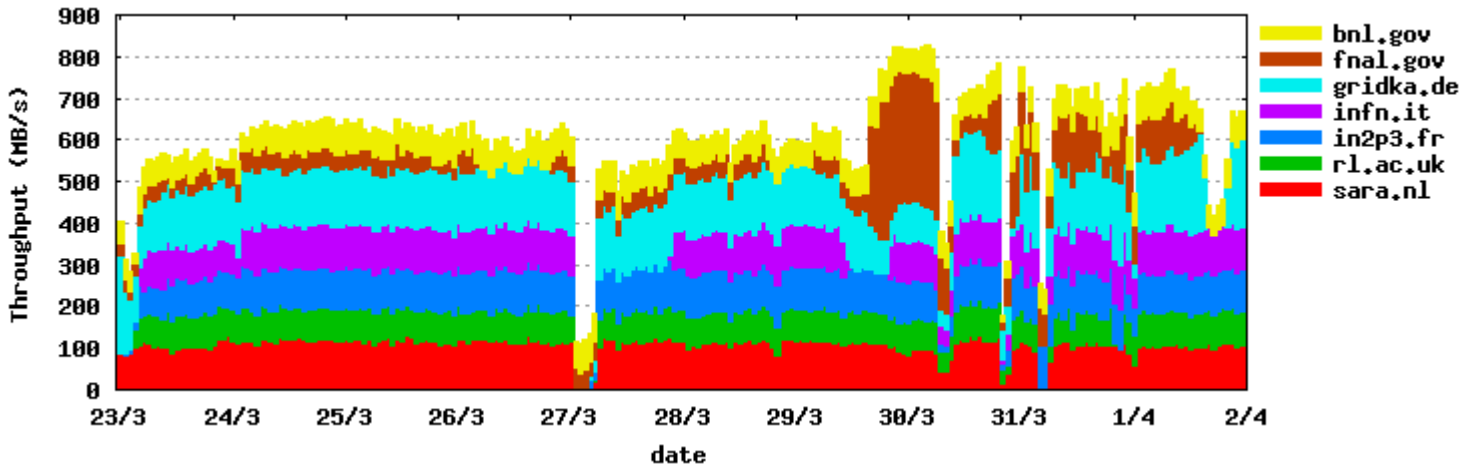
T0/T1 Network Meeting

"Network for the Service Callenges"

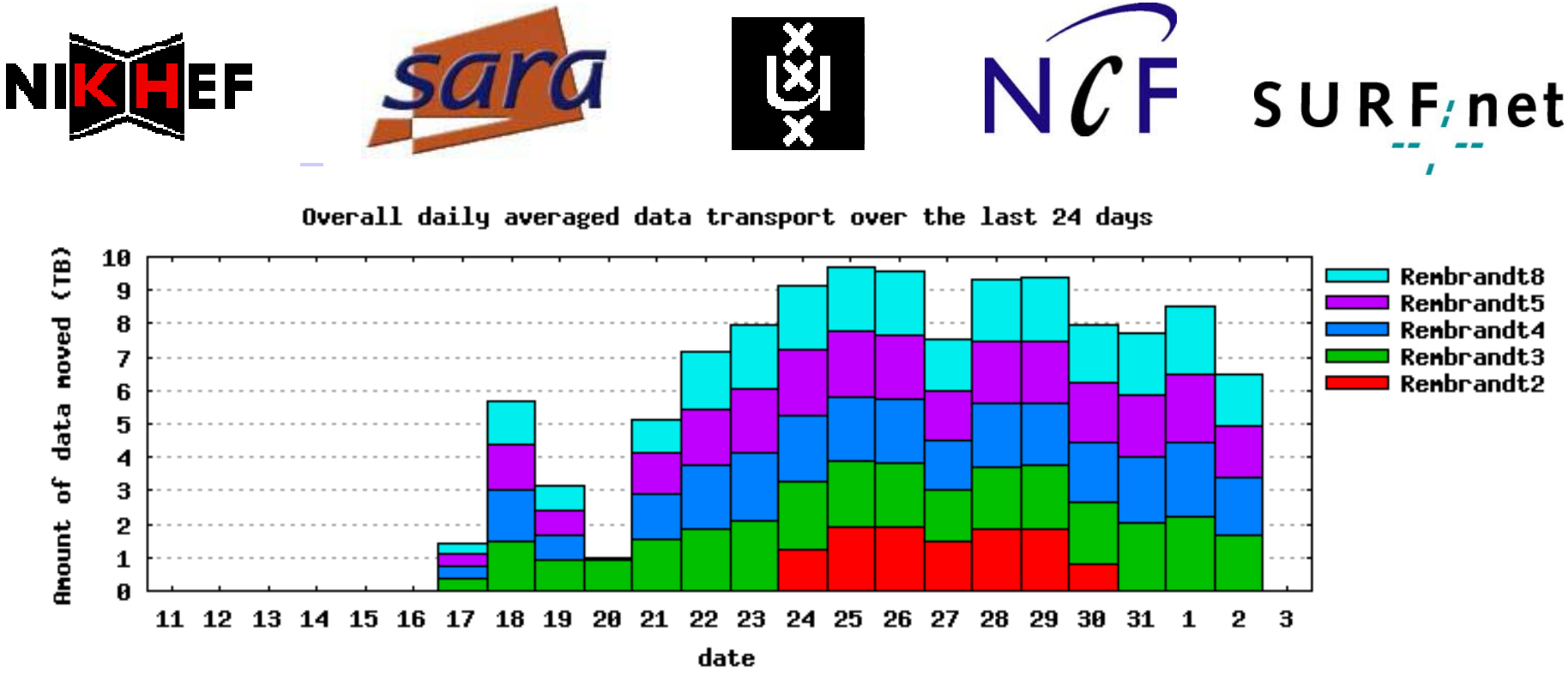
Kors Bos, NIKHEF, Amsterdam

Amsterdam, 8th April 2005



LCG Service Challenge II

"The Dutch Contribution"

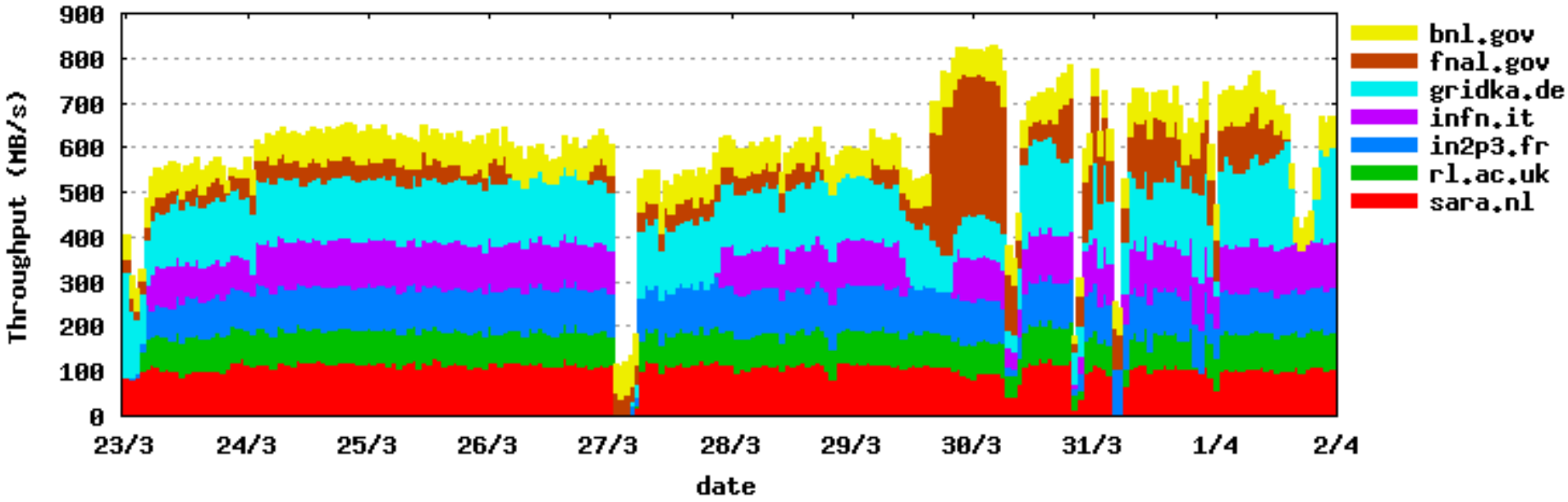


Service Challenge 2 Summary

- Service Challenge 2
  - Throughput test from Tier-0 to Tier-1 sites
  - Started 14th March
- Set up Infrastructure to 7 Sites
  - BNL (Upton, NY), CCIN2P3 (Lyon), CNAF (Bologna), FNAL (Chicago), GridKa (Karlsruhe), RAL (Didcot, UK), SARA (Amsterdam)
- ~100MB/s to each site
  - At least 500MB/s combined out of CERN at same time
  - 500MB/s to a few sites individually
- Two weeks sustained 500 MB/s out of CERN

SC2 met its throughput targets

- >600MB/s daily average for 10 days was achieved - Midday 23rd March to Midday 2nd April
  - Not without outages, but system showed it could recover rate again from outages
  - Load reasonable evenly divided over sites (given network bandwidth constraints of Tier-1 sites)



Division of Data between sites

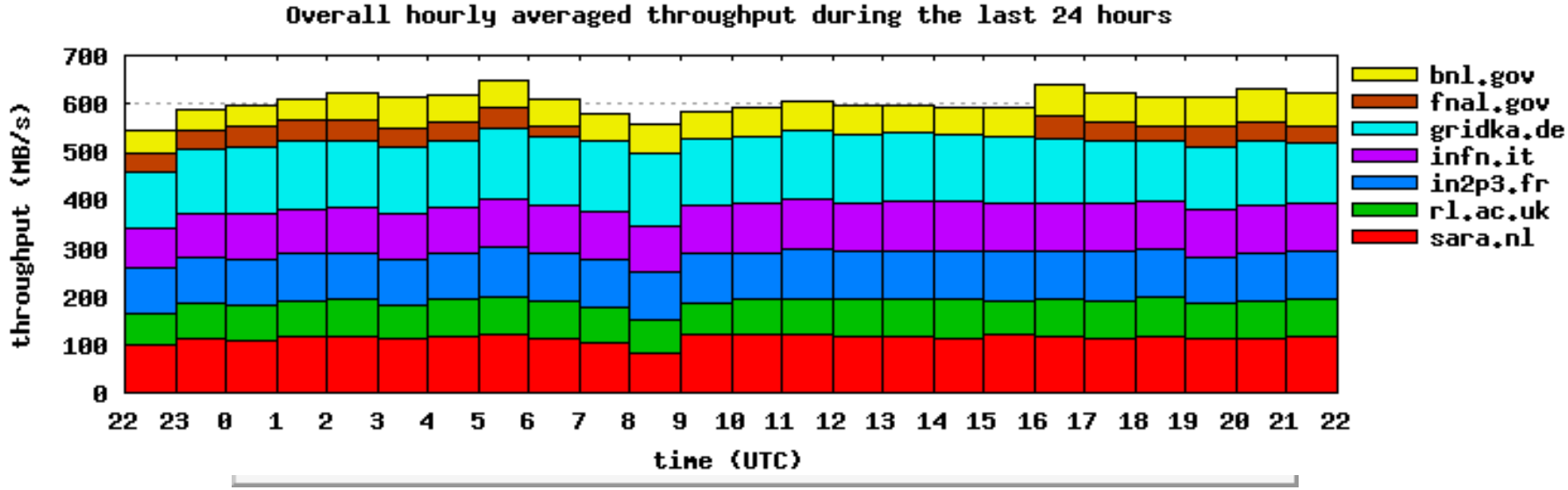
Site	Average throughput (MB/s)	Data Moved (TB)
BNL	61	51
FNAL	61	51
GridKA	133	109
CCIN2P3	91	75
CNAF	81	67
RAL	72	58
SARA	106	88
TOTAL	600	500

Storage and Software used

- Most sites ran Globus gridftp servers
  - CCIN2P3, CNAF, GridKa, SARA
- The rest of the sites ran dCache
  - BNL, FNAL, RAL
- Most sites used local or system-attached disk
  - FZK used SAN via GPFS
  - FNAL used production CMS dCache, including tape
- Load-balancing for gridftp sites was done by the RADIANT software running at CERN in push mode

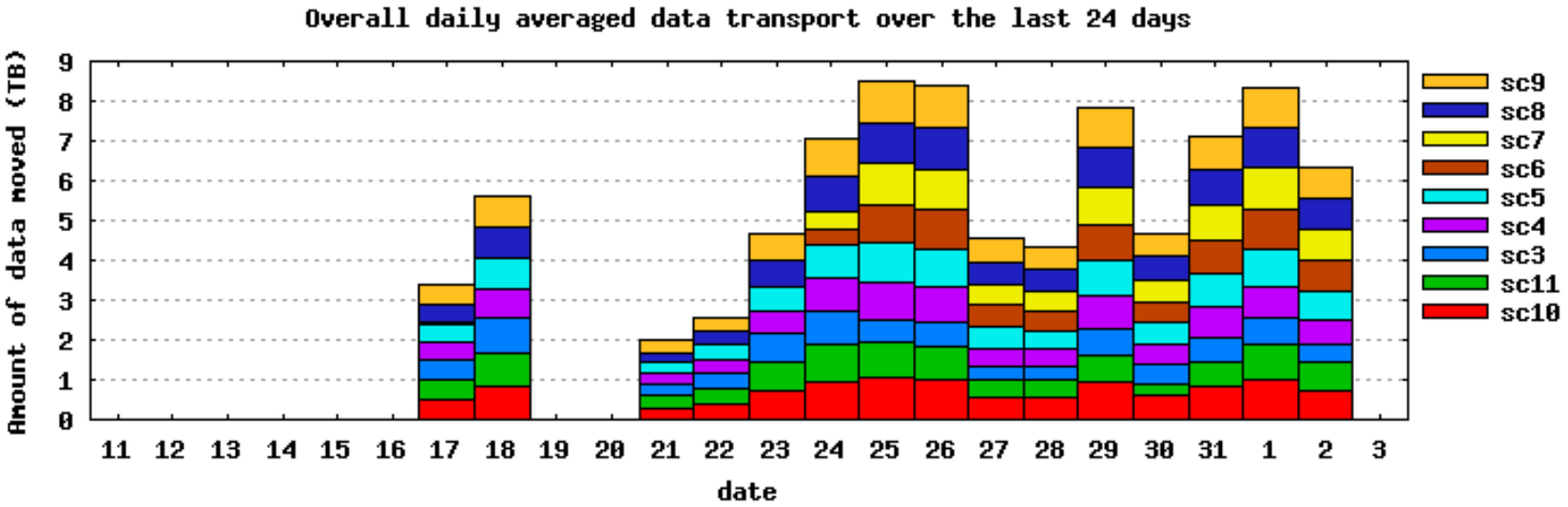
Monitoring @ CERN

- MRTG Graphs of network usage out of cluster
- LEMON monitoring of cluster
  - CPU usage
  - Disk usage
  - Network usage
- Gridftp logfile monitoring

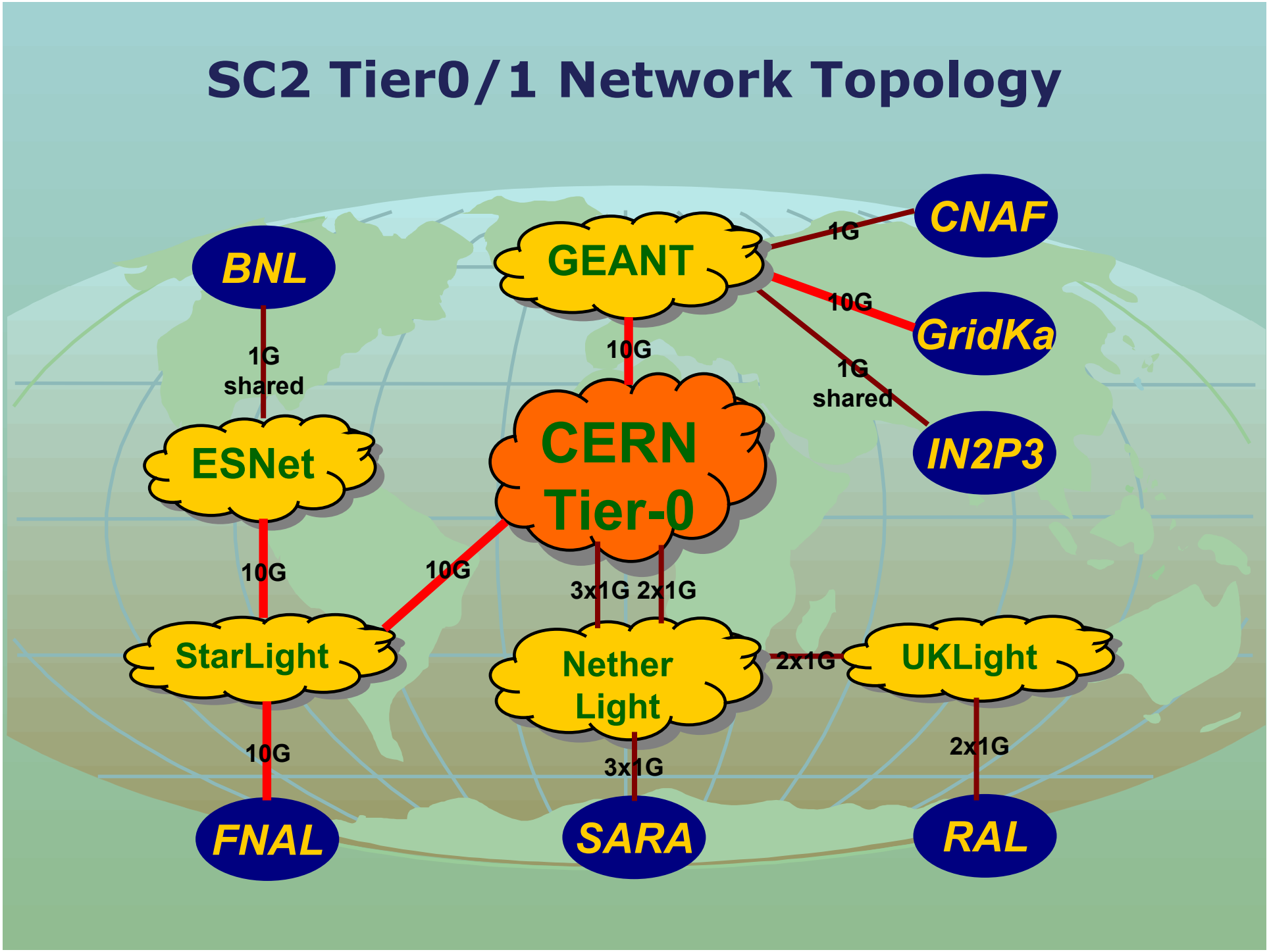


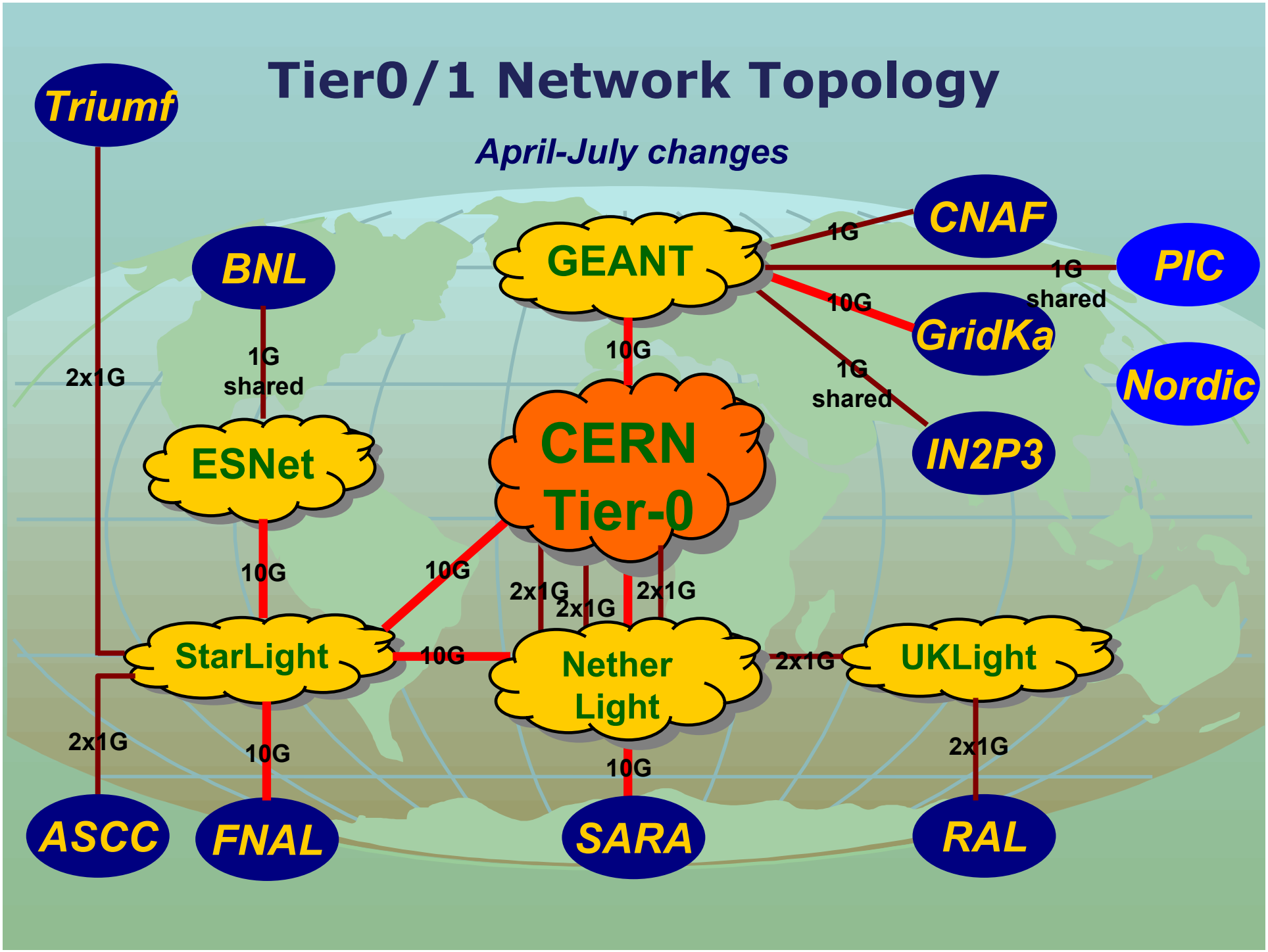
Tier-1 Monitoring

- Gridftp Monitoring (http://challenge.matrix.sara.nl/SC/)
- Main monitoring tool used during the SC2
  - Hourly throughput per site
  - Hourly throughput per host
  - Daily throughput per site
  - Daily throughput per host



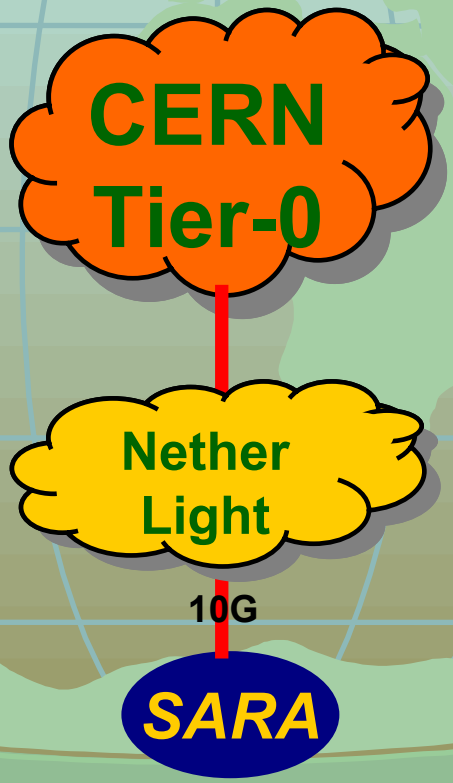
SC2 Tier0/1 Network Topology




Tier0/1 Network Topology
April-July changes
Triumf
BNL
GEANT
CNAF
PIC
GridKa
Nordic
IN2P3
ESNet
CERN
Tier-0
StarLight
Nether
Light
UKLight
ASCC
FNAL
SARA
RAL
2x1G
1G
shared
1G
10G
1G
shared
1G
shared
10G
10G
10G
2x1G
2x1G
2x1G
10G
2x1G
2x1G
10G
10G
2x1G

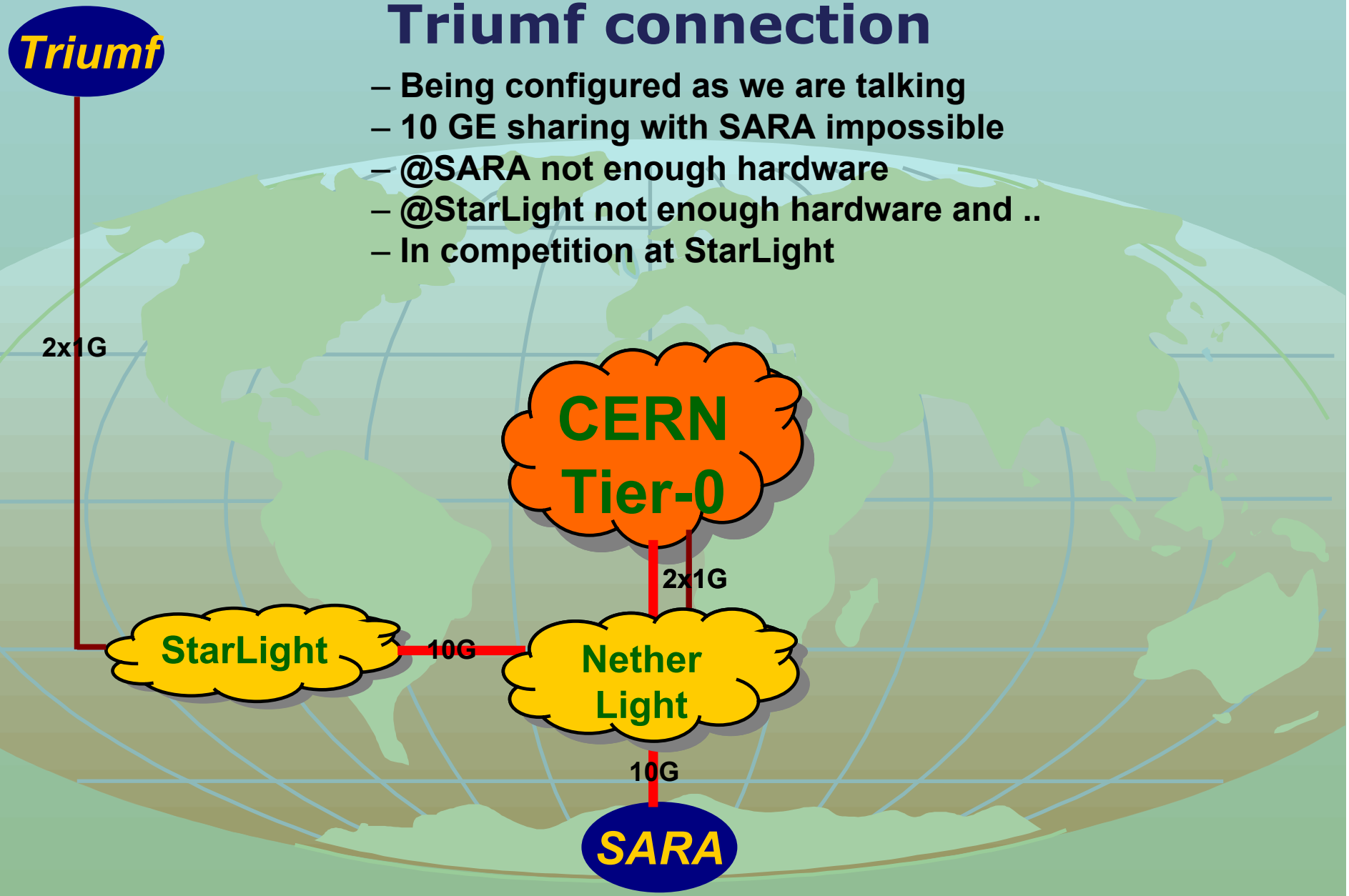
SARA connection

- Being configured as we are talking
- @SARA Force10 switch on loan from UvA
- @CERN need WAN PHY to LAN PHYconversion
- Switch on loan from Foundry to Canadians
- Only works until end-July … what then?



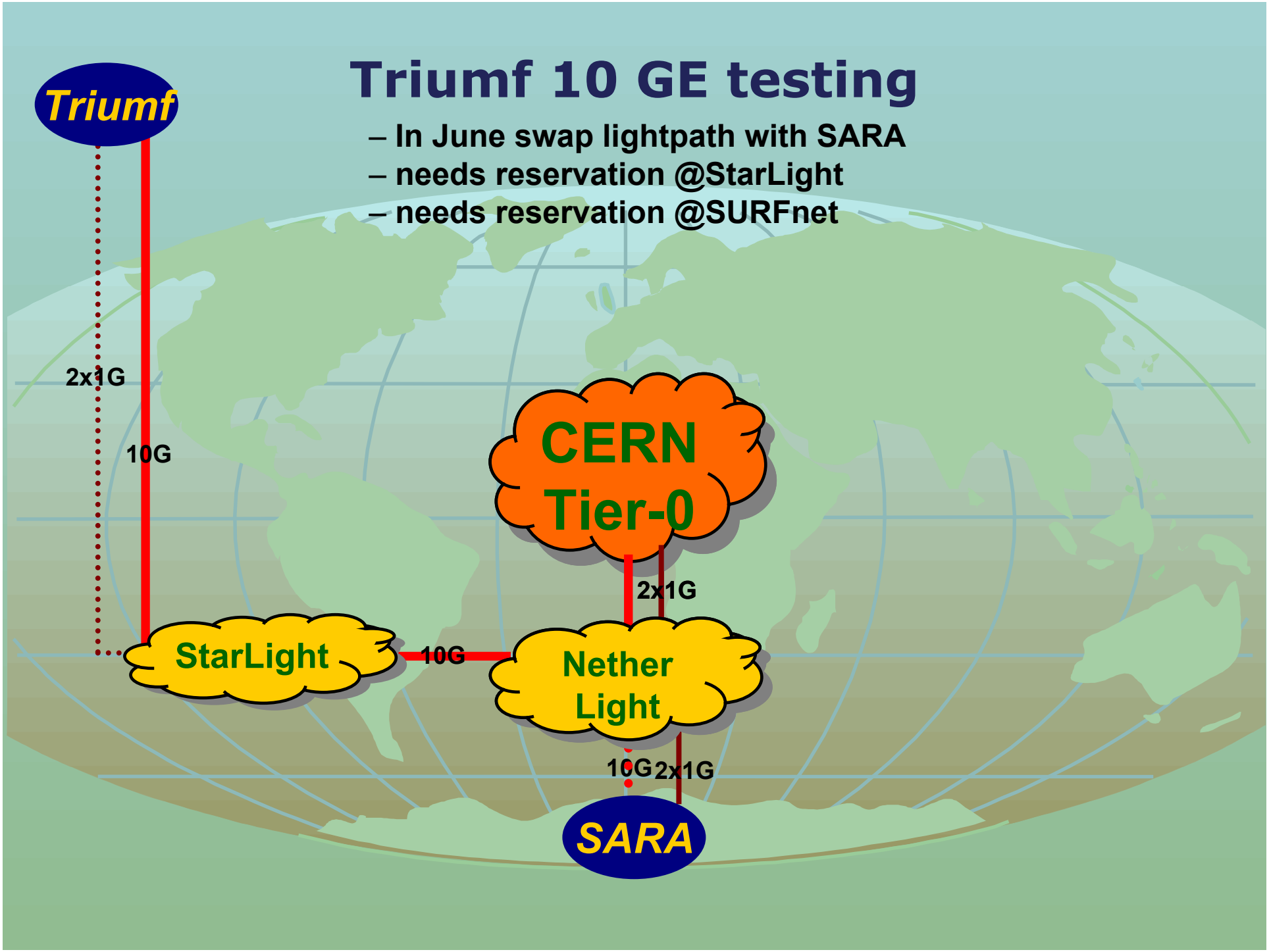
Triumf connection

- Being configured as we are talking
- 10 GE sharing with SARA impossible
- @SARA not enough hardware
- @StarLight not enough hardware and ..
- In competition at StarLight



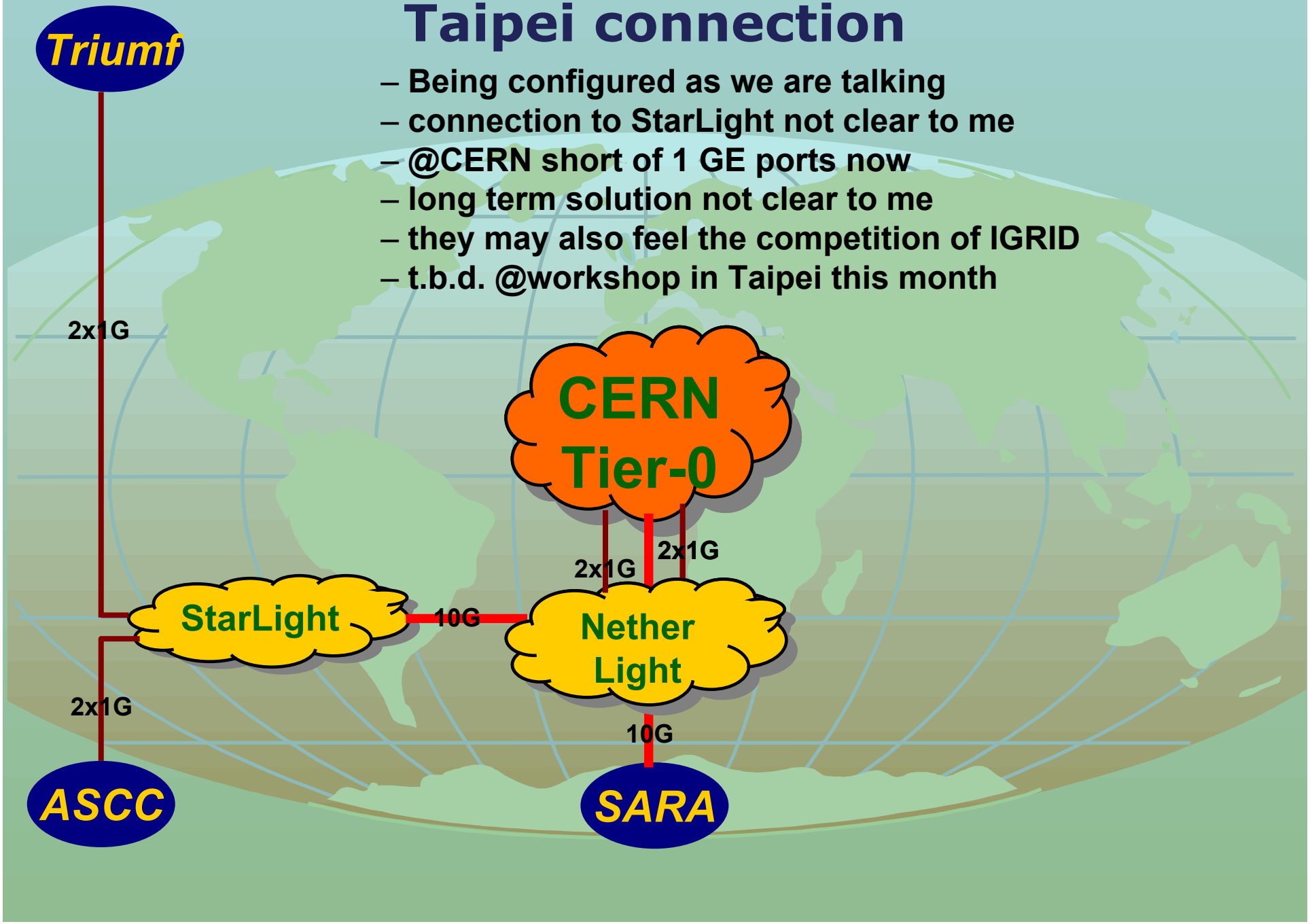
Triumf 10 GE testing

- In June swap lightpath with SARA
- needs reservation @StarLight
- needs reservation @SURFnet



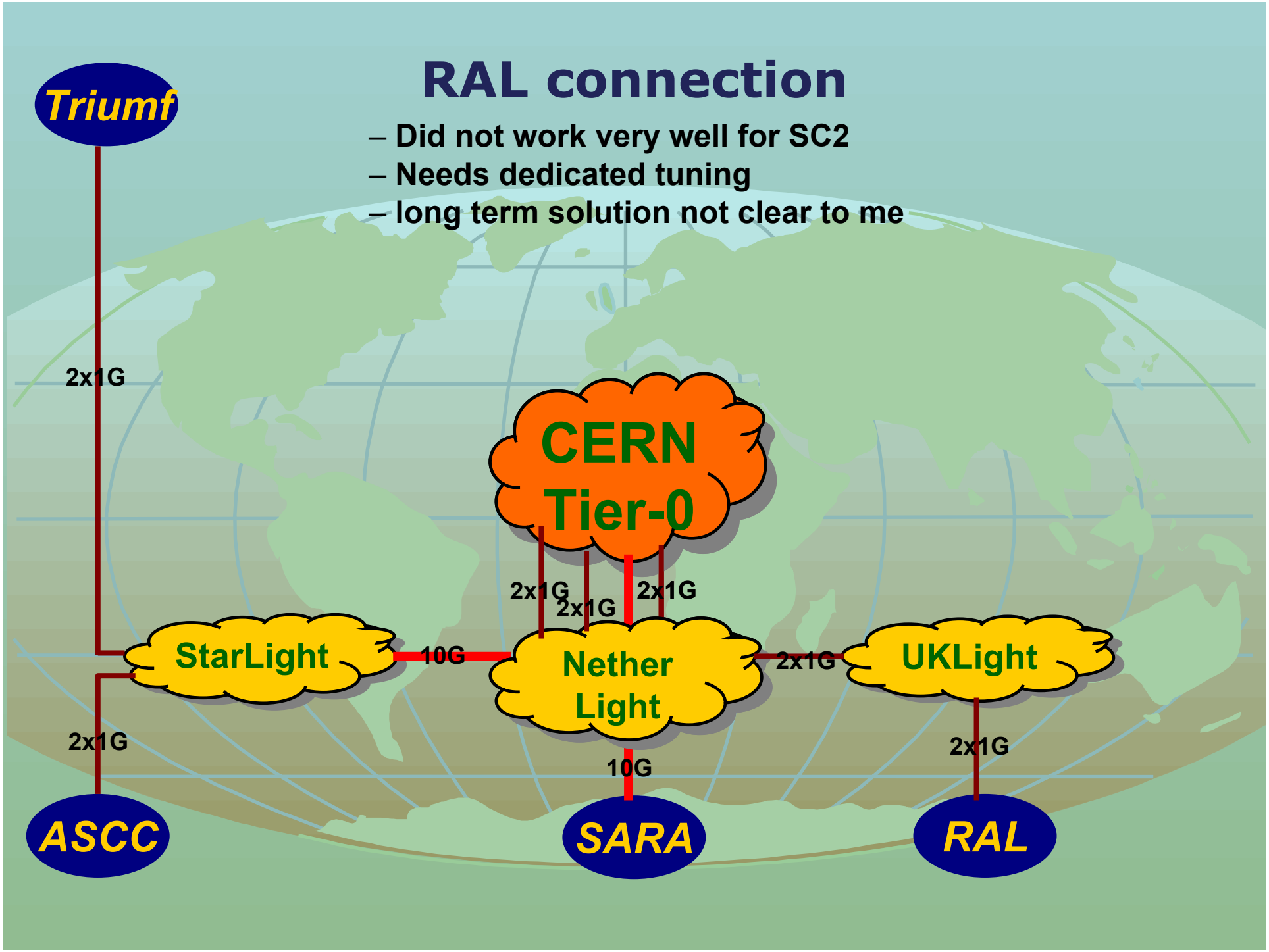
Taipei connection

- Being configured as we are talking
- connection to StarLight not clear to me
- @CERN short of 1 GE ports now
- long term solution not clear to me
- they may also feel the competition of IGRID
- t.b.d. @workshop in Taipei this month



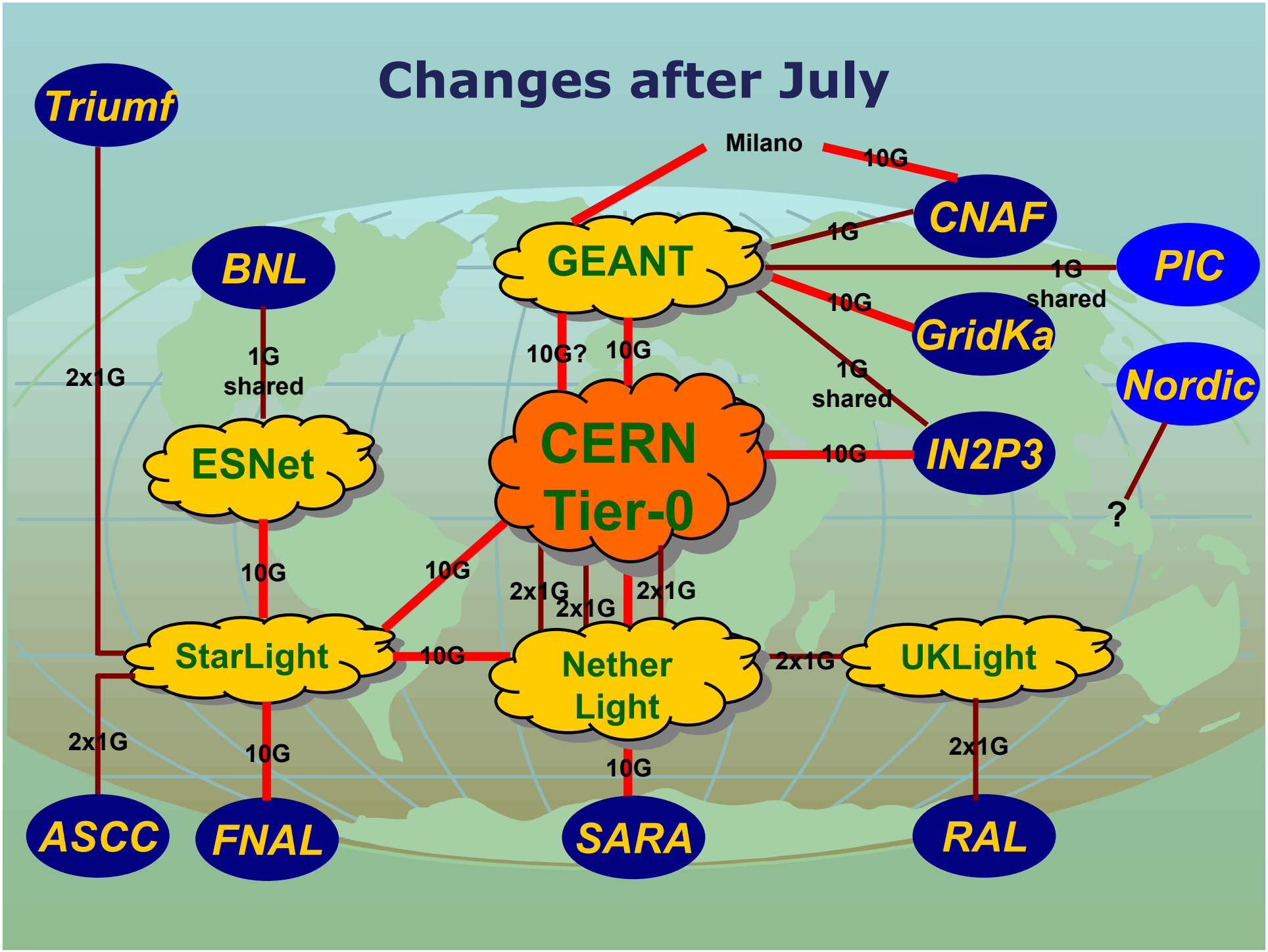
RAL connection

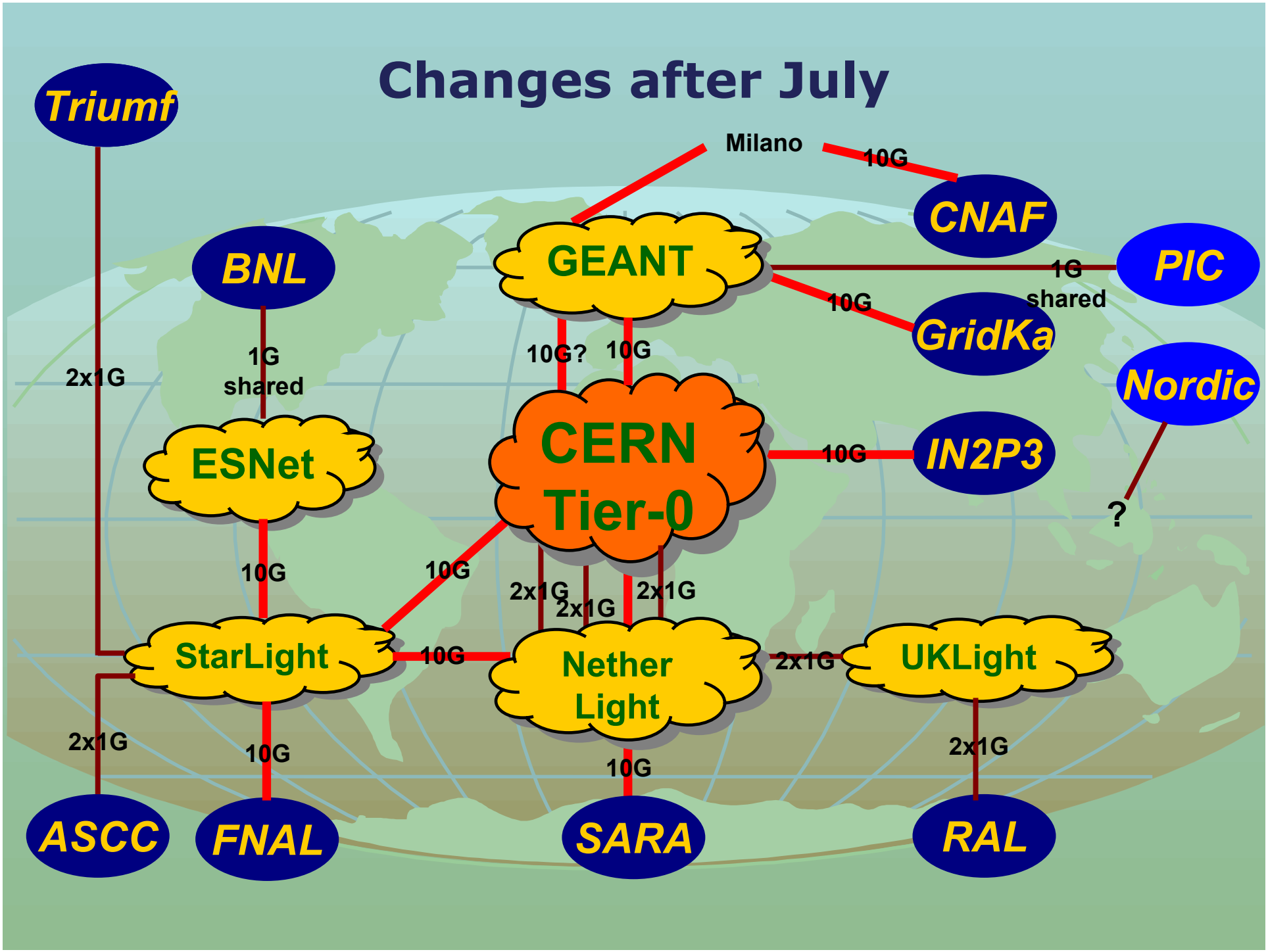
- Did not work very well for SC2
- Needs dedicated tuning
- long term solution not clear to me



Network outlook for SC3 (July)

- **Dedicated 10 GE connections**
  - Fermilab (StarLight)
  - GridKa (GEANT)
  - SARA (SURFnet)
- **Dedicated n x 1 GE connections**
  - CNAF (GEANT) n=1
  - CCIN2P3 (GEANT) n=1
  - RAL (UKERNA, SURFnet) n=2
  - Taipei (?, SURFnet) n=2
  - Triumf (Canari, StarLight, SURFnet) n=2
- **Shared 1 GE connections**
  - BNL (ESNET, StarLight)
  - PIC (GEANT)
- **Not connected yet**
  - Nordic Federation


Changes after July
Triumf
Milano
10G
CNAF
1G
GEANT
PIC
1G
shared
BNL
10G
GridKa
2x1G
1G
shared
10G?
10G
1G
shared
Nordic
CERN
Tier-0
ESNet
10G
IN2P3
?
10G
10G
2x1G
2x1G
2x1G
StarLight
10G
Nether
Light
2x1G
UKLight
2x1G
10G
2x1G
10G
ASCC
FNAL
SARA
RAL


Changes after July
Triumf
Milano
10G
CNAF
BNL
GEANT
PIC
1G
shared
10G
GridKa
2x1G
1G
shared
10G?
10G
Nordic
CERN
Tier-0
ESNet
10G
IN2P3
?
10G
10G
2x1G
2x1G
2x1G
StarLight
10G
Nether
Light
2x1G
UKLight
2x1G
10G
2x1G
10G
ASCC
FNAL
SARA
RAL

Network outlook for SC4 (December)

- **Dedicated 10 GE connections**
  - Fermilab (StarLight)
  - GridKa (DFN, GEANT)
  - SARA (SURFnet)
  - **CCIN2P3 (Renater)**
  - **CNAF (GEANT)**
- **Dedicated n x 1 GE connections**
  - RAL (UKERNA, SURFnet) n=2
  - Taipei (XYZnet, SURFnet) n=2
  - Triumf (Canari, StarLight, SURFnet) n=2
- **Shared 1 GE connections**
  - BNL (ESNET, StarLight)
  - PIC (GEANT)
- **Not connected ?**
  - Nordic Federation

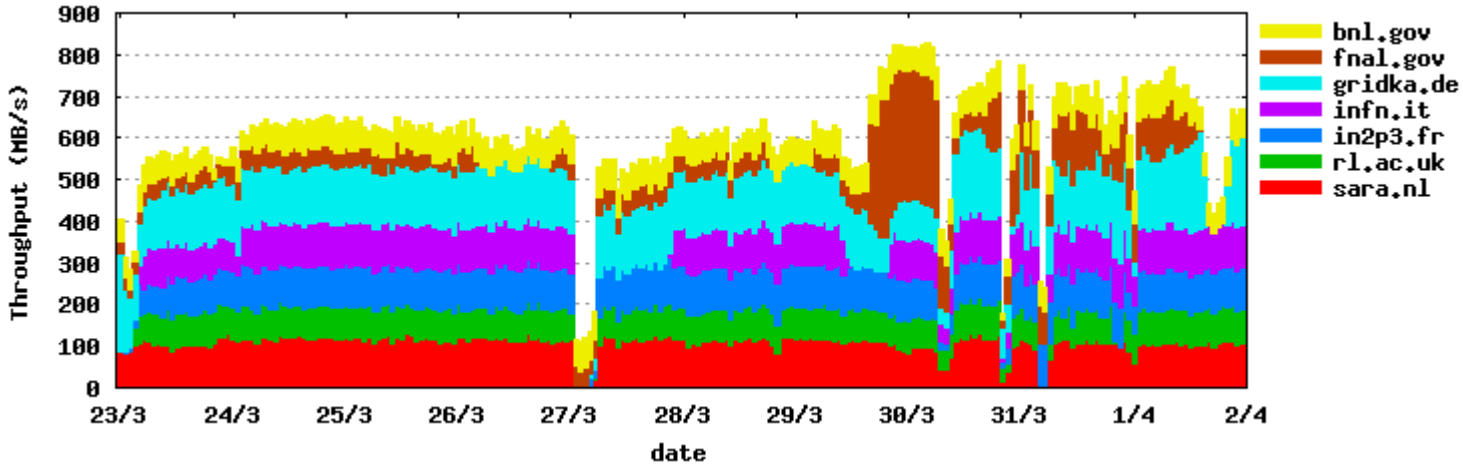
Summary

- **SC2 met it's throughput goals**
  - An improvement from SC1, but not a service yet
  - Monitoring; Outages can be understood and controlled
  - 2 sites have dedicated 10G and 5 dedicated n x 1G links
- **SC3 in July**
  - More T1's involved; write/read to/from tape
  - New Software gLite transfer software, SRM service
  - 3 sites have dedicated 10G and 5 dedicated n x 1G links
- **SC4 end 2005**
  - All T1's, full model at reduced rate
  - Experiment's involvement
  - 5 sites have dedicated 10G and 3 dedicated n x 1G links
- **Worries ?**
  - 2 T1 sites on shared links, 1 T1 not connected at all
  - Do we have sufficient network hardware (switches, converters,..)

T0/T1 Network Meeting

"Network for the Service Callenges"

END



Some backup slides

Service Outages (1/2)

- Progress page kept in SC wiki of
  - All tunings made to the system
  - All outages noted during the running
  - Any actions needed to cleanup and recover service

    <http://service-radiant.web.cern.ch/service-radiant/wiki/ow.asp?ChallengeSC2Progress>
- No real 24x7 service in place
  - Manual monitoring of monitoring webpages
  - Best-effort restart of service
  - Also at Tier-1 sites – problems communicated to service challenge teams, but this was not a 24x7 coverage

Service Outages (2/2)

- Capacity in the cluster meant that we could recover from one site not being active
  - Other sites would up their load a bit automatically due to gridftp stream rates increasing
  - Only thing that killed transfers were CERN outages
- We did not do any scheduled outages during the SC
  - No procedures for starting a schedule outage
  - If we had done one to move to managed network infrastructure, it would have removed some of the scheduled ones

Outage Breakdown - CERN

- Mxproxy instability
  - Locks up under high load
    - Understood by developers of myproxy
    - Can be handled by watchdog job
  - Particular problem on restart after network outage
- CERN LAN Network outages
  - Had 2 long-ish outages (~12 hours)
  - Issue with being on non-managed network hardware
- Database quota limit hit
  - Tablespace was being monitored but not quota
  - Quota monitoring added
- Database load problems
  - Caused intermittent drops in throughput due to new jobs not being scheduled
  - In-memory hob queues in the transfer agents meant these we're a big problem

Individual site tests

- Being scheduled right now
  - Sites can pick days in next two weeks when they have the capacity
  - 500MB/s to disk
  - 60MB/s to tape
- FNAL is running 500MB/s disk tests right now