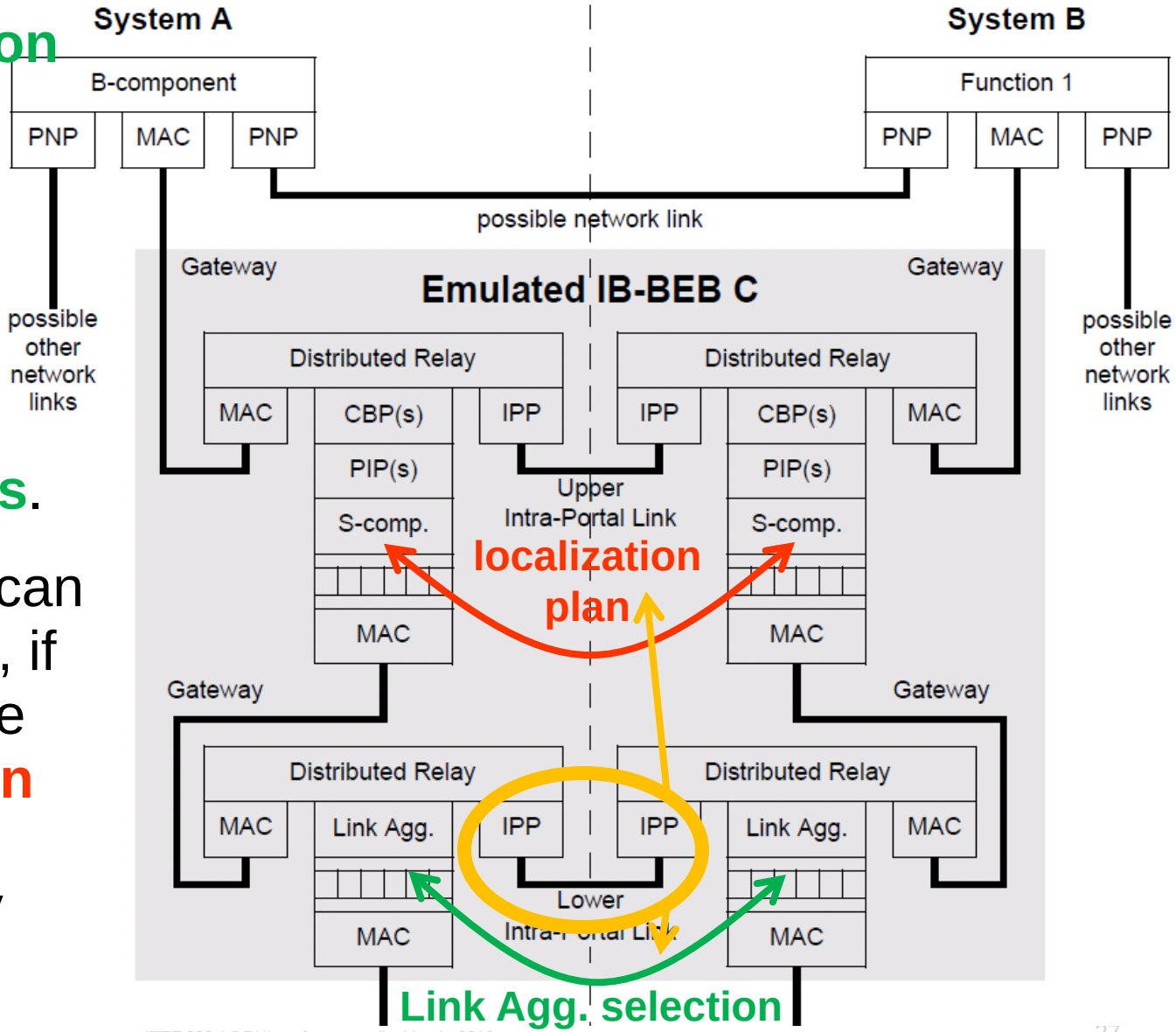
Conflicting conversation assignments

- The **network's fault recovery protocols** can dictate the criteria for assigning conversations to the upper **Gateways**.
- The **Upper IPL** can resolve conflicts, if any, between the **localization plan** and the **Gateways** by shuffling data.



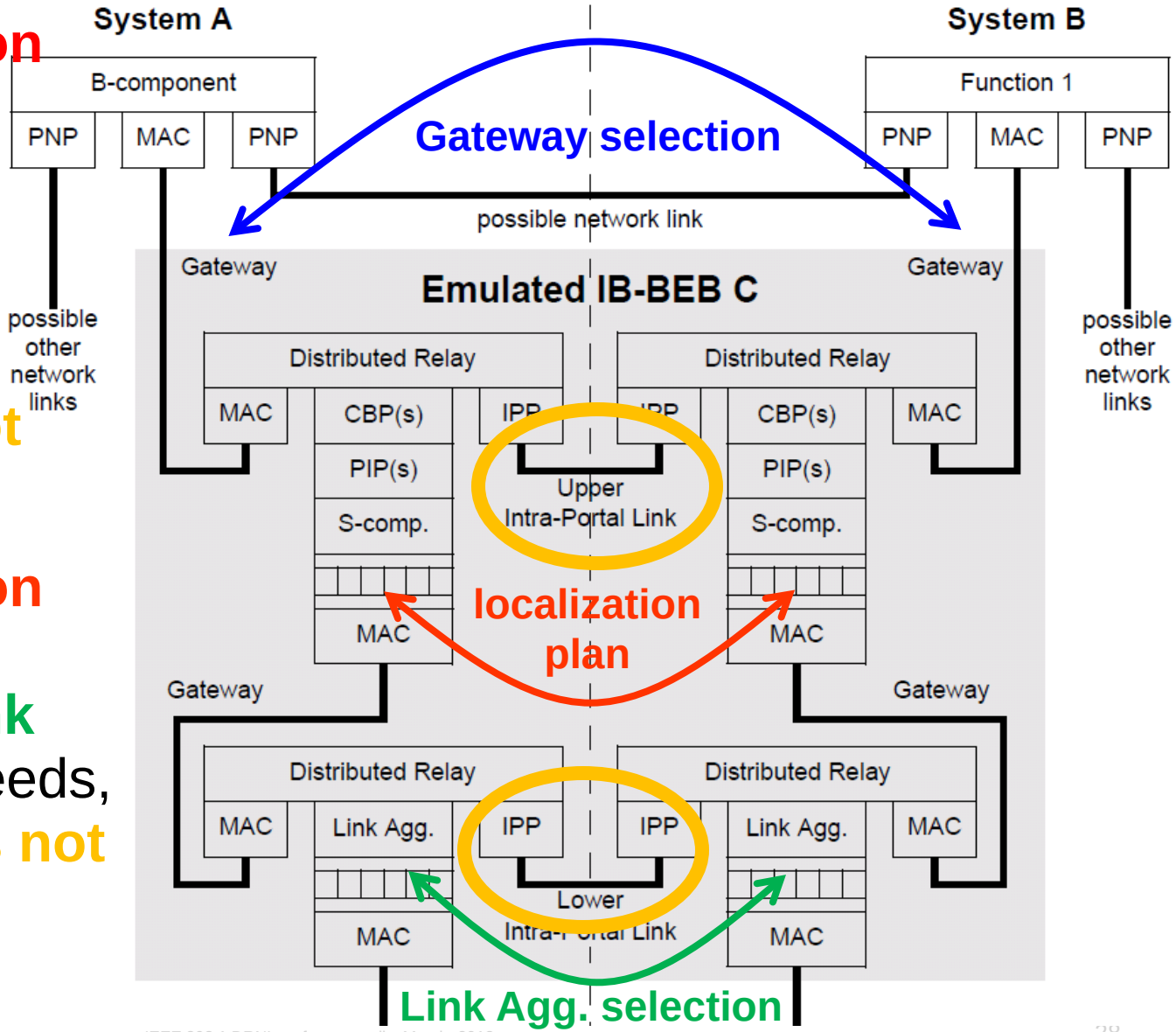
Conflicting conversation assignments

- **Link Aggregation** can, from the other network, dictate the criteria for assigning conversations to **physical links**.
- The **Lower IPL** can resolve conflicts, if any, between the **localization plan** and **Link Aggregation** by shuffling data.



Conflicting conversation assignments

- If the **localization plan** always matches the needs of **Gateway selection**, the **upper IPL is not needed**.
- If the **localization plan** always matches the **Link Aggregation** needs, the **lower IPL is not needed**.



REFERENCE model vs. VISSERS model

VISSERS model == REFERENCE model

Protection switching

- The Distributed Relay works in the same space as the B-component. There is no conflict between the TESI/Segment protection routes using the B-space IPL and those using the network link. So, in the data plane at any given moment, **it really doesn't matter** whether you call the link between A and B a **network link** or a **B-space IPL**. Vissers calls it a network link.
- It is useful, however, to make this distinction, because events in the DRNI result in **dynamic** changes to the routes taken ESPs and/or Segments. These dynamic changes are local to the Portal, and hidden from other systems in the Network, whereas ESP/Segment routes are **not** dynamic in the rest of the network. So, the difference between a "network link" and a "B-space IPL" is that the use of the former is **static and controlled by the network** (in terms of the **emulated System C**), and the latter is **dynamic and controlled by the DRNI.**
- By this definition, the Vissers model **does** have a B-space IPL.

VISSERS model == REFERENCE model

Dynamic Protocols

- When SPB or MSTP is controlling the network, and the upper Distributed Relay operates in B-space, Haddock's time sharing plan for controlling the use of a physical link shared between the network and the DRNI works fine. In the data plane, there is, again, no real distinction between the network link and a B-space IPL.
- And again, it is the emulated System C that makes it clear what is controlled by the network protocols (the network link) and what is controlled by the DRNI (the B-space IPL).

VISSERS model == REFERENCE model

- It is true that, in the data plane, there is no distinction between a (required, not optional) network link and a B-space Intra-Portal Link. But, it is still a useful fiction.

  It enables the network control protocol to operate on the basis of a relatively static emulated System C. Without this fiction, events in the DRNI may have to be exported to the state of the network control protocol, **at least** within the Portal Systems.

  One could, for example, omit the “network link” from the repertoire of links available to the network protocol, and the physical link can be entirely in the hands of the DRNI.

- That is, **the IPL/network link fiction enables the emulation of System C, and thus the clean separation of the network control protocol(s) from the DRNI**.

REFERENCE model vs. FINN model

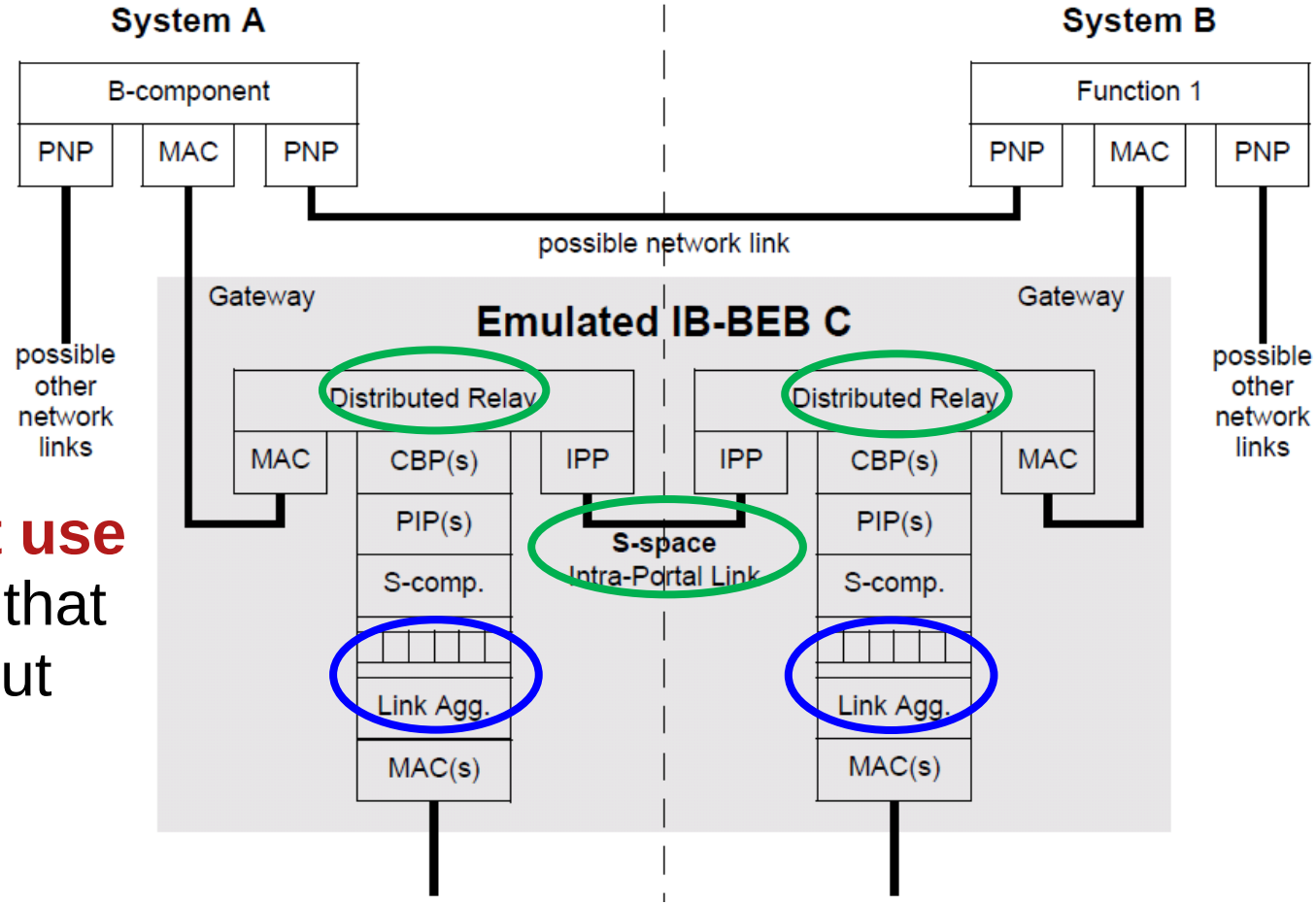
FINN model

- Because, in the Finn model, neither the CBP functions nor the Physical Link S-VLAN assignments are **always** aligned with the Gateways, it is possible for a frame to traverse both IPLs on its way up or down the stack. This excess IPL usage does not seem to be present in the Vissers or Haddock models.
- But, as we will see, this penalty need not be large, and localizing a CBP has advantages, as well.

REFERENCE model vs. HADDOCK model

REFERENCE Model vs. HADDOCK model

- The **Link Aggregation** layer **could be rolled up**, and Haddock does so.
- **But we will not use this picture** so that we can talk about choices more explicitly.



Comparing models Part 1

Localization Plan selection

This selection seems to me to drive most of the differences!

Finn model:

- The Localization Plan is chosen **solely** to avoid creating and implementing standards for the Share model for CFM distribution. Only a combination of Per CBP and Per Segment localization plans will work.

Vissers model:

- The Localization Plan is chosen to minimize intra-Portal traffic, while reducing Shared CFM. Several localization plans will work.

Haddock model:

- The Localization Plan is chosen to match the needs of Link Aggregation, in order to avoid creating and implementing a second Intra-Portal Link. The Per S-VLAN localization plan works, of course.

Gateway selection

All models:

- If the network protection is MSTP or SPB, then upper Gateways are selected by B-VLAN. If a Gateway is lost, the B-VLAN must move to another Portal System, else the B-VLAN and its traffic are lost.
- If the network uses TESI Protection, upper Gateways are selected by ESP, and never move; a lost Gateway is a set of lost ESPs and their services.
- If the network uses Segment Protection, upper Gateways are selected by Segment, and never move; a lost Gateway is a set of lost Segments and their services.

Distributed Relay / IPL usage

Finn model:

- The upper IPL distributes data between CBPs and Gateways as needed. The lower IPL distributes data between CPB/PIPs and Physical Links. Both must be distinguished somehow from the network link.

Haddock model:

- The upper IPL is used, and in some cases, data on the upper IPL must be distinguished somehow from data on the network link. The upper Distributed Relay, if excessive intra-Portal traffic is to be avoided, operates on the S-VID. The lower IPL is not used.

Vissers model:

- The upper IPL is used for data, but there is never any confusion between IPL data and network link data (but see below, CFM distribution). The upper Distributed Relay operates in B-space. The lower IPL is also used for data, and some means is required to differentiate it from data on the network link and/or upper IPL. The lower Distributed Relay operates in S-space.

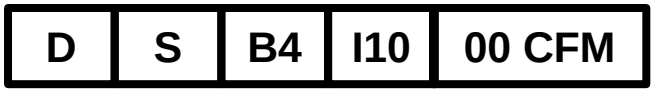
CFM distribution

- The three models differ significantly with regard to CFM issues.
- We will look at three MEP positions, and see what the differences among the models are:

ESP/Segment MEPs:

DRNI per-S-VLAN MEPs

TESI MEPs:



CFM distribution: ESP/Seg MEPs

Finn model:

- The localization plan always places **whole CBPs together** in one Portal System; a **CBP is never split** across Portal Systems.
- If Segment Protection is used, it would be possible for a CBP to be split among multiple Segment pairs. In this case, the closure of the interconnected Segments and CBPs must **all** reside in a single Portal System.
- Since a CBP or Segment is never distributed, an ESP MEP or a Segment **MEP is never distributed**, and the **Standby mode** for these MEPs is always possible.

CFM distribution: ESP/Seg MEPs

Haddock and Vissers models:

- The localization plan can split CBPs.
- The **Shared** model for ESP and Segment CFM distribution must be implemented.
- At least the **Haddock** model can split a single TESI across Portal Systems, thus requiring the **Shared** model for TESI CFM, as well. The **Vissers** model could keep a TESI MEP in a single Portal System, and thus use the **Standby** model.

CFM distribution: S-VLAN MEPs

Finn model:

- In order to avoid Shared MEPs, the S-VLAN MEPs are above Link Aggregation. not per-MAC. They operate in Standby mode.

Haddock/Vissers models:

- Per-MAC-per-S-VLAN MEPs are used. These can operate in Standby mode when talking to another per-VLAN Aggregation System or Portal, but must be Shared MEPs if talking to another kind of conversation distribution, such as per-IP-5-tuple.

CFM distribution: TESI MEPs

Finn model:

- **Always Standby**, because a CBP is never distributed.

Vissers model:

- Not clear to me. I think **always Standby**.

Haddock model:

- Since a TESI can have multiple S-VLANs that can be split across the Portal Systems, **Shared MEPs are required**.

Summary

Summary

model	IPL usage	IPL encaps	Virtual CFM	IPL usage	Per-SVID MEPs
Finn	upper+lower	upper+lower	Standby for all	More*	per S-comp
Vissers	upper+lower	lower only	Shared ESP/Seg./ SVID, Standby TESI	Less	per MAC or per S-comp
Haddock	upper only	upper only†	Shared for all	Less	per MAC

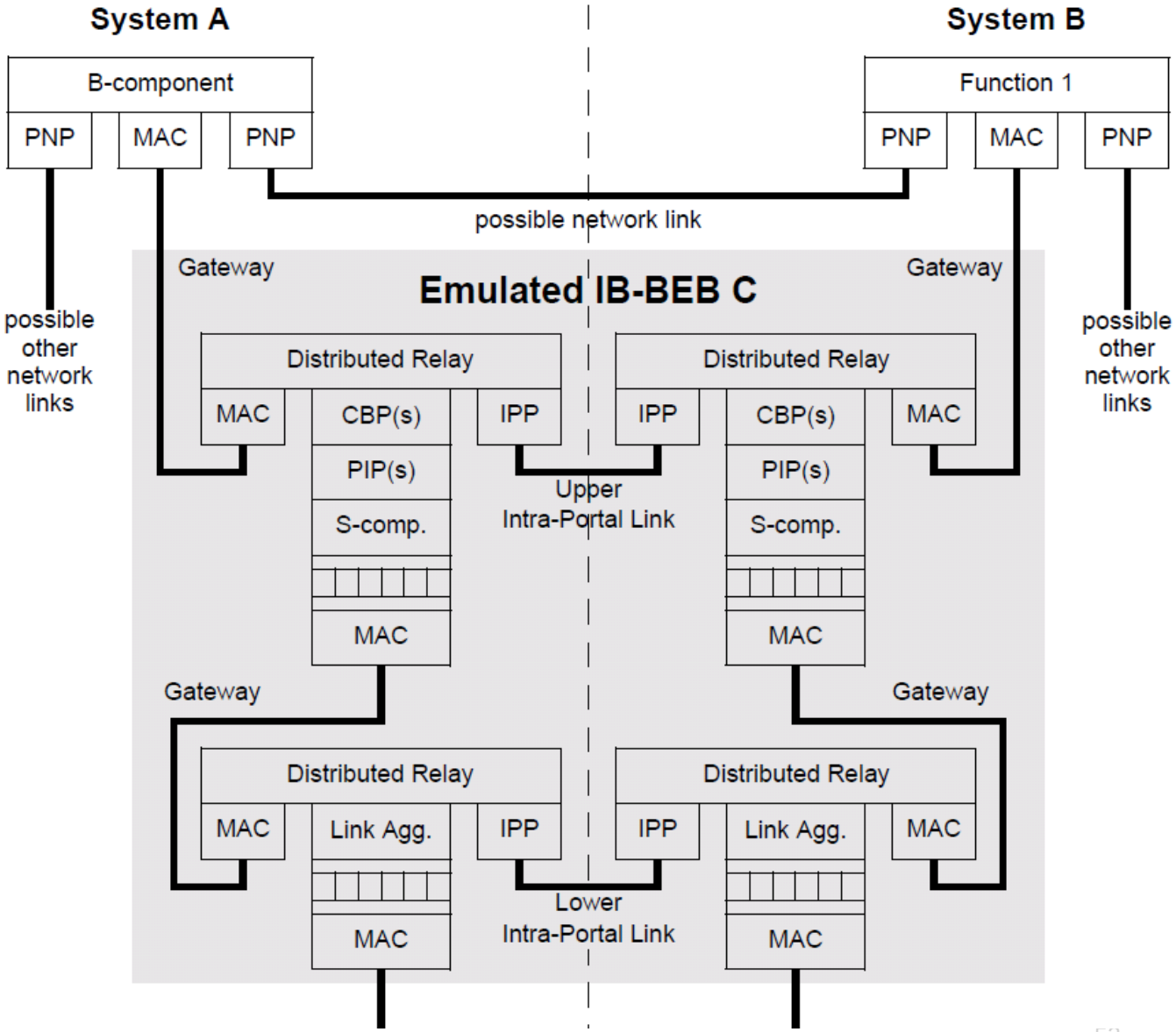
***** Excess upper IPL usage can be mostly eliminated with proper configuration of CBPs.
† A channel equivalent to a control-only lower IPL can be needed for S-VLAN MEP coordination.

- We clearly must create/modify a protocol to support the distributed Link Aggregation state machines.
- We clearly must supply handles so that hot standby movement of other state machines is possible.
- We have a choice whether to create/modify a protocol to support distributed CFM state machines.
- **We cannot decide how to proceed until we have looked at what it takes to do a Shared MEP.**

Footnote: IPL bandwidth in Finn model

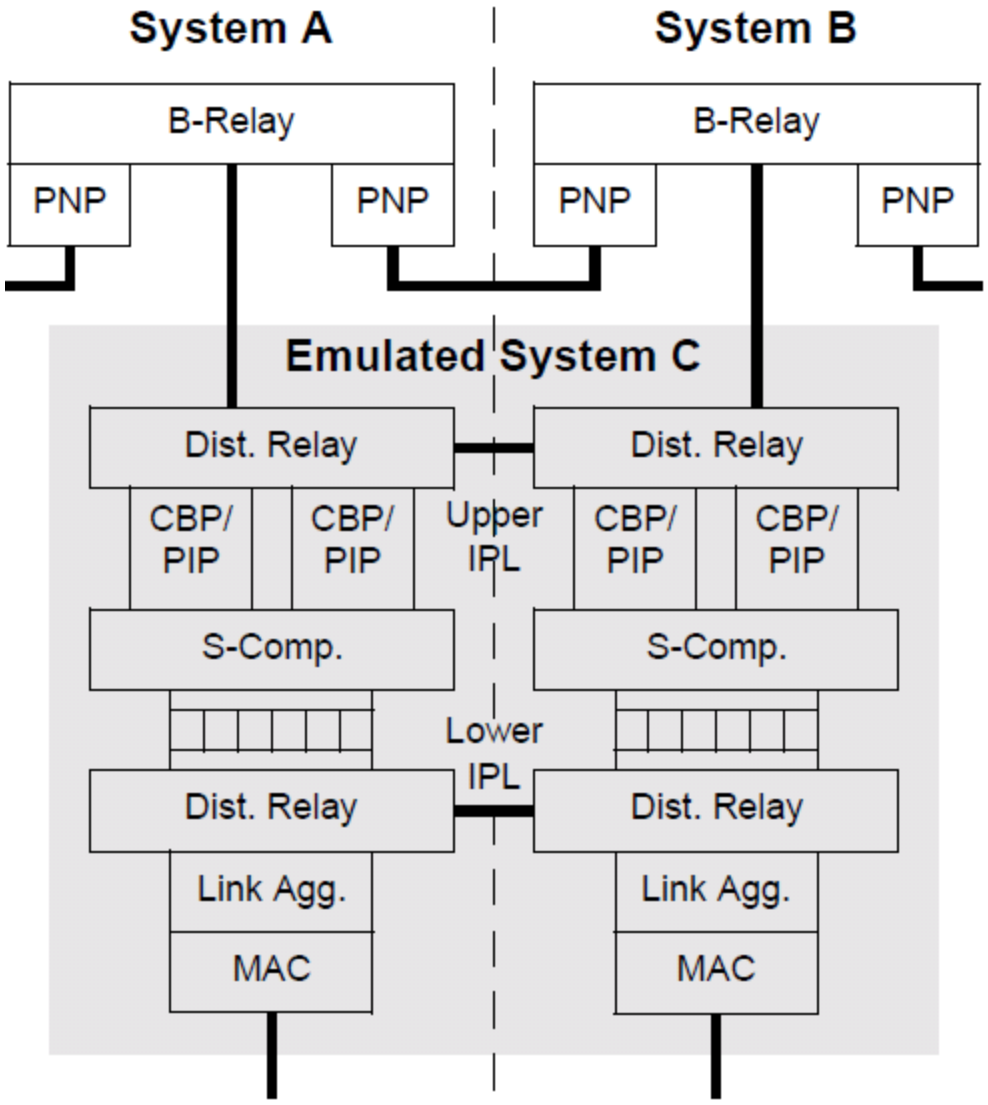
Finn model: IPL bandwidth

- Let us use a simpler version of the reference diagram:



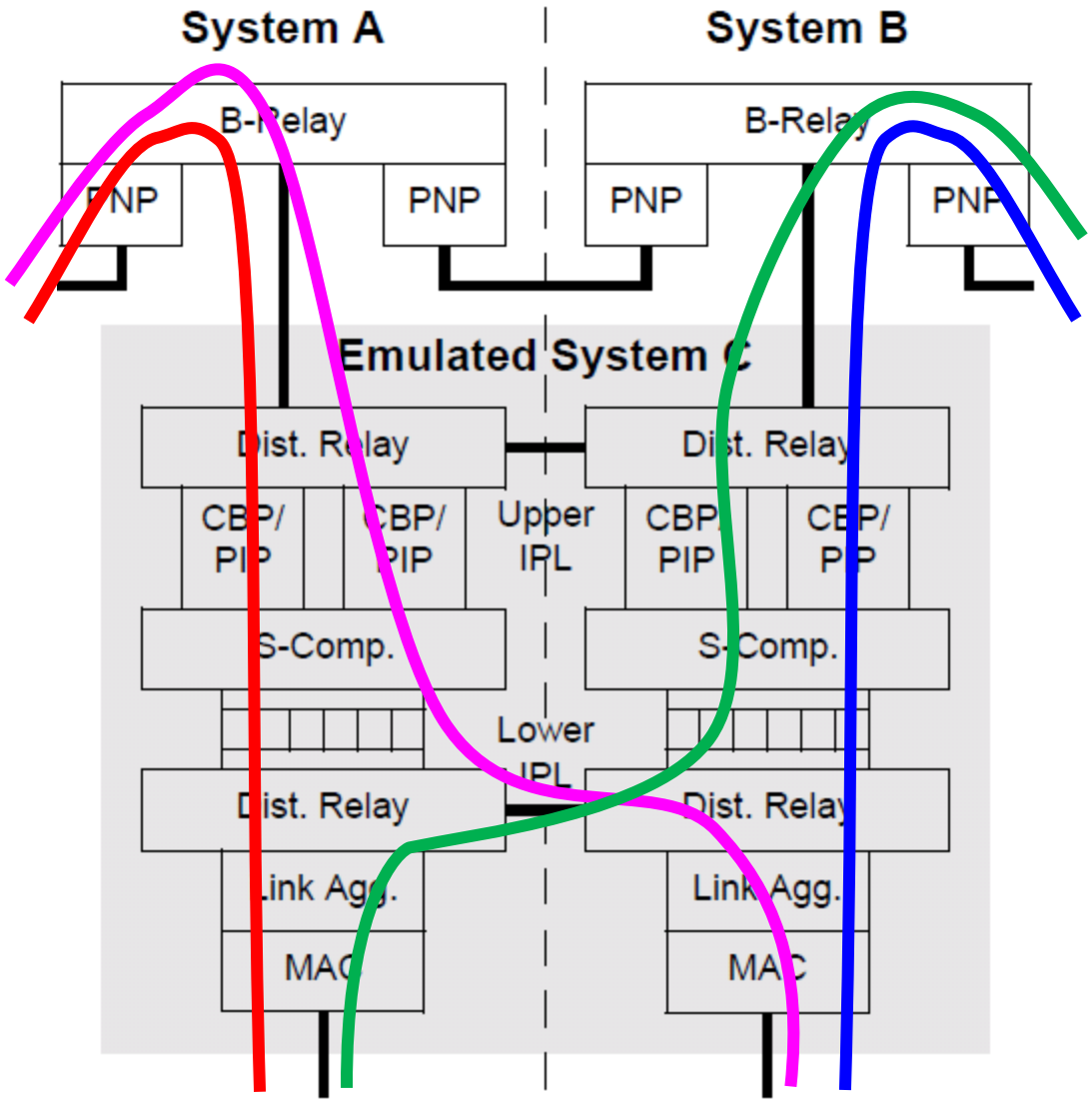
Finn model: IPL bandwidth

- Let us use a simpler version of the reference diagram, but one that shows four CBP/PIP pairs.
- (We'll worry about the lower MEP placements later.)



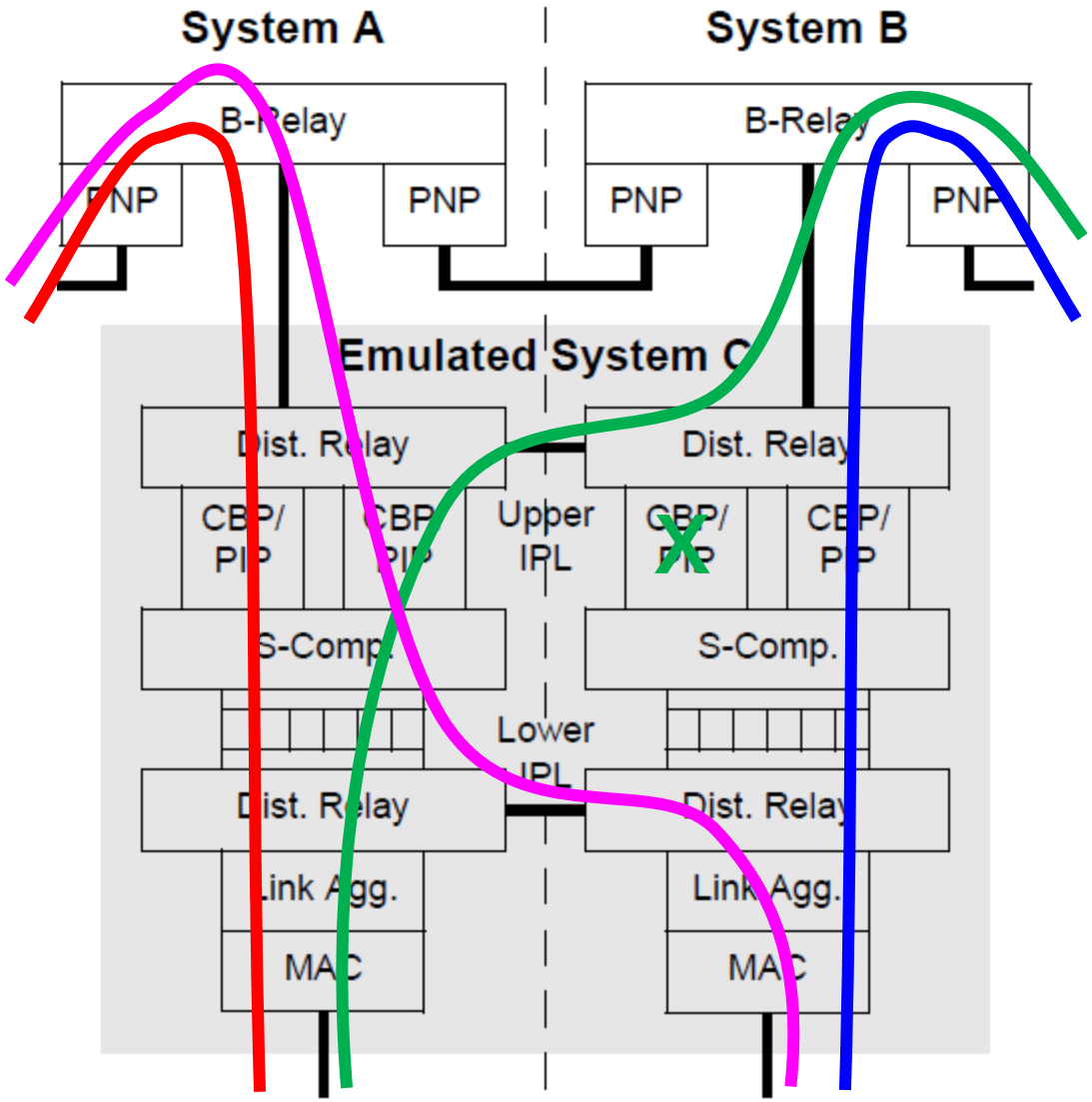
Finn model: IPL bandwidth

- You would normally configure three or four CBPs for:
  - **Left flows**
  - **Right flows**
  - **A-B criss-cross**
  - **B-A criss-cross**
- **Note that there is no excess IPL traffic.**



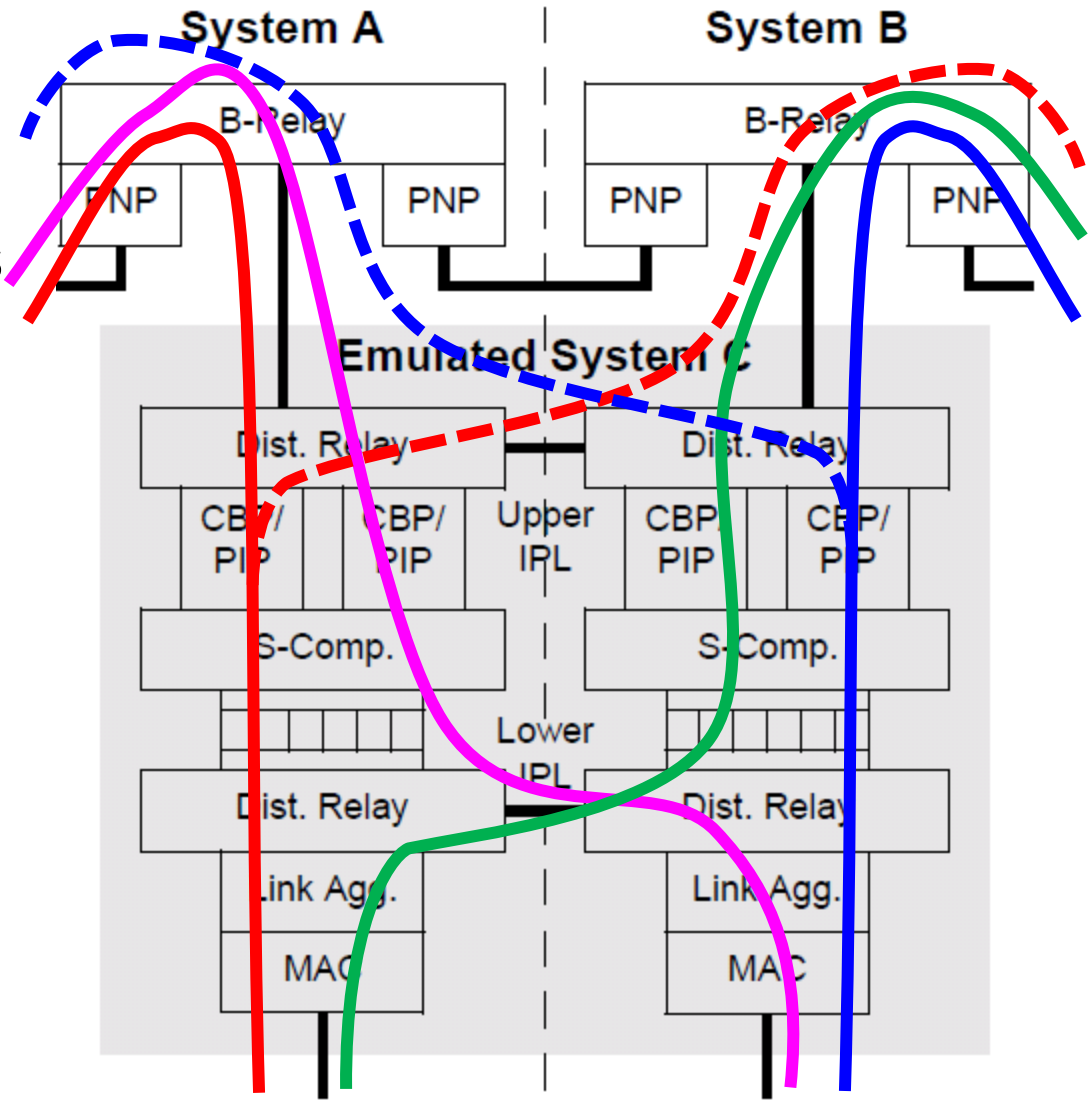
Finn model: IPL bandwidth

- (One **criss-cross** CBP is sufficient, but two give more flexibility.)



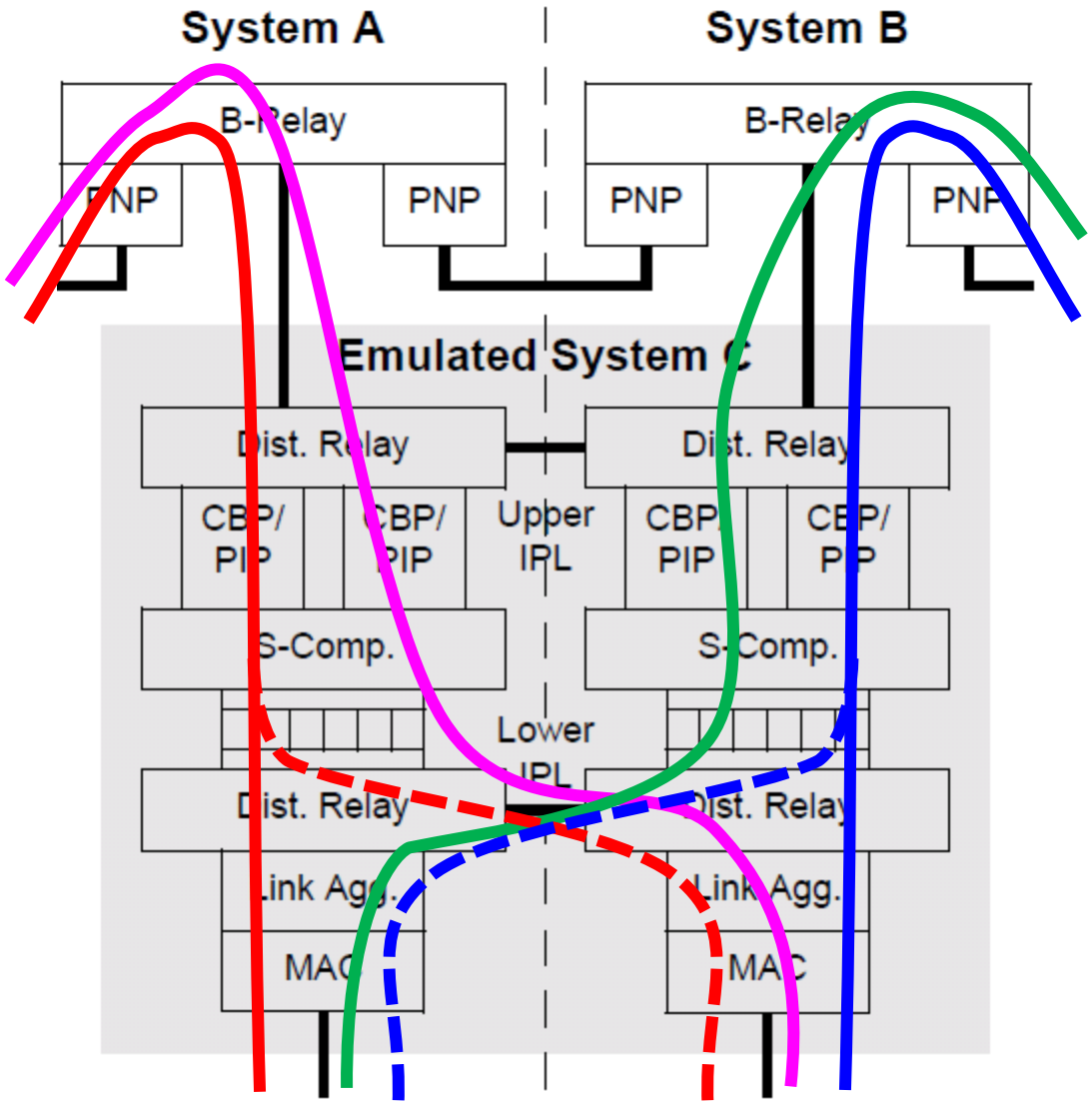
Finn model: IPL bandwidth

- A failure of one of the (presumably many) ESPs in the **Left** or **Right** CBPs would trigger the use of an alternate ESP that would use the Upper IPL.
- **This is not "excess bandwidth".** This is necessary for any plan.



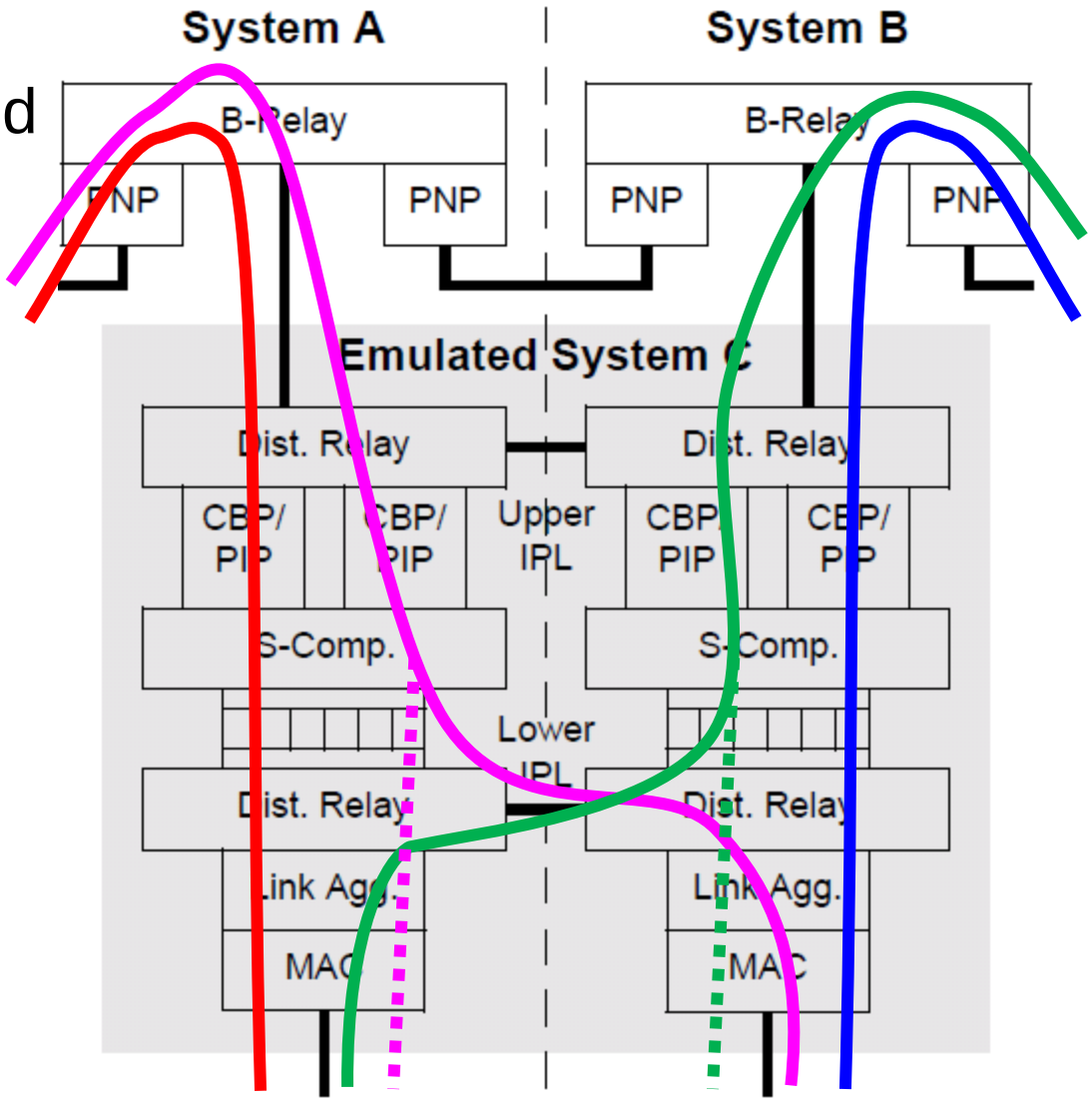
Finn model: IPL bandwidth

- Similarly, a failure of a DRNI link or a movement of an S-VLAN to another DRNI link on the **Left** or **Right** CBPs does not generate any excess IPL bandwidth.



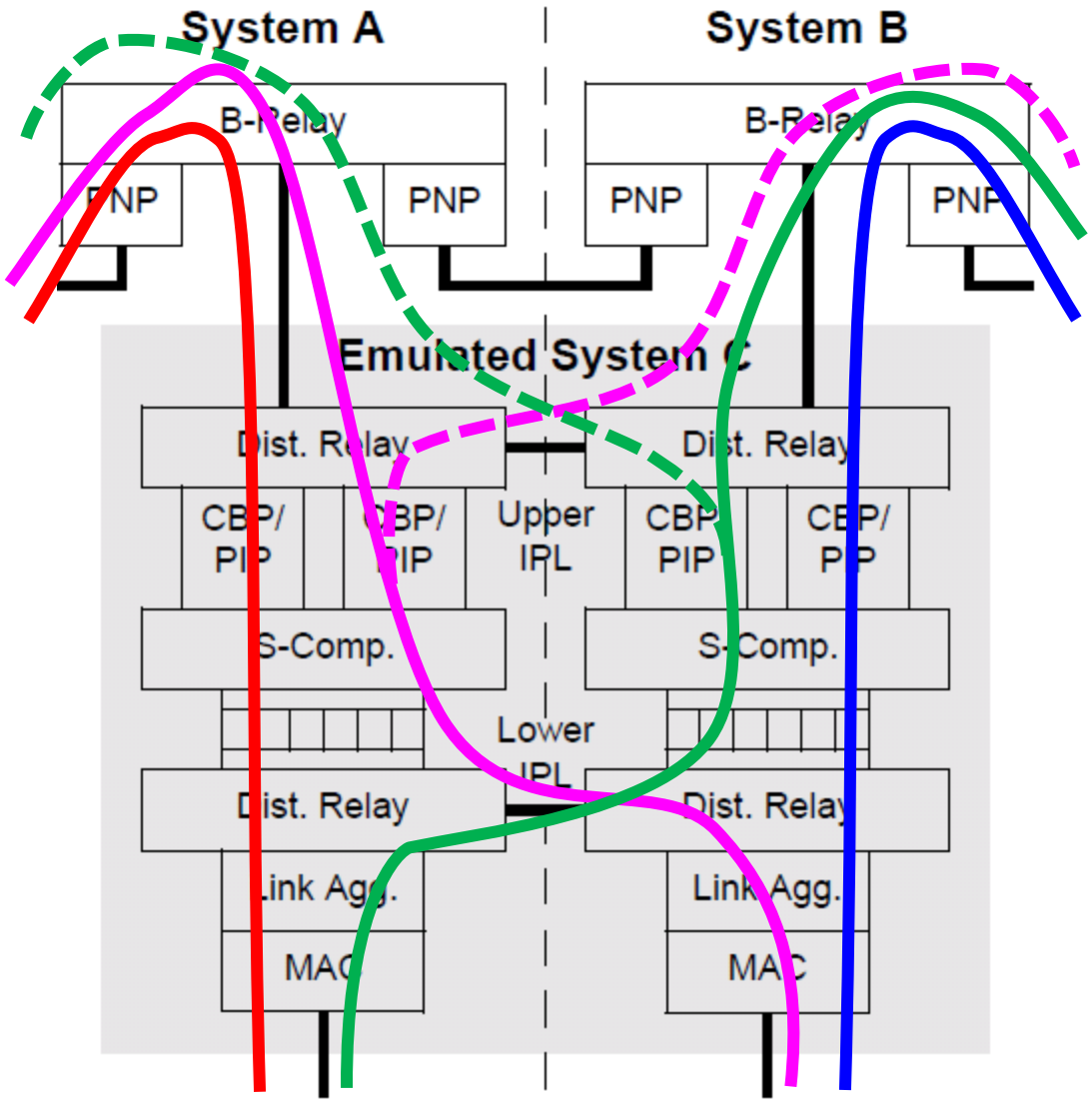
Finn model: IPL bandwidth

- A failure at the “better” end of a **criss-cross** link actually **improves** the situation.



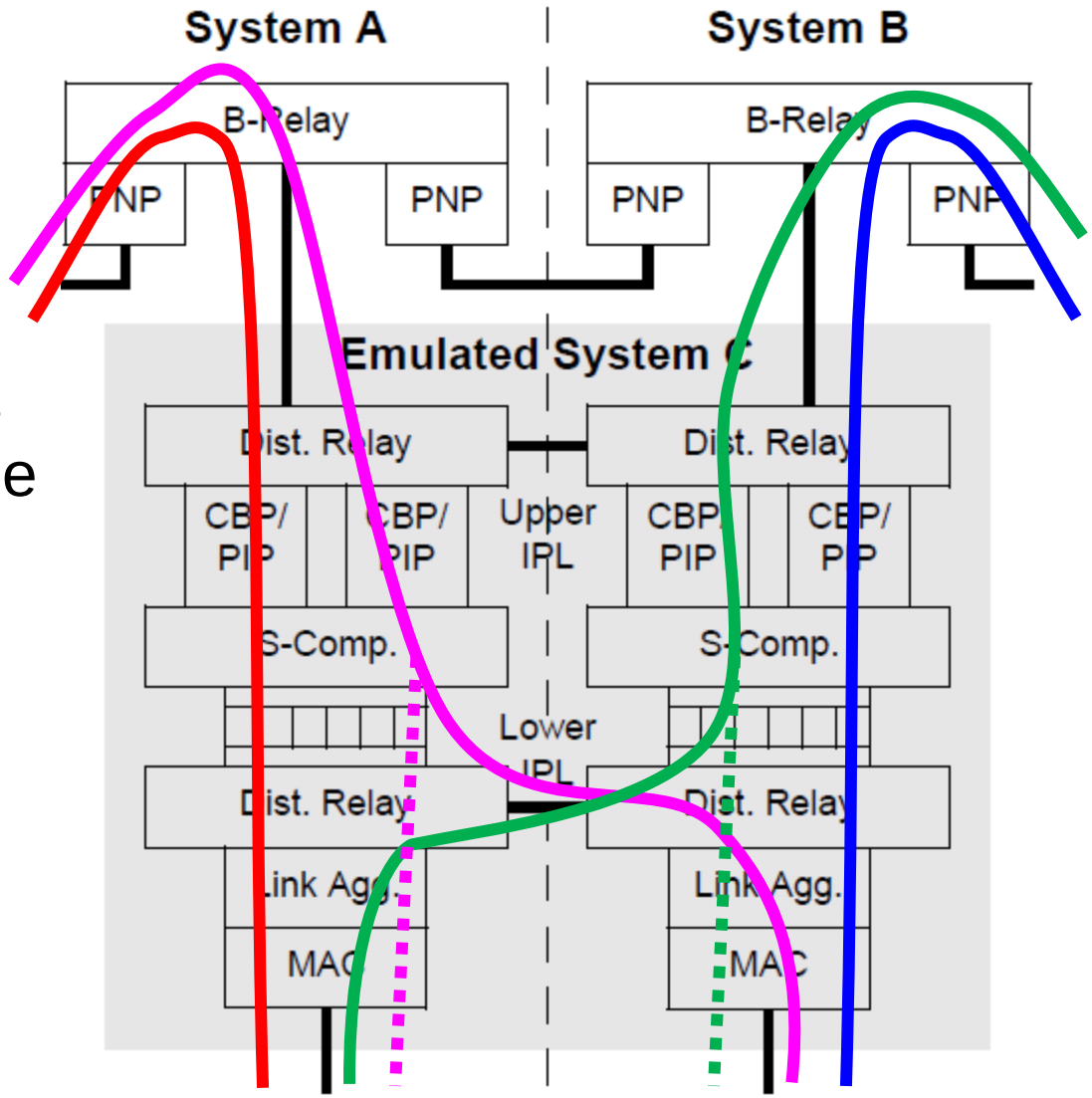
Finn model: IPL bandwidth

- **Only** a failure at the "worse" end of a **criss-cross link** creates **excess IPL bandwidth**, in that frames in the failed service make two trips between the Portal Systems.



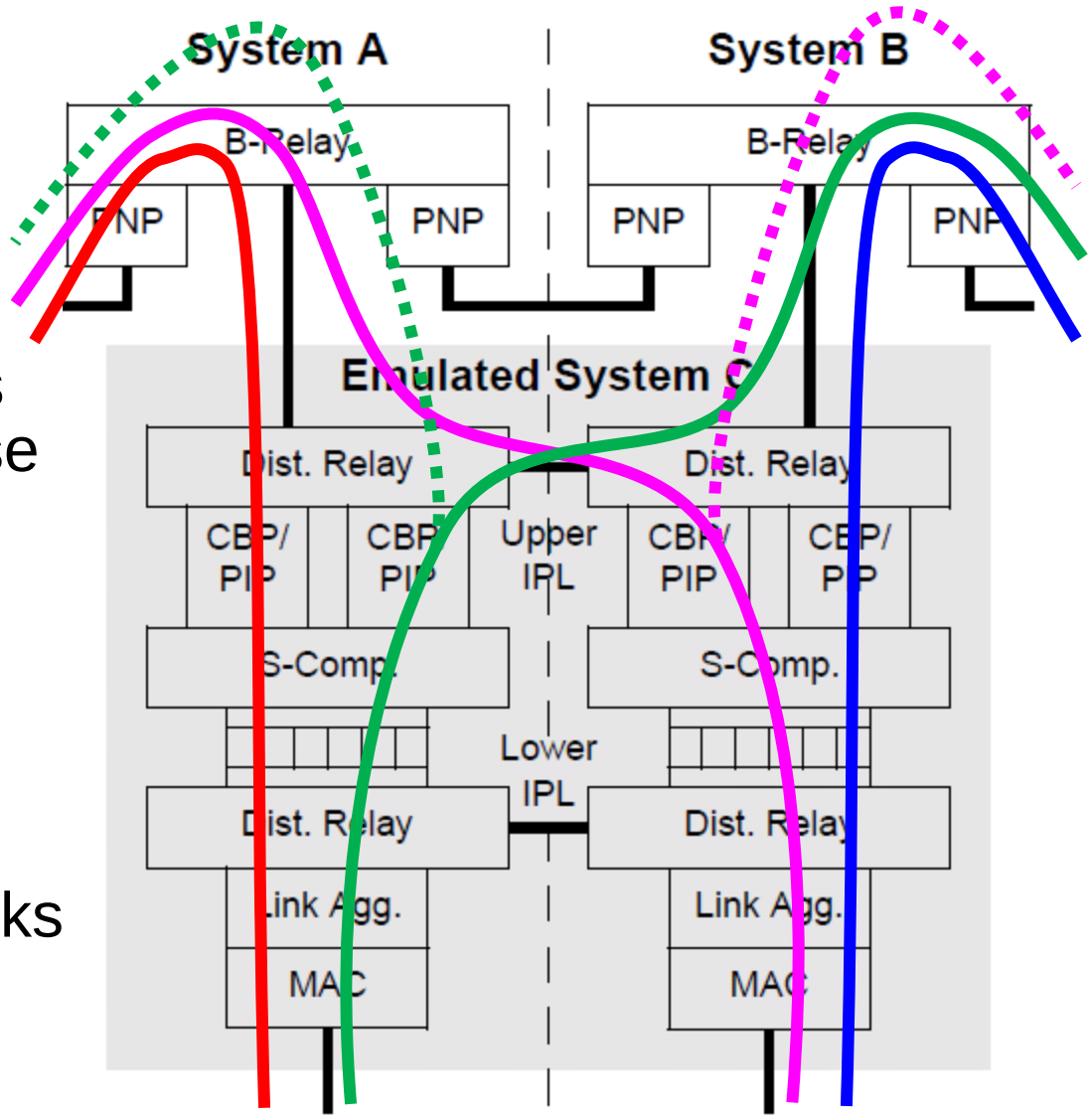
Finn model: IPL bandwidth

- If one expects far more failures at one end that at the other end (network end vs. LAG end), then one should arrange one's "**criss-cross**" CPBs to use the end that is expected to fail most often.
- Failures at the expected end improve, rather than hurt, IPL usage.



Finn model: IPL bandwidth

- If one expects far more failures at one end that at the other end (network end vs. LAG end), then one should arrange one's "**criss-cross**" CPBs to use the end that is expected to fail most often.
- Failures at the expected end improve, rather than hurt, IPL usage.
- Presumably, the DRNI links fail more often, so **this is not the typical picture**.



Finn model: IPL bandwidth

- The other models avoid this situation, at the cost of Shared MEPs.
- However, we should note that, because there are no Shared mode MEPs, and because the sharing protocol necessarily adds to the failover time, the **failover times** of the Finn model do not require any more changes in state or alterations of the forwarding tables than for the non-DRNI case, and hence **will be faster** than for the other models.