The Science Cloud – Towards a Research Data Ecosystem for the next Generation of Data-intensive Experiments and Observatories

711. WE-Heraeus-Seminar

12 Jan - 15 Jan 2020
at the Physikzentrum Bad Honnef/Germany

WILHELM UND ELSE
HERAEUS-STIFTUNG

Introduction

The Wilhelm und Else Heraeus-Stiftung is a private foundation that supports research and education in science with an emphasis on physics. It is recognized as Germany’s most important private institution funding physics. Some of the activities of the foundation are carried out in close cooperation with the German Physical Society (Deutsche Physikalische Gesellschaft). For detailed information see https://www.we-heraeus-stiftung.de

Aims and scope of the 711. WE-Heraeus-Seminar:

The next-generation of physics and astronomy flagship projects require high-throughput exascale computing solutions. Examples are the

High-luminosity upgrade of the Large Hadron Collider at CERN,
Square Kilometre Array radio telescope,
Large Synoptic Survey Telescope, or
Einstein Telescope.

The projects have in common that they heavily rely on robust and scalable solutions for their data challenge. These solutions will make best use of progress in the performance and sustainability of the IT-hardware, but most of all of intelligent data pipelines. In the future, raw data will be heavily processed reducing the volume under a controlled loss of information to facilitate the long-time archival of the data and to make the data accessible to the widely distributed user communities. Among the data processing tools, machine-learning algorithms could play a potentially important role for reducing data but retaining the encoded information. New technologies and computer architectures as well as a standardized approach to the programming of data pipelines are key to a highly performant research data infrastructure serving the needs of the diversity of experiments and observatories.

The seminar encourages discussions among experts from science and industry with young scientists how to meet these challenges, addressing the long-term vision of community efforts for the advancement of fundamental science. The seminar also has a message to the scientific community: More efforts are needed to support the career of young researchers who gear up to tackle the gigantic data challenges.

Scientific Organizers:

Dr. Andreas Haungs	KIT Karlsruhe (KCETA), Germany E-mail: andreas.haungs@kit.edu
Prof. Dr. Karl Mannheim	Julius-Maximilians-Universität, Würzburg, Germany E-mail: mannheim@astro.uni-wuerzburg.de
Prof. Dr. Matthias Steinmetz	Leibniz-Institut für Astrophysik, Potsdam, Germany E-mail: msteinmetz@aip.de

Program

Sunday, 12 January 2020

16:00 – 20:00	Registration	
from 18:00	*BUFFET SUPPER / Informal get together*	
20:00 – 21:00	Dieter Kranzlmüller	**Are we ready for the next level of Big Data?**

Monday, 13 January 2020

08:00 – 08:50	*BREAKFAST*	
08:50 – 09:00	Andreas Haungs Karl Mannheim Matthias Steinmetz	**Welcome and Organization**
09:00 – 09:45	Alex Szalay	**Lessons from the Sloan Digital Sky Survey**
09:45 – 10:30	Martin Kümmel	**Data pipelines for Euclid**
10:30 – 11:00	*COFFEE BREAK*	
11:00 – 11:45	Frossie Economou	**How Big Data missions like LSST drive new models of how we build our systems - and our teams**
11:45 – 12:30	Michiel van Haarlem	**Science Data Centres for Radio Astronomy: from LOFAR to SKA**
12:30 – 12:40	**Conference Photo** (in front of the Lecture Hall)	
12:40	*LUNCH*	

Program

Monday, 13 January 2020

14:00 – 14:45	Stefan Schlenstedt	**The Cherenkov Telescope Array Data Management Model**
14:45 – 15:30	Kay Graf	**Handling of Neutrino Telescope Data**
15:30 – 16:00	*COFFEE BREAK*	
16:00 – 16:45	Volker Gülzow	**Computing Challenges for the HL-LHC**
16:45 – 17:30	Steffen Hauf	**Data Challenges at the European XFEL – 1B/s to 10GB/s**
17:30 – 18:30	**General discussion**	
18:30 – 18:45	Stefan Jorda	**About the Wilhelm and Else Heraeus Foundation**
19:00	*DINNER*	

Program

Tuesday, 14 January 2020

08:00	*BREAKFAST*	
09:00 – 09:45	Stefan Hachinger Luigi Iapichino	**Withnessing the Convergence of HPC and Data Analytics from a Supercomputing Centre Perspective**
09:45 – 10:30	Susanne Pfalzner	**Knowledge Gain in the Age of HPC and Big Data**
10:30 – 11:00	*COFFEE BREAK*	
11:00 – 11:45	Achim Streit	**Enabling Data-Intensive Computing & the EOSC**
11:45 – 12:30	Hermann Kohlstedt	**Bio-inspired Information Processing: The Future of Artificial Intelligence?**
12:30	*LUNCH*	
14:00 – 14:45	Ingolf Wittmann	**IT infrastucture of the futute – Envision the future of Computing !**
14:45 – 15:30	**Poster-Flash I**	
15:30 – 16:00	*COFFEE BREAK*	
16:00 – 16:45	**Poster Flash II**	
16:45 – 18:45	**Poster Session**	
19:00	*HERAEUS DINNER at the Physikzentrum* *(cold & warm buffet, with complimentary drinks)*	

Wednesday, 15 January 2020

08:00	*BREAKFAST*	
09:00 – 09:45	Kai Polsterer	**Accessing complex structures with unsupervised and deep-learning techniques**
09:45 – 10:30	Stefanie Walch-Gassner	**Simulating has dynamics in galaxies: a 3D view of star formation and feedback**
10:30 – 11:00	*COFFEE BREAK*	
11:00 – 11:45	Johannes Schemmel	**Brain Inspired Computing**
11:45 – 12:30	**Poster Prize**	
12:30	*LUNCH*	
14:00 – 14:45	Martin Brennecke	**The Machine & Memory-Driven Computing**
14:45 – 15:30	Hermann Heßling	**Memory-based computing for astronomical applications**
15:30 – 16:00	*COFFEE BREAK*	
16:00 – 16:45	Katharina Morik	**Physics and machine learning**
16:45 – 17:30	Seminar conveners	**Summary**

End of Seminar / Departure

Posters

Posters

01	Markus Blank-Burian	**WWU Cloud - Open Source based Cloud Services at the University of Münster**
02	Harry Enke	**FAIR in astronomy context**
03	Anastasia Galkin	**Daiquiri - Python based framework for the publication of scientific databases**
04	Zohreh Ghaffari + Catalina Sobrino Figaredo	**A new multi-band optical image pipeline for the Magellan 6.5 m telescope**
05	Niraj Kandpal	**Molecular Outflow Detection in G327: A 3D Approach**
06	Lars Künkel	**Searching Pulsars Using Neural Networks**
07	Jörn Künsemöller	**Metadata and User-Provided Data in the LOFAR Long Term Archive**
08	Man I Lam	**PyParadise: A simultaneous pipeline of stellar and gas kinematics**
09	Abhishek Malik	**Exoplanet detection using Machine Learning**
10	Jan Mayer	**Entering NeuLAND: Analysis workflow preservation for a fair FAIR**
11	Martin Müller	**Data Infrastructure at the University of Cologne’s Institute for Nuclear Physics**
12	Annika Oetjens	**Past planetary engulfment as a possible explanation for observed high stellar rotation on metal poor main sequence stars**
13	Oleksandra Razim	**Towards reliable photometric redshifts with machine learning methods**

Posters

14	Benedikt Riedel	**Cloud burst for Multi-Messenger Astrophysics**
15	Bernd Schleicher	**Machine Learning in Cherenkov Astronomy**
16	Kilian Schwarz	**The PAHN-PaN consortium at the NFDI**
17	Kseniia Sysoliatina	**Astronomy meets big data: Improving the Milky Way model with the billion-star surveyor Gaia**
18	Alexander Trettin	**From 2D to 3D to Graph Networks: Representing Detector Geometries in Neural Networks**
19	Meetu Verma	**Classification of high-resolution solar Hα spectra using t-SNE**
20	Christoph Welling	**NuRadioMC and NuRadioReco – A Software Framework for the Radio Detector Community**
21	Tobias Winchen	**Prototypes for the Next Generation of Computing Backends in Radio-Astronomy**
22	Martin Wolf	**BoostNumpy: Big Data Processing in C++ with Python convenience**

Abstracts of Lectures

(in alphabetical order)

The Machine & Memory-Driven Computing

M. Brennecke

Hewlett Packard Enterprise, Hamburg, Germany

The future of computing: Technical background and reasoning to harness a Memory-Driven Computing architecture & The Machine research project.

Data explosion, exponentially parallel mass data processing, NP hard problems and the end of Moore's Law require new concepts for the future of standardized and open computing infrastructures. Leaving the von-Neumann-model behind, a different architecture is required to build future computing devices of any size and scale. Using a memory-centric system and software model instead, with a non-proprietary photonics-based bus, enables the effective usage of specialized units like ASICs, FPGAs or any future upcoming processing technology. All of this within a standardized non-proprietary architecture and – in the future – realized with commonly acquirable infrastructure components.

A few years back Hewlett Packard Labs, the research facility of Hewlett Packard Enterprise (HPE), started working on the theory of a massive-scale computer system, using light instead of electricity to couple in-system components and utilizing a fundamental different approach for data and memory handling. Since then, the research project has evolved into the first ready-to-use prototype components, the open Gen-Z standard and promising discussions on future usage and problems to be solved.

Get a view into industry research on memory-driven computing, the research project The Machine, the work of the German Center of Neurodegenerative Diseases (DZNE, Alzheimer's research) on HPE's first prototype and the current state of technology development from Hewlett-Packard Labs.

How Big Data missions like LSST drive new models of how we build our systems - and our teams

Frossie Economou[1] **and William O'Mullane**[1]

[1]*Large Synoptic Survey Telescope, AURA, Tucson, USA*

The Large Synoptic Survey Telescope [1] will undertake an optical survey that is expected to result in approximately 400 Petabytes of data holdings after its 10 year mission. As we build software services to curate, process and enable scientific discovery for thousands of astronomers, we find ourselves in a regime where traditional astronomical services do not scale. Using as an example the architecture of the LSST Science Platform, we discuss the implications for how the engineering of data services is evolving towards technologies and techniques common outside of academia. Mindful of the meeting's focus on young researchers, we will also discuss our experience with building high-performing teams suitable for this kind of work, and what it means for young people preparing to enter the field, and the institutions who want to employ them.

References

[1] I. Zeljko, Astrophysical Journal, **873**, page (year)

Handling of Neutrino Telescope Data

K. Graf[1]

[1]ECAP, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Neutrino telescopes are constructed and operated by international collaborations in remote locations around the Earth, deep under the surface: ANTARES[1] and KM3NeT[2] in the Mediterranean Sea, Baikal-GVD[3] in Lake Baikal, and IceCube[4] in the Antarctic ice shelf. The detector configurations differ, based on the science goals pursued and the natural environment they are constructed in. As large three-dimensional arrays of photosensors, they share, however, the same fundamental challenges for research data handling.

This talk will provide an overview on the different projects and their scientific goals, with the focus on the data challenges. It will sketch the currently employed tier-like data infrastructure model and the data management plans of the experiments. The data pipelines and the data formats leading from low-level, large data sets to high-level event tables, and the distribution of data and resources for processing and analysis will discussed.

In addition to the currently employed robust models, the scalability of the data infrastructure solutions needs to be taken into account by the experiments under development or construction. Also, the implementation and integration of new technologies, e.g. AI accelerators, is pursued. The software development, testing and deployment infrastructure goes hand-in-hand with the data infrastructure and should also allow for flexible and scalable deployment of the data processing and analysis software, e.g. through a fully integrated development system with continuous integration and deployment. Cloud-based scenarios are now under discussion both for the data processing and the analysis pipelines, especially also in the context of open (access) data.

Thus, the current implementation of data handling in neutrino telescopes will be discussed as well as the opportunities and challenges for using cloud-based data ecosystems.

References

[1] J.A. Aguilar et al., Nuclear Inst. and Meth. in Phys. Res. **A 656**, p. 11 (2011)

[2] S. Adrián-Martínez et al., J. Phys. G, **43 (8)**, 084001 (2016)

[3] A.D. Avrorin et al., EPJ Web Conf. **207**, 01003 (2019)

[4] F. Halzen and T. Gaisser, Ann. Rev. Nuc. Part. Sc. **64**, p. 101 (2014)

Computing Challenges for the HL-LHC

Volker Guelzow

DESY, Hamburg, Germany

In Run 2 Atlas and CMS collected about 160 fb**-1 at centre of mass 13 TeV. With the HL-LHC the computing challenge is driven by the increased event complexity (~200 proton interactions/bunch rather than 60 during Run-2)) and the expected increase in event rate of a factor 5 to 10 with respect to Run-2, both data and Monte-Carlo. Based upon today's computing models for ATLAS and CMS this would require an increase of approximately 20 times the resources available today.

Some fraction of these requirements can be compensated through the hardware developments over the years, A today as realistic considered assumption is about 10% for CPU power and 15% for storage capacity. New technologies like intensive GPU usage and maybe "in memory" computing or Quantum Computing may help but will not solve the problem entirely. Today, Software Frameworks like Gaudi, CMSSF etc are in use and have been identified for improvement.

In 2017, the HEP community has produced the Community White Paper (CWP), under the aegis of the HEP Software Foundation (HSF) [1] and the results were picked up in the briefing book for the European Strategy [3]. It covers the entire spectrum of activities that are part of HEP computing and the idea of a European Data Science Institute for Particle Physics [2]. Key elements are:

●Software: today's code does not use modern architectures good enough

●Algorithmic improvements: For HL-LHC Monte-Carlo simulations and reconstruction algorithms must be improved significantly

●Event generators:As the precision of the experiments increases the generators need to simulate higher-order effects.

●Reducing data volumes: With a direct effect on costs for example with "nano AOD"

●Managing operations costs: The conepct of a "data-lake" where few large centres manage the long-term data

●Optimising hardware costs: There is an opportunity to reduce storage cost also by more actively using tape

References

1. //hepsoftwarefoundation.org
   Community white paper: https://arxiv.org/abs/1812.07861
2. M. Pierini,
   https://indico.cern.ch/event/765096/contributions/3295512/attachments/1785106/2906008/A_European_Data_Science_Institute_for_Particle_Physics.pdf
3. Briefing book final :
   http://cds.cern.ch/record/2691414/files/Briefing_Book_Final.pdf

Withnessing the Convergence of HPC and Data Analytics from a Supercomputing Centre Perspective

S. Hachinger[1] and **L. Iapichino**[1]

[1]Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities (LRZ), Boltzmannstr. 1, Garching b. München, Germany

In this decade, computational scientists and data specialists have withnessed large leaps forward of the technologies in their scientific environment. The Leibniz Supercomputing Centre (LRZ, Garching) currently hosts the most powerful HPC system in the EU, SuperMUC-NG. Nowadays, operating such leadership-class computing infrastructure needs a clear strategy and dedicated services to manage massive amounts of data. On our systems, we support large simulation projects not only from the computing side, but also with the analysis and visualisation of extremely large data sets. As an example, the novel "Magneticum Web Portal" allows users to openly access the data products of the cosmological simulation suite "Magneticum", and to trigger customized analysis tasks on dedicated HPC resources. Such integrated approaches to Scientific Computing and Data include a strategy to make supercomputing data at Leibniz Supercomputing Centre FAIR (Findable, Accessible, Interoperable, Reusable). These efforts meet the growing request by researchers, institutions and funding agencies for making research transparent and visible both to the scientific community and the public.

Data Challenges at the European XFEL – 1B/s to 10GB/s

S. Hauf[1]

[1]*European XFEL GmbH, Schenefeld, Germany*

The European XFEL is a X-ray free electron laser near Hamburg, Germany. It has started operation in September 2017. As of summer 2019 all three beam lines are operational, providing six experimental end stations to facility users. The facility is unique both in terms of its peak brilliance of 10^{33} photons/s/mm^2/mrad^2 at 0.1 % bandwidth, as well as its time structure, with photon pulses delivered at up to 4.5 MHz repetition rate, and organized into so-called trains of 2700 pulses repeating at a frequency of 10 Hz [1, 2].

The AGIPD, LPD and DSSC Mpixel 2D detectors are designed to resolve this time structure, providing image acquisition rates in the MHz regime. They output data at ~10 GB/s [3]. Fast digitizers based on custom FPGA developments furthermore sample data at MHz pulse resolution, and are used e.g. in beam diagnostics [4]. Additional 2D detectors, such as Jungfrau, FastCCD, ePIX100a and pnCCD operate at the 10Hz train frequency. Typical data rates for such equipment are 10-1000 MB/s. For all instrumentation and beamline equipment auxiliary data is produced on Beckhoff PLCs or through manufacturer specific interfaces. This “slow” data is frequently event driven with update rates in the low Hz regime resulting in data rates of a few B/s to MB/s.

To ingest these varying data types and rates, and to provide operator control interfaces for the facility, the Karabo SCADA system [5] has been developed in-house and is in use at the photon beamline systems and the instrument stations.

The current status of these data-related services at the European XFEL, and their availability to facility users [6] will be presented. Current and upcoming “data challenges”, and lessons learned from two years of user operation will be shown.

References

[1] M. Altarelli, *European X-ray Free Electron Laser*, European XFEL GmbH, Technical Design Report (2006).

[2] T. Tschentscher, *Photon Beam Properties at the European XFEL*, European XFEL GmbH Technical Report TR-2011-006 (2012).

[3] Kuster, Markus, et al. "Detectors and calibration concept for the European XFEL." *Synchrotron Radiation News* 27.4 (2014): 35-38.

[4] Grünert, Jan, et al. "X-ray photon diagnostics at the European XFEL." *J. Synchrotron Rad.* 26.5 (2019).

[5] Hauf, Steffen, et al. "The Karabo distributed control system." *J. Synchrotron Rad.* 26.5 (2019).

[6] Fangohr, Hans, et al. "Data analysis support in Karabo at European XFEL." (2018): TUCPA01

Memory-based computing for astronomical applications

E. Buchholz[1] and H. Heßling[1]

[1] *University of Applied Sciences (HTW) Berlin, Berlin, Germany*

The upcoming Square Kilometre Array (SKA) is a next-generation radio telescope distributed over three continents and will consist of thousands of antennas. Due to its exceptional resolution power, high-resolution images of the Universe will be generated where the size of single images may be of the order of one Petabyte [1].

Current computing architectures are not designed for analyzing huge data objects of such size (“data monsters”). Analyzing data stored on disks takes up rather a long time (memory-wall problem). Memory-based computing provides a change in paradigm: it replaces the current processor-centric by a memory-centric architecture.

Hewlett Packard Enterprise (HPE) developed a memory-driven computing prototype (with 160 Terabytes main memory) that is used since 2016 at the DZNE Bonn for genomic research [2]. Our University is cooperating most recently with HPE and has access to a new prototype in the HPE labs, the so-called *sandbox*. The talk presents first results of studies with the sandbox.

The resolution power of sensors is increasing strongly in many areas. In medicine, for example, digital slide scanners generate high-resolution images of tissue sections up to the Terabyte-range. Traditionally, these images are analyzed by complex workflows on single workstations. Porting image processing workflows to a computing cluster has to cope with several challenges. In particular, the fundamental Divide and Conquer method of parallel computing cannot be applied directly [3]. This may also be an issue when astronomical workflows are applied to subsets of high-resolution images of the Universe.

References

[1] P. Diamond: *Big Data from the SKA: data intensive science*. Conference *Big Data made in Germany*, June 29-30 2017, Berlin (Germany). http://bigdata.htw-berlin.de/17/slides/1.2_Diamond.pdf

[2] M. Becker et al.: *Memory-driven computing accelerates genomic data processing*. *bioRxiv*, 519579, doi:10.1101/519579 (2019).

[3] M. Strutz, H. Heßling, P. Hufnagl: *A Gray-box Testing Method for Divide&Conquer in Image Processing*. IEEE Big Data 2019, December 9-12 2019, Los Angeles (USA).

Bio-inspired Information Processing: The Future of Artificial Intelligence?

Hermann Kohlstedt

Kiel University
Faculty of Engineering
Institute for Electrical Engineering and Information Technology
Chair of Nanoelectronic
Kaiserstr. 2, 24143 Kiel, Germany

Information processing in biological nerve system is characterized by highly parallel, energy efficient and adaptive architectures in contrast to clock driven digital Turing machines. Even simple creatures outperform supercomputers when it comes to pattern recognition, failure tolerant systems and cognitive tasks. Fundamental building blocks leading to such remarkable properties are neurons as central processing units, which are (with variable strengths) interconnected by synapses to from a complex dynamical three dimensional network. The field of neuromorphic engineering aims to mimic such biological inspired information pathways by electronic circuitries [1]. Up to today, this approach is hindered by an inadequate understanding how to link the information pathways on the local, synaptic and neuron level to the global functionality of the entire brain network [2]. Pulse-coupled oscillators combined with memristive devices are an interesting approach to mimic basal information processes of nervous systems [3].

In the talk I will show restrictions of conventional IT and present alternative computing architectures, which are currently under investigations. The challenges and possible limitations of bio-inspired computing approaches will be addressed.

References

[1] G. Indiveri, et al., Neuromorphic Silicon Neuron Circuits, Frontiers in Neuroscience. 5 (2011) 73. https://doi.org/10.3389/fnins.2011.00073.

[2] D.S. Bassett, O. Sporns, Network neuroscience, Nat Neurosci. 20 (2017) 353–364.

[3] M. Ignatov, M. Hansen, M. Ziegler, H. Kohlstedt, Synchronization of two memristively coupled van der Pol oscillators, Appl. Phys. Lett. 108 (2016) 084105. https://doi.org/10.1063/1.4942832.

Are we ready for the next level of Big Data?

Dieter Kranzlmüller

LRZ, Garching, Germany

Data pipelines for Euclid

M. Kümmel

LMU Faculty of Physics, Scheinerstr. 1, 81679 München, Germany

The **Euclid** satellite [1,2] is an **ESA** mission scheduled for launch in 2022. It will observe an area of 15,000 deg^2 with two instruments, the Visible Imaging Channel (**VIS**) and the Near IR Spectrometer and imaging Photometer (**NISP**). Ground based imaging data in *griz* from ground based surveys such as the Dark Energy Survey (**DES**), **Pan-STARRS** and **LSST** complement the **Euclid** data to enable photo-z determination. The mission investigates the distance-redshift relationship and the evolution of cosmic structures by measuring shapes and redshifts of galaxies and clusters of galaxies out to redshifts ~2.

In this presentation I will discuss and illustrate the various pipelines that are currently being implemented to handle the expected several Petabyte of data. The **Euclid** data reduction is in the responsibility of the Euclid Science Ground Segment (**SGS**), which is split up into several Organizational Units (**OU**) and Science Data Centers (**SDC**). The 10 **OU**'s are each responsible to prototype one or several pipelines to execute a certain reduction step, such as reducing the **VIS** data or measuring the galaxy shapes for weak lensing. The nine national **SDC**'s then implement the pipelines into the Euclid framework and run the optimized versions in the production environment. The entire pipeline system is centrally orchestrated by the Coordination & Orchestration System (**COORS**). The Euclid Archive System (**EAS**) has a central role to store metadata, distribute the data across the **SDC**'s and interface all ground system components.

In order to test all the different components and aspects of the **SGS** we conduct a series of so called **Science Challenges**. In a Science Challenge the various pipelines are run on simulated data. In each Science Challenge the survey area and the number of simulated effects are increased, with the final goal to eventually reduce in the **SGS** the full complexity of the data to the required accuracy.

References

[1] Laureijs, et al. 2011, ArXiv e-prints. 1110.3193

[2] https://www.euclid-ec.org/

Physics and machine learning

Katharina Morik

Technische Universität Dortmund, Germany

Knowledge Gain in the Age of HPC and Big Data

Susanne Pfalzner

Forschungszentrum Jülich , Jülich Supercomputing Centre, Jülich, Germany

Accessing complex structures with unsupervised and deep-learning techniques

K. L. Polsterer[1]

[1]*HITS gGmbH, Heidelberg, Germany*

The amount and size of astronomical data-sets was growing rapidly in the last decades. Now, with new technologies and dedicated survey telescopes, the databases are growing even faster. VO-standards provide an uniform access to this data. What is still required is a new way to analyze and tools to deal with these large data resources. E.g., common diagnostic diagrams have proven to be good tools to solve questions in the past, but they fail for millions of objects in high dimensional features spaces. Besides dealing with poly-structed and complex data, the time domain has become a new field of scientific interest. By applying technologies from the field of computer sciences, astronomical data can be accessed more efficiently. Machine learning is a key tool to make use of the nowadays freely available datasets.

This talk exemplarily presents how to create an explorative access to large astronomical data-set and how to utilize proper scoring rules, to train deep architectures more efficiently.

References

[1] Galvin, T. J., Huynh, M., Norris, R. P., Wang, X. R., Hopkins, E., Wong, O. I., Shabala, S., Rudnick, L., Alger, M. J., & Polsterer, K. L. (2019), PASP, 131, 108009.

[2] Polsterer, K. L., & Taylor, M. B. (2017), Astronomical Data Analysis Software and Systems XXV, 512, 485.

[3] Crawford, E., Norris, R. P., & Polsterer, K. (2017), Astronomical Data Analysis Software and Systems XXV, 512, 109.

[4] Kügler, S. D., Gianniotis, N., & Polsterer, K. L. (2016), MNRAS, 455, 4399.

Brain Inspired Computing

J. Schemmel[1]

[1]*Heidelberg University, Heidelberg, Germany*

Brain Inspired or Neuromorphic Computing, as a realization of Non-Turing, in-memory, event-based computing, will allow us to overcome the power wall our CPU-centric CMOS technology is facing. But that does not mean that the era of Turing-based computing will come to an end soon, or that Turing-based computing does not have its place in the neuromorphic world. This talk will shortly summarize how the Heidelberg BrainScaleS-2 accelerated analog neuromorphic architecture balances Turing and Non-Turing computing to combine power efficiency with the necessary flexibility and programmability, thereby reducing the resource requirements of AI and extending it by recent insights from neuroscience. These bio-inspired AI technologies may be beneficial for the data challenges the next generation of science instrumentation is facing. Possible applications of the BrainScaleS technology in the areas of edge computing and high-energy physics will be presented.

The Cherenkov Telescope Array
Data Management Model

Stefan Schlenstedt[1]

[1]CTAO, Heidelberg, Germany

The Cherenkov Telescope Array (CTA) will be the first ground-based observatory for gamma-ray astronomy at very-high energies. CTA, as an open observatory, will pose particular challenges to the computing and software during science operations, for example the proposal handling and scheduling and user support. CTA will produce several PB of raw data per year stored in a bulk data archive. These data will be processed and reduced to a much smaller volume of data stored in the CTA science archive. The latter data will be publicly accessible to the worldwide astronomical and particle physics communities. An overview of plans for managing, processing, and dissemination of the data is presented.

Enabling Data-Intensive Computing & the EOSC

A. Streit[1]

[1]Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Transforming data into knowledge is considered to be the 4th paradigm in science next to theory, experiment and simulation and enables the engineering of digital futures in science, industry and society. Federated e-infrastructures, data analytics, AI/ML and cloud computing are key technologies in supercomputing & big data, which can be summarized by "data-intensive computing". Although increasingly required and applied in many scientific disciplines, an efficient and scalable use of these technologies requires close cooperation between informatics and mathematics research, professional operation of IT infrastructures, research software engineering and the domain science disciplines.

The talk will give an overview on the activities of the Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT), Germany in this topical domain. Practical examples of interdisciplinary research, development and support results will be given stemming from activities at KIT, in Baden-Württemberg, in the Helmholtz Association as well as in the international context are given.

Finally, an introduction and overview on the idea, the evolutionary history and some current activities towards realizing the European Open Science Cloud (EOSC) are presented.

Lessons from the Sloan Digital Sky Survey

A.S.Szalay[1]

The Johns Jopkins University, Baltimore MD 21218, USA

Science is changing. While in the past many experiments followed each other incrementally, today's large science project are getting ever larger, soon reaching the point then the next experiment can only built by the whole world. This will fundamentally change the data lifecycle: while the incremental sequence of experiments results in a short useful life of the data, the emergence of the "ultimate world-wide instruments" means that the data generated by these are not likely to be superseded for several decades, they are here to stay. Furthermore, the community is using these data with the expectation of open, accessible and sustainable smart services, often involving a large database with an intuitive visual interface and additional programming APIs.

Many large-scale astronomy projects created such unique high-value data sets at a cost of hundreds of millions of dollars. While the projects are active, they are mandated to create highly sophisticated, smart archives, used by a large community. However, there are no significant efforts about maintaining these data beyond the point when their instruments reach the sunset. This will soon lead to a serious data loss unless there is a focused effort to come up with long-term policies and sustainability solutions. Built upon lessons learned from the Sloan Digital Sky Survey, we outline the challenges involved and make a few recommendations toward a sustainable data infrastructure. While many of the problems are more general than astronomy, it is likely that the first real challenges in this space will impact current astronomical surveys and require immediate attention.

Science Data Centres for Radio Astronomy: from LOFAR to SKA

Michiel van Haarlem (ASTRON, The Netherlands)

Modern digital radio telescopes produce enormous volumes of data. This is due to the large numbers of antennas or dishes, the wide bandwidths, high frequency resolution and high bit-rates. The Square Kilometre Array project is an international effort to build the world's largest radio telescope. The raw sensor output data rate is of SKA will be of order 1 petabit per second. This provides engineers and scientists with the challenge of transporting, processing and reducing vast amounts of radio astronomy data, firstly from the array of antennas and telescopes in the field through intermediate processing facilities to data processing centres on a national scale. Once the data has been processed - to the level of science ready data products - they are to be distributed and archived by a global network of SKA Regional Centres. The design of the SRC network is taking shape now, fuelled in Europe by H2020 projects like AENEAS, EOSC-Hub and ESCAPE. Until SKA is operational, in the second half of this decade, precursor and pathfinder telescopes play an important role in preparing for the future. A prime example is the International LOFAR Telescope, which has pioneered many of the technologies employed in the design of the SKA-Low telescope in Australia, and is already producing data at a rate of 7 petabytes per year. LOFAR data and software will be used extensively in the coming years to prepare for SKA. I will present the challenges and plans to set up and operate the global SRC network and the initiatives to tackle these challenges.

Simulating has dynamics in galaxies: a 3D view of star formation and feedback

Stefanie Walch-Gassner

University of Cologne, I. Physics Institute, Köln, Germany

IT infrastructure of the future
Envison the future of Computing!

Ingolf Wittmann

IBM Deutschland GmbH

Cognitive computing is on everyone's lips as the limitations of Moore's Law, so there are new technologies and approaches necessary to make cognitive solutions a reality. This is also affecting HPC environments, instead of brut force simulation intelligence is coming into the game with new AI approaches. For that there is a need for HPC neuromorpic computing based on accelerators like GPUs, FPGAs, neuromporphic chips, or quantum computing. The presentation will point out based on real examples how IT environments can benefit from such solutions and technologies to drive cognitive solutions and machine/deep learning including future processor technologies and compute architectures which still will allow the immense increase in performance for HPC environments.

Abstracts of Posters

(in alphabetical order)

WWU Cloud - Open Source based Cloud Services at the University of Münster

M. Blank-Burian[1]

[1]University of Münster, Münster, Germany

At the University of Münster (WWU), we provide cloud services for scientific and non-scientific applications at no cost to all faculty members (WWU Cloud). Our services are based on OpenStack as an IaaS Platform using Ceph as storage backend. The deployment of OpenStack is facilitated by using a bare metal Kubernetes Cluster, which itself is deployed via Airship Armada. Using this toolstack we provide VMs as well as CephFS/NFS/SMB shares to our users. This newly developed cloud platform serves as a blueprint for some partner universities in NRW [1]. Together, these clouds can host multi-cloud services for high availability and georedundant storage.
At the WWU, we also operate a Kubernetes cluster within VMs for use by our IT staff. It is used to provide services like JupyterHub, BinderHub, Gitlab CI runner and RDM software. We plan to extend this cluster using Istio and EdgeFS to provide a multi-cloud solution.

References

[1] R. Vogl, J. Hölters, M. Ketteler-Eising, D. Rudolph, M. Blank-Burian, H. Angenent, C. Schild, S. Ost., EUNIS 2018, 61 (2018)

FAIR in astronomy context

Harry Enke

Leibniz Institute for Astrophysics Potsdam (AIP), EScience & SuperComputing, Potsdam, Germany

The FAIR principles are currently considered as the guidance for good data for the digital age.

I will discuss two questions:

Is astronomy data already FAIR data?

Is making data FAIR already sufficient for the scientific purposes?

Daiquiri - Python based framework for the publication of scientific databases

A. Galkin[1] , **J. Klar**[2] **and H. Enke**[1]

[1]Leibniz Institute for Astrophysics Potsdam (AIP), Potsdam, Germany
[2] independent

At Leibniz Institute for Astrophysics Potsdam (AIP)[1] we host, curate, and publish terabytes of astrophysical data using the Daiquiri framework. Dedicated web applications allow scientists from all around the world to run SQL queries and get their desired data in reasonable time. In the last two years, Daiquiri was completely re-written in Python and received major updates - upload of VOTables, VO TAP access and many more features. Daiquiri has been developed in close cooperation with scientists and with support for collaborations in mind. This poster will introduce the key features of the Daiquiri framework and the technologies behind it. All components are Open Source software and available on GitHub [2].

Today, the publication of research data plays an important role in astronomy and astrophysics. Dedicated surveys like RAVE or massive simulations like Millennium and MultiDark are initially planned to release their data for the community. But also individual scientific projects strive to publish their data as a key requirement demanded by the funding agencies. Web sites are used as entry point to the publication of the data. Most of the web sites are tailor made for the particular case and are therefore not easily transferable to future projects.

At Leibniz-Institute for Astrophysics Potsdam (AIP), we gained experience with both the maintenance and the development of such applications. It became, however, apparent that the amount and the complexity of the applications constitutes a major challenge for maintenance expenses and scalability. In order to address these issues, we developed the Daiquiri framework, which is particularly designed to allow for different highly customizable web applications based on a common easily maintainable code base.

Since 2013, Daiquiri has enabled researchers from all over the world to access data from the RAVE survey, the APPLAUSE database[3] or the CosmoSim the cosmological simulations database.

References

[1] E-Science at AIP https://escience.aip.de/

[2] Daiquiri on GitHub https://github.com/django-daiquiri/daiquiri

[3] Archives powered by Daiquiri https://gaia.aip.de, https://www.plate-archive.org

A new multi-band optical image pipeline for the Magellan 6.5 m telescope.

Zohreh Ghaffari[1], Catalina Sobrino Figaredo[1], Martin Haas[1], Rolf Chini[1], Steve Willner[2]

[1]Astronomisches Institut Ruhr-Universität Bochum, Universitätsstraße 150, D-44801 Bochum, Germany
[2] Harvard-Smithsonian Centre for Astrophysics, Cambridge, MA, USA.

We present a new image reduction pipeline for PISCO, the Parallel Imager for Southern Cosmology Observations, attached to the 6.5 m Magellan telescope at the Las Campanas Observatory, Chile. PISCO obtain simultaneous g, r, i, z band images with a pixel size of 0.2" and 5' X 8' field of view.

Our pipeline package performs all basic standard reduction steps on the raw images of each CCD and removes several instrumental contaminations. We also apply astrometry, implement photometric calibration and construct deep co-added images in each band. We show the procedure of reducing LCO images from raw data to the final results and illustrate the quality of our data reduction by comparison with PANSTARRS.

Special emphasis is placed on a high-fidelity photometric calibration. This is indispensable for our research purpose to study the evolution of galaxy clusters around 3C radio sources in the early universe (z > 1).
The reduction allows for the reliable detection of faint sources down to r = 27 mag. For four 3C fields, a cross-match with Hubble Space Telescope catalogues in 2' X 2' field of view demonstrates the exceptional depth and significance of our source catalogue.

References

[1] Brian Stalder et al., "PISCO: the Parallel Imager for Southern Cosmology Observations", Proc. SPIE **9147**, Ground-based and Airborne Instrumentation for Astronomy V, 91473Y (2014)

[2] Z. Ghaffari et al., "Galaxy overdensities around 3C radio galaxies and quasars at 1<z<2.5 revealed by Spitzer 3.6/4.5µm and Pan-STARRS", Astronomische Nachrichten, **338**, pp. 823-840 (2017)

Molecular Outflow Detection in G327: A 3D Approach

Niraj Kandpal[1] , A. Sanchez Monge[1], Peter Schilke[1]

[1] I. Institute Physics Cologne

Email : kandpal@ph1.uni-koeln.de

Molecular outflows are an energetic mass-ejection phenomenon associated with very early stage of stellar evolution. Depending on whether the flow is moving towards or away from us they can be assigned as blue and red shifted. Molecular outflows can be a useful tool for understanding the underlying formation process of stars of all masses, as they provide a record of mass-loss history of the system .

Usual methods of outflow detection involve 2D contour analysis . We here present 3D analysis of the ALMA data in the molecular line SiO(5-4) at 217 GHz for G327 which is a star forming region with many outflows . We found that analyzing the data using usual 2D methods, it was not possible to disentangle the outflows. While with 2D contour analysis, we could find only around 21 outflows, we could find around 42 outflows using 3D contour analysis. By generating the 3D skeletal structure of our data, we can also relate the blue and red shifted outflows and the find the direction of the outflows in 3D.

We have developed a semi automatized way to calculate the outflow parameters(mass, momentum, energy, mass loss rate) by using ellipsoid as a mask and rotating it to calculate the flux at each outflow lobe. Using Monte Carlo simulation from Kong et. el. 2019 we found the relation between outflow direction and G327 filament which seems to imply that most outflows are perpendicular to the filament.

References

[1] Kong et. el. 2019

[2] Scoville et. el. 1986

Searching Pulsars Using Neural Networks

Lars Künkel[1], Joris P. W. Verbiest[1,2] and Rajat M. Thomas[3]

[1]Universität Bielefeld, Bielefeld, Germany
[2]Max Planck Institute for Radio Astronomy, Bonn, Germany
[3]University of Amsterdam , Amsterdam, Netherlands
E-mail: lars.kuenkel@uni-bielefeld.de

We are developing a pulsar search pipeline that utilises neural nets to detect pulsars. Pulsar signals consist of very faint, periodic pulses that have been dispersed by the interstellar medium. Pulsar searching has been traditionally plagued by creating millions of false pulsar candidates. Increases in telescope sensitivity force us to build increasingly sophisticated classification systems. Most of the machine learning effort that has been put into pulsar surveys so far has been put into classifying pre-computed pulsar candidates while we propose a pipeline than can be trained on survey observations directly in an end-to-end fashion.

Specifically out architecture combines a convolutional neural network with typical search techniques like the FFT and the fast folding algorithm (FFA) to create a pulsar detection pipeline that can detect faint pulsars while having a low number of false positives.

We show that convolutional layers can be trained to dedisperse pulsars without knowing the pulsar dispersion measure (DM). This approach is competitive with dedispersing the pulsar using the real pulsar DM.

Here we present preliminary results demonstrating the power of this approach and its competitiveness compared to existing methods in regards to detection sensitivity and false positive rates.

Metadata and User-Provided Data in the LOFAR Long Term Archive

J. Künsemöller[1], H.A. Holties and G.A. Renting[2]

[1]Bielefeld University, Bielefeld, Germany
[2] ASTRON, Dwingeloo, Netherlands
E-Mail: jkuensem@physik.uni-bielefeld.de

The International LOFAR Telescope (ILT) is a large low frequency interferometric radio telescope, distributed throughout several European countries. It operates the LOFAR Long Term Archive (LTA) to archive and serve its large and growing dataset. In three physical locations (SURFsara/Netherlands, FZ Jülich/Germany, and PSNC/Poland), the LTA currently stores about 49 Petabyte in 10 million dataproducts.

For discovery and data access, metadata is kept in a separate catalog database and describes each dataproduct and its full provenance in detail. To deal with a changing datamodel over time, the LTA follows principles of the Open Archival Information System (OAIS) model. When adding a dataproduct to the archive, the ILT control software provides a Submission Information Package (SIP) in XML format with information about the original measurements and every processing step applied. Each SIP gets validated against a strict and versioned schema before the data is stored and its metadata is added to the catalog. System changes are reflected in both an updated version of the SIP schema as well as in the database itself.

Adding user-provided derived dataproducts to the LTA poses a challenge concerning the consistency and cross-linking of dataproducts. The user has to provide a valid SIP that not only completely describes the entire genesis of the new dataproduct, but also has to correctly refer to any existing dataproducts in the LTA that it was derived from. We provide services and a Python module to request information and LTA identifiers on the base data in form of a base SIP. The users can then extend that by programmatically adding the processing steps that they applied to create the derived dataproduct. Unique LTA identifiers can be linked to custom user-specified labels for reference by the providing user. We further provide tools to validate and visualize the outcome on the user-end. Each catalog entry is further associated with a user-specific identifier to allow filtering in data discovery and for potential rollbacks.

References

[1] G. A. Renting, H. A. Holties: *LOFAR Long Term Archive*, Proc. ADASS XX, ASP Conf. Ser. 442, 49 (2011)

[2] H. Holties, A. Renting, Y. Grange: *The LOFAR long-term archive: e-infrastructure on petabyte scale*, Proc. SPIE 8451, 451-456 (2012)

PyParadise: A simultaneous pipeline of stellar and gas kinematics

Man I Lam[1], **Bernd Husemann**[2], **Omar S Choudhury**[1],

and

C. Jakob Walcher[1]

[1] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany

[2] European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching b. München, Germany

PyParadise is a state-of-the-art code, which is based on MCMC. It derives stellar population, stellar kinematics, gas kinematics at the same time. In this poster, we are going to present the fitting results based on mock and observed data. We also compare our pipeline to some existed methods, and conclude that our method is stable in velocity measurement up to z~1.

Exoplanet detection using Machine Learning

A. Malik[1]

[1]Universitäts-Sternwarte, Ludwig-Maximilians-Universität, München, Germany

We introduce a machine learning based technique to detect exoplanets using the transit method.

Machine learning and deep learning techniques have proven to be very useful in various scientific research areas. We would like to exploit some of these methods to improve the conventional algorithm based approach used in astrophysics today to detect exoplanets.

We used popular time-series analysis library 'TSFresh' to extract features from lightcurves. For each lightcurve, we extracted 789 features. These features capture information about the characteristics of a lightcurve. We used these features later to train a tree-based classifier using a widely used machine learning tool 'XGBoost'.

To train this model, we used data from the K2 mission. We removed all the known sources and randomly injected planet transits. Our model is trained to detect these planet transit signals.

The current test data has many cases where signal-noise ratio is very low as planet injection was completely random. This model was able to correctly predict with an accuracy of ~85 % i.e. it was able to classify planet vs non-planet signal correctly in 85% of the cases. It was able to identify planet lightcurves with a precision of 0.78.

This approach can be easily applied to different kinds of lightcurves and is able to detect single as well as multi-planet transit signals.

Entering NeuLAND: Analysis workflow preservation for a fair FAIR

J. Mayer[1] and A. Zilges[1]

[1]Institute for Nuclear Physics, University of Cologne, Germany

It is a long way from the data recorded in an experiment to the tables, figures, and values in a publication, a way that is not always obvious to others – maybe not even within the same collaboration. One might even encounter situations where a PhD left after analyzing his experiment, but before publishing. Maybe analysis scripts, dozens of ROOT-C-Macros with thousands of LoC, give segmentation faults instead of reproducing the right values and must be fixed with weeks of effort. Not even to mention applying an existing analysis pipeline to new data.

Workflows should be implemented and stored such that they are **f**indable, **a**ccessible, **i**nteroperable, and **r**e-usable (fair). Here we present ideas and open questions for workflow preservation at the example of simulation and event reconstruction with machine learning for NeuLAND, the New Large Area Neutron Detector. This high-resolution, large-acceptance time-of-flight spectrometer is a key piece for Reactions with Relativistic Radioactive Beams (R^3B) at FAIR, the Facility for Antiproton and Ion Research.

Data Infrastructure at the University of Cologne's Institute for Nuclear Physics

J. Mayer[1], M. Müller[1] and A. Zilges

[1]*Institute for Nuclear Physics, University of Cologne, Germany*

At the university of Cologne, the institute of nuclear physics oversees its own computing and data storage resources. The system was set up and is still maintained by masters and Ph.D. students, instead of relying on the university's central computing service. For a rather small institute at a quite large university, this unusual approach has nevertheless served well since its inception over 40 years ago.

However, developments in data acquisition and detector physics as well as an increasing amount of collaborations have led to ever larger amounts of data to be transferred, stored, and processed. In addition, new mindsets and paradigms, like compliance with the fair-principles or workflow preservation, might render this approach inadequate in the future. Here, the NFDI initiative could provide both the required concepts and processing power to level up our data analysis processes.

Past planetary engulfment as a possible explanation for observed high stellar rotation on metal poor main sequence stars

A.Oetjens[1 ,2] , M.Bergemann[2] and L.Carone[2]

[1]Ruprecht-Karls Universität , Heidelberg, Germany
[2] Max-Planck-Institute for Astronomy, Heidelberg, Germany

Gyrochronology - the analysis of stellar rotation - is a standard astronomical method to determine the ages of stars. Since the fundamental (Skumanich 1972) relationship that predicts that main-sequence stars spin down as $t^{-0.5}$, this canonical method has been widely used to provide a diagnostic of stellar ages (Angus et al. 2019). However, recently advances in age tagging of stars by asteroseismology have revealed severe discrepancies with ages derived by gyrochronology (e.g. Sahlholdt et al. 2019). In this work, we explore whether the tension could be explained by the tidal interaction of a massive planetary companion with its host star.

We develop a new model that describes the evolution of angular momentum of main-sequence stars in the presence of a Jupiter-mass planet. Our model relies on analytical recipes for angular momentum evolution (Bouvier 1997), tidal friction (Privitera et al. 2016), and stellar spin-up (Carone 2012).

We apply this model to an ensemble of synthetic star-planet configurations.

We find that, in most cases, the dynamical evolution of the star-planet system leads to a gradual spin-up of the main-sequence star from a few km/s up to 40 km/s. But the time it takes for planet to be engulfed by the star depends on the initial orbit, mass and metallicity of the system.

We compare our results with the observed sample of ~32,000 stars with accurate measurements of [Fe/H], masses, and rotation periods (Huber et al. 2014, McQuillan et al. 2014). The data suggest that ~10% of metal-poor old main-sequence stars exhibit very high rotation rates, in stark contrast with gyrochronology models.

In our scenario, this can be naturally explained by the interaction of a star with its planet. We are able to confine the mass-orbit parameter space of the planet, before the engulfment happened. Our model provides a viable alternative to more complex and poorly-understood scenarios, such as the inefficiency of magnetic breaking. In summary, our results suggest that the observed distributions of rotation periods and metallicities of stars in the Galaxy can be explained by the dynamical interaction in a star-planet system, and it may also explain the discrepancies between gyrochronology and asteroseismology.

Improving reliability of photometric redshifts using machine learning methods

O. Razim[1] and G. Longo[1]

[1]*Department of Physics, Strada Vicinale Cupa Cintia, 21, 80126, University Federico II, Napoli, Italy*

Upcoming wide-field surveys, such as LSST and Euclid, will require precise and reliable photometric redshifts (photo-z) for fulfilling their goals [**1**,**2**]. Two most popular methods for obtaining photo-z, machine learning algorithms and Spectral Energy Distribution (SED) fitting, rely on spectroscopic samples for training and calibration correspondingly [**3**]. Spectroscopic catalogs are usually much shallower than photometric ones. This difference and contamination of spectr-z catalogs are a possible source of biases.

In this work, we use unsupervised machine learning method called Self-Organizing Maps (SOM) [**4**] to improve the reliability and precision of photo-z catalogs. We use COSMOS2015 catalog [**5**], which provides photometry and SED photo-z, to calculate machine learning photo-z with Multi-Layer Perceptron with Quasi-Newtonian Algorithm (MLPQNA) [**6**,**7**]. After that, we train SOM and determine photometric and spectr-z outliers for each cell of the map. Removing these outliers with different threshold coefficients allows us to select subsamples of the catalog for which we have the most reliable photo-z. The improvement of standard deviation of residuals can be as good as from ~0.05 to ~0.02, with the reduction of the percentage of outliers from ~2% to ~0.2%.

References

[1] R. Laureijs et al., Euclid Definition Study Report (2011)

[2] LSST Science Collaboration et al., LSST Science Book, Version 2.0 (2009)

[3] M. Salvato et al., Nature Astronomy Volume 3, p. 212-222 (2019)

[4] T. Kohonen, Biological Cybernetics volume 43, pages 59–69 (1982)

[5] C. Laigle et al., The Astrophysical Journal Supplement Series, Volume 224, Issue 2, article id. 24, 23 pp. (2016).

[6] S. Cavuoti et al., Astronomy & Astrophysics, Volume 546, id. A13, 8 pp.

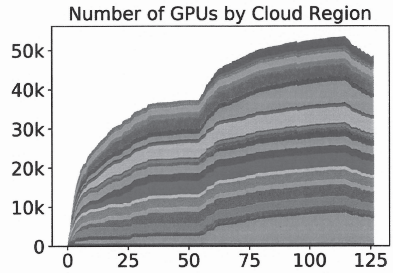
[7] M. Brescia et al., The Astrophysical Journal, Volume 772, Issue 2, article id. 140, 12 pp. (2013).

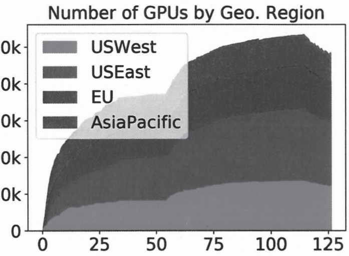
Cloud burst for Multi-Messenger Astrophysics

Igor Sfiligoi[1], Frank Wuerthwein[1], Benedikt Riedel[2] and David Schultz[2]

[1] University of California San Diego, La Jolla CA 92093, USA
[2] University of Wisconsin - Madison, Madison WI 53715 USA

As we approach the Exascale era, it is important to verify that the existing frameworks and tools will still work at that scale. Moreover, public Cloud computing has been emerging as a viable solution for both prototyping and urgent computing. Using the elasticity of the Cloud, we have put in place a pre-exascale HTCondor setup for running scientific simulation in the Cloud, with the chosen application being IceCube's [1] photon propagation simulation [2]. This was not a purely demonstration run, but it was used to produce valuable and much needed scientific results for the IceCube collaboration. In order to reach the desired scale, we aggregated GPU resources across 8 GPU models from many geographic regions across Amazon Web Services, Microsoft Azure, and the Google Cloud Platform. Using this setup we reached a peak of over 51k GPUs corresponding to almost 380 PFLOP32s, for a total integrated compute of about 100k GPU hours. In this paper we provide the description of the setup, the problems that were discovered and overcome, as well as a short description of the actual science output of the exercise.





References

[1] M.G. Aartsen et. al., The IceCube Neutrino Observatory: Instrumentation and Online Systems, JINST 12 (2017) no.03, P03012, DOI: 10.1088/1748-0221/12/03/P03012, arxiv: 1612.05093

[2] Dmitry Chirkin, Photon tracking with GPUs in IceCube, NIMA, Volume 725, 2013, Pages 141-143, DOI: https://doi.org/10.1016/j.nima.2012.11.170.

Machine Learning in Cherenkov Astronomy

B. Schleicher[1] and D. Dorner[1]

[1]*Julius-Maximilians-Universität Würzburg, Würzburg, Germany*

Being a data-intensive and analysis-intensive field, Cherenkov astronomy has several use cases for machine learning methods. For example, in the background supression, it can be used to differntiate between gamma rays and cosmic rays as primary particle of the shower that produced the observed Cherenkov light. Also for reconstructing the origin or the energy of the events, machine learning can be applied. The challenge for all these use cases is that for training the methods, simulated data are needed. If the simulated events do not describe the real data correctly, the machine learning methods do not provide equally good results on the real data. On the other hand, generative adversarial networks might help to reduce the mismatch between real and simulated data. Furthermore, machine learning methods are interesting for the high-level analysis, e.g. for studying light curves and predicting the behavior of a source. Therefore systematic studies of variability and periodicity can profit from machine learning approaches. The long-term goal is to predict the flux for variable sources and coordinate multi-wavelength observations and studies based on this.

The PAHN-PaN consortium at the NFDI

K. Schwarz[1] **and T. Schoerner-Sadenius**[2]

[1]GSI, Darmstadt, Germany
[2] DESY, Hamburg, Germany

The aim of the German national research data infrastructure (NFDI) is to systematically help in managing scientific and research data. The NFDI will bring multiple stakeholders together in a coordinated network of consortia tasked with providing science-driven data services to research communities.
One of the consortia planning to contribute to the NFDI is the PAHN-PaN consortium.

The PAHN-PaN communities have always been at the forefront of technological developments. Today, due to the development of new accelerators, new observatories and experiments, and new detectors with increased resolutions and higher event rates, PAHN physics is experiencing a rapid increase of data rates and volumes and also a more diverse access sharing. This boost of data leads to ever increasing demands on data analysis power and methods, and on data management capabilities.
The goal of the PAHN-PaN Consortium is to develop solutions for the data challenges and to help setting up the structures necessary for this endeavour. These structures will facilitate the exploitation of synergies within the consortium, easy transfer of knowledge and technology to and from neighbouring consortia and communities, and the establishing of relevant services for PAHN-PaN and the entire NFDI. These goals are pursued in dedicated task areas and cross-cutting topics:

Details are presented in a poster contribution.

References

[1] https://www.pahn-pan.de/

Astronomy meets big data: Improving the Milky Way model with the billion-star surveyor Gaia

K. Sysoliatina[1] and A. Just[1]

Astronomisches Rechen-Institut (ARI), Mönchhofstr. 12-14,
69120 Heidelberg, Germany
E-mail: Sysoliatina@uni-heidelberg.de

The ESA's mission Gaia has already mapped about 1.6 billion of stars in the Milky Way (DR2, [1]) that corresponds to about 200 Tb of raw data during its five-year nominal mission. For most of these stars five astrometric parameters (positions, proper motions and parallaxes) are known, and a 7.2-million subset of bright stars additionally contains radial velocities, thus providing us with the full 6D dynamic information. The high quality and abundance of these data strongly stimulate the development of the existing Galactic models.

In this study, we view the Milky Way within the framework of the Just-Jahreiß model that solves Poisson's equation for an axisymmetric steady-state disk in the presence of gaseous, dark matter and stellar halo components [2]. To achieve the best consistency between our model and the new Gaia data, we adapt its parameters within the Bayesian approach by performing a parallelized MCMC sampling of the vertical number density profiles and W-velocity distributions of the different disk stellar populations in the Solar neighbourhood. The eight to ten fitting parameters describe the shape of the disk star formation rate, age-vertical velocity dispersion relation and the local surface density normalization. Our fitting approach imposes a strong constraint on the maximum model-to-data comparison time for a single combination of fitting parameters, which in its turn leads to special constraints on the data selection. As a result, we define easily modelable magnitude-complete data samples de-reddened with the local extinction map [3], of about 1 million stars in total. For the final comparison we further bin these stars with respect to their distance to the Galactic plane and W-velocity.

To sum up, the efficient exploration of the parameter space even of a relatively simple Galactic model, requires reduction of the large initial catalogue to much smaller set of its statistical properties.

References

[1] Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1

[2] Just, A. & Jahreiß, H. 2010, MNRAS, 402, 461

[3] Lallement, R., Capitanio, L., Ruiz-Dern, L., et al. 2018, A&A, 616, A132

From 2D to 3D to Graph Networks: Representing Detector Geometries in Neural Networks

A. Trettin[1]

[1]DESY, Platanenallee 6, 15738 Zeuthen, Germany

Convolutional neural networks (CNNs) have proven themselves to be powerful tools enabling computers to represent and recognize the content of images. They derive their power from exploiting translational invariance in 2D and 3D space. The signatures of particle interactions inside Cherenkov detectors such as IceCube also exhibit translational invariance, but the geometry of the detector seldom follows that of a regular square grid expected from a CNN. We present methods developed by the IceCube Collaboration to represent a non-square or even completely irregular detector geometry in such a way that CNNs can be used to recognize and reconstruct particle signatures in the detector.

Classification of high-resolution solar Hα spectra using t-SNE

M. Verma[1], G. Matijevič[1], C. Denker[1], A. Diercke[1,2], C. Kuckein[1], E. Dineva[1,2], H. Balthasar[1], I. Kontogiannis[1], P.S. Pal[1,3]

[1]Leibniz-Insitut für Astrophysik Potsdam (AIP), Potsdam, Germany
[2]Universität Potsdam, Institut für Physik und Astronomie, Potsdam, Germany
[3]University of Delhi, Bhaskaracharya College of Applied Sciences, Delhi, India

Starting mid-2020, the 4-meter Daniel K. Inouye Solar Telescope (DKIST) will become operational. With five post-focus instruments equipped with large-format detectors, the expected annual data volume is around 3 PB. Data include not only images but imaging spectropolarimetric data as well as high-precision full-Stokes spectra. This amount of data cannot any longer be inspected by astronomers. We propose a framework to classify solar spectra using t-distributed Stochastic Neighbor Embedding (t-SNE) to speed up basic spectral inversions. This exploratory study is based on high-spectral resolution Hα spectra obtained with the echelle spectrograph of the Vacuum Tower Telescope (VTT) located at Observatorio del Teide, Tenerife, Spain. The Hα spectral line is a well-studied absorption line, revealing properties of the highly structured and dynamic solar chromosphere. Typical features with distinct spectral signatures in Hα include filaments/prominences, bright active-region plages, superpenumbrae around sunspots, surges, flares, Ellerman bombs, filigree, and mottles/rosettes, among others. t-SNE is a machine learning algorithm, which is used for nonlinear dimensionality reduction. In this application, it projects the Hα spectra onto a two-dimensional map, where it is easy to classify them according to results of Cloud Model (CM) inversions, i.e., with respect to optical depth, Doppler width, line-of-sight velocity, and source function of the cloud material. Initial results of t-SNE indicate its strong discriminatory power to separate quiet-Sun and plage profiles from those that are suitable for CM inversion. In addition, the identified classes are linked to chromospheric features, the impact of seeing conditions on the classification is assessed, the projection of new Hα spectral data (different observing time and target) onto the 2D t-SNE maps is evaluated to optimize CM inversions, and representative Hα spectra are determined as input for deep neural networks speeding up the CM inversion.

NuRadioMC and NuRadioReco – A Software Framework for the Radio Detector Community

Christoph Welling[1,2] **for the RNO-G Collaboration**

[1]DESY, Platanenallee 6, 15738 Zeuthen, Germany
[2] Friedrich-Alexander Universität Erlangen-Nürnberg, Schloßplatz 4, 91058 Erlangen

In order to be able to detect neutrinos with energies beyond 10PeV, it is necessary to build detectors that can monitor many cubic kilometers of detector material. One approach to solving this problem is the detection of radio signals emitted by particle cascades from neutrino interactions in glacial ice. With construction of the first discovery-scale neutrino radio array starting in the coming summer, a software framework for the event simulation and reconstruction is needed. As deployment continues over several years, the detector layout is likely subject to change due to lessons learned from previously deployed stations, and future detectors may opt for entirely different configurations. Additionally, a radio neutrino detector is also able to detect the radio signals from cosmic-ray induced air showers, making them both a background and an important calibration signal. For these reasons, NuRadioMC and NuRadioReco aim to not limit themselves to a specific experiment, but to be flexible enough to serve a variety of radio experiments, including existing prototype in-ice neutrino detectors like ARIANNA and ARA and even air shower detectors like AERA and LOFAR. This is accomplished by a flexible detector description and data structure as well as a reconstruction process in which tasks are split up into individual modules that can be combined as needed. Modern software tools are used to speed up calculations, for data visualization and to manage the development process. We present the software's capabilities, the challenges arising from this flexibility and our ways of overcoming them.

Prototypes for the Next Generation of Computing Backends in Radio-Astronomy

A. Bansod[1], E. Barr[1], M. Heininger[1], S. P. Sathyanarayanan[1], T.Winchen[1], G. Wieching[1], J. Wu[1]

[1]Max Planck Institute for Radio Astronomy,
Auf dem Huegel 69, 53121 Bonn, Germany
E-mail: tobias.winchen@rwth-aachen.de

Traditional backends in radio-astronomy are comprised of highly specialized analog and digital hardware as data processing components to satisfy the transfer and processing requirements imposed by the extreme data rates. Only recent developments in high speed networking and PCIe mounted accelerator cards make new backend designs possible that are based on commercial-off-the-shelve (COTS) computing hardware. Here, we discuss the Effelsberg Direct Digitization (EDD) project and the FBFUSE and APSUSE instruments at the MeerKAT Radio Telescope as corresponding prototypes. In these we reduced the analog components in the signal chain as much as possible and perform processing of the digitized signals on COTS computing systems hosting GPUs and FPGA based systems. The resulting flexible backend designs are build on current industry standard hard- and software and can be deployed at different telescopes with non-to-minimum customization. Here we discuss the design and implementation of these next generation backends that may serve as template for future computing backends in radio-astronomy.

BoostNumpy: Big Data Processing in C++ with Python convenience

M. Wolf[1]

[1] *Technische Universität München, Physik Department,*
James-Franck-Str. 1, 85748 Garching, Germany

Efficient processing of big data can only be achieved with algorithms implemented in a compiled language like C++. However, for convenient steering of these compiled algorithms an interpreted scripting language like Python is desired. This contribution presents the meta-programming library "BoostNumpy" [1] that serves as interface between C++ and Python by utilizing the Boost.Python library [2] and the numpy software package [3] for high-performance big data storage management and processing.

References

[1] M. Wolf, github.com/martwo/BoostNumpy

[2] D. Abrahams and R. W. Grosse-Kunstleve, *Building Hybrid Systems with Boost.Python*. C/C++ Users Journal, (July 2003)

[3] S. van der Walt, S. C. Colbert and G. Varoquaux, *The NumPy Array: A Structure for Efficient Numerical Computation*, Computing in Science & Engineering, **13**, 22-30 (2011)