High Level Design of Serbian Scientific Computing Tier-1 (SSC-T1) Center

Purpose, Project Description, Design with Scientific Guidelines, Organization, Roadmap

Vladimir Rekovic for Serbian Tier-1
13. 09. 2024.

Table of contents:

Abstract

The Project of the deployment and the exploitation of the Worldwide LHC Computing Grid in the Republic of Serbia to support high energy physics experiments CMS at CERN assumes establishing of a Tier-1 computing center. The center will be hosted by the State Data Center in Kragujevac, Serbia. In this document the case argument for creation of an advanced scientific computing facility is presented and a design of the SSC-T1 center is described. The future extensions of the center are discussed in scope of a development of broader scientific community as well.

Acknowledgments

The author of this document is grateful for his experience within the Offline and Computing group of the CMS Collaboration, particularly to James Letts, Liz Sexton Kennedy, Tommaso Boccali, Danilo Piparo, and to Lucia Sylvestris former CMS Deputy Spokesperson, and Simone Campana and Alessandro Di Girolamo from CERN for their comments and suggestions in helping make this project possible. Lots of material in this document is a result of collaboration and many helpful discussions we had. Many thanks also go to CERN IT staff for sharing their thoughts regarding present technologies (Marteen, Andreas, Enrico, Vlado, Eric, Luca, Ben, Arne). Many thanks to Danilo Savic the director of State Data Center in Kragujevac for his support and introduction to the SDC-KG facility. The author is exceptionally grateful to the two funding agencies of the Republic of Serbia, the Ministry of Science, Technological Development and Innovation and the Ministry of Information and Telecommunication, for their visionary work and support in recognizing the importance of this project.

1. Introduction

On December 9th 2023, CERN and the Republic of Serbia (Ministry of Information and Telecommunication and Ministry of Science, Technological Development and Innovation) signed a Memorandum of Understanding for the deployment and the exploitation of the Worldwide LHC Computing Grid (WLCG) in the Republic of Serbia to support high energy physics experiments at CERN, in particular to support the CMS experiment. The signing of the MoU was a prelude and expressed Serbia's definitive commitment to the project of establishing a Tier-1 computing center in Serbia, which is to be physically hosted in the State Data Center in Kragujevac (SDC-KG), Serbia. In this document I will refer to the future Serbian WLCG Tier-1 center ias Serbian Scientific Computing Tier-1 (SSC-T1)

There are two distinct phases of the SSC-T1 Project: the first phase (Phase-1) consists of construction and commissioning, and the second phase (Phase-2) which comprises the period of full operations and full functioning of the computing center. The Phase-1 is expected to last 12 months and was originally planned to start at the beginning and be completed by the end of year 2024. In this phase the design of the center is developed, and all the ingredients needed for the full functioning of the center serving the CMS experiment are put in place. These include: the team organization and personnel, installation of hardware, software, and middleware, and the integration and testing tasks completed. The operations phase, the Phase-2, will begin immediately afterwards when SSC-T1 is put online and fully connected to the WLCG infrastructure. This moment marks the beginning of the *probation period*, typically of one year, in which the center actively participates in CMS experiment's organized exercises, managed by CMS Offline & Computing, performs global computing and data transfer tasks, fully involved in collaborative workflows with other WLCG centers, Tier-0 at CERN and CMS's Tier-1 and Tier-2 centers around the globe. After successfully completed tasks and a satisfactory performance evaluation the center is declared a "valid WLCG Tier-1 center" and enters in a full-fledged operations mode.

In this document I describe the design of SSC-T1, the very first step of the Phase-1. The main necessary guidelines are provided for the future construction and development of the center. The design is principally motivated by the scientific goals SSC-T1 is planned to achieve in the field of experimental high energy physics while keeping in consideration the scientific requirements and guidelines of the CMS detector The scientific goals of this project are in line with the country's recently increased active involvement in CERN, a full member state since 2019, with active participation of Serbian scientists in LHC experiments and international collaborations, but also in the context of the country's long-lasting cooperation with CERN starting back in 1954 and the very beginning of CERN of which the Federal Republic of Yugoslavia was one of the twelve founding members.

This paper documents a design proposal of a SSC-T1 computing center to be developed and established in Serbia, in the State Data Center in Kragujevac. The design is motivated by the scientific needs of the CMS experiment as well as the strategic goals of enlarging the scientific ecosystem in Serbia in the field of experimental particle physics and strengthening its collaboration with CERN. Using particle physics as a primer will introduce the scientific/technology field of high-throughput high data volume computing in Serbia.

Section 2 describes the scientific argument, the purpose and the project description. In section 3 the full design of the project is given, covering all subsystems: computing farm, disk-based storage, archival storage, network, and monitoring, a summary of the major hardware.

A structure of the original organization and the type of personnel needed for a SSC-T1 site is discussed in the Section 4. are motivated by and derived from the list of services a standard Tier-1 center is expected to be able to provide to support to offline computing of the CMS experiment at CERN. In section 5, an original road map

is presented which assumed start of the project in January of 2024. Given that the project is at least 6 months behind, the gantt-chart at the end of the document needs to be properly adjusted.

2. Purpose, Scientific Argument, and the Project Description

Serbian researchers have participated in experiments at CERN since the very beginning of this international organization, with Federal Republic of Yugoslavia being one of the twelve founding members ratifying the founding of CERN in 1953. After a few decades spent in an observer status, the Republic of Serbia returned to the status of full membership in 2019. Since, the presence of Serbia at CERN has increased, both through the means of direct contribution but also through the increased number of Serbian researchers participating in the LHC experiments, also growing in number of groups and areas of research. Only in the LHC experiments there are three large active Serbian groups, two participating in the CMS experiment and one in the ATLAS experiment. Historically, the groups have contributed largely in the commissioning and operations, data analysis, but since recently the contribution has expanded to the areas of detector development. However, in the field of computing at CERN the involvement of Serbia has been limited and a more substantial scientific effort had been lacking.

This project of establishing a WLCG Tier-1 center in Serbia is designed with a goal to improve exactly this aspect of the scientific cooperation of Serbia with CERN. Putting in place an advanced and modern computing facility is Serbia to support high energy physics experiments at CERN, a Tier-1 center, will provide mutual benefits for the two parties. Being member of a world grid through the Tier-1, the country will functionally connect to the global scientific framework at the highest level. Designing, constructing, and operating the center is Serbia will induce a substantial knowledge transfer to the country through a collaborative work within WLCG, through sharing experiences, good practices, innovation and joint projects. This requires a mobilization of the local human resources to join the computing international collaboration the WLCG, and in this process stimulate a growth of the country's local expertise in this scientific discipline.

In high energy experiments with LHC at CERN the unprecedented number of particle collisions are recorded by detector instruments and are made available for further scientific studies only through very advanced computational workflows and numerical processing of the data. Given that very large data sets are recorded, their processing often requires complex computing hardware systems. Robust computing is an indispensable component of the successful scientific program of the LHC experiments and probably even more so for the future experiments in high energy physics where collecting even larger data sets are envisioned. Computing is one of the pillars in this scientific field. In case of improper functioning, computing can act as a bottleneck and stall the operation of the experiment, causing irrecuperable time losses, data losses, inflicting large losses of resources and even be detrimental in the experiment's lifecycle. This project of SSC-T1 is of high importance with high international visibility, and it therefore requires careful planning, secured commitment, qualified personnel and development of expertise. Enlarging the scope of scientific activities in Serbia by adding a respected computing facility such as Tier-1 will contribute to the growth of the particle physics community in the country. By complementing the most common fields of research, data analysis and theory, with the field of scientific computing will add to the community's overall strength and versatility. This Project will be starting with a considerable knowledge transfer but will continue through work in an international collaboration of CERN Tier centers, which will both provide for new training and learning opportunities. New research opportunities are also expected to open within the country and internationally. Joint scientific ventures between physicists and computer scientists should start to emerge already within this project and should continue to grow as the center commissioning and operations progress.

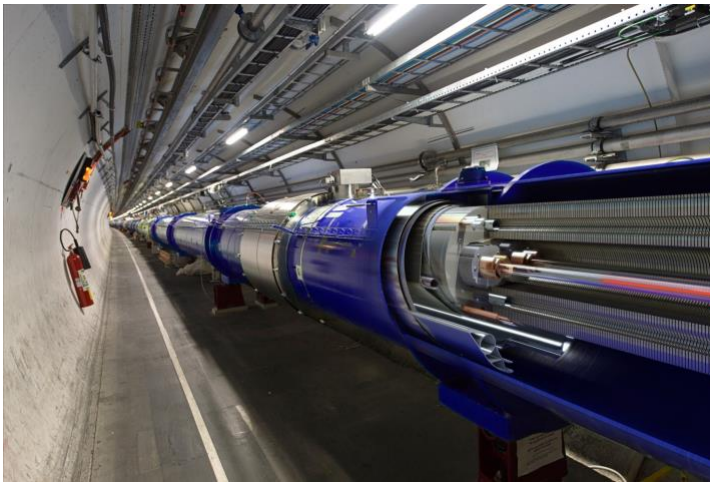

Figure 1 A section of LHC 27 km accelerator at CERN.

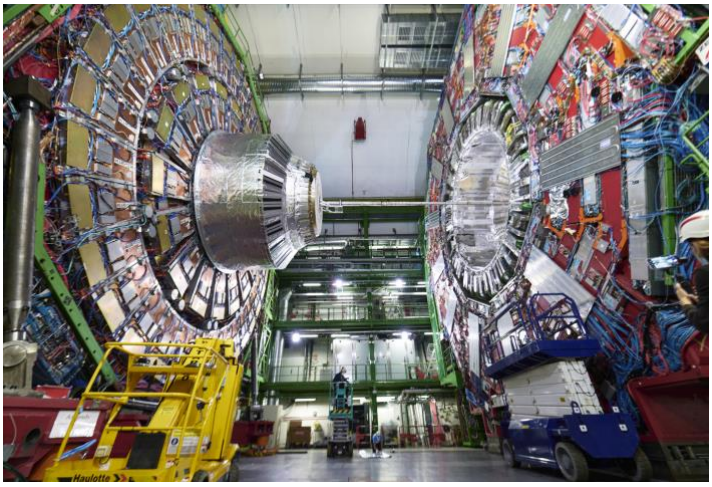

Figure 2 The CMS detector in the experimental cavenr, one of four particle detectors for collisions in LHC accelerator.

The approved scientific program for hadron collider experiments at CERN consists of ongoing LHC accelerator which will be succeeded by a High Luminosity LHC (HL-LHC) era in the period 2030 to 2040 and the upgraded detectors. The projected increase in both quantity and complexity of the future experimental data from the upgraded CMS detector (prototypes are already available) is far beyond the current computing capacities. The CMS experiment has already identified a need for improvement in both resources and technologies. Given that the R&D processes take many years, the Upgrade Project Office for Software and Computing of the CMS collaboration is already now calling for efforts to search for inventive and novel solutions in computing. A symbiotic research and shared expertise in the areas of experimental physics and computer science are highly desirable in this case and are expected to be fruitful in providing novel ideas and might even provide essentially valuable solutions to the problem the CERN experiments are facing. A natural platform for such research activities to occur are functioning computing centers, furnished with adequate hardware which is well connected to realistic scenarios providing relevant data, and with well-established scientific goals.

A factor of ten more data is expected to be collected by CMS during the High Luminosity LHC which is starting in 2030. A properly functioning SSC-T1 would represent Serbia's important contribution to the successful operation of the CERN experiment in the scientifically challenging years to come. The design of SSC-T1 targets to cover for approximately 15% of the CMS experiment's Tier-1 computing needs, data processing and storage, which are expected to double by the start of the HL-LHC. This should provide resources to contribute considerably to the smooth operation of the experiment.

The central aim of the SSC-T1 Project is to build a new computing center that would serve computing needs of an LHC particle physics experiment (CMS) on the LHC accelerator at CERN. The center will process data using advance technologies and scientific applications developed by experimental physics scientific community, serve as a custodial center for storing and sharing of data for the wider CERN experiment community.

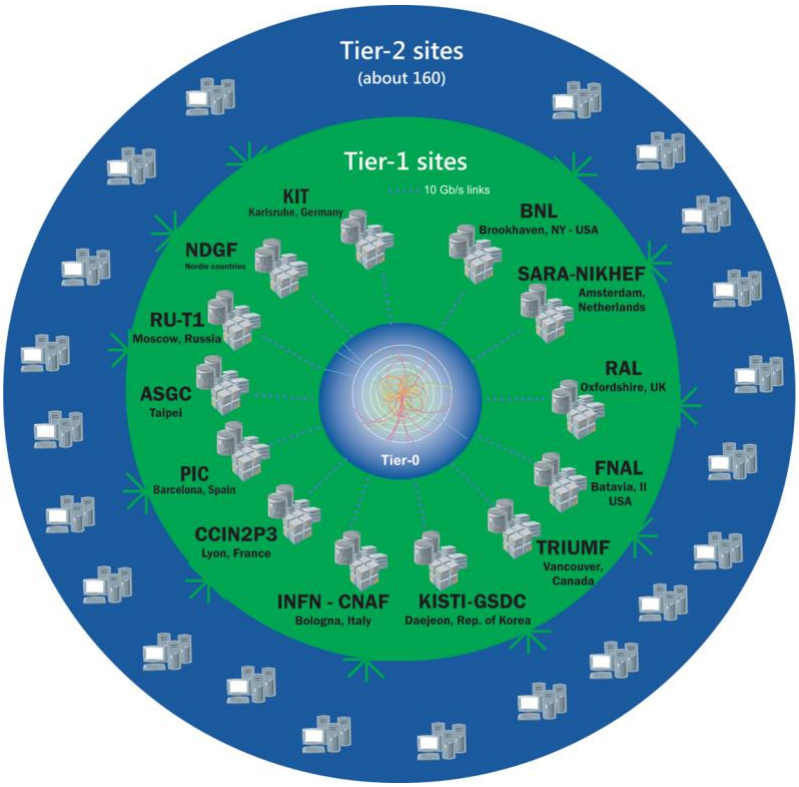

Figure 3 WLCG Tier-0, Tier-1, and Tier-2 sites (as of June 2014)

Such a center is unprecedented in the country and the region of Southeast Europe and therefore does not have a broad community in this field. However, Serbia does have a very vibrant IT community and a long tradition of scientific collaboration with CERN. This project is envisioned to cross that gap and establish a platform for the particle physics and computer science to meet. A special attention of the Project is given to the educational aspect and building of scientific and technological expertise. The knowledge sharing and learning will be achieved through the WLCG collaborative work and through trainings targeting two different groups of the SSC-T1 personnel. The first group, which is made of mid-level experts, will receive a practical training through collaboration within the WLCG, with experts from the existing CMS Tier-1 centers, supplemented by educational visits of personnel to existing functioning Tier-1 centers, and in trainings led by visiting experts and conducted locally in SSC-T1. These training activities are expected to bring the team up to speed in a period of 1 year, when SSC-T1 could start participating in global CMS data exercises. The second approach has a long-term return-on-investment. It targets university faculty and graduate and undergraduate students in related fields from the universities in Serbia who will be assisting Experts in their work and will together participate in the project SSC-T1. Through involvement in the construction, commissioning, and operations of the center the students will obtain invaluable experience and necessary skills for the future work through a necessary hands-on training. The trained students would be a great source of Fellows and CAT-A dev-ops personnel for the CMS and other LHC experiments at CERN.

The central part of the Project is to construct the SSC-T1 which will operate using advanced technology for distributed computing and will serve the computing needs of the CMS experiment at CERN. An equally important aspect of the Project is of a strategic nature, to build a local expertise in the country and develop an ecosystem for the high-end technology large data management and distributed computing. At this moment the technology originally developed for high energy physics is matured enough to be able to successfully support other sciences, if a need arises with not much difficulty adjusted to serve them. Fields like astrophysics, atmospheric science and biophysics are good clients to learn about large scale data management, how to move,

store, and efficiently process large volumes of data. The center will be the first of this type in the country and it is meant to become a research platform attracting local talent from universities and research institutes in Serbia who would be able to take part in R&D for HEP and further contribute to WLCG infrastructure development.

SSC-T1 will provide a platform to enable research in computing for high energy physics scientific field in the following areas

- Possible contributions are instrumental and might turn out to be essential to face the challenge of managing, processing and serving large amounts of experimental data of particle collisions in current and future experiments at CERN.
- Possible topics are 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.

Being plugged into CMS international collaboration, and in partnerships with other Tier-1's, CERN as well as other sites worldwide and industry will open the following 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.

3. Design

Almost all of twenty some WLCG Tier-1 centers have evolved from a smaller size computing centers that have organically grown to larger and more complex Tier-1 centers. Although the experience gained in the process was invaluable as the centers evolved, this scenario was not without challenges. While undergoing multiple upgrades spanning a long period of time, in some cases over twenty years, the computing centers were usually forced to continue using large fraction of existing hardware and had to face complications while integrating hardware of very different generations. One of the main requisites of Tier-1 center is the commitment of uninterrupted availability, which in practical terms means an operational center with data "online" available to the experiments they were serving. Sometimes, the funding limitations forced the centers to evolve in a direction with minimal costs. As an outcome, some Tier-1 centers evolved into systems that are non-trivial to manage and require a lot of in-house expertise. A system that is starting from scratch and developing into a Tier-1, as it is the case with the Serbian Tier-1, is an ideal position to choose an optimal development path. As there are no compatibility restrictions coming from any existing hardware (modulo physical size and power consumption limitations) the new center has a privilege to be able to select optimal hardware architecture while maximizing its potential scientific impact.

This is an opportunity for putting a creative design in place that addresses multiple goals of the project at hand while taking full advantage of the most modern available technology. The central goal of the project is serving as Tier-1 computing center for the CMS experiment and learning about the experiences of the functioning Tier-1 centers and knowing their recommended practices is an invaluable asset in the design process. While answering to the current needs of the experiment and pledges committed to WLCG, the design must bear in mind expectations of future expansions of capacities of about twenty percent annually. This aspect must be taken in consideration as it can affect provisioning of physical space and planning for data traffic capacity.

Furthermore, the data center SDC-KG is a modern and already functioning facility, home of various state and commercially owned data-serving centers. In September of 2023 it was awarded a certificate of highest European standard EN-50600 for the state for the functioning facility and the infrastructure. This greatly simplifies the construction of SSC-T1 in many major aspects, mechanical, electrical, and telecommunications engineering, which would otherwise be both labor, financially, and time intensive.

3.1. Data Processing Functionalities of the CMS Experiment

Processing of the LHC experiments data is tailored to the nature of the data that the four detectors collect and their data acquisition models. To better understand this let us discuss the basics of the compute model of the CMS experiment..

A collision in the LHC accelerator, which is colloquially called "event" occurs every 25 nanoseconds. Although the signal from every collision event is sensed by the electronics of the CMS detector, not every event is saved offline. As a matter of fact most of the events are dropped. Through the real-time (online) data selection system, called trigger, only a subset of events which the trigger labels as interesting are chosen, and their data are recorded on the permanent storage for further analysis (offline). The data from *non-triggered* events are considered as non-interesting and are dropped and lost forever. The frequency of triggered and therefore saved LHC events was typically on the order of 1 kHz to a few kHz lately, it will go up to 10-20 kHz in the HL-LHC era. The data size of a typical CMS LHC selected event is around 2 MB, while due to increased number of interactions and an upgraded detector with increased granularity the data size of a HL-LHC event will be 7.4 MB. This translates to about 10 PB of offline storage capacity needed at Tier-0 in one day of uninterrupted operations.

The data collected directly from the detector, called "RAW" data contain the maximal available information of the signals that the detector registered in each of its smallest units (pixels, strips, crystals, chambers) when a collision occurred. This is the detector-level information. But the RAW data also contain the information needed to retrieve the full and detailed description of the status of every subcomponent of the detector at the time of the recorded collisions. These data are quickly and promptly processed at the CERN computing center Tier-0 and a rough description of particles produced in a collision together with all particles' properties are provided for each event. In this processing step called "prompt reconstruction" the new information about the particles involved in the event are obtained using reconstruction algorithms in the custom software developed by the CMS collaboration, called the CMSSW. The prompt reconstruction has been reserved for the Tier-0 facility, but more recently, in 2023 and 2024 the CMS experiment has been leveraging the close-by Tier-1 centers resources for heavy-ion reconstruction, and it might become a wider practice in the future.

The new data called "reconstructed" data are saved for each event in the format called "RECO". The reconstruction step is generally very compute and memory intensive process, and depending on the complexity of the event, that is to say the number and type of particles produced in the collision event, and a very large granularity of the detector information, reconstructing a single event can take from a few seconds to minutes even on most modern processor. This is extremely demanding and large compared to nanoseconds that it takes for the next collision to occur in the center of the detector.

The full and more "precise" reconstruction of the collision events needs additional time for the important and most updated detector alignment and calibration information to arrive. For all realistic purposes it is mandatory that these subsequent re-processing steps are offloaded to the dedicated computing centers, like Tier-1, where the data are stored and processed in an optimal way without a danger of presenting a bottleneck for the ongoing operations of the data taking experiment.

A Tier-1 center performs many data processing steps of experimental data (there are many steps and the reconstruction is only one of them) using the custom software appropriate for the detector it serves, CMSSW in case of the CMS detector.

A typical Tier-1 center has several main functions while serving high energy physics experiment. These functions are almost always continuously performed in parallel and are expected to be available without the intermission. The **first function** of a WLCG Tier-1 center is to store and re-process the data from the LHC experiments. These data principally contain the information about the real collisions of particles produced in the LHC accelerator, and that were recorded by the real hardware of the particle detector.

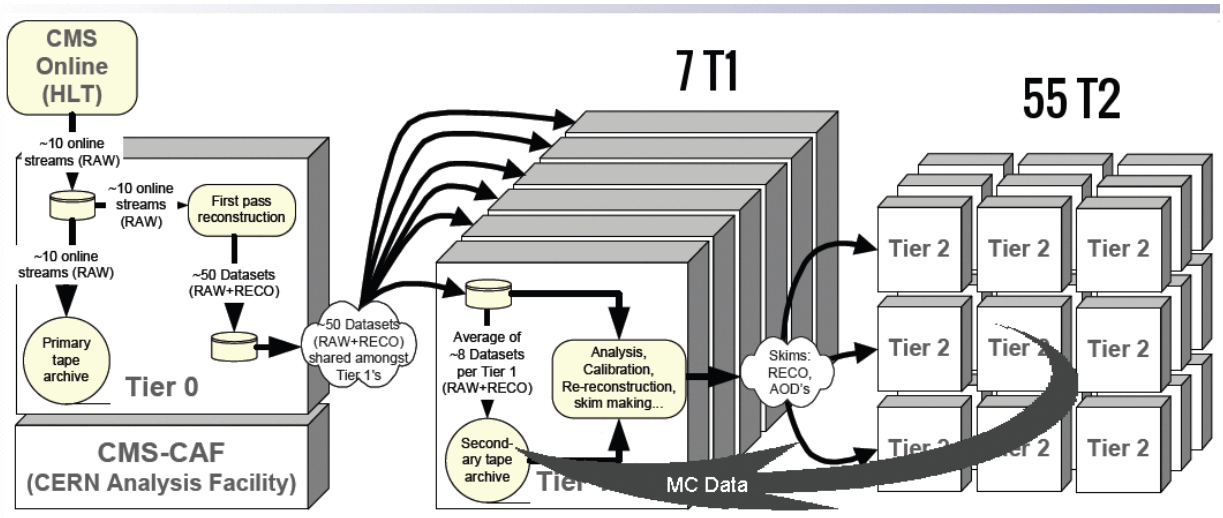

Figure 4 Schematic flow of real and MC event data in the CMS Computing Model. Not all connections are shown - for example peer-to-peer connections between Tier-1's. Credits CMS experiment.

The detectors are custom-made, very complex 3-dimensional cameras (see Figure 2) containing heterogeneous components, but also very fast (capable of taking tens of millions of snapshots per second) specialized in recording signals of passage of highly energetic elementary particles, typical for products of collisions occurring inside the LHC accelerator. Equally important in the scientific analyses, sometimes a factor of two lager in quantity are the data of computer-simulated collisions of accelerated particles, of the collision products, and their interactions with virtual representations of the real detectors. The numerical data obtained in these Monte Carlo simulations give important information to the scientists by describing the hypothetical physical processes which could occur or could have occurred in a real experiment. Large data sets of both the recorded "real" data and the simulated "Monte Carlo" data require intensive processing and robust computing systems, so the **second functionality** of the Tier-1 is to produce the Monte Carlo. This very critical part of the experiment's activity is almost exclusively performed in Tier-1 centers.

The **third function** is to make the processed data available and share to the other, numerous, smaller computing centers, Tier-2s, for further and often less intensive processing for scientific analyses.

The **fourth function** of the Tier-1 centers is custodial in nature, to keep and archive all the previously recorded and processed data throughout the experiment's life cycle.

3.2. Requirements and Input Parameters

Withing the SDC-KG, a dedicated space is allocated to the SSC-T1, a room with surface dimensions of 12 x 6 m, and 20 computing racks. The room is already provided from SDC-KG with all the infrastructure services including electrical power, cooling (air), telecommunications dark fiber, surveillance systems and monitoring. The 20 racks of 42U to maximally 50U are installed and connected to the services, ready to be filled with hardware. The current physical limitation for hardware space is given by the physical volume of the room and the maximal number of racks that can fit. However, in the very near future an additional space in the data center can become available in the additional modules of the data center which are being constructed and are expected to be completed in 2025.

An important input in developing design of the system are the pledges for the expected capacity of the future SSC-T1 and which are stipulated in the signed MoU, which refer to computing and data storage capacities. These values were derived as the average for the six Tier-1 centers located around the world, currently comprising the CMS Tier-1 computing network. The targeted computing capacity of the center is 170 kHS23, which if translated in CPU power corresponds to approximately 24 k cores. The expected data archiving on tape capacity is 30 PB, expected to double by the start of High Luminosity LHC era, starting in year 2030.

The third consideration which must go in the architecture design is the need for flexibility and extensibility of the center to serve other scientific disciplines in the future. One example is BioPark, a future research center in the Republic of Serbia, envisioned as the regional center in the field of bio-science. The extension of the SSC-T1 system to be connected to BioPark's data storage system is shown in Figure 1.

The fourth and a very important component of SSC-T1 is its being a platform for research. To attract the interest of the local talent in the fields of scientific computing and computer engineering, the center must be designed with care so that it can be readily available for scientific research without interfering or disturbing its nominal operations of serving the needs of the CMS experiment. Research and development of novel solutions for technologies of distributed computing, high performance computing, data authentication, security, are some of the topics for local and international collaborative scientists that are attractive and for which the designed system mush have capacities.

The fifth element in the design of the SSC-T1 is the adaptability to the heterogeneous computing. The projected growth of the computing needs of the experiments in the HL-LHC area and the recent results of R&D in the existing Tier-1 centers point towards the need to evolve computing data centers from homogenous (computing only on CPUs) to heterogeneous (mix of CPUs, GPUs, ARMs processors, etc) to achieve improved power to cycle ratio in processing high energy physics data.

3.3. Architecture Topology

Although the interdependence and interconnection among the parts of the system exists, the design and the corresponding needed hardware needed for the SSC-T1 center can be divided in four areas. These four areas are computing, warm storage, cold storage, and network, which are discussed in detail in the following slides. These systems are pictorially represented in Figure 5. For the baseline (Stage 1) configuration with the capacities of the center as listed in the MoU.

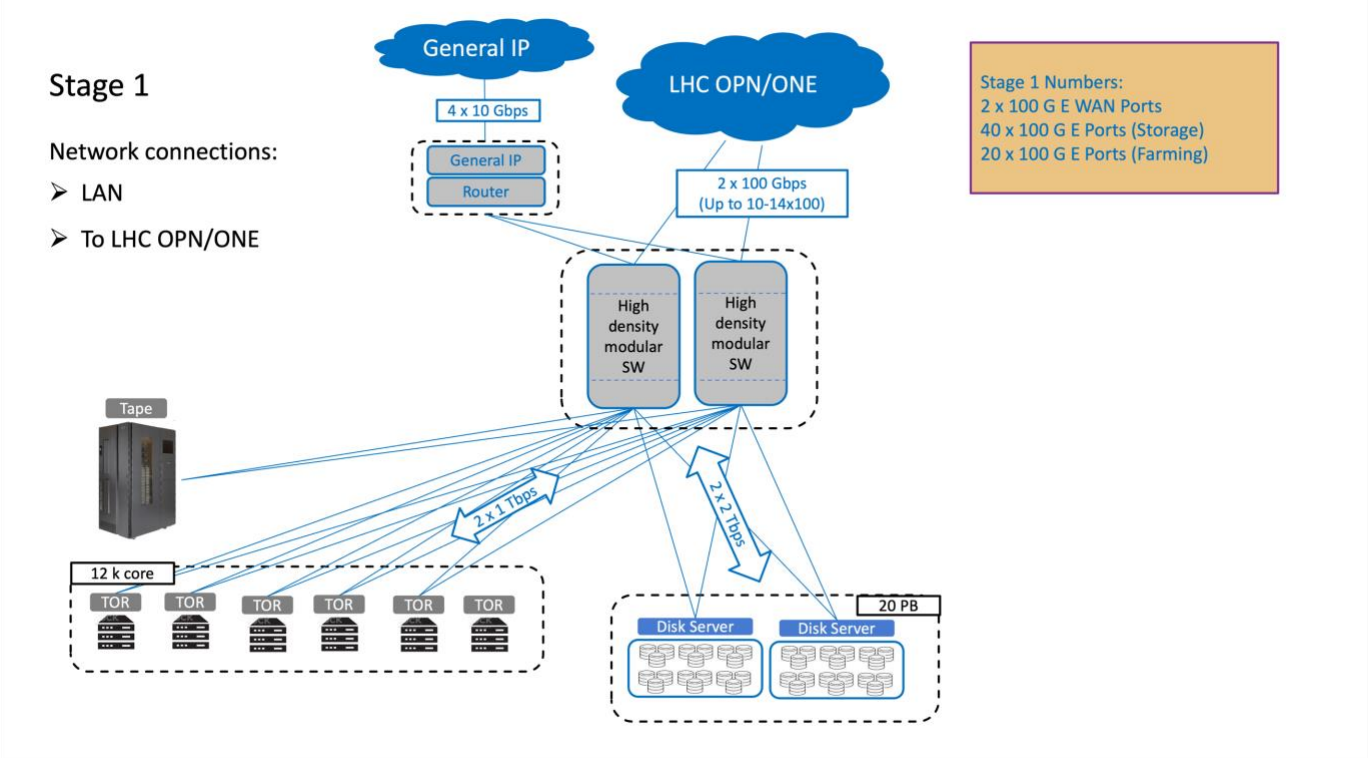

Figure 5 Stage1 system architecture of the Serbian Scientific Computing Tier-1 for CMS Experiment at CERN, at State Data Center in Kragujevac. Computing is performed in CPU nodes with total of 12 k cores installed in 6 racks with TOR switches each connected to the system central high density modular network switch, with total of 2x1 Tbps throughput capacity. Worm storage is composed of HDD disks with total of 20 PB installed in 4 racks with servers, connected to the network switch with 2x2 Tbps throughput capacity. System for achieving data is composed of a tape library of 30 PB capacity served by a robot, with 10 Gbps network connection capacity

The ultimate extended configuration of the Tier-1 centers (Stage-2) where the capacities are virtually doubled is shown in Figure 6. The computing capacity is increased from 12 to 24 cores occupying 12 racks and disk-based storage increased even up to 40 PB still fitting in 4 racks. The core switches being modular and expandible in capacity would be populated but additional network cards. These would handle additional LHCOPN traffic needed during HL-LHC era to cover the increased Tier-0 output rate and for the increased internal data traffic between the disk-based storage and the compute farm. The tape library capacity will also be expanded by installing additional frames to total of 5, and by upgrading to the High-Density license. By the beginning of HL-LHC it is expected that the LTO tape cartridge technology will advance to at least double or even quadruple amount of data storable per cartridge. In that case it might not even be needed to add additional frames to the library but instead adopt new technology in terms of tape cartridges and tape drives.

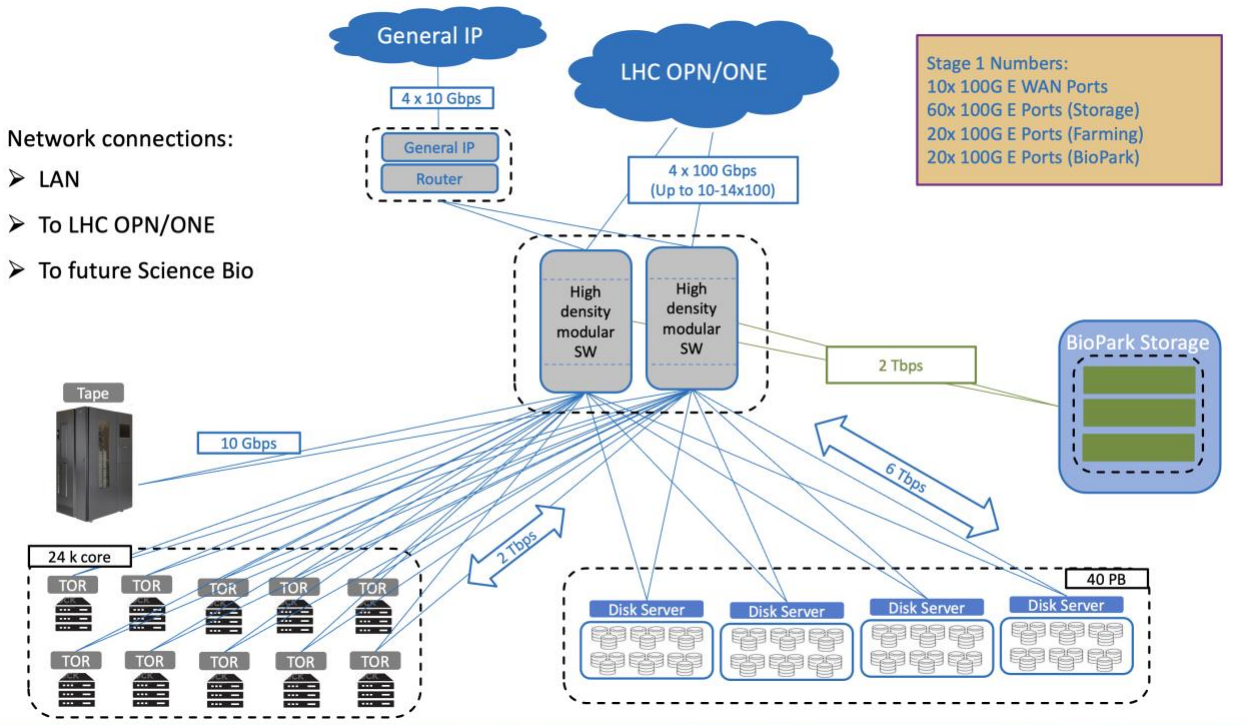

Figure 6 Stage2 system architecture of the Serbian Scientific Computing Tier-1 for CMS Experiment at CERN, at State Data Center in Kragujevac. Computing is performed in CPU nodes with total of 24k cores installed in 10 racks with TOR switches each connected to the system central high density modular network switch, with total of 2 Tbps throughput capacity. Worm storage is composed of HDD disks with total of 40 PB installed in 4 racks with servers, connected to the network switch with 6 Tbps throughput capacity. System for achieving data is composed of a tape library of 20 PB capacity served by a robot, with 10 Gbps network connection capacity.

3.4. Computing Farm System

As it is the case with all four LHC experiments, the needs of the CMS experiment are evaluated for each data-taking year (starting in April typically) and subsequently approved within relevant CERN Scientific Committees. In the case of computing the needs of the experiments are approved by the Computing Resource Scrutiny Group (C-RSG) withing the Resources Review Boards (RRB) In the latest published report of the C-RSG (link), in the year 2023 about 4.1 MHS06-years of CPU utilization 24-7 were executed for all four LHC experiments together, out of which 30% (about 1,176 kHS06-years) went on the CMS. In terms of pledged CPU capacities for all six Tier-1s together serving CMS and pledging resources, there is in increase in total compute capacity pledge from 930 kHS06 in 2023 to 1,030 kHS06 in 2024. The SSC-T1 contribution targets an average of the six CMS Tier-1s, and pledges 170 kHS06 for 2024, as specified in the signed MoU.

The existing T1 centers have grown their capacities over the lifetime of the ongoing experiments they support, as the operating experiments continue to collect data resulting in a continuous growth of recorded data volumes, the offered services, the technologies used, and the capacities of the T1 centers evolved as well. This occurred either to match the new needs of data processing or to profit from new technologies and benefit from novel solutions made available either commercially or developed within the WLCG collaboration. Historically, computing in the existing WLCG Tier-1 centers is most predominantly done on the farms of CPU servers. However more recently, in the last few years, other technologies commercially available on the market are being integrated in the centers as well, examples being GPU and ARM processors. The incorporation and the exploitation of GPUs, FPGAs, ARM processing technologies for running CMS software based applications is not a “plug-and-play” step, but a considerable amount of expert level effort is required both in system architecture and software engineering.

The main task of the Tier-1 computing farm is to process CMS experiment’s data by executing CMS software in jobs scheduled by the CMS workflow management system. Translation of computing requests to a managed complete workflow of executable computing jobs (parallel or sequential) optimized to run on the local computing hardware is performed by middleware technologies suitable for high-compute needs of HEP experiments. In addition to providing job queueing mechanism, this middleware provides scheduling policy, priority mechanism and monitoring, all critical for successful high-throughput computing. The execution of the jobs is performed on the computing hardware, a farm of servers hosting commercially available processor units.

Middleware - HTCondor

Almost all WLCG Tier centers use one technology solution, the HTCondor, developed by the University of Winsconsin. HTCondor is an open-source high-throughput computing software batch framework, which is well-exercised and broadly used in WLCG Tier-1, Tier-2 and Tier-3 centers. Apart for LHC experiments HTCondor is used by many organizations and scientific experiments. The documentation is readily available, and the user community is large.

HTCondor Compute Entrypoint (CE) serves as the door that forwards resource allocation requests (RAR) onto the local compute resources. It provides authentication and authorization of remote clients and interacts with the batch-system layer. A CE host is made up of a thin layer of CE software installed on top of the software that submits to and manages RARs in your local batch system.

Hardware

A typical Tier-1 computing farm consists of a set of mutually connected processor servers, nominally placed in computing racks, and connected to disk storage system through a local network. A typical configuration of a

computing rack is shown on the Figure 1. A computing rack is populated by server units (multiples) and network switch units (one or multiple)

Switches. The recommended configuration of a rack is with two dedicated switches performing the tasks of data routing and of management control, separately. (a) TOR switch routes data to servers for processing (data network) via optical fiber links, and (b) dedicated Baseboard Management Controller (BMC) switch connecting to servers for routing monitoring system health data, managing power, updating firmware, and providing remote access for troubleshooting the servers (management network). Data throughput from TOR to servers is much larger compared to that of from BMC, reflected in very different minimal capacity of the links to CPU servers of 25 G (optics) and 1G (ethernet) respectively. BMC can be a non-expensive switch, unlike the TOR switch needs to be able to handle large throughputs.

CPU Servers. Computing for hadron collider experiments is typically CPU intensive and is not IO limited. As the data processing tasks are numerous in types, the CMS requirements on computing hardware are driven by the most resource-hangry type of processing jobs to be executed on the processing units and by the size and format of data being processed. The scientific data collected in the CMS experiment are organized in the units of "events", where an "event" represents the full information collected in a single LHC beams crossing (bunch crossing) where colliding protons interact, and in which collision products are recorded and readout by the CMS detector, in real collisions, or a MC simulated proton collisions, their products, and the interactions of the products with a simulated CMS detector and emulated read-out, in case of MC data. En event comprises the information from 16 million channels in the CMS subdetector system all combined by the event builder into a "raw data" format. The CMS raw event size varies depending on the type and number of proton-proton interactions in a LHC bunch-crossing and its nominal size is about 2 MBs in case of 60 interactions. For the Phase-2 upgraded CMS detector and HL-LHC upgraded accelerator the raw event size is expected to be 8 – 10 MB. Depending on a type of a processing job, a typical CPU takes anywhere between 100 ms to a few seconds in case of busy events. Especially CPU intensive jobs are Monte Carlo simulation of interactions of particles with the detector material performed through GEANT software.

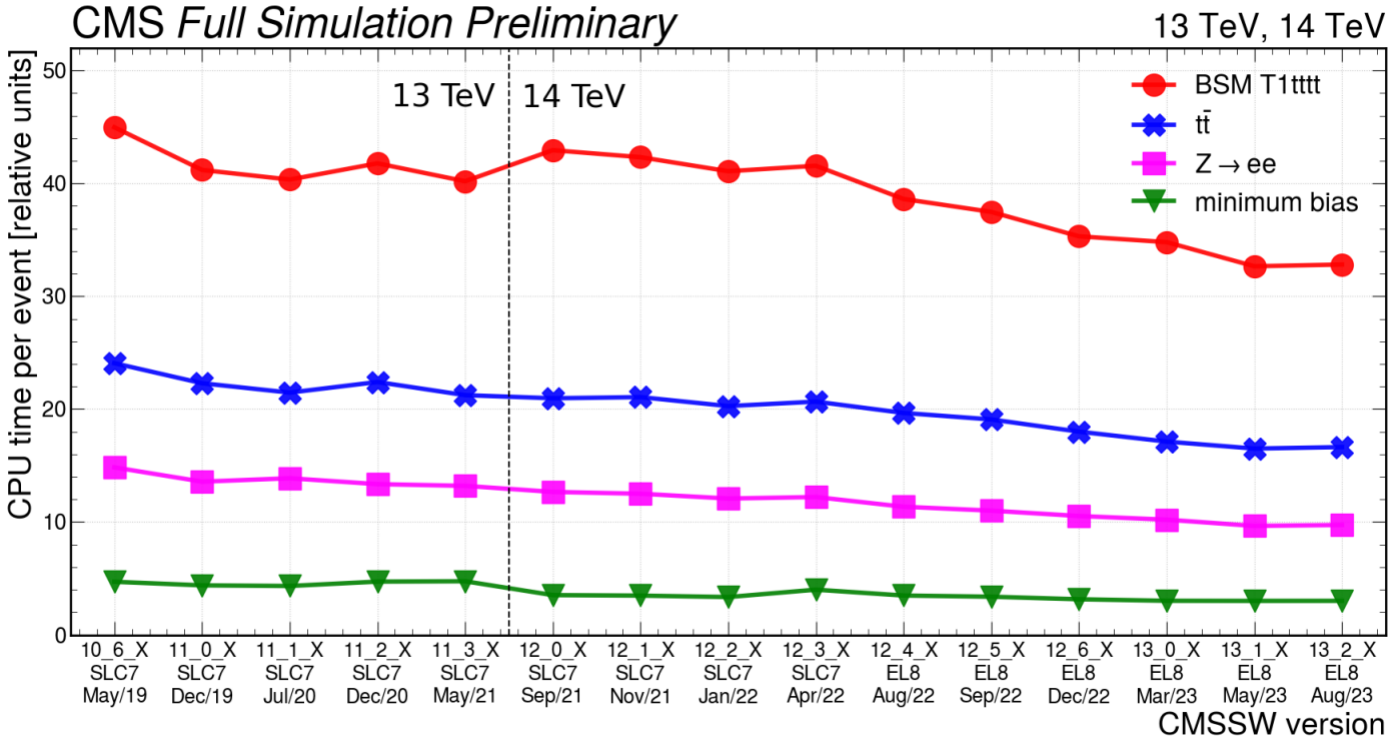

Figure 7 Historical trends of Full Simulation CPU time performance of Run-2/13 TeV CMSSW (10_6_X to 11_3_X) and Run-3/14 TeV CMSSW (12_0_X to 13_2_X) data of minimum bias, ttbar, BSM T1tttt (pp -> gluino + gluino, gluino -> ttbar + lightest neutralino) and Z -> ee processes. The average CPU run time per event in relative units of the event simulation is shown for 500 events on single threaded jobs. Main improvements are connected with the Geant4 migration from 10.4 to 10.7 (CMSSW 11_3_X) and to 11.1.1 (CMSSW 13_1_X), the change of the computing platform operating system from CentOS 7 (SLC7) to Alma Linux 8 (EL8) (CMSSW 12_4_X) and the usage of LTO (Link time optimization) build method (CMSSW 13_0_X). During > 4 years between the versions 10_6_X and 13_2_X the CPU time has improved for the processes BSM T1tttt by 27 %, ttbar by 32 %, Z -> ee by 32 % and minimum bias by 36 %. Credit for figure and caption CMS experiment twiki pages.

To account for event processing time and maximum event size, the fact that CMS data processing is multithreaded , the CMS recommends a nominal amount of RAM for CPU servers in Tier-1. For multi-core CPUs, with typically double-threaded cores, 2 GB RAM per core is required, but more memory (3 GB) is very beneficial and useful for smooth operations.

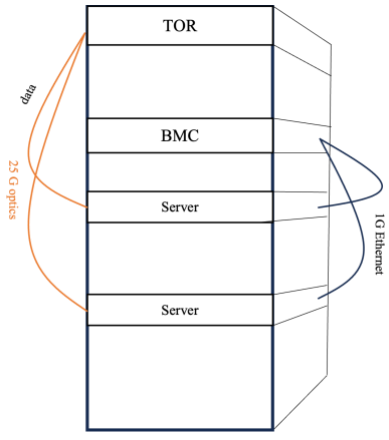

Figure 8 Computing rack with TOR switch, Basebord Management Controller (BMC) and CPU servers

Processor hardware for computationally intensive tasks
Existing T1 centers are diverse in terms of vendors of compute servers and the complementing hardware configurations deployed in the computing farms, but a vast majority of processors are CPUs with x86 architecture. WLCG evaluates computing performance of all the existing servers' configurations by monitoring an execution of standard CMS computational jobs processing experimental data (available locally or remotely) and benchmarks each of the configurations.
At the moment of writing this document, servers with latest AMD CPUs have been shown to have excellent benchmark performance and currently the best value for money.

For the latest AMD CPU processors are available with cores manufactured in 7 nm or 5 nm MOSFET technologies, ZEN3 and ZEN4 architectures respectively. For CPU-intensive tasks the architecture of choice falls on the latter. Recently available Genoa AMD processors are available with up to 96 cores, but the most suitable solution in terms of value-for-money and the optimal cooling power costs, are the 64-core CPUs. Therefore the suggested CPU server configuration consists of AMD 4th generation EPYC processor platform "Genoa" with 64-cores (128 threads) and 12 channels for DDR SDRAM memory.

The CMS Offline and Computing Group optimal recommendation for computing farm servers suggests minimal RAM memory of 3 GB / core and 30 GB of scratch disk space. The memory requirement for AMD EPYC Genoa 64-core processor is met with a configuration of 12 DIMM memory sticks of 32 GB of DDR5 memory each.

Survey of the market is needed to obtain the best value-for-money, but server configuration with dual socket (two CPUs) are not motivated by any needs of the CMS Offline Computing and probably do not make much sense. A wild estimate of a single-socket CPU server configuration is about 15,000 CHF / unit for orders of 100 units.

Note: When procuring the CPU servers, ex. AMD EPYC Genoa 64 core it is highly desirable to obtain a full Baseborad Management (BM) license as well.

3.5. Disk Based Storage System

The targeted capacity of SSC-T1 of disk-based storage is 15 PB as listed in the MoU pledge for 2024, which translates to 18 PB for 2025, assuming 20% increase per year. For the sake of planning for this High Level Design and we consider a system of 20 PB capacity.

Disk storage system is made of disk servers, hard discs and network. We discuss each of the components.

Disk servers can be similar in configuration the CPU servers used in compute farms, except that they can host lower-end CPUs, 32-core instead of 64-core CPUs.

Hard disks are enterprise level HDD discs of Continuous Magnetic Recording (CMR) type, of typical capacity of 16 TB to 20 TB (recommended 18 TB). Note, Shingled Magnetic Recording disks (CMR) cannot be used. Currently tray enclosures with space for 60 or even up to 106 discs fitting in 4U rack are available on the market. With of 18 TB HDD disks such 4U trays could provide Disk storage capacity of 1TB to 1.9 TB.

The number and architectural organization of disk servers is very much a function of file system chosen as the number of servers can vary, We discuss this in the following subsection.

3.5.1.Technology Solutions for Disk Based Storage

The choice of Disk storage solution plays a role while designing the architecture of the disk storage system. There are two flavors of storage platforms which we consider here, EOS and CEPH, or more precisely EOS-on-CEPH. If a plan is to have OpenStack, Containers, Cloud in the Tier-1 center than having CEPH is a solution. Another option is a solution with dCache, which needs to be evaluated.

EOS

The CERN Disk Storage System EOS is a free, standalone disk-only storage with in-memory namespace, few millisecond read/write open latency, which runs well on low cost discs in JBOD configuration. Being developed at CERN and being used at CERN and across WLCG Tier-1 centers it has a considerable user support community and robust documentation. It is implemented with XRootD framework providing versatile remote access protocol. It offers a number of authentication methods (KRB5, X509, OIDC, shared secret, and JWT and proprietary token authorization) used in WLCG.
It is a solution that rans on a “bare metal” with servers running on Linux

- Management node (MGM)
- Database QuarkDB (QDB) saves the configuration (namespace metadata). QDB runs on Flash or SSD running 3 copies, i.e. on 3 nodes with decently large amount of memory.
- Disk server - run FST deamons, one FST deamon per HDD discs.
- MGM and QDB can, and usually do, run on the same node

CEPH

CEPH is an open-source software-defined storage platform providing a storage service for Objects (S3), Blocks, and Files (CephFS) that can run on low-cost hardware. CephFS client is part of Linux, very performant and

stable. An important feature of CEPH platform is high scalability, meaning that storage capacity can be easily added-on as the storage system grows.

Deamons keep track of of the internal state of the cluster and performs repair of the system. When a hard disk becomes faulty the daemons Instantaneously realize it and take actions.

- Monitor maintains map of cluster state (equivalent to MGMs in eos)
- MGR deamon manager of tasks, keeping track of runtime metric
- OSD deamon runs on disk servers and performs I/O on a discs. One OSD deamon runs per one Disk (like FST in case of eos)
- MDS stores meta-data for CephFS
- Rados Gateway (RGW) deamon is needed in case the storage supports objects.

In the past, OSD ran on Disk servers, and MGR and Mon deamon on virtual machines. It is however possible to run MGR, Mon, and OSD deamons all on Disk servers. In Replica-3 mode means every data has 3 copies, which means that at least 4 racks are needed. In case of failure of say 1 rack, the system network capacity must be configured so that enough traffic is possible to perform the needed repair.

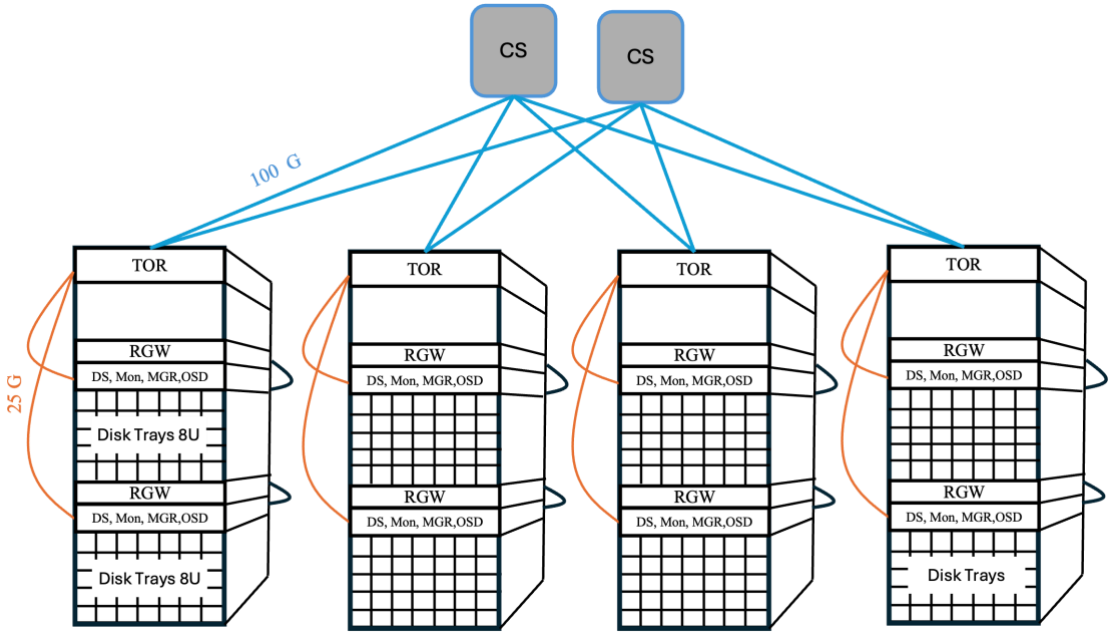

Figure 9 CEPH storage with 4 racks

Crash Map maps the storage representation in CEPH where the data are stored
Crash Rule – a configurable lit of rules which among other things defines a failure domain (e.g. a single rack)

Scientific argument for EOS on CephFS

At CERN, CEPH is used as storage for CERN IT services, and EOS for storage of physics data.
However, there is an option is to install EOS on top of CephFS (reference a presentation at CHEP23). The key motivation that can urge a HEP data storage center to implement such a solution is the possibility to implement a complete virtualization of the storage environment. High availability functionality of EOS can relatively easily be delegated to CEPH, while providing the Access and Authentication functionality needed for a WLCG Tier-1 center. Implementing this symbiotic storage solution might be of high importance for future development

of HEP storage technology, especially in the context of the ongoing research in heterogeneous computing with CPUs and GPUs that HEP community is pursuing. CERN is putting a substantial R&D effort to adjust the main analysis software package ROOT for maximizing performance on modern hardware (including GPUs) and in distributed computing environments. The Serbian Tier-1 center would be an ideal place.

Note, CephFS can also be used on a fraction of EOS storage area and be tuned to a particular use case.

3.6. Archival Storage System

Archival storage is composed of an automated library of magnetic tapes for a long-duration storage of data. It is a low performance but high-capacity system that acts as a secondary storage. It provides reliable permanent data storage and functions as back-up layer to the primary disk storage. The pledged archival capacity of SSC-T1 is 30 PB, but it is expected to double the capacity by the HL-LHC era, or even triple by the end of the data taking period of the Phase-2 upgraded CMS detector. Therefore, a care must be taken to ensure the extensibility of the original archival system design to be able to support CMS experiment needs.

Currently in WLCG Tier-1s two archival technologies implemented and used, IBM and Spectral. Here we discuss a configuration with IBM technology as a use case, which offers various configuration capacities.

Archival storage system is composed of a tape library (frames with magnetic tape cartridges), robot (tape drives), and tape servers (CPU server nodes) which act as "data movers" and are connected to data source via local optical network, as shown in the Figure 10. Number of needed servers serving the data is at least a half the number of tape drives.

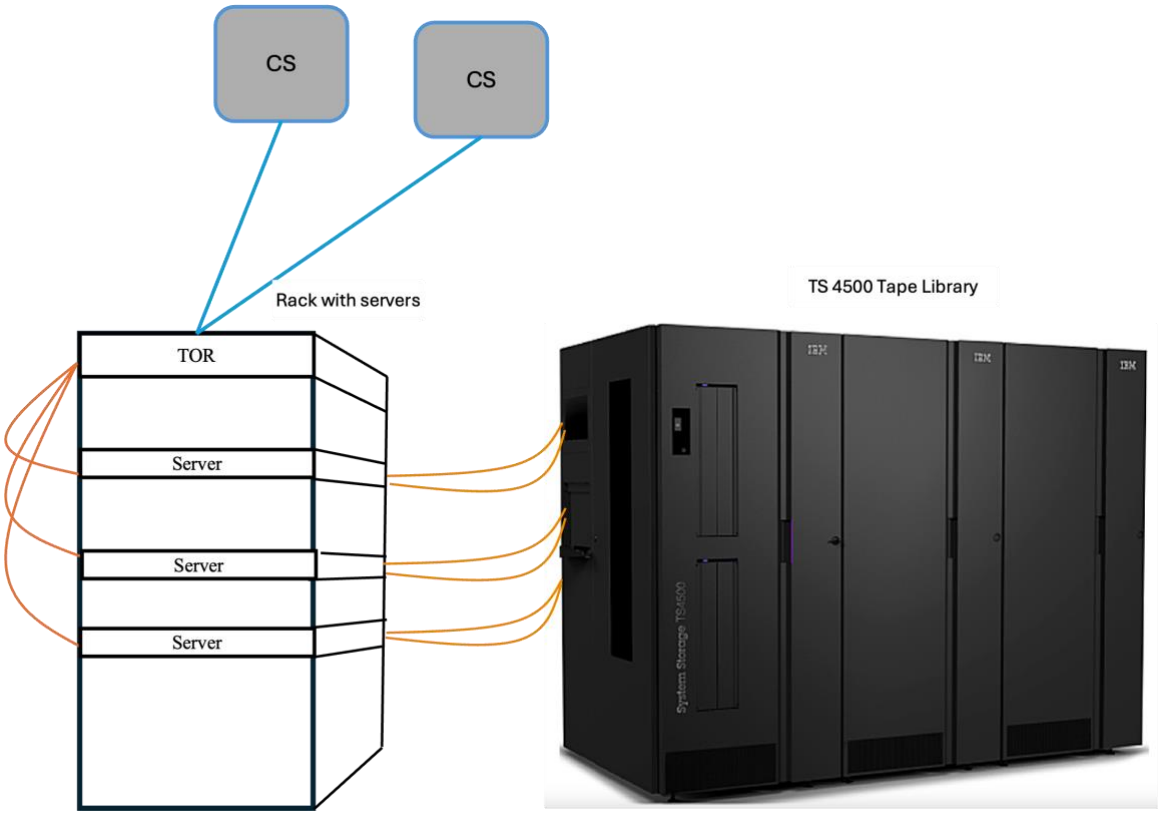

Figure 10 Archival Storage System:On the right, the tape library IBM TS4500 in frame configuration for a library configuration L55-D55-S55. Tape servers are not visible and are inside of the library. On the left a schematic drawing of a rack with data servers units each connected to two tape drives. 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)

Library. Considering a use of typical low-cost LTO9 tape cartridge with capacity of 18 TB, a library of about 5,000 slots fully populated would have at lease 90 PB capacity, which is well above the baseline design requirements but would correspond to the ultimate expanded state of the Tier needed for the HL-LHC era. This can be achieved with IBM TS 4500 tape library constructed of "frames" (computing rack size cabinets) assembled in a flexible configuration that can provide for any targeted capacity and needed scalability. The library would be implemented with flexible growth in mind, starting with a baseline configuration of frames,

L55-S55-D55. Growth is possible from both the right and the left of the central library (L) frame allowing for spatial flexibility. The L-frame hosts the initial tape drives (the robot) and tape slots. Expansions of capacity are performed by physically attaching similar additional drive (D) frame(s) hosting tape drives and tape slots, or by attaching simple storage (S) frame(s) hosting only tape slots, which is doable without disruption or deconfiguration of of the existing library.

Available frames are:
L55 Frame (Library Frame) is a base frame, host up to 12 tape drives and 730 tapes (LTO-9)
D55 Frame (Dive Frame) is an expansion frame, host up to 12 tape drives and 774 tapes
S55 Frame (Storage Frame) is a storage-only frame, host up to 1054 tapes

Tapes.
Currently, LTO Ultrium 9 technology is available for tapes data cartridges that can store of up to 18TB (native) and up to 45TB (2.5:1 compressed) data, and tape drives with transfer speeds of 400 MB/s (native), 1,000 MB/s (compressed 2.5:1) if SAS interfaced.

Tape Drives.
IBM tape drives for LTO technology (eg. ULTRIUM-TD9, ULT3580-TD9) have native data I/O rate of 360-400 MB/s and 750-1000 MB/s for compressed data. For a baseline configuration.

The baseline plan is it to start with a smaller configuration, for example L55-D55 or L55-D55-S55 and 6 drives, with a small density available with "Base" license that features 400, 500, and 660 tape cartridge slots. Considering a library configuration L55-S55-D55 with 6 tape drives, the estimated minimal time needed for writing of 15 or 30 PB of data from servers to archival storage assuming zero server/network latency is about 76 or 151 days for data in native format and 30 or 60 days for compressed data, respectively. Having an ultimate expansion of the archival storage in mind for late HL-LHC era, the IBM TS4500 library with expanded configuration S55-L55-S55-D55-S55 and High-Density license would have host capacity of 4,666 tapes of LTO-9 technology and 24 tape drives. This corresponds to maximal data transfer speed of the robot of 9.4 GB/s and 23.4 GB/s for native and compressed data, respectively.

Although the LTO Program has a roadmap with a planned LTO-14 tape cartridge surpassing 1 PB within the next decade is too far in future, recently LTO vendors have announced the LTO10 technology available by the end of 2024, with double storage capacity and transfer rates of LT0-9. This will be taken in consideration later.

Physical size. Library footprint for L55-D55-S55 and S55-L55-S55-D55-S55 configurations, including service clearance has a minimum with of 3.053 m and 4.562 m.

Power. Total power consumption and cooling requirements for a library configuration L55-D55-S55 is between 2.7 kW and 4.2 kW when the drive has no tape cartridge loaded and when the drive is actively reading and writing to the tape, respectively.

Servers. Tape servers are CPU based worker nodes which are physically located on computer racks and are connected to the tape drivers via data cables. The servers run the archival storage system's middleware, acting as "data-movers", so these nodes need not be high-end CPU servers. Configurations with cheaper CPUs with 2 ports of 16 Gbps, 128 GB RAM, 1 TB SSD suffice. Each of two ports is connected to one library tape drive via dedicated FC or SAS cable, enabling data transfer speeds of at least 8 Gbps.

Middleware. CERN Tape Archive (CTA) is a mature middleware solution developed at CERN used for custodial copy of all physics data at CERN. CTA is implemented as a tape-backend to EOS and is used in the archival storage systems of many WLCG Tier-1's as well.

3.7. Network

Tier-1 centers contain computing farms that execute CPU intensive software applications in very high-throughput workflows. To properly design the network it is necessary to understand the computational and custodial task of a Tier-1 center serving a CMS experiment.

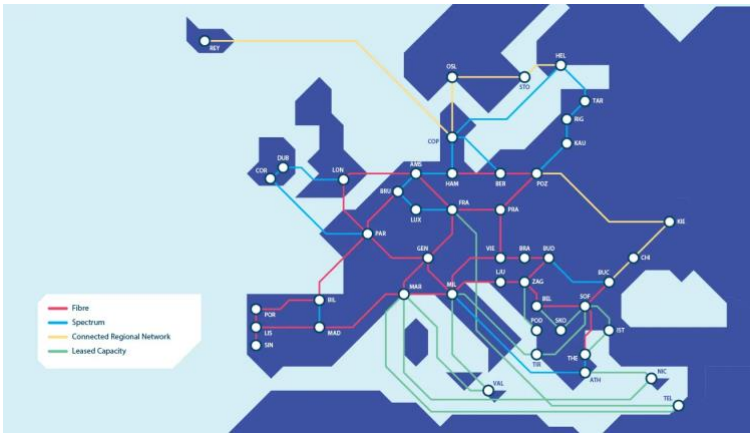

Figure 11 GEANT Network infrastructure as of 2024.

External Network – SSC-T1 connection to WLCG via AMRES
Serbia is well integrated in GEANT and collaboration of European National Research and Education Networks (NRENs). Network infrastructure of Serbia's connectivity to Geneva and rest of Europe is in formidable state.

The capital of Serbia, Belgrade to CERN Teir-0 connectivity is possible through the existing Dark Fiber infrastructure, which is managed by Serbian Academic Network, AMRES and the Ministry of Information and Telecommunications of the Republic of Serbia. The SDC-KG to Belgrade DarkFiber infrastructure is in place as well.

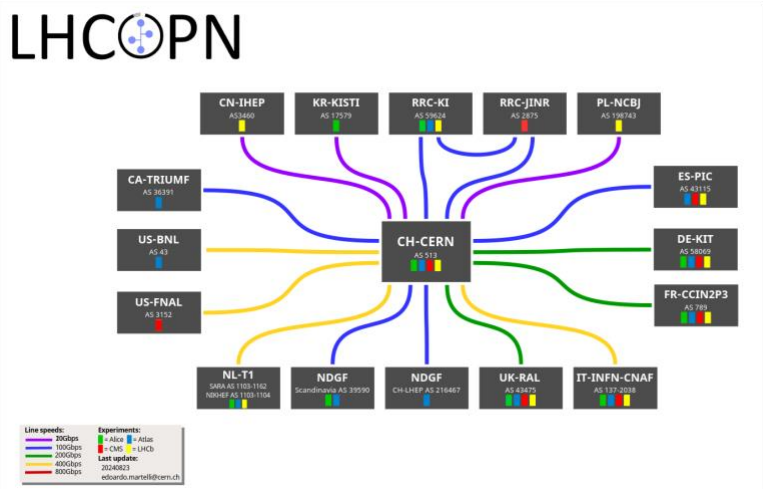

Figure 12 LHCOPN Network showing mutual connections between CERN (Tier-0) and all Tier-1 sites.

Data traffic between WLCG computing centers Tier-0/1/2/3 occurs through LHCOPN and LHCONE networks. LHCOPN is a dedicated private IP network used to connect Tier-0 and all Tier-1 centers. The LHCOPN consists of any T0-T1 or T1-T1 link which is dedicated to the transport of WLCG traffic and whose utilization is restricted to the Tier-0 and the Tier-1s. Full documentation is readily available on a twiki page here. The T0-T1 connectivity varies across the Tier-1s, and ranges from 20 Gbps to 400 Gbps at the moment, but some

with plans to upgrade to 1 Tbps in near future.

Given that the SSC-T1 is envisioned as a mid-size Tier-1 center the envisioned LHCOPN connectivity is 200 Gbps via DarkFiber but can be easily expanded if a need arises, likely to the level of 1-1.4 Tbps. The hardware needed within the SSC-T1 to connect to LHCOPN is a L3-type router, and a choice of technology is open and can be chosen in the wide variate of market available solution (CYSCO, Juniper, Huawei, etc).

On the other hand, a dedicate network LHCONE is differently constructed. It essentially provides a collection of access locations that are effectively entry points into a network that is private to the LHC T1/2/3 sites. LHCONE is not designed and cannot be used as an alternative to LHCOPN but its purpose is to complement it. The entry point for SSC-T1 can be provided by the state telecommunication provider (eg. POP via Amsterdam to CERN).

Internal network within SSC-T1.

As shown in the SSC-T1 architecture schematic drawing in Figure 5, the internal network is managed through core switches (CS). There are two CS to provide redundancy of the system.
The core switches need to be able to manage and provide the internal traffic flow between the three subsystems (compute farm, disk-based storage, and archival storage) and the traffic flow between LHCOPN and LHCONE and the three subsystems. The baseline capacity of a CS must support baseline LHCOPN traffic of 200 Gbps but it must contain needed expandability to support possible future traffic of 1-1.4 Tbps

Traffic to Internet is performed via a dedicated router with a public IP with baseline capacity of 10-40 Gbps. Router technology is highly preferably with IPv6 capability for multitude of advantages over IPv4 technology: better monitoring, better privacy, end-to-end data encryption, no need for NAT, and the list can go on. Further technical details are left to the network specialists to define.

In the rest of this chapter we discuss the network capabilities and needs of each of the three internal subsystems

Compute network
Events with particle collison data are processed, where a single event is processed on a single thread. Taking an average event data size and average time it takes to process a single CMS event, the nominal processed data throughput on a single thread is about 5 MB/s. For a farm of 11.5 thousand cores or 23 thousand threads per farm with processing throughput of 5 MB/s/thread the total compute throughput per farm is about 100 GB/s or about 1 Tb/s. This capacity is used by 180 servers located in 6 racks with 30 CPU servers each, or spread across 12 racks with 15 servers each. In either case the rest of the racks can be populated by less power demanding disk servers

The central switch network capacity for data transfer to CPU farm of must be about 1 Gbps. This number linearly scales with the number of threads, so for a farm twice as big the network capacity is 2 Gbps. Two core switches which are envisioned for redundancy therefore must be able to handle this amount of data traffic.

Disk Based Storage System network
This part of network must allow for data traffic arriving from outside the center, either from Tier-0 or from other Tier-1 sites, and for the traffic within the center, either to/from compute farm or archival storage system.

The traffic throughput from Disk storage to/from compute farm is 2 Tbps to allow for maximal processing throughput of 1 Tbps of data arriving through each of the two core switches. Maximal traffic to/from outside world is about 2 Tbps.

3.8. Monitoring and Accounting of Hardware

SSC-T1 will be located in the SDC-KG facility, as one of the clients using the existing infrastructure. Monitoring of the infrastructure parameters used by the SSC-T1 will be performed by the SDC-KG with the already existing services offered to the other users (government, industry clients).

Monitoring of the SSC-T1 will need a complete dedicated system supported with the 24-7 hyper-care which will ensure the Tier-1's maximum availability of services. A detailed discussion on this topic, in terms of the hardware and implemented services is deferred to later documents to be provided by the technical team. Here we discuss major aspects that are of over-all concern for a Tier-1.

One of the key aspects of a successfully and efficiently operating Tier-1 system is a thorough monitoring of its functions, services and hardware. In terms of the functions, the most critical monitoring is that of the computing tasks and storage, whose performance monitoring is also accessible by the CMS which oversees the deployment of central workflows. All of these services need be monitored locally and properly reported as well. SSC-T1 weekly meetings where reports on batch, EOS, CEPH, tape library are envisioned.

The details of the interfacing and procedures for reporting to WLCG and the CMS experiment are well established and occur at a lesser rate. However, CMS collaboration Computing and Operations Group holds weekly meetings where Tier-1 facility services are discussed, and technical representatives of each Tier-1s are expected to participate.

In terms of hardware accounting a robust local inventory system needs to be put in place, which will cover

- location of hardware
- date of installation
- replacement and warranty due date

An inventory software with GUI (readily available commercially) containing all the necessary information will be installed and clearly written procedures to be followed are to be put in place. The replacement team which can be outsourced, say to the SDC team, would be responsible for replacement of most of the hardware, like disks, motherboards, batteries, PSU, network interface cards, always with an agreement of the relevant subsystem manager. A ticketing system will be in place for the full tracking of actions and responsibilities.

Sufficient replacement stock is envisioned to be available in house, so that a quick replacement is possible. The inventory of the replacement stock is therefore also needed, to keep track of the Return Merchandise Authorization.

Given that the SDC-KG is home to many clients other than SSC-T1 that all share parts of its infrastructure it is important to stay informed about the data center plans long- and short-term for actions that can have an effect on operations, such as preventative maintenance. SDC-KG infrastructure related issues are to be discussed within SSC-T1 and properly communicated to CMS and WLCG collaborations.

3.9. Hardware Summary

The architecture of the storage, compute and network has been developed, and is summarized in Table 2, together with the proposed technology solutions. This is still a subject of study, and the final decisions are expected to be made in the beginning of year 2024, once the Coordination Team and Core teams are established.

Hardware	Capacity		**Configuration Solution**	**Technology Solution**
Disks storage	15 PB		20 PB in HDD discs (2 racks) 200 TB in SDD discs	Huawei 18 TB discs
Archival storage	30 PB		Tape library Tape drives Tape cartridges Tape servers	IBM TS4500 TS1160 tape drive LTO 9 tape technology AMD 32-core
Computing farm	11.5 k CPU cores		180 CPU servers @ 64 cores	AMD EPYC Genoa 64-Core 192 GB DDR5
Network	2 Core Switch @7 Tbps		128 ports of 10/25 G E, SFP 300 ports of 100 G E, QSFP	Huawei Cloud Engine 12800
	6 TOR Switch@200 Gbps		Uplink: 6 x 40/100 GE QSFP28 Downlink: 48 x 10/25 GE SFP28	Huawei Cloud Engine 6863
	Optic Fibres	Disk	2 CS x 50 SU x 40 Gbps	TBD
		Archive	2 CS x 12 drives x 10 Gbps	
		Compute	2 CS x 6 TS x 160 Mbps	

Table 1. Hardware for SSC-T1, required capacity, configuration solution, proposed technology solution. For reference, approximate cost of such hardware in the world or other WLCG Tier-1 centers is given in the last column.

3.10. Power consumption for Stage 2 (dual capacity of Stage 1)

For the over-all design of the SSC-T1 we consider power consumption of the main hardware presented in this document. To ensure that the desired expansion to the capacity needed for the start of HL-LHC are, here we present the accounting for each of the subsystem, assuming the Stage 2 hardware, doubling the capacity of Stage 1, the initial stage in 2024/2025. For the sake of this study, we don't assume any development of the technology or computing performance of the CMSSW software. We also don't assume any hardware improvement of hardware, and we simplistically assume only a doubling of the needed resources. Therefore the result of this accounting study is meant to be conservative. We list the power consumption for each of the subsystem separately.

- 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 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

- 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
  - < 15 racks + 5 rack size library

- Total power w/ cooling
  - 154 kW

4. Team and Organization

The standard functioning state of the center assumes constant scheduling and execution of large number of CPU and IO intensive computing jobs, and transfer of large amounts of data within or from and to the center. Sustainability and resilience require a highly skilled and trained personnel to operate Tier-1 center, but above all, a considerable amount of overlap between personnel's expertise.
A proposal for a Team structure in put in place, intended to work on setting-up the Tier-1 and its services, and will ensure hyper-care of the system once it becomes operational.
The Team is composed o total of 21 members: organized in the Coordination Team and the Core Team, and shown in the Organigram in Figure 2.

- **Scientific and Technical Coordination team** consists of a Scientific Director and Technical Director that report to the Funding Agencies.
- **Advisory Board** consists of CMS Offline & Computing Coordinators, WLCG Project Manager(s), and Project Leaders of other WLCG Tier-1 centers. This board is appointed by the Coordination Team.
- **Core Team** consists of 6 Experts and 12 Junior Specialists, together covering 6 core macro areas which encompass the major functions and responsibilities within in this Tier-1 computing center. These include: **Network**, **Compute**, **Disk Storage**, **Archive Storage**, **Software & Middleware Support**, and **Fabric**. For each macro area there is 1 FTE Expert plus typically 2 FTE Junior Specialists dedicated. To ensure a large overlap of expertise in the Core Team, the work of each Expert has 0.6/0.4 FTE responsibility shared between at least two macro areas. Clear lines of communication and collaboration between these macro areas are critical to ensure the smooth operation of the computing center and effective integration with the larger WLCG infrastructure.

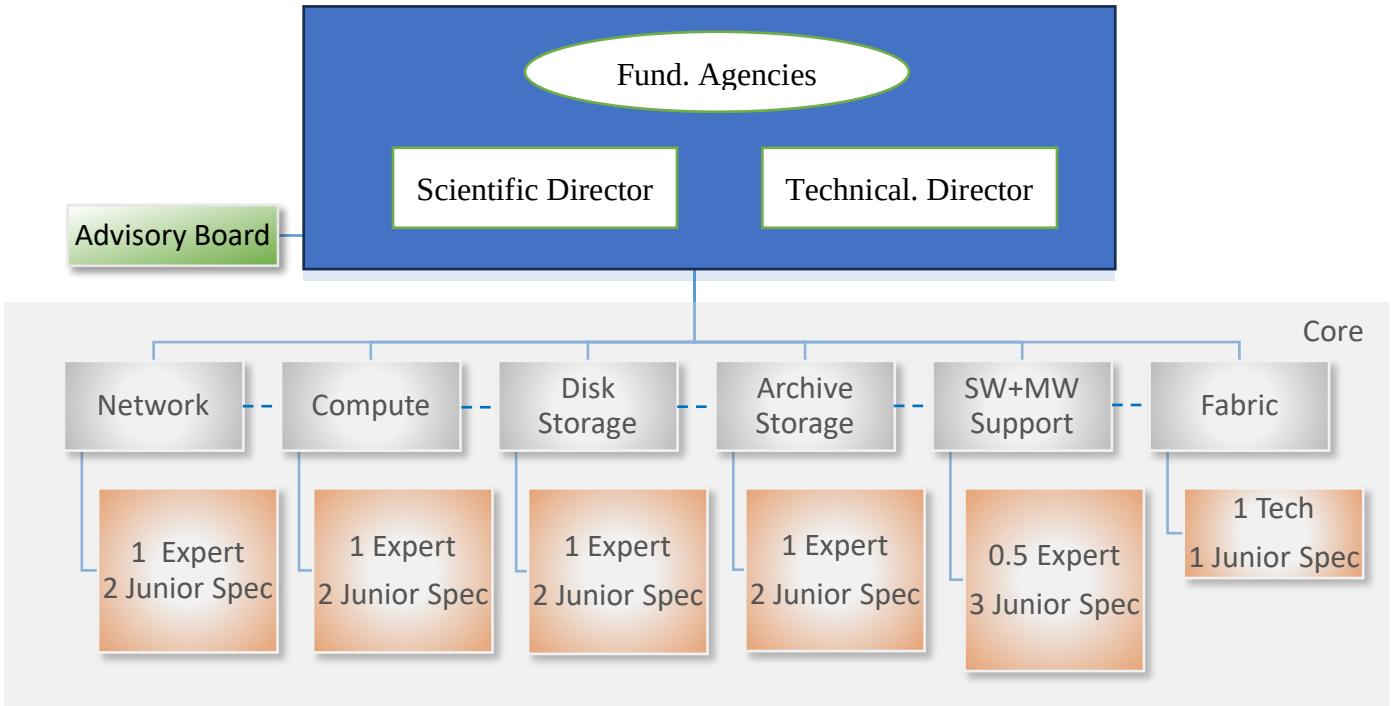

Figure 2. Organigram of Serbian Scientific Computing Tier-1 for CMS Experiment at CERN, at State Data Center in Kragujevac. Team of 21 consisted of senior experts, experts, and junior specialist is organized in Coordination Team and Core Team, the latter covering with overlap the the six macro areas. The team is

complemented with an external Advisory board.

5. Team Personnel Description (slightly adjusted from original proposal)

Members of the **Coordination Team** are Scientific Director and Technical Director:

- **Science Director** Level: Senior, Leadership
  Responsible for a strategic vision for the Tier-1 center, understanding its role within the broader research ecosystem and aligning it with the goals of the high-energy physics community. Develops and oversees the comprehensive project plan for the Tier-1 computing center, including timelines, milestones, and resource allocation.
  Appoints members of the Scientific Team consisting of particle physicists and computer scientists. Appoints members of Advisory Board. Leads and inspires a multidisciplinary team, fostering a collaborative and innovative work environment. With the rest of the Coordination Team participates in the design of the overall architecture of the CMS Tier-1 computing center. Oversees the study of technology survey and solutions.
  Develops budget for the computing center project and presents it to Funding Agencies. Identifies potential risks to the project's success and develops strategies for risk mitigation. Navigates and adapts to changes in project scope, technology, or external factors coming both from the experiment and WLCG, ensuring that the computing center remains agile and responsive to evolving needs.
  Ensures effective communication and collaboration with various stakeholders, including CMS experiment teams, other WLCG centers, and international collaborators. Ensures knowledge transfer to the Team and secures its professional connections with the outside community. Engages with universities and develops opportunities for scientific research. Strong leadership and interpersonal skills are essential for building and maintaining these relationships.
  Reports to the Funding Agencies, Data Center and the WLCG periodically.
- Qualifications: Needs to have a good knowledge of experimental particle physics and extensive experience in team leadership in data taking projects of high energy physics experiment, in particular, offline computing, the CMS experiment at LHC in CERN. Knowledge of CMS data model is required. Ph.D. in experimental high energy physics with at least 10 years of experience in team coordination.

- **Technical Director (TC)** and **Deputy Technical Coordinator (DTC):** Level: Senior/Mid-level.
  Strongly motivated individual. Oversees the integration of various technical components within the computing center, ensuring seamless communication and functionality. Participates in the design of the overall architecture of the computing center, including hardware, software, and network infrastructure. Manages the day-to-day operations of the computing center, including server installation, configuration, and maintenance. Handles the management of disk and tape storage infrastructure, ensuring data integrity, availability, and efficient retrieval. Robust interconnection and hand-shake of the parts of the system is the responsibility of this expert. Manages monitoring solutions to track system performance, resource utilization, and technical services. Ensures the security of the computing center, implementing measures to protect against cyber threats and ensuring compliance with relevant standards. In charge of integrating the computing, storage, and network sections of the center and insures proper functioning of the system as whole. Participates in the weekly meetings of CMS Collaboration within a working group of CMS Offline Computing, giving reports on status of the Serbian Tier-1. Serves as administrator of the Virtual Organization (VO) of the center associated with CMS Experiment, and performs operations of the LHC grid related tasks. The role of Technical Coordination is critical for the effective functioning of the Tier-1 computing center, as it involves aligning technical aspects with the needs of experiments, ensuring seamless interoperability with the larger grid infrastructure, and optimizing resource utilization.

Coordination within this macro area is vital for the success of the Tier-1 center in supporting high-energy physics research.
- **DTC** shares 0.5 FTE with SW & MW tasks, one of the 6 macro areas in the Core Team.
- Qualifications: Ph.D. in particle physics, computer science, electrical engineering (IT) or related fields with at least 5 years of experience in operations and offline computing. Knowledge of the CMS experiment, CMS data-taking and CMS data model is a plus.

Core Team consists of sub-teams covering 6 macro areas of the Project: **Network**, **Compute**, **Disk Storage**, **Archive Storage**, **MW & SW**, and **Fabric**. Each sub-team consists of one Expert and usually two Junior Specialist (student). A full detailed job description of members of the Core Team are not provided here on purpose and are instead left to be designed together with the Coordination Team once it is formed. Below is a strategic vison of the 6 macro areas and the personnel related to them.

- **Network**

    - **Expert** (Level: Mid-Level), is responsible for designing and maintaining the external and internal network of the center. The center is required to have internal network redundancy while maintaining efficient configuration allowing for large data traffic and high number of computing job requests. Responsibility of this expert is to contribute to the overall design of the network between the center and other Tier-1 and Tier-2 centers, in collaboration with national research networks and international network organizations. During operations, transfer of data-volumes in multiple directions are to be overseen. Ability to diagnose problems, find and implement solutions to problems in data flow in a large computing center are expected. Holds relevant industry certifications that can demonstrate expertise and proficiency like:
        - Cisco Certified Network Professional (CCNP) or Cisco Certified Internetwork Expert (CCIE)
        - Juniper Networks Certified Internet Associate (JNCIA) or Juniper Networks Certified Internet Professional (JNCIP)
        - CompTIA Network+
        - Certified Information Systems Security Professional (CISSP) for a broader understanding of network security aspects.

      Education: Masters in computer science IT, electrical engineering or related fields with 5 years of experience in designing, implementing, and managing large-scale network infrastructures is crucial.
    - **Junior Specialists (Computer Science Students):** Assistant, role TBD.
      Education Level: Enrolled in a relevant undergraduate or graduate program.

- **Disk Storage**

    - **Expert** (Level: Mid-Level), is in charge of the filesystem needed to store large amounts of data on disk. Will be implementing and maintaining software solutions for creation, deletion, modification of the files and managing their access, security and the resources used by them. The expert's will be responsible for the installation of a distributed filesystem and the operation of tools ensuring low latency, high availability, strong authentication, and multiple reproduction schemas as well as multiple access protocols and features. This expert is in charge of databases used to store physics data and associated meta-data. Ability to work on novel solutions and designs are expected.
      Qualifications: Masters in computer science, IT, electrical engineering or related fields with 5 years of experience in filesystem management or development.
    - **Junior Specialists (Computer Science Students):** Assistant, role TBD.
      Education Level: Enrolled in a relevant undergraduate or graduate program.

- **Archive Storage**
  - **Expert** (Level: Mid-Level), is responsible for proper functioning of archive library and serving of archived data to the disk storage. Installation and maintenance of the middleware technology solutions adopted in the center are expected, including drivers and support services. Collaborate with experts in other Tier-1 centers on finding new solutions. Good working knowledge of file systems is expected and previous experience in archival storage systems is preferred.
    Qualifications: Masters in computer science, IT, electrical engineering or related fields with 5 years of experience in filesystem management or development.
  - **Junior Specialists (Computer Science Students):** Assistant, role TBD.
    Education: Enrolled in a relevant undergraduate or graduate program.

- **Compute**
  - **Expert** (Level: Mid-Level), is in charge of computing environment of the center. Will be designing and implementing solutions for automated scheduling of computing jobs executed locally on the servers of the center. Maintaining of the job-scheduling system will be a large fraction of responsibility, including identifying anomalous state of the system, diagnosing and solving problems. Work on optimization and improvement of the system, including multi-threaded, parallel processing, and heterogeneous computing. Collaborate with experts in other Tier-1 centers on finding new solutions. Participate in computing campaigns advertised by the CMS experiment.
    Qualifications Masters in computer science, IT, electrical engineering or related fields, with 5 years of experience in system administration.
  - **Junior Specialists (Computer Science Students):** Assistant, role TBD.
    Education: Enrolled in a relevant undergraduate or graduate program.

- **SW & MW**
  - **Expert** (Level: Mid-Level). Manages software and middleware locally residing in the center. CMSSW is an extremely large software package, a collection of software that the CMS experiment uses to acquire, produce, process, monitor and analyze its data, written in C++ but the configuration and manipulation is written in the Python language. Knowledge of CMS core software and CMS data model is an advantage. Installs and manages middleware. Implements software security measures. Possible work on adaptation of code to run on heterogeneous computing platforms.
    Qualifications: Masters in computer science, IT, electrical engineering or related fields, with at least 5 years of experience in system administration and management of large software packages.
  - **Junior Specialists (Computer Science Students):** Assistant, role TBD.
    Education: Enrolled in a relevant undergraduate or graduate program.

- **Fabric**
  - **Expert** (Level: Tech) is responsible for managing the physical infrastructure, including hardware, data center facilities, storage, and networking components. The qualifications for a Fabric Expert would encompass a combination of education, certifications, and practical experience. Here are general qualifications that would be relevant for a Fabric Expert in this context:
  - **Junior Specialists (Computer Science Students):** Assistant, role TBD.
    Education: Enrolled in a relevant undergraduate or graduate program.

6. Original Roadmap for Construction and Commissioning Phase

The original roadmap of the activities for the Construction and Commissioning of SSC-T1 center is shown in Table 3. which needs to be adjusted for the delay introduced for the administrative matters. In this document we don't include a discussion on the Operations phase, which will be added as the Phase-1 progresses.

1. Appoint Coordination Team, 2 mo
2. Define Technology, 2 mo
    a. Study within Coordination Team, 1 mo
    b. Technical Proposal already sent to D.Savic in SDC (Done)
        i. Computing nodes
        ii. Network switches
        iii. Disks
        iv. Archive Library
    c. Converge, 1 mo
3. Build Core Team, 3 mo
    a. Agree on job descriptions
    b. Start job announcement and look for talent
    c. Announce opportunities for undergrad and grad students in universities
4. Procurement, depending on technology, 3-6+ mo
    a. Huawei storage discs and Network Switches, 3 mo
    b. DELL servers, 4-6 mo
    c. IBM Tape Library (Archive Storage), 6+ mo
5. Set-up HW + MW + Testing + Integration, 7 mo
    a. Disk
    b. Computing
    c. Network
    d. Archival Storage Library

Author:
Vladimir Rekovic

12.09.2024.

Appendix: Construction and Commissioning of SSC-T1 (Phase-1)

Table 3. Original baseline schedule for the Project SSC-T1 center, Construction and Commissioning phase (Phase-1), for year 2024

	January	February	March	April	May	June	July	August	Sept	Octob	Novem.	Decem.
Coordination Team	PL	TC, DTC										
Define Technology		Study	Converge									
Core Team		Expert Search			Expert Hire		Student Hire					
Disk Storage			Procurement			Set-up	MW & Test					
Network			Procurement			Set-up		MW & Test				
CPU Servers			Procurement					Set-up	MW & Test			
Archive Storage			Procurement						Set-up	MW & Test		
Integration								Link HW, install MW, Test				