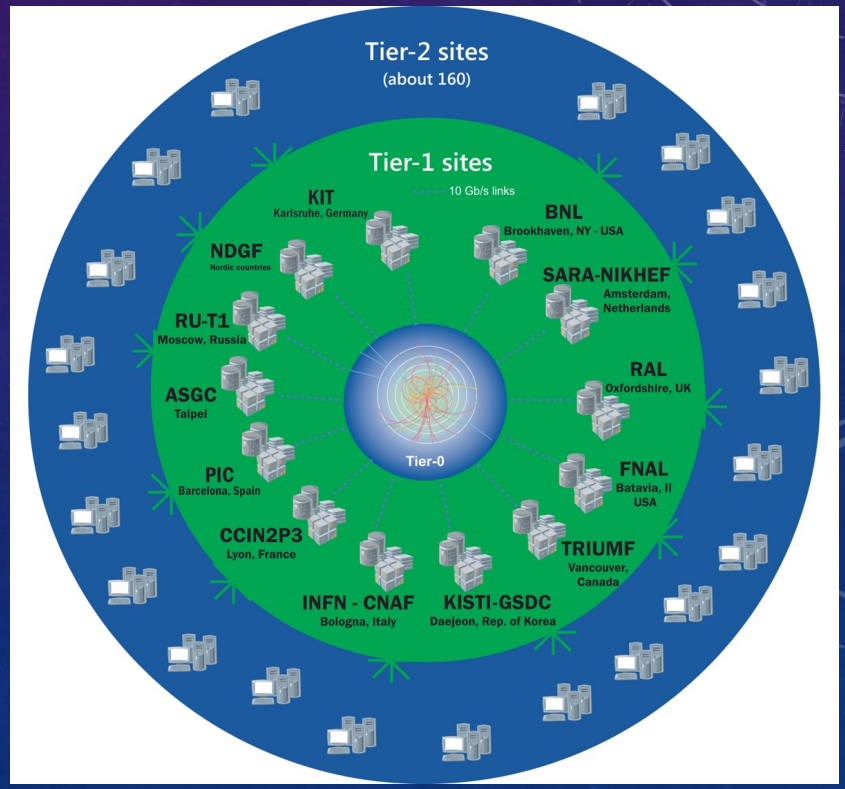
SERBIAN CMS TIER-1 HIGH LEVEL DESIGN

VLADIMIR REKOVIC
VINCA INSTITUTE, UNIVERSITY OF BELGRADE

16.09.2024
SERBIAN TECHNICAL VISIT TO CERN

WLCG AND SSC-T1

- Serbian Scientific Compute – Tier 1 (SSC-T1) oneclients of the State Data Centers-KG
- Infrastructure of SDC-KG well in place, highest standards
- MoU signed, now in the process of implementing it



SCIENTIFIC LOOK AT SSC-T1

- Most of T1s have evolved from already existing HEP computing centers , and their designs often were restricted by legacy systems,
- SSC-T1 is starting from scratch and has a unique opportunity to freely develop into an optimal scientific and research platform serving
  - CMS Experiment (for now)
  - Advanced computing technology innovation
- It is a learning experience for the Serbian HEP and IT community
  - So start simple

TIER-1 PHYSICS FUNCTIONALITIES

Functions for experiment computing

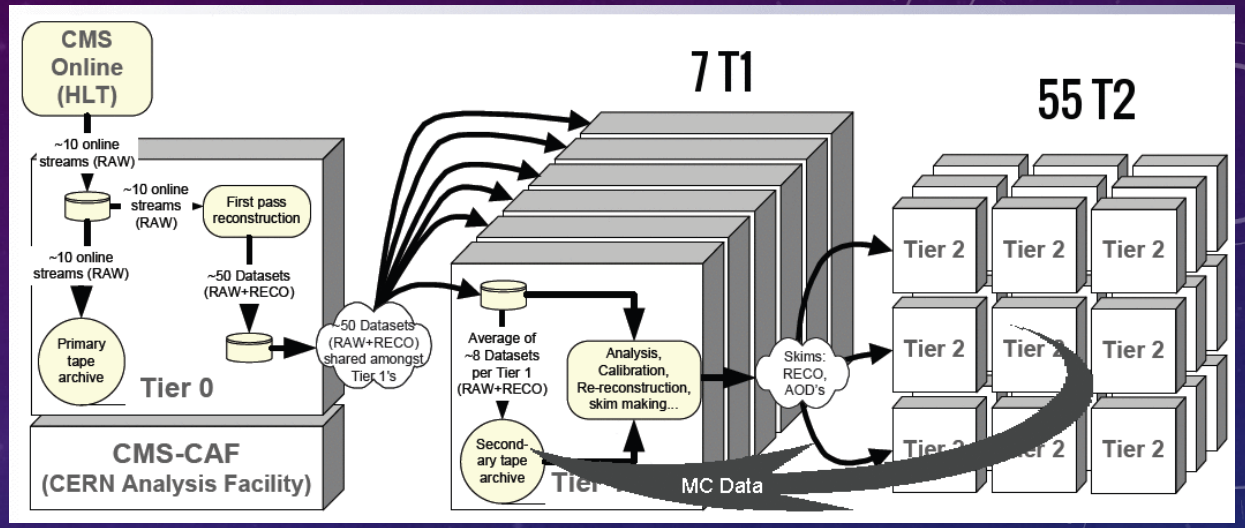
- data processing
  - Real and MC on processing units (CPU, GPU…)
- data storage.
  - Warm (disk-based)
  - Cold or archival (tape-based)

Stored data:

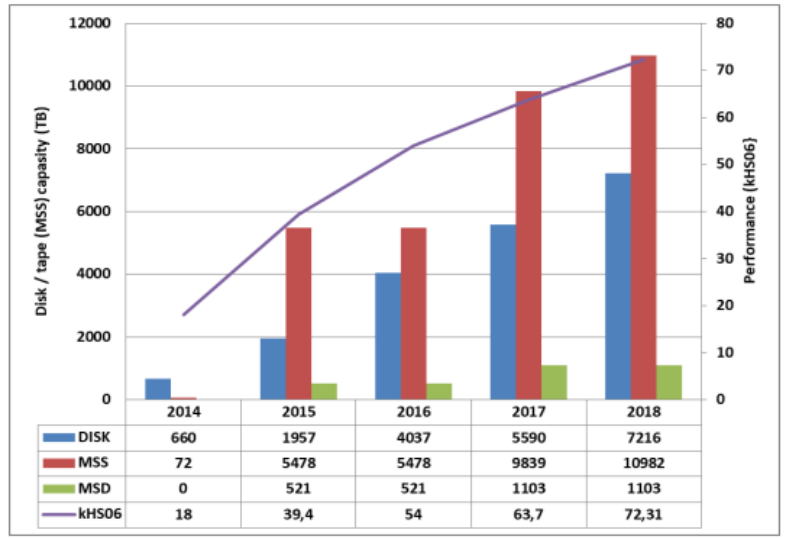
- raw data from T0 and locally produced MC data
- AOD production and storage
- physics data used in end user analysis.

Requirements:

- scalability up to hundreds PB,
- high availability (for data taking and processing),
- data-intensive processing with high I/O performance,
- data portability to Grid service.



Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018), Dubna, Moscow region, Russia, September 10 - 14, 2018

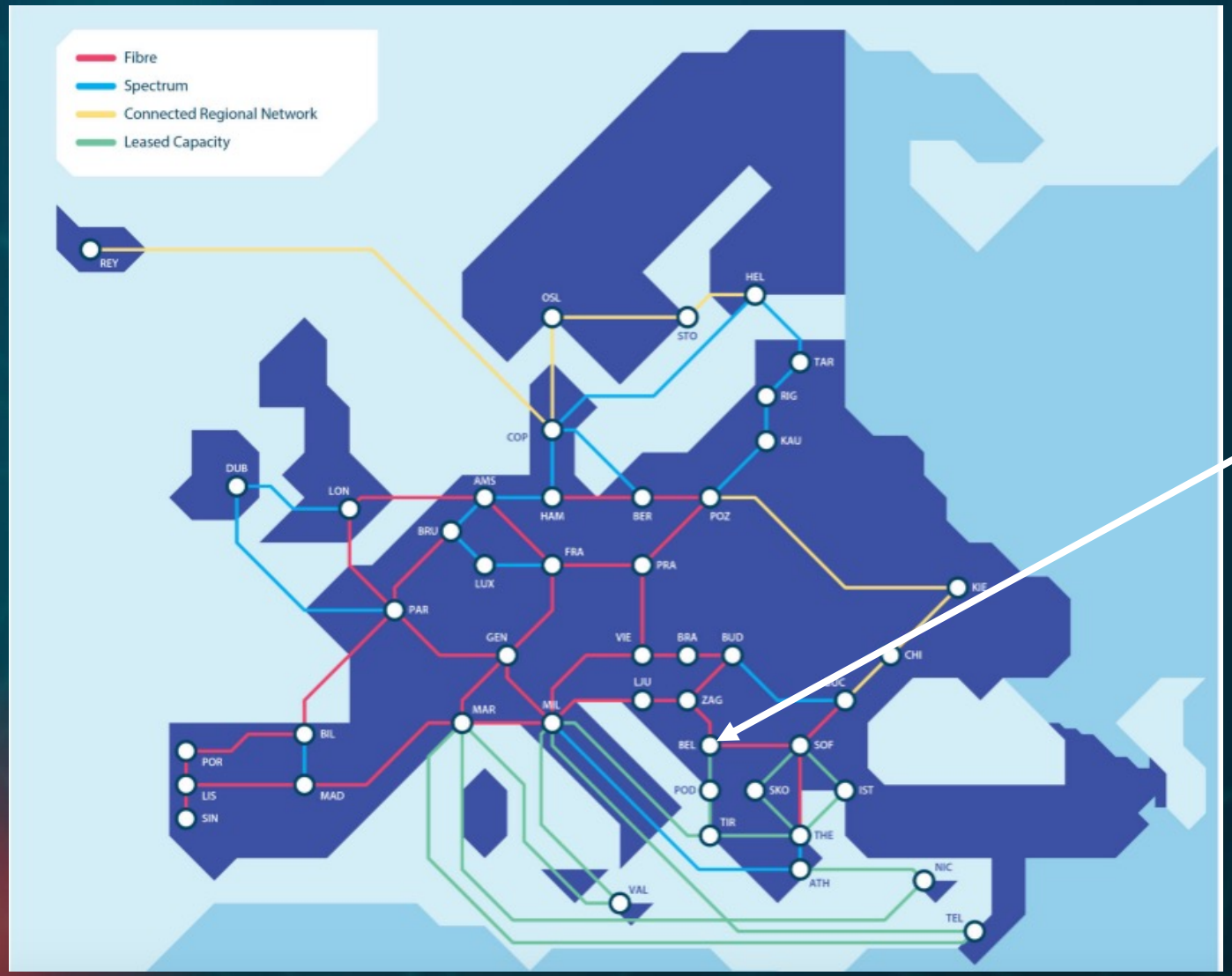


	2014	2015	2016	2017	2018
DISK	660	1957	4037	5590	7216
MSS	72	5478	5478	9839	10982
MSD	0	521	521	1103	1103
kHS06	18	39,4	54	63,7	72,31

Figur

Example: JINR (RU) CMS Tier-1

SERBIA NETWORK CONNECTIVITY



Serbia well integrated in GEANT.
- collaboration of European National Research and Education Networks (NRENs)
- Academic Network of Serbia (AMRES)

Connection to CERN good.
- Belgrade and DC connected to CERN via DarkFiber

Discussed connectivity requirements w/ CERN experts and assessed by DC experts
- LHC-OPN (OK, 1 or few x 100 G via DarkFiber)
- LHC-ONE (OK, via State TC, via Amsterdam,CERN)

All confirmed by DC as doable.

SSC-T1 (SERBIAN SCIENTIFIC COMPUTING – TIER1)

SSC-T1 Goal:
-average size capacity T1
- provide 15% of CMS T1's capacity

First proposal SSC-T1 2023
Now we are towards 2024/2025
Take 2024 numbers at least.

The Request

CMS		'23 Approved Request - Spring '22	'24 Preliminary Request - Fall '22	'24 Final Request Spring '23	Increase wrt '23 Abs.	Increase wrt '23 Perc.
CPU [kHS06]	Tier-0	720	750	980	260	36%
	Tier-1	800	860	930	130	16%
	Tier-2	1,350	1,500	1,600	250	19%
	Total	**2,870**	**3,110**	**3,510**	**640**	**22%**
Disk [PB]	Tier-0	45	52	54	9	20%
	Tier-1	98	108	122	24	24%
	Tier-2	117	130	149	32	27%
	Total	**260**	**290**	**325**	**65**	**25%**
Tape [PB]	Tier-0	228	293	320	92	40%
	Tier-1	316	370	380	64	20%
	Total	**544**	**663**	**700**	**156**	**29%**

J. Letts, D. Piparo, CRB, March 15, 2023

15

SSC-T1 (SERBIAN SCIENTIFIC COMPUTING – TIER1)

SSC-T1 Goal:
-average size capacity T1
- provide 15% of CMS T1's capacity

First proposal SSC-T1 2023
Now we are towards 2024/2025
Take 2024 numbers at least.

Or even better pledged 2025 numbs
15% is
CPU - 165 kHS06
Disk – 20 PB
Tape – 65 PB

VERY CLOSE TO MoU pledge

The Request 2025

CMS O&C

CMS		'24 Approved Request - Spring '23	'25 Preliminary Request - Fall '23	Increase wrt '24 Abs.	Increase wrt '24 Perc.
CPU [kHS23]	Tier-0	980	1,180	200	20%
	Tier-1	930	1,100	170	18%
	Tier-2	1,600	1,900	300	19%
	Total	**3,510**	**4,180**	**670**	**19%**
Disk [PB]	Tier-0	54	64	10	19%
	Tier-1	122	142	20	16%
	Tier-2	149	175	26	17%
	Total	**325**	**381**	**56**	**17%**
Tape [PB]	Tier-0	320	420	100	31%
	Tier-1	380	452	72	19%
	Total	**700**	**872**	**172**	**25%**

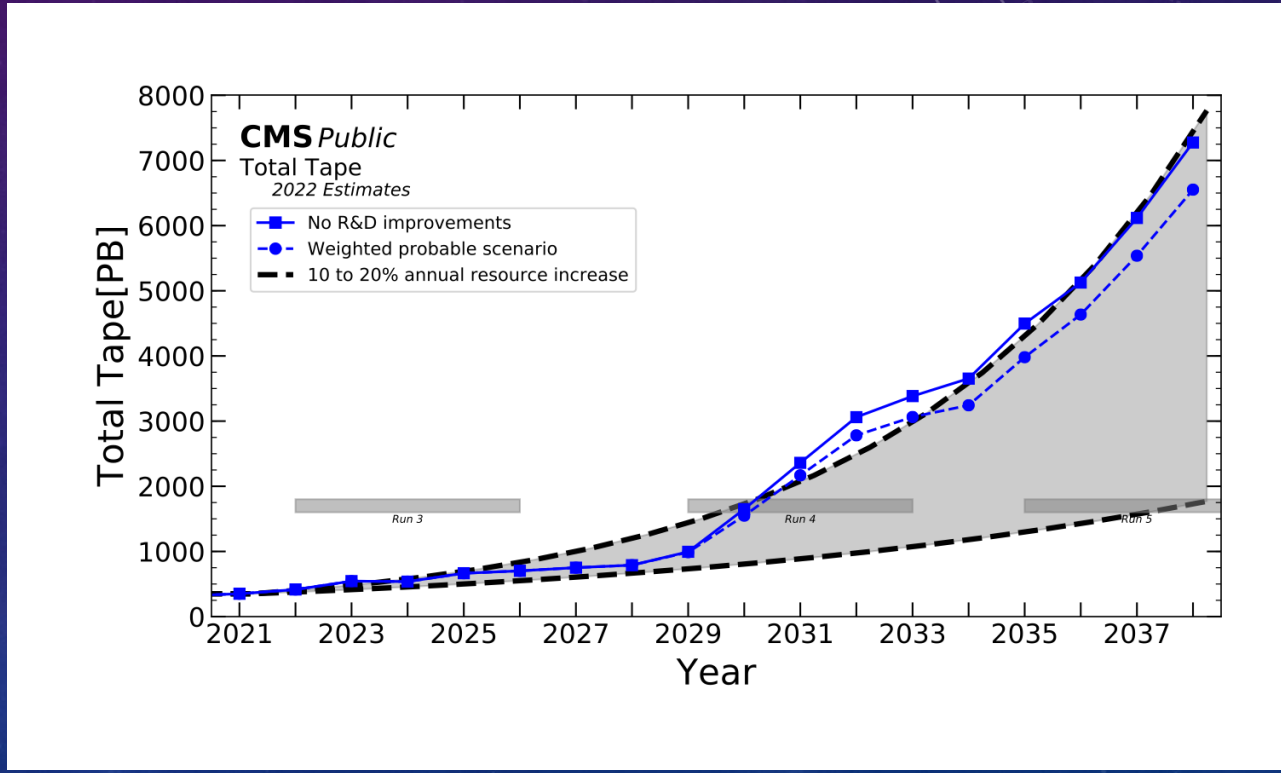
J. Letts, D. Piparo, CRB, March 15, 2023

15

NEED FOR R&D
SSC-T1 SCIENTIFIC CONTRIBUTION TO HEP COMPUTING

- Simple doubling of T1 resources would not suffice
- SSC-T1 would like to contribute in scientific R&D to provide improvements. Potential areas
  - Computing
    - CMSSW
    - ROOT
    - Alternative processors
  - Disc-Storage
    - CEPH for HEP storage

- VERY IMPORTANT aspect of the design is to create a platform so that scientist and students can contribute



HL-LHC (2030-2040) INCREASE IN DATA RATE AND VOLUME

CMS TDR-022 Phase 2 DAQ-HLT TDR

	LHC	HL-LHC	
CMS detector	Run-2	Phase-2	
Peak ⟨PU⟩	60	140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size	2.0 MB [a]	5.7 MB [b]	7.4 MB
Event Network throughput	1.6 Tb/s	23 Tb/s	44 Tb/s
Event Network buffer (60 seconds)	12 TB	171 TB	333 TB
HLT accept rate	1 kHz	5 kHz	7.5 kHz
HLT computing power [c]	0.5 MHS06	4.5 MHS06	9.2 MHS06
Storage throughput	2.5 GB/s	31 GB/s	61 GB/s
Storage capacity needed (1 day)	0.2 PB	2.7 PB	5.3 PB

SCIENTIFIC MISSION FOR DEVELOPMENT

- SSC-T1 will provide a platform to enable research in computing for high energy physics scientific fields:
  - Possible contributions perhaps essential to face the challenge of managing, processing and serving large amounts of HEP and future experiments at CERN.
  - Topics: network and data management, computer security, cold to warm storage, data caching, data compression, innovative algorithms for data processing (also on heterogeneous platforms), novel data center architectures, incorporation of HPC resources for high throughput processing, energy efficient computing, monitoring and big data analysis, AI driven decision making and failure pre-emption.
- Plugged into CMS international collaboration w/other Tier-1's, CERN opens opportunities
  - Train students, exposing them to operational tasks and big data management tools
  - Share best practices, potentially exchange staff
  - Participation in software/computing R&D: opportunity for Ph.D. topics
  - Privileged access to CMS data and software
- Potential cross-fertilization at the host lab/university, e.g. usage of HEP tools and infrastructure for other data intensive analysis/processing use cases, and other scientific fields.
- Impact can be reflected in several conference presentation and journals where we can publish results.

SERBIA CMS TIER-1

- Target 10% of CMS Tier-1s total capacity
- Build Tier-1 in 2 stages
- Stage 1 – initial configuration, 2024
  - 170 kHS06 = 12 kCores = 24 kThreads
  - 20 PB disc storage
  - 40 PB tape storage

} 6 racks + 2 racks + tape library

- Stage 2 – reach adiabatically ’25/ ‘26/ ‘27/ ‘28/ ‘29
  - look for new solutions to spare space in the mini module – capacity 15 racks + tape library)
    - 350 kHS06 = 25 kCores = 50 kThreads
    - 40 PB disc storage
    - 80 PB tape storage

} 10 racks + 4 racks + tape library expanded

PHYSICS NEEDS FOR SERBIA TIER-1

Computing needs :

- 5 MB/s/ thread

Stage 1

- For 24 k threads
  - 960 Mbit /s ~ 1 Tbps
  - x 2 core SW = 2 Tbps

Stage-2:

- 42 k threads
  - 1,900 Mbit /s ~ 2 Tbps
  - x 2 core SW = 4 Tbps

Disc Storage needs:

- 2-3 MB /s/TB (read/write limit)

Stage 1

- For storage of 20 PB,
  - Or, 20 SU @ 1.2 PB
    - 20 x 100 G (100 G @ SU)
  - X 2 core SW = 40 x 100G

Stage-2:

- Disc storage of 40 PB
  - 40 x 100 G, X 2 core SW = 80 x 100 G
- For storage of 100 PB
  - 80 G, X 2 core SW = 160 x 100 G

Tape requirement :

- 10 Gbps / drive

Stage 1

- Tape library 30 PBs
  - 12 drives x 10 Gbps
  - x 2 core SW

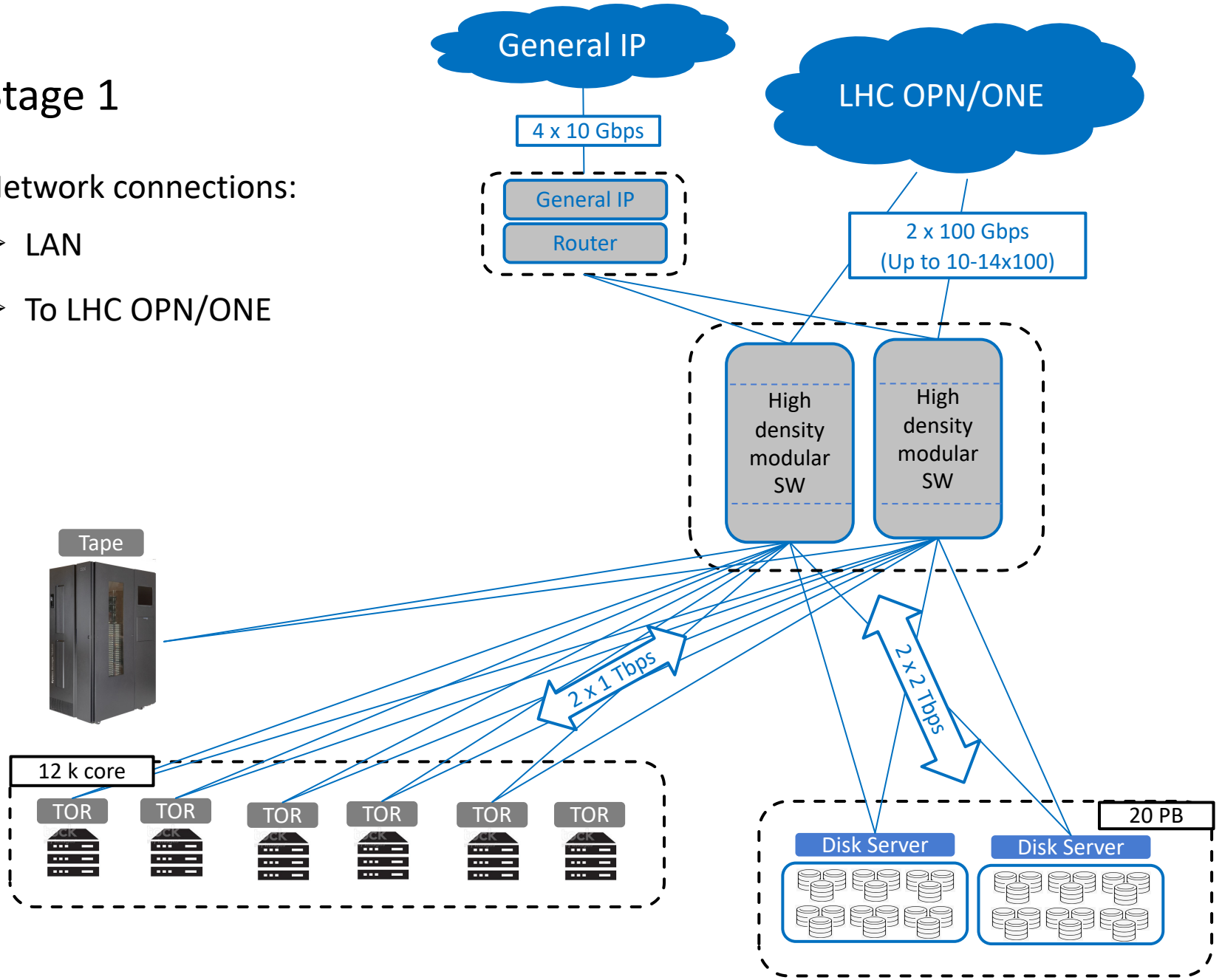
Stage-2:

- Tape library 80 PBs
  - 30 drives x 10 Gbps
  - x 2 core SW

Stage 1

Network connections:

- LAN
- To LHC OPN/ONE

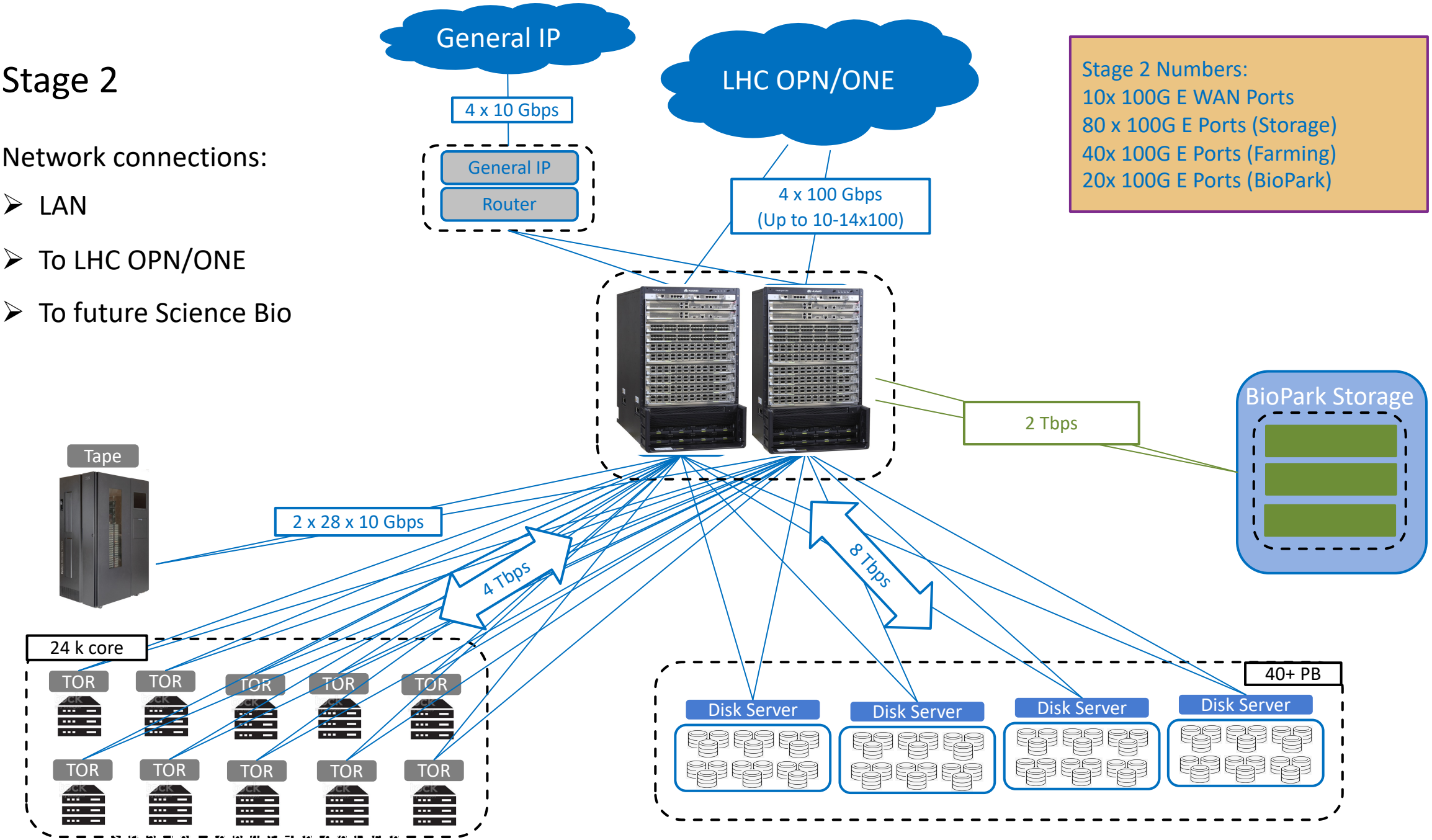


Stage 1 Numbers:
2 x 100 G E WAN Ports
40 x 100 G E Ports (Storage)
20 x 100 G E Ports (Farming)

Stage 2

Network connections:

- LAN
- To LHC OPN/ONE
- To future Science Bio

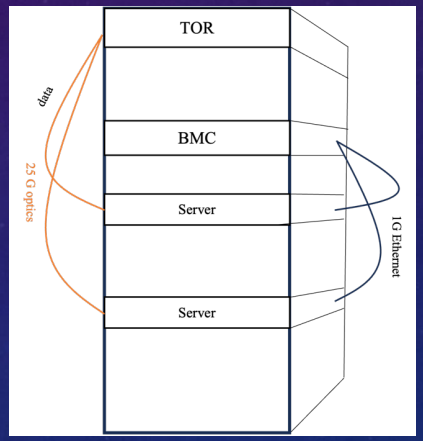


STAGE-1 IMPLEMENTABLE?

- In this presentation we don't present a detailed technical plan but rather discuss the overall architecture motivated by the physics needs and discuss what is needed . More detail is provided in the HLD document posted in the agenda,
  - Physics motivated utilities
    - Processing with HTCondor technology
      - CMS experimental data – Reconstruction
      - CMS MC – generation – simulation – reconstruction - reprocessing
    - Storage
      - EOS + CEPH (limited) technologu
    - Archival
      - Cern Tape Arcive (CTA)
    - Network +Authentication
      - Through EGI or OSG confedrations

COMPUTING

- Computing is done on CPU servers
- Motivated by the CMS Event model (size, content)
- 2-3 GB RAM per core
- 20 GB disk storage per core
  - For 64 core –we need 192 GB RAM + 4 TB discs
- T1 Jobs executed and managed via grid in the Pilot project



ROBOT - TAPE LIBRARY

System Storage (Robot) IBM TS4500

- TS1160 tape drives LTO 9 (of 10) tape technology
  I/O rate of 360-400 MB/s and 750-1000 MB/s (compressed)data.
- Take entry-level configuration(Base licence)
- can be placed in any active position, so the library can grow from both the right and the left side of the first L frame for floorspace flexibility

Tape library:

Stage 1 Storage with Base License 1560 slots:

LTO-9 = 27 PB (now)

LTO-10 = 54 PB (end of 2024)

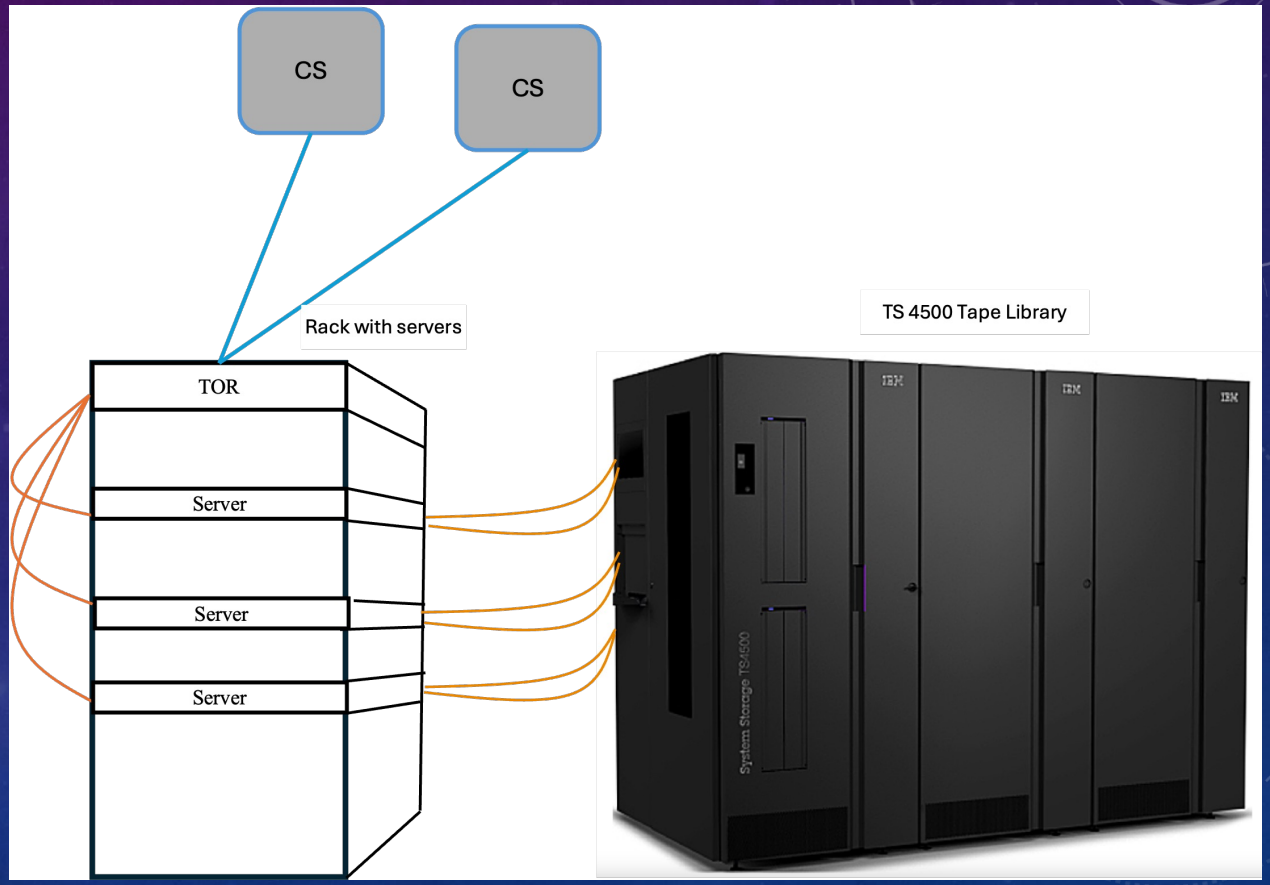
Stage 2 with HD License 5000 slots :

LTO-10 = 160 PB

TAPE LIBRARY – STAGE 1 CONFIGURATION

- Archival Storage System:
  On the right, the tape library IBM TS4500
- Frame configuration L55-S55-D55
- Entry level license

- Tape servers (6) are not visible and are inside of the library
- On the left a schematic drawing of a rack with data servers units each connected to two tape drives. The rack is connected via TOR to the Tier-1s core switches which manage the network of futher connections to disc-based storag, and network with core switches (CS)



CORE SWITCHES

- System: consist of 2 identical Core Switch Routers.
  - Installable in 19" 42U racks
- Modular, expendable

- Capacity of each module
  - 48 ports of 10/25 G E without using breakout cables
  - 36 ports of 100 G E without using breakout cables
- Minimum capacity available at time of delivery (assume 50% free)
  - 48 ports of 10/25 G E, SFP type
  - 72 ports of 100 G E, QSFP type

- Maximum theoretical total capacity of each Switch - ultimate expansion
  - At least 125 ports of 100 G E
  - At least 16 ports of 400 G E

- 10 Gbps Card 48 x 10 GE (180 – 242 W)

1. Forty-eight 10GE optical ports

- 100 Gbps Card 36 x 100 GE (525 – 764 W)

1. Thirty-six 100GE optical ports

POWER CONSUMPTION - STAGE 2

- Farm:
  - CPU servers, 360 U, 100-130 kW
    - @ 280 or 360 W for AMD EPYC 9534 or 9554
    - <12 racks
- Disk Storage
  - Data servers 120 U 8 kW
    - 3-4 racks @ 40U / 3 kW
- Network
  - Core Switches, 32 U or 16U, 10 kW
    - 2 x 16U / 5 kW or 16 U @ 10 kW
    - < 1 rack

32 U (2 x 16 U)
2 x 5 kW

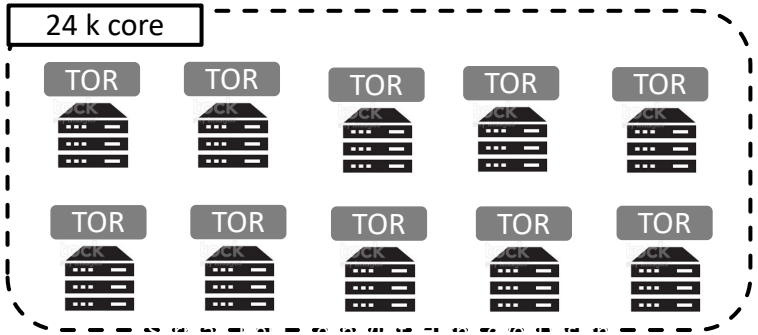
Or

16 U
10 kW

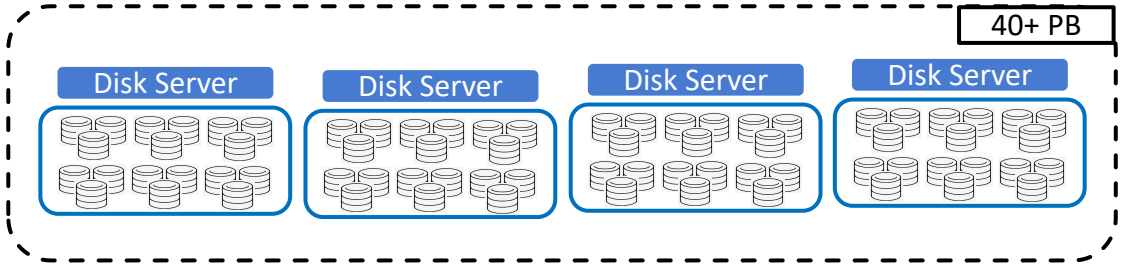
- Tape Library S55-L55-S55-D55-S55
  - Tape drives for 80 PB, 3.1 kW,
    - 24 drives @ 130 W
    - 5 rack size library
  - Tape servers 12 U, 2.4 kW
    - < 1 rack
- Total space
  - 512 U in < 15 racks + 5 rack size library

Total power w/ cooling
- 154 kW

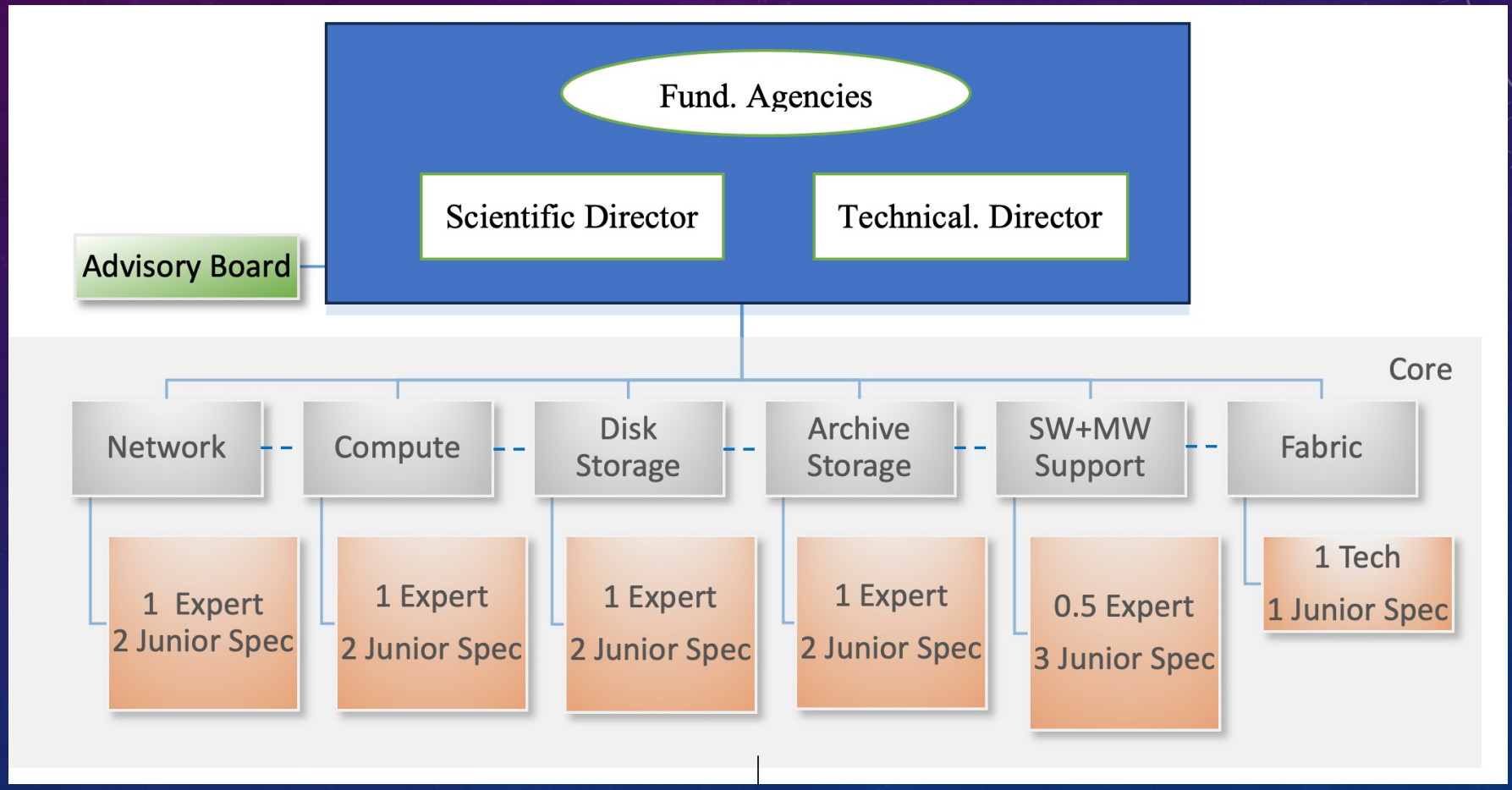
12 racks



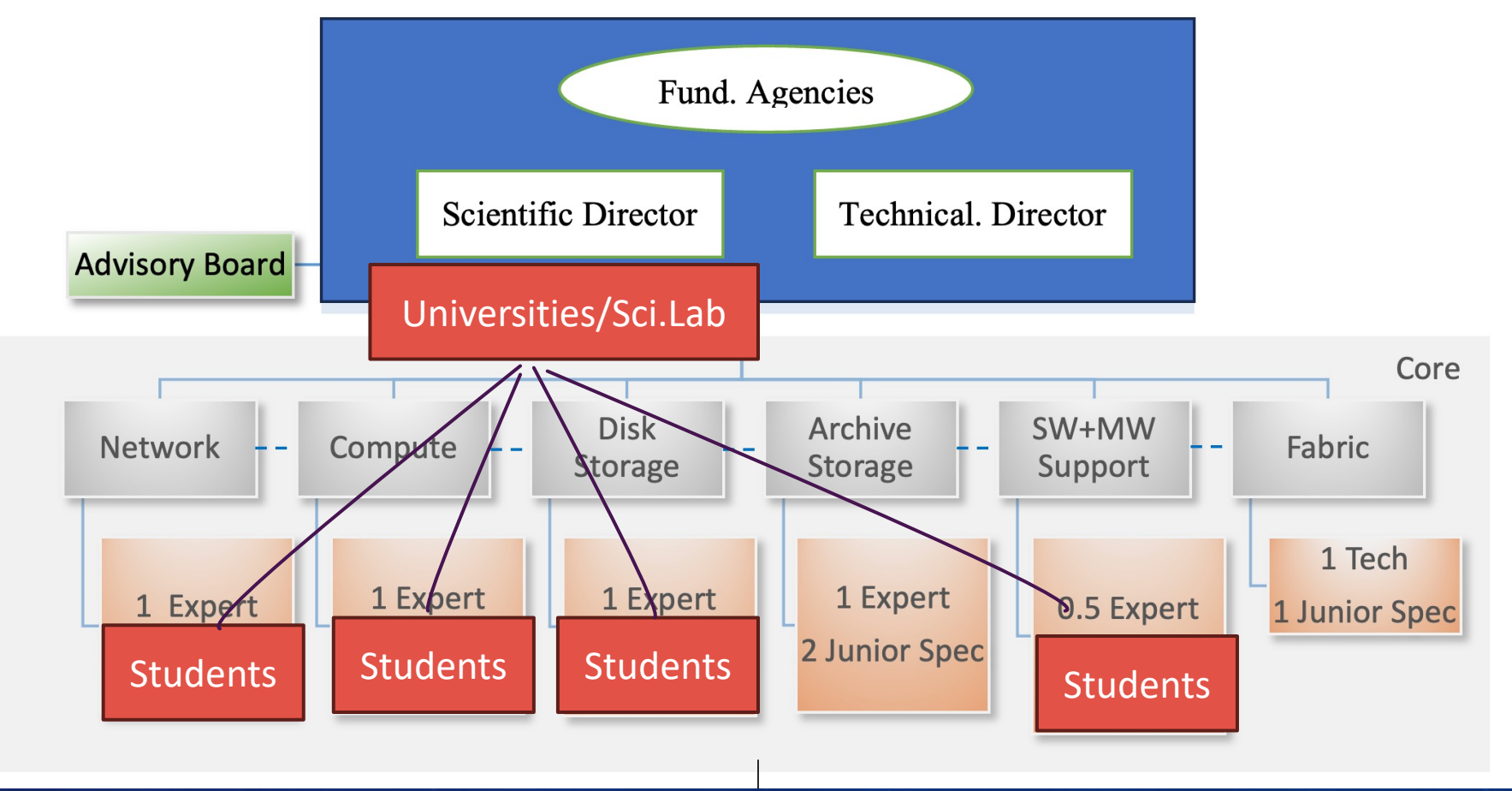
4 racks @ 30 U / 2 kW
8 kW



SSC-T1 ORGANIGRAM



SSC-T1 ORGANIGRAM WITH SCIENCE



MONITORING AND INTERFACE WITH CMS

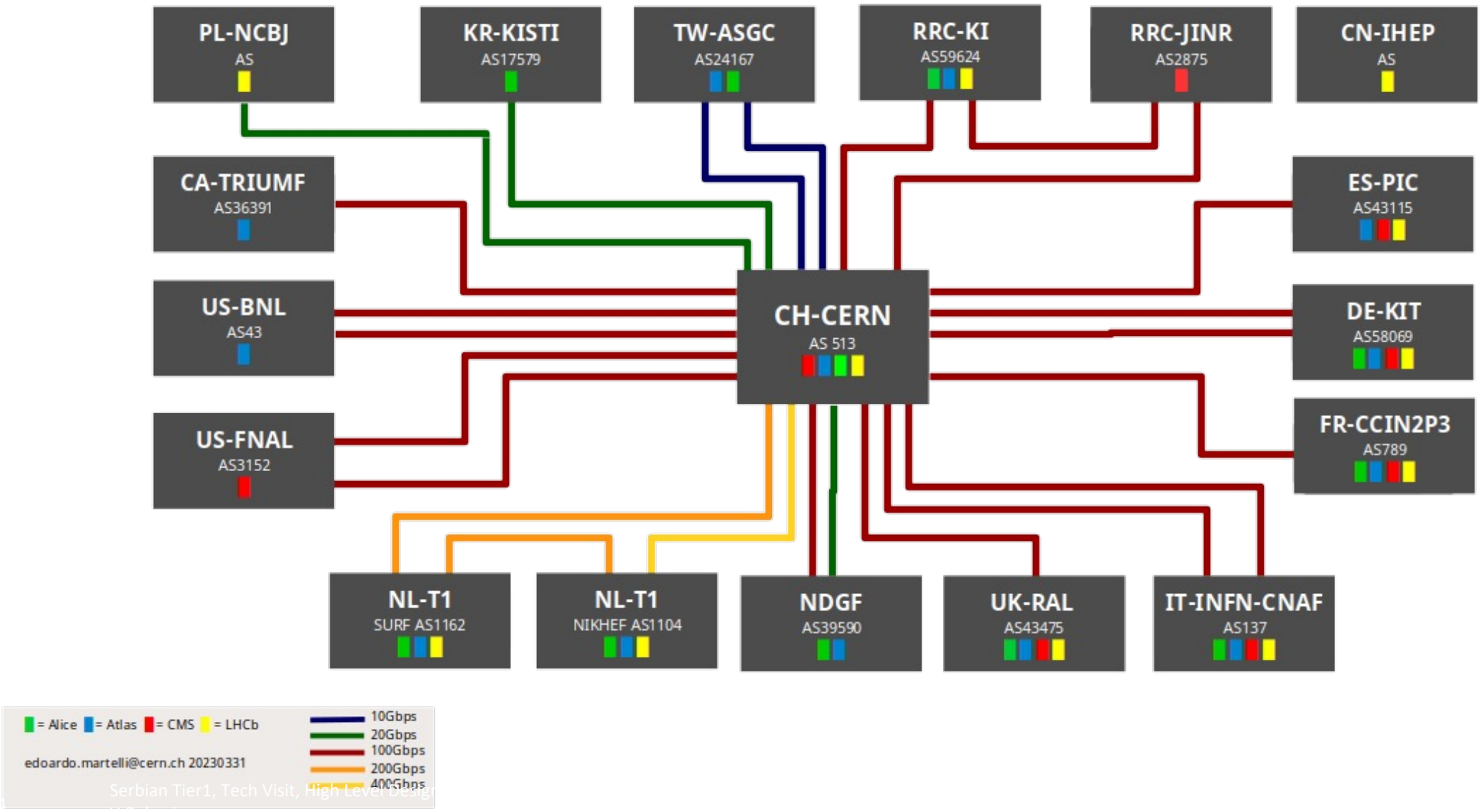
- Local Monitoring will be done with commercially available solutions (Services, Inventory)

  Very important a continuous communication with the experiment
- CMS – Offline and Computing group – science
- CMS Computing Resource Board – science
- CMS Facility Services group - technical
- CMS – Condor Weekly Meetings

BACK UP

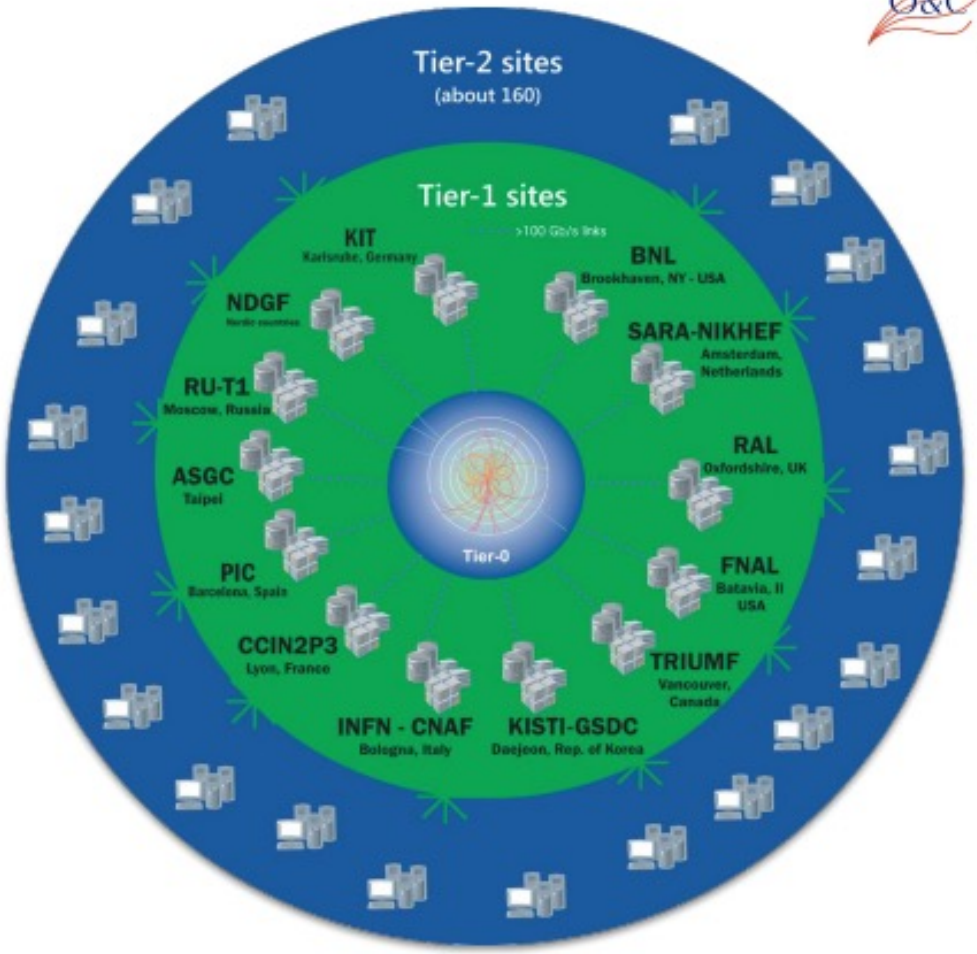
LHCOPN

LHC Optical Private Network - Topology



Our Tier-1's

- **Seven sites**: CCIN2P3 (FR), CNAF (IT), JINR (RU)*, KIT (DE), FNAL (US), PIC (ES), RAL (UK)
- Tier-1's are not necessarily larger than Tier-2's
  - e.g. T2_IN_TIFR pledges 40% more CPU than T1_IT_CNAF
- Tier-1 sites supports a varying range of the CMS needs:
  - FNAL 40%, PIC 4%
- **They are all equally important: they are the pillars of our computing model and data processing**
  - Unique combination of compute, cold, storage, warm storage and high network connectivity
  - High level of support and availability



** Presently used opportunistically and not for RAW custodial storage as per CERN Management and WLCG directives.*

CONNECTIVITY, LHCOPN : T0 – T1

- Dedicated network, provides links from CERN (T0) to T1s
  - Current T0-T1 mostly 100G and well above
- It is used for transfer of experiment RAW data, SIM data, RECO data
- Need L3, 100G minimum
  - Belgrade connected to Geneva (CERN) via dark fiber (OK)
  - Note: by 2029 T0-T1 expected to go 1 Tbps
- T1 must have public IP, with IPv6

- LHCOPN documentation link

Primary information

- Foundations
- LHCOPN AUP
- Routing architecture
- IP addresses
- Routing policies
- Link's information and Site names
- Technical contacts
- Operational contacts

CONNECTIVITY, LHCONE: T1 – T2

- Network for communication between T1 and T2s
  - When LHCOPN broken, used as backup
- Managed through GEANT network (link)
- Local T1 and country's network provider responsibility to cover costs
- CERN can lend connection Geneva(CERN)- Amsterdam (Research Park) for 1 year

LHCONE L3VPN Service

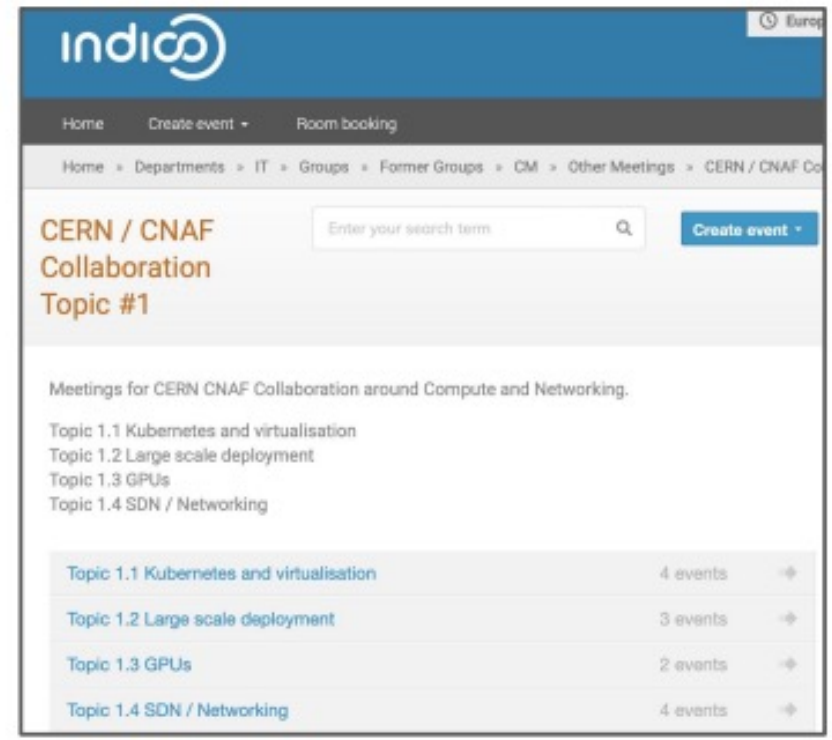
- **L3VPN Service**, **Maps**
- How to connect to the LHCONE L3VPN
- LHCONE AUP
- How to provide a regional instance of the LHCONE L3VPN backbone (for Network Operators)
- MTU recommendations

Tier-1 Collaborations with CERN

- Collaborating with CERN also means collaborating with the Member States and the HEP field at large

Some Examples:

- CNAF (IT) developments of token-based authentication: IAMS [CHEP talk]
  - Deploying industry-standard solutions for Grid authentication in WLCG
- All LHC Tier-1's: Data [talk] and Network scale challenges for HL-LHC
- CERN Tape Management Software (CTA): Fermilab, RAL [talk]
- Storage Technologies:
  - Fermilab: EOS deployment
  - CNAF: EOS and Ceph Integration with K8s [talk]
- Linux Distributions after CentOS:
  - Fermilab - "Future Linux Committee" [talk]
- CERN OpenLab
- Technology transfer, knowledge sharing



SERBIA CERN TIER-1 @ STATE DATA CENTER

SERBIA AT CERN

- In CMS,
  - Total of about 20 researches, engineers, technicians
  - 2001 “Construction”, 2007 “Maintenance and Operations”
  - Phase-II Upgarde MoUs, now **two institutions**
    - **Vinca Institute, University of Belgrade ->** Level-1 Trigger
    - Faculty of Physics, University of Belgrade (FP) -> ECAL Barrel

- In CERN since 1953,
  - Yugoslavia (Serbia) one of the 12 CERN founding members
  - Since 2019, Serbia again Full Member of CERN
  - Groups in CMS (2), ATLAS, ILC, CLIC, ISOLDE, NA61

- In **Vinca** group lots of experience in CMS Core software and TriDAS
  - Design of OpenHLT framework, data streams (HLTMON), Smart prescales
  - DataFormats , EventContent, Emulation workflows, MC, data reprocessing, code optimisation
  - Level-1 Trigger (Phase-1, Phase-2)
- Convenerships (L3/L2)
  - HLT Menu, HLT Performance, Level-1 Trigger Menu, Level-1 Trigger Offline Software

STATE DATA CENTER OPENED IN KRAGUJEVAC

Kragujevac, university town, and one of industrial centers of Serbia

137 km auto-route from Belgade, 1.5 hr comfortable drive

Državni data centar u Kragujevcu

State Data Centar

in Kragujevac since 2022

- the most modern facility of this type in the region built at highest standards
- intended for the storage of data and computing, for use by state, state universities and research institutions, industry.

Certification

ISO 9001:2015
ISO 27001:2013
ISO 27701:2019

EN 50600 class 4 completed: 08/2023

The highest level of data center security certification that meets EU standards.

STATE OF THE ART COMPUTING CENTER IN SERBIA

- TIER 4 according to TIA942 standard
- Providing services with satisfying conditions
  - Connectivity - carrier neutral w/ 2 access points (2N red.)
  - Availability - 24/7 monitoring and surveillance
  - Retainability – guaranteed space and expandability
- Suitable candidate for CERN Tier-1 center

- 8 IT mini-modules with 20 rack cabinets
- 4 UPS rooms
- 2 technical rooms
- Maneuver hall

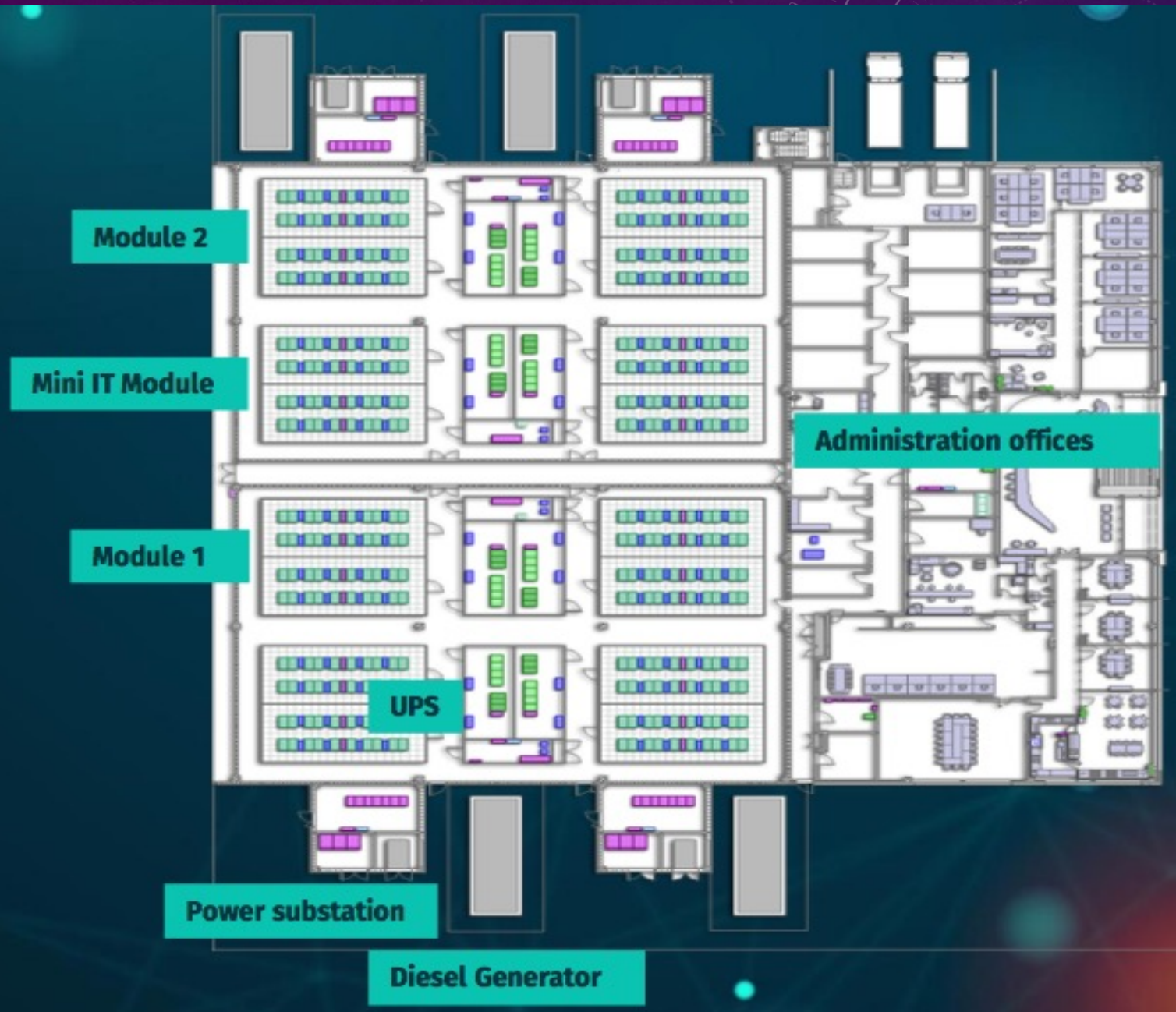
Each IT module has two Power substations and two diesel generators
Each IT module contains 160 rack cabinets

Module configuration
Within each mini module there are:

- 20 Rack cabinets
- 8 In row units
- 2 MDU cabinets

Inside the UPS room there are:
- UPS and batteries
- Air conditioning cabinets

Technical rooms contain:
- Cabinets of energy, EMP and BMS
- Air conditioning cabinets
- Distribution cabinet FD



MINI IT MODULE (20 RACKS)

The internal dimensions of individual Mini IT modules are 800mm x 1200mm x 42U (WxDxH) (APC AR3350 – More technical information can be found - LINK)

Mini IT modules have the following certificates:

- Protection of IT systems from fire and heat according to standard EN1047-2 for 60 minutes
- Protection against unauthorized entry according to standard LPS1175 SR1 to SR5
- Fire resistance for 90 minutes according to EN 1634-1 (for doors)
- 90-minute fire resistance according to EN 1364-1 (for panels)
- Protection against water and dust penetration at the IP66 leve according to the EN60529 standard
- Protection against the penetration of corrosive gases and smoke according to the standard EN1634-3

CMS DELEGATION VISIT TO SERBIA

- July 17, CMS delegation visited Serbia: L. Silvestris (Spokesperson Office), D. Piparo (Offline & Computing Office), V.R. (host, Vinca TL)
  - State Data Center
  - Ministry (FA)
- V.R. presented proposal of technical solution for CMS Tier-1 to DC experts (TCO, Network experts, architecture expert)
  - Capacity, Architecture, CPU, Disk Storage, Tape Storage, Network, Power
  - Initial discussions have already started on details of proposed hardware solution
    - To be continued in working meetings and re-iterate
    - Re-iterate with CMS/CERN experts to converge and identify person-power and eventual tasks