Lightweight Network Network Interface

Using Link Aggregation

Rev. 1

Norman Finn

nfinn@cisco.com

ENNI: Heavy or light?

- There are at least two distinct methods we can pursue for defining an Ether NNI:
- **Heavy**:

  A Buffer Network is built along the lines suggested in new-nfinn-buffer-networks-0310-v01.pdf with an explicit data encapsulation.

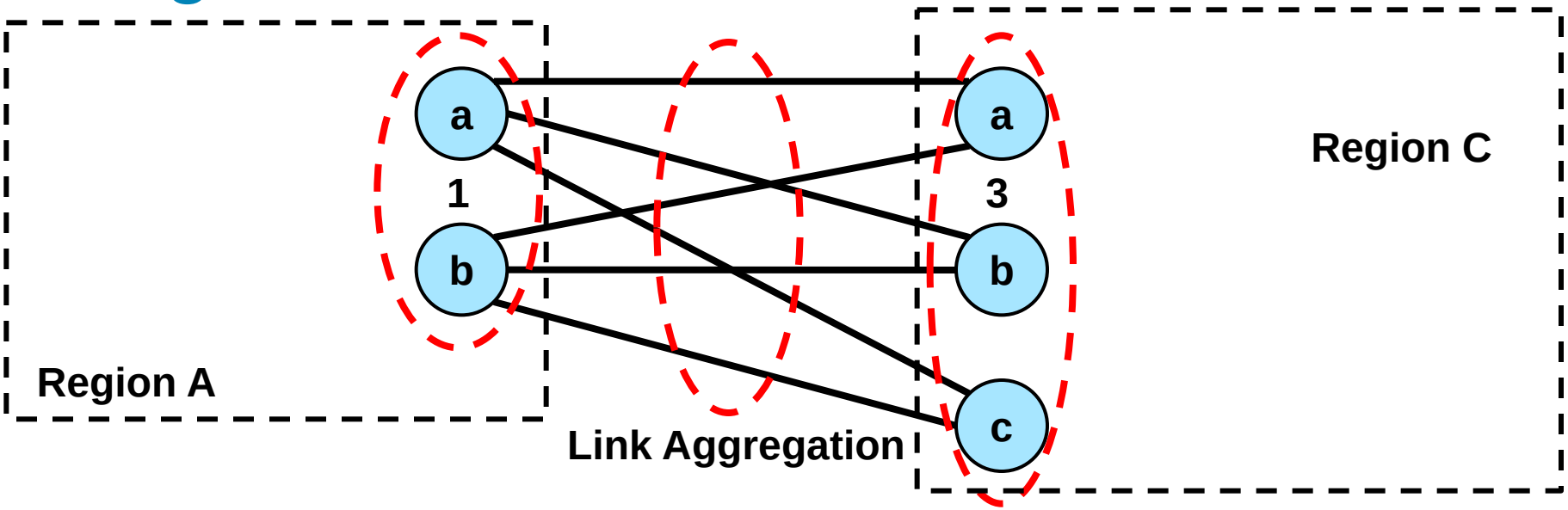
- **Light**:

  Buffer Network is built using “virtual nodes,” i.e. the multiple physical Nodes of each Portal cooperate to give the appearance of a Portal consisting of a single Node. This present document is new-nfinn-light-nni-0710-v01.pdf.

- Each method has its advantages and drawbacks, and all of the drawbacks can be addressed. This is a classic engineering decision.

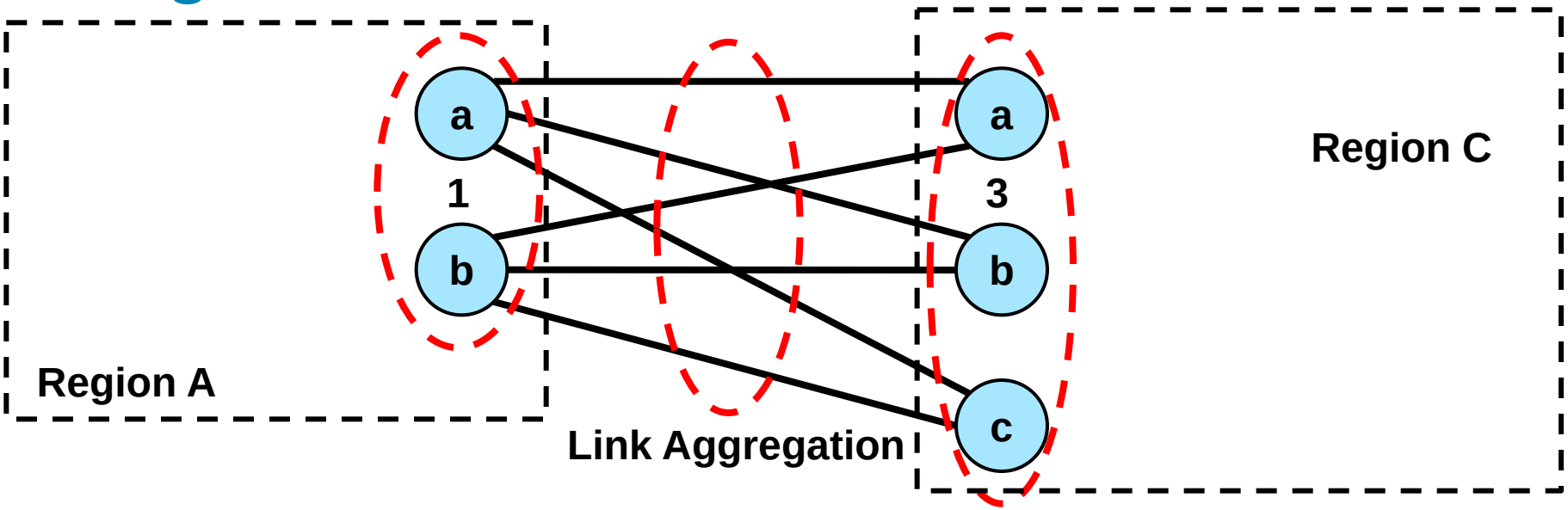
Light ENNI

Light ENNI



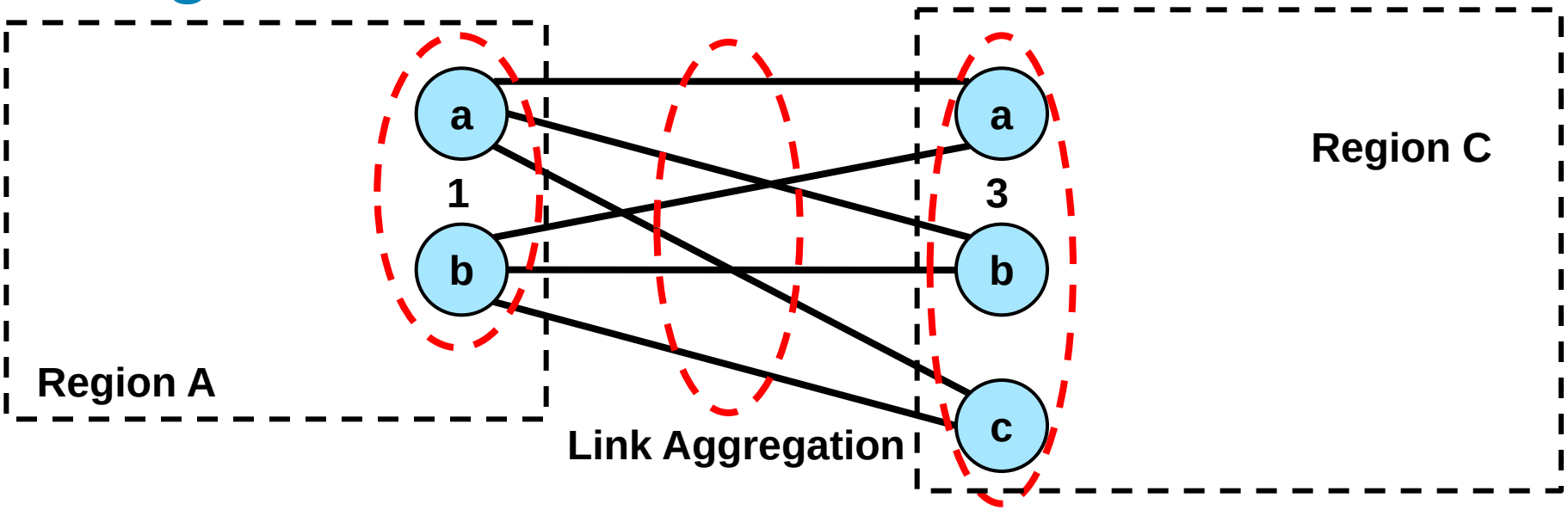
- The Terminal Nodes in each Region appear, to the other Region, to be a single Terminal Node (bridge, switch, or whatever).
- All of the inter-Region Links are combined into a single Aggregated Link using LACP.
- Links among Nodes in the same Region are invisible and irrelevant to the ENNI.

Light ENNI



- The means by which the Virtual Terminal Nodes are implemented does **not** need to be standardized; this author sees no requirement for 3a, 3b, and 3c to come from three different vendors.
- The choice of physical link is always up to the transmitting Virtual Terminal Node, and the receiving Virtual Terminal Node must live with the choice.

Light ENNI

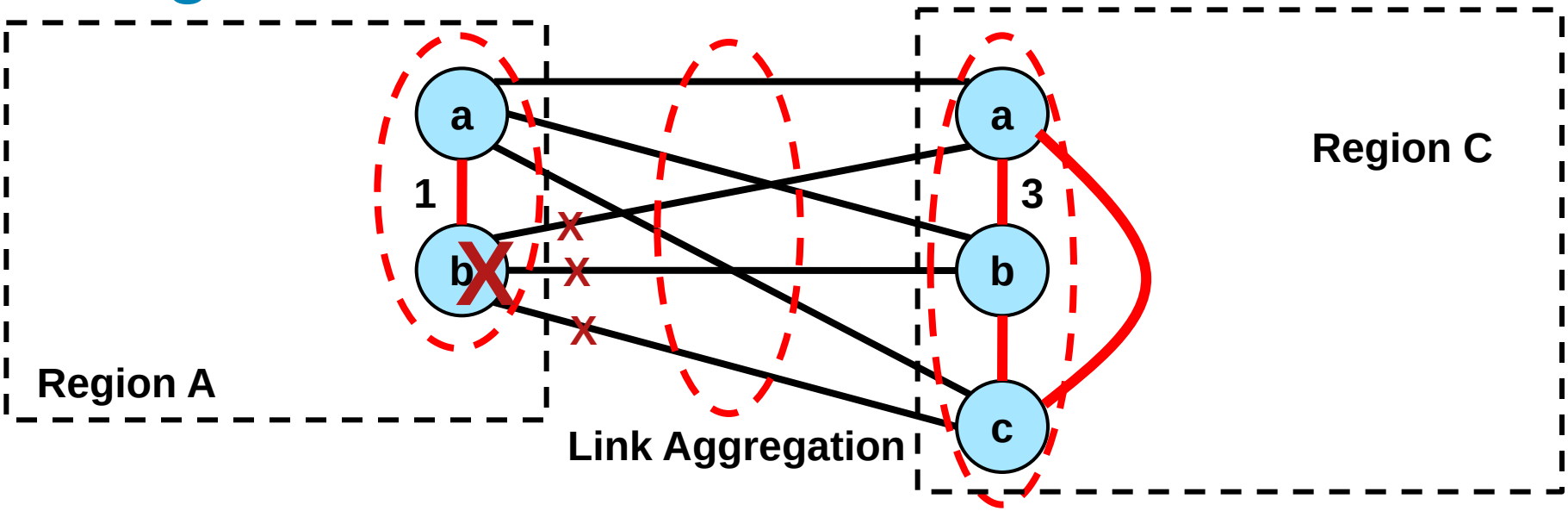


- Physical level CFM can be used to improved failure detection time for the physical links.
- Obviously, the two Regions have to agree on a data encapsulation, but a 1:1 service encapsulation translation can be performed at either (or both) ends, and no encapsulation-dependent CFM is required.

Light ENNI – issues

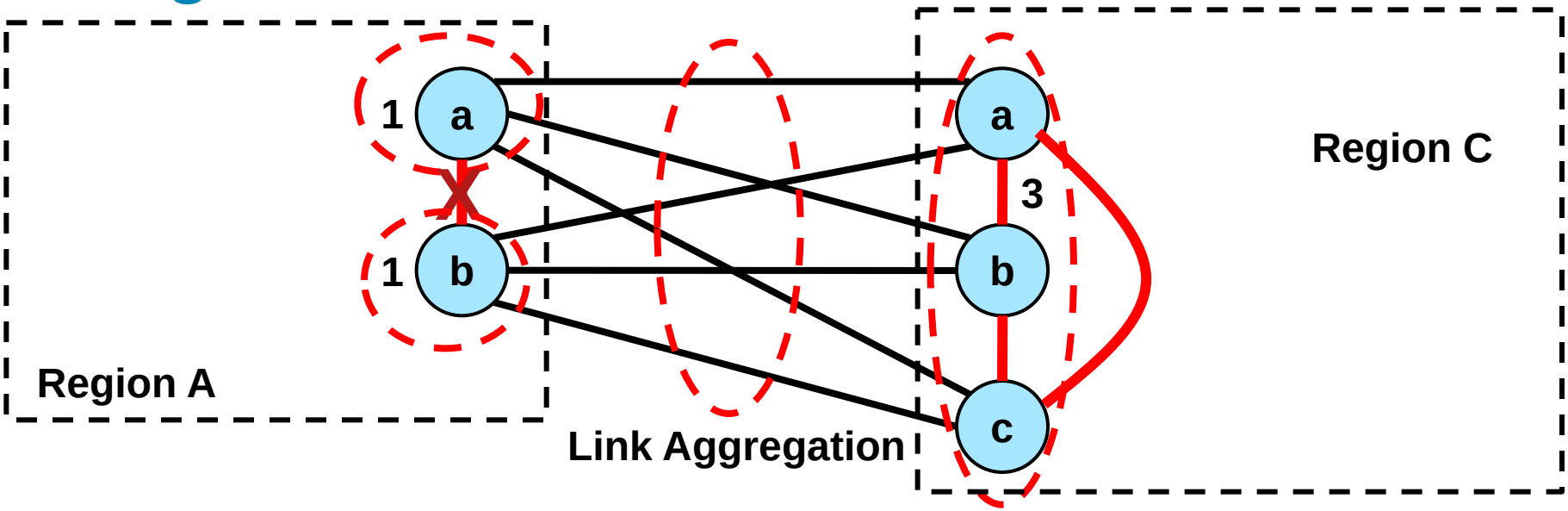
- Clearly, service-based physical link selection is preferred to other methods, e.g., hashing the IP 5-tuple.
- For efficiency of routing, a means (perhaps LACP extensions) should be provided for one Region to express a preference (not a demand) for which link should be used for which service.
- For optimum maintainability, it we should provide a means (perhaps LACP extensions) for the two Virtual Terminal Nodes to agree to use the same physical link for both directions for a given service (or bundle of services).

Light ENNI – issues



- The physical components of a Virtual Terminal Node appear to be a single Node to (at least) the other Region, and perhaps to other Nodes internal to the Region, as well.
- If one component of a Virtual Terminal Node **fails** (say 1b) then its attached Links fail, but the remaining Nodes (1a in this case) continue to function; recovery is quick.

Light ENNI – issues



- If the **Link** between two components of a Virtual Terminal Node (e.g. 1a-1b) fails, both components can takeover the Node's identity, but act independently (the "**split brain**" scenario), with disastrous results.
- For this reason, "inter-VTN links" are made extra-reliable, and in some implementations, are assumed to be failure-proof.

Light ENNI – issues

- We cannot (in the author's opinion) design a network standard around "failure proof links".
- Since we are assuming that LACP is being used to establish Aggregated Links between Virtual Terminal Nodes, we could **enhance LACP** so that the devices connected to a Virtual Terminal Node can assist the VTN in detecting a "split brain" scenario.
- But, split brain detection is necessarily a hippity-hippity-hop operation, involving multiple Nodes; there is no equivalent to the (3c –) 3b – 5b – 1b (– 1a) Maintenance Association described for the Heavy ENNI. Split brain detection will be **slower** than MA failure detection.

Light ENNI – issues

- Recovery from the split brain is up to the implementation:

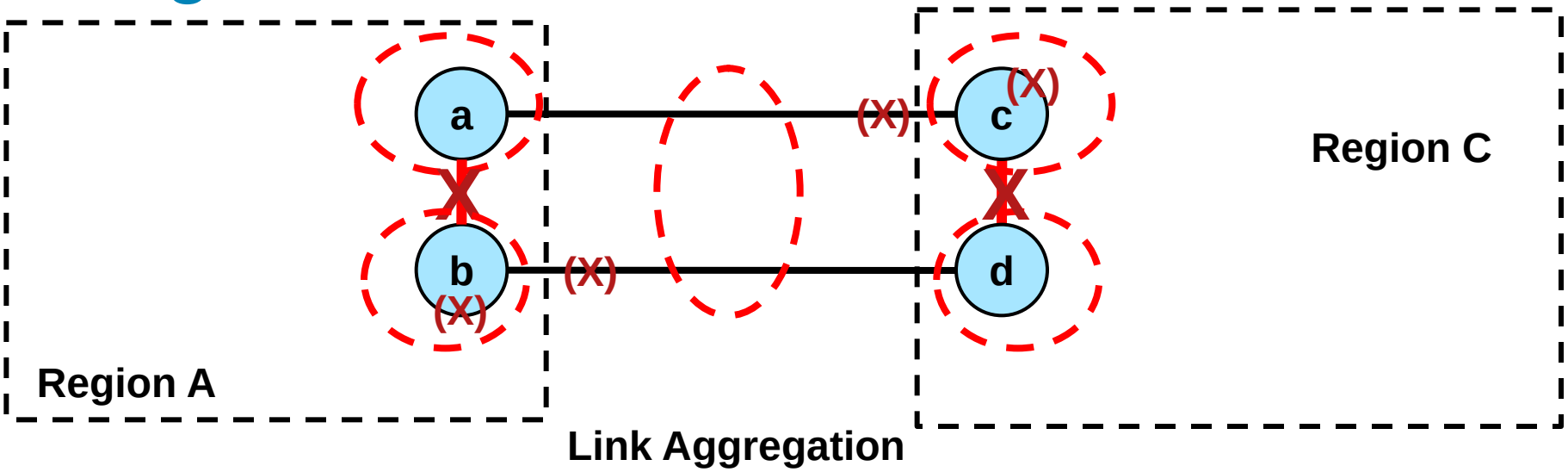
  Some implementations may have no issues with a split brain.

  Some implementations may shut down an isolated secondary component of the virtual node.

  Some implementations may change identities to become two separate devices (equivalent to shut down for the ENNI, since the “light” scheme requires a single virtual node).

- Signaling the recovery choice can be handled with current LACP, e.g., by removing Links to one of the physical Nodes from the aggregation.

Light ENNI – issues



- Suppose the recovery method for “split brain” is that the secondary device shuts down.
- If **a** and **d**, above, are “master” nodes, then if both inter-VTN links fail (as shown), the ENNI would fail.
- Indicating in LACP which is the master node would enable the administrators to make **a** and **c** the master nodes, so that the **a—c** link would remain operational.

Common issues

Common issue: bundling preferences

- We cannot express link preferences for thousands of services in an LACP or CCM PDU; some kind of “bundling” is necessary across the ENNI.
- We can say that the bundling is handled by configuration, and that both administrations must get the configuration right. This seems risky.
- We probably need to define a protocol (or use an existing one) that allows each side to express its bundling preferences, and to tie the LACP or CCM signals to a particular expression.
- We may or may not provide an automatic means of resolving differences in bundling preferences.
- A transport protocol is likely required to carry this much information.

Common issue: service information

- We may or may not provide a means (or use an existing one) of exchanging information about each service passed across the ENNI. Such information could include:

  The existence of a service with a particular service identifier.

  QoS parameters for a service such as EIR/CIR rates, latency requirements, or connectivity priority.

  The global service identifier (used in CCMs) for the service.

  Membership of a VID in a root/individual/group VID set for a rooted multipath service.

  MIRP “I need to receive this service” registrations.

Protocol changes

LACP protocol changes: BDID

- Specify bundles by configuration. The mapping of service ID to Bundle Number is a large database.
- Just like MSTP, a Bundling Database has a “Bundling Database Identifier (BDID)” consisting of a name, a revision number, and a hash function, so that it is very unlikely to accidentally think you are in sync with your neighbor when you are not.
- Just like MSTP in BPDUs, the BDID is carried in every LACPDU.

LACP protocol changes: Bundle prefs

- Along with the BDID is a list of my preferred bundles for this physical link. This list is in the form of a vector of *n*-bit “weight” values.
- Weight values are compared mod 8 as:

```
If (signExtend((myWeight – yourWeight) & (2^n – 1))> 0)
   lwin()
else
   youWin();
```

- This way, “arms races” caused by raising weight values can go on indefinitely: -1 > -3 > 3 > 0, etc. for mod 8.
- My vector says, for each Bundle, “This is my bid for receiving the Bundle on this physical link.”

LACP protocol changes: Bundle prefs

- Each LACPDU also carries, separately for each Bundle, an “I really want the allocation to be symmetrical” bit.
- If neither wants symmetrical, we should supply bundle on each other’s preferred link.
- If either wants symmetrical, both should use the winning bid to choose the link.
- Ties are broken by comparing system ID TLV as an alphanumeric value.

LACP protocol changes: Master cookie

- *One possible means – many more will work!*
- Each LACPDU includes a “Master Physical ID” TLV indicating which physical box is the “master” of the Virtual Terminal Node of the physical box sending the LACPDU.
- If I am in an aggregation with another system and see more than one different Master Physical IDs coming from the other system, I set a “Different Physical IDs Received” bit in *all* of my Master Physical ID TLVs.
- The persistence of a Different Physical IDs Received bit in my received Master Physical ID TLVs indicates I am part of a system with multiple masters – a split brain.

NNI criteria list

How the NNI via LACP fares with criteria

- Protect a single service (VLAN) or a group of services (VLAN).
  - We express bundle preferences in LACPDU.
- Protect against any single failure or degradation of a facility (link or node) in the interconnected zone.
- Support interconnection between different network types (e.g. CN-PBN, PBN-PBN, PBN-PBBN, PBBN-PBBN, etc.)
  - This method works even for MPLS!
- Provide sub-50 ms fault recovery
  - Probably. To be determined.

How the NNI via LACP fares with criteria

- Provide a clear indication of the protection state.
- Avoid modifying the protocols running inside each of the interconnected networks
- Maintain an agnostic approach regarding:
  - the network technology running on each of the interconnected networks, and
  - any protection mechanism deployed by each of the interconnected networks
- Allow load-balancing between the interfaces that connect the networks to ensure efficient utilization of resources.

How the NNI via LACP fares with criteria

- The effects of protection events in one network must not affect other networks.
- The effects of protection events in the interconnected zone on the topology of the related attached networks should be minimized.
  - Tie-in between changes in one Region and bundling preferences are buried in the implementation. This criterion is a goal for the implementation.
- Design the interconnected zone in a way that will ensure determinism and predictability.
  - This is an implementation requirement, now.
- There can be at least one failure in every provider cloud, and at least one failure in every interconnect cloud, and connectivity will still be maintained

How the NNI via LACP fares with criteria

- Support topologies with more than two nodes and more than two inter-cloud links, so that equipment can be taken down and replaced without a period of unprotected operation.
- Control packets cannot be 1:1 with customer services; that is, some kind of bundling is necessary in order to support thousands of services.
  - Bundles take care of this
- The bundling of services for protection purposes (e.g. MST instances) can be completely different in different service provider clouds.

How the NNI via LACP fares with criteria

- The NNI protects services, not parts of services.
  - We must make this type of Aggregation mandatory.
  - Presence of Bundling TLV announces this.
- If one service provider cloud becomes split into multiple disjoint clouds, it cannot depend on the interconnect cloud or any adjacent service provider cloud to provide connectivity among its parts.
- We cannot assume an ultra-reliable link.

  Cookies!
- It must be possible to ensure the use of the same link in both directions for every service.

How the NNI via LACP fares with criteria

- Inter-domain coordination should be minimized.
- Support asymmetrical links -- not all the same speed or cost
  - Symmetry is an arbitrary imposition, at present.
  - No need to specify *how* split over unequal speed links is to be made.
- Do we support an encapsulation scheme in the interconnect cloud, or is the ENNI independent of the encapsulation?
  - Independent!

How the NNI via LACP fares with criteria

- Do we assume that the bandwidth (or other Traffic Engineering parameter) of the interconnect cloud is adequate for all of the services, or do we do something special if it is insufficient?
  - Good question! To Be Determined
- Do we need protocol for conveying service creating/deletion or traffic engineering requirements between Service Providers?
  - Good question! To Be Determined