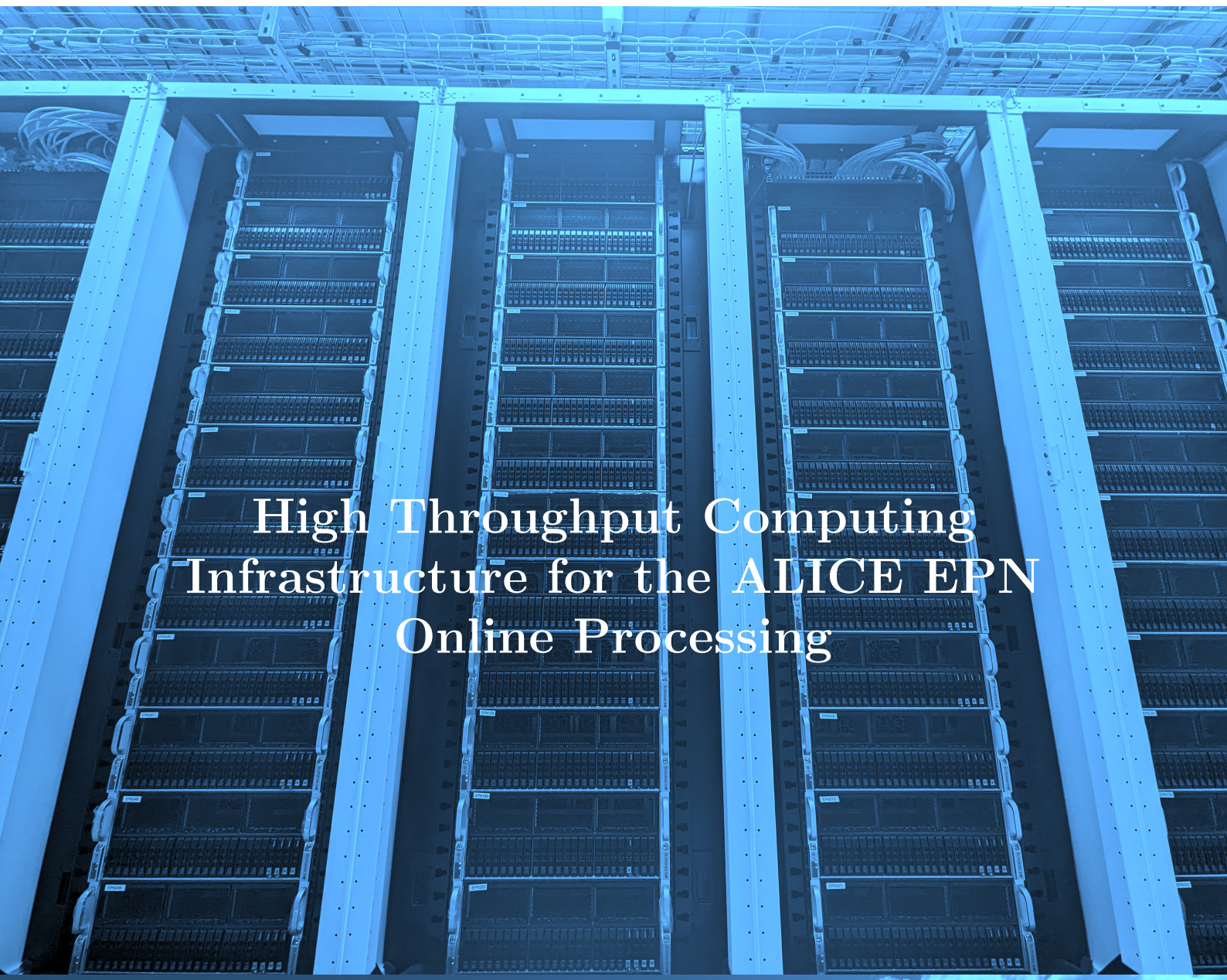


High Throughput Computing Infrastructure for the ALICE EPN Online Processing

Johannes Lehrbach

High Throughput Computing Infrastructure for the ALICE EPN Online Processing

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Informatik und Mathematik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Johannes Lehrbach

Frankfurt (2023)
(D30)

Vom Fachbereich Informatik und Mathematik
der Johann Woldgang Goethe - Universität
als Dissertation angenommen

Dekan	Prof. Dr. Martin Möller
Gutachter	Prof. Dr. Volker Lindenstruth
	Prof. Dr. Udo Kebschull
Datum der Disputation	

Abstract

A Large Ion Collider Experiment (ALICE) is a high-energy physics experiment, designed to study heavy ion collisions at the European Organization for Nuclear Research (CERN) Large Hadron Collider (LHC). ALICE is built to study the fundamental properties of matter as it existed shortly after the big bang. This requires reading out millions of sensors with high frequency, enabling high statistics for physics analysis, resulting in a considerable computing demand concerning network throughput and processing power. With the ALICE Run 3 upgrade [14], requirements for a High Throughput Computing (HTC) online processing cluster increased significantly, due to more than an order of magnitude more data than in Run 2, resulting in a processing input rate of up to 900 GB/s. Online (real-time) event reconstruction allows for the compression of the data stream to 130 GB/s, which is stored on disk for physics analysis.

This thesis presents the implementation of the ALICE Event Processing Node (EPN) compute farm, to cope with the Run 3 online computing challenges. Building a Data Centre tailored to ALICE requirements for the Run 3 and Run 4 EPN farm. Providing the operational conditions for a dynamic compute environment of a High Performance Computing (HPC) cluster, with significant load changes in a short time span, when starting or stopping a data-taking run. EPN servers provide the required computing resources for online reconstruction and data compression. The farm includes network connectivity towards First Level Processors (FLPs), requiring reliable throughput of 900 GB/s between FLPs and EPNs and connectivity from the internal InfiniBand network to the CERN Exabyte Object Storage (EOS) Ethernet network, with more than 100 GB/s.

The results of operating the EPN computing infrastructure during the first year of Run 3 LHC collisions are described in the context of the ALICE experiment. The EPN farm was delivering the expected performance for ALICE data-taking. Data Centre environmental conditions remained stable during the last more than two years, in particular during starting and stopping runs, which include significant changes in IT load. Several unforeseen external circumstances lead to increasing demands for the Online Offline System (O2). Higher data rates than anticipated required network performance to exceed the initial design specifications, for the throughput between FLPs and EPNs. In particular, the high throughput from an internal EPN InfiniBand network towards the storage Ethernet network was one of the challenges to overcome.

Contents

1 Introduction

This thesis describes the implementation of the Event Processing Node (EPN) compute infrastructure for A Large Ion Collider Experiment (ALICE). This includes the implementation of the Container Data Centre (CDC), which is internally called Counting Room 0 (CR0), housing the EPN servers, as well as the computing infrastructure itself, e.g. networking components and services.

The Large Hadron Collider (LHC) at European Organization for Nuclear Research (CERN) is currently the largest particle accelerator, providing a unique opportunity for fundamental physics research. ALICE is one of four big LHC physics experiments at the CERN. ALICE is designed to analyse heavy-ion Lead - Lead (Pb-Pb) collisions, to study the properties of Quark-Gluon Plasma (QGP). QGP is the assumed state of matter in the universe shortly after the big bang, at very high temperatures and pressures, before particles were formed. Chapter 2 gives an overview of the ALICE experiment.

LHC operation is organized in run periods of several years, intermitted by stops for maintenance and upgrades. During the LHC Long Shutdown 2 (LS2), between 2019 and 2021, there was a major ALICE upgrade, to enable data taking at much higher interaction rates of 50 kHz Lead - Lead (Pb-Pb) during Run 3 [14]. This required a paradigm shift from a triggered detector readout towards continuously running detectors, sending out data in a continuous stream. Compared to the previous Run 2 data rates, Run 3 data rates increased by more than 10x. This increase affects the whole processing chain. The input rate to the EPN cluster is estimated 900 GB/s during the foreseen 50 kHz Pb-Pb collision rate. More than 90% of the data comes from one detector, the Time Projection Chamber (TPC) [35]. EPNs are performing detector calibration and data compression in real time (online). This includes a partial event reconstruction of all detectors and in particular full track reconstruction for the TPC detector. With respect to TPC raw data, reconstructed events can be stored more efficiently by discarding the raw data and storing the track information, as well as the clusters associated with the tracks. By storing only the cluster differences towards the tracks they belong to, data can be compressed more efficiently. The resulting data rate to disk, after processing and compression on EPNs, is ca. 130 GB/s, without affecting the later physics analysis. Chapter 3 describes the ALICE Run 3 computing infrastructure requirements, for the upgraded system.

Chapter 4 provides a selection of available technology choices for the compute infrastructure. The implementation of the compute infrastructure, tailored to the ALICE Run 3 requirements is described in Chapter 5. This includes the Data Centre, which had to be built due to the lack of existing rack space and cooling for the increased equipment requirements. A new Data Centre was built close to the experiment, at the surface of Point 2 (P2). Besides the servers with Graphics Processing Units (GPUs), providing

the required computing power, a focus is on the EPN network for the High Throughput Computing (HTC) cluster. In the current production system, data is pushed with more than 100 GB/s to disk, from the internal InfiniBand network towards disks of the CERN Exabyte Object Storage (EOS) Ethernet.

LHC operation resumed in 2022, which gave the opportunity to evaluate the EPN computing infrastructure during realistic ALICE data-taking conditions. This allowed to verify that the EPN computing infrastructure fulfils the Run 3 requirements. Results of the EPN infrastructure performance at several milestones are shown in Chapter 6. There were no high-rate Pb-Pb collisions in 2022. The overall performance during Synthetic runs with simulated 50 kHz Pb-Pb Monte Carlo (MC) was as expected and demonstrated that the EPN computing infrastructure is fulfilling the specified Run 3 requirements. The first high-rate Pb-Pb collisions are expected in October 2023.

2 The ALICE Experiment at CERN

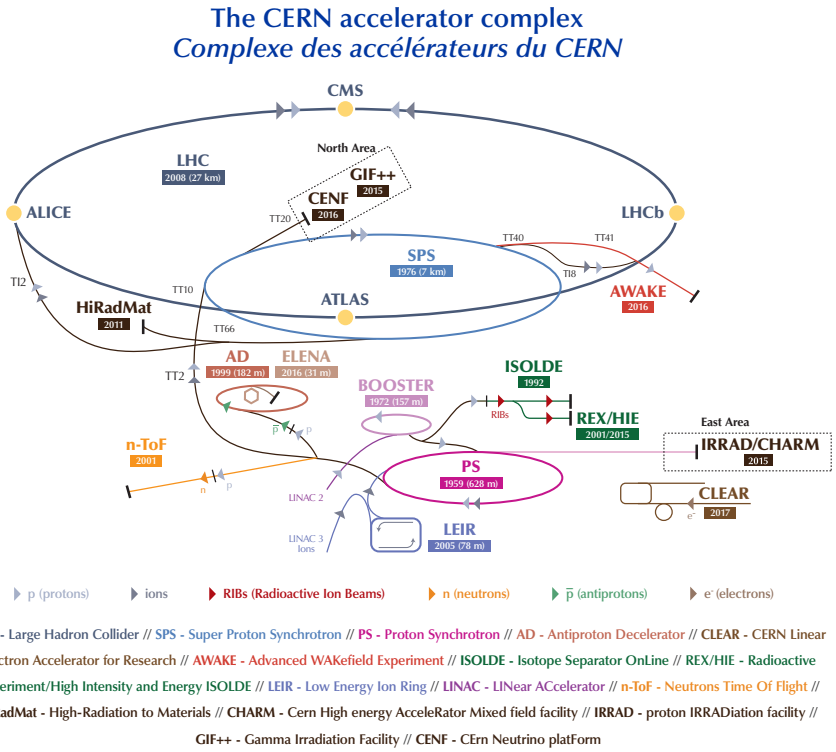


Figure 2.1: Overview of the CERN accelerator complex [34]. The largest ring, the Large Hadron Collider, is currently the biggest existing particle accelerator. ALICE is one of the four experiments located on the LHC ring.

The European Organization for Nuclear Research (CERN) is a research facility located in the Geneva region, at the border between Switzerland and France. It is a research facility in Europe focused on fundamental physics research. Figure 2.1 shows the Large Hadron Collider (LHC) accelerator complex, including all stages. Accelerating charged particles to the final energy requires several steps. To get positively charged protons, hydrogen atoms are stripped off their electron with an electric field. The protons are then accelerated by an alternating electrical field in Radiofrequency (RF) cavities [13]. The protons are injected in the LINAC2 (see Figure 2.1) and accelerated to 50 MeV. From LINAC2, the protons are injected into Proton Synchrotron Booster (PSB), which accelerates them to 1.4 GeV. From the PSB, proton beams are transferred to the Proton

Synchrotron (PS) and accelerated to 25 GeV. From the PS the beams are fed into the Super Proton Synchrotron (SPS), which brings the proton beams to 450 GeV. The SPS transfers the beams to the Large Hadron Collider (LHC) and accelerated to the top energy of 6.5 TeV [33].

There are two particle beams circulating in the collider, one moving clockwise and the other counterclockwise. The collision energy of pp collisions is the sum of the two beam energies and therefore up to 14 TeV at the design energy of 7 TeV per beam [33]. The LHC has four interaction points, where the two beams cross each other and particles collide. A Large Ion Collider Experiment (ALICE) is one of the four big experiments, located at the LHC interaction point at CERN Point 2, recoding collision data for physics analysis.

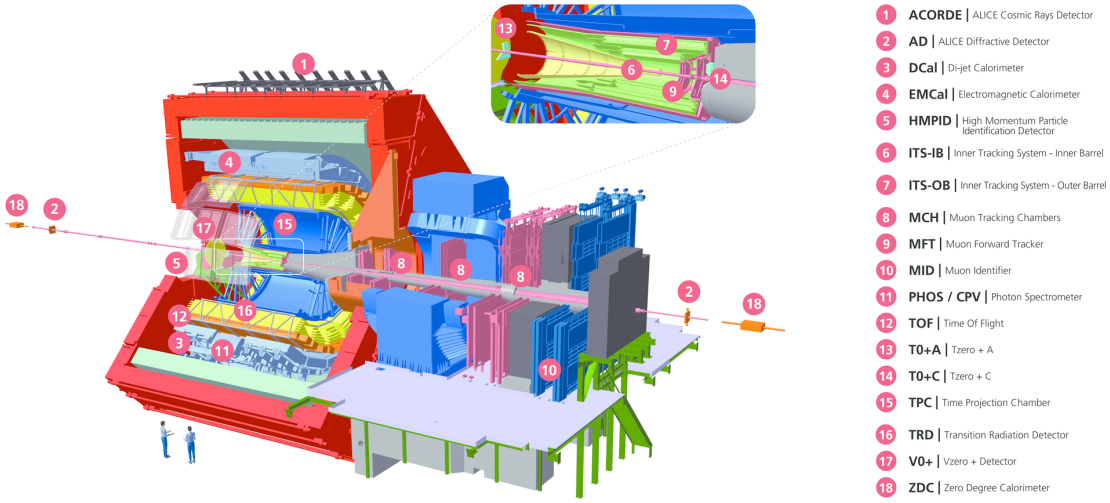


Figure 2.2: The ALICE experiment schema for Run 3 with all it's individual detectors [16]

The overall goal of ALICE is the study of heavy ion Lead - Lead (Pb-Pb) collisions at the LHC, to understand the properties of the strong force by studying matter in a state of extreme temperature and density. To get the best possible analysis opportunity, ALICE aims to capture as many Pb-Pb collisions as possible. The Run 3 experiment design is targeting continuous data taking at 50 kHz Pb-Pb interaction rate, compared to a triggered data taking between 2 - 3.5 kHz during Run 2. With the Run 3 upgrade, pp and Proton - Lead (P-Pb) collision rates of up to 200 kHz should be continuously recorded during online data taking. This increased event rate allows to gather more statistics and get a better understanding of rare processes.

Some detectors, like the TPC as a drift detector, require continuous readout to record the measurements of all collisions at high interaction rates. Additional complexity during the event reconstruction stems from the fact that multiple collisions can overlap in the TPC, which must be disentangled during the tracking [11]. This more than 10x increase in Pb-Pb collision rate required several changes to the experiment itself, which was done during Long Shutdown 2 (LS2) [14], as well as the complete readout and processing

system.

The experiment consists of several detectors with different sensor types, to measure specific parameters of the collisions. Figure 2.2 shows a schematic ALICE overview, after the LS2 upgrade, with all its detectors. For ALICE in Run 3, the overall detector readout design changed from a triggered running mode, only sending out data after a trigger identified an interaction of interest, to a continuous running mode in which the main detectors continuously send data and therefore capture all collisions. This came together with an upgrade of several detectors as well as changes in the readout schema introducing the Common Readout Unit (CRU), a PCIe card receiving detector data via up to 24 optical links. Some of the detectors were not upgraded significantly and are still running in a triggered mode.

The most important Run 3 upgrades, described in the respective Technical Design Reports (TDRs), were ITS [20], TPC [35] [15], MFT [19], FIT and the readout and trigger system [5], to enable data taking at higher Pb-Pb collision rates [21]. In addition, online and offline systems were restructured, to create the Run 3 Online Offline System (O2) [11]. In addition to this restructuring the compute capabilities needed to be upgraded significantly. With the Run 3 detector upgrades, the O2 system has to digest up to 900 GB/s, sent from the FLPs to the EPNs for online event reconstruction and compression of the data. The detailed compute requirements for Run 3 are described in Chapter 3.

3 ALICE Run 3 Computing Requirements

The ALICE upgrade during the Long Shutdown 2 (LS2) [14] also included an upgrades Online Offline System (O2). The increased data rates of ALICE Run 3 upgrade, as well as the paradigm shift towards continuous running tremendously increased the O2 requirements compared to the previous Run 2. From the compute perspective an increase in recorded collisions of more than 10x also increases the corresponding data volume compared to Run 2.

For ALICE Run 3 a new O2 software framework was developed. This development aimed to unify the code-base and to use the same software for online data taking and offline physics analysis. The whole ALICE O2 design for Run 3 was specified in the Technical Design Report (TDR) for the O2 computing system [11] in 2015. In this early planning stage there were a lot of estimates made, which were refined later on. In particular since the detector development was not yet finalized and several performance factors still needed refinement. There was an update of the O2 specification in 2019, the evolution of the O2 system [17]. Both documents combined were the basis for planning of the Run 3 EPN computing infrastructure.

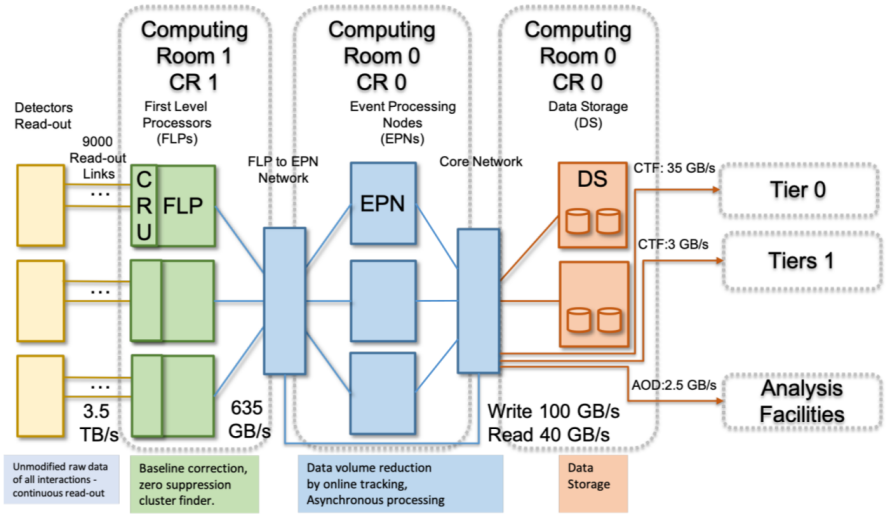


Figure 3.1: Schematic O2 Data flow as shown in "Evolution of the O2 system" [17]. During online data taking the dataflow is directed from left towards right. Offline reprocessing loads data from disk to EPNs e.g. for calibrations and stores it back on disk.

3.1 EPN Data Centre Requirements

Power and cooling requirements for the new O2 farm were estimated for the TDR in 2015. It was evident that the space, weight and cooling constraints for 1500 EPN servers could not be satisfied in the existing IT facilities in Counting Room 2 (CR2), used by the ALICE High Level Trigger (HLT) during Run 2. A new Data Centre was required, on the surface of Point 2, close to the ALICE experiment. The new data centre should have enough space to house the EPN servers, together with the required storage for the disk buffer of 60 PB and all necessary infrastructure and service nodes for the EPN cluster. Following Point 2 naming schema the new Data Centre is called Counting Room 0 (CR0).

IT estimates were done on a rack level and 54 U racks were chosen in the exemplary overview. For the 1500 EPNs 34 racks with a cooling power of 50 kW each was estimated. Including storage and all other items this summed up to a total of 45 racks with 54 U each, with a total of 2086 kW power consumption [11]. TDR specifications for the EPNs assumed a power consumption per U of almost 1 kW. Table 3.1 shows the technical specification of the tender documents [25], which was derived from the TDR requirements of CR0.

Table 3.1: Data Center power and cooling requirements as specified in the Container Data Centre (CDC) tender [25]:

CDC Facility		Rack height units (U)	Rack width (cm)	Power Height Unit (W/U)	Total Power (MVA)	Max. Footprint (m^2)
Alice O^2		2300	60	1000	2.1	27x24
LHCb	Med. Density	1300	80	550	0.6	35x16
	High Density	2700	60	700	1.4	
Neutrino Platform		650	60	650	0.4	22x9

Data Centre planning was done on the TDR estimates of 2015, since the updated numbers were not yet available in 2017, when the contract was awarded. The significant reduction in the number of servers done in 2020, see table 3.2 was not evident at the time of writing the technical specification for the Data Centre.

3.2 EPN Online Computing Requirements

Table 3.2 shows the evolution of EPN computing power requirements, based on knowledge at the respective time. Early on there were a lot of interpolations from Run 2

	TDR 2015	Update 2019	2020	2022	2023
EPNs	1500	750	250	280	350
GPUs : CPU cores	2 : 64	2 : 20	8 : 64	8 : 64	8 : 64 (MI50)
					8 : 96 (MI100)

Table 3.2: Evolution of the amount of EPN servers needed, from the TDR to running experiences in 2022, including number of GPUs and amount of physical CPU cores per server [11] [17] [14].

software and assumptions about hardware developments in the future, including a lot of safety margins, assuming 3000 GPUs and 96'000 CPU cores [11]. The evolution of the O2 system [17] reduced this estimate to 1500 GPUs and 30'000 CPU cores.

More than 90% of the processing time is consumed by TPC reconstruction. Most of the TPC reconstruction tasks were ported to run on GPUs [17]. A significant effort was also dedicated towards code improvements for both CPU and GPU. Together with a significant increase in computing power per GPU this could reduce the total amount of servers needed to an estimated 750 EPNs in 2019.

Since a large fraction of the overall compute is done on GPUs, servers with as many GPUs as possible have the best performance per cost ratio. Server hardware availability motivated the step to 8 GPUs servers in 2020, again reducing the number of required servers and therefore network ports significantly.

These updated numbers were not available when the Data Centre was tendered. With the knowledge of 2020, the amount of rack-space, as well as the overall required cooling capacity could have been more precisely tailored to the final requirements.

3.3 Network Requirements

	TDR 2015 (GB/s)	Update 2019 (GB/s)	Data Taking 2022 (GB/s)
CRU input rate	1095	3500	3500
FLP to EPN rate(total)	500	635	900*
FLP to EPN	400	570	830*
EPN to EOS	90	100	130

Table 3.3: Data rate evolution from the TDR to running experiences in 2022 [11] [17] [14].
* with the final TPC data format

Figure 3.1 shows the schematic overview of the O2 system as of 2019 [17], with data rate estimates as of 2019. The data flow for online data taking is directed and goes from the detector input on the left side of the picture towards the right side, to storage. The overall goal of the O2 system design was to reduce the data as early and as much as possible [11]. Planning of the TDR assumed zero suppressed data being sent from all detectors toward the O2 system. Zero suppressed data discards all sensor values below a specified threshold. The idea is that only sensor channels with an active signal are read out. For the TPC, the detector contributing more then 90 % of the raw data, zero suppression was not possible on the Front End Cards (FECs) and needed to be shifted to the CRU Field Programmable Gate Array (FPGA) [17]. With a better modelling of detector data, in particular TPC clusters compression rates the data rate from FLPs to EPN was increased from 500 GB/s to 635 GB/s. Data taking in 2022 showed that the new TPC detector was sending more data than anticipated and that the data rate will be 900 GB/s. The data rate from the EPNs to disk on the EOS was estimated to be up to 90 GB/s in the TDR, was updated to 100 and got now increased again to 130 GB/s due to an increased amount of TPC clusters found in 2022 compared to the initial estimates. Table 3.3 summarizes the data rate evolution from TDR toward the first year of operation with LHC collisions.

4 State of the Art

This chapter describes a selection of available technology choices, which were taken into consideration for the implementation of the EPN computing infrastructure. The O2 TDR provided a basic technology framework considered early in the planning phase. Besides the constraints of the ALICE experiment (see Chapter 3), ease of implementation, flexibility, as well as budget constraints, had an impact on the choices made.

4.1 Data Centre Infrastructure

Data Centre infrastructure for large installations can be challenging, depending on regional constraints. Large installations in the order of several megawatts of power consumption require proper connectivity to the power grid. A good location, which does not disturb surrounding residential areas can be important as well, since cooling infrastructure can be noisy. Outside connectivity, usually via optical fibres, plays an important role to get data in and results out of the Data Centre. Depending on the cooling solution, a good connection to water and potentially drainage is needed. Data Centres are often planned in industrial areas, which often already provide most of the mentioned requirements and are usually far enough away from residential areas, to allow some noise emissions.

ASHRAE definitions of the environmental envelopes for IT equipment operation [6] are the dominant specification hardware suppliers are using to design and build equipment. The most used envelope is ASHRAE A2, which allows supply air temperatures during operation between $10°C$ and $35°C$ with temperature changes of up to $5°C$ during any 15-minute window. This then specifies the cooling system requirements, to always ensure the operational environmental envelope. Power densities are trending upward in recent years, making it more challenging to provide appropriate cooling, in particular if power density is not homogenous and there are large differences in cooling requirements for a specific area. Avoiding hot spots and making sure the cooling reaches the equipment, e.g. by limiting bypass air, is increasingly important with higher power densities.

Equipment availability for specific use cases also plays into the Data Centre's operational requirements, e.g. the usage of standard 19" rack equipment vs. specific hardware with different rack requirements. In particular for server hardware with multiple GPUs as accelerators in a single chassis, equipment is only widely available for the ASHRAE A2 envelope. All GPU servers which were taken into consideration for the ALICE Run 3 EPN farm, as well as previously used Run 2 servers, are limited to operational temperatures of $10 - 35°C$ [7] [26] [44].

Environmental considerations play an increasingly important role. In recent years the focus was mainly on Power Usage Efficiency (PUE), limiting energy overhead for cooling.

Data Centres' fraction of the total energy consumption, to satisfy the rapidly increasing computational demand is hiking. Projects to recover the heat of a Data Centre for heating buildings are trying to reduce the impact on the primary energy requirements. Other ideas focus on the power infrastructure impact of Data Centres, in combination with renewable energies, proposing to detach Data Centres as large consumers from the grid and run them on their backup generators, to compensate for power fluctuations and stable the power grid.

4.1.1 Data Centre Cooling Concepts

There are different cooling concepts for Data Centres, which demonstrated good performance. Air and water based cooling concepts are the two predominant ways to implement Data Centre cooling. Immersion cooling can allow high power densities as well, but is still a more niche cooling concept, requiring specialized hardware. The most widely used concepts are described in the following sections.

Another important point is the cooling efficiency with respect to the overall equipment density it can cool effectively. In particular for highly integrated systems, the power density inside the racks did increase significantly in the last decade. For the majority of IT equipment, the ASHRAE A2 envelope is still the definition of allowed operational conditions. This means that the temperature at the equipment inlet should not exceed $35°C$ and not be lower than $10°C$. The ASHRAE A2 recommended environmental envelope for all air-cooled systems is between $18°C$ and $27°C$.

Free air cooling - direct and indirect

In recent years free air cooling got increasing attention, with an increased focus on energy efficiency. The basic idea behind this is to use the environmental conditions, in particular in areas with a moderate climate, where temperatures over the course of a year are not extremely high. In case temperatures remain significantly below $35°C$ all year long, the temperature difference between the exhaust air of the equipment and the outside temperature can be enough to efficiently cool the supplied air toward equipment inside ASHRAE A2.

The clear advantage of these climate conditions is the abundance of sufficient cool outside air. All that is required are big fans to move the air toward the equipment. This results in a very competitive PUE, having a low energy overhead for cooling. In the simple case of the environmental temperatures being lower than the desired supply air temperature, the outside air is just mixed with the exhaust air of the equipment, to get the desired temperatures. In particular for direct free air cooling, appropriate filters are required to ensure the operation of the equipment. In particular in areas with large vegetation, pollen can otherwise be a serious problem for the equipment, clogging fans and quickly deteriorating the cooling efficiency. In other areas, dust particles can be the dominant source of these kinds of problems.

In the climate zones, where most Data Centres are built, it will get too warm in the summer to have the whole cooling only relying on the outside air. In case the temperature

difference between the equipment exhaust air and the outside is not sufficient, a common way to ensure the cooling is by spraying water into the airflow. Phase transitions of different states of matter require a large amount of energy, which therefore can be used to cool down the environment around. This principle is used to cool down the supplied air to the equipment with evaporation cooling from the energy needed for the phase transition of water from liquid to gas.

For direct free air cooling, this is often realized by spraying water in front of the filter stage or directly onto the filters, to lower the air temperature to the required temperatures. Though one must be careful that the air humidity remains within the operational limits of the ASHRAE A2 envelope. Another potential issue when evaporating larger amounts of water is the caused by deposits of minerals in the water, remaining as leftover from the phase transition of the water. This can result in a mineral layer on the areas the waters are sprayed, e.g. the filters and therefore impact the functionality, in this example, the filtering capacity. This comes on top of the problems particles in the air can create, e.g. pollen, dust etc.

For the indirect free air cooling, the outside air is not directly supplied to the equipment but used to cool down the supply air with a heat exchanger to the desired temperatures. External and internal air is therefore separated and is usually not mixed. Warmer internal air transports the heat towards colder outside air via the metal surface of the heat exchanger. This process is not the most efficient way to transport heat energy and therefore needs large airflow rates. It is not possible to cool down the supply air to the exact outside temperatures with this concept, the remaining temperature difference depends on the system itself, the available surface areas and airflow rates. Heat exchangers are structures of material with a high thermal conductivity and a very large surface, to maximize heat transport. Thin metal structures have proven to be performing well for that use case, in particular Aluminium is widely used for its favourable material features with acceptable costs.

The concept of direct or indirect free air cooling only works in countries with the right climate and does not work in areas with very high temperatures, in particular in combination with high humidity. The availability of sufficient water for evaporation can be a crucial factor for this concept to work. Air can only absorb a certain amount of water vapour before being saturated and water condensing on surfaces. The higher the air temperatures, the larger the amount of water vapour air can absorb. This can lead to problems inside the cooling system itself, in case there are colder surfaces in contact with warm and humid air.

One potential issue is the mediocre heat capacity of the air itself. To achieve high power densities, a significant amount of air needs to be pushed through the system to get the heat out. This can be challenging for filter systems and the required pressure and speed to get sufficient air inside the Data Centre for direct free air cooling. For the indirect free air cooling filter requirements are often lower, since the air is not passing the IT equipment itself but the required airflow still has to pass by the heat exchanger in the internal as well as the external air circuit.

Cooling via Rack Back-Coolers

One popular way to cool IT equipment is to run water through heat exchangers mounted at the back doors of the racks. Warm air from the equipment passes through the heat exchanger and gets cooled down, keeping the room temperature stable. Depending on the required cooling capacity, the water needs to be sufficiently cold, to provide enough temperature difference in the heat exchanger to get the desired cooling effect. The server fans push the air naturally towards the heat exchanger in the back of the rack. Most of the back coolers have a few fans integrated, to assist with the airflow.

The water flowing through the heat exchanger is usually several degrees colder than the set point of the supply air for the equipment. Water has a much bigger heat capacity than air, so the actual water flow rates through the heat exchanger are usually moderate. Water allows the movement of significant amounts of heat energy easily over almost any distance via pipes.

Extraction of the heat energy is usually done via big cooling towers, evaporating water and using the energy loss from the phase transition to cool down the remaining water. The cooled water from the cooling tower is then pumped towards the Data Centre.

This cooling concept is well established and can be implemented in a very efficient way.

Water Cooling Inside the Racks

Modern servers with high power requirements on relatively small footprints are difficult to cool efficiently with air, due to the limited air heat capacity and the therefore required air flows needed for heat transfer. One solution to achieve higher cooling efficiency for this highly integrated hardware is to use water cooling inside the servers themselves for heat transport from the CPUs and all other silicon with high thermal load, e.g. GPU and other hardware accelerators. Due to the higher heat capacity of water compared to air, transferring heat from the heat sink to water is more efficient than with air. This allows much smaller heat sinks of all water-cooled components in the servers. By removing big volumes for the heat sink on top of the chips, even higher degrees of integration can be achieved, by packing chips even denser inside the chassis.

Usually, water-cooled servers are highly integrated into the rack, with a separate piping system inside the rack providing the required water flow. For these installations, it is crucial to achieve perfect water tightness, to prevent any damage to the installed hardware through leaks. The advantage of this is that in principle almost any power densities can be cooled, by providing the required water flows and temperatures. This way, the heat energy from the silicon dies can be efficiently moved towards big cooling towers. Another advantage is that the die temperatures can remain lower than with typical air cooling. This can allow higher power limits for the silicon.

One potential drawback is the hardware availability and compatibility of all server components. Water cooling requires a high integration level and therefore preferably a single-vendor solution. Most server hardware is still air-cooled and therefore not inherently compatible with water-cooled rack systems. It is of course possible to modify

servers, to allow water cooling, but this is time-consuming and costly. With increasing power densities and higher integration, water cooling gets more and more attention and the server hardware options with integrated water cooling are increasing.

4.1.2 Redundancy Considerations (n+1)

High availability is one of the biggest concerns of Data Centre operations. The downtimes need to be as low as possible since service non-availabilities have an increasing economical impact. To increase the uptime as much as possible toward 100 %, a common concept is n+1 redundancy of critical systems required to run the Data Centre. This can be a secondary power connection to the Data Centre, so the interruption of one power feed has no impact on the operation and equipment can still run on the redundant power line. In the case of cooling towers or other equipment, n+1 means that one complete part can fail, without any effect on operations. This is important for maintenance when infrastructure equipment often needs to be shut down for some time, to perform the required maintenance actions.

Data Centres usually have an Uninterruptible Power Supply (UPS) system and power generators installed, to allow operations even when the external power grid goes down. The UPS system usually consists of batteries, providing power until the standby power generators are running.

For the ALICE Data Centre, a full redundancy with UPS and power generators is not implemented. In case of a failure of the external power grid, large parts of the ALICE experiment and the LHC are temporarily without power and are therefore not operational. Only critical systems have an UPS with backup power generators. For a Data Centre in this particular use case, it is therefore acceptable to lose a large fraction of the computing hardware in case of power grid failure, when recovery is fast (in the order of a few hours) and the Data Centre is operational at the time the experiment is able to restart data taking. Sacrificing an UPS system of 2 MW, as well as power generators on standby, has a significant impact on the overall budget and was the only reason the new Data Centre could be built.

4.2 Available Standard Server Hardware

This section gives an overview of standard server hardware, which was considered for the ALICE EPN farm. It is focusing on widely available hardware, which is compatible with the Data Centre's infrastructure and environmental specification, standard 19" racks with an ASHRAE A2 envelope. Therefore, water-cooled equipment, mainframes and more specialized hardware are not included, since this is already excluded by design.

4.2.1 Main Processors

From a market share perspective, the two big manufacturers of server CPUs are Intel and AMD, with Intel having the biggest fraction of server CPUs. Arm also provides a server CPU, which is pushing into the server market.

The CPU determines the server hardware platform and is one of the first important choices to make. During the planning and hardware procurement phase, Arm platforms were not yet widely available and therefore are not discussed in detail. This leaves just two server platforms to choose from, Intel Xeon-based systems and AMD EPYC based ones.

For the ALICE software, a combination of CPU cores and GPU was required, to run efficiently. As the first baseline scenario servers with two GPUs and 20 CPU cores were the starting point for the hardware selection [17]. To optimize for costs and save on infrastructure servers with as many GPUs as possible were investigated. The detailed requirements are discussed in Chapter 3.

The AMD EPYC CPUs with the second-generation Zen architecture was an interesting possibility. With up to 64 CPU cores it is possible to choose the appropriate core count for the software. 128 PCIe-Lanes per CPU allow to connect up to eight x16 PCIe interfaces and therefore allow up to eight GPUs in a single server. In a dual-socket server, 64 PCIe lanes of each CPU are used to connect the two CPUs and are therefore no longer available, limiting the overall system to 128 PCIe lanes for GPU, as the single socket server. However, with the 2nd generation of the AMD EPYC, mainboard manufacturers got the option to reduce the amount of PCI lanes used to connect both CPUs from 64 to 48. This reduces the bandwidth between both CPUs, but increases the amount of available PCIe lanes to 160 and therefore supporting up to 10 PCIe x16 interfaces. This change allowed servers with up to 8 GPUs, together with an additional PCIe x16 slot for a network interface and up to four NVMe drives.

Intel Xeon CPUs were available with up to 28 cores per CPU and dual-socket solutions were available. The Intel Xeon at the time did not yet support PCIe gen 4 and only supported PCIe gen 3, with up to 48 lanes per CPU, for a total of 96 lanes in a dual socket server. Our initial baseline scenario was therefore possible with a Xeon server platform. However, an Intel Xeon server did not have sufficient PCIe lanes to support 8 GPUs in a single dual-socket server. With the Intel server platform, it was possible to have a server with 4 GPUs, an additional PCIe slot for a network interface as well as 4 NVMe drives. Compared to a dual-socket AMD server, twice the number of servers would be needed to house the same amount of GPU.

Single CPU systems vs dual CPU systems is another typical choice for standard server hardware. With the second generation of the AMD EPYC the dual-socket solution with the reduced interconnect, providing additional PCIe lanes to connect GPUs provides additional flexibility for the server manufacturers to tailor the hardware to the customers' compute requirements.

4.2.2 Hardware Trends In High Performance Computing

Two times per year the TOP500 list is updated, listing the fastest 500 clusters, which submitted Linpack benchmark results. Linpack measures the 64 Bit Floating Point Operations performance of a system. Since submission of benchmark results is voluntary and requires effort to run the benchmark itself, not all clusters are listed though. However it is prestigious to be one of the top performers in the list, in particular for scientific

projects and there is therefore incentive to get the benchmark running, in case there is a change to get the project listed.

The compute performance increase in the last decade was stunning, by now the first Exaflop cluster (10^{18} Floating Point Operations per second) is listed (since June 2022). The first Petaflop system was listed in 2008, this means an increase of a factor 1000 in 14 years.

Besides the TOP500 list, the GREEN500 is giving a measure of compute power per watts. The bigger systems require significant electrical energy in the the order of Megawatts. The environmental footprint of the big systems in the TOP500 is an increasingly important point, the economic impact due to the electricity prices and the total cost of ownership is also a very important aspect, the electricity costs over several years is a significant part of the overall required budget.

Most large clusters, as listed in the TOP500, are homogeneous and all nodes of the cluster are build from the same hardware [1]. This makes it much easier to administer and operate, since node configuration doesn't need to differ for different hardware. Another trend for supercomputers is the usage of GPUs as hardware accelerators, in particular for new clusters listed towards the top of the list.

The ALICE EPN system is following this trend and was designed with a single server type, each processing node having the same specs. The detailed implementation is described in Chapter 5

Hardware Accelerators

In recent years a shift towards hardware accelerators has taken place, particularly in the context of machine learning applications for neural networks. The usage of GPUs as hardware accelerators was particularly effective for the training of neural networks, due to the problem structure, which allows efficient GPU utilization for these computations.

GPUs can accelerate a wide variety of algorithms, in particular when they can be parallelized by splitting the problem into small independent subsets. The size of these computationally independent subsets has to be small enough to be processed by one of the many small GPU cores. The software infrastructure to enable good GPU utilization is very important, to allow algorithms to run efficiently on GPUs. The two major software frameworks used are ROCm for AMD and CUDA for NVIDIA GPUs. Experienced developers are needed to design algorithms utilising the full computing capabilities of the GPU architecture.

With the rise of deep learning a new generation of hardware accelerators did rise quickly in popularity, () were developed by big players in the industry, like Google. A caveat is that the TPU hardware accelerators are often not freely available and can only be used via cloud services. GPUs are still widely used for the training of neural networks, and are the most commonly used hardware accelerators, granting the big gains in speed-up.

FPGAs as hardware accelerators have the advantage of being extremely flexible and can be perfectly tailored to a specific problem. However, this can also be a disadvantage, since it requires significant effort from experts with hardware expertise, to design the

tailored algorithm for the FPGAs logic. Therefore, it is not usable out of the box and needs time to implement the required algorithms in firmware.

Energy Efficiency

FLOPS per Watt is one possible way to measure the efficiency of compute clusters, as used in the Green500 list. Energy costs are a significant part of the budget of a compute cluster and there is an increasing focus on limiting energy consumption, while further enhancing compute performance. One way, which proved very successful in this regard was the usage of hardware accelerators for certain parts of the algorithms, which are easily split into smaller sub-problems and therefore easy to parallelize on in optimized hardware, e.g. GPUs. Additional efforts to use accelerators like TPUs for machine learning or more general FPGAs as extremely flexible and reprogrammable hardware accelerators.

CPUs are designed to have a large instruction set, with a pipelined program execution. This gives good computational flexibility but sometimes has drawbacks in terms of efficiency. GPUs have smaller ALUs with more limited capabilities but can perform these simpler instructions much more energy efficiently. If the problems allow the usage of hardware accelerators, this is one way to significantly impact the energy efficiency, with respect to the performed FLOPS per Watt.

4.3 Network Options

In HPC clusters, the network is one of the central parts, which is of utmost importance for certain computing scenarios. Communication between nodes needs to be fast and often low latency, to allow the best computing performance and reduce wait times of nodes. Idle hardware is one of the biggest enemies of efficient computation and high system throughput. The network capabilities did grow in an almost exponential manner, with data rates doubling every couple of years. The high signal frequency gets increasingly problematic and often limits the distance of the network connection. In particular electrical signals need to be properly shielded, to allow transmissions in the hundreds of gigabit per second.

Looking at the TOP500 statistics, there are two dominant network families used in the listed Data Centres, Ethernet and InfiniBand. Omnipath, custom network interconnects or proprietary network solutions sum up to less than 15 % of the listed systems [1].

Underlying Technology

Modern network connections with 200 Gbit/s allow only cables of 2m via copper. Larger cable length require thicker cables, which can get quite stiff and more difficult to arrange inside the racks due to larger bending radii. One trend in recent years was to optimize cable length by installing switches in the middle of the racks and pulling cables through the sides of the rack, to add neighbouring rack equipment to the same switch. This increasingly replaces the concept of top-of-the-rack switches, which was predominant in

the past but requires longer cables for neighbouring racks. Greater distances can be covered with fibres, which can usually bridge distances in Data Centres of up to 100 m. Copper cables for cabling inside the rack has still a significant cost advantage over fibres, explaining the demand for copper cables for smaller distances. In general, fibres have a small diameter and are more fragile, but also easier to pull inside the racks and have a much smaller impact on the airflow, not blocking too much area in the back of the racks. Connections between switches are usually larger distances and therefore fibres are the only way to realize these inter-switch connections. Another trend in recent years is the splitting of ports, to allow a bigger number of connected equipment to a single switch. This can be extremely useful in case a single node does not require the full link capacity and can be operated at slower speeds, without any impact on the target computation. In such a case, Ethernet can currently split a single network port into four connections, each with $\frac{1}{4}$ of the total throughput. InfiniBand can currently split a single network port into two connections, each with half of the bandwidth. A lot of cost optimisation scenarios in Data Centres make extensive usage of this port splitting, to reduce the number of required switches or to connect more equipment in the racks to a single switch. Inter switch connections with the full available link speed can reduce the number of connections towards the backbone and therefore minimize the size of the usually expensive backbone overall as well. Extensive port splitting wherever possible seems to be even more tempting in the future, with even faster line rates and probably will remain one of the focal points for cost optimisations.

Faster link speeds were usually achieved by higher frequency signals. Since it gets increasingly difficult to handle high frequencies, other ways to increase the bandwidth are taken into consideration. In the past digital signals for networking were almost always binary, so basically encoding either 0 or 1 on the transmission line. For the 200 Gbit/s and 400 Gbit/s networking generation, this concept was changed from NRZ to PAM4, transmitting two bits in parallel by encoding everything in four different logical levels instead of just two. The signal-to-noise ratio is usually much worse by using four signal levels instead of two, reducing the eye of the signal, which allows the decoding without errors, to $\frac{1}{3}$ compared to NRZ. The signal quality plays an increasingly important role in the usage of PAM4. To achieve an acceptable Bit Error Rate (BER) with the reduction of decoding margins, mitigation techniques like Forward Error Correction (FEC) gain increasing importance. The FEC signal is encoded with error-correcting code, to not only detect bit errors but also correct them. This introduces overhead on the transmission line, and reduces the available total bandwidth with respect to the line-rate.

5 Architecture of the ALICE EPN Run 3 Requirements

This Chapter describes the requirements for the ALICE EPNs for Run 3. The data rates coming into the compute farms increased significantly from Run 2 to Run 3, as described in Chapter 3. Compared to the Run 2 HLT [40], the Run 3 EPN cluster input rate increased by more than 10x [14], from up to 48 GB/s to an estimated 900 GB/s, with a direct impact on network and computing requirements.

These significantly increased requirements for Run 3, of roughly one order of magnitude, required a new and more performant infrastructure. Run 2 HLT IT infrastructure supported up to 200 kW IT load and roughly 10 kW cooling per rack. The Data Centre Section 5.1 describes a 10x increase in overall cooling capacity for Run 3, for a total of 2.1 MW, while increasing the density by 3x to approximately 30 kW per rack in the new Run 3 Data Centre, to accommodate the increased compute demand.

Section 5.2 focuses on the server hardware. For the increased computing requirements the number of GPUs per server was maximised and increased to eight per EPN server, compared to a single GPU per server during Run 2.

With the larger data rates, network requirements increased significantly as well. Input data rates into the farm are more than 10x higher during Run 3, as are the output rates towards disk. The Run 3 network is described in Section 5.3. Moving data directly from an internal InfiniBand network to the Ethernet storage network added additional complexity to the setup compared to Run 2. Utilising newer generations of the network also increased the infrastructure constraints for copper and fibre connectivity, by increasing line rates from InfiniBand HDR (56 GBit/s) to InfiniBand HDR (200 GBit/s).

5.1 ALICE EPN Data Centre (CR0)

The Data Centre for the ALICE online processing farm was designed and build in a modular way. This in principle enables to tailor and scale the installation to actual needs and allowed tendering together with the CERN colleagues from LHCb and NP, each experiment with slightly different requirements for their Data Centre [25]. This chapter focuses on the ALICE Data Centre and the experiment's computing infrastructure requirements. As one of the technical contacts from the ALICE side between May 2017 and September 2019, my responsibility was to ensure all ALICE IT requirements are met by the new Data Centre.

During the design phase of a Modular Data Centre (MDC) one has flexibility and can easily scale the Data Centre to specific IT needs. Future extensions of additional modules usually have to be planned ahead of time, to have the necessary connections

in place and the space available. Once the Data Centre is built it is not easily possible to do significant modifications of the Data Centre, e.g. add additional IT containers for extra equipment.

For ALICE the Modular Data Centre (MDC) consists of four containers for the IT equipment and one utility container, providing the required power and water infrastructure necessary to run the IT containers. Reflecting the container implementation of modules, Container Data Centre (CDC) is synonymously used to describe the MDC.

For modern HPC workloads tailoring the hardware to the software requirements is an important part of the planning phase, to keep costs in check and not to waste unused resources. Since the compute requirements for ALICE, as specified in the Technical Design Report [11], were clear and would not change significantly for the experiment's duration, the new Data Centre could be fully tailored to the ALICE Run 3 needs.

Figure 5.1: Installation of the power and utility container at P2 as the first step of the Data Centre installation in 2018.

The tender itself was kept open with respect to the cooling solution and Data Centre design, to allow different technical implementations. Automation's [8] offer was the winning bid of an European wide tendering process. After the tender decision was done ALICE had a detailed look at the offer and asked for several modifications, to better suit the infrastructure needs. In the initial ALICE design specifications a minimum of 2300 U and 60 cm racks was required for the new Data Centre. The winning offer consisted of

5 containers with 21 60 cm wide and 42 U high racks and a total cooling capacity of 564 kW per container. This offer therefore extended the available rack space to 4400 U and could provide 2.82 MW cooling capacity. This largely exceeded the estimated rack space and was including roughly 30 % more cooling capacity then estimated [11]. Budget considerations motivated the design change to only 4 IT-Containers, since this was still inside the specifications with respect to cooling capacity and rack space. The smaller installation saves ALICE a considerable amount of money and gets closer to the initial cost estimates. Together with LHCb ALICE decided to request additional modifications of the IT-Containers. Extending them by 2 m still fits on the envisaged area footprint and could accommodate a few more racks into each container. Increasing height allowed to install 48 U racks, and not only 42 U, sacrificing space above the racks. This was easily possible, since the required space for cabling was still available after this change. ALICE decided that it would be most convenient for equipment installation and operation to have enough space for cabling and to use 80 cm wide racks instead of 60cm, even though this meant a 25 % reduction of racks per container (18 vs 24). Four IT-Containers with 18x 48 U 80 cm racks still provide plenty of rack space in that configuration after deciding on higher racks and longer containers. This was one of the first optimizations for the Data Centre usability and would provide more flexibility for hardware installation and cabling. The Data Centre was built from 4 IT-Containers each with one row of 18 racks, each rack 80cm wide 48 U high, providing 3456 U of total rack space for equipment. This is roughly 1.5x the estimated requirements [11] and therefore providing good margins, while easing installation with wider racks. The provided cooling capacity with four Containers was 2.25 MW and also still above the requirements.

A big advantage of a CDC is flexibility combined with the ability to pre-assemble everything [8]. This means most of the required facilities are already inside the containers at delivery. In particular power infrastructure, racks, cable trays and all the things like lighting, electrical power outlets, etc. can already be assembled before delivery. This minimises the required work on site and reduces the overall time until the Data Centre is operational. The container as a whole is transported with a special transport to the destination and then put in place by a large crane. Figure 5.1 shows the power container placement with a crane at Point 2 in 2018, which was the first installation step of the Data Centre containers. Solid foundations are a requirement for the whole Data Centre, to ensure a good and stable position for all containers of the CDC. Civil engineering included concrete foundations for each container, trenches for services (power, water, network connectivity, drainage), additional piping preparations, cable ducts for later use of arbitrary connectivity, pavements between the containers to ease accessibility, etc. All civil engineering works at Point 2 were coordinated and supervised by ALICE technical coordination. These civil engineering works were done well in advance, to give time for the concrete foundation's hydration and curing. All that is left to do after delivery and installation of the containers is to provide the external connections e.g. power, water, drainage, etc. The required works can be done in a relatively short time span, and the Data Centre containers are therefore quickly usable after delivery. However installation of the actual IT equipment, providing network connectivity and in particular all outside connection, usually takes time to establish. For ALICE it took several months after the

first IT-Container delivery to finalize external fibre installations and to get the Data Centre fully into production with the vertical slice set-up as the first test system.

Figure 5.2: CR0 Data Centre view after two IT-Containers were installed at Point 2.

One of the technical requirements ALICE specified was a cooling capacity of up to 1 kW per rack height Unit [11]. This should allow servers with multiple GPUs. However, the average cooling power per rack in the final solution is significantly below 1 kW for the initially proposed 21 x 42 U as well as the modified design 18 x 48 U, only around 600 W per U, see 5.1 [9]. The increased rack space together with a good peak cooling power per rack height unit gives us the required flexibility for the hardware choices. The difference between average and peak cooling capacity motivated intense testing, to verify a reliable cooling of up to 1 kW per U, without creating hot spots. This was one of the challenges during the commissioning phase, detailed test results are described in chapter 6.

$$\frac{\text{Total cooling power per container}}{\text{Height units per container}}$$
$$\text{Initial proposal: } \frac{564}{882} = 639\frac{W}{U} \tag{5.1}$$
$$\text{Final solution: } \frac{525}{864} = 607\frac{W}{U}$$

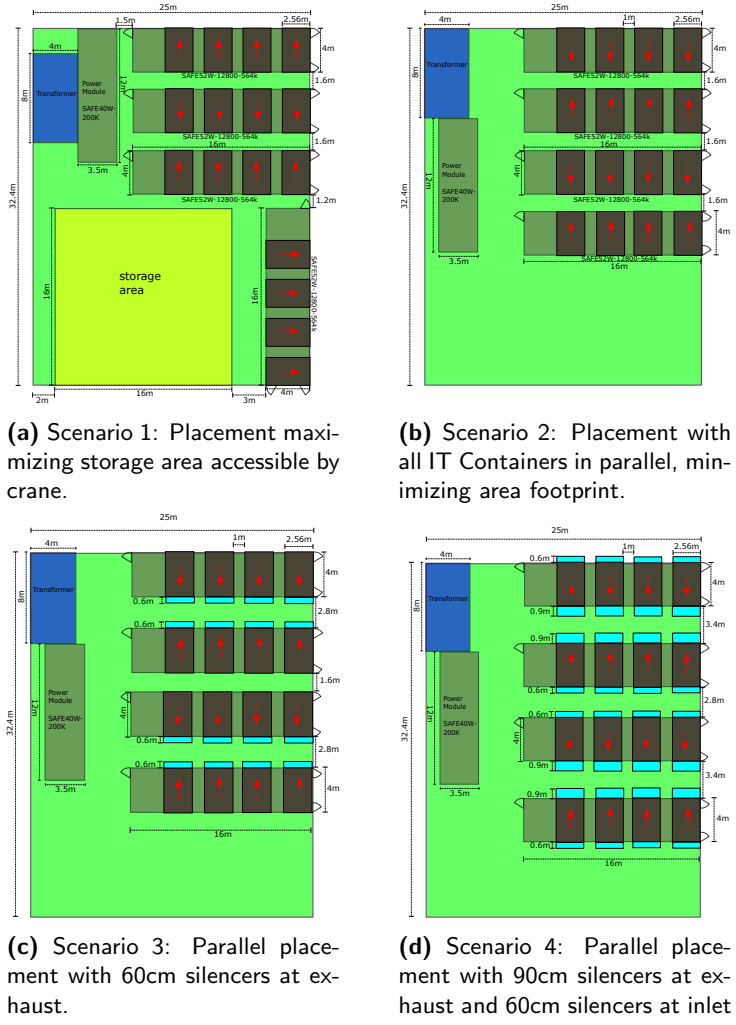
An advantage of higher racks is the potential server proximity to the Top of the Rack (ToR) switches enabling a better network integration. High speed networks with connection speeds of more than 100 Gbit/s have a limited range with passive copper cables (DAC) [38], due to the high frequency of the electrical signal. Higher frequencies suffer more losses e.g. increasing insertion loss and attenuation from increasing resistance of the cable (skin effect) [31][29] and therefore the length of Direct Attached Copper (DAC) cables is more limited. For the 200 Gbit/s InfiniBand network the maximum length of DAC is 2 m for all HDR cables. In case the link speed is reduced to 100 Gbit/s port speed, using 3 m EDR DAC is an option. Network connectivity to a ToR via DAC for 200 Gbit/s and above is limited to the rack housing the ToR switch and its neighbours, in particular if wider racks are chosen (racks above 60 cm width) and cables are routed above or below the racks.

Data Centre Noise Considerations

Another required modification for the ALICE Data Centre, with respect to the initially offered design, was the addition of silencers for the cooling units, to significantly reduce the noise generated by the Data Centre. CERN has multiple sites close to residential areas. In particular Point 2, where ALICE as well as the new Data Centre is located, is in the vicinity of residential buildings. CERN internal policy requires noise levels of new infrastructure to be 10 dB lower than the surrounding already existing buildings, e.g. the cooling infrastructure and transformers at Point 2. A noise level of -10 dB corresponds to $\frac{1}{10}$ of the existing noise level. This should ensure that new buildings at Point 2 do not increase the noise level at the site boundary and in particular at the premises of close residential buildings.

Internal sound sources are well attenuated by the Data Centre structure. The IT equipment itself and internal fans therefore only contributed to the noise level inside the Data Centre. The cooling units' fans driving the external air circuit are the main contributors to the overall noise emissions of the Data Centre and are located close to the cooling units exhaust, see figure 5.6. The target was therefore to reduce the fan noise level reaching the outside as much as possible, without hindering air flow too much.

ALICE technical coordination mandated an external consulting company to provide noise simulations for the Data Centre, to ensure the new Data Centre installation stays within the CERN objective of -10 dB. The noise simulations were done for multiple scenarios with respect to the container orientation and silencer configuration. Simulation assumed the worst case, maximum fan speeds, for the highest possible sound level. I provided the different placement scenarios shown in figure 5.3 for the simulation input.



(a) Scenario 1: Placement maximizing storage area accessible by crane.

(b) Scenario 2: Placement with all IT Containers in parallel, minimizing area footprint.

(c) Scenario 3: Parallel placement with 60cm silencers at exhaust.

(d) Scenario 4: Parallel placement with 90cm silencers at exhaust and 60cm silencers at inlet

Figure 5.3: Data Centre configurations used for noise simulation:

Figure 5.3 shows the simulated Data Centre configurations and Table 5.1 summarises it's impact on the noise level at the sites' premises and the closest residential areas [23]. The resulting noise levels were too high for the internal policies for the two container placement scenarios without silencers (Scenario 1 and 2 in figure 5.3). Both configurations with 60 cm and 90 cm silencers (Scenario 3 and 4 in figure 5.3) were inside the internal policies of a -10 dB target [23]. It was decided to go with a 60 cm silencer for the exhaust of each cooling unit, however adjusting the distance between the containers to allow future modifications with the larger silencers at the suction and exhaust. A distance greater than 1.5 m between cooling units outlet openings facing each other is required to prevent any cross-talk of the units facing each other. This distance had to be adapted when using the silencers. In this example the configuration with 60 cm adds the same length to the outlet, supernatant of the container and therefore adds 1.2 m towards

Scenario	Res1	Res2	Res3	PL1	PL2	PL3
1	20.4	30.0	34.3	24.7	33.7	37.9
2	20.0	27.5	32.0	22.6	33.6	37.0
3	13.1	19.8	26.2	15.5	30.1	32.4
4	6.2	14.7	17.9	8.5	23.7	23.9
Cern target	27.8	25.8	27.8	50.0	47.8	49.9

Table 5.1: Simulated scenarios from figure 5.3. Results for closest residential areas (Res#) and property limits (PL#). Noise levels too high at two residential areas for the scenarios without silencers.

a total of 2.7 m minimum distance between two containers on the exhaust side. The containers at Point 2 have a distance of 3.9 m between the containers on the exhaust side. Initial tests of the prototype cooling unit at the manufacturers test-bed, as described in chapter 6, raised the impression that noise levels were higher than anticipated. ALICE technical coordination invited the consulting company to witness following tests and measure the resulting noise for different fan speed configurations. These measurements were used by the company to verify expected noise levels of the new Data Centre at different loads [24] and confirmed the noise level from the first simulation.

During the commissioning, noise measurements confirmed that the Data Centre noise emissions are matching simulations and are well within the set noise limits. Regular tests are done to ensure that everything remains inside acceptable noise levels.

5.1.1 IT Equipment Environmental Requirements

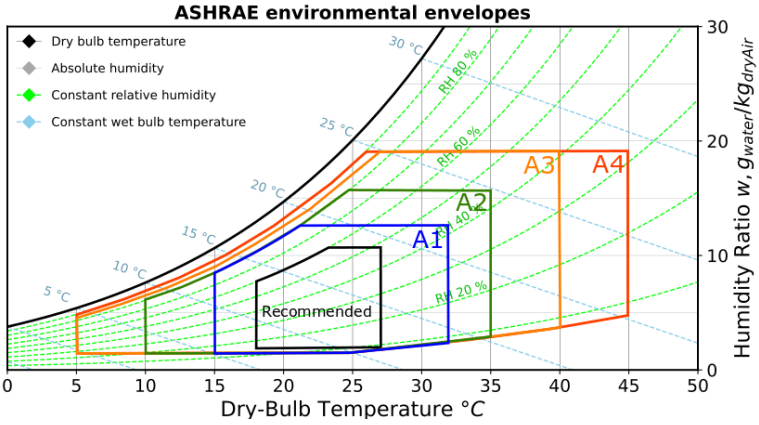


Figure 5.4: ASHRAE environmental envelopes for IT equipment according to TC99 [22] [6]. The servers are specified for the ASHRAE A2 envelope (dark green box). Recommended range for all environments indicated in black.

One of the Data Centre challenge is the potential power density of multi GPU systems,

which can be challenging to cool appropriately with traditional air cooling. Especially when the load is not distributed equally among racks and the locally required cooling differs significantly. This can easily lead to hot spots, if the cooling system is not designed optimally. One of the challenging parts for the control system of the cooling is the very specific use case and the experiments run concept. In this context a run starts by enabling the detector read out and feeding the data through the system, including the whole processing chain. This results in ramping up full computational load in seconds after getting the data from FLPs and also stopping rather abruptly when the run is stopped and detectors are no longer read out and therefore no more data is pushed to the EPN farm any longer. Compute load in the cluster is directly correlated to data coming from the detectors and can quickly change. This extreme behaviour can bring the power consumption from idle to full power and back in less than half a minute, requiring the cooling system to be reactive and fast to steer towards the actual needs at any point in time. A failure to steer appropriately can quickly amount to peak supply air temperatures or lead to oscillations of the supply air temperature over a longer time. Chapter 6 gets into details of the control system commissioning and testing, to ensure appropriate environmental conditions in the data rooms. Common IT equipment is usually specified to work inside the ASHRAE A2 environment, between $10°C$ and $35°C$ server inlet temperatures, as described in figure 5.4. It is possible to get equipment with a wider operational temperature range, though this is often specialized equipment for telecommunication purposes or other highly specific use cases. Equipment for wider operational temperature ranges, as specified in ASHRAE A3 and A4 [6], therefore tends to be significantly more expensive. In particular for highly integrated equipment requiring significant airflows for cooling, the A2 environmental envelope is often the only available operational specification since local hotspots inside the server can easily become problematic for higher supply air temperatures and high integration levels. In particular from a cost and hardware availability perspective the Data Centre was planned and designed for ASHRAE A2. Increasing Data Centre supply air temperatures can significantly increase the power efficiency of the cooling and is one of the parameters to tune for a better Power Usage Efficiency (PUE) [10]. The average PUE for European Data Centres in 2016 was 1.64 [10]. For CR0 one of the important considerations was energy efficiency, with a target PUE of below 1.1, significantly below the average Data Centre at the time. Besides environmental benefits of energy reduction, the driving motivation are electricity expenses, which are a significant cost factor of Data Centre operations. More and more equipment is available for a wider operational temperature range and there might be a shift for future Data Centres to focus more on ASHRAE A3 and A4. With the Data Centre tailored to the ALICE EPN Run 3 needs, the choice of the specific server hardware remained flexible. The equipment selected is described in chapter 5.2.1.

5.1.2 Cooling

The cooling design was done by the supplier with respect to the technical specification of the requirements. Each IT-Container has 4 independent cooling units mounted on top (see Figure 5.5), directly above the racks. Figure 5.6 shows the general airflow schema of

Figure 5.5: IT Container with 4 cooling units side view.

the cooling cycle as well as an outside overview of a whole IT-Container. Total cooling capacity of the installed IT-Container is 525 kW ($\frac{1}{4}$ of the total Data Centre cooling of 2.1 MW). One row of 18 racks is installed below the cooling units with a strict hot aisle separation, to direct the supply air flow through the equipment and avoid bypass air. There is only a single cold & hot aisle in the container.

Figure 5.6 shows the working principle of the cooling. A big air-to-air heat exchanger is the very core of the cooling units. Internal and external air circuits are completely separated and the plate heat exchanger is where the heat transfer from inside to outside is taking place. Cooling 525 kW per container, with four cooling units, require a temperature difference of at least $8°C$ between outside and supply air temperature to run in dry mode at maximum load, just passing enough outside air through the heat exchanger for the required cooling effect. For the Data Centre, a set-point of $27°C$, the upper end of the recommended temperature range for ASHRAE A2 [6], as indicated in figure 5.4 was chosen. This means for outside temperatures above $19°C$ air alone is not enough to achieve the maximum cooling. In such a case water is sprayed on top of the heat exchanger. Cooling units have a water reservoir located below the heat exchanger, from which the water is pumped upwards and sprayed on top. The water reservoir also collects the remaining water, which is not evaporated immediately and drops down again. Evaporation energy required for the phase transition of water from liquid to gas adds the additional cooling effect required. For higher outside temperatures more water is sprayed and evaporated. This adiabatic mode can consume a significant amount of water, evaporating more than $150\frac{L}{h}$ per cooling unit under worst case weather and load conditions [43]. Geneva climate does not require additional mechanical cooling (e.g. direct expansion cooling using a compressor) for summer temperatures, the adiabatic mode is sufficient to achieve the necessary cooling capacity for all weather conditions.

The outside air passes through a bag filter, to remove pollen and other particles, to

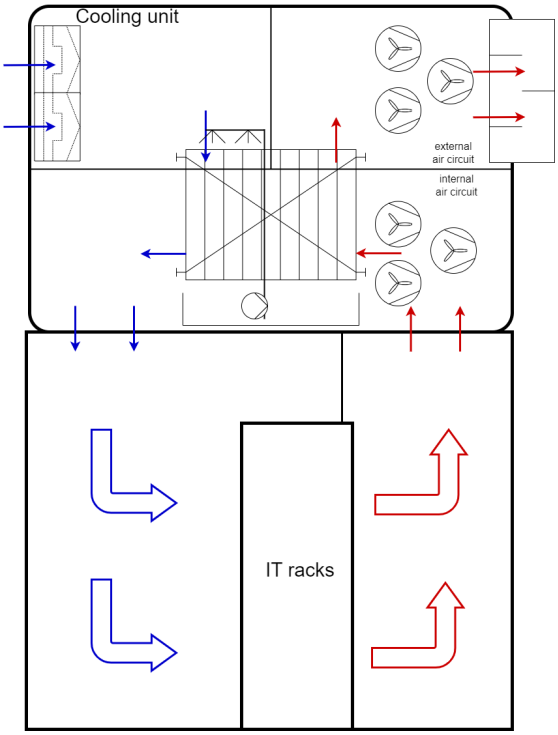


Figure 5.6: Internal IT-Container airflow

keep the heat exchanger and the water clean. A silencer at the exhaust dampens fan noise to reduce the emitted sound level. The units are mounted on top of the containers and directly connected to a cold and hot aisle, without any additional ducts. Cooling unit fans are sucking air from the hot aisle and push it through the heat exchanger towards the cold aisle. This concept requires a strict hot aisle separation, to ensure the supply air is directed through the IT equipment, without too much bypass air.

Since all cooling units are steered individually there is potential cross-talk inside the container between neighbouring cooling units. This allows shutting down one cooling unit for maintenance during operation, while neighbouring units take over the cooling load. During hot summer periods shutting off one unit would reduce the total cooling capacity. Since the installed peak IT load is around 75 % of the maximum cooling capacity a failure of one out of four cooling units should not impact ALICE operation. For more moderate weather conditions, which are not extremely hot, cooling units have sufficient head-room to compensate for the loss in capacity without problems. Maintenance is also scheduled during LHC technical stops, which means there are no collisions and therefore less compute load on the Cluster. Scheduled like this, Data Centre maintenance has therefore no impact on the experiment and is transparent for the data taking campaign. This is an important factor due to the immense operating costs of the LHC including ALICE. It is a priority to gather as much experiment data as possible during stable beams.

5.1.3 Current IT Installation

The first containers of the new Data Centre were delivered end of 2018, well before the start of ALICE Run 3. The full Data Centre was completed in Q4 2019 and available for operations. The milestones defined in the evolution of the O2 system [17] describe the installation and commissioning process in detail, with a first version of a vertical slice system available in July 2019. This required the new Run 3 network connectivity and the Data Centre to be in place and it was therefore required to have the infrastructure installed ahead of time. Servers from the previous Run 2 were used to build the vertical slice system, to verify the software and the complete readout and processing chain. The vertical slice was planned to allow continuous testing and integration during the commissioning phase, parallel to the detector installations. This allowed us to verify the O2 system was working according to the design specifications.

To capture technology advances which usually lead to compute performance improvements, it was decided to purchase the final IT equipment for the EPN Cluster at a later stage. The first Run 3 servers with the final hardware arrived end of 2020 and were operational beginning of 2021.

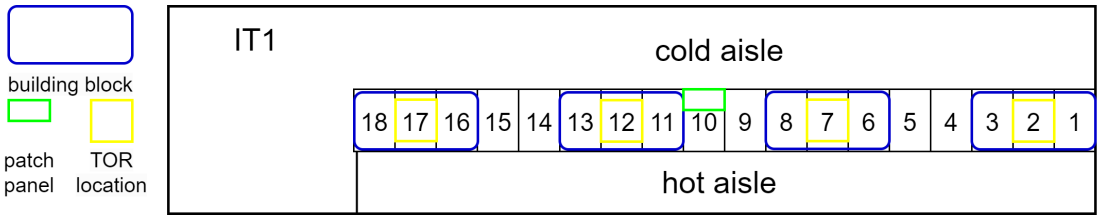


Figure 5.7: Schematic server installation in IT1

Figure 5.7 shows my proposed layout of the IT installation in the container IT1. The layout planning was done with building blocks of three neighbouring racks, to make good use of the InfiniBand network switches. Figure 5.15 shows the current layout of one of the building blocks. The network switches are installed in the middle rack, to keep the cable in a manageable length. For InfiniBand HDR the maximum length for DAC splitter cables is 2 m. Servers which are further away can be connected by 3 m DAC EDR cables, increasing the amount of needed ports compared to splitter cables. To maximize the servers connected to a single switch the distance and the usage of splitter cables is important. Since the network detailed planning was done after the procurement of the containers and the rack modifications were not easily possible the network switches are located at the top of the middle rack. In case of openings in the side wall, between all racks, it would have been possible to install the switches in the middle of the rack and therefore reaching more servers with 2 m splitter cables. The range limitations were not yet known in the planning phase of the containers and these optimization was not easily possible after the containers were built and delivered. Since the servers have InfiniBand HDR-100 HCAs using EDR cables, they do not have any performance penalty and can easily be used alongside HDR splitter cables, using more switch ports. With 3 m cables we can reach the full neighbouring racks of the ToR switches.

Connectivity to the core switches is provided via 10 200 Gbit/s HDR links via the patch panel placed in the middle of the container IT1, to optimize the patch cable length required to connect each Top of the Rack. Additional building blocks can be easily added, as long as there are free ports on the core switches. This results in a scalable design, until the core switch ports are fully utilized. To ease controls of the cooling system, IT equipment rack layout was done as balanced as possible, to balance the load as equally as possible throughout the containers. Each container has four cooling units on top, to provide the required cooling capacity. The number of building blocks per container is four as well. To balance this in the most optimal way, building blocks inside the containers are positioned just below one of the cooling units. This leads to gaps between the building blocks and can require shuffling around equipment, in case of an extension of the existing farm and the addition of building blocks. The benefit of one cooling unit having to cool exactly one building block far outweigh this inconvenience, especially after first load tests showed significant temperature oscillations in the cold aisle. Figure 5.8 shows the layout of the central container, IT2. In addition to the servers it contains most of the infrastructure required to run all the services, the core network switches, as well as the connectivity towards the FLPs and EOS.

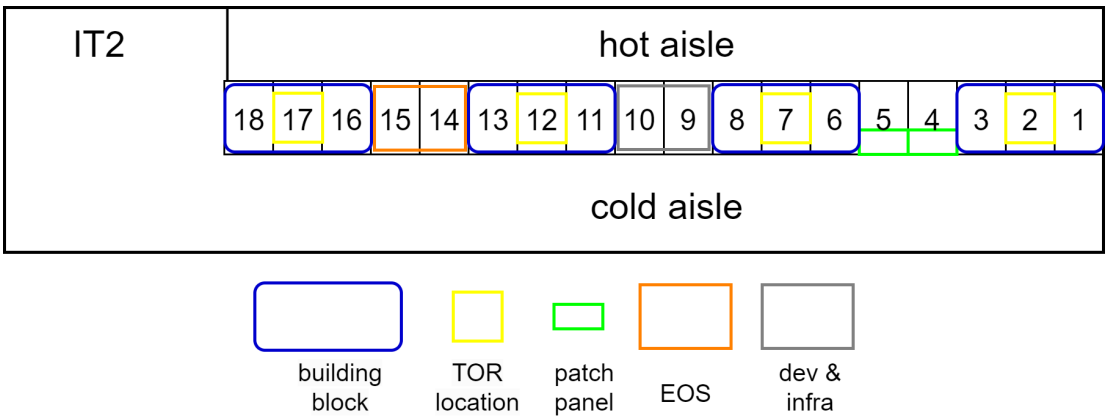


Figure 5.8: Schematic server installation in IT2 including backbone for the data network, EOS instance at Point 2 and services.

The experiences with the current layout are positive. The cooling system is well behaved in this configuration since the load distribution is equally distributed along the building blocks. Due to the network topology and the requirements for data throughput it is important to schedule nodes equally throughout the cluster, by using roughly the same amount of nodes per building block. So even if the servers are used for different jobs the load distribution remains balanced throughout the containers, as well as each building block.

During the first stress testing with the new detectors for high collision and therefore data rates it got increasingly clear that the available computing was not sufficient for the real time processing requirements and that the experiment probably cannot run with the expected rate without dropping data due to a lack of real-time processing and as a result

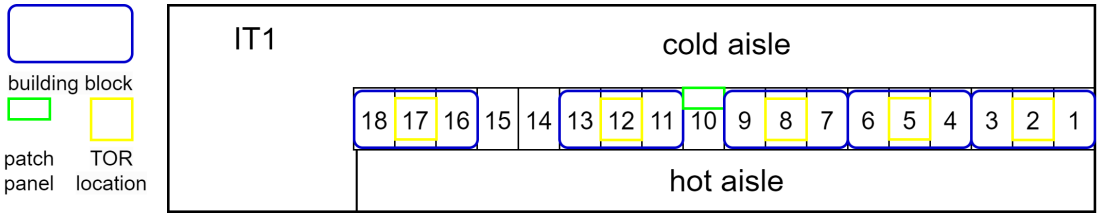


Figure 5.9: Schematic server installation in IT1 including additional building block for 30 new servers.

of insufficient buffers. During the time of writing the installation was already extended with a new building block of 30 additional servers in IT1 anticipating the first Pb-Pb run end of 2022. Due to a potential energy shortage and the large energy consumption of the LHC, CERN decided to stop operations earlier in 2022, to significantly reduce energy consumption and locally relieve the situation a bit. This changed schedule led to postponing the Pb-Pb run to 2023. It was therefore not yet possible to check the real performance for high rate Pb-Pb data taking from the cluster. To have additional processing margin the planning to extend the server farm by another 70 servers was finalized end of 2022 and the procurement processing was started in Q1 2023. The initial 250 compute servers were increased to 280 with the aim to be at 350 before the Pb-Pb campaign October 2023. Figure 5.9 shows the additional building block. To fit another block, moving one of the already existing building blocks by one rack was required, to have three consecutive free racks for the new block. This means the IT load is no longer perfectly balanced underneath all cooling units, so half of the units have more IT load in their suction area than the other half. This does not impose any issue for the cooling system. After tuning and in particular fixing the input temperature measurement for the cooling unit control system, as described in Chapter 6.1, stability is not affected by an uneven load distribution inside the racks.

For the additional 70 servers another building block will be required and in addition will add three or more additional servers to existing building blocks to fill up all available rack space.

5.1.4 Energy Efficiency

One of the central constraints for the new Data Centre was the CERN policy to have highly energy efficient infrastructure for new buildings. For the Data Centre the resulting constraint was to have a PUE of less than 1.1. The PUE as defined in equation 5.2 is the total IT load plus all the additional power used by the Data Centre for cooling, water treatment, lights, losses of the power distribution etc., divided by the total IT load.

$$\begin{aligned} \mathrm{PUE} &= \frac{P_{\text{Total Data Centre power consumption}}}{P_{\text{Total IT load}}} \\ &= \frac{P_{\text{Total IT load}} + P_{\text{Data Centre Infrastructure}}}{P_{\text{Total IT load}}} \end{aligned} \tag{5.2}$$

A PUE of 1.1 therefore means that the total overhead for all the Data Centre operation, including the cooling system should be designed to only generate an overhead power consumption of 10 % of the actual IT load. This level of power efficiency is not easy to reach and requires significant effort. At the time of planning average PUE of existing Data Centres was around 1.6 [10]. In the technical specification for the Data Centre, the operational window for the power efficiency is constricted between 50 % and 100 % of the maximum IT load, which is common practice, to make it more feasible to achieve and to prevent static consumers to dominate the equation for extremely low load scenarios.

ALICE did not impose any limitations on the cooling system design other than the required total cooling capacity per U and the PUE [25]. Design as well as the implementation was up to the bidding companies and was evaluated for conformity by us afterwards. There were multiple concepts introduced by the offers like back-coolers in the rear rack door. The winning offer uses an air-to-air heat exchanger in combination with an adiabatic mode, evaporating water to cool at higher temperatures. Main energy consumers of the cooling system are internal and external fans. Fan power consumption increases with fan speed cubed and it is therefore usually very power intensive to operate fans closer to their maximum speed. To keep energy consumption in check the fan speed in the cooling system is capped at 72 % at the external and 80 % internal circuit. Equation 5.3 illustrates the power reduction of the now capped fans compared to them running uncapped at 100 %. Reducing the maximum fan speed by just 20 % reduces peak power consumption by almost half. The imposed fan speed limits guarantee that the cooling will draw less than 10 % of the maximum IT power, so 52.5 kW per IT-Container or 13.125 kW per cooling unit and the Data Centre remains below a PUE of 1.1. Having a stable control system with a steady fan speed is more efficient than constantly over-steering by increasing the fans more than necessary and reducing them afterwards, due to the non-linear power increase for faster fan speed. Table 6.1 shows the achieved PUE during the load tests in 2019. For all load scenarios above 50 % of the maximum capacity, the Data Centre remains below 1.1. During this tests the internal fans were set to a minimum speed of at least 60 %, which created a significant overhead for low load scenarios.

For an ideal fan, power consumption scales cubic with the rotational speed:

$$\text{Fan power reduction at 80 \% vs 100 \%} = (\frac{4}{5})^3 = \frac{64}{125} = 51.2\ \% \tag{5.3}$$

$$\text{Fan power reduction at 72 \% vs 100 \%} = (\frac{18}{25})^3 = \frac{5832}{15625} = 37.32\ \% \tag{5.4}$$

One of the advantages of the design is the closed internal air circuit, preventing any major air exchange with the outside and guaranteeing a good air quality, with low risk of contamination for example with pollen, dust etc. There is however an air intake, which is opened in a regular interval as soon as the lights are on. This is to allow safe working conditions at all times in compliance with safety rules. The incoming air for the fresh air intake is filtered by a fine grained F7 filter.

5.1.5 Water Treatment for the Cooling System

One of the potential problems with the evaporation of large amounts of water for cooling is scaling due to calcium, magnesium and other minerals solved in the water and left inside the system after evaporating the water. These scaling effects can significantly reduce the cooling capacity of the whole system by growing an isolating layer on the surface of the heat exchanger. To reduce this risk one can apply multiple protective measures. One of the options was to use softened water, reducing calcium and magnesium and therefore significantly reducing the scaling effects in the system. For the system it was decided that this is not sufficient and to built an RO water treatment plant, to get an even better water quality for the cooling system. The conductivity of the RO water is usually below 50 microsiemens and constantly monitored by the RO plant. The low conductivity is directly indicating a low amount of residual minerals and is therefore contributing to keep the cooling system clean and to prevent any scaling. Figure 5.10 shows the whole RO installation, including the activated carbon filter and softener pre-stages. In addition this reduces the potential biofilm significantly, since the nutrition value of the water for bacteria growth is extremely low. The incoming air is filtered before running through the heat exchanger preventing bigger particles to come into the system. The maximum production of RO water is up to 3.5 $\frac{m^3}{h}$, which is sufficient to cool up to 2 MW IT load during the warmest recorded Geneva weather conditions. Since currently less than 1 MW of the total cooling capacity is used, there remains a comfortable margin in this regard. The production of RO water is done in batches of approximately 300 L. After each batch the reject water is flushed to the drain and there is a reverse water flow through the membrane, to keep it clean and prevent clogging. To ensure a constant water flow towards the cooling units the installation has a buffer tank of 2 m^3, which is directly fed from the RO plant.

In case of a failure of the RO water treatment plant there is an automatic bypass from the softener stage to the buffer tank. This failure has to be detected as soon as possible, to prevent too much soft water being used. The amount of solved particles contained in the soft water can otherwise be problematic and lead to scaling effects due to the high amount of water evaporated. One cooling unit can evaporate 3 m^3 of water on a hot summer day, in case the Data Centre is operated close to the maximum load. Worst case calculations for extremely hot weather even get to around 4 m^3 water evaporated in a single cooling unit during one day. Water in the reservoir of each cooling unit gets flushed after 48 h, to replace the water with fresh RO water.

Figure 5.10: Water treatment installation in the infrastructure container. The leftmost blue bottle is an activated carbon filter, to remove chlorine. Next to it is a salt reservoir for the softener stage (two blue bottles), replacing calcium and magnesium with sodium ions to protect the reverse osmosis membrane. In the center of the picture there are six reverse osmosis tubes, filtering large particles from the water. Small buffer tank (blue) to the right provides clean reverse osmosis water to clean the membranes. Big black buffer tank to the right keeps approximately 2000 L reverse osmosis water for the cooling units.

5.2 Server and Compute Infrastructure

The overall design goal of the EPN cluster was to be as independent from external services as possible. This should ensure that the experiment can take data even if external services are temporarily not available. In addition the cluster should be as isolated from the campus network as possible, to avoid any interference from the outside. Critical services e.g. for name resolution, IP assignment and everything related to provisioning are running locally, as part of the cluster infrastructure. External access is possible and the local DNS forwards to the CERN nameservers, to resolve non-local addresses. A cluster local The Foreman [45] instance is used for bare metal Operating System installation. It also acts as an inventory of the hardware and is used to update DNS and DHCP. The advantage of a local setup is full control of everything. However it is significant work to keep the complete infrastructure up to date, so this comes at a cost.

So far all the monitoring and logging infrastructure is kept local to the cluster. Some metrics are forwarded to external databases in addition. It is foreseen to switch to central CERN databases instead, to save administrative work.

In practice ALICE is relying on EOS storage [12] for data taking, which is located in the CERN IT Data Centre at Meyrin. A smaller EOS instance was installed in the IT-Container at Point 2, as fail-over solution, in case the long range network connection to the IT Data Centre fails. This non-local dependency for storage also relies on the CERN DNS. Even though the system was designed to minimize external dependencies, running completely independent of cluster external services as the HLT in Run 2 [30] is not possible at the moment.

5.2.1 Run 3 EPN Computing Hardware

To find appropriate hardware for the ALICE computing needs multiple server prototypes were assessed. Table 5.2 shows the models which were taken into consideration and from which prototypes were successfully build. The pre selection was already focusing on servers supporting a maximum number of hardware accelerators, since a significant amount of the software was planned to run on GPUs. In the evolution of the O2 system [17], which was the base for prototyping efforts, the synchronous reconstruction needs are explained in detail. 93 % of the compute time is spend on TPC reconstruction, which was ported to run on GPUs and runs mostly with single precision float operations. Early estimations already expected that a single GPU can replace 40 CPU cores [17], leading to a significant cost reduction by utilizing GPU for 90 % of the required computation. As a result systems supporting a maximum amount of GPUs are favourable, to leverage potential cost benefits. For the final hardware a single AMD MI-50 replaces 80 CPU cores during online data taking and 55 CPU cores during offline reconstruction [42].

The Supermicro AS-4124GS-TNR [44] was integrated into the vertical slice test system at CERN Point 2, as part of the prototype verification process. This server has two 32 core AMD EPYC CPUs, 512 GB RAM and 8 AMD MI-50 [3] GPUs with 16 GB memory each.

Since the vertical slice test system already consisted of more than 100 old servers,

automation of the software installation and configuration process was required to keep the system in a consistent state. The prototype setup at CERN was done as part of this thesis, in particular system configuration and tuning. Automation was done via The Foreman [45] for provisioning the servers, a bare metal Operating System (OS) installation, finalizing the configuration with Ansible [28]. The prototype integration into the vertical slice system was one step in the configuration verification process. With the prototype integration, all further tests on this prototype were also testing the baseline cluster configuration was performing as expected. Minimizing manual interventions by automation was one of the priorities during planning and set-up of the Cluster. This was also helpful during the prototyping, since the desired server configuration could be recreated automatically.

In 2020 several problems running the software on the EPN server prototype were discovered, most of them were related to the GPU kernel driver and resulted in bug reports and several fixes by AMD. During this iterative process several tests were done with patched drivers and optimized configurations, to provide test environments within the production prototype system. The automated server setup with Ansible helped to get the system into a defined state, by applying patches, rebuilding the kernel drivers and as a last step recreating the Initramfs to apply the changed configuration at boot time when the updated kernel was loaded.

Table 5.2: Server models from successful EPN prototypes as presented during the PRR: [32]

Manufacturer	Model	Rack units (U)	Number of CPUs	Max. GPUs	PCIe gen
GigaByte	G482-Z51	4U	2	7	gen4
GigaByte	G242-Z10	2U	1	4	gen3
GigaByte	G292-Z40	2U	2	8	gen4, PCI bridges
GigaByte	G292-Z42	2U	2	8	gen3
ASUS	ESC4000A-E10	2U	1	4	gen4
Supermicro	AS-4124GS-TNR	4U	2	8	gen4

For the compute hardware review the latest up to date performance requirements were taken as the basis for judging size and feasibility of the compute concept. In particular reconstruction performance on the GPUs was of interest, to estimate the overall number of GPUs and therefore servers needed in the whole cluster.

Several GPUs were benchmarked by PDP colleagues. GPUs were one of the major cost drivers of the whole EPN farm. Table 5.3 shows models on which the software benchmark was running, to determine the performance and therefore the amount of GPUs needed

Table 5.3: Tested GPU models and relative performance in comparison to a Nvidia RTX 2080 TI, as presented during the PRR by David Rohr: [42]

Model	Performance
NVIDIA V100s	122.7 %
NVIDIA Quadro RTX 6000 (active)	105.8 %
NVIDIA RTX 2080 TI	100 %
NVIDIA Quadro RTX 6000 (passive)	96.1 %
NVIDIA RTX 2080	83.5 %
AMD Radeon 7	71.2 %
AMD MI50	67.8 %
NVIDIA GTX 1080	60.1 %
NVIDIA T4	59.3 %

for Run 3. Previous experiences with consumer GPUs in a server environment showed fan problems with actively cooled GPUs. The usual server airflow is different from a desktop PC. In addition the GPUs are usually placed right next to each other in multi-GPU installations, blocking the normal airflow to the side from actively cooled GPUs. In the past, a constant strong airflow perpendicular to the fan air direction caused massive fan failures and required the removal of active cooling components [41]. To prevent GPU modifications only passively cooled models were considered feasible for envisaged HPC environment, which ruled out most consumer models. In addition driver licences agreements can also prohibit using consumer GPUs in a Data Centre environment. Budget constraints required to get the optimal cost-performance ratio for the software requirements. Internal analysis of different GPU quotes showed that AMD MI-50 provided the cost optimal performance for us.

The estimate at this point was a minimum of 1570 AMD MI-50 was needed to fulfil the ALICE software computing requirements. This estimate was done under the assumption that the AMD MI-50 with 32 GB memory has the same performance as the AMD MI-50 with 16 GB memory, which was intensively tested in the prototype setup. The total number of servers was 250 amounting to 2000 GPUs, which should give us a sufficient margin for the runs [42]. The actual number of required AMD MI-50s increased due to three unforeseen effects. Fixes in the ROCm driver slightly degraded the performance and increased the number to roughly 1630 GPUs. With the 32 GB model, benchmarks from PDP colleagues showed a 14 % performance decrease in case the full 32 GB GPU memory is used. Increasing the minimum number of GPUs to 1840 (or 230 out of 250 EPNs). Larger TPC data added another 30 % to the GPU compute requirements.

All prototype servers were based on AMD Rome CPUs. The software stack has a large amount of processes and makes good use of a multi core processor. The advantages of the AMD compared to the INTEL CPUs were the amount of available cores per CPU, which leads to a better scalability of the ALICE software application. Most importantly the number of available PCIe lanes per server and therefore the number of PCIe x16

slots for GPUs. Another difference in favour for the AMD Rome CPU was the PCIe gen 4 capabilities. One of the main driving factors for the hardware decision were budget constraints. The possibility to integrate as many GPUs as possible in a single server was therefore a very welcome option, which was explored in detail. The advantage of highly integrated servers is that infrastructure can be saved on the networking side. It is possible to plan the core network with less ports, since the overall boxes to connect are significantly less in case of multiple GPUs per server. This also contributed to significant cost savings for the whole Run 3 IT installation. From the tested GPUs in table 5.3 the AMD MI-50 was providing the best performance per cost and was therefore chosen for the EPN servers. The 32 GB version with larger memory allows in principle to transfer bigger time-frames into the GPU memory. As for all components, the performance per price metric was one of the most important factors and here the AMD MI-50 was top as well.

The EPN server choice was the Supermicro AS-4124GS-TNR. It was the only server model at that time supporting up to 8 GPUs with 16 PCIe gen 4 lanes and additionally another x16 gen 4 for a networking interface, without using any PCIe bridge chips on the mainboard. In the hardware-software co-design the server fulfilled all the requirements and can be efficiently utilized to provide the required compute resources. From a performance to cost ratio this server provided the optimal solution for the requirements.

SuperMicro AS-4124GS-TNR

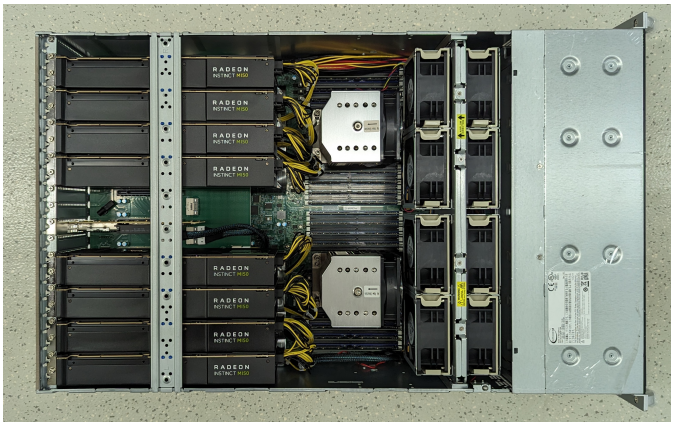


Figure 5.11: Picture of the EPN server hardware used to build the cluster. 8 AMD MI-50 GPUs, 2 AMD EPYC 7452 32 core CPUs, 512 GB memory (16x 32 GB), HDR-100 HCA.

At the time of prototyping, the server hardware the Supermicro AS-4124GS-TNR was chosen among the candidates listed in table 5.2 for the production system. Figure 5.11 shows an EPN server internals. It is a dual socket server with two AMD EPIC Rome CPUs with 32 cores each, providing the required CPU based compute for the part of the software which is not running on one of the eight AMD MI50 GPU. Each GPU has

32 GB memory, allowing us to copy full Time Frame with 128 Heart Beat Frame onto the GPU. All EPN have 512 GB main memory, 16 modules with 32 GB each, to utilize the available memory bandwidth of the 8 memory channels per AMD EPYC CPUs. Since only one RAM module per channel is installed doubling the memory is easily possible. An InfiniBand HDR-100 HCA provides the required network bandwidth to receive the data and ship it to storage after processing. At the time the decision was made the Supermicro AS-4124GS-TNR was the only server supporting 8 GPU connected with 16 PCIe Gen 4 lanes each plus an additional PCI x16 connection for a network interface.

5.2.2 General Cluster Setup

To get the software running in a performant way it was required to invest time to tune the configuration of the operating system, in particular the kernel modules.

Memory allocation was one of the focal points in the configuration optimization. The servers have 512 GB memory and the software has to be able to allocate almost all of the available RAM. In particular a large fraction of memory has to be accessible by the GPUs. This was not possible with the default OS and driver configuration. For this to work it was necessary to set the user limit for memory allocation (memlock ulimit) to unlimited, so users can in principle allocate all memory, at least from the user limits configuration.

For the GPU kernel drivers it was required to increase the Graphics Translation Table via the amdgpu gttsize parameter to 524280, allowing the memory mapping for the GPU to span the entire available RAM. The default setting only allowed 256 GB GPU memory to be mapped, which was not sufficient to run the software with 8 GPUs in parallel. In addition it was needed to enable the kernel drivers system memory limit via the amdgpu no_system_mem_limit option and increase Translation Table Manager page limit via amdttm pages_limit to 1024000000. These settings have to be set during the initial loading of the kernel drivers in the Initramfs and therefore need to be part of the image, which is loaded at boot time. This means that the image has to be rebuild every time the configuration is changed.

To avoid over subscription of the available RAM and to protect operating system processes to be killed by the out of memory killer (oom-killer), cgroup limits were set to 490 GB, leaving 22 GB to the OS and background daemons. The rather generous 22 GB memory for the operating system seemed required to prevent occasional termination of vital system processes by the oom-killer, which then required a restart of the whole system to get back into a good operational state. This limit ensures only user processes are affected by the oom-killer and guarantees the system health of the worker nodes. On top of this cgroup protection there is another limit set by the Slurm scheduler limiting allocated Slurm resources.

An important feature for debugging is to enable core-dumps, to provide debugging information in case the software crashes. One potential problem is the size of core-dumps due to the large memory and processes which can be several Gigabytes. A sequence of crashes can therefore quickly fill up the disks and it was better to limit the total size available for core-dumps to prevent this issue.

To have the time on all cluster nodes decently synchronized the chrony daemon is used to query the time from CERN stratum 2 time servers, which are available via the general purpose network. All the cluster nodes are therefore in the stratum 3 domain. This gives us a sufficiently precise time reference for the needs.

For user access the centrally CERN-hosted Lightweight Directory Access Protocol is configured, which provides CERN user and e-group membership information. The group memberships determine to which nodes the users have access to. Using a centrally managed user database has many advantages, in particular since there is a number groups and persons which need different level of access, e.g. to investigate logs or to perform tests. Dedicated local users configured on the servers exist as special purpose accounts or system users running dedicated daemons/services.

Mostly for user convenience a shared NFS directory for the user home directories as well as the dedicated cluster toolkit and scratch space is configured on each server. This prevents local copies of commonly used scripts and can improve user experience. The NFS server is one of the local services of the cluster, which is only available inside the farm. CERN has developed Cern VM File System, which provides an NFS like shared file system. The ALICE CVMFS directory is mounted on all the nodes. This directory is mostly used by the offline analysis and contains the latest ALICE software.

5.3 Network

In the O2 system the design of the whole data-flow is done in a way that data is only send once and is fully self contained. This means that a fully assembled TF contains all the required data and no further communication between EPNs is needed to share additional information. In figure 5.12 the data-flow is from the left (detectors) to the right (Data Storage). This clear direction makes it easy to get the aggregate data rates from the connected servers and simplifies the network design.

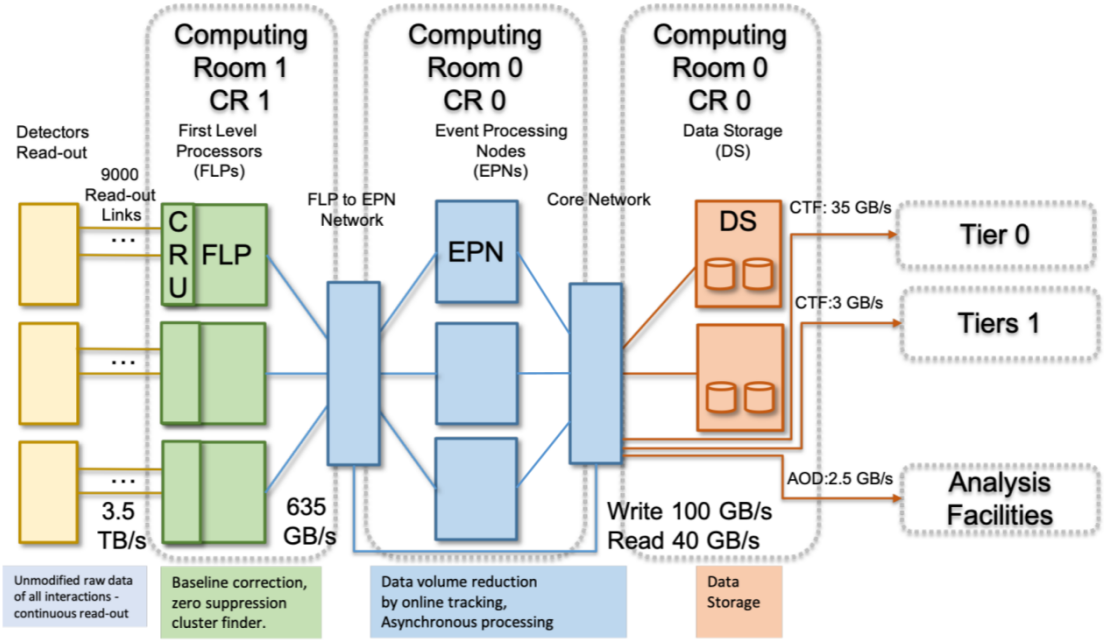


Figure 5.12: Schematic O2 Data flow as described in "Evolution of the O2 system" [17].

All EPNs are equal and failing hardware can easily be compensated by excluding the server and all remaining EPNs receiving more data. As long as the overall computing capacity is sufficient the system will continue working normally. The Run 2 HLT as the predecessor cluster had dedicated input and output nodes. Input nodes were connected to specific detector links, which meant that any failure of one input node had a direct impact on data taking and all input nodes were required at all times. The Run 3 schema is far more resilient and single hardware failures have no immediate impact on runs.

5.3.1 FLP to EPN network connectivity

Table 3.3 shows the evolution of the Run 3 data rate estimates. TPC as the main contributor injects 90 % of the overall data volume, while other detectors have lower data rates compared to the TPC. This is also resulting in a high amount of detector links and therefore a high number of FLPs. From the network perspective this means

that the FLPs reading out TPC need more bandwidth towards EPN than other detectors. To keep the network as simple as possible FLPs from a single detector are all connected to the same Top of the Rack (ToR). For TPC there are 144 input servers, which are equally distributed to four ToRs.

In total there are 72 links between CR1 and CR0, with a total capacity of 14.4 TBit/s or 1.8 TB/s. This raw bandwidth is the theoretical maximum throughput between the two clusters. 60 of these links are connected to the TPC ToRs, providing 12 $\frac{Tbit}{s}$ = 1.5 $\frac{TByte}{s}$ theoretical throughput. The installed line rate is therefore roughly three times the specified required throughput of 635 GB/s. The amount of network links between FLPs and EPNs were optimized for costs, while fulfilling the requirements and providing contingency.

The data-flow is designed to drop data, in case buffers are full and data is not sent out after a short time. Getting close to the limit input servers with a higher load would inevitably drop data and would lead to incomplete TFs. During the tests with the TPC detector it was possible to achieve up to 1.2 TB/s without any loss of data, which corresponds to 80 % of the theoretical maximum throughput and is more than a factor two of the initially required design value for the TPC of 570 GB/s [17]. This gives evidence that the whole network tailored to the specific needs of the experiment is set-up in an optimal way.

5.3.2 Network Topology

ALICE requires a fast network connectivity, which can reliably provide the necessary bandwidth between the FLP and EPN. InfiniBand provides several useful features, which are helpful in the overall system design. InfiniBand networks combine high bandwidth with low latency. In particular the option to write directly into the host memory of the target server via RDMA is a very useful tool, to avoid the CPU overhead of TCP streams. Data Distribution utilizes RDMA to transfer the detector data in the form of STFs from the FLP memory directly to the memory of the EPNs. A credit based flow control helps with congestion management of the FLP to EPN many-to-one transmission schema.

Core of the network is a backbone consisting of five InfiniBand QM8700 managed switches with 40 ports each. Figure 5.13 shows an schematic overview of the data network. The network layout opted for a very flat network topology with only two layers for the main part of the data network. The number of core switches has significant impact on overall costs. For the ALICE case, five core switches provide an optimal balance between costs and flexibility for further upgrades, with free ports to connect additional switches.

The O2 system is split into multiply parts, which are installed at different locations. The FLPs in CR1 are located less than 100 m apart from the EPNs in CR0. This still allows us to use short range fibre connections via patch panels, which have a range limit of 100 m with OM4 fibres. This is an important cost factor, since long range transceivers are significantly more expensive than short range ones. The network core is located in CR0, close to the incoming trunks, to minimize the length of the overall connection and

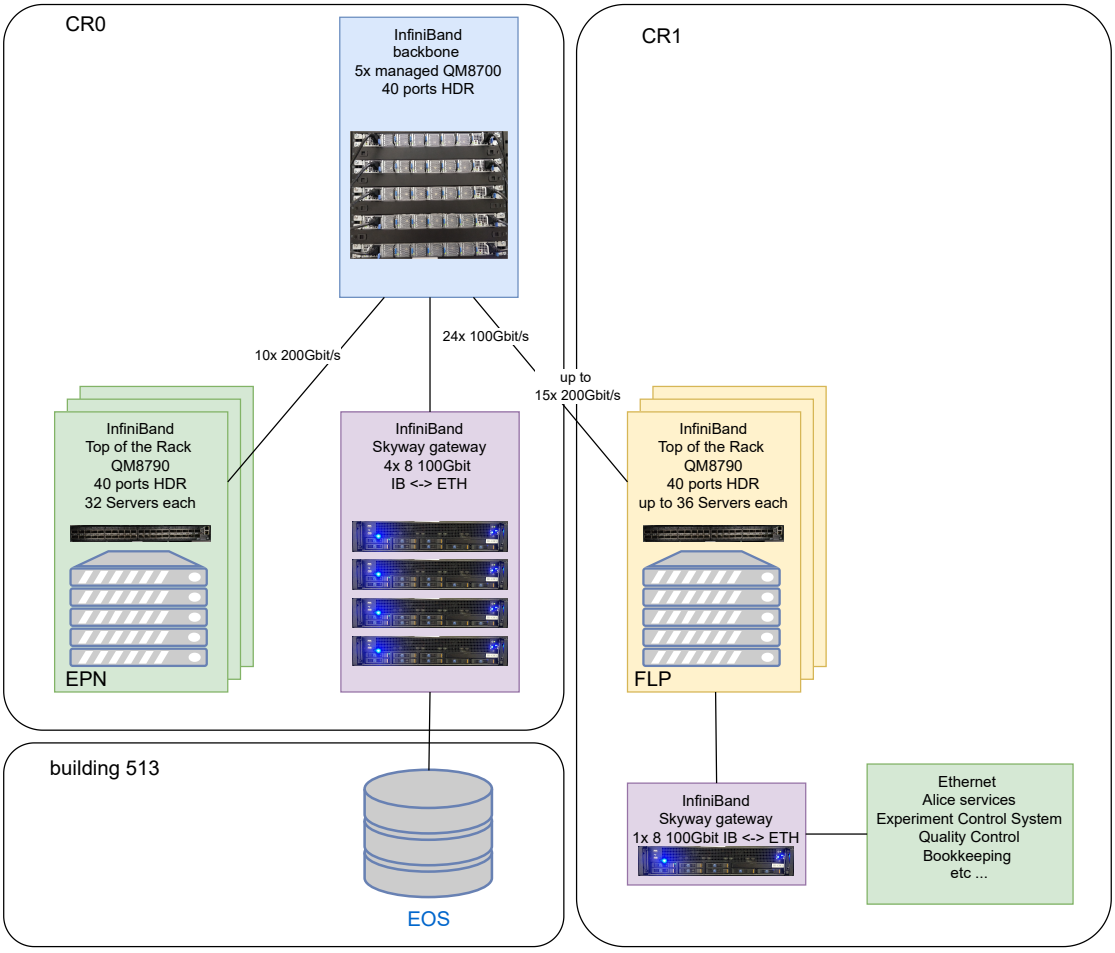


Figure 5.13: Initial data network connectivity schema on the building block level.

ensure to remain within the 100 m limit. EPNs are much closer to the core switches, even when connecting the more distant IT-Container IT4 the overall length is around 50 m or less, which gives us margin with the short range links. All FLP and EPN ToRs are directly connected to the core network.

$$\text{switch blocking factor} = \frac{\text{total bandwidth of connected servers}}{\text{total uplink bandwidth to backbone}} \tag{5.5}$$

$$\text{blocking factor TPC switch} = \frac{36 * 100\ Gbit/s}{15 * 200\ Gbit/s} = \frac{3.6T}{3T} = 1.2 : 1 \tag{5.6}$$

$$\begin{aligned}\text{blocking factor ITS, MFT, TOF, FIT, EMC switch} \\ = \frac{27 * 100\ Gbit/s}{6 * 200\ Gbit/s} = \frac{2.7T}{1.2T} = 2.25 : 1\end{aligned} \tag{5.7}$$

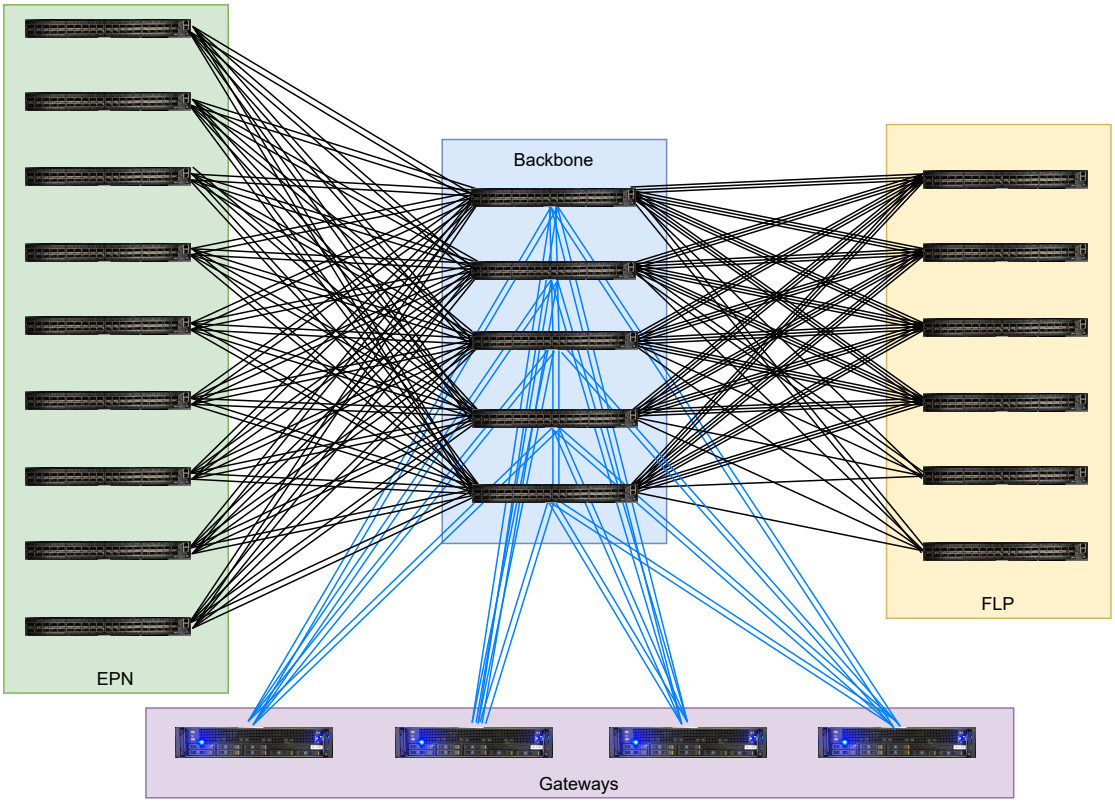


Figure 5.14: Switch interconnectivity of the InfiniBand data network. Blue connections 100 Git/s, black connections 200 Gbit/s. The figure is showing the current status at the time of writing with a 9th building block added in the 280 EPN configuration and a 4th IB to ETH gateway. Another building block will be added for the next extension of the cluster to 350 EPNs.

$$\begin{aligned}\text{blocking factor TRD, MCH, MID, PHS, HMP, CPV, CTP switch} \\ = \frac{31 * 100\ Gbit/s}{5 * 200\ Gbit/s} = \frac{3.1T}{1T} = 3.1:1\end{aligned} \tag{5.8}$$

On the FLP side there are 6 building blocks each with one InfiniBand QM8790 switch, connecting in total roughly 200 FLPs to the InfiniBand network. The majority (approximately 90 %) of the data is coming from the 144 TPC FLPs. 4 of the 6 ToR switches are dedicated to the TPC FLPs connectivity, using 60 of the 72 total uplinks to the InfiniBand backbone. This means each of the 4 ToR switches connecting TPC FLPs has 15 uplinks witch 200 Gbit/s and a theoretical maximum throughput of 3 Tbit/s towards the backbone. The remaining 12 uplinks are equally distributed between the two remaining ToR switches, connecting all other detectors. The total bandwidth between FLP and EPN is 14.4 Tbit/s. This design was chosen to implement a cost efficient network with respect to the required network throughput as defined in the evolution of the O2 system [17].

On the EPN side there are 8 building blocks connected to the backbone. Each building block is connected with 10 200 Gbit/s links to the core, two links to each of the 5 core switches. The total throughput from the backbone to each ToR is 2 Tbit/s, summing up to 16 Tbit/s from the backbone to the EPNs. Figure 5.14 shows all links connected to the core switches.

Network Connectivity to the EOS Storage

The storage called EOS in figure 5.13 is not located close to the experiment at Point 2 but in another Data Centre, managed by Cern IT colleagues, at Meyrin. This requires a roughly 6 km connection with enough throughput to push 100 GB/s to disk. EOS is also not part of the InfiniBand network, the disk servers are connected via Ethernet. This requires to bridge the InfiniBand domain towards Ethernet with gateway switches translating between InfiniBand and Ethernet. These gateway switches are directly connected to the core InfiniBand switches as well as the Ethernet routers routing to the other Data Centre. Getting acceptable performance between the two different networks via the gateway switches was one of the challenges, which had to be resolved to achieve the required storage performance.

5.3.3 EPN building block

Figure 5.15 shows the building bocks of three racks. In the current configuration each building block supports 35 servers, but currently only 32 servers are installed. Each Top of the Rack (ToR) switch is connected via 10 200 Gbit/s links to the InfiniBand core switches. Each of the servers is connected with 100 Gbit/s. The use of splitter cables, which split one 200 Gbit/s switch port into two 100 Gbit/s server ports where possible, reduces the number of switch ports needed and therefore is one of the cost optimisation measures that were implemented. A more optimized solution with the switches in the middle of the racks was not easily possible. Since HDR splitter cables have a maximum length of 2 m and the racks forced routing the cables around the top of the rack, it was required to use 3 m EDR cables to connect servers installed in the lower U of the racks. With 3 m EDR one switch port per server is used, which is the trade-off in this schema. The layout optimized for the maximum amount of splitter cables possible, to reduce the number of needed ports.

$$\frac{\text{EPN aggregated bandwidth}}{\text{switch to core aggregated bandwidth}} = \frac{3.2T}{2T} = 1.6 : 1 \tag{5.9}$$

The blocking factor, as defined in equation 5.5, of an EPN ToR is less than 2. With respect to the data-flow the blocking factor of the EPN ToRs is not too important, since the processing time is longer than the network transfer of the incoming data. Since the data-flow is directed the relation between total bandwidth of the FLPs and EPNs towards all core switches is more important. With the initial design there was roughly 10 % more output bandwidth from the core network switches to the EPNs available

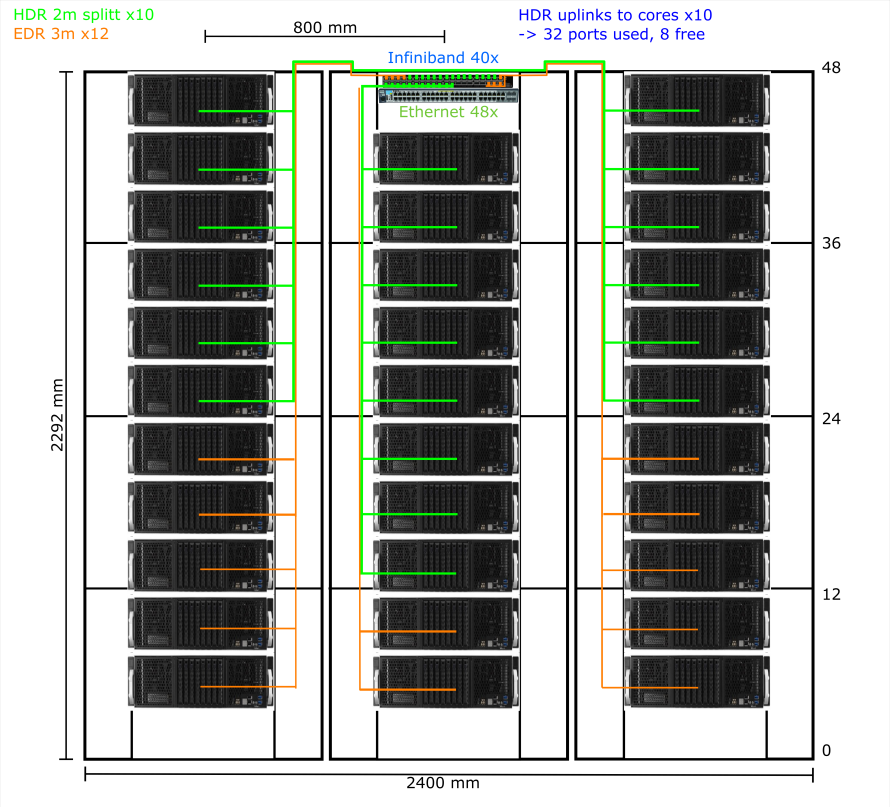


Figure 5.15: Building block for network connectivity, including detailed cabling

than input bandwidth from the FLPs. This already guarantees the non-blocking data path from the input stage to the processing stage of the system. Equation 5.10 shows this relation for the initially planned 8 building blocks, the upgraded 9 and planned 10 building block setup. Each extension giving more bandwidth overhead to distribute between the processing nodes.

$$\frac{\text{FLP switches aggregated bandwidth}}{\text{EPN switches aggregated bandwidth}} \tag{5.10}$$

$$\text{8 EPN building blocks: } \frac{14.2T}{16T} = 0.9 : 1 \tag{5.11}$$

$$\text{9 EPN building blocks: } \frac{14.2T}{18T} = 0.79 : 1 \tag{5.12}$$

$$\text{10 EPN building blocks: } \frac{14.2T}{20T} = 0.71 : 1 \tag{5.13}$$

There is still the option to install three additional servers per building block in the lowest 4 U of the racks. U 1-4 are still possible to connect with 3 m cables to the Top of the Rack switch, however there is no more margin in that case and the cable route is rather tight, which is the reason the lowest four U were unused in the first server

installation. Since the core switches don't have many free ports left, to add additional building blocks, it was decided to fill up the racks with the upcoming upgrade. This way two free ports per core switch remain and leave flexibility for the future.

Due to increased O2 compute requirements, cluster extension options were investigated. The system is designed to scale well and to give future flexibility. As long as there are free network ports on the core switches it is easy to add additional building blocks. In the end of 2022 30 additional servers were installed in a 9th building block, to increase the available compute power and enable us to cope with a slightly higher number of TPC clusters. The additional servers were integrated seamlessly into the existing installation. The scaling of the compute requirements is linear with respect to the amount of clusters in the data. A roughly ten percent increase in available servers therefore enables us to cope with ten percent more clusters. The overall effects in the real data were significantly more than 10 %, so a discussion started to buy additional 70 servers, to increase the compute resources even further. The overall size of the EPNs farm therefore increased from initially 250 servers to 280 and then 350 total. In relative terms this corresponds to a roughly 50 % increase in compute power, to cope with the unexpected increases of the data volume and amount of clusters as well as to regain margin, to get flexibility back.

5.3.4 Network Tuning

One of the central parts of the InfiniBand network is the subnet manger. The subnet manager is the heart of the network, registering all connected devices and assigning LIDs. The LIDs are used to build the routing table of the network and keep track of which device is connected where.

Traditionally the traffic is routed via one route between two LIDs. In the case of multiple links e.g. between switches a dedicated path is chosen by the subnet manager. This can lead to imbalances of the inter switch link utilization. One of the features of the subnet manager is adaptive routing for Connect-X 6 devices (HDR), which was enabled in the network. Adaptive routing tries to balance the traffic between all inter switch links equally, distributing the traffic towards the less utilized link on the fly. Adaptive routing adds more latency, which is not critical for our application. In this case it led to a better utilization of all the inter switch links and increased the overall throughput between FLP and EPN in the network.

5.3.5 Gateway Switches

A significant part of the network traffic has to cross from the internal InfiniBand network to the Ethernet domain of the CERN IT Data Centre in Meyrin, where the disc storage is located. The data rate to disk is approximately 100 GB/s for Pb-Pb data taking. There are multiple options to bring the traffic towards the Ethernet domain. One would be to have servers with InfiniBand as well as Ethernet interfaces and let the servers act as gateways between the two networks. To get the required throughput several machines would be required and the configuration, especially to balance between multiple

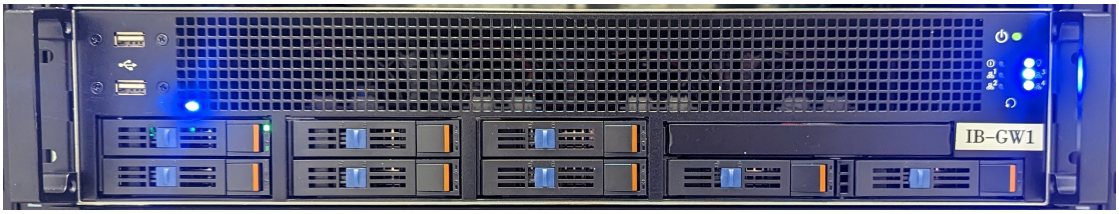


Figure 5.16: 2 U Skyway InfiniBand to Ethernet gateway with theoretical throughput up to 800 Gbit/s [39].

interfaces can be tricky. The choice was made to get a new gateway switch from Mellanox / NVIDIA called Skyway. These appliances are similar to a server housing 8 dual port HCAs, with one port configured as InfiniBand and the other port configured as Ethernet. The appliance translates all incoming traffic on one of the ports to the other network and sends it out via the second port. The configuration as a switch allows an easy set-up of a Link Aggregation Group (LAG), combining all links of the gateway to a single logical link on the InfiniBand as well as one on the Ethernet side. Furthermore Skyway gateways provide an easy configuration option for a High Availability (HA) set-up, which provides a variety of advantages. The HA configuration is a multi chassis LAG, combining links of a single appliance to a Link Aggregation Group of multiple gateways, including hot fail-over in case one appliance is temporarily not available. This offers the option to have a single static route configured on all servers, pointing to the IP of the LAG as a next hop for all traffic towards the other domain. The multi chassis LAG of the Skyway HA cluster has itself a single next hop as the default route configured to the Ethernet domain. In general the routing configuration can be simpler if there is only one dedicated next hop to configure. Equal Cost Multiple Path (ECMP) configuration was also considered, which would utilize multiple Ethernet routers and would have a build-in redundancy via the redundant routes. With ECMP multiple next hops are configured with the same priority. A local decision on the router is made, to determine which of the routes a package gets forwarded to. The traffic gets load balanced over all existing ECMP routes. The configuration for ECMP was not implemented in time and this functionality could not be utilized. This means that the configuration options on the Ethernet side were limited to have a Multi-Chassis Link Aggregation Group or a big router, which can support the required large LAG over 24 or 32 links in case of 3 or 4 skyways. In this case CERN IT decided to go with a large router, supporting the required LAG without going into a multi chassis configuration, since the multi chassis configurations tend to depend on the vendor and can be slightly different depending on the hardware.

5.3.6 Gateway Link Balancing

One of the challenges for a high throughput network environment between multiple networks is the traffic balancing. There are widely used hashing algorithms, which can map sessions to a specific link of a Link Aggregation Group. The better traffic is balanced

across all available links of the LAG, the closer to the theoretical link limit one can go without experiencing problems like delays in the transmission or package drops due to buffer limitations.

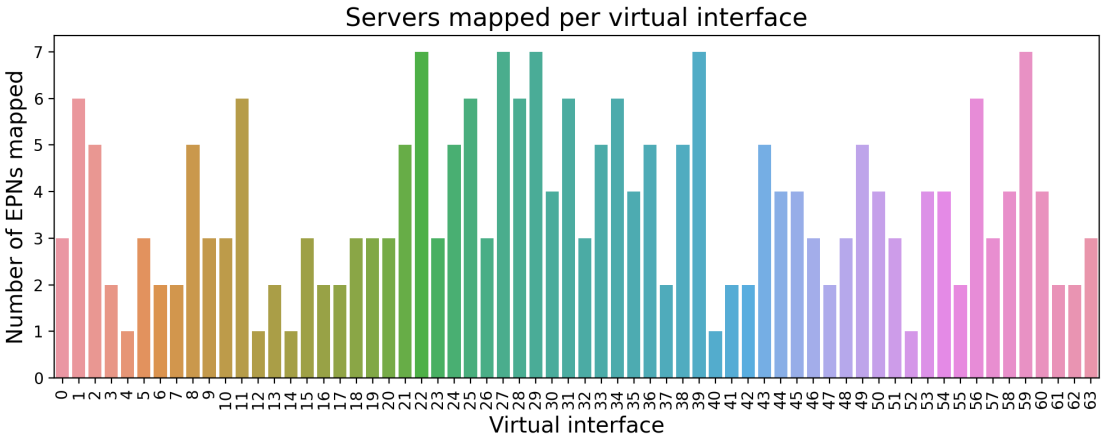


Figure 5.17: Number of EPNs assigned to one of the 64 virtual interfaces (firmware v8.1.2000, default hashing).

In this case the load balancing via the gateways got significant attention because of observed throughput limitations. There are two levels of mapping involved in the way the gateways distribute traffic between links. This is slightly different from a traditional hashing over a LAG, e.g. selecting to hash over the IP and MAC address of the source as well as the destination. During boot time the Skyway creates 64 virtual interfaces, which are distributed equally over the available links and used to balance the traffic. The exact number of virtual interfaces created changed in the past and depends on the firmware used. In the current firmware version 8.1.5002 64 virtual interfaces are initialized. In previous firmware versions up to 256 virtual interfaces were created. The creation on the gateway takes significant time and introduces a delay on the start-up of minutes. This seems to scale linearly, with 64 virtual interfaces it's around 4 minutes, with 256 roughly 15 minutes. The gateway assigns LIDs of the current InfiniBand subnet to one of the virtual interfaces, hashing over the interface IP. When the server sends an ARP request to the IP of the LAG only the gateway interface responsible for the virtual interface the HCA is mapped to answer the request. Traffic from a specific HCA is therefore always routed via the same gateway link. Since the servers only have one InfiniBand HCA this means the whole traffic to storage from a single server is always taking the same gateway link. The balancing on that level is more course grained than with other hashing algorithms also factoring in multiple properties.

Figure 5.17 visualizes the hashing results and shows the mapping of EPNs to the virtual interfaces created on the gateways. It clearly visualises the potential issue with a random bases mapping of just one property. Some of the virtual interfaces have more LIDs assigned than others and for 64 virtual interfaces the deviation from the mean can be large. In case the traffic is equally distributed between all servers and each server is

sending roughly the same amount of data, the difference in the number of LIDs assigned to a virtual interface inevitably leads to an imbalance in the link utilization.

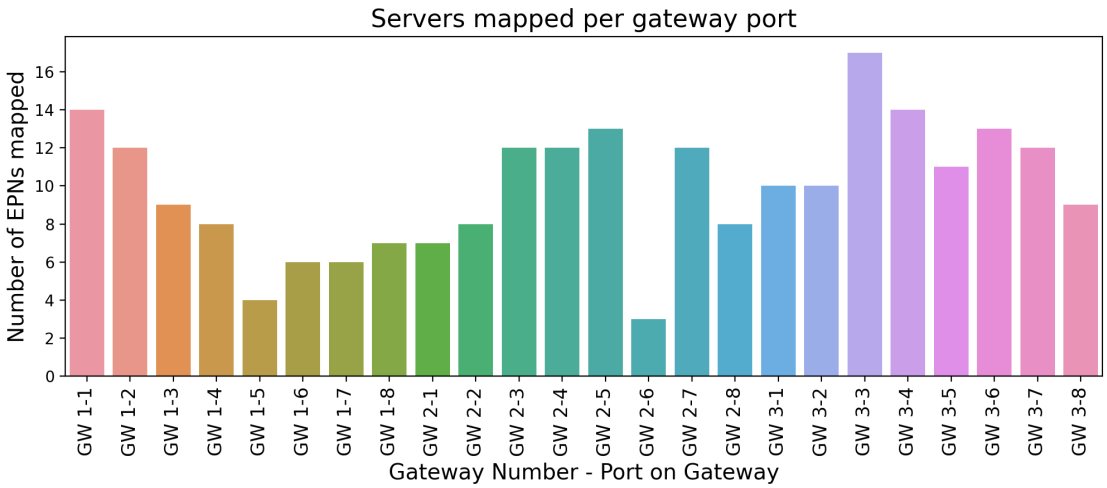


Figure 5.18: Number of EPNs sending traffic via one of the 24 specific ports of the gateway HA cluster (firmware v8.1.2000, default hashing).

The second mapping stage is the distribution of the virtual interfaces to the actual physical links. Each physical link is responsible for a certain set of virtual interfaces. In case one link goes down the gateway can simply distribute the virtual interfaces of the unavailable link to the other ports, which are still active. This grants flexibility in the HA set-up, to always ensure that the traffic can be routed in case of any failure of the physical links. Figure 5.18 shows how many EPNs are mapped to a specific link. Since the data is equally distributed across all EPNs this imbalanced mapping over the physical links inevitably leads to an unbalanced traffic distribution.

The Skyway firmware creates 64 virtual interfaces and maps them equally to the physical links. This is a large problem in the set-up with three gateways and contributes to significant imbalances. In this specific case and three active gateways, with 8 links each, so a total of 24 100 Gbit/s links from InfiniBand to Ethernet are part of the LAG. Since 64 is not a multiple of 24 it is impossible to achieve an equal distribution of virtual interfaces to physical ports. In this set-up one physical link is therefore responsible to handle two or three virtual interfaces. The low amount of virtual interfaces distributed across the physical links therefore leads to a significant difference in the mapping, which makes an equal link utilization across all gateways and all links extremely difficult. A potential solution would be to configure the number of virtual interfaces to be a multiple of the number of physical links, which is currently not possible. Figure 5.18 shows the resulting mapping, the number of EPNs sending traffic over the 24 gateway links. The imbalance is caused by the random assignment of the LIDs to the virtual interfaces, as well as the fact that either two or three virtual interfaces are mapped to one physical link. The link with the lowest number of EPNs (3) and the highest (16) differ by a factor of > 5. This makes it impossible to utilize all gateway links close to line-rate

and therefore reduces the overall available bandwidth to EOS storage. A major issue with the oversubscribed links is that package drop occurs when throughput is getting closer to the line rate. Retransmission then amplifies the problem and can easily create a situation where the data from a subset of servers is not transmitted fast enough and is buffered locally on disk. This introduces additional problems. Since the mapping is static the same EPNs are affected and therefore filling the disk. If this continues long enough to fill the disk completely the EPNs are no longer usable for processing until space is freed. The draining is usually extremely slow, since the overloading of links persists even after the run already stopped. In the tests, draining of the disks can take as long as the actual run lasted, in certain situations even longer than the data-taking time. This is not compatible with continuous data taking over long periods of stable beams, since there are usually only short breaks between data-taking.

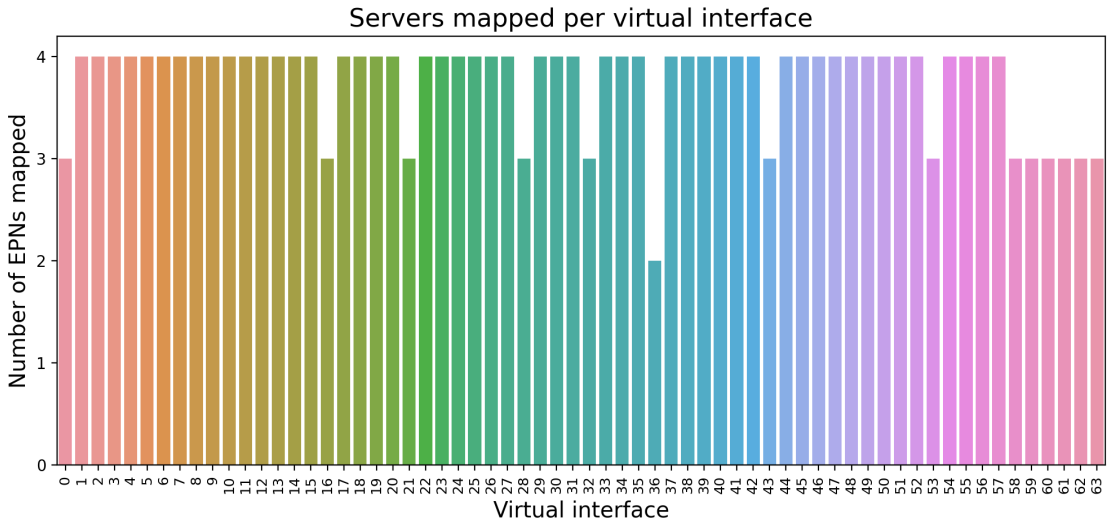


Figure 5.19: Number of EPNs mapped to each of the 64 virtual interfaces with firmware version 8.1.3050 with modulo hashing.

The benefit of hashing algorithms which hash over multiple properties is that they often take advantage of the number of streams per server, e.g. routing different streams via different links depending on their destination. It therefore distributes the traffic from a single server via multiple links, which can result in a much more fine grained level of balancing, which scales well with the number of involved servers on both sides.

In this particular use case a strict round robin distribution of servers and therefore IPs to the virtual interfaces would lead to a more balanced distribution of the number of HCAs handled by each virtual interface, see figure 5.19. This could in principle increase the overall balancing and therefore the link utilization as lead to a higher total throughput. The default hashing algorithm is optimized for very large and inhomogeneous systems and would work well if we have significantly more EPNs.

To overcome this issue different options were feasible. The first one would be to reconfigure the servers and assign the IPs in a way that the hashing algorithm, which

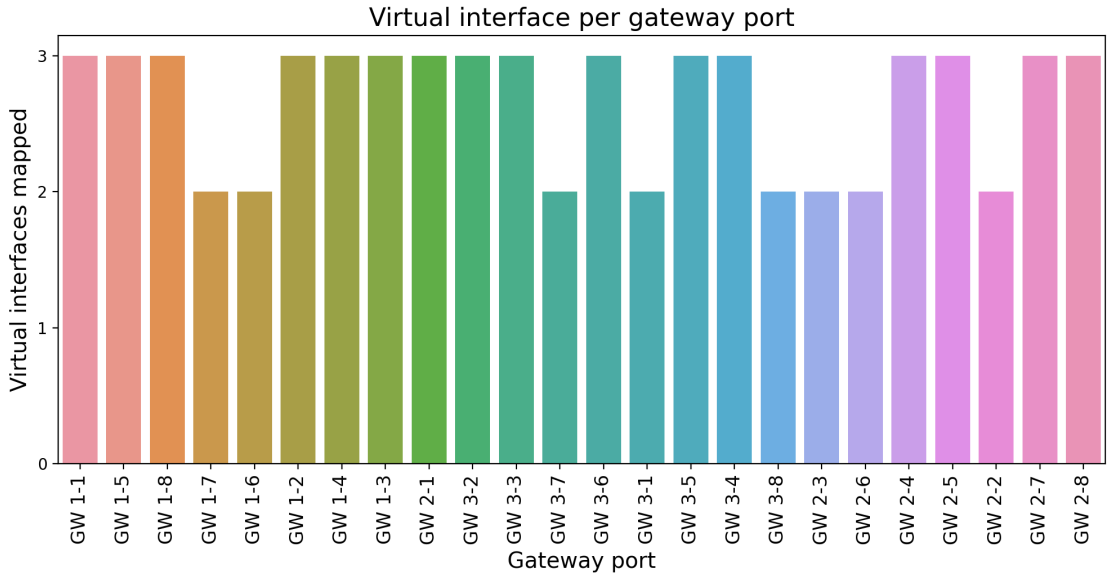


Figure 5.20: Distribution of the 64 virtual interfaces over the 24 gateway ports with firmware version 8.1.3050 with modulo hashing.

was not known exactly, map the servers equally to the available ports. In this way it would be possible to get an equal distribution of traffic via all links. This seems to be a rather tedious way to improve the throughput towards storage and implies also effort toward other services, to adjust the names and IPs in various places. The feature to use InfiniBand Pkeys was introduced with a firmware update of the Skyway gateways. Pkeys allows us to use an additional virtual layer on the InfiniBands interfaces and assign multiple IPs to the network card. This then allows us to keep the current IP configuration for the internal data transport inside the InfiniBand network between FLPs and EPNs, which is a great benefit since it would not require to adjust that part. The virtual layer on top of that would only be used for the transfer toward Ethernet. The IPs of that virtual layer need to be chosen in a way that the hashing algorithm distributes the servers equally over all links. With virtualisation it would be possible reduce the overall impact on the existing set-up and on top have an isolation of the different traffic streams. A significant caveat in this regard is the modification of the network configuration, which requires significant testing and potential downtime. Since the balancing problems got most apparent only after significant testing with real data taking conditions, it was too late to perform any major network modification. Since ALICE aims to maximize data taking during LHC collisions any long downtime of the network was not acceptable.

The second option discussed was a improved hashing algorithm, to provide a good distribution of the existing IPs over all links. This option does not require any modifications on the cluster side, but development work on the manufacturer side. One significant caveat of this solution is that it was relying on a very special firmware version and can not easily upgrade with the official firmware release. Significant support from the manufacturer with the custom hashing, combined with the lack of time to modify the

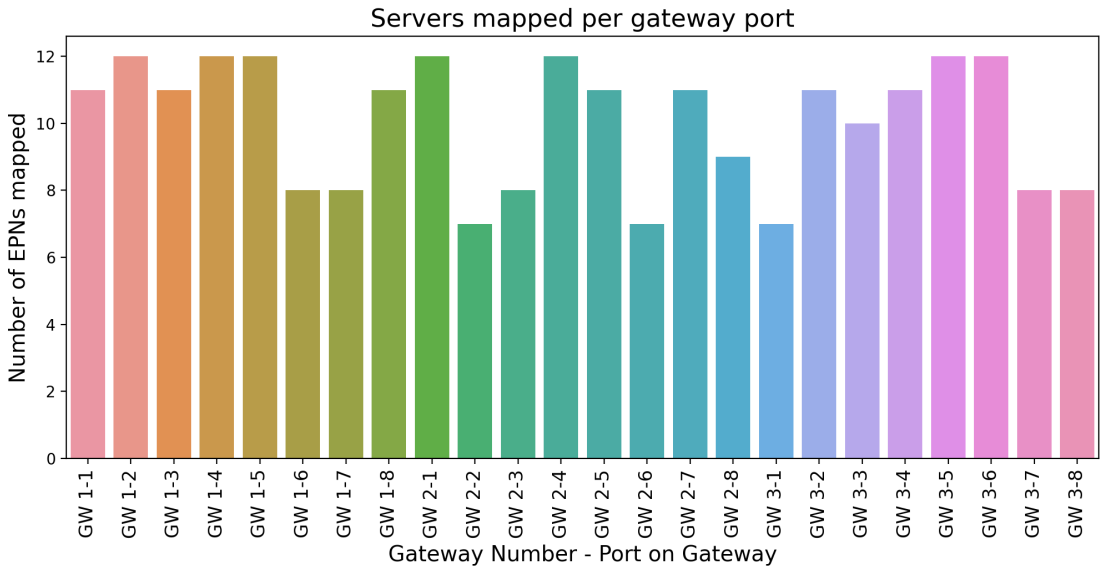


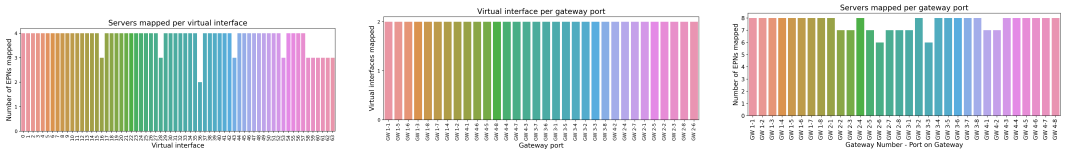
Figure 5.21: Distribution of the EPNs over all 24 gateway ports with firmware version 8.1.3050 with modulo hashing.

network configuration during LHC beam time, the custom hashing approach was chosen. By now the feature made it to the upstream firmware and it is possible to choose between the default hashing algorithm, which has a rather poor balancing performance in the particular use case and the modulo policy, which interprets the source IP as integer and performs a modulo operation for the mapping. The modulo policy is a round robin distribution of the servers to the physical links of the gateways, since the servers have a continuous IP range counting upwards. The very first overview was done with 229 out of 250 servers, leading to a slight gaps in the distribution. Figure 5.19 shows the distribution of the EPN IPs between all virtual interfaces. The few interfaces with only 2 EPNs mapped were due to the missing EPNs, the majority was between 3 and 4 IPs per virtual interface and therefore close to the perfect distribution. The improved hashing therefore solved the IP mapping imbalance.

Figure 5.21 shows the new distribution of all available EPNs during the test across all gateway links. The new hashing improved the balancing situation significantly. The expected congestion over the gateway ports should be greatly reduced. The tests pushing data can be found in the benchmark section 6.3.

Balancing in a Four Gateway Setup

Figure 5.21 illustrates the problem of the planned three gateway set-up. The fixed number of 64 virtual interfaces, which is not configurable, in combination with using 24 ports across all three gateways. Even though the balancing of all EPNs across the virtual interfaces is now perfect with the modulo mapping as shown in 5.19, the number of EPNs sending via a specific port still varies, since the ports responsible for three virtual



(a) EPN mapping to the 64 virtual interfaces

(b) Virtual interfaces mapped to gateway ports

(c) Number of EPNs per gateway link

Figure 5.22: 4 gateway distributions with modulo hashing

interfaces will inevitably have more servers mapped to the port than those with only two virtual interfaces. Due to the lack of configuration options, the only way to resolve this issue is to operate with a port number equal to any divisor of 64. This is the case for 1, 2, 4 and 8 gateway set-ups. In this case the HA cluster was expanded and added a 4th gateway to the system, to overcome the imbalances for good. Figure 5.22 shows that the mapping of EPNs to the virtual interfaces does not change, as expected the fixed hashing policy maps the IPs deterministically to each virtual interface. The number of IPs and therefore EPNs mapped to each gateway port changed now, since the number of virtual interfaces per port can be perfectly distributed and each port has two virtual interfaces assigned. Having four gateways instead of three increases the theoretical bandwidth by 1/3. The good balancing in addition gives a better link utilization and prevents link over utilization even further. The tests described in 6.3 showed that the current setup can achieve maximum EOS write speeds with this improved set-up and are therefore no longer limited by the network transfer speed but by the write speed to disk, which is around 200 GB/s and therefore twice the initially envisaged storage rate of up to 100 GB/s.

6 Results and Benchmarks

During the whole planning and procurement process continuous checking and testing of critical systems and infrastructure was done, to ensure the required performance for Run 3 operations. In this Chapter the verification process and the results are described in detail, highlighting a significant increase compared to the predecessor, the Run 2 HLT, as outlined in Chapter 5. This is topically split into the Data Centre performance, in particular the cooling system, the server performance and the network.

The Data Centre infrastructure was one of the focus areas early on, to ensure a good operational environment for the IT equipment. One of the concerns regarding the cooling was to potentially have hot spots and cause problems for multi-GPU IT equipment. This led to extensive testing and continuous improvements until the results of the steering of the cooling system was satisfactory.

During the whole software development the performance on the target hardware was benchmarked, to ensure the processing is fast enough for the real-time requirements. There was a constant benchmarking of algorithm execution time for simulated data, to ensure the expected compute requirements are met and Pb-Pb data can be processed at the foreseen rates.

Network performance got a lot of attention and there were continuous efforts to improve the performance and throughput, to ensure the expected data rates can be handled and to have some margins to handle unexpected spikes. In particular the crossing of different technologies was a challenge, to get data from the InfiniBand network into the Ethernet domain.

The infrastructure planning was based on educated guesses, often extrapolating from the Run 2 experiences. Verification of these estimates and continuous adoption in case of updates was required, to get a working system. It turned out that some of the assumptions about detector responses were incomplete and simulations were missing some effects, which contributed significantly towards the data rates as well as processing needs. Therefore, there was a constant effort to try to utilize the existing infrastructure as best as possible. The EPN system scales very well and the increased computing demand, due to these external reasons, can be easily provided by adding additional servers.

6.1 Data Centre Performance and Test Results

The new Data Centre is considered part of the critical infrastructure running the experiment. Without the Data Centre operating it is not possible to store experiment data. In particular for the data rates produced by the TPC online reconstruction is crucial for

reducing the rates to about 100 GB/s for Pb-Pb, allowing to store everything on disk. Reliable Data Centre operation, availability of sufficient servers as well as the network is crucial and a stable operational environment inside the Data Centre contributes strongly to this. From the very beginning cooling was one of the focus points, to ensure stable operations for the IT equipment, in particular the control system steering the cooling. It was important to closely follow the developments, since the cooling system was newly designed for the CDC and there was therefore no experience operating them in an HPC environment. The resulting performance of the Data Centre for operations turned out to be as expected in the end. However, there were a few surprises along the way and some challenges to overcome towards the final results. In the following chapters most important steps are explained towards stable operation in a highly demanding compute environment, with respect to power densities and dynamic load changes. As technical contact it was my responsibility to ensure the performed tests were as close to the real use case as possible and that the IT equipment requirements are met.

Figure 6.1: Fully assembled Data Centre at P2, with 4 IT-Containers in place. The total cooling capacity of 2.1 MW split equally between the 4 containers. 250 EPNs are housed in IT1 and IT2, the two rightmost IT-Containers in the picture [14].

6.1.1 Design of the First Cooling Tests

One of the first tests CERN closely followed was the factory testing at the manufacturer premisses in Spain. One of the company halls was used for testing and contained a test-bed, on which the cooling units could be tested with different input parameters. Figure 6.2 shows a schematic overview of the test set-up. Large resistor banks created static load. The amount of heat generated by the resistor banks was one of the adjustable parameters and was usually set to the nominal maximal cooling capacity of the unit,

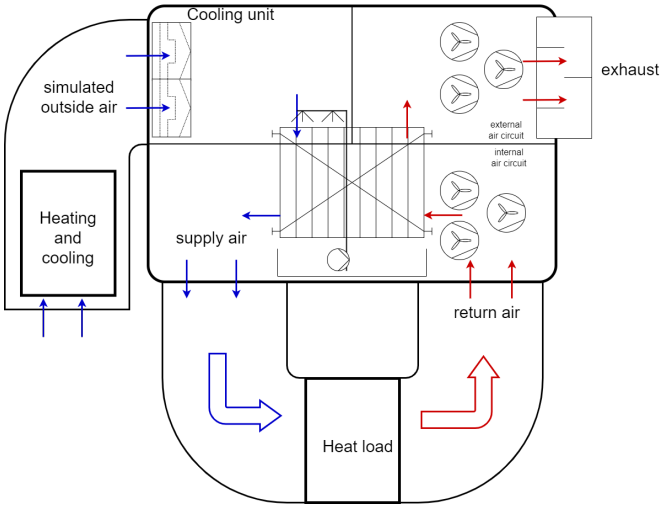


Figure 6.2: Schema of the cooling unit test bed. The heat load was created by large resistor banks placed in the internal air circuit of the cooling unit. An external heating and cooling system could provide different input temperatures, simulating a variety of weather scenarios [27].

to be as close to the worst case as possible. An external cooling and heating system was able to provide a specific air temperature for the external circuit of the cooling system, therefore mimicking changing weather conditions and also providing another adjustable parameter to simulate worst case conditions or to trigger a transition from dry to adiabatic mode. This was therefore the first benchmark of the system, verifying the cooling capacity of one single unit at different simulated weather conditions, including the expected worst case temperatures for the Geneva area deduced from weather data of the past 35 years, up to $38°C$. This first testing was a great opportunity to get a first impression of the performance numbers of the future cooling system, with one of the first newly developed units, so basically still a form of prototype. This opportunity allowed to give first feedback, so the whole system could be accustomed more towards the needs before delivery to us.

As this could be considered a prototype, there were multiple issues discovered in the very first tests. One of the steering challenges was the transition between dry and adiabatic mode. The first version of the control system provided by the vendor had a binary steering of the adiabatic mode, either running the water pump at maximum speed and therefore spraying the maximum amount of water on the air-to-air heat exchanger or having the pump off and not spraying any water at all. Since the external fans were usually close to or at the maximum fan speed at the cross-over point when the adiabatic mode got triggered this resulted in pushing a large amount of air through the whole system, while starting to spray as much water as possible. The resulting increase in the cooling capacity was very substantial and there was an immediate drop in the supply air temperature, turning off the adiabatic mode shortly after activation at temperatures close to the cross-over point. This over-steering was creating oscillations, toggling the

adiabatic mode very frequently. The transition between dry and adiabatic mode was one of the things to be improved after the first tests.

This first test did already show that the cooling unit can provide the required cooling capacity for the expected Geneva weather conditions and keep the supply air inside ASHRAE A2 at high external temperatures [6]. The time in adiabatic mode increases as the outside air gets warmer. At the same time higher external temperatures also reduce the temperature drop of the supply air when activating the adiabatic pump. A solution to mitigate the problem partially was a more fine grained steering of the water pump, starting slower and increasing the water volume when needed. A smoother transition significantly reduced the oscillations but did not completely avoid them. The control system core functionality, to determine fan speeds and water flow is realized with a Proportional Integral Derivative (PID) controller, a feedback loop with three control inputs. The proportional P is a scaling factor for the difference between the measured value and the configured target value. The integral I is scaling the integrated differences between measurements in the past and the target over a defined time window. The differential D is scaling the rate of change of the actual value with respect to the target in a defined time window. The PID control parameters needed further tuning, in particular when installing multiple units on top of a single container.

Careful analysis of the test data showed a steep increase in air humidity of the internal air circuit when the adiabatic mode was activated and water was sprayed on the heat exchanger. This was surprising, since the cooling system is designed in a way that there is no air exchange with the outside, the internal air circuit is completely isolated. The test bed was designed to mimic that very behaviour. The internal air circuit of the cooling system had the heat banks inside a closed loop and was supposed to be completely isolated from the external air. It was worrisome that the relatively fast rise in inside humidity was coinciding with the activation of the adiabatic system. The heat exchanger was designed to be completely water and air-tight, the external and internal air should not mix and remain completely isolated. The first potential explanation was that the sealing between air circuits was not fully tight at the test-bed and the humidity increase was due to some unwanted cross-talk between the inside and outside air. Due to the amount of water evaporated over a longer time in adiabatic mode would increase the humidity level inside the hall the test-bed was located. The increase however was so significant that the suspicion was immediately directed towards the heat-exchanger itself. It turned out to be the correct assumption, some production issues with the first batch of heat-exchanger allowed water to come into the internal air circuit through micro cracks. The capillary effect was contributing to the problem and the heat-exchangers of the first batch of cooling units had to be completely replaced, to ensure tightness and avoid water getting inside the Data Centre area. The manufacturing process of the heat-exchanger was changed and additional quality control ensures that all units build afterwards don't have defects which would allow cross-talk of any kind. None of the units delivered to ALICE did show issues. Early testing paid off in this regard and the problem was spotted early enough not to affect any of the delivered cooling units. Humidity values inside the Data Centre are continuously read out and fed into the container monitoring.

6.1.2 Factory Acceptance Test

After several test rounds and incremental improvements on a single unit, the factory acceptance test was the first opportunity to test a full container equipped with load-banks at the supplier's premises. This gave the first impression of the cooling units interplay, since there is a single cold and hot aisle per container and four units providing the required cooling power. With the amount of air pushed through the internal air circuit this inevitably leads to some cross-talk of neighbouring cooling units and sharing the cooling load of the racks in between cooling units.

The test was planned for two days, to run the IT-Container with the maximum load for at least one temperature cycle of one day. The load was simulated by heat resistors and was kept static at the specified maximum cooling, to get an idea of worst case cooling performance for the weather conditions during the test. Temperatures during the Factory Acceptance Test were not too demanding for the whole system. The system demonstrated the expected cooling capacity, even though the steering would need further improvements.

Seeing the control setup and possible parameter space to steer the cooling, the ability to set the most relevant parameters remotely via Modbus was increasingly important. This should ease the Data Centre setup and make it easier to ensure all cooling units always run with the same settings.

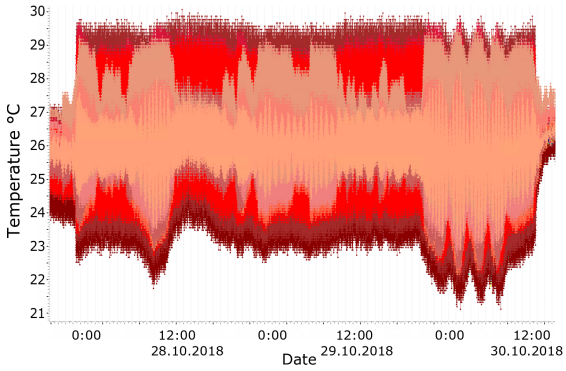
6.1.3 Site Acceptance Testing of IT-Containers IT1 and IT2

For the first two containers the testing program was consisting of a full load scenario over two days. This very basic test gives a good impression of how the whole system reacts towards changing weather situations. The temperature changes during the day can be easily above 10°C, in the summer even towards 20°C difference between the lowest temperature at night and the hottest during the day. Fixing one parameter, the load, however makes the steering significantly easier and this test is just a first indication of the stability in varying outside conditions.

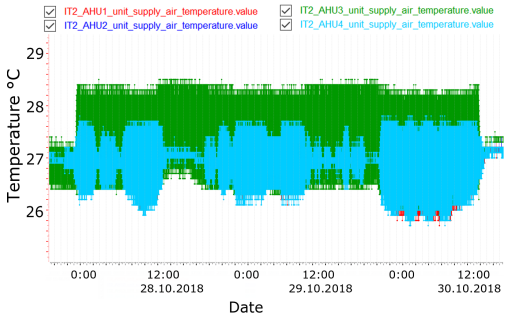
To get meaningful test results, sufficient temperature sensors in front of the rack are required, since local temperatures can differ a few degrees. In all of these test cases a minimum of three temperature sensors every second rack was installed, to get a good overview and to spot local differences. A particular challenge was to get a good working control system at all possible weather and load conditions. A PID parameter set working at a certain outside temperature range can behave significantly differently in case the outside temperatures changed for the next test.

One concerning observation was that the whole system was tuned toward the current weather conditions for the test. Expectations were that the steering can cope with the whole spectrum of Geneva weather and therefore does not need to be tuned specifically for a certain day.

After initial installation and setup by the company, the acceptance procedure on site was started. There were multiple tests with the load banks, since the initial performance did not fully meet the specifications. For the tests a static maximum load scenario, with

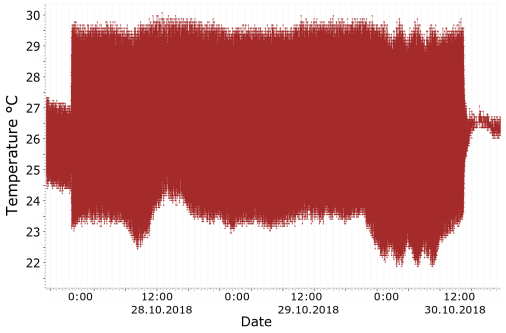


(a) Cold Aisle temperature profile during site acceptance testing of IT-Container 2. Overlay of different sensors, each colour is a distinct sensor.

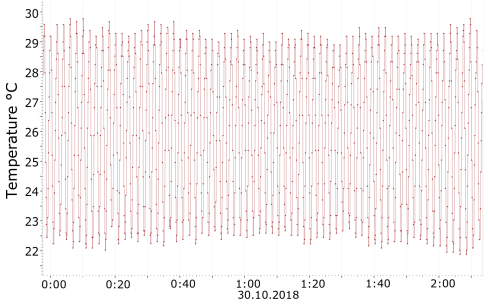


(b) Temperature readings of the cooling units.

Figure 6.3: IT2 528 kW load test October 2018



(a) Single sensor with the worst oscillations.



(b) Oscillations have a 2.5 to 3 minutes period.

Figure 6.4: IT2 528 kW load test October 2018

528 kW heat load active in the container was chosen. The cold aisle temperature profile measured at the rack doors for one of the tests is shown in Figure 6.3. It was clear that the oscillations measured in front of the rack were outside the ASHRAE A2 envelope. Amplitudes of the temperature measurement of a single sensor were up to approximately 8°C and therefore significantly above the allowed 5°C in an arbitrary 15 minute window [6]. Figure 6.4 shows the sensor with the highest oscillation amplitude during these particular tests. Temperature peaks are between 2.5 and 3 minutes apart and are a clear indication of a control issue. In this particular test one of the cooling units (AHU3) was malfunctioning during the test.

Testing in IT-Container 1 was done afterwards, to ensure full functionality of the container after delivery as well. The steering system was constantly tuned, to improve the response of the cooling system. The resulting temperature profile of the sensors in front of the racks can be seen in 6.5 and shows much better results. The oscillations were inside ASHRAE A2 [6] and the performance was therefore in the acceptable range for us in a static maximum load scenario. It was clear that further tuning for dynamic

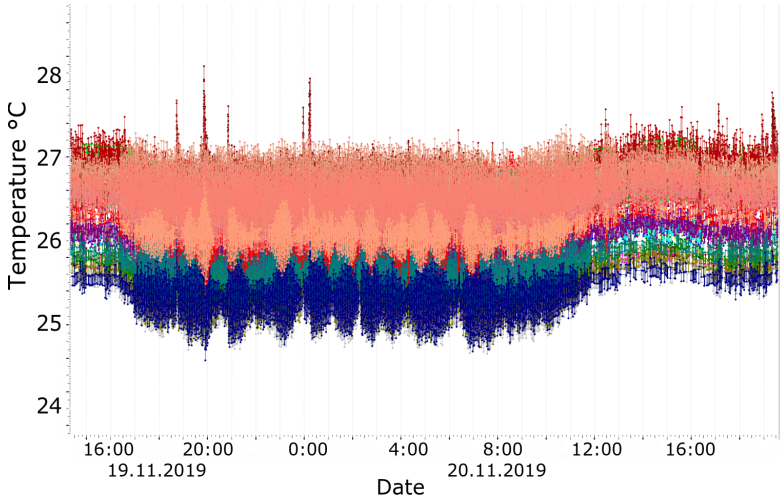


Figure 6.5: Cold Aisle temperature profile during site acceptance testing of IT-Container 1

load scenarios was required to get the desired performance of the cooling system.

One of the concerns was again the humidity changes in the data room. After experiencing the heat exchanger tightness problems it was essential to make sure that this problem was solved for good and that there is no risk to get water inside the data room through the heat exchanger. During the tests the outside humidity was tracked without any delay, which indicated a significant cross-talk between inside and outside air. The first cooling units had some caps between individual parts of the casing, which allowed air exchange and in certain conditions even rainwater coming into the cooling unit. Rainwater entering the units slowly trickled down and eventually made it through the supply air duct, which is located just above the cold aisle above the IT racks. This caused severe operational risks and required modifications on the delivered units, to ensure water tightness and conformity to the specified air leakage rates. For all subsequent cooling units the modifications for additional sealing on the roof and sides of the casing were done in the factory, before delivery.

6.1.4 Site Acceptance Testing of IT-Containers IT3 and IT4

After the experiences during the first testing campaign and due to the lack of time to do additional tests for dynamic load scenarios, the efforts to understand the system toward changing load scenarios during the site acceptance testing of IT3 + IT4 were prioritized. These tests clearly showed weaknesses of the control system. It was not possible to tune the system to a stable working point during changing load scenarios in a way that the ASHRAE A2 specification was met. The cold aisle temperature oscillations were beyond the allowed 5°C in 15 minutes, measured with temperature sensors attached to the rack doors. Figure 6.6 shows one of the performed tests with quickly switching on and off a significant amount of load banks, to simulate operational conditions during

LHC data taking, which introduces quick IT load changes. The cooling system was usually over-steered and took some time to stabilize afterward. Multiple attempts to tune the controls to be more resilient and avoid temperature changes above $5°C$ in the cold aisle.

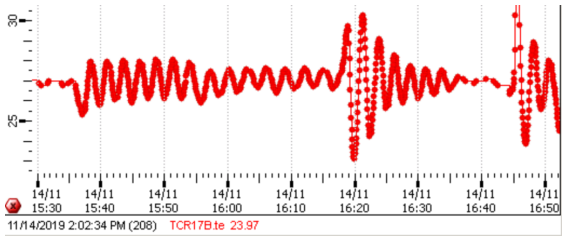


Figure 6.6: Cold aisle temperature sensor during quick load changes. Steps: 15:30 100 %, 15:50 75 %, 16:16 22 % 16:45 100 %

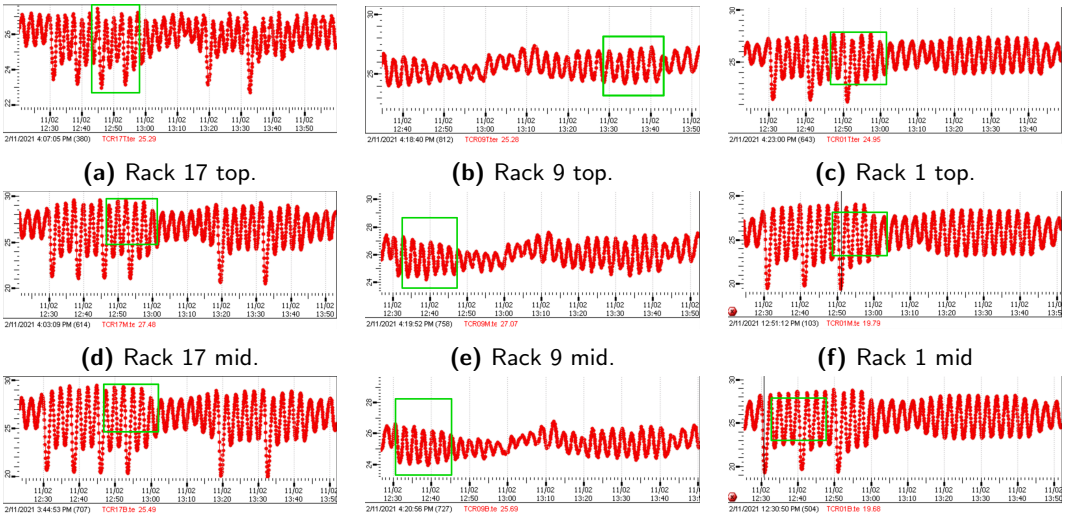
Total power consumption (kW)	570	498	426	359	291	215	130
Infrastructure overhead (kW)	32.8	26.3	22.0	21.4	20.6	20.5	19.6
Heat load approx. (kW)	537	471	404	337	270	194	110
% of nominal max.	102	90	77	64	52	37	21
PUE ($\frac{\text{Total Power}}{\text{Heat Load}}$)	1.061	1.056	1.054	1.063	1.076	1.105	1.178

Table 6.1: Power consumption and PUE during site acceptance testing 2019. Power measurement for total power at the main distribution board is not reliable below a total power consumption of 100 kW.

Figure 6.1 shows the power efficiency during the test procedure, to verify the Data Centre operates within the specification of a PUE below 1.1 for IT load of at least 50 % of the maximum load. This part of the test was successful. The Power Usage Efficiency test was done at temperatures around $20°C$, at the cross-over from dry to adiabatic mode. Since the adiabatic mode is triggered at maximum external fan speed, energy consumption is highest just before activation of the adiabatic mode. Due to the limited fan speeds, the Data Centre stays below 1.1 for load scenarios between 50 % and 100 % of the nominal maximum load. In this particular case the minimum internal fan speed was 60 %, which is high for the lower load tests and not only creates significant overpressure in the cold aisle but also adds a static overhead for the power consumption impacting the PUE. Fan speed related power consumption is discussed in Chapter 5.1.4.

6.1.5 Testing with the Final Load Profile

After the hardware decision for the servers was finalized in 2020 and it was possible to measure power consumption, fan speeds, etc., with the final equipment beginning of 2021. With the exact hardware in place the Data Centre infrastructure and in particular the cooling could be tested under realistic conditions.



(a) Rack 17 top. **(b)** Rack 9 top. **(c)** Rack 1 top.

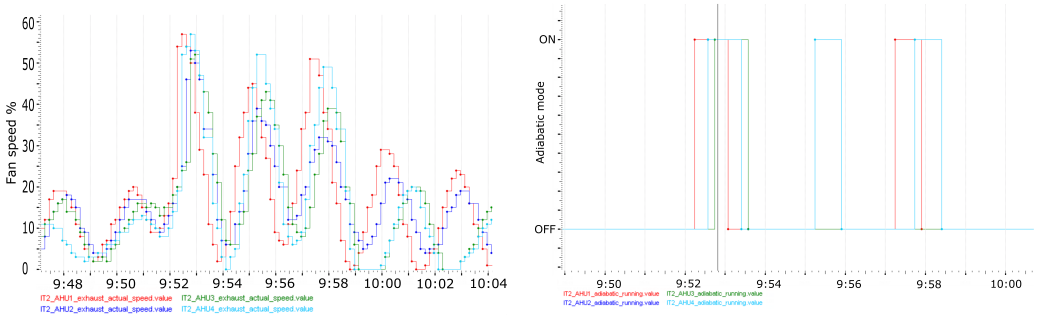
(d) Rack 17 mid. **(e)** Rack 9 mid. **(f)** Rack 1 mid

(g) Rack 17 bottom. **(h)** Rack 9 bottom. **(i)** Rack 1 bottom.

Figure 6.7: IT2 cold aisle temperature with EPNs without additional tuning of the control system (11.02.2021). Green rectangles represent the ASHRAE A2 window, $5°C$ in 15min. Middle racks are okay, the outer racks are clearly outside the specified environmental envelope.

It was quickly clear that with the final server hardware the Data Centre encountered the same issues with the cooling steering process as before encountered during the dynamic load tests with heat dummies. Figure 6.7 shows the cold aisle temperature during the first tests. These tests started DGEMM benchmarks on the GPUs with the ROCm validation suite [4], to simulate the start of a run. This resulted in a significant spike in power consumption and therefore cooling requirements. There were several corner cases, which could trigger a significant temperature swing of the server supply air. Depending on the weather and current load scenarios these temperature swings could be above 5°C and therefore violate the operational environment of the servers, which was specified to be the ASHRAE A2 envelope.

One scenario which could cause such severe swings was triggering the adiabatic mode in low outside temperature environments. When dynamically changing the load significantly in a short time, e.g. simulating a start of the run and therefore quickly changing from idle to full load, the steering was unable to cope with that in an orderly fashion. The adiabatic mode was almost always activated under these load swings, even at temperatures close to 0°C outside temperature. The control system itself had a minimum fan speed of 55 % set (external maximum fan speed 72 %) when entering the adiabatic mode. This resulted in an immediate ramping up of the outside fans from approximately 20 % to 55 %, causing a significant cooling increase and in this case a massive overshoot. There was no water sprayed, since the water gets completely drained for outside temperatures below $5°C$ to avoid freezing. The increase in cooling capacity was completely realized by the increased airflow. 55 % fan speed represents $\frac{3}{4}$ of the maximum fan speed



(a) Fan speeds during the test, jump when adiabatic mode is triggered the first time

(b) Adiabatic mode on/off

Figure 6.8: Fan speeds during a test going from idle to full load and back, approximately 8 minutes full load, between 9:50 and 9:58.

set by the control system. Supply air temperatures dropped significantly as a result and the outside fans were slowed down again, sometimes even turned off for a short time and then starting another cycle of supply air temperatures increasing, adiabatic mode triggering a cooling overshoot. A single load transition could easily get us in an oscillating meta-state, which would not recover until the operating conditions changed significantly again e.g. load or outside temperatures were different. Figure 6.8 shows the fan speed response including enabling the adiabatic mode, causing a jump in outside fan speed.

The measured temperature swings were bigger on the outskirts of the cold aisle. In the middle of the container the temperatures tended to be more stable. The cooling units however are steered individually, and it should therefore not make any difference if the controls work as expected. Figure 6.7 shows temperatures in the middle inside a $5°C$ window according to the ASHRAE A2 specification, the sides of the cold aisle on the other hand can suffer significant temperature swings of more than $10°C$. The outermost sensors clearly show that the oscillations are not dampened and therefore don't stop without an external change. These oscillations are the result of a delayed response of the control system and a clear over-steering afterwards. The larger temperature drops are a result of triggering the adiabatic mode and therefore a jump in the outside fan speed. In certain scenarios this can be triggered periodically and does not automatically stop after a certain time. The control system stabilizes after the load was turned off, even though there are still visible oscillations. This behaviour was strongly dependent on outside temperature and the parameters could be tuned to be more stable for a certain temperature range. The overall goal was a good operation during the whole Geneva weather cycle. Constant parameter tuning for a specific temperature and load scenario was not a very appealing prospect, since the outside conditions, as well as the load profile, could change significantly and often. Changes from idle to full load as well as everything in between can be occurring multiple times per hour, depending on the current running scenarios of the experiment. It was simply not feasible to constantly adjust this in operation.

For several tests the parameters of the control system were adapted to the specific

test. The proportional and integral part of the PIDs was usually tuned to a specific load and weather combination and would work in some operational window. This was problematic since the weather conditions change significantly over the course of a year and the EPNs cluster operates in a wide range of different load scenarios. The differential factor of the PIDs was not used during early tests. The system was tuned during the site acceptance tests at temperatures in the mid 20 °C and therefore tended to over-steer at much lower temperatures. This was one of the reasons why it was possible to trigger an oscillating meta-state of the cooling system with quick load changes during cold weather, e.g freezing and lightly negative temperatures in the beginning of a year. The whole system was not reacting fast enough and in particular very fast load changes got the control system to its limits.

6.1.6 Fixing Cooling Controls and Stability with Tuned Settings

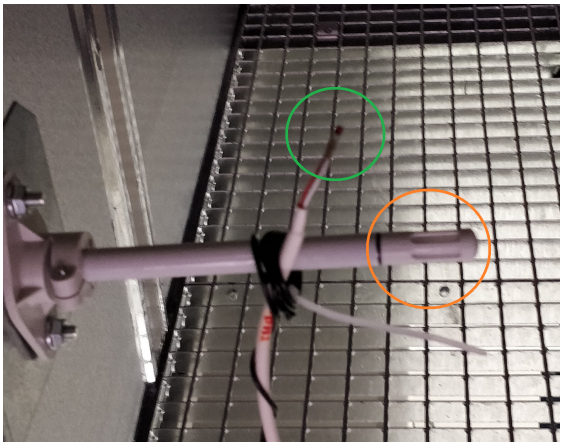


Figure 6.9: Additional temperature sensor installed (green circle) close to the cooling units temperature input sensor (orange circle).

During testing it was concluded that the major problem in the control loop of the cooling system was the input temperature measurement of the cold aisle conditions, which was neither fast nor precise enough to allow good control of the whole system. This was not immediately obvious, since the technical specification of the utilized temperature sensors showed a response time $t_{90\ \%} = 15\ s$, corresponding to a time constant of $\tau < 5s$. This means that for a temperature change of $1°C$ the sensor would capture 90 % of in 15 seconds, in this case at least measure $0.9°C$ of that $1°C$ change in 15 seconds. A time constant of $\tau < 5\ s$ means that 63.2 % of the temperature change is captured in less than 5 seconds. This should have been sufficiently fast to steer the cooling precisely enough.

However, by placing additional sensors next to the input sensors of the cooling unit significant differences in the measured temperature got apparent, which could not be explained at first. Figure 6.10 shows the observed differences between the input of the control system and the secondary measurements. For this comparison, an additional

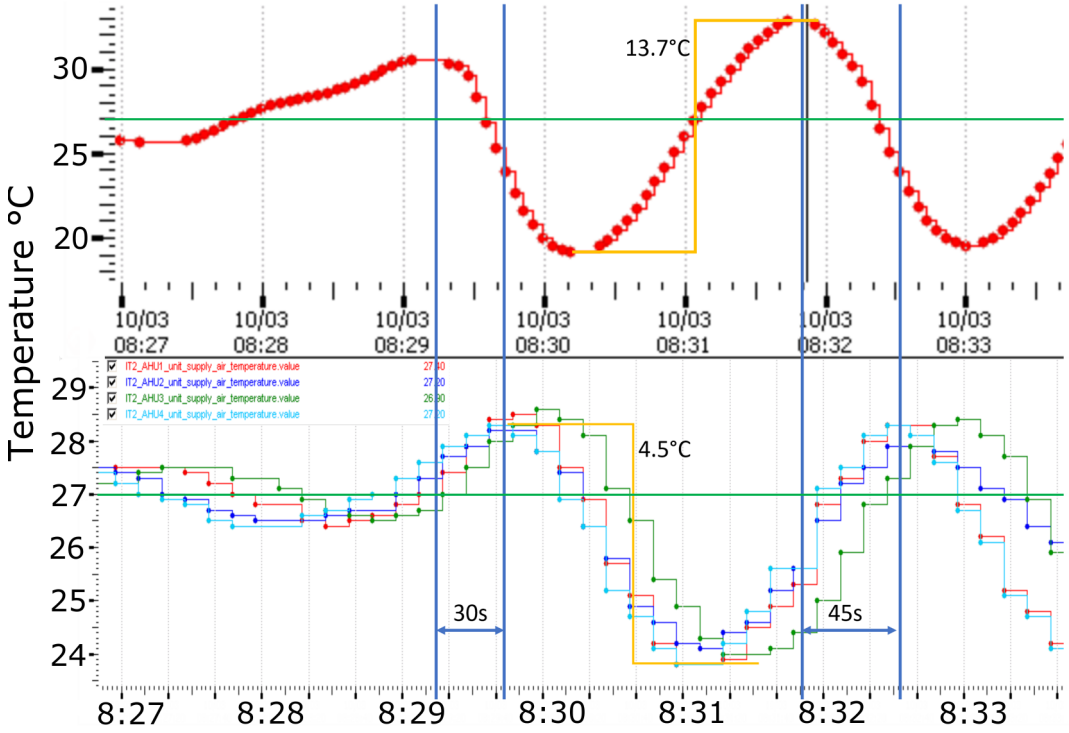


Figure 6.10: Temperature measurement a reference sensor (top) and the AHU4 input sensor (turquoise, bottom) in comparison. The two sensors mounted close to each other show a significantly different temperature profile (sensor placement see figure 6.9).

temperature sensor was installed (red circle in 6.9) close to the sensor of the unit, which is used as input to steer the cooling system. There were significant differences between the two sensor readings. The additional sensor measured a significantly lower minimum temperature as well as a higher peak temperature. The second observation was a time shift in the values, the maxima of the second sensors seemed to be shifted against the temperature sensor of the unit by more than 30 seconds. At first glance it looked like a moving average filter was applied to the input, to smooth the measurements. This however was a false lead, there was no electronic filter applied to the input at all.

The problem was not on the electronic part of the sensor at all but on the mechanical side. There are protective caps installed around the temperature sensors, to protect them from dust and small particles. The sensor caps acted as a thermal insulator, detaching the sensor from the direct airflow and therefore making a direct measurement impossible. The air inside the sensor caps was only slowly converging towards the actual temperature of the airflow and therefore acted similar to an electronic moving average filter. Even though the sensors themselves were fast enough to capture temperature change, the protective caps counteracted this and caused severe problems when fast temperature changes happened.

Figure 6.11: Protective caps of the temperature sensors. Left - paper inlet removed, right - original version.

After removing the paper filter inside the protective caps and therefore allowed the airflow to touch the sensor, the measurements of the second sensor and the one used by the cooling unit were immediately in sync. The time shift was gone and the values for the measured temperature matched. The cooling unit was also instantly more responsive to changes in the IT load, without changing the parameters of the controller. Figure 6.11 is showing the clear difference between the sensor caps with the filter paper obstructing the airflow and without. Figure 6.12 shows the improved input measurements after the modifications, now well aligned with the additional sensor readings.

Especially for a highly dynamic environment, in which the load can change significantly over a short time due to the run concept, the temperature measurement precision of the cooling steering input is of the utmost importance. This applies to the absolute value of the temperature as well as the time delay after which a changed temperature is measured. For a good working control of the system inputs on which the steering is calculated play a crucial role. The control system is unable to compensate for flawed input values. This is a principal problem of the steering challenge. The response of the system can only work if it gets reliable information about the environment to steer. In particular, the time offset of the measurements contributed to an oscillating behaviour of the system, always trying to steer on something which is already 30 s or more in the past. Trying to tune the control system with the problematic input measurements was not leading to a stable solution for all operational conditions.

6.1.7 Results With Fixed Control Inputs

The biggest improvement in stability was achieved when the differential part of the fan control PIDs was tuned, after fixing the temperature measurements. Due to previous experiences with other installations, the cooling system manufacturer strongly advised against using the differential part and warned about potential instabilities, which was reasonable with the typical slow and unprecise temperature measurement of the unit.

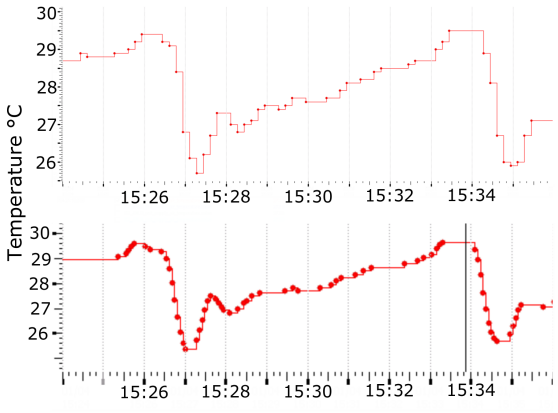
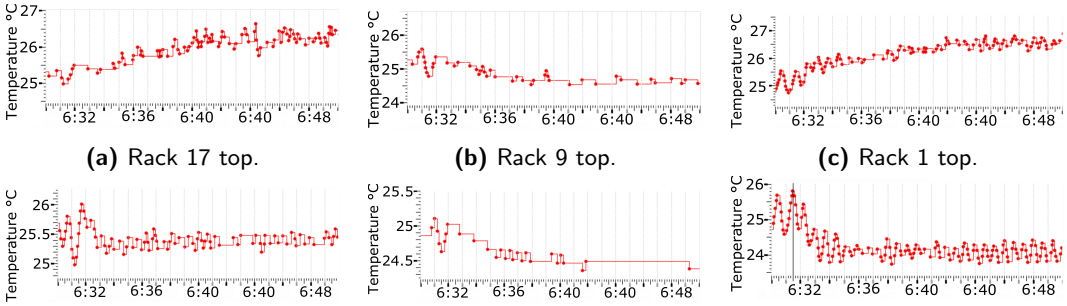


Figure 6.12: Comparison between sensor values of the cooling units and control measurement, after removing filters inside protective caps. Top: sensor of the cooling unit, bottom: additional temperature sensor installed close to the cooling unit sensor. Both sensors show now the same behaviour.



(a) Rack 17 top. **(b)** Rack 9 top. **(c)** Rack 1 top.

(d) Rack 17 bottom. **(e)** Rack 9 bottom. **(f)** Rack 1 bottom.

Figure 6.13: IT2 cold aisle temperature at full load with production EPNs.

However, in theory the differential part is crucial for the steering in a highly dynamic HPC environment, with large load changes in a very short amount of time. From the ALICE run concept it is possible to move from idle during the time no data is incoming toward full load when data is arriving with up to 1 TB/s in roughly 30 seconds. There is of course some inertia in the servers itself, so temperature changes are slightly delayed and are not instant. Nevertheless, this highly dynamic scenario turned out to be challenging to control without the differential part. Figure 6.13 shows the supply air temperatures of the servers during a transition from idle to full load. After introducing a proper differential factor in the PID, the cold aisle temperatures remained stable during all tested load transitions. Figure 6.14 shows external fan stability after tuning. The overall stability of the whole cooling system was significantly improved.

Figure 6.15 shows the improvements after using the differential part of the PIDs for fan control. There are still some apparent oscillations, however the amplitude was reduced by a factor of 5-10x in certain scenarios. The remaining residual control problem, in this

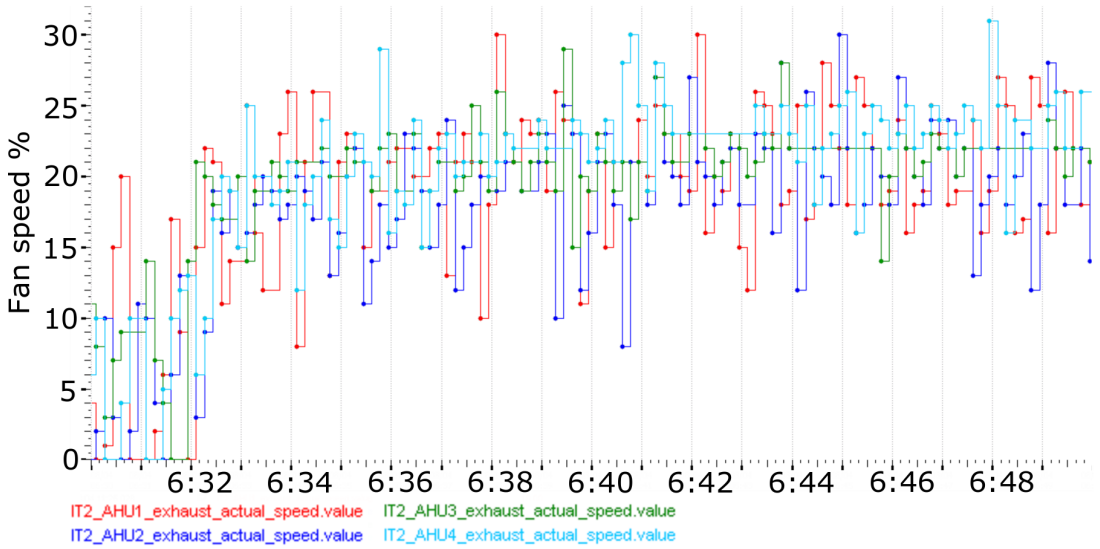


Figure 6.14: Outdoor fans after control tuning, transition from idle to full load at $0°C$ outside temperature. No more big jumps and almost constant speeds indicate a more stable system.

case, was due to the low outside temperature. Figure 6.16 shows the external fan speed. Due to the outside conditions, the external fans can be turned off completely (0 % fan speed). This corner case still leads to the oscillations, since there is still a threshold remaining to turn the fans on again.

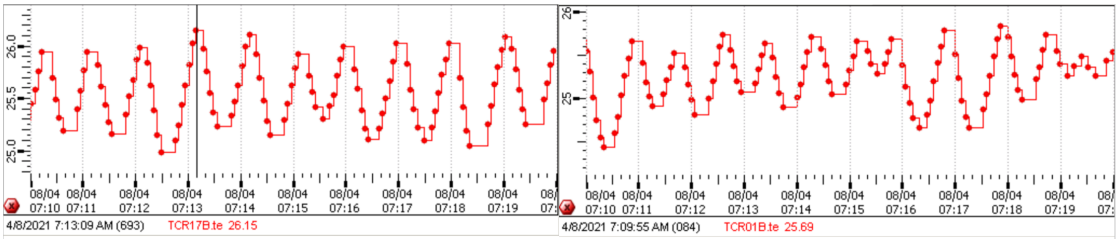


Figure 6.15: Temperature profile in front of the racks after tuning with the differential parameter, using the full PIDs for the fan steering.

The overall stability however is acceptable and oscillations of less than 1°C are barely noticeable for the servers themselves. The final tuning step and slight modifications on all the PIDs parameters got us to good operational conditions, well inside the ASHRAE A2 envelope. Quick and significant load changes of a 5-6x increase of the load between idle and full load as well as the corresponding decrease going back to idle are very well compensated by the control system. This means that no matter how the cluster is operated and regardless of the outside weather the control system manages to keep the supply air temperature very stable and precisely at the set-point. To see a significant change in supply air temperature now needs some extraordinary conditions, like one

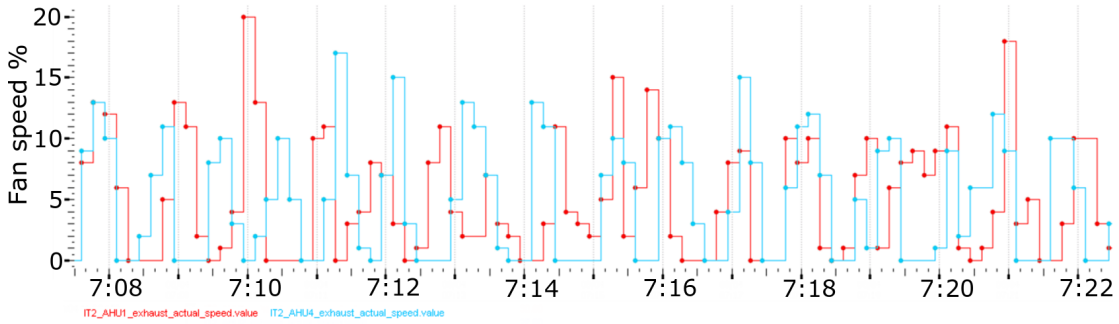


Figure 6.16: Outside fan speeds after tuning with the differential parameter, using the full PIDs for the fan steering.

cooling unit missing, either due to failure or during maintenance. However, even then the cooling system is operating inside the ASHRAE A2 envelope and can safely operate the cluster in the foreseen operational window.

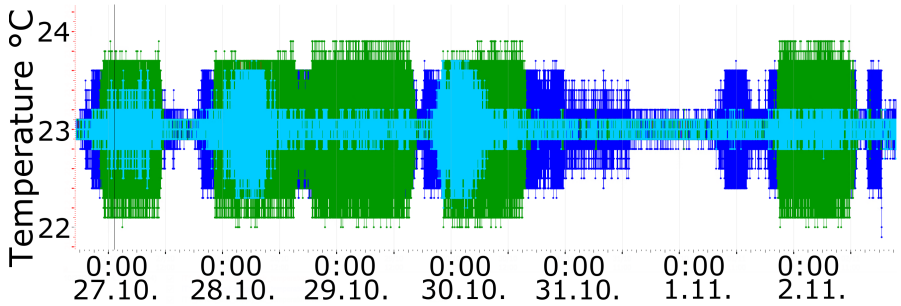


Figure 6.17: Supply air profile during pilot beam collisions (very low interaction rate and therefore low IT load). Temperatures stable around the set point of $23°C$. Individual cooling units represented by different colours. Blue: AHU2, green: AHU3, turquoise AHU4. The temperatures of AHU1 are more stable and therefore fully covered by the other units.

Figures 6.17, 6.18 and 6.19 show the cooling system performance for some of the important milestones. In particular during the pilot beam and the first stable beams at top energy, the cooling system remained in a narrow window. During the high intensity scans there were a few occasions when the adiabatics kicked in and compute load decreased at the same time, which resulted in a temperature drop of up to $4°C$, which is still inside the ASHRAE A2 envelope. This shows there is still further room to optimize the cross-over from dry to adiabatic mode.

6.1.8 Full Weather Cycle

By now the Data Centre demonstrated performance during a full weather cycle, while operating in the nominal data taking conditions during ALICE Run 3. The Data Centre

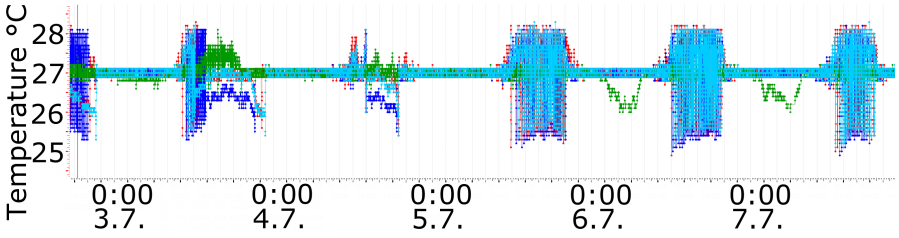


Figure 6.18: Supply air profile during first stable beam collisions (low interaction rate and therefore low IT load). Activation of adiabatic mode can still lead to a temperature drop of roughly $3°C$. Cold aisle set point of $27°C$.

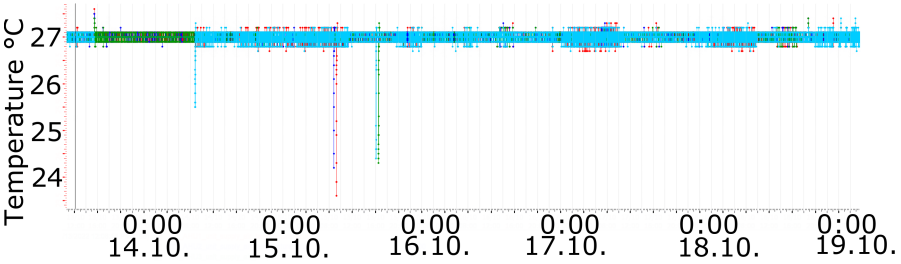


Figure 6.19: Supply air profile during high-rate tests (very high interaction rate and therefore close to maximum IT load). A few temperature drops of up to $4°C$ when adiabatic is started. Cold aisle set point of $27°C$.

performance now, after tuning everything to the installed hardware needs, are good. There were no further adjustments needed and the steering works for all weather conditions. The actual IT load is also no longer an issue and the system steers very well in idle, full load and every scenario in between. Quick load shifts, e.g. the start-up of a run which corresponds to a quick transition from idle to full load are also not causing any issues and supply air temperatures are reliably in the ASHRAE A2 envelope. The overall goal of stable operations without regular control adjustments was therefore achieved.

6.1.9 Partial Running Challenges

During the build-up of the compute infrastructure, the IT-Containers were operating at low load and therefore significantly reduced cooling requirements. Since two out of four units still provided much more cooling capacity than was needed for this scenario, two units were disabled to save electricity. In the tested scenario even one unit would have been enough, but it was decided to run with two for redundancy considerations, to have some margin in case of potential problems. This way of operating the Data Centre container worked well during the summer when the outside temperatures were above $20°C$. However, when temperatures dropped below $10°C$, condensation occurred in the external air circuit of the cooling units. Figure 6.20 shows the upper compartment, so the nominal outside air circuit of the disabled cooling unit. The louver is closed during the shutdown procedure of the cooling units, to protect the unit. During normal

Figure 6.20: Condensation inside an offline cooling unit when operating with two out of four units at a low load scenario during cold weather conditions.

operation this part of the unit gets heated through the Data Centre heat and is therefore significantly warmer than the outside temperatures. The casing however gets cooled down and provides a perfect surface for condensation. This behaviour prevents shutting off any cooling units during cold weather conditions. The units have to be on, so some air is passing through the external air circuit to minimize temperature differences in this area and to prevent condensation.

In the meantime, there were some efforts to insulate potential cold spots, to prevent condensation at other places, close to the heat exchanger. No more condensation was observed after insulating the most critical areas.

6.1.10 Fail-over Testing and Maintenance Interventions

Data Centre operations shall be possible even if components fail or are temporarily shut down for maintenance. An important test was the simulated failure of one power line. Since the cooling units are only fed by one power line and an Automated Transfer Switch (ATS) will switch the cooling units to the second line automatically. This leads to a short interruption of the cooling units, since this is effectively a power cut, in case the primary line goes down. For ALICE all cooling units are connected to a single ATS and are therefore always running on the same power line and will go down all together. It was verified that even under full load, the cooling units come back fast enough to ensure that the supply air temperature remains manageable and operation is not impacted. Figure 6.21 shows the resulting temperatures measured at the rack doors. A rise slightly above $35°C$ occurs for less than a minute and recovers quickly to the normal set-point.

During maintenance one cooling unit is regularly shut down, to do the required in-

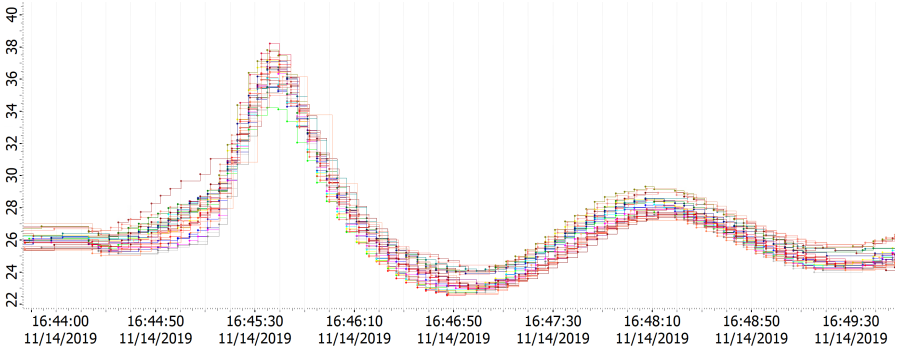


Figure 6.21: Cold aisle temperatures during a simulated failure of one power feed, with a short cooling interruption and automatic switch over to the secondary feed during a maximum load scenario.

spections, cleaning or replacing filters. Data Centre operation is not impacted by this and the IT-Container continues running with the remaining cooling units during these interventions. There are however safety related tests, which don't allow operations in parallel and result in planned downtimes of the Data Centre. Internal guidelines require an annual test of the emergency power cut, to ensure this safety measure works as expected. A red button or lever at the entrance of each Data Centre room will cut all power in the Data Centre room, including UPS. The emergency power cut infrastructure is on multiple levels of the supply chain and result in several power cuts during the annual testing procedure. Usually, the equipment was shut down in advance for this test and turn off the circuit breakers in the racks, to prevent damages from repeated power cuts. This planned downtime is a particularity of CERN. It is scheduled at the beginning of the year, during the winter shutdown of the LHC. Since it is coordinated and performed everywhere at the experiment site at the same day, ALICE is off during that day and the overall impact on the experiment is therefore limited.

6.2 EPN Server Performance

The EPN servers as described in 5.2.1 were constantly used during the commissioning phase, as soon as they were available beginning of 2021.

Already on the hardware prototype the software was benchmarked with simulated data of the most stressful scenario, data taking at 50 kHz Pb-Pb [42]. The aim was always to ensure enough processing power for the data taking goals. The hardware prototype was used to determine how many GPUs would needed to be procured, which of course depended on the actual model. 250 servers with a total of 2000 AMD MI-50s were purchased and installed [14], see chapter 5.2. On the software side there were continuous tests, to ensure there are no regressions and the processing remains fast enough. When the GPU estimates were done the software was not yet completely ready and several processing steps were still to be ported from CPU to GPU. A very useful tool for benchmarking is called Full System Test (FST), which is the approach to have testbed testing as much of the software functionality as possible in a single test. The Full System Test (FST) is developed by PDP as part of their software verification process. For the FST a dataset of simulated MC data is used and all the processing steps are done, as it would be done during the data taking phase of the experiment. This way it is possible to spot problems early on, before using the software in the production environment. This was also the main benchmark to ensure the computing performance was sufficient. Having a fixed data set, which is the base of every software test makes it also possible to compare the performance of different software versions. However, until the end of 2022, there were significant changes in the simulations and a few times in the data format itself, which makes a direct comparison of the results impossible and also indirect comparisons somehow difficult. The changes are of course required, to incorporate the learnings during the first data taking phase of the experiment and to simulate the real detector responses as closely as possible. Figure 6.22 shows the processing time (wall time) of a single Time Frame during a few runs of the Full System Test on a few different servers.

$$\text{Processing time} <= \frac{\text{Time Frame Rate}}{\text{Number of EPNs}} \tag{6.1}$$

$$<= \frac{88}{250\ s} \tag{6.2}$$

$$<= 2.84 \text{ seconds to output a TF on a single EPN} \tag{6.3}$$

The average time to process a single time frame on an EPN must not exceed a certain time, to be able to process all the data in real-time. HBFs correspond to a LHC orbit of approximately 89,4 μs. TFs with 128 HBFs contain therefore approximately 11 ms of data [14]. This results in a time frame readout rate of approximately 88 Hz. The available processing time of a single TF on one GPU is roughly 22.7 seconds, if all 2000 GPUs are assumed working and included in the run. This corresponds to 2.84 seconds to output another TF on each EPN, to reach 88 Hz TF rate.

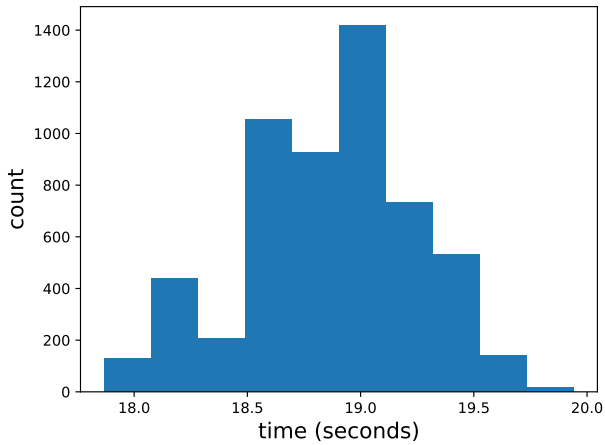


Figure 6.22: Result of a short Full System Test (FST) run. Performance is in the expected window below 20 s per Time Frame (TF).

$$\text{Processing time per TF} <= \frac{\text{Time Frame Rate}}{\text{Number of GPUs}} \tag{6.4}$$

$$<= \frac{88}{2000\ s} \tag{6.5}$$

$$<= 22.7 \text{ seconds to process a TF on a single GPU} \tag{6.6}$$

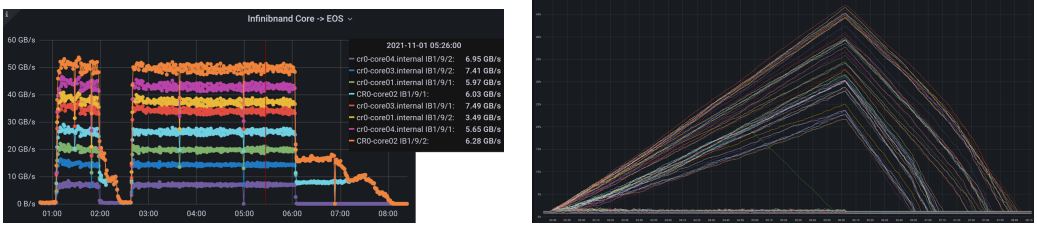
It's unreasonable to assume all compute resources are available during any Pb-Pb run with TPC. There can always be some hardware problems on a few servers or the requirement to run a test for other detectors in parallel. Therefore 230 EPNs are used as a reference for Pb-Pb processing at the nominal 50 kHz rate, which reduces the time to process a single TF to 20.9 s per GPU, see equation 6.7.

$$\text{Processing time per TF} <= 20.9 \text{ seconds to process a TF on a single GPU} \tag{6.7}$$

Figure 6.22 shows that the EPNs are perfectly able to process the simulated data in the required time of less than 20.9 s per TF.

In 2022 the LHC restarted and the remaining ALICE commissioning was done with beam, so while taking data which is usable for physics analysis at a later stage. It turned out that the actual data rate and in particular the number of clusters found for the TPC is higher than in simulations. This has a negative effect on the processing time since the number of clusters roughly scales linearly towards the GPU processing requirements. To compensate the increased computing requirements, due to unforeseen external reasons, 30 additional EPNs were already added to the farm. Another 70 servers are being integrated, to ensure sufficient processing margins for 50 kHz Pb-Pb data taking.

6.2.1 EPN Performance With Real Data



(a) Throughput via the IB-ETH gateway during the pilot beam after reducing the amount of streams and changing congestion control to BBR

(b) Fill status of the buffer disk during the run. Throughput to storage was not enough to push raw data rates for several EPNs

Figure 6.23: Pilot beam monitoring.

The pilot beam was the first test with LHC collisions for the newly upgraded detectors and therefore an important milestone. The first collisions delivered the first realistic detector data at LHC running conditions. This included all negative effects on the data such as radiation induced bit errors, misbehaving detector hardware etc. leading to data corruption. Since there was no real collision data available beforehand, the software was mostly tested with simulated Monte Carlo data. Several unexpected cases of data corruption were not yet fully handled in software, which led to crashes and the loss of an EPN every few minutes. The software was improved daily during the 6 days of pilot beam data taking. Additional checks were included, in particular to the raw decoders, to protect against all sorts of data corruption seen at these first tests. This increased EPN stability significantly. Towards the end losing an EPN due to a software crash only occurred every other hour.

ALICE decided to store all raw data from the pilot beam collisions and not just the Compressed Time Frames. This was required to enable further tests and verifications of the whole reconstruction chain since it enables direct comparison of the reconstructed data with the raw data and allowed additional tests. This resulted in data rates to disk of roughly 60 GB/s and was therefore already 60 % of the expected maximum Pb-Pb data rates. The high data rate to disk was one of the big challenges for the EPNs, caused by the availability of InfiniBand to Ethernet gateways. Only one of the gateways was available at the time. Even though the theoretical maximal throughput is 100 GB/s via eight 100 Gbit/s links, there were significant struggles to transfer the data coming from the detectors towards storage. The gateway hashing at the time was still the default hashing, which gives an unbalanced allocation in our case, see Chapter 5.3.6. In the beginning the installation barely achieved 30 GB/s. This was actually only roughly 50 % of the data rate which was required to push to EOS.

The EPN have an additional secondary hard-disk, which is a dedicated local data buffer. This local buffer is 3.84 TB per EPN and sums up to roughly 1 PB across the whole farm. At a rate of 100 GB/s this allows theoretically to buffer approximately 2.5h of experiment data.

During the pilot beam, the insufficient network throughput to EOS storage was compensated by the local EPN buffer-disks. This allowed the experiment to run without limitations. However drain periods were required to empty the local buffers. The total running time during the pilot beam was therefore limited by the local buffering capabilities. Delays in data taking and LHC fill cycles provided time slots without any incoming data towards the EPNs. These downtimes were used to shuffle around EPNs and temporarily removed some of them with high disk utilization. Due to the unbalanced link allocation only a subset of EPNs were affected by the low network throughput, the remaining EPNs were transferring the data fast enough. Allowing the draining of disks this way prevented any EPN from running full during the pilot beams and did therefore not impact operations. In parallel the network throughput to disk was improved.

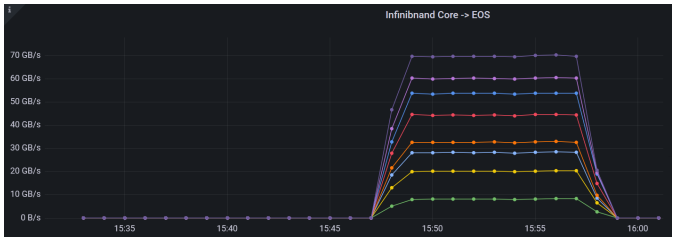


Figure 6.24: Throughput test EPNs to EOS with Iperf via a single gateway.

One of the encountered problems was caused by scalability effects. Even though micro benchmarks with the tool to copy data beforehand from the EPNs towards EOS looked promising, things were significantly different once data was transferred with more than 100 EPNs simultaneously. The number of parallel streams per EPN was configured to 20, which worked well with a smaller subset of concurrent EPNs. With 100 or even 150 and more EPNs, this summed up to be more than 2k parallel streams over just 8 network links. Congestion on the links led to package drops and therefore limited the total throughput to roughly 30 GB/s. To mitigate this issue the amount of parallel streams per EPN were significantly reduced to 4. This already helped to increase the data rate and get to a more stable system. In addition the congestion control algorithm was changed from the default cubic to Bottleneck Bandwidth and Round-trip propagation time (BBR). This improved the situation significantly and the system was able to transfer up to 50 GB/s towards storage and therefore 50 % of the available line rate, see figure 6.23.

Between the pilot beam and the first stable beams there was more than half a year. Continuous improvements of the whole computing chain, including stress testing and replacing problematic hardware ensured stable running of the infrastructure. Software stability was significantly improved by PDP colleagues compared to the pilot beam test and software crashes were no longer an issue. The interaction rates were low and only four bunches were injected in the LHC, with two bunches colliding at ALICE. This led to low data rates for the cluster and therefore everything went smoothly. There were no surprises and the software was stable, the whole chain was working well, including reconstruction and writing to disk. The network toward EOS was significantly improved

by adding an additional Skyway in an HA cluster extending the throughput between InfiniBand and Ethernet domains, see chapter 5.3.6. This meant that no local buffering was taking place, since the available bandwidth was much greater than the actual data rates at this low interaction rate test.

6.2.2 Verification at Full Rate

Verification of the whole experiment was a challenging task. Nominal data taking conditions during pp collisions are far away of the Pb-Pb scenario. Nominal pp interaction rates at ALICE are low compared to other experiments, but can be increased by the LHC. A pp interaction rate of 3.8 MHz produces the same amount of detected particles inside the TPC detector as 50 kHz Pb-Pb. 3.8 MHz pp interaction rate is used to simulate 50 kHz Pb-Pb like conditions for the whole readout and processing chain, since the TPC produces more than 90 % of the data. During the high rate tests, the ALICE interaction rate was increased in steps from 500 kHz to 3.8 MHz, to get close to particle detection rates occurring during Pb-Pb collisions. This was one of the key commissioning exercises and did reveal some surprises. The TPC data rates were significantly higher than expected and the data rates did not match the simulations. This problem was apparent throughout the whole chain and multiple bottlenecks were unveiled during this exercise. The network as well as the processing of the system seemed to be insufficient to digest all the data and required to tackle this problem from different angles.

Table 6.2: Detector data rates during 4 Mhz P-P test without data distribution and without processing [2].

Detector	TPC	ITS	TOF	MFT	FIT	EMC
FLP readout (GB/s)	1660	56.8	3.74	23.6	3.1	0.65
Detector	TRD	MCH	PHS	MID	HMP	CPV
FLP readout (GB/s)	4.36	14.5	-	0.9	-	0.004

Table 6.2 shows detector data rates at 4 MHz pp collisions, without Data Distribution on the FLP and without processing on the EPNs. Incoming TPC data on the FLPs is too high, meaning the detector data rate on a single FLP exceeds the line rate of the InfiniBand network. This is due to the intermediate data format and will be fixed in the future. Without sending the data over the network TPC data rates were 1.6 TB/s. Dropped data on the CRU readout level led to inconsistent and corrupted data, which could not be properly processed and therefore a test with processing was not possible at this interaction rate. These high rate tests will be repeated regularly in 2023 with the final TPC data format. With the new TPC data format, data rates will be comparable to the estimated baseline scenario, on which the computing requirements of the whole

processing chain were based. The numbers from the high rate tests were used to update all estimates, to get more realistic numbers for 50 kHz Pb-Pb. The total TPC data rate between FLP and EPN will be 1.67x of the baseline assumption around 900 GB/s with the new data format. This is still manageable with the current network, since the network demonstrated it can handle 1.2 TB/s from the TPC FLPs, see figure 6.27, and therefore does not need any modifications on the system. This increase in data rate is mostly due to a higher number of non zero suppressed ADC value, which is 1.9x of the baseline assumption. From the processing point of view the algorithms scale linearly with the amount of clusters found in the raw data. Clusters are increased by 1.51x of the baseline assumptions and therefore require 51 % more GPUs. The number of tracks found is also 1.4x of the simulations. The CTF size also increases with the amount of tracks and clusters associated to physics, a precise estimate for this increase was not yet available at the time of writing. From the compute side 30 additional EPN were already purchased, when first indications of bottle-necks did arise, which already adds 12 % processing power. 70 more servers are ordered and to be included in the cluster. This time with AMD MI-100, which are 30 % faster than the current AMD MI-50 as well as more CPU cores and increased memory. These 70 will give an additional compute increase equal to 91 of the previous servers.

In the current state the system was verified to perform with detector data, including processing up to 2 MHz pp. Figure 6.27 shows stable data rates during one of these 2 MHz pp for more than 2 hours, indicating no network or processing bottlenecks. Tests with higher interaction rates which included TPC were hitting network limitations and were therefore not yet stable. This nevertheless already proved the increased network demands can be accommodated with the current system. It also indicates that the cluster is overall performing very well and can be easily modified to fulfil all unforeseen supplemental compute requirements, by extending the EPN farm with additional servers.

6.3 Network Performance

This section describes the InfiniBand network performance of the EPN farm during ALICE operation. Continuous testing with simulated Pb-Pb data during the whole commissioning ensured the whole O2 system is fulfilling the experiments demands. High rate pp tests were creating higher data rates than simulated data and were demonstrating the networks capabilities above the initial design specifications.

6.3.1 FLP to EPN Connectivity

To test the whole data taking chain, including all systems, an infrastructure was established to inject simulated Monte Carlo (MC) data as early as possible into the readout chain. Since the CRUs don't have any memory, where a larger data set could be loaded for replay, the first feasible stage to inject the data is the CRU driver, reading the data from a file. To be as flexible as possible there is a possibility to scale down the rate at which the data is injected. The nominal time-frame rate is 88 Hz, which corresponds to the design data rate injected into the whole system. The replay of MC data can be done

with all detectors included in the test, or just a subset required for a specific test. Since the TPC is contributing more than 90 % of the total data volume any meaningful test regarding the data rates has to include TPC. ALICE labels these kind of tests synthetic runs, referring to simulated data as source. Figure 6.25 shows TPC data rates between FLPs and EPNs for 60 Hz, approximately 68 % of the nominal rate of 88 Hz.

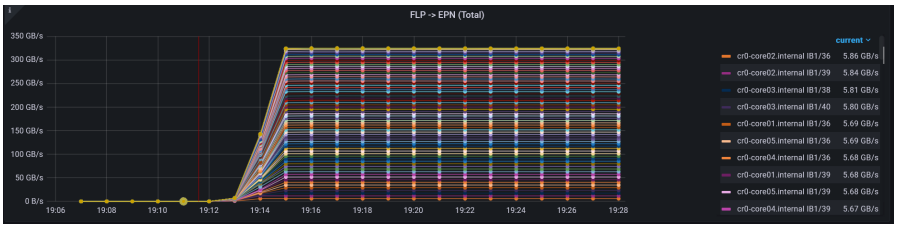


Figure 6.25: Monitoring of the traffic between FLPs and EPNs during a test run with 60 Hz Pb-Pb data replay.

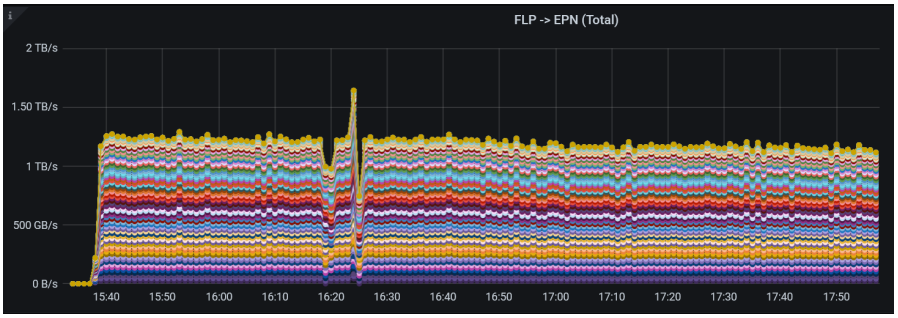
Another way to test the whole system is to inject noise data or physics data that was recoded beforehand. This is usually more effort to set up and requires several detector experts to be involved, compared to the replay of simulated data from the FLPs which is by now fully automated and can be started easily via the Alice Experiment Control System (ALIECS) GUI.

To achieve the required data rates on the EPN side it is important to balance the selected EPNs in a given run over all available InfiniBand ToRs switches. If the first X EPNs would chosen in a run, corresponding to the first $\frac{X}{32}$ building blocks, traffic would not be well distributed over all available links and can create bottlenecks, which reduce the overall throughput between FLP and EPN. Figure 6.26 shows a test run with MC data and is a good example of EPNs selected in a run leading to a decent balanced traffic across the InfiniBand ToRs.

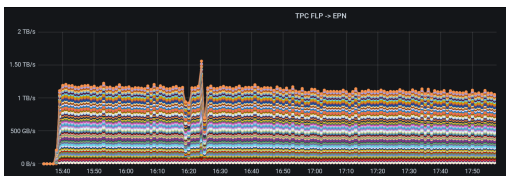
Figure 6.26: Distribution of the traffic from FLPs to EPNs across the InfiniBand ToRs during a Pb-Pb test run with 60 Hz time-frame rate and 200 EPNs (88 Hz nominal).

In this tests three of the ToRs got slightly less data than the remaining five. This can happen in case some processing on the EPNs gets stuck or crashes. Data distribution

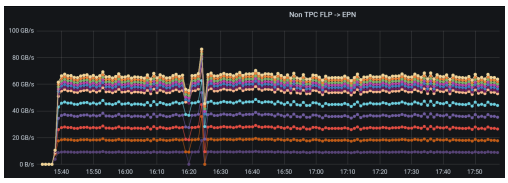
pushes data only to the EPN which have sufficient free buffer space and can handle another time-frame. This can lead to slight imbalances in case a few EPNs get stuck and can not receive more data.



(a) Total Data from FLPs.



(b) Data coming from TPC.



(c) Data from all other detectors

Figure 6.27: Data rates between FLPs and EPNs during a 2 MHz pp test. The dips as well as the spike are a monitoring artefact.

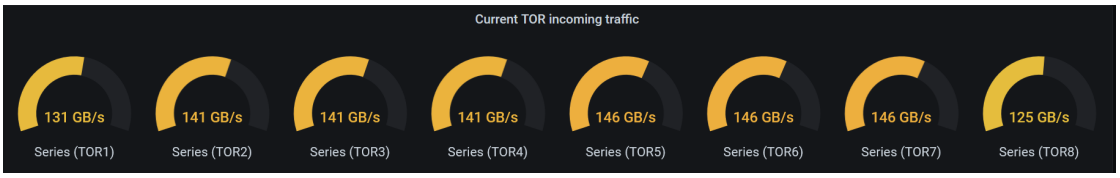


Figure 6.28: Incoming data rate for each building block (line rate to cores 250 GB/s). The last building block got less data during the test.

The network capabilities were verified during the high rate tests, described in Chapter 6.2.2. In case the TPC was included in these tests, the maximum possible rate at which the network could still keep up was 2 MHz pp collisions, generating approximately 1.1 TB/s TPC data towards the EPNs. Multiple tests were done to determine possible network limitations and to understand how far the current installation could be pushed. It was successfully shown that the interconnect between FLPs and EPNs can deliver around 1.25 TB/s using the normal experiment chain, which corresponds to twice the expected TDR value of 635 GB/s. The adjusted expectations with the final TPC data format is less than 1 TB/s for 50 kHz Pb-Pb, which is significantly higher than the TDR value but within the network capabilities. Figure 6.27 shows one of the 2 MHz pp tests, performed in November 2022. For this test only 214 out of 250 EPNs were used. However, the overall achieved throughput was 1.2 TB/s from all included detectors,

approximately 1.1 TB/s from the TPC and around 80 GB/s from all other detectors. Figure 6.28 shows the data flow distribution on the EPN building block level. Traffic is usually almost equally distributed between the building blocks. The last building block has 26 servers instead of 32, since there are 250 servers (instead of 256) and therefore usually gets slightly less data. Processing during this test was enabled, this therefore did not only verify the network but also demonstrated enough processing was available to handle the amount of data, without introducing back-pressure into the chain.

6.3.2 EPN to Storage Connectivity

With a new hashing algorithm available for the gateways, tests were done to verify that the hashing of EPN IPs to the gateway links was improved. These first tests were done with a 3 gateway set-up. Figure 6.29 shows the throughput to storage from the InfiniBand network to Ethernet. The installed theoretical bandwidth is 300 GB/s and the first micro benchmark shows stable throughput of 150 GB/s, so 50 % of the total available line-rate. The expected maximal write speed to the EOS storage instance is 200 GB/s. Internal throughput tests inside the Ethernet network of the Data Centre did show a maximum achievable write rate of roughly 190 GB/s, without crossing from InfiniBand to Ethernet. With the very first test performance was therefore still quite a bit away from the maximum achievable storage rate and further investigation towards tuning and alternatively adding bandwidth through additional gateways was initiated.

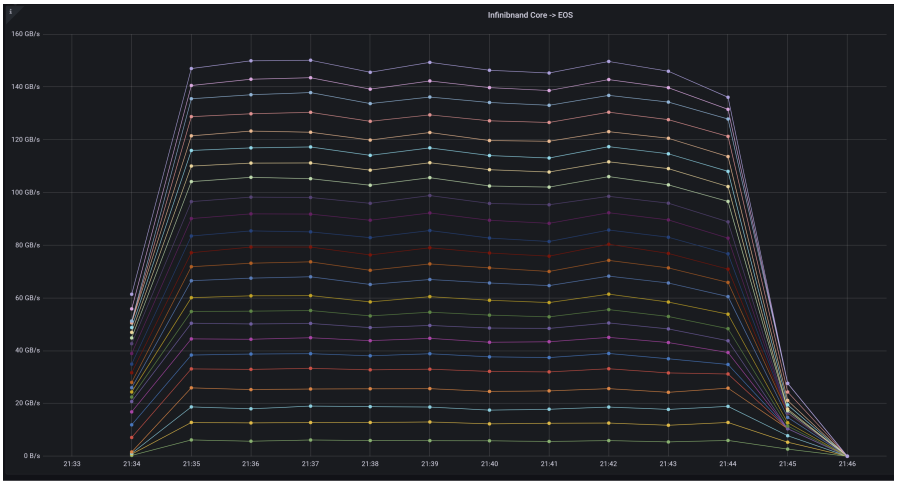


Figure 6.29: Synthetic test with the new firmware 8.1.3050 and the improved hashing, pushing data to EOS

Figure 6.30 shows the traffic balancing over all three gateways as well as the individual links. Balancing challenges with three gateways are described in chapter 5.3.6 gateway link balancing.

The good throughput with three gateways however was only possible with Iperf as micro benchmark. During nominal data taking the achieved throughput is only approxi-

Figure 6.30: Link balancing with the new firmware 8.1.3050 with improved hashing during benchmarking.

mately 100 GB/s with the three gateway setup, as shown during the redundancy testing, see Figure 6.33. With the increased bandwidth demand this is not sufficient for ALICE Run 3. The gateway HA cluster was extended from three to four gateways, to provide sufficient throughput and additional margins.

The throughput limitations seem to be largely due to the amount of parallel streams and congestion towards the storage servers. As soon as there are dropped packages the overall throughput falls off and the link utilization can drop well below 50 %. A possible mitigation in this regard is to limit the allowed throughput per stream and have a well defined number of parallel streams. Since the system is dynamic and data taking runs can consist of a different number of EPNs a static configuration seems to be suboptimal as well. The overall throughout per EPN can also significantly change, depending on the interaction rate of the LHC. A dedicated configuration for pp and Pb-Pb to tune for the very different data rates can provide additional performance improvements. One of the issues to determine the best configuration is the structure of EOS. A dedicated head-node determines how to distribute the clients transferring to the disk servers. This by itself represents another balancing layer in the whole transfers. On top of this the current design sends all the data to a single disk server, which computes the Reed-Solomon encoding for 10+2 redundancy and then transfers 11/10 of the data to other disk servers. There is therefore additional traffic inside the EOS cluster, which can potentially saturate some of the internal links. In case of dropped packages congestion control will quickly reduce the overall throughput.

6.3.3 Gateway Upgrade - Performance with 4 Gateways

To overcome potential bandwidth limitation for the throughput to EOS, the InfiniBand to Ethernet gateway HA cluster was extended from the initially planned three to four gateways. Uncertainties of the actual data rate to disk during Pb-Pb data taking and the wish to have a higher margin in particular in case of any failure were motivating

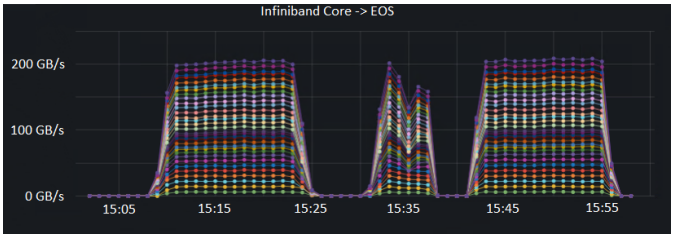


Figure 6.31: EOS throughput via 4 gateways with random generated data reaching storage limit, as presented during the technical board.

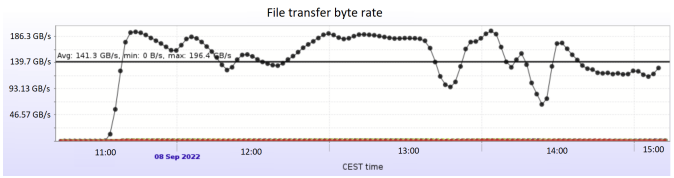


Figure 6.32: EOS throughput rate measured by epn2eos transfer tool during a synthetic run with a large fraction of raw data stored, to push the system to it's limit.

this upgrade. Since the data path to disk is crucial to store experiment data, risk minimization was a priority. Figure 6.31 shows the achieved throughput with four gateways and randomly generated data pushed to EOS. This test was done with perfectly balanced EPN to gateway port distribution with modulo IP hashing. Figure 6.32 shows the achieved throughput using the whole readout machinery, to verify that everything works also in combination with the whole readout chain. The throughput to EOS is in both cases close to the storage limit of approximately 190-200 GB/s.

The four gateway HA cluster successfully demonstrated to deliver the required throughput to EOS, including sufficient margins. The test with four gateways showed that the network was no longer the limiting factor, transferring data up to the maximum write speed of the storage system. Figure 6.32 shows the throughput verification with the full ALICE system, by storing a fraction of the raw data in addition to CTFs.

6.3.4 Gateway Redundancy and Fail-over Testing

The network installation was planned in a way that it would be possible to lose one of the core switches without impacting operations. In the case one of the ToRs would fail, the connected part of the servers will be unavailable until the failing switch is replaced by one of the spares. In case of an EPN ToR failure there is a limited impact on the operations of the ALICE experiment, since all EPNs are equal and a fraction of the processing power is lost. In case of an FLP ToR failure at least a good fraction of detector data is missing, depending on the switch even multiple detectors with lower data rates grouped on a single switch. This complete missing of the data would stop operations in most cases until the switch is replaced with a spare, which just has to be swapped with the problematic one.

For the InfiniBand to Ethernet gateways are configured in a HA cluster and planned for a seamless operation in case of a singe failure. This was tested with a four gateway High Availability cluster by sending an IPMI power off to the BMC of one of the boxes during a run with nominal data rates and look at the potential impact on the ongoing run and the transfer of data to storage in particular. However, there were some potential issues spotted, which require further investigations.

Figure 6.33 shows the data rates from the ALICE alimonitor website [18]. The data rates to EOS look stable at first sight, even though one gateway is missing in the set-up. The test run was started at 11:40:01 and stopped at 12:19:20. There is a reduction in rate visible at the time one of the gateways was suddenly powered down at 11:51, around 10 minutes into the run. The peak towards the end of the run, when the gateway was again booted up fully and back in operation, indicates that there was some local buffering occurring during the test.

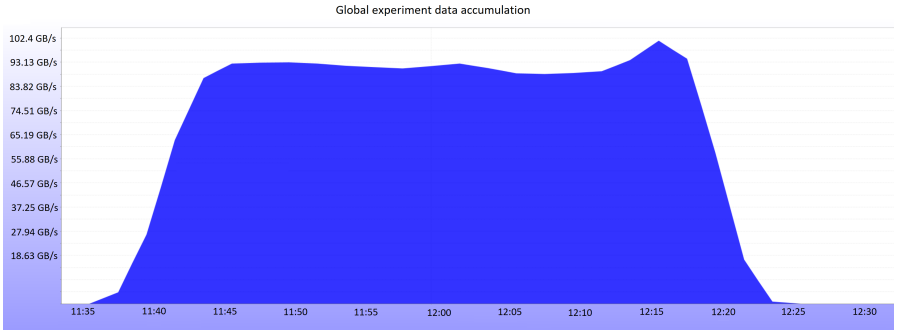


Figure 6.33: Data rates to EOS during a gateway redundancy test run with Monte Carlo Data. One gw powered down via IPMI at 11:51.

Figure 6.34 shows the number of files locally buffered on each EPN. The default settings during a test allows for up to 4 different streams per server, so anything below indicates the data is pushed out faster than it is incoming and the amount of streams is not fully utilized. At the time on of the gateways suddenly dropped out the number of files buffered on the EPNs slowly increased, even though the overall data rate seemed stable.

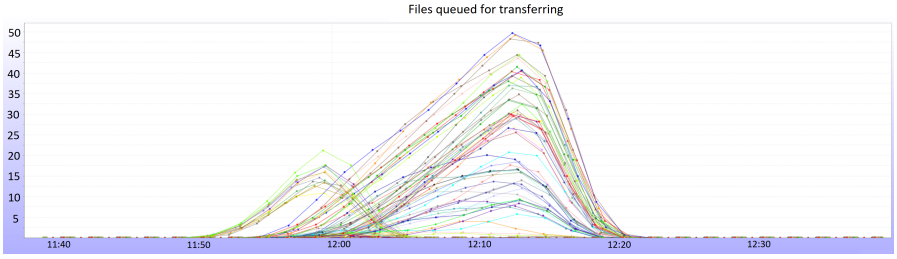


Figure 6.34: Queued files during the gateway redundancy tests as an indication of local buffering of data. One gw powered down via IPMI at 11:51.

The data rates per EPN show the impact of one failing gateway in the most direct way, see figure 6.35. As soon as one gateway drops out some transfers slow down as the distribution over the gateway links changes and things are dynamically reallocated. As soon as the 4th gateway was back again the throughput of the slow EPNs spiked, quickly draining the links. The two EPNs which moved to zero throughput after start of the run got stuck in the processing due to software issues. They were therefore no longer producing data to be shipped toward storage, that was not a network effect but some software issue to be investigated and fixed. In general some of the EPNs have a slightly lower throughput to storage after the gateway is missing. Three gateways are at the limit of being able to reliably transfer approximately 100 GB/s.

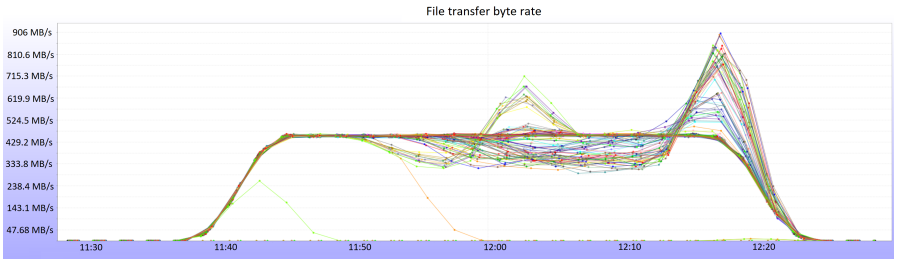


Figure 6.35: Data rates per EPN during the gateway redundancy test. One gw powered down via IPMI at 11:51.

The test was successful as no drop in data rate was encountered and the remaining HA cluster was still able to push the data to EOS without any interruption. Powering on the gateway and integrating it again into the HA cluster also did not impact the operations, traffic via the links was redistributed as soon as the gateway was back. Booting the gateways takes some time, in particular the generation of the virtual interfaces, so there is a delay in the order of 5-10 minutes to get the functionality back. There was still an overall impact and local buffering, so a continuous operation with a three gateway configuration is not possible and will require draining of the local buffers between runs.

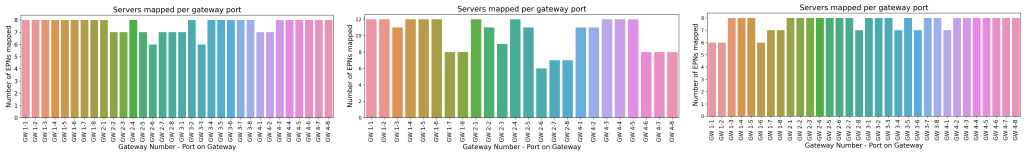


Figure 6.36: Mapping of EPNs per gateway ports during the redundancy tests. Left: distribution before power off of gateway 3. Middle: distribution in three gateway operations. Right: distribution after reboot and restored operation with all 4 gateways.

The mapping of EPNs to the gateway ports looked close to perfect in the four gateway configuration, before the redundancy test started. Not all EPNs were available during this test, which explain the few ports with a lower amount of interfaces mapped. Figure 6.36 shows the mapping before, during and after the power off of gateway number 3.

Figure 6.36 shows the clear unbalances as described in Chapter 5.3.6.

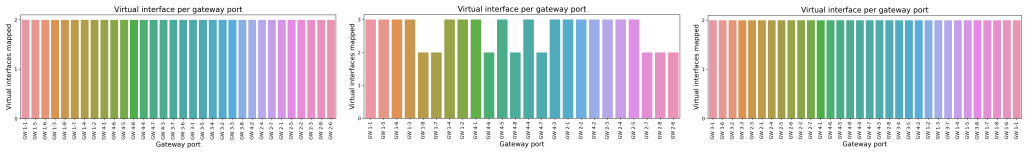


Figure 6.37: Mapping of virtual interfaces per gateway port during the redundancy tests. Left: distribution before power off of gateway 3. Middle: distribution in three gateway operations. Right: distribution after reboot and restored operation with all 4 gateways.

When the 4th gateway is back the distribution is again close to optimal. However, the virtual interfaces are not necessarily mapped to the same links as before and therefore the intrinsic mapping changes. Some EPNs now send via a different link. The overall balance is still maintained with respect to the balancing before powering off one gateway. The mapping of each interface to a dedicated virtual interface does not change during the fail-over procedure. Only the mapping of the virtual interfaces to the individual links is affected.

The redundancy tests clearly showed the balancing improvements with the four gateway HA cluster. The possible throughput is significantly impacted by the imbalance and limits the transfer to storage to approximately 100 GB/s in the three gateway setup. This demonstrates the importance of upgrade to four gateways, to ensure the bandwidth requirements to storage are met reliably and the network does not create any bottleneck.

7 Summary

The commissioning efforts in 2021 did end successfully and the Data Centre (DC), network and the compute farm are fully operational [14]. During 2022 it was demonstrated that everything is ready for data taking and that the overall compute and network performance is as expected. Network tests showed that the network exceeds its design values by a factor of more than 2 and can handle the increased requirements. The Data Centre was up to the challenges of the demanding run concept and did not cause any downtime for ALICE operations and was always operational, except during a forced downtime for the CERN General Emergency Stop (AUG) tests.

Tailoring the computing hardware and the network to the exact software requirements was one of the most important principles to build a working system within the budget constraints. This starts with the Data Centre allowing maximum flexibility in hardware choices and designing a scalable system. It is also reflected in the network design, exactly mapping the experiments technical requirements to the topology. In particular the server choices with multiple GPUs were driven by the software needs and built for a good resource utilization, not wasting memory or compute due to over-dimensioned components being idle all the time.

Since the whole systems scales nicely, it was possible to cope with all the challenges encountered along the way, including additional requirements.

7.1 Data Centre

The Data Centre is fully operational and tuned to the local weather conditions. The four IT-Containers provide a cooling capacity of 525 kW each, amounting to the nominal and tendered cooling capacity of 2.1 MW. The Data Centre therefore provides the required cooling for a HPC cluster. The Data Centre is able to cool up to 1 kW per rack height unit with an average of roughly 600 W per U over the whole container reliably, without creating hot spots. In an 48 U rack 29 kW of IT load can therefore be housed and cooled efficiently.

Some rather minor changes in the control system had tremendous impact on stability of the overall cooling. It was proven once more that the inputs of any control loop are of utmost importance. If the measurements fed into the control system are not reflecting the actual situation inside the Data Centre, e.g. have some delay or any kind of filtering, it is almost impossible to achieve a good steering. This is particularly important when the load can significantly change in a short time. The ALICE Run 3 concept has very pronounced changes in load whenever a run is started or stopped. This can result in a significant load change from idle to full load in a very short time. The idle load of EPN servers is around 500 W, while a full load scenario is roughly 2.7 kW per server. This

can lead to power jump of more than 5x in less than 30 sec. There is some thermal capacity in the container itself, smoothing the transition a bit. However, the cooling system has to react quickly and rather precisely to this kind of changes in the IT load.

The changes done after intensive testing, in particular removing an unnecessary filter from the temperature sensor, to get a more precise and faster response and a much better input to the steering were highly effective. The overall stability increased tremendously, in particular during highly volatile load scenarios with quickly changing IT load as shown in chapter 6.1. The slower measurement, which had a similar behaviour as a moving average filter, led to significant oscillations and was not suitable for cooling the HPC farm.

Data Centre operation was demonstrated through a full year weather cycle. The Data Centre controls are set up in a way that the parameters are robust enough to cope with all weather and load scenarios. There is therefore no need to adjust settings for certain temperature rages, as it was initially required right after delivery of the IT-Containers. From an operational standpoint the increased stability of the controls have a significant impact on the quality of life for the operations team, since it avoids constant monitoring and adjusting of certain thresholds.

The Data Centre infrastructure allowed us to plan with multi GPU servers as Event Processing Node (EPN). The main goal of this integration was to achieve the required compute capacity for the ALICE Run 3 with the limited budget. The main focus was therefore to efficiently use GPUs for as much processing as possible, since the expected cost - performance ratio for the compute farm using GPUs as hardware accelerators in the servers appears optimal for this particular use-case. The achieved integration with respect to the compute capacity was even higher than initially planned. This was possible with the hardware platform of Supermicro, allowing us to put 8 GPUs as well as one network card in one 4U server. All 8 GPUs plus the additional NIC are connected with 16 PCIe GEN4 lanes and therefore provided the maximum possible PCIe bandwidth. This would allow the installation of a 200 Gbit/s InfiniBand HDR HCA, instead of the currently utilized HDR-100 HCA.

The increased integration level of the servers resulted in a much more compact compute farm than initially anticipated 4 years earlier in the TDR. Since the Data Centre was planned for the initial specification of the TDR, with more servers, there is still unused rack space and spare cooling capacity in the new Data Centre, which is not required for the Run 3 production farm. The Data Centre therefore provides significant headroom for additional upgrades.

Currently the installed peak IT load per container is around 75-85% of the cooling capacity of the containers at full load. The Data Centre therefore still provides some margin, to continue operation without any impact on the equipment, e.g. in case of a fan failure or even a complete loss of one cooling unit.

7.2 Servers

The compute performance corresponds exactly to the expectations and the cluster achieves the same performance as the prototype system, with the expected scaling for all servers. The performance was continuously evaluated and verified with synthetic runs. In the context synthetic runs are referring to test runs in which the data is not coming directly from the detectors but injected on all FLPs into the processing chain. This means that simulated MC data is read from a file and then processed by the readout software as if the data would arrive directly from the detectors via the CRUs. For all following parts of the processing chain incoming data is looking the same as if it would be produced directly from the detectors. Regular tests with simulated data are used to confirm that the processing speed is still fast enough, to cope with the expected Pb-Pb data rates. This is important since the software is continuously improved and it has to be ensured that new features don't reduce the overall processing capacity.

In particular the GPU performance is important, since the majority of the reconstruction is done on the hardware accelerators. Continuous testing of new driver versions and cross-checking if the performance would improve or if there are regressions is a regular maintenance task. In particular the Full System Test has proven to be a valuable benchmark for continuous testing, as shown in 6.2. The FST shows that hardware-software co-design leads to a good utilization of all available compute resources. During these single node tests of the software memory utilization is above 90 %, CPU utilization 60 % and GPU utilization 80 %. During the high rate tests the network inbound traffic was approximately 50 % of the line rate on each EPN.

During high rate interaction tests more clusters were found in the data, increasing the reconstruction compute requirements. To compensate 30 additional servers were already added to the cluster, which were seamlessly integrated into the already existing system. Furthermore 70 servers are currently ordered and will be integrated as well.

7.2.1 Adoption to the increased network requirements

The main detectors are completely new in Run 3 and in particular the TPC, which produces approximately 90% of the data has a completely new sensor technology. The simulated behaviour has to be adjusted with the actually observed data, to get the simulation as close as possible to the real detector response. There was a big gap between the simulated and observed data rates at high interaction rates, in particular during tests with pp rates creating an equivalent amount of particles as the expected Pb-Pb collision rate.

The TPC as main data contributor sends more ADC values than expected and therefore increases the network requirements significantly. The currently available TPC data format is bigger than the final data format, as well as the initial format used for early studies of the O2 system. This resulted in limitations during high rate testing, which will only be resolved with the final raw data format. The existing network will be able to handle the increased data rate, with the new TPC raw data format, without any problem and can absorb more than 2x the design value as shown in Chapter 6.3.

7.3 Network

The InfiniBand network is working well and providing the expected performance and even exceeding the throughput requirements defined in the TDR. Since the actual data rates were significantly higher than simulated data rates, the network had to cope with rates more than two times higher than planned. In particular during high rate tests with a preliminary data format data taking ran into network limitations.

The initially envisaged 635 GB/s between FLPs and EPNs was already reached with 650 kHz pp collision rate. This is due to the intermediate data format and will be significantly reduced when the final TPC data format will be used. During tests with higher interaction rates TPC quickly reached 1 TB/s network traffic towards the EPNs. A tuning campaign, to improve network settings and in particular tuning the readout buffer configuration on the FLP side improved the situation significantly. Overall stable throughput rates of more than 1.2 TB/s with TPC alone were achieved. The TPC as main contributor of the data is connected to the backbone with 12 TBit/s and was able to utilize above 80% of the available line rate in a many-to-many transmission schema. The remaining detectors only have 2.2 TBit/s connectivity to the backbone and the traffic there is not very well balanced over the two ToRs.

Stable and fast throughput between FLPs and EPNs is crucial for the whole system. Due to the available memory and buffer sizes transmission delays can lead to buffers running full. Very short spikes can still be de-randomized with the available memory, delays in the order of a few seconds can bring the whole system to a halt, since the extreme data rates immediately fill the buffers. In principle it is possible to recover from this and to drop data in a synchronized manner, to enable recovery. In practice however it is more complicated, since several components get significantly slower with full buffers, so in reality a stop and restart is required, if too many buffers at different stages run full.

RDMA is a particularly useful way to transfer data between FLP and EPN, due to the low CPU overhead, high throughput and low latency data transmission. On top of that InfiniBand RDMA is offloaded to the HCA and does therefore not consume many CPU resources. The Data Distribution [36] [37], uses RDMA as the way to distribute the data from FLPs to EPNs. InfiniBand HDR also provides extremely useful features like adaptive routing, which improved the balanced utilization of multiple switch-to-switch links significantly, since it enables dynamic load balancing between the available links.

The whole system ran into several limitations due to the increased rate. For the TPC FLPs it is possible to get roughly 15 GB/s detector data into memory via the CRU. The network bandwidth per FLP is only 100 GBit/s (or 12.5 GB/s) and it was therefore impossible to ship out all detector data via the network toward EPNs. The server hardware of the FLPs made it impossible to expand on the network, without adding additional servers, due to the lack of additional PCIe slots for a secondary network card. The 16 lanes PCIe gen 3 has an inherent throughput limitation of 128 GBit/s and can therefore not support a 200 Gbit/s network adapter with a single slot. Expanding the network infrastructure was therefore not possible, also due to budget and time constraints. Tuning as much as possible was the only way to temporarily try working around the high data

rates, until this will be solved at the source by using the final and smaller TPC raw data format. Network utilization in this scenario of above 80 % is a very good result and the maximum achievable with acceptable manpower investment. In particular a wide range of potential scenarios makes it very difficult if not impossible to tune for everything and usually requires some trade-off.

7.3.1 EOS Storage Connectivity

Crossing from InfiniBand to Ethernet involved some challenges along the way. In particular during the commissioning and early testing the gateway unavailability was an issue, since it was not yet generally available and we only had a single box in the beginning. Link balancing and unequal link utilization were additional challenges along the way, which were overcome with modified firmware of the manufacturer.

The TDR estimated a data rate up to 100 GB/s towards EOS storage. During the first high rate tests, it was clear that these rates might not be enough. To increase throughput the number of gateways was increased from three to four, to gain additional margins enabling a higher storage rate. In addition there was another tuning campaign, in particular focusing on the balancing via all available links with the manufacturer. This resulted in a solid margin of the available throughput and tests achieved to write up to the storage limit of 200 GB/s with the new setup and were therefore no longer network but disk-bound. During the fail-over tests, powering down one of the gateways roughly 100 GB/s throughput were demonstrated, which corresponds to the design value of the storage connection. In normal operations there is therefore a roughly 2x margin and the system manages to transfer as much as the storage can absorb. In case of a failure the balancing gets significantly worse and the HA cluster looses much more that 25 % of the total throughput. Nevertheless, even in the failure case the initial design rate of 100 GB/s is achieved. Overall, the network towards storage is in very good shape for Pb-Pb data taking.

8 Outlook

ALICE is ready for the first Pb-Pb run at 50 kHz interaction rate, to see the system in its design environment. Extensive testing did show that the EPN project is in good shape and ready to provide the needed computing and network infrastructure. However, due to the unexpected excess in computing requirements compared to the TDR, it was necessary to add additional EPNs to the system. The Data Centre still has plenty of excess cooling capacity for additional extensions of computing resources. The network is currently at the limit of supported nodes and would require extending the core switches and potentially re-cable and rebalance the network for further extensions, due to a lack of free ports. In particular the throughput between FLPs and EPNs is not easily extendible and would require a network re-design and in addition some re-ordering of the Common Readout Unit on the FLP. Further significant changes to the overall system, therefore, come with a large price tag and require significant manpower to implement.

For the Data Centre itself, the power efficiency is one of the points which could still be improved and was investigated recently by a colleague. In particular, during the idle phases, the cooling system had a significant overhead due to the fast-running internal fans, which has already been reduced. More dynamic steering, with a lower base value, can significantly improve efficiency during downtimes between runs. It might be an interesting project to train a neural network to replace the current PID controllers, since other big data centres reported improvements in Power Usage Efficiency with AI controls. PIDs however have proven to be working very well for that particular application and can provide a very good efficiency if tuned correctly.

Regarding the servers themselves it is always possible to upgrade the hardware, with the fast improvements in this area. New server hardware is usually not only faster but also more energy efficient and therefore also lowering the operational costs.

The network advances are still incredible, in particular the available InfiniBand speeds doubled already during the time of the writing. In Q2 2021 the next generation, InfiniBand NDR was available, including ConnectX-7 as the next generation of HCAs. The plans to double the speed again with XDR are in the pipeline. One of the biggest issues with the new network generation however is to feed them properly on the server side, since 16 lanes PCIe gen 5 or 32 lanes PCIe gen 4 are needed to achieve the required throughput into the HCA and the available server hardware to support these is still limited.

The processing largely relies on GPUs and the technological advances in this area are remarkably fast. The compute increases with each generation can improve the processing capacity significantly. However, it is not sufficient to replace a single component, a significant increase in GPU processing capability needs a larger memory to buffer more data, a faster network as well as more CPU cores - to reach a significantly higher

throughput with the software stack.

For the upcoming years of Run 3, the project is focused on an operation of the EPN cluster and contribute to the successful data taking of the ALICE Experiment, for which the infrastructure is well prepared. The Data Centre infrastructure can still be used for Run 4. Since there are no detector changes planned during the next long shutdown the compute requirements will not change significantly for Run 4.

Acknowledgements

This thesis would not be possible without the support of many people. Many thanks to everyone, who supported me on this journey.

I want to take the opportunity to thank my whole family from the depths of my heart for their unconditional support during all these years. In particular for the last years, which were not always easy. This thesis would not have been possible without your strong support.

My deepest gratitude to my supervisor, Prof. Dr. Volker Lindenstruth, for the opportunity to work for the ALICE experiment at CERN. Many thanks for the trust in my work, all the support during the last decade, and the freedom to follow my ideas and to implement them in a big collaboration.

Thanks to Prof. Dr. Udo Kebschull for the recommendation towards the ALICE HLT project, as well as all the support, which enabled this unique opportunity.

I feel fortunate and blessed that I had the opportunity to work for ALICE. This was an incredible experience, which taught me a lot over the last decade. Seeing my work in production for a big experiment is an incredibly rewarding experience. Many thanks to all my colleagues from the HLT project, who taught me a lot, especially in the beginning when everything was new and at times overwhelming. In particular, all the colleagues at FIAS were always very helpful and willing to discuss problems whenever I felt stuck. Despite stressful times, thanks to the great team spirit and mutual support, I still enjoyed the work. Also, a big thanks to everyone from the successor project, EPN, for taking over a lot of my work and allowing me to put everything into writing.

Thanks to the ALICE collaboration and everyone I worked closely together at CERN over the past years, for this great experience. I feel very fortunate for several friendships that have grown over the years.

My gratitude also goes to all the friends, who held me accountable over the years and grounded me, whenever I needed it the most. Thanks for keeping me on track and helping me to keep pushing through some challenging times.

Bibliography

[1] Top 500. *Top 500 list - NOVEMBER 2022*. URL: https://www.top500.org/lists/top500/list/2022/11/ (visited on 04/24/2023).

[2] Alice. "Intensity scan 15/10/22". Internal document, unpublished. 2022. URL: https://codimd.web.cern.ch/SfrEkAg1SeWuvrnTP9-trg?both.

[3] AMD. *AMD Radeon Instinct™ MI50 Accelerator (32GB))*. URL: https://www.amd.com/en/products/professional-graphics/instinct-mi50-32gb (visited on 04/24/2023).

[4] AMD. *Introduction to ROCm Validation Suite Guide*. URL: https://docs.amd.com/bundle/ROCm-Validation-Suite-Guide-v5.3/page/Introduction_to_ROCm_Validation_Suite_Guide.html (visited on 04/24/2023).

[5] P Antonioli, A Kluge, and W Riegler. *Upgrade of the ALICE Readout & Trigger System*. Tech. rep. CERN-LHCC-2013-019, ALICE-TDR-015. Presently we require a LHCC-TDR reference number at a later stage we will fill the required information. Sept. 2013. URL: https://cds.cern.ch/record/1603472.

[6] ASHRAE. *2021 Equipment Thermal Guidelines for Data Processing Environments*. Tech. rep. URL: https://www.ashrae.org/file%20library/technical%20resources/bookstore/supplemental%20files/referencecard_2021thermalguidelines.pdf.

[7] ASUS. *Asus ESC4000 G2S specifications*. URL: https://www.asus.com/commercial-servers-workstations/esc4000_g2s/specifications/ (visited on 04/24/2023).

[8] Automation. *Automation Datacenter Facilities*. URL: https://datacenter.automation.be/ (visited on 04/24/2023).

[9] Automation. "D150153 - technical description SAFE IT V5 - SAFE60W-18R800-528KW". Internal document, unpublished. 2017.

[10] Maria Avgerinou, Paolo Bertoldi, and Luca Castellazzi. "Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency". In: *Energies* 10.10 (2017). ISSN: 1996-1073. DOI: 10.3390/en10101470. URL: https://www.mdpi.com/1996-1073/10/10/1470.

[11] P Buncic, M Krzewicki, and P Vande Vyvre. *Technical Design Report for the Upgrade of the Online-Offline Computing System*. Tech. rep. CERN-LHCC-2015-006, ALICE-TDR-019. Apr. 2015. URL: https://cds.cern.ch/record/2011297.

[12] CERN. URL: https://eos-docs.web.cern.ch/ (visited on 04/24/2023).

[13] Cern. *How an accelerator works*. URL: https://home.cern/science/accelerators/how-accelerator-works (visited on 04/24/2023).

[14] ALICE Collaboration. *ALICE upgrades during the LHC Long Shutdown 2*. 2023. arXiv: 2302.01238 [physics.ins-det].

[15] The ALICE Collaboration. *Addendum to the Technical Design Report for the Upgrade of the ALICE Time Projection Chamber*. Tech. rep. CERN-LHCC-2015-002, ALICE-TDR-016-ADD-1. Feb. 2015. URL: https://cds.cern.ch/record/1984329.

[16] The ALICE Collaboration. *ALICE Figure - 3D ALICE Schematic RUN3 - with Description*. URL: https://alice-figure.web.cern.ch/node/11220 (visited on 04/24/2023).

[17] The ALICE Collaboration. *Evolution of the O2 system*. CERN, 2019. URL: https://edms.cern.ch/document/2248772/1.

[18] The ALICE Collaboration. *MonALISA Repository for ALICE*. URL: http://alimonitor.cern.ch/display (visited on 04/24/2023).

[19] The ALICE Collaboration. *Technical Design Report for the Muon Forward Tracker*. Tech. rep. CERN-LHCC-2015-001, ALICE-TDR-018. 2015. URL: http://cds.cern.ch/record/1981898.

[20] The ALICE Collaboration. *Technical Design Report for the Upgrade of the ALICE Inner Tracking System*. Tech. rep. CERN-LHCC-2013-024, ALICE-TDR-017. 2014. DOI: 10.1088/0954-3899/41/8/087002. URL: http://cds.cern.ch/record/1625842.

[21] The ALICE Collaboration. *Upgrade of the ALICE Experiment: Letter of Intent*. Tech. rep. CERN-LHCC-2012-012, LHCC-I-022, ALICE-UG-002. Geneva: CERN, 2014. DOI: 10.1088/0954-3899/41/8/087001. URL: http://cds.cern.ch/record/1475243.

[22] ASHRAE Technical Committee. *Mission Critical Facilities, Data Centers, Technology Spaces, and Electronic Equipment*. Tech. rep. ASHRAE, 2016. URL: https://tpc.ashrae.org/FileDownload?idx=c81e88e4-998d-426d-ad24-bdedfb746178.

[23] dBVib Consulting. "ETUDE D'IMPACT ACOUSTIQUE D'UN DATA CENTER". Internal document, unpublished. 2017.

[24] dBVib Consulting. "ETUDE D'IMPACT ACOUSTIQUE D'UN DATA CENTER, Campagne de Mesures Acoustiques site STULZ Madrid". Internal document, unpublished. 2018.

[25] U Fuchs et al. *Technical Specification for the Supply, Installation and Commissioning of Container Data Centres (CDC)*. Tech. rep. CERN internal document, unpublished. Sept. 2016.

[26] Gigabyte. *Gigabyte G482-Z51 Specifications*. URL: https://www.gigabyte.com/Enterprise/GPU-Server/G482-Z51-rev-100#Specifications (visited on 04/24/2023).

[27] Stulz Group. *CyberHandler 2*. URL: https://stulztecnivel.com/en/air-handling-units/cyberhandler-2/ (visited on 04/24/2023).

[28] Red Hat. *Ansible community documentation*. URL: https://docs.ansible.com/ (visited on 04/24/2023).

[29] testBernard Hyland. *Impact of Cable Losses*. URL: http://www.maximintegrated.com/an4303 (visited on 04/24/2023).

[30] J Lehrbach et al. "ALICE HLT Cluster operation during ALICE Run 2". In: *J. Phys.: Conf. Ser.* 898.8 (2017), p. 082027. DOI: 10.1088/1742-6596/898/8/082027. URL: https://cds.cern.ch/record/2296653.

[31] Yunqing Li et al. "High-Speed Transmission Cable Performance- Simulations and Measurements". In: *2020 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)*. 2020, pp. 1–3. DOI: 10.1109/NEMO49486.2020.9343442.

[32] Volker Lindenstruth. *O2 – EPN PRR Hardware Selection*. Internal document, unpublished. URL: https://indico.cern.ch/event/937858/contributions/3977407/attachments/2087881/3507828/2020-08-21_EPN_PRR_ALICE.pdf (visited on 04/24/2023).

[33] Ana Lopes and Melissa Loyse Perrey. *FAQ-LHC The guide*. CERN Document Server. 2022. URL: https://cds.cern.ch/record/2809109.

[34] Esma Mobs. "The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018". In: (2018). General Photo. URL: https://cds.cern.ch/record/2636343.

[35] Robert Münzer. "Upgrade of the ALICE Time Projection Chamber". In: *Nucl. Instrum. Methods Phys. Res., A* 958 (2020), 162058. 4 p. DOI: 10.1016/j.nima.2019.04.012. URL: https://cds.cern.ch/record/2712926.

[36] Gvozden Nešković. *Data Distribution and Load Balancing for the ALICE Online-Offline (O2) System*. URL: https://indico.cern.ch/event/587955/contributions/2935761/attachments/1678768/2701788/CHEP18_WP5_O2_Data_Dist_rev1.pdf (visited on 04/24/2023).

[37] Gvozden Nešković. *Design of the data distribution network for the ALICE Online-Offline (O2) facility*. URL: https://indico.cern.ch/event/773049/contributions/3474331/attachments/1937535/3211410/2019-11-05_CHEP19_rev1.pdf (visited on 04/24/2023).

[38] Nvidia. *InfiniBand DAC Cables*. URL: https://www.nvidia.com/en-us/networking/infiniband/direct-attach-copper-cables/#:~:text=InfiniBand%20HDR%20DACs%20reach%20up,can%20reach%20up%20to%204m. (visited on 04/24/2023).

[39] Nvidia. *NVIDIA MLNX-GW User Manual for NVIDIA Skyway Appliance v8.1.5002*. URL: https://docs.nvidia.com/networking/display/MLNXGWv815002 (visited on 04/24/2023).

[40] "Real-time data processing in the ALICE High Level Trigger at the LHC". In: *Computer Physics Communications* 242 (2019), pp. 25–48. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2019.04.011. URL: https://www.sciencedirect.com/science/article/pii/S0010465519301250.

[41] David Rohr. personal communication. June 19.

[42] David Rohr. *ALICE Reconstruction during Run 3 and Synchronous Reconstruction Full System Test*. Internal document, unpublished. URL: https://indico.cern.ch/event/937858/contributions/3945815/attachments/2087839/3511711/2020-08-21_EPN_PRR.pdf (visited on 04/24/2023).

[43] Stulz. "Life Cycle Cost Estimation - Stulz CyberHandler 2, CH2-S4-ADB-SB/RB Customized CERN". Internal document, unpublished. 2018.

[44] Supermicro. *A+ Server 4124GS-TNR (Complete System Only)*. URL: https://www.supermicro.com/en/Aplus/system/4U/4124/AS-4124GS-TNR.cfm (visited on 04/24/2023).

[45] theforeman.org. *Foreman Manual*. URL: https://theforeman.org/manuals/ (visited on 04/24/2023).

Glossary

ACO Acorde (ACO) is one of the detectors of ALICE during Run 2. It was constructed on top of the ALICE magnet, to detect cosmic rays.

ADC Analogue to Digital Converter (ADC) is an electronic component, which converts an analogue signal e.g. a sensor value to a digital value to enable digital processing. In the ALICE context ADC often refers to the digital signal value and not the electronic component.

AHU In this thesis Air Handling Units (AHUs) is used synonymously for cooling units. The AHUs push large quantities of air through the internal air circuit, matching the server air flow inside the Data Centre. The outside air circuit is steered to control the supply air temperture to match the configured set-point temperature. If needed water is evaporated for additional cooling at higher outside temperatures.

AI Artificial Intelligence (AI) is a popular form of machine learning and often refers to learning with neural networks. Due to recent successes in the field it gained popularity and is used in more and more areas to replace traditional algorithms.

ALICE A Large Ion Collider Experiment (ALICE) is one of the four big experiments at CERN located at one of the collision points of the LHC, focusing on heavy ion research..

Alice ECS Alice Experiment Control System (ALIECS) is the interface with which the data taking of the Alice experiment is controlled and steered. It is used to start and stop data taking runs and controls which detectors are included and which software workflow is used.

AlmaLinux AlmaLinux (ALMA) is a free Red Hat Enterprise Linux clone, which was binary compatibel to Red Hat Enterprise Linux. It is one of the new projects to provide a free Red Hat Enterprise Linux version, which started after CentOS Stream was announced.

ALU The Arithmetic Logic Unit (ALU) is a circuit performing arithmetic operations on two operands and providing the result as output. It is one of the core components of CPUs and GPGPUs.

AMD Advanced Micro Devices (AMD) is one of the big CPU and GPU manufacturers.

AMD EPYC 7452 AMD EPYC Processor is the AMD server processor with the Zen architechture, used in the EPN servers.

AMD MI-100 Advanced Micro Devices (AMD) MI-100 GPU (AMDMI100) server GPU used in the EPN servers, with 32 GB memory and a maximum FP32 performance of 23.1 TFLOPs.

AMD MI-50 Advanced Micro Devices (AMD) MI-50 GPU (AMDMI50) server GPU used in the EPN servers, with 32 GB memory and a maximum FP32 performance of 13.3 TFLOPs.

Ansible Ansible is an open source project for configuration management [28]. In the EPN project it is used to ensure the servers are configured to the exact needs, without manually setting up the configuration. Automated configuration ensures that all nodes are configured the same.

AOC Active Optical Cable (AOC) is a fibre cable of a specific length, with directly attached transceivers.

AOD Analysis Object Data (AOD) is the output of the physics analysis of specific CTFs. The AOD contains all data relevant for further physics analysis of the recorded collisions after calibration and reconstruction and is stored persistently on tape.

ARP Address Resolution Protocol (ARP) specifies how network interfaces discover the MAC address address corresponding to the IPs they want to send data to.

ASHRAE The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) is an industry association establishing standards for their field. ASHRAE is releasing guidelines for specific purposes e.g. cooling of data centers or heating and cooling buildings.

ASHRAE A2 The ASHRAE A2 envelope (ASHRAEA2) is referring to an environmental envelope for IT equipment, specifying the working conditions of the equipment. Most Data Centre IT hardware is specified to work in the A2 envelope.

ASIC Application-specific integrated circuit (ASIC) is referencing to a silicon chip which is designed for a specific application and usually highly specialized and optimized. It is usually very efficient for the application it was designed for but has no flexibility to perform different tasks.

ATS Automated Transfer Switch (ATS) is an electrical component which is connected to two power lines and automatically switches to the other line in case the primary line fails.

AUG General Emergency Stop (AUG) (French: Arrêt d'Urgence Général) is a CERN safety procedure to check all emergency power cut installations. This is a planned test, usually once per year, resulting in a series of power cuts to ensure the safety infrastructure is working at all times.

AVX Advanced Vector Extensions (AVX) are vector units able to perform SIMD instructions to a 128 bit vector.

AVX256 Advanced Vector Extensions (AVX256) are vector units able to perform SIMD instructions to a 256 bit vector.

AVX512 Advanced Vector Extensions (AVX512) are vector units able to perform SIMD instructions to a 512 bit vector.

BBR Bottleneck Bandwidth and Round-trip propagation time (BBR) is a congestion control algorithm developed by Google. Compared to the Linux default cubic congestoin contol it ramps up transmission more aggressively when package drop occurs. This can increase overall throughput of links with multiple streams.

BER Bit Error Rate (BER) is a metric to describe the signal quality of a connection. It measures the typical error rate after error correcting measures of $\frac{\text{Bit Errors}}{\text{Transmitted Bits}}$. The lower the number the better the quality of the transmission line.

BMC The Baseboard Management Controller (BMC) is a microcontroller on the Mainboard of a server, which provides remote control capabilities and some basic monitoring for the server. For example it provides usually a remote console, often monitor forwarding, control of the power button, sensor readout for temperatures, voltages etc. as well as a basic health status and a system event log.

CCDB Conditions and Calibration Data Base (CCDB), the data base to store run related information, which is required for reconstruction. This includes environment and detector information, e.g. temperatures, included/excluded sensors and a lot more.

CDC Container Data Centre (CDC) used to describe a modular Data Centre during the ALICE O2 tender process. The modular design concept with containers allows easy scalable installations, with good flexibility.

CentOS CentOS (CentOS) is a free Red Hat Enterprise Linux clone, which was binary compatibel to Red Hat Enterprise Linux before the introduction of CentOS Stream.

CERN Conseil européen pour la recherche nucléaire - European Organization for Nuclear Research is an international research centre located close to Geneva. It is one of the leading physics laboratories.

CPU A Central Processing Unit (CPU) the heart of most computers and servers, as the central and often main processing hardware, executing the program code and returning results.

CPV Charged-Particle Veto (CPV) is one of the detectors of ALICE, helping to identify photons with PHS by suppressing charged particles.

CR0 Counting Room 0 (CR0) is the ALICE internal name of the Run 3 Data Centre. It is located on the surface close to the ALICE experiment at Point 2.

CR1 Counting Room 1 (CR1) is a container hanging in the experiment shaft at P2, above the ALICE experiment. It is located at the first level below ground and was housing the ALICE Run 2 DAQ. During ALICE Run 3 it is home of the FLP farm.

CR2 Counting Room 2 (CR2) is a container hanging in the experiment shaft at P2, above the ALICE experiment. It is located at the second level below ground and was housing the ALICE Run 2 HLT.

CRAC Computing Room Air Conditioning (CRAC) is a cooling unit placed in the computing room of a Data Centre, which uses mechanical cooling e.g. via a compressor to cool down the air of the computing room.

CRAH Computing Room Air Handler (CRAH) is a cooling unit placed in the plenum of a Data Centre, which cools down the air via cooling coils. Cold water is pushed through the cooling coils and fans create a sufficient air flow passing the cooling coils, to cool down the air to the target temperature. CRAH units require external cold water connectivity.

C-RORC The Common Read-Out Receiver Card (CRORC) is a custom PCI card with an FPGA developed for ALICE Run 2. The C-RORC has three QSFPs to connect up to 12 optical links.

CRU The Common Readout Unit (CRU) is a custom PCI card developed by LHCb and named PCI40, with an FPGA supporting up to 48 bidirectional optical links. In the ALICE use case up to 24 input and output links are used.

CTF Compressed Time Frame (CTF) is the output of a processed Time Frame. It does no longer contain any RAW data, only the reconstructed and compressed information e.g. tracks through the whole detector for TPC. The output of multiple TFs can be accumulated in a single CTF file.

CTP Central Trigger Processor (CTP) is distributing timing and trigger information for ALICE.

DAC Direct Attached Copper (DAC) refers to copper network cables with directly attached transceivers.

DAQ Data Acquisition (DAQ) was the ALICE Run 2 equivalent of the FLP and was reading out the ALICE detectors.

Data Centre The Data Centre (DC) refers to the infrastructure to provide the IT operational environment, e.g. power, racks, cooling and in this thesis usually refers to the CR0 CDC.

Data Distribution Data Distribution (DD) is a part of the O2 Software, which is responsible to build Sub Time Frame (STF) on the FLPs, send them to the EPNs and build the full time-frame on the EPN.

DGEMM Double-precision General Matrix Multiply (DGEMM) is often used as a benchmark to measures the compute performance of GPUs by multiplying big matrices and measuring how many operations can be done per second.

DHCP A Dynamic Host Configuration Protocol (DHCP) server assigns Internet Protocol (IP) addresses to the network interfaces of servers.

DNS Domain Name System (DNS) is the protocol to translates server names to Internet Protocol addresses.

DWDM Dense Wavelength Division Multiplexing (DWDM) is used to send light of different colors via the same optical fiber. This way multiple network links can be transmitted via a single fiber with light of different wavelengths, significantly increasing the bandwidth of a single fiber.

ECMP Equal Cost Multiple Path (ECMP) is referring to a network configuration which allows to set multiple routes with equal costs to a target network. This is very useful to route via multiple different switches e.g. for redundancy or load balancing purposes.

ECS The Experiment Control System (ECS) or Alice ECS is the central system to start and stop runs. It is determining the run configuration, e.g. which detectors are included, which software components are running.

EDR Enhanced Data Rate (EDR) is an InfiniBand (IB) generation, providing a maximum throughput of 100G bit/s on a single link.

EMC EMCAL - Electromagnetic Calorimeter (EMC) is one of the detectors of ALICE measuring the energy of particles.

EOS Exabyte Object Storage (EOS) exabyte scalable storage system developed at CERN.

EPN Event Processing Node is an ALICE project, part of the O2 system. As integral part of the data taking chain the EPNs provide the compute resources required to do a real-time event reconstruction, enabling data compression to enable data storage.

Ethernet Ethernet is a network technology widely used for the interconnection of different IT components.

eV electron Volt (eV) is a common energy unit, widely used for energy levels on an atomic scale. It is defined as the kinetic energy of one electron accelerated by a potential difference of 1V.

FDR High Data Rate (HDR) is an InfiniBand (IB) generation, providing a maximum throughput of 56 Gbit/s on a single link. Each link consists of four lanes with 14.0625 GBit/s. The name is referring to the frequency of the lanes.

FEC This abbreviation has two distinct meanings depending on the context. In the detector context FEC Front End Card (FEC) is usually an electronic readout board directly connected to the sensors of the detectors and then forwarding the data through some data link.
In the networking and computing context Forward Error Correction (FEC) describes the encoding of a data stream, which allows the correction of bit errors occurring during transmissions. One example would be the GBT which encodes 80 data bits into 120 bit words and has 32 bit of FEC as overhead, allowing the correction of up to 16 consecutive corrupted bits in each 120 bit frame.

FIAS Frankfurt Institute of Advanced Studies is a non profit foundation focusing on interdisciplinary research. FIAS has close ties to Goethe University Frankfurt and is located at campus Riedberg.

FIT Forward Interaction Trigger (FIT) is a set of ALICE detectors, which can be used as a trigger input for other detectors.

FLOPS Floating Point Operations per second (FLOPS) refers usually to the number of 64 bit floating points operations per second as a indicator of the compute performance.

FLP First Level Processor is one of the ALICE projects part of O2 system. It is the input stage of the experiment, receiving the data from the detectors. On the FLPs the data is aggregated as Sub Time Frame (STF) and then send to the EPNs.

FP64 64 Bit Floating Point Operations (FP64) is the an operation in an ALU with two 64 bit floating point numbers as input and one 64 bit floating point as output.

FPGA Field Programmable Gate Array (FPGA) is a chip which can be (re-)programmed by loading a firmware. This allows a very flexible use, since the algorithm it performs can be modified any time.

FST Full System Test (FST) describes the effort to provide a test-bench for the O2 Software, to test all available processing functionality with simulated (MC) data. It can be a bit misleading, since it is not testing the complete system, but runs everything on a single server. However, it includes the Data Distribution part of Time Frame building as well as the processing itself. It is also used as a benchmark to determine the compute performance of the software.

GBT GigaBit Transceiver (GBT) is a radiation hard optical link developed at CERN.

GEM Gas Electron Multiplier (GEM) is the current sensor technology used for the TPC.

Goethe University As a German university, founded 1914 in Frankfurt am Main Goethe University Frankfurt was named after Johann Wolfgang von Goethe in 1932. It is one of the larger German universities, with about 43.000 students currently.

Google is a large company and technology leader developing several new technologies, known for it's search engine.

GPGPU General Purpose Graphics Processing Unit (GPGPU) refers to the usage of GPU as hardware accelerators for arbitrary problems. Some of the GPUs high performance accelerators do no longer have a video output, and are purely used for computation.

GPU Graphics Processing Unit (GPU) is a hardware accelerator, initially designed to accelerate video output towards one or multiple monitors. GPUs grew significantly more powerful over time and are now also used to accelerate other computations, in particular machine learning applications.

HA High Availability (HA) is a term used in IT for a specific redundancy configuration. In a HA setup usually one component can fail and the remaining part of the configuration takes over seamlessly, without any service interruption.

HBF Heart Beat Frame (HBF) is a sub unit of a Time Frame, splitting a TF in smaller junks. Initially a TF contains 128 HBFs and was now changed to 32 HBFs.

HCA Host Channel Adapter (HCA) is the common abbreviation for an InfiniBand network card and therefore a synonym for an InfiniBand Network Interconnect Controller (NIC).

HDR High Data Rate (HDR) is an InfiniBand (IB) generation, providing a maximum throughput of 200 Gbit/s on a single link.

HIP Heterogeneous Interface for Portability (HIP) is the AMD programming environment for HPC GPU programming.

HLT The High Level Trigger (HLT) is usually referring to a compute cluster which does an online event selection. The HLT is usually an extension of the lower level hardware triggers. For ALICE the Run 2 HLT was used to compress the data and was not actually selecting or discarding any events.

HPC High Performance Computing (HPC) is referring to compute problems which can't be solved on a single server and need a large number of powerful and closely interconnected nodes to solve the problem.

HTC High Throughput Computing (HTC) refers to large compute system which process a high data volume and therefore require good network connectivity.

ICT Information and Communication Technology (ICT) is an extended Information Technology (IT) abbreviation with an emphasis on the communication part.

IDC Integrated Digital Current (IDC) is the sum of the TPC ADC values of the last second.

InfiniBand InfiniBand (IB) fast network with low latency.

Initramfs Initial RAM FileSystem (INITRAMFS) loads the root file system at boot time and allows user space operations early on during the boot process. It can provide configuration parameters for kernel modules, to configure drivers e.g. for GPUs.

Internet Protocol Internet Protocol (IP) describes routing of Internet Protocol addresses through a network.

IPMI The Intelligent Platform Management Interface (IPMI) standard defines the interface between Baseboard Management Controller (BMC) and the server hardware. It is commonly used to remotely control servers, e.g powering on a server or shutting it down.

IPoIB IP over InfiniBand (IPOIB) allows applications to send IP traffic over the InfiniBand network, without using the native InfiniBand verbs.

IT Information Technology (IT) is a widely uses abbreviation for everything that includes information processing.

ITS Inner Tracking System (ITS) is one of the detectors of ALICE, as the name implies it is the innermost detector build around the collision point. It is used to determine the exact vertex of the collision and is used as the seed for further reconstruction and tracking..

LAG Link Aggregation Group (LAG) is referring to a configuration which virtually combines multiple ports of a network to a single logical port.

LEP The Large Electron-Positron Collider (LEP) is the predecessor of the LHC at CERN. LEP was an accelerator colliding electrons and positrons at 209 GeV.

LHC Large Hadron Collider is a big facility in a 27 km tunnel located close to Geneva at CERN to accelerate particles close to the speed of light and to collide them with very high energy at four interaction points. Currently the largest particle accelerator worldwide, reaching the highest collision energies ever achieved so far.

LHCb The Large Hadron Collider beauty (LHCb) experiment is one of the four big experiments at the LHC.

LID Local Identifier (LID) assigned by the InfiniBand subnet manager to define the routing inside the InfiniBand network.

Linpack Linpack Benchmark (LINPACK) is used to measure the FP64 performance of a system by performing matrix multiplications.

LS2 The Long Shutdown 2 (LS2) was planned to be a long shutdown of the LHC, during which upgrades of the machine and the experiments are possible. The initial plan was to have LS2 from the end of 2018 until 2021. Due to several delays during the Covid pandemic, LS2 was extended and the LHC only restarting mid of 2022.

MAC address Media Access Control Address (MAC) is the hardware address of a NIC used to identify connected network card to a switch.

MC Monte Carlo (MC) is in the thesis context mostly referring to Monte Carlo simulations. In a wider context Monte Carlo usually refers to algorithms or processes using random sampling to obtain results. In this context it refers to the simulation of physics events using random seeds for the event and simulating the detector response given the randomized input.

MCH Muon Chamber (MCH) is one of the detectors of ALICE used for Muon detection.

MDC Modular Data Centre (MDC) refers to an modular design approach for Data Centres, which is scalable to the needs of the project. The idea is that modules can be added to the design, to increase the IT capacity. The flexibility often only exists during the design phase, before building the Data Centre and potential extensions already need to be reflected in the planning.

MELLANOX is a company focusing on network equipment. It is the major provider of InfiniBand hardware. MELLANOX recently got bought by NVIDIA and is now rebranded to Nvidia Networking.

MFT Muon Forward Tracker (MFT) is one of the detectors of ALICE in the Muon arm.

MLAG Multi-Chassis Link Aggregation Group (MLAG) is referring to a configuration which virtually combines multiple ports from multiple switches/routers of a network to a single logical port.

NFS Network File System (NFS) is a shared file system, which can be mounted over the network. It therefore provides a common file system for multiple servers.

NIC Network Interconnect Controller (NIC) is referring to the network port(s) of a computer or server. It's a generic term for any controller providing network connectivity and can be cable based or wireless.

NP This abbreviation has two distinct meanings depending on the context. The first Nondeterministic Polynomial time (NP) is a complexity class in computer science complexity theory and is the class of problems for which a nondeterministic algorithm exists which terminates in polynomial time.
The second Neutrino Platform (NP) is a neutrino experiment at CERN.

NRZ Non Return to Zero (NRZ) is a common way to transmit a digital signal. On the electrical level it has two significant voltages representing 0 and 1, usually a positive voltage for 1 and a negative voltage for 0. It avoids to use zero voltage to encode any data, to make sure there is no misinterpretation of a disconnected transmission line.

NUMA Non Uniform Memory Access (NUMA) describes memory locality for multi processor environments. In a system with multiple processors memory modules are connected to a single memory channel on a single CPU. This means that access from the connected CPU is faster than from all other CPUs. Using the memory connected to the CPU for processes running on the same CPU can have significant performance benefits.

NVIDIA NVIDIA manufacturer of GPUs and network equipment. MELLANOX, NVIDIA InfiniBand network is used for the EPN data network.

O2 Online Offline System (O2), ALICE Run 3 project representing all Online and Offline Systems: FLP, EPN and PDP.

O2TDR Online Offline Technical Design Report is a reference to the TDR of the new ALICE O2 system [11] .

offline Offline processing refers to all ALICE processing done after data taking.

OM4 Optical Mode 4 (OM4) is a classification of optical fibres according to ISO/IEC 11801. OM4 fibres are optimized multi mode fibres for 850 nm lasers and used mostly for short range optical connections inside data centres.

Operating System The Operating System (OS) provides an interface to operate the compute hardware. In the EPN project Linux is used as server operating system. CERN uses a free Red Hat Enterprise Linux clone for the majority of it's Grid compute resources. In recent years CentOS was used as a binary compatible clone, before CentOs Stream was introduced. Now there is a shift to Alma Linux.

P8 Point 8 is one of the LHC access points and one of the four points the accelerated particles are collided. It is the location of the LHCb experiment and is at the French part of Cern at Ferney Voltaire, close to the Geneva airport.

PAM4 Pulse Amplitude Modulation 4-Level (PAM4) is a way to encode two bits with a single pulse by using four different signal levels, which are then interpreted as any possible two bit zero and one combination. It is used to increase the data rate, without increasing the frequency of the transmission signal.

Pb-Pb Lead - Lead is the chemical abbreviation of lead and refers to the type of particles accelerated and collided at the interaction points of the LHC. The analysis of lead-lead collisions is the main focus of ALICE.

PCIe Peripheral Component Interconnect Express (PCI) is a fast bus interface to connect components to the CPU. It is the most common interface to connect GPUs and Network cards and provide high throughput toward the main processing unit.

PDP Physics and Data Processing is one of the ALICE projects part of the O2 system. The project is responsible to provide the software required for the online event reconstruction, compression as well as the reprocessing of the data during the asynchronous physics analysis.

PDU In the data center environment a Power Distribution Unit (PDU) usually refers to the distribution of the electrical power inside a rack. It has a similar functionality then a common power strip. In CR0 the PDU is connected to the 400V 3-Phase power board in the container and provides C13 outlets for the servers.

PHS PHOS - Photon Spectrometer (PHS) is one of the detecors of ALICE for photon detection.

PID Proportional Integral Derivative (PID) controller is a feedback loop, able to utilize three different steering parameters to produce a control output.
Proportional P = difference between actual value and target
Integral I = integrated differences between measurements in the past and target over a defined time
Derivative D = rate of change of the actual value with respect to the target in a defined time window.

Pkey The Partition Key (PKEY) is similar to an Ethernet VLAN and allows to partition the InfiniBand network into independent subnets. This can be extremely helpful to isolate parts of the network and get logical separation.

Point 2 Point 2 is one of the LHC access points and one of the four points the accelerated particles are collided. It is the location of the ALICE experiment and is at the French part of CERN at Saint Genis Pouilly.

pp proton - proton refers to the type of particles accelerated and collided at the interaction points of the LHC. Protons and Neutrons are part of the nucleus. A single proton represents the nucleus of a hydrogen atom, without the electron and is positively charged. P-P is the main operational mode of the LHC.

p-Pb Proton - Lead refers to the type of particles accelerated and collided at the interaction points of the LHC. During p-Pb a proton beam is collided with a lead ion beam in the LHC.

PRR Production Readiness Review (PRR) is a CERN procedure before big purchases. Experts from different experiments or projects review the design choices and comment on potential improvements.

PUE Power Usage Efficiency (PUE) is a metric of the efficiency of a system. In the Data Centre context it is specified by the total power consumption of the data center divided by the power consumption of the IT equipment. It is a metric of the overhead for cooling and all the remaining infrastructure.

QGP Quark-Gluon Plasma (QGP) is a state of matter of our universe shortly after the big bang. It requires extremely high temperatures and densities. It acts like a perfect fluid, in which Quarks and Gluons can move freely and are not confined in hadrons.

QSFP The Quad SFP (QSFP) is a connector for four optical links in a single connector.

RAW Raw Data (RAW) is the data delivered directly from the detector in the O2 RAW format.

RDH Raw Data Header (RDH) is the header providing all required information to decode the RAW data. There are multiple versions, the most up to date version is RDH 7.

RDMA Remote Direct Memory Access (RDMA) describes the ability to write directly to the Memory of another server via the InfiniBand network.

Red Hat Enterprise Linux Red Hat Enterprise Linux (RHEL) is a widely used Linux distribution, which is supported by almost all big hardware vendors. It is based on Fedora.

RO Reverse Osmosis (RO) water treatment refers to a method of filtering water with a membrane. The water is pushed through a semi-permeable membrane with high pressure. The water molecules are able to pass, bigger molecules are held back and flushed out with the reject water. With this method one can produce a very good quality of water with an extremely low conductivity.

ROCm ROCm Open Software Platform for GPU Compute (ROCm) is the AMD software platform for HPC and AI.

Run 2 Run 2 (RUN2) is referencing the second running period of the ALICE experiment from 2015 to 2018.

Run 3 Run 3 (RUN3) is referencing the third running period of the ALICE experiment, initially planned for 2021 to 2024 and now shifted and extended from 2022 to 2026.

Run 4 Runr4 is referencing the fourth running period of the ALICE experiment, planned for 2029 to 2032.

SAMPA (SAMPA) ASIC used on the TPC FEC to convert the analog sensor signal to a digital output, which is then sent to the CRU via GBTs.

SHM Shared Memory (SHM) refers to a specific memory segment which is shared among multiple processes, allowing these processes to access the data directly. To move data between processes it is sufficient to forward the pointer to the shared memory address instead of copying the data explicitly to another memory segment owned by the other process.

SIMD Single Instruction Multiple Data (SIMD) is one way to parallelize, applying a single instruction to a vector of input data, e.g. multiplying multiple elements of a vector by a scalar factor and producing one output vector consisting of the individual multiplication results.

Skyway Skyway is a gateway switch from NVIDIA, able to route traffic between an InfiniBand subnet and an Ethernet domain.

SLA A Service Level Agreement (SLA) defines the availability expectations of a service or system. It usually also defines a specific time in which problems have to be resolved. It is common to have penalties defined in case the time to resolve a problem is exceeded. The SLA is often part of a maintenance contract, which assures that all required preventive and corrective measures which are necessary for operations are done.

STF Sub Time Frame (STF) is partial data from a Time Frame (TF), containing only data from a specific detector link. The TF building assembles all Sub Time Frames (STFs) into a complete TF.

Synthetic ALICE Synthetic runs are referring to test runs with simulated Monte Carlo (MC) data. The simulated data is injected into the data taking chain on the FLPs and not coming directly from the detectors. This allows testing the whole system with realistic data rates, without relying on LHC collisions.

TCP Transmission Control Protocol (TCP) is a communication protocol to send messages over a network. It controls the flow of packages and ensures the successful data transmission.

TDR Technical Design Report describes detectors or subsystems of the experiment during the planning phase. After approval the systems are built according to the TDR, usually with small modifications in case of unforeseen circumstances.

TF Time Frame (TF) is a logical delimiter of the data stream for a certain time-window. A TF contains the data of all detectors, which are included in a run. TFs are the data unit on which the online event processing is done, the output of multiple TFs is aggregated in a CTF.

The Foreman The Foreman (TheForeman) is a open source software for life-cycle managment. It provides provisioning, configuration and some monitoring features and can be used to automate cluster tasks [45].

TOF Time of Flight Detector (TOF) is one of the detectors of ALICE, measuring time extremely precisely.

TOP500 The Top 500 List list is an attempt to rank the fastest supercomputers in the world, comparing the compute performance via the Linpack benchmark. It is updated twice a year and in 2022 the first Exascale cluster was listed. Since the cluster operators need to submit the benchmarking results themselves, the list is not complete and misses clusters who don't want to invest the benchmarking efforts or don't want to be listed.

.

TPC Time Projection Chamber is one of the most important detectors of ALICE for particle tracking. A big cylinder filled with gas with a high electric field. Charged particles ionize the gas and the electrons drift to the end plates of the cylinder and are detected by the sensors.

TPU Tensor Processing Unit (TPU) is hardware accelerator, highly specified to perform tensor (vector) computations. With the rising popularity of deep learning and the extensive need to accelerate the training of neural networks TPUs spread quickly, in particular in cloud services.

TRD Transition Radiation Detector (TRD) is one of the detecors of ALICE, located just outside the inner barrel, used for particle identification.

UPS Uninterruptible Power Supply (UPS) refers to a system, which provides power for some time in case of a power cut. Most of the UPS systems are batteries, providing power for a few minutes. It can also be some flywheel, connected to a generator, which keeps spinning and creates power in case of a power cut for a small amount of time. Often the UPS systems are used to bridge the time until a diesel generator is active and can provide power, in case of a longer power cut.

VLAN VLANs are a convenient way to tackle the complexity of growing networks with virtualization. VLANs are used to logically separate network segments by introducing a virtual network layer, which separates a part of the network from the rest. There can be multiple VLANs configured on a single network port, so there is a convenient way to separate traffic for different purposes, e.g. data from monitoring or management.

WDM Wavelength Division Multiplexing (WDM) is used to allow the transmission of multiple signals over a single fibre. Different wavelength of light (different colors) are multiplexed to be transmitted simultaneously and de-multiplexed again on the other side. This technology is widely used in telecommunication services to increase the capacity of expensive long range fibre connections.

Acronyms

ACO Acorde.

ADC Analogue to Digital Converter.

AHU Air Handling Unit.

AI Artificial Intelligence.

ALICE A Large Ion Collider Experiment.

ALIECS Alice Experiment Control System.

ALMA AlmaLinux.

ALU Arithmetic Logic Unit.

AMD Advanced Micro Devices.

AMDEPYC AMD EPYC Processor.

AMDMI100 Advanced Micro Devices (AMD) MI-100 GPU.

AMDMI50 Advanced Micro Devices (AMD) MI-50 GPU.

ANSIBLE Ansible.

AOC Active Optical Cable.

AOD Analysis Object Data.

ARP Address Resolution Protocol.

ASHRAE American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAEA2 ASHRAE A2 envelope.

ASIC Application-specific integrated circuit.

ATS Automated Transfer Switch.

AUG General Emergency Stop.

AVX Advanced Vector Extensions.

AVX256 Advanced Vector Extensions.

AVX512 Advanced Vector Extensions.

BBR Bottleneck Bandwidth and Round-trip propagation time.

BER Bit Error Rate.

BMC Baseboard Management Controller.

CCDB Conditions and Calibration Data Base.

CDC Container Data Centre.

CentOS CentOS.

CERN European Organization for Nuclear Research.

CPU Central Processing Unit.

CPV Charged-Particle Veto.

CR0 Counting Room 0.

CR1 Counting Room 1.

CR2 Counting Room 2.

CRAC Computing Room Air Conditioning.

CRAH Computing Room Air Handler.

CRORC Common Read-Out Receiver Card.

CRU Common Readout Unit.

CTF Compressed Time Frame.

CTP Central Trigger Processor.

CVMFS Cern VM File System.

DAC Direct Attached Copper.

DAQ Data Acquisition.

DC Data Centre.

DD Data Distribution.

DGEMM Double-precision General Matrix Multiply.

DHCP Dynamic Host Configuration Protocol.

DNS Domain Name System.

DWDM Dense Wavelength Division Multiplexing.

ECMP Equal Cost Multiple Path.

ECS Experiment Control System.

EDR Enhanced Data Rate.

EMC EMCAL - Electromagnetic Calorimeter.

EOS Exabyte Object Storage.

EPN Event Processing Node.

ETH Ethernet.

eV electron Volt.

FDR Fourteen Data Rate.

FEC Front End Card.

FIAS Frankfurt Institute of Advanced Studies.

FIT Forward Interaction Trigger.

FLOPS Floating Point Operations per second.

FLP First Level Processor.

FP64 64 Bit Floating Point Operations.

FPGA Field Programmable Gate Array.

FST Full System Test.

GBT GigaBit Transceiver.

GEM Gas Electron Multiplier.

GeV Giga Electronvolt.

GOETHE Goethe University Frankfurt.

GOOGLE GOOGLE.

GPGPU General Purpose Graphics Processing Unit.

GPU Graphics Processing Unit.

GTT Graphics Translation Table.

HA High Availability.

HBF Heart Beat Frame.

HCA Host Channel Adapter.

HDR High Data Rate.

HDR-100 High Data Rate - 100G.

HIP Heterogeneous Interface for Portability.

HLT High Level Trigger.

HPC High Performance Computing.

HTC High Throughput Computing.

IB InfiniBand.

ICT Information and Communication Technology.

IDC Integrated Digital Current.

INITRAMFS Initial RAM FileSystem.

IP Internet Protocol.

IPMI Intelligent Platform Management Interface.

IPOIB IP over InfiniBand.

IT Information Technology.

ITS Inner Tracking System.

LAG Link Aggregation Group.

LDAP Lightweight Directory Access Protocol.

LEP Large Electron-Positron Collider.

LHC Large Hadron Collider.

LHCb Large Hadron Collider beauty.

LID Local Identifier.

LINAC Linear Accelerator.

LINPACK Linpack Benchmark.

LS2 Long Shutdown 2.

MAC Media Access Control Address.

MC Monte Carlo.

MCH Muon Chamber.

MDC Modular Data Centre.

MELLANOX Mellanox.

MeV Mega Electronvolt.

MFT Muon Forward Tracker.

MLAG Multi-Chassis Link Aggregation Group.

NFS Network File System.

NIC Network Interconnect Controller.

NP Neutrino Platform.

NRZ Non Return to Zero.

NUMA Non Uniform Memory Access.

NVIDIA NVIDIA.

O2 Online Offline System.

O2TDR Online Offline Technical Design Report.

OFFLINE Offline processing.

OM4 Optical Mode 4.

OS Operating System.

P2 Point 2.

P8 Point 8.

PAM4 Pulse Amplitude Modulation 4-Level.

Pb-Pb Lead - Lead.

PCI Peripheral Component Interconnect Express.

PDP Physics and Data Processing.

PDU Power Distribution Unit.

PHS PHOS - Photon Spectrometer.

PID Proportional Integral Derivative.

PKEY Partition Key.

P-P proton - proton.

P-Pb Proton - Lead.

PRR Production Readiness Review.

PS Proton Synchrotron.

PSB Proton Synchrotron Booster.

PUE Power Usage Efficiency.

QGP Quark-Gluon Plasma.

QSFP Quad SFP.

RAM Random Access Memory.

RAW Raw Data.

RDH Raw Data Header.

RDMA Remote Direct Memory Access.

RF Radiofrequency.

RHEL Red Hat Enterprise Linux.

RO Reverse Osmosis.

ROCm ROCm Open Software Platform for GPU Compute.

RUN2 Run 2.

RUN3 Run 3.

RUN4 Run 4.

SAMPA .

SHM Shared Memory.

SIMD Single Instruction Multiple Data.

SKYWAY Skyway.

SLA Service Level Agreement.

SPS Super Proton Synchrotron.

STF Sub Time Frame.

SYN Synthetic.

TCP Transmission Control Protocol.

TDR Technical Design Report.

TeV Tera Electronvolt.

TF Time Frame.

TheForeman The Foreman.

TOF Time of Flight Detector.

TOP500 Top 500 List.

ToR Top of the Rack.

TPC Time Projection Chamber.

TPU Tensor Processing Unit.

TRD Transition Radiation Detector.

TTM Translation Table Manager.

UPS Uninterruptible Power Supply.

VLAN Virtual Local Area Network.

WDM Wavelength Division Multiplexing.

Zusammenfassung

Das ALICE Experiment

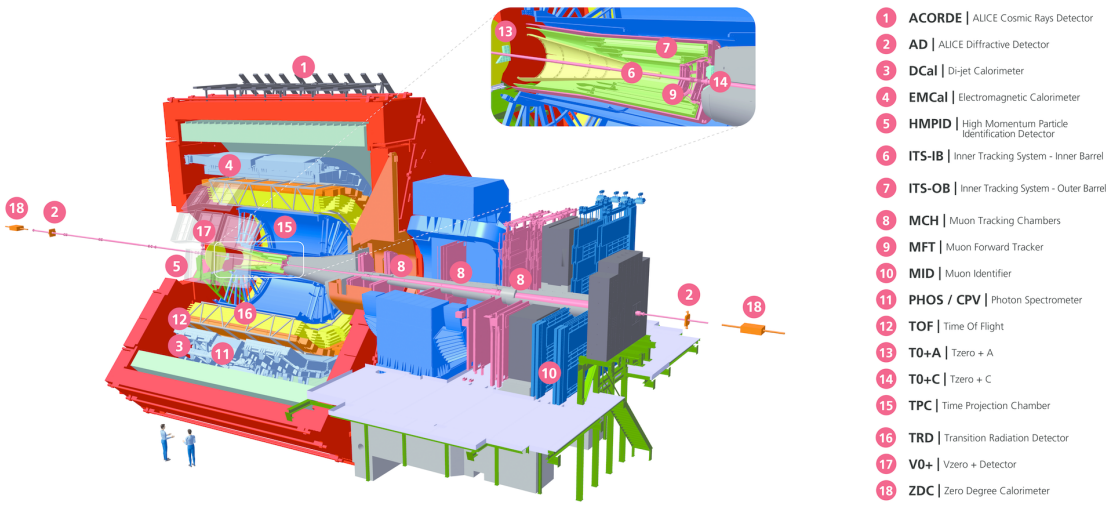


Figure 1: Schematische Darstellung aller Detektoren des ALICE Experimentes im dritten Lauf der Datennahme [16]

ALICE ist eines der vier großen Physik experimente des Teilchenbeschleunigers Large Hadron Collider (LHC) am europäischen Kernforschungszentrum (CERN). Das ALICE Experiment wurde für Studien von Schwerionenkollissionen (Blei-Blei) konzipiert, um die Eigenschaften des Quark-Gluon Plasma (QGP) besser zu bestimmen, ein Zustand von Materie, der kurz nach dem Big Bang vorherrschend war. Der Teilchenbeschleuniger LHC wird in regelmäßigen Abständen für Wartungsarbeiten und technische Verbesserungen ausgeschaltet. Während der zweiten langen Betriebspause wurde das ALICE Experiment verbessert und für eine höhere Kollissionsrate vorbereitet. Hierfür wurden einige Detektoren und die jeweilige Elektronik teilweise oder komplett ausgetauscht. Bild 1 zeigt den schematischen Aufbau des Experimentes mit allen Detektoren, nach den Verbesserungen für die nächste Phase der Datenaufzeichnung (Run 3). Im Vergleich zum vorherigen Lauf des Experimentes (Run 2) werden die Kollissionen nicht mehr von den Detektoren ausgewählt, sondern alle Kollissionen werden kontinuierlich aufgezeichnet. Dies erhöht die aufgezeichneten Kollissionen von 2-3.5 kHz im zweiten Lauf von ALICE auf 50 kHz im dritten Lauf des Experimentes. Diese mehr als 10x Steigerung bedeutet, dass deutlich mehr Daten verarbeitet und gespeichert werden müssen. Die benötigte Netzwerk- und Rechenleistung skaliert mit den Daten und Eventraten und

muss demenentsprechend ebenfalls 10-mal leistungsfähiger werden. Hierfür wurde eine neues Online Offline System (O2) geplant, um die Physikdaten während der Datennahme in Echtzeit (online) verarbeiten und komprimieren zu können (Schema siehe Bild 2). Die Serverinfrastruktur soll ebenfalls für die Physikanalyse im Anschluss (offline) genutzt werden. Das O2 System ist in die Eingangsknoten und die Ereignisverarbeitungsknoten unterteilt. Die Detektoren des ALICE Experiments senden Sensordaten an die Eingangsknoten. Die Eingangsknoten aggregieren die Daten eines bestimmten Zeitfensters und leiten diese dann an einen Ereignisverarbeitungsknoten weiter. Hierbei bekommt ein Server alle Daten aller Eingangsknoten von einem spezifischen Zeitfenster. Die Ereignisverarbeitungsknoten rekonstruieren die Kollissionen, um die Daten komprimieren zu können und unwichtige Daten, etwa von rauschenden Sensoren besser unterdrücken zu können. Die Ereignisverarbeitungsknoten reduzieren das Datenvolumen von ca. 900 GB/s auf ca. 130 GB/s und schicken die komprimierten Daten zum Festplattenspeicher, für eine spätere Physikanalyse [14].
In dieser Dissertation wird der Aufbau eines neuen Rechnenzentrums und der Netzwerk- und Serverinfrastruktur für die Ereignisverarbeitungsknoten EPNs beschrieben, um die gestiegenen Anforderungen für den ALICE Run 3 zu ermöglichen.

Anforderungen an die EPN Recheninfrastruktur

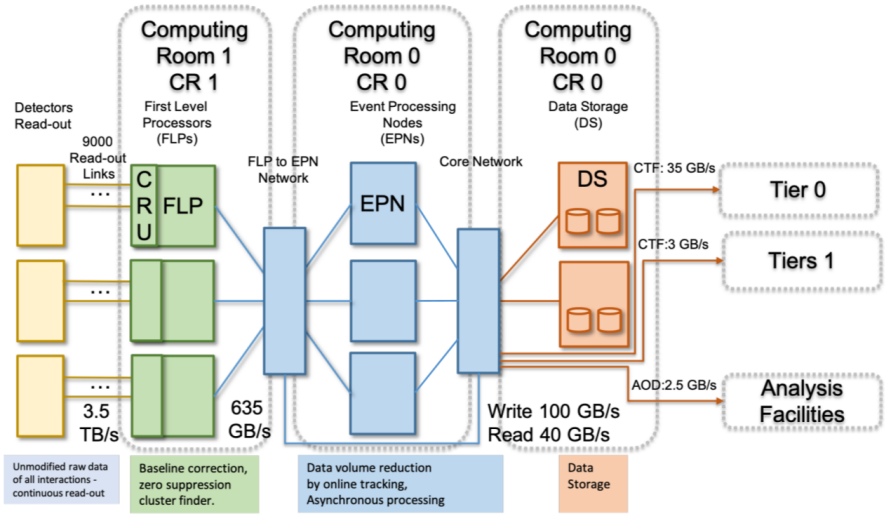


Figure 2: Schematische Übersicht des Datenflusses für das Online Offline System (O2) System im dritten Lauf des ALICE Experimentes [17].

Bild 2 zeigt eine schematische Übersicht des O2 Systems für den dritten ALICE Lauf [17], mit den Schätzungen der Datenraten von 2019. Mit der erneuten Inbetriebnahme des LHC und dem Beginn des dritten ALICE Laufes hat sich gezeigt, dass die Schätzungen aus 2019 nach oben korrigiert werden müssen. Die beobachteten Datenraten mit den neuen Detektoren belaufen sich auf bis zu 900 GB/s von den Ein-

gangsknoten (FLP) zu den Ereignisverarbeitungsknoten (EPN) und bis zu 130 GB/s zum Festplattenspeicher [14].

Die Infrastruktur für den Betrieb des EPN IT Equipments war nicht ausreichend, um die benötigte Leistung der Server zur Verfügung zu stellen. Die notwendige Kühlung, das Platzangebot für Server Racks und das erwartete Gewicht waren mit der Run 2 Infrastruktur nicht möglich. Deshalb wurde ein neues modulares Rechenzentrum ausgeschrieben, um ausreichend Platz für mindenstens 2300 Rack Units IT equipment und mindestens 2.1 MW Kühlleistung zur Verfügung zu stellen.

Die Arbeit beschreibt den Aufbau des Rechenzentrums mit den Herausforderungen die hohe Leistungsdichte der Server zuverlässig zu kühlen. Desweiteren wird der Aufbau der EPN Server Farm beschrieben, der die notwendige Rechenleistung für die Eventrekonstruktion und Datenkompression für den ALICE Run 3 zur Verfügung stellt. Das Netzwerk für den hohen Datendurchsatz spielt eine besondere Rolle, insbesondere der Transfer der komprimierten Daten zum Festplattenspeicher war eine Herausforderung.

Run3 Implementierung der ALICE EPN Anforderungen

Das neue Rechenzentrum wurde mit Containern realisiert. Es gibt insgesammt 4 IT-Container, die jeweils 864 Rack Einheiten und 525 kW Kühlung bereit stellen. Ein zusätzlicher Container stellt die Infrastruktur für Stromverteilung und Wasseraufbereitung zur Verfügung. Der Vorteil einer modularen Container Lösung ist eine hohe Flexibilität und die Möglichkeit einen Großteil der benötigten Infrastruktur bereits vor Lieferung in die Container zu integrieren. Das Rechenzentrum ist nach dem Anschluss von Strom, Wasser, Abwasser und Netzwerk einsatzbereit.

90% der Daten kommen von einem Detektor, der Zeitprojektionskammer TPC. Die benötigte Rechenleistung für die Rekonstruktion der Spuren während der Datennahme beträgt über 90 % der Gesammtrechenleistung. Um die Berechnungen der TPC Rekonstruktion möglichst effizient zu gestalten wurde der überwiegende Teil auf Grafikkarten portiert [17].

Server mit möglichst vielen Grafikkarten sind deshalb ein kosteneffizienter Weg die Rechenleistung für die Eventrekonstruktion des ALICE Experimentes für Run 3 zu realisieren. Für die EPN Farm wurde der 4 U Supermicro Server AS-4124GS-TNR verwendet, mit zwei AMD EPYC 7452 CPUs und 512 GB RAM. Der Vorteil dieses Servers ist, dass 8 GPU mit je 16 PCIe Leitungen unterstützt werden und zusätzlich ein InfiniBand Hostadapter für den Datenaustausch mit 16 PCIe angeschlossen werden kann [44].

Für die benötigte Netzwerkbandbreite wurde ein InfiniBand Netzwerk genutzt. Für den Datentransfer zwischen FLPs und EPNs werden die Daten mit Remote Direktem SpeicherZugriff (RDMA) aus dem Hauptspeicher des FLP direkt in den Hauptspeicher der EPN transferiert. Für den Datentransfer von den EPNs zum Festplattenspeicher werden InfiniBand nach Ethernet Netzwerkübergangsstellen benötigt, da der Datenspeicher im Ethernet Netzwerk von CERN IT installiert ist. Hierfür werden Mellanox Skyway Netzübergangseinheiten verwendet, die jeweils bis zu 800 Gbit/s zwischen InfiniBand und Ethernet transferieren können. Vier Skyways in einem Hochverfügbarkeitscluster

stellen die benötigte Bandbreite zur Verfügung, um mit über 100 GB/s auf den Festplattenspeicher zu schreiben.

Ergebnisse aus dem ersten Jahr von ALICE Run3

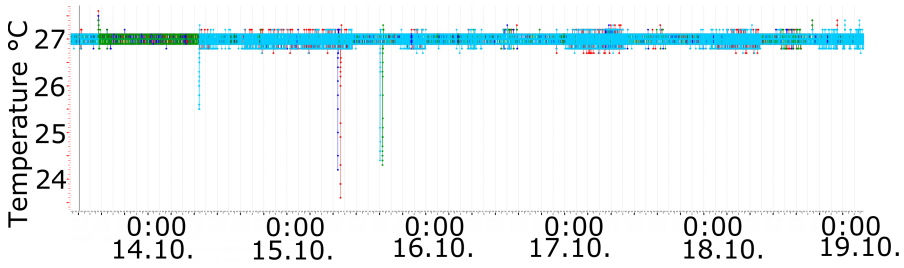


Figure 3: Temperatur profil der Zuluft während einer Reihe von Tests mit hohen Kollisionsraten und somit großen Lastwechseln zwischen Leerlauf und Volllast. Zieltemperatur von $27°C$, Temperaturabfall möglich, wenn ein Wechsel von Volllast auf Leerlauf mit einem Anschalten des adiabatischen Modus zusammenfällt.

Die größte Herausforderung für das Kühlsystem sind neben der hohen Leistungsdichte von bis zu 1 kW per Rack Höheneinheit die großen Leistungsschwankungen im Betrieb. Wenn das ALICE Experiment die Datenaufnahme startet, bekommen alle Server in weniger als 30 Sekunden Daten geliefert und die Rechenleistung ändert sich von Leerlauf in Volllast. Die Leistungsaufnahme der Server springt in dieser Zeit von ca. 500 W im Leerlauf, auf bis zu 2.7 kW unter Volllast. Dies bedeutet ein Sprung in der Leistungsaufnahme um das 5 bis 6-fache und eine entsprechende Steigerung der erforderlichen Kühlleistung. Die Steuerung der Kühlung muss dementsprechend schnell und präzise auf die entsprechende Leistung regulieren. Nach vielen Tests mit Wärme-Dummies und weiteren Optimierungen mit den finalen Servern wurde die Steuerung soweit optimiert, dass sie stabil unter allen Wetter- und Lastbedingungen die Zuluft auf die gewünschte Temperatur reguliert. Bild 3 zeigt die Stabilität der Kaltluft während eines Tests mit hoher Rate.

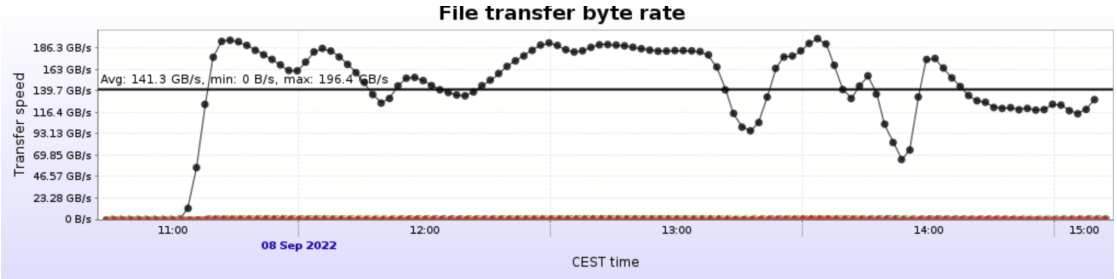


Figure 4: Datendurchsatz von den EPNs zum Festplattenspeicher EOS während eines Tests mit höherem Anteil gespeicherter Rohdaten, um den maximalen Durchsatz zu bestimmen.

Die Netzwerk Bandbreite zum Festplattenspeicher EOS war spezifiziert mit 100 GB/s, mit den höheren Datenraten werden bis zu 130 GB/s erwartet. Bild 4 zeigt den Durchsatz während eines Tests bei dem ein Teil der Rohdaten gespeichert wurde. Es wurden bis zu 190 GB/s erreicht, das Limit des Festplattenspeichers. Der benötigte Durchsatz wird demnach zuverlässig erreicht. Bild 5 zeigt den Durchsatz zwischen FLPs und EPNs während eines Tests mit 2 MHz pp Kollissionen. Die benötigten Datenraten für ALICE Run 3 werden auch hier zuverlässig erreicht.

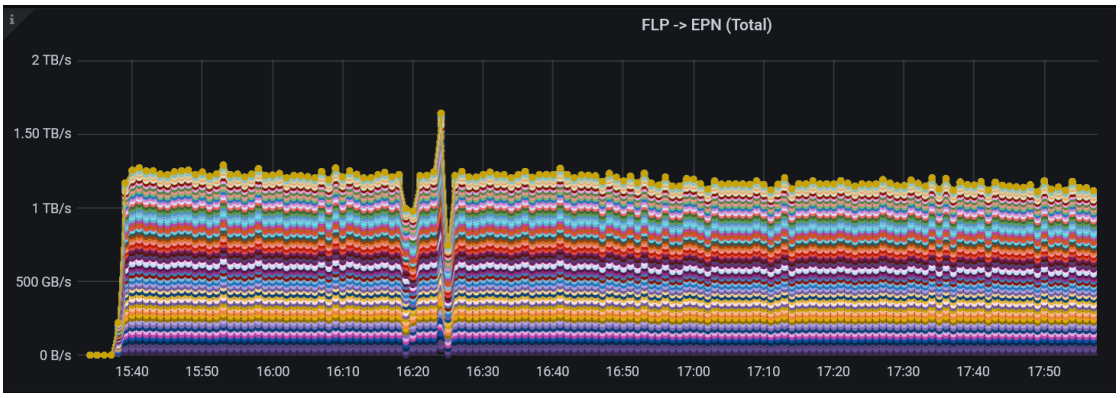


Figure 5: Datenraten zwischen FLPs und EPNs während eines Tests mit 2 MHz Kollissionsrate. Netzwerkdurchsatz ist stabil, die Eingangsdatenrate in den EPN Cluster wird demnach problemlos verarbeitet.

Die EPN Farm kann die ALICE Run 3 Datenraten zuverlässig bearbeiten, ohne Daten zu verwerfen oder den Datenstrom drosseln zu müssen. Tests mit simulierten Daten für 50 kHz Blei-Blei Kollissionen haben erfolgreich gezeigt, dass die EPN Farm den Ansprüchen für die ALICE Run 3 Rekonstruktion genügen. Bisherige Datennahmen in 2022 haben gezeigt, dass einige Annahmen nicht komplett waren und die Datenrate der TPC, die über 90 % der Daten erzeugt, höher ist als erwartet. Die EPN Farm wurde deshalb bereits Ende 2022 um 30 Server erweitert. 2023 werden dem EPN Cluster 70 zusätzliche Server hinzugefügt, um die zusätzlich benötigte Rechenleistung bereit zu stellen.