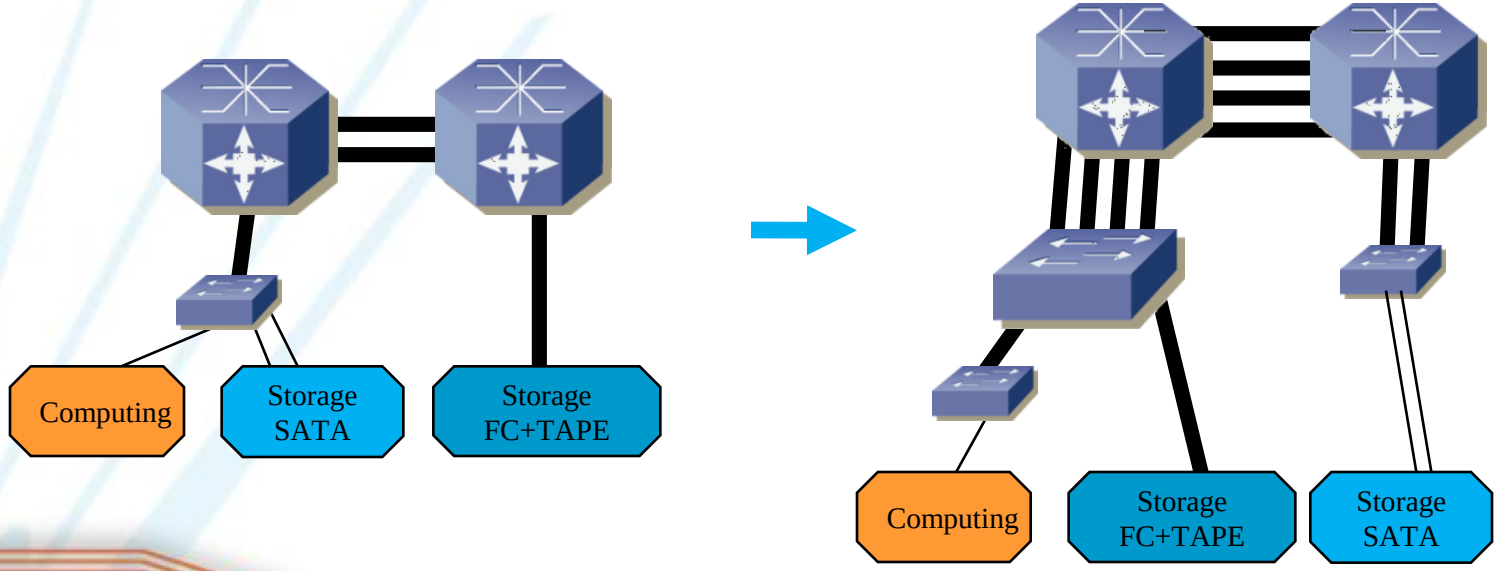
New network infrastructure at CCIN2P3

Guillaume Cessieux – CCIN2P3 network team
Guillaume . Cessieux @ cc.in2p3.fr

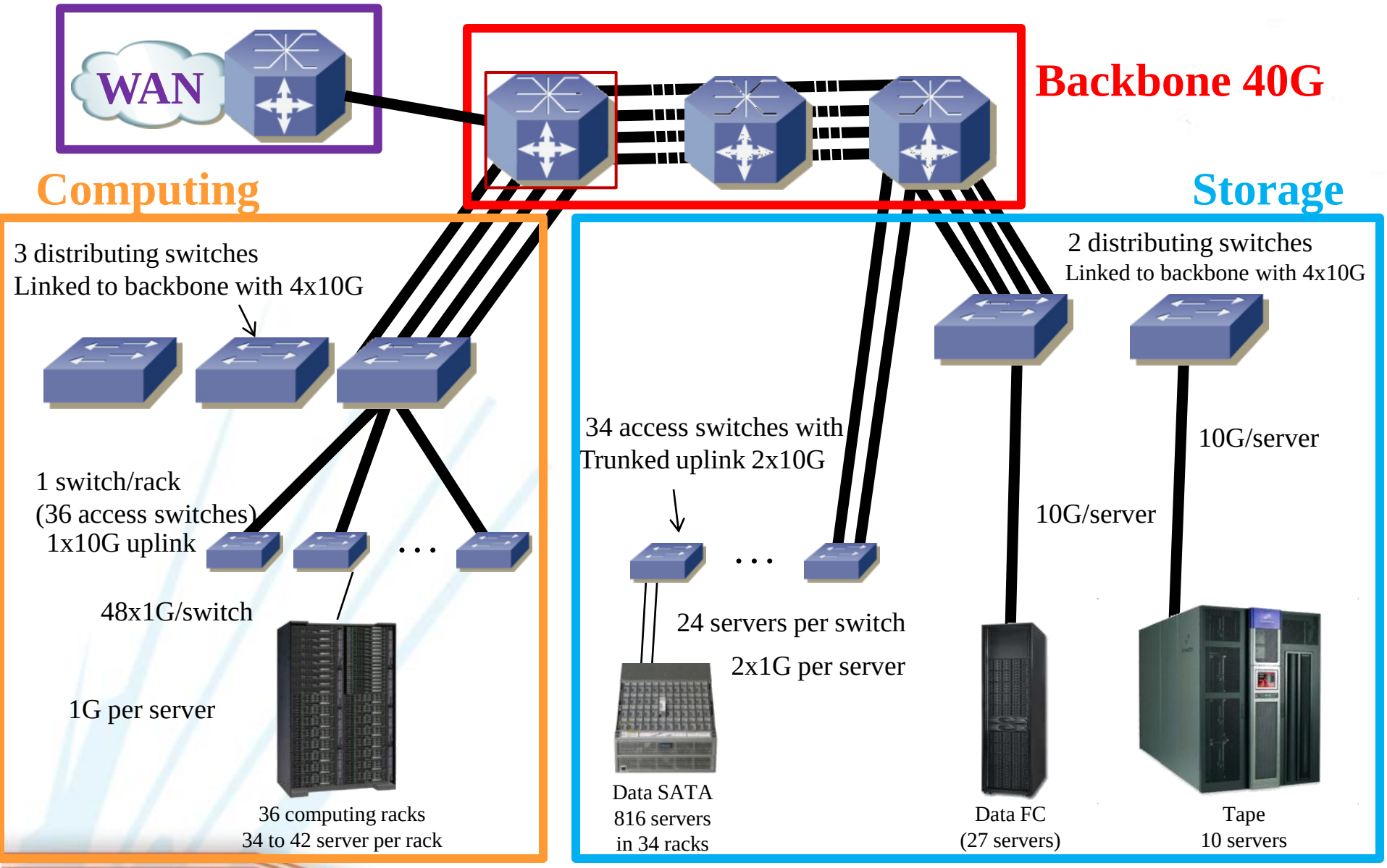
HEPiX Fall 2010, Cornell University, 2010-11-03

Latest major network upgrade: Mid 2009

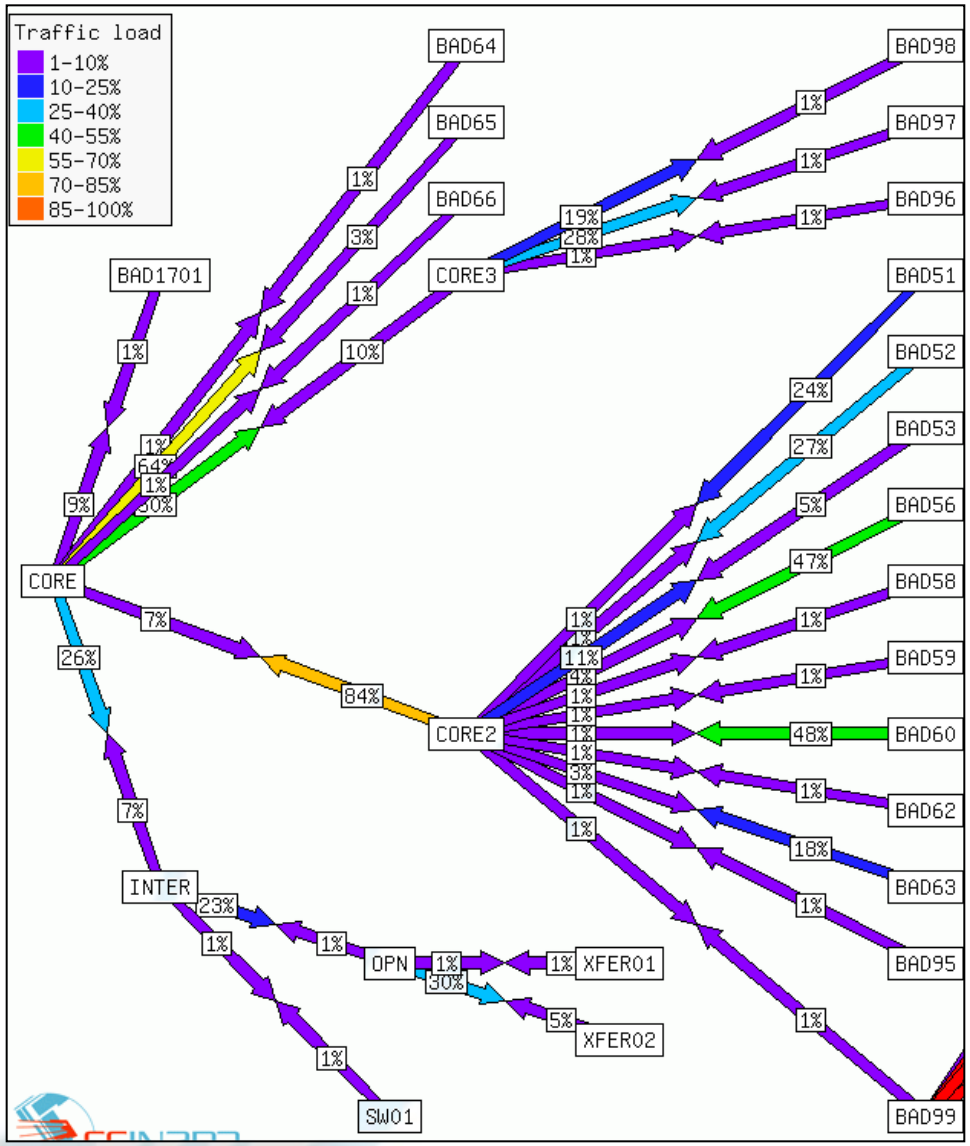
- Backbone 20G → 40G
  - No topology change, only additional bandwidth
- Linking abilities tripled
  - Distribution layer added

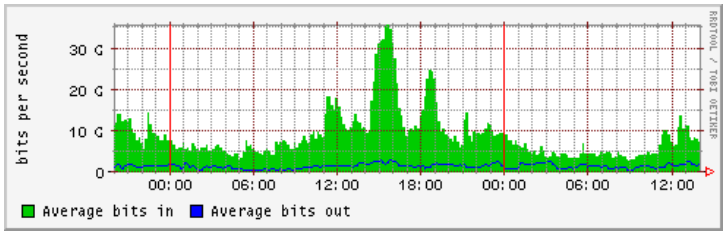


Previous network architecture

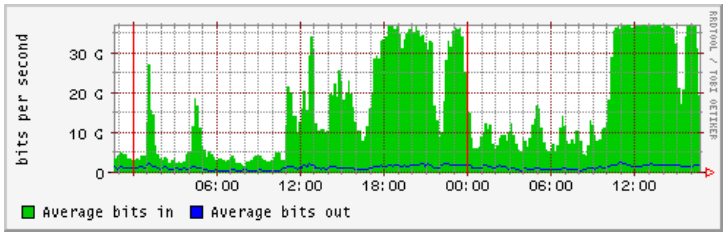


Reaching limits (1/2)

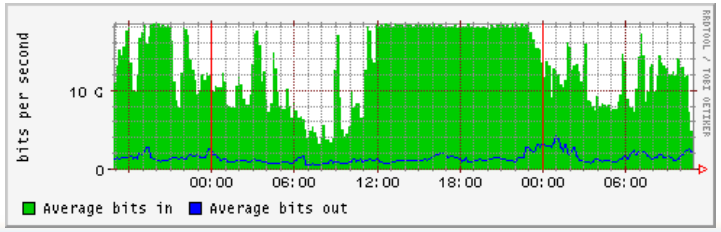




Same 40G path 1 year later



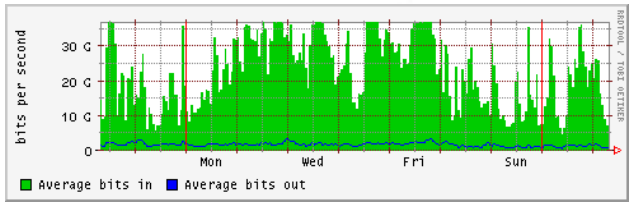
20G uplink of a distribution switch:



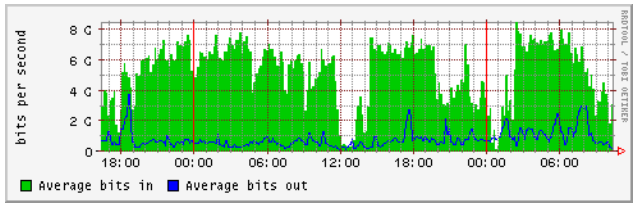
Reaching limits (2/2)

- Clear traffic increase
  - More hosts exchanging more
  - More remote exchanges

Usage of one 40G backbone etherchannel

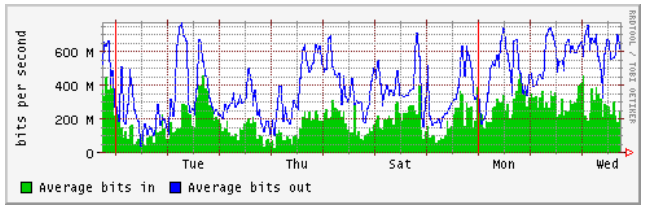


10G direct link with CERN



- Limits appearing
  - Disturbing bottlenecks
  - Long path before being routed

Sample host average: 620M on 1G



- Upcoming challenges
  - New computing room, massive data transfers, virtualization, heavy Grid computation

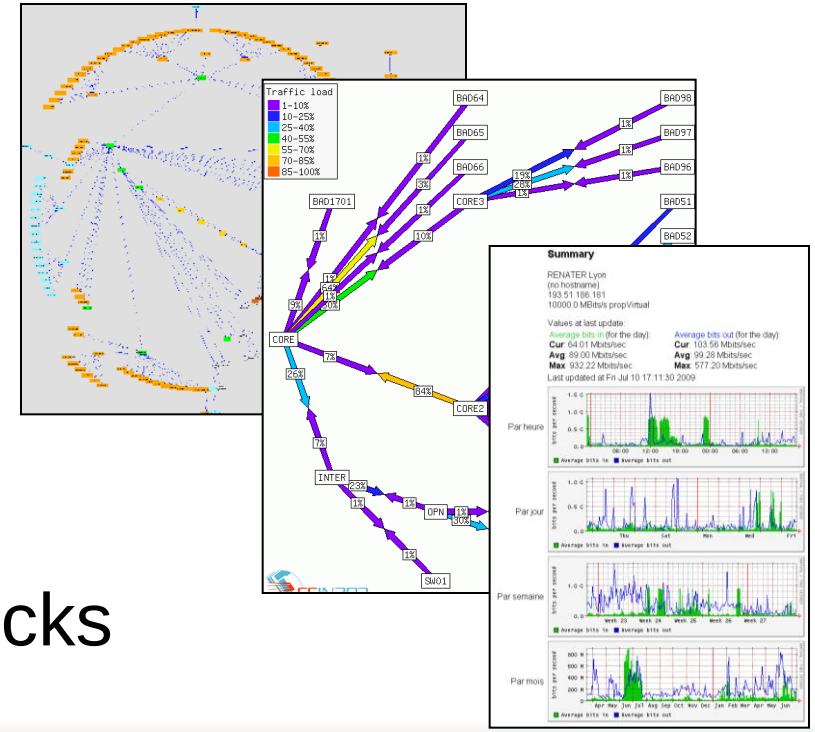
Complete network analysis performed

- Inventory
  - 20 network devices found
    - Thanks discovery protocols…
  - Software and features not harmonised

```
switch>show cdp neighbors
Capability Codes: R - Router, T - Trans Bridge, B - Source Route Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater, P - Phone,
                  D - Remote, C - CVTA, M - Two-port Mac Relay

Device ID        Local Intrfce    Holdtme    Capability  Platform  Port ID
s1.in2p3.fr.
                 Ten 2/1          146        R S I  WS-XXXXX-  Ten 1/50
s2.in2p3.fr.
                 Ten 3/3          130        R S I  WS-XXXXX-  Ten 7/6
s3.in2p3.fr.
                 Ten 3/4          150        R S I  WS-XXXXX-  Ten 6/6
```

- Topology
  - A map worth anything



- Usage
  - Traffic patterns, bottlenecks

Requirements for new network architecture

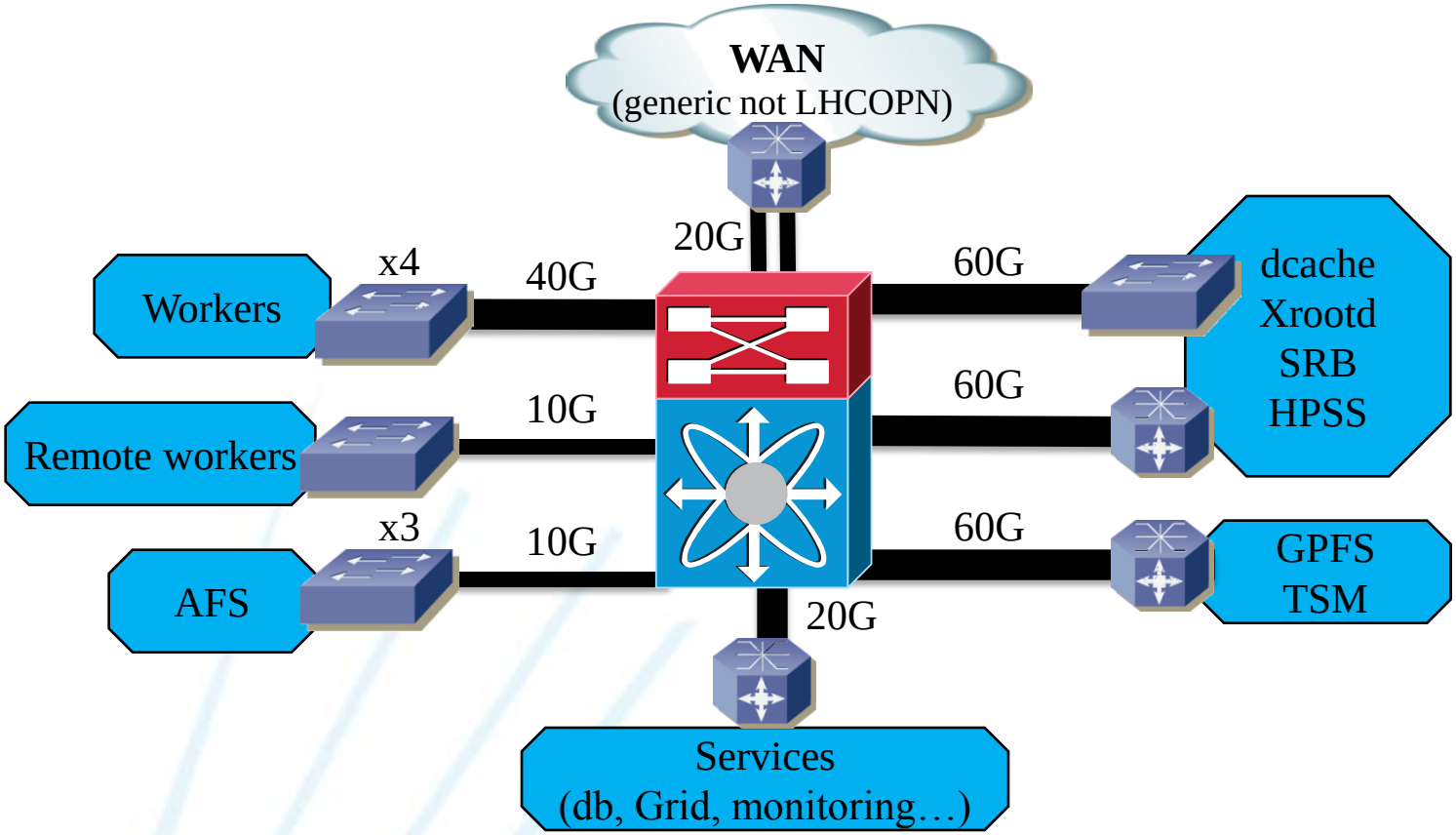
- More bandwidth!
- Able to scale for next years
  - Allowing non disruptive network upgrade
  - Particularly with new computing room

- Ease exchanges betwen major functional areas
- As usual: Good balance between risks, costs and requirements

Main directions

- Target non blocking mode
- No physical redundancy, meshes etc.
  - “A single slot failure in 10 years on big Cisco devices”
  - Too expensive and not worth for us
  - High availability handled at service level (DNS…)
  - Big devices preferred to meshed bunch of small
- Keep it simple
  - Ease configuration and troubleshooting
  - Avoid closed complex vendor solutions
    - e.g things branded “virtual”, “abstracted”, “dynamic”

New network architecture

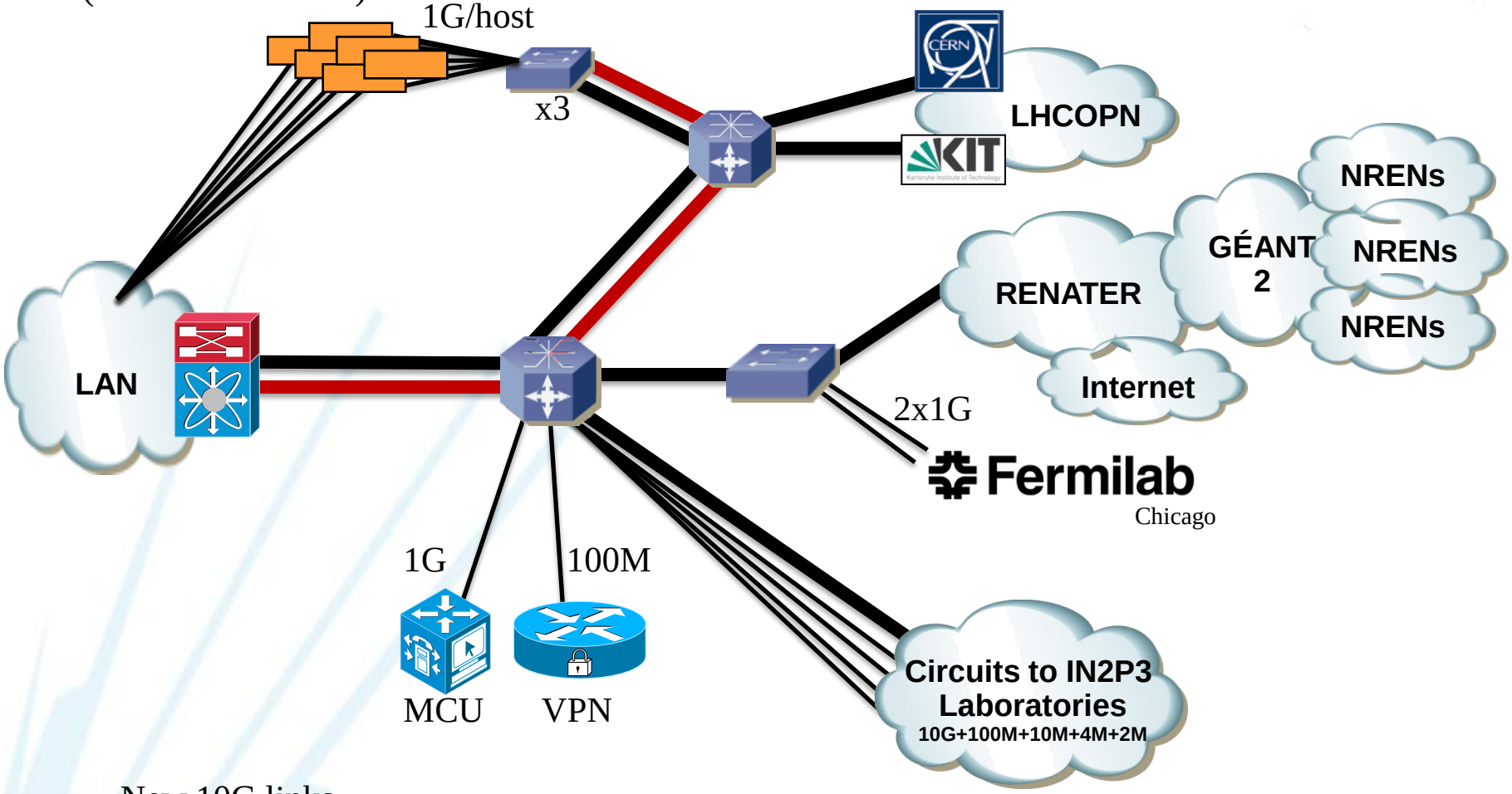


From 160G often shared to 400G wirespeed

Area	Old bandwidth	New bandwidth
AFS	30G **shared**	30G
GPFS TSM	40G **shared**	60G
Dcache Xrootd srb HPSS	40G **shared**	120G
Workers	40G **shared**	170G
WAN	10G	20G

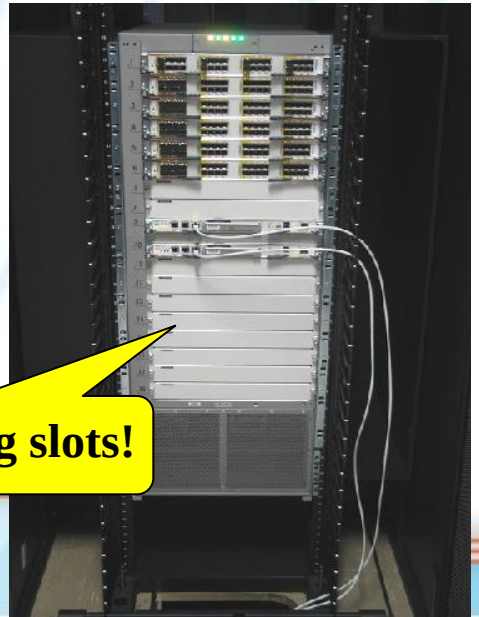
WAN upgrade

Dual homing of hosts doing massive data transfers (dcache for WLCG)

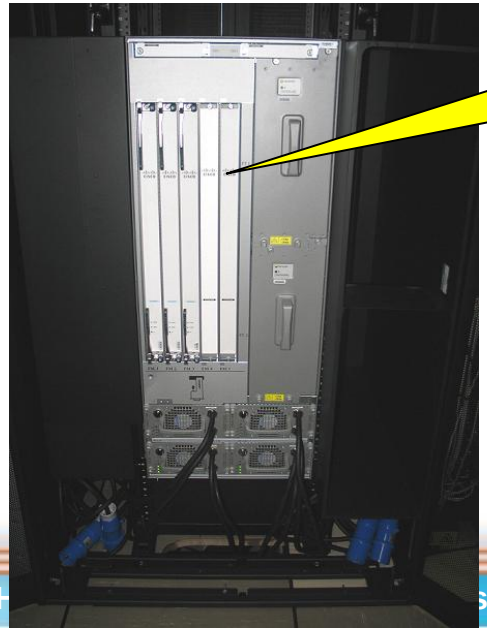


Main network devices and configurations used

- New core device: Nexus 7018
  - High density device, really scalable
  - Very modular: slots & switching engine cards
  - 80G backplane per slot (8x10G non blocking)
    - Initial configuration: 6 slots, 3 switching engines (3x48G)
  - This device is **vital**
    - 4 power supplies on 2 UPS, 2 managements slots



10 remaining slots!



2 extra switching engines possible

Compatibility check is done:

Mod	*boot*	*Impact*	*Install-type*
1	*yes*	***non-disruptive***	*rolling*
2	*yes*	***non-disruptive***	*rolling*
3	*yes*	***non-disruptive***	*rolling*
4	*yes*	***non-disruptive***	*rolling*
5	*yes*	***non-disruptive***	*rolling*
6	*yes*	***non-disruptive***	*rolling*
9	*yes*	***non-disruptive***	***reset***
10	*yes*	***non-disruptive***	***reset***

Main network devices and configurations used

6513

- 24x10G (12 blocking)
  + 96x1G
  + 336x1G blocking (1G/8ports)
- 48x10G (24 blocking)
  + 96x1G

6509

- 64x10G (32 blocking)

Core & Edge

4900

16x10G

Distribution

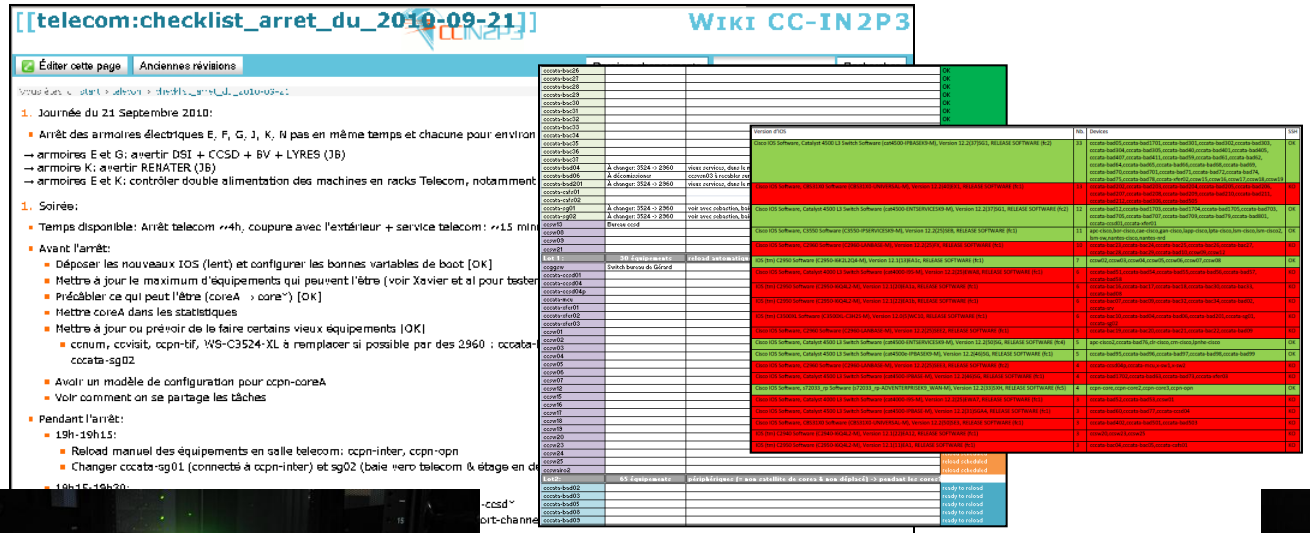
4948

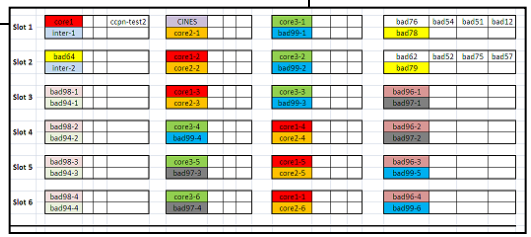
48x1G + 2x10G

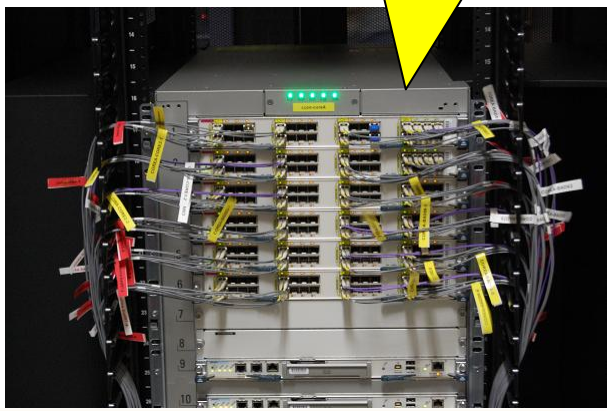
Access

A 3 months preparation

- Testing, preconfiguring, scripting, checklist
- Optical wiring: Fully done and tested before

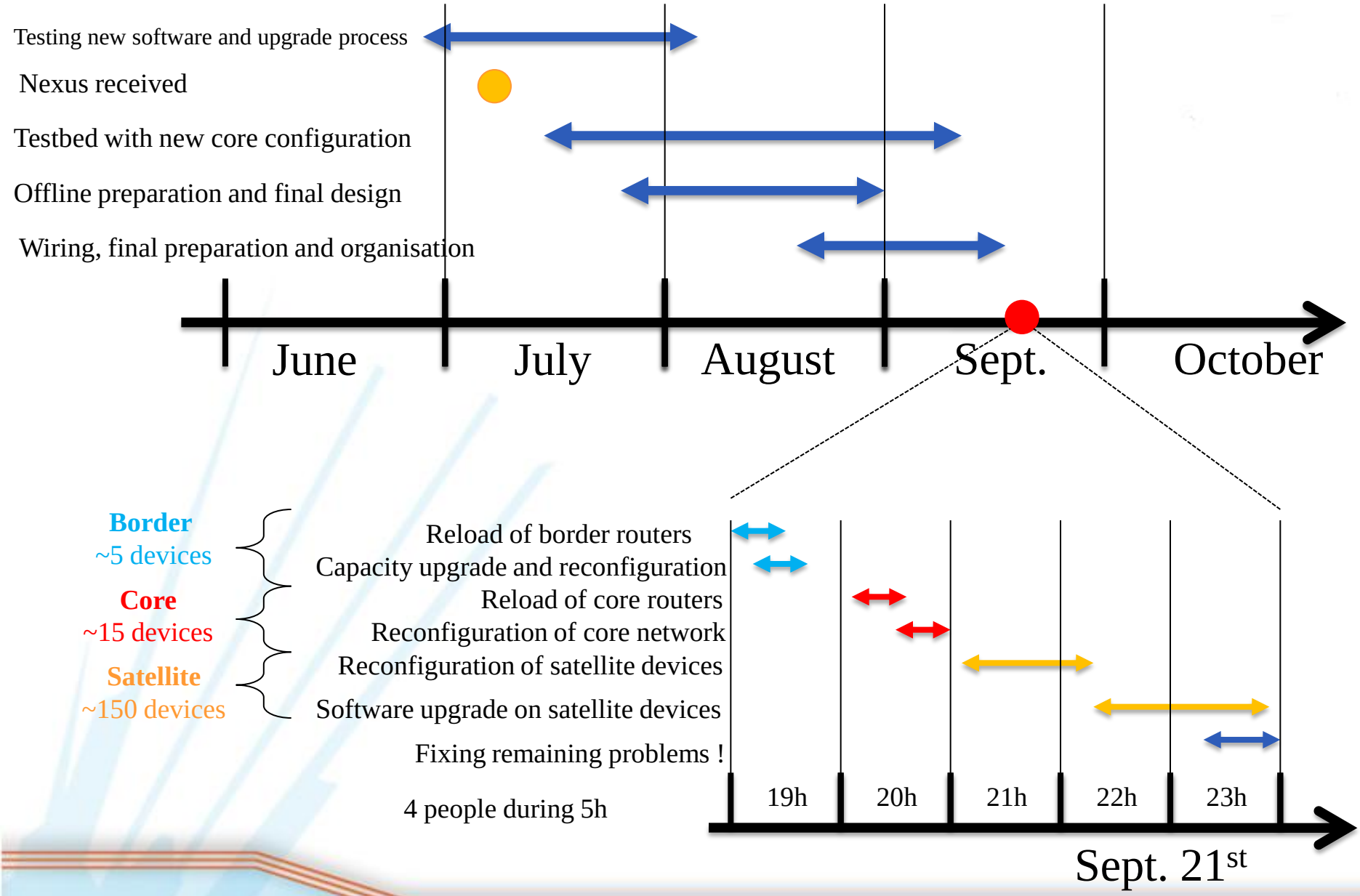






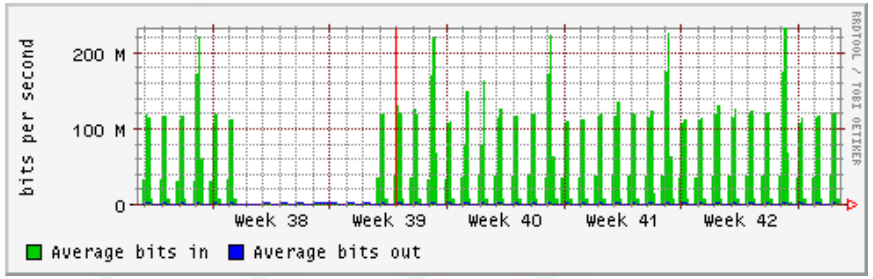
~60 new fibres
> 1km of fibre deployed!

Timeline


Testing new software and upgrade process
Nexus received
Testbed with new core configuration
Offline preparation and final design
Wiring, final preparation and organisation
June
July
August
Sept.
October
Border
~5 devices
Core
~15 devices
Satellite
~150 devices
Reload of border routers
Capacity upgrade and reconfiguration
Reload of core routers
Reconfiguration of core network
Reconfiguration of satellite devices
Software upgrade on satellite devices
Fixing remaining problems !
4 people during 5h
19h
20h
21h
22h
23h
Sept. 21st

Feedbacks

- We get no major surprise!
  - Heavy testing phase was fruitful
  - Main issue: Some routes not correctly announced
    - Not detected nor understood, but workaround found



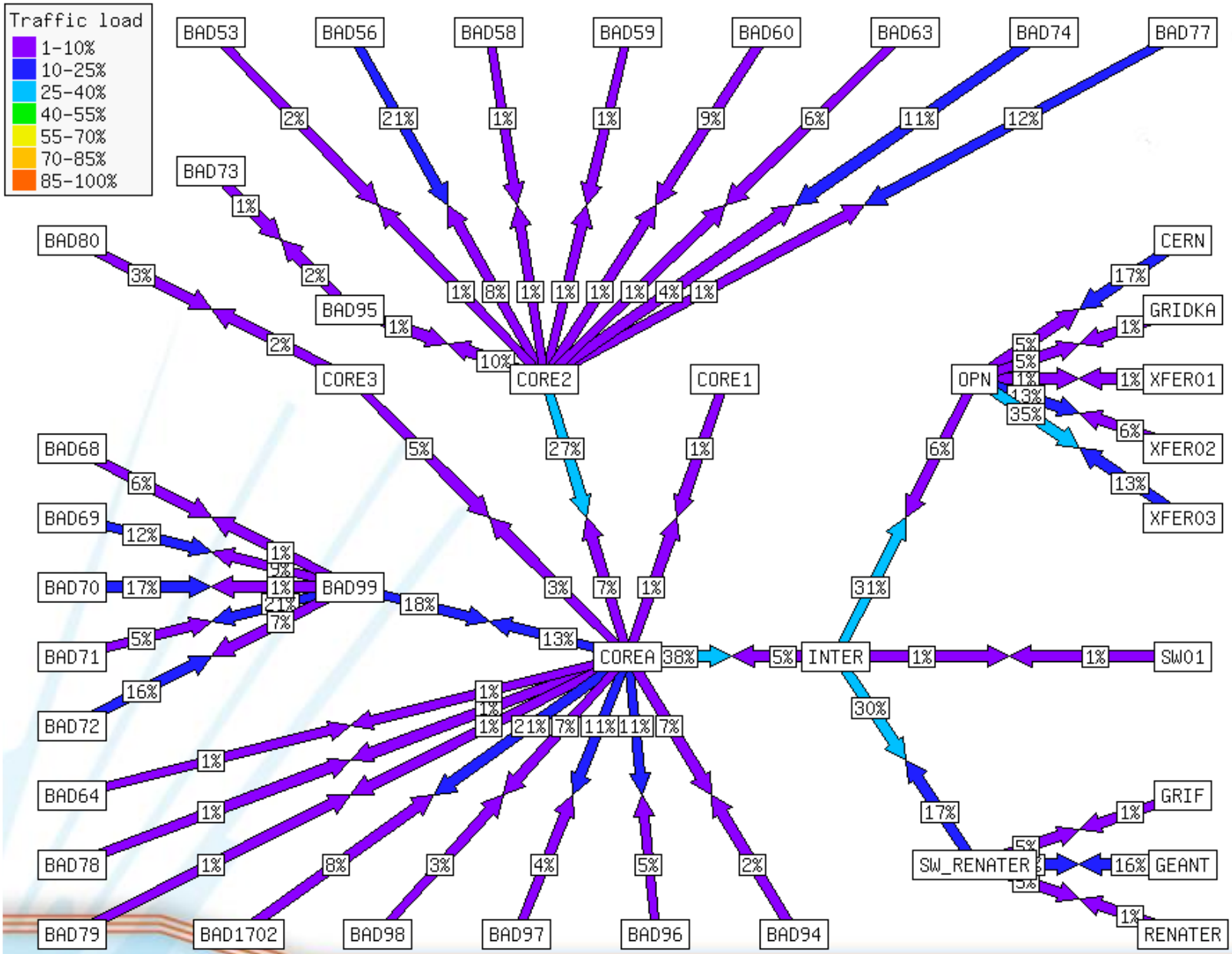
- Routing resilience hiding such problem
- Snapshot routing tables with a traceroute to each route!

- Keep monitoring, but deactivate alarms
  - Spare 800 SMS and 6k e-mails to each team members
- Do not parallelize actions too much
  - Hard to isolate faults or validate actions

Key benefits

- Increased core capacity by 2.5
- Isolated areas, delivering wire speed, removed bottlenecks, shortened paths
- Seamless capacity upgrade now possible
- Harmonised softwares and features on 170 devices
  - From 37 different versions to 13

Current status

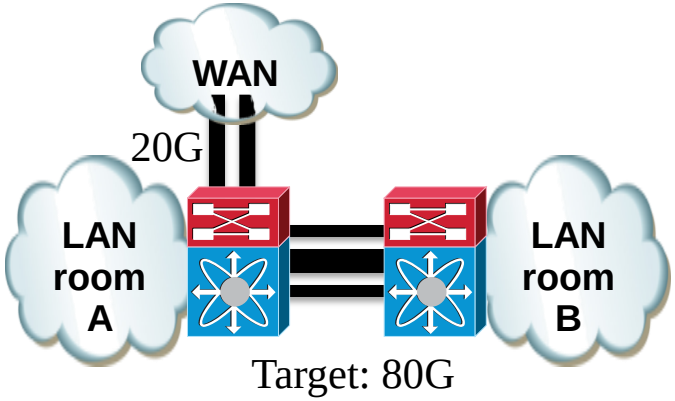


Upcoming

- **100G** tests

This part was for eyes only

- Plug new computing room
  - Switch as far as you can
    - 80G trunk to a similar architecture
    - Not autonomous, this is an extension



Conclusion

- Flat to starred network architecture
  - Closely matching our needs
- Average network usage down from 60% to ~15%
- Ready to face traffic increase for some more time
  - How long?