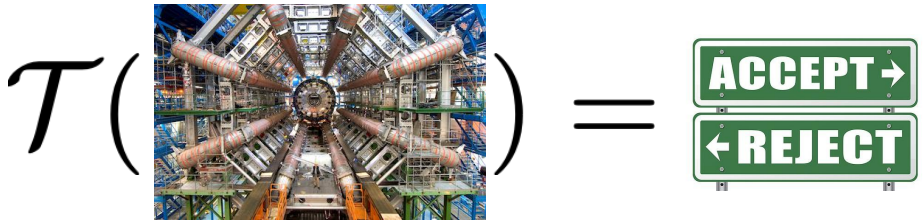
arXiv:2203.07620

Innovations in Trigger and Data Acquisition (TDAQ) systems for next-generation physics facilities

Authors:
Rainer Bartoldus (SLAC),
Catrin Bernius (SLAC),
David W. Miller (Chicago)

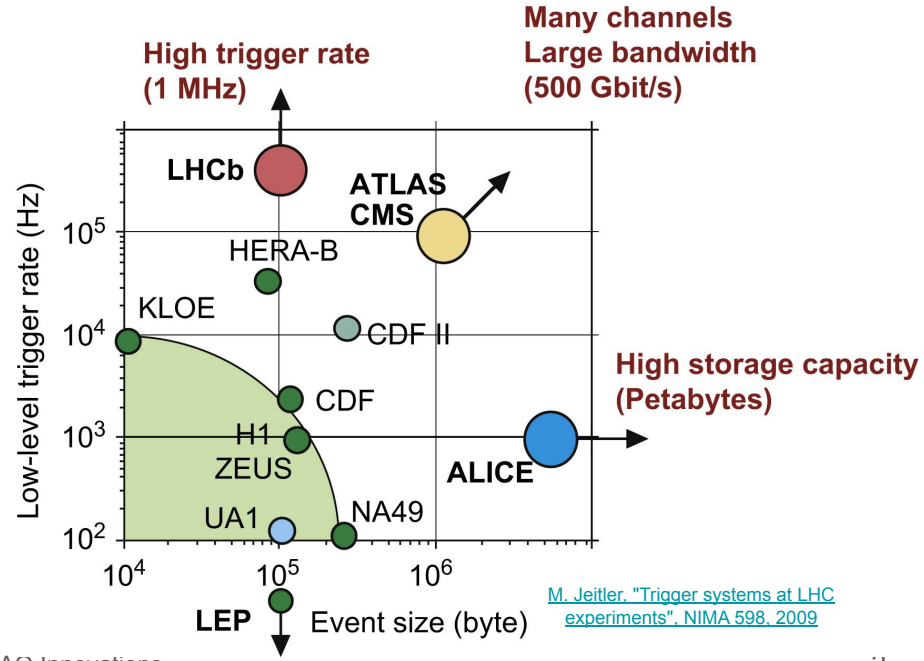
Introduction



TDAQ systems are a **critical interface between detector instrumentation and computing systems** at current and future physics facilities

Direct impact on the **volume and quality of the data** collected and the physics that can be extracted from them

⇒ Innovative solutions to the challenges faced by these facilities are essential

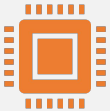


M. Jeitler, "Trigger systems at LHC experiments", NIMA 598, 2009

Three main themes for future facilities

Future TDAQ systems will have to leverage currently existing and available hardware, firmware, and software in new ways.

→ Use old stuff in new ways



TDAQ developers will have to find ways to adopt and incorporate the unique technical requirements of physics facilities into new hardware concepts.

→ Get new stuff to do what we want

The TDAQ community must ensure that the people, knowledge, and experience required to build, commission, and operate new TDAQ systems are fully supported and retained in the field.

→ Build & train a knowledge-based community

Challenges for next-generation TDAQ systems

Identify three sets of challenges and opportunities for addressing the needs of TDAQ systems of next-generation physics facilities based on the three themes:

1. Novel applications and uses of current generations of commodity and custom hardware, firmware, and software for TDAQ systems;

2. Confronting state-of-the-art devices and TDAQ paradigms with the specialized needs of future low-latency, high-throughput, and high-performance physics facilities;

3. Building, integrating, commissioning, and operating large-scale, heterogeneous, and dynamic TDAQ systems at future facilities
    - including the acquisition and retention of the required domain knowledge and technical expertise.

Challenges faced by current & future facilities (1)

Particular challenges at ever-evolving operating conditions of particle accelerators

- Higher energies, intensities, and data rates
- Signatures of interest become more and more challenging to analyze and select in high particle density environments

Example: *Fast track reconstruction at the trigger level*

- Valuable for searches for long-lived particles (e.g. displaced vertices and disappearing tracks) and exotic processes (e.g. *R*-parity violating SUSY)
- Speed achieved by embedding algorithms in highly-parallelized and potentially heterogeneous computing architectures

⇒ Usage of *heterogeneous* digital and analog solutions (CPUs, GPUs, FPGAs, ASICs, SOCs, …) together with advanced algorithms (Artificial Intelligence (AI), Machine Learning (ML))

See the next 3 talks by K. Hahn, A. Kotwal, T. Homes on track triggering as well as yesterday's Machine Learning talks by N. Tran, D. Shih, and others

Challenges faced by current & future facilities (2)

Upgraded experimental designs lead to expansion of both fiducial detector volumes and channel multiplicities by orders of magnitude, presenting some of the following challenges:

- Signal extraction in the **low-energy regime** challenging due to the enormous amount of additional, background-like contributions ***(e.g. DUNE Far Detector)***
- **High data rates**, require a scalable architecture of the DAQ system as well as specialized algorithms that are employed for signal identification and reconstruction ***(e.g. Project 8)***
- Computational approaches can be limiting data analyses ***(e.g. Multi-messenger astrophysics facilities, MMA)***
    - Challenges include the characterization of an **extremely high number of data channels** from multi-modal and heterogeneous detectors/sensors to provide low-latency data quality evaluation, optimizing and fitting data in high-dimensional space, …
    - Limiting factor for **low latency estimates** of gravitational wave signal significance to enable electromagnetic follow-up of gravitational wave sources

Heterogeneous and dynamic systems

Heterogeneous computing can help address the needs of high-throughput and high performance compute power together with machine learning algorithms which are well-suited to run on these new computing architectures

- **LHCb** experiment already using **fully GPU-based** implementation of first level trigger
- Various design proposals with **heterogeneous usage** for upgrade of **Mu2e** experiment (Mu2e-II)
- Novel **physics-inspired neural network image recognition** techniques for differentiating signals, for example of Higgs bosons, from the background physics processes using **MPSoC and FPGA** devices
- **FPGA-based artificial intelligence** inference in triggered detectors
- Concept of **self-driving trigger system** to autonomously and continuously learn to select, filter and process data

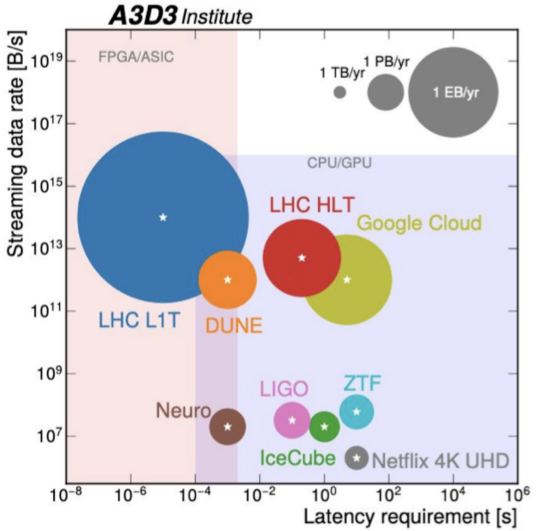


Figure produced by A3D3 team (NSF institute for "Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery")

Hyper-dimensional DAQ and processing systems at scale

Need for specialised solutions for:

- Extremely high-rate applications and high bandwidth requirements
- Highly tailored or varying temporal data structures
- Data integrity and robustness
- Reproducibility requirements

⇒ **Further push the envelope of industry standard devices and systems** to confront and challenge the commodity solutions with the specific needs of low-latency, high throughput and high performance physics facilities

Examples discussed in the paper and the following slides:

- Asynchronous first level trigger systems
- Edge computing devices for detectors
- Streaming DAQ

Example: Asynchronous first-level trigger systems

Increasingly challenging to distribute and synchronize stable, low-jitter, high-frequency clocks over very large systems of thousands of optical links

- Timing precision to reach O(10 ps) and number of boards greatly increases with granularity

Asynchronous system concept: event-driven instead of clock-driven

- Tag data with time marker only at detector front-end
- Then transport and process the data asynchronously up to the first level trigger decision
  - Same already done for higher-level triggers
- Event builder runs at full rate from detector: changes from *clock-driven* to *event-driven*
- Some portions of LHC first-level triggers already asynchronous
  - Trigger optical links not multiple of machine frequency
  - Time de-multiplexing scheme serves 1 bunch crossing of data to 1 individual trigger board
- **Blurs the lines between first-level (FPGAs) and higher levels (CPUs and now also GPUs)**

A set of FGPAs, CPUs and GPUs used to execute a mix of very fast and very complex algorithms

Example: Streaming DAQ (1)

Traditional TDAQ architectures, for decades, based on these fundamental assumptions:

- It is impossible to read out the detector at full rate
- Even if it were possible, one could not process the data at full rate
- Even if that could be done, it would create unsustainably large datasets for offline computing

But: ***continued advances in computing, networking and storage technologies have challenged these assumptions***

Most of today's TDAQ systems have **pipelined and triggered detector readout**

- Detector signals are digitized at fixed clock rates, written to **circular buffers**
- System reads part of these data, processes them, decides whether to keep or discard the buffer
- If the buffer is to be kept, full readout is initiated over the DAQ path and sent to the next stage

Hugely successful in working around rate and bandwidth limitations, but known shortcomings

Example: Streaming DAQ (2)

Pipelined mode limitations:

- Latency (processing time) limited by depth of front-end buffers (electronics)
- Trigger can become very complex to achieve rate reduction
- Readout data frames arrive randomly to due trigger latency (and jitter)
- Downstream event building has to collect and arrange all contributions while data are in motion

Known downsides:

- Trigger decision introduces selection biases
- Doesn't deal well with event pileup (what is an "event" becomes a trigger artifact)
- Performance ultimately constrained by latency and the need to execute within deadtime path
- Data reduction comes at a loss of information

Alternative: Streaming mode

- Read out many parallel continuous streams of data, encoded with their original time and location
- Flow of data can be reduced by applying thresholds and zero-suppression locally

Example: Streaming DAQ (3)

Streaming mode advantages:

- Many streams can be aggregated into single link or transport channel
  - Reduction can be substantial because detectors with high number of channels typically have low occupancy
  - Aggregation is time-based and agnostic to data payload: can use generic components
- Streams can be translated with one type in and another type out
  - E.g. clusters from hits on the fly
- Event building now takes place when the data are at rest
- Further event processing, filtering (or High-Level “triggering”) can be done asynchronously
  - Not limited by the front-end electronics, but the depth of the intermediate storage
- Unbiased raw data samples for calibration and monitoring without the need for special runs
- Leverages high-performance computing for a much simpler architecture
  - no complex custom-hardware trigger, no separate trigger path, but a deterministic time-ordered data transport system with tiered storage and parallel processing

Concept can be taken to the next level (HLT) for data scouting and Trigger-Level Analysis

Installation, commissioning, integration, and operations

Bringing the innovative detector-related research, development and design and construction to life

- The transition from final testing to their installation bring new challenges for new components
- Upgrades of today's facilities bears additional challenge of getting new components to work with existing ones

It is immensely important to **invest adequate effort and pay equal attention** to the installation, commissioning, integration and operation

Without these crucial efforts, **data-taking will be compromised and science goals will not be met**.

- Experience has shown that lack of attention can lead to substantial and costly delays, failure to reach design goals and insufficient performance, all at the cost of the physics output.
- High-efficiency operations demand substantial level of effort and has to be maintained as long as the experiment is running.

Building and retaining domain knowledge & technical expertise

> *"[Shifters and experts] wanted for hazardous journey. Small wages, bitter cold, long months of complete darkness, constant danger, safe return doubtful. Honour and recognition in case of success."*
>
> – Alleged advertisement by Sir Ernest Shackleton for the voyage of the *Endurance* to Antarctica

It is critical for our field to **recognize the importance of the highly skilled people** that are carrying out the deployment part of the innovation process

- Without the technical and operational skills contributed by a wide variety of personnel, it is not possible to deliver discovery science at physics facilities
- Availability and recruitment is problematic due to the challenging prospects of **career progression and promotion**
- **Long-term retention** is a serious problem for building a community of experts who are able to operate the experiment effectively
- Scientists who take such opportunities, e.g. to engage in operational efforts, must be considered an **integral part of the science programs and the facilities themselves**

HEPAP subpanel on “International Benchmarking”

U.S. DEPARTMENT OF ENERGY | Office of Science

Charge to HEPAP on “International Benchmarking”

March 7th, 2022

Office of High Energy Physics

Presenter: Glen Crawford

Elements of the Charge. III

How can programs and facilities be structured to attract and retain talented people? What are the barriers to successfully advancing careers of scientific and technical personnel in particle physics and related fields, and how can U.S. funding agencies address those barriers? A complete answer to these questions must address how we can ensure that we are recruiting, training, mentoring, and retaining the best talent from all over the world, including among traditionally underrepresented groups within the United States.

Comments:

- Particle physics has much work to do in this space. We appreciate input from HEPAP on these issues, not only in terms of policy approaches to address barriers, but also how we can better engage with communities and institutions that have been underserved.

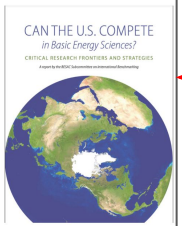
International Benchmarking HEPAP Subpanel

A core tenet of the 2014 P5 Report is that particle physics is fundamentally a global enterprise

Co-chaired by Bonnie Fleming and Patty McBride

Charge Summary

- How can the U.S. particle physics program maintain critical international cooperation in an increasingly competitive environment for both talent and resources?
- Identify key areas where the U.S. currently has, or could aspire to, leadership roles in HEP via its unique or world-leading capabilities, or leading scientific and technical resources.
- How can programs and facilities be structured to attract and retain talented people?



Report due at Fall 2022 HEPAP meeting

11

From [JoAnne Hewett’s talk on Sunday]

Final Remarks

- We appear to be on a technology threshold where innovations and capabilities are beginning to shift long held paradigms of TDAQ systems
- Those innovations will be essential for the success of future physics facilities
- Current experiments may have pushed scales to practical limits
- Whereas previous leaps were to big data and massive CPU farms near the control rooms, we now see an explosion of real-time processing and heterogeneous, accelerated computing
- These innovations can only be brought to life if we acquire new domain knowledge, and if the next-generation facilities attract and retain highly skilled people and support their careers

Current version of the paper: arXiv:2203.07620

Google doc for comments: https://drive.google.com/file/d/1cNLej2n8cnKZJdRvK0Avw4JmISzjQ7sM/view

Indico page to sign for support: https://indico.slac.stanford.edu/event/7171/

Our suggestions for recommendations

- The field should explore and invest in paradigm-shifting approaches that can bring TDAQ systems to new levels by leveraging new technologies and creating partnerships to meet the goals and challenges of next-generation physics facilities.
- Programs and facilities should be structured and funded to identify, acquire, and retain the people and the knowledge required to make the leap to a new generation of technology.

Current version of the paper: arXiv:2203.07620

Google doc for comments: https://drive.google.com/file/d/1cNLej2n8cnKZJdRvK0Avw4JmISzjQ7sM/view

Indico page to sign for support: https://indico.slac.stanford.edu/event/7171/