An Evaluation of Options for Balancing Trans-Atlantic LHCONE Links
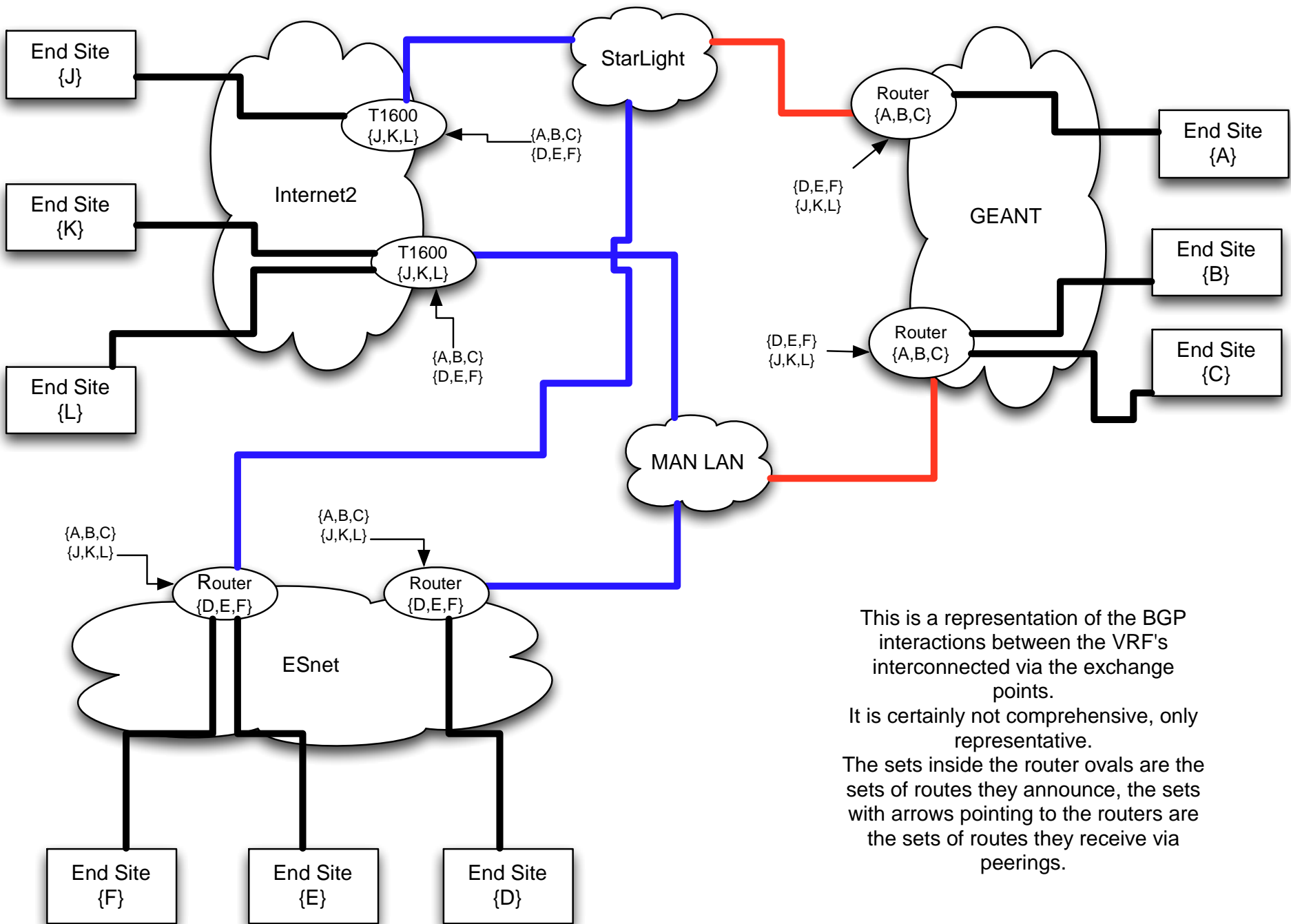
Dale Finkelson
Joe Mambretti
Mike O'Conner
Tom Knoeller

INTERNET 2

Background

- LHCONE is an overlay network.
  - Resides on existing networks and existing Trans-Atlantic links.
  - Interconnects for these networks and connections to the circuits via the established exchange points.
  - Implemented in the Context of Other LHC Network Paths (Not a Major Consideration of This Presentation But Important for Future Discussions)
- There are many links available.
  - 3 10G circuits in NY - MANLAN
  - 2 10G circuits in Chi - StarLight
  - 2 10G circuits in DC – WIX
  - Others provided by other communities.

Problem

- Much of the traffic flow is between Europe and the US (Primarily From Europe To the US).
  - Peerings exist at all 3 peering point between European and US VRF's.
- How can these traffic flows be balanced across these links?
  - Goal is to not congest some circuits while starving others.
  - This in an environment with mixed and highly varied traffic flows.

INTERNET 2

End Site {J}

T1600 {J,K,L} ← {A,B,C} {D,E,F}

Internet2

End Site {K}

T1600 {J,K,L}

End Site {L}

{A,B,C} {D,E,F}

StarLight

Router {A,B,C}

End Site {A}

{D,E,F} {J,K,L}

GEANT

End Site {B}

{D,E,F} {J,K,L} → Router {A,B,C}

End Site {C}

MAN LAN

{A,B,C} {J,K,L}

Router {D,E,F}

{A,B,C} {J,K,L}

Router {D,E,F}

ESnet

End Site {F}

End Site {E}

End Site {D}

This is a representation of the BGP interactions between the VRF's interconnected via the exchange points.
It is certainly not comprehensive, only representative.
The sets inside the router ovals are the sets of routes they announce, the sets with arrows pointing to the routers are the sets of routes they receive via peerings.

Routing between VRF's

- Routing between the VRF's is handled by BGP.

- BGP will be employed to communicate reachability between the NSPs in the LHCONE layer3 overlay network.

- The path selection tools in BGP are designed to find efficient paths or paths with least cost.
  - Metrics might be AS-path length for instance.

INTERNET

BGP

- The problem is that the default behavior of BGP will almost certainly lead to imbalances in utilization of the various paths.
  - BGP does not look at path utilization in determining path selection.
  - BGP primarily determines reachability, not the speed, length or any other factors of any given path.

INTERNET2

BGP Tools

- There are some tools available.
  - MEDS (Mulit-Exit Discriminators)
  - AS padding
  - Local Prefs
  - Restricted announcements

MEDS

- Used to show peers which inbound path is preferred when several are available.

- A means of relating internal topology to peers for selecting inbound paths.

- Potentially useful but alone will not assist in balancing.

  - It may turn out that everyone prefers one set of paths for instance.

  - Requires coordination between all the NSP's.

AS Padding

- Used by the receiving site to show a preference for one link over another.
  - One may be faster or cheaper or have some other desirable characteristic.
- Any given site prefix might be announced to GEANT at each of the 3 IXP's with different AS padding.
  - For that prefix this will force traffic down a specific link.

INTERNET

AS Padding

- Can this help with load balancing?
  - It could but it may require
    - Substantial initial manual monitoring and doing traffic analysis based on that monitoring.
  - At any given moment it might do as much harm as good.
    - At present there are no automated tools to adjust this based on actual traffic.
  - This approach might also lead to increased asymmetry.

Local Prefs

- If there are multiple exit points from the AS, the local preference attribute is used to select the exit point for a specific route.

  - For instance Internet2 could prefer Starlight for a specific prefix and WIX for a different one.

- This certainly could affect overall traffic patterns.

INTERNET 2

Local Prefs

- Questions:
  - Can this approach actually scale?
  - Does it have to scale substantially?
  - Will sites have the time and information to make the on-going continuous adjustments needed?
  - Is LHCONE likely to be sufficiently predicable that this is likely to succeed?
  - Would an initial trial of this approach be worthwhile?

INTERNET2

Restricted announcements

- It is certainly possible to only announce some prefixes on some links.
  - Very possibly would do some good.
  - Would require a fair amount of co-ordination and communication between all of the organizations

INTERNET2

Short Term vs Long Term

- With a limited number of sites participating some of these tools might well be useful.

- They do require a lot of monitoring, analysis and traffic engineering.

- In the longer term, these tools do not seem to hold the promise of solving the problem.

- However, in the short term, experimenting with one of these approaches may be useful to assess potentials for the eventual approach

- In the short term, ideal utilization is not required

- A short term approach could be used to assess longer term requirements and solutions

INTERNET

Summary

- LHCONE NSPs will peer at multiple locations. Otherwise, MEDs are ineffective except for complex proprietary implementations.

- Capacity planning at peering points is the most effective load balancing strategy.

- MEDs will be exchanged between all NSPs and should be employed in favor over AS path pre-pending whenever possible.

- Measurement needs to be performed continuously and be readily available for capacity planning purposes, PerfSonar etc.

- A short term approach/assessment may be useful to assist in decision making for the longer term solution

INTERNET2