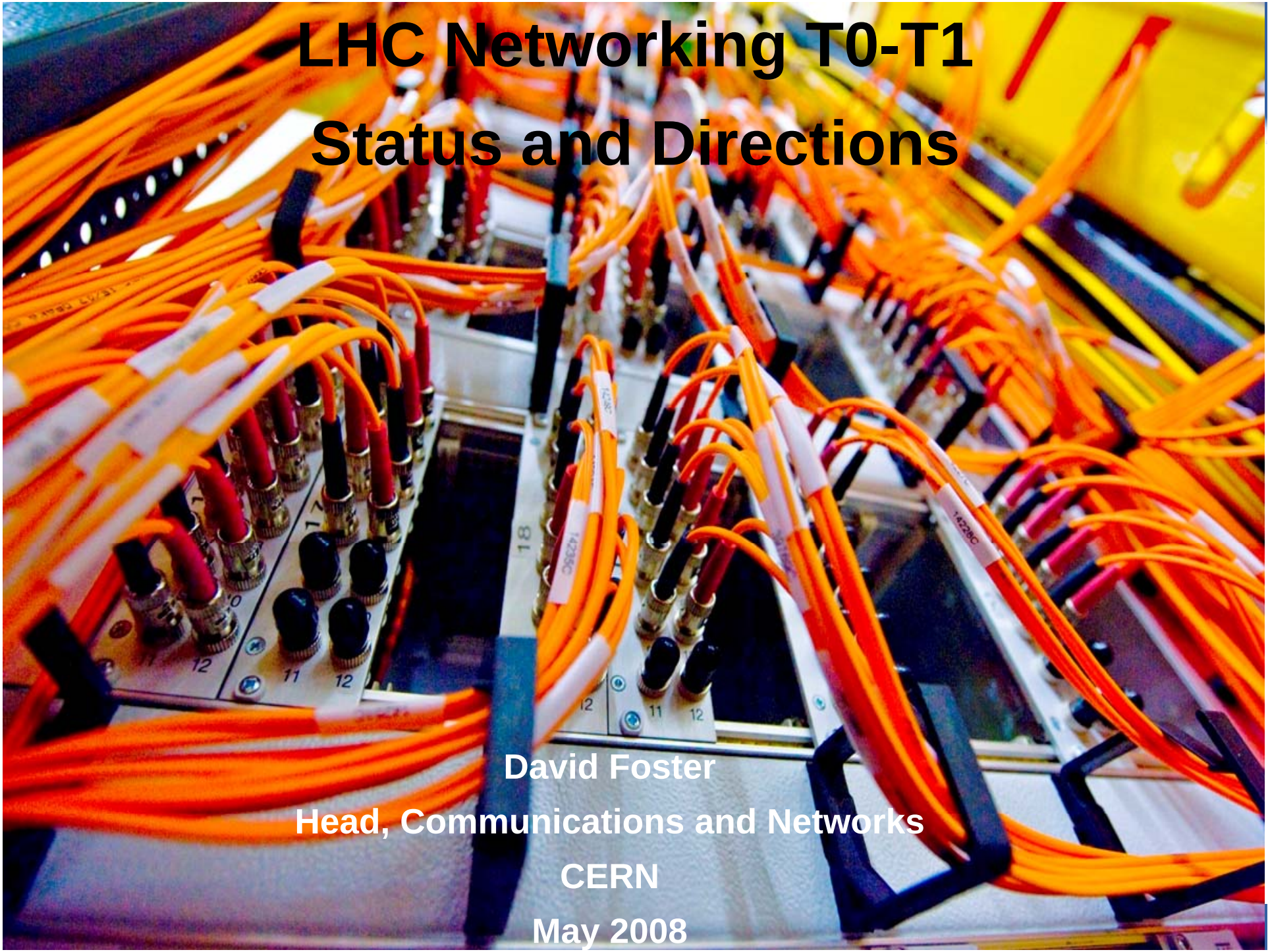
LHC Networking T0-T1
Status and Directions



David Foster

Head, Communications and Networks

CERN

May 2008

Acknowledgments

- ✓ Many presentations and material in the public domain have contributed to this presentation

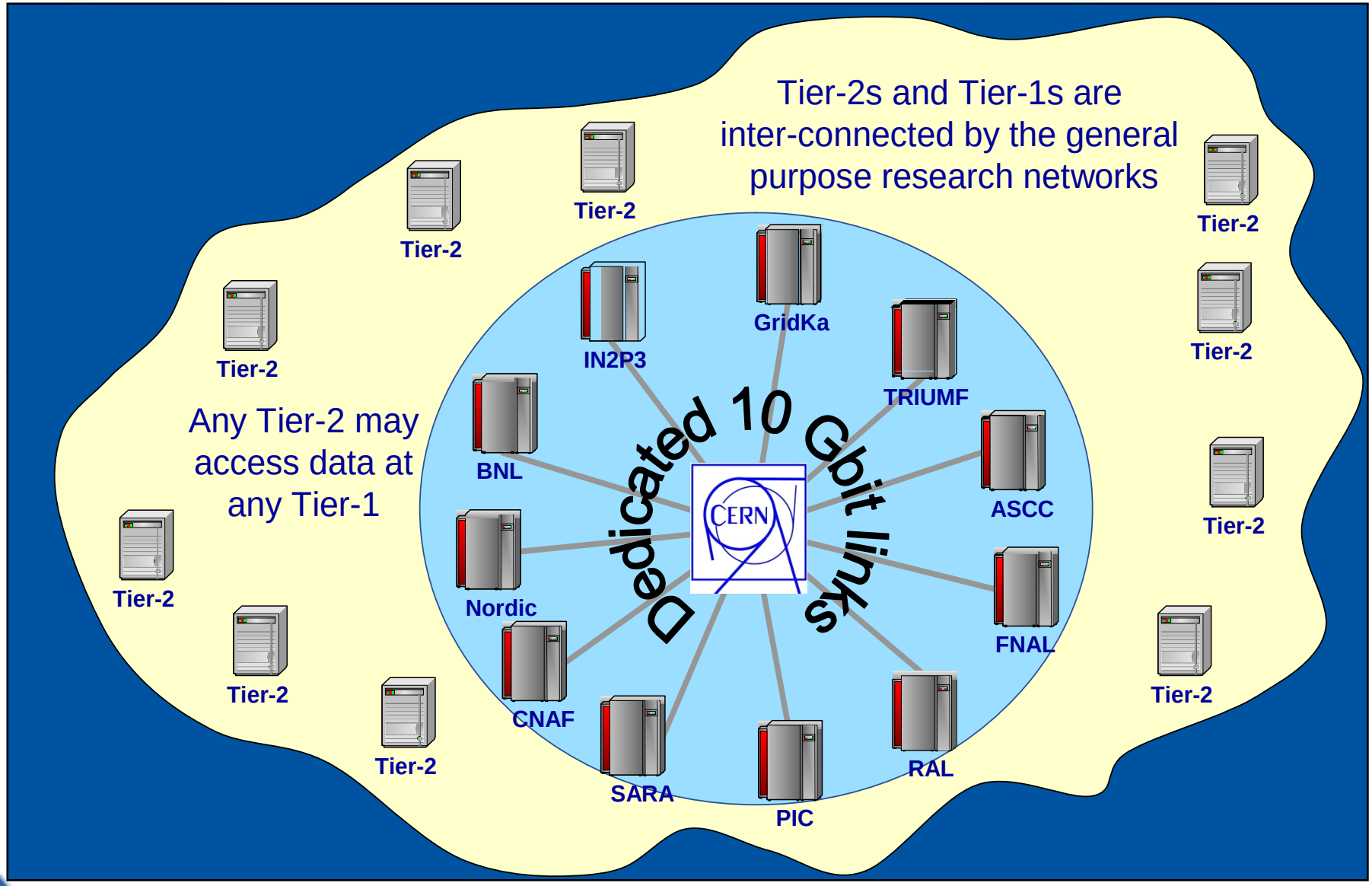
Over-provisioned packet networks are useful

Packet interference can be a problem

The Beginning ...

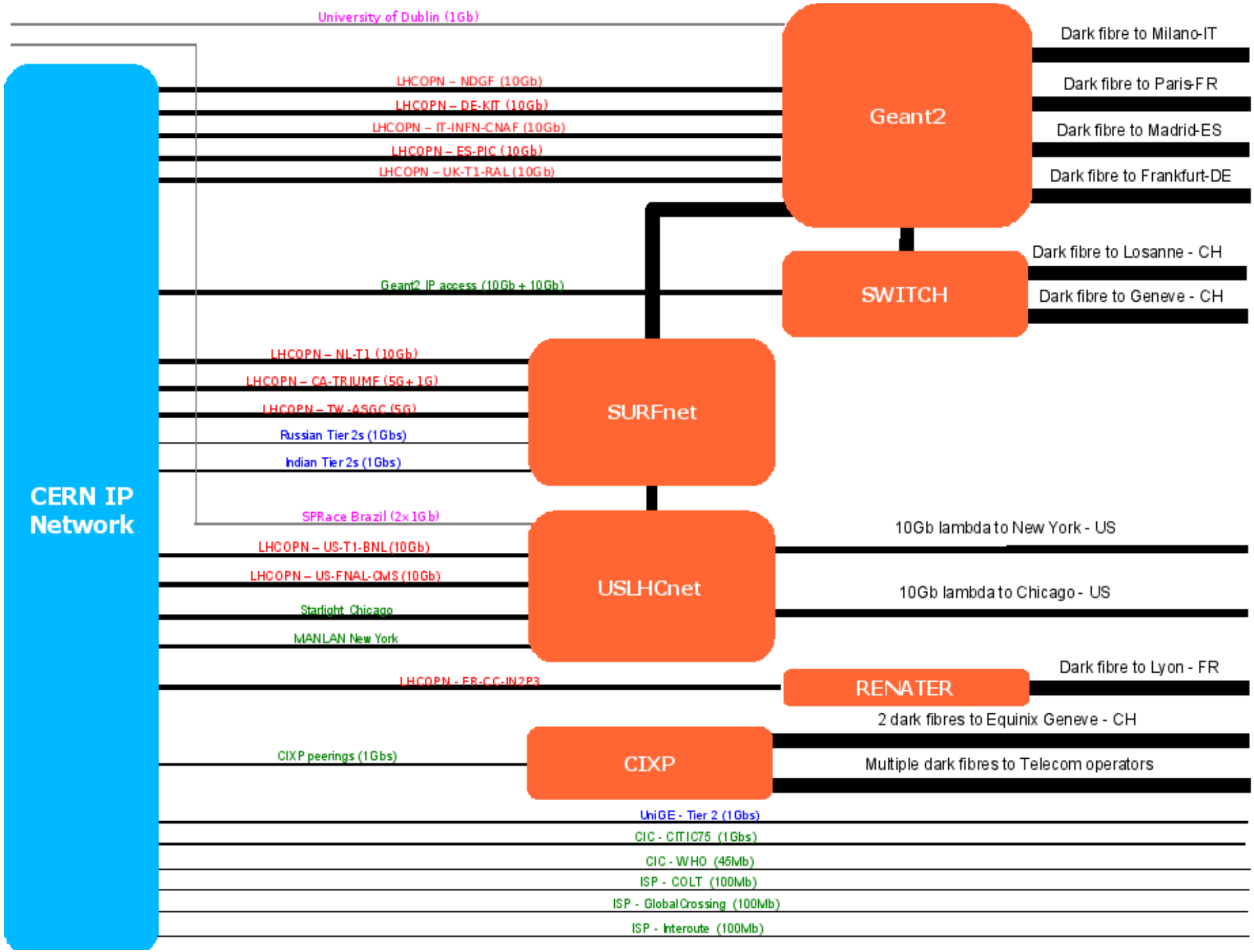
- Essential for Grid functioning to distribute data out to the T1's.
  - Capacity must be large enough to deal with most situation including "Catch up"
- OPN conceived in 2004 as a "Community Network"
  - Renamed as "Optical Private Network" as a more descriptive name.
  - Based on 10G as the best choice for affordable adequate connectivity by 2008.
    - 10G is (almost) commodity now!
    - Considered by some as too conservative - Can fill a 10G pipe with just (a few) pc's!
- Simple end-end model
  - This is not a research project, but, an evolving production network relying on emerging facilities.

LHCOPN Architecture 2004 Starting Point

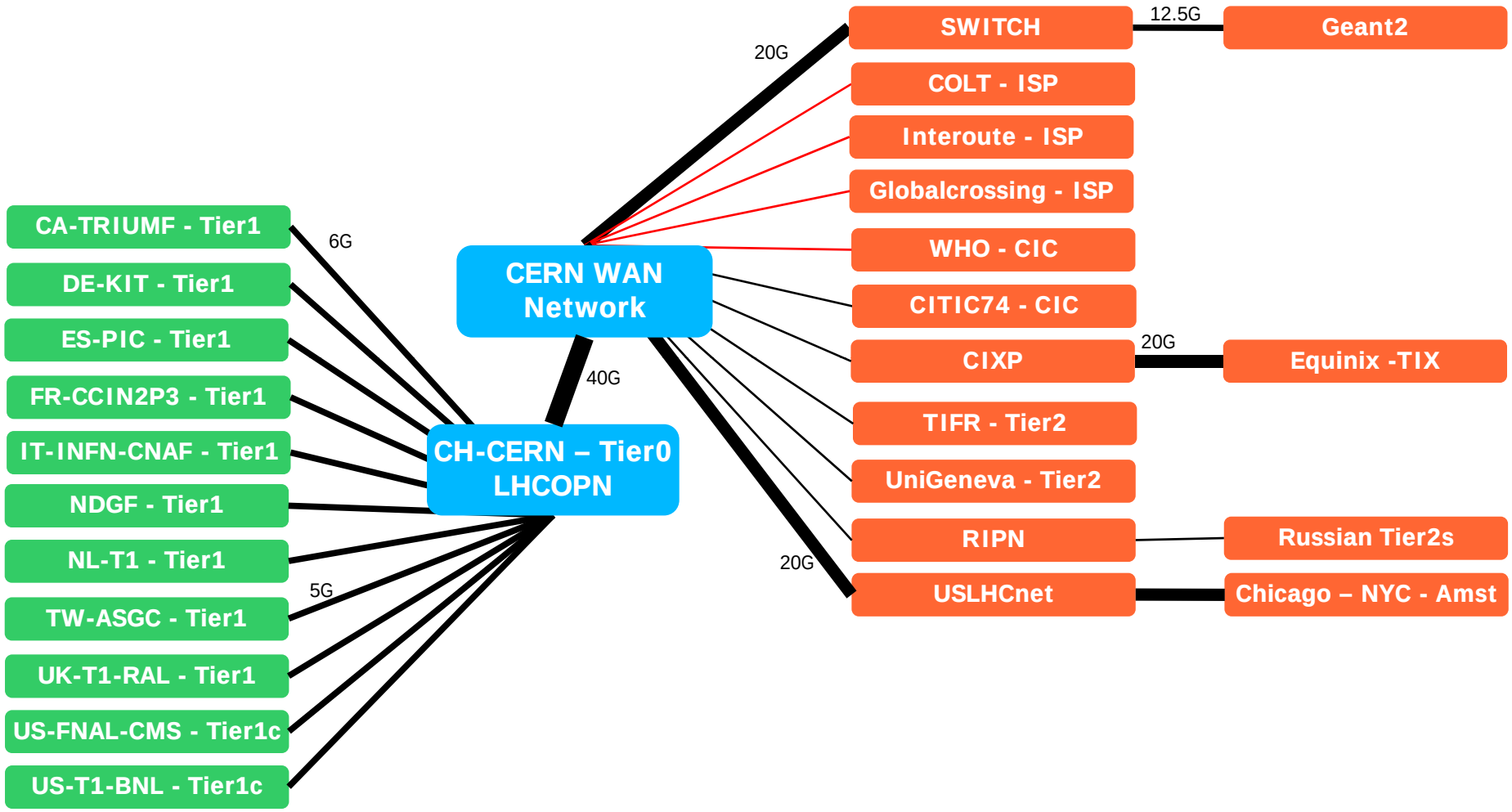


Hybrid Networking Model

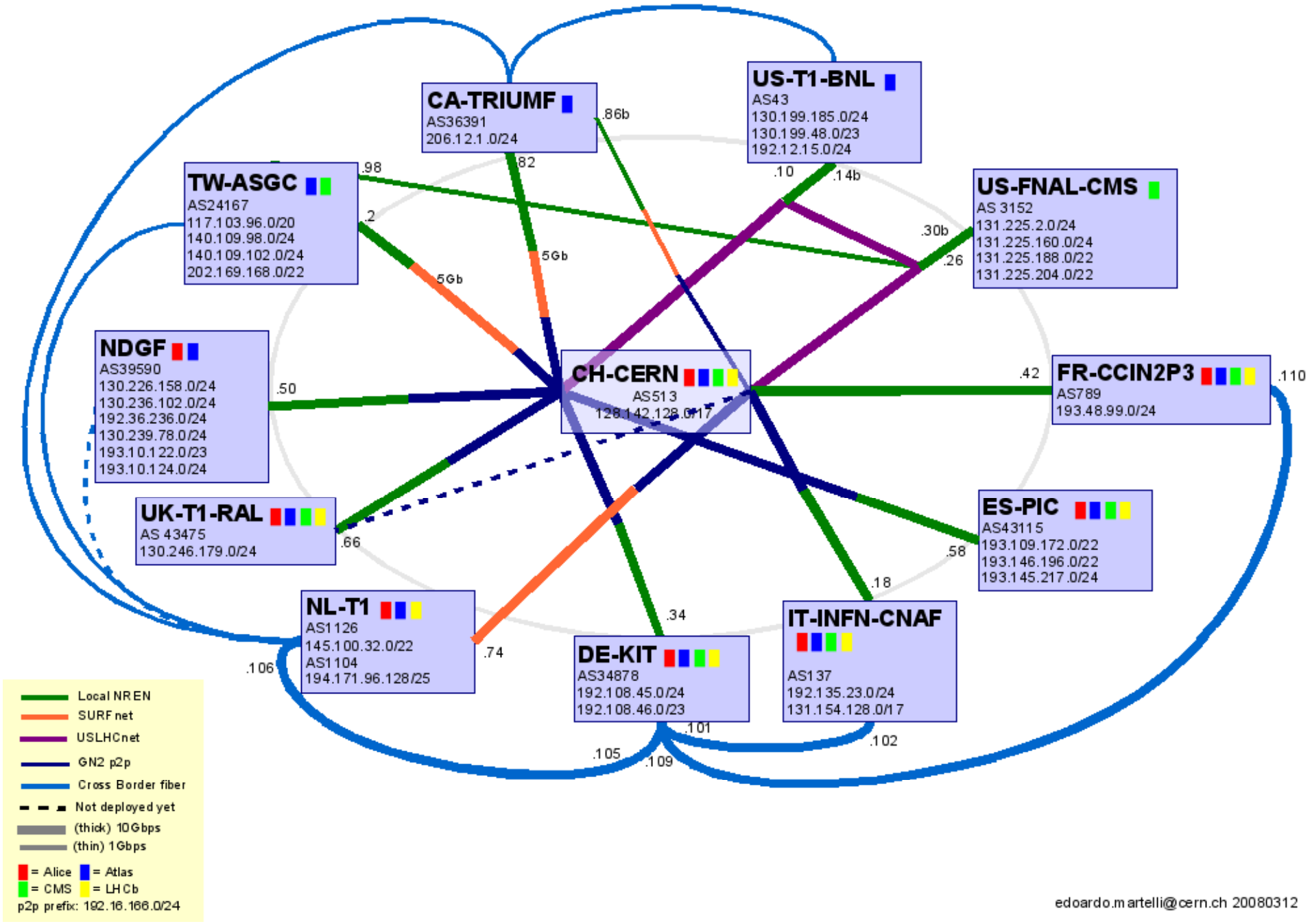
- Infrastructure is provided by a number of initiatives:
  - GEANT-2
  - Commercial Links
  - Coordinated Infrastructures (USLHCNet, GLIF)
  - NRENS + Research Networks (ESNet, I2, Canarie etc)
- Operated by the community
  - “Closed Club” of participants
  - Routers at the end points
  - Federated operational model
- Evolving
  - Cross Border Fiber links playing an important role in resiliency.


University of Dublin (1Gb)
CERN IP Network
LHCOPN – NDGF (10Gb)
LHCOPN – DE-KIT (10Gb)
LHCOPN – IT-INFN-CNAF (10Gb)
LHCOPN – ES-PIC (10Gb)
LHCOPN – UK-T1-RAL (10Gb)
Geant2
Dark fibre to Milano-IT
Dark fibre to Paris-FR
Dark fibre to Madrid-ES
Dark fibre to Frankfurt-DE
Geant2 IP access (10Gb + 10Gb)
SWITCH
Dark fibre to Losanne - CH
Dark fibre to Geneve - CH
LHCOPN – NL-T1 (10Gb)
LHCOPN – CA-TRIUMF (5G+1G)
LHCOPN – TW-ASGC (5G)
Russian Tier 2s (1Gbs)
Indian Tier 2s (1Gbs)
SURFnet
SPRace Brazil (2x1Gb)
LHCOPN – US-T1-BNL (10Gb)
LHCOPN – US-FNAL-CMS (10Gb)
Starlight Chicago
MANLAN New York
USLHCnet
10Gb lambda to New York - US
10Gb lambda to Chicago - US
LHCOPN - FR-CC-IN2P3
RENATER
Dark fibre to Lyon - FR
2 dark fibres to Equinix Geneve - CH
CIXP peerings (1Gbs)
CIXP
Multiple dark fibres to Telecom operators
UniGE - Tier 2 (1Gbs)
CIC - CITIC75 (1Gbs)
CIC - WHO (45Mb)
ISP - COLT (100Mb)
ISP - GlobalCrossing (100Mb)
ISP - Interoute (100Mb)

CERN IP connectivity



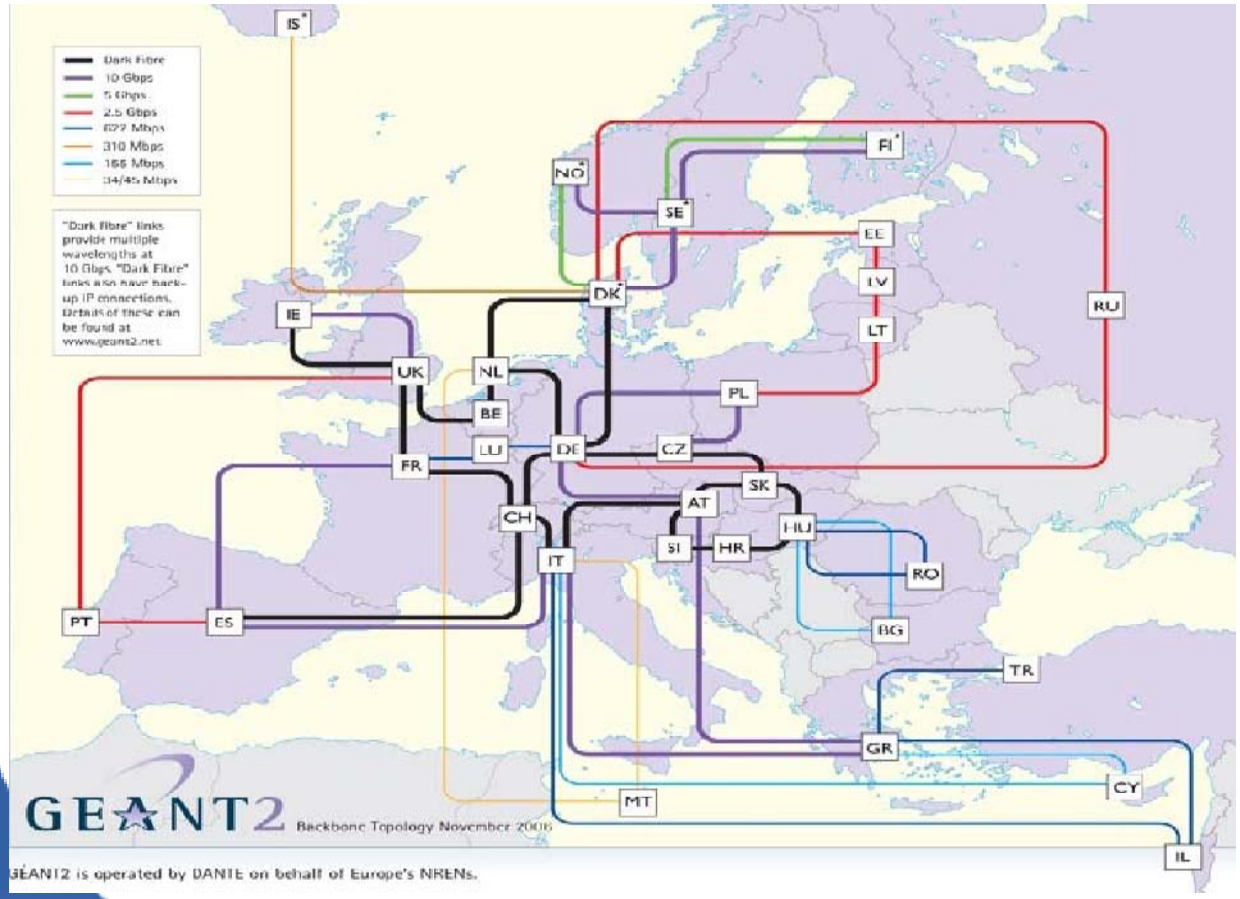
LHCOPN – current status



edoardo.martelli@cern.ch 20080312

GÉANT2:
Consortium of 34 NRENs

22 PoPs, ~200 Sites
38k km Leased Services, 12k km Dark Fiber
Supporting Light Paths for *LHC, eVLBI*, et al.



Dark Fiber Core Among
16 Countries:

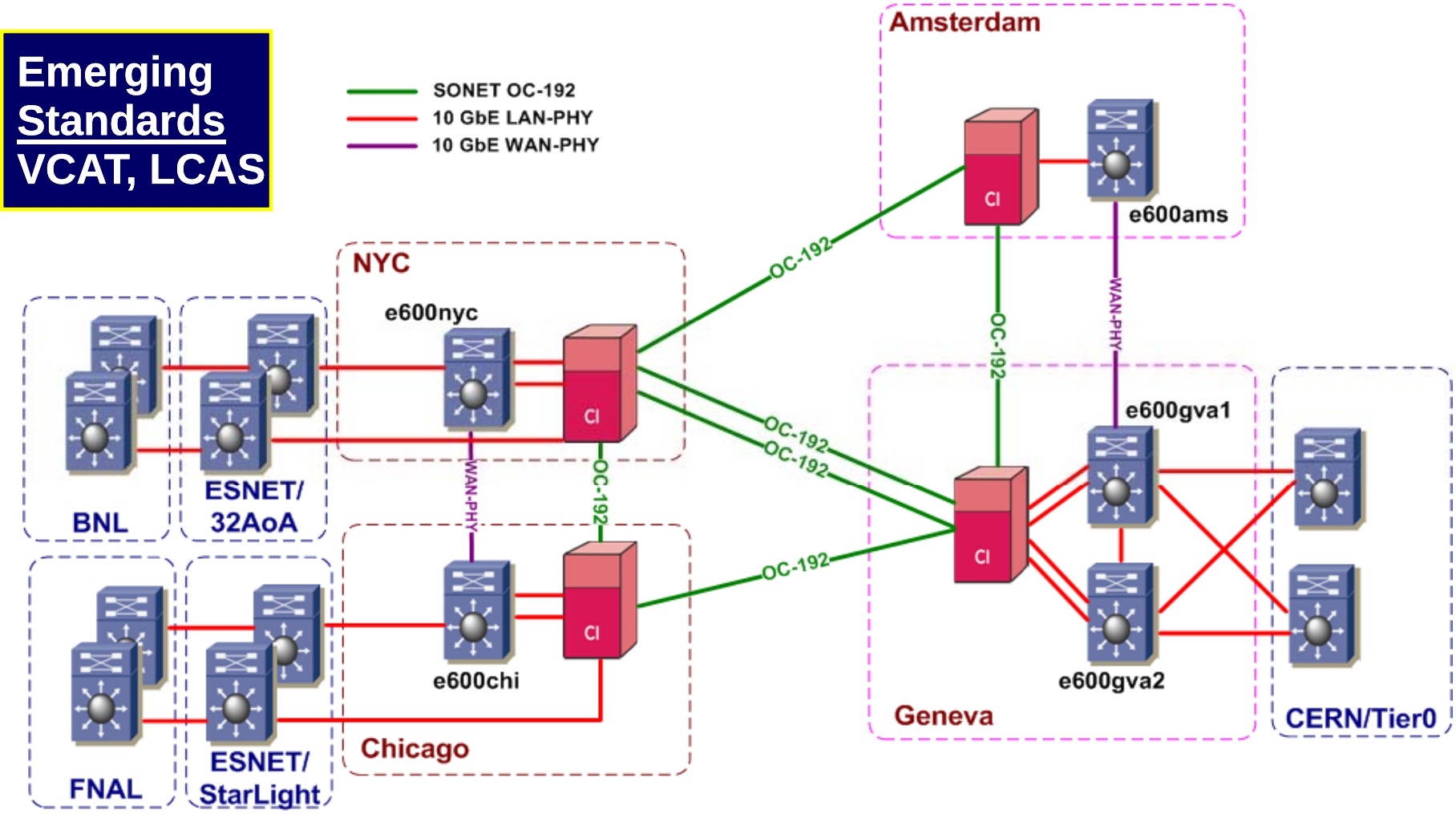
- Austria
- Belgium
- Bosnia-Herzegovina
- Czech Republic
- Denmark
- France
- Germany
- Hungary
- Ireland
- Italy,
- Netherland
- Slovakia
- Slovenia
- Spain
- Switzerland
- United Kingdom

Multi-Wavelength Core (to 40λ) + 0.6-10G Loops

H. Doebbeling

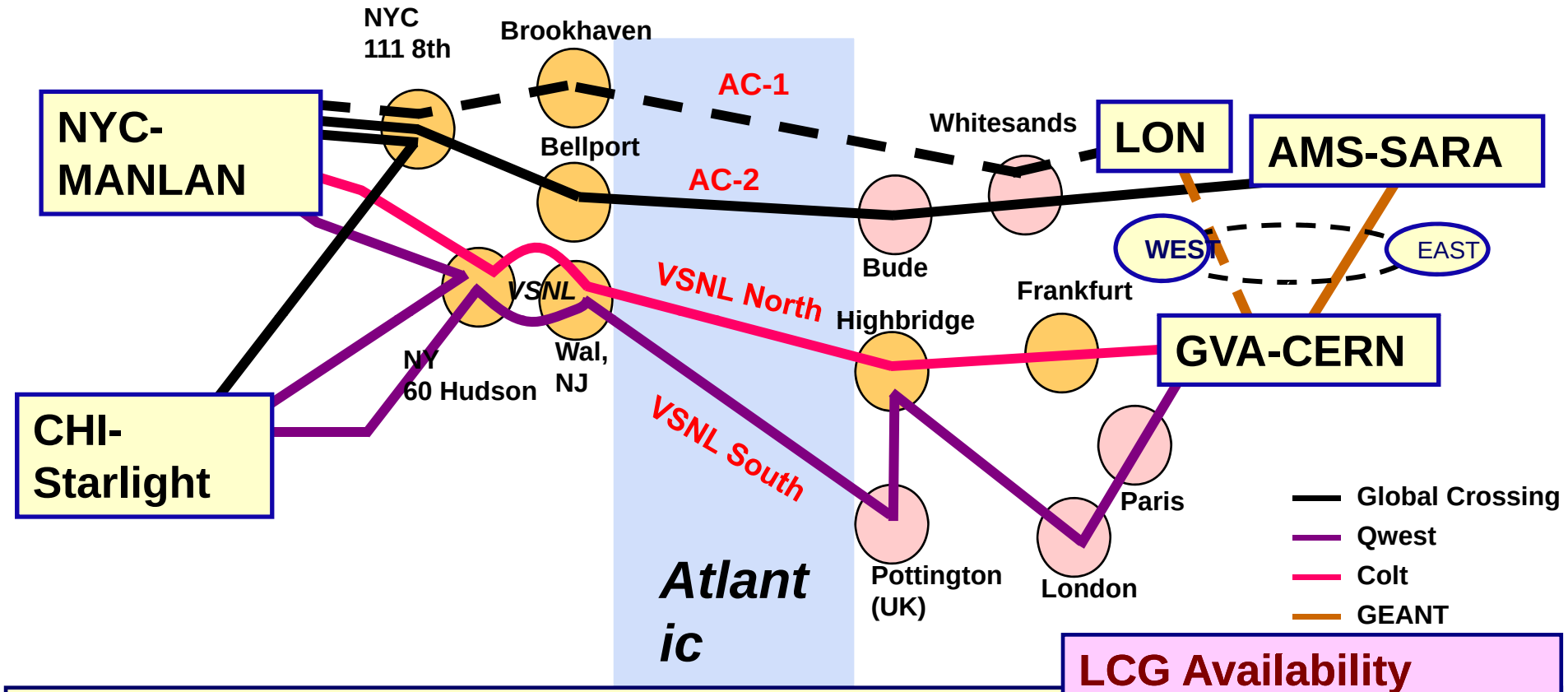
USLHCNet Planned Configuration for LHC Startup

Emerging Standards
VCAT, LCAS



Robust fallback at layer 1 + next-generation hybrid optical network: *Dynamic* circuit-oriented network services with BW guarantees

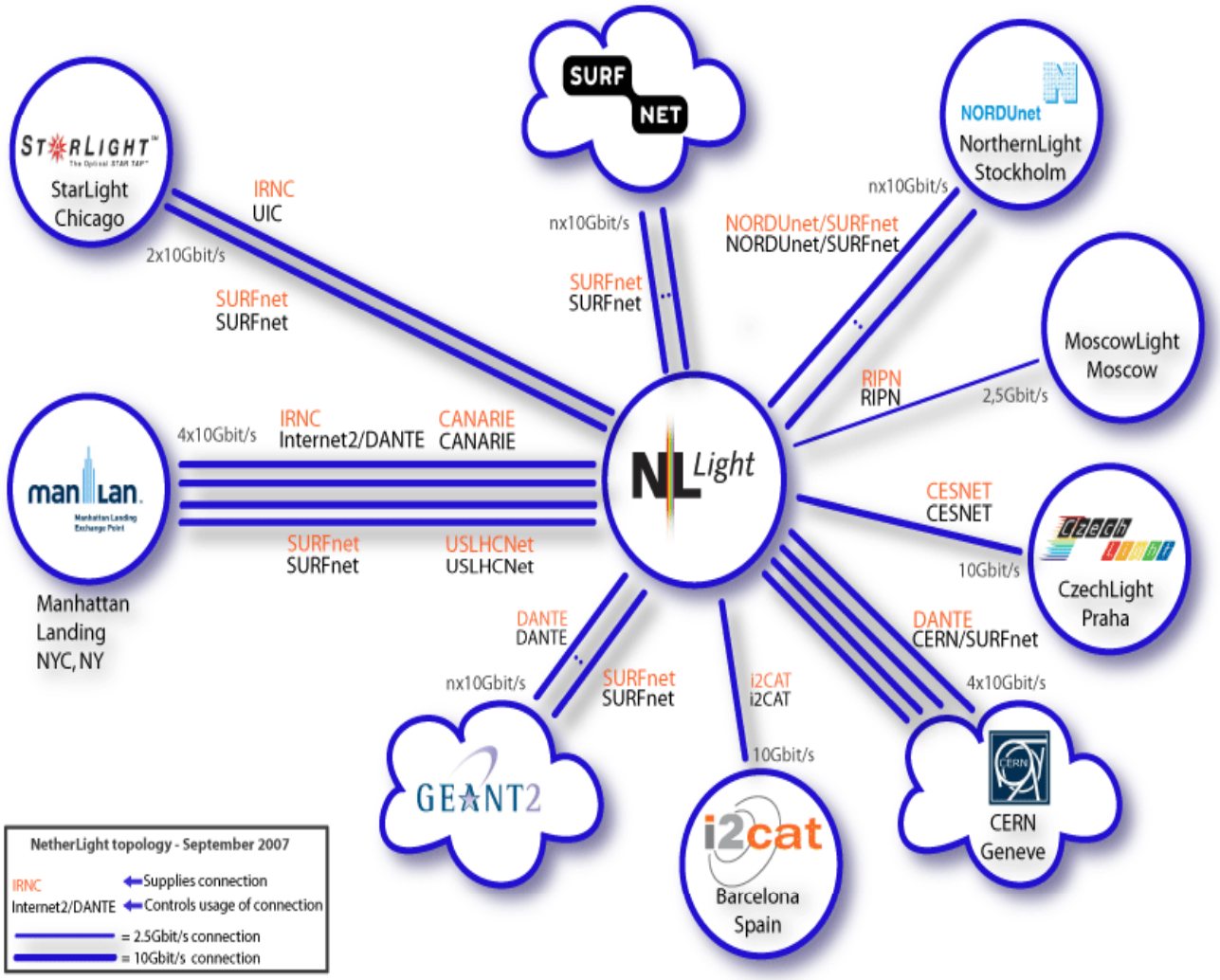
US LHCNet in 2008: Increased Reliability



LCG Availability requirement: 99.95%

- New tender process completed in October
- We were able to improve on the pricing, path diversity and SLAs
- GC NYC-LON circuit will be cross-connected to the GEANT LON-GVA circuit to make a NYC-GVA circuit

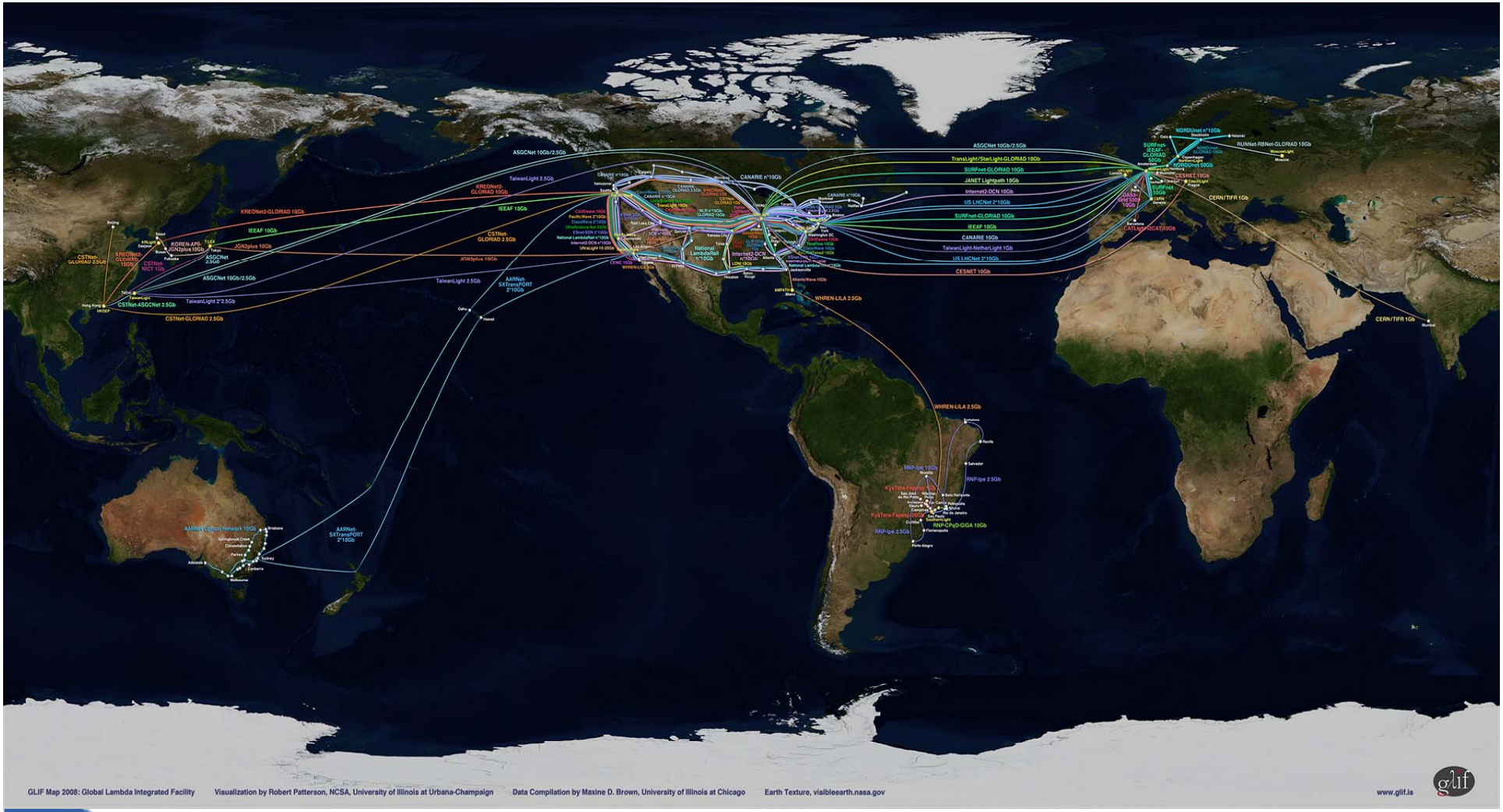
GLIF Open Lambda Exchanges (GOLE)



- AMPATH - Miami
- CERN/Caltech – Geneva+U.S.
- CzechLight - Prague
- HKOEP - Hong Kong
- KRLight - Daejoen
- MAN LAN - New York
- MoscowLight - Moscow
- NetherLight - Amsterdam
- NGIX-East – Wash. D.C.
- NorthernLight - Stockholm
- Pacific Wave (L.A.)
- Pacific Wave (Seattle) - Pacific Wave (Sunnyvale)
- StarLight - Chicago
- T-LEX - Tokyo
- UKLight - London

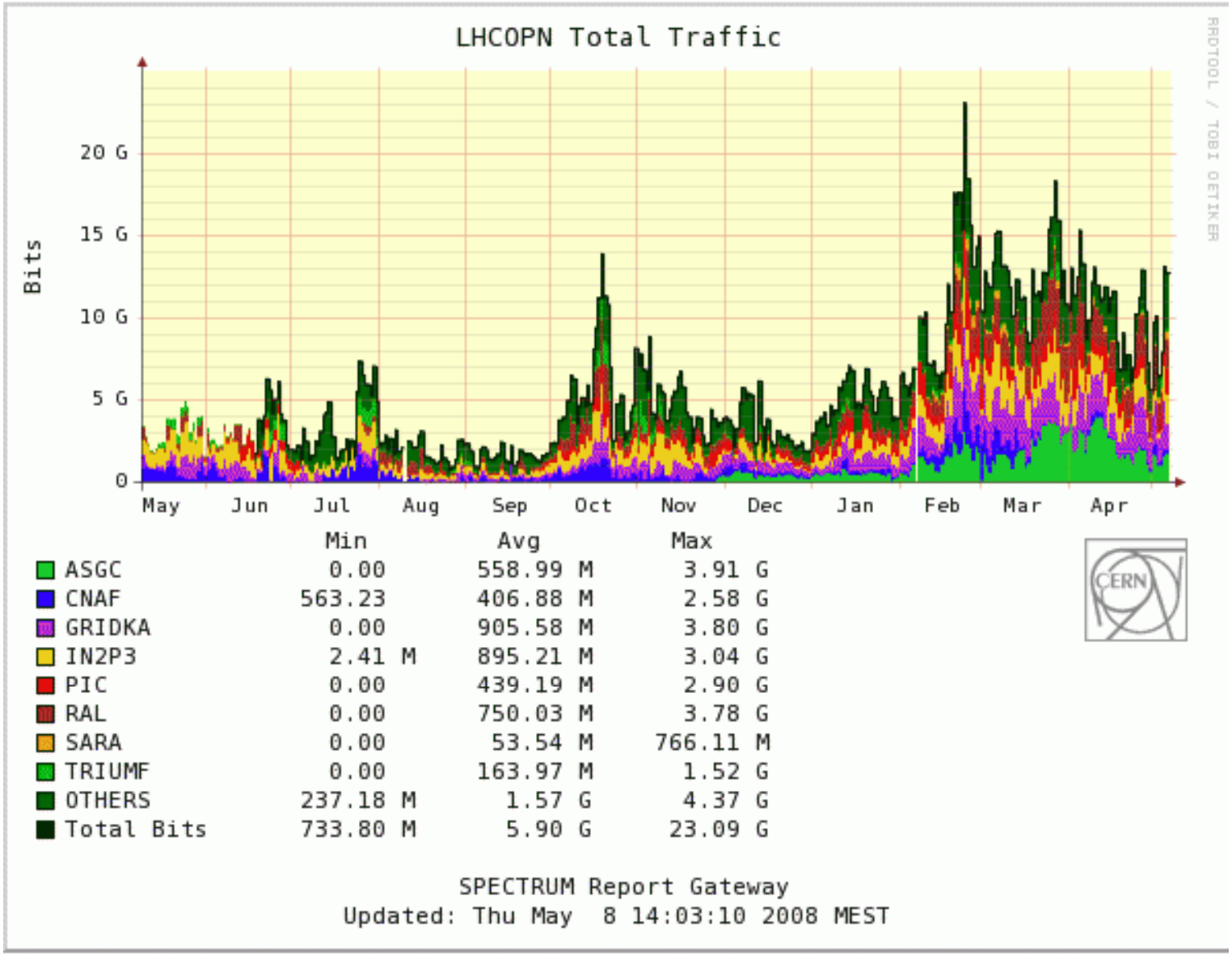
14

Global Lambda Integrated Facility
World Map – May 2008



Visualization courtesy of Bob Patterson, NCSA/University of Illinois at Urbana-Champaign.
Data compilation by Maxine Brown, University of Illinois at Chicago. Earth texture from NASA.

Traffic Statistics



	Min	Avg	Max
ASGC	0.00	558.99 M	3.91 G
CNAF	563.23	406.88 M	2.58 G
GRIDKA	0.00	905.58 M	3.80 G
IN2P3	2.41 M	895.21 M	3.04 G
PIC	0.00	439.19 M	2.90 G
RAL	0.00	750.03 M	3.78 G
SARA	0.00	53.54 M	766.11 M
TRIUMF	0.00	163.97 M	1.52 G
OTHERS	237.18 M	1.57 G	4.37 G
Total Bits	733.80 M	5.90 G	23.09 G

Current Situation

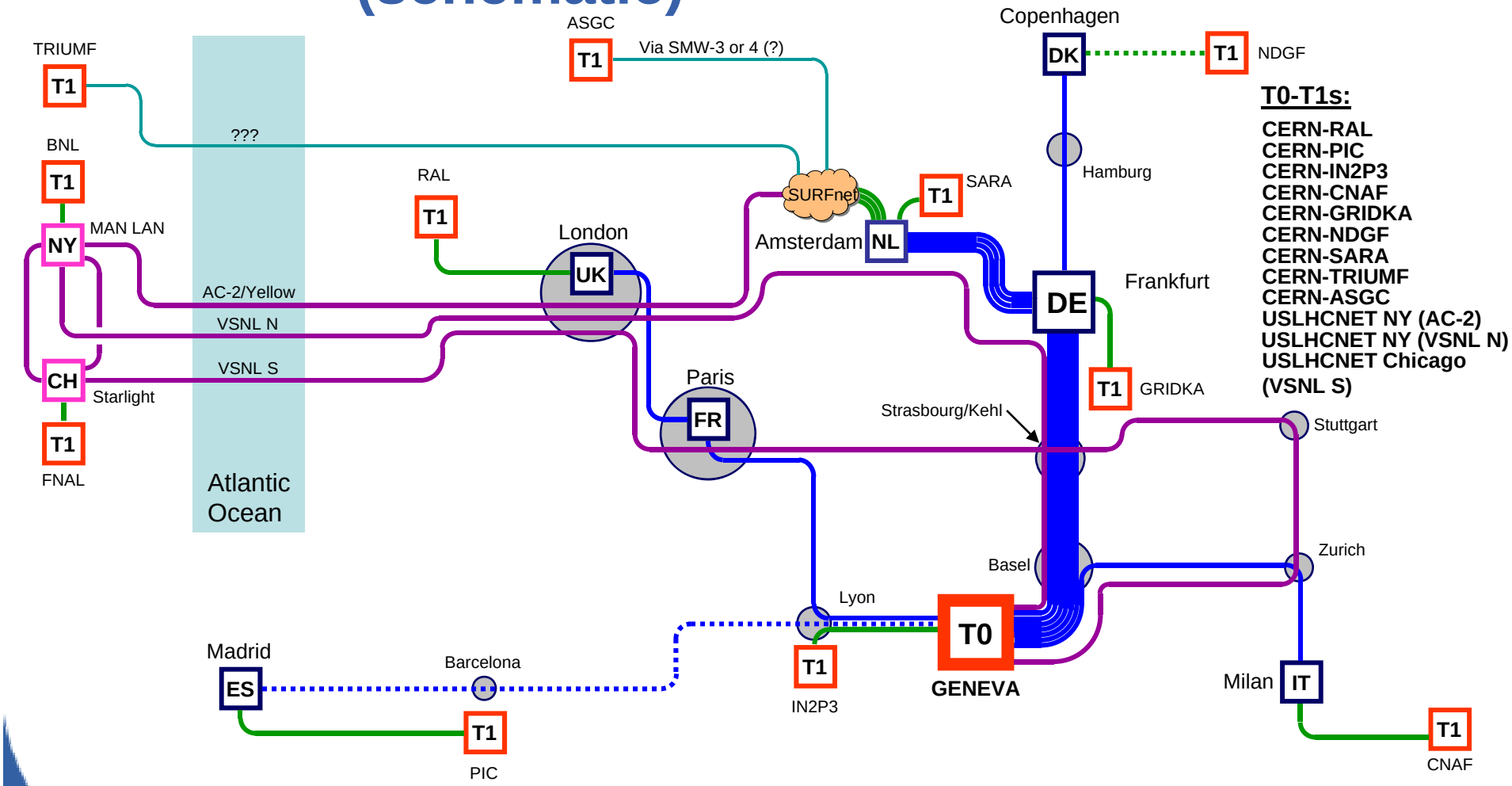
- T0-T1 Network is operational and stable.
  - But, **“The first principle is that you must not fool yourself, and you're the easiest person to fool.”** Richard Feynman
- Several areas of weakness
  - Physical Path Routing
  - IP Backup
  - Operational Support
  - Monitoring

Physical Paths

- Dante analysed the physical path routing for the OPN links.
- The network had been built over time, taking in each case the most direct (and cheapest!) wavelength path in the GEANT network.
- Analysis showed many common physical paths of fibers and wavelengths.
- Re-routing of some wavelengths has been done.
  - More costly solution (more intervening equipment)
  - especially the path from Amsterdam -> CERN
  - 5x10G on this path.

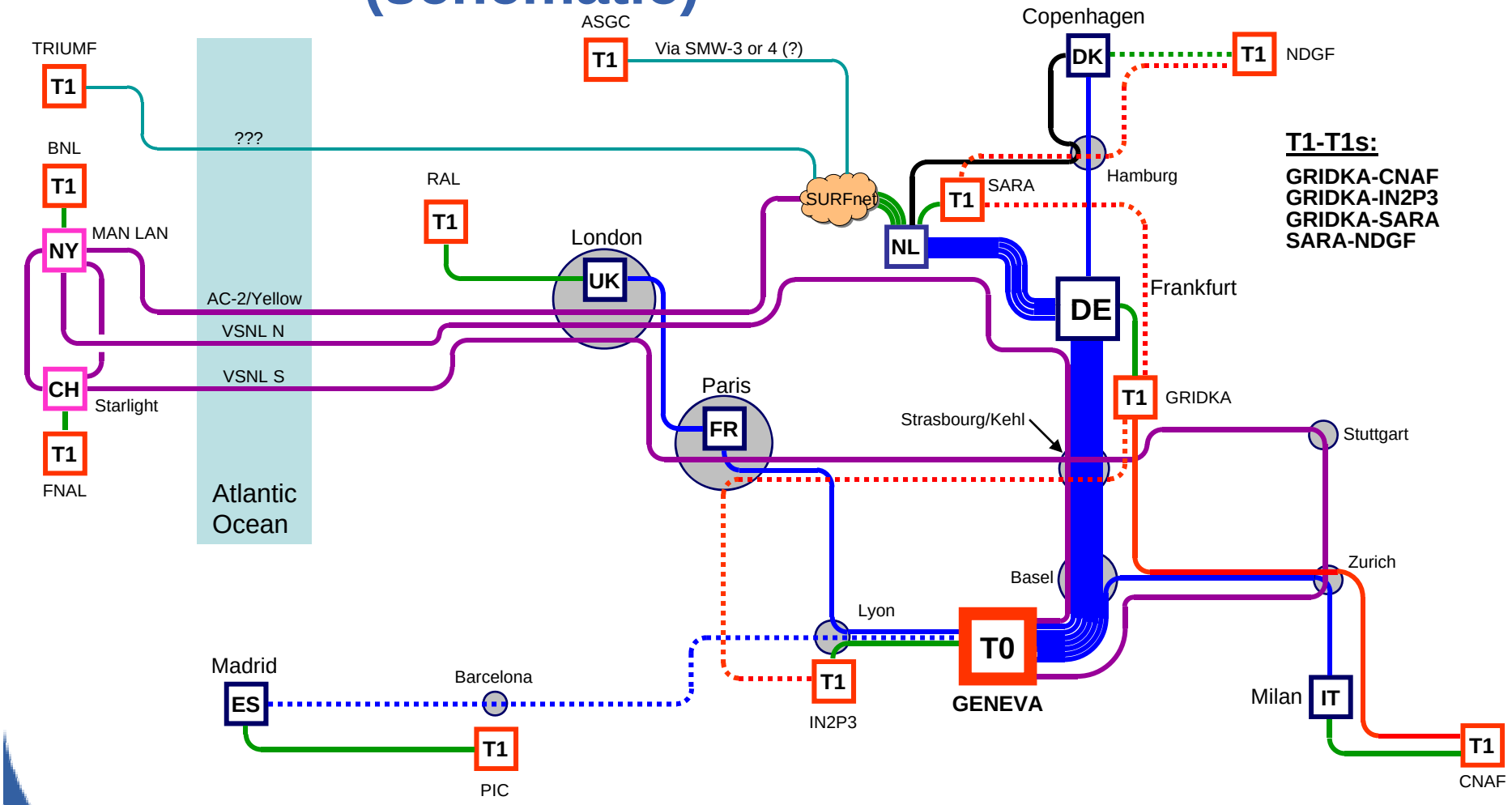
T0-T1 Lambda routing (schematic)

Connect. Communicate. Collaborate



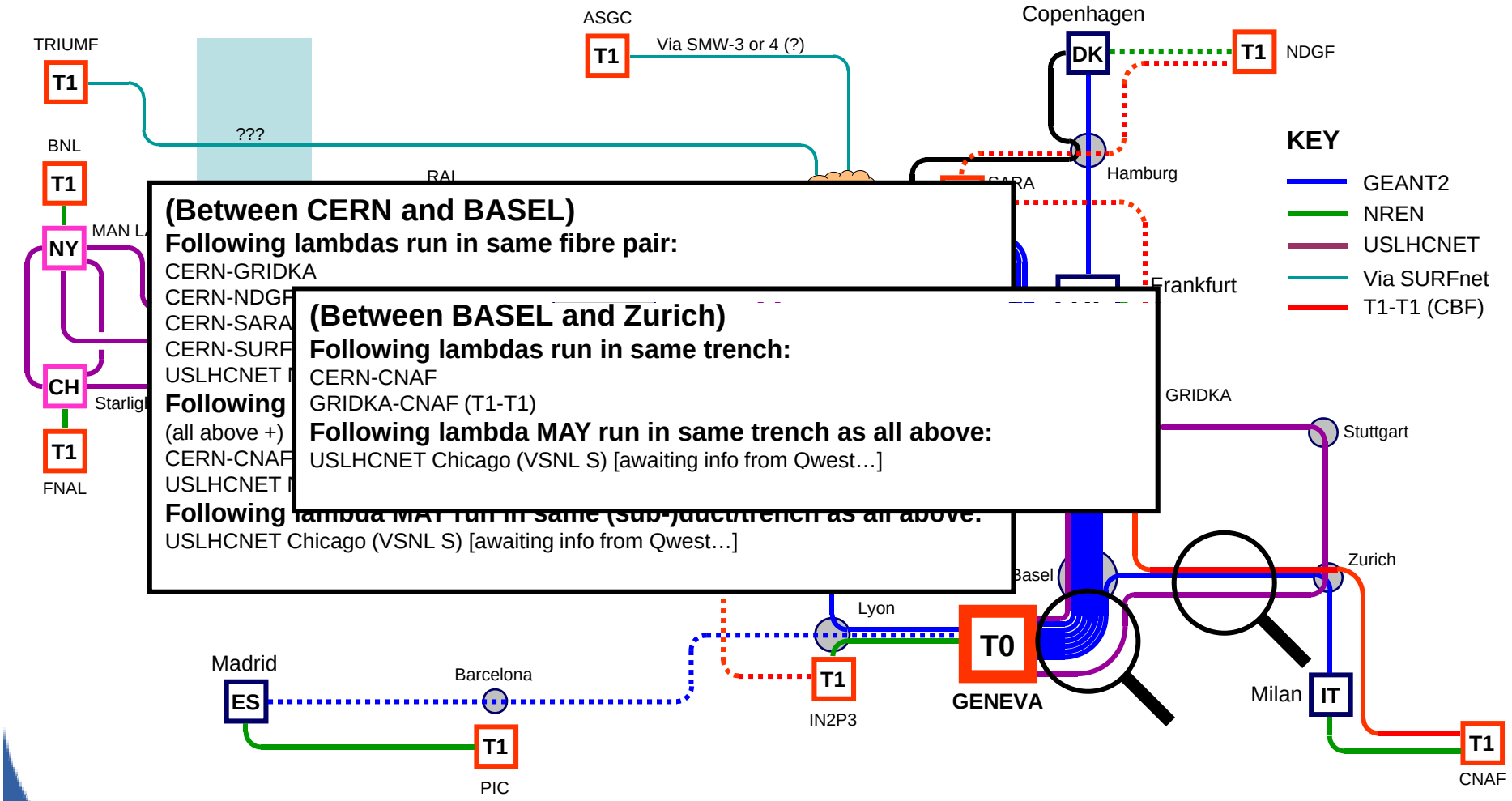
T1-T1 Lambda routing (schematic)

Connect. Communicate. Collaborate



Some Initial Observations

Connect. Communicate. Collaborate



IP Backup

- In case of failures, degraded service may be expected.
  - This is not yet quantified on a “per failure” basis.
- The IP configuration needs to be validated
  - Some failures have indeed produced successful failover.
  - Tests executed this month (9th April)
- Some sites still have no physical backup paths
  - PIC (difficult) and RAL (some possibilities)

Structured Backup Tests
9th April

Site	Primary down		CERN-R01 down		CERN-R02 down	
CA-TRIUMF	via backup	[1G]	via backup	[1G]	OK	[10G]
DE-KIT	via NL-T1 (asym risk)	[10G/2]	via NL-T1 (asym risk)	[10G/2]	OK	[10G/4 (10G/3)]
ES-PIC	unreachable	[0G]	unreachable	[0G]	OK	[10G]
FR-IN2P3	via DE-KIT	[10G/2]	OK	[10G/2]	via DE-KIT	[10G/4 (10G/3)]
IT-INFN-CNAF	via DE-KIT	[10G/2]	OK	[10G/2]	via DE-KIT	[10G/4 (10G/3)]
NDGF	via GN2-IP (will use NL-T1)	[(10G/2)]	n.a. (will use NL-T1)	[(10G/2)]	n.a. (will be OK)	[(10G/2)]
NL-T1	via DE-KIT	[10G/2]	OK	[10G/2]	via DE-KIT	[10G/4 (10G/2)]
TW-ASGC	via backup link	[2.5G]	via backup link	[2.5G]	OK	[10G]
UK-T1-RAL	unreachable	[0G]	unreachable	[0G]	OK	[10G]
US-CMS-FNAL	via backup link	[10G/2]	OK	[10G/2]	via backup link	[10G/2]
US-T1-BNL	via backup link (BGP bug)	[10G/2]	via backup link (BGP bug)	[10G/2]	OK	[10G/2]

Real Fiber Cut Near Chicago 24th April

Site	Status	
CA-TRIUMF	Primary up. Backup down.	[5G]
DE-KIT	Not affected	[10G]
ES-PIC	Not affected	[10G]
FR-IN2P3	Not affected	[10G]
IT-INFN-CNAF	Not affected	[10G]
NDGF	Not affected	[10G]
NL-T1	Not affected	[10G]
TW-ASGC	Primary up. Backup down.	[10G]
UK-T1-RAL	Not affected	[10G]
US-CMS-FNAL	Primary and backup down. Reachable via Geant2-IP and ESnet	[10G shared]
US-T1-BNL	Primary up. Backup down	[10G]

Real Fiber Cut (DE-CH) Near Frankfurt 25th April

Site	Status	
CA-TRIUMF	Primary and secondary down. Reachable via BNL	[1G]
DE-KIT	Primary down. Reachable via NL-T1, CNAF, IN2P3	[10G/2]
ES-PIC	Not affected	[10G]
FR-IN2P3	Not affected	[10G]
IT-INFN-CNAF	Not affected	[10G]
NDGF	Primary down. Reachable via Geant2-IP	[10G shared]
NL-T1	Not affected	[10G]
TW-ASGC	Primary down. Reachable via backup link	[10G shared]
UK-T1-RAL	Not affected	[10G]
US-CMS-FNAL	Reduced bandwidth on the primary link. Backup down	[10G shared]
US-T1-BNL	Primary down. Reachable via backup in Chicago	[10G shared]

Operational Support

- EGEE-SA2 providing the lead on the operational model
  - Much initial disagreement on approach, now starting to converge. Last OPN meeting concentrated on “points of view”
    - The “network manager” view
    - The “user” view (“Readiness” expectations)
    - The “distributed” view (E2ECU, IPCU, GGUS etc)
    - The “grass roots” view (Site engineers)
    - The “centralised” view (Dante)
  - All documentation is available on the Twiki. Much work remains to be done.

Evolving Operational Model

- Need to identify the major operational components and orchestrate their interactions including:
  - Information repositories
    - GGUS, TTS, Twiki, PerfSonar etc.
  - Actors
    - Site network support, ENOC, E2ECU, USLHCNet etc.
    - Grid Operations.
  - Processes
    - Who is responsible for which information?
    - How does communication take place?
      - Actor <-> Repository
      - Actor <-> Actor
    - For what purpose does communication take place?
      - Resolving identified issues
      - Authorising changes and developments
- A minimal design is needed to deal with the major issues
  - Incident Management (including scheduled interventions)
  - Problem Management
  - Change Management

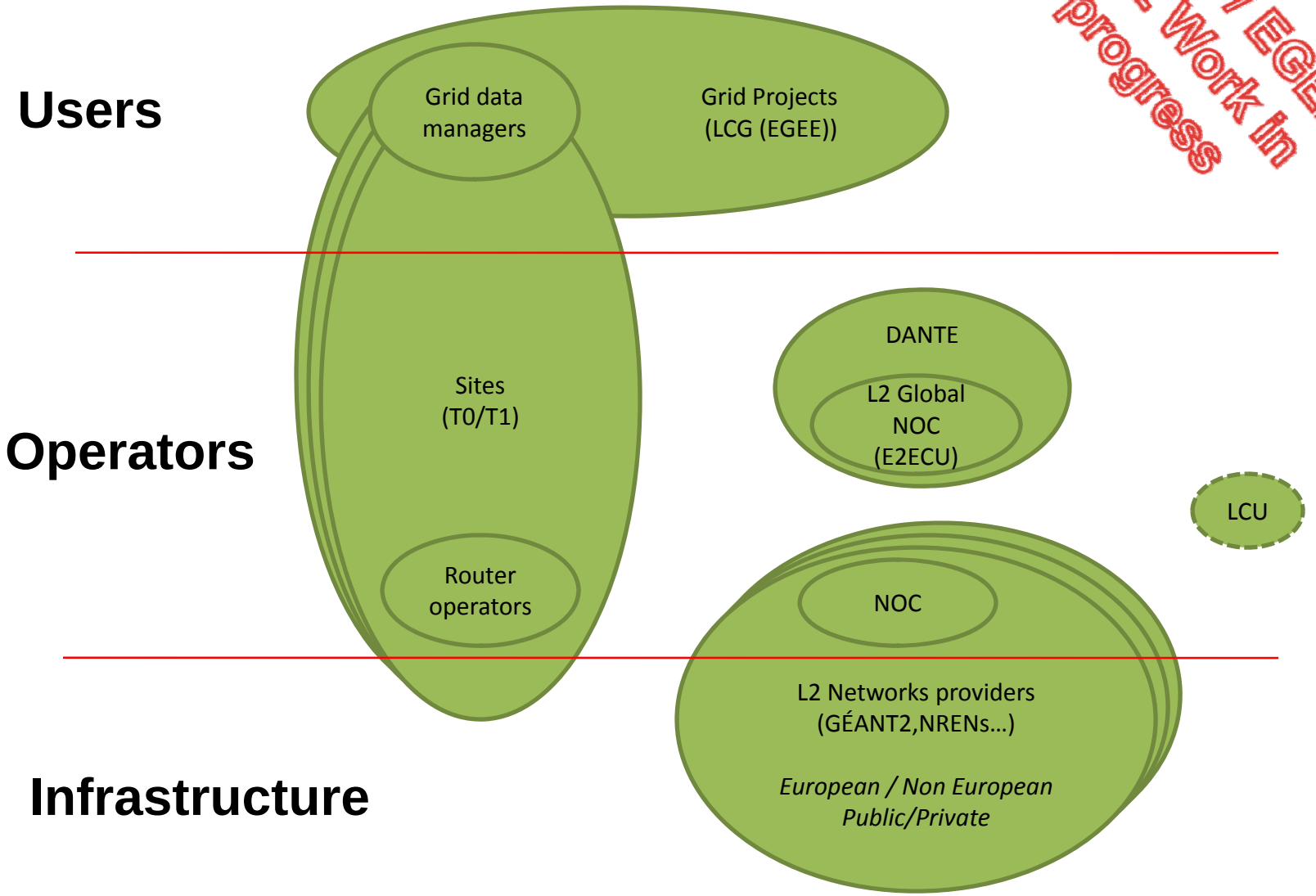
In Practical Terms ….

(provided by Dan Nae, as a site managers view)

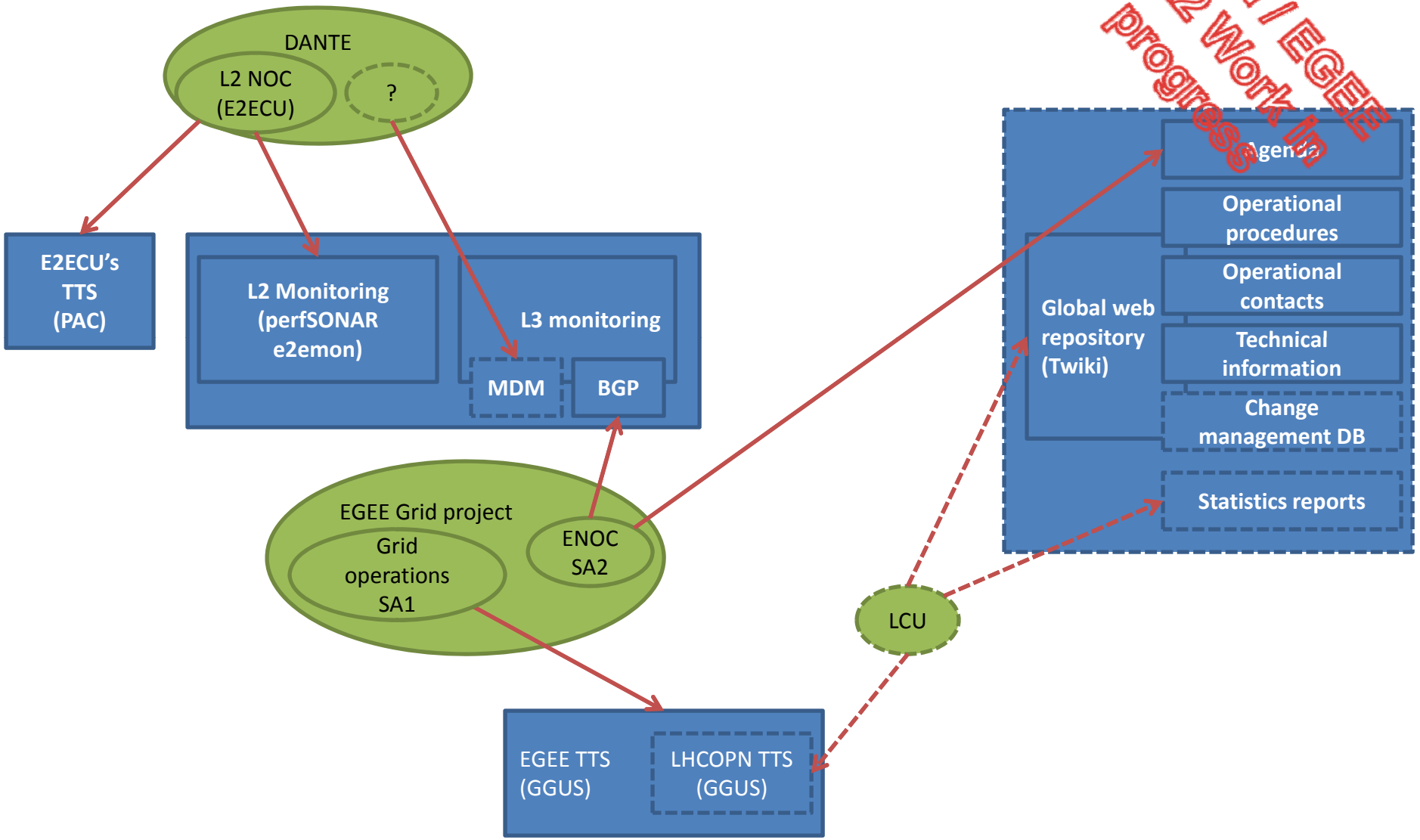
- An end-to-end monitoring system that can pin-point reliably where most of the problems are
- An effective way to integrate the above monitoring system into the local procedures of the various local NOCs to help them take action
- A centralized ticketing system to keep track of all the problems
- A way to extract performance numbers from the centralized information (easy)
- Clear dissemination channels to announce problems, maintenance, changes, important data transfers, etc.
- Someone to take care of all the above
- A data repository engineers can use and a set of procedures that can help solve the hard problems faster (detailed circuit data, ticket history, known problems and solutions)
- A group of people (data and network managers) who can evaluate the performance of the LHCOPN based on experience and gathered numbers and can set goals (target SLAs for the next set of tenders, responsiveness, better dissemination channels, etc)

LHCOPN Actors

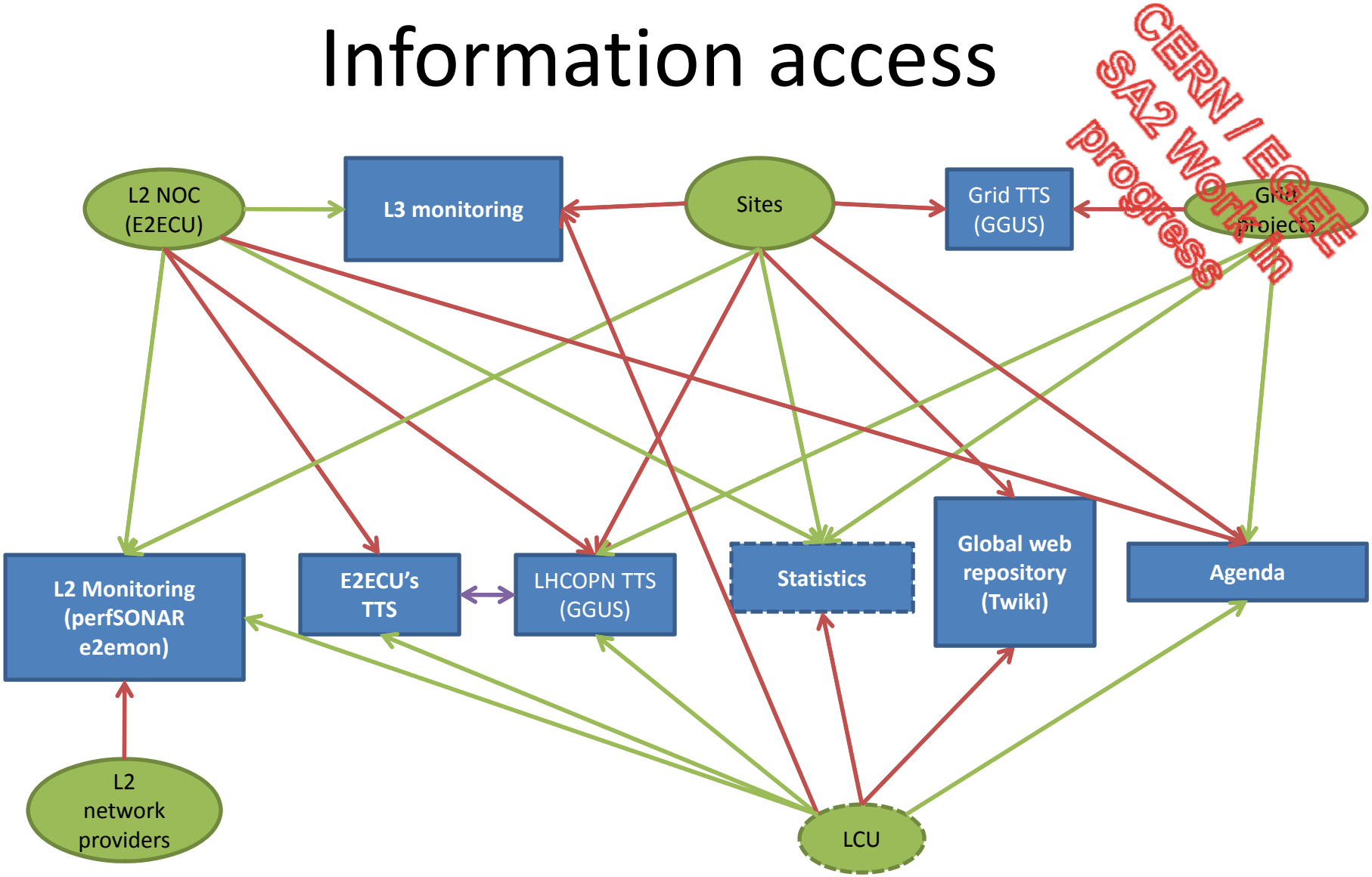
CERN / EGEE SA2 Work in progress



Actors and information repositories management



Information access



A → B A reads B

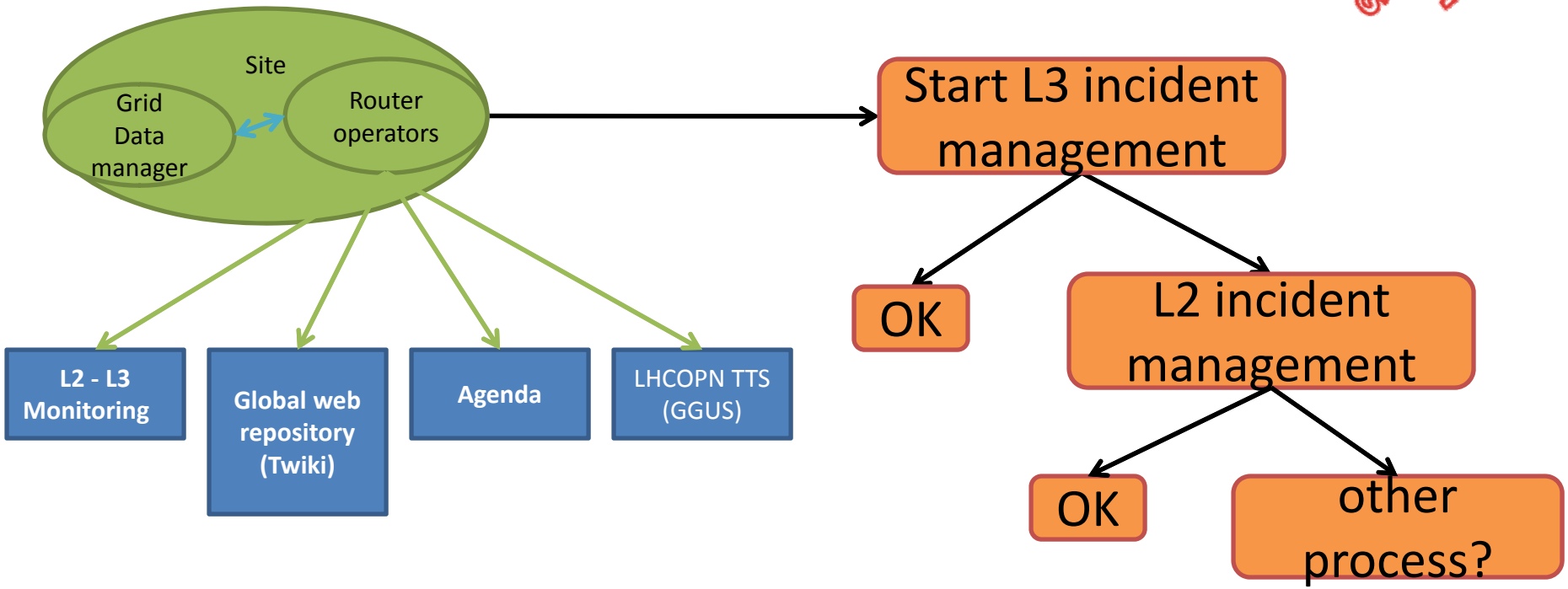
A → B A reads and writes B

A ↔ B TT exchange between A and B

Trouble management process

problem cause and location unknown

CERN / EGEE SA2 Work in progress



A → B A reads B

A ↔ B A deals with B

A → B A notifies B

Basic Link Layer Monitoring

- Perfsonar very well advanced in deployment (but not yet complete). Monitors the “up/down” status of the links.
- Integrated into the “End to End Coordination Unit” (E2ECU) run by DANTE
- Provides simple indications of “hard” faults.
- Insufficient to understand the quality of the connectivity

E2emon Link Status

Mon. Link Local Name	E2E Link ID	Topology Point A	Role	Topology Point B	Role	Link Type	Oper. Status	Admin. Status	Time Stamp
S513-C-BE12	CERN-PIC-LHCOPN-001	CERN-T0	EP	GEANT2-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:23+02:00
S513-C-BE2	CERN-CNAF-LHCOPN-001	CERN-T0	EP	GEANT2-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:30+02:00
S513-C-BE3	CERN-SARA-LHCOPN-002	CERN-T0	EP	NETHERLIGHT-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:50+02:00
S513-C-BE4	CERN-RAL-LHCOPN-001	CERN-T0	EP	GEANT2-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:23+02:00
S513-C-BE5	CERN-ASGC-LHCOPN-003	CERN-T0	EP	NETHERLIGHT-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:18+02:00
S513-C-BE6	CERN-TRIUMF-LHCOPN-002	CERN-T0	EP	NETHERLIGHT-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:19+02:00
S513-C-BE7	CERN-IN2P3-LHCOPN-001	CERN-T0	EP	RENATER-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:17:08+02:00
S513-C-BE9	CERN-GRIDKA-LHCOPN-001	CERN-T0	EP	GEANT2-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:19+02:00
S513-C-RE1-VLAN	CERN-FERMI-LHCOPN-002	CERN-T0	EP	USLHCNET-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:33+02:00
S513-C-RE10-VLAN	CERN-BNL-LHCOPN-002	CERN-T0	EP	USLHCNET-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:52+02:00
S513-C-RE6	CERN-TRIUMF-LHCOPN-001	CERN-T0	EP	NETHERLIGHT-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:17:11+02:00
S513-C-VE1-VLAN	CERN-FERMI-LHCOPN-001	CERN-T0	EP	USLHCNET-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:50+02:00
S513-C-VE2-VLAN	CERN-BNL-LHCOPN-001	CERN-T0	EP	USLHCNET-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:25+02:00
S513-E-EE1	CERN-NDGF-LHCOPN-001	CERN-T0	EP	GEANT2-GEN	DP	ID Part.Info	Up	Normal Oper.	2008-05-07T09:16:57+02:00

E2emon detail

Domain	DFN				GEANT2					CERN	
Link Structure	EP	◄——►	DP	◄·····	·····►	DP	◄——►	DP	◄·····	·····►	EP
Type	EndPoint	Domain Link	Demarc	ID Part.Info	ID Part.Info	Demarc	Domain Link	Demarc	ID Part.Info	ID Part.Info	EndPoint
Local Name	DFN-FZK23	DFN-GE10/HUA0674_FRA_FZK	DFN-FRA82	DFN-GE10/ANWD_KA1006_FRA_0GEANT	DFN-LHC_ter1.15.fra-gen.fra	GEANT2-FRA	fra-gen_LHC_CERN-DFN_06006	GEANT2-GEN	CERN-LHC_ter1.15.fra-gen.gen	S513-C-BE9	CERN-T0
State Oper.	-	Up	-	Up	Up	-	Up	-	Up	Up	-
State Admin.	-	Normal Oper.	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-	Normal Oper.	Normal Oper.	-
Timestamp	-	2008-05-07T09:19:00+01:00	-	2008-05-07T09:19:00+01:00	2008-05-07 T05:31:23.0+0100	-	2008-05-07T05:31:23.0+0100	-	2008-05-07 T05:31:23.0+0100	2008-05-07 T09:22:19+02:00	-

Monitoring

- Coherent (active) monitoring is a essential feature to understand how well the service is running.
  - Many activities around PerfSonar are underway in Europe and the US.
- Initial proposal by Dante to provide an “appliance” is now largely accepted.
  - Packaged, coherent, maintained installation of tools to collect information on the network activity.
  - Caveat: Service only guaranteed to end of GN2 (Macrh 2009) with the intention to continue in GN3.

Initial Useful Metrics and Tools

(From Eric Boyd I2)

Network Path characteristics

- Round trip time (perfSONAR PingER)
- Routers along the paths (traceroute)
- Path utilization/capacity (perfSONAR SNMP-MA)
- One way delay, delay variance (perfSONAR owamp)
- One way packet drop rate (perfSONAR owamp)
- Packets reordering (perfSONAR owamp)
- Achievable throughput (perfSONAR bwctl)

Issues, Risks, Mitigation

- OPN is fundamental to getting the data from CERN to the T1's.
- It is a complex multi-domain network relying on infrastructure provided by:
  - (links) NREN's, Dante and commercial providers
  - (IP) T1's and CERN
  - (operations) T1's, CERN, EGEE and USLHCNet
- Developing a robust operational model is a major ongoing piece of work.
  - Define responsibilities. Avoid "finger pointing loops"
  - Need to separate design from implementation
  - Need to combine innovation and operation
    - Be robust, but not too conservative

Harvey
Newman

HEP Bandwidth Roadmap for Major Links (in Gbps): US LHCNet Example

Year	Production	Experimental	Remarks
2001	0.155	0.622-2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH DWDM; GigE Integ.
2003	2.5	10-20	DWDM; 1 + 10 GigE Integration
2005-6	10-20	2-10 X 10	λ Switch; λ Provisioning
2007-8	3-4 X 10	~10 X 10; 100 Gbps	1st Gen. λ Grids
2009-10	6-8 X 10	~20 X 10 or ~2 X 100	100 Gbps λ Switching
2011-12	~20 X 10 or 2 X 100	~10 X 100	2nd Gen λ Grids Terabit Networks
2013-5	~Terabit	~MultiTbps	~Fill One Fiber

Paralleled by ESnet Roadmap for Data Intensive Sciences

Science Lives in an Evolving World

- New competition for the “last mile” giving a critical mass of people access to high performance networking.
  - But asymmetry may become a problem.
- New major investments in high capacity backbones.
  - Commercial and “dot com” investments.
  - Improving end-end performance.
- New major investments in data centers.
  - Networks of data centers are emerging (a specialised grid!)
  - Cloud computing, leverages networks and economies of scale – *its easier (and cheaper) to move a bit than a watt.*
- **This creates a paradigm change, but at the user service level and new business models are emerging**
  - Multimedia services are a major driver. (YouTube, IPTV etc.)
  - Social networking (Virtual world services etc)
  - Virtualisation to deliver software services – Transformation of software from a “product” to a “service”
- **Sustained and increasing oil prices should drive demand for networked services even more in the coming years.**


Simple solutions are often the best!