Networks for ATLAS Trigger and Data Acquisition

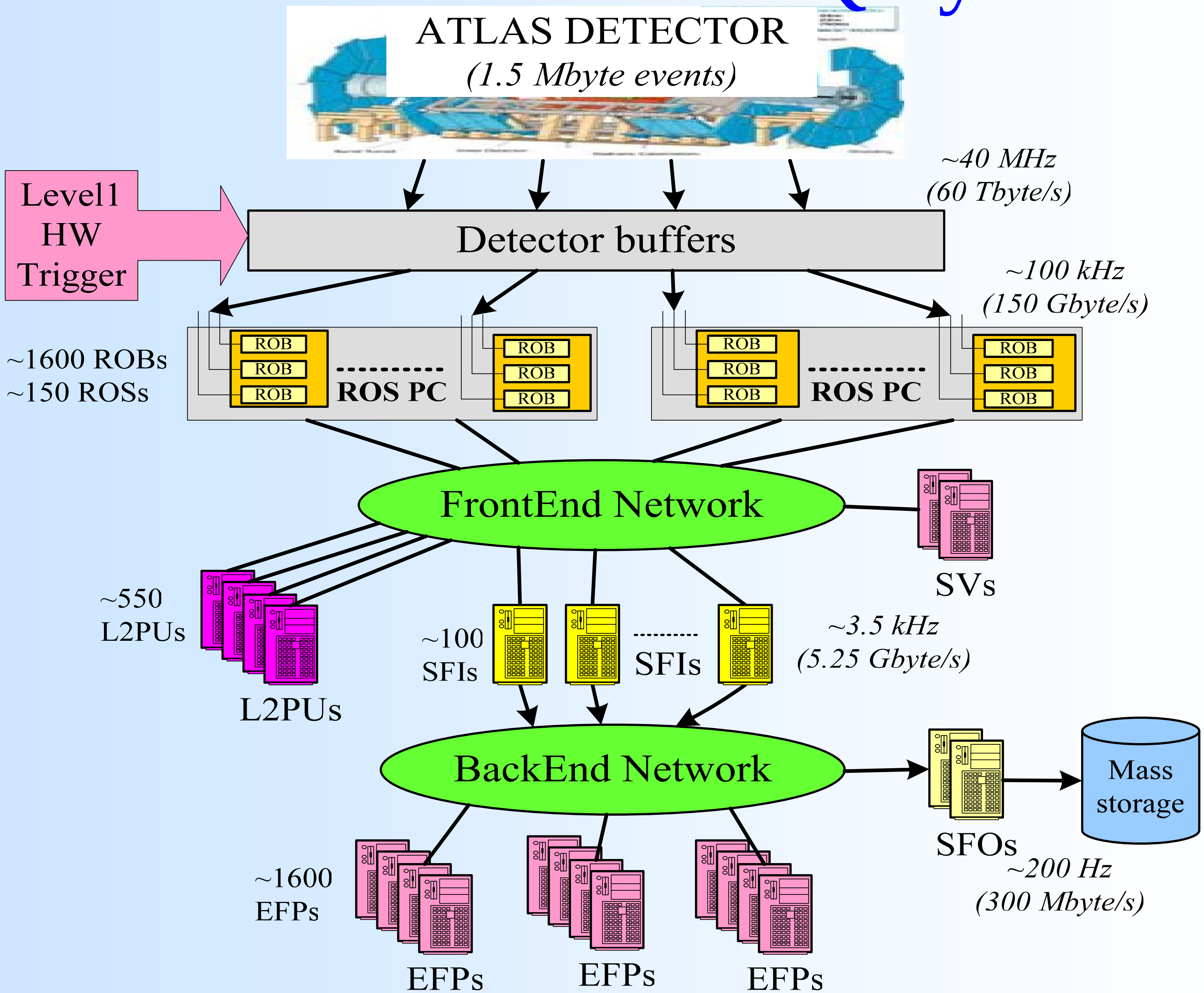
S. Stancu[*†], C. Meirosu[†‡] , M. Ciobotaru[*†], L. Leahu[†‡], B. Martin[†]

* University of California, Irvine
† CERN, Geneva
‡ “Politehnica” University of Bucharest

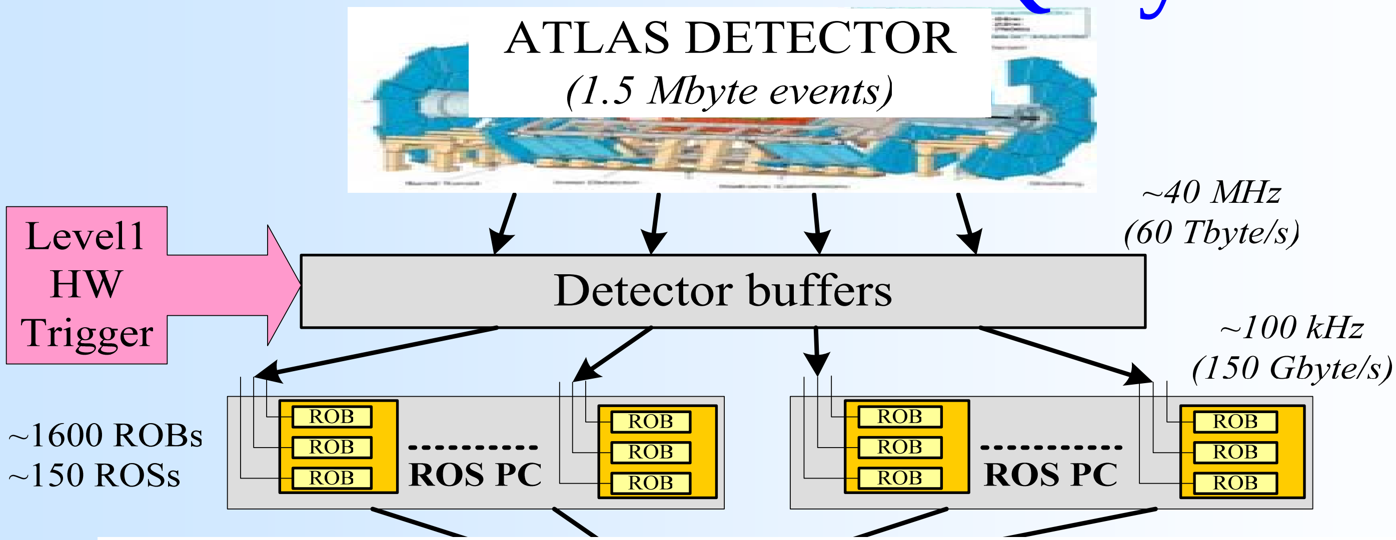
Outline

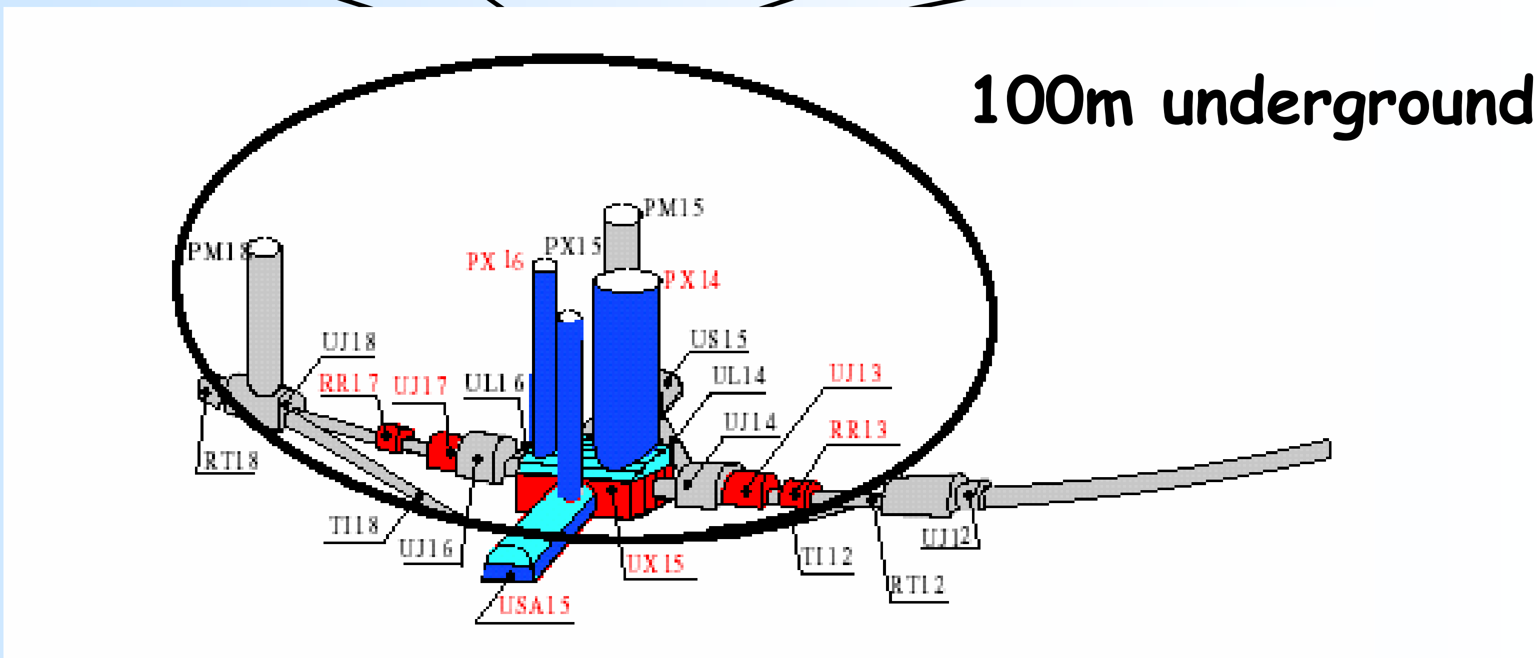
- Overview of the TDAQ system and networks
- Technology and equipment
- TDAQ networks:
  - Control network - no special bandwidth requirement
  - Dedicated data networks:
    - FrontEnd network – high bandwidth (~100Gbit/s cross-sectional bw.) and minimal loss
    - BackEnd network – high bandwidth (~ 50Gbit/s cross-sectional bw.)
- Sample resiliency test
- Management/installation issues
  - Dedicated path for management and monitoring
  - Automatic topology/connectivity check
- Conclusions

The ATLAS TDAQ System

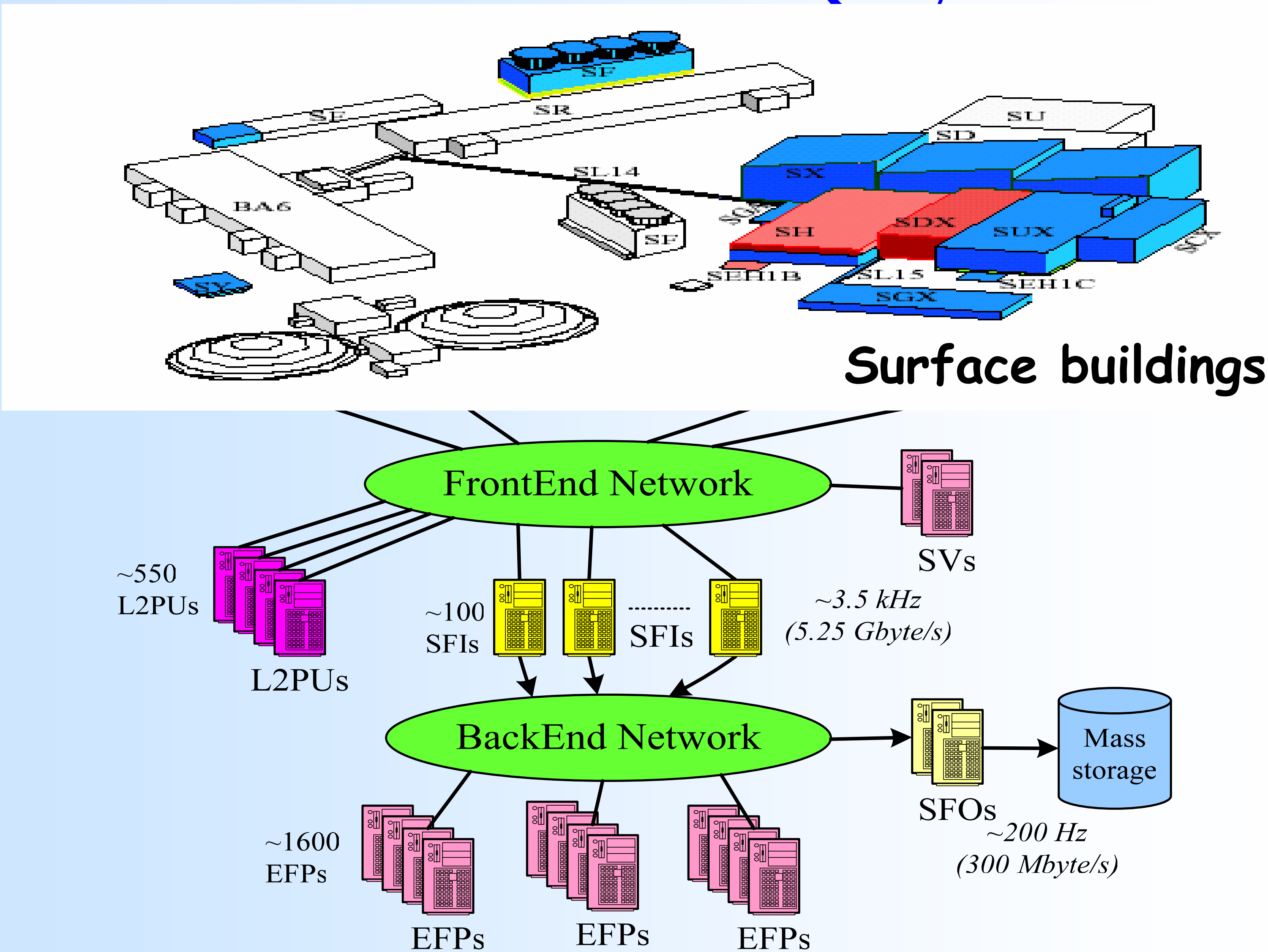


The ATLAS TDAQ System

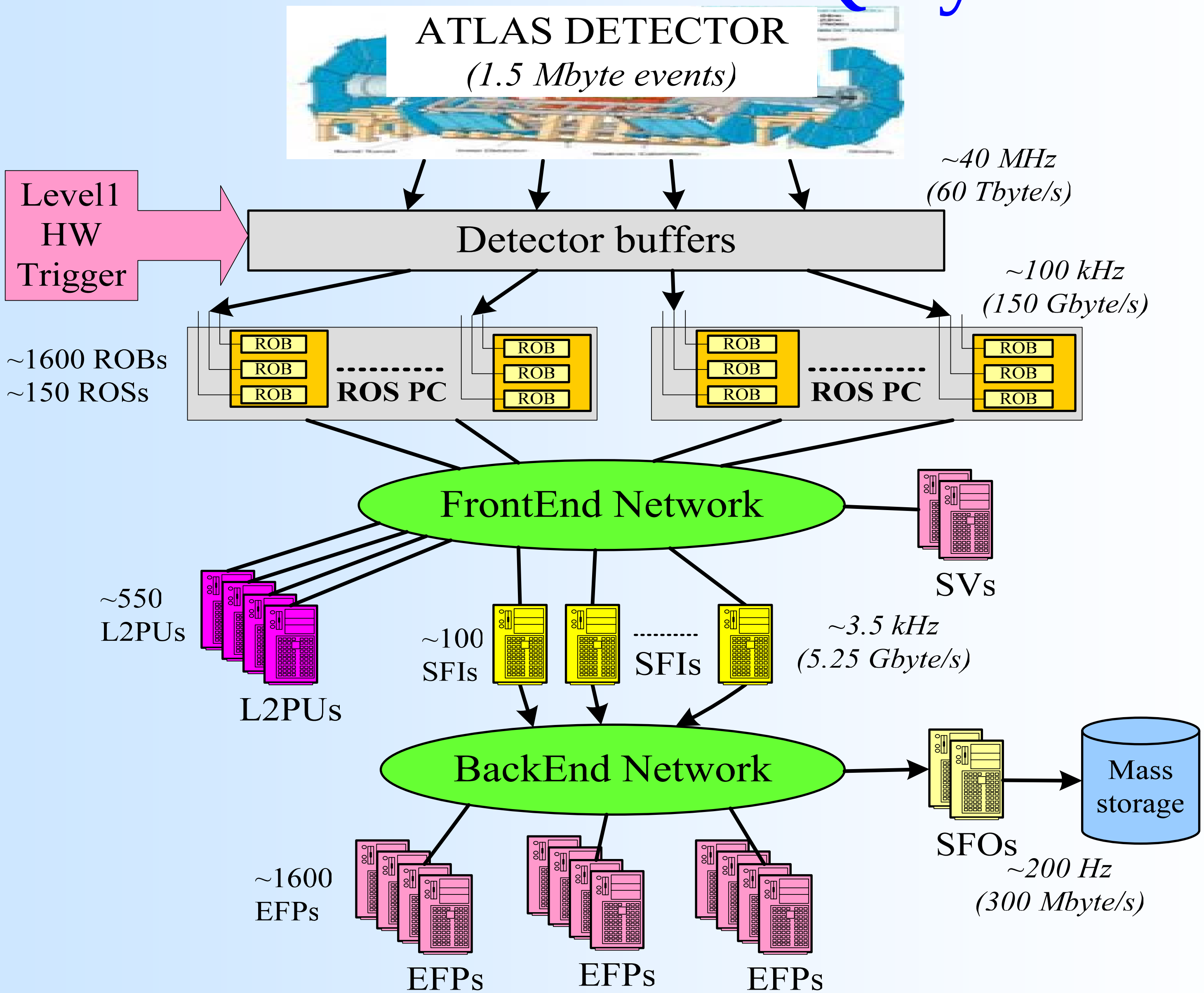




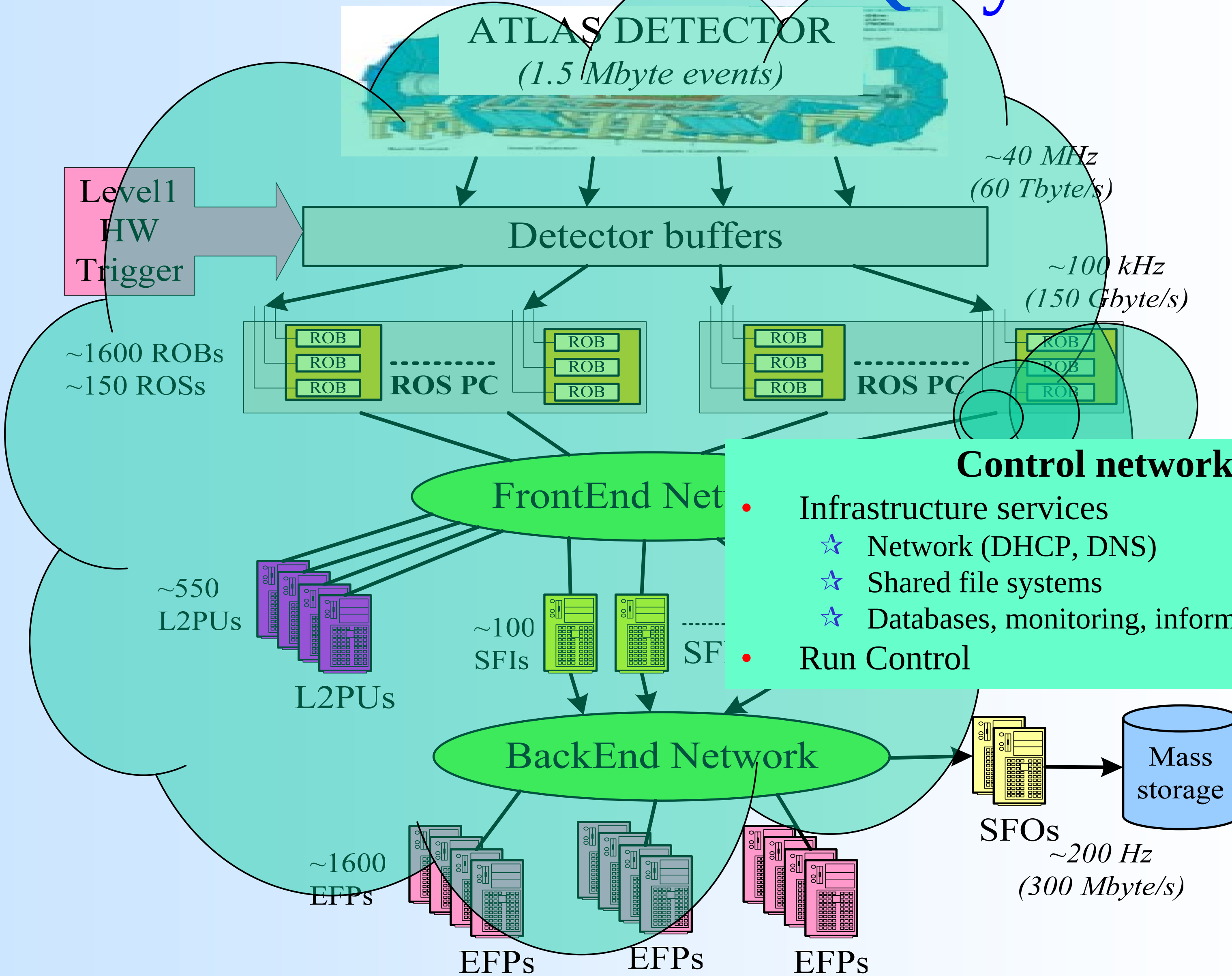
The ATLAS TDAQ System



The ATLAS TDAQ System



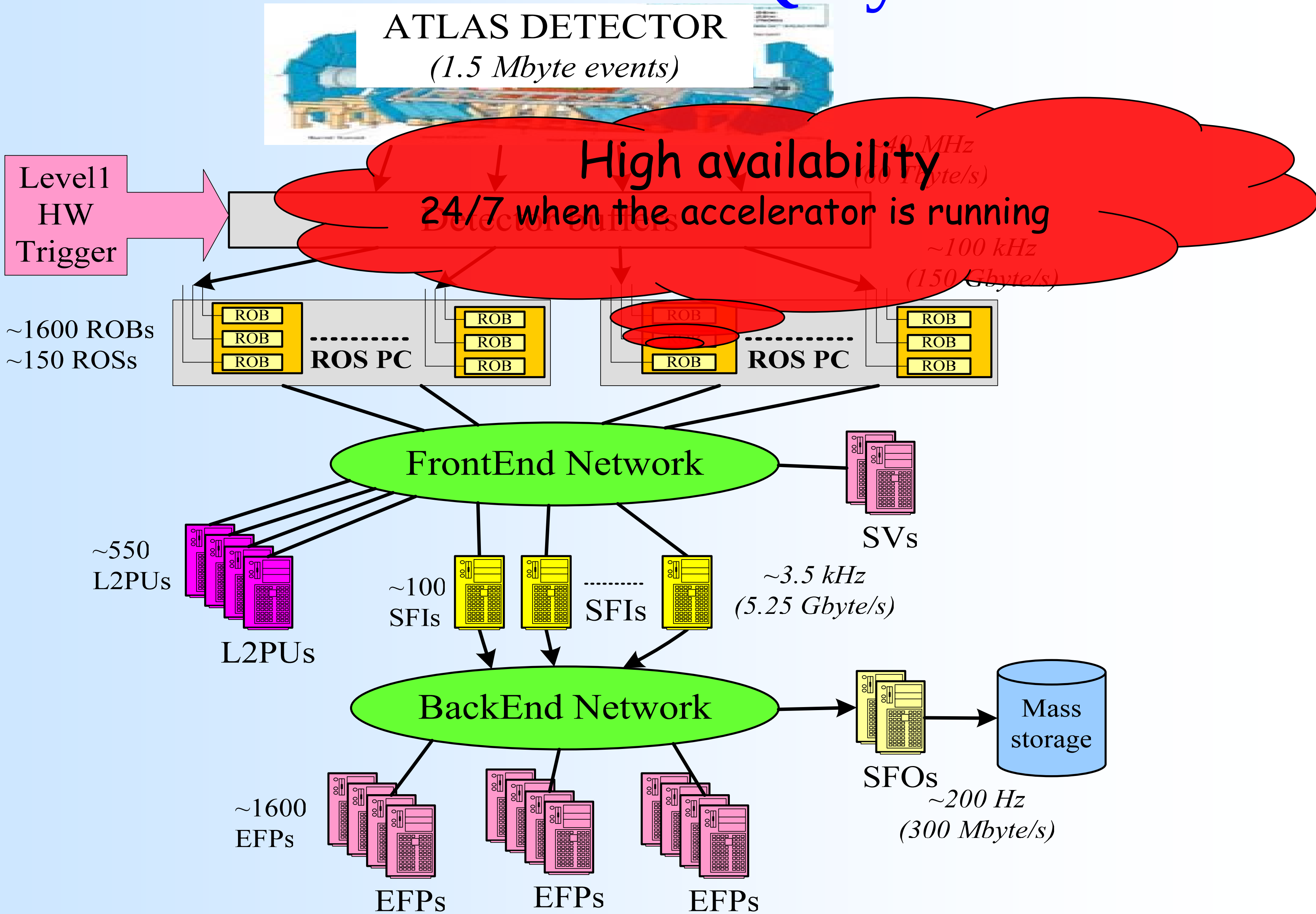
The ATLAS TDAQ System



Control network

- Infrastructure services
  - ☆ Network (DHCP, DNS)
  - ☆ Shared file systems
  - ☆ Databases, monitoring, information service, etc.
- Run Control

The ATLAS TDAQ System



Technology and equipment

- *Ethernet* is the dominant technology for LANs
  - TDAQ’s choice for networks (see [1])
    - multi-vendor, long term support, commodity (on-board GE adapters), etc.
  - Gigabit and TenGigabit Ethernet
    - Use GE for end-nodes
    - 10GE whenever the bandwidth requirements exceed 1Gbit/s
- Multi-vendor Ethernet switches/routers available on the market:
  - Chassis-based devices (~320 Gbit/s switching)
    - GE line-cards: typically ~40 ports (1000BaseT)
    - 10GE line-cards: typically 4 ports (10GBaseSR)
  - Pizza-box devices (~60 Gbit/s switching)
    - 24/48 GE ports (1000BaseT)
    - Optional 10GE module with 2 up-links (10GBaseSR)

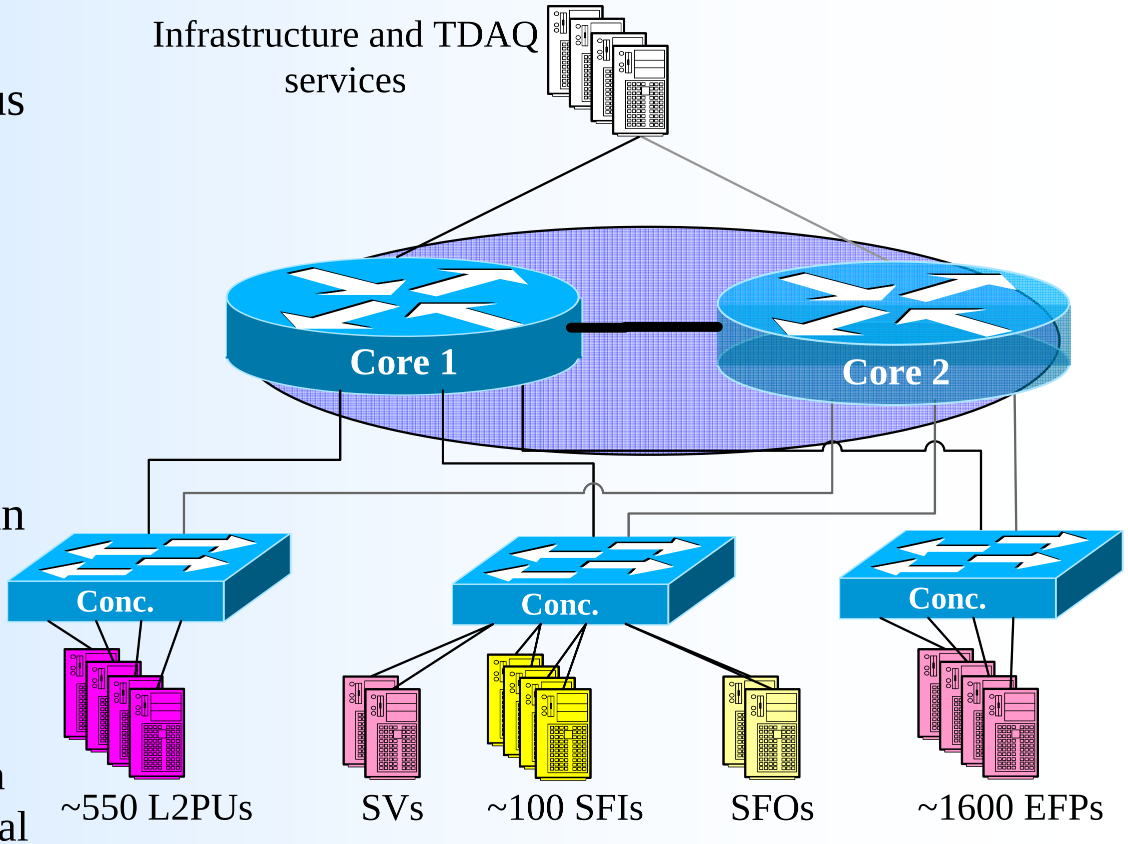
Resilient Ethernet networks

- What happens if a switch or link fails?
  - Phone call, but nothing critical should happen after a single failure.

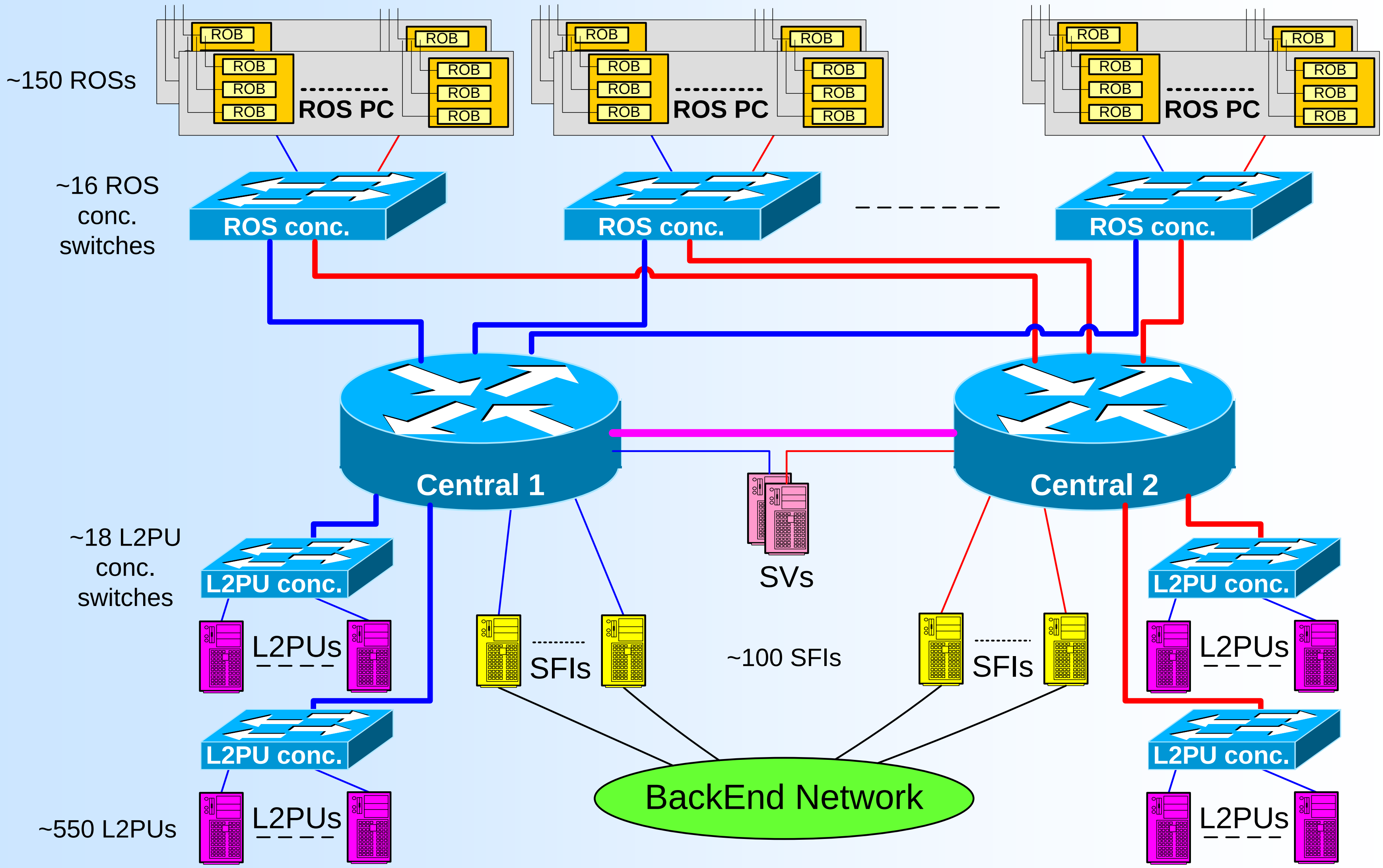
- Networks are made resilient by introducing redundancy:
  - *Component-level redundancy*: deployment of devices with built-in redundancy (PSU, supervision modules, switching fabric)
  - *Network-level redundancy*: deployment of additional devices/links in order to provide alternate paths between communicating nodes.
    - Protocols are needed to *correctly* (and *efficiently*) deal with multiple paths in the network [2]:
      - Layer 2 protocols: Link aggregation (trunking), spanning trees (STP, RSTP, MSTP)
      - Layer 3 protocols: virtual router redundancy (VRRP) for static environments, dynamic routing protocols (e.g. RIP, OSPF).

Control network

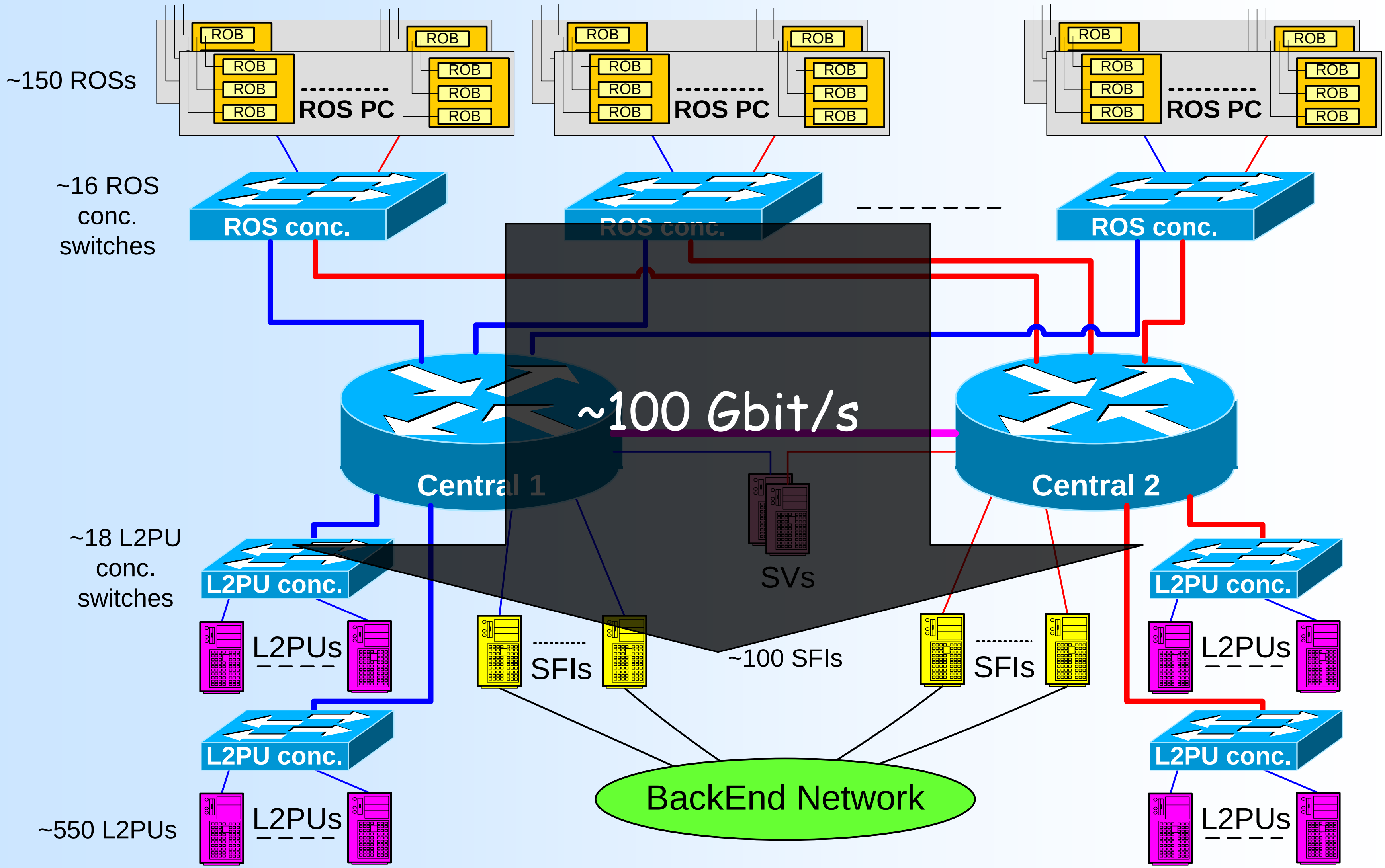
- ~3000 end nodes
- Design assumption: the instantaneous traffic does not exceed 1 Gbit/s on any segment, including up-link.
- One device suffices for the core layer, but better redundancy is achieved by deploying 2 devices.
- A rack level concentration switch can be deployed for all units except for critical services.
- Layer 3 routed network
  - ☆ One sub-net per concentrator switch
  - ☆ Small broadcast domains → potential layer 2 problems remain local.



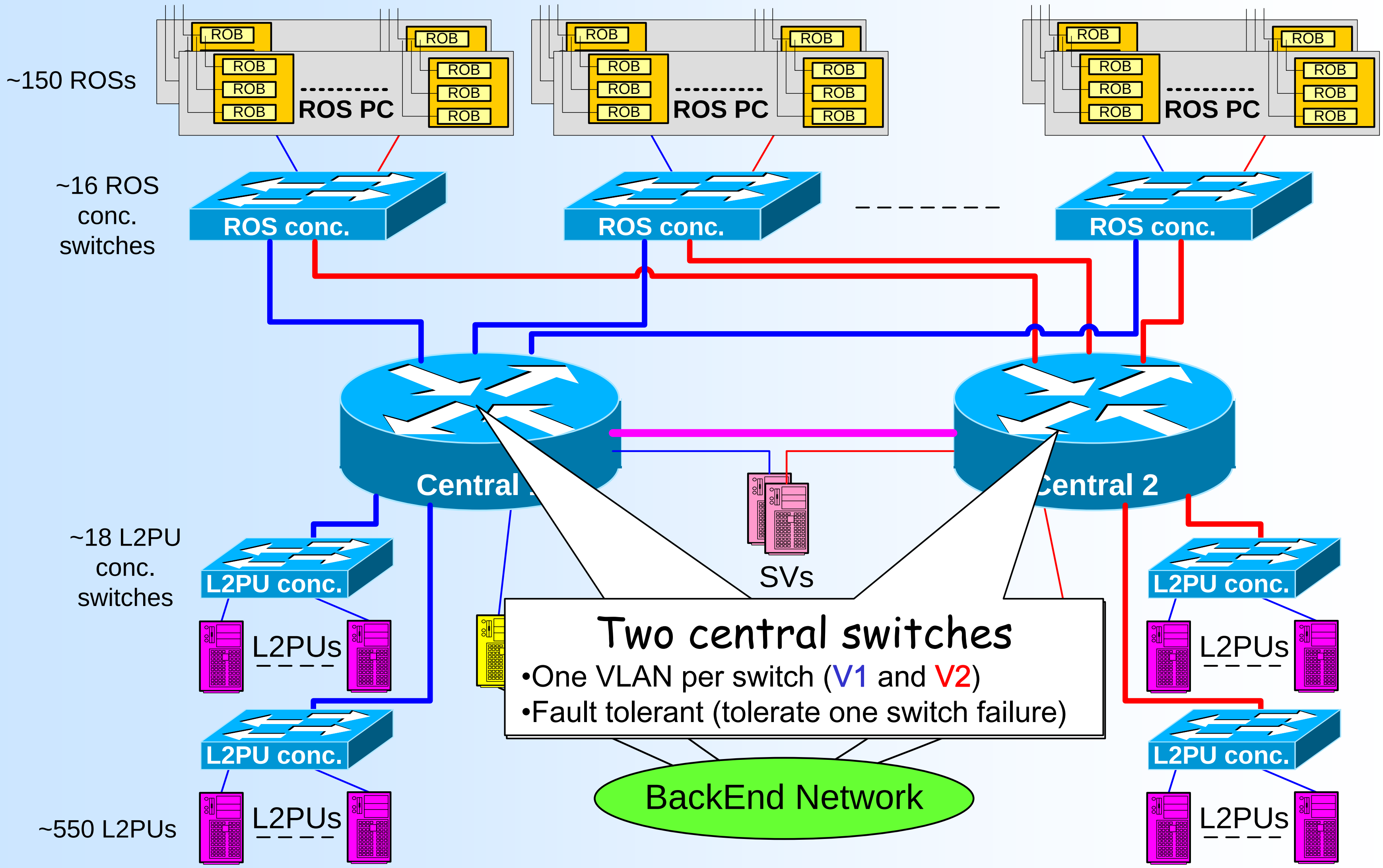
FrontEnd network (see [3])



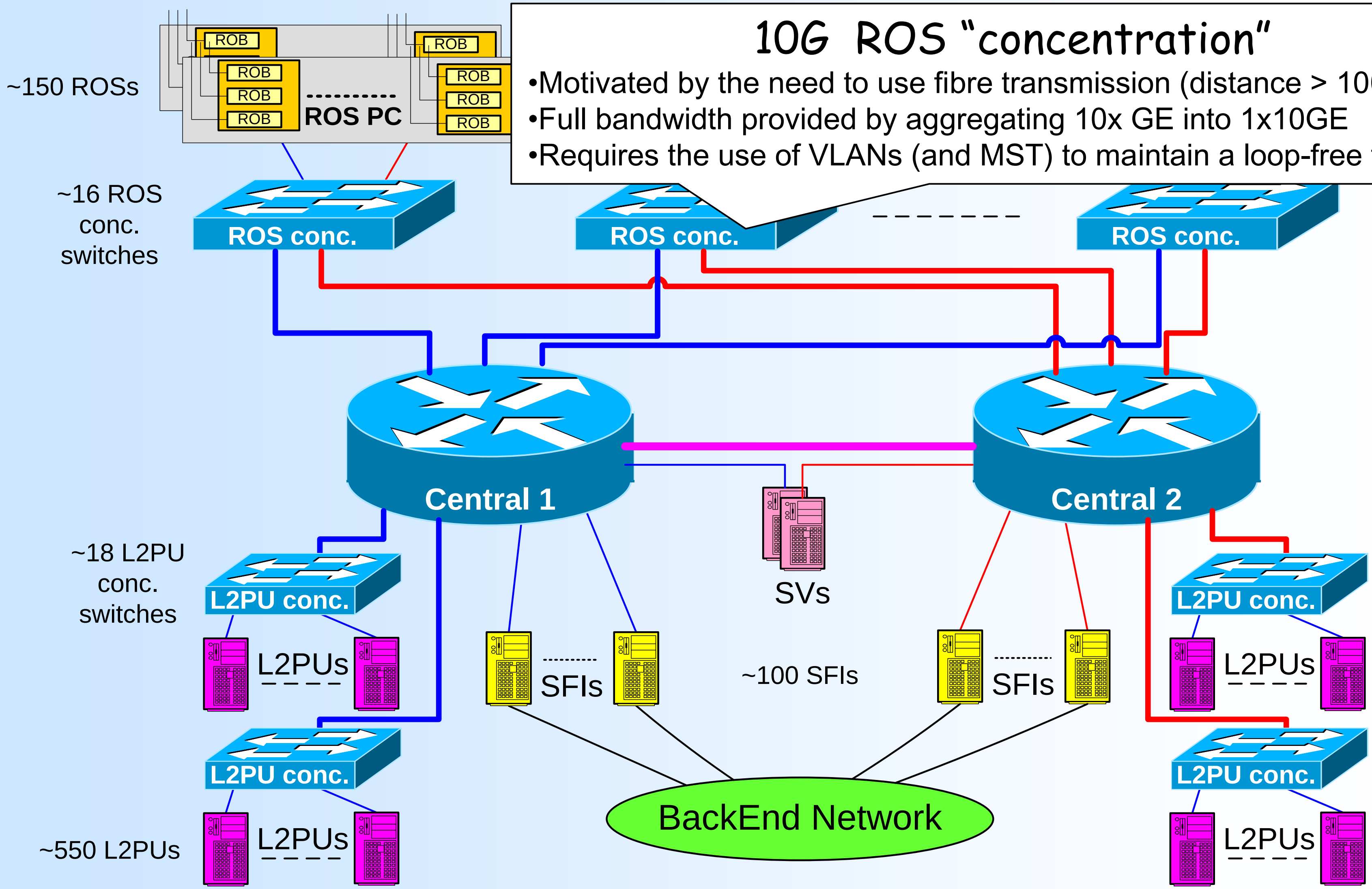
FrontEnd network



FrontEnd network



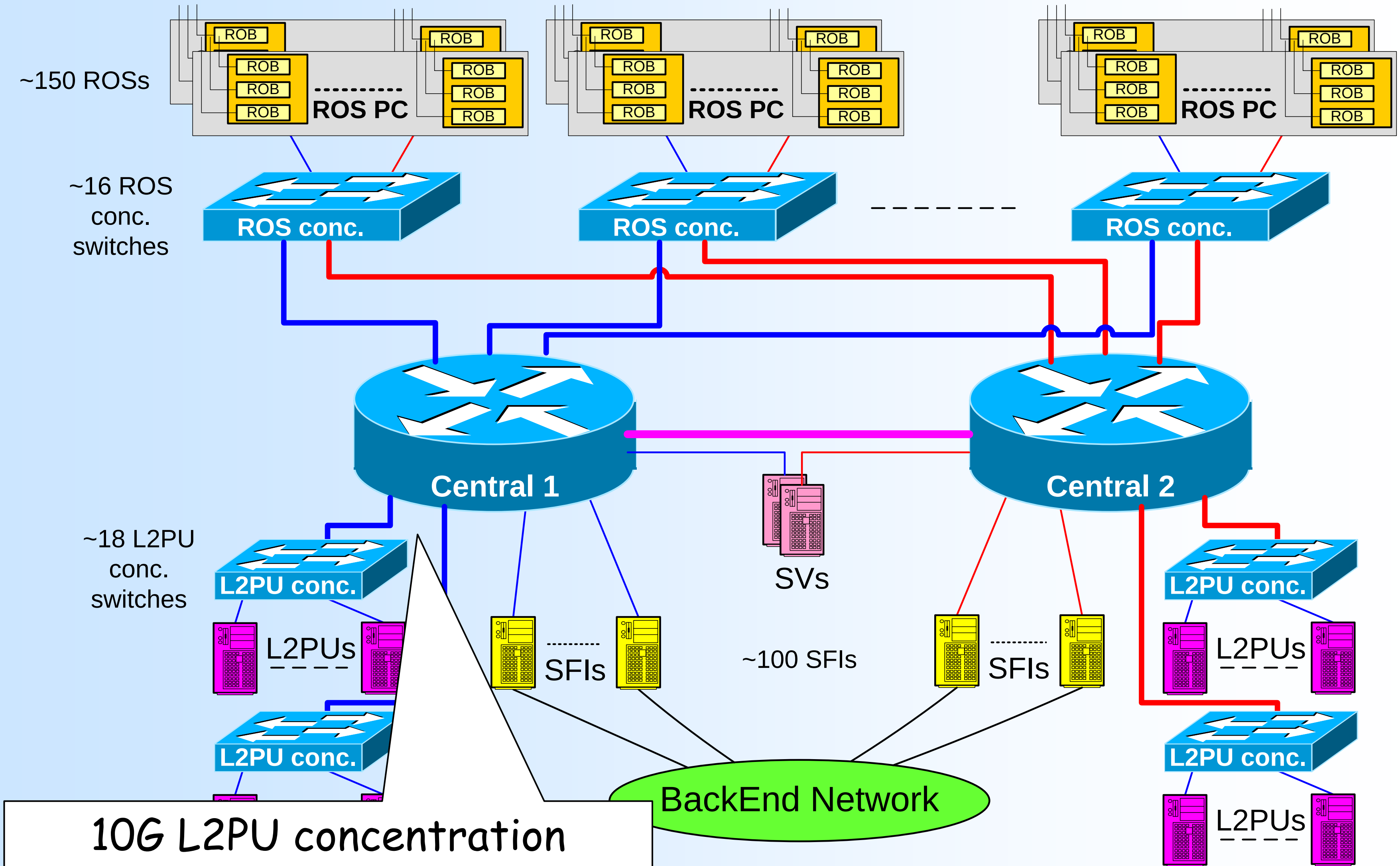
FrontEnd network



10G ROS "concentration"

- Motivated by the need to use fibre transmission (distance > 100m)
- Full bandwidth provided by aggregating 10x GE into 1x10GE
- Requires the use of VLANs (and MST) to maintain a loop-free topology

FrontEnd network

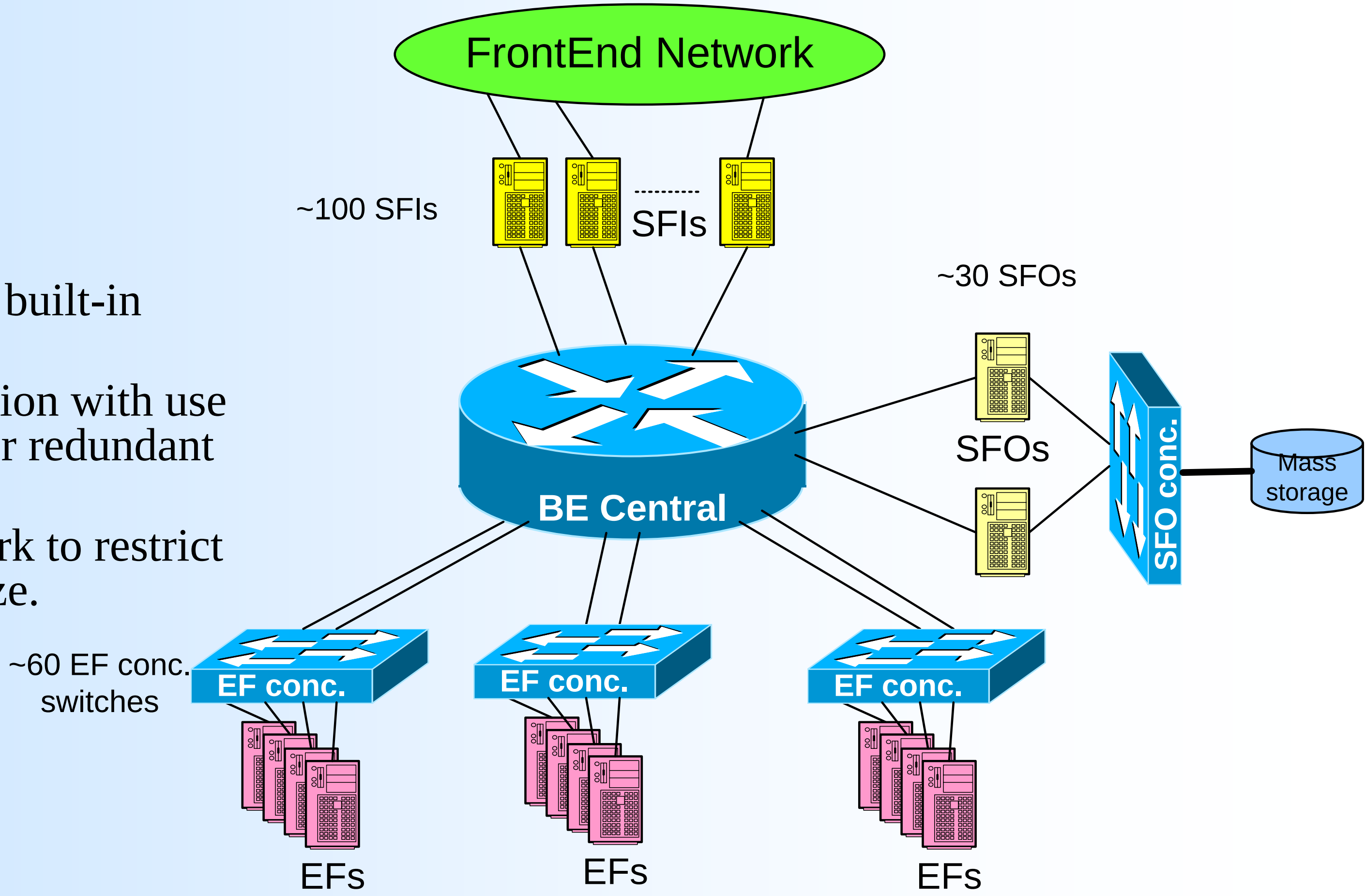


10G L2PU concentration

- One switch per rack

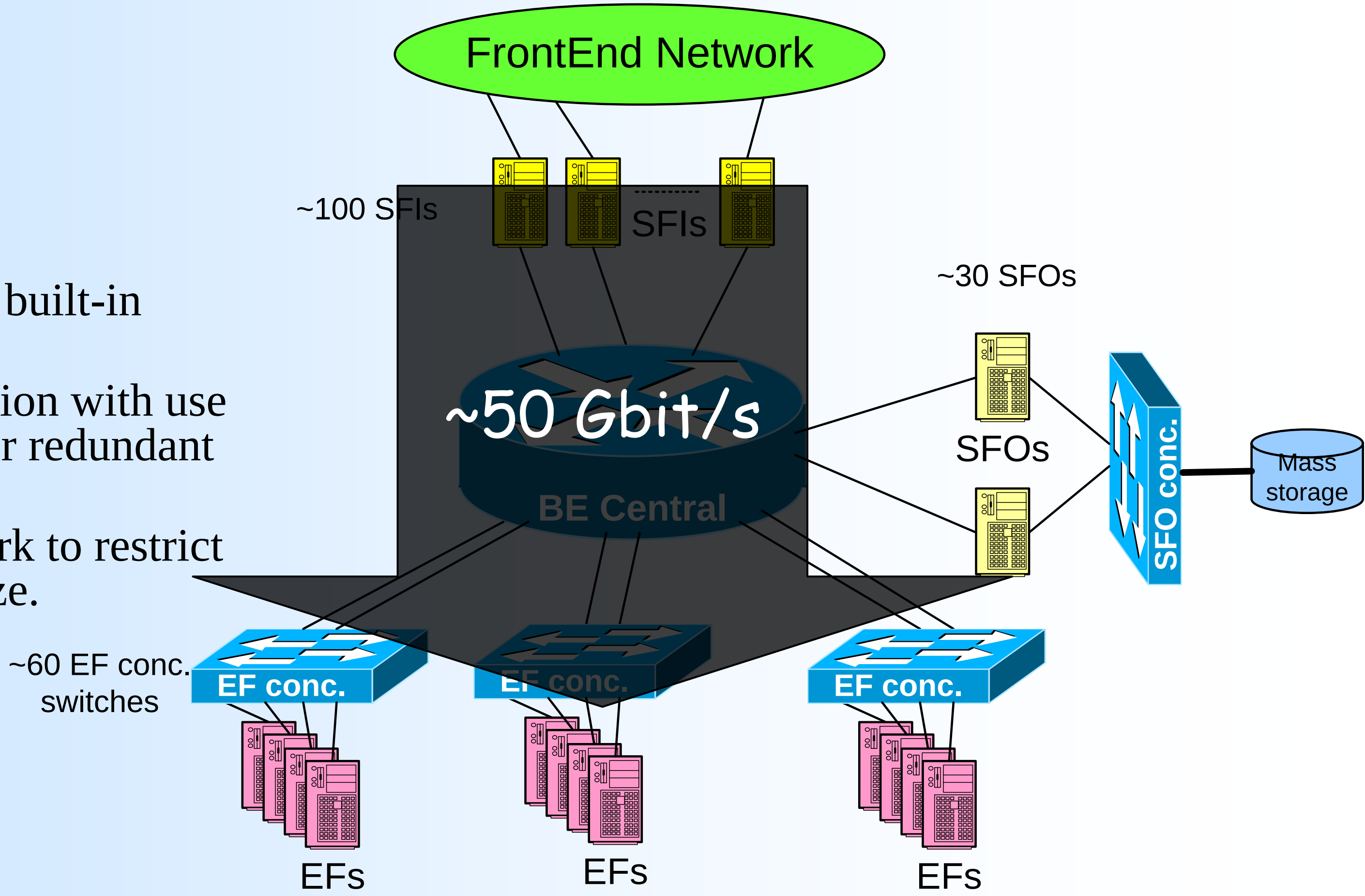
BackEnd network

- ~2000 end-nodes
- One core device with built-in redundancy
- Rack level concentration with use of link aggregation for redundant up-links to the core.
- Layer 3 routed network to restrict broadcast domains size.



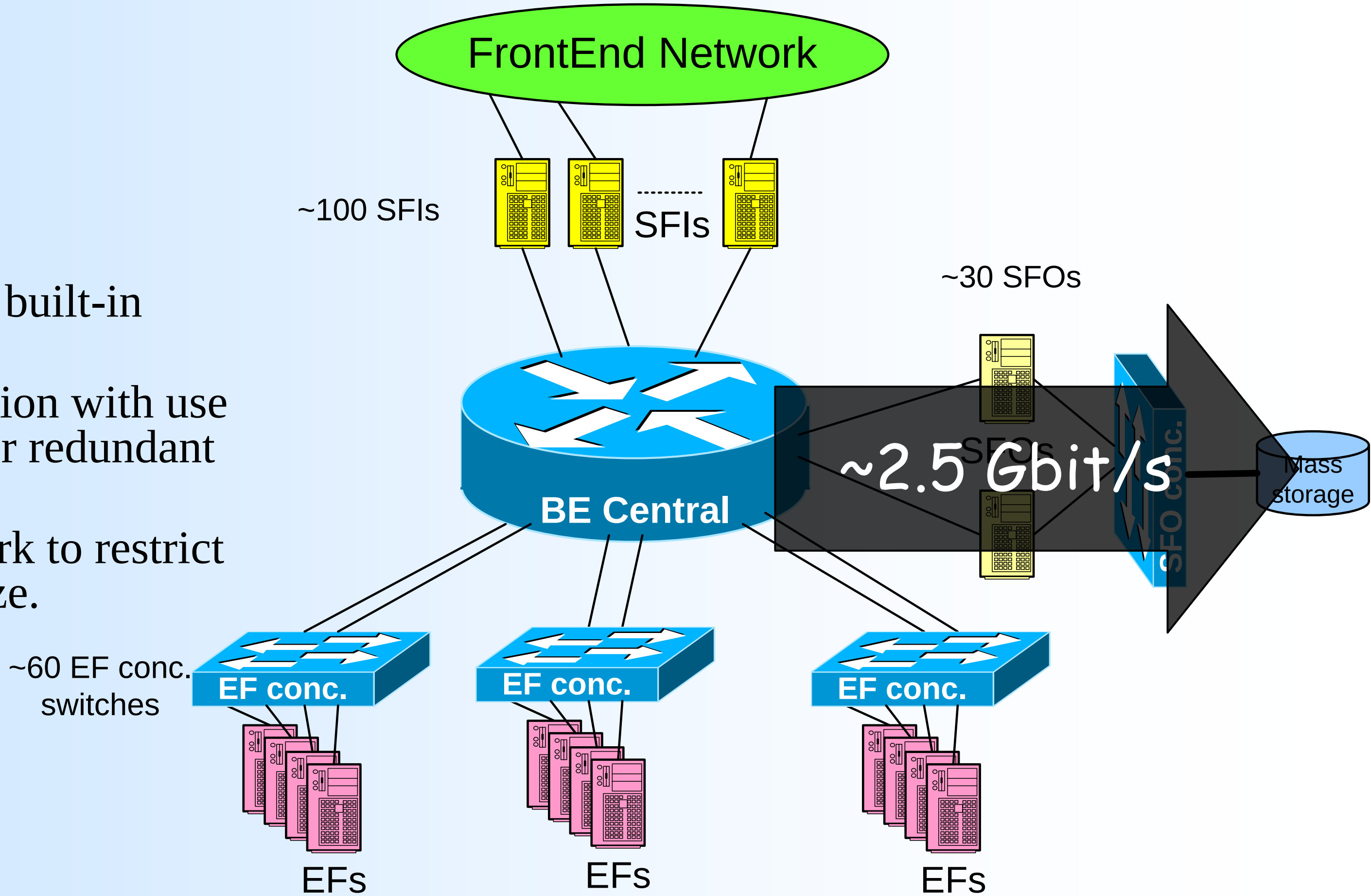
BackEnd network

- ~2000 end-nodes
- One core device with built-in redundancy
- Rack level concentration with use of link aggregation for redundant up-links to the core.
- Layer 3 routed network to restrict broadcast domains size.



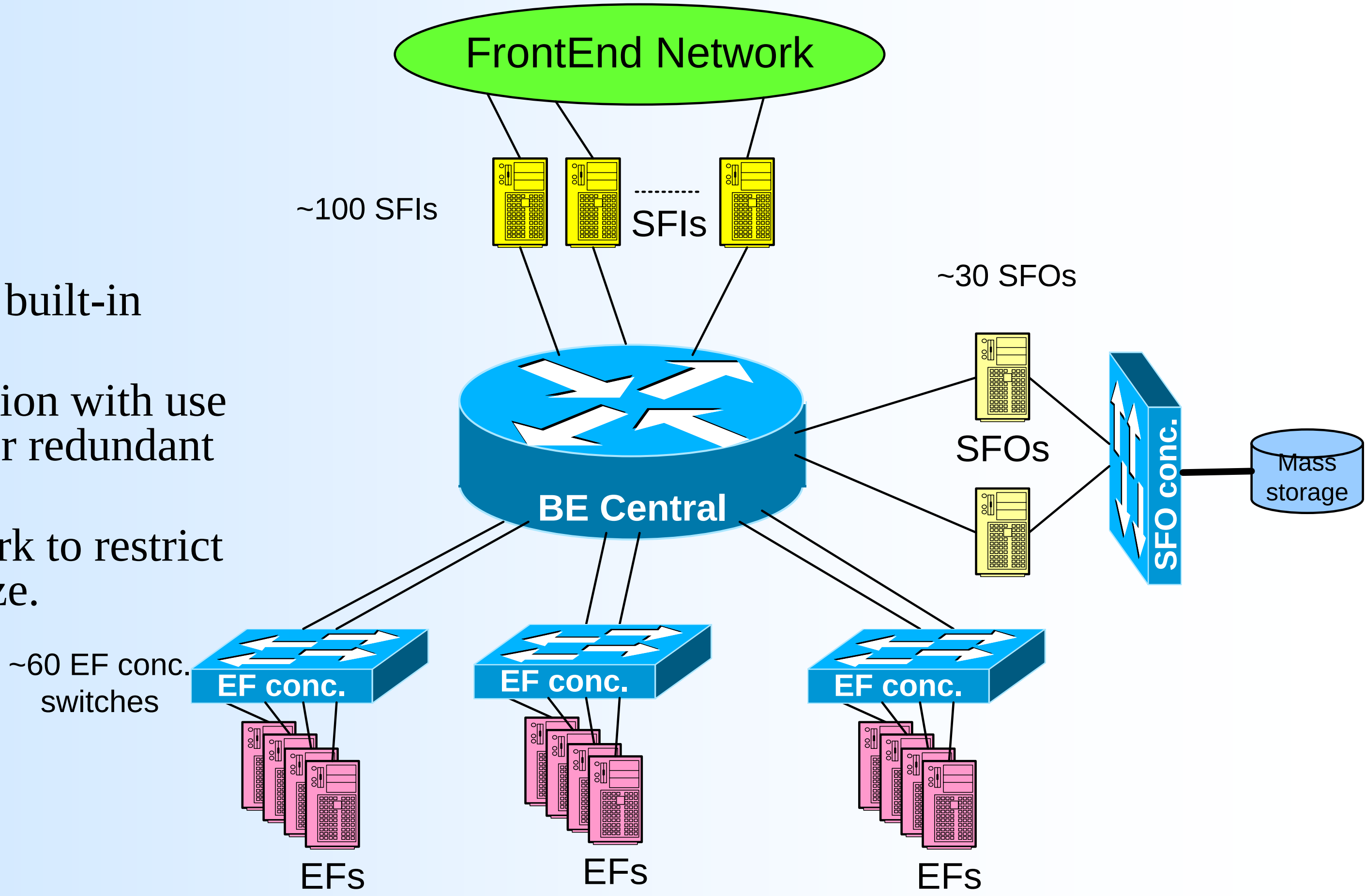
BackEnd network

- ~2000 end-nodes
- One core device with built-in redundancy
- Rack level concentration with use of link aggregation for redundant up-links to the core.
- Layer 3 routed network to restrict broadcast domains size.



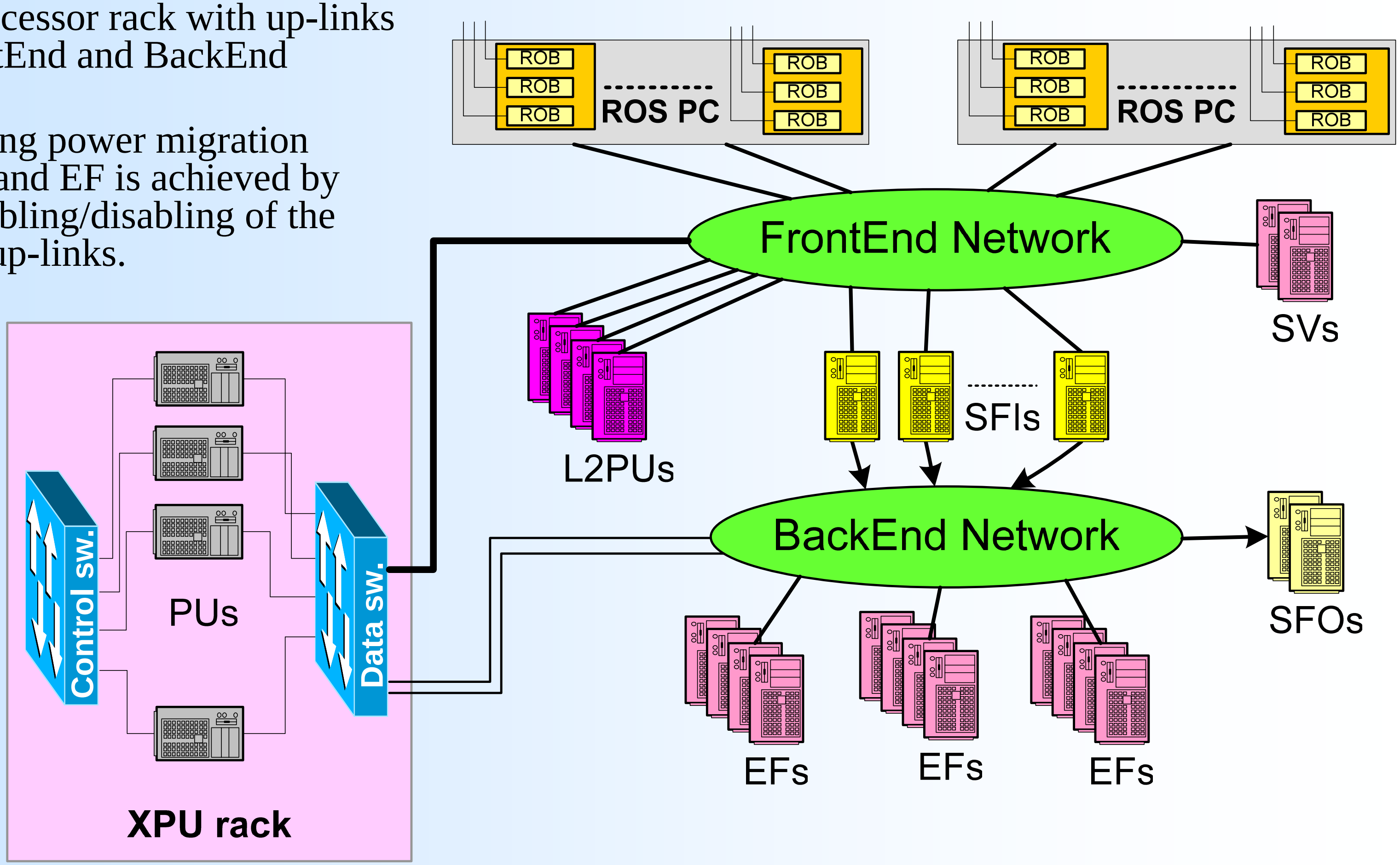
BackEnd network

- ~2000 end-nodes
- One core device with built-in redundancy
- Rack level concentration with use of link aggregation for redundant up-links to the core.
- Layer 3 routed network to restrict broadcast domains size.

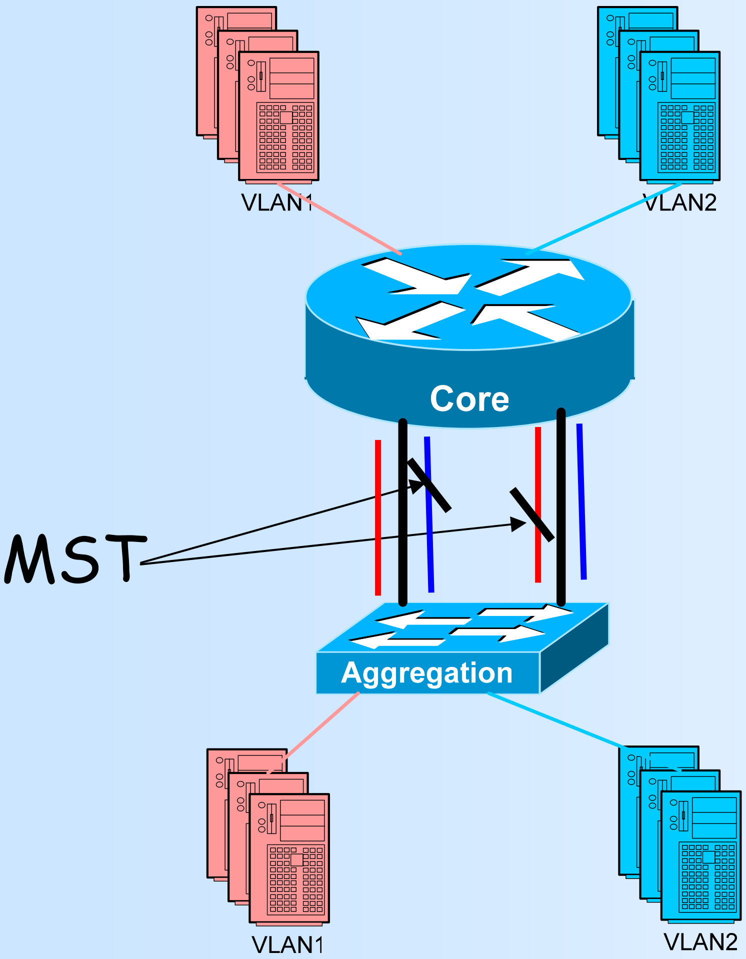


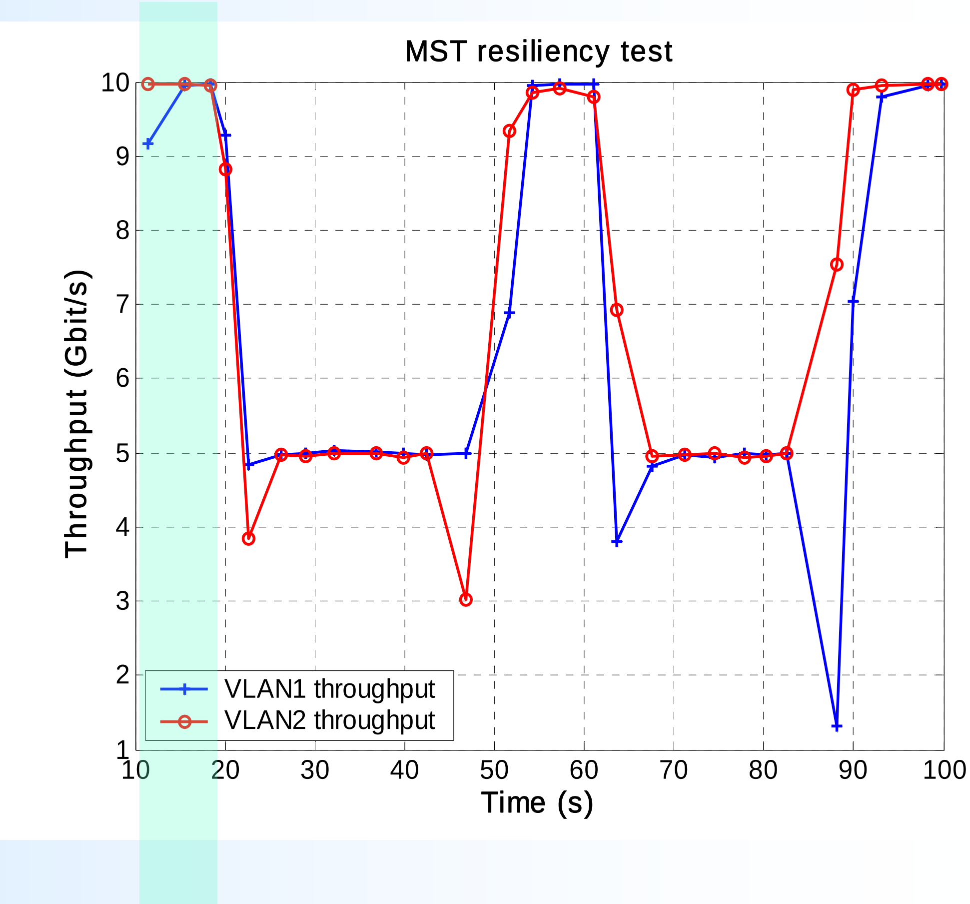
Interchangeable processing power

- Standard processor rack with up-links to both FrontEnd and BackEnd networks.
- The processing power migration between L2 and EF is achieved by software enabling/disabling of the appropriate up-links.

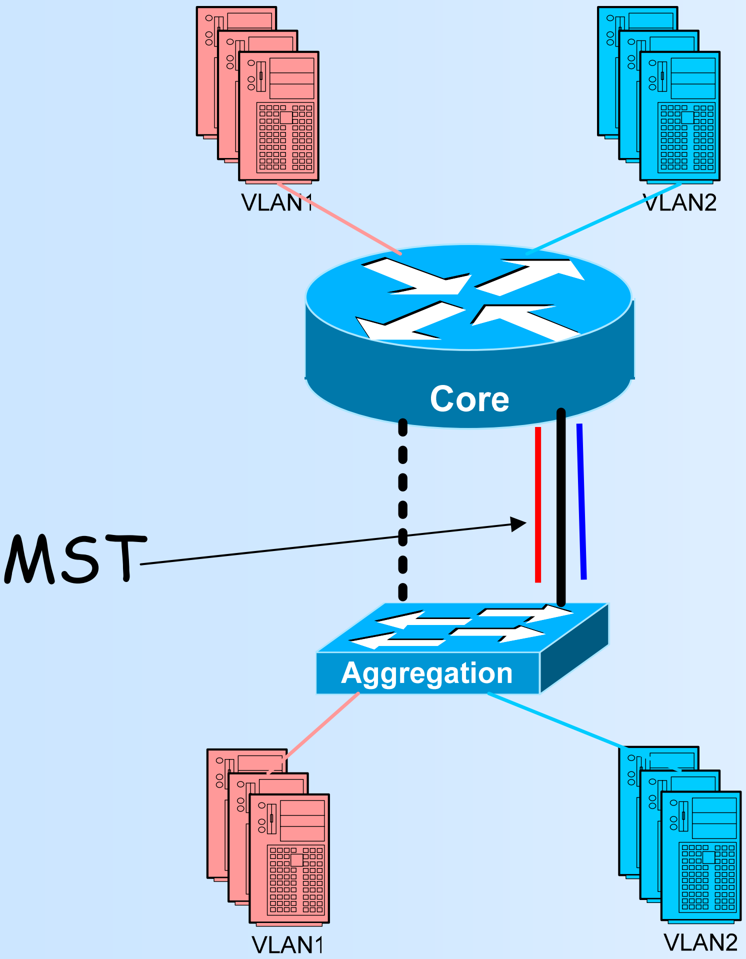


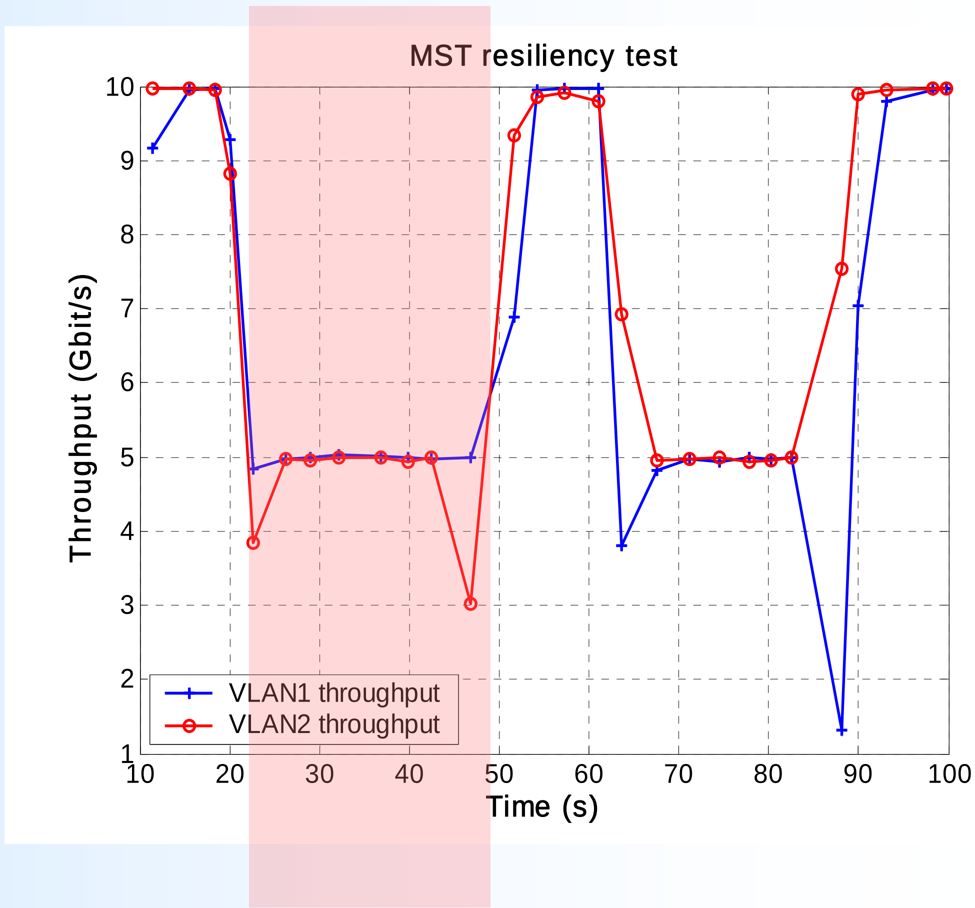
Sample resiliency test (see [4])



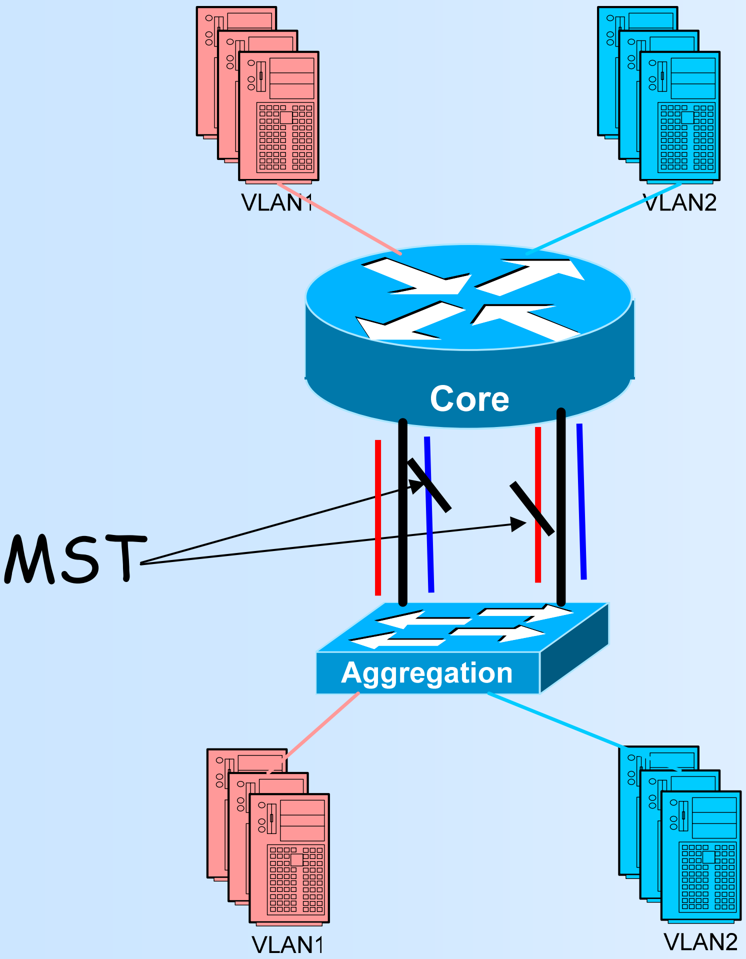


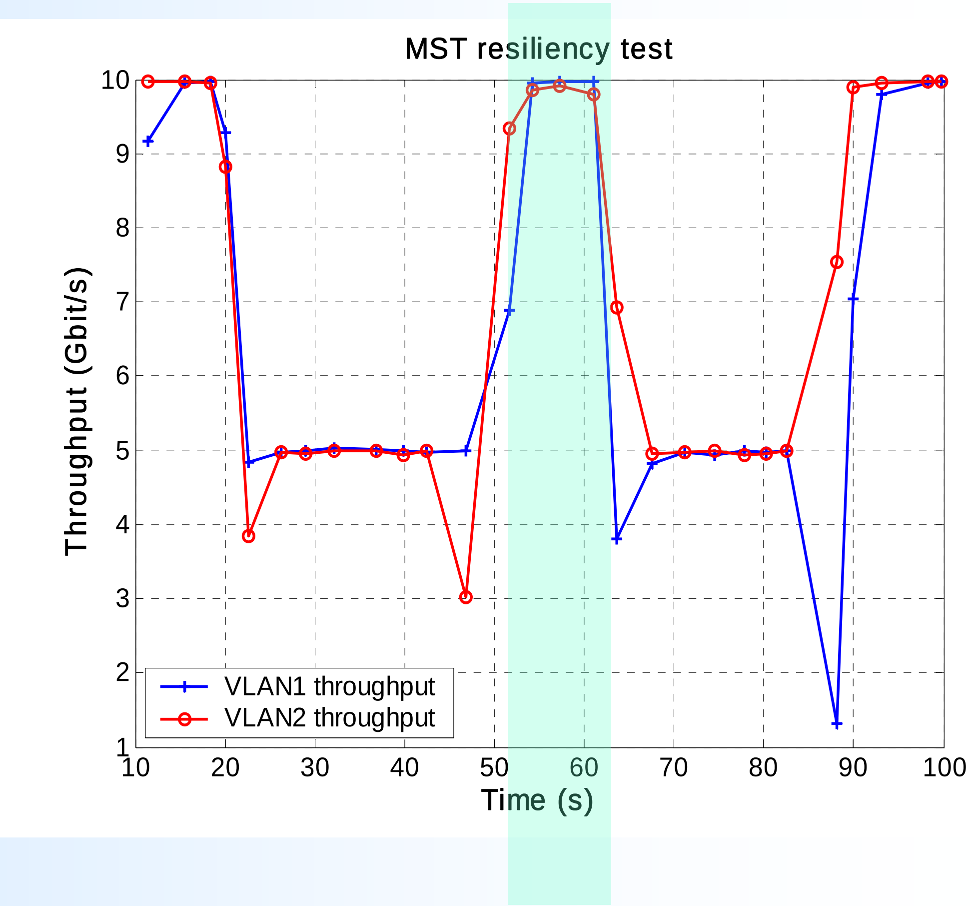
Sample resiliency test



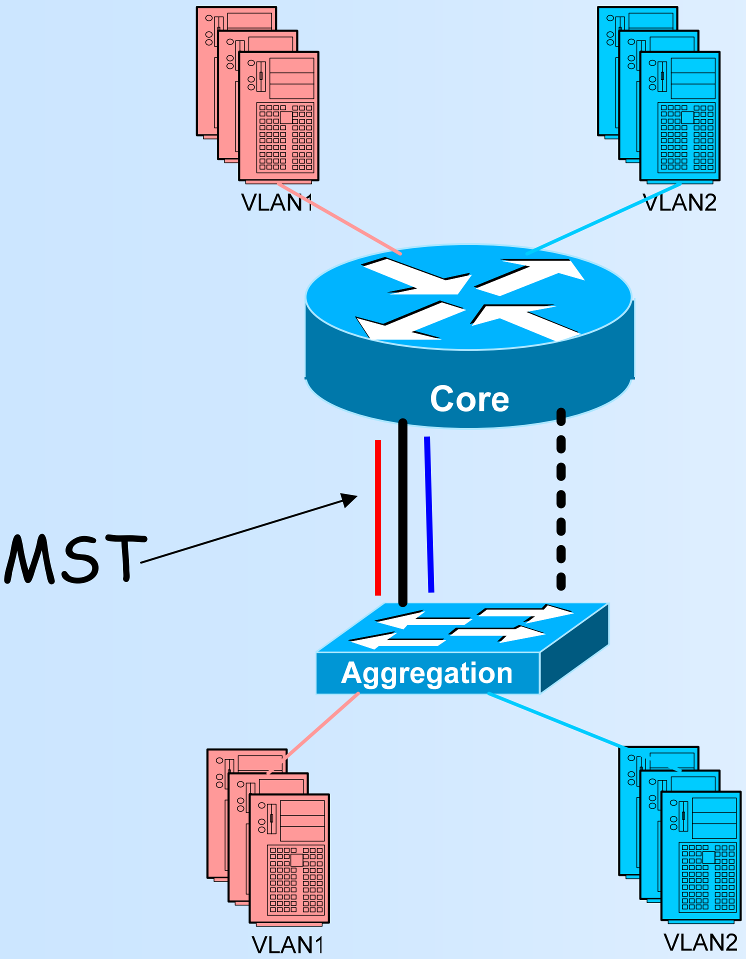


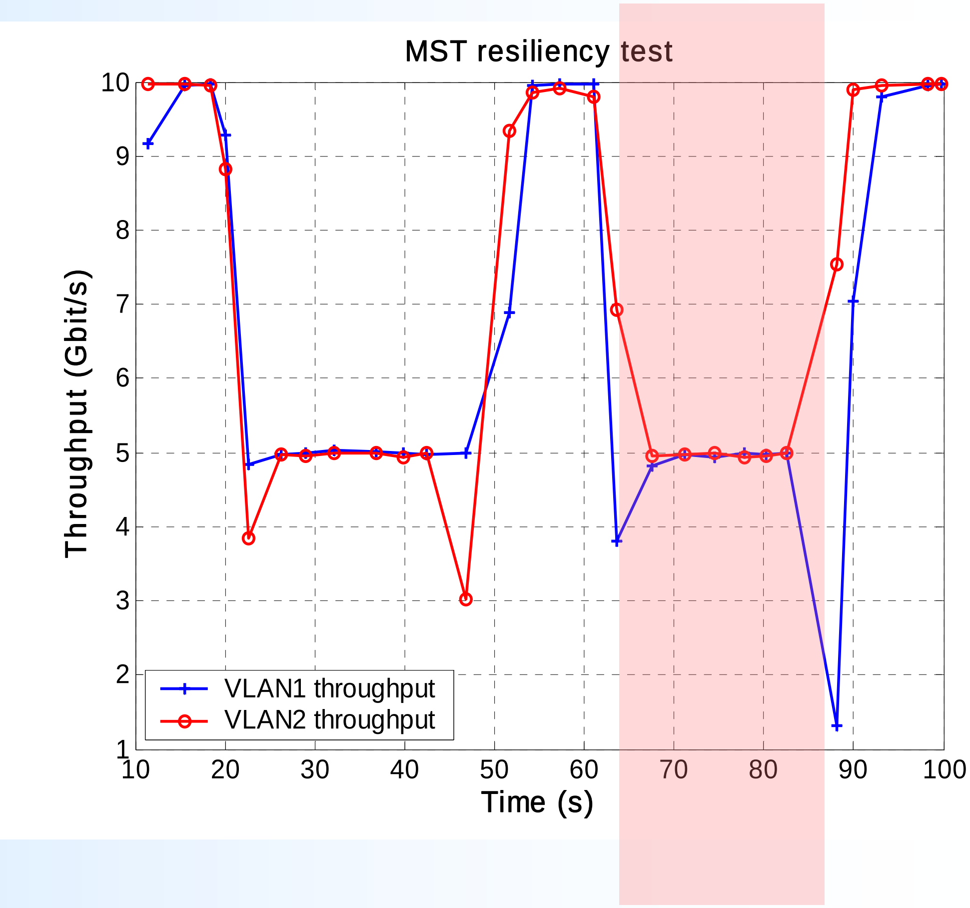
Sample resiliency test



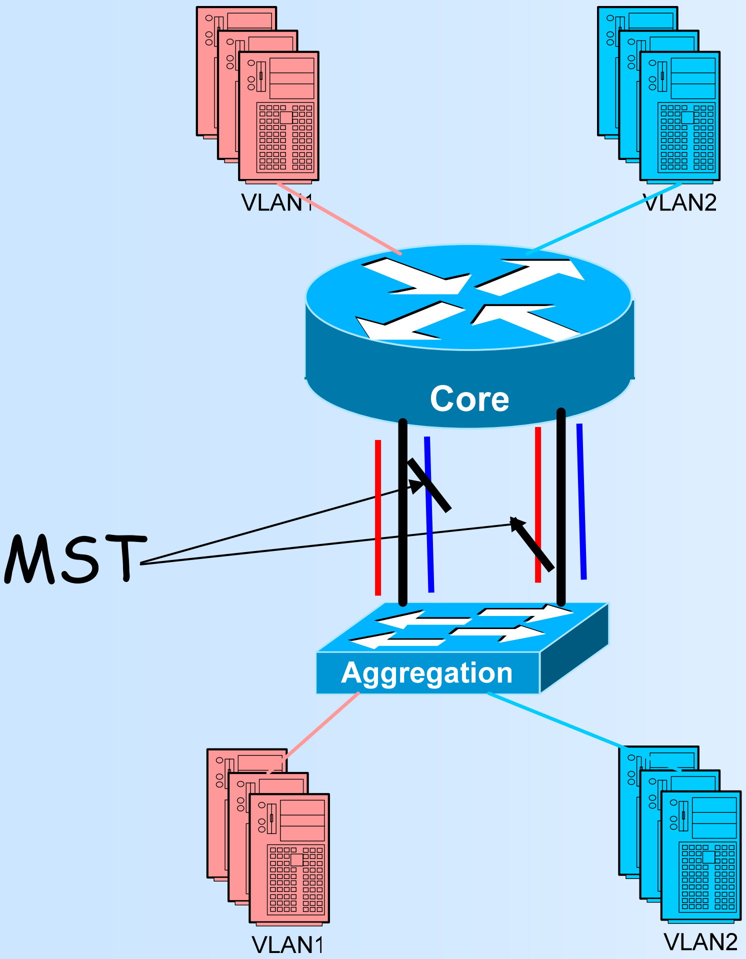


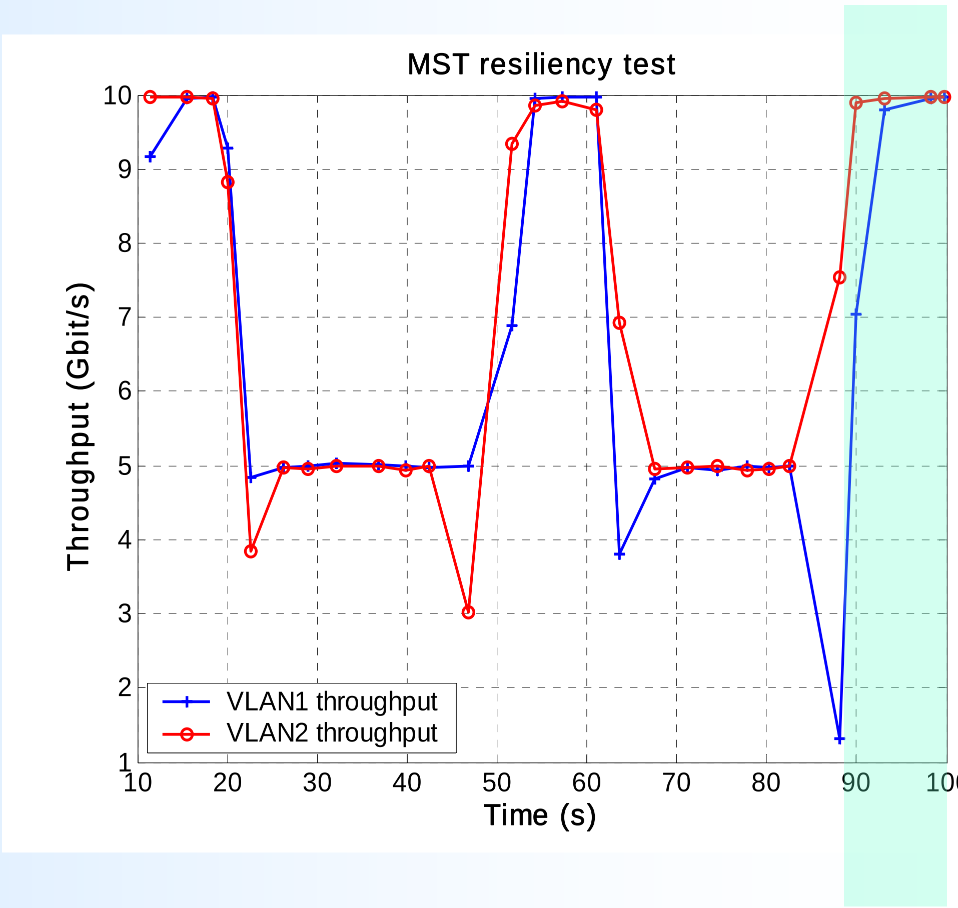
Sample resiliency test





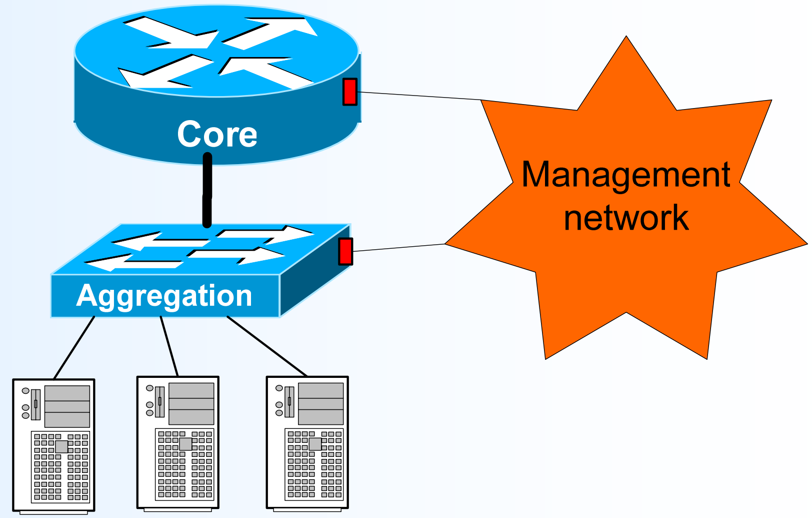
Sample resiliency test



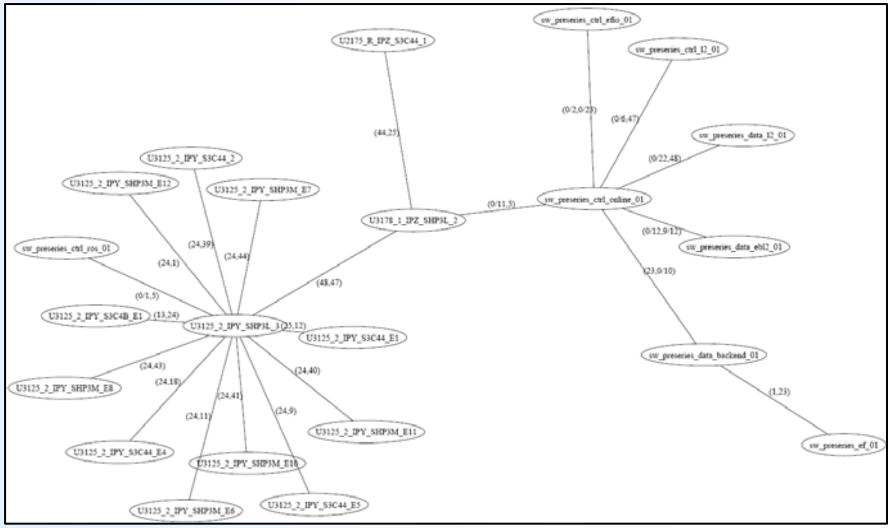


Installation/management issues

- Dedicated path for management
  - ☆ Each device will have an "out of band" interface dedicated for management.
  - ☆ A small layer 2 network will be used to connect to the "out of band" interfaces of devices



- Automatic topology discovery/check
  - ☆ Maintaining accurate active cabling information in the installation database is tedious
  - ☆ Developed a tool which constructs the network topology based on the MAC address table information
  - ☆ To do: interface the tool with the installation database.



Conclusions

- The ATLAS TDAQ system (approx. 3000 end-nodes) relies on networks for both control and data acquisition purposes.
- Ethernet technology (+IP)
- Networks architecture maps on multi-vendor devices
- Modular network design
- Resilient network design (high availability)
- Separate management path
- Developing tools for automatic population/cross-checks of installation data-bases.
- Network operation → see Catalin Meirosu's talk [5].

References

[1] S. Stancu, B. Dobinson, M. Ciobotaru, K. Korcyl, and E. Knezo, "*The use of Ethernet in the Dataflow of the ATLAS Trigger and DAQ*" in Proc. CHEP 06 Conference

[2] T. Sridhar, *"Redundancy: Choosing the Right Option for Net Designs"*

http://www.commsdesign.com/showArticle.jhtml?articleID=25600515

[3] S. Stancu, M. Ciobotaru, and K. Korcyl, "*ATLAS TDAQ DataFlow Network Architecture Analysis and Upgrade Proposal" in* Proc. IEEE Real Time Conference 2005

[4] Cisco whitepaper, "*Understanding Multiple Spanning Tree Protocol (802.1s)"* http://www.cisco.com/warp/public/473/147.pdf

[5] C. Meirosu, A. Topurov, A. Al-Shabibi, B. Martin, "*Planning for predictable network performance in the ATLAS TDAQ"*, CHEP06 Mumbay, February 2006.