USATLAS Networking: Status and Discussion

Shawn McKee/University of Michigan

US ATLAS Tier2/Tier3 Facilities Meeting

March 7th, 2011

Harvard Medical Campus

Status and Planning

- There are a number of interesting and relevant network-related activities underway:
  - The DYNES project is beginning to deploy its distributed virtual instrument for constructing dynamic circuits
  - The LHCOPN community has begun an effort (called “LHCONE”) to support Tier-2 and Tier-3 networking for the LHC
  - Within USATLAS we have almost completed full deployment and configuration of our perfSONAR infrastructure
  - All our sites now have 10GE WAN (or better) connectivity, in time for the next LHC run.
  - We have a new paradigm for data access and distribution: much more dynamic and much more grid-like.
  - This year will likely be very interesting in terms of physics reach for ATLAS --- if not discoveries at least strong hints should come out

Current USATLAS Network Status

- As noted we have almost completed a working perfSONAR installation amongst all our sites. We have made significant progress in "hardening" the installations
- New NAGIOS monitoring from Tom Wlodek/BNL has been very helpful in isolating the remaining issues.
- All Tier-2 sites (and some Tier-3s) have 10GE WAN
  - UTA still has a bottleneck of 2x1GE for its gridftp servers
- This quarter we have two network related items for sites to certify: "Green" perfSONAR matrices and new Loadtest

Network Monitoring

- We have a few tools for 'network' monitoring:
  - Hiro's regularly scheduled DDM transfers to each site (see https://www.usatlas.bnl.gov/dq2/throughput)
  - perfSONAR deployment monitored by BNL's Nagios (see https://nagios.racf.bnl.gov/nagios/cgi-bin/prod/perfSonar.php?page=100)
  - FTS monitoring (see http://www.usatlas.bnl.gov/fts-monitor/ftsmon)
- There has been a intensive effort on the perfSONAR front to get robust, appropriately configured instances operating at all our sites. Basically there now.
- Having the perfSONAR data collected correctly and consistently is a necessary but not sufficient step.

perfSONAR Latency Monitoring

Status of perfSONAR Latency Matrix

Status as of: March 7, 2011, 11:50 am	0	1	2	3	4	5	6	7	8
0:psmsu01.aglt2.org (AGLT2)	-	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK
1:psum01.aglt2.org	OK	-	OK	OK	OK	OK	OK	OK	OK

The rows of this table represent SOURCE nodes for a test while the columns represent DESTINATION nodes.
Each cell in the table represents a source-destination LATENCY test via OWAMP (600 UDP packets/test) tests, 1/minute.
The metric we are plotting is the packet loss between the source and destination averaged over the last 30 minutes.
Each cell contains the result of two tests:

The upper result is the loss measured in the test initiated from the source end.
The lower result is the loss measured in the test initiated from the destination end.

An 'OK' (green) result is when the average packet loss is less than 2 out of 600 packets.
A 'WARNING' (orange) result is when the average packet loss >=2 but < 10 out of 600 packets.
A 'CRITICAL' (red) result is when EITHER the test is not defined or the packet loss >= 10 out of 600 packets.
An 'UNKNOWN' (brown) result may indicate any other test outcome, including but not limited to: uncomprehensible test output, no response, test timed out etc.

5:atlas-npt1.bu.edu (NET2)	OK OK	OK OK	OK OK	OK OK	CRIT OK	-	OK OK	OK OK	OK OK
6:netmon1.atlas-swt2.org (SWT2)	OK OK	OK OK	OK OK	OK OK	OK OK	CRIT OK	-	OK OK	OK OK
7:ps1.ochep.ou.edu (SWT2)	OK OK	OK OK	OK OK	OK OK	OK OK	CRIT OK	OK OK	-	OK OK
8:psnr-lat01.slac.stanford.edu (WT2)	OK OK	OK OK	OK OK	OK OK	OK OK	CRIT OK	OK OK	OK OK	-

perfSONAR Throughput Monitoring

Status of perfSONAR Throughput Matrix

Status as of: March 7, 2011, 11:50 am	0	1	2	3	4	5	6	7	8
0:psmsu02.aglt2.org (AGLT2)	-	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK	OK OK
1:psum02.aglt2.org (AGLT2)	OK	-	OK	OK	OK	OK	OK	OK	OK

The rows of this table represent SOURCE nodes for a throughput test while the columns represent DESTINATION nodes.
Each cell in the table contains the result of two versions of a BWCTL throughput test for the specified source and destination.
Tests are configured to run by BOTH the source and destination once every 4 hour period.
The upper link in each cell represents the results of the throughput test initiated from the SOURCE end.
The lower link in each cell represents the results of the throughput test initiated from the DESTINATION end.
A cell is OK (green) if the measured bandwidth (averaged over all measurements in the last 24 hours) is >= 100 Mbits/sec.
A cell is WARNING (yellow) if the measured bandwidth (averaged over all measurements in the last 24 hours) is >=10 Mbits/sec and <100 Mbits/sec.
A cell is CRITICAL (red) if the measured bandwidth is not available (no test defined?) or is <10 Mbits/sec (averaged over all tests in the last 24 hours)

5:atlas-npt2.bu.edu (NET2)	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK UNKN	OK OK	UNKN OK
6:netmon2.atlas-swt2.org (SWT2)	OK UNKN	OK OK	OK OK	OK OK	OK OK	OK OK	-	OK OK	OK OK
7:ps2.ochep.ou.edu (SWT2)	UNKN OK	OK OK	OK OK	OK OK	WARN WARN	OK OK	OK OK	-	OK OK
8:psnr-bw01.slac.stanford.edu (WT2)	OK OK	OK OK	OK OK	OK OK	OK OK	OK UNKN	OK OK	OK OK	-

DYNES and USATLAS

- As noted DYNES has begun deployment. (OSG talk Tues)
- For USATLAS we have 13 sites:
  - Boston, Chicago, Harvard, Indiana, Illinois, Michigan, Oklahoma, U Penn, SMU, Tufts, UTD, UTA and UC Santa Cruz
- DYNES will deploy in 3 phases (~10-12 sites per phase)
  - Phase 1 – April to July
  - Phase 2 – August to October
  - Phase 3 – October to December
- After Phase 3 completes there will be a push to fill out our DYNES site list. For USATLAS: SLAC/WT2 and MSU?

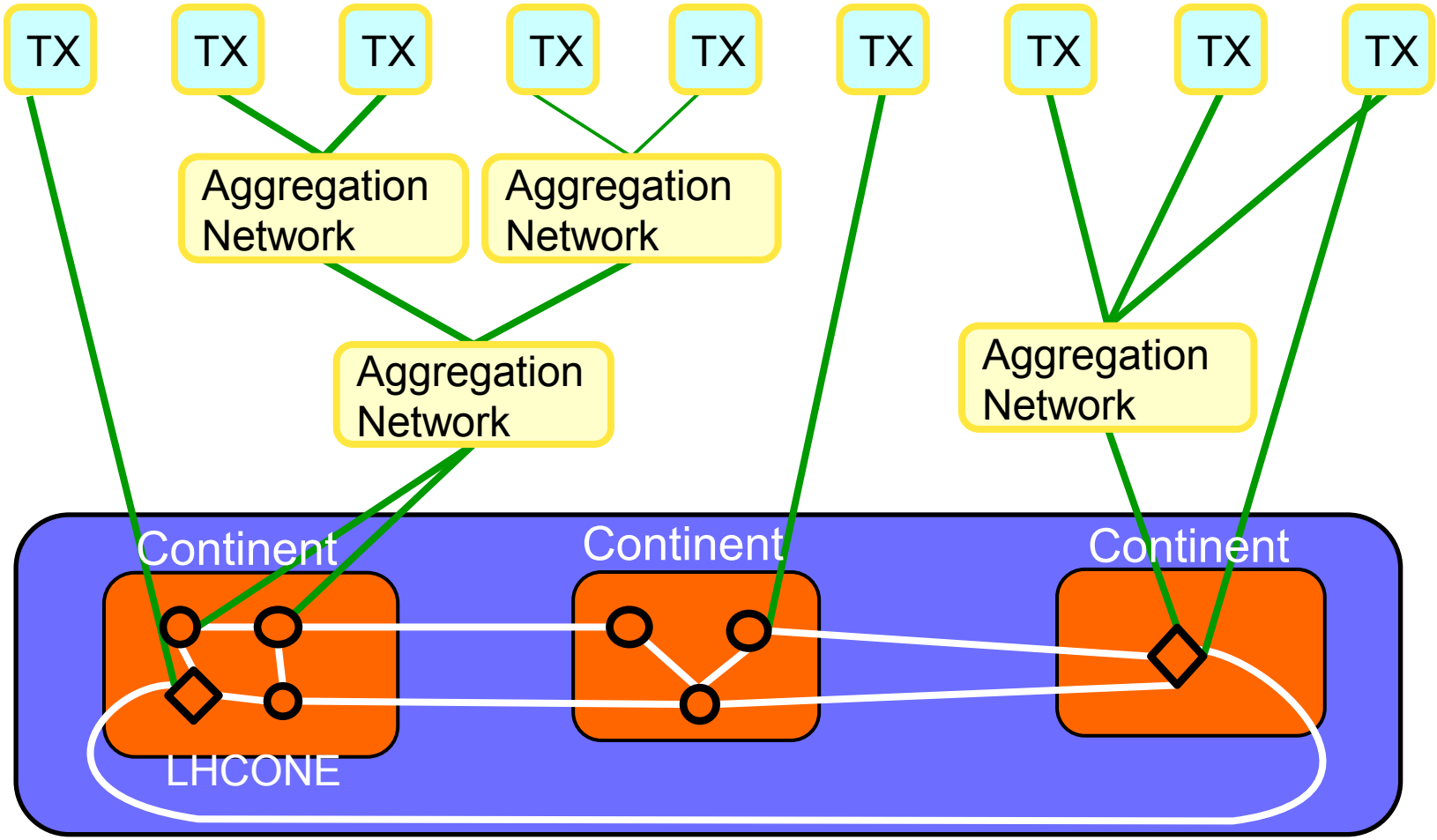
LHCONE Proposal/Project

- LHCONE - **LHC O**pen **N**etwork **E**nvironment
- Results of LHC Tier-2 network working group convened summer 2010. A merger of 4 “whitepapers” from the CERN LHCT2 meeting in January 2011
- LHCONE builds on the hybrid network infrastructures and open exchange points provided today by the major R&E networks on all continents
- Goal: To build a global unified service platform for the LHC community
- By design, LHCONE makes best use of the technologies and best current practices and facilities provided today in national, regional and international R&E networks

LHCONE 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

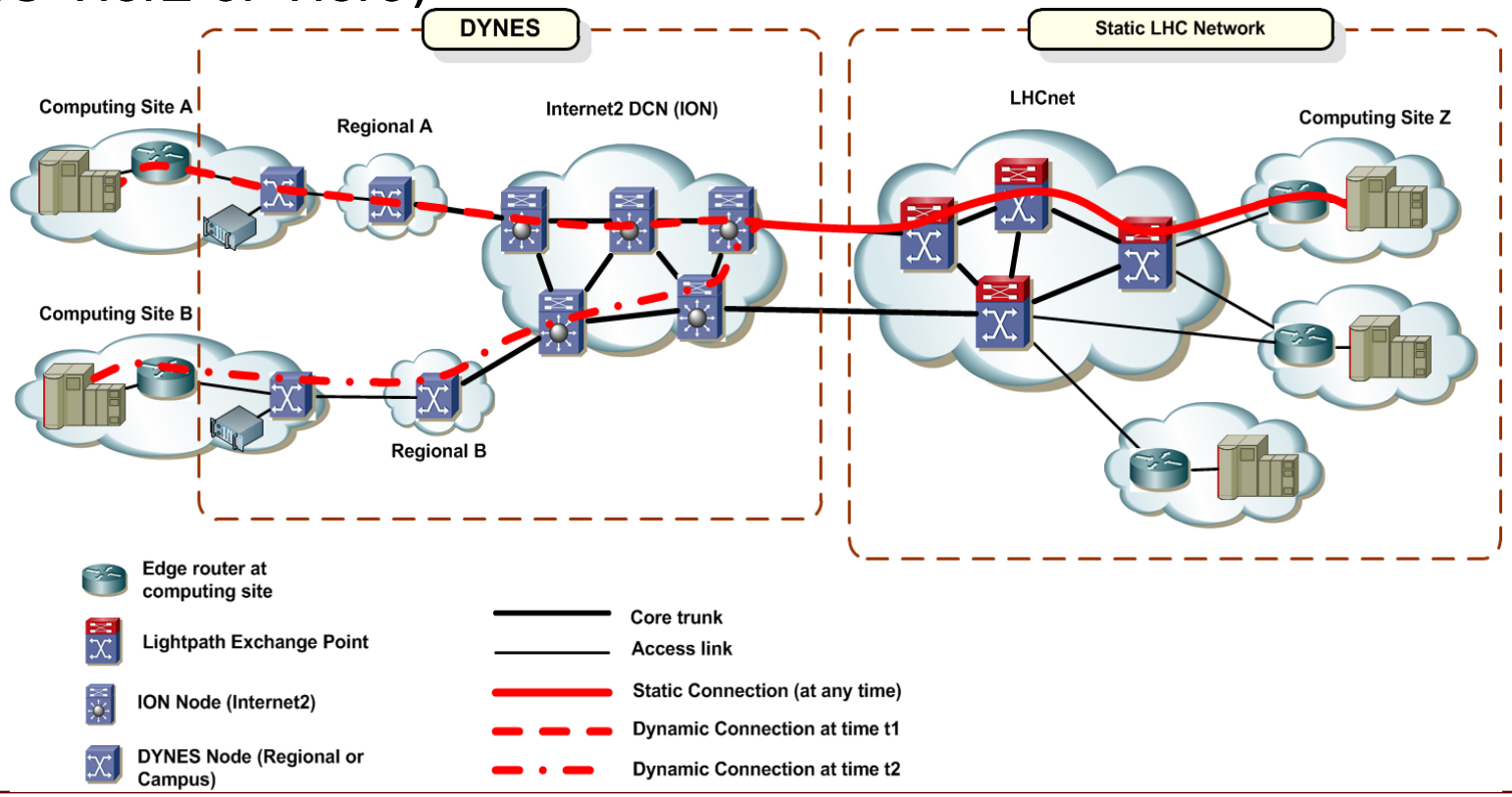
LHCONE High-level Architecture, Example



Single node Exchange Point Distributed Exchange Point

How Can the DYNES Project be Leveraged in LHCONE?

- The Internet2 ION service currently has end-points at two GOLEs in the US: MANLAN & StarLight; + I2 "Distributed OEP" Proposed
- 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)



Near-term Steps for Working Group

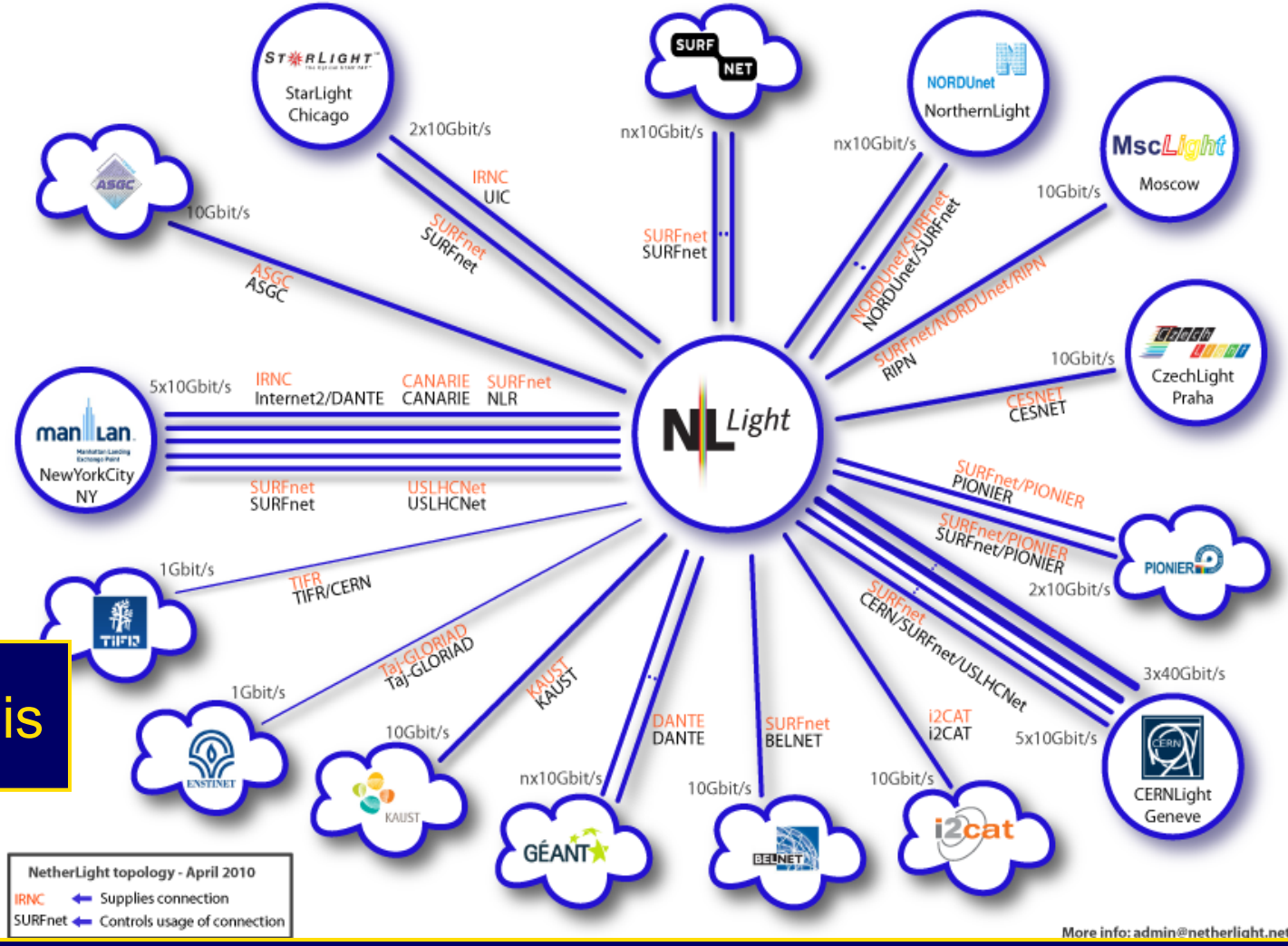
- ‘Green’ latency and throughput matrices in Nagios
- Re-validate site throughputs
- Begin active alerting based upon perfSONAR results
  - Lots of work here. How do we get a system capable of alerting when there are real issues but not also generating lots of false positives?
- Get baseline values between sites for future reference
- Explore some network related tunings/settings are our sites
  - TCP congestion protocol? HTCP, BIC ?
  - Is autotuning active for our DDM transfers (SRM/FTS/Gridftp/OS all potentially involved)
- Plan for DYNES and LHCONE…start discussion here!

Questions / Discussion?

Starting Discussion Questions

- How can we get needed network monitoring ATLAS(LHC)-wide? (How best to prepare for our “new” DDM model?)
- How best to utilize the perfSONAR data we are now capturing?
  - **Goals:** Alert on problems, Set baseline expectations, Quickly localize problems, Differentiate end-site vs network issues, others?
- What preparation-for and participation-in LHCONE is needed from USATLAS?

Open Exchange Points: NetherLight Example
3 x 40G, 30+ 10G Lambdas, Use of Dark Fiber

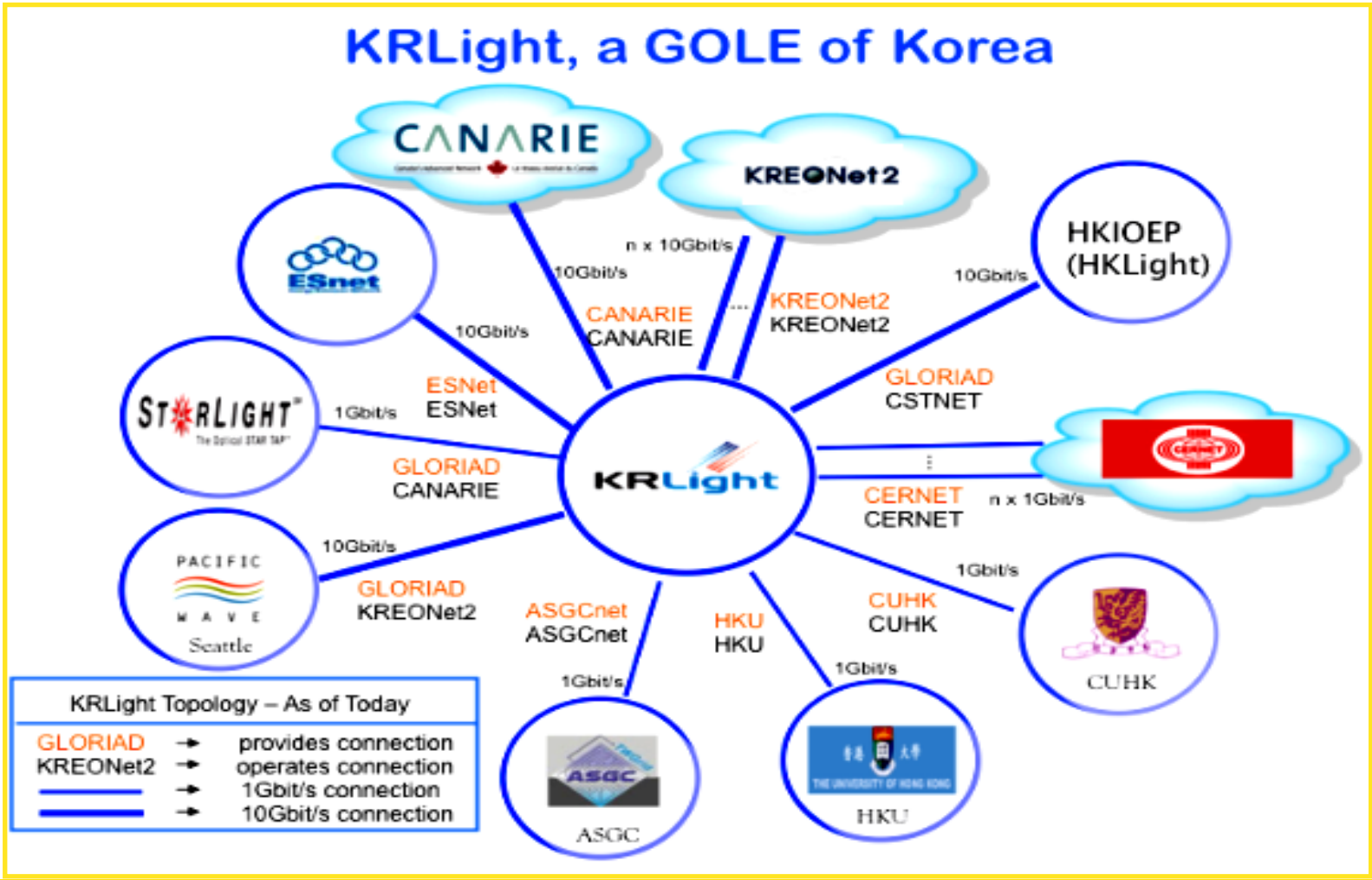


www.glif.is

Convergence of Many Partners on Common Lightpath Concepts

Internet2, ESnet, GEANT, USLHCNet; nl, cz, ru, be, pl, es, tw, kr, hk, in, nordic

Open Exchange Points: KRLight in Korea



Convergence of Many Partners on Common Lightpath Concepts

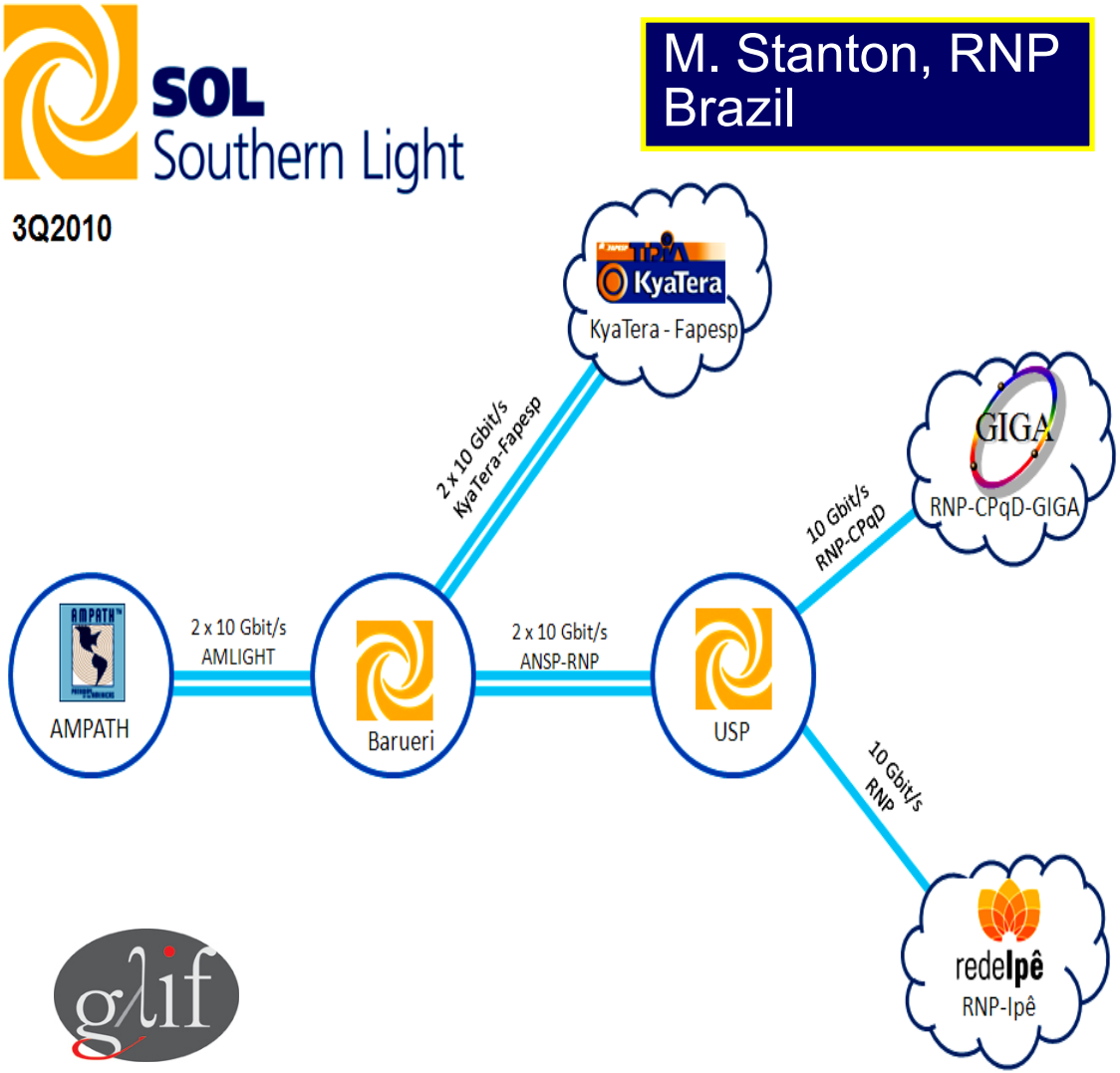
Here kr, hk, tw, cn, ca, us, Gloriad

SouthernLight

Latin American Open Exc

Figure shows the current config-uration of the SouthernLight GOLE.

M. Stanton, RNP Brazil



Additional GLIF resources include the multigigabit core of the Ipê network,
to be greatly extended in early 2011,
the experimental GIGA network, operated jointly by RNP and CPqD,
and the KyaTera network in São Paulo state.

RNP has committed itself to demon-strate an interoperable dynamic circuit service, and it is planned to deploy an experimental service in the next upgrade of the Ipê network in 2011. Such a facility will greatly enhance RNP's capability to manage the widespread use of end to end circuits. RNP was able to carry out experimental studies and is on

Southern Light is Recognized as a GOLE by

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 + 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 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