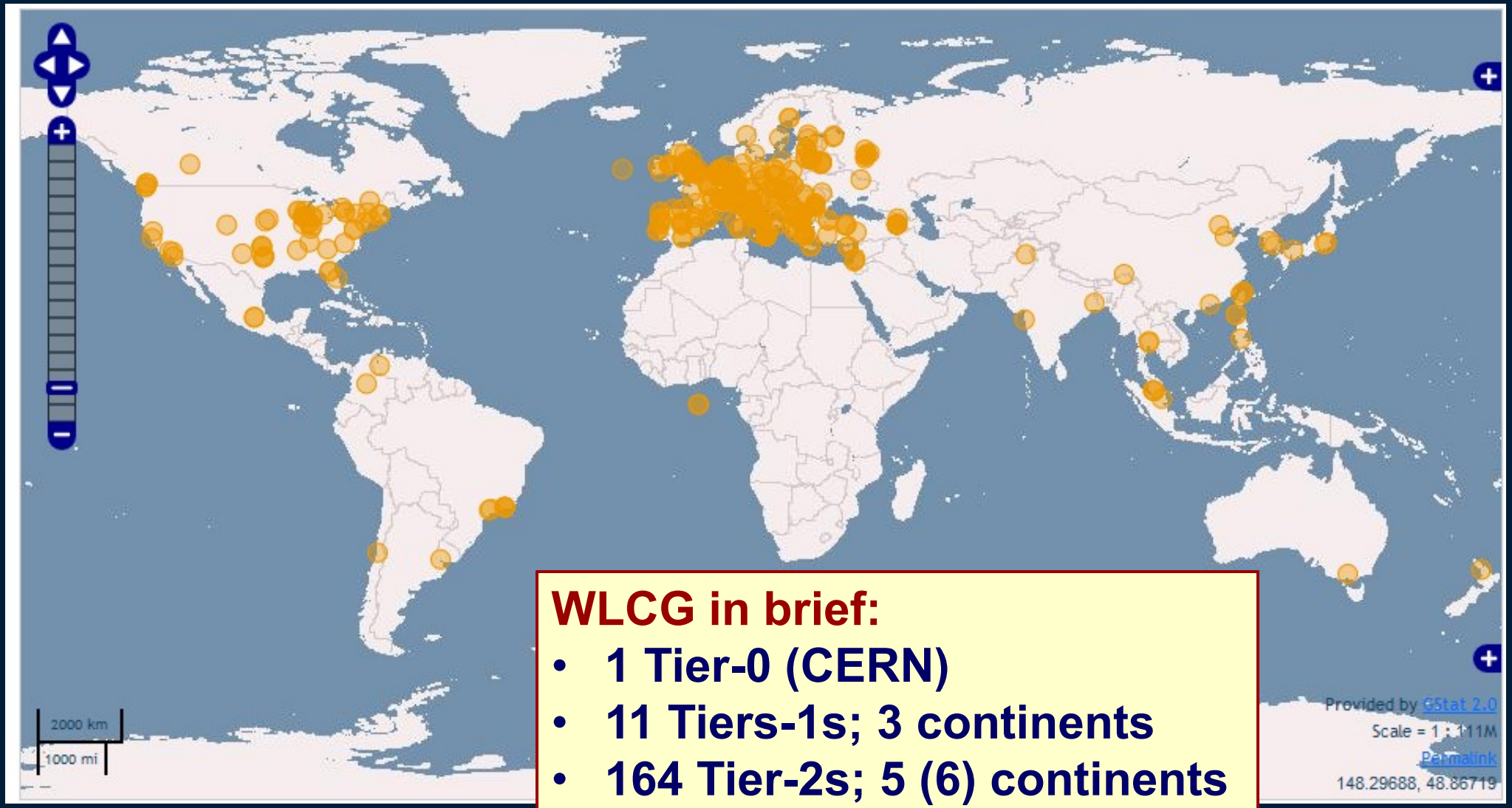
LHC Open Network Environment
LHC*ONE*

Artur Barczyk
California Institute of Technology
LISHEP Workshop on the LHC
Rio de Janeiro, July 9th, 2011

Path to LHCONE

- Started with Workshop on Transatlantic Connectivity for LHC experiments
  - June 2010 @ CERN
- Same time as changes in the computing models were being discussed in the LHC experiments
- Experiments provided a requirements document (Oct 2010)
  - Tasked LHCOPN with providing a proposal
- LHCT2S group was formed from within the LHCOPN
- LHCT2S Meeting in Geneva in January 2011
  - Discussion of 4 proposals, led to formation of a small working group drafting an architectural proposal based on these 4 documents
- LHCOPN Meeting in Lyon in February 2011
  - Draft architecture approved, finalised as “v2.2”
- LHCONE meeting in Washington, June 2011

LHC Computing Infrastructure



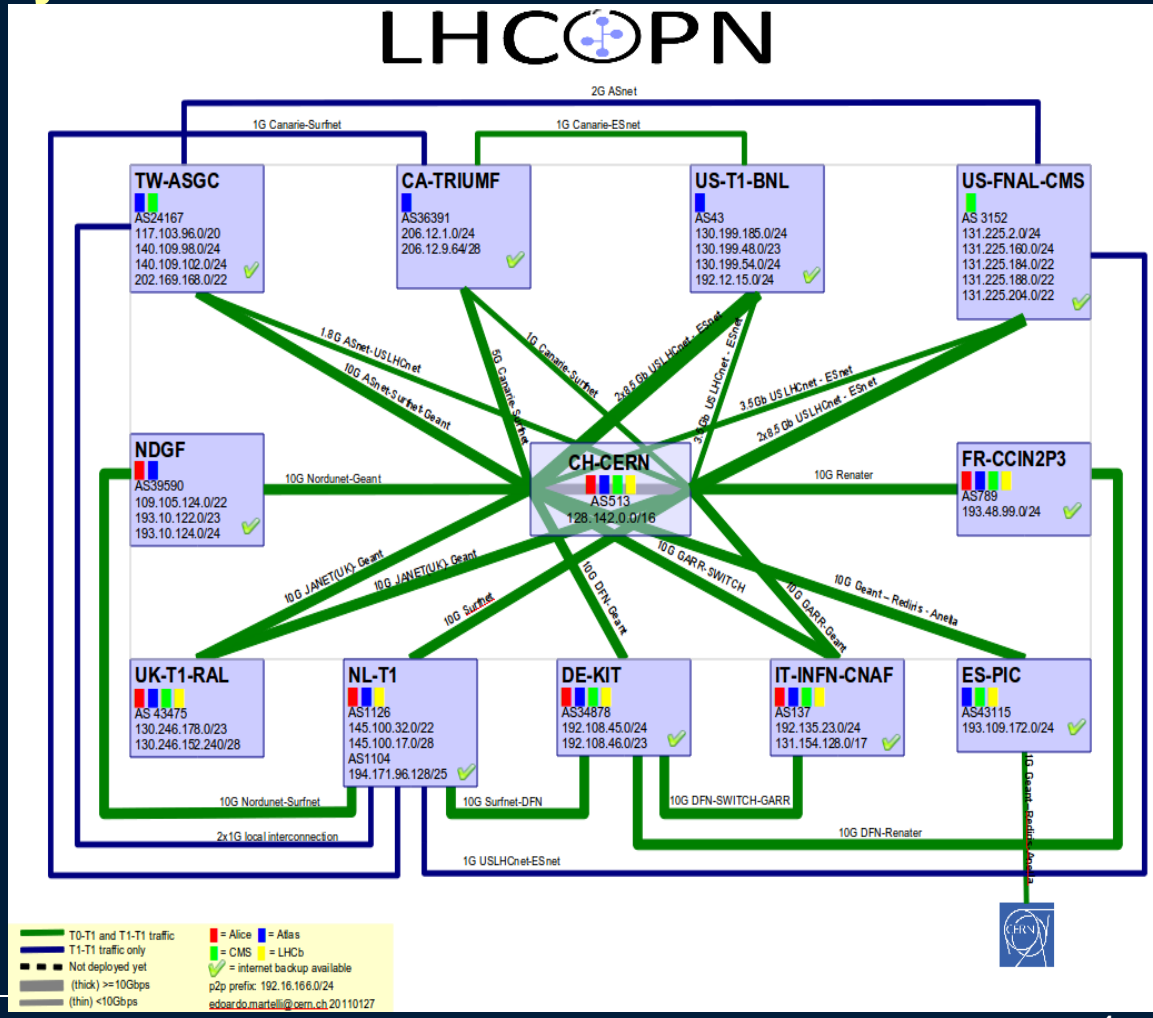
WLCG in brief:

- 1 Tier-0 (CERN)
- 11 Tiers-1s; 3 continents
- 164 Tier-2s; 5 (6) continents

Plus O(300) Tier-3s worldwide

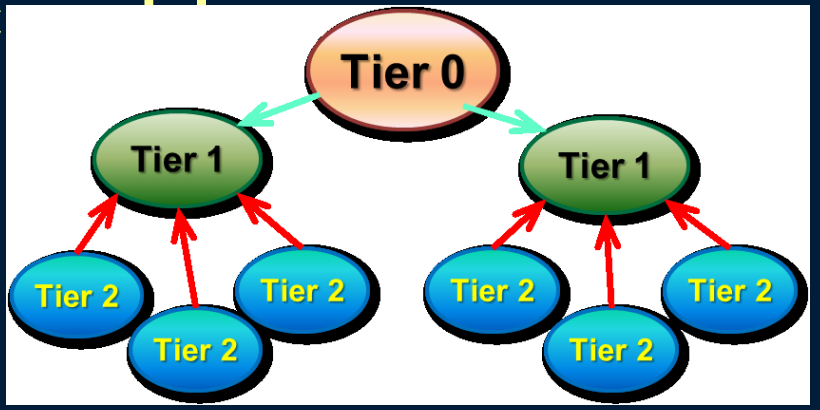
The current LHCOPN

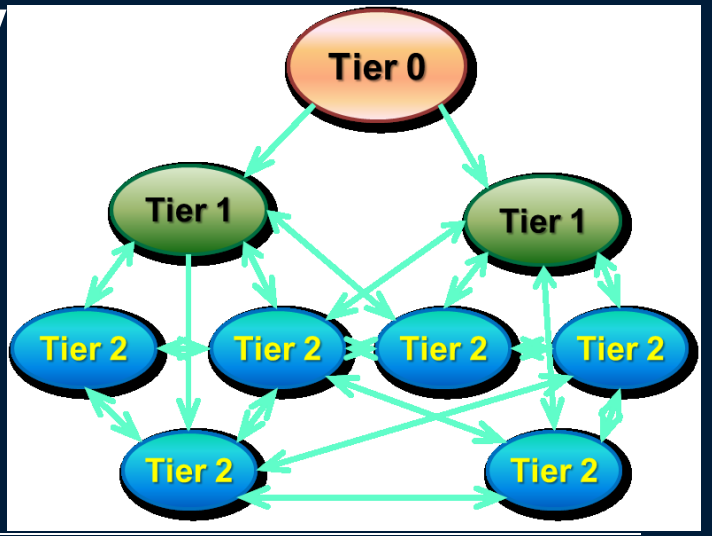
- Dedicated network resources for Tier0 and Tier1 data movement
- 130 Gbps total Tier0-Tier1 capacity
- Simple architecture
  - Point-to-point Layer 2 circuits
  - Flexible and scalable topology
- Grew organically
  - From star to partial mesh
  - Open to technology choices
    - have to satisfy requirements
- Federated governance model
  - Coordination between stakeholders
  - No single administrative body required



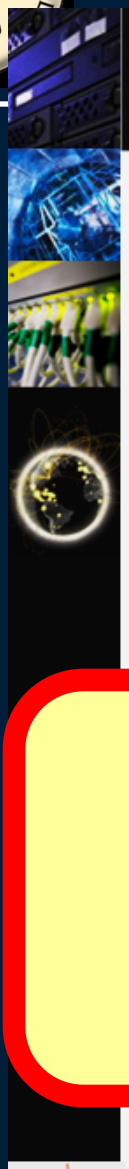
Moving to New Computing Models

- Moving away from the strict MONARC
- 3 recurring themes:
  - Flat(ter) hierarchy: Any site can use any other site as source of data
  - Dynamic data caching: Analysis sites will pull datasets from other sites "on demand", including from Tier2s in other regions
    - Possibly in combination with strategic pre-placement of data sets
  - Remote data access: jobs executing locally using data cached at a remote site in quasi-real time
    - Possibly in combination with local caching
- Expect variations by experiment





Implications for networks



- Hierarchy of Tier 0, 1, 2 no longer so important
- Tier 1 and Tier 2 may become more equivalent for the network
- Traffic could flow more between countries as well as within (already the case for CMS)

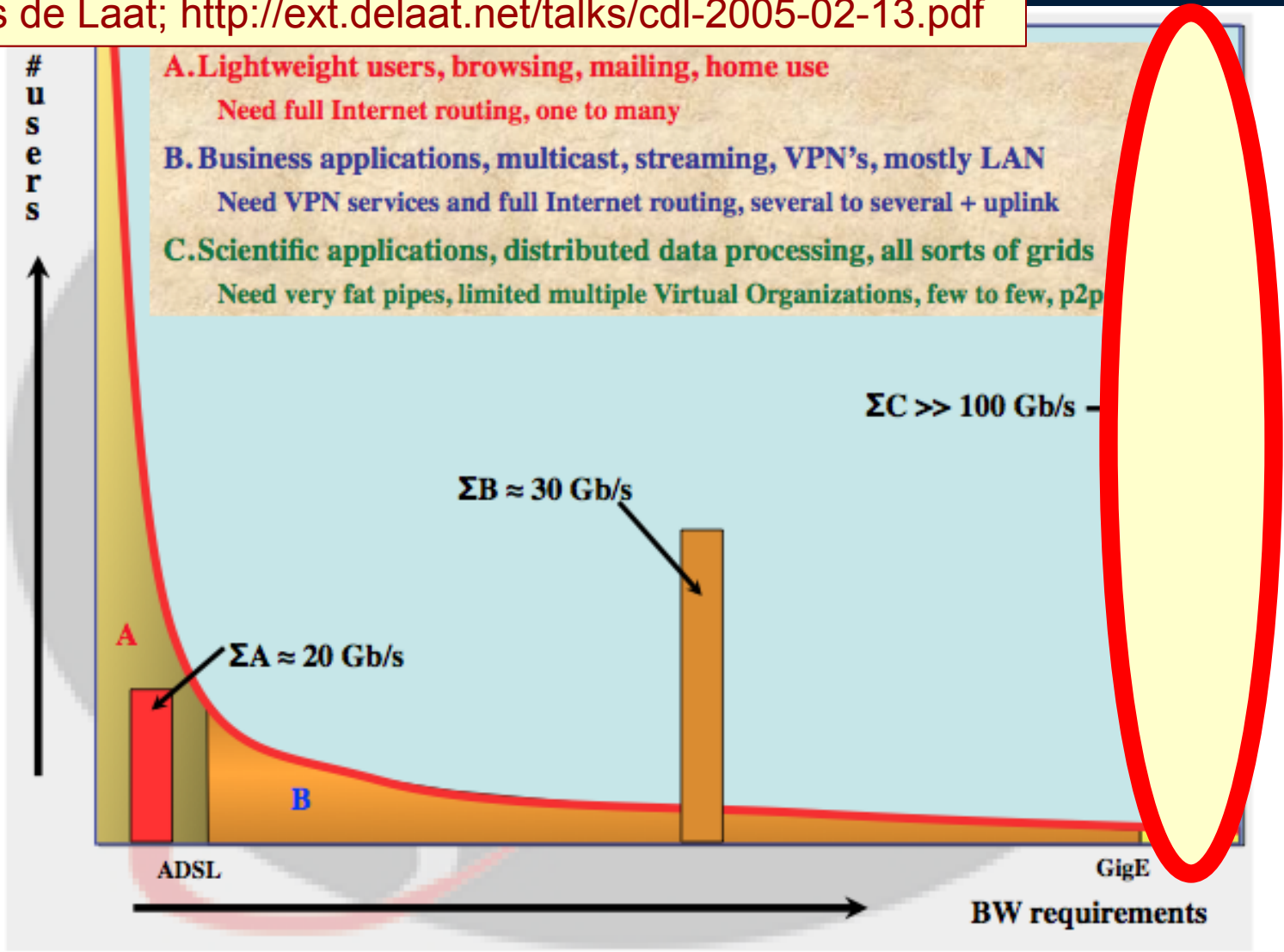
Ian Bird, CHEP conference, Oct 2010

Why LHCONE?

- Next generation computing models will be more network-intensive
- LHC data movements have already started to saturate some main (e.g. transatlantic GP R&E) links
  - Guard against "defensive actions" by GP R&E providers
- We cannot simply count on General Purpose Research & Education networks to scale up
  - LHC is currently *the* power-user
  - Other science fields start creating large data flows as well

Characterization of User Space

Cees de Laat; http://ext.delaat.net/talks/cdl-2005-02-13.pdf



This is where LHC users are

LHC*ONE*

HTTP://LHCONE.NET

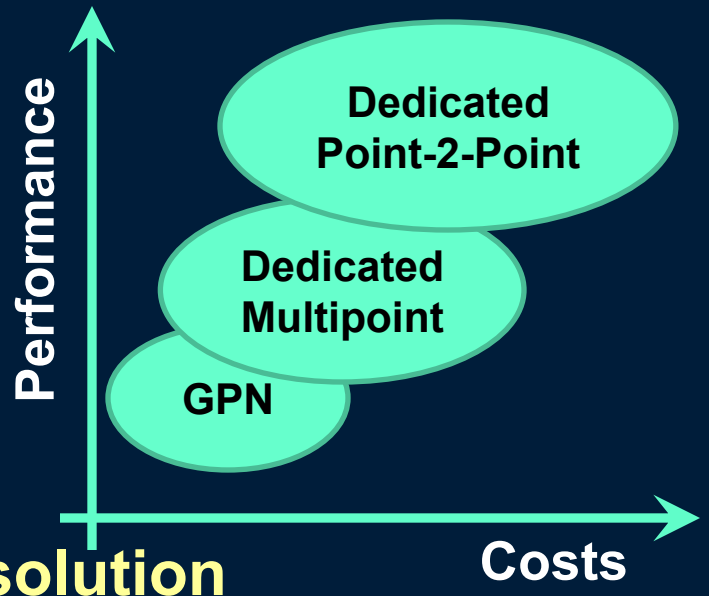
The requirements, architecture, services

Requirements summary (from the LHC experiments)

- **Bandwidth:**
  - **Ranging from 1 Gbps (Minimal site) to 5-10Gbps (Nominal) to N x 10 Gbps (Leadership)**
  - **No need for full-mesh @ full-rate, but several full-rate connections between Leadership sites**
  - **Scalability is important,**
    - sites are expected to migrate **Minimal → Nominal → Leadership**
    - Bandwidth growth: Minimal = 2x/yr, Nominal&Leadership = 2x/2yr
- **Connectivity:**
  - **Facilitate good connectivity to so far (network-wise) under-served sites**
- **Flexibility:**
  - **Should be able to include or remove sites at any time**
- **Budget Considerations:**
  - **Costs have to be understood, solution needs to be affordable**

Some Design Considerations

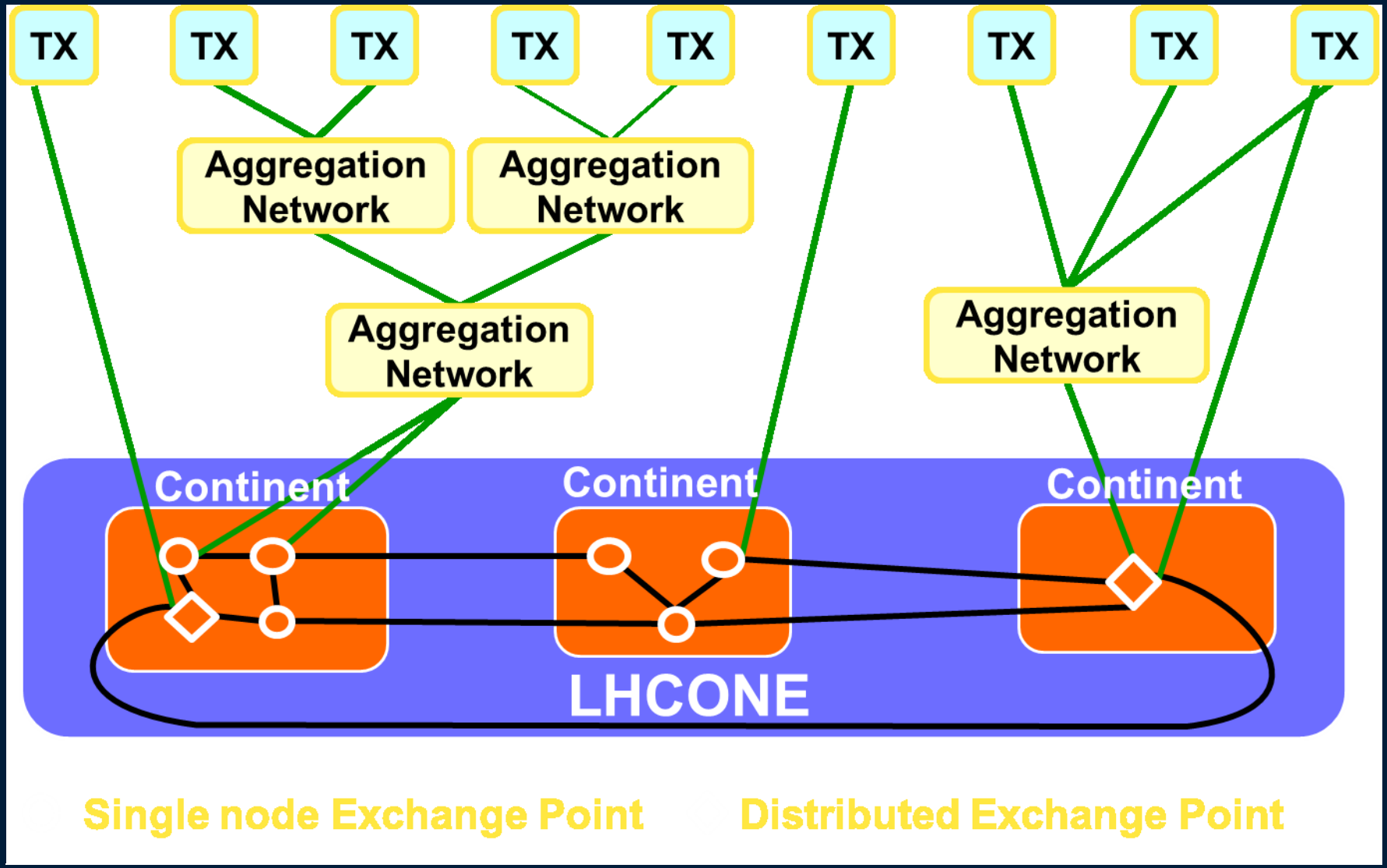
- So far, T1-T2, T2-T2, and T3 data movements have been mostly using General Purpose Network infrastructure
  - Shared resources (with other science fields)
  - Mostly best effort service
- Increased reliance on network performance → need more than best effort
  - Separate large LHC data flows from routed GPN
- Collaboration on global scale, diverse environment, many parties
  - Solution to be Open, Neutral and Diverse
  - Agility and Expandability
    - Scalable in bandwidth, extent and scope
- Allow to choose the most cost effective solution
- Organic activity, growing over time according to needs



LHC*ONE* Architecture

- **Builds on the Hybrid network infrastructures and Open Exchanges**
  - **To build a global unified service platform for the LHC community**
- **LHC*ONE*'s architecture incorporates the following building blocks**
  - **Single node Exchange Points**
  - **Continental / regional Distributed Exchanges**
  - **Interconnect Circuits between exchange points**
    - Likely by allocated bandwidth on various (possibly shared) links to form LHCONE
- **Access method to LHCONE is chosen by the end-site, alternatives may include**
  - **Dynamic circuits**
  - **Fixed lightpaths**
  - **Connectivity at Layer 3, where/as appropriate**
- **We envisage that many of the Tier-1/2/3s may connect to LHCONE through aggregation networks**

High-level Architecture, Pictorial

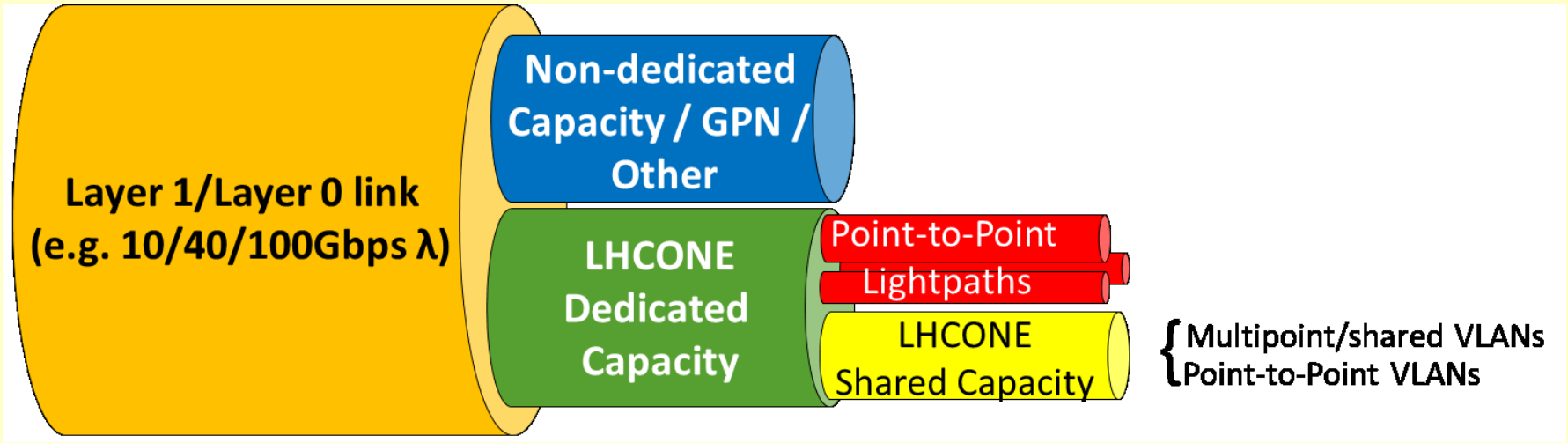


LHC*ONE* Network Services
Offered to Tier1s, Tier2s and Tier3s

- **Shared Layer 2 domains: separation from non-LHC traffic**
  - IPv4 and IPv6 router addresses on shared layer 2 domain(s)
  - Private shared layer 2 domains for groups of connectors
  - Layer 3 routing is between and up to the connectors
    - A set of Route Servers will be available
- **Point-to-point layer 2 connections: per-channel traffic separation**
  - VLANS without bandwidth guarantees between pairs of connectors
- **Lightpath / dynamic circuits with bandwidth guarantees**
  - Lightpaths can be set up between pairs of connectors
- **Monitoring: perfSONAR archive**
  - current and historical bandwidth utilization and availability statistics
- **This list of services is a starting point and not necessarily exclusive**
- **LHCONE** does not preclude continued use of the general R&E network infrastructure by the Tier1s, Tier2s and Tier3s - where appropriate

Dedicated/Shared Resources

- LHCONE concept builds on traffic separation between LHC high impact flows, and non-LHC traffic
  - Avoid negative impact on other research traffic
  - Enable high-performance LHC data movement
- Services to use resources allocated to LHCONE



- Prototype/Pilot might use non-dedicated resources, but need to be careful about evaluation metrics

The Case for Dynamic Circuits in LHC Data Processing

- Data models do not require full-mesh @ full-rate connectivity @ all times
- On-demand data movement will augment and partially replace static pre-placement → Network utilisation will be more dynamic and less predictable
- Performance expectations will not decrease
  - More dependence on the network, for the whole data processing system to work well!
- Need to move large data sets fast between computing sites
  - On-demand: caching
  - Scheduled: pre-placement
  - *Transfer* latency is important
- Network traffic in excess of what was anticipated
- As data volumes grow rapidly, and experiments rely increasingly on the network performance - what will be needed in the future is
  - More bandwidth
  - More efficient use of network resources
  - Systems approach including end-site resources and software stacks

Issues With Demanding Users

- There are more and more of them.
- The swamping of IP infrastructures with traffic from "well connected sites"
  - Occurs when the capability of a site are approaching that of the routed IP network.
  - Looks like a "denial of service" to the other users.
- Solution 1: Build a bigger routed IP network.
  - A big investment to solve a problem for relatively few users.
  - All domains in any end-end path must do the same.
  - Only temporary, new users will come with bigger requirements.
- Solution 2: Give the sites "what they need when they need it".
  - May be considered as "Just in time provisioning"
  - Has led to the circuit approach.

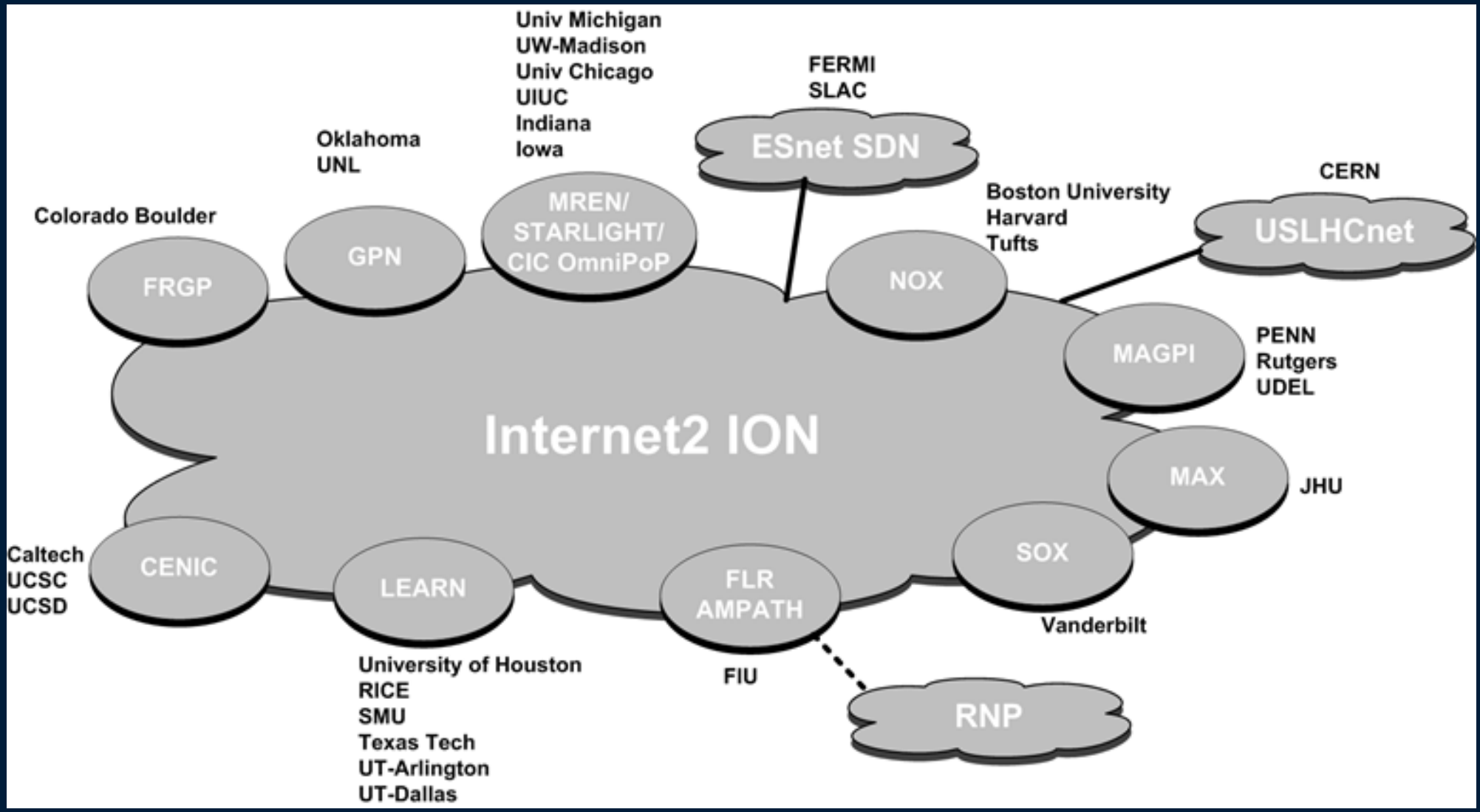
David Foster, CERN

David Foster; 1st TERENA ASPIRE workshop, May 2011

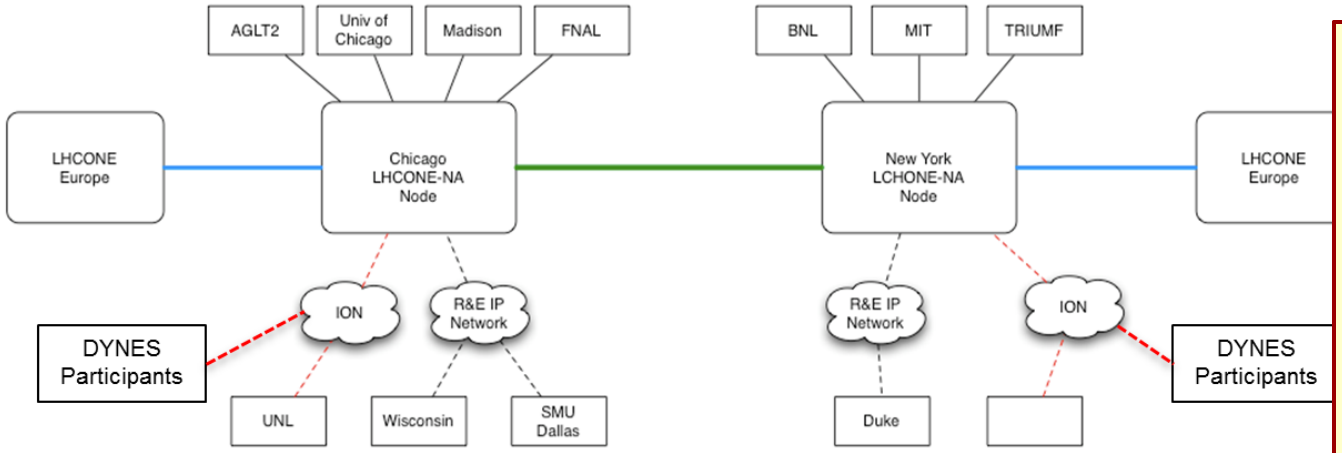
Dynamic Bandwidth Allocation

- Will be one of the services to be provided in LHCONE
- Allows to allocate network capacity on as-needed basis
  - Instantaneous ("Bandwidth on demand"), or
  - Scheduled allocation
- Dynamic Circuit Service is present in several networks
  - Internet2, ESnet, SURFnet, US LHCNet
- Planned (or in experimental deployment) in others
  - E.g. GEANT, RNP, …
- DYNES: NSF funded project to extend hybrid & dynamic network capabilities to campus & regional networks
  - In first deployment phase; fully operational in 2012

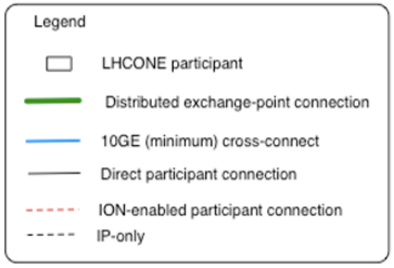
DYNES Deployment Topology

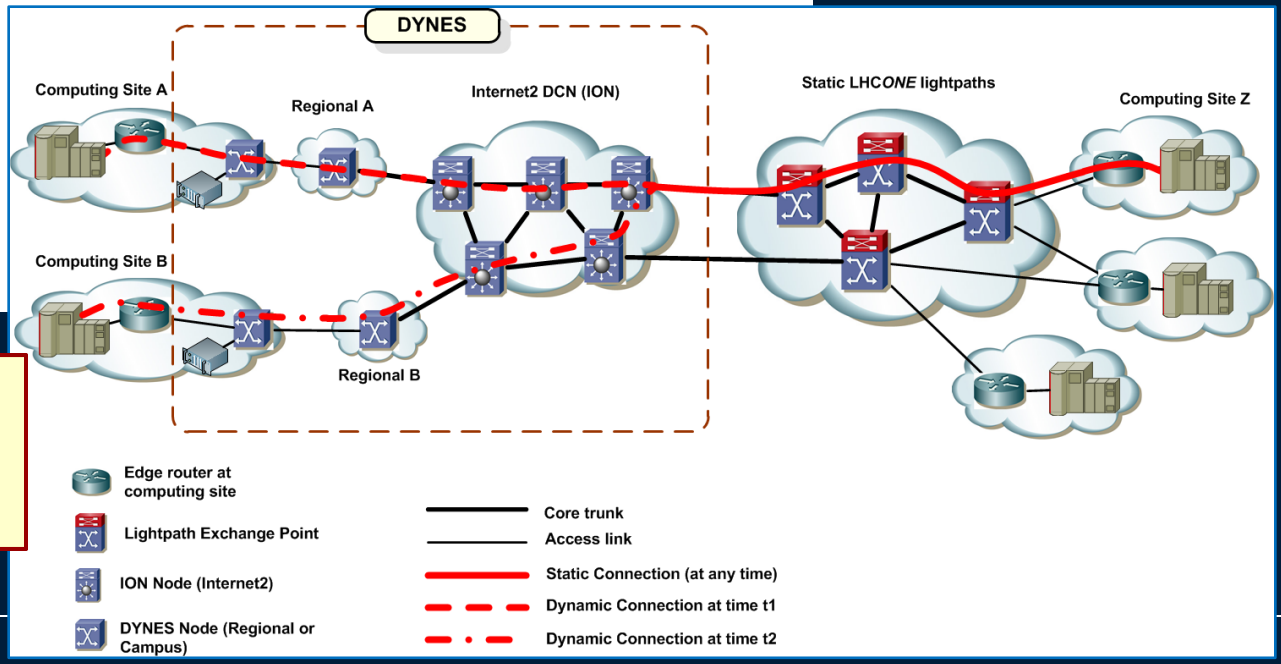


LHCONE + DYNES



- DYNES Participants can dynamically connect to Exchange Points via ION Service
- Dynamic Circuits through and beyond the exchange point?
- Static tail?

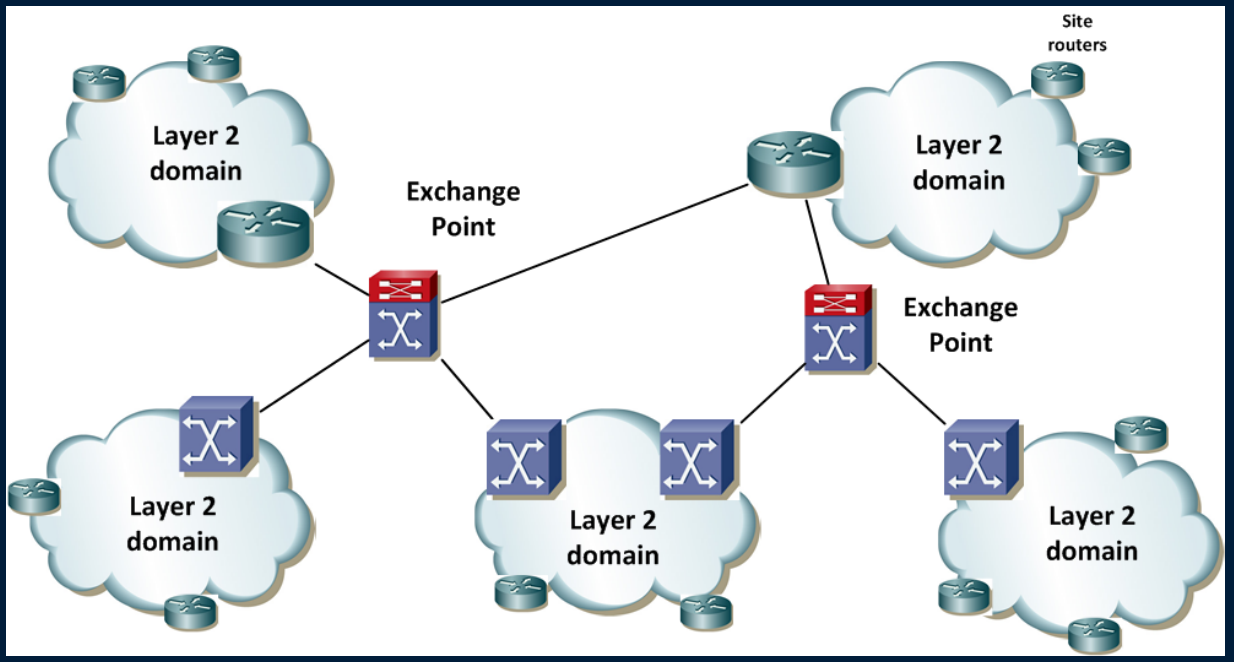




- Hybrid dynamic circuit and IP routed segment model?

LHCONE Pilot Implementation

- To include a number of sites identified by the CMS and Atlas experiments
- It is expected that LHCONE will grow organically from this implementation
- Currently operational: multipoint service using
  - 4 Open Exchange Points
    - CERNLight, Netherlight, MANLAN and Starlight
  - Dedicated core capacity
    - SURFnet, US LHCNet
  - Route server at CERN
- Architecture working group is finalizing inter-domain connectivity design
  - GEANT+ 4 NRENs
  - Internet2, ESnet
  - Other Open Exchanges
  - Connections to South America and Asia



Summary

- LHCONE will provide dedicated network connectivity for the LHC computing sites
  - Built on the infrastructure provided by the R&E Networks
  - Collaborative effort between the experiments, CERN, the networks and the sites
- Will provide 4 services
  - Static point-to-point
  - Dynamic point-to-point
  - Multipoint
  - Monitoring
- Pilot is currently being implemented
- LHCONE will grow organically according to requirements and funding

THANK YOU!

http://lhcone.net

Artur.Barczyk@cern.ch

EXTRA SLIDES

LHC*ONE* Policy Summary

- **LHCONE policy will be defined and may evolve over time in accordance with the governance model**
- **Policy Recommended for LHCONE governance**
  - Any Tier1/2/3 can connect to LHCONE
  - Within LHCONE, transit is provided to anyone in the Tier1/2/3 community that is part of the LHCONE environment
  - Exchange points must carry all LHC traffic offered to them (and only LHC traffic), and be built in carrier-neutral facilities so that any connector can connect with its own fiber or using circuits provided by any telecom provider
  - Distributed exchange points: same as above + the interconnecting circuits must carry all the LHC traffic offered to them
  - **No additional restrictions can be imposed on LHCONE by the LHCONE component contributors**
- **The Policy applies to LHCONE components, which might be switches installed at the Open Exchange Points, or virtual switch instances, and/or (virtual) circuits interconnecting them**

LHC*ONE* Governance Summary

- Governance is proposed to be similar to the LHCOPN, since like the LHCOPN, LHCONE is a community effort
  - Where all the stakeholders meet regularly to review the operational status, propose new services and support models, tackle issues, and design, agree on, and implement improvements
- Includes connectors, exchange point operators, CERN, and the experiments; 4 working groups
  - Governance, Architecture, Operations, Stakeholders
- Defines the policies of LHCONE and requirements for participation
  - It does not govern the individual participants
- Is responsible for defining how costs are shared
- Is responsible for defining how resources of LHCONE are allocated