Dynamic Networks, Intelligent Systems

Artur Barczyk

InGrid Workshop

Mumbai, April 2nd, 2012

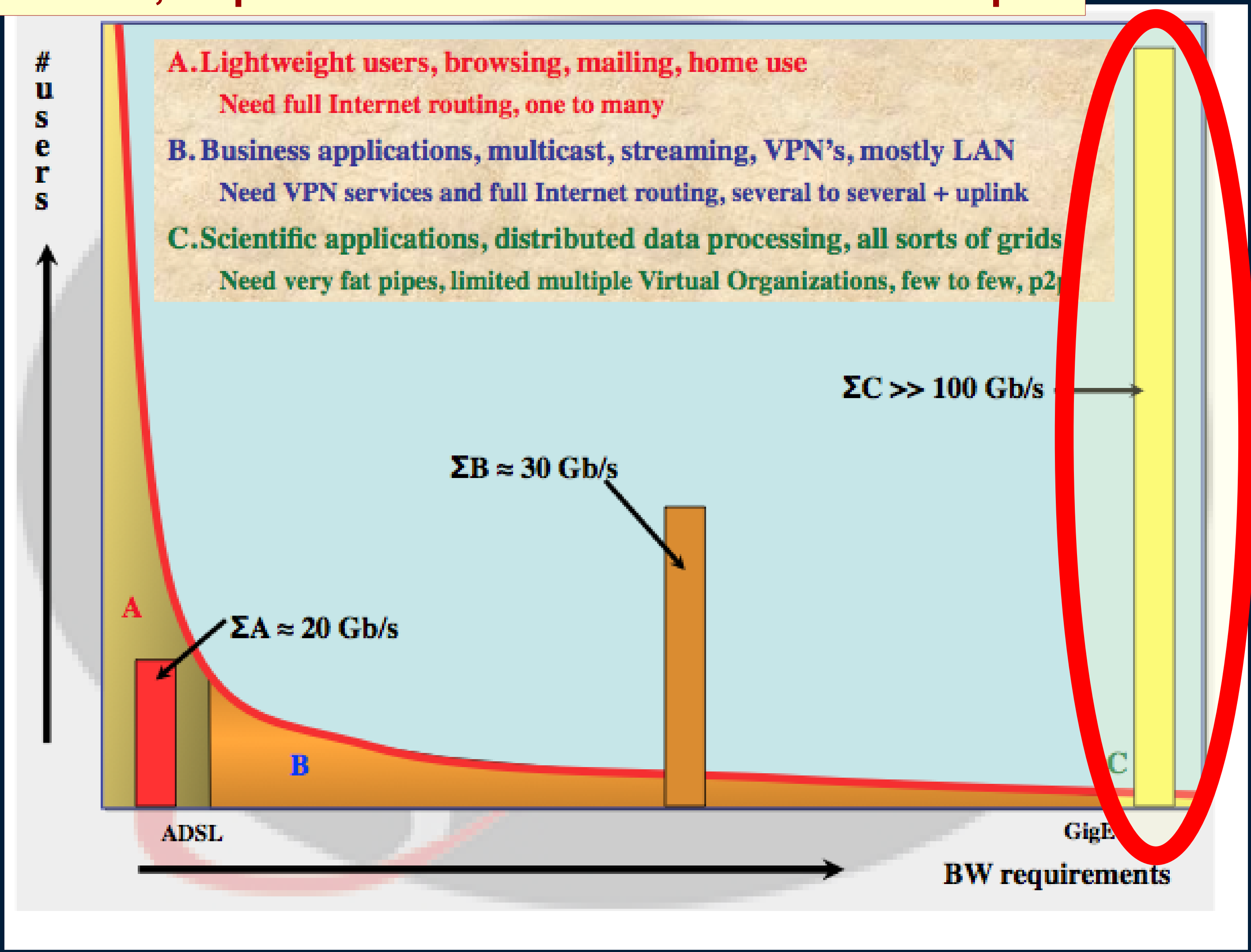
Outline

- **Introduction and Motivation**
- **Network Virtualization**
- **Software-Defined Networking**
  - OpenFlow
- **Dynamic bandwidth allocation (Dynamic Circuits)**
  - Example: DYNES project
- **OGF standards: OGF NSI, NML, NMC**
- **Pervasive Monitoring**
- **LHC Open Networking Environment (LHCONE)**
  - Global virtualized infrastructure for the LHC

Introduction

Demanding Users - Characterization of User Space

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



This is where LHC users are

Now also: Genomics, Earth Sciences, Radio-astronomy, …

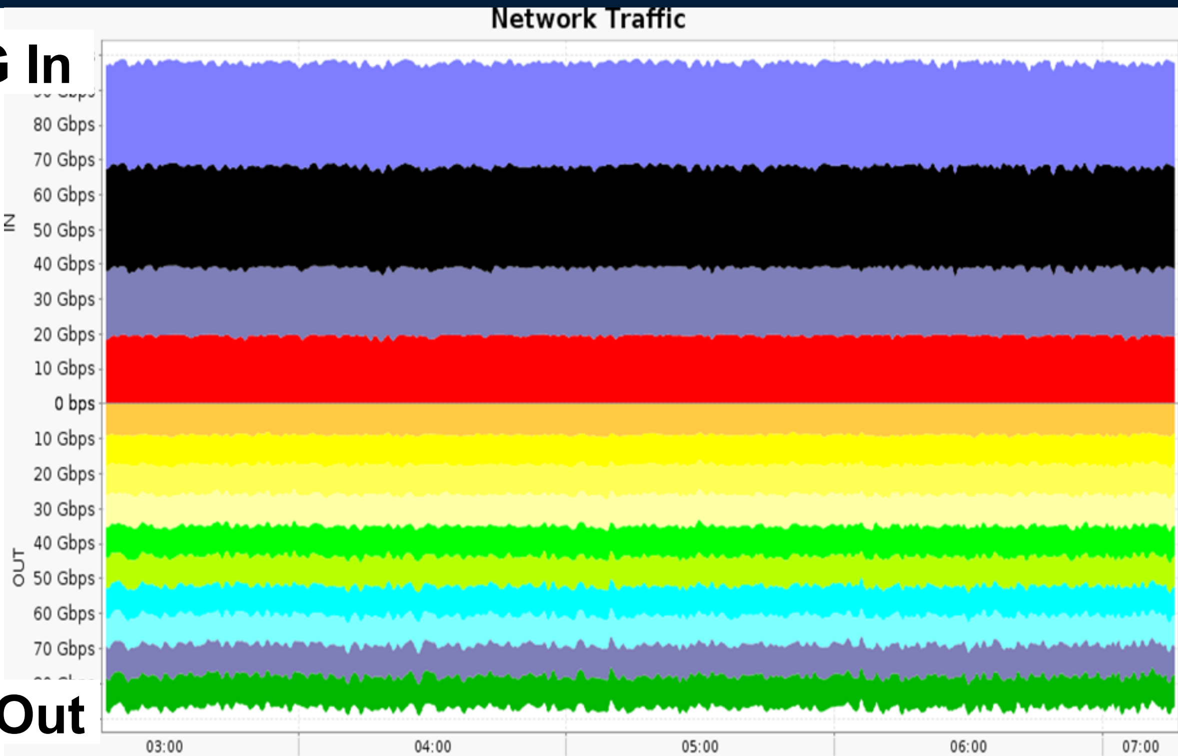
NB: Users = VOs or computing sites

A Few Well Connected Sites…
(Demo at SuperComputing 2011)

University of Victoria

Caltech and Univ of Victoria demo at SC11:
Data exchange at 100Gbps between two Tier2-sized sites (Seattle and Victoria)

100G In



100G Out

2 Petabytes/Day Stable Flow
Using 4 PCIe Gen3 and 2 Gen2 Servers in Caltech SC Booth

FDT: Fast Data Transfer

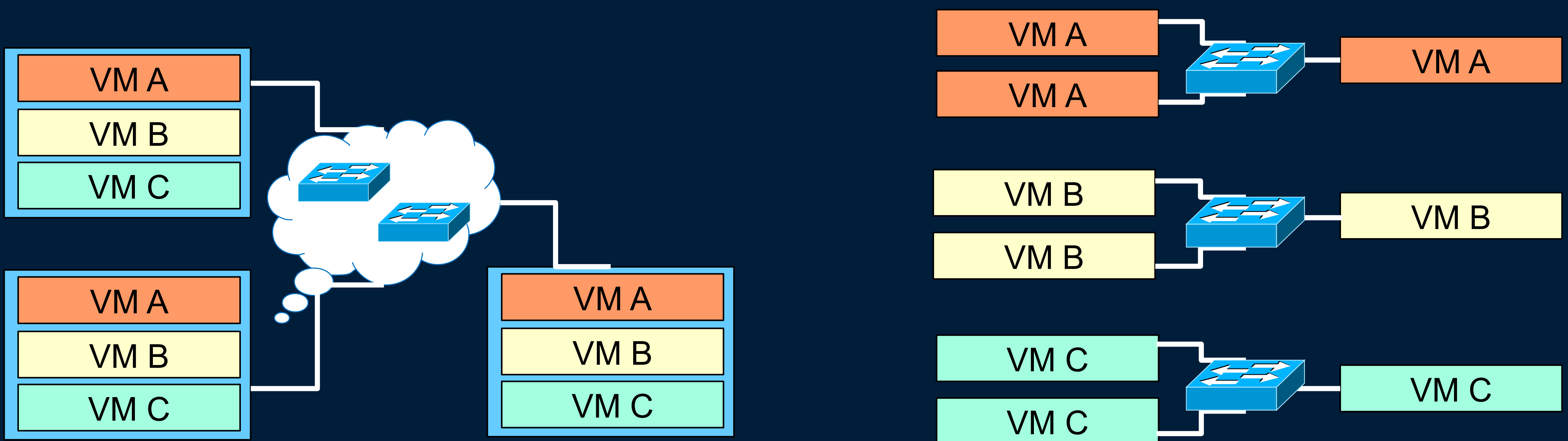
- **Easy to use open source Java app.**
- **Uses asynch. Multi-threaded system to achieve smooth, linear data flow:**
  - Streams a dataset (list of files) continuously through open TCP socket
  - Sends buffers at rate matched to the capability of each end to end path
  - Independent R/W threads per drive

Network Virtualization (in the WAN)

- **Cannot build a separate network for each user community, but need to accommodate the varying needs of each of them**
- **Overprovisioning obviously not a long-term solution**
- **Virtualized networks provide logical separation of traffic from different sources or organizations**
- **If coupled with the right technology, can provide bandwidth guarantees**
- **Different technologies, at various network layers**
  - Layer 1, Layer 2, Layer 3 VPNs (Virtual Private Networks)
  - Virtual Routing and Forwarding (VRF), MPLS
  - Dynamic Lightpaths (aka dynamic circuits, BoD, …); developed by R&E networking community:
    - OSCARS, DRAC, UCLP, AutoBAHN, …
  - Layer 3 virtualization: Mantychore, Federica (for network research)
  - Layer 2: VLAN, Carrier Ethernet E-LAN
- **Target: deterministic transfer performance**
  - Throughput, jitter

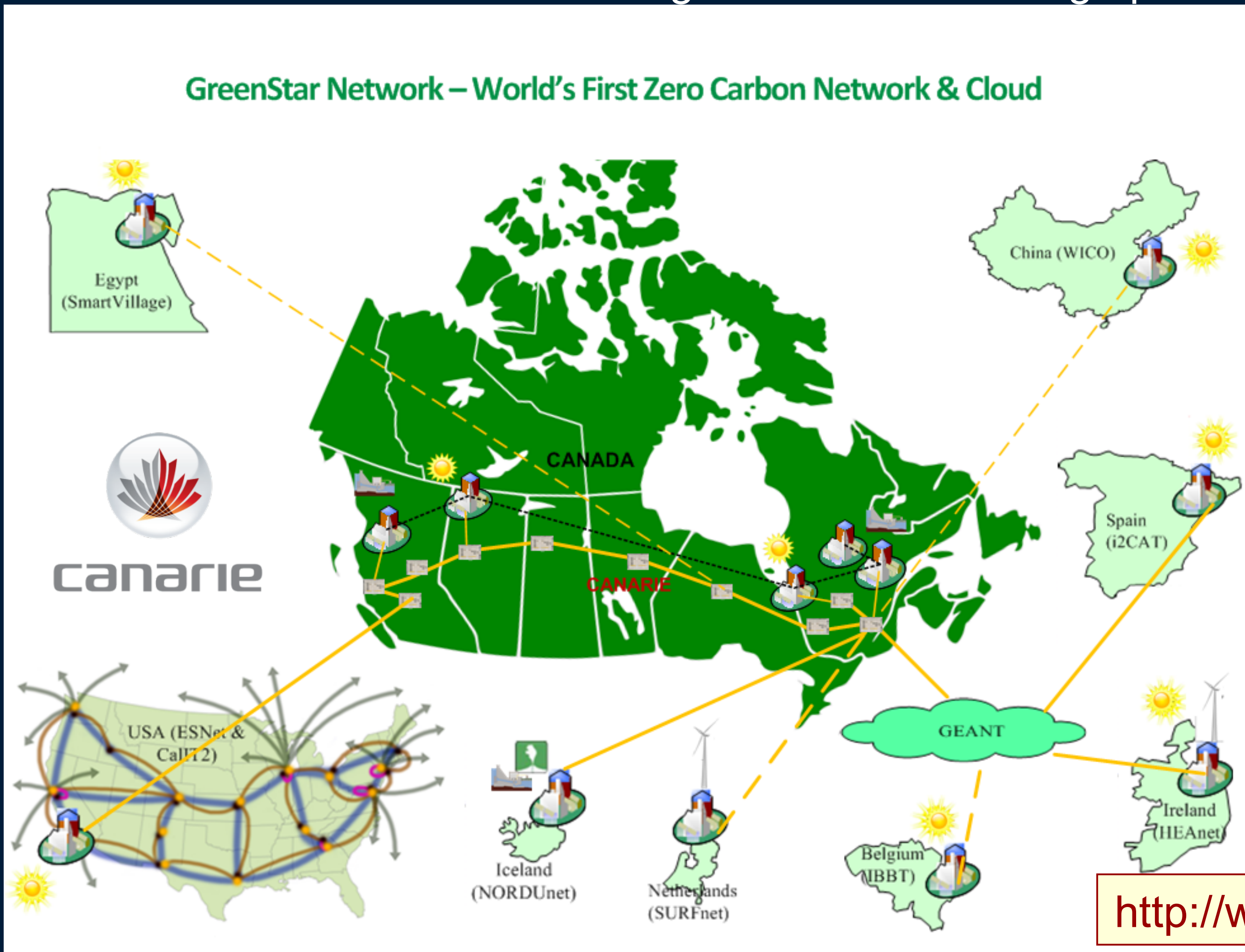
Network Virtualization in the Data Center

- **Server Virtualization is a well-known concept by now**
  - Separate (virtual) server functionality from hardware
- **Poses new challenges on the network:**
  - Multiple applications/services share same server hardware, mixed flows
  - Multiple tenants require traffic separation and service quality guarantees
  - VM Mobility vs addressing vs
- **Virtual networks: separate (virtual) network services from hardware instances**

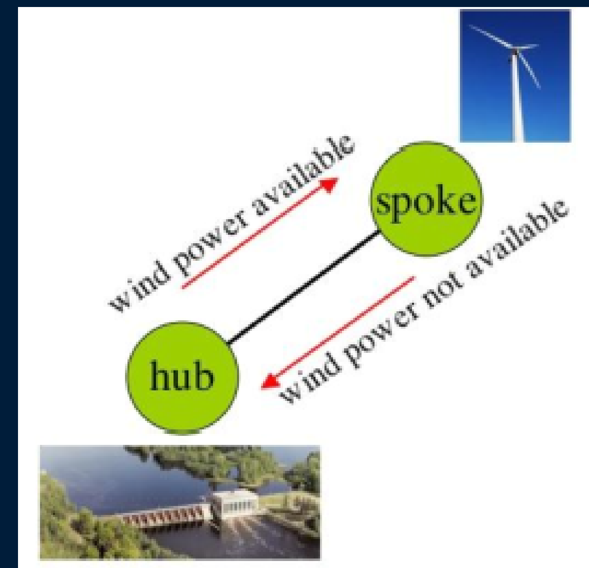


WAN Virtualization Example: GreenStar Network

- **Developing green ICT based on resource virtualization and IaaS concepts**
  - Move Virtual Machine images and data over Lightpaths



"Follow the Wind"
"Follow the Sun"



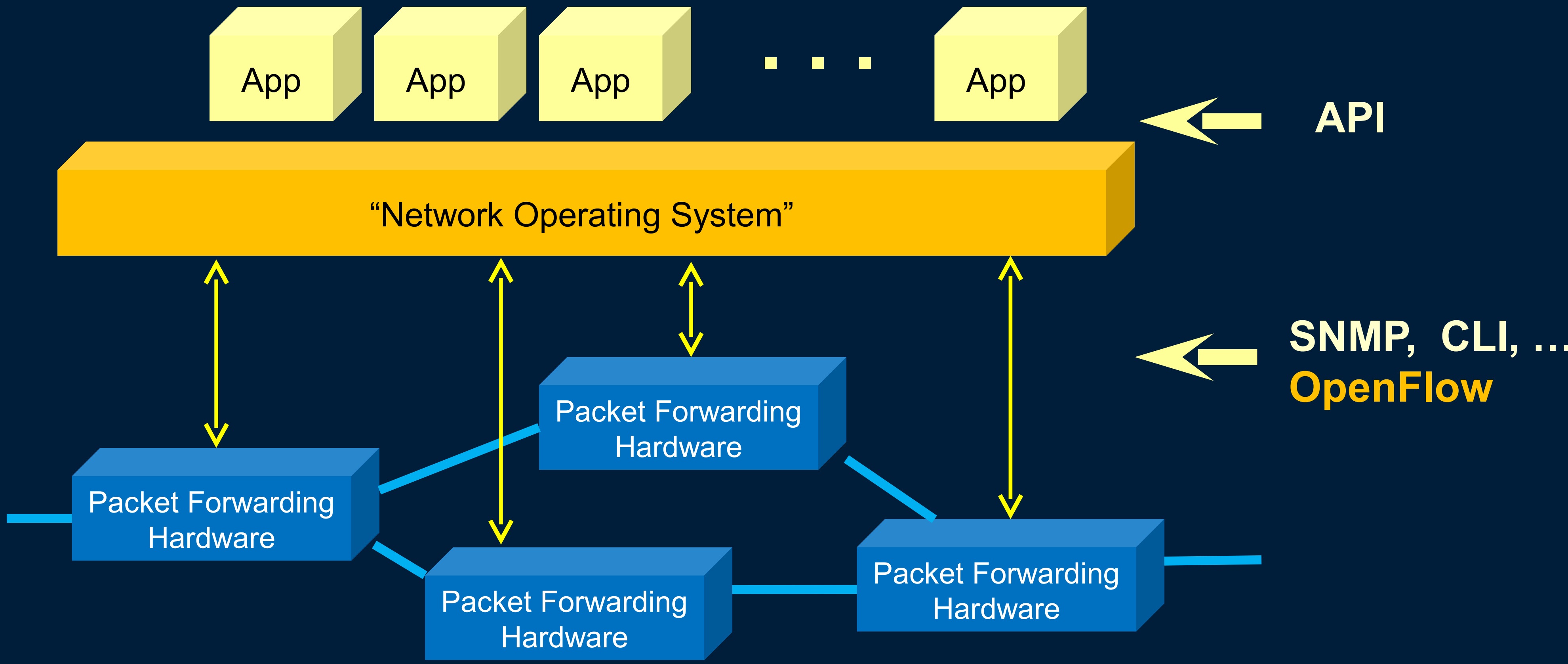
Dedicated network bandwidth crucial for transparent service migration!

http://www.greenstarnetwork.com/

Software Defined Networking

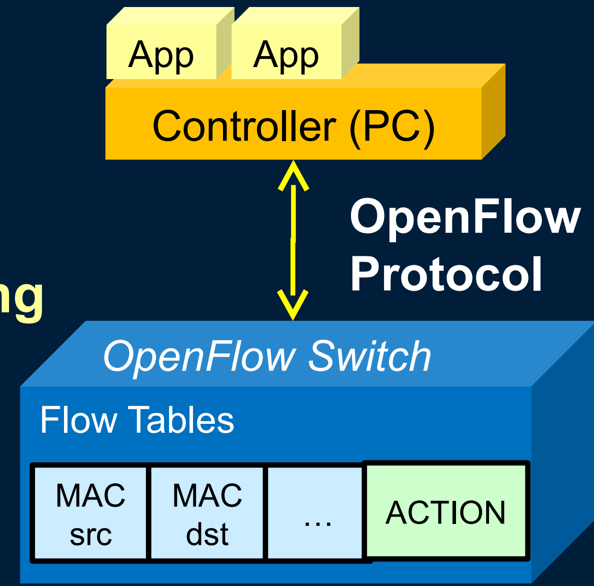
Software Defined Networking

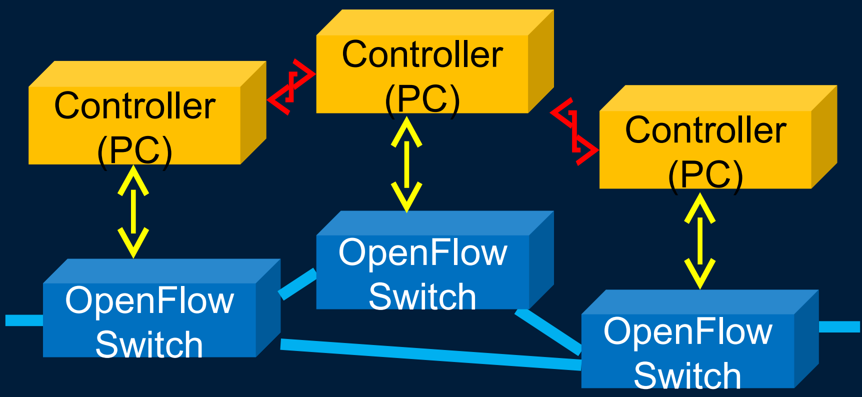
- SDN Paradigm - Network control by applications; provide an API to externally define network functionality

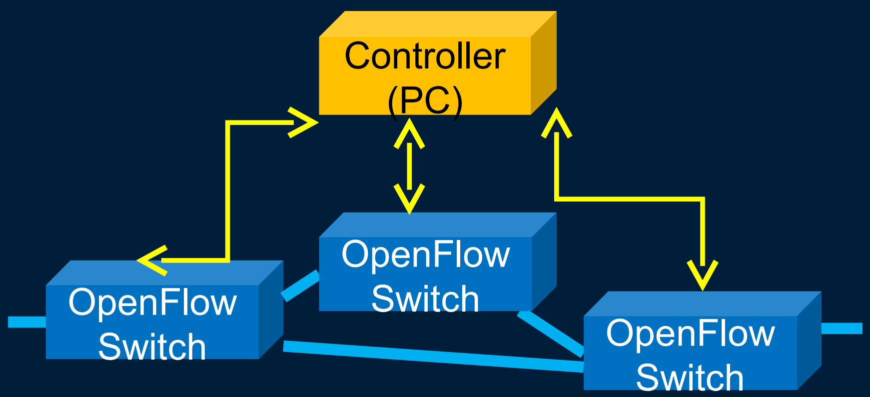


OpenFlow

- **Standardized SDN protocol**
  - Open Networking Foundation (https://www.opennetworking.org/)
- **Let external controller access/modify flow tables**
- **Allows separation of control plane and data forwarding**
- **Simple protocol, large application space**
  - Forwarding, access control, filtering, topology segmentation, load balancing, …
- **Distributed or centralized**
- **Reactive or pro-active**







Dynamic Circuits

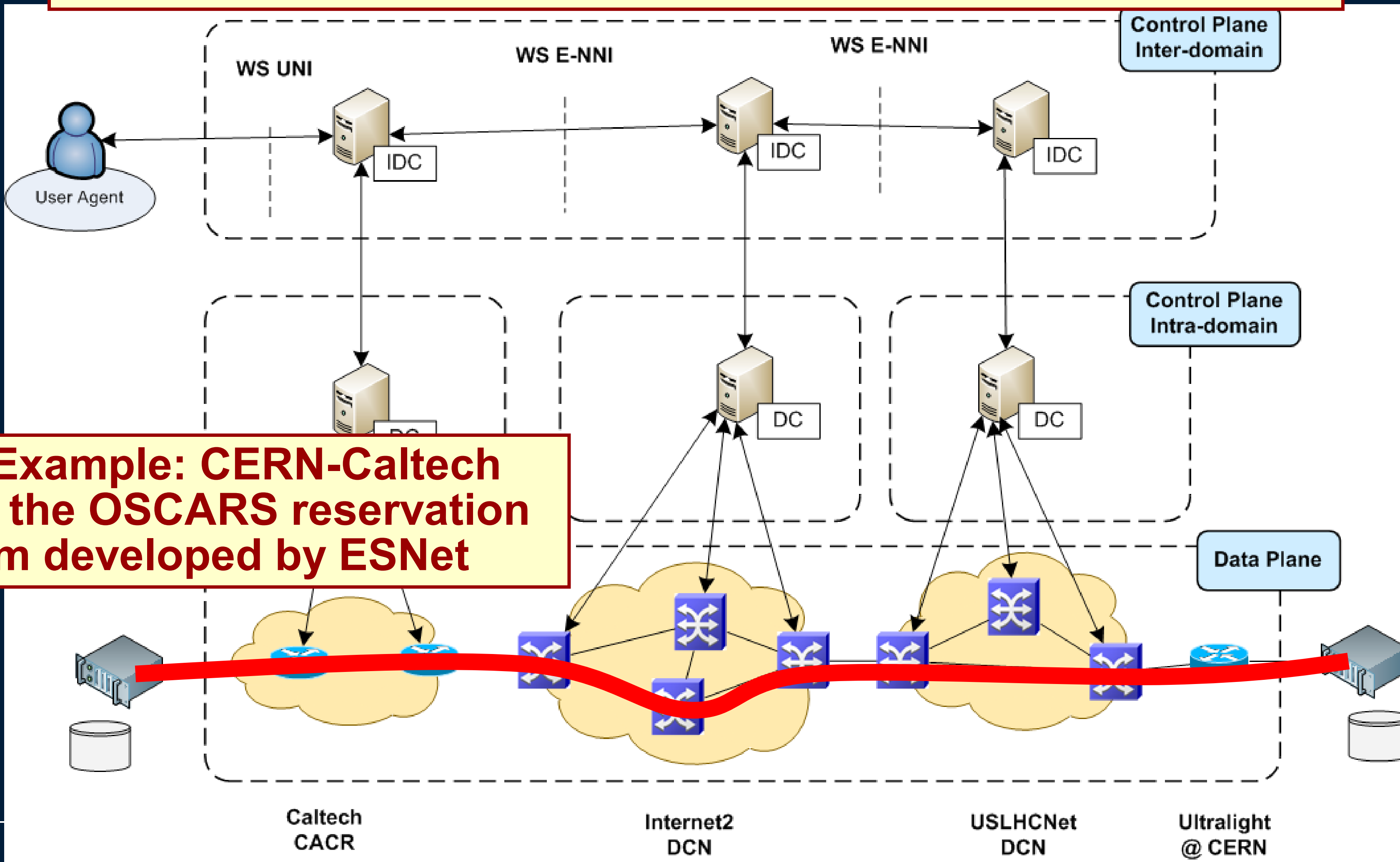
- **Aka Bandwidth On Demand, Dynamic Lightpaths, ….**

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
- **Significant effort in R&E Networking community**
  - Standardisation through OGF (OGF-NSI, OGF-NML)
- **Dynamic Circuit Service is present in several advanced R&E networks**
  - SURFnet (DRAC)
  - ESnet (OSCARS)
  - Internet2 (ION)
  - US LHCNet (OSCARS)
- **Planned (or in experimental deployment)**
  - E.g. JGN (Japan), GEANT (AutoBahn), RNP (OSCARS/DCN), …
- **DYNES: NSF funded project to extend hybrid & dynamic network capabilities to campus & regional networks**
  - In first deployment phase; fully operational in 2012

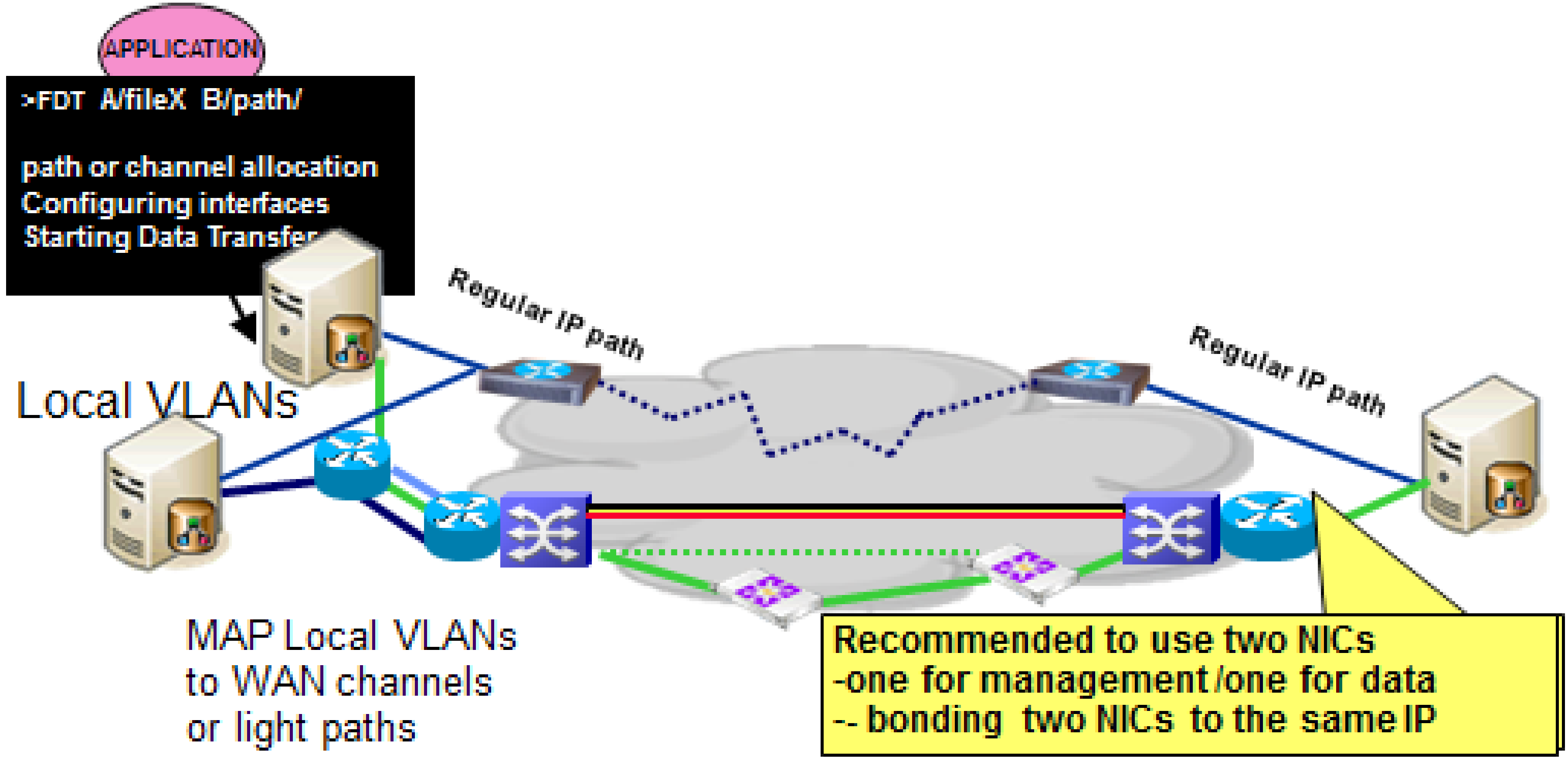
Dynamic Circuits: On-demand Point-to-Point Layer-2 Paths

Bandwidth requested by "User" Agent (application or GUI):
- **Scheduled**
- **On-demand**

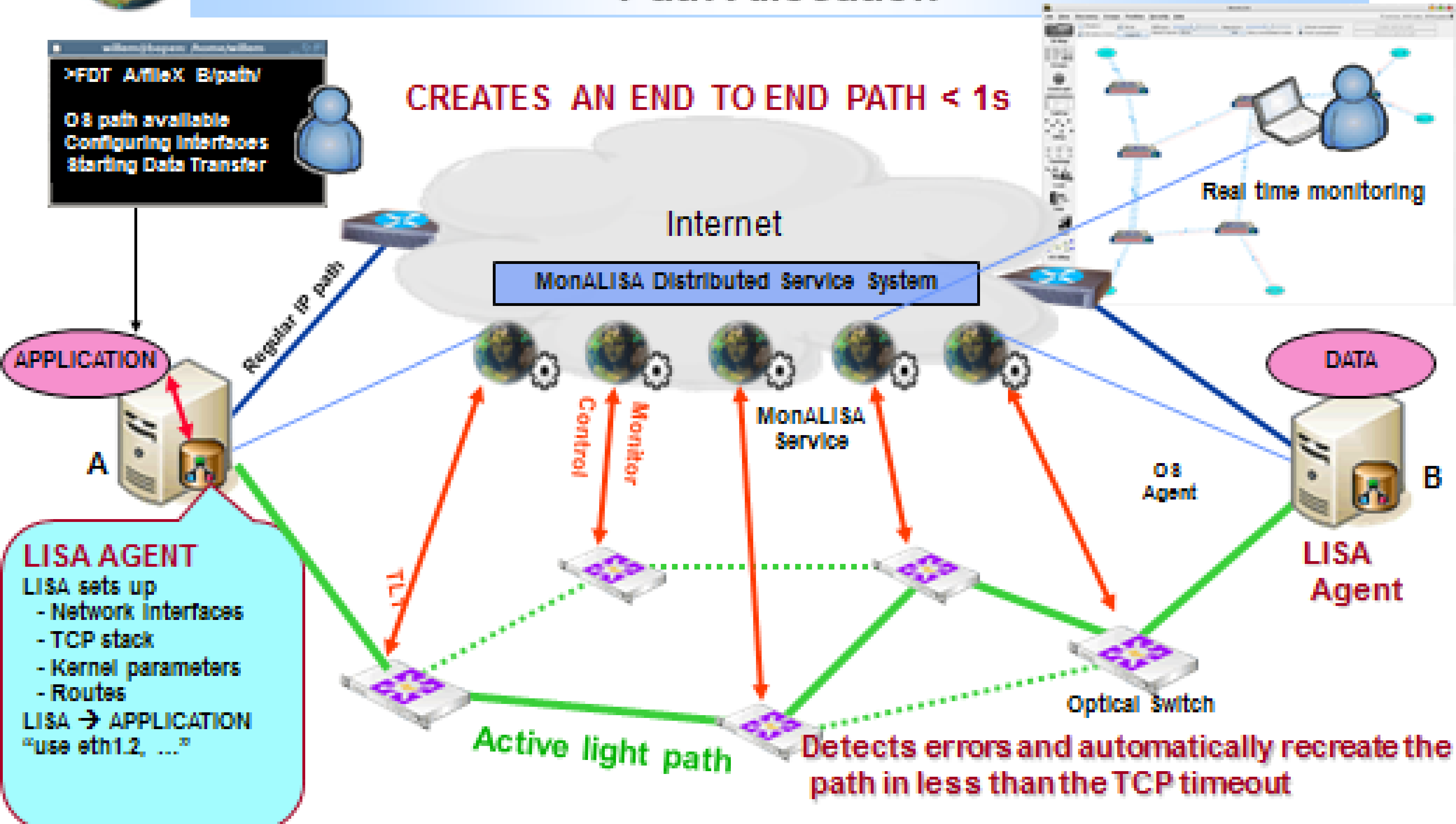


2009 Example: CERN-Caltech using the OSCARS reservation system developed by ESNet

“On-Demand”, Dynamic Circuits Channel and Path Allocation



"On-Demand", End to End Optical Path Allocation



DYNES

- The early dynamic circuit adopters…

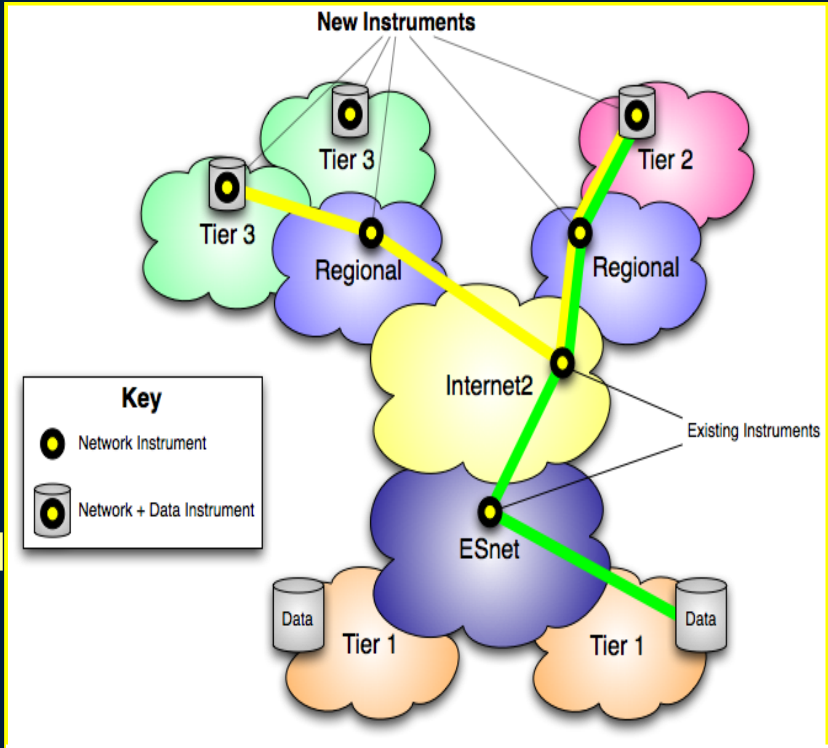
US Example: DYNES Project

- **NSF-funded project: DYnamic NEtwork System**
- **What is it?**
  - A nationwide cyber-instrument spanning up to ~40 US universities and ~14 Internet2 connectors
  - Extends Internet2s ION service into regional networks and campuses, based on ESnet's OSCARS implementation of IDC protocol
- **Who is it?**
  - A collaborative team including Internet2, Caltech, University of Michigan, and Vanderbilt University
  - Community of regional networks and campuses
  - LHC, astrophysics community, OSG, WLCG, other virtual organizations
- **The goals**
  - Support large, long-distance scientific data flows in the LHC, other leading programs in data intensive science (such as LIGO, Virtual Observatory, and other large scale sky surveys), and the broader scientific community
  - Build a distributed virtual instrument at sites of interest to the LHC but available to R&E community generally

http://www.internet2.edu/dynes

DYNES System Description

- **AIM: extend hybrid & dynamic capabilities to campus & regional networks**
  - A DYNES instrument must provide two basic capabilities at the Tier 2s, Tier3s and regional networks:
    1. Network resource allocation such as bandwidth to ensure transfer performance
    2. Monitoring of the network and data transfer performance
- **All networks in the path require the ability to allocate network resources and monitor the transfer. This capability currently exists on backbone networks such as Internet2 and ESnet, but is not widespread at the campus and regional level**
  - In addition Tier 2 & 3 sites require:
    3. Hardware at the end sites capable of making optimal use of the available network resources

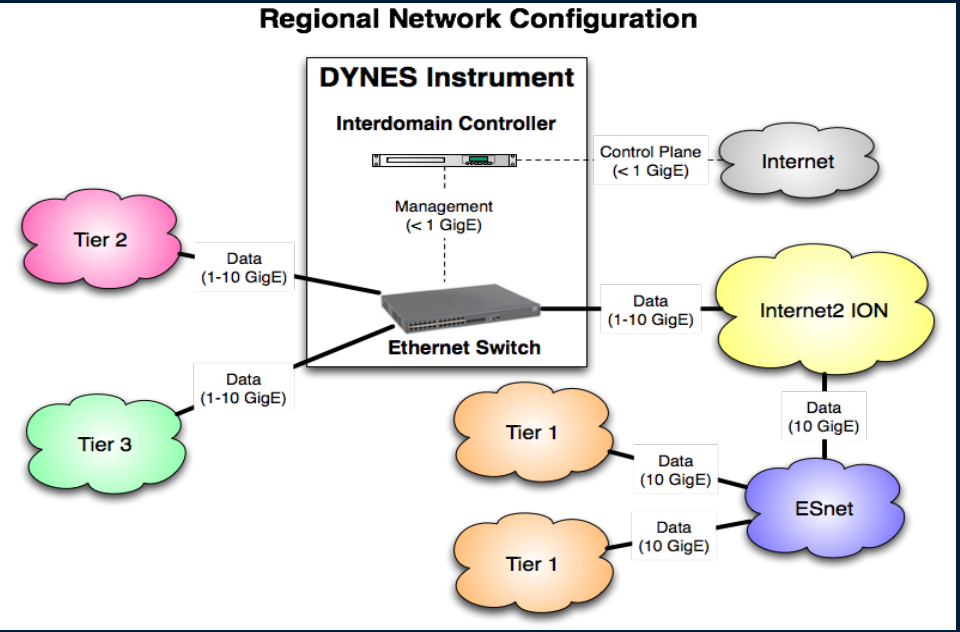


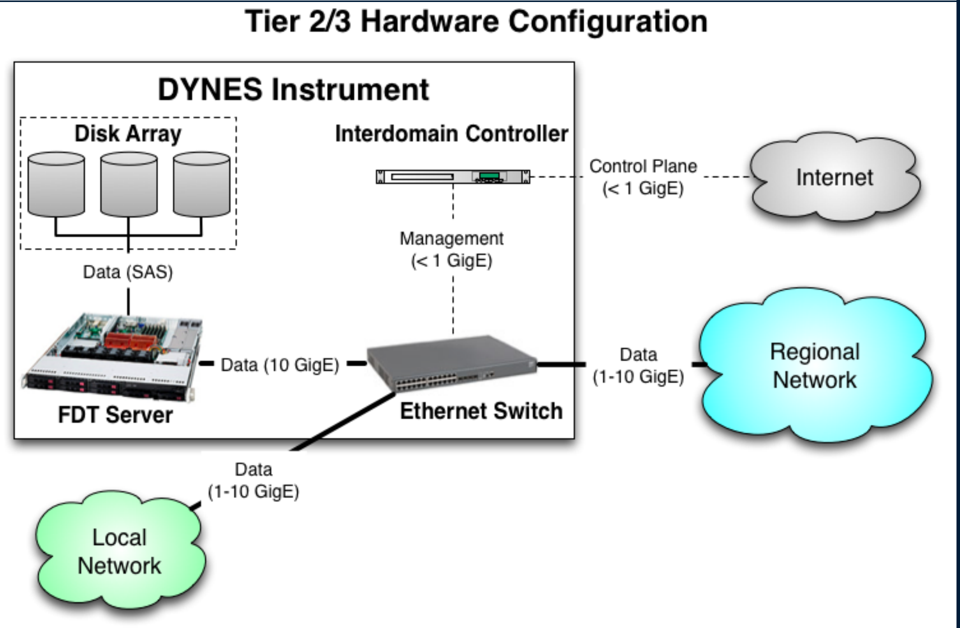
Two typical transfers that DYNES supports: one Tier2 - Tier3 and another Tier1-Tier2.

The clouds represent the network domains involved in such a transfer.

DYNES: Instrument Design

- **Regional networks require**
  1. An Ethernet switch
  2. An Inter-domain Controller (IDC)
- **The DYNES (sub-)instrument at a Tier2 or Tier3 site in addition includes**
  3. A Fast Data Transfer (FDT) server. Sites with 10GE throughput capability will have a dual-port Myricom 10GE network interface in the server
- **The configuration of the IDC consists of OSCARS, DRAGON, and perfSONAR. This allows the regional network to provision resources on-demand**

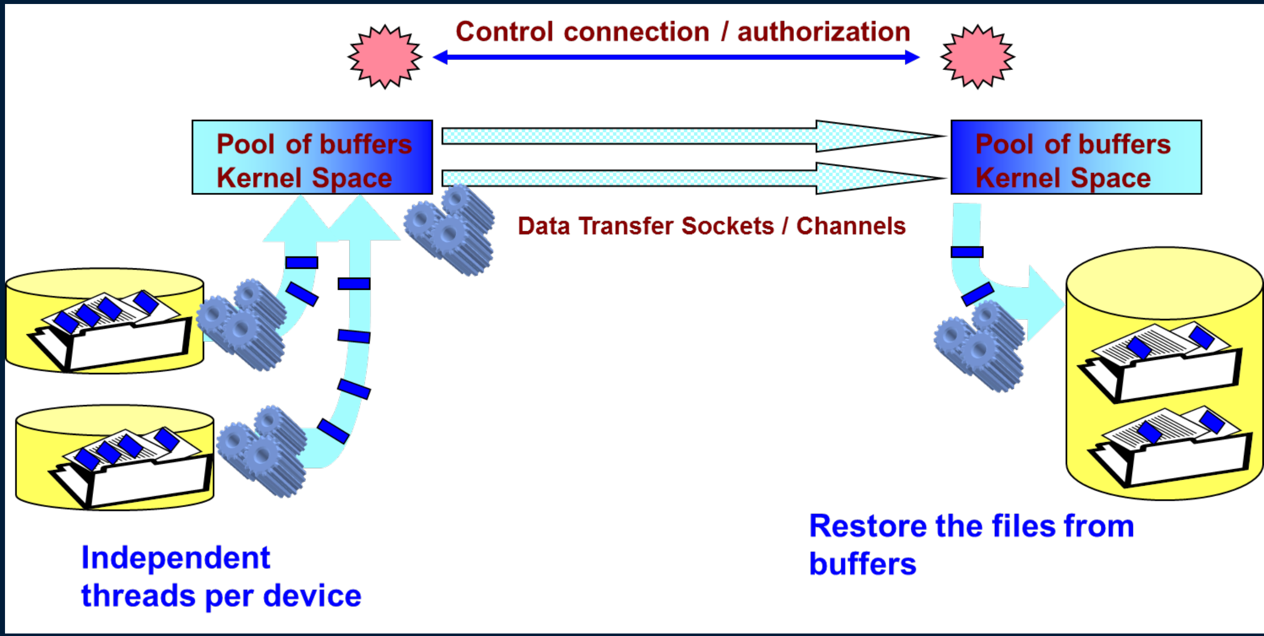




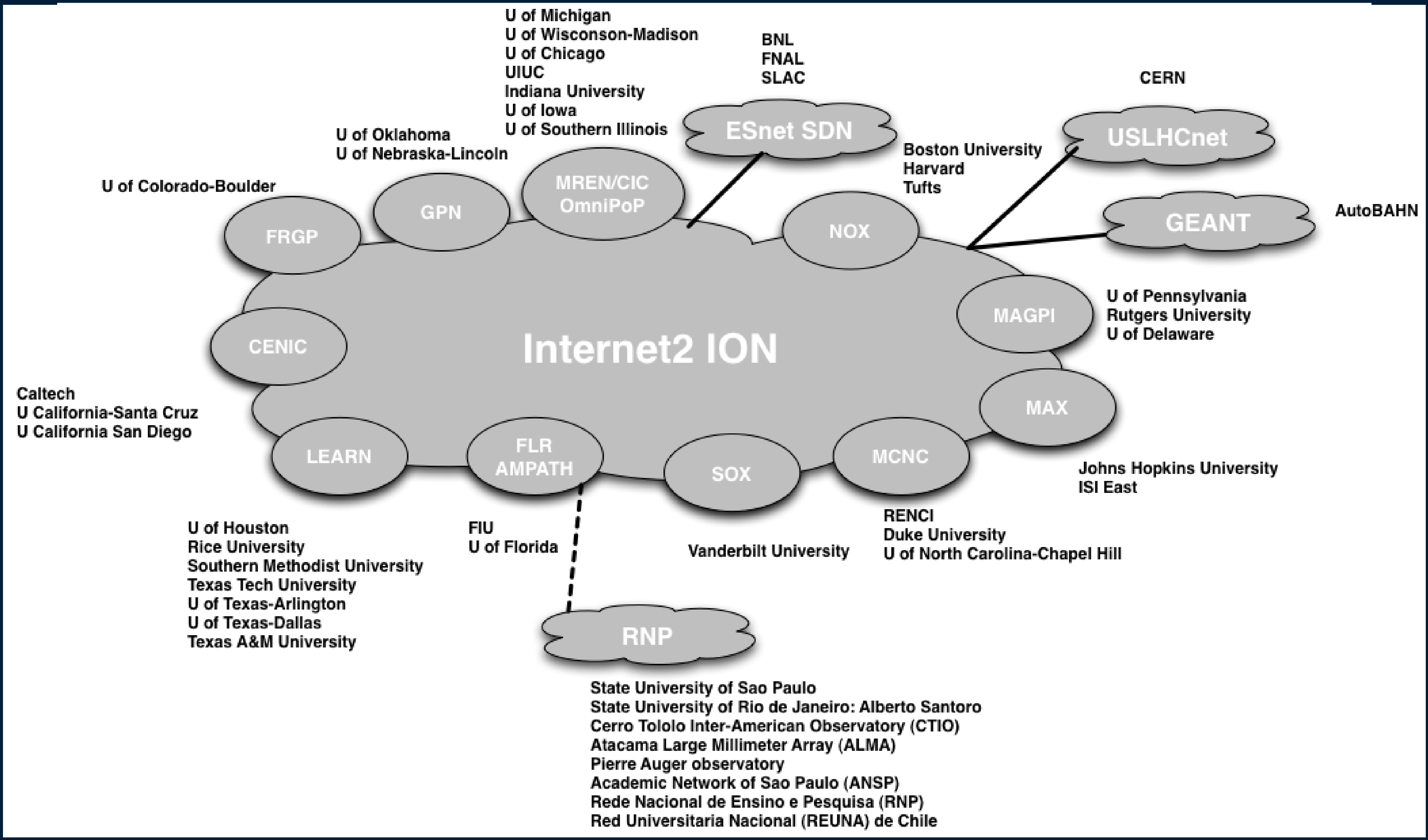
Fast Data Transfer (FDT)

- **DYNES instrument includes a storage element, FDT as transfer application**
- **FDT is an open source Java application for efficient data transfers**
- **Easy to use: similar syntax with SCP, iperf/netperf**
- **Based on an asynchronous, multithreaded system**
- **Uses the New I/O (NIO) interface and is able to:**
  - stream continuously a list of files
  - use independent threads to read and write on each physical device
  - transfer data in parallel on multiple TCP streams, when necessary
  - use appropriate size of buffers for disk IO and networking
  - resume a file transfer session

FDT uses IDC API to request dynamic circuit connections



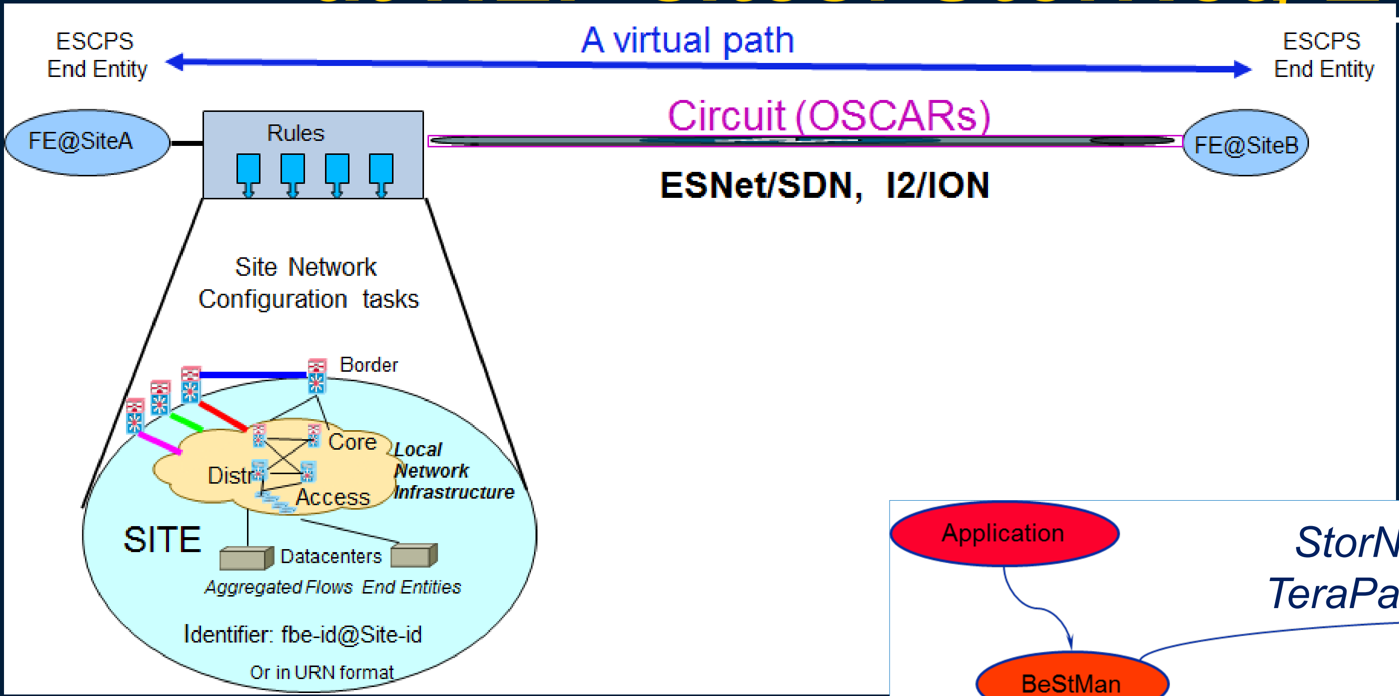
DYNES Current Topology



The Case for Dynamic Provisioning in LHC Data Processing

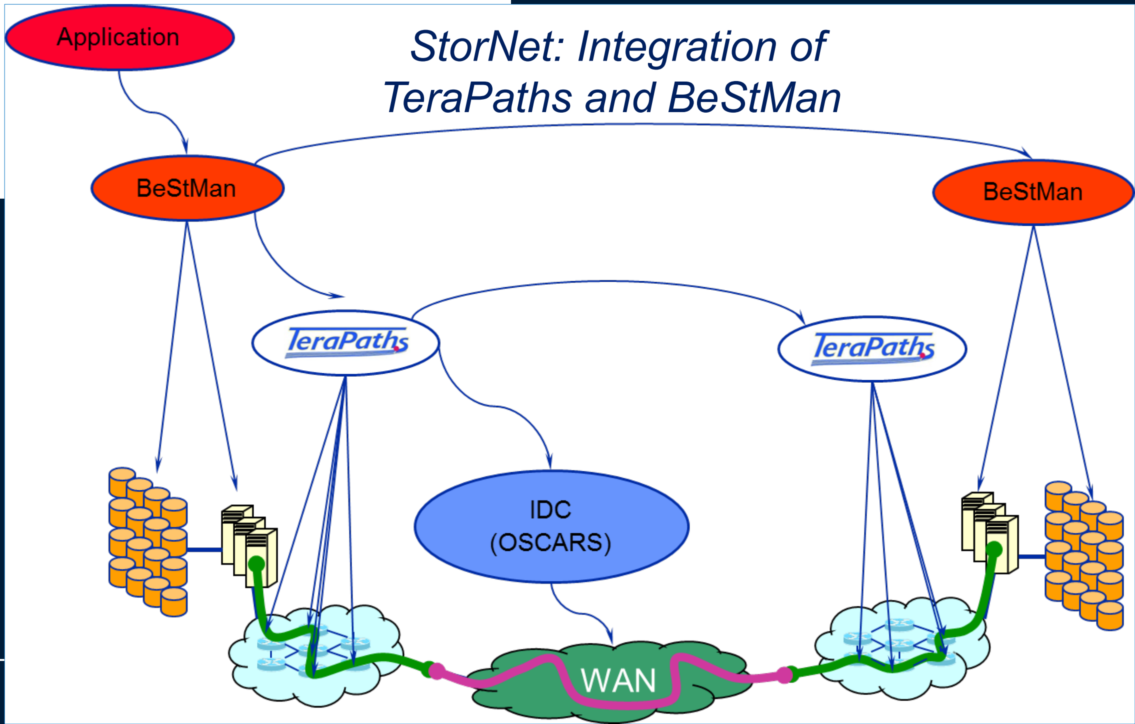
- **Data models do not require full-mesh @ full-rate connectivity @ all times**
- **On-demand data movement will augment and partially replace static pre-placement · Network utilisation will be more dynamic and less predictable**
- **Performance expectations will not decrease**
  - More dependence on 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 important for workflow efficiency
- **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
- **Note: Solutions for the LHC community need global reach**

Development of Dynamic Circuits at HEP sites: StorNet, ESCPS



Building on previous developments in and experience from the TeraPaths and LambdaStation projects

- **StorNet – BNL, LBNL, UMICH**
  - Integrated Dynamic Storage and Network Resource Provisioning and Management for Automated Data Transfers
- **ESCPS – FNAL, BNL, Delaware**
  - End Site Control Plane System



OGF Standards

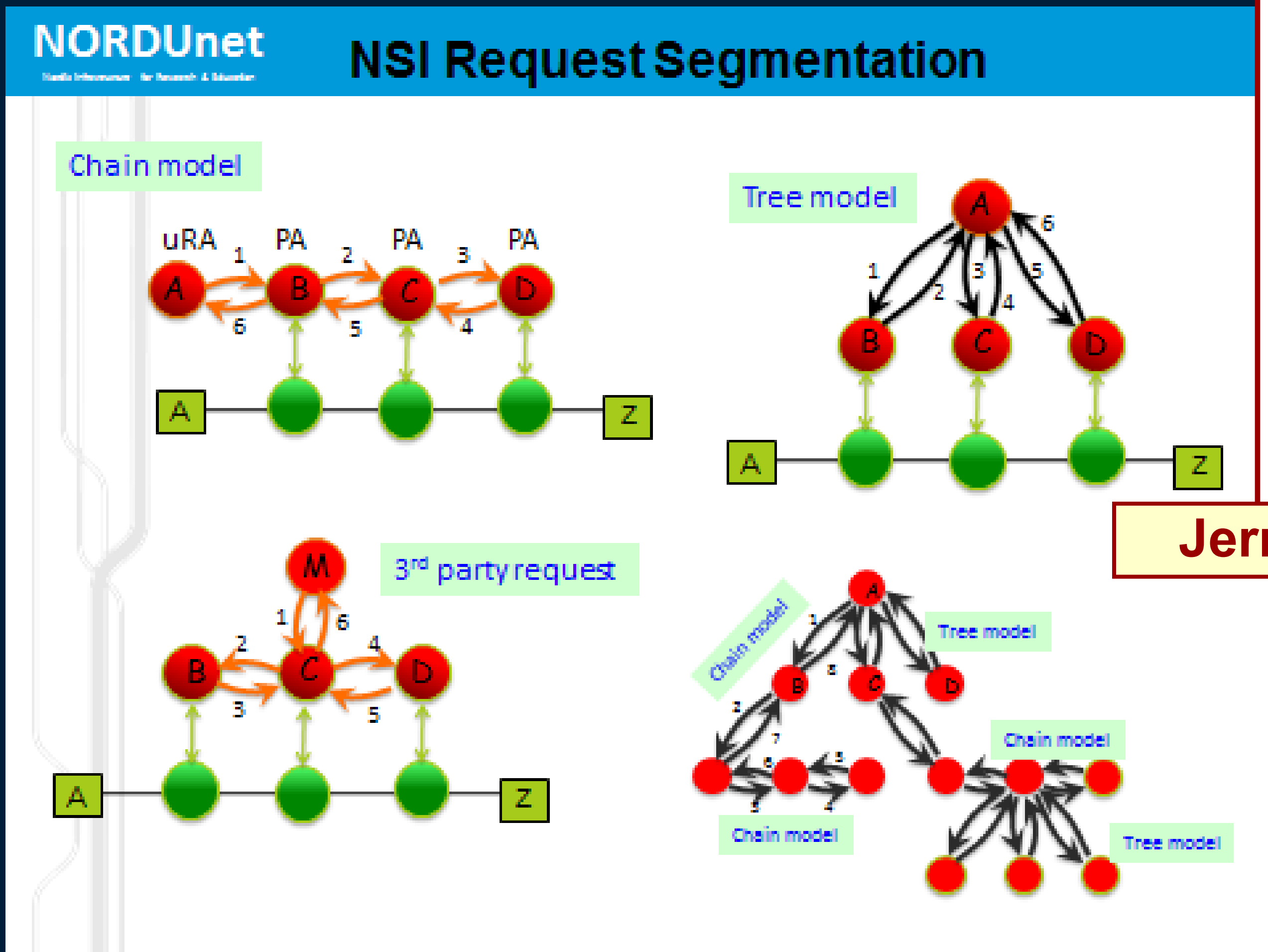
- Related to dynamic provisioning

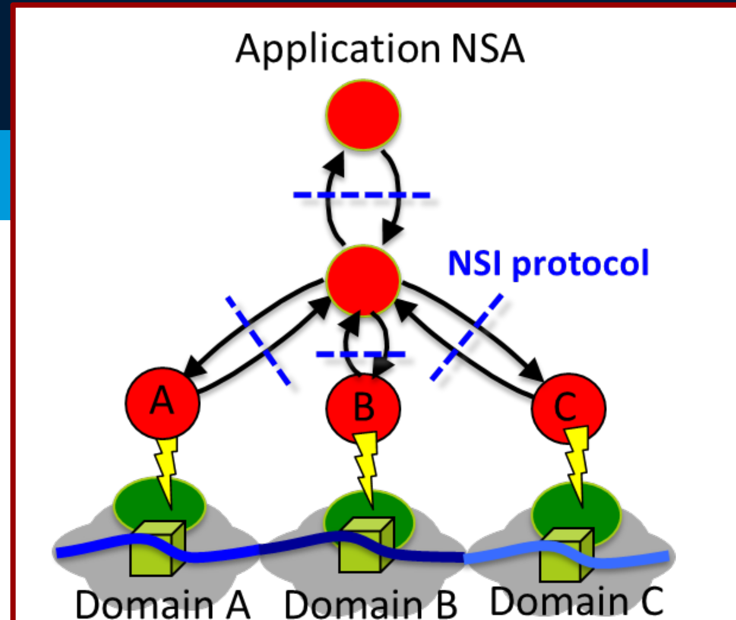
OGF Standards in Working

- **Standards provide interoperability**
- **NSI: Network Services Interface**
  - *The Network Service Interface Working Group will provide the recommendation for a generic network service interface that can be called by a network external entity such as end users, middleware, and other network service providers.*
- **NMC: Network Measurement and Control**
  - *The purpose of the Network Measurement and Control Working Group is to standardize the XML-based protocols that are currently in use in the perfSONAR project to control network measurement infrastructure*
- **Also important, although probably less exposed to users:**
- **NML: Network Mark-up Language**
  - *The purpose of the Network Mark-up Language Working Group is to combine efforts of multiple projects to describe network topologies, so that the outcome is a standardised network description ontology and schema*

OGF NSI Framework

- Aiming at definition of a Connection Service in a technology agnostic way
- Network Service Agent (NSA)
- High-level protocol

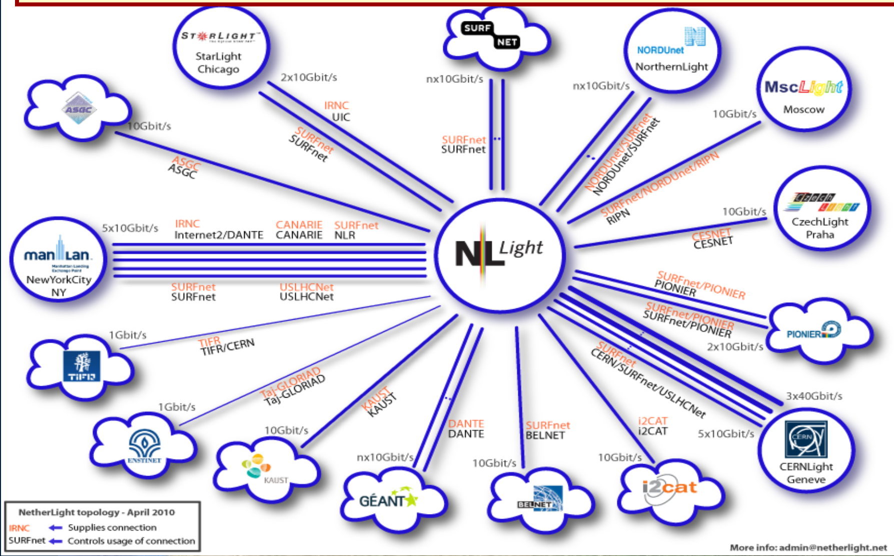




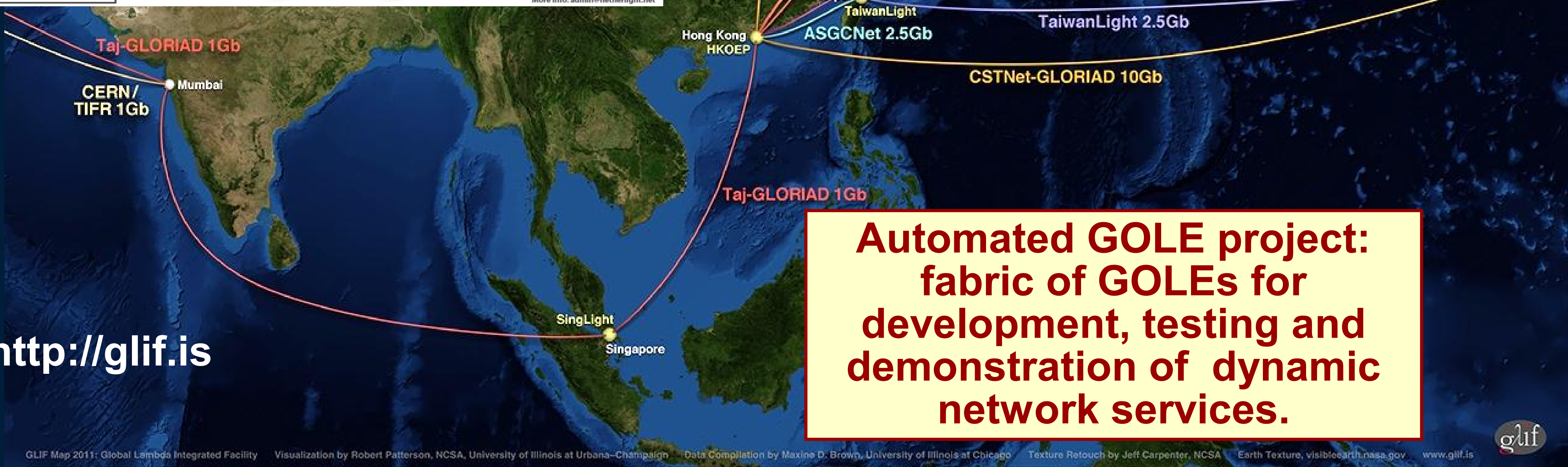
Jerry Sobieski, NORDUnet

GLIF Open Lightpath Exchanges

GOLE Example: Netherlight in Amsterdam



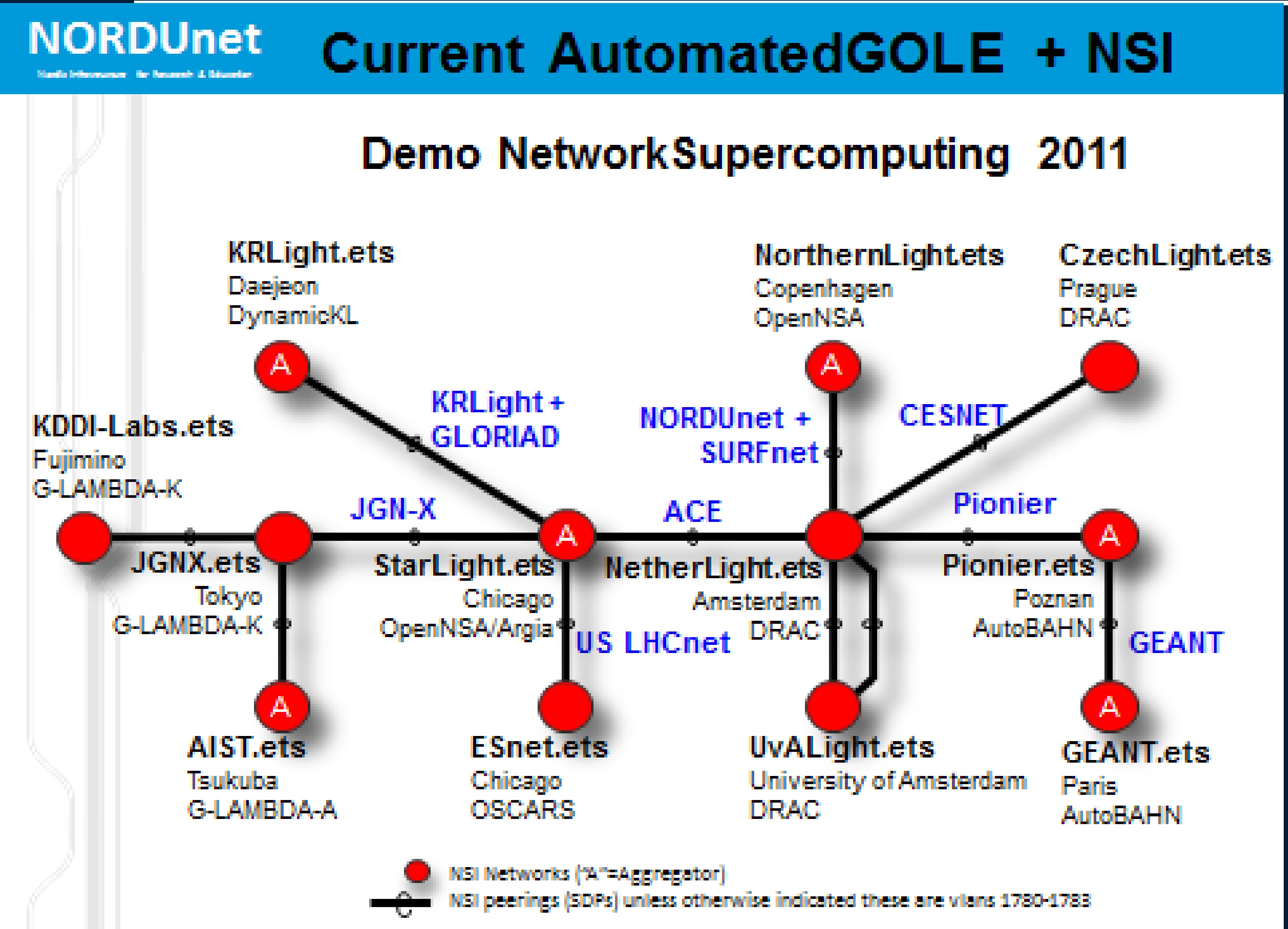
Exchange Points operated by the Research and Education Network community
http://glif.is



http://glif.is

Automated GOLE project: fabric of GOLEs for development, testing and demonstration of dynamic network services.

NSI + AutoGOLE demonstration



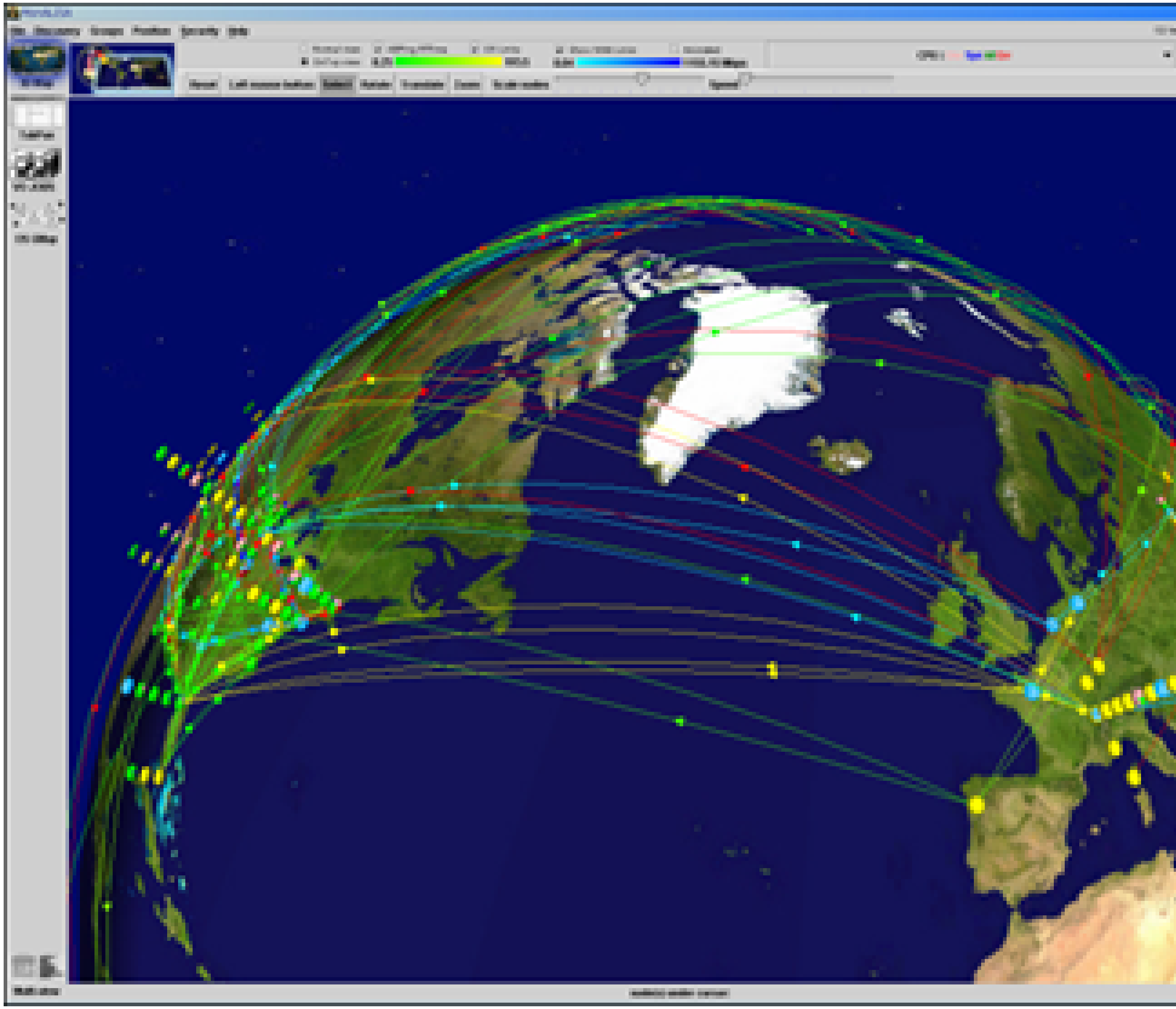
Jerry Sobieski, NORDUnet

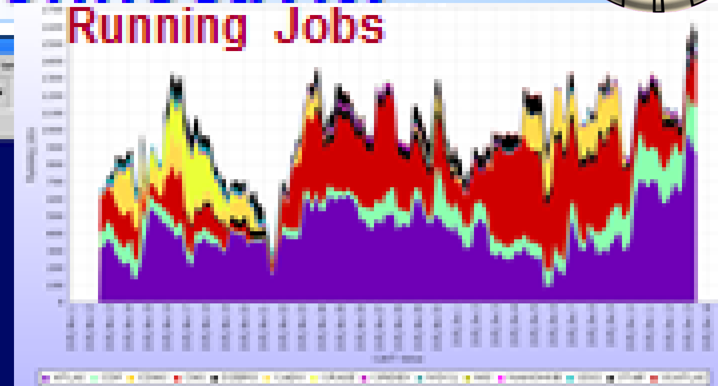
Software Implementations

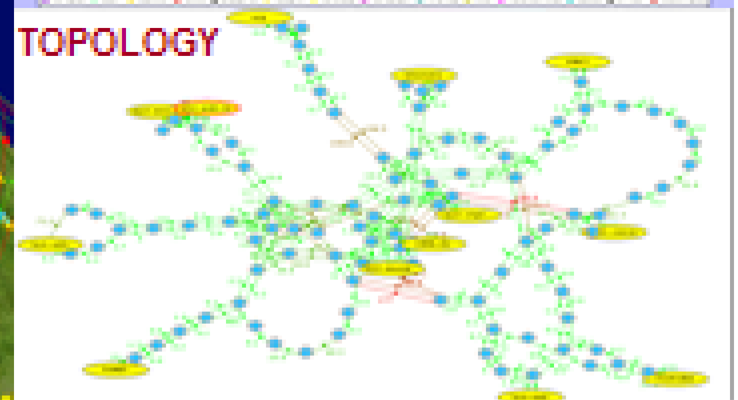
- **OpenNSA** – NORDUnet (DK/SE/)
- **OpenDRAC** – SURFnet (NL)
- **G-LAMBDA-A** - AIST (JP)
- **G-LAMBDA-K** – KDDI Labs (JP)
- **AutoBAHN** – GEANT (EU)
- **DynamicKL** – KISTI (KR)
- **OSCARS*** – ESnet (US)

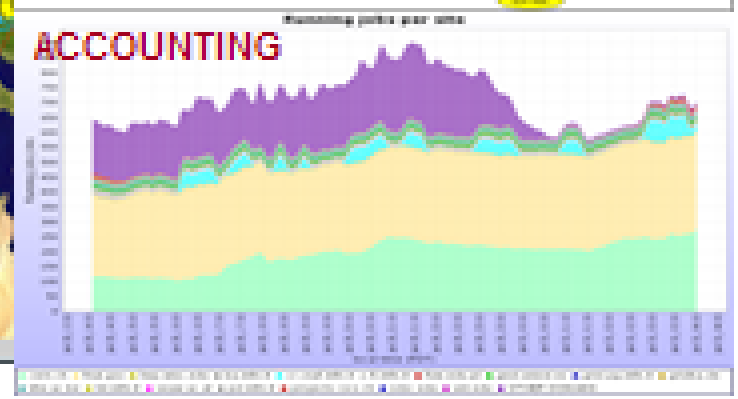
Pervasive Monitoring of End-to-end Systems

Monitoring Grid sites, Running Jobs, Network Traffic, and Connectivity

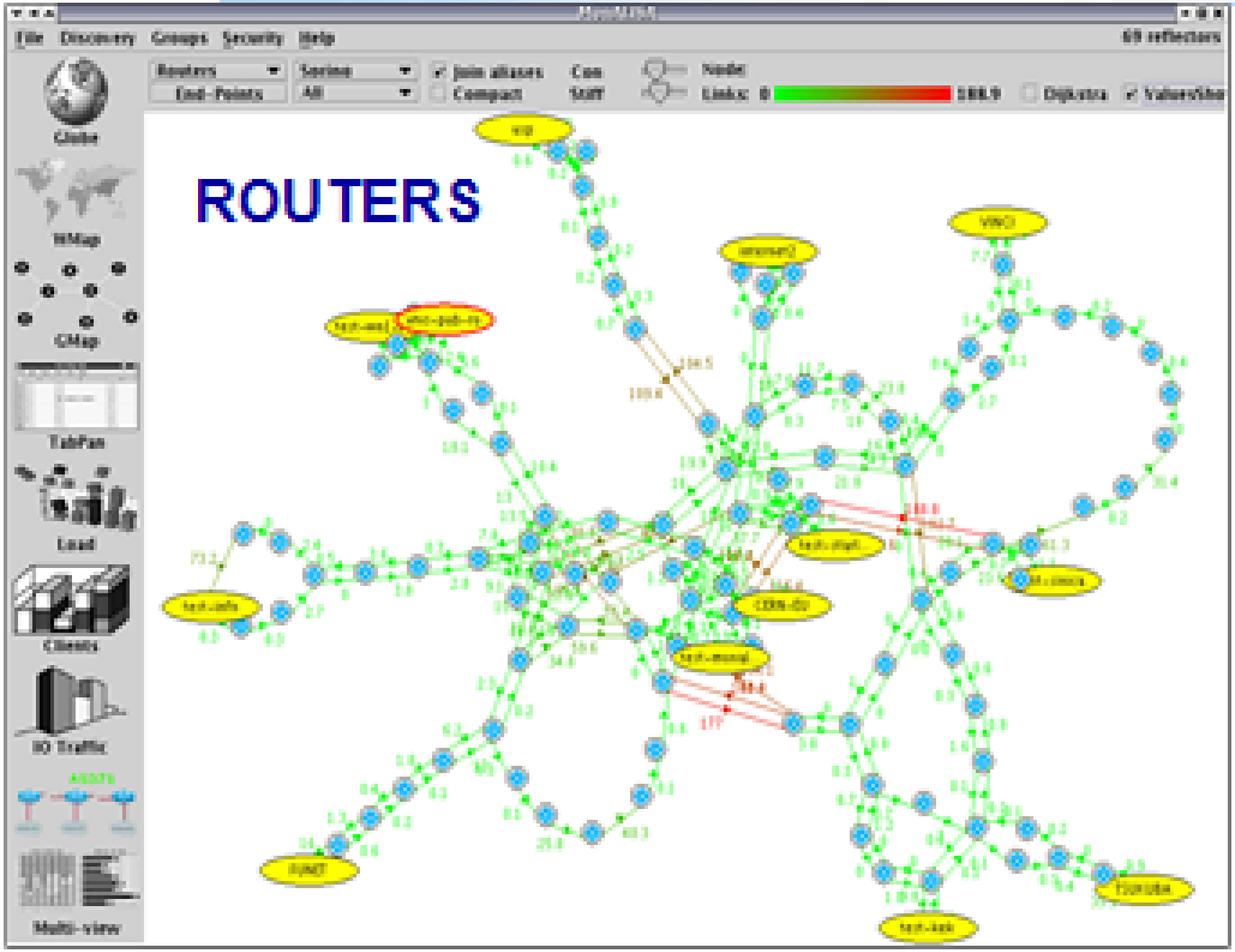




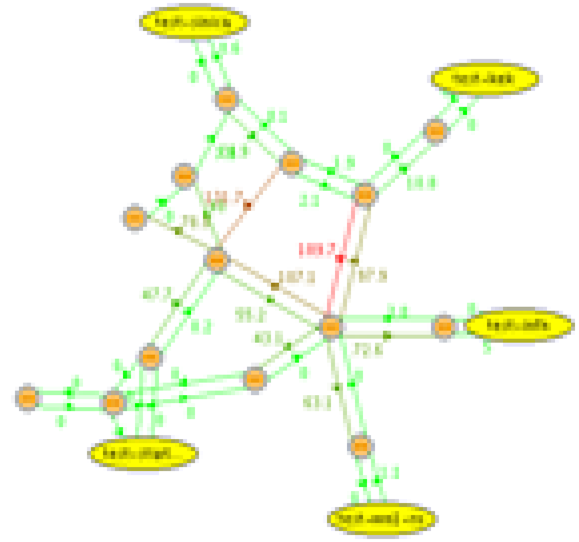




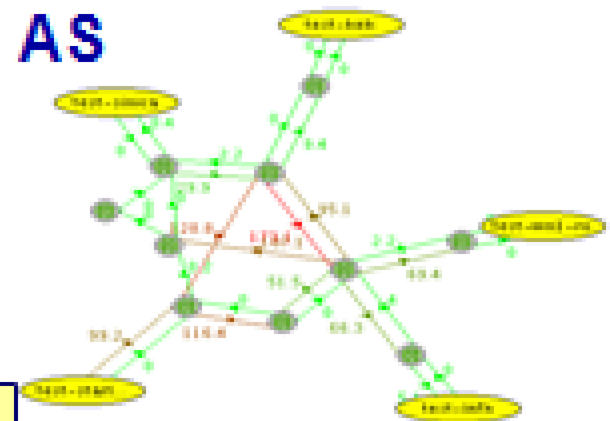
Monitoring Network Topology (L3), Latency, Routers



NETWORKS

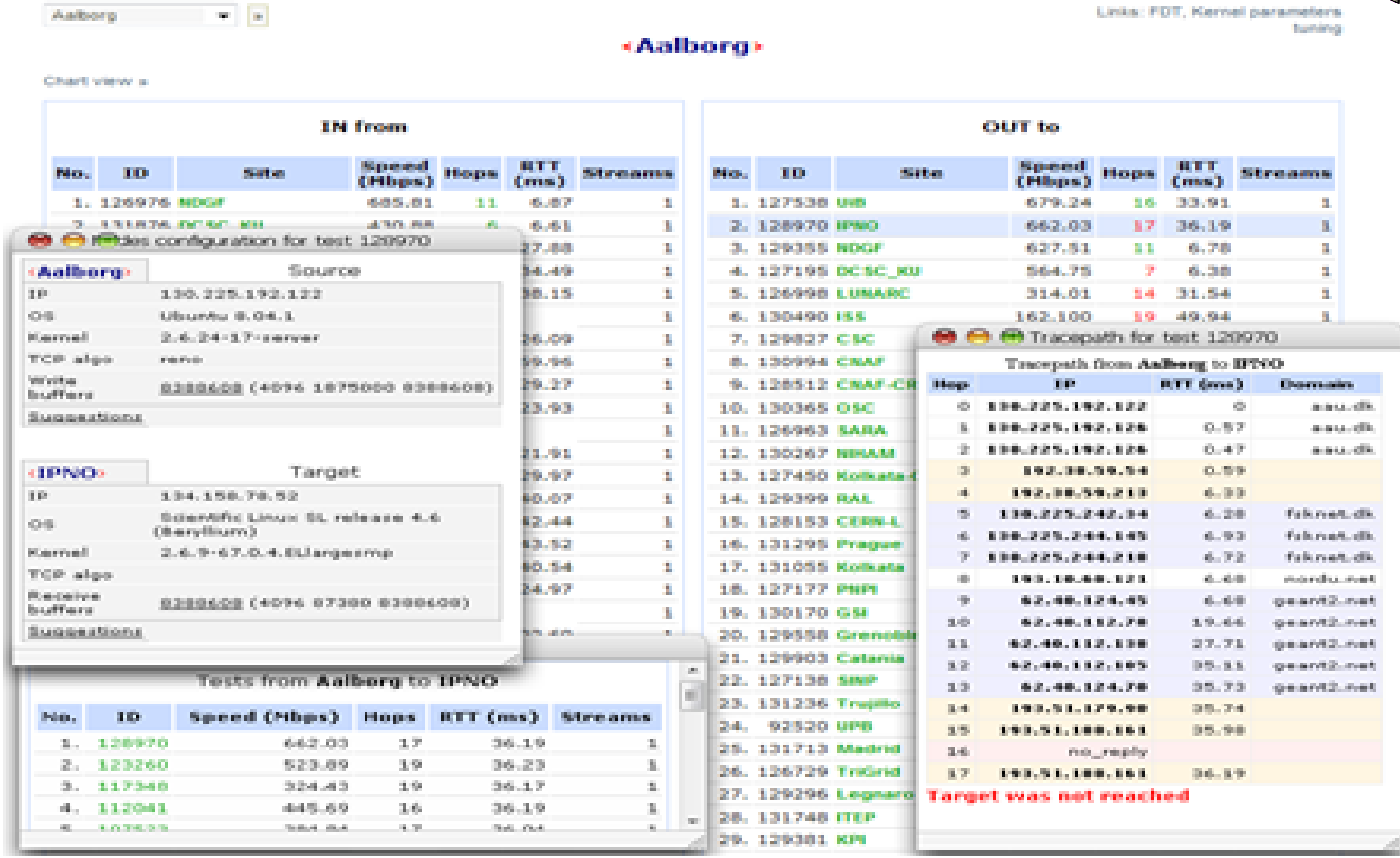


AS

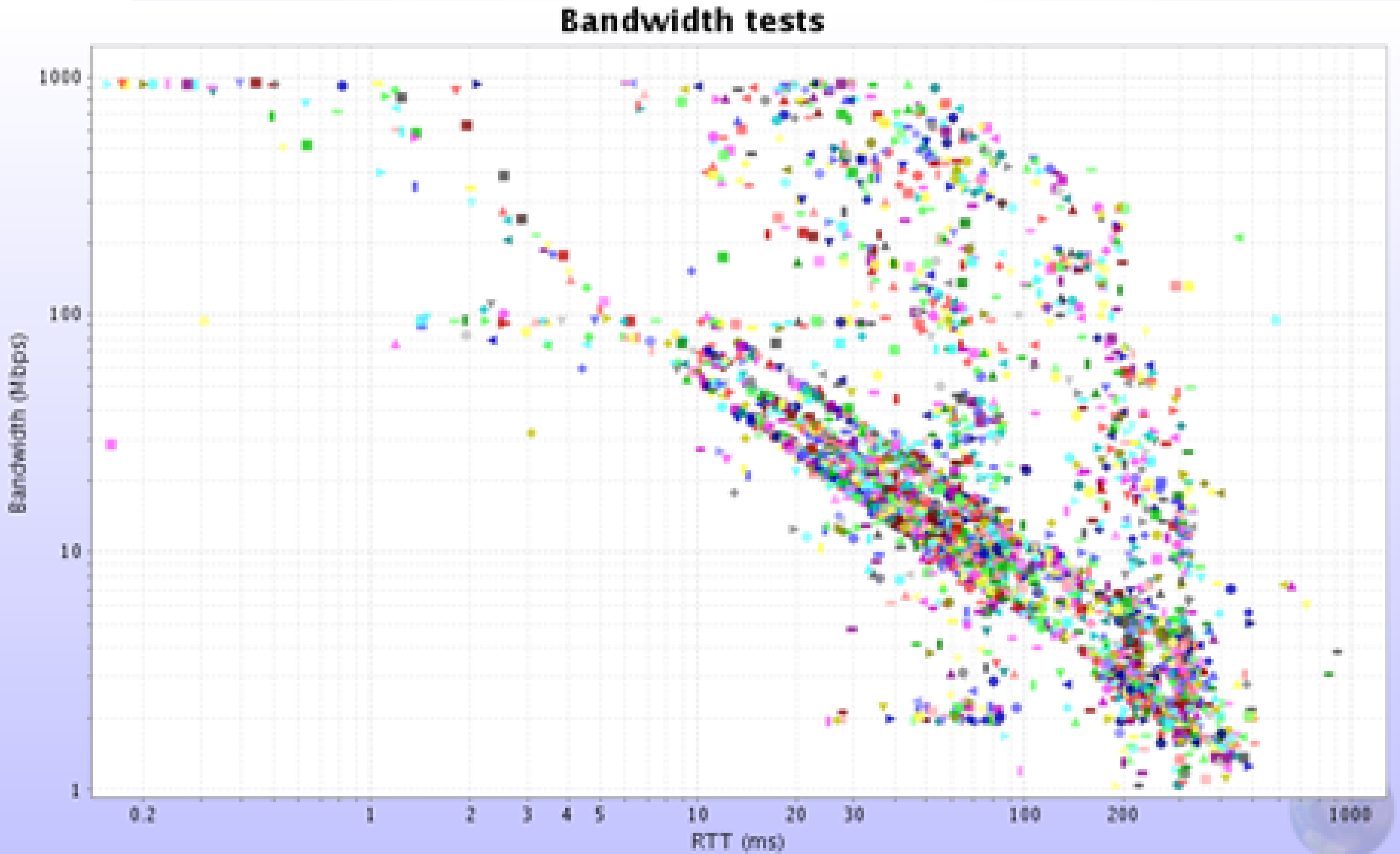


Real Time Topology Discovery & Display

Active Available Bandwidth measurements between all the ALICE grid sites



Active Available Bandwidth measurements between all the ALICE grid sites

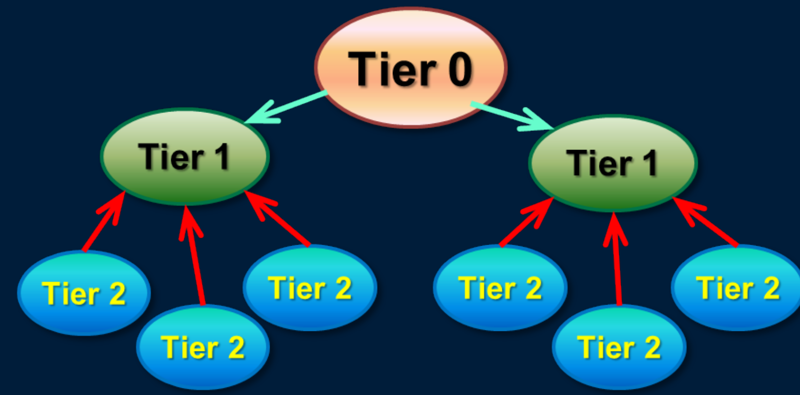


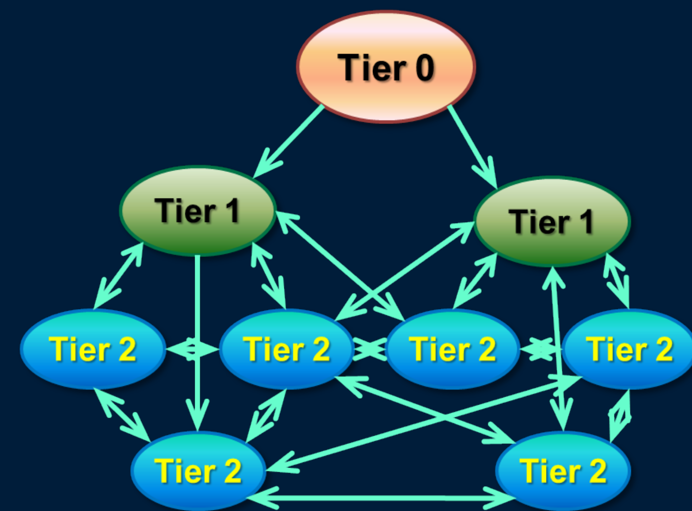
LHCONE

- LHC Open Network Environment

Computing Models Evolution

- **Moving away from the MONARC model**
- **Introduced gradually since 2010**
- **3 recurring themes:**
  - Flat(ter) hierarchy: Any site can use any other site as source of data
  - Dynamic data caching: Analysis sites pulling 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**
- **Increased reliance on network performance**





LHC Open Network Environment LHCONE

- **So far, T1-T2, T2-T2, and T3 data movements have been using General Purpose R&E 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 R&E GPN
- **Collaboration on global scale, diverse environment, many parties**
  - Solution has to be Open, Neutral and Diverse
  - Agility and Expandability
    - Scalable in bandwidth, extent and scope
- **Organic activity, growing over time according to needs**
- **Services being constructed:**
  - Multipoint, virtual network (logical traffic separation and TE possibility)
  - Static/dynamic point-to-point Layer 2 circuits (high-throughput data movement)
  - Monitoring/diagnostic

http://lhcone.net

LHCONE Future Development

- **2011 has seen an early prototype deployment**
  - Single VLAN, multiple domains, several LHC sites in Europe, US, Canada, India, Mexico
  - Operational challenge of global Layer 2 solution

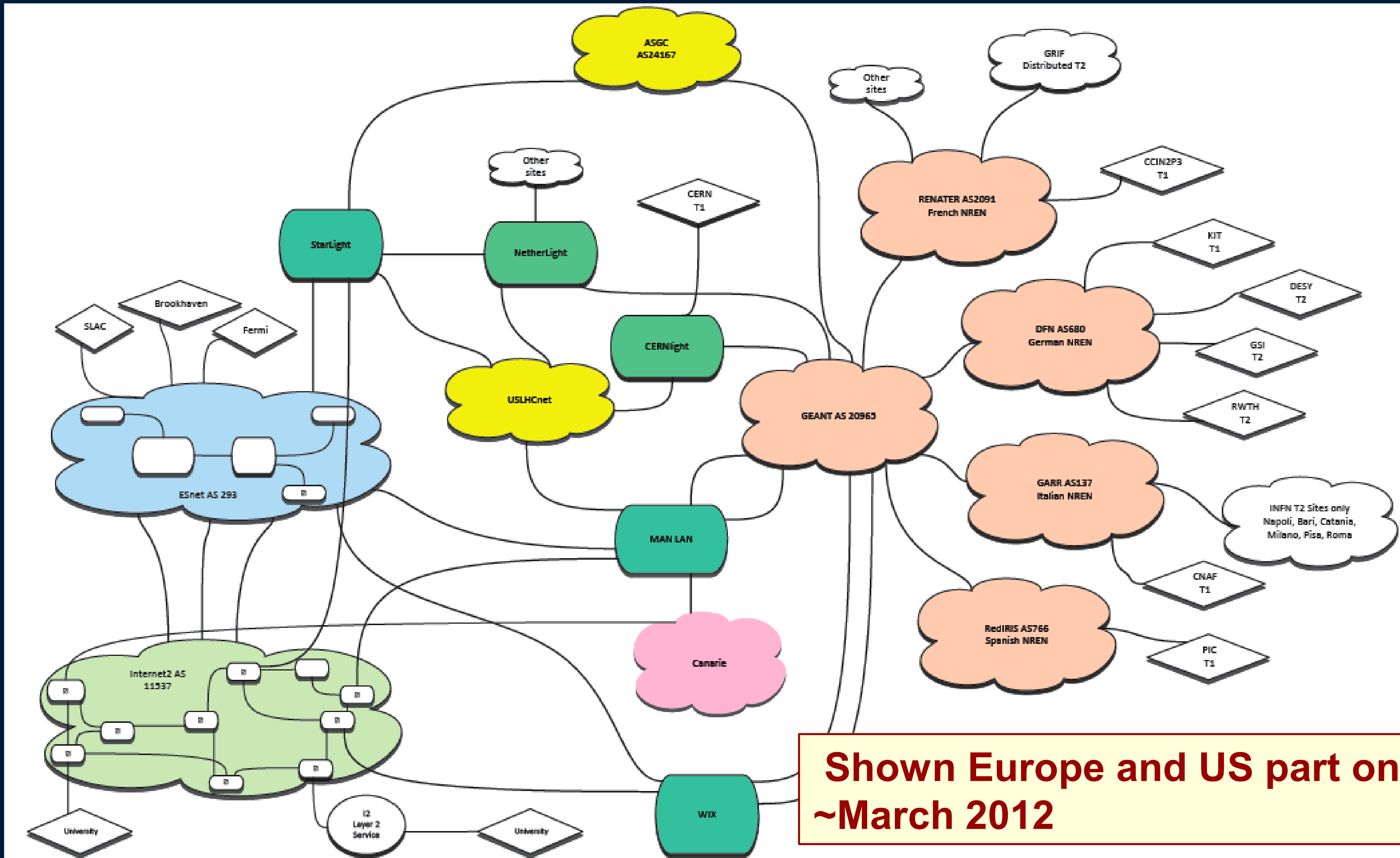
- **Fork in the path forward:**
  - A solution for “now”
    - To makes sure the immediate needs are satisfied
  - A long-term view at the LHC shutdown time scale
    - Leveraging next generation technologies
    - Requires some R&D investment to assure global scalability

- **LHC time scale:**
  - 2012: LHC run will continue until ~November
  - 2013-2014: LHC shutdown, restart late 2014
  - 2015: LHC data taking at full nominal energy (14 TeV)

LHCONE future activities

- **The Amsterdam Architecture workshop (Dec. 2011) has defined 5 activities:**
  1. VRF-based multipoint service: a “quick-fix” to provide the multipoint LHCONE connectivity as needed in places today
  2. Layer 2 multipath: evaluate use of emerging standards like TRILL (IETF) or Shortest Path Bridging (SPB, IEEE 802.1aq) in WAN environment
  3. Openflow: There was wide agreement at the workshop that SDN is the probable candidate technology for the LHCONE in the long-term, however needs more investigations
  4. Point-to-point dynamic circuits pilot
  5. Diagnostic Infrastructure: each site to have the ability to perform end-to-end performance tests with all other LHCONE sites
- **Plus, overarching:**
  6. Investigate what impact (if any) LHCONE will have on the LHC software stacks and data operations procedures

LHCONE – Multipoint Service L3VPN Implementation



Shown Europe and US part only, ~March 2012

Summary and Conclusions

- **Whatever the computing models of the future will look like, they will be more and increasingly dependent on network performance**
- **Network bandwidth will hardly be infinite**
- **Software-Defined Networking: application driven networks**
  - OpenFlow: standard backed by academia and industry alike
- **Dynamic network provisioning**
  - OGF standards in making
- **Efficient data movement for the LHC requires systematic end-to-end approach, including end-systems**
  - Adequate backbone and regional capacity
  - Well-connected end-sites
  - Network-awareness needs integration in the workflow
  - Better than best effort services are needed for determinism in workflows
- **LHCONE services to be implemented until LHC restart (late 2014):**
  - Multipoint data service
  - Dynamic point-to-point data service
  - Monitoring and diagnostics

Thank You!

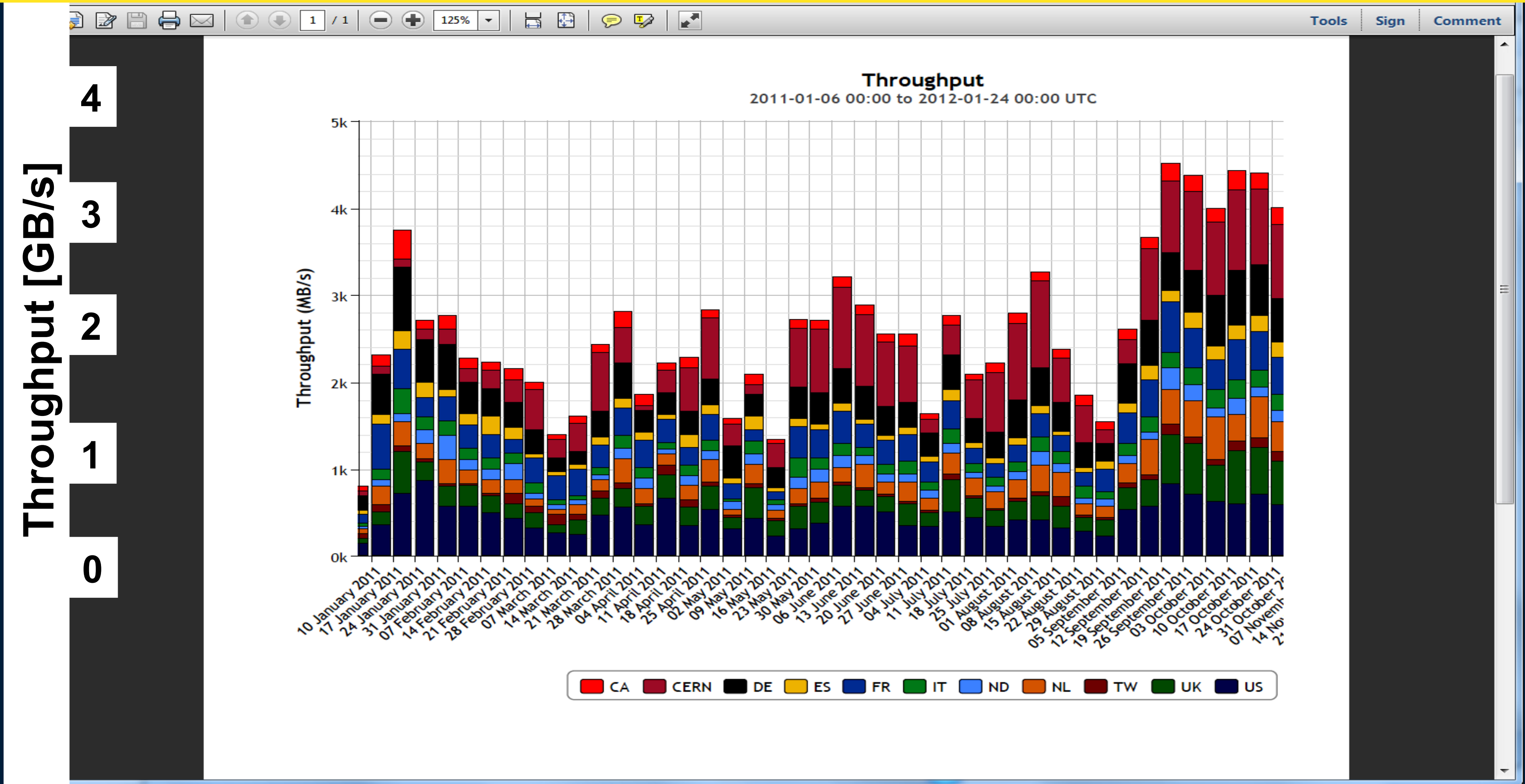
- Artur.Barczyk@cern.ch

EXTRA SLIDES

ATLAS Data Flow by Region: Jan. – Nov. 2011

~2.8 Gbytes/sec Average, 4.5 GBytes/sec Peak

> 100 Petabytes Transferred During 2011



Global Ring Network for Advanced Applications Development

Genomics on GLORIAD Network