Computing Plan

August 29, 2019

Executive Summary

This document presents the computing plan for the sPHENIX experiment. Its form follows the computing plans for similar experiments at the LHC in that it aims to present estimates for all aspects of the computing hardware and software development needed to read the data from the detectors, perform calibrations, reconstruct tracks, and to analyze physics results suitable for publication. In response to past reviews of sPHENIX computing, this document provides milestones to be met for sPHENIX to process data using achievable CPU and storage procurements. It also discusses software topics where sPHENIX may benefit from close collaboration with the newly formed Nuclear and Particle Physics Software (NPPS) group at BNL.

The key findings of this plan are that the reconstruction needs of sPHENIX can be met with computing system of 100k CPU-cores and 17 PB of disk in the first year of running, rising to 200k CPU-cores and 40 PB of disk in year three. A capacity of 686 PB of tape is sufficient for the base three year run plan, while a total of 1300 PB of tape accomodates a full five year run plan. Each year of the base sPHENIX run plan provides unique data, so we prioritize reconstruction with a small, fixed latency in part to fully ensure the correctness of data taken by the experiment in a given year. The computing plan assumes an average reconstruction time of 24 sec per minimum bias Au+Au event, a significant improvement over current performance, but reasonable given the performance now being achieved by event reconstruction codes at the LHC. Work on tracking optimization is one task identified for collaboration with the NPPS group. Another task is integration of sPHENIX software with distributed computing tools.

Although sPHENIX has the advantage of a large local infrastructure capable of meeting its core computing needs, we revisit a previous cost comparison between BNL and various cloud-computing options, concluding that the private sector solutions are still too expensive. However, the use of both container-based and grid-capable workflow tools opens up additional resources on the Open Science Grid and High Performance Computing facilities at other national laboratories, especially for the generation and analysis of simulation events.

Contents

Chapter 1

Introduction to sPHENIX Computing

1.1 Physics Goals

All aspects of sPHENIX are driven by the experiment's scientific objectives, and the computing plan is no different. As stated in the original proposal [1] and Conceptual Design Report [2] its goal is to probe the quark-gluon plasma (QGP) created at RHIC at different length scales by measuring jet and heavy quark probes of the QGP medium.

Table 1.1: Run plan for sPHENIX [3].

Year	Species	E [GeV]	Phys. Wks	Rec. L	Samp. L	Samp. L All-Z
year-1	Au+Au	200	16.0	7/nb	8.7/nb	34/nb
year-2	p+p	200	11.5	—	48/pb	267/pb
year-2	p+Au	200	11.5	—	0.33/pb	1.46/pb
year-3	Au+Au	200	23.5	14/nb	26/nb	88/nb
year-4	p+p	200	23.5	—	149/pb	783/pb
year-5	Au+Au	200	23.5	14/nb	48/nb	92/nb

Table 1.1 shows an overview of the sPHENIX run plan. This is described in more detail in [3] and forms the basis for all subsequent data volume estimates. The luminosity numbers in Table 1.1 are based upon projections by the Collider-Accelerator Division for the period spanning 2022–2027 [4]. These projections incorporate appropriate times for cool-down, set-up, ramp-up, and warm-up when changing collision species. The recorded luminosity assumes a vertex cut of $|z| < 10$ cm and a DAQ rate of 15 kHz for all systems, mostly consisting of min-bias triggers for Au+Au and small (1–2 kHz bandwidth) samples of level-1 triggers for p+p and p+Au running. RHIC live-time is assumed to be 60% and sPHENIX live-time is assumed to be 60% for the first two years and 80% thereafter. The sPHENIX proposal provides many examples of jet and heavy-quark projected statistics for

a 20-week Au+Au run under similar circumstances. We note here that for 0–20% central Au+Au collisions direct photon and heavy quark hard processes above a p_T cut of 20 GeV occur for approximately one in 10^6 events, leading to rare-event samples of order 10^5 recorded from the three years of Au+Au running. Accounting for binary collision scaling ($N_{\rm coll} = 770$), the corresponding sampled luminosity for rare-events for p+p and p+Au is lower by factors of 5 and 10 respectively. Anticipated rates for the Y(1S) state are of order 10^3 with reductions by factors of 4 and 7.5 for the 2S and 3S states, respectively.

1.2 Computing Needs

To calculate the computing needs we follow the general approach used to develop the computing models for the LHC Experiments [5]. To varying degrees computing models for the LHC experiments describe the data workflows and provide estimates for computing and storage needs based on event rates and sizes, calibration needs, reconstruction and simulations times, and data size and replication needs. The LHC computing models also specify approximate estimates for effort levels required for software development. Our approach to developing the sPHENIX computing model is similar, but with a few notable exceptions. By design, the LHC relies upon a distributed computing framework, the Worldwide LHC Computing Grid (WLCG), beginning with an initial copy of raw data being sent to distributed sites, whereas the RHIC experiments traditionally rely upon a central computing resource, the RHIC and ATLAS Computing Facility (RACF). The LHC experiments also work with higher data rates requiring specific approaches to merging online and offline tasks for ALICE [6] and significant levels of pile-up for CMS and ATLAS. For sPHENIX, pile-up will also turn out to be an important factor affecting TPC event sizes and reconstruction times.

The following chapters describe specific aspects of the computing model. Chapter 2 describes the sPHENIX model for online computing and calibrations. Chapter 3 describes offline computing and analysis, and the computing model for simulations is presented in Chapter 4. Chapter 5 provides a summary of all CPU, storage, and effort level projections. The use of private sector and opportunistic resources is considered in the Appendix.

Chapter 2

Online Computing

2.1 sPHENIX Data Acquisition

The sPHENIX data acquisition is designed to to achieve a 15 kHz data accept rate with a livetime greater than 90% in a high-multiplicity environment. These estimates are based on the RHIC Collider-Accelerator Department (C-AD) Projections [4]. Compared to the luminosity achieved in 2014, we expect an increase of up to about a factor of two of the rates of interactions which take place within a z-vertex range $|z| < 10\,\text{cm}$ for Au+Au collisions at 200 GeV. The $|z| < 10\,\text{cm}$ vertex is inside the coverage of the sPHENIX tracking system. There are ongoing discussions with C-AD about optimizing the beam conditions for sPHENIX.

While the network and storage system could be configured for higher event rates, the 15 kHz working point has been chosen based on an estimated overall fixed readout time of 6 μs per event and a DAQ livetime target of better than 90%. The livetime is defined as the ratio

$$\text{livetime} = \frac{\text{number of triggers accepted by the DAQ system}}{\text{number of triggers offered to the DAQ system}} \tag{2.1}$$

If the triggers are truly randomly spaced in time, an approximate relationship between the number of triggers offered to the DAQ system (N_{offered}) and the accepted triggers (N_{accepted}) per second is

$$N_{\text{offered}} = \frac{N_{\text{accepted}}}{1 - \tau \times N_{\text{accepted}}} \tag{2.2}$$

where τ is the average duration of the system not being able to accept a new trigger. With $\tau = 6\mu$s and $N_{\text{accepted}} = 15000/\text{s}$, we get a livetime of 91%.

In the case of Au+Au collisions, we expect to record minimum bias triggers mostly (i.e. a

simple interaction trigger), and expect to collect about 100 billion events in a typical 22-week running period. In $p+p$ and $p+A$ collisions, more selective triggers utilize both calorimeter systems, EMCal and HCal.

2.2 The Data Acquisition Mode of Operation

The online computing and calibration procedures need to match the mode of operation of the data acquisition system, and accommodate the structures of files, storage servers, and network bandwidths. We describe the pertinent characteristics of the data acquisition system here.

The sPHENIX data acquisition system organizes the data in *runs*. A typical run typically lasts one hour, but can be shorter or longer, as required. A run is meant to represent an amount of data that can be conveniently analyzed. All controllable conditions must stay the same for the duration of that run. For example, if one was to change the gain setting of a detector, one would end the ongoing run, change the gain, and start a new run. In this way, there is a well-defined point where the new gain setting will take effect.

There are other changes that cannot be controlled, such as the tripping of a power supply, or other, more subtle changes that affect the performance of a detector. If that is a significant change that requires a repair or other intervention, one would again end the run, restore the desired conditions, and then start a new run. However, there are often small changes that can be corrected within a short period of time of about a minute. A typical example is a trip of just one or a few individual channels of a bias supply that only requires a (possibly automated) reset of the channels in question. In that case, one would continue taking data, and merely account for the fact that a certain number of events have been taken under non-standard conditions. This is implemented by defining *luminosity blocks* that last about two minutes. An unusual condition such as a bias channel reset would then invalidate one or more of such blocks. We would exclude a certain amount of data from the later analysis and can account for that loss of events, but the run can continue.

The instantaneous data rate varies over time as the RHIC luminosity changes over the duration of a RHIC store. With the exception of the online and near-line monitoring systems, the online computing processes generally need to accommodate the longer-term *average* data rates, not the peak rates. Table 2.1 shows a breakdown of the envisioned average data rates per subsystem, estimated from HIJING Monte Carlo and plausible expectations for noise. This is the rate that we send to long-term storage, for Au+Au collision in Year-1.

The online and near-line monitoring systems alert the shift crews of significant failures, such as a bias voltage change that changes a channel's gain, on a timescale of about one minute. Other types of error conditions or problems are flagged on timescales of about 5 to 10 minutes. A typical error condition is a too high data volume from a detector system, which can be due to wrong threshold or pedestal values loaded, or due to conditions invalidating those.

The average data rate is the rate at which we send data to the long-term storage system,

Table 2.1: The estimated average data rates from select subsystems in Au+Au collisions at 200 GeV in Year-1. The TPC data rates are proportional to rate of collisions and for Au+Au the rate is projected to grow to 170 GBit/s by Year-5.

subsystem	data size
TPC	100 GBit/s
MVTX	20 GBit/s
Calorimeters	8 GBit/s
INTT	7 GBit/s
	135 GBit/s

which can be as much as 50% lower than the peak rate. Generally speaking, longer averaging times yield lower average values, as this allows us to include the time it takes to set up a new RHIC store, short-term interruptions in data taking, and also the time set aside for RHIC machine development and beam experiments, when sPHENIX does not take data.

The principal mechanism to level the data rate is to temporarily store the incoming data on the *buffer boxes*, and transfer previously taken data to storage (Fig. 2.1). This approach also allows us to ride out short-term outages of the storage system at the RHIC Computing Facility, up to an envisioned duration of about 60 hours. The temporary storage that the buffer boxes provide is designed to provide at least 72 hours of storage for Au+Au data, which have the highest data volumes.

Figure 2.1 shows schematically the readout of the various detectors systems. A number of calorimeter front-end modules send their data, via "Data collection Modules" (DCM2s), to Sub-Event Buffers (SEBs). The equivalent of the SEB for the tracking detectors are the *Event Buffering and Data Compressor* machines (EBDC), which receive the data from the MVTX, the INTT, and the TPC through FELIX FPGA cards.

No further online event building is envisioned. Each SEB and EBDC writes a dedicated file on one of the buffer box disks, resulting in about 60 individual files being written concurrently.

The buffer boxes also provide limited (and managed) access to the most recent data. The standard operation mode will be carried over from the previous PHENIX experiment. It uses three independent file systems on each of the buffer boxes. One set of file systems is being written to by the data acquisition system, while the data from the most recently filled file system are transferred to storage. The remaining file system then holds data that have already been transferred and are available for near-line analysis, event filtering, and calibration processes. When the written-to file system reaches its maximum fill status, the file system with the oldest data is erased, the data from the just-written file system are getting transferred, and the next-oldest data are made available for calibration processes. In this way, the system cycles through the file systems in a round-robin fashion.

Segregating the data accesses in this way prevents performance degradation that would

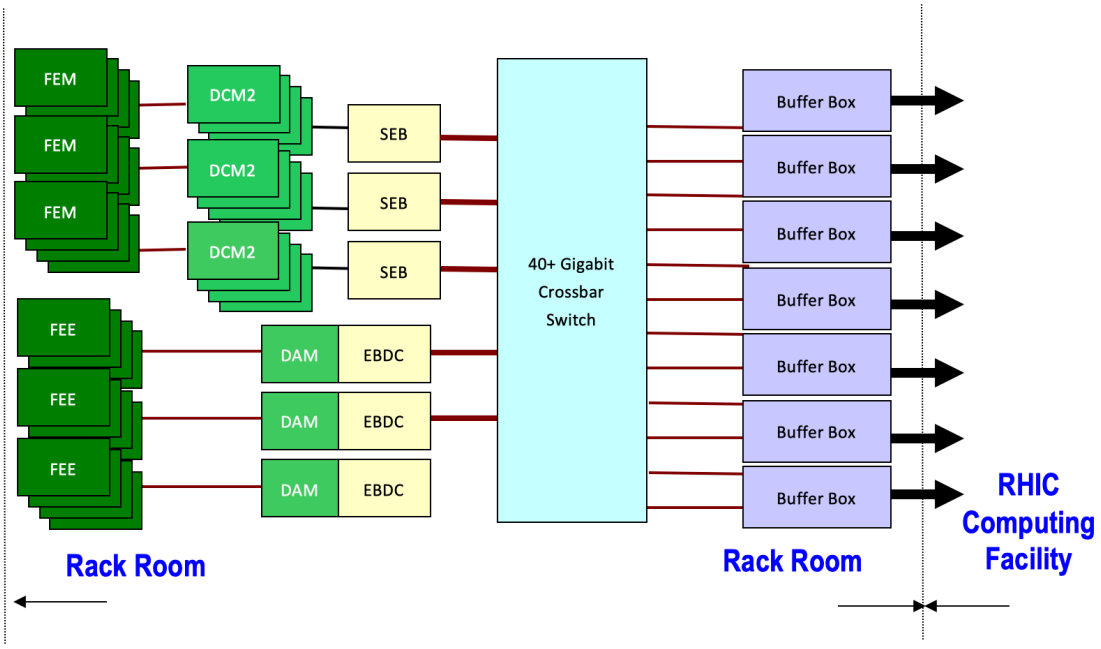

Figure 2.1: Overview of the data acquisition design. The data from the calorimeters and the minimum bias detector (MBD) are digitized in the Front-End Modules and zero-suppressed and packaged in the Data Collection Modules. The TPC, INTT, and MVTX use different front-end electronics that send the data to Event Builder and Data Compressor" (EBDC) computers. The data are then transmitted to the *Buffer Boxes*, from where the data are transferred to a long-term storage system.

occur by unmanaged concurrent write and read accesses, especially to the currently active data acquisition file system.

It is foreseen that all but the aforementioned online monitoring processes run on compute nodes in the RACF. The available network bandwidth is large enough to support the transfer of a second copy of the files in addition to the copy already transferred to long-term storage. This is important because it is not guaranteed that all transferred files can also be made available on general-purpose disks in an efficient manner.

Chapter 3

Offline Computing

3.1 Offline Software

The sPHENIX offline software is built around the Fun4All framework shown schematically in Fig. 3.1. It is a lightweight ROOT based modular system developed by PHENIX which has been in use since 2003. It supports reconstruction, analysis, embedding, and simulations, and it has been the workhorse for PHENIX since its inception. It has processed many PBs of raw data and is now mainly used during analysis, processing PBs/month via the PHENIX analysis taxi [7]. In 2011 the capability to run GEANT4 based simulations was added. In 2015 PHENIX and sPHENIX software became separate efforts followed by a major cleanup and modernization of the framework. Git became the code management system of choice, the code is maintained on github under `https://github.com/sPHENIX-Collaboration/coresoftware`. sPHENIX adopted coding conventions based on the conventions of ATLAS augmented with the lessons learned from the experience with PHENIX software. Software changes are introduced via pull requests. Continuous integration which runs code checking tools (cppcheck, valgrind) and some benchmark analysis verifies the integrity of the pull request. Each pull request is then subject to an informal code review by the repository administrators before it is accepted.

More CPU intensive software checks with insure and coverity take prohibitively long during continuous integration. Those are performed daily using the current software.

sPHENIX employs multiple daily builds of its software for different purposes. In addition, a tagged archival build is done weekly so changes can easily be tracked down to commits from a given week. The libraries are published in cvmfs from which they are available in RACF. For outside use a separate cvmfs volume is available which contains a selection of builds. This enables sPHENIX to use opportunistic resources on the OSG for detector simulations and will open the possibility to run on HPC clusters.

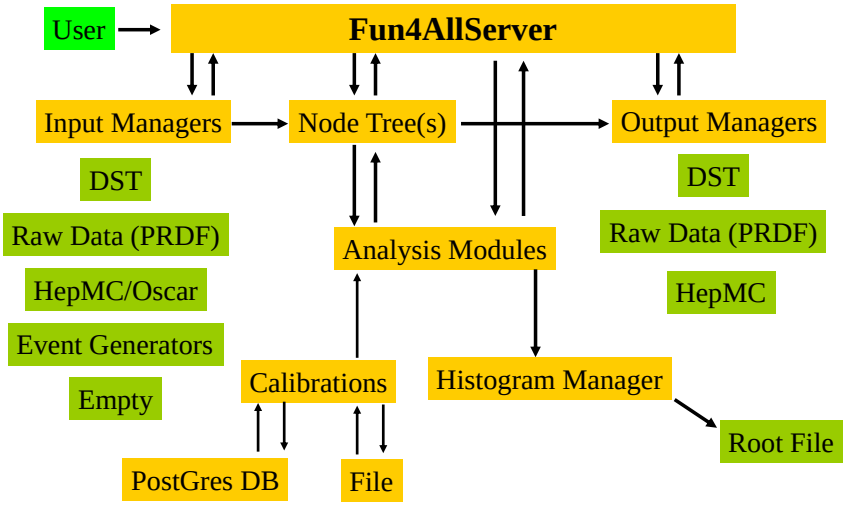

Figure 3.1: Basic structure of the Fun4All framework: The user interacts with the server using CLING macros. It can read/write multiple formats. Reconstruction/Analysis modules are called in the order in which they were registered. Global objects are stored in a tree structure (Node Tree) which can be made persistent in form of a ROOT file. Calibrations can be loaded from a PostgreSQL data base or from files, which enables sPHENIX to use CPU resources which do not have networked database connectivity.

3.2 Offline Event Building

As pointed out in the online section the bulk of the events will not be assembled online but event building will be part of the reconstruction. This approach has two main consequences:

- Due to the large number of streams reconstruction directly from tape is not feasible
- The number of events in a given file is much too high to reconstruct all of them within one job in a reasonable time ($\approx$ 24 hours)

At the current time it is not clear if all files contain the same number of events or files will be written to optimize their size for tape storage/retrieval. PHENIX has been very successful in this regard using tape drives at close to line speed and sPHENIX intends to follow that approach. The design of the offline event builder does not make any assumptions with regard to this. But the events (or in the case of streaming readout, the time slices) are written in ascending order which reduces the complexity of the event building dramatically. A proof of principle that the framework can deal with multiple sources for raw data has been done.

The reconstruction of the data requires a large disk buffer, so all files can be made readily available for the reconstruction. The events will be assembled from the input files and the resulting complete events will be send to clients for reconstruction. This assembly process will handle a complete run and a subset of machines will work exclusively on the reconstruction of this data. After reconstruction the assembled raw event will be discarded. For remote processing the events would be assembled locally at RACF into a file which would then be shipped to the remote site for reconstruction.

3.3 Offline Workflow

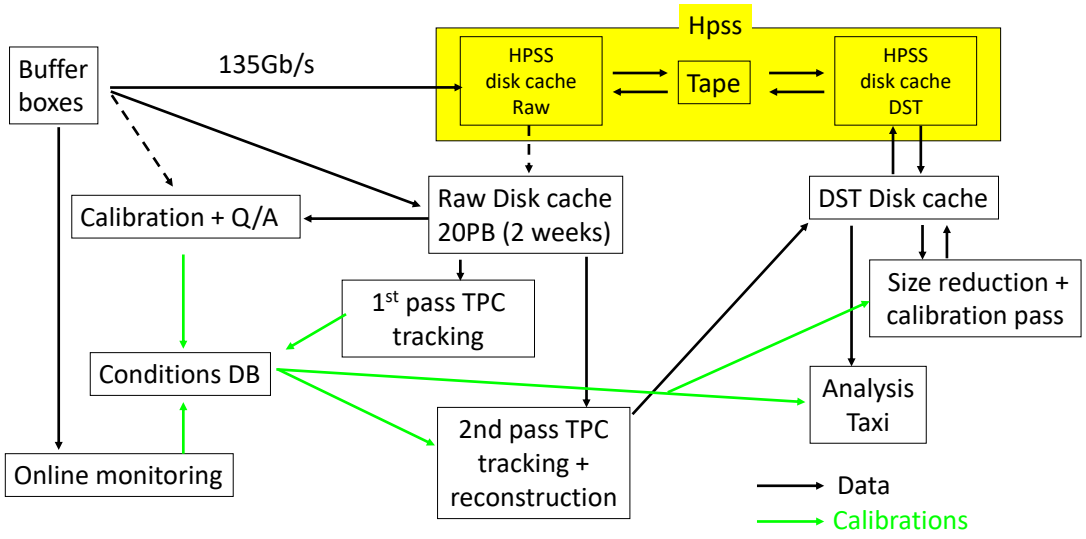

Figure 3.2: Reconstruction flow: In steady state, the buffer boxes will transfer data at a rate of 135 Gb/s to the HPSS based mass storage. Online monitoring processes verify the integrity of the data and flag detector problems. Those processes are running either locally on counting house resources or in RACF, a summary for each run will be stored in the conditions data base. Calibration and QA processes triggered when a run is completely transferred to RACF will extract calibrations and verify the data. Runs which pass this QA and have a complete set of calibrations will be reconstructed using the raw data from the 20 PB raw data cache. The output is stored in a dedicated DST cache and saved to HPSS. A second pass later will reduce the volume of this data. A coordinated analysis approach (Analysis Taxi) will be used to run the user analysis

The most important objective is to store the data safely in the HPSS mass storage. The design shown in Fig. 3.2 makes sure that the immediate reconstruction does not interfere with this. sPHENIX has the bandwidth to transfer the data twice from the counting house to the computing center. This scheme keeps the archiving of the raw data and the reconstruction

separate. It is unlikely that the computing facility provides enough bandwidth to allow simultaneous writing and reading of data. Therefore data which cannot be reconstructed within a short period of time will have to be removed from the buffer space and processed after the data taking has finished.

There is significant uncertainty about the effort and time needed to calibrate the TPC data. For the other detectors the current assumption is that it will take about two weeks to extract initial one time calibrations (e.g. detector alignment, calorimeter tower by tower calibrations). Based on experiences from PHENIX online calibrations of its calorimeters which are needed on a run by run level they can be finished within a day of data taking. The duration of these initial calibrations determines the size of the buffer space needed if sPHENIX wants to be able to reconstruction all data immediately. At the envisioned data rate (for the third year) this translates to about 20 PB of buffer space. During the first year this accommodates about one third of the total predicted data volume which would provide a sufficiently large data sample to refine and finalize the calibrations and reconstruction schemes for the TPC data. Once the data is calibrated up to a level which allows one to run the CPU intensive pattern recognition a first pass over the data will be performed. Enough information will be saved to allow the application of final calibrations. This approach enables the early start of the analysis while final calibrations are being worked out. A second pass where these calibrations are applied is meant to reduce the data volume which will ease analysis passes over the full dataset. Rare events will be copied to separate data streams for dedicated fast analysis. At the current time the final size of the data has large uncertainties and it is not clear if it is possible to keep them disk resident for analysis passes. Given the amount of data sPHENIX will process, we envision the use of an Analysis Taxi approach similar to that employed by PHENIX, where jobs will be combined to run over a given dataset. This enables an optimal use of the computing resources while at the same time allows users to spend their time on analysis rather than computing problems.

3.4 Track Reconstruction

Each subsystem in sPHENIX will have its own computing requirements for calibrations and tracking or cluster reconstruction. We focus here on the TPC calibrations and tracking because they are the primary driver for the offline computing requirements in sPHENIX.

3.4.1 Distortion Correction Calibrations

Most calibrations to be derived during the fixed latency period are non critical in terms of storage space or processing time, like for example detector alignment, calorimeter tower by tower calibrations and gain and thresholds for the silicon based detectors. However a few effects related to the continuous readout scheme of the TPC in conjunction with the high collision rate and consequently high particle occupancy in the detector necessitate special attention.

The high charge load in the TPC under high rate Au+Au running conditions result in significant space charge distortions in the drift volume. Current estimates of the magnitude of these distortions suggest effects of around 3(8) mm at inner(outer) surface of the detector with expected local fluctuation of the expected distortion in the ≈2% range. In order to achieve the 150 μm absolute spatial resolution needed to provide the momentum resolution necessary to fulfill the physics goals of the sPHENIX experiment the average space charge distortion as well as the local fluctuations have to be corrected.

The track distortions averaged over short periods of time, on the order of seconds to minutes, will trace the instantaneous running conditions and can be derived by reconstructing particles and calculating the residuals in the TPC with respect to the trajectories constrained by the inner and potentially outer tracking detectors. The measured residuals will be turned into an average correction map addressing the time local space charge distortions as well as static $E \times B$ distortions and drift velocity variations.

The local fluctuations of the space charge distortions, as experience by the ALICE collaboration has shown, needs to be addressed in reasonably fine time and space granularity, e.g. every 5 ms with the TPC drift volume split into up to 2 million voxels. With a fine granularity like this and the finite interaction rate the recorded particle statistics are not sufficient to derive the corresponding correction. Instead the local charge density in the detector due to ion backflow has to be estimated by monitoring the charge produced in each time slice at the detector readout level. The continuous readout of the TPC allows to monitor these "digital currents" and store them in the raw data stream irrespective of the triggers selecting the events to be stored in the data stream.

During the fixed latency calibration loop the TPC charge frames based on the digital currents recorded in 5 ms time intervals have to be inverted into a correction map of the local fluctuations of the space charge distortions. The inversion process can be reasonably CPU heavy. In the ALICE case of the finest applicable granularity this can take up to 90 sec per 5 ms time frame.

Based on these constraints we propose a three step reconstruction workflow:

- **Initial Reconstruction** Reconstruct charged particle trajectories in the TPC and inner and outer detectors. Calculates residual maps based on the constrained trajectories

- **Calibration Loop** Collect the residual maps and turn them into correction maps for the average corrections for the TPC. Invert the digital current charge maps and build correction maps for the local space charge fluctuations in the TPC. Collect all corrections based on external measurements for all detector subsystems.

- **Full event reconstruction** Reconstruct particle trajectories based on fully corrected TPC clusters. Apply all corrections. Build particle flow candidates based on charged track and calorimetry information.

3.4.2 Charged Particle Tracking

The envisioned track reconstruction algorithm for the sPHENIX event reconstruction is based on the a seeded Kalman filter algorithm comprised of the following steps:

- Local cluster reconstruction in all sub detectors
- Track seeding in the outer TPC layers. Currently a 5-dimensional Hough transform is employed to locate clusters from helical hit patterns in the TPC. Due to run time performance considerations this will soon be replaced by an algorithm based on a nearest neighbor search based on geometric indexing using R-trees for fast cluster access.
- Track seeds are propagated outside-in from the TPC to the inner silicon based detectors by a Kalman filter [8] based pattern recognition algorithm.
- Iteration of the first two steps using looser seeding criteria in subsequent iterations.
- Clusters belonging to the same track are fit using a Kalman-filter-based generic track-fitting toolkit [9], to extract track parameters including displacement at the vertex and the momentum vector at vertex.
- Primary and secondary vertices are reconstructed based on the reconstructed particle trajectories.

3.4.3 Reconstruction Time Estimates

The sPHENIX recontruction code is under rapid development and improvement, so we project final reconstruction performance estimates of the reconstruction time for Au+Au events under high pileup conditions on the concrete experience gained by current CERN experiments faced with similar reconstruction challenges. To estimate the time needed to seed the track reconstruction in the TPC we rely on the performance of the ALICE track reconstruction algorithms designed for the upcoming data taking campaign in Run 3 of the LHC. Employing an algorithm based on cellular automata and a simplified Kalman filter ALICE is able to reconstruct min bias Pb+Pb events in the TPC alone in 1 sec per event using a single CPU core. One min bias event in ALICE consists of 500k clusters on average and ALICE observes a linear scaling of the CPU performance with the number of clusters per event.

In sPHENIX min bias Au+Au collisions with a store average pileup corresponding to an interaction rate of 100 kHz have 400k clusters on average. Assuming an algorithm with similar performance in sPHENIX the track seeding in the TPC should also be able to execute in about 1 sec CPU time. To estimate the time needed to propagate the tracks in the TPC to the inner tracking system and to perform the final momentum we can use an estimate based on the ACTS tracking package. The ACTS package is based on the ATLAS track reconstruction code and using a detector geometry similar to the sPHENIX detector

layout ACTS is able to perform a track propagation through the magnetic filed and detector material in 50 steps (layers) in 0.5 ms per track. Assuming track multiplicity of up to 1500 tracks per event and up to 4 full propagation iterations to fit a given track we estimate that the Kalman track propagation and fitting can be done in about 3 sec per event.

Based on this input we consider a 5 sec average time budget sufficient to reconstruct particle trajectories in the sPHENIX detector. Note that the calibration considerations most likely will require the charged particle reconstruction to be executed twice for each event.

To estimate the full event reconstruction time we consider another 5 sec per event for calibration purposes and an additional 5 sec for the generation of particle flow analysis objects. This estimate is based on experience by the CMS experiment with the CPU time needed to generate particle flow objects is of similar order of magnitude as the pure track reconstruction time. The total average reconstruction time is consequently 15 sec for the full event reconstruction plus 5 sec for initial particle trajectory reconstruction need to derive the space charge distortions.

3.4.4 Data Volumes and CPU resources

At the envisioned data rate (for the third year) this translates to about 20 PB of buffer space. During the first year this accommodates about one third of the total predicted data volume which would provide a sufficiently large data sample to refine and finalize the calibrations and reconstruction schemes for the TPC data. Once the data is calibrated up to a level which allows one to run the CPU intensive pattern recognition a first pass over the data will be performed. The assumption is that a processing speed of 15 sec/evt is achievable, based on experiences by ATLAS and ALICE. Estimates for CPU-cores required to keep up with the data stream at this processing rate are given in Chapter 5. Enough information will be saved to allow the application of final calibrations. This approach enables the early start of the analysis while final calibrations are being worked out. A second pass where these calibrations are applied is meant to reduce the data volume which will ease analysis passes over the full dataset. Rare events will be copied to separate data streams for dedicated fast analysis. At the current time the final size of the data has large uncertainties and it is not clear if it is possible to keep them disk resident for analysis passes. Given the amount of data sPHENIX envisions to use an Analysis Taxi approach similar to what is employed by PHENIX where jobs will be combined to run over a given dataset. This will enable an optimal use of the computing resources while at the same time allows users to spend their time on analysis rather than computing problems.

3.5 Calorimeter Processing

The workflow for processing the calorimeter data is shown in Fig. 3.3. Only active towers' data are recorded in the data stream, which consists of approximately 20% of the EMCal towers and up to 100% of the HCal towers in Au+Au collisions. This leads to 8000 active

towers on average per event. The processing time and storage need are estimated as following:

1. Raw data are first unpacked in memory, producing 16 $\times$ 14-bit ADC samples per active channel. This step is relatively fast, 2.5 μs per tower and 20 ms per event.

2. Then a signal-shape template fit is performed to extract both pedestal and peak amplitude that represents the tower energy. A possible second fit may be performed to separate electromagentic and hadronic components with different time-domain functional forms. In the recent beam tests, such fit with two free parameters of pedestal and peak consumes in average 470 μs per channel, which projects to on average 3.6 s for each full sPHENIX calorimeter event in Au+Au collisions. This step represents the most time-consuming step in the calorimeter analysis, which is likely to be further speed up with look-up tables and fit procedure tuning.

3. Then tower-level calibration is applied to convert the signal amplitude in the ADC units to the tower energy in the EM energy scale, which consumes on average 30 ms per event. The output tower energy are stored in the DST. Recent test with simulated data shows the storage object occupies approximately 50 kb per event after the DST file level compression.

4. Then clustering is performed which costs approximately 30 ms per event using the shower-template cluster finder as imported from the PHENIX experiment. A cluster-level geometry-dependent energy scale calibration will be applied at this stage for the candidate EM-shower clusters, which are expected to be very fast. The storage object of clusters consumes approximately 20 kB per event on the DST for Au+Au collisions.

5. Calorimetry jet finding is the last step of calorimeter-only reconstruction. With the FastJet package [10], this step uses 260 ms in the most challenging central Au+Au events and the storage object consumes less than 8 kB per event. A jet-level energy scale calibration will be applied at this stage too, for which CPU consumption should be negligible. The calorimeter information will also be used as the input the particle flow object and particle flow jet finding too, which is discussed in the last section.

In summary, the calorimeter-only reconstruction stage consumes approximately 4 sec per event in Au+Au collisions, which is dominated by the ADC time-series fitting step. The output is dominated by the calibrated tower objects that occupies approximately 80 kB per event in the Au+Au collisions. Both CPU and storage consumption for the p+p and p+A collisions will be much lower, which approximately scale with the number of active channels passing the zero-suppression in the DAQ.

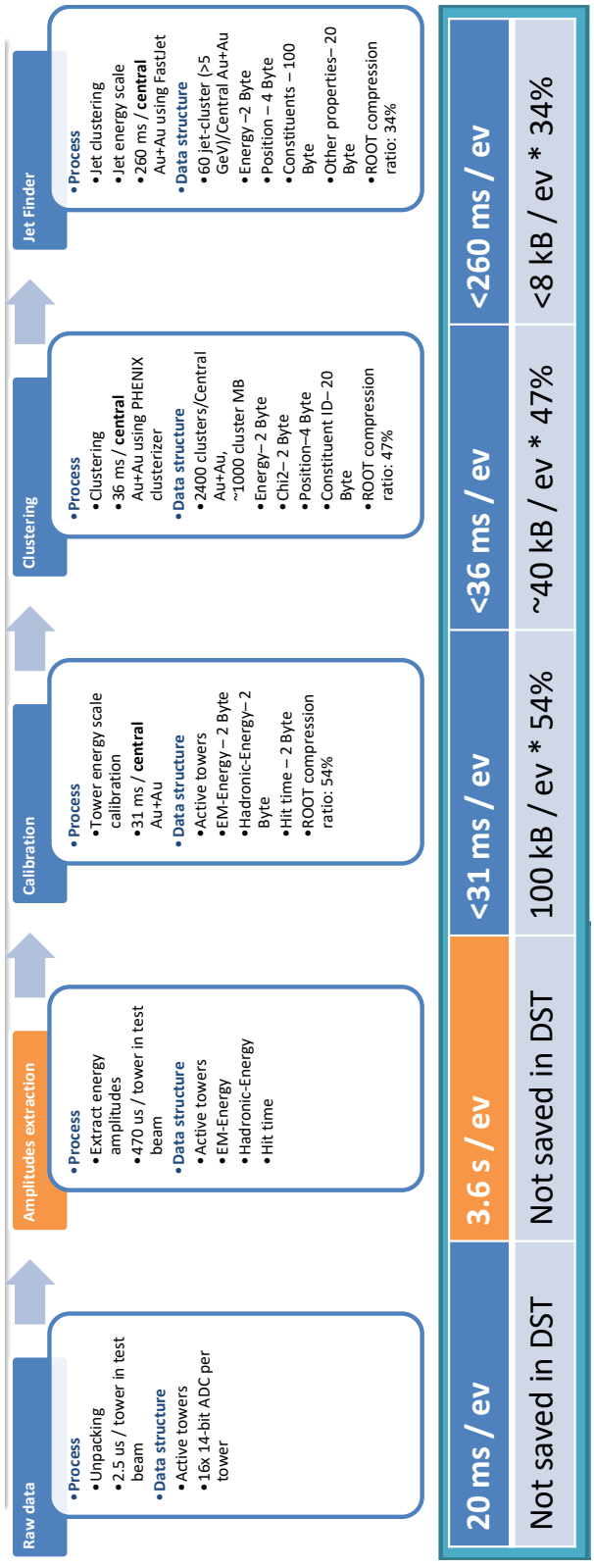

Figure 3.3: Processing times and compression factors for calorimeter data.

Chapter 4

Simulations

4.1 Introduction

As shown in Fig. 4.1, the sPHENIX detector is described in detail in a GEANT4-based simulation and evaluation environment. The simulation function is built into the reconstruction framework as discussed in Chapter 3, which allows flexible configuration of the simulation in the macro-level and present data to reconstruction modules in a consistent way as the unpacked real data. Multiple event generators are supported including PYTHIA, HIJING, and generic HepMC records. Following the GEANT4 tracking, the energy deposition is digitized to detector hits with light production, diffusion, and noise modeling. The resulting hits are used in offline reconstruction and evaluated with the simulation truth information. The geometry and magnetic field in the simulation are also transferred to the reconstruction stage for consistent use in algorithms such as the Kalman Filter [9].

The simulation is validated via multiple generations of calorimeter test beams to control the systematic uncertainties in the simulated calorimeter response [11]. The reproducibility for the calorimeter simulation is checked daily and for each new Pull Request via sPHENIX continuous integration[1], while the QA for tracking simulation is still in development.

In this chapter, we will focus on the sample estimation and the resulting computing needs for sPHENIX simulation.

4.2 Simulation sample requirement

The simulation sample need has been collected from the four sPHENIX physics topical groups (TGs). For each sample, the event category and statistics requirement were estimated based experience with similar studies at RHIC and LHC. Multiple simulation campaigns are planned before and with the sPHENIX data taking to build up analysis expertise with

[1] sPHENIX continuous integration, `https://web.racf.bnl.gov/jenkins-sphenix/` (sPHENIX login required)

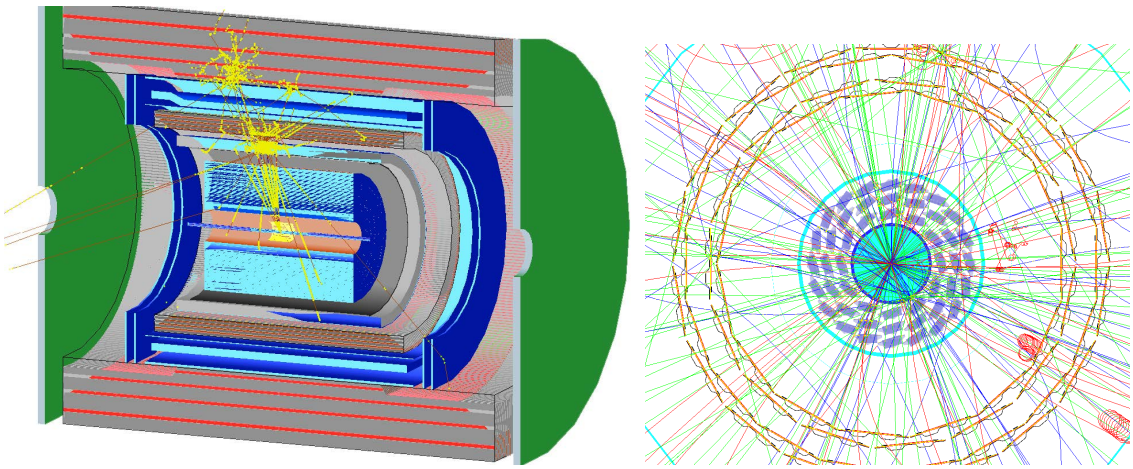

Figure 4.1: Left: GEANT4 event display of a $p_T \sim 30$ GeV/c b-jet showering in the sPHENIX detector. Right: a zoom-in beam view of near the silicon trackers (MVTX in blue and INTT in yellow with detailed support structure implemented) with a full event simulation of $p + p \to D^0 + X$ collision.

the sPHENIX detector, to extract the simulation input to the final analysis (e.g. acceptance and efficiency) and to determine its systematic uncertainties.

4.2.1 Photon and jet studies

For photon and jet studies, two major categories of simulations are required:

- Signal samples that consists of hard-QCD events and their embedding into heavy ion collisions. Their statistics are driven by the precision required to sufficiently quantify the detector effects (such as efficiencies and kinematic resolutions) and allow simulation-based corrections.
- Fake jet background sample that consists of the heavy ion collisions, which are required to determine the fake jet yield and used in the embedding.

This subsection will discuss the requirement for both categories.

4.2.1.1 Signal events

Photon and jet signal events are generated with hard-QCD p+p event generators. Based on past iterations of studies, we plan for 1M jet events in four kinematics slices: 1) minimal biased (no truth jet filtering), and truth jet filtering of 2) 10–20 GeV, 3) 20–40 GeV and 4) $>$ 40 GeV. For photon events, we will plan for 1M events in two photon-truth-p_T slices: 10-30 GeV/c and $>$ 30 GeV/c.

Then we consider the statistics requirement based on the following factors:

- The above signal categories give 6M signal events in each simulation setting.
- Separate samples for p+p and and four Au+Au centralities in 0–10%, 10–20%, 20–40%, and $> 40\%$, and therefore an overall five event categories
- Physics models comparisons, which will be critical in evaluating jet-related uncertainties: for the jet studies we consider PYTHIA vs. HERWIG vs. a quenched MC generator (such as [12]). And for the direct photon we would compare PYTHIA and SHERPA generators. This leads to a $3\times$-multiplier

The above multiplicative factors leads to $6\text{M}\times 5 \times 3 = 90\text{M}$ events per simulation campaign for the photon and jet signal sample.

4.2.1.2 Heavy ion background

The heavy ion collisions has hundreds to thousands of particle per-collision and they take relative long time for full detector GEANT4 simulation ($O(1\text{ hr})$/event). However, it is critical to carry out a substantial simulation of the Au+Au sample for both estimation of rare background, such as the fake jet (this section) and fake track background for embedding studies (next subsection).

The fake jet rate was estimated in a calorimetry-based jet reconstruction study [13] as shown in Figure 4.2. In the most demanding case, i.e. $R = 0.4$ anti-k_T jets in 0-10% most central Au+Au collisions, the expected fake jets will dominate in $E_T < 35$ GeV region. In the rarest fake jet region, $30 < E_T < 35$ GeV, the fake rate is $O(10^{-5})$jet/GeV/event. In order to produce 100 fake jets in this region, we need to simulate 2M central Au+Au events. Considering five-iteration studies with multiple generation of simulation setup, this set the requirement for 10M Au+Au simulation samples. The simulation time need for the Au+Au events in the calorimeters is about 2000 s and the storage is about 0.2 GB/event.

4.2.2 Heavy flavor studies

Heavy flavor (HF) production at sPHENIX is observed in both HF resonances in $p_T \lessapprox 15$ GeV/c and HF tagged jets in $p_T \gtrapprox 15$ GeV/c. The jet sample discussed in the last sub section can be reused for studying HF jet tagger. Meanwhile, the dedicated simulation sample for the HF resonance study would required, that include p+p and p+p embedding events to study acceptance efficiency for the HF signal, as well as to quantify its background.

For characterization for the signal HF samples, we expect 100M PYTHIA p+p event with truth-HF particle event filtering and cross section adjusted with the $\hat{q}$ weighting to the forth order. And for background combinatorial tracks, we expect to use a large single track sample to tune a fast HIJING simulator. The single track sample size are on the order of 100 kinematic bins with 10k events per bin for each of six long-lived particle species,

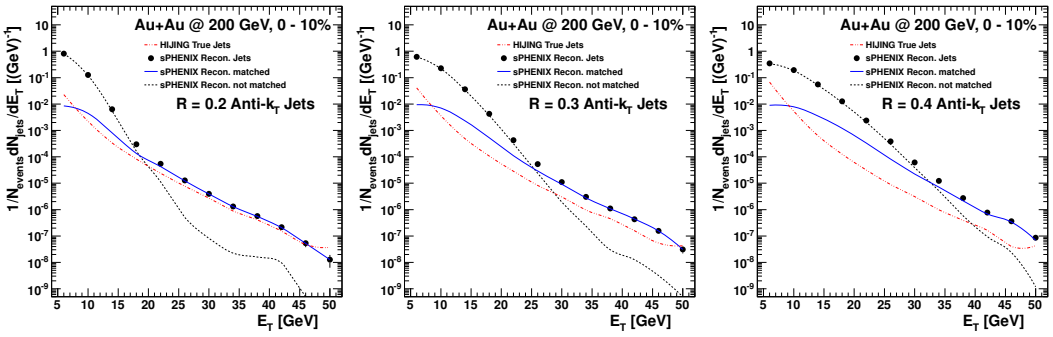

Figure 4.2: E_T-spectra for true HIJING jets (red line) and reconstructed jets (black points) [13]. The reconstructed jets are further divided into those which are matched to a true HIJING jet (blue line) and those which are not matched to a true HIJING jet ("fake jets", black line). Shown are results for 0-10% central HIJING events using anti-k_T jets with $R = 0.2$ (left), $R = 0.3$ (middle) and $R = 0.4$ (right).

i.e. 6M single tracks. These p+p and single track events should also be embedded in Au+Au collisions in four centrality bins, which leads to a factor of four multiplication factor. We plan to have such a data sample well ahead of the first sPHENIX run for analysis exercises and for preparation for initial data analyses. Additional iterations of simulation of equivalent statistics should be planned after each data taking period for the final analysis of sPHENIX data with the simulation setting adjusted for the corresponding run period (e.g. apply the actual dead-map and alignment).

4.2.3 Cold QCD studies

The leading drive for the simulation samples for the sPHENIX Cold QCD studies is the p+p events used in the background study for low-p_T direct photon in the proposed forward detectors (fsPHENIX). Despite the direct photon signals are rare at 10 μb, the softer QCD events at 30 mb could mimic the direct photon signal in the calorimeter response via rare hadronic decay to isolated photons and rare hadronic showers in the EMCal. Quantitatively, for the direct photon below $p_T = 5$ GeV/c, we found softer QCD events with $\hat{q} < 2$ GeV/c still give the dominant contribution to the direct photons up to $p_T \sim 4$ GeV/c. Therefore for lower p_T photon studies, it would be safer to run without the $\hat{q}$ threshold beyond the PYTHIA minimal.

We considerable multi-$\hat{q}$ binning cover from low-p_T to high-p_T background. The Lowest bin has no $\hat{q}$-cut at 30 mb. The number of events is defined by the desired statistics for the background analysis. At higher pT, the total integrated luminosity for the background simulation should approximately matching the direct photon simulation. The following lists the events for each simulation campaign for the whole fsPHENIX setup:

- Simulate direct photon signal (e.g. PYTHIA, $MSEL = 10$) with $\hat{q}$ ¿0 (no threshold), 100M events. With cross section of 10 μb, this sample represents an integrated lumi-

nosity of 10 pb^{-1}.

- Simulate MB QCD events (e.g. PYTHIA, $MSEL = 1$) with $\hat{q} > 0$ (no threshold), 1B events. With cross section of 28 mb, this sample represents an integrated luminosity of 0.036 pb^{-1}

- Simulate MB QCD events (e.g. PYTHIA, $MSEL = 1$) with $\hat{q} > 5$ GeV, 1B events. With cross section of 0.17 mb, this sample represents an integrated luminosity of 6 pb^{-1}

Such study, in particular the background sample generation, could greatly benefit from a well tuned and validated fast simulator as discussed in Section 4.4.

4.2.4 Upsilon studies

For Y signal event characterization, we plan to simulate 3M Upsilon events which represents 100 times that of the data statistics. Then a large sample is planned for simulating its di-electron background that include 100M of HF and Drell-Yan events. Last a sample of single track events are needed for reproducing the combinatorial background that include 1M single particles in six long-lived hadron species and photons. We plan for the above event sets embedded in both p+p pile up background and Au+Au collisions in four kinematics bins that leads to a factor of five multiplication factor. This leads to total of 0.5B events simulatoin.

4.3 Schedule and computing resource requirement

The major simulation productions are grouped in multiple campaigns aggregating overlapping sample from topical groups. The timeline is detailed in Table 4.1 and the sum of resource are listed in Table 4.2. Before the data taking starts, we plan to execute two campaigns providing data sample for building analysis expertise and build the initial correction tables for the preliminary analysis of the sPHENIX data. As suggested by the physics topical groups in Section 4.2, each campaign include 1B p+p, p+p embedding and 0.5B p+A events in the sPHENIX detector as well as 1B p+p events in the fsPHENIX detector configurations. These simulation production are scheduled to complete well ahead of the first sPHENIX run (completes in Q1 2022). After each sPHENIX run, we plan to run another set of simulation tuned to the setting of the corresponding runs for the final analysis.

In the pre-PD2/3 design stages, we have successfully carried out multiple iterations of Million-event level Au+Au simulation and embedding campaign. The result event file are shared among detector and physics study groups for further analysis. For these new large simulation dataset with PBs of output, we plan to execute organized analysis similar to the practice for the experiment dataset, which allows user to submit git-version-tagged analysis module to reduce the simulation DST data for further analysis on user disks. We

also plan to generate and host standard small evaluation file on disk[2] for tutorial and quick validation of basic features.

4.3.1 Milestones

Following the schedule in Table 4.1, the milestones for the simulation are listed below:

1. 2020 First simulation campaign
   (a) Deploying a realistic TPC response simulation
   (b) Final setup review
   (c) 10%-Sample review
   (d) Completion of statistics (use as input for data challenge)
2. 2021 Second simulation campaign
   (a) Deploy speed-optimized tracker simulation
   (b) Final setup review
   (c) 10%-Sample review
   (d) Completion of statistics (use as input for data challenge)
3. 2023 Q4 First p+p to Au+Au embedding production 1B events
   (a) Import production data condition to simulation
   (b) Final setup review
   (c) 10%-Sample review
   (d) Completion of statistics

4.4 Explorations

The primary computing facility for sPHENIX simulation is the RACF. Meanwhile, the simulation software is packaged in an sPHENIX Singularity container[3], which is validated to produce identical simulation results offsite. Therefore, the simulation computing is well suited for opportunistic distributed grid computing, as further discussed in Appendix A.

The primary approach for the sPHENIX simulation is running the full event via full detector GEANT4 simulation, which are expected to produce the most reliable result as validated by the sPHENIX beam tests and global GEANT4 tuning. However, full detector simulation leads to high computing resource consumption as shown in Table 4.2 and to limitations in the rare background samples. Meanwhile, we also plan to explore reliable fast simulation

[2] Example is the jet evaluation NTuple produced by the `JetEvaluator`
[3] sPHENIX simulation via Singularity container: github.com/sPHENIX-Collaboration/Singularity

Table 4.1: Timeline for major central simulation productions

Year	2020				2021				2022				2023				2024				2025			
Quarter	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
sPHENIX Run													AA				pp/A				AA			
p+p collisions [B]		1				1		1											1					
A+A collisions [M]	6				6				6						6								6	
p+A collisions [B]				0.5																0.5				
p+p →Au+Au embed [B]			1				1									1								1

Table 4.2: Simulation resource use summary

Collision System	Au+Au	p+p	p+p embedding	p+A
Events	30M	4B	4B	1B
CPU load [s/event]	2000	100	100	300
Storage [MB/event]	200	1	2	2
Total CPU [10^4-core$\times$day]	70	460	460	350
Total storage [PB]	6	4	8	2

framework and algorithms. We consider both traditional fast simulation framework, such as DELPHES3 [14], as well as recent development of the Machine Learning-based algorithms, such as CaloGAN [15] and DijetGAN [16]. We estimate this work would require an additional 0.5 FTE to the computing effort.

Chapter 5

Projections

In this chapter we summarize all projections that are part of the sPHENIX computing model, including computing, storage, and effort. Section 5.1 summarizes all assumptions needed, such as event sizes and reconstruction time. This includes s those assumptions that have been specified previously in the document. The calculations are presented in Section 5.2.

5.1 Assumptions

Average events sizes are listed in Table 5.1. The event sizes year are slightly reduced due to detectors that are not yet included. The DAQ rate is assumed to be 15 kHz for all collisions.

System	Size (MB)
p+p	1.6
p+Au	1.4
Au+Au$_{\mathrm{yr1}}$	1.7
Au+Au$_{\mathrm{yr3}}$	2.0
Au+Au$_{\mathrm{yr5}}$	2.3

Table 5.1: Average size for raw event for p+p, p+Au, and Au+Au.

Table 5.2 shows the reduction factors for each file type as well as the number of copies that will need to be stored on tape and disk. The number for PDRF disk copies is set to 0.1 to represent the expectation that raw data files will each be buffered on disk for approximately two weeks out of a 20-week run. The table also includes an entry for the number of versions, for example it is anticipated that each DST will exist in an initial and final version. We assume that there will be a highly compressed (by at least a thousand) post-DST format that has not yet been specified. For now we assume a factor of a thousand compression and ten versions for the various flavors and revisions that will be produced by the physics working

groups.

File type	reduction factor	versions	tape copy	disk copy
PRDF	1	1	1	0.1
DST	0.3	2	1	0.1
post-DST	0.01	5	1	1

Table 5.2: Data size reduction factors and number of copies stored on tape and disk.

5.2 Calculations

Table 5.3 shows the tape and disk storage needs for sPHENIX in units of petabytes All projections scale linearly with the projected data rates. Each data type is multiplied by the reduction factor and number of versions and copies given in Table 5.2. Projections for simulations and calibrations are not yet included.

Storage (PB)	year-1	year-2	year-3	year-4	year-5
Raw data	80	144	192	154	221
Disk storage	17	30	40	32	46
Tape storage	132	238	317	253	364
Tape cumulative	132	369	686	940	1304

Table 5.3: Storage needs

Table 5.4 shows current estimates for sPHENIX computing requirements. As with the storage projections, all results scale linearly with projected data rates multiplied by processing times for each system. As described above, we assume a rate of 15 sec per core for track reconstruction for a minimum bias Au+Au event, and 5 sec for calibration. We also assume that reconstruction will keep pace with data collection, implying that a 20 week run will be reconstructed in 20 weeks with a small time-lag. There is no reason to reconstruct at a rate that exceeds the rate of data collection, and a rate that is slower by more than a factor of 2 would not complete by the next start of next beam. Reconstruction times for p+Au and p+p are obtained by scaling the Au+Au times by the total events sizes in Table 5.1 until more accurate estimates can be obtained. The projections for calorimetry are based on a 4 s per core estimate for Au+Au, similarly scaled by event size for p+Au and p+p. Overall, CPU needs are driven primarily by the need to keep pace with track reconstruction for Au+Au, which lead to estimates in the range of 100k-200k CPU-cores. These numbers are in direct proportion to the reconstruction times, implying that tracking optimization will be a high priority for sPHENIX over the next few years.

Projected computing and storage needs simulations are approximately an order of magni-

CPU-cores	year-1	year-2	year-3	year-4	year-5
calorimetry	1.6E+4	2.4E+4	3.2E+4	2.4E+4	3.2E+4
calibration	1.9E+4	2.0E+4	4.0E+4	2.4E+3	4.0E+4
reconstruction	5.8E+4	7.9E+4	1.2E+5	1.2E+5	1.2E+5
analysis	1.9E+3	2.4E+3	4.0E+3	4.0E+3	4.0E+3
total (no ana)	9.3E+4	1.2E+5	1.9E+5	1.3E+5	1.9E+5

Table 5.4: Computing needs

Calendar year	2020				2021				2022			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Simulation	simulation Campaign 1				simulation Campaign 2					installation pause		
TPC	digitization				fast drift simulator							
TPC	improve drift and ExB simulation				implement full slow simulation							
CAL	detailed simulation validation				fast shower lib such as caloGAN							
Offline	EB proto-test						EB full-test ???					
TPC	fast tracking / ACTS (5 sec/ev)				faster tracking (< 5sec/ev)				continue tracking optimization			
TPC	data compression				geometry/alignment							
TPC	distortion calibration				data challenge							
MVTX/INTT	integration / DB contstruction				data challenge							
CAL	fast ADC Time-series fit/clustering				cluster optimization		data challenge		calib/reco infrastructure for production			
Software framework	Memory Optimization				Multi/Hyper Threading				Data Compression			
Distributed Computing	Integrate toolset				Testing (Campaign 2)							
ML Applications	Detector fast simulator				fast simulator and tracking							
Analysis Framework	Framework design/data structures				Distributed Analysis model				Storage/Performance Optimization			
Data Quality Monitoring (DQM)	Data Harvesting and DB storage				GUI + processing loops				Reference creation + evaluation			
Code Release Validation	Validation cycles / Provenance				DQM Integration							

Figure 5.1: Gantt chart for sPHENIX software tasks through 2022.

tude lower than the those for track reconstruction. To generate 200 M Au+Au at a rate of 2000 s per event over a 10 week period would require only 3000 cores, and a 1 B sample of p+p events 100 s per event requires a farm of 16,0000 cores. These events could be readily generated at the other DOE computing facilities or during beam-off at RACF.

5.3 Effort Projections

Figure 5.1 shows approximate timelines for upcoming sPHENIX software tasks. The tasking includes improvements to simulation of the TPC and calorimeters, critical improvements to TPC tracking and calibration and a significant number of software framework tasks. Note that the online tasks are primarily managed through sPHENIX Project, and therefore are not included in this table.

The effort required to perform these and other tasks is shown in Table 5.5. Here we include some online projects that are not fully captured by the sPHENIX Project, as well as effort needed for Simulations, software tasking for the TPC and calorimeters, and the software framework. Those projects that are well suited for collaboration with the Nuclear and Particle Physics Software (NPPS) Group are identified in this table. We request a total of 5 FTE from the NPPS for work on sPHENIX software tasking, with the remainder coming from within the sPHENIX Collaboration.

Project Description	Effort	Name/Affiliation
Online Computing: Develop/Maintain tools	1.0	M. Purschke
Trigger Calculations	0.2	Univ. Colorado
0.2 FTE per subsystem	1.0	general
Simulation: Develop/Maintain tools	1.0	J. Huang
Integration	0.5	
Fast Simulator	0.5	NPPS
0.2 FTE per subsystm	1.0	general
TPC/Tracking : Reconstruction/Calibration	1.0	C. Roland
Tracking optimization	1.0	NPPS
Distortion corrections	1.0	NPPS
Vertex Finding	1.0	A. Frawley
Calorimeter : Fast ADC/Clustering	1.0	
Calorimeter Calibration	1.0	
Simulation validation	1.0	
Jet-calibraton (w/ TPC)	1.0	.
Software Framework : Develop/Maintain tools	1.0	C. Pinkenburg
Release Q&A	0.5	NPPS
Database	0.5	NPPS
Distributed workflows	0.5	NPPS
GPU-acceleration	0.5	NPPS
Total	17.2	(5 NPPS)

Table 5.5: Effort needs

Appendix A

Opportunistic Resources

A.1 DOE Facilities

In addition to resources in the private sector, the Department of Energy supports a number of high performance computing facilities that can be accessed opportunistically through a variety of mechanisms at no cost to sPHENIX.

A.1.1 NERSC

The National Energy Research Scientific Computing Center (NERSC) has hosted a Tier-2 PDSF cluster for ALICE. It currently hosts a 30 PF Cray XC40 system soon to be upgraded to the new Perlmutter supercomputer. Allocations are granted through the INCITE (check this) and through requests from LBNL PIs affiliated with DOE-SC projects. It supports a large High Performance Storage System (HPSS) for archival storage, similar to all the computing facilities mentioned in this section.

A.1.2 ORNL

The Oak Ridge Leadership Computing Facility (OLCF) is home to Summit, the worlds fastest supercomputer at 200 PFs, along with Titan (27 PF) and a set of smaller clusters designed for code-porting, and pre/post-processing for jobs on the larger supercomputers. Allocations are granted through the INCITE, ALCC and Directors Discretion processes.

A.1.3 LLNL

The Livermore Computing (LC) Facility is home to the worlds second fastest supercomputer, Sierra, at 125 PF, along with several smaller HPC systems and a wide variety of commodity clusters. Although the clusters primarily support LLNL's stockpile science

mission, allocations are also available through Grand Challenge proposals submitted by local investigators. From 2010–2015 LLNL hosted a WLCG Tier-2 cluster for the ALICE Collaboration for the local heavy-ion group, and allocations in the ranging from 10^5–10^9 core-hrs have been achieved for calculations pertaining to nuclear structure, nuclear fission, and the equation of state of the quark-gluon plasma. A proof-of-principle simulation of 1 M sPHENIX Au-Au events has been performed using the singularity container.

A.2 Grid Computing

While utilization of the BNL computing resources could in principle be the most effective for rapidly processing incoming sPHENIX data, the use of available grid resources should not be overlooked. The use of these resources can be enabled by adoption of technologies, such as PANDA and RUCIO, that are widely used by the LHC communities (especially ATLAS). It has been noted that the PHENIX effort has not focused on use of grid resources, but there is substantial experience in the STAR and ATLAS groups, and it is a major focus of the new HENP computing support group at BNL.

A.2.1 Open Science Grid

The Open Science Grid (OSG) was founded in 2004 and is funded by DOE and NSF as a consortium of universities and national laboratories across the US, which provides resources for scientific computing. Any member institution is allowed to use available resources opportunistically via a fair share algorithm. We have consulted with senior members of RACF, and we are of the understanding that the available resources do not depend on the contributions of a specific institution but just on general availability. Thus, one might expect several thousand cores to be available steady-state, providing possibilities for Monte Carlo and physics analysis.

A.2.2 Worldwide LHC Computing Grid

Conversely, the worldwide LHC computing grid, which includes other institutions worldwide, does not seem to provide as direct a path to steady-state resources. sPHENIX does not have a large base of European collaborators, who provide much of the LHC computing outside the US, and we are not aware of large contributions from Asia, compared to what is available in Europe and the US.

Appendix B

Private Sector Resources

B.1 Private sector

In 2017 the RHIC and ATLAS Computing Facility (RACF) performed a study comparing the cost-effectiveness of various paths for deploying an additional 7PB of storage and 5,000 CPU-cores [17]. The final comparison was between re-purposing an existing building to host the resources locally and the use of Amazon Cloud Services. The study found a cost differential of $3.3M for storage and $0.5–0.6M for the CPU, in favor of the local solution. However, prices used in this study were reported in July 2016, and changing prices in the private sector lead us to revisit this comparison using pricing models in the following sections. When considering costs for data storage, we also include the cost of data egress, which was not included in [17].

B.1.1 Off-site storage

The private offerings for off-site storage are numerous and cover a wide range of requirements. Generally, the storage offerings can be classified by their latency for the first byte read: For hot storage, these latencies are comparable to local disks and allow for rapid processing, cold and ultra-cold storage are optimized for scenarios where the data is read rarely, like backups. Table B.1 gives an overview over current offerings.

A crucial point of the cost analysis are traffic fees. While most offerings include ingress traffic for no extra cost, data egress to the internet is often costly. For example, the egress of 1 PB of data out of Amazon AWS costs more than $50k. An exception to this is Wasabi, where ingress and egress traffic is included. At $5900/PB/month, this is factor of 4 less than current Amazon pricing, and a factor of 4.5 compared to the RACF study [17]. However, if we repeat the previous comparison for just a 7 PB procurement over 3-years, this cloud option is $1500k compared to $1300k based on a local BNL procurement prior to 2017.

A recent development in the storage provider space are tape-library replacement offerings, like the recently introduced AWS Glacier deep archive and announced offerings from GCE.

With costs around $1k/PB/month and high egress costs, they are currently too pricey for off-site backup as an insurance against loss of the on-site tape library, but could potentially be used for long term archival of smaller data sets from analyses. We also expect a reduction in cost in the future when this market segment further develops.

Offering	Storage cost per PB·month	Ingress traffic per PB	Egress traffic per PB
AWS S3	$21k	$0	$50k
AWS Glacier	$4k	$0	$52.5k
AWS Glacier Deep Archive	$0.99k	$0	$52.5k
GCE regional	$20k	$0	$80k
GCE nearline	$10k	$0	$80k
GCE coldline	$7k	$0	$80k
GCE announced tape replacement	$1.2k	$0	N/A
Backblaze	$5k	$0	$10k
Wasabi	$5.9k	$0	$0
OVH	$2.3k	$11k	$11k

Table B.1: Overview over current private sector storage offerings. (Note: Microsoft Azure, Alibaba and IBM cloud offerings are similar to AWS or GCE, but more expensive.)

B.1.2 Off-site computing

The sheer amount of configuration options for private sector off-site computing makes it unfeasible to give a comprehensive overview here. Table B.2 lists a small selection.

The cost per CPU-core is slightly more expensive than the $0.02 per-core-hr pricing for c4 quoted in [17], presumably due to hardware and memory improvements. However, these offerings are still prohibitively expensive due to associated data egress costs, as we do not expect any of the CPU-intensive tasks to produce small output data sets.

A possible exception is the training of machine learning algorithms like deep neural networks. Here, large training sets produce a small output, the weights of the network. Larger networks need to be trained on multi-GPU systems, and the rental of such configurations from cloud providers can be very cost effective. For example, a $25/h p3.16xlarge instance at AWS offers 8 Tesla V100 cards, each priced at >$6k.

Cloud offerings typically offer reduced pricing for preemptible/spot instances, and larger reductions for long term engagements. In the latter case, dedicated root-server rentals from bigger hosting companies like Hetzner become an option, with typically smaller prices and included traffic, but more operational overhead.

Note that cloud computing options warrant continued study as costs decline, but it behooves us to first consider other DOE resources that are available at no cost to our project.

Offering	Memory per vCPU	Cost per core-hr	Comment
AWS a1	2GB	$0.0255	
AWS t2	4GB	$0.0376	
AWS c5	2GB	$0.0425	faster CPU
AWS p3	7GB	$0.3825	includes GPU
GCE n1	3.75GB	$0.0475	
GCE n1 highmem	6.35GB	$0.0592	
GCE n1 highcpu	0.9GB	$0.03545	
Hetzner cloud cx51	4GB	$0.00075	shared CPU
Hetzner cloud ccx31	4GB	$0.015	dedicated CPU

Table B.2: Overview over current private sector compute offerings. A vCPU typically corresponds to a hyperthread. Hetzner cloud offerings include 20TB traffic per node ($1.2k/PB for additional traffic), for AWS and GCE see above.

Bibliography

[1] A. Adare et al. An Upgrade Proposal from the PHENIX Collaboration. *arXiv*, 2014. arXiv:1501.06197. 1.1

[2] sPHENIX Collaboration. sPHENIX Conceptual Design Report. https://indico.bnl.gov/event/4640/attachments/18495/23200/sphenix-conceptual-design.pdf, 2018. 1.1

[3] J. Nagle and D. Perepelitsa. sPHENIX five-year running scenario and luminosity projections. https://indico.bnl.gov/event/4788/, 2018. 1.1

[4] BNL Collider-Accelerator Division. Rhic collider projections fy2017 - fy2027. www.rhichome.bnl.gov/RHIC/Runs/RhicProjections.pdf, 2017. 1.1, 2.1

[5] ALICE, ATLAS, CMS, and LHCb. Update of the computing models of the WLCG and the LHC experiments. CERN-LHCC-2014-014 / LCG-TDR-002, 2014. 1.2

[6] ALICE Collaboration. Upgrade of the online-offline computing system. CERN-LHCC-2015-006, 2015. 1.2

[7] C. Pinkenburg. Analyzing Ever Growing Datasets in PHENIX. In *Proceedings, International Conference on Computing in High Energy and Nuclear Physics (CHEP2010)*, 2011. 3.1

[8] R. Fr?hwirth. Application of kalman filtering to track and vertex fitting. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 262(2):444 – 450, 1987. URL: http://www.sciencedirect.com/science/article/pii/0168900287908874, doi:https://doi.org/10.1016/0168-9002(87)90887-4. 3.4.2

[9] Johannes Rauch and Tobias Schl?ter. GENFIT - a Generic Track-Fitting Toolkit. *J. Phys. Conf. Ser.*, 608(1):012042, 2015. arXiv:1410.3698, doi:10.1088/1742-6596/608/1/012042. 3.4.2, 4.1

[10] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012. arXiv:1111.6097, doi:10.1140/epjc/s10052-012-1896-2. 5

[11] C. A. Aidala et al. Design and Beam Test Results for the sPHENIX Electromagnetic and Hadronic Calorimeter Prototypes. *IEEE Trans. Nucl. Sci.*, 65:2901–2919, 2018. doi:10.1109/TNS.2018.2879047. 4.1

[12] Korinna C. Zapp. JEWEL 2.0.0: directions for use. *Eur. Phys. J.*, C74:2762, 2014. arXiv:1311.0048, doi:10.1140/epjc/s10052-014-2762-1. 4.2.1.1

[13] J.A. Hanks, A.M. Sickles, B.A. Cole, A. Franz, M.P. McCumber, et al. Method for separating jets and the underlying event in heavy ion collisions at the BNL Relativistic Heavy Ion Collider. *Phys. Rev.*, C86:024908, 2012. arXiv:1203.1353, doi:10.1103/PhysRevC.86.024908. 4.2.1.2, 4.2

[14] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, and M. Selvaggi. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014. arXiv:1307.6346, doi:10.1007/JHEP02(2014)057. 4.4

[15] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev.*, D97(1):014021, 2018. arXiv:1712.10321, doi:10.1103/PhysRevD.97.014021. 4.4

[16] Riccardo Di Sipio, Michele Faucci Giannelli, Sana Ketabchi Haghighat, and Serena Palazzo. DijetGAN: A Generative-Adversarial Network Approach for the Simulation of QCD Dijet Events at the LHC. 2019. arXiv:1903.02433. 4.4

[17] Costin Caramarcu, Christopher Hollowell, William Strecker-Kellogg, Antonio Wong, and Alexandr Zaytsev. The role of dedicated data computing centers in the age of cloud computing. *J. Phys.: Conf. Ser.*, 898(8):082009, October 2017. B.1, B.1.1, B.1.2