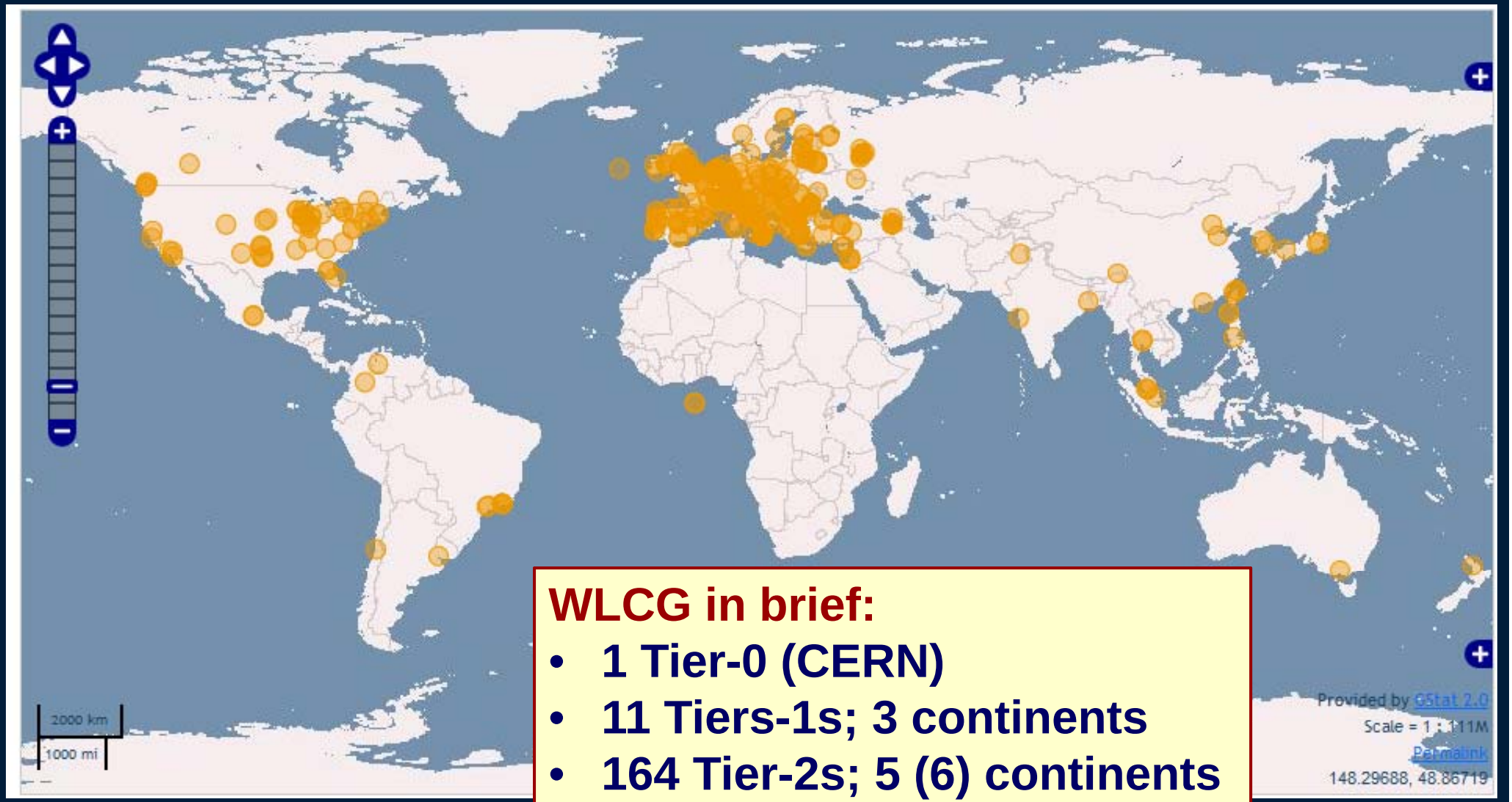
LHC Open Network Environment
LHC*ONE*

Artur Barczyk
California Institute of Technology
GLIF Technical Working Group Meeting
Hong Kong, February 25th, 2011

LHC AND WLCG
FIRST YEAR OF LHC RUNNING

From the network perspective

LHC Computing Infrastructure



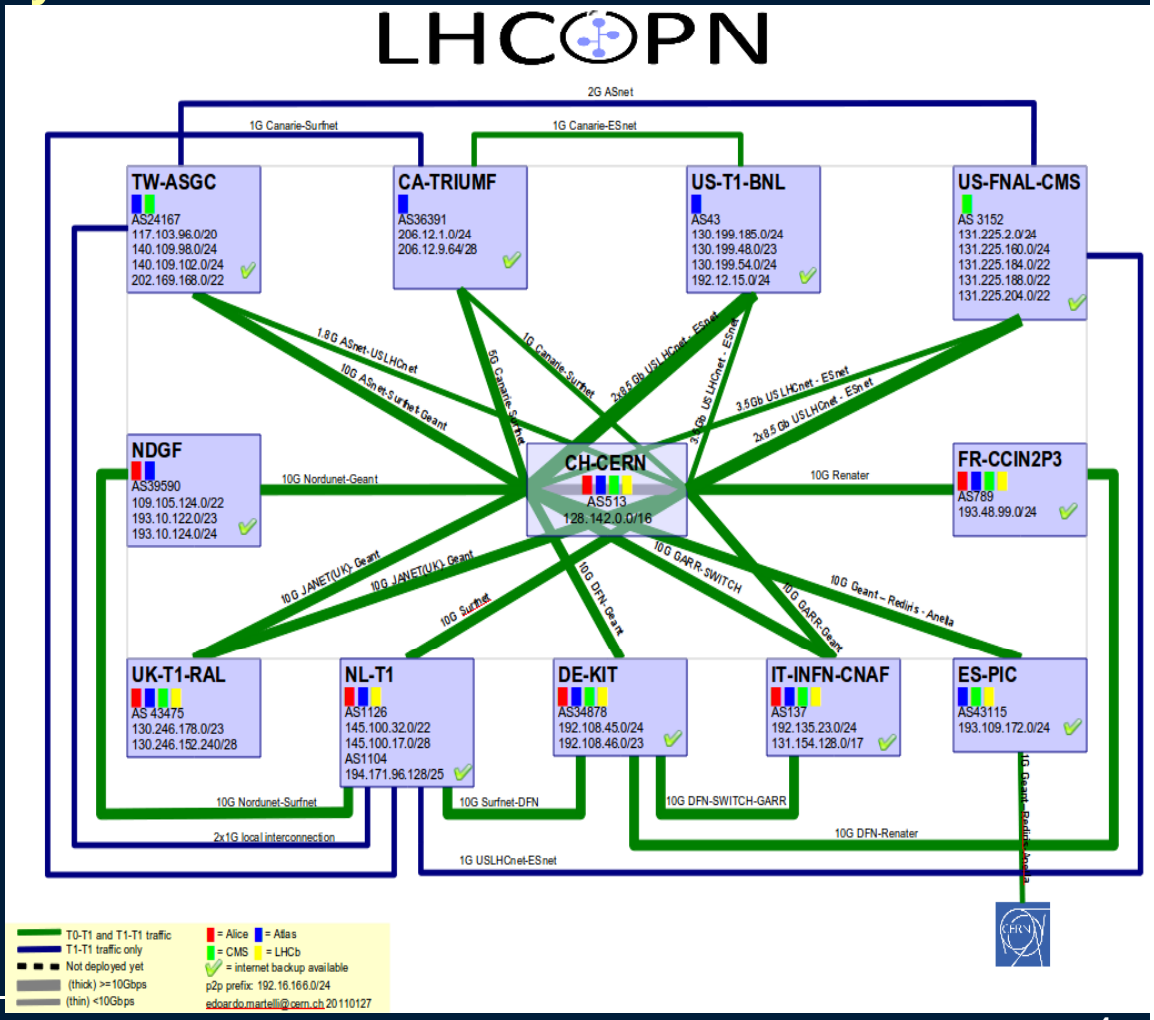
WLCG in brief:

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

Plus O(300) Tier-3s worldwide

The LHCOPN

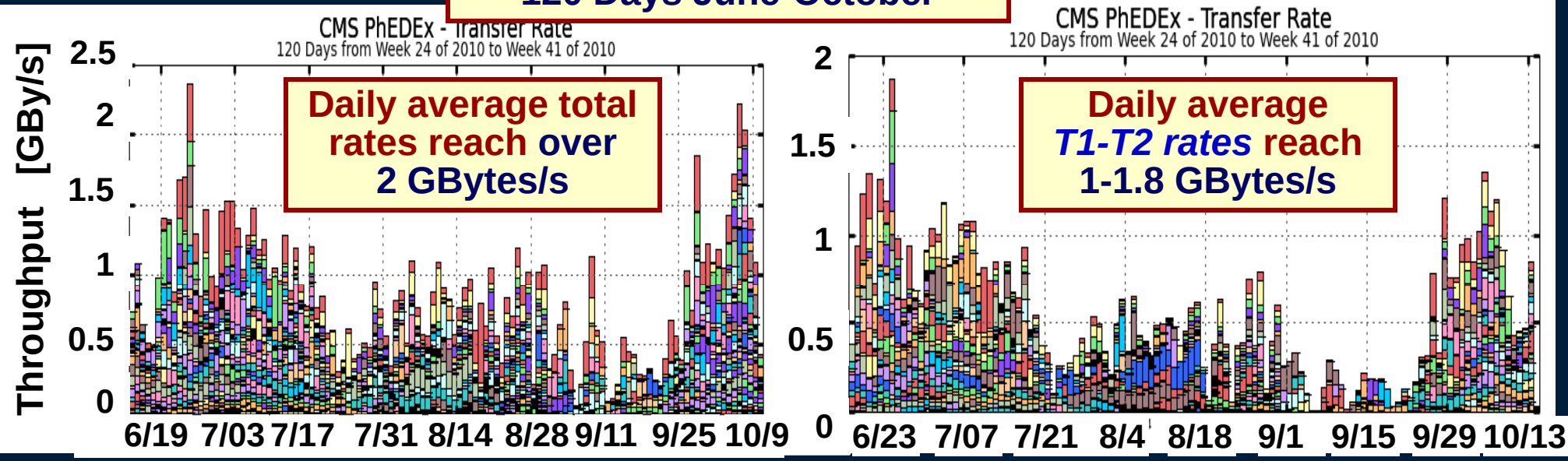
- Dedicated network resources for Tier0 and Tier1 data movement
- 130 Gbps total Tier0-Tier1 capacity
- Simple architecture
  - Point-to-point Layer 2 circuits
  - Flexible and scalable topology
- Grew organically
  - From star to partial mesh
  - Open to technology choices
    - have to satisfy requirements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

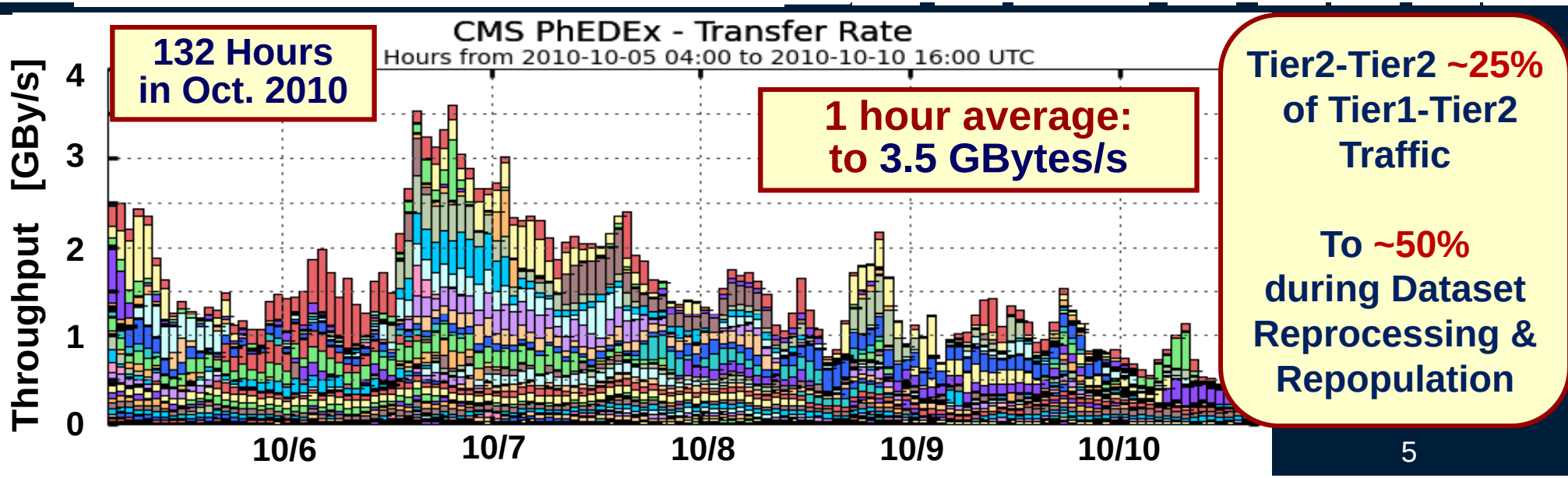


CMS Data Movements

(All Sites and Tier1-Tier2)

120 Days June-October



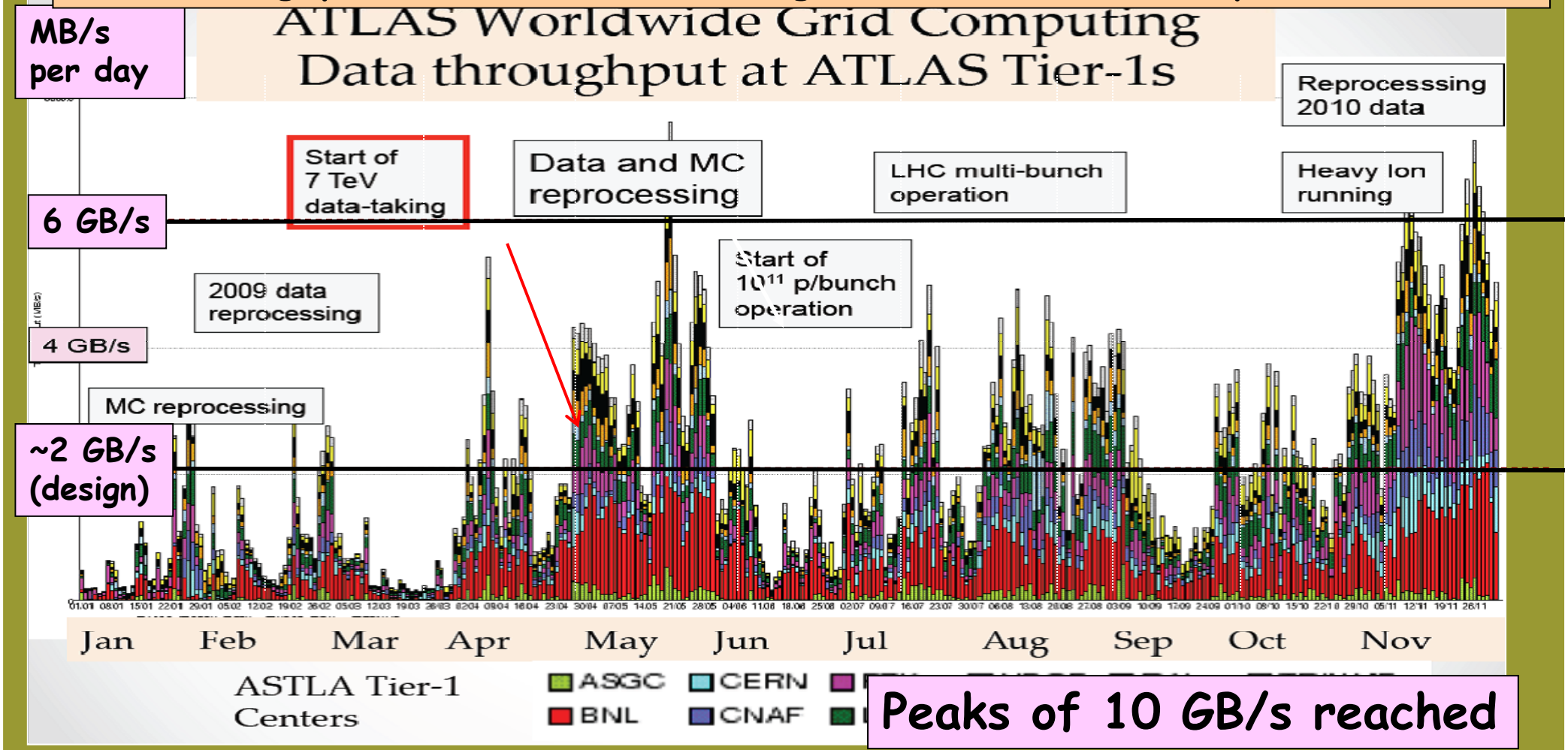


Tier2-Tier2 ~25% of Tier1-Tier2 Traffic

To ~50% during Dataset Reprocessing & Repopulation

Worldwide data distribution and analysis (F.Gianotti)

Total throughput of ATLAS data through the Grid: 1st January → November.



Peaks of 10 GB/s reached

Grid-based analysis in Summer 2010: >1000 different users; >15M analysis jobs

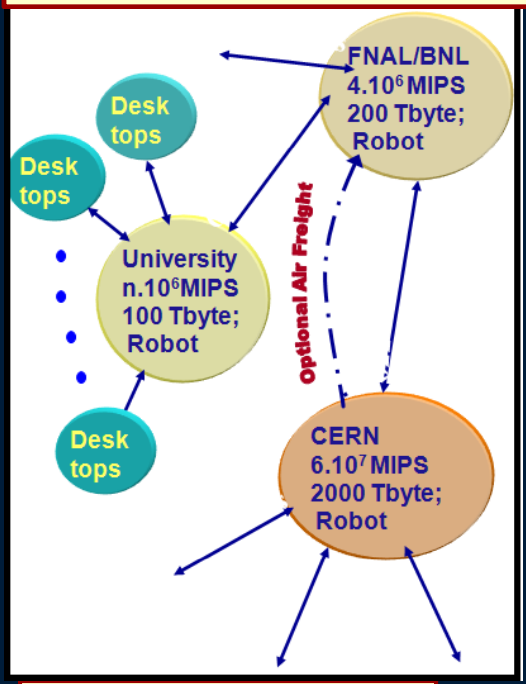
The excellent Grid performance has been crucial for fast release of physics results. E.g.: ICHEP: the full data sample taken until Monday was shown at the conference Friday

LHC EXPERIMENTS' DATA MODELS

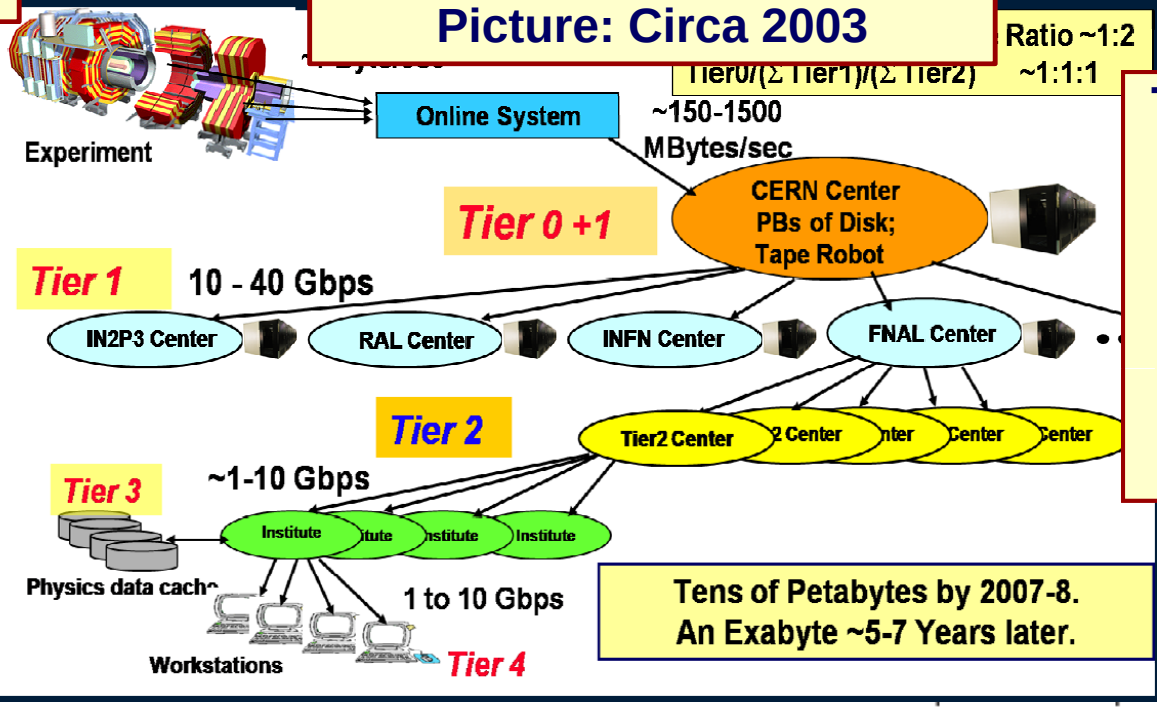
Past, present and future

Past Data Models

Circa 1996



The Evolving MONARC Picture: Circa 2003

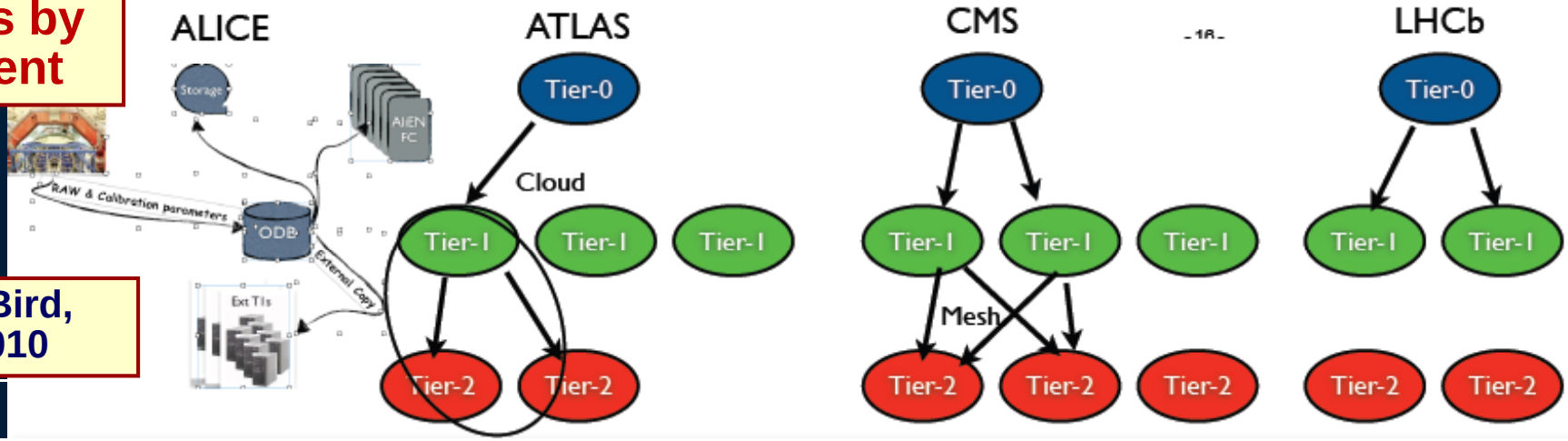


The models are based on the MONARC model

Now 10+ years old

Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

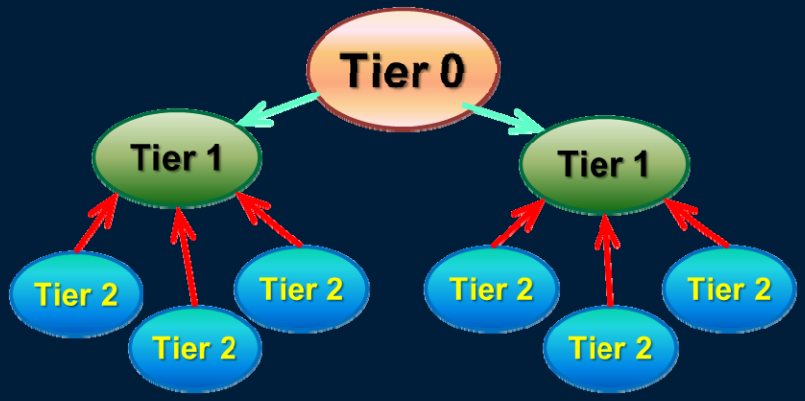
Variations by experiment

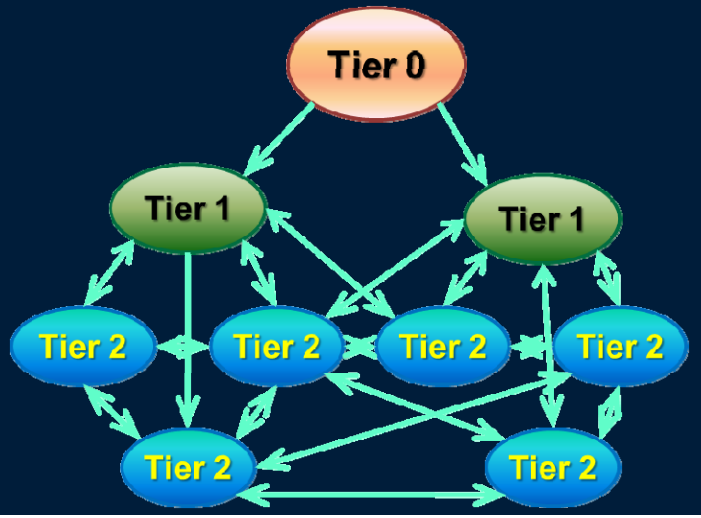


From Ian Bird, ICHEP 2010

The Future is Now

- **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)
- Network bandwidth (rather than disk) will need to scale more with users and data volumes
- Data placement will be driven by demand for analysis and not pre-placement

Ian Bird, CHEP conference, Oct 2010

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**

Design Inputs

- By the scale, geographical distribution and diversity of the sites as well as funding, only a federated solution is feasible
- The current LHC OPN is not modified
  - OPN will become part of a larger whole
  - Some purely Tier2/Tier3 operations
- Architecture has to be Open and Scalable
  - Scalability in bandwidth, extent and scope
- Resiliency in the core, allow resilient connections at the edge
- Bandwidth guarantees → determinism
  - Reward effective use
  - End-to-end systems approach
- Core: Layer 2 and below
  - Advantage in performance, costs, power consumption

LHC*ONE* Design Considerations

- **LHCONE complements the LHCOPN by addressing a different set of data flows: high-volume, secure data transport between T1/2/3s**
- **LHCONE uses an open, resilient architecture that works on a global scale**
- **LHCONE is designed for agility and expandability**
- **LHCONE separates LHC-related large flows from the general purpose routed infrastructures of R&E networks**
- **LHCONE incorporates all viable national, regional and intercontinental ways of interconnecting Tier1s, 2s and 3s**
- **LHCONE provides connectivity directly to Tier1s, 2s, and 3s, and to various aggregation networks that provide connections to the Tier1/2/3s**
- **LHCONE allows for coordinating and optimizing transoceanic data flows, ensuring optimal use of transoceanic links using multiple providers by the LHC community**

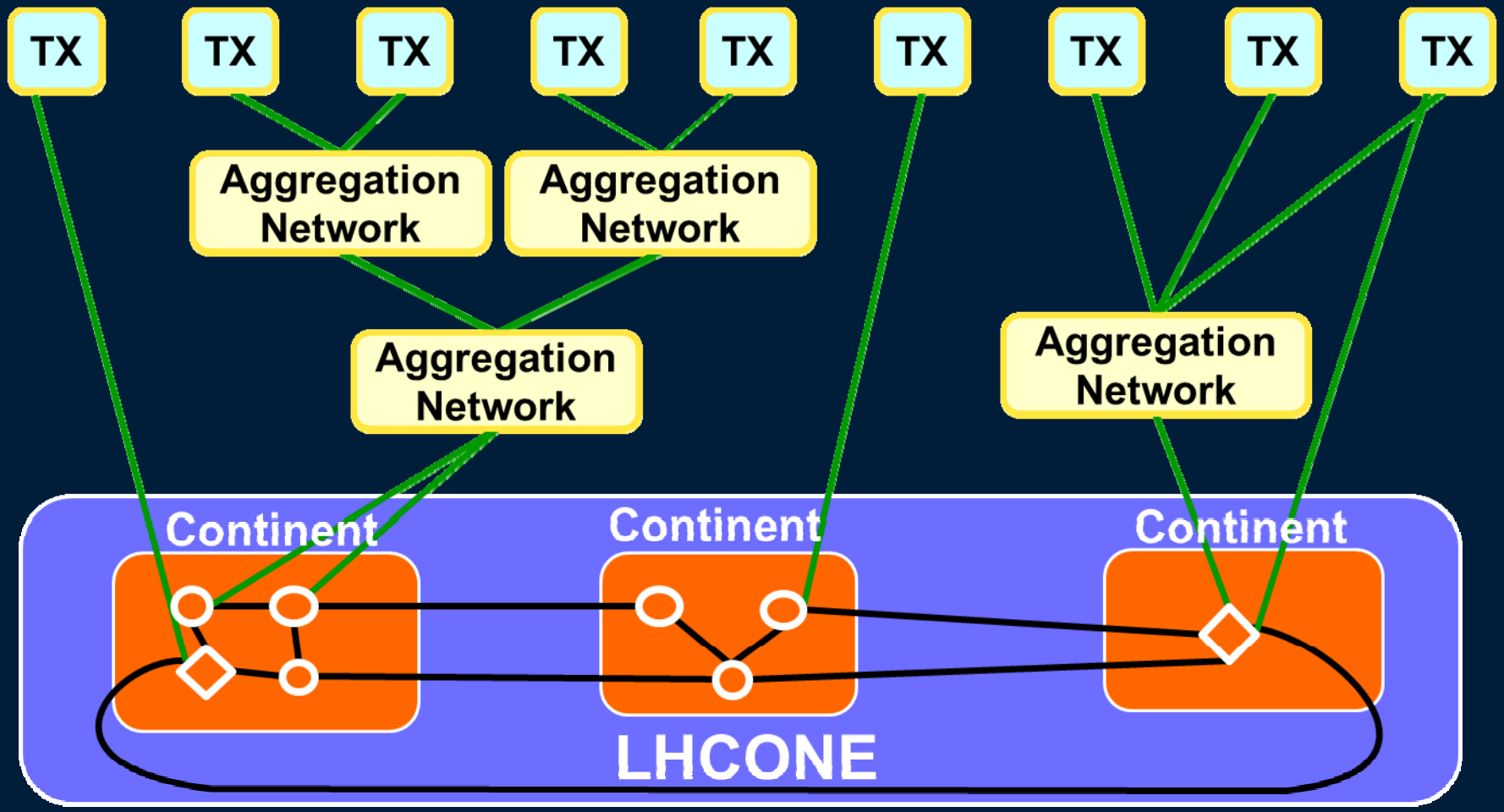
LHC*ONE* Architecture

- Builds on the Hybrid network infrastructures and Open Exchanges
  - As provided today by the major R&E networks on all continents
  - To build a global unified service platform for the LHC community
- Make best use of the technologies and best current practices and facilities
  - As provided today in national, regional and international R&E networks
- LHC*ONE*'s architecture incorporates the following building blocks
  - Single node Exchange Points
  - Continental / regional Distributed Exchange Points
  - Interconnect Circuits between exchange points
- Continental and Regional Exchange Points are likely to be built as distributed infrastructures with access points located around the region, in ways that facilitate access by the LHC community
  - Likely to be connected by allocated bandwidth on various (possibly shared) links to form LHCONE

LHC*ONE* Access Methods

- Choosing the access method to LHCONE, among the viable alternatives, is up to the end-site (a Tier1, 2 or 3), in cooperation with site and/or regional network
- Alternatives may include
  - Dynamic circuits,
  - Dynamic circuits with guaranteed bandwidth
  - Fixed lightpath(s)
  - Connectivity at Layer 3, where appropriate and compatible with the general purpose traffic
- We envisage that many of the Tier-1/2/3s may connect to LHCONE through aggregation networks

High-level Architecture, Example



Single node Exchange Point Distributed Exchange Point

LHC*ONE* 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

LHC*ONE* Policy Summary

Details at http://lhcone.net

- **It is expected that 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, in a form to be determined.
- 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

Details at http://lhcone.net

LHC*ONE* Implementation Guidance

- **Access Switches**
  - Devices that provide the LHCONE Layer2 Ethernet connectivity with 1G and 10G Ethernet ports
  - 40G, 100G Ethernet ports are expected to be available in the future
  - Access switches are expected to be located at the Exchange Points
- **Access Links**
  - Ethernet-framed point-to-point links connecting a connector's device to one of the LHCONE Access Switches
  - Links are purchased and operated by the connectors and are not under the responsibility of LHCONE
  - Any connector may optionally connect to two (or even more) different Access Switches, for resiliency reasons

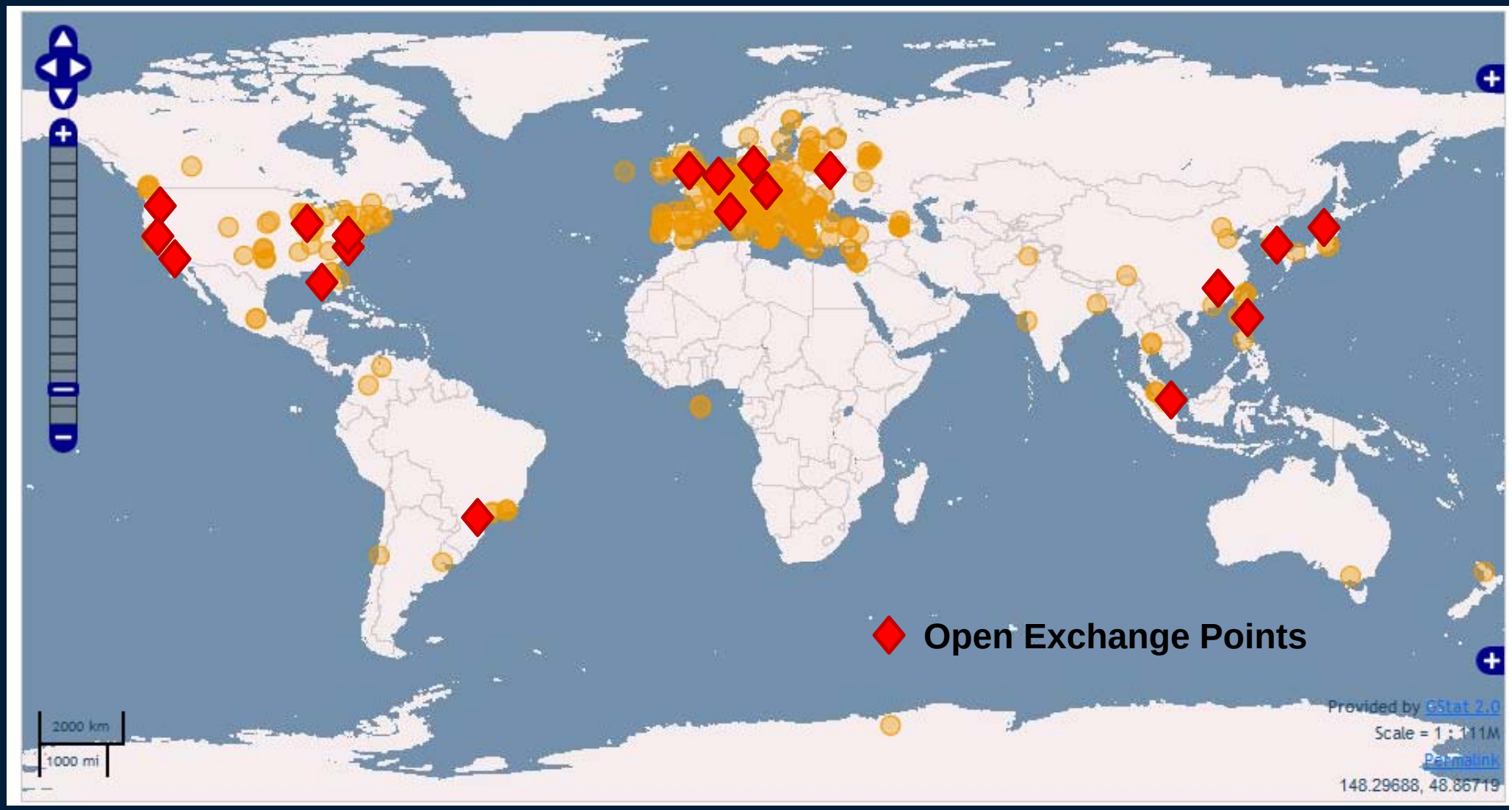
Next Steps

- **Prototype implementation (Seed)**
  - CMS & Atlas to prepare a use case with ~10 “Leadership” Tier2s (Week 8)
  - Identify BW targets, metrics for success
  - Small engineering group to work out prototype design (Week 12)
  - Implementation to start after Week 12
- **Follow-up roadmap**
  - In parallel with prototype implementation
    - Refine governance model
    - Refine service and policy definitions
    - Refine architecture
  - Gather information (“RFI/RFP”)
    - implementation details, time scales, cost estimates
- **LHCONE will grow “organically”, as needs arise**
  - and where funding is available, or is made available

Organic Growth

- The LHCONE Prototype will be open to participation from the start
  - Allow to include any site from day one
- Reflect the immediate need of the LHC community
  - Experiments are in the process of moving to the new computing models (Process started in Summer 2010)
  - LHC to restart data taking in March 2011, will continue throughout 2012
- Support LHC computing operations at global scale from day one
- Support immediate needs of important sites like IHEP (Beijing), UNAM (Mexico) and others
- Open Exchange Points in all world regions to play an important role; e.g. HKOEP, T-LEX, TWLight, KRLight

WLCG and Today's Open Exchanges

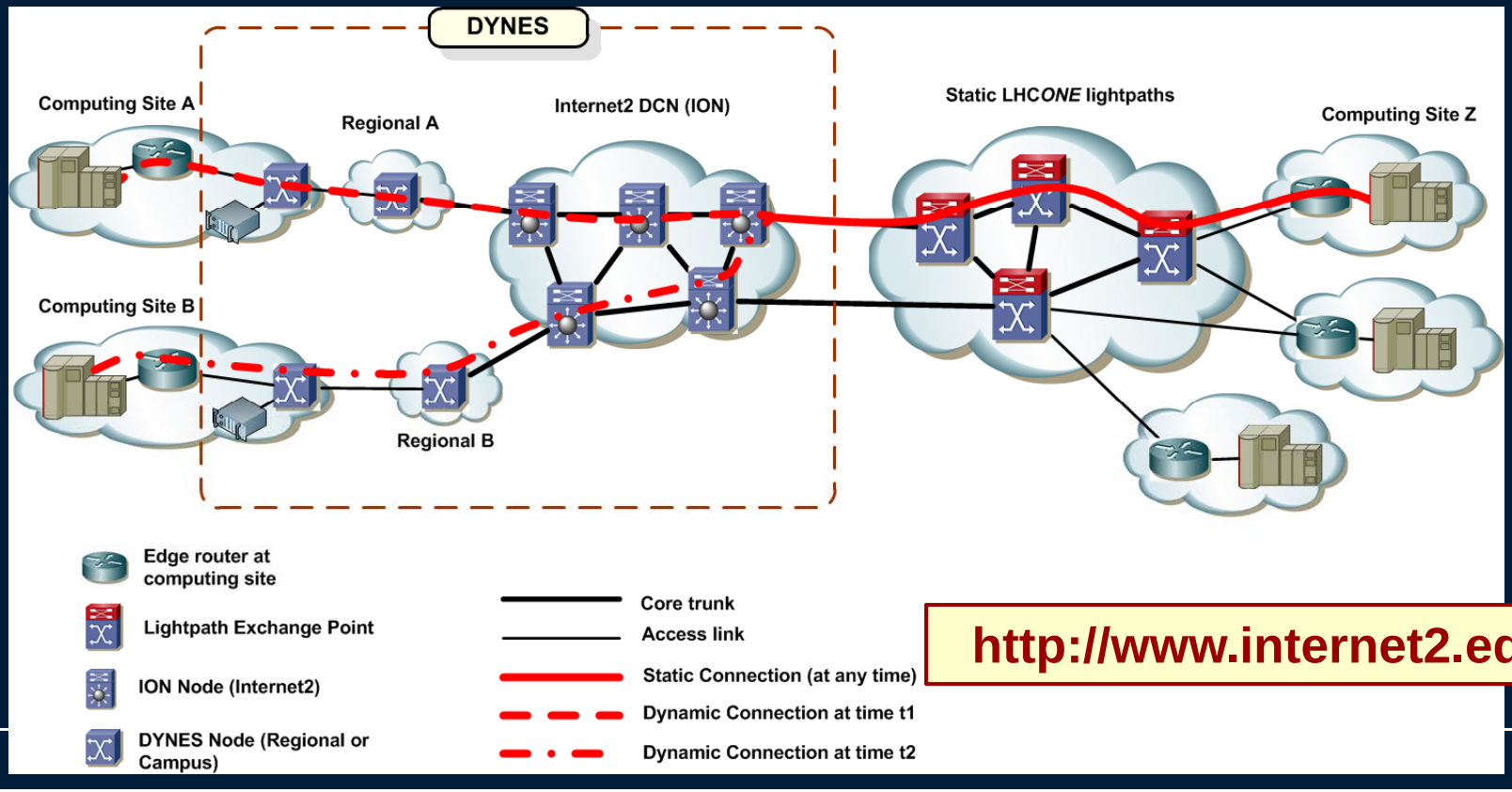


How do End-Sites Connect? A Simple Example

- **A Tier2 in Asia needs 1 Gbps connectivity (each) to the ASGC Tier1, 2 sites in Europe and 2 in the US**
  - 5 x 1G intercontinental circuits is cost-prohibitive
- **The Tier2 could however afford a 1-2 Gbps (e.g. EoMPLS) circuit to next Exchange Point (e.g. HKOEP, KRLight, TaiwanLight, T-LEX)**
  - Through aggregation network or a direct connection
- **The Exchange Point connects to other Exchange Points, e.g. Starlight, NetherLight and has a connection to e.g. ASGC Tier1**
- **Static bandwidth allocation (first stage):**
  - **Tier2 has a 1Gbps link in a shared VLAN, peers only with selected sites**
  - **Bandwidth is allocated by the exchange points to fit the needs**
- **Dynamic allocation (early adopter + later stage):**
  - **The end-site has a 1Gbps link, with configurable remote end-points and bandwidth allocation**

Early Dynamic Circuits: LHCONE + DYNES

- The Internet2 ION service currently has end-points at two GOLEs in the US: MANLAN & StarLight
- A static Lightpath from any end-site to one of these GOLE sites can be extended through ION to any of the DYNES sites (LHC Tier2 or Tier3)



http://www.internet2.edu/dynes

Summary

- **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**
- **Prototype/Seed implementation interconnecting an initial set of sites to start soon**
  - **Organic growth; Key Role of NRENs (also in Asia!)**

THANK YOU!

http://lhcone.net

Artur.Barczyk@cern.ch