Rationalizing Tier 2 Traffic and Utilizing the Existing Resources (T/A Bandwidth)

What We Have Here is a Traffic Engineering Problem

LHC Tier 2 Technical Meeting – CERN
13 January 2011

William E. Johnston
(wej@es.net)

Chin Guok, Joe Metzger, Kevin Oberman, Chris Tracy, et al

Rationalizing Tier 2 Networking

- The view of the problem from the U.S. is somewhat different than from Europe
- In the U.S.
  - The two Tier 1 centers are large: Fermilab holds about 40% of the CMS data and Brookhaven holds 100% of the ATLAS data
  - Fermi is on a dark fiber ring to StarLight and currently has 50G configured to the ESnet core
  - BNL currently has 40G to ESnet/MAN LAN , and within two months will be on an ESnet owned dark fiber ring
  - All of the Tier 2 centers are connected to either StarLight or MAN LAN with dedicated 10G circuits, and several have their own fiber
  - ESnet handles all Tier 2 traffic (world-wide) to and from Fermi and BNL
    - ESnet sees all transatlantic traffic headed to Fermi and BNL
    - ESnet sees none of the U.S. Tier 2 out-bound traffic as that is university traffic that is handled by the Regionals, Internet2, NLR, etc.

Rationalizing Tier 2 Networking

- Therefore, Tier 2 traffic within the U.S. is not now, nor is it every likely to be, an issue
  - Tier 3 traffic is still largely uncharacterized
- However, Tier 2 traffic across the Atlantic (in both directions) requires careful attention
  - the way the IP networking across the Atlantic is currently structured results in most general traffic (including almost all LHC non-OPN traffic) to use a small number of paths – the same paths used by most other R&E traffic
  - this situation will get better as the ACE infrastructure comes on-line, but Tier 2 traffic will be ramping up at the same time

The Need for Traffic Engineering – Example

- The LHC community has developed applications and tools that enable very high network data transfer rates over intercontential distances
- This is necessary in order to accomplish their science
- On the LHC OPN – a private optical network designed to facilitate data transfers from Tier 0 (CERN) to Tier 1 (National experiment data centers) – the HEP data transfer tools are essential
  - These tools are mostly parallel data movers – typically GridFTP
  - The related applications run on hosts that have modern TCP stacks that are appropriately tuned for high latency WAN transfers (e.g. international networks)
  - the Tier 2 sites use the same highly tuned WAN transfer software

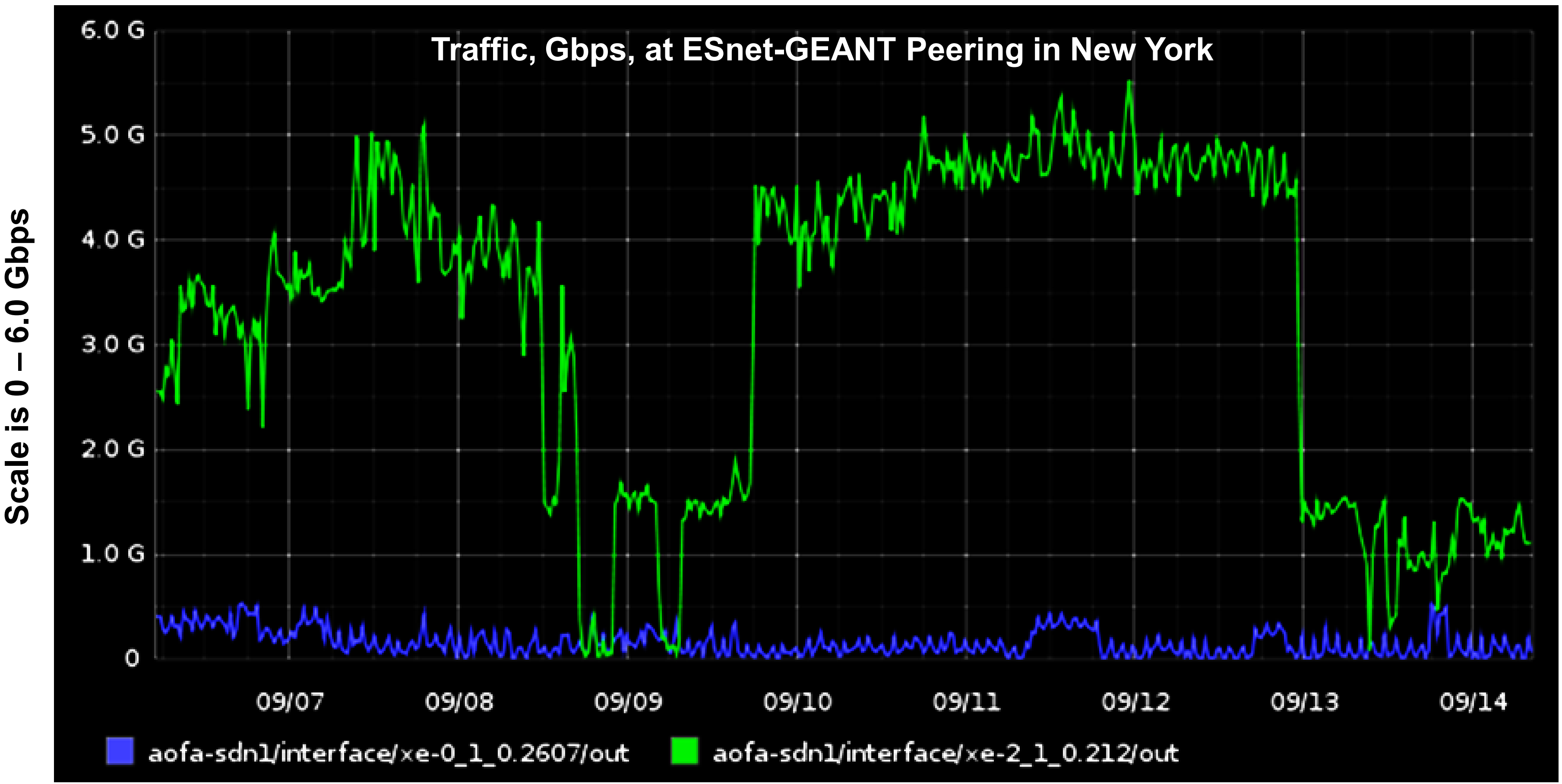
The Need for Traffic Engineering – Example

- Recently, the Tier 2 (mostly physics analysis groups at universities) have abandoned the old hierarchical data distribution model
  - Tier 0 -> Tier 1 -> Tier 2, with attendant data volume reductions as you move down the hierarchy

  in favor of a chaotic model
  - get whatever data you need from wherever it is available
- This has resulted in enormous site to site data flows on the general IP infrastructure that have never been seen before apart from DDOS attacks

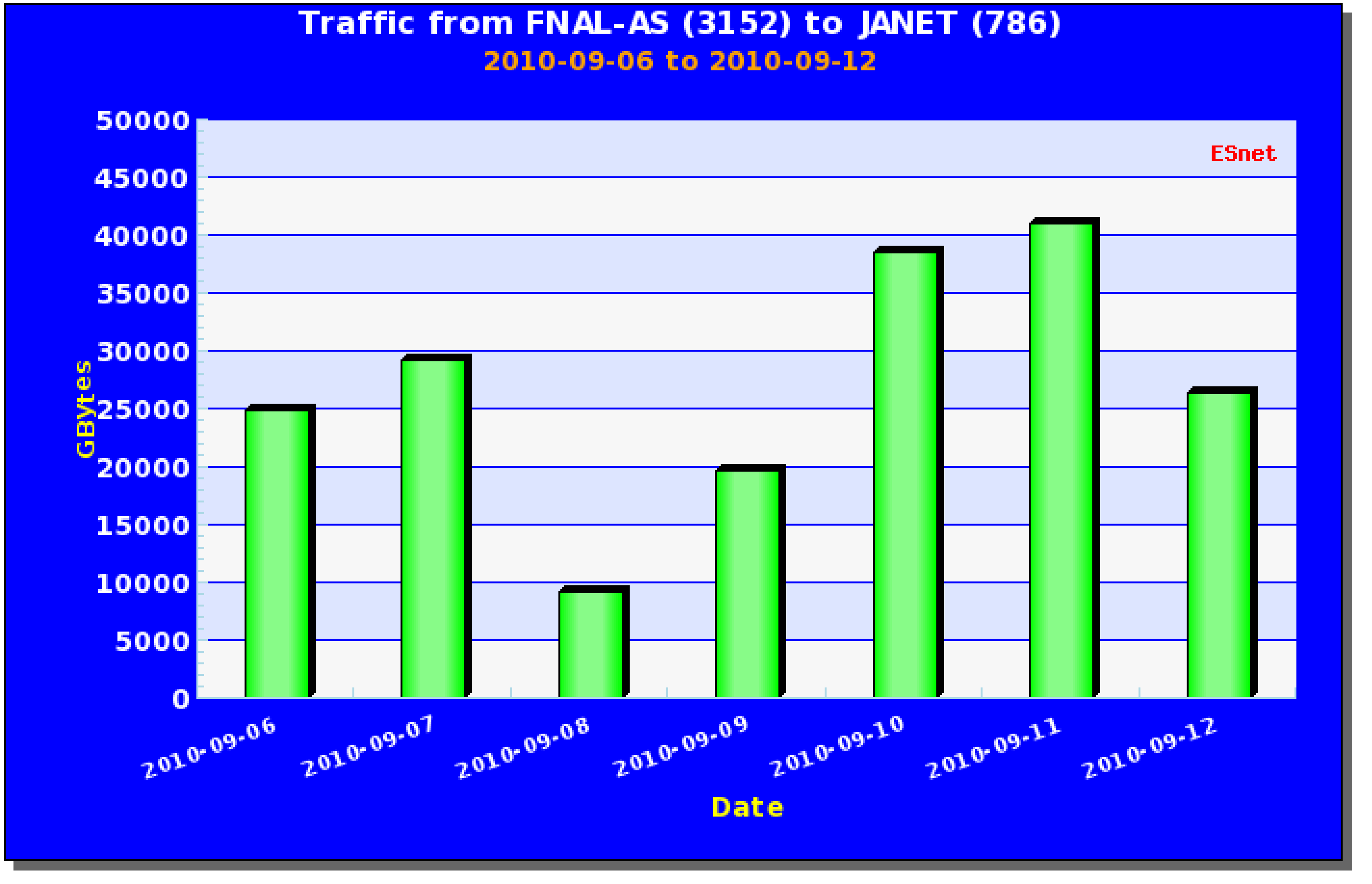
The Need for Traffic Engineering – Example

- GÉANT observed a big spike on their transatlantic peering connection with ESnet (9/2010)
  - headed for Fermilab – the U.S. CMS Tier 1 data center
- ESnet observed the same thing on their side



The Need for Traffic Engineering – Example

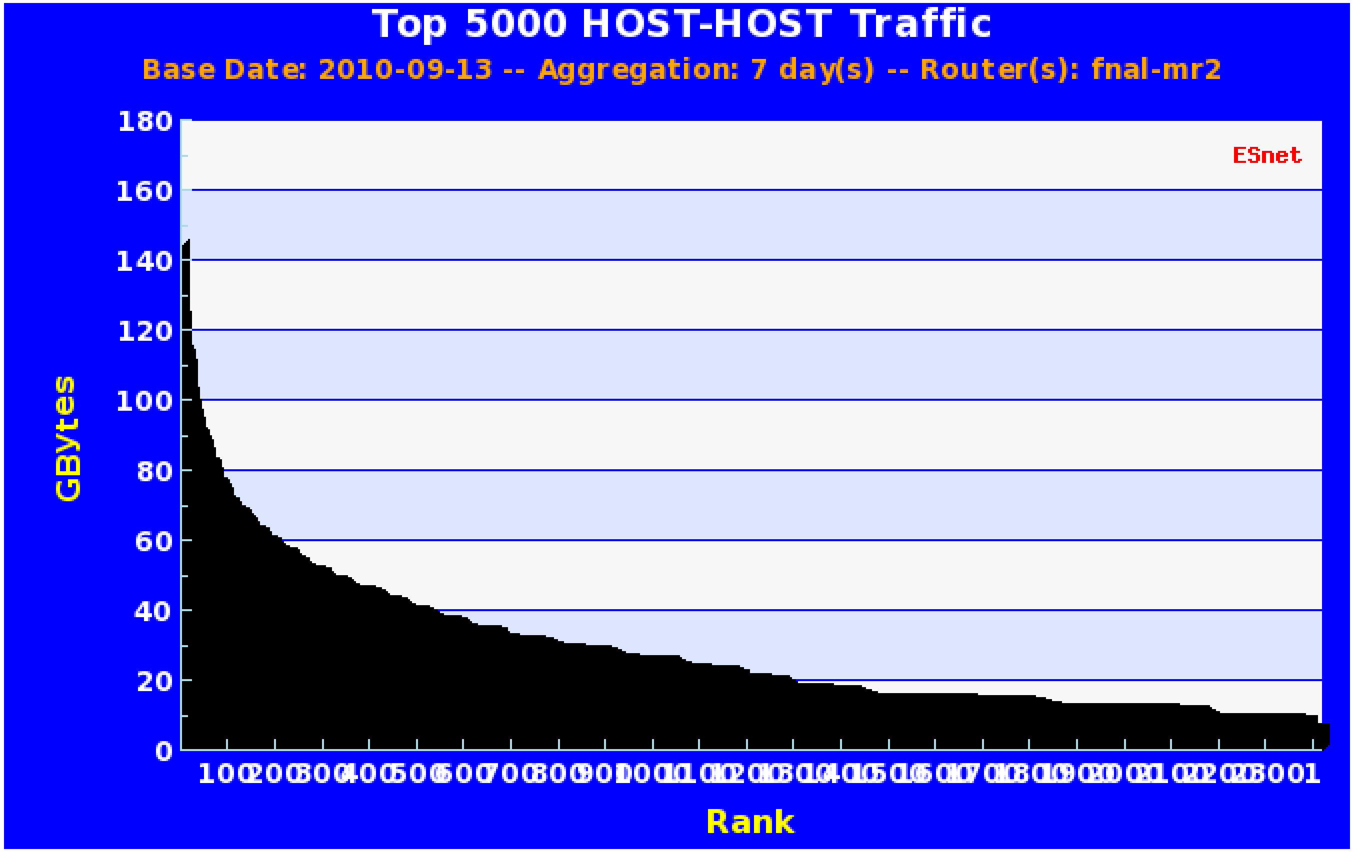
- At FNAL is was apparent that the traffic was going to the UK



- Recalling that moving 10 TBy in 24 hours requires a data throughput of about 1 Gbps, the graph above implies 2.5 to 4+ Gbps of data throughput – which is what was being observed at the peering point

The Need for Traffic Engineering – Example

- Further digging revealed the site and nature of the traffic
- The nature of the traffic was – as expected – parallel data movers, but with an uncommonly high degree of parallelism: 33 hosts at the UK site and about 170 at FNAL



The Need for Traffic Engineering – Example

- This high degree of parallelism means that the largest host-host data flow rate is only about 2 Mbps, but in aggregate this data mover farm is doing 860 Mbps (seven day average) and has moved 65 TBytes of data
  - this also makes it hard to identify the sites involved by looking at all of the data flows at the peering point – nothing stands out as an obvious culprit
- THE ISSUE:
- This clever physics group is consuming 60% of the available bandwidth on the primary U.S. – Europe general R&E IP network link – for weeks at a time!
- This is obviously an unsustainable situation and this is the sort of thing that will force the R&E network operators to mark such traffic on the general IP network as scavenger to ensure other uses of the network

The Need for Traffic Engineering – Example

- In this case marking the traffic as scavenger probably would not have made much difference for the UK traffic (from a UK LHC Tier 2 center) as the net was not congested
- However, this is only one Tier 2 center operating during a period of relative quiet for the LHC - when other Tier 2s start doing this things will fall apart quickly and this will be bad news for everyone:
  - For the NOCs to identify and mark this traffic without impacting other traffic from the site is labor intensive
  - The Tier 2 physics groups would not be able to do their physics
  - It is the mission of the R&E networks to deal with this kind of traffic
- There are a number of ways to rationalize this traffic, but just marking it all scavenger is not one of them

Transatlantic Networking

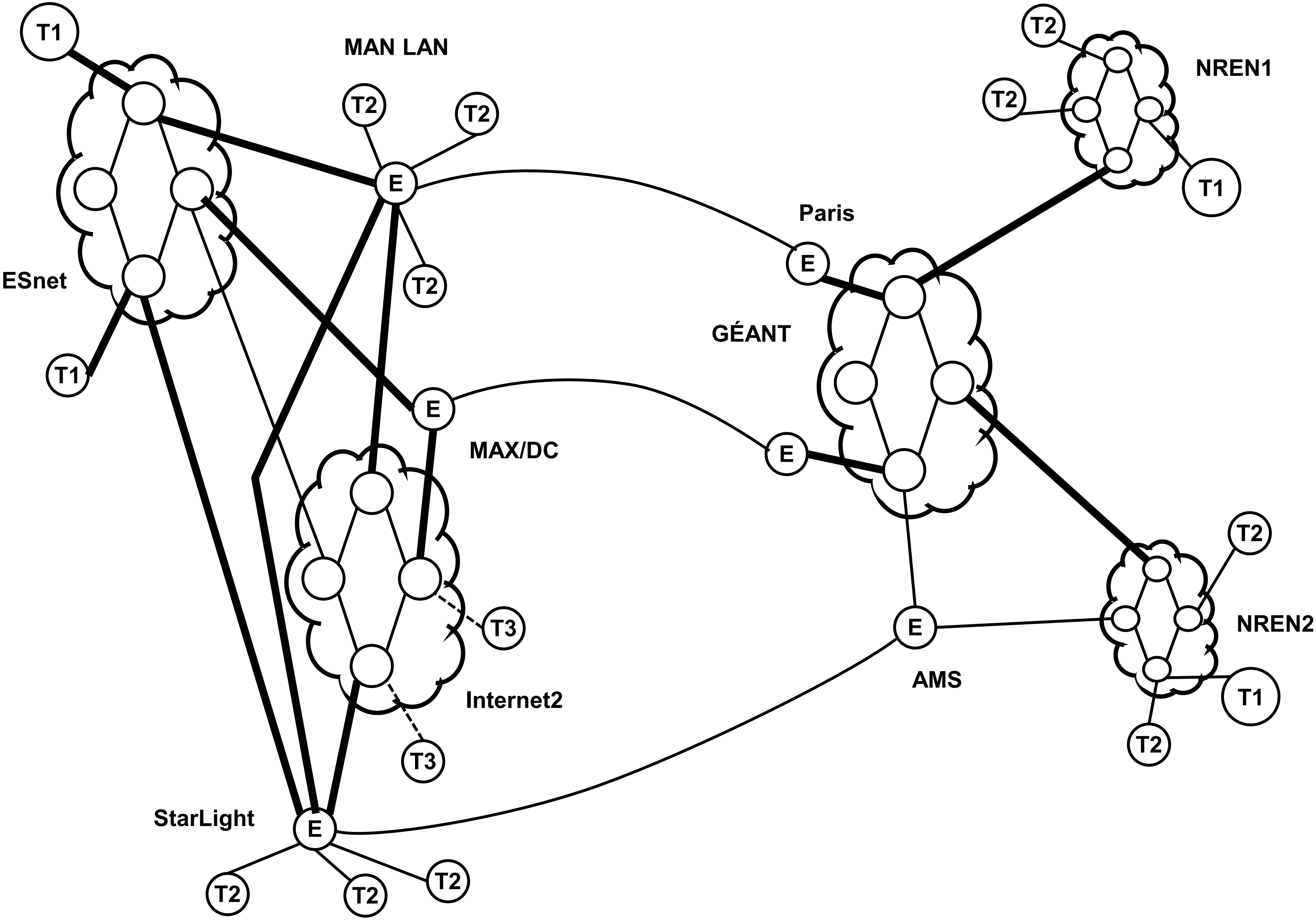
- The transatlantic capacity available to the community is probably sufficient in the near term if it used optimally
- Current R&E T/A circuits (John S. Graham, Global Research NOC)

Number	Endpoints		Owner	Operator	Purpose
	USA	Europe			
1	New York	Amsterdam	Indiana University	SURFNet	Geant IP Peerings
2	Washington	Frankfurt	Geant	Geant	
3	New York	Paris	Geant	Geant	Lightpaths
4	New York	London	Internet2	Internet2	Lightpaths
5	New York	Amsterdam	CANARIE	SURFNet	Lightpaths
6	New York	Amsterdam	SURFNet	SURFNet	Lightpaths
7	New York	Amsterdam	NLR	SURFNet(?)	Unknown
8	New York	Amsterdam	NorduNet	NorduNet	IP Peerings
9	New York	??	SINET	SINET	IP Peerings

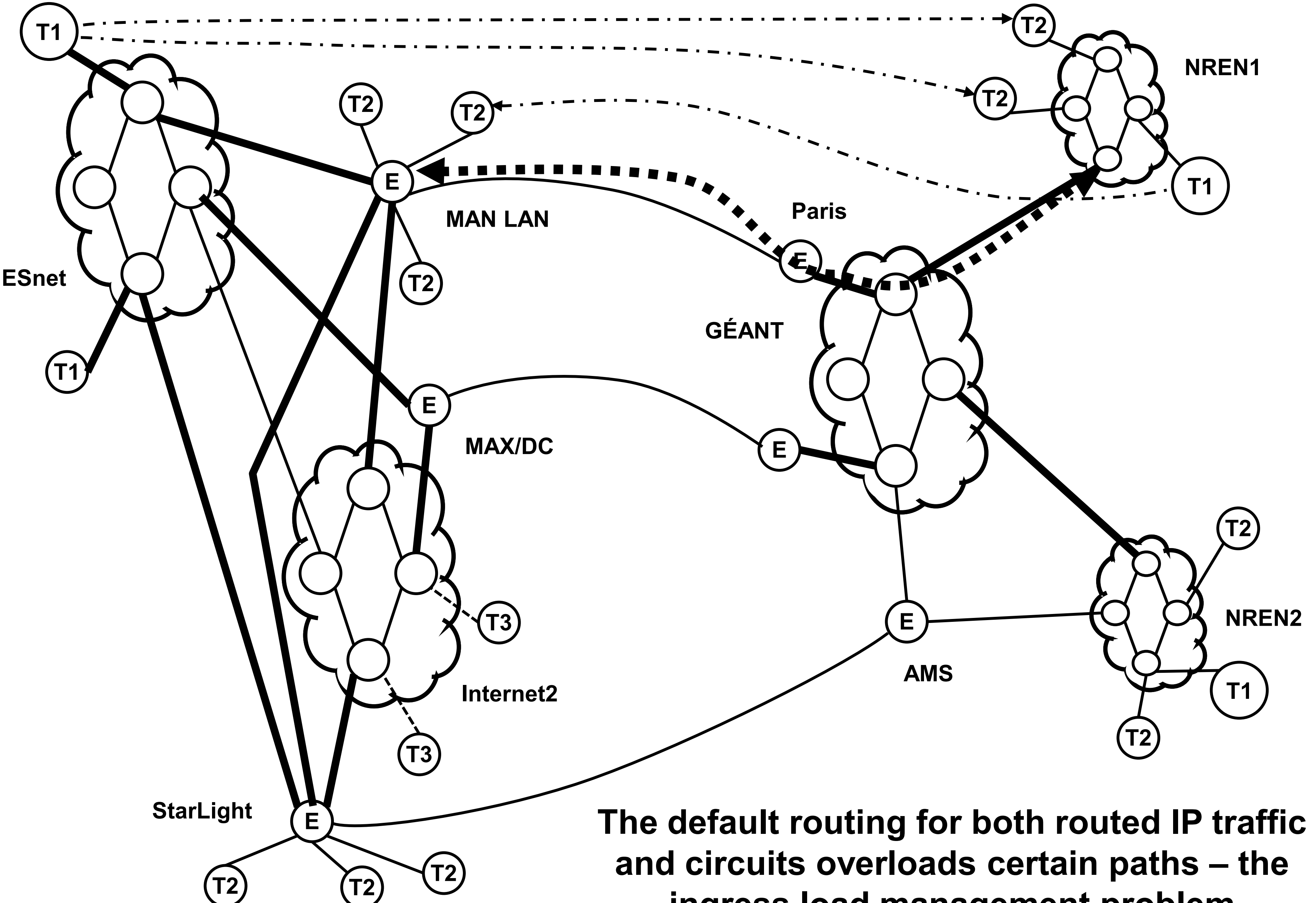
Transatlantic Networking

- The question is how to optimize the use of the available capacity
  - satisfy the LHC needs while accommodating all other R&E traffic at the same time
    - this point is critical because the available non-OPN capacity is funded for the benefit of the entire R&E community, not just the LHC

Roughly Today's Situation

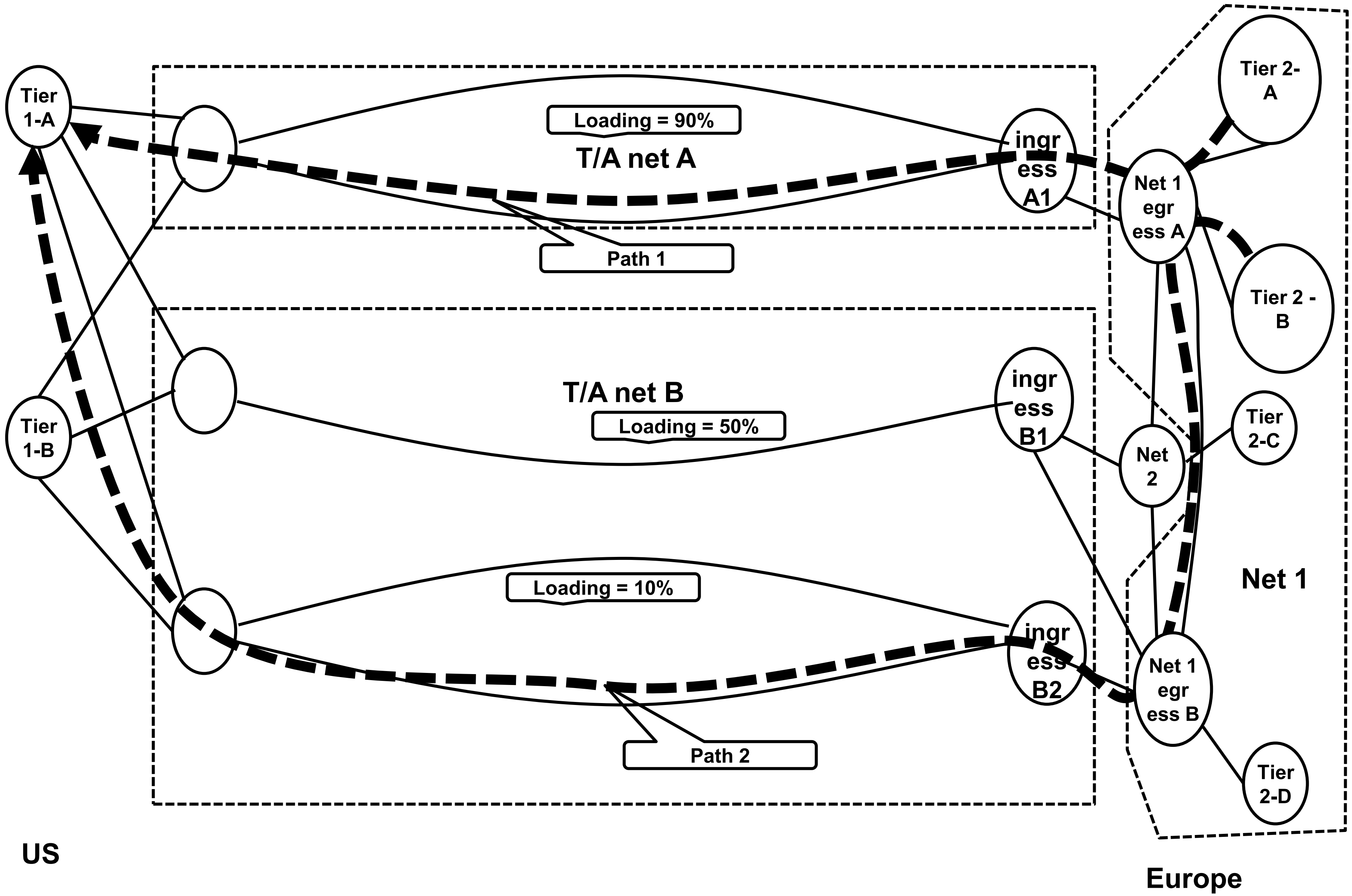


The Problem



The default routing for both routed IP traffic and circuits overloads certain paths – the ingress load management problem

What you would like to do is spread the load to avoid congestion (e.g. at the ingress of net A)



Traffic Engineering – Routed IP Traffic

- There are several ways that one could address this for IP traffic in a federated infrastructure such as we have now
- In a single domain such as net B, the operator can use MEDs that are dynamically established from transatlantic path loadings to direct traffic to a less loaded ingress point
  - e.g. ingress B1 vs. B2 in the figure
- In the case of balancing across several independent domains (e.g. net A and net B) then the source must redirect traffic away from a congested (though perhaps closer) ingress point
  - This should be able to be done with BGP local preferences to control the exit path from an edge router
    - the local perfs would have to established and changed dynamically based on the loading of several available paths (e.g. forcing net 1, egress A traffic away from net A, ingress A1 and routing it to net B, ingress B2 – which is not the default route)

Traffic Engineering- Circuit Approaches

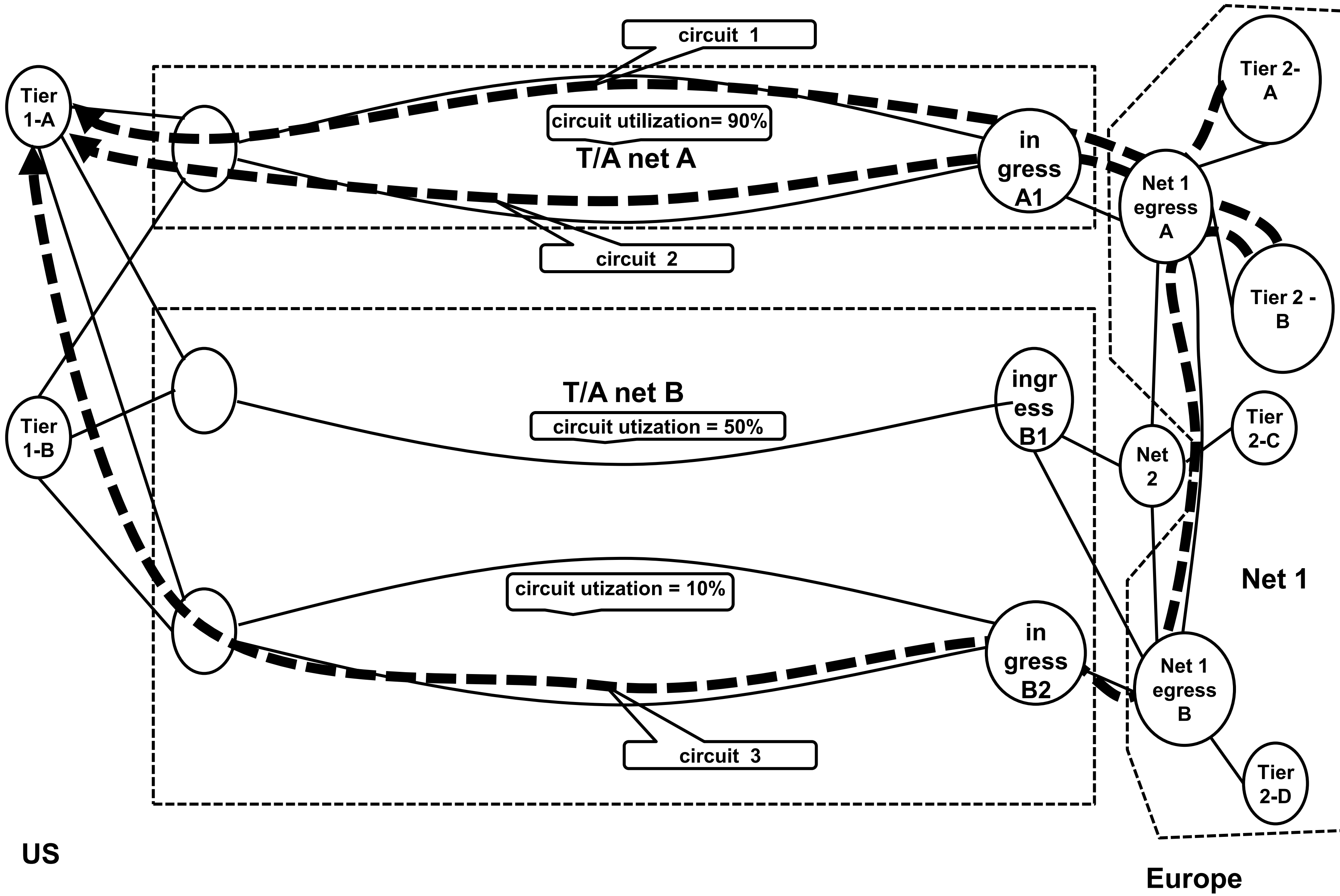
- One way to rationalize Tier 2 traffic is to set up virtual circuits that have guaranteed, but at the same time controlled, bandwidth that is isolated from general traffic, from the Tier 2 sites to the Tier 1 data centers
  - The number of such combinations per Tier 2 is probably relatively small (10s at most) due to the access patterns arising from the nature of the Tier 2 analysis interests and distribution of data in the Tier 1 centers
- This sort of multi-domain traffic engineering is what is done for almost all of the U.S. Tier 2 centers for accessing the U.S. Tier 1 centers
  - all of these circuits and all of the U.S. LHC OPN circuits are based on OSCARS virtual circuits
- With caveats, the DICE IDC protocol has the capability to do this in the international network arena

Traffic Engineering- Managing IDC Circuit Paths

- The situation with circuits is similar to the IP traffic problem: How to avoid “congestion” (fully committed paths is the circuit version of “congestion”)
- The inter-domain IDCs have a global view of available topology, but not the current state of utilization, so cannot route around congestion
- In the next figure, with net A at full capacity, a successful circuit request must find and use a longer than normal circuit between T2-A and T1-A, which the current version of the IDC will not do automatically

Traffic Engineering- Managing IDC Circuit Paths

Tier 2-A and Tier 2-B circuits have exhausted the default IDC route capacity. Any further circuits will have to take non-default paths (e.g. circuit 3).



Traffic Engineering- Managing IDC Circuit Paths

- Even though the inter-domain IDCP cannot find a path from T2-A to T1-A, the path exists (“circuit 3”)
- The hop-by-hop circuit path is defined by an MPLS construct called an Explicit Route Object (ERO)
- The IDC can return the ERO to the user, and the user can modify it and use the modified version to define a path (assuming it represents a valid path)
- Currently perfSONAR can return path utilization information on a by-node basis
  - this information can be used to manually modify an ERO to represent an alternate path that is not “congested” (i.e. has capacity for the requested circuit)
  - however, perfSONAR cannot report on temporal circuit commitments on the path – this is being worked on
- This sounds like a “heavy weight” approach, and it is, but not impractically so if the circuit will be long-lived, as almost all production circuits are
  - DICE group is looking at tools to simplify the process
  - automation of the process is an active research topic that is seeing some progress

Traffic Engineering as a Solution for Tier 2 T/A Traffic

- At some level, adequate transatlantic R&E capacity is sufficient for the near future, if it can be managed in a federated way that distributes the LHC load across the available capacity
- Tools exist to accomplish – or at least to prototype – this approach
- Can a suitable federation be established?
  - Probably if an acceptable governance model can be agreed to that addresses capacity sharing and operational cooperation