INFN Tier-1 status

Report on the flooding event
Network evolution

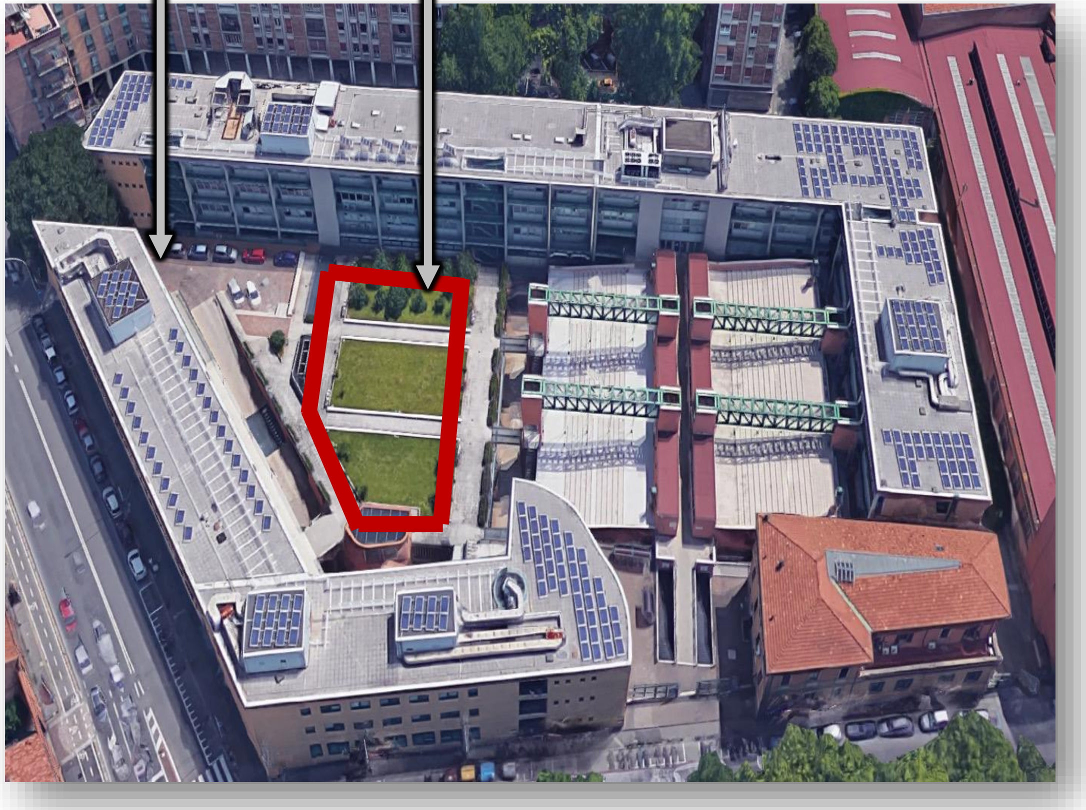
Stefano Zani (stefano.zani@cnaf.infn.it)

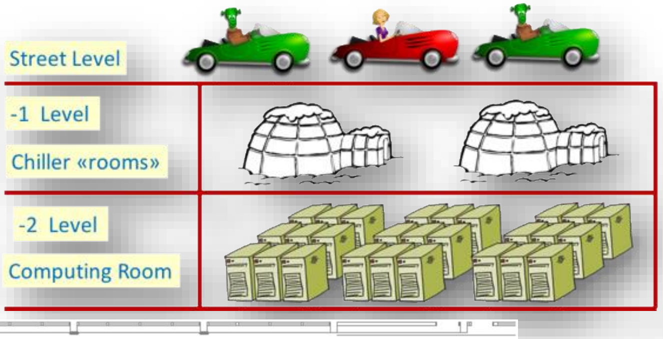
LHC OPN Meeting - RAL, Mar 6 2018

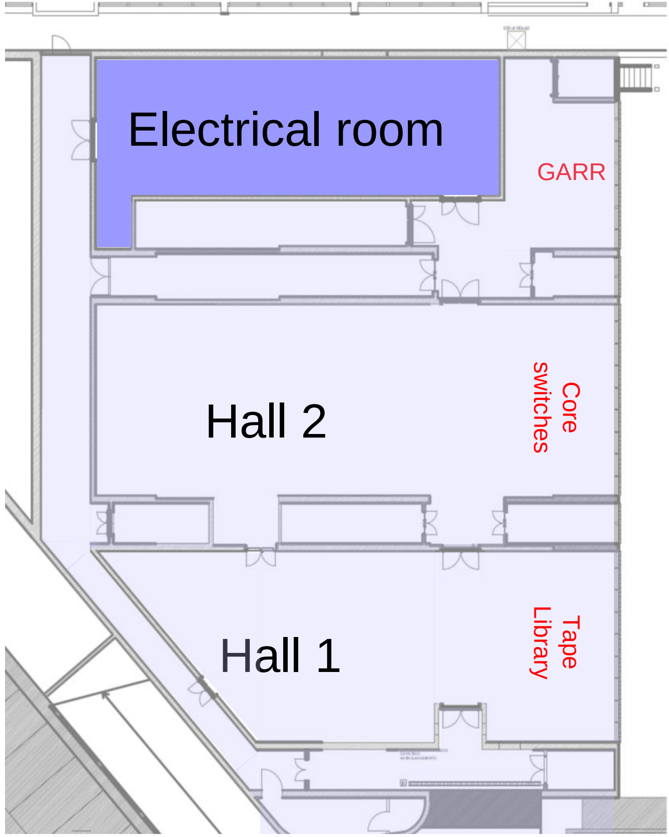
The Tier-1 location

Transformers

Electrical room

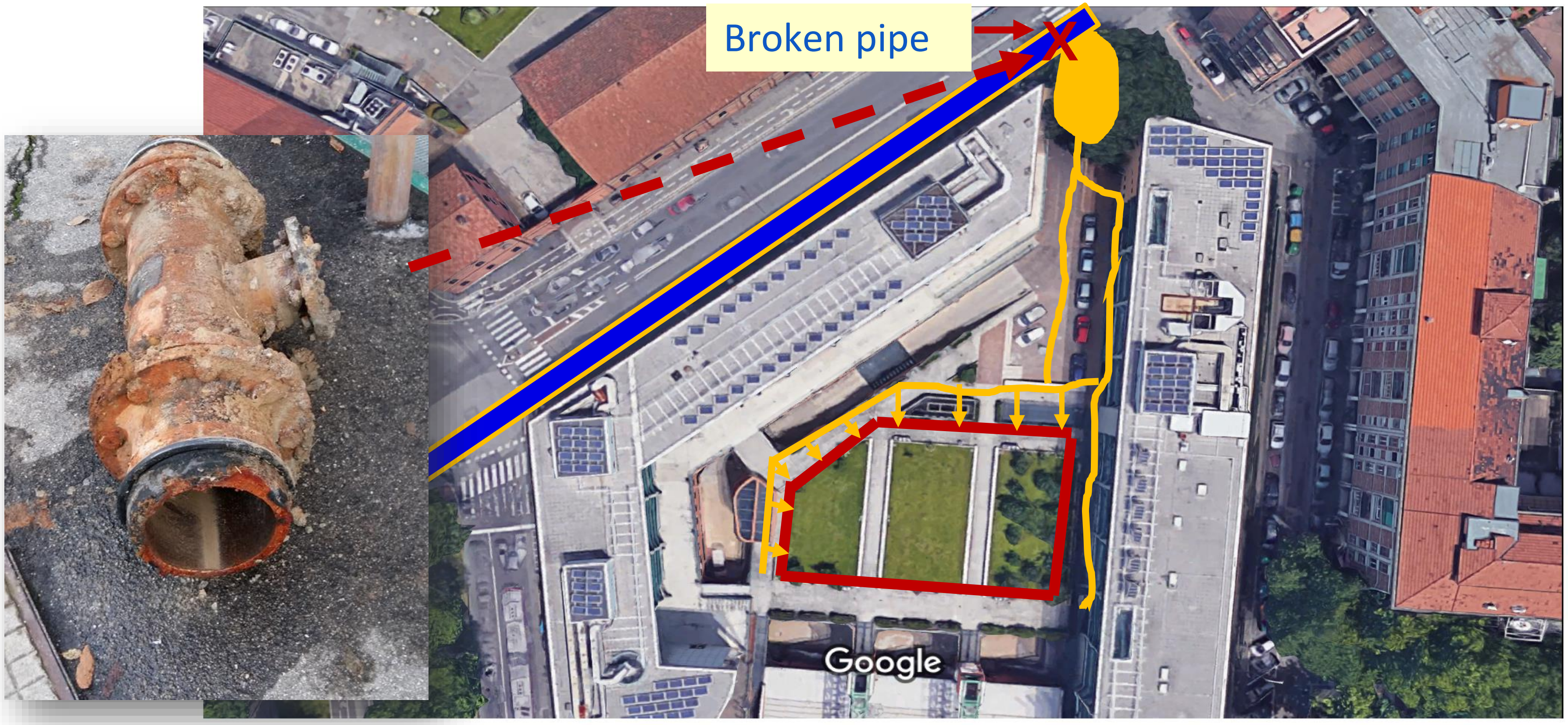






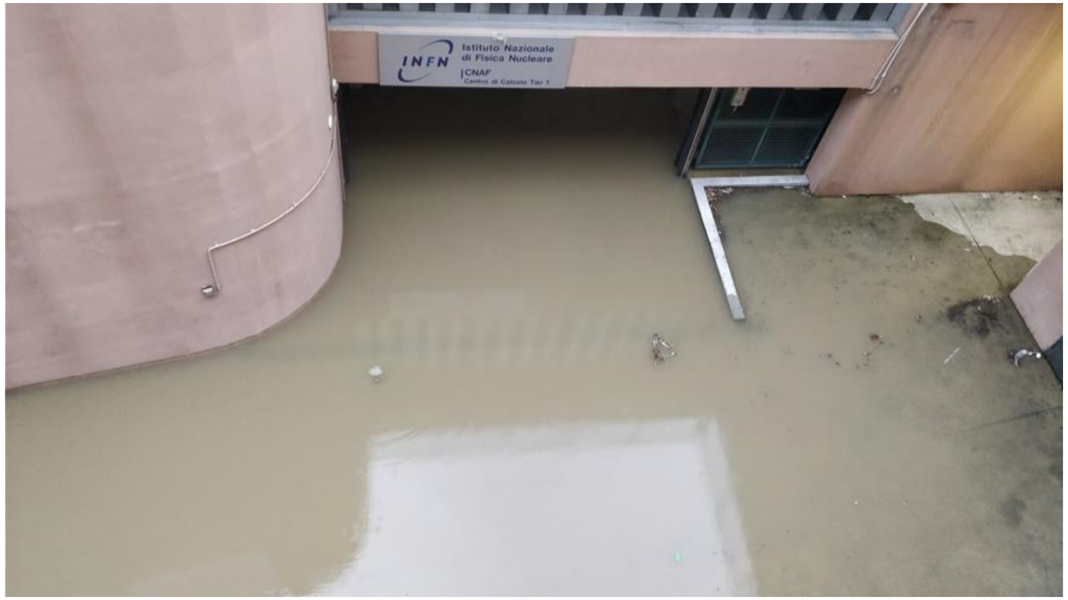
11/9: the flood

- The flood happened on November 9 early in the morning
  - Breaking of one of the main water pipelines in Bologna
  - Also the road near CNAF seriously damaged



Immagini ©2017 Google,Dati cartografici ©2017 Google 10 m

The Tier-1 entrance that morning



All Tier-1 doors are watertight
Height of water outside: 50 cm
Height of water inside: 10 cm (on floating floor) for a total volume of ~500 m^3

First inspection damage estimation

- Nearly all the electrical equipment in the electrical room damaged by the water
  - Both power lines compromised
- The two lower units of all racks in the IT halls submerged
  - Including the two lowest rows of tapes in the library
- The 3 Core Switch/Routers and the General IP Router were installed starting from unit 3! (Safe for few centimeters)

Damage to IT equipment (1)

- **Computing farm**
  - □ ~150 main boards (34 kHS06) are now lost (~10% of the total 2018 capacity of 320 kHS06)
- **Library** (Oracle T10000) and HSM system
  - □ 1 drive damaged
  - □ Cleaning needed (completed)
  - □ Recertification needed (completed)
- **Tapes**
  - □ 136 tapes damaged on 6500 total tapes (Mainly CMS)
  - □ “wet” tapes are being recovered in Oracle lab
    - Very slow process (ONGOING, ETA: end of April)
    - Prioritization based on requests from experiments

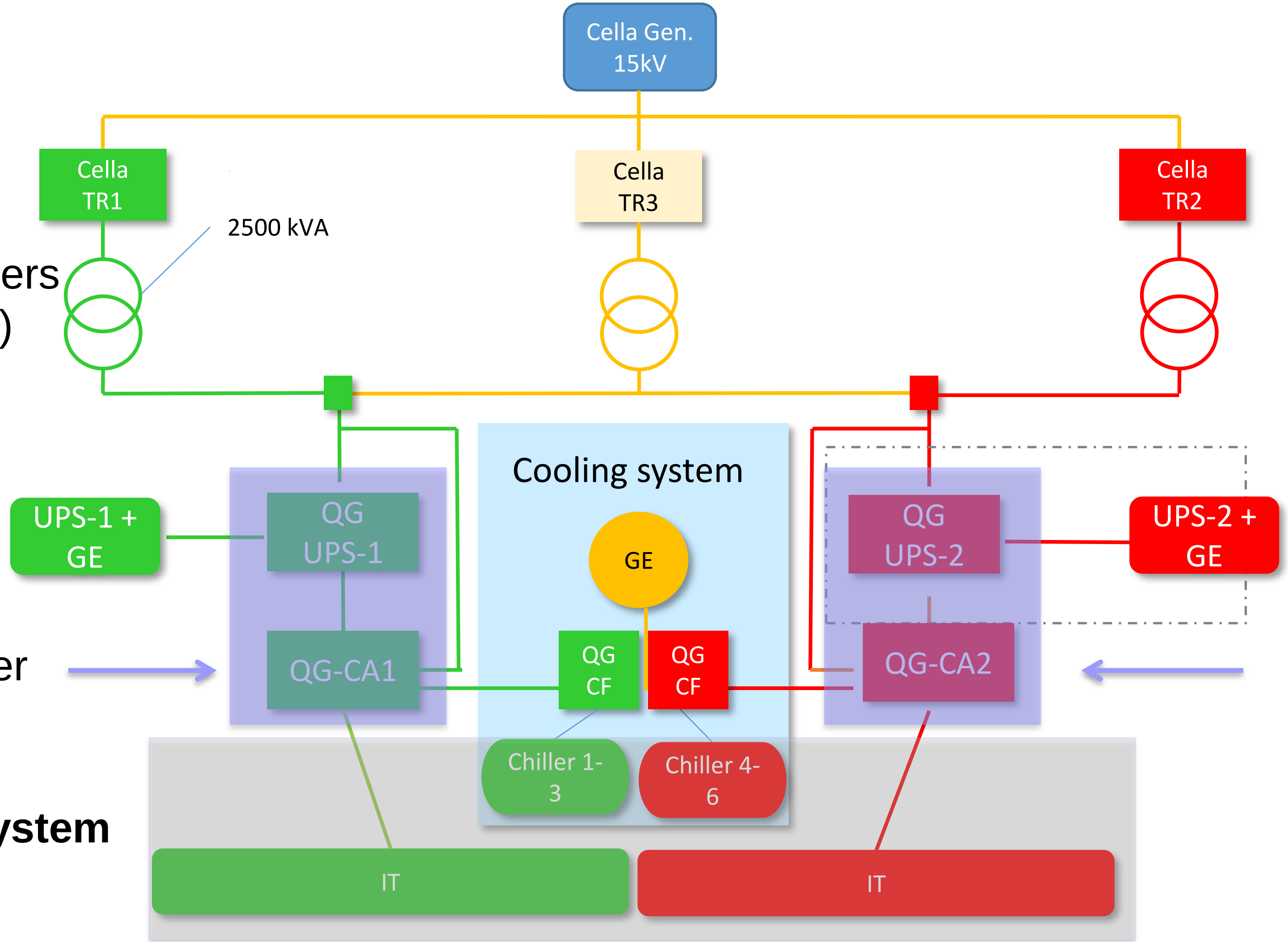
Damage to IT equipment (2)

- Nearly all **storage disk systems** impacted (All the JBOD installed in lowest Unit)
  - 11 DDN JBODs (LHC, AMS)
    - RAID parity affected
  - 2 Huawei JBODs (all non-LHC experiments excepting AMS, Darkside, Virgo)
  - 2 Dell JBODs including controllers (Darkside and Virgo)
    - Most critical - 2 trays out of 5 went underwater.
  - 4 disk-servers (4 Alice) + 4 TSM-HSM servers

System	PB	JBODs	Disks	Involved experiments
Huawei	3.4	2	150 x 6 TB	Astro-particle and nuclear experiments excepting AMS, Darkside e Virgo
Dell	2.2	2	120 (48) x 4 TB	Darkside and Virgo
DDN 1,2	1.8	*4*		ATLAS, Alice and LHCb
DDN 8	2.7	*2*		LHCb
DDN 9	3.8	2		CMS
DDN 10, 11	10	3+2	252 x 8 TB	ATLAS, Alice and AMS
Total	**23.9**	**17**	**~4 PBytes**	

Power Center configuration before the flood

ElectricalTransformers
(Energy Supplier)



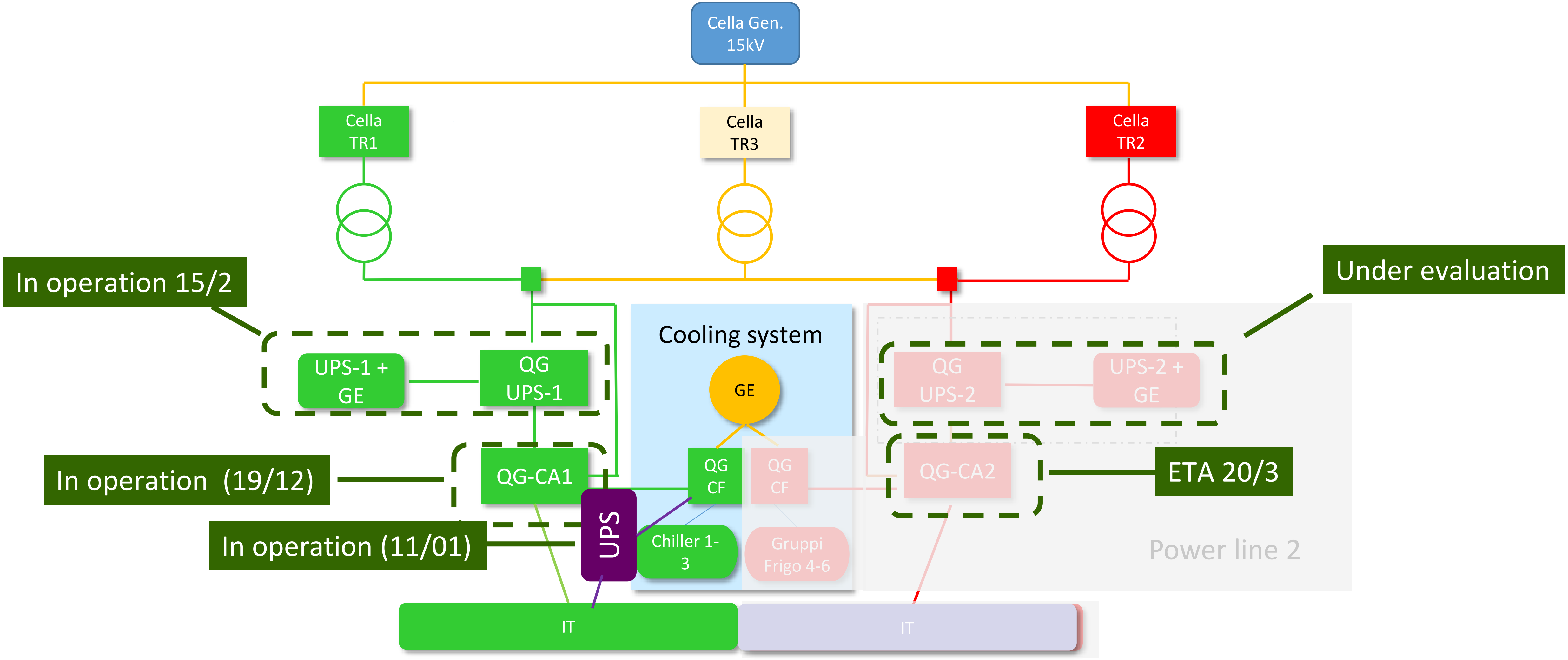
Damaged by water

Damaged by water

**All the electrical system
Affected !**

QG=Electric Panel
GE=Generator

Present Power Center status



In operation 15/2

In operation (19/12)

In operation (11/01)

Under evaluation

ETA 20/3

Cooling status and recovery

- Cooling
  - 3 chillers (out of 6) in operation since Jan 15
  - To turn on the other 3 chillers is needed the 2nd electrical line operational
    - Limit the farm power
  - In-row cooling systems checked and recertified

IT services recovery and Halls cleaning

- Basic services
  - □ **IT services** (non scientific computing) immediately moved outside CNAF (Recovered in 2 to 5 Days)
  - □ **The General IP connectivity** in the area restored 2 days after the flood (Also thanks to GARR effort !)
- Halls
  - □ Data center dried over the first week-end and cleaned from dust and mud completed during the first week of December
- In the meanwhile all activity to recover wet IT equipment have been done:
  - □ Cleaning and drying disks, servers, switches (using oven when appropriate)
  - □ IT components to be replaced have been ordered
- Deep inspection of the data center to understand the flow of the water
  - □ Now understood: water broke in through various “sources” on the perimeter and below the datacenter (Actions to improve the insolation of the perimeter are ongoing)

Storage recovery status

System	Type	Status	Readiness
Alice	Tape buffer	OK	**PRODUCTION**
	Disk	Parity ok	**PRODUCTION**
Atlas	Tape buffer	OK	**PRODUCTION**
	Disk	parity ok	**PRODUCTION**
CMS	Disk + tape buffer	**Degraded parity: raid5 in few LUNs, raid 6 in the others** **Disks to be replaced**	**PRODUCTION**
LHCb	Disk + tape buffer	Data to be moved to the new system	**Moving data to new storage ongoing.** **Almost ready for production**
AMS	Disk	Parity ok	**PRODUCTION**
Virgo	Disk + tape buffer	OK	**PRODUCTION**
Darkside	Disk	OK	**PRODUCTION**
Astro-particle experiments	Disk (3.4PB)	Maintenance intervention	**Recovered 1/3 of files**

About 95% Disk subsystems have been recovered.

All LHC Disk storage is back in production as Virgo and Dark Side .

But… There is the risk of 2.2 PB data loss on other astro particle subsystems

Farm recovery

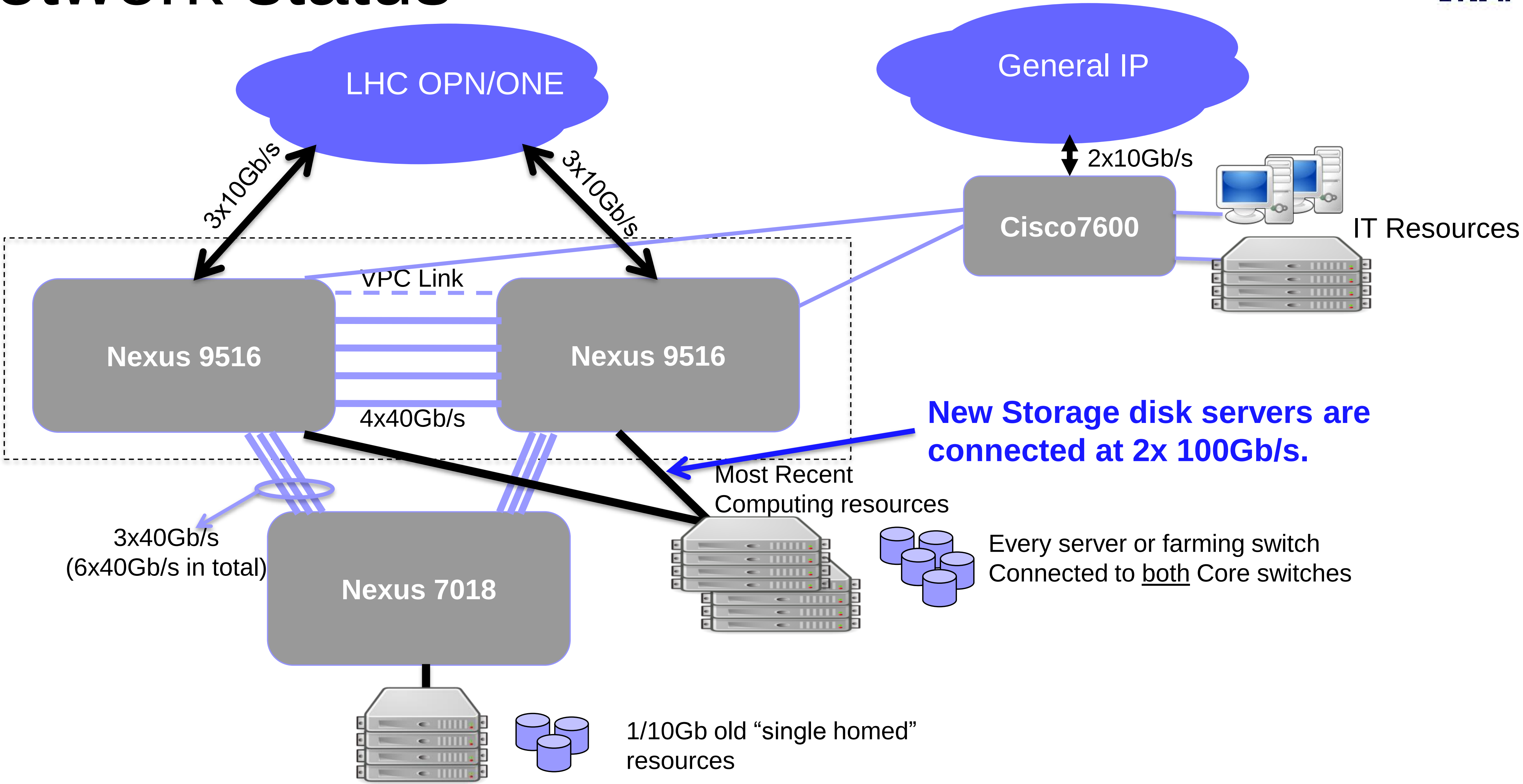
- Test and recovery of farm services completed
  - LSF masters, CEs, squids etc…
- Performed upgrade of WNs
  - Middleware, security patches (i.e. meltdown etc..)
- Part of the local farm powered on (only 3 chillers in production)
  - ~140 kHS06 at the moment (out of ~200kHS06 available)
  - Exploiting the CNAF farm elastic extension to provide more computing power
    - Remote farm partition in Bari-RECAS (~20 kHS06) – In production
    - Installing the CNAF-CINECA farm extension (~ 170 kHS06)

Network status

- Core switches (2x Cisco Nexus 9516) upgraded in December
  - All Fabric changed with (N9K-C9516-FM-E) and each CORE expanded with 32 x 100G Ethernet port modules (N9K-X9732C-EX)

- Necessary to install the new storage.
  - New Disk Servers are connected at 2x100Gb/s to the core switches. (Installation done - commissioning phase)
- Necessary for DCI to CINECA (Remote farm extension) (Done)
  - 4x100Gb/s Ethernet extension (upgradable to 12x100Gb)
- Necessary to Upgrade TIER1 WAN Access from 6x10Gb to 2x100Gb– (Q1-Q2 2018)

Network status



New Storage disk servers are connected at 2x 100Gb/s.

Every server or farming switch
Connected to both Core switches

1/10Gb old “single homed” resources

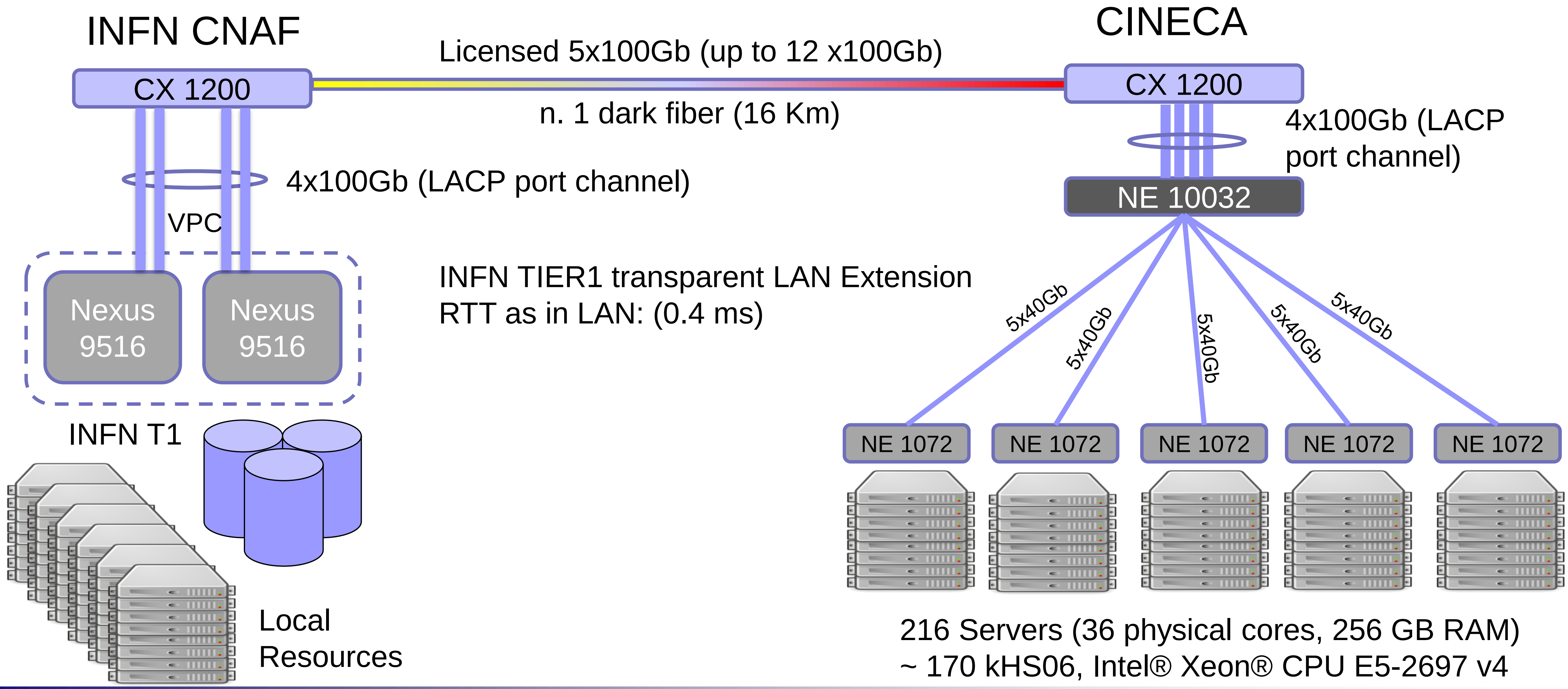
INFN T1 farm extension to CINECA

- CINECA is the largest Scientific HPC Center in Italy and the (14# TOP-500) it is located in Bologna 16 Km far from INFN T1
- Since last year there is a collaboration agreement between INFN and CINECA (INFN entered in CINECA board)
- The agreement for both HPC and HTC. In this framework a subset of Marconi A1 partition has been refurbished to be configured as (remote) Tier1 farm
- So we Extended our LAN to Cineca in order to use those nodes “as they were local” to INFN T1 Computing Center.

- DCI with Infinera Cloud Express© 1200 (CX1200):
  - □ 1.2 Tb/s in 1 U Device (12x100Gb Ethernet interfaces)
  - □ 5 ,5 µs Delay
  - □ Up to 27 Tb/s on 1 pair of fibers (stacking CX1200)
  - □ CX1200 can be configured by CLI or managed via DNA and SDN control (API Driven)

CX1200

T1 extension: CNAF- CINECA Data Center Interconnection
(In collaboration with GARR)



INFN CNAF

CX 1200

Licensed 5x100Gb (up to 12 x100Gb)

n. 1 dark fiber (16 Km)

CINECA

CX 1200

4x100Gb (LACP port channel)

VPC

Nexus 9516

Nexus 9516

4x100Gb (LACP port channel)

NE 10032

INFN TIER1 transparent LAN Extension
RTT as in LAN: (0.4 ms)

5x40Gb

5x40Gb

5x40Gb

5x40Gb

5x40Gb

NE 1072

NE 1072

NE 1072

NE 1072

NE 1072

INFN T1

Local Resources

216 Servers (36 physical cores, 256 GB RAM)
~ 170 kHS06, Intel® Xeon® CPU E5-2697 v4

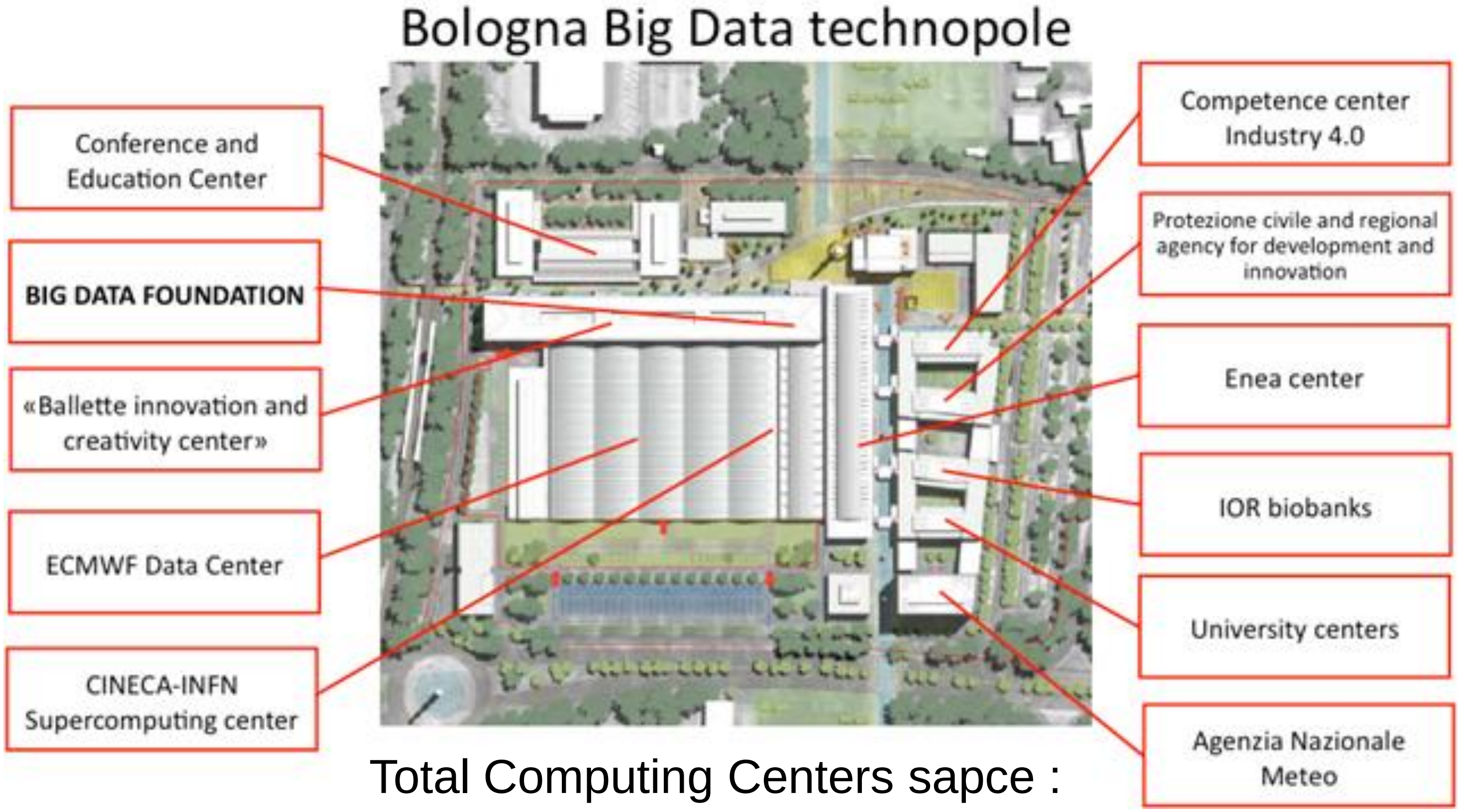
Search for a new location for the Tier-1

The goal: provide a new location for the INFN Tier-1 to take into account future expansion.

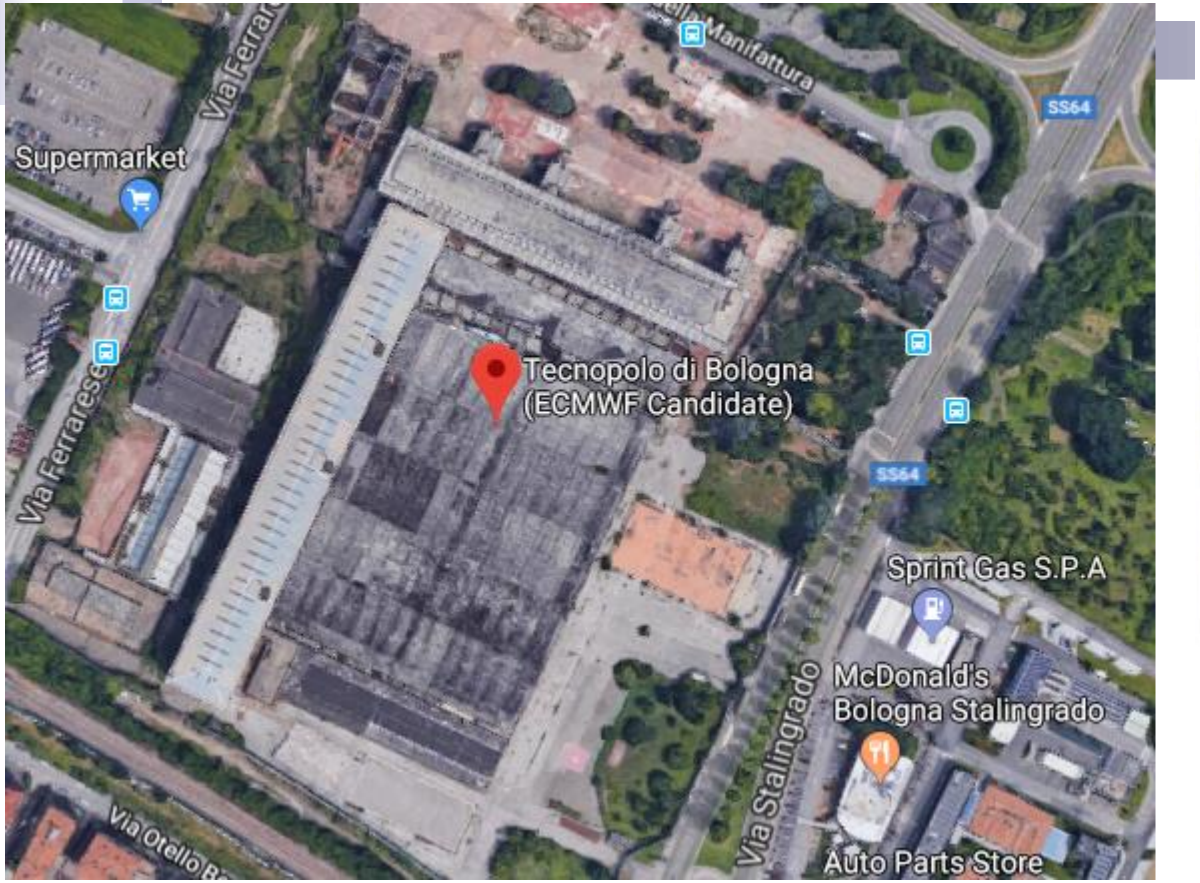
ECMWF center will be hosted in Bologna from 2019 in the “Tecnopolo” area.

Possibility to host in the area also:
- INFN Tier-1 (3000 sqm)
- CINECA computing center (3000 sqm)

Already allocated 40 M€ from the Italian government to refurbish the area.
Looking for extra budget for INFN & CINECA


Bologna Big Data technopole
Conference and Education Center
BIG DATA FOUNDATION
«Ballette innovation and creativity center»
ECMWF Data Center
CINECA-INFN Supercomputing center
Competence center Industry 4.0
Protezione civile and regional agency for development and innovation
Enea center
IOR biobanks
University centers
Agenzia Nazionale Meteo

Total Computing Centers sapce : 21000 sqm


ella Manifattura
Via Ferrar
Supermarket
SS64
Tecnopolo di Bologna
(ECMWF Candidate)
Via Ferrarese
SS64
Sprint Gas S.P.A
McDonald's
Bologna Stalingrado
Via Stalingrado
Via Orello Bo
Auto Parts Store

Tecnopolo

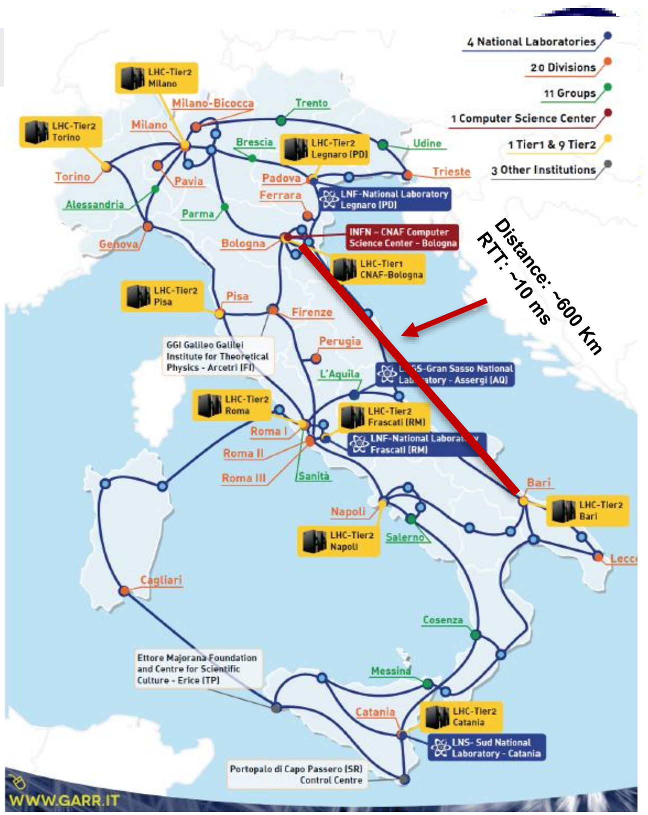
Thank you!

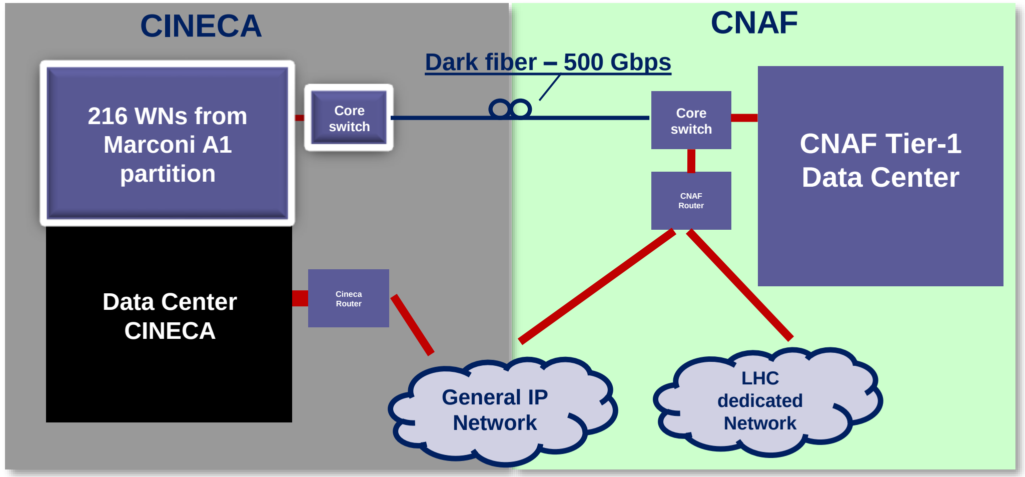
Backup Slides

Farm remote extensions

- ~13% of CPU resources pledged to WLCG experiments are located in Bari-RECAS data center
  - Transparent access for WLCG experiments
  - Similar to CERN/Wigner extension
  - 20 Gbps VPN
  - Disk cache provided via GPFS-AFM
- In 2018 ~170 kHS06 will be provided by CINECA
  - Setup on going
  - 500 Gbps (→ 1.2 Tbps) VPN ready
  - No disk cache, direct access to CNAF storage
    - Quasi-LAN situation

- Participation to HNSciCloud project
- Tests of opportunistic computing on a commercial cloud providers (Aruba, Azure)





CINECA

CNAF

Dark fiber – 500 Gbps

216 WNs from Marconi A1 partition

Core switch

Core switch

CNAF Tier-1 Data Center

CNAF Router

Data Center CINECA

Cineca Router

General IP Network

LHC dedicated Network

Summary

- One electrical line recovered
  - □ UPS on this line available end of this week
  - □ It allows to switch on the storage and part of the farm (due to cooling constraints)
- Farm services recovered
  - □ Missing resources (34 kHS06) provided by remote sites
    - INFN-Bari already in production (24kHS06), CINECA during 2018
- Storage for CMS, ATLAS, ALICE, AMS is ready and moved to production
- We need to find a temporary solution for LHCb
  - □ All data has to be moved to a new Storage system not yet ready
- Serious issues with astro-particle experiments storage
  - □ Not for AMS, VIRGO - in Production