Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# Online Network for protoDUNE Single Phase

Geoff Savage
protoDUNE Single Phase

08-Mar-2017

# Functions

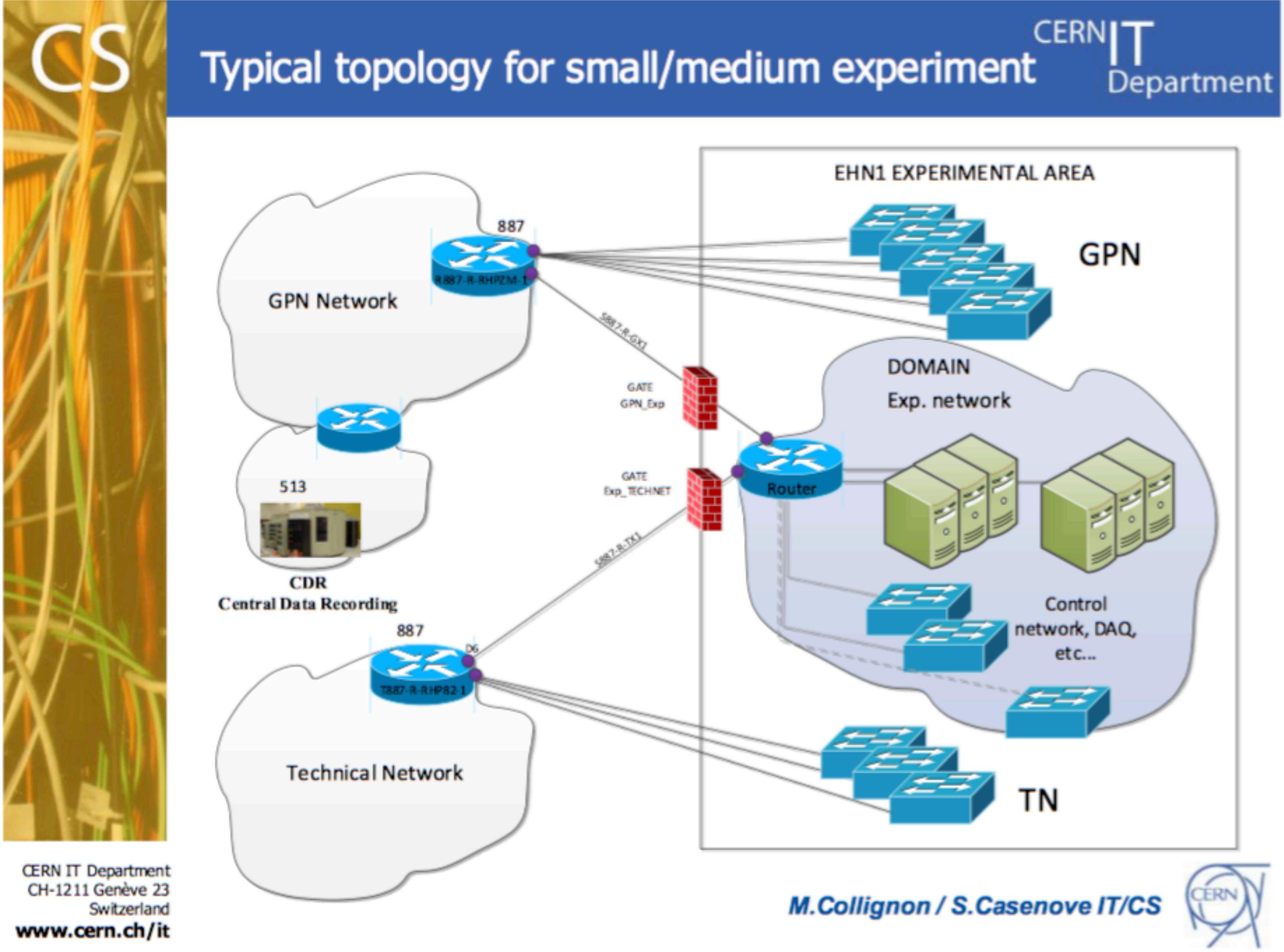
- Network must support these computing functions
- DAQ (RCE, WIB, SSP, timing/trigger, computers, software)
  - Monitoring
  - Data flow
  - Configuration
- Databases
- Slow controls
- Beam instrumentation
- Online monitoring
- Interactive - login, web server, control room, run control
- Data quality monitoring (offline at EHN1)
- Offline data transfer to EOS

# Subnets (a proposal)

- Experiment = DAQ + Slow Controls+ Online Monitoring???
  - Probably want a combined experiment network
  - DAQ - Data flow and configuration
  - Slow controls with database
    - WIB has only one network interface which must support DAQ and slow controls functions
    - Monitor cold electronics via WIB
    - Removes need for a gateway
  - Online monitoring
    - Removes need for a gateway
- Beam instrumentation with database
- Data quality monitoring
- Interactive
  - Login, web server, control room, run control
  - Monitoring
  - Configuration
  - Databases
- EOS

# From CERN Networking

- Gateways and Switches
  - Gateways are part of the network function
    - No computer needed
    - Best provided by HP network switches
    - I will leave gateways off of diagrams for simplification
  - Brocade network switch for performance networking
- WiFi – already installed in EHN1?
- Next slide shows a typical network topology for a small/medium experiment at CERN
  - M. Collignon and S Casenove are from CERN networking
  - Experiment controls access lists in gateways using web interface

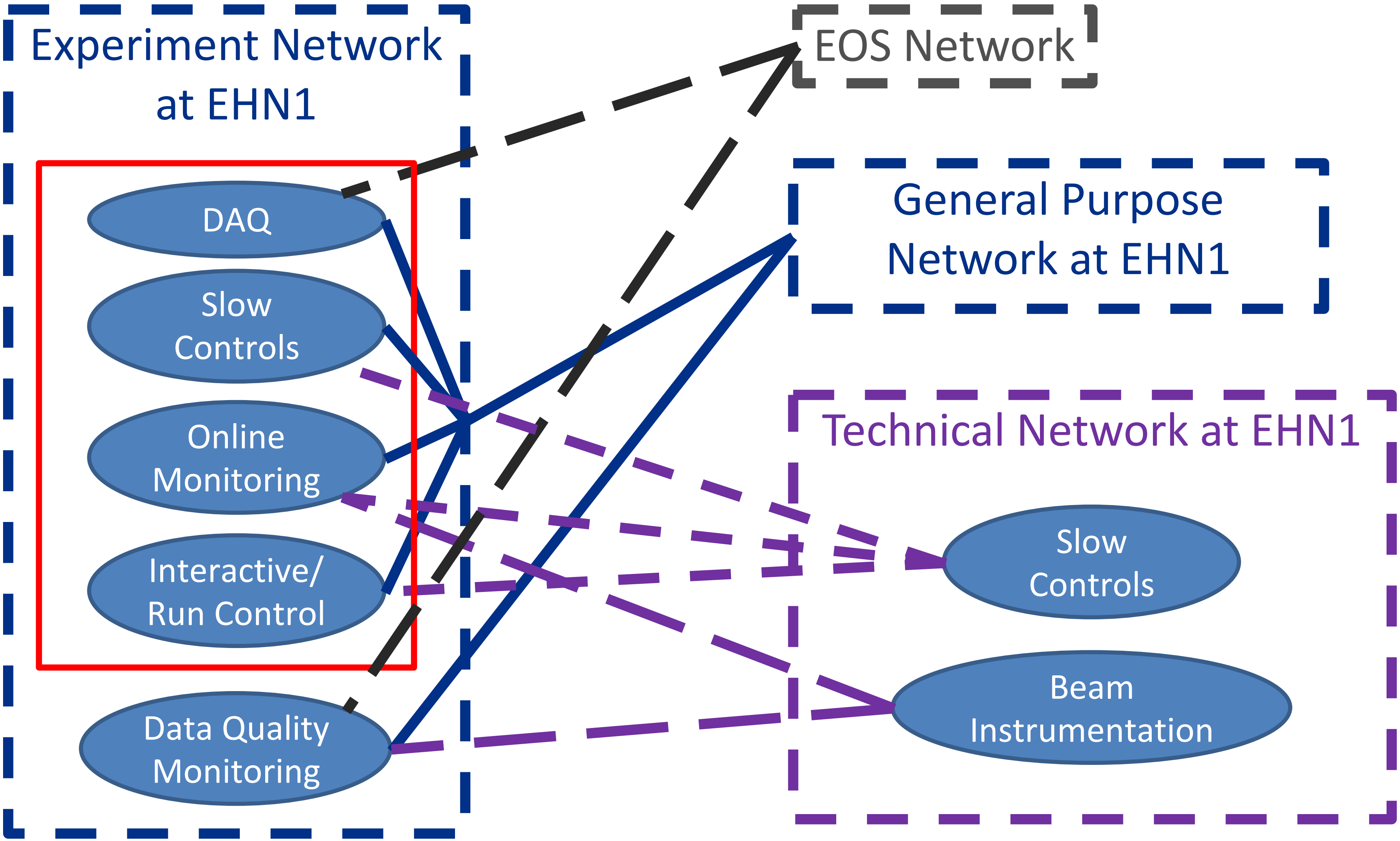


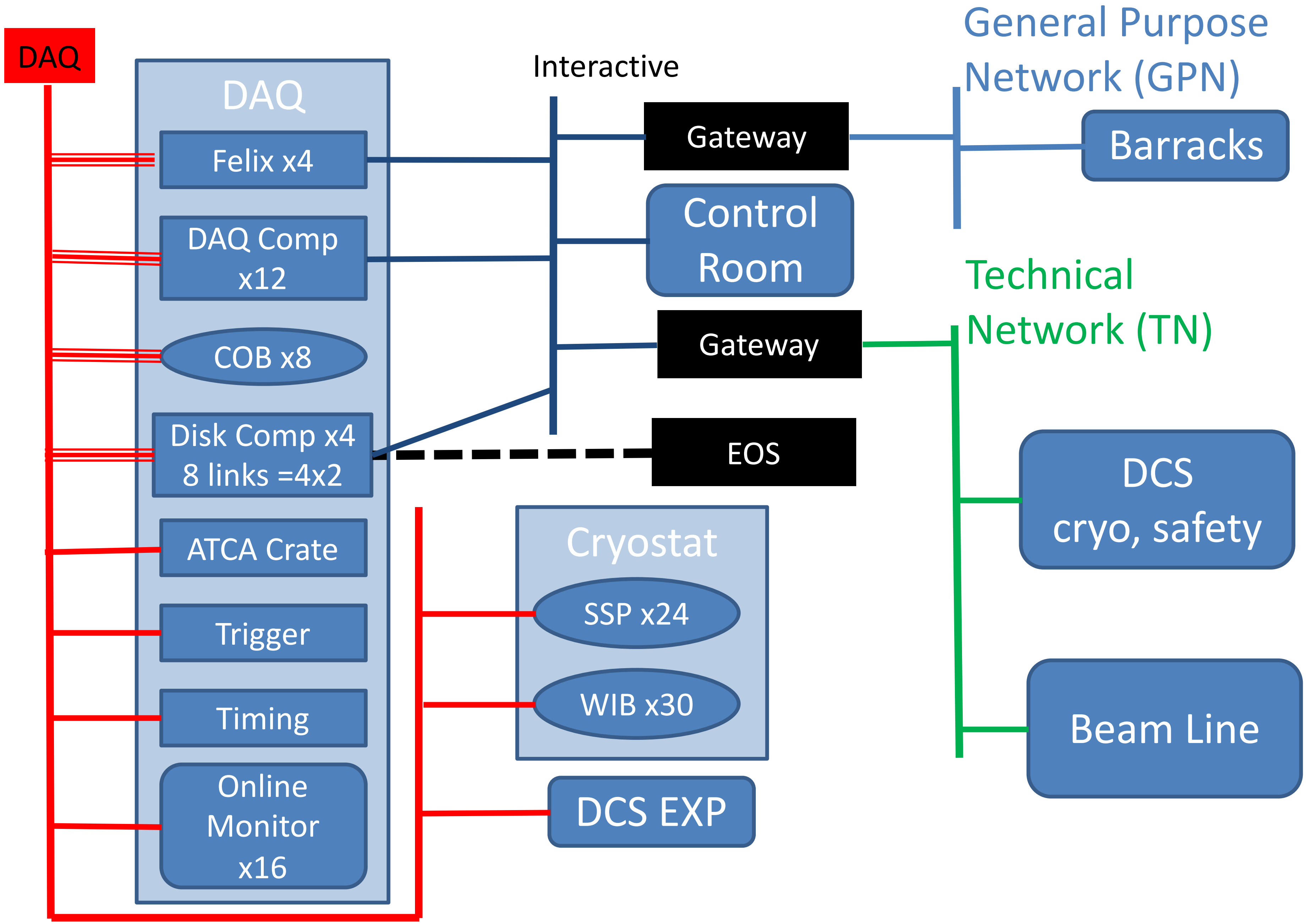
# Typical topology for small/medium experiment

CERN IT Department
CH-1211 Genève 23
Switzerland
**www.cern.ch/it**

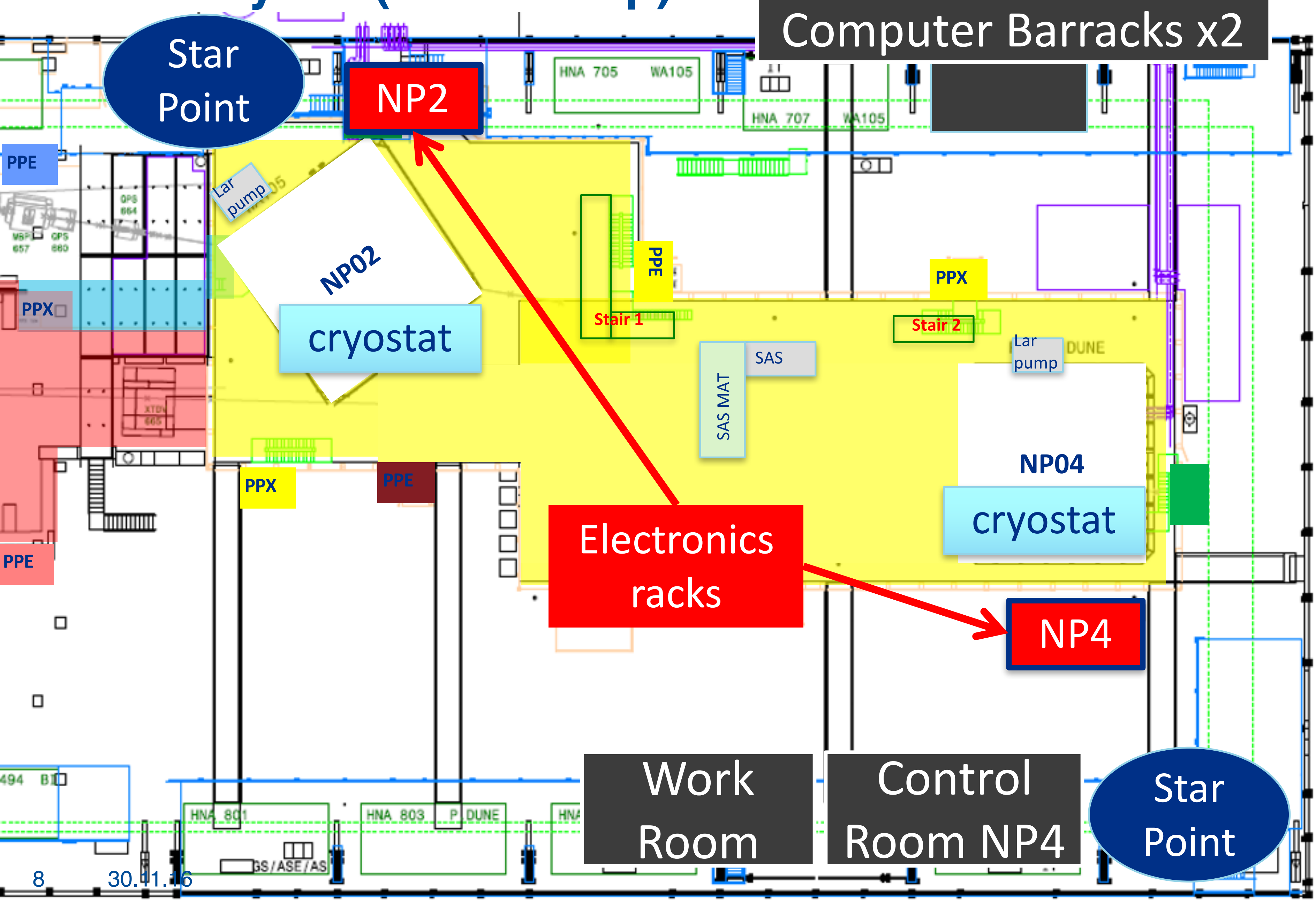
***M.Collignon / S.Casenove IT/CS***

# Network Overview at EHN1




DAQ
DAQ
Felix x4
DAQ Comp x12
COB x8
Disk Comp x4 8 links =4x2
ATCA Crate
Trigger
Timing
Online Monitor x16
Interactive
Gateway
Control Room
Gateway
EOS
Cryostat
SSP x24
WIB x30
DCS EXP
General Purpose Network (GPN)
Barracks
Technical Network (TN)
DCS cryo, safety
Beam Line

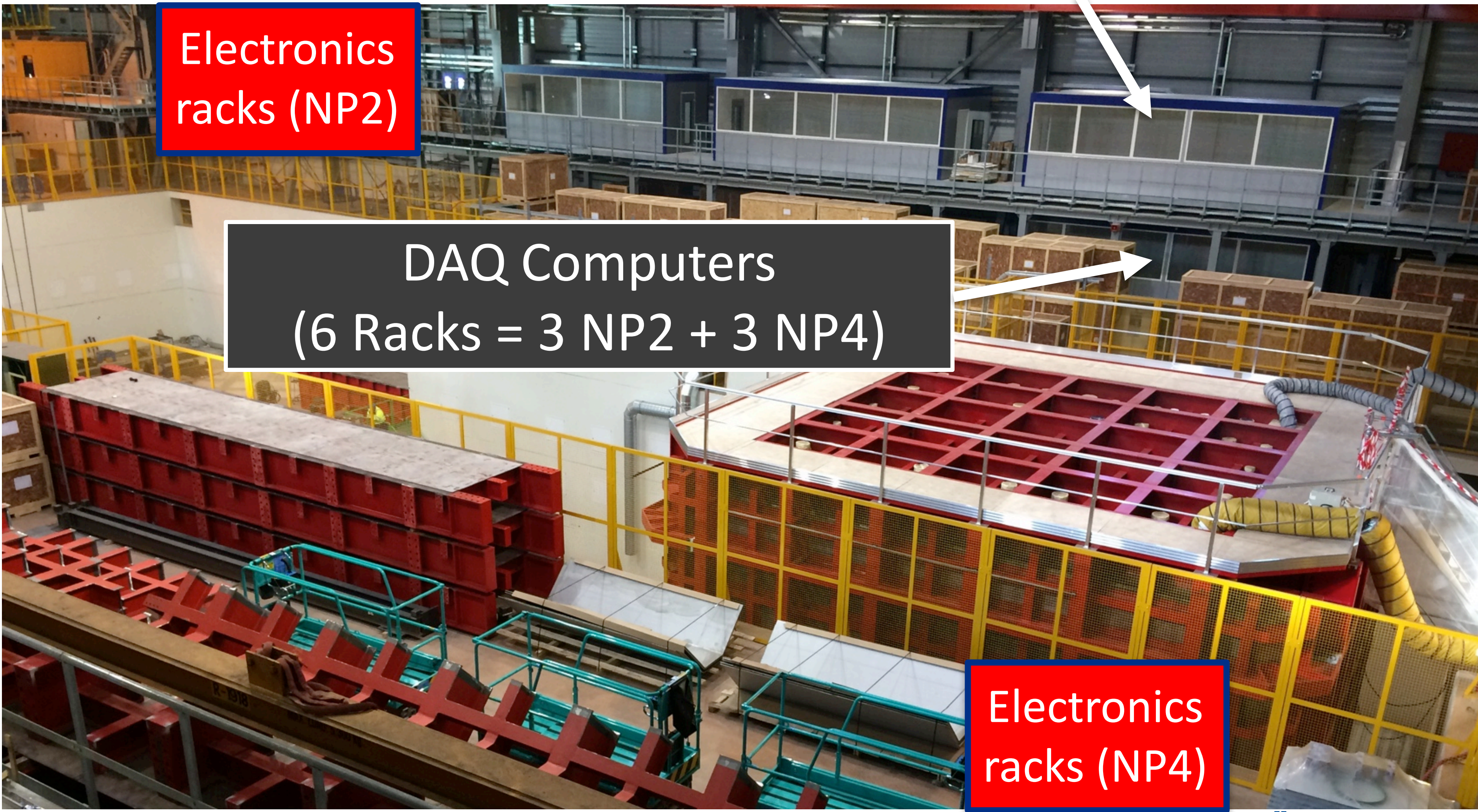
# EHN1 Layout (Jura at top)



# Cryostat Connections

- Fibers from barracks to top of cryostat
  - No AC power for network switches on top of cryostat
  - Fibers needed for electrical isolation
- Six feed throughs
  - One fiber bundle per feed through (144 fibers in each bundle)
  - Nine 1Gb ethernet connections (= 4 SSP + 5 WIB)
    - WIB = Warm Interface Board
    - SSP = SiPM Signal Processor
  - Other fibers are for data coming from the cold electronics and timing/trigger
- 54 (=9x6) fibers from cryostat
  - Feed two network switches
  - Each switch has four 10 Gb uplinks
  - Need eight 10 Gb uplinks

# Locations (looking toward Jura)



Offline Computers (12 Racks)

Electronics racks (NP2)

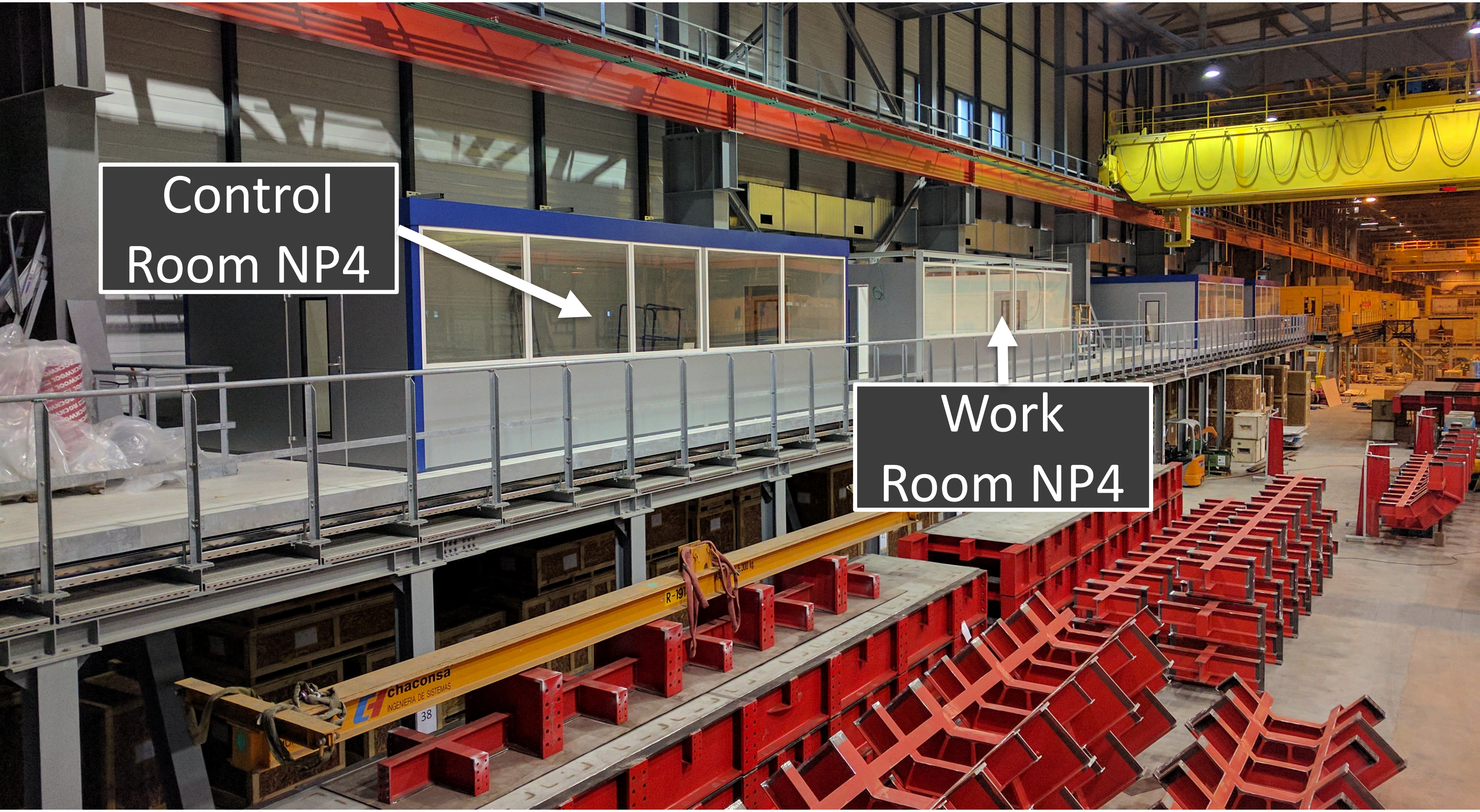
DAQ Computers (6 Racks = 3 NP2 + 3 NP4)

Electronics racks (NP4)

# Electronics Racks

- Fiber to racks – 20 racks for NP4
- Configurations
  - Each rack has a small switch with fiber uplink and copper within rack
    - 20 fiber uplinks and 20 small switches
    - Preferred by noise prevention experts
  - One switch with fiber uplink in a rack then distribute to small switches in each rack via fiber
    - One fiber uplink and 20 small switches
  - One switch with fiber uplink in a rack then distribute to all racks with copper
    - One fiber uplink and one switch
- Connect to nearest star point?

# Locations (looking away from Jura)



Control Room NP4

Work Room NP4

# Barracks

- Standard layout
  - Layout calls for 10 network connections
    - 5 on front wall and 5 on back wall
    - Distributed evenly along with power plugs
  - Is there a standard layout for barracks?
  - Giovanna suggested reviewing NA62 networking.
- Offline computers (12 racks)
  - I'm guessing here
  - 30 1U computers per rack
  - 1 Gb copper
  - 360 = 30 * 12

# DAQ Components

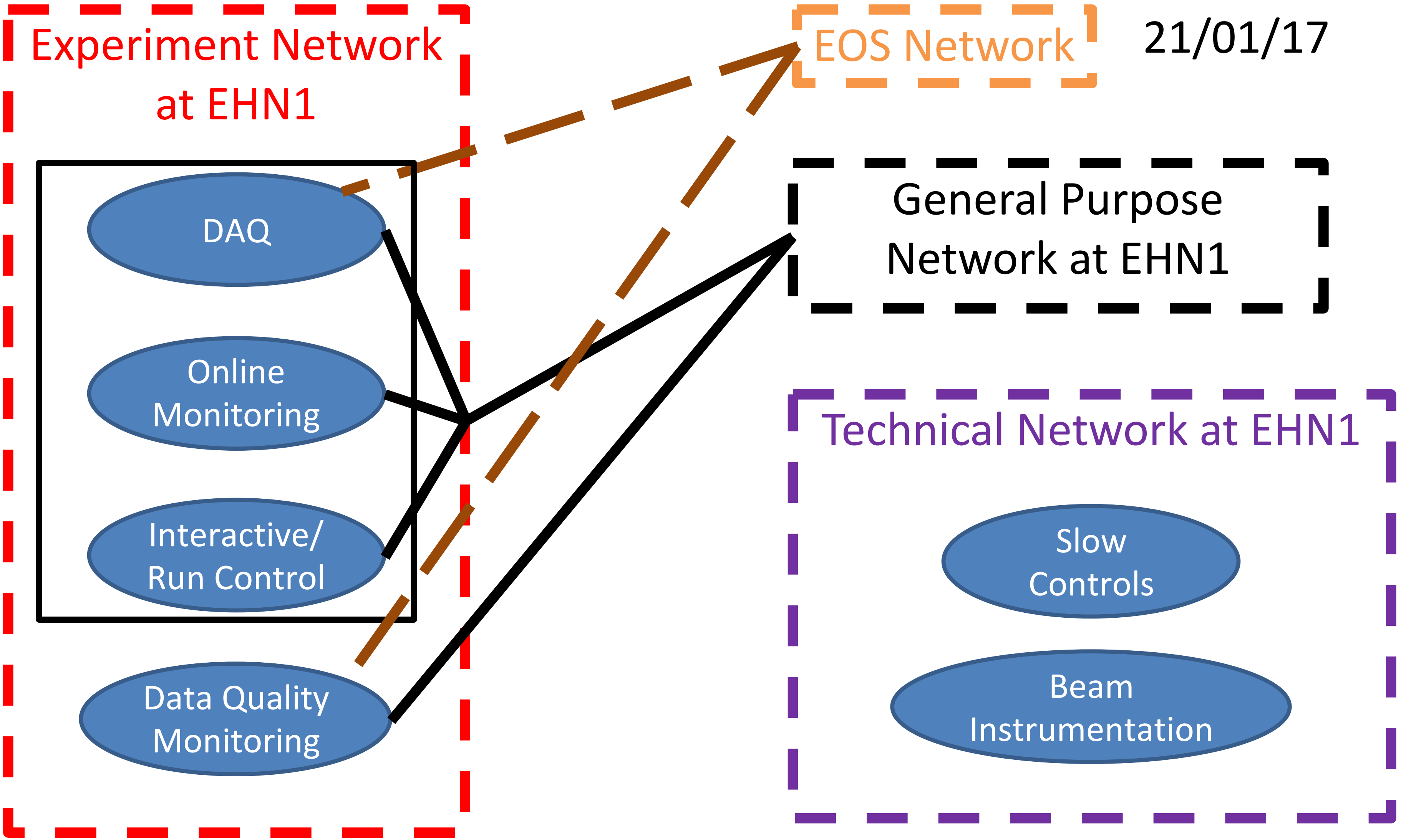
- 10 Gb connections (40 =4+12+8+8+8)
  - 4 Felix
  - 12 DAQ computers
  - 8 COBs
  - 8 Disk computers
  - 8 uplinks to 1Gb connections
- 1 Gb connections (70 =24+30+16)
  - 24 SSP
  - 30 WIB
  - 16 Online monitor
- Summary?
  - One 48 port 10Gb Brocade switch
  - Two 48 port 1 Gb switches

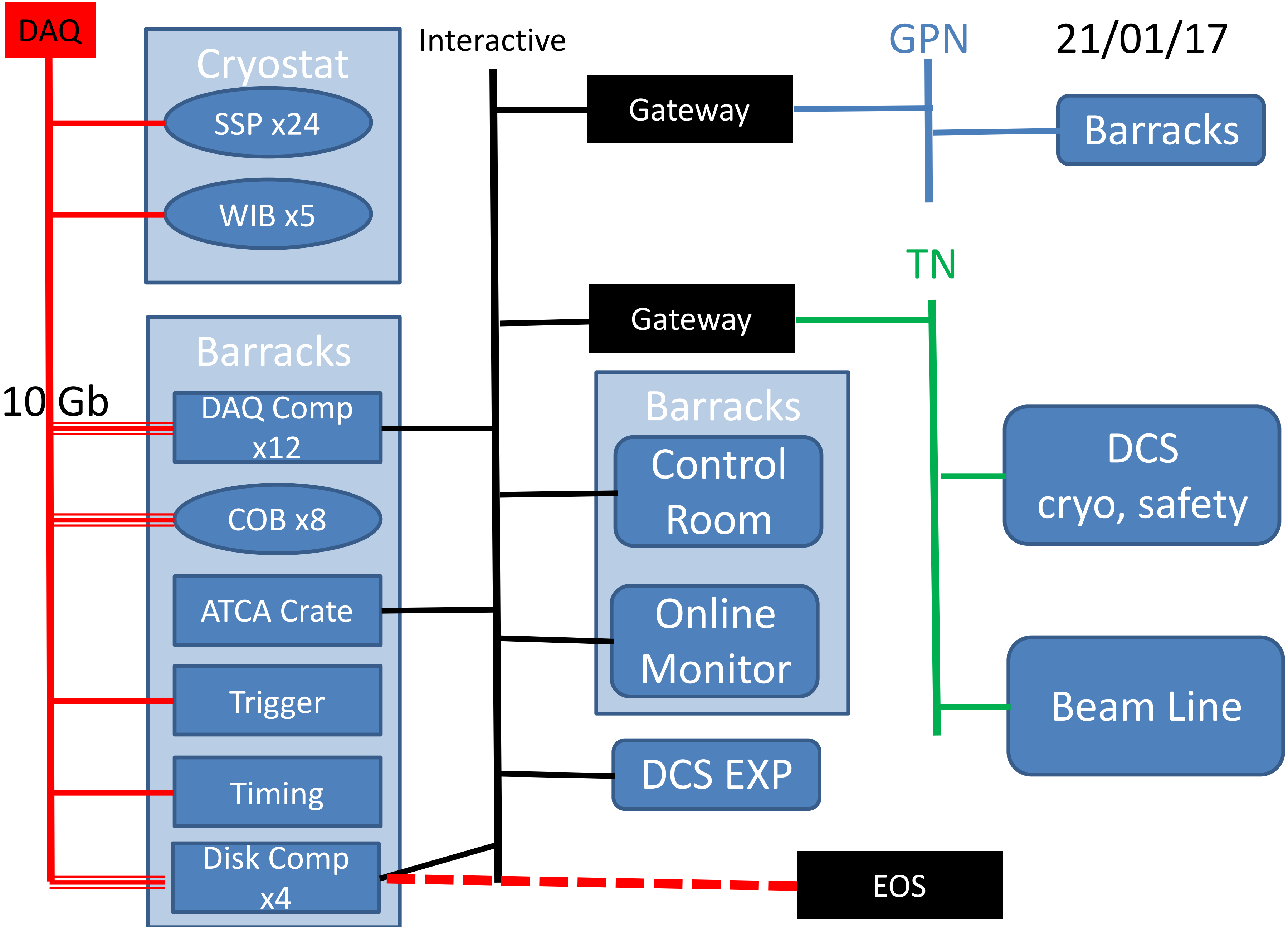
# EOS

- Four disk writer computers
  - 10 Gb link for each computer from EOS network
  - Network switch?
- Plan is to bring data back from EOS to the offline network at EHN1 on the same link
- Combination with NP2 and NP4?

# BACKGROUND

# Networks at EHN1




DAQ
Cryostat
SSP x24
WIB x5
Interactive
Gateway
GPN
21/01/17
Barracks
TN
Gateway
Barracks
10 Gb
DAQ Comp x12
COB x8
ATCA Crate
Trigger
Timing
Disk Comp x4
Barracks
Control Room
Online Monitor
DCS cryo, safety
Beam Line
DCS EXP
EOS

# Data Rates

- From Giovanna,
- Thanks for the summary. I wanted to comment on a point in the twiki, but apparently have no edit access.
- The event size is 250 - 300 MB without compression, but we aim at achieving a compression level of ~4. In addition, the 25 Hz rate applies only to the in-spill time, thus even with 100% efficiency the number should be divided by 4. Thus, normally we will be writing in the order of 40 TB/day.

# Network Questions from Giovanna (Fall 2016)

- Physically separate networks for data flow and control traffic. Is this correct?
- For servers this just implies having 2 network interfaces (1Gbps for control, 10 Gbps for data).
- How about custom devices?
  - Do they have dual interfaces or a single one to connect to the board readers?
  - If they have a single one, then we are sort of obliged to put them on the control network.
  - Is 1 Gbps enough ?
- How about the monitoring nodes?
  - Do we want them on both networks or only on the control network?
- Is it OK to share what we call control network with the slow control?

# RCE and Timing System Answers

- From Matt Graham
  - For the RCEs, there are 10GbpsE links out the front of each board and then a single 1 Gbps link to the ATCA shelf manager and these two should be on separate networks so you don’t get weird looping.
- From Dave Newbold
  - For the timing system, there is no distinction between control, slow control, and DAQ traffic, as they are all dealt with by the same UDP block in the firmware anyway - so there's not much point having two network interfaces. The DAQ traffic should be very low anyway.
  - If, for other reasons, we want two interfaces this will need some work. The FPGA board we are planning to use only has one physical interface (at 1Gbps), so it would require a hardware change. Not a insurmountable problem, but it would cost some time.

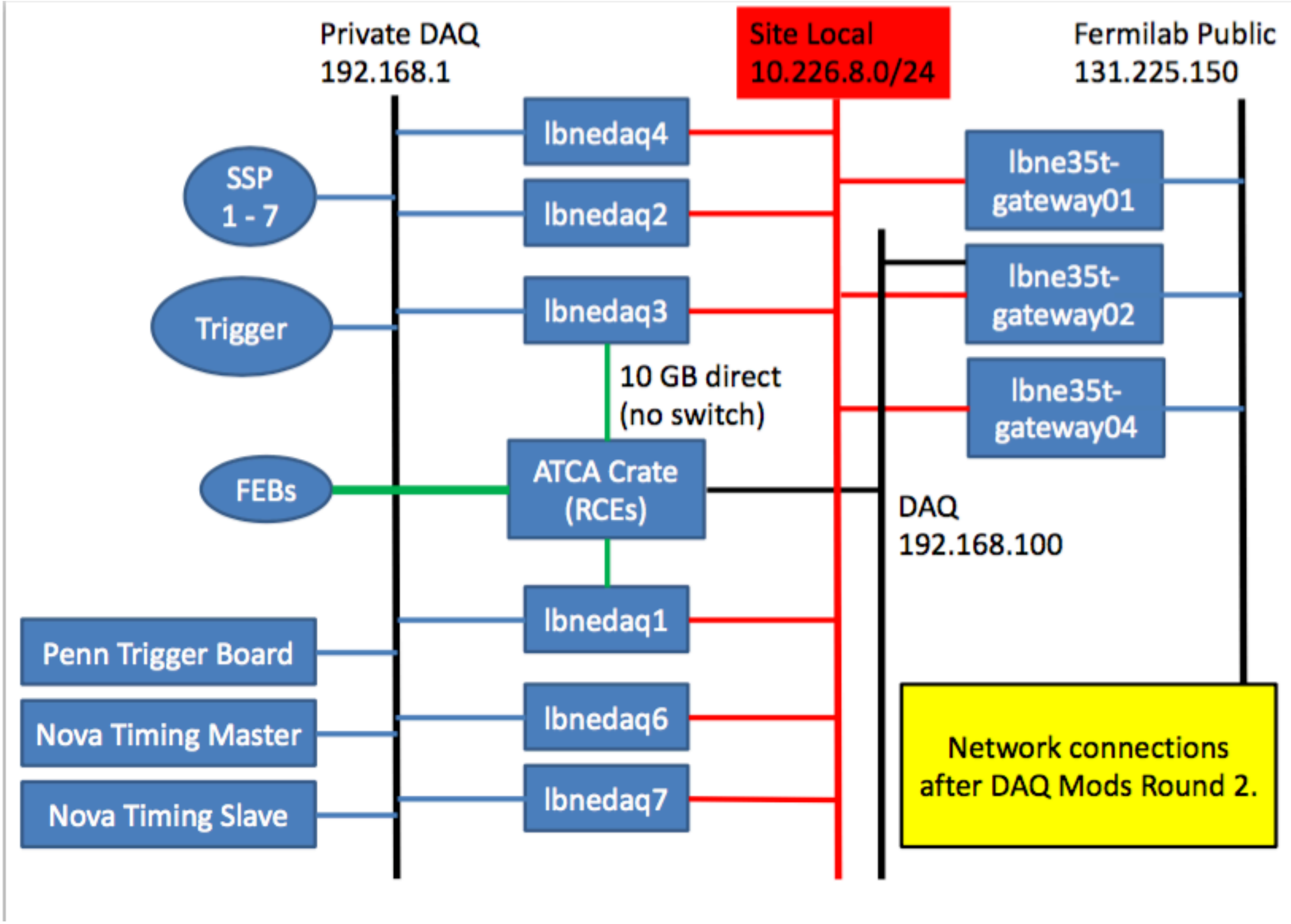
# Trigger Board Answers

- From Nuno Barros
- Concerning the CTB, having two separate physical ethernet connections requires either significant work or some "creative adjustments".
- The core of the board is a MicroZed SoM which has a single Gigabit ethernet socket. To add a second socket we would have to do it by implementing the second ethernet connection almost from scratch (hardware and firmware), which puts even more stress on an aggressive schedule. We could potentially use a USB-to-Ethernet adapter to add a second, independent, ethernet connection, but I haven't yet tried this (although through a quick search found reports of successful attempts doing this).  The board itself runs linux, which allows to have multiple ethernet sockets.
- On 35t we used this to keep 2 separate connections (one for data, one for control). As the data volume produced by the board was very modest, this posed no constraints, except that we had to be in a single network. In any case, considering the expected data volume,  1Gbps should be well more than enough for the CTB.

# 35 Ton Description

- From Geoff Savage
- The network architecture was developed by Tim Nichols in consultation with Fermilab networking.
- At 35Ton there were three subnets.  See attachment.
  - Public - access from offsite via kerberos
  - Site Local - access to resources at Fermilab
  - Private - no access
- There was one glitch which occurred that did not impact data flow.
  - I did not get the site local network configured when the computers were moved from the test stand at DAB to PC4 where the 35 Ton prototype detector was located.
  - We were unable to modify the IP address of the shelf manager to a site local address so we left it on the private network instead of site local.
- Fermilab performs security scans on the public and site local network so the private network was needed for the systems that can't be kerberized, Penn trigger board, Nova timing boards, SSPs, ... .
- Network access to the systems was via gateway computers that also served other roles, data transfer to offline storage and run control for example.
  - For protoDUNE I recommend using the gateway computers only as gateways.

# 35 Ton Network - Actual



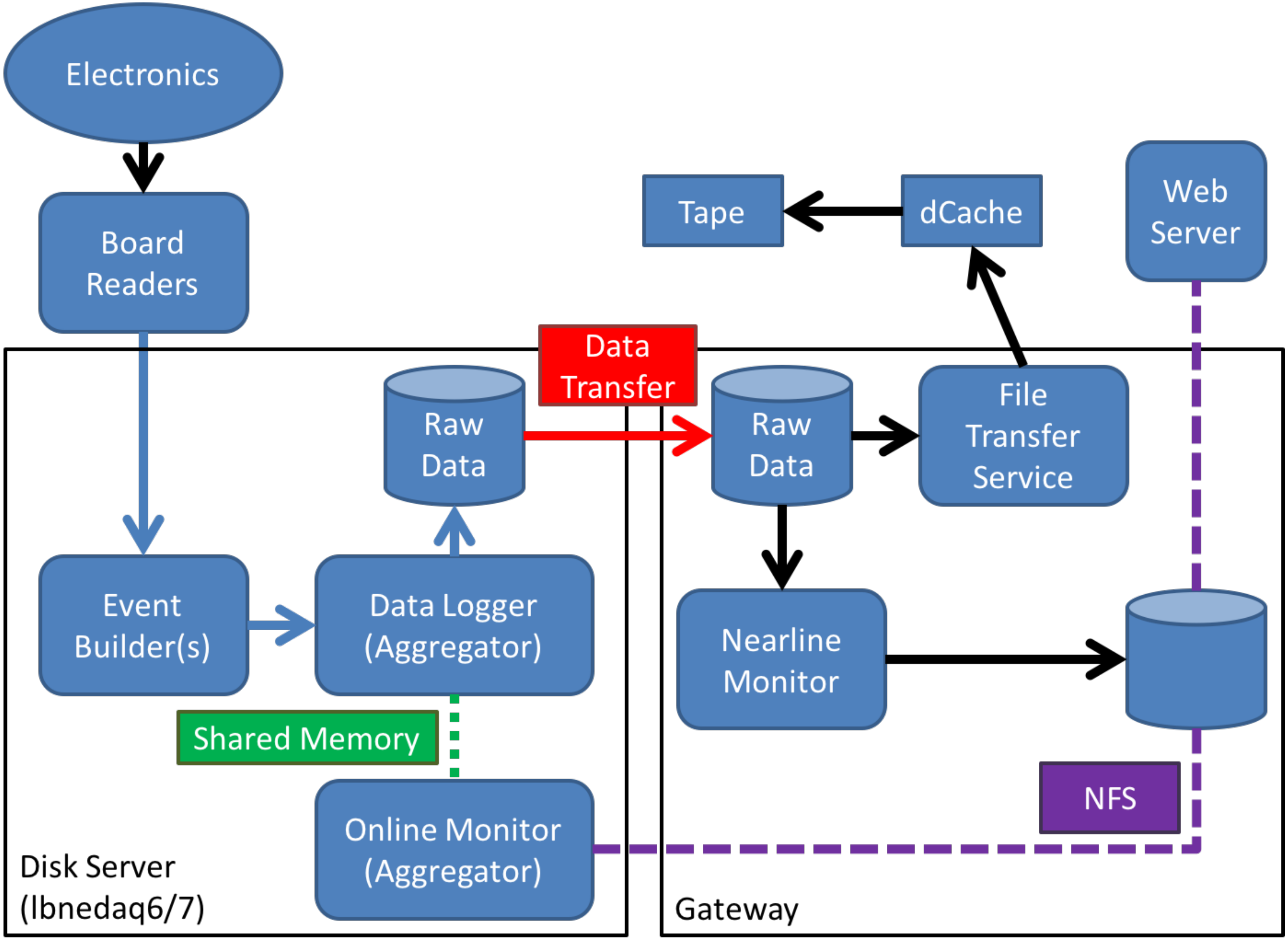
# Background from Geoff Savage

- Definitions:
  - DAQ network - 10 Gb move data at expected rate of 3 GByte/s (24 Gbit/s) to disk servers
  - Control network - 1 Gb for logins, computer monitoring (ganglia), artDAQ (error message, write logfiles)
  - Private network - 1 Gb to move data from SSP, trigger, timing to DAQ computer (where board readers run).
  - DAQ computer - where board readers execute.
  - Disk server - where event builder run and write data to disk
- Restrict access to DAQ computers and units by requiring authentication through gateways.
- Trigger, timing, ssp units can't be exposed to security scans. So a private network is needed to move data between these units and the DAQ computers.
- The DAQ computers need at least three interfaces - control, DAQ, private (to receive data from ssp, trigger, timing).
- The Disk servers need at least two interfaces - control and DAQ.

# Some Considerations (from Geoff Savage)

- RCE
  - At 35 Ton there was a direct 10Gb connection to the DAQ computers.
  - Do we want to make this connection via a switch at protoDUNE?
  - If we use a direct connection to the DAQ computer then the DAQ computer needs two 10 Gb interfaces.
  - One to get the TPC data from the RCE and one to move the data to the disk server.
- How do we connect to the monitoring and offline file transfers?
  - Events are built and stored on the disk servers.
  - The offline file transfers need to be fast to utilize the full 20 Gb/s to the outside.
  - We can send data to monitoring without writing the events to disk.
  - Is 1 Gb fast enough?
  - This is probably not the control network so we need an additional monitoring network.
- Where do we perform data sanity checks with Beam Instrumentation data?
  - To do these you would need all the data and could not drop events as has been discussed.
- Do we put some computers in the protoDUNE networks and the Beam Instrumentation network?
  - We did this at DZero with the Accelerator Division network.

# 35ton Data Flow



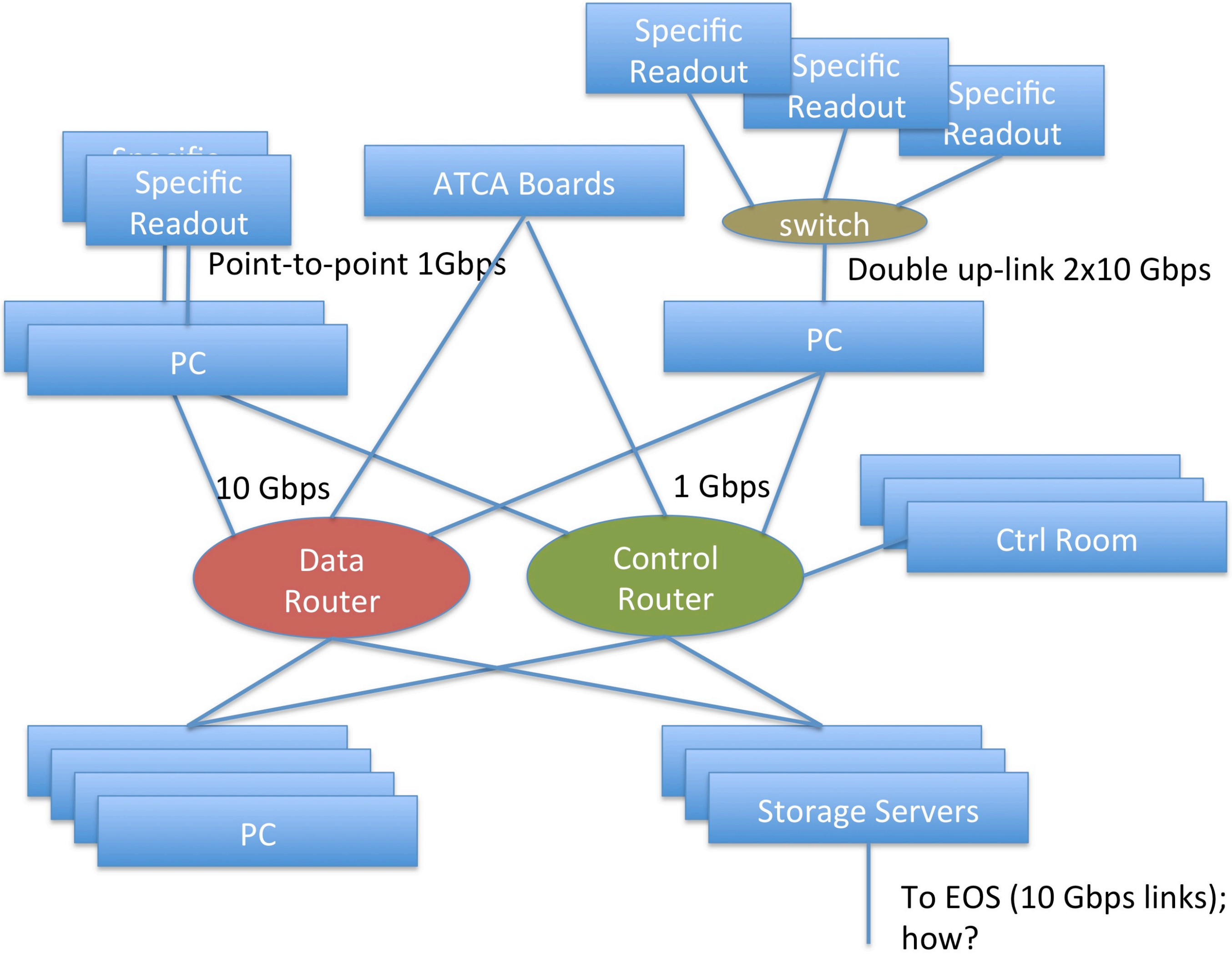
# Data Transfer (from Tom Junk)

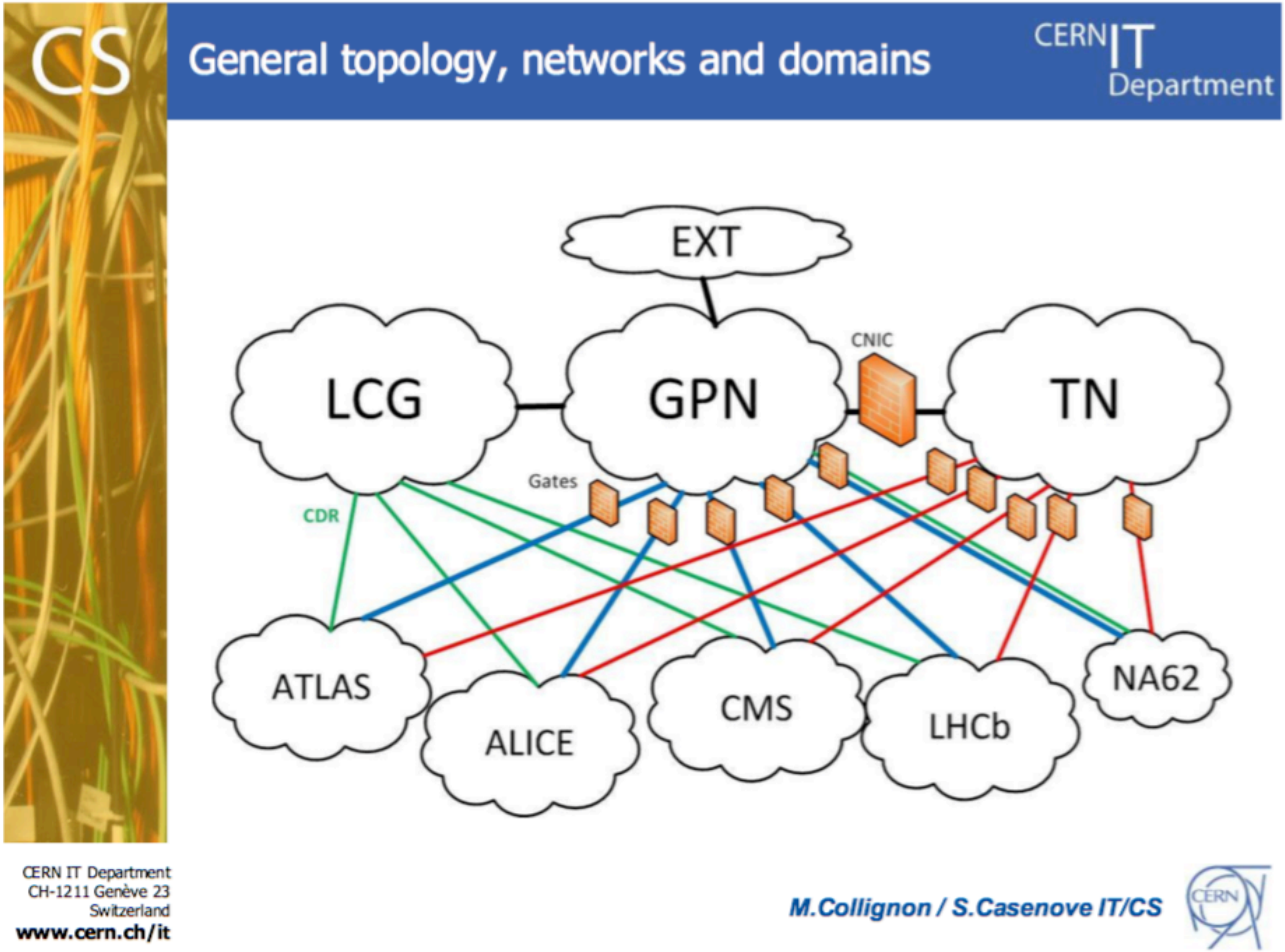
- lbne35t-gateway02:/home/lbnedaq/trj/dtr35t1.sh
- It is run once/hour by a cron job that starts dtranslog35t.sh which just directs the output to a logfile.
- It checks to see if there's already one running, and if not, ls's the files on lbnedaq6, skipping the files with no run and subrun number. It runs a checksum on lbnedaq6 (just cksum; I've benchmarked others but they're all cpu bound and adler32 isn't installed on lbnedaq6). We could think of just skipping the checksum step as per our discussion, though sometimes we may want to make sure a file gets across not because it's cosmics but because it was an intentional test of something so we should try not to drop data.
- It doesn't delete files — just mv's them to a directory of finished files. A separate script in that directory gwtdeleter.sh loops over the transferred files and deletes the ones it finds in enstore. I misspoke at the meeting in that it looks for enstore in the file location but doesn't (yet) check for a tape label, which should be of the form (*@*) and should be easy to check for.
- The copy step from daq6 to gateway02 is certainly a step that can be optimized.
- I think we are currently limited by gateway02's disk write speed but with better hardware we may move that bottleneck elsewhere.

# CERN Networks (Karol Hennessey)

- Briefly about CERN networks:
  - CERN general network : includes your office computer but also some batch systems, and general services
  - Only part of it is public - Cern terminal services, lxplus (which is a batch system but also doubles as an ssh gateway - kerberos is used here), and web servers.
  - All the public stuff is under control of CERN IT.
  - Your office computer is not publicly accessible from outside.
- Experiment specific domain : behind a gateway usually.
  - Can have many subnets inside the expt domain.
  - Inside the domain I think we still have to follow CERN IT rules, but we don't have to ask permission for setting up our own servers etc.
  - So I don't think it changes too much from what you had in mind from Fermilab.
  - The main difference as a user would be that I have to go through two gateways to get inside the protodune domain.ssh lxplus.cern.ch, and then from there ssh pdgw.cern.ch.

# DAQ Network Proposal (from Giovanna)





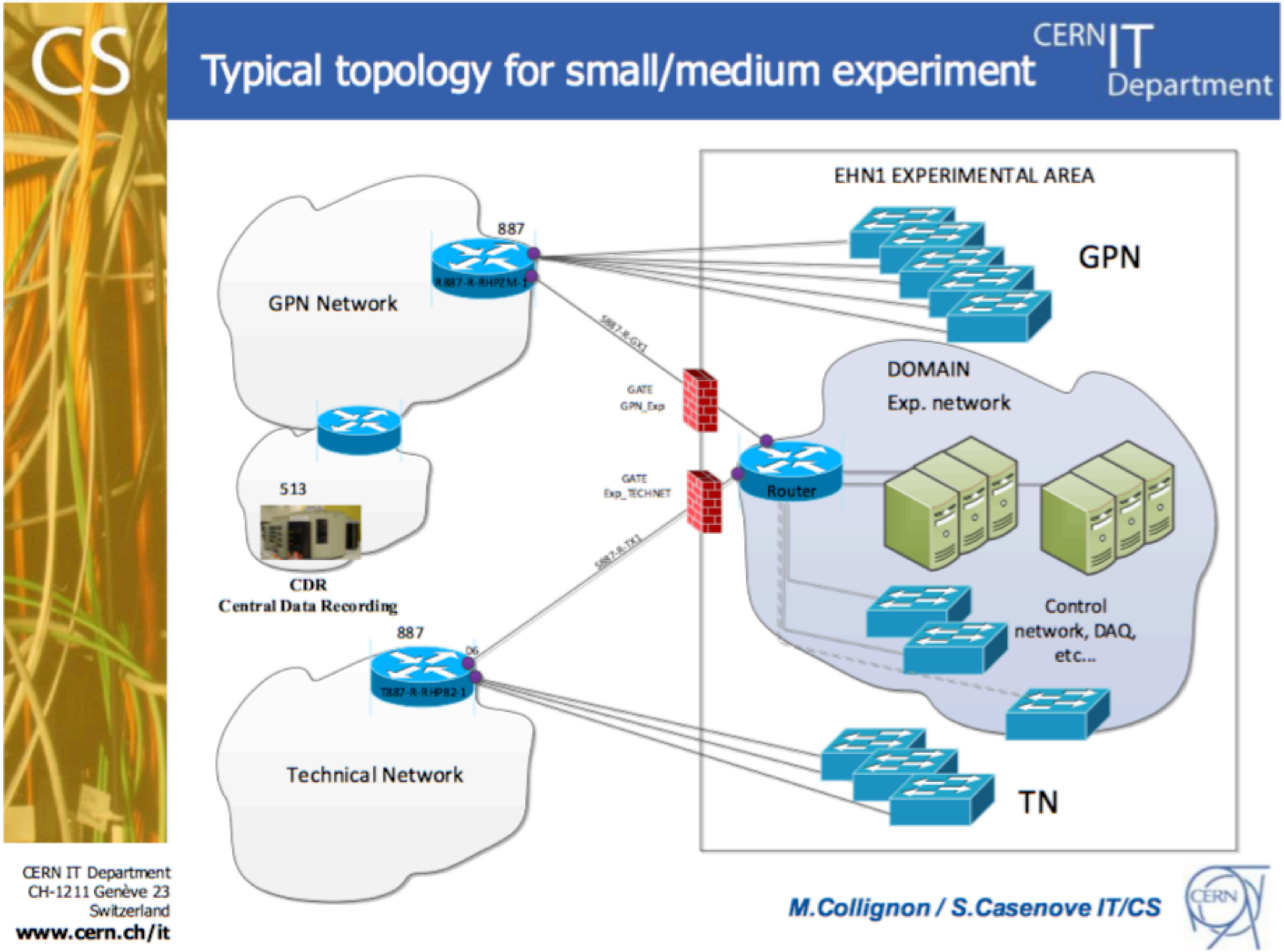
CS

## General topology, networks and domains

CERN IT Department

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

*M.Collignon / S.Casenove IT/CS*



# Typical topology for small/medium experiment

CERN IT Department
CH-1211 Genève 23
Switzerland
**www.cern.ch/it**

*M.Collignon / S.Casenove IT/CS*