T0-T1-T2 networking

Vancouver, 31 August 2009
LHCOPN T0-T1-T2 Working Group

Requirements from the experiments

T0-T1 traffic

1. **For all the experiments, the most important thing is to save a second copy of the RAW data to tape at the Tier-1's. The first copy is on tape at CERN**. This makes the T0-T1 links different from all the others. **This also says something about the importance of the T0-T1 backup paths.** Any additional features on the OPN should not endanger this prime functionality.

T1-T1 traffic

2. The latest STEP exercise has shown that for ATLAS and also for CMS **the peak rates between T1's are just as high or even in excess of the T0-T1 rate**. It has never become a problem though during the test because it is less time critical than the T0-T1. Moreover there is the freedom to go to round-robin rather than point to point mode because it is all about distributing the same data to all T1's.

T1-T2 traffic

3. There is a difference between the ATLAS and the CMS model for T1-->T2: where **for CMS the T2 should be able to get its data from each T1**, **for ATLAS the most important traffic is between the T2's and the T1 in the same "cloud"**. So the ATLAS model is more hierarchical and the CMS model is more of a mesh. ATLAS has some out-of-the-cloud T1-T2 traffic but it is less important.

T0-T2 traffic

4. **ATLAS has only a few (four: Rome, Munich, Michigan, Geneva) calibration data streams between the T0 and T2's**. These data are time critical but of moderate rate (<50 MB/s). Calibrations (muon and trigger) are done at those sites and processing in the T0 depends on the results of those calibrations being send back to CERN in a timely fashion.

Special T2s

5. For ATLAS some T2's are more important than others: we have officially ~60 T2's but 50% of our analysis gets done in the ~10 best T2 sites. **We would like to have the option of having those T2's better served but must keep in mind that the list of "golden T2's" may vary with time (matter of months)**. For T1-T2 traffic, the same holds as for the T1-T1: the rate may be high, higher than any of the other channels and even more so for CMS than for ATLAS.

Problem definition

Issues

T0, T1 and T2 sites need to transfer large amount of data among them. Traffic patterns may vary.

Connectivity via normal internet upstreams may be limited in bandwidth and expensive, thus not suitable for LCG data transfer.

Links between pairs of sites may already exist, but they've a limited scope and are not well exploited.

Requirements

Let T0/T1/T2 sites exchange traffic in the most flexible and economical way:

- Maximize exploitation of available wavelengths and dark fibres

- Reduce costs

- Reliable network configuration

Constrains

- High cost of last mile and long distance connectivity

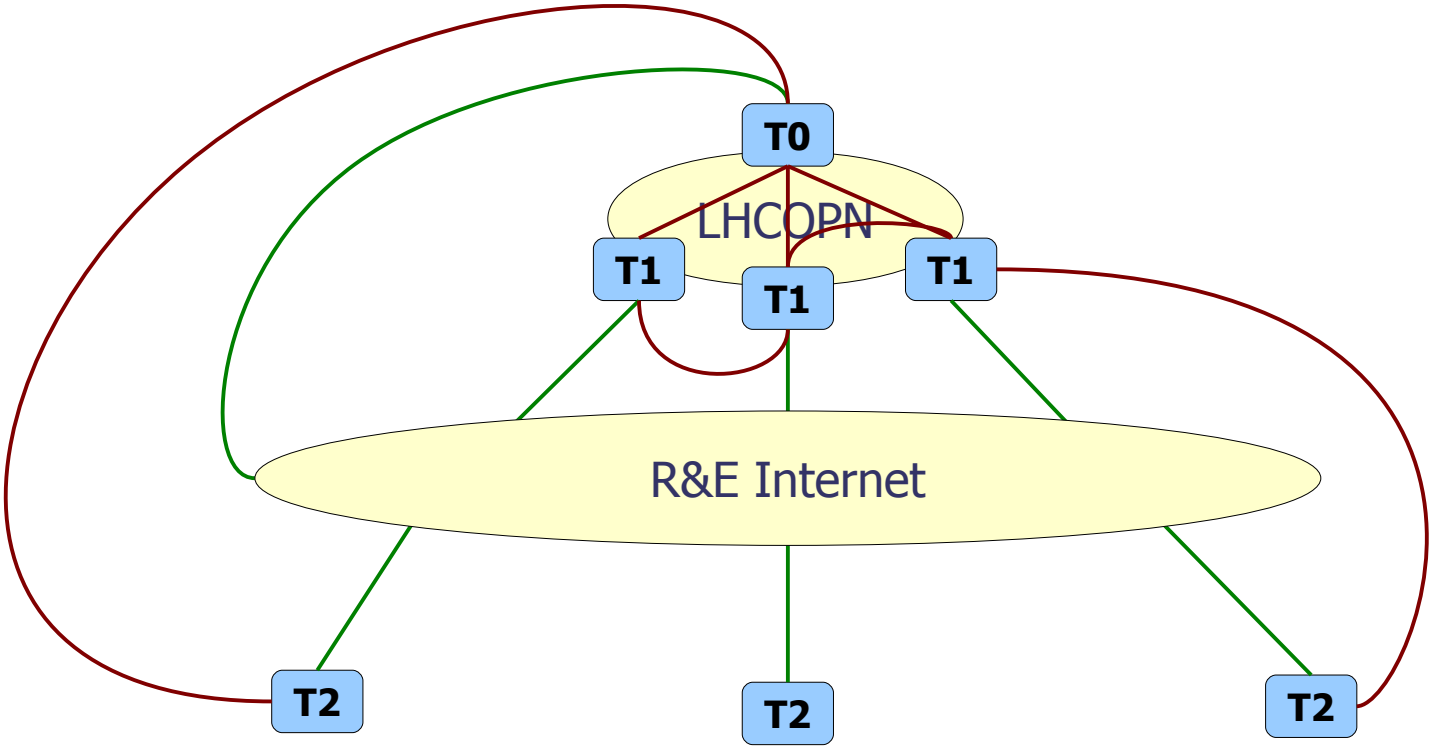
- Heterogeneous network domains

- Limited network know-how at Tier2s

- Do not replace the LHCOPN.

Current status

T0-T1-T2 Interconnections today



Possible solutions

Model A: Dynamic circuits

Model A – Dynamic circuits

Point-to-point circuits are dynamically provisioned between any pair of sites, whenever needed, with the optimal bandwidth, just for the necessary duration.

The WLCG applications communicate their needs to the Network Control Plane, which implements the requested circuits.

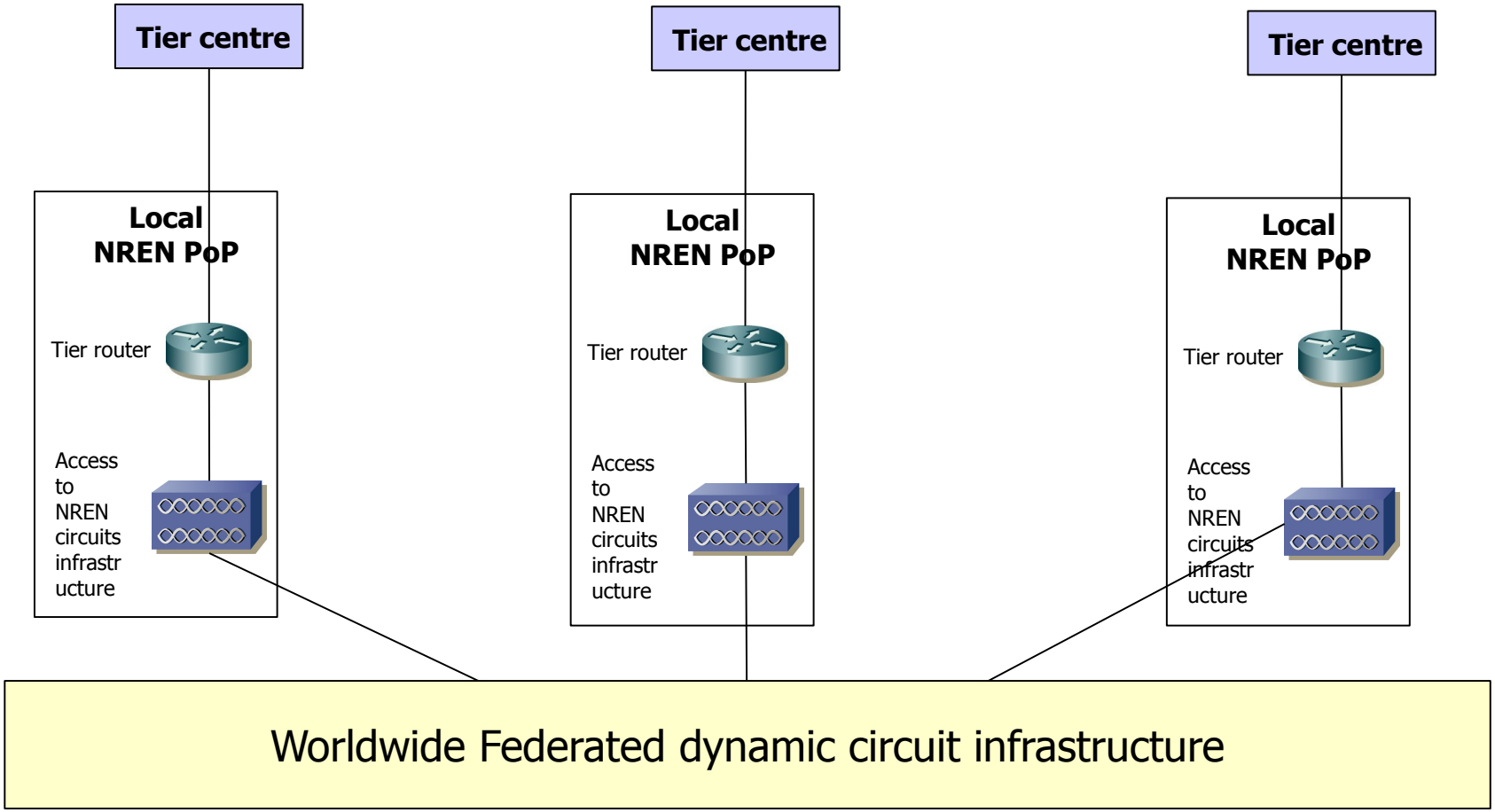
Model A – In practice

Each site co-locates its border router in its NREN PoP and buys one or two access circuits (primary and backup) from its promises to the border router

All the R&E networks are interconnected in a worldwide federated dynamic circuit infrastructure

An API lets the WLCG applications requests circuits between pairs of site's border-routers.

Model A – Dynamic circuits



Model A – Pro and Cons

Pro

- Best use of R&E Networks bandwidth
- Sites pay for long distance links only when necessary

Cons

- Not straightforward routing configuration of sites' routers
- lot of coordination needed among R&E Networks
- May not be possible to connect any pair of sites
- WLCG applications must know which circuit to ask

Model B: Internet eXchange Point

Model B – Internet eXchange Point

The WLCG community builds and maintain a distributed exchange point infrastructure with access switches in few strategic locations.

The sites connect to the IXP infrastructure and peer with any other site is needed to communicate with.

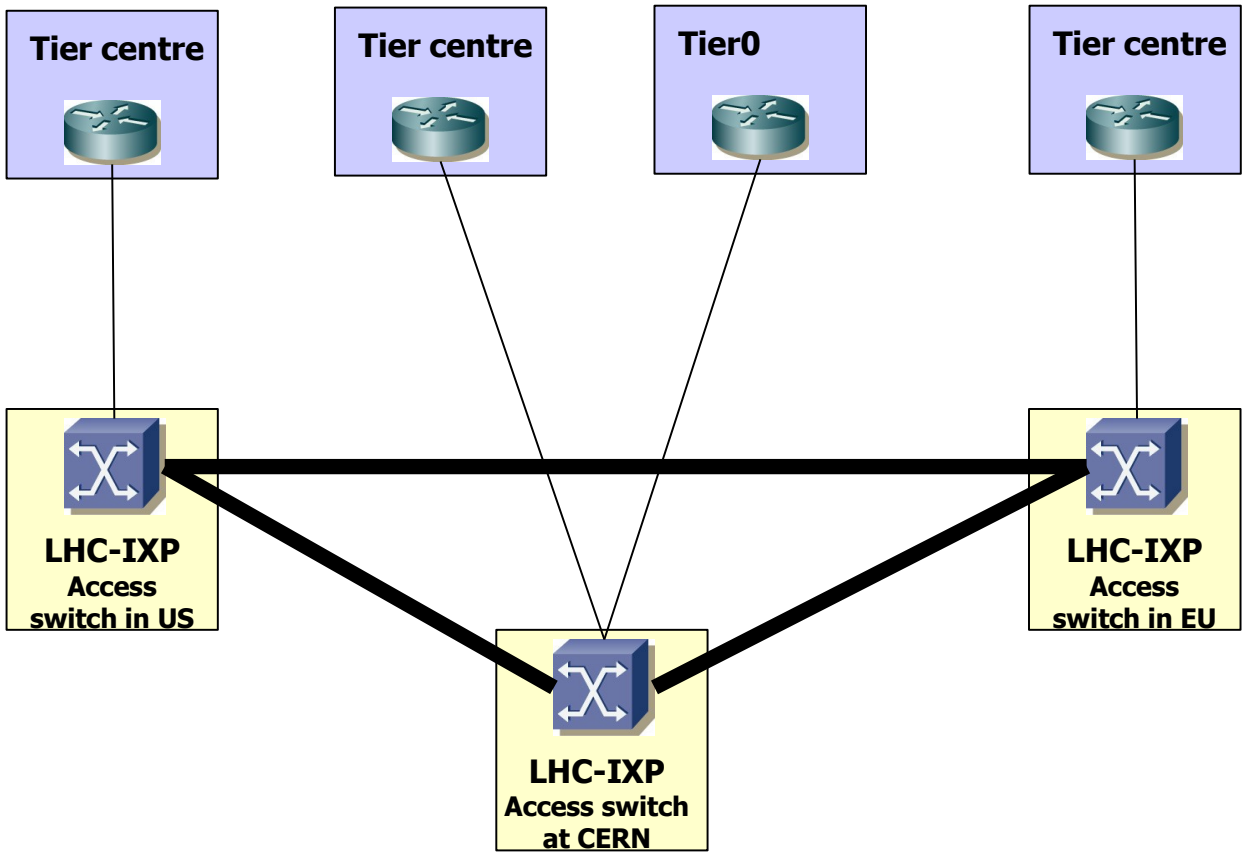
Model B – In practice

The LHC distributed IXP has access switches in few strategic locations (one or two in Europe, one in North America, maybe one in Asia). The switches are interconnected with enough bandwidth.

The sites buy one or more circuits from their promises to the access switches and terminate them to their border routers.

All the sites' routers will reside in the same IP network, being able to reach any other site connected to the IXP infrastructure and to establish ad hoc routing policy with any of them.

Model B – Internet Exchange



Model B – Pro and Cons

Pro
- effective and scalable L3 configuration
- best use of sites' access links
- easy to connect pair of sites independently of their location

Cons
- cost and maintenance of IX infrastructure
- permanent cost of long access links for sites

Model C: Lightpath eXchange

Model A – Lightpath exchange

Point-to-point circuits are dynamically provisioned between pairs Lighpath exchange access points, whenever needed, with the optimal bandwidth, just for the necessary duration.

Each site has to connect to one Lightpath access points.

The WLCG applications communicate their needs to the Network Control Plane, which implements the requested circuits.

Model C – In practice

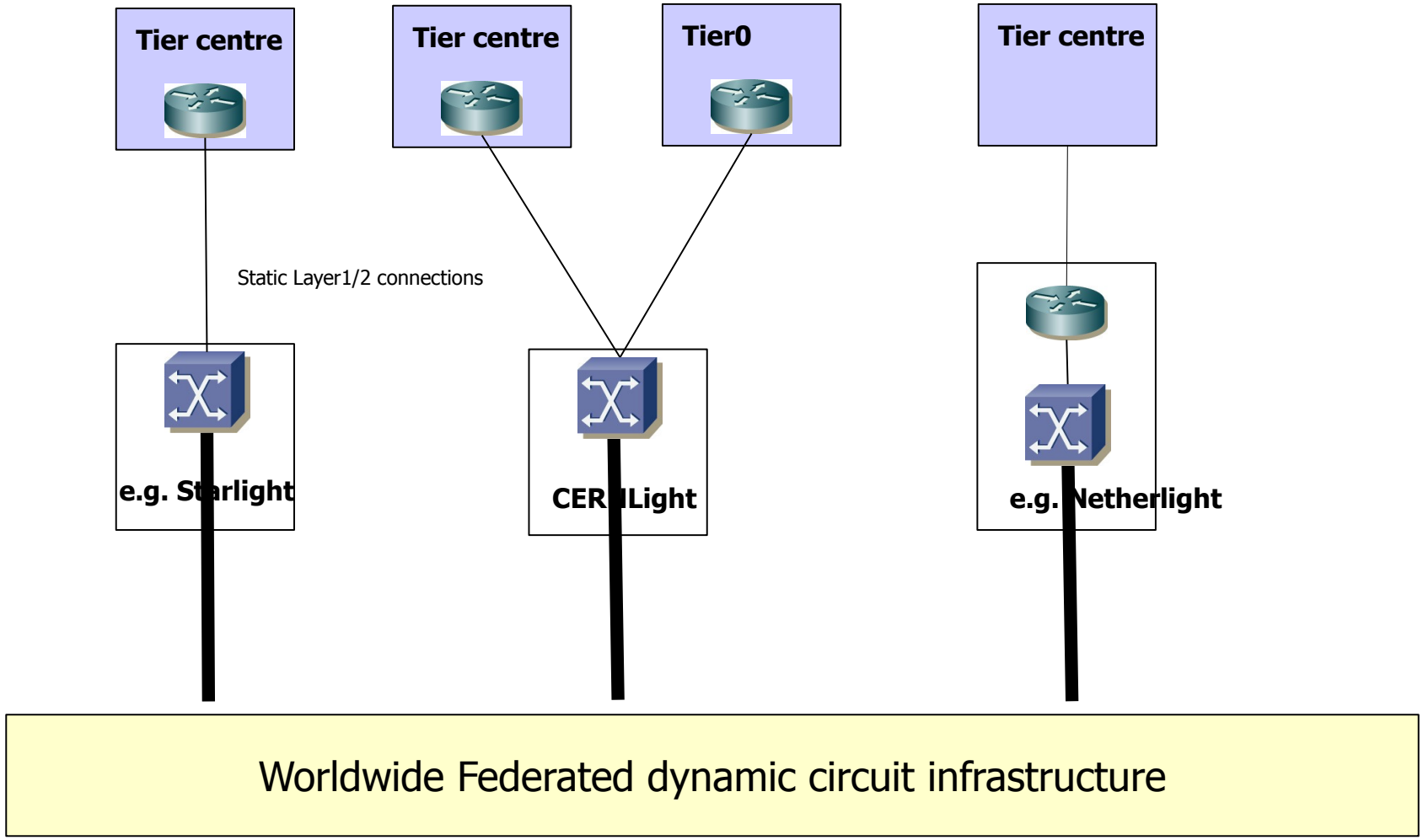
Few R&E Network Operators build a Dynamic Circuits Infrastructure that connects several key locations(access points).

The sites connects their network to one access point of this infrastructure.

Dynamic circuits are built between pairs of sites, according to the requests coming from the WLCG applications.

(similar to Model A, but with longer tails to the sites and a more compact Dynamic Circuit Infrastructure with few access points)

Model C – Lightpath Exchange



Model C – Pro and Cons

Pro
- best use of R&E Networks' bandwidth
- Sites pay for long distance links only when necessary
- Agile Dynamic Circuits Infrastructure

Cons
- Not straightforward routing configuration of sites' routers
- Only sites permanently connected to an access point can be interconnected.
- WLCG applications must know which circuit to ask.

Model D: VLANs

Model D – VLANs

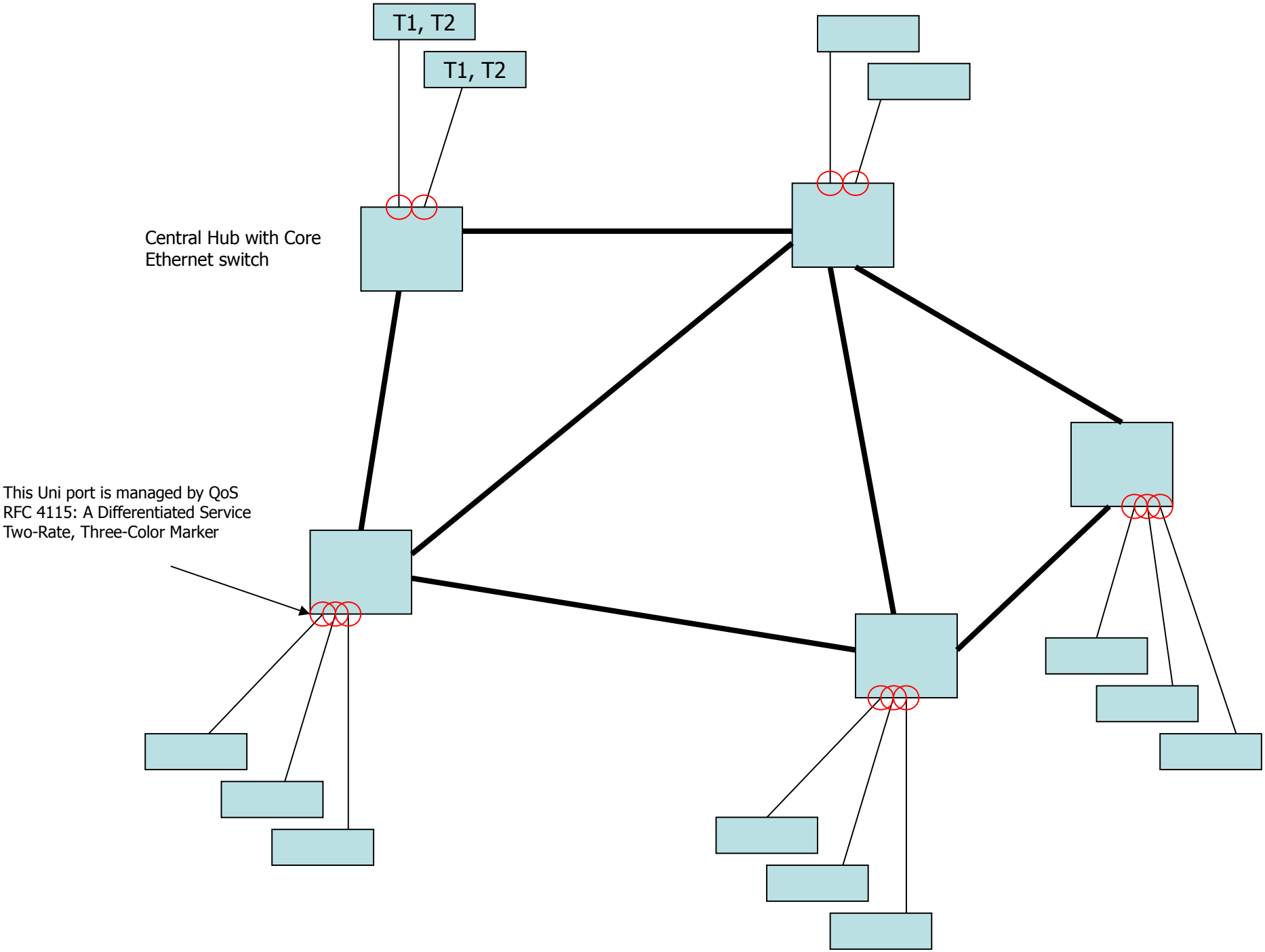
It is a centrally hubbed network for shared bandwidth on 10G (transatlantic) long distance links. Typically T1 and T2s will hub to these locations using dedicated links. Within the core network a mesh of VLANs should be configured to address the service needs of T1 and T2. With the QoS setting on the UNI port (say 100M CIR, and 1000M EIR), one can control bandwidth per T1 T2 and hopefully prevent congestion.

This is the static approach, but could be improved if the QoS of the UNI can be changed on the fly by human or application.

An intelligent protection mechanism is required, with per VLAN STP as a bare minimum. But the IS-IS conversion of PLSB may be very interesting as well.

This VLAN model is a not carrier grade approach, but may scale just with a limited number of end sites. Otherwise a carrier Ethernet / MPLS approach may be more applicable.

Model D – VLANs



Opinions?