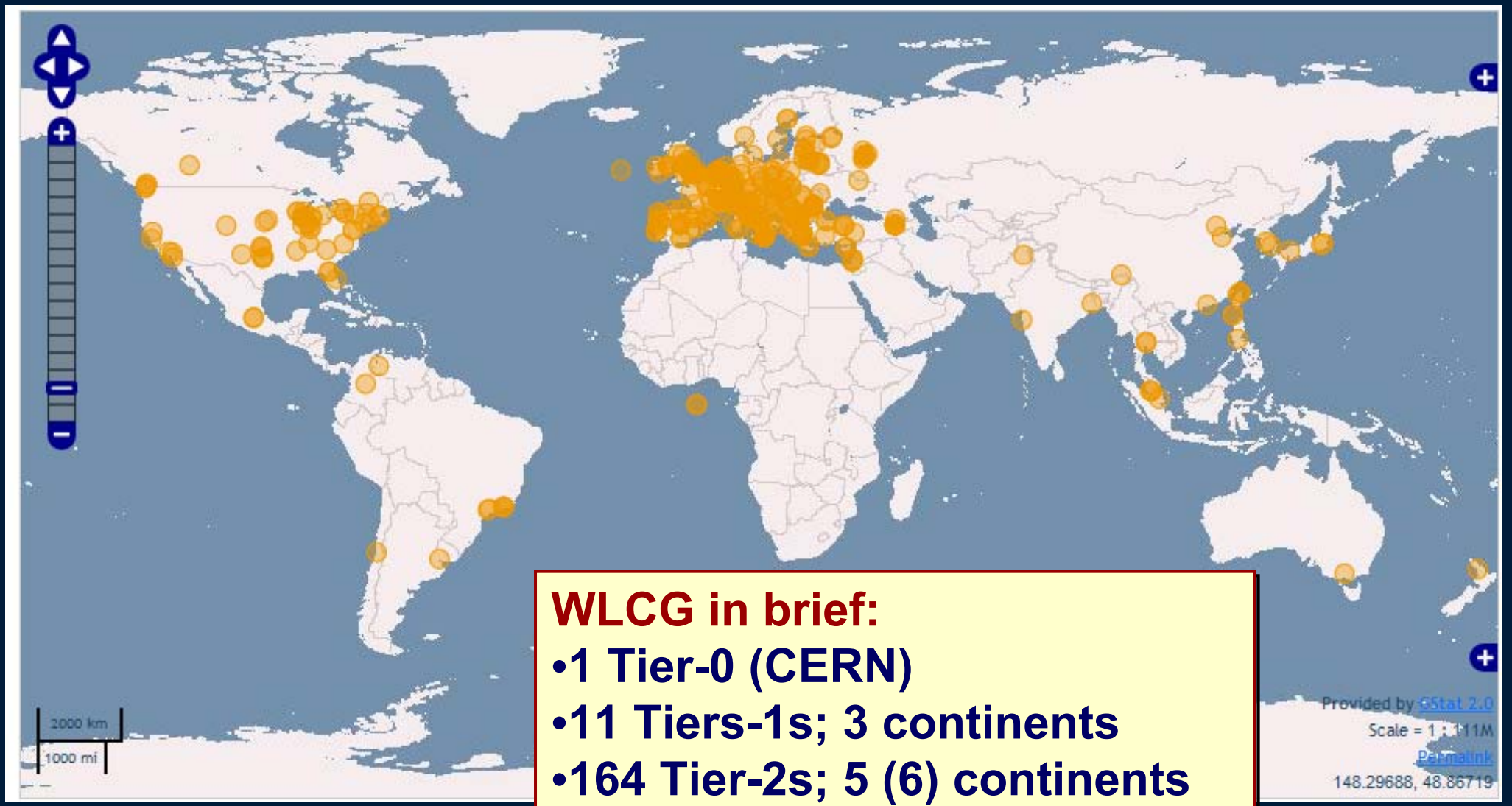
LHC Open Network Environment
LHC*ONE*

Artur Barczyk
California Institute of Technology
11th Annual Global LambdaGrid Workshop
Rio de Janeiro, September 13th, 2011

INTRODUCTION

A bit of background

LHC Computing Infrastructure



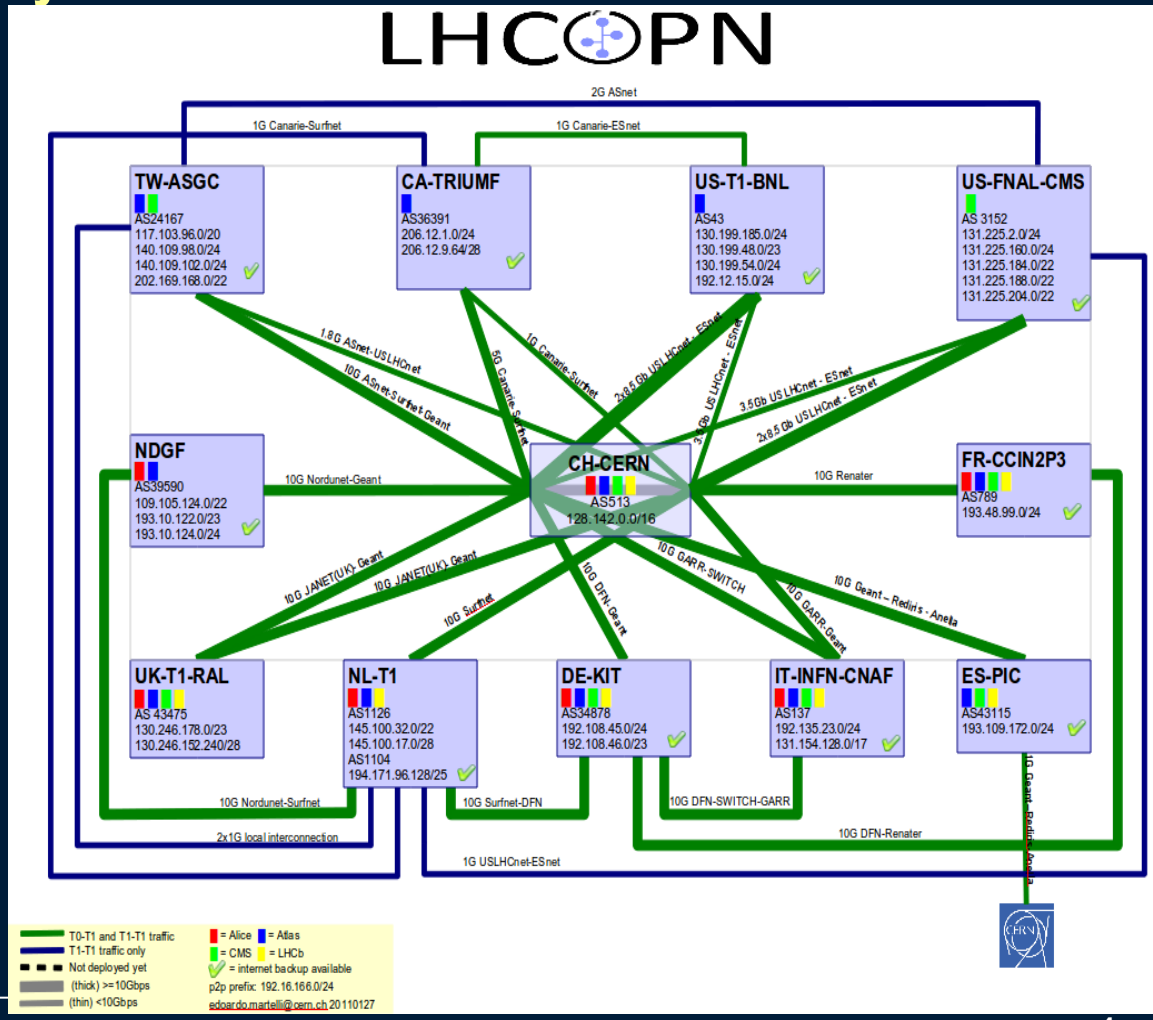
WLCG in brief:

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

Plus O(300) Tier-3s worldwide

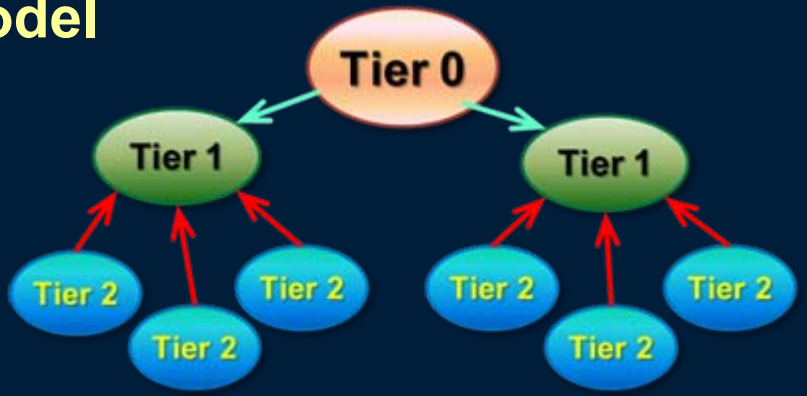
The LHCOPN – Serving Tier0 and Tier1 sites

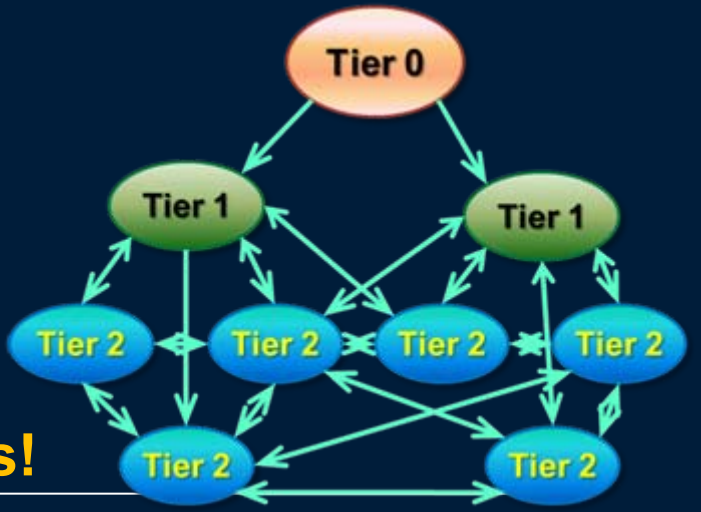
- Dedicated network resources for Tier0 and Tier1 data movement
- 130 Gbps total Tier0-Tier1 capacity
- Simple architecture
  - Point-to-point Layer 2 circuits
  - Flexible and scalable topology
- Grew organically
  - From star to partial mesh
  - Open to technology choices
    - have to satisfy requirements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



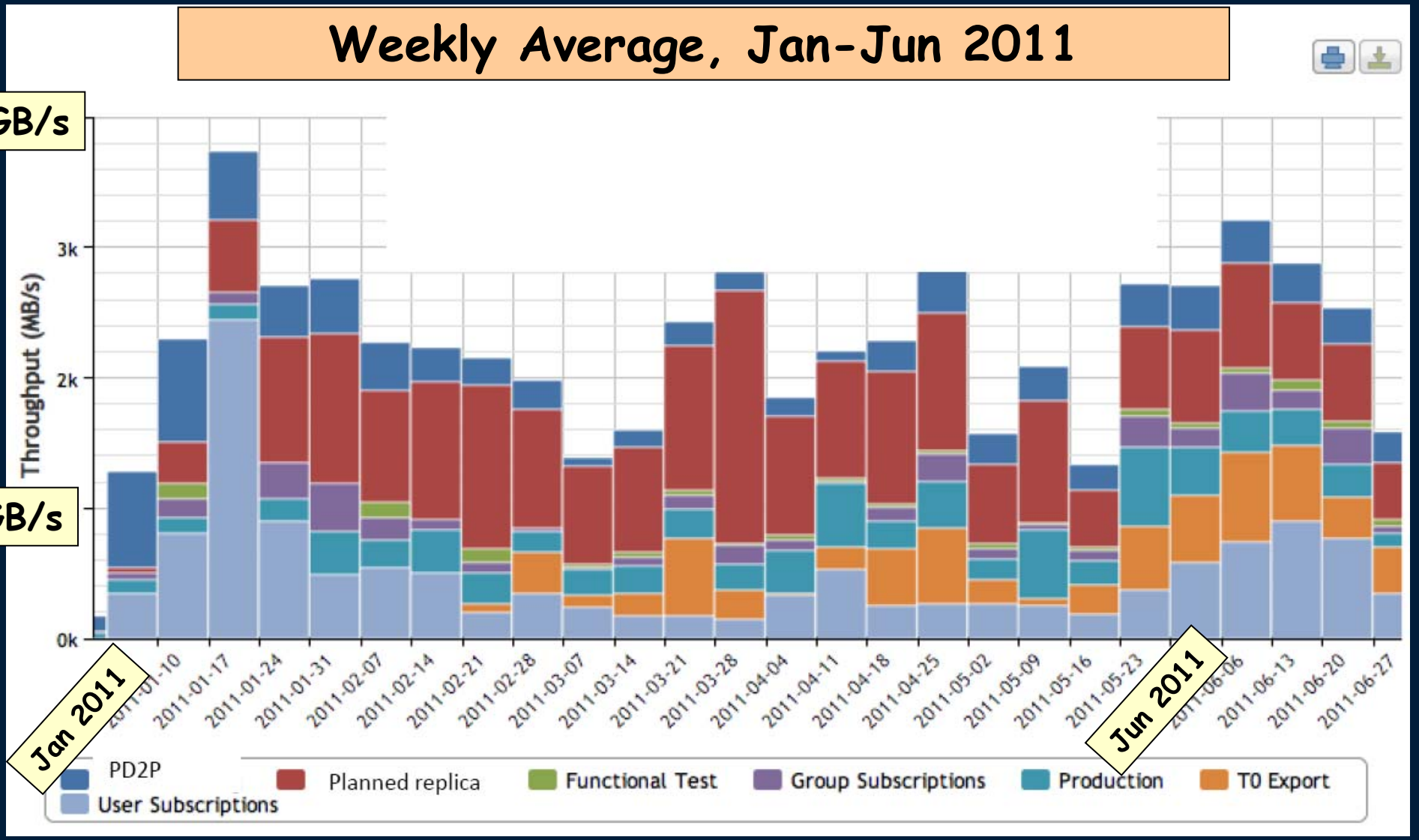
Moving to New Data Models

- Moving away from the strict MONARC model
  - Gradually progressing since 2010
- 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
- Variations by experiment
- But: LHCONE connects only Tier0 and Tier1s!

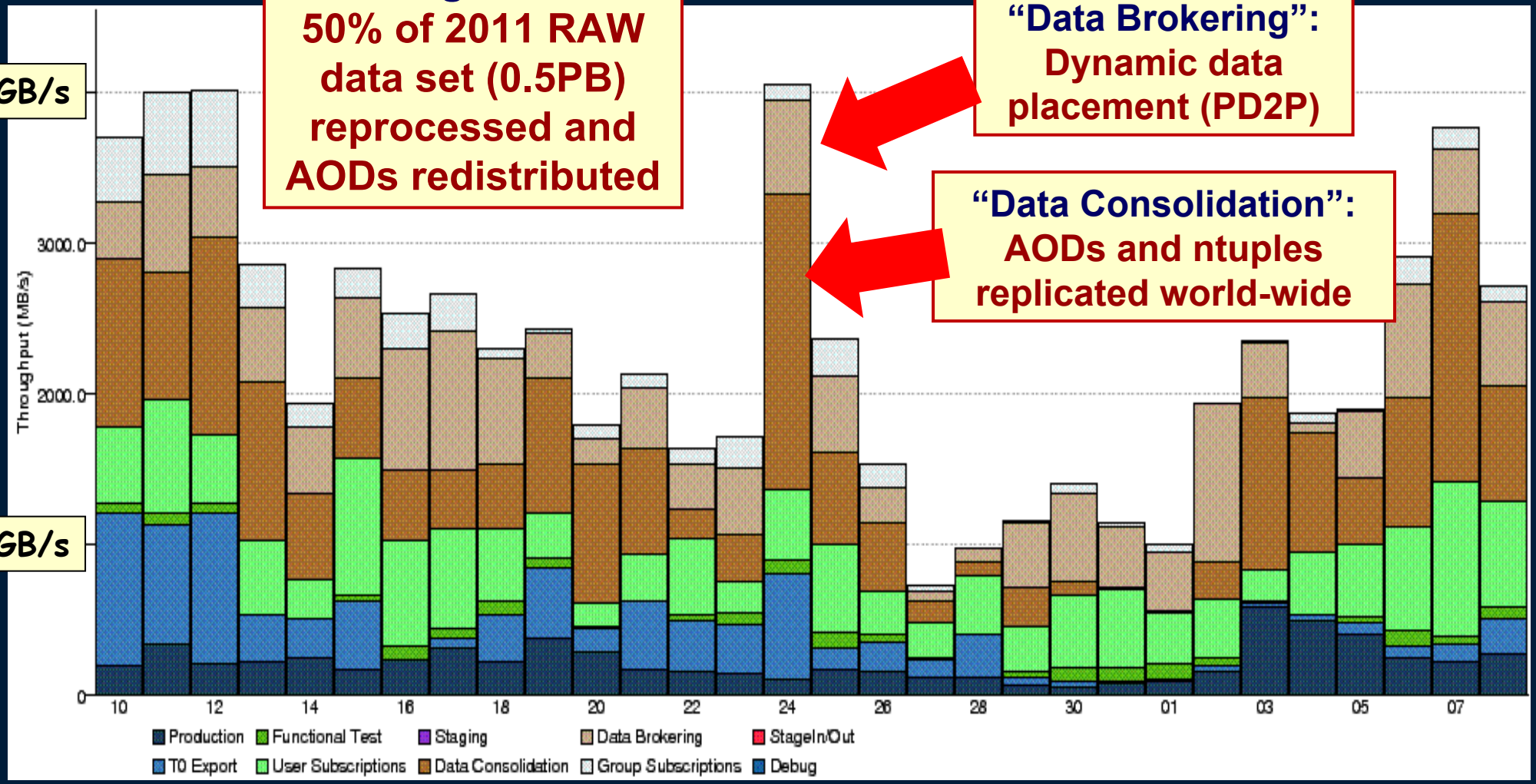




Atlas 2011 Data Movement

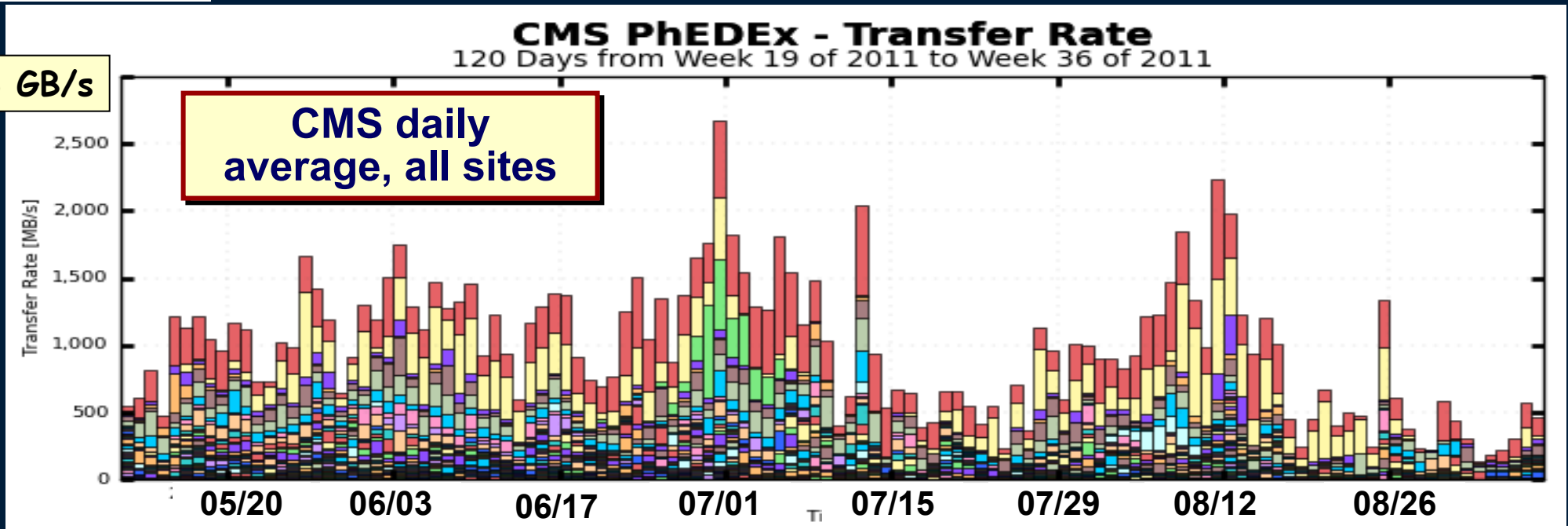


Atlas 2011 Data Movement
Daily Averages, All Sites



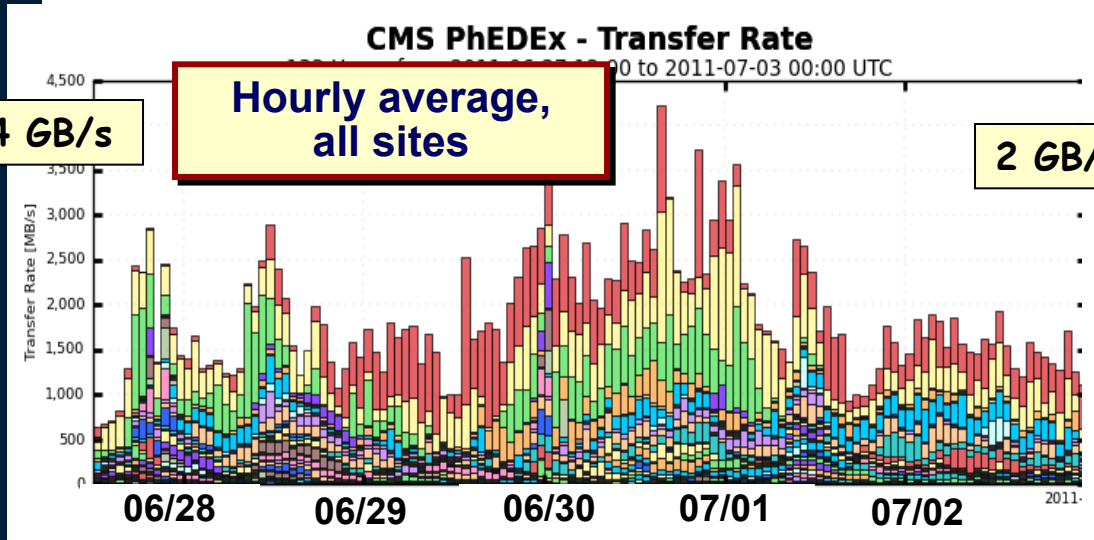
CMS Data Movements

(All Sites, Tier1-Tier2)



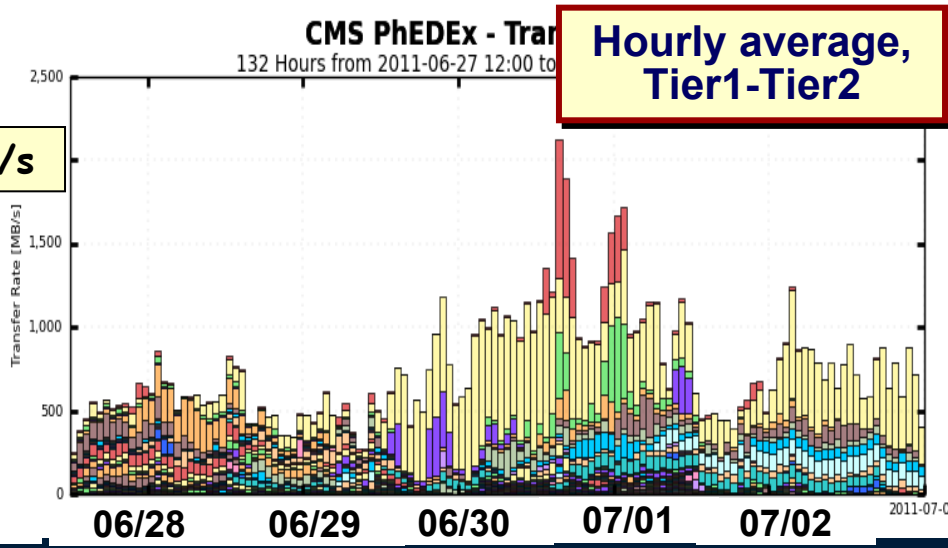
3 GB/s

CMS daily average, all sites



4 GB/s

Hourly average, all sites



2 GB/s

Hourly average, Tier1-Tier2

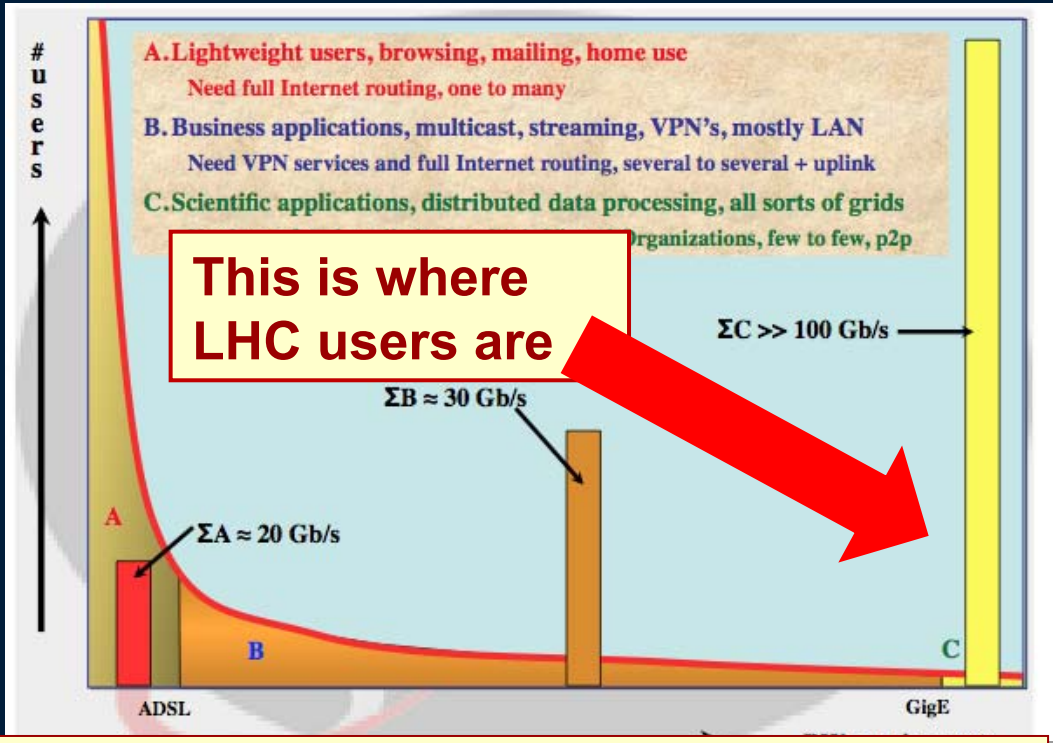
LHCONE

HTTP://LHCONE.NET

The why, what, how…

Motivation for LHCONE

- LHC data movements to/from Tier2 and Tier3 sites generate heavy flows in the R&E General Purpose Networks
  - Possible negative impact on other users
- New LHC computing models; more network reliance
- Integration with LHC computing models and operations
- Main target benefits for users:
  - Predictability in end-to-end data movement
- Main benefits for networks:
  - Avoid negative impact on non-LHC users
  - Capacity planning and traffic engineering based on user requirements



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

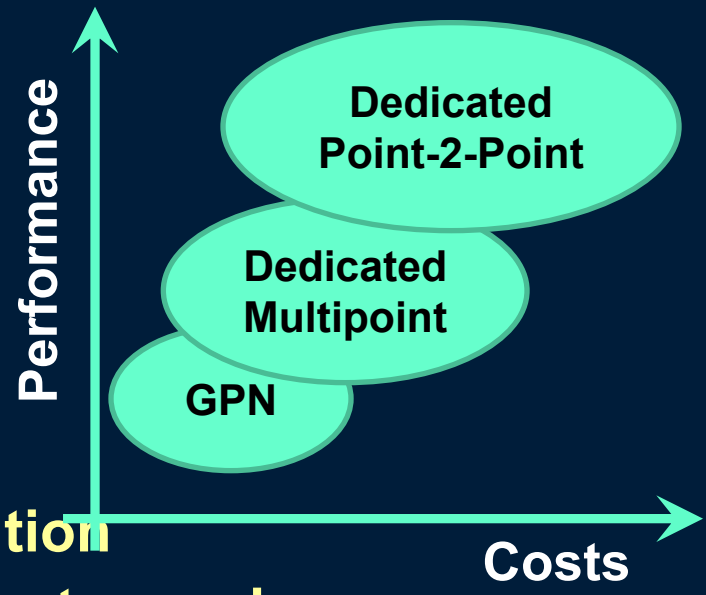
Requirements summary (from the LHC experiments)

- **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**

"Bos-Fisk" requirements paper available at http://lhcone.net

Some Design Considerations

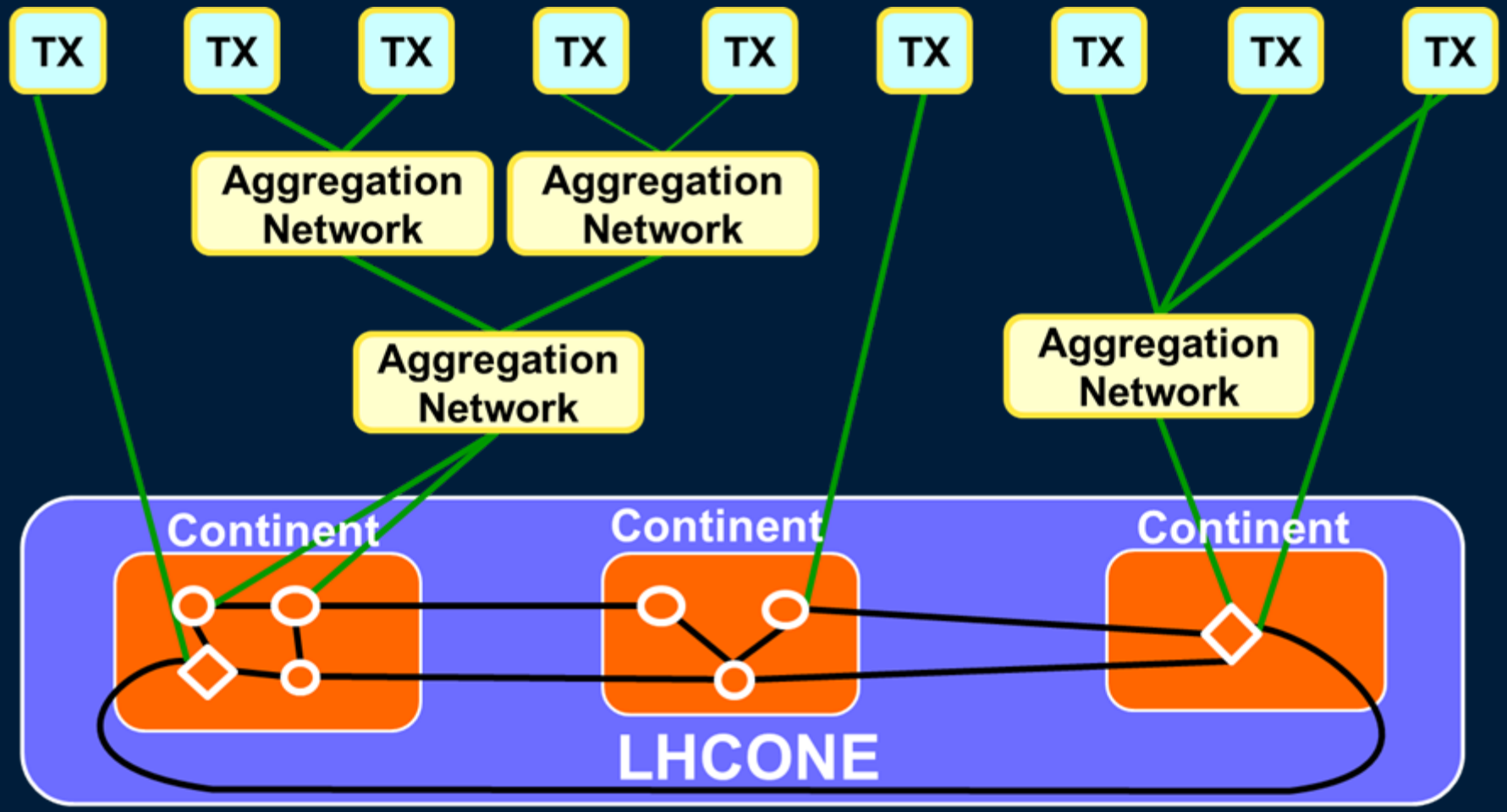
- So far, T1-T2, T2-T2, and T3 data movements have been using General Purpose Network infrastructure
  - Shared resources (with other science fields)
  - Mostly best effort service
- Increased reliance on network performance → need more than best effort
  - Separate large LHC data flows from routed GPN
- 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



LHCONE 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, Example

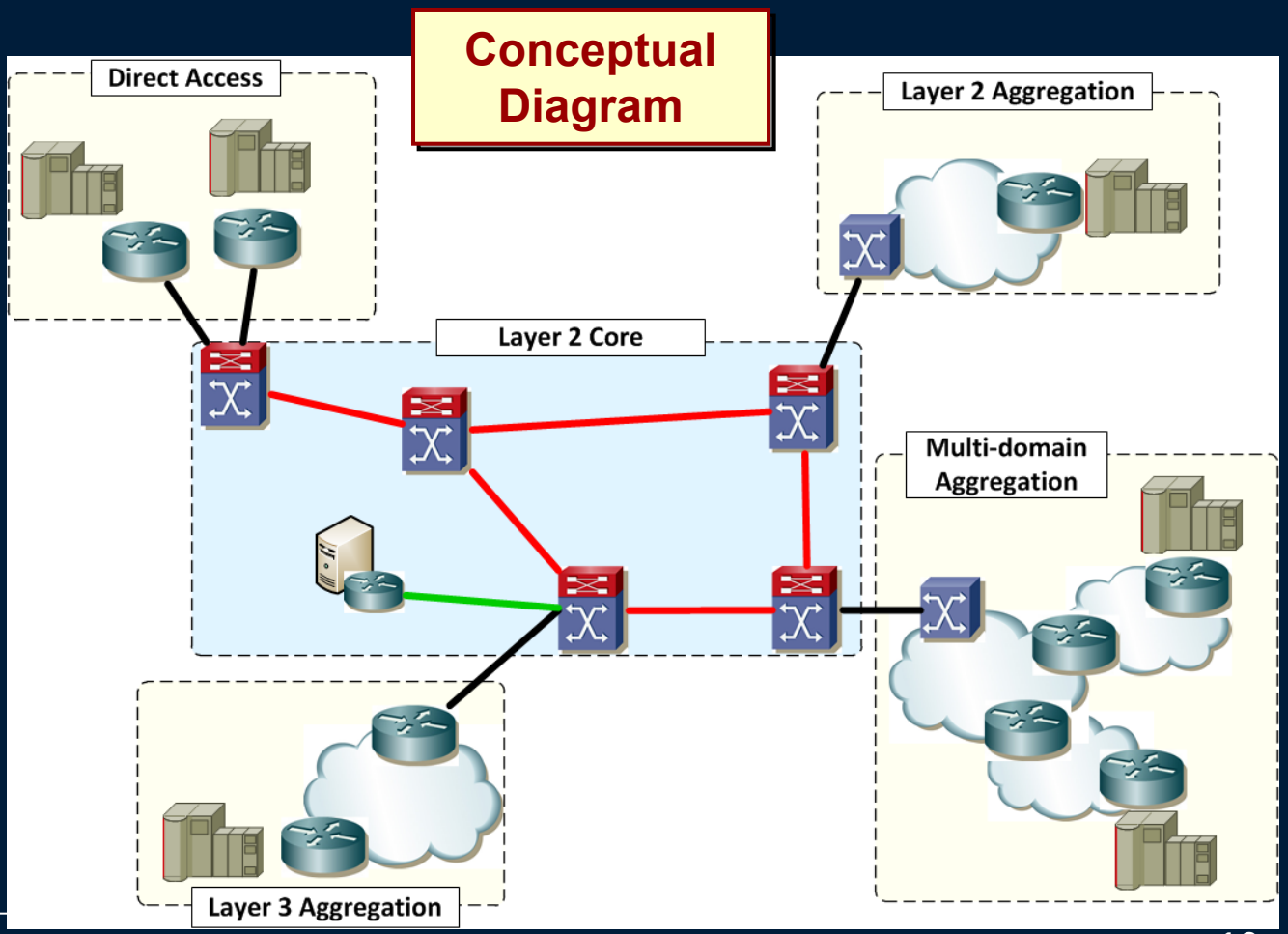


LHCONE Network Services
Offered to Tier1s, Tier2s and Tier3s

- **Shared Layer 2 domains (private VLAN broadcast domains)**
  - IPv4 and IPv6 addresses on shared layer 2 domain including all connectors
  - Private shared layer 2 domains for groups of connectors
  - Layer 3 routing is up to the connectors
    - A Route Server per continent is planned to be available
- **Point-to-point layer 2 connections**
  - VLANS without bandwidth guarantees between pairs of connectors
- **Lightpath / dynamic circuits with bandwidth guarantees**
  - Lightpaths can be set up between pairs of connectors
  - Circuit management: DICE IDC & GLIF Fenius now, OGF NSI when ready
- **Monitoring: perfSONAR archive now, OGF NMC based when ready**
  - Presented statistics: current and historical bandwidth utilization, and link availability statistics for any past period of time
- **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

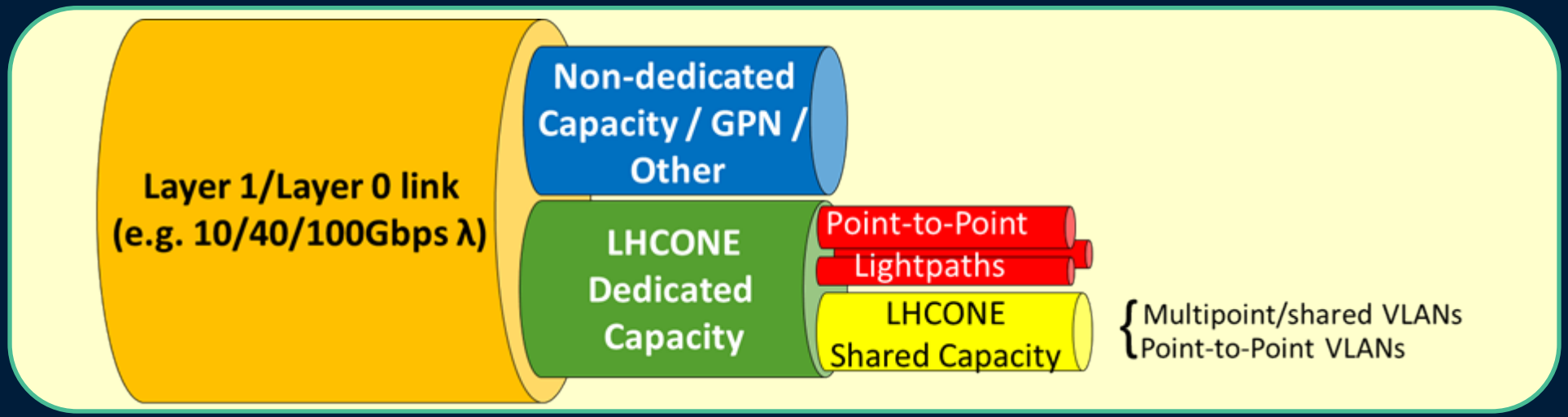
Multipoint service: Switched Core, Routed Edge

- Following "switch where you can, route where you must"
- LHCONE Layer 2 Core interconnects end-sites, possibly through aggregation networks
- At Layer 2:
  Tree topology,
  STP enabled to guard against misconfigurations
- Two VLANs implemented for resiliency and use of multiple paths
- Route Servers: Simplified Control Plane connectivity



Dedicated or Shared Resources?

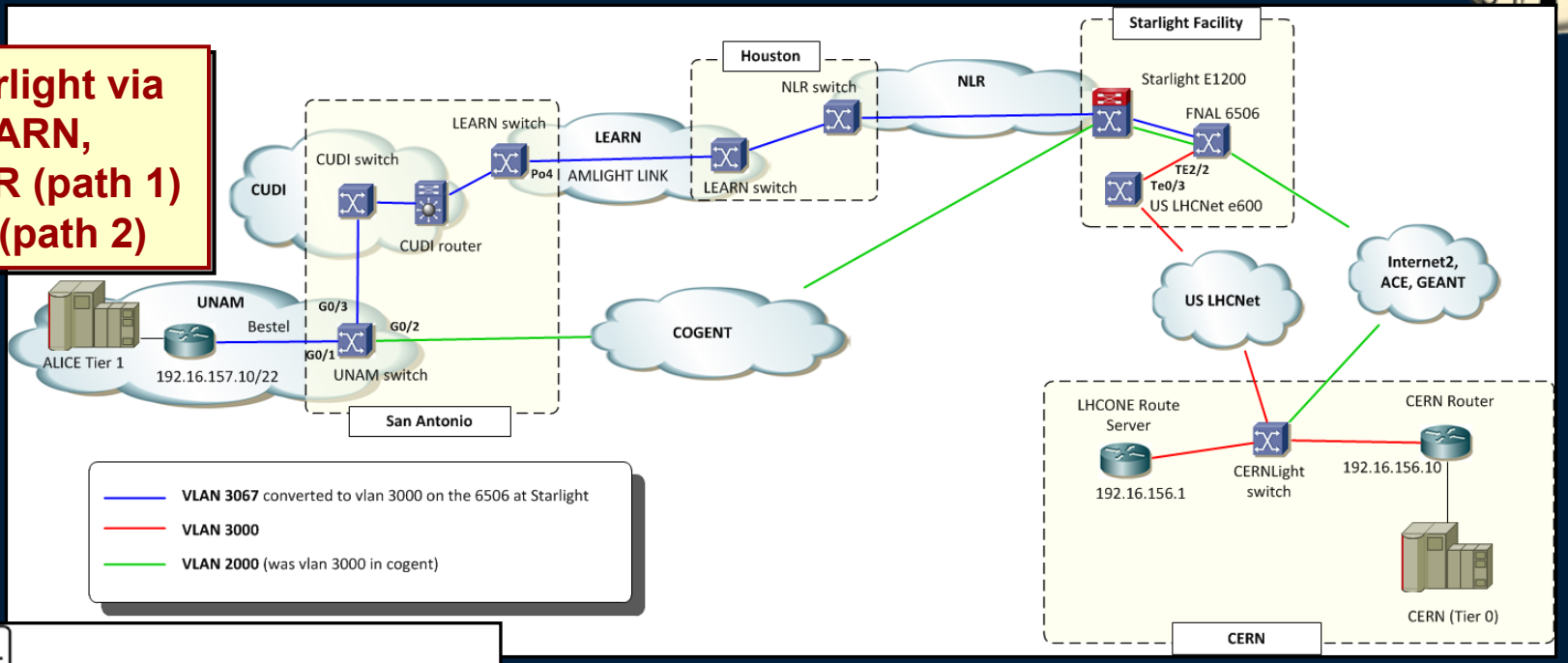
- LHCONE concept builds on traffic separation between LHC high impact flows, and non-LHC traffic
  - Avoid negative impact on other research traffic
  - Enable high-performance LHC data movement
- Core: Services to use resources allocated to LHCONE

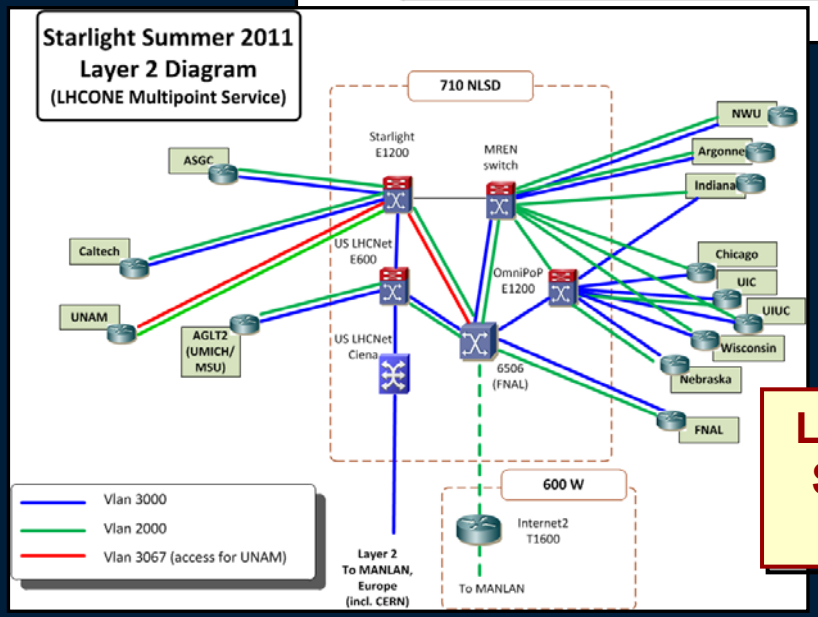


- Pilot implementation of multipoint service will use non-dedicated resources
  - but need to be careful about evaluation metrics

Some Implementation Examples

UNAM to Starlight via CUDI, LEARN, AMLIGHT, NLR (path 1) and Cogent (path 2)

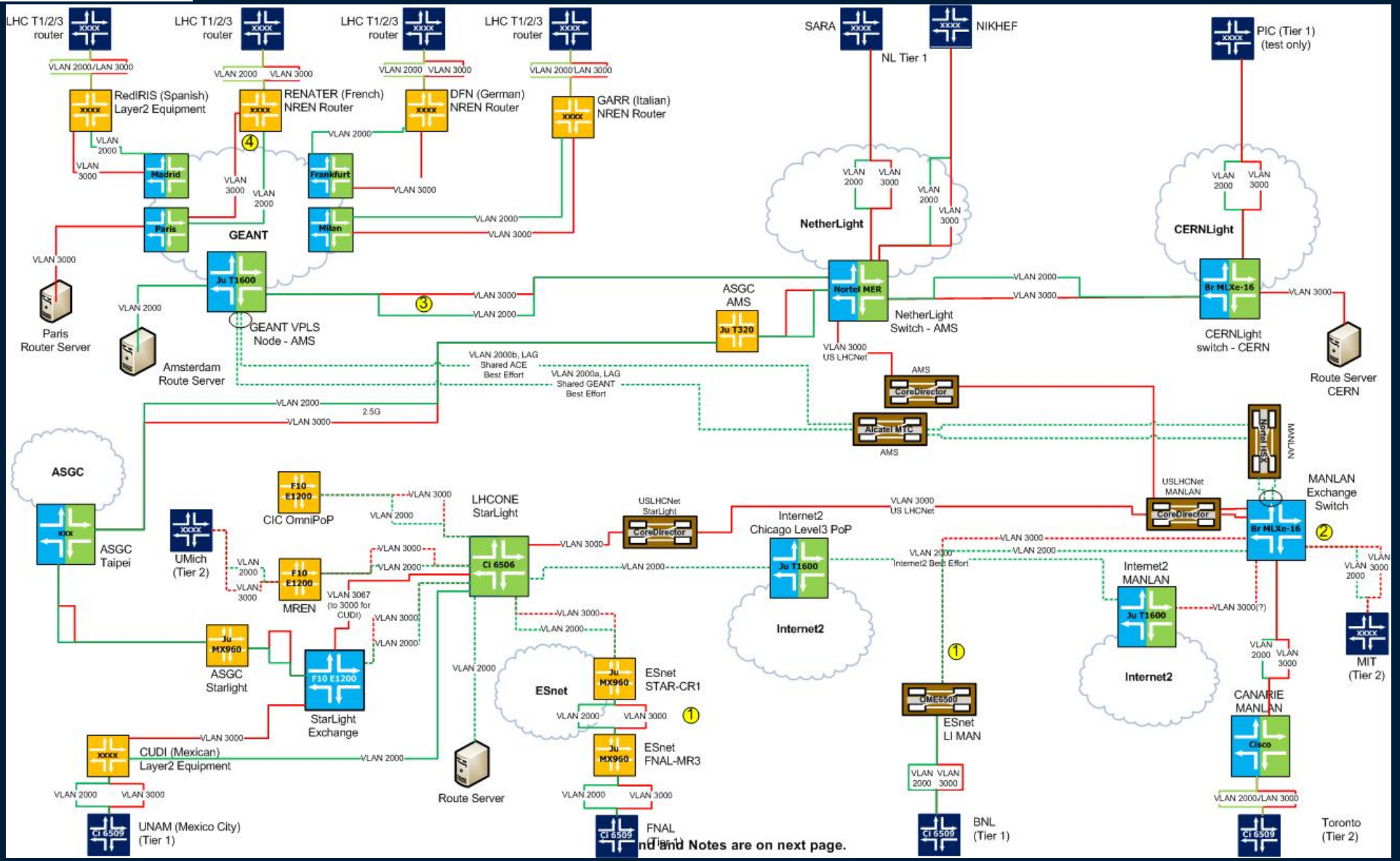




LHCONE implementation at Starlight: using dedicated central switch

Device-level Diagram (Current)

[Thanks to Bill Johnston/ESnet]



OUTLOOK BEYOND PILOT IMPLEMENTATION

Possible Future Directions

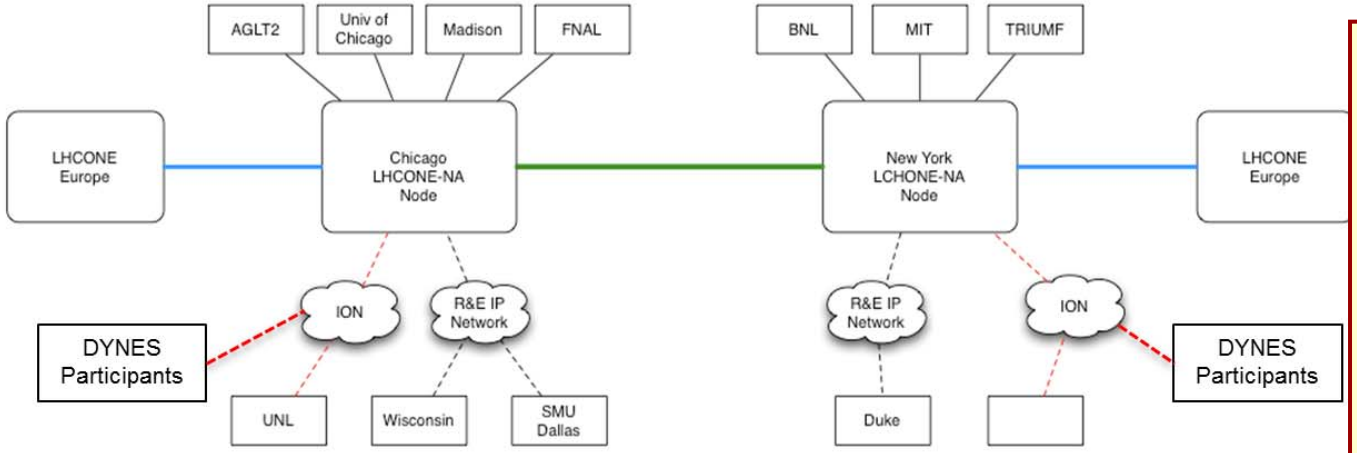
Addressing Scalability and Resiliency

- **Loop avoidance at Layer 2 poses constraints on the topology**
- **Efficient use of multiple paths, e.g. transatlantic**
  - **One VLAN per path does not scale well – it's a temporary solution**
- **Several approaches can be thought of**
  - **TRILL or SPB: Very interesting concept, providing multipath at Layer2**
    - Being developed for data centers; applicable to WAN?
  - **Proprietary multipath implementations (e.g. Cisco FastPath)?**
    - Would require same-vendor equipment at all exchange points
  - **OpenFlow with customized controller software?**
    - Support at future GOLE implementations?
    - Maturity in production environment?
  - **…**

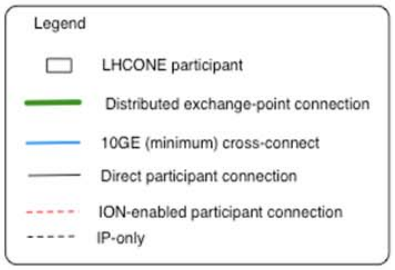
Point-to-point service

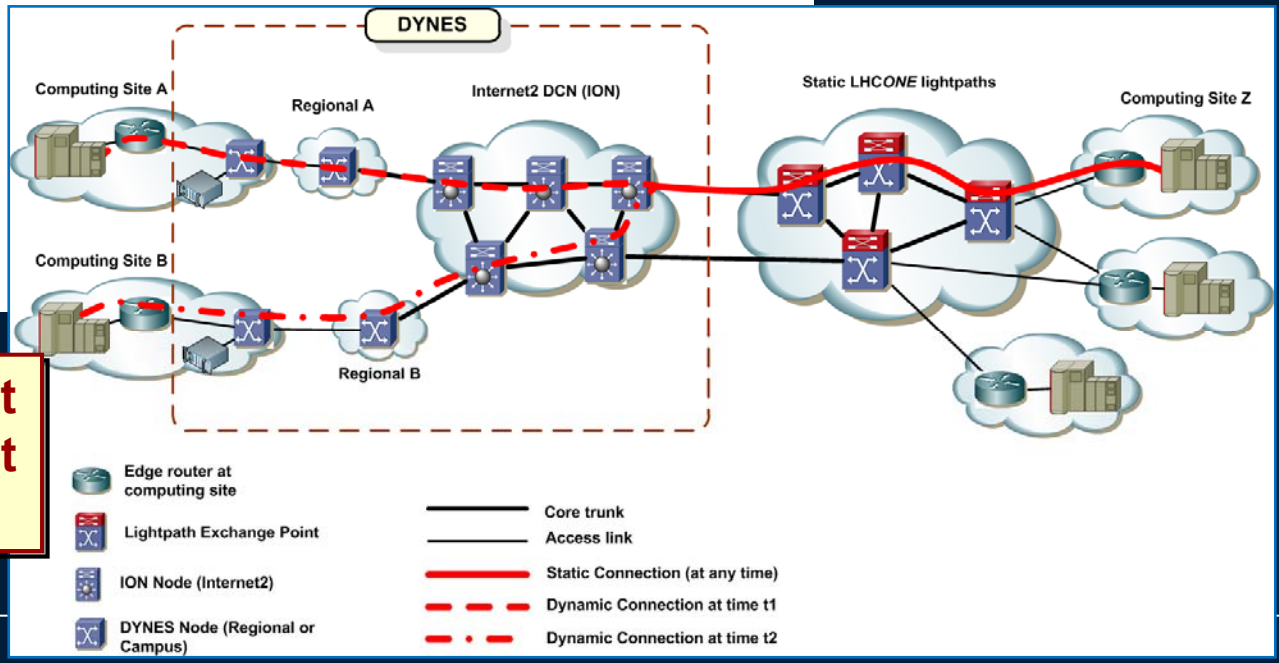
- Static or dynamic Lightpaths
- Interconnecting end-site's or aggregation networks' border routers
- Advanced users/sites: Layer 2 connections between end-systems
- Dynamic circuits still need buy-in from the user community
  - LHC computing is a global system, needs a global solution
  - Projects like DYNES start with national footprint
  - Projects (past and current) within the LHC community
    - LambdaStation, Terapaths, StorNet, ESCPS
- The LHC computing and data models changes the role of the network
  - "Making *it* work" was main priority in the past
    - "Network is not a problem"
  - Optimization of performance and resource utilization is addressed now
    - "Network is WLCGs most reliable resource"
- Standardization (OGF) is important for wide-scale adoption

Dynamic Lighpaths
DYNES + LHCONE



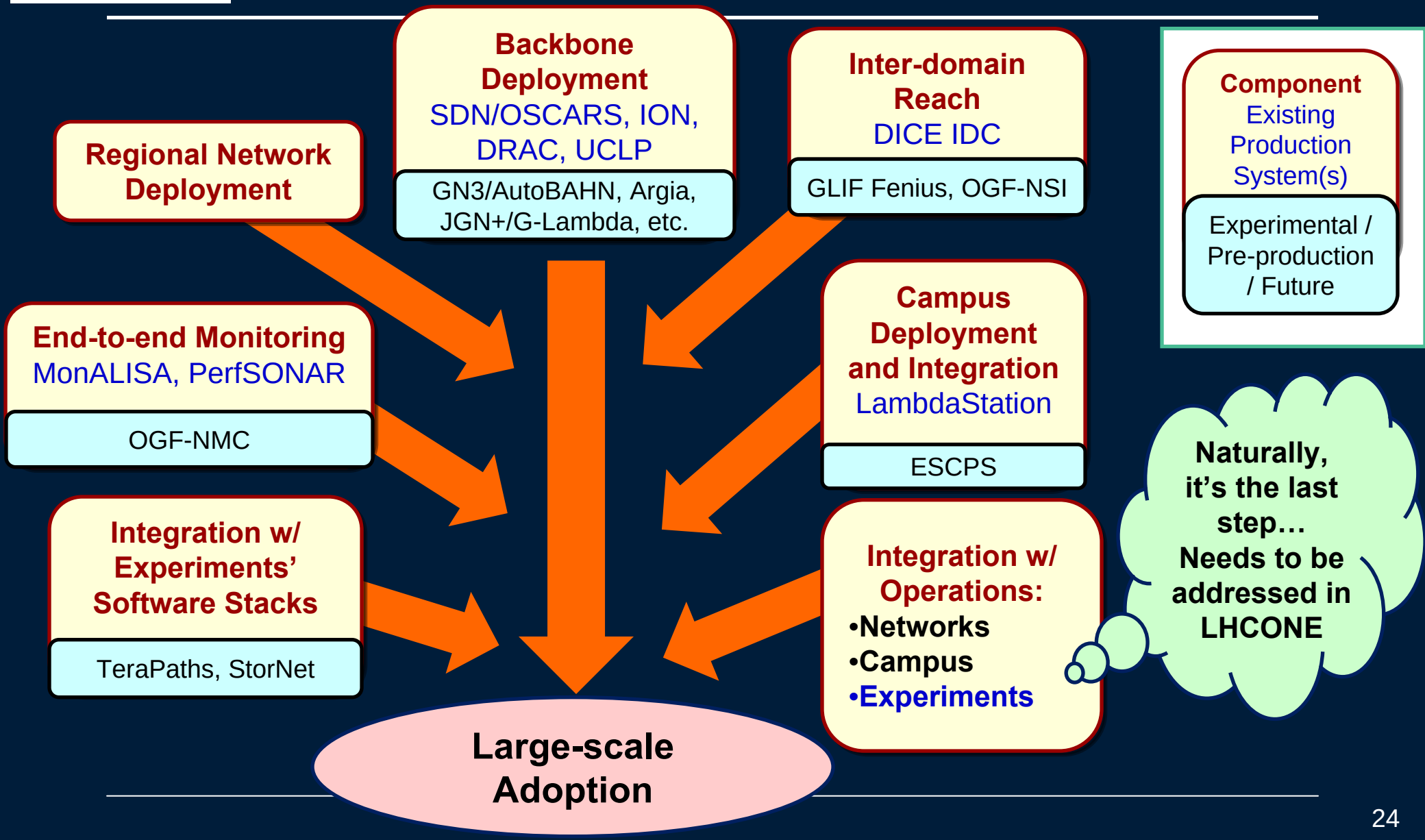
- DYNES Participants can dynamically connect to Exchange Points via Internet2 ION Service
- Dynamic Circuits through and beyond the exchange point?
- Static tail?





- Hybrid dynamic circuit and IP routed segment model?

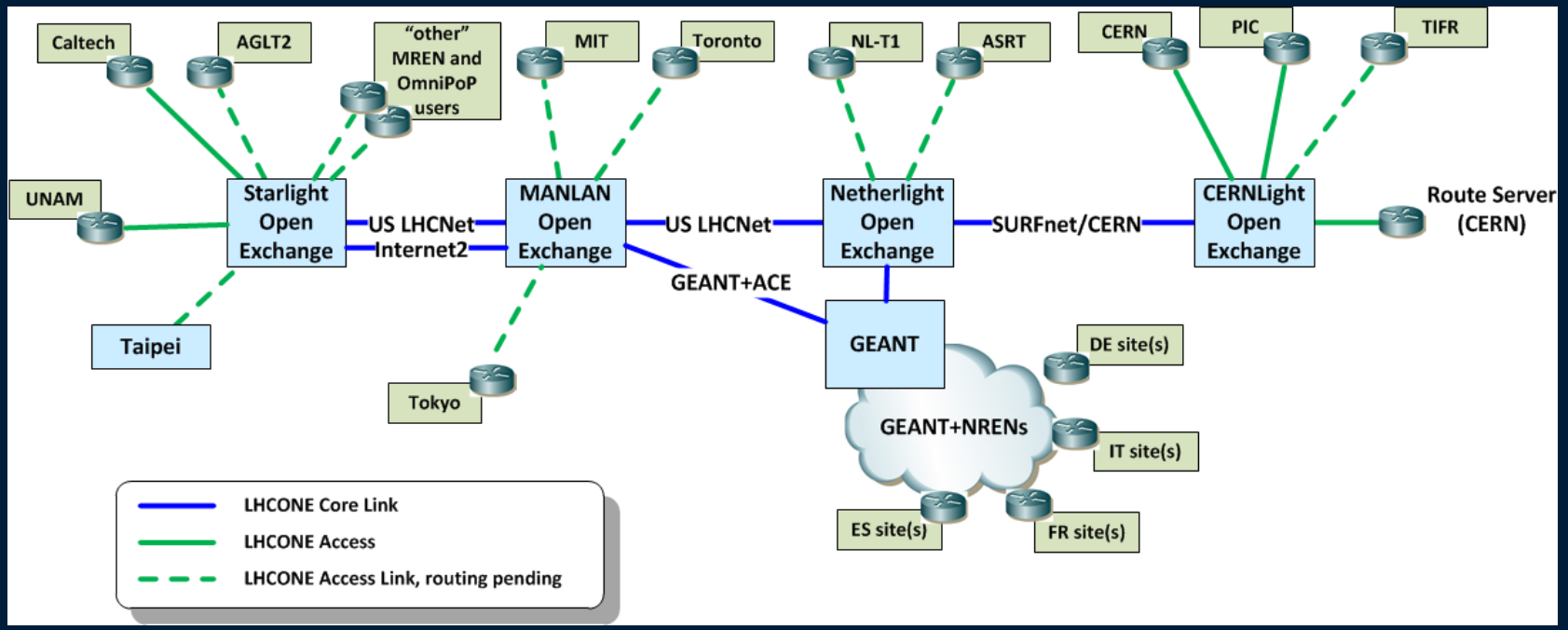
Towards Large Scale Dynamic Circuits in LHC Data Processing



LHCONE AND GLIF

LHCONE and GLIF

- LHCONE today is present at 4 GOLEs:
  - Starlight, MANLAN, Netherlight, CERNLight
- Will grow and potentially use other GLIF resources
  - GOLEs in South America, Asia, Africa
- Strong mutual interest, collaborative effort



Is GLIF important to LHCONE?
And if so… why?

- LHCONE is open (participation), neutral (policy) and diverse (technology and scope)
- LHCONE is designed as a global-scale overlay on existing open exchanges, with LHC specific policy
  - This policy does not imply changes to exchange point policies
- LHCONE constructs production services using GLIF and other resources
- GLIF is…
  - "…an international virtual organisation that promotes the paradigm of lambda networking"
    - Lightpaths as enabling technology for LHC high-throughput data movement
    - Enabling predictable end-to-end transfer latency
  - "…interested in developing application-empowered networks, in which the networks themselves are schedulable Grid resources"
    - LHCONE: Empowering the (LHC) user community

GLIF: Empowering the User

- LHCONE is a user-driven activity on a global scale
- The GLIF approach empowers the LHC computing community together with the R&E networking partners to construct services customized to its needs
  - No "one-size fits all" services
  - No need for centralized funding (Open ≠ Free) and governance
    - Often encountered situation in global science projects
- The administratively independent but coordinated resources in GLIF are key enablers for a flexible yet powerful solution on a global scale
  - GLIF is unique in this respect
- The collaboration between LHCONE and GLIF partners could be a model for current and future global science projects

Quo Vadis?

- **LHCONE sparked discussions within GLIF on policies, meaning of "open", governance, …**
- **GLIF is a perfect match for LHCONE, last but not least thanks to the open nature of collaboration**
  - **But that was easy, HEP is a "special" community**
- **LHCONE targets production grade services, which require support in several domains:**
  - **Technical**
    - Does the Facility provide the technologies for a global scale infrastructure?
    - Deployment of new technologies, provided by or compatible with GLIF resources?
  - **Operational**
    - Do resources provide interfaces for integrated operation on global scale?
  - **Policy**
    - Can the LHC community rely on the availability of resources (e.g. access to "any" GOLE? Do we need "any"?)

Summary (I)

- LHCONE is a robust and scalable solution for a global system serving LHC's Tier1, Tier2 and Tier3 sites' needs
  - Fits the new computing models
  - Based on a switched core with routed edge architecture
  - IP routing is implemented at the end-sites
- Core consists of sufficient number of strategically placed Open Exchange Points interconnected by properly sized trunks
  - Scaling rapidly with time as in requirements document
- Initial deployment to use predominantly static configuration (shared VLAN & Lightpaths),
  - later predominantly using dynamic resource allocation
- Pilot implementation interconnecting an initial set of sites has started
  - Organic growth

Summary (II)

- LHCONE is an overlay on existing open exchanges
  - With LHC specific policy
- Some resources are/will be dedicated
- It will grow organically from the pilot implementation
- Starting with the multipoint service addressing connectivity and traffic separation, building out to use of dedicated lightpaths
- Building on GLIF resources, LHCONE is an open, neutral and diverse solution for the LHC networking needs
  - On global scale
- LHCONE could be a model for other large-scale scientific and research collaborations

THANK YOU!

http://lhcone.net

Artur.Barczyk@cern.ch