EPJ Web of Conferences **214**, 02017 (2019) https://doi.org/10.1051/epjconf/201921402017
CHEP 2018

Data Preparation for NA62

Paul Laycock[1,*] on behalf of the NA62 Collaboration[**]

[1]CERN, Geneva, Switzerland

Abstract. In 2017, NA62 recorded over a petabyte of raw data, collecting around a billion events per day of running. Data are collected in bursts of 3-5 seconds, producing output files of a few gigabytes. A typical run, a sequence of bursts with the same detector configuration and similar experimental conditions, contains 1500 bursts and constitutes the basic unit for offline data processing. A sample of 100 random bursts is used to make timing calibrations of all detectors, after which every burst in the run is reconstructed. Finally the reconstructed events are filtered by physics channel with an average reduction factor of 20, and data quality metrics are calculated.

Initially a bespoke data processing solution was implemented using a simple finite state machine with limited production system functionality. In 2017, the ATLAS Tier-0 team offered the use of their production system, together with the necessary support. Data processing workflows were rewritten with better error-handling and I/O operations were minimised, the reconstruction software was improved and conditions data handling was changed to follow best practices suggested by the HEP Software Foundation conditions database working group. This contribution describes the experience gained in using these tools and methods for data-processing on a petabyte scale experiment.

[*]e-mail: paul.james.laycock@cern.ch

[**]R. Aliberti, F. Ambrosino, R. Ammendola, B. Angelucci, A. Antonelli, G. Anzivino, R. Arcidiacono, M. Barbanera, A. Biagioni, L. Bician, C. Biino, A. Bizzeti, T. Blazek, B. Bloch-Devaux, V. Bonaiuto, M. Boretto, M. Bragadireanu, D. Britton, F. Brizioli, M.B. Brunetti, D. Bryman, F. Bucci, T. Capussela, A. Ceccucci, P. Cenci, V. Cerny, C. Cerri, B. Checcucci, A. Conovaloff, P. Cooper, E. Cortina Gil, M. Corvino, F. Costantini, A. Cotta Ramusino, D. Coward, G. D'Agostini, J. Dainton, P. Dalpiaz, H. Danielsson, N. De Simone, D. Di Filippo, L. Di Lella, N. Doble, B. Dobrich, F. Duval, V. Duk, J. Engelfried, T. Enik, N. Estrada-Tristan, V. Falaleev, R. Fantechi, V. Fascianelli, L. Federici, S. Fedotov, A. Filippi, M. Fiorini, J. Fry, J. Fu, A. Fucci, L. Fulton, E. Gamberini, L. Gatignon, G. Georgiev, S. Ghinescu, A. Gianoli, M. Giorgi, S. Giudici, F. Gonnella, E. Goudzovski, C. Graham, R. Guida, E. Gushchin, F. Hahn, H. Heath, T. Husek, O. Hutanu, D. Hutchcroft, L. Iacobuzio, E. Iacopini, E. Imbergamo, B. Jenninger, K. Kampf, V. Kekelidze, S. Kholodenko, G. Khoriauli, A. Khotyantsev, A. Kleimenova, A. Korotkova, M. Koval, V. Kozhuharov, Z. Kucerova, Y. Kudenko, J. Kunze, V. Kurochka, V.Kurshetsov, G. Lanfranchi, G. Lamanna, G. Latino, P. Laycock, C. Lazzeroni, M. Lenti, G. Lehmann Miotto, E. Leonardi, P. Lichard, L. Litov, R. Lollini, D. Lomidze, A. Lonardo, P. Lubrano, M. Lupi, N. Lurkin, D. Madigozhin, I. Mannelli, G. Mannocchi, A. Mapelli, F. Marchetto, R. Marchevski, S. Martellotti, P. Massarotti, K. Massri, E. Maurice, M. Medvedeva, A. Mefodev, E. Menichetti, E. Migliore, E. Minucci, M. Mirra, M. Misheva, N. Molokanova, M. Moulson, S. Movchan, M. Napolitano, I. Neri, F. Newson, A. Norton, M. Noy, T. Numao, V. Obraztsov, A. Ostankov, S. Padolski, R. Page, V. Palladino, C. Parkinson, E. Pedreschi, M. Pepe, M. Perrin-Terrin, L. Peruzzo, P. Petrov, F. Petrucci, R. Piandani, M. Piccini, J. Pinzino, I. Polenkevich, L. Pontisso, Yu. Potrebenikov, D. Protopopescu, M. Raggi, A. Romano, P. Rubin, G. Ruggiero, V. Ryjov, A. Salamon, C. Santoni, G. Saracino, F. Sargeni, V. Semenov, A. Sergi, A. Shaikhiev, S. Shkarovskiy, D. Soldi, V. Sougonyaev, M. Sozzi, T. Spadaro, F. Spinella, A. Sturgess, J. Swallow, S. Trilov, P. Valente, B. Velghe, S. Venditti, P. Vicini, R. Volpe, M. Vormstein, H. Wahl, R. Wanke, B. Wrona, O. Yushchenko, M. Zamkovsky, A. Zinchenko.

1 Introduction

The NA62 experiment [1] is designed to measure the extremely rare kaon decay $K^+ \to \pi^+ \nu\bar{\nu}$ with a precision of 10%. It utilises the high intensity 400 GeV/c beam of protons from the CERN Super Proton Synchrotron which impinges on a beryllium target producing a 75 GeV/c secondary charged hadron beam containing 6% kaons. The NA62 experimental apparatus is shown in figure 1 (figure and caption taken from [1]).

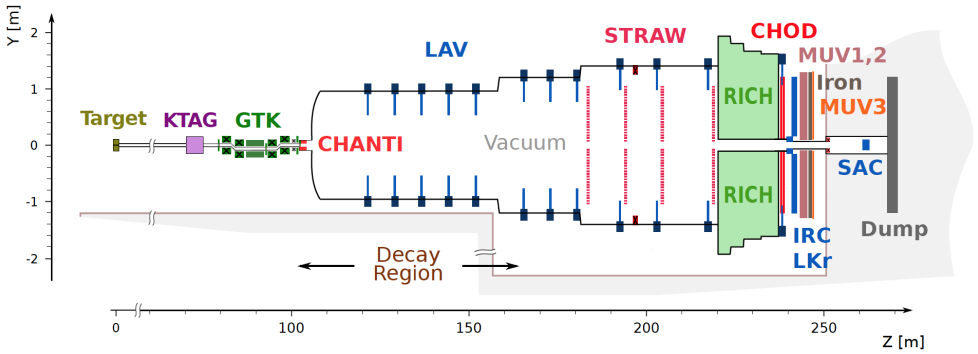


Figure 1. Schematic vertical section through the NA62 experimental setup. The main elements for the detection of the K^+ decay products are located along a 150 m long region starting 121 m downstream of the kaon production target. Useful K^+ decays are detected in a 65 m long decay region. Most detectors have an approximately cylindrical shape around the beam axis. An evacuated passage surrounding the beam trajectory allows the intense (750 MHz) flux of un-decayed beam particles to pass through without interacting with detector material before reaching the dump.

The rate of kaon decays in the decay region is 5 MHz, providing a challenge for data acquisition and triggering documented elsewhere in these proceedings [2]. To put this in the context of other experiments, a modified version of Figure 1 from [3] is shown in figure 2. Despite the small event size, the very large event rate places NA62 at the same challenging frontier as the LHC experiments in terms of data volume. Moreover, the event rate is very difficult to tame with triggers, with only modest reductions possible to the level of 100 kHz, meaning that offline data preparation is challenging with around one billion events produced per day, with a volume of 10 TB. The time matching between the incoming kaon, identified by the KTAG detector matched to a track in the GTK pixel detector, and the outgoing charged particle identified in the RICH detector, must be kept at the level of 100-150 ps with a matching efficiency greater than 99% if the ambitious goals of the experiment are to be realised, requiring a detailed understanding of the detectors and careful yet automated calibrations.

1.1 Data preparation workflow

A schematic of the main components of the data preparation workflow employed by NA62[1] is shown in figure 3. It begins with an iterative (4-step) calibration procedure which performs a large part of the timing calibration, using a randomly selected subset of the data. The data are collected in units of bursts of protons from the CERN Super Proton Synchrotron of around 3 s duration, one burst is stored in one raw data file of several GB in size; sequential bursts with similar data-taking conditions are organised into data acquisition runs of typically

[1]This workflow was accurate at the time of presentation at CHEP 2018.

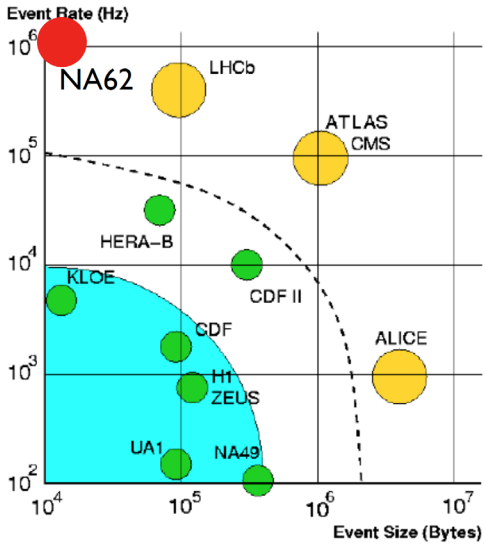


Figure 2. First level trigger event rates and sizes for various experiments, including the NA62 experiment which is shown as the red circle.

1500 bursts. The iterative calibration sample size is nominally 100 bursts[2] which provides sufficient statistics for the calibration of the majority of the detectors. However, the most precise detectors (KTAG, GTK and RICH) can be calibrated with even higher precision and thus the bulk processing step which follows the iterative calibration first re-determines the timing calibration for these detectors before performing the full event reconstruction of all detectors with the optimal conditions. The fully reconstructed data volume is larger than the raw data volume and is thus far too large for physics analysis when considering several years of data. Therefore NA62 performs offline event selections on the fully reconstructed data to produce several (around 10) filtered datasets which are the input to physics analysis. Data quality (DQ) must also be performed on the fully reconstructed dataset, together with any other tasks that require the fully reconstructed data information for every event. Once these key tasks are finished, the fully reconstructed data is deleted from disk.

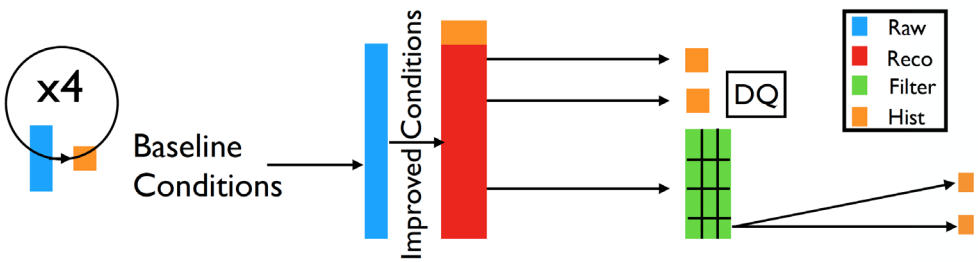


Figure 3. A schematic of the main components of the data preparation workflow for NA62.

[2]Simple quality criteria are applied based on the size of the raw event data file for a given burst. If fewer than 100 bursts remain in the calibration sample then manual intervention is required, but calibration will normally be successful with fewer than 100 bursts.

2 Production system

The initial attempts to tackle the data preparation workflow achieved only moderate success, resulting in only a fraction of the data being processed for physics analysis, despite the data preparation procedures being technically correct and having a simple finite state machine to provide retry functionality. The workflows initially used to analyse one burst did not scale well and in some cases were entirely unsuitable for the resources used for large scale computing at CERN. Furthermore, NA62 had started taking data without a conditions database and instead relied (and continues to rely at the time of writing) on plain ascii text files, several per detector, for the storage of the majority of its conditions data. The simple production system that had succeeded for commissioning purposes needed to be replaced with something that could address those issues.

2.1 NA62 Tier0

In 2017, NA62 embarked on the work needed to interface their reconstruction and analysis code with a state of the art production system running on CERN computing infrastructure, drawing on the experience of other CERN experiments, particularly ATLAS. The ATLAS Tier-0 production system [4] was designed for exactly the kind of data preparation problems that NA62 faced, and benefited from several years of experience and optimisation operating with an order of magnitude more data. It was also designed to be experiment and payload agnostic, which is the hallmark of a reusable production system, and with very few exceptions[3] the same production system code was used by both ATLAS and NA62.

Adapting the NA62 reconstruction code and workflows required effort, in particular the separation of conditions data from software and adaptation to follow best practices for conditions data handling from the HSF Community White Paper [5]. Further work on the software was required to allow a full configuration of the software to be defined outside of a built release. All of the executables required for the full data preparation workflow were reviewed and repackaged to respect these constraints. Quite remarkably, the NA62 software experts then decided to go further than meeting the baseline requirements, aiming to fully optimise their software in terms of robustness and resource usage. Memory usage was minimised to respect the 2GB limit of the HTCondor [6] compute nodes at CERN[4] and the software was given robust error handling to allow $\approx$100% of bursts to be processed[5].

A new Python transform package, NA62Transform, provided the interface between the production system and the NA62 reconstruction and analysis software. It implemented all of the individual data preparation workflow steps, optimised the usage of resources, in particular I/O, and took care of error handling, reporting error codes back to the production system. Several lightweight and dedicated daemons were written in Python that implemented various orchestration procedures and hand-offs, including mechanisms to help manage and prioritise workload, and to orchestrate the staging of data from tape. The NA62 Tier-0 production system configuration completed the setup, defining each of the steps in the data preparation workflow. As the ATLAS Tier-0 production system accepts parameters in its configuration language, changes in e.g. NA62 reconstruction release were handled by changing one parameter in the global configuration script. The NA62 Tier-0 production system is shown in figure 4, together with the main components of CERN CPU and storage infrastructure that it uses, namely HTCondor, EOS [7] (disk storage) and CASTOR [8] (tape storage).

[3] These exceptions could also have been dealt with, but were a handful of lines of code.

[4] NA62 had a fair share quota on HTCondor and jobs requiring more than 2GB memory resulted in allocating more nodes per job.

[5] If a burst was corrupt, the software would detect this and finish cleanly with an appropriate error code, as opposed to crashing.

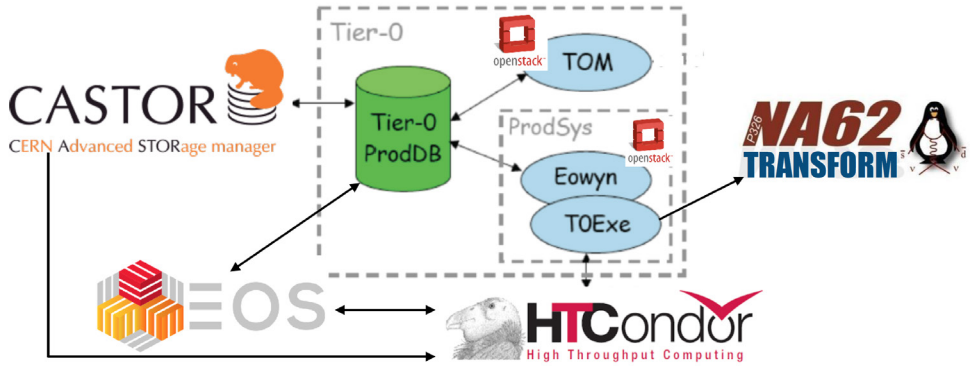


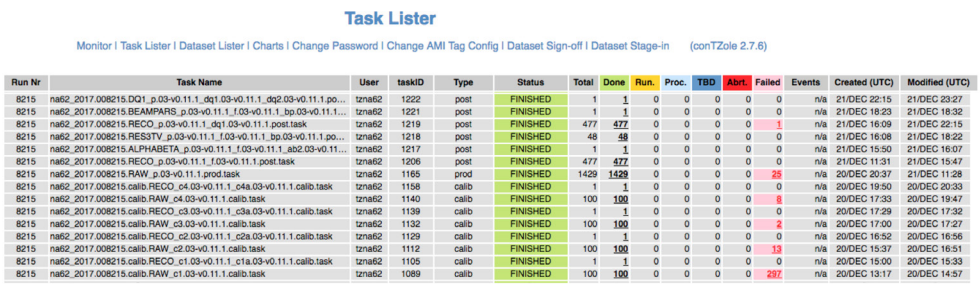
Figure 4. The NA62 Tier-0 production system, together with the CERN CPU and storage components with which it interacts. The Tier-0 comprises an Oracle database backend (ProdDB), the Tier-0 Manager (TOM) which creates new jobs in ProdDB based on the data preparation workflow configuration, and the ProdSys subsystem. The ProdSys subsystem looks for jobs in ProdDB, submits executables (T0Exe) to a batch system (in this case HTCondor) and monitors the status of the batch system. Finally, job management on the HTCondor worker nodes is performed by T0Exe which performs file input and output management, runs the NA62 reconstruction or analysis code, captures errors and produces job reports. For more details see [4].

2.2 Puppet-managed infrastructure

Critical tools for monitoring and operation of the production system are graphical user interfaces and dedicated machines to run the required services. CERN provisions virtual machines using OpenStack [9] and uses the Puppet configuration management system [10] to configure services[6]. Puppet configurations for the ATLAS Tier-0 production system services were refactored to be experiment agnostic, allowing the same modules to be used with a handful of parameters needing to be defined for each of ATLAS and NA62. A screenshot of the NA62 Tier-0 production system web interface is shown in figure 5, each step in the workflow is represented by one row. The number of job failures is evident, as is the ultimate success of the retry functionality. Many of the cells in the display are interactive and allow the user / operator to drill down to see error codes and log files and then take the appropriate action. This interactivity is another crucial component of the success of the NA62 Tier-0 production system, and the ATLAS Tier-0 production system on which it is based.

The same approach allowed ATLAS and NA62 to share Puppet configurations for the web interface to the database backend of the production system. After refactoring the common Django web application module, the remaining configuration was also reduced to a few lines of code. Although the majority of the work could be performed using the Python API of the production system, inspecting the database contents was crucial both for debugging and gaining understanding of the system. Data-mining of information from the database also proved to be extremely useful, e.g. analysing job run times by job type allowed the workflow steps and retry strategies to be optimised and thus optimise throughput. Finally, dedicated virtual machines were configured to run these web services and the production system itself.

[6]Puppet is a popular open source configuration management system used to configure machines, it is very useful for creating many machines with identical configurations, or in cases where machines with exact configurations are required and must be easily recovered.



Task Lister

Monitor | Task Lister | Dataset Lister | Charts | Change Password | Change AMI Tag Config | Dataset Sign-off | Dataset Stage-in (conTZole 2.7.6)

Run Nr	Task Name	User	taskID	Type	Status	Total	Done	Run.	Proc.	TBD	Abrt.	Failed	Events	Created (UTC)	Modified (UTC)
8215	na62_2017.008215.DQ1_p.03-v0.11.1_dq1.03-v0.11.1_dq2.03-v0.11.1.po...	tzna62	1222	post	FINISHED	1	1	0	0	0	0	0	n/a	21/DEC 22:15	21/DEC 23:27
8215	na62_2017.008215.BEAMPARS_p.03-v0.11.1_f.03-v0.11.1_bp.03-v0.11.1...	tzna62	1221	post	FINISHED	1	1	0	0	0	0	0	n/a	21/DEC 18:23	21/DEC 18:32
8215	na62_2017.008215.RECO_p.03-v0.11.1_dq1.03-v0.11.1.post.task	tzna62	1219	post	FINISHED	477	477	0	0	0	0	1	n/a	21/DEC 16:09	21/DEC 22:15
8215	na62_2017.008215.RES3TV_p.03-v0.11.1_f.03-v0.11.1_bp.03-v0.11.1.po...	tzna62	1218	post	FINISHED	48	48	0	0	0	0	0	n/a	21/DEC 16:08	21/DEC 18:22
8215	na62_2017.008215.ALPHABETA_p.03-v0.11.1_f.03-v0.11.1_ab2.03-v0.11...	tzna62	1217	post	FINISHED	1	1	0	0	0	0	0	n/a	21/DEC 15:50	21/DEC 16:07
8215	na62_2017.008215.RECO_p.03-v0.11.1_f.03-v0.11.1.post.task	tzna62	1206	post	FINISHED	477	477	0	0	0	0	0	n/a	21/DEC 11:31	21/DEC 15:47
8215	na62_2017.008215.RAW_p.03-v0.11.1.prod.task	tzna62	1165	prod	FINISHED	1429	1429	0	0	0	0	25	n/a	20/DEC 20:37	21/DEC 11:28
8215	na62_2017.008215.calib.RECO_c4.03-v0.11.1_c4a.03-v0.11.1.calib.task	tzna62	1158	calib	FINISHED	1	1	0	0	0	0	0	n/a	20/DEC 19:50	20/DEC 20:33
8215	na62_2017.008215.calib.RAW_c4.03-v0.11.1.calib.task	tzna62	1140	calib	FINISHED	100	100	0	0	0	0	8	n/a	20/DEC 17:33	20/DEC 19:47
8215	na62_2017.008215.calib.RECO_c3.03-v0.11.1_c3a.03-v0.11.1.calib.task	tzna62	1139	calib	FINISHED	1	1	0	0	0	0	0	n/a	20/DEC 17:29	20/DEC 17:32
8215	na62_2017.008215.calib.RAW_c3.03-v0.11.1.calib.task	tzna62	1132	calib	FINISHED	100	100	0	0	0	0	2	n/a	20/DEC 17:00	20/DEC 17:27
8215	na62_2017.008215.calib.RECO_c2.03-v0.11.1_c2a.03-v0.11.1.calib.task	tzna62	1129	calib	FINISHED	1	1	0	0	0	0	0	n/a	20/DEC 16:52	20/DEC 16:56
8215	na62_2017.008215.calib.RAW_c2.03-v0.11.1.calib.task	tzna62	1112	calib	FINISHED	100	100	0	0	0	0	13	n/a	20/DEC 15:37	20/DEC 16:51
8215	na62_2017.008215.calib.RECO_c1.03-v0.11.1_c1a.03-v0.11.1.calib.task	tzna62	1105	calib	FINISHED	1	1	0	0	0	0	0	n/a	20/DEC 15:00	20/DEC 15:33
8215	na62_2017.008215.calib.RAW_c1.03-v0.11.1.calib.task	tzna62	1089	calib	FINISHED	100	100	0	0	0	0	297	n/a	20/DEC 13:17	20/DEC 14:57

Figure 5. A screenshot of the web interface to the NA62 Tier-0 production system.

3 Conclusions

The NA62 experiment, despite its small event size, has challenging data processing requirements for both online [2] and offline. The data preparation workflows required to guarantee timing cross-calibration at the level of 100 ps are complex, with a data volume that demands a state of the art production system. NA62 has benefited from the work of previous experiments, particularly ATLAS, to produce the NA62 Tier-0 production system that meets those demands. The production system uses a new NA62Transform Python package that optimises the data preparation workflows, especially in terms of I/O. Together with many improvements in the NA62 reconstruction and analysis software including proper error handling, reduced memory consumption and separation of both conditions data and configuration from the software, offline processing of NA62 data can now be performed efficiently and accurately.

References

[1] E. Cortina Gil *et al.* [NA62 Collaboration], JINST **12** (2017) no.05, P05025 doi:10.1088/1748-0221/12/05/P05025, arXiv:1703.08501 [physics.ins-det].

[2] M. Boretto *et al.* [NA62 Collaboration], proceedings of CHEP 2018.

[3] P. J. W. Faulkner *et al.* [GridPP Collaboration], J. Phys. G **32** (2006) N1. doi:10.1088/0954-3899/32/1/N01

[4] M. Elsing, L. Goossens, A. Nairz and G. Negri, J. Phys. Conf. Ser. **219** (2010) 072011. doi:10.1088/1742-6596/219/7/072011

[5] A. A. Alves, Jr *et al.*, arXiv:1712.06982 [physics.comp-ph].

[6] D. Thain, T. Tannenbaum, and M. Livny, Concurrency and Computation: Practice and Experience, Vol. 17, No. 2-4, pages 323-356, February-April, 2005

[7] A. J. Peters and L. Janyst, J. Phys. Conf. Ser. **331** (2011) 052015. doi:10.1088/1742-6596/331/5/052015

[8] J. P. Baud, B. Couturier, C. Curran, J. D. Durand, E. Knezo, S. Occhetti and O. Barring, eConf C **0303241** (2003) TUDT007, cs/0305047 [cs-oh].

[9] Website: https://www.openstack.org/

[10] Website: https://puppet.com/