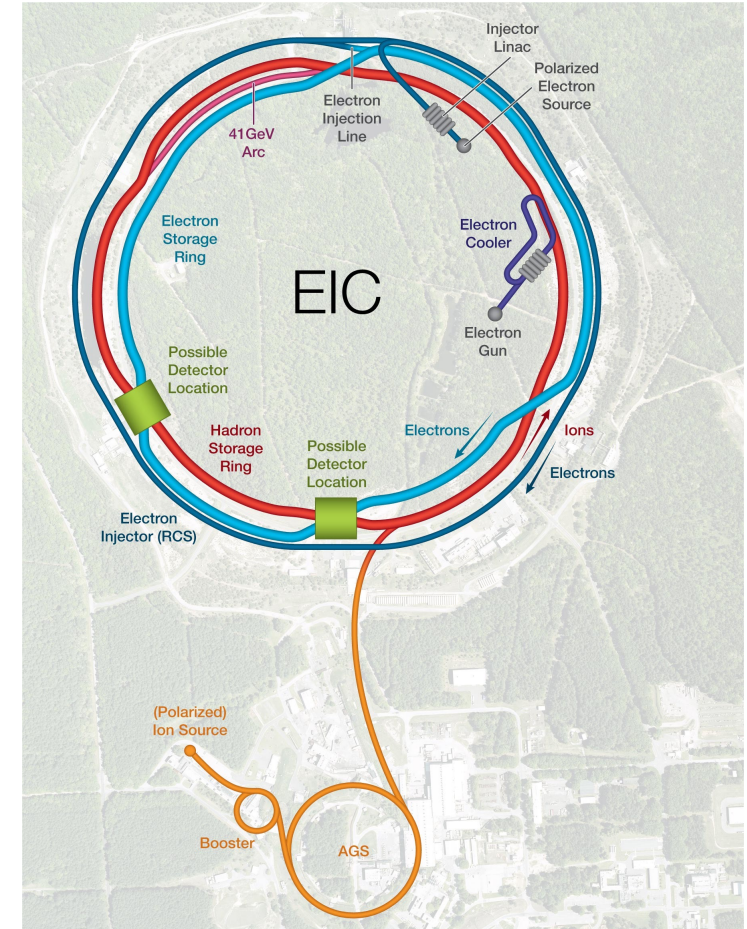


EIC in the Streaming/AI Era

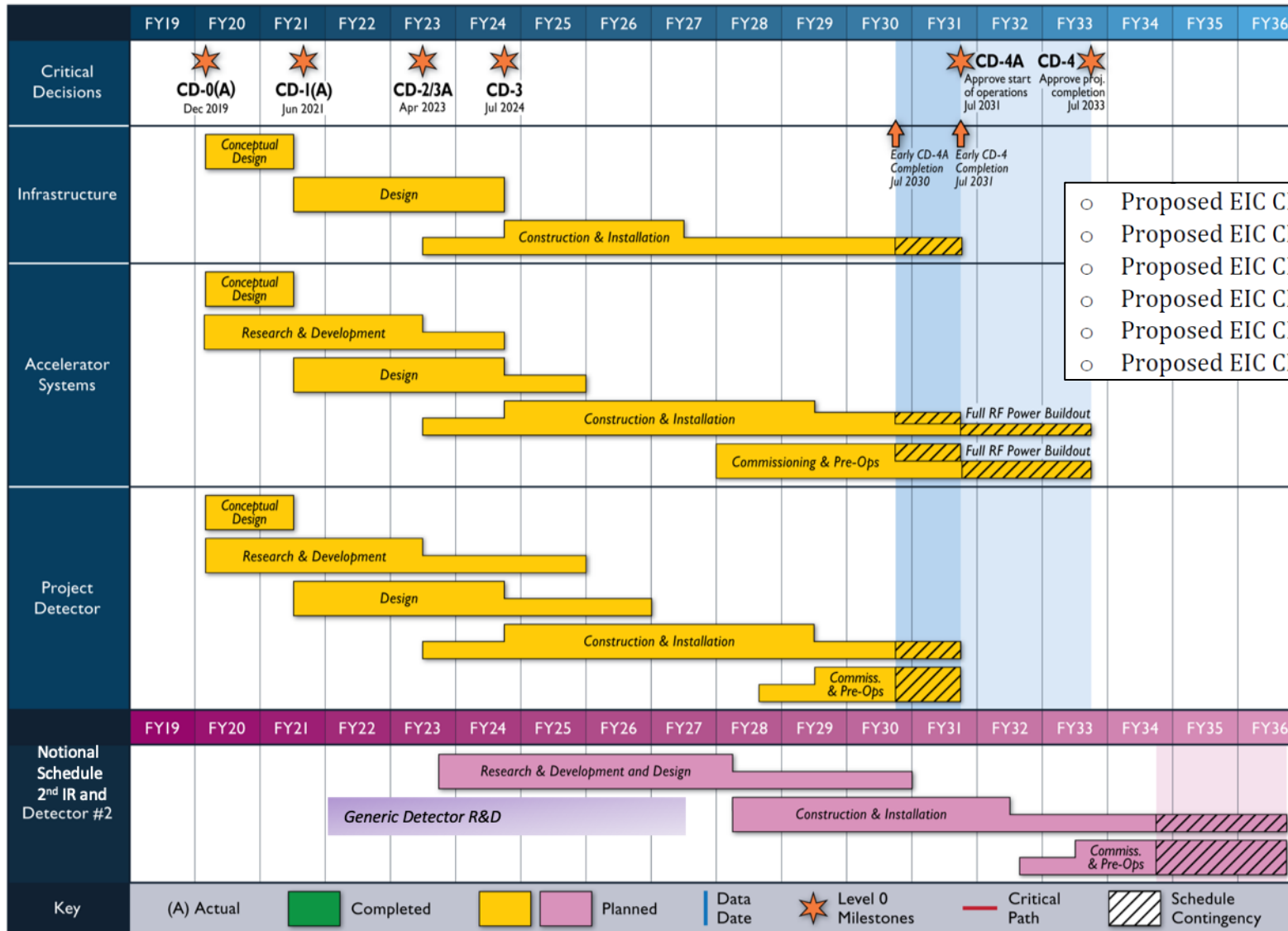
Rolf Ent, EIC Project, Co-Associate Director, Experimental Program



A JLab-centric view of this community effort.



EIC Timeline – Operations Start a Decade from Now



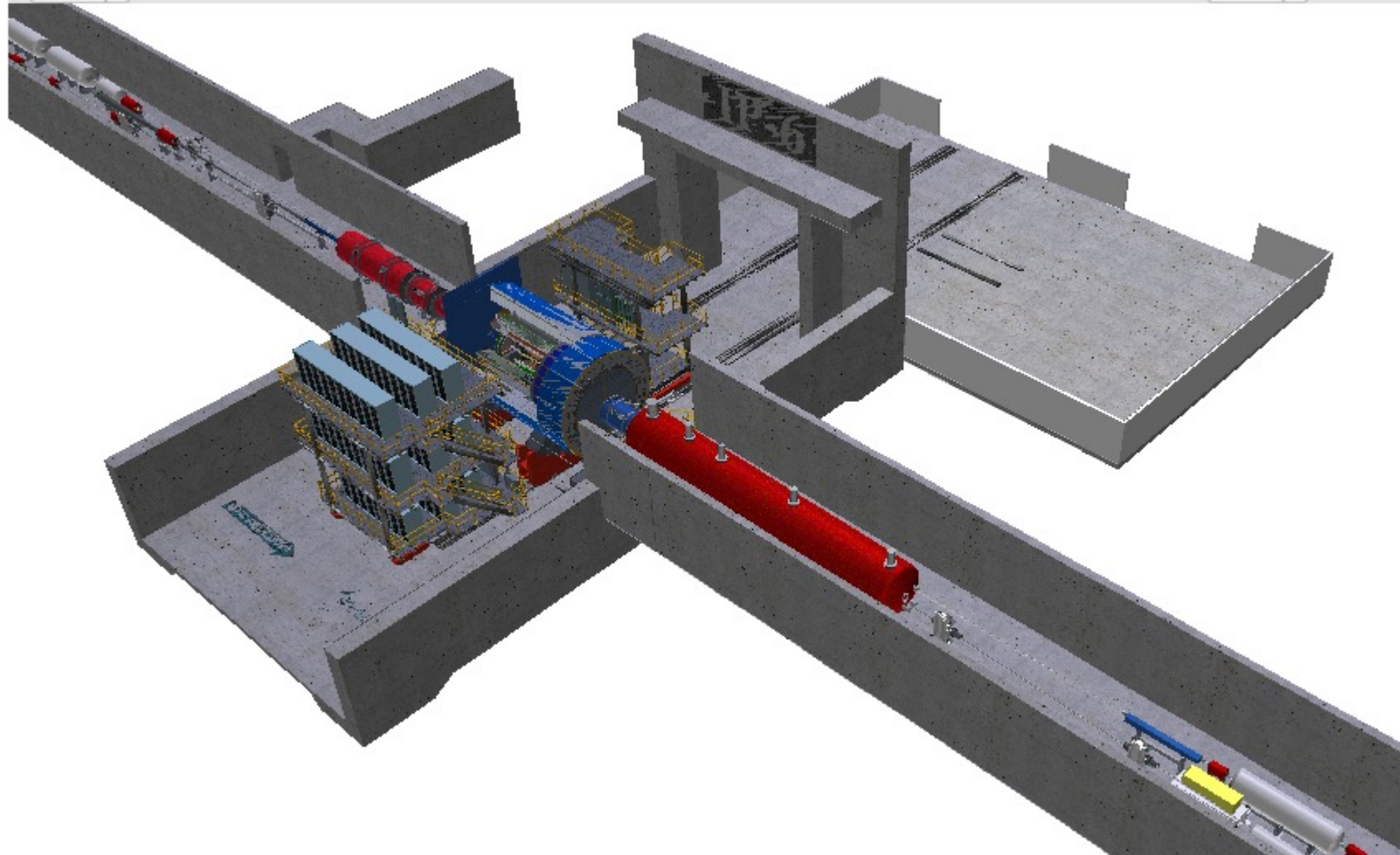
- Proposed EIC CD-2/3A (for EIC planning only) - 2nd Quarter FY2024
- Proposed EIC CD-3 (for EIC planning only) - 3rd Quarter FY2025
- Proposed EIC CD-4A Early Finish (for EIC planning only) - 3rd Quarter FY2031
- Proposed EIC CD-4A (for EIC planning only) - 3rd Quarter FY2032
- Proposed EIC CD-4 Early Finish (for EIC planning only) - 3rd Quarter FY2032
- Proposed EIC CD-4 (for EIC planning only) - 3rd Quarter FY2034

Schedule in flux to take account of FY22 actuals and the FY23 outlook, this will mean an expected 9-month delay of forthcoming CD dates).

EIC: Fully Integrated Detector/Interaction Region++

Far-Backward Region ~ 40 meter

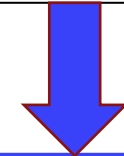
Far-Forward Region ~ 40 meter



EIC science: ALL particles count!

Many particles with $\beta = 1$, but in the far-forward region @ 30 m distance also many particles with $\beta = 0.5$ or so

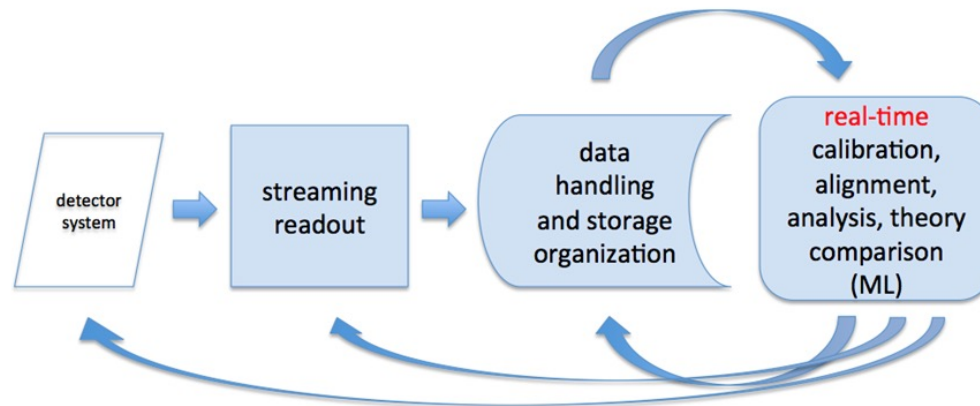
→ $\Delta t = 200$ ns



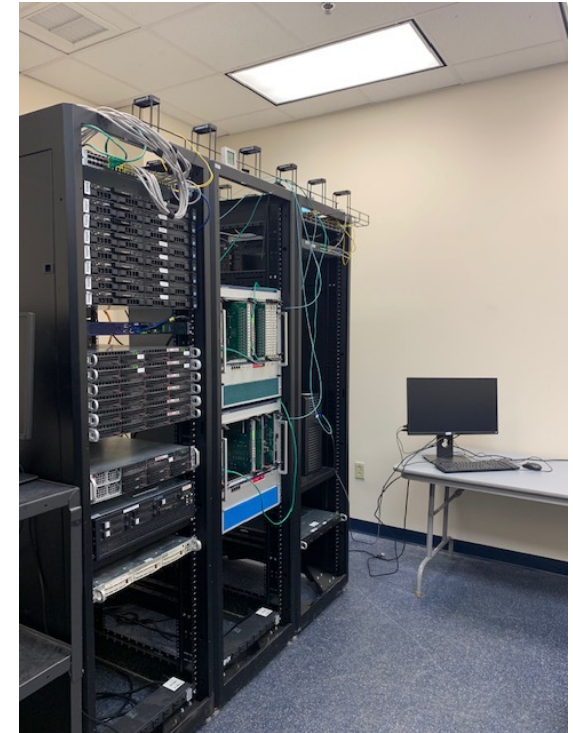
Ultimate EIC Streaming Model: integrate machine, detector, electronics, data acquisition, software and physics analysis.

Streaming Readout at the EIC

- January 2017 (remote, small) Streaming Readout I @ JLab & MIT
- January 2018 (in-person) Streaming Readout II @ MIT
- May 2018 Submitted first eRD23 SRO proposal (PI's Jan Bernauer, Marco Battaglieri)
- November 2018 "Grand Challenge in Readout and Analysis for Femtoscale Science" (Boehnlein, Ent, Yoshida)



MOMENTUM!!!!



Innovation in Nuclear Data Readout and Analysis (INDRA) Facility towards streaming readout, right next to DAQ lab & computing facility

November 2018 INDRA Facility Ready

Lots of progress, by 2020 streaming was the default assumption in the EIC Yellow Report!

May 2022 Streaming Readout X

Grand Challenge in Readout and Analysis for Femtoscale Science

Grand Challenge in Readout and Analysis for Femtoscale Science

Amber Boehnlein, Rolf Ent, Rik Yoshida

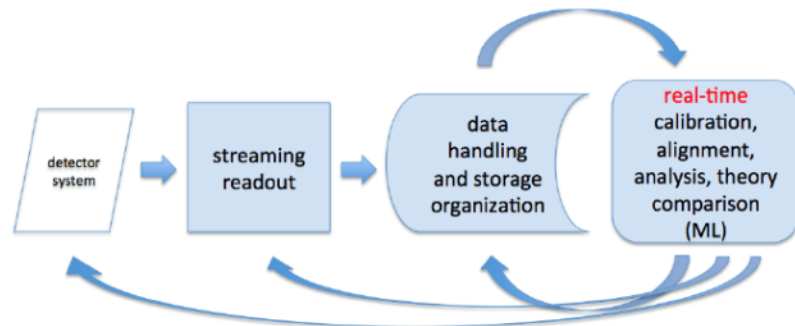
November, 2018

Introduction

Micro-electronics and computing technologies have made order-of-magnitude advances in the last decades. Combined with modern statistical methods, it is now possible to analyze scientific data to rapidly expose correlations of data patterns and compare with advanced theoretical models. While many existing nuclear physics and high-energy physics experiments are taking advantage of these developments by upgrading their existing triggered data acquisition to a streaming readout model, these experiments do not have the luxury of an integrated systems from DAQ through analysis. Hence, we aim to remove the separation of data readout and analysis altogether, taking advantage of modern electronics, computing and analysis techniques in order to build the next generation computing model that will be essential for probing femto scale science.

Integrated Whole-Experiment Model

An integrated whole-experiment approach to detector readout and analysis towards scientific output is summarized in the following figure.



Key Elements

An integrated whole-experiment approach to detector readout and analysis towards scientific output will take advantage of multiple existing and emerging technologies.

Amongst these are:

- “Streaming readout” where detectors are read out continuously.
- Continuous data quality control and calibration via integration of machine learning technologies.
- Task based high performance local computing.
- Distributed bulk data processing at supercomputer centers.
- Modern statistical methods that can detect differences among groups of data or associations among variables even under very small departures from normality.

Existing and Proposed Efforts

Several of the current LDRD proposals as well as separate on-going efforts naturally fit into the framework of the integrated whole-experiment model of data handling and analysis. They are

- Jefferson Lab EIC science related activities
 - Web-based Pion PDF server
- Jefferson Lab and related part of the Streaming Consortium proposal to the EIC Detector R&D committee including
 - Crate-less streaming prototype
 - TDIS streaming readout prototype
 - EM Calorimeter readout prototype
 - Computing workflow - distributed heterogeneous computing
- LDRD proposals
 - JANA development 2019-LDRD-8
 - Machine Learning MC 2019-LDRD-13
 - Streaming Readout 2019-LDRD-10

Grand Challenge

Develop a proof of concept of quasi-instantaneous high-level nuclear physics analysis based on modern statistics from a self-calibrated matrix of detector raw data synchronized to a reference time, without intermediate data storage requirements, with production systems developed for late stage 12 GeV analysis and the Electron Ion Collider. We propose organizing some of the LDRD proposals and other exploratory work around these themes to achieve proof of concept.

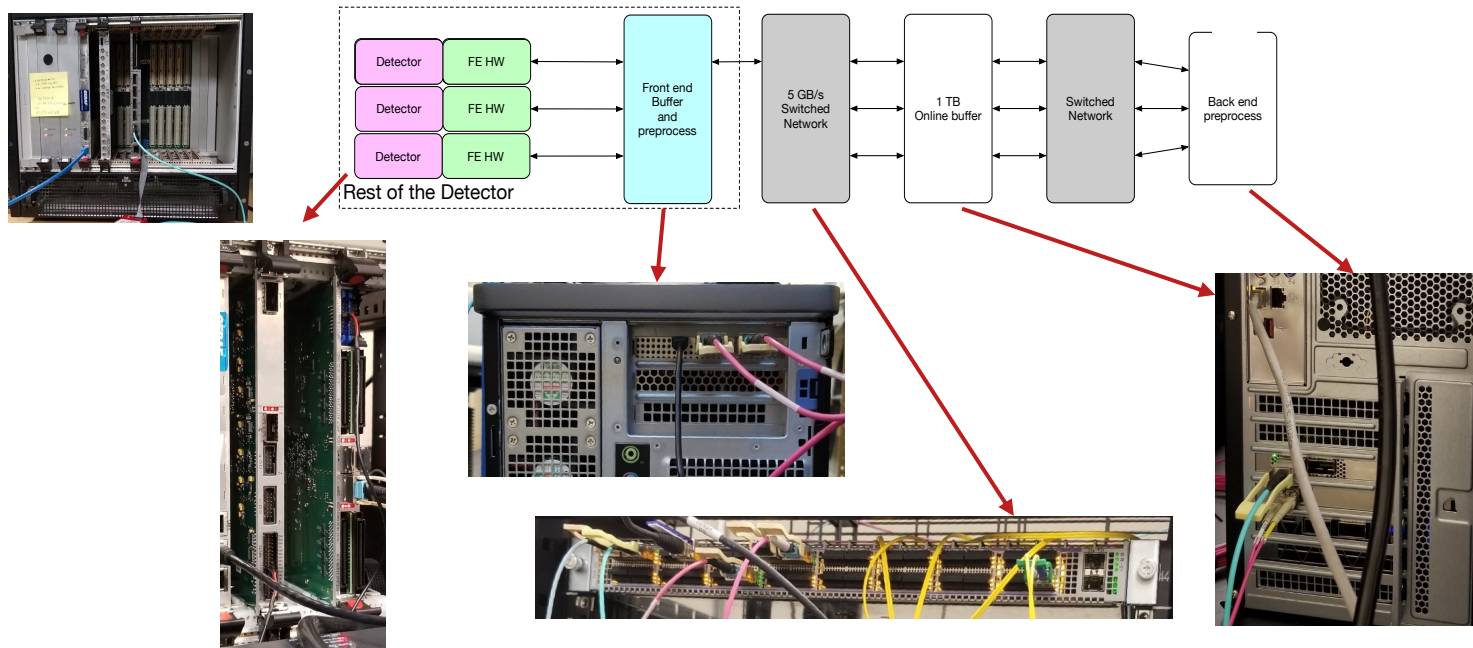
See the buzzwords!!!

- Streaming
- Calibration/ML
- Distributed Computing
- Heterogeneous
- Statistical Methods

What Do We Need for Streaming Readout?

- A streaming DAQ design has several key elements :
 - A data source outputting on fiber.
 - A front end buffer and data processing system with FPGA.
 - A high speed low latency network.
 - An online compute resource to buffer and process data.
 - A timing system.

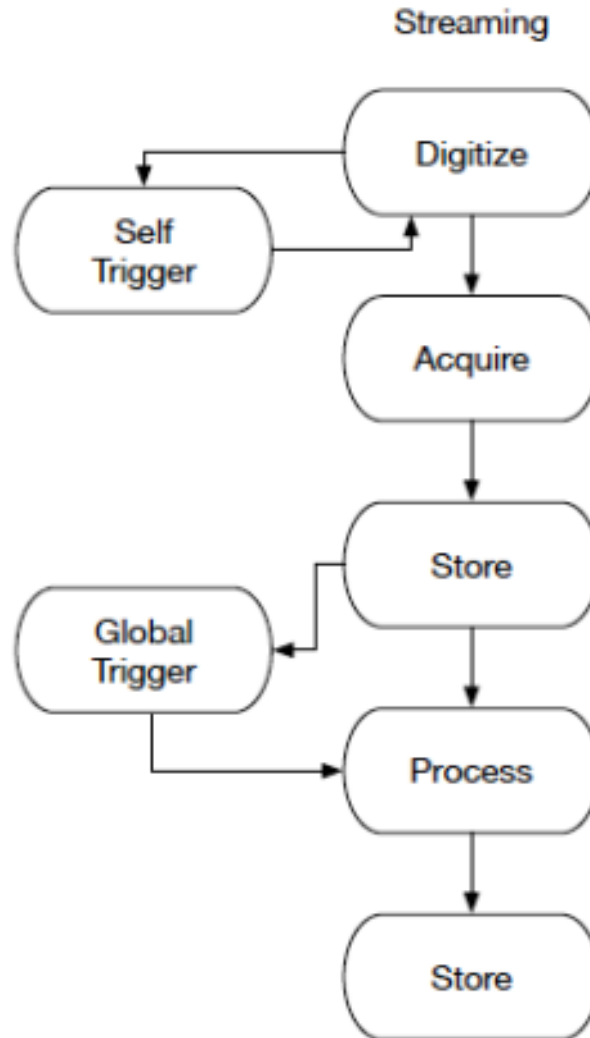
Slide of Graham Heyes to introduce what was going on in the INDRA lab (2019) – all pieces already being worked on...



Benefits of Streaming Readout

Based on slide of Marco Battaglieri

A HIT MANAGER receives hits from FEE, order them and ship to the software defined trigger



A software-defined trigger re-aligns in time the whole detector hits applying a selection algorithm to the time-slice

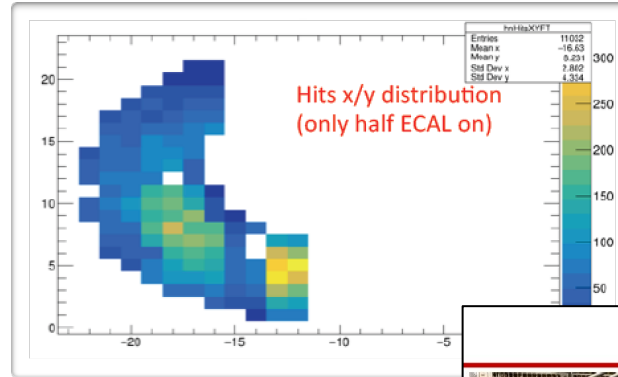
- the concept of 'event' is lost
- time-stamp is provided by a synchronous common clock distributed to each FEE

All channels continuously measured and hits streamed to a HIT manager (minimal local processing) with a time-stamp

- + All channels can be part of “the trigger”, no bias
- + Simplification of readout: No custom trigger hardware and firmware to implement & debug
- + Enables sophisticated tagging/filtering algorithms
- + Allows use of high-level programming languages
- + Ease of scalability
- + Takes advantage of emerging technologies
 - Allows use of available AI/ML tools
 - Allows use of heterogeneous computing
- + Allows rapid turnaround of physics data

Jefferson Lab Streaming Readout Highlights (2021 slide)

- Streaming: Data streams off the detector at a fixed rate.
 - Flexibility of Streaming allows for more online processing
- Activities:
 - Proof of concepts for CLAS12 High Intensity Running
 - Calorimeter and GEMs
 - Streaming concepts in development for SOLID, Tagged DIS
 - Active EIC Streaming Consortium
- INDRA Lab: Streaming Testbed operational
 - Streaming capable hardware, GEM detector, DAQ servers
- ERSAP system (Environment for Real-time Streaming And Processing)
 - Design Requirement Document in review
- *Moving from Proof of Concept towards production design and implementation*



CLAS12 Streaming test with online reconstruction

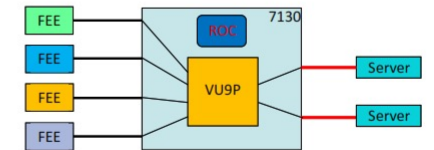
Approach: use commercial hardware where we can (slide from Streaming Readout VIII)

Arista 7130 – FPGA-Based “switch”

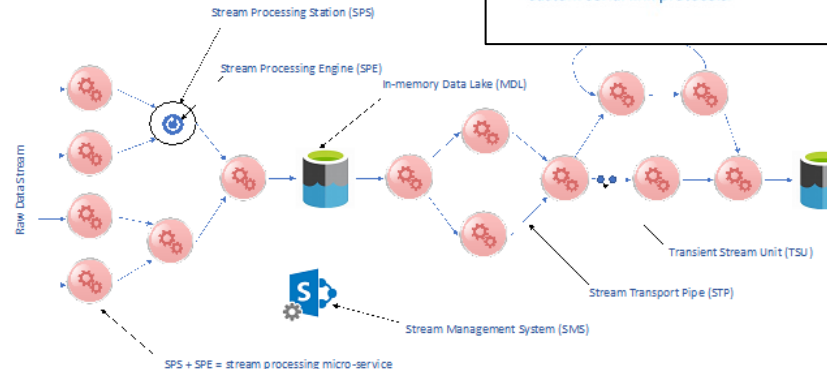


- Virtex Ultrascale+ VU9P-3 FPGA
- 48 SFP+ Ports can be mapped to 60 application ports directly on the FPGA
- 32 GB (4x8GB) DDR4-2200 RAM
- JTAG and Gen 2 PCIe x8 access to FPGA by on board Intel x86 CPU running Linux.
- Available Vendor application support including - port aggregation and high resolution timestamps.
- Development kits for full access to FPGA resources and custom user applications.
- All ports can support standard 10Gb ethernet or custom serial link protocols.

- Next commercial hardware option we will be working with. It is kind of a VTP on steroids.
- Potentially useful for aggregating serial links from different front-end detector electronics and presenting them as standard ethernet streams for back-end processing.



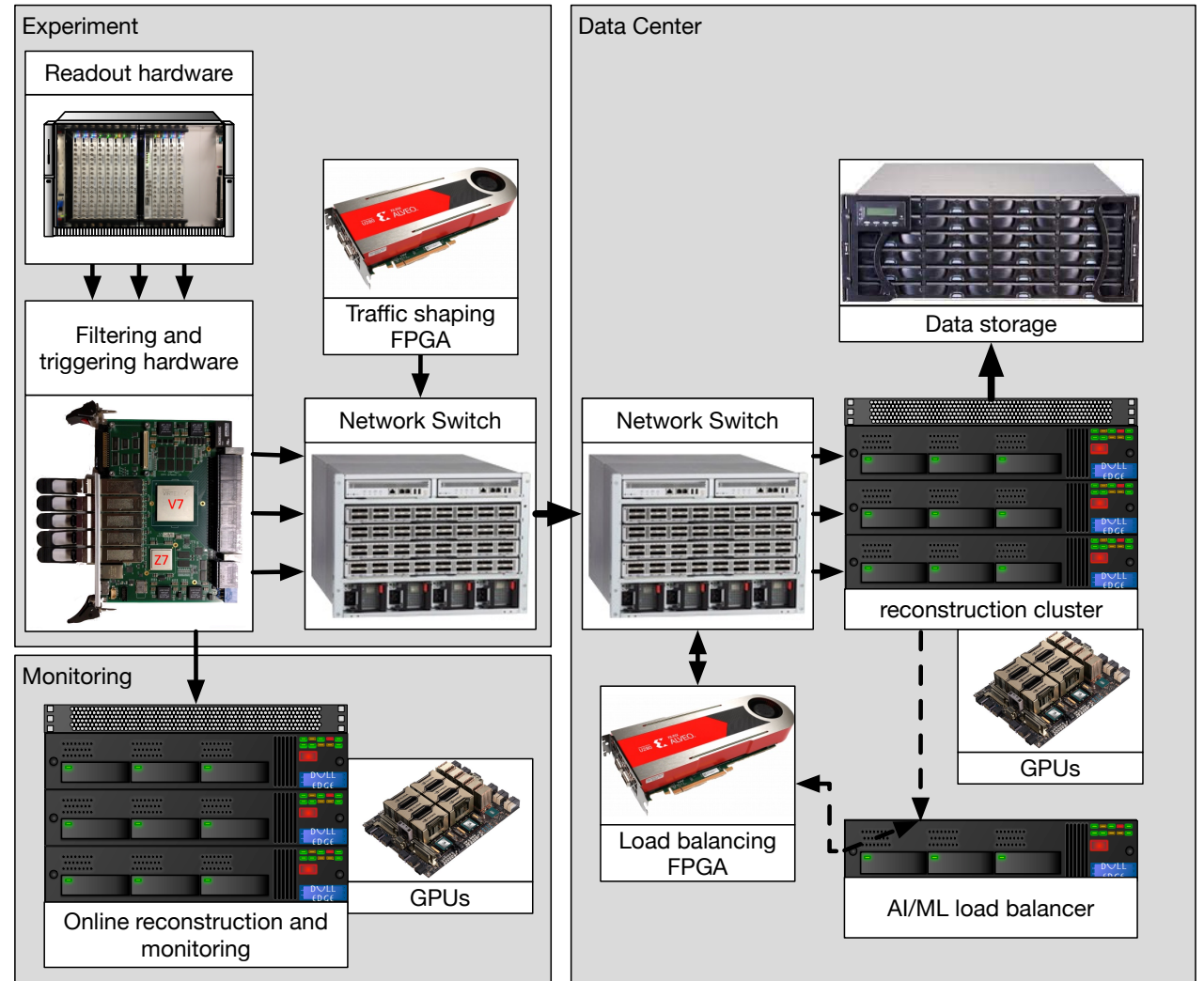
Jefferson Lab



Environment for Real-Time Streaming and Processing (ERSAP) for use by all Halls (and others)

Jefferson Lab Grand Challenge – To Be Continued

- Expanding production deployment of streaming readout, real time calibration, and management of data
 - Keep strong track record of accomplishments
 - Synergistic with ASCR funded activities
- Goal is full SRO, and use of AI/ML for real time calibration and expedient data/theory/simulation comparisons
- Production real time applications for 12 GeV program
 - Opportunistic detector testing in all four Halls:
<https://www.jlab.org/eiccenter/detector-testing>
 - Augmenting the 12 GeV program and moving towards EIC detectors.
- Building digital twin capacity for EIC future
 - Project eAST: Turnkey Geant4 based simulation for EIC detector design with synergies with Medical Physics Initiatives



Streaming Readout – Modular Approach is Key

ERSAP

- Reactive, event-driven datastream processing framework that implements microservices architecture
- Provides basic stream handling services (stream aggregators, stream splitters, etc.)
- Adopts design choices and lessons learned from TRIDAS*, JANA, CODA and CLARA

ERSAP approach is a level of modularity that leads to an agile framework that can evolve rapidly over time

**TRIDAS = TRiggerless Data Acquisition System, Developed for KM_3NET*

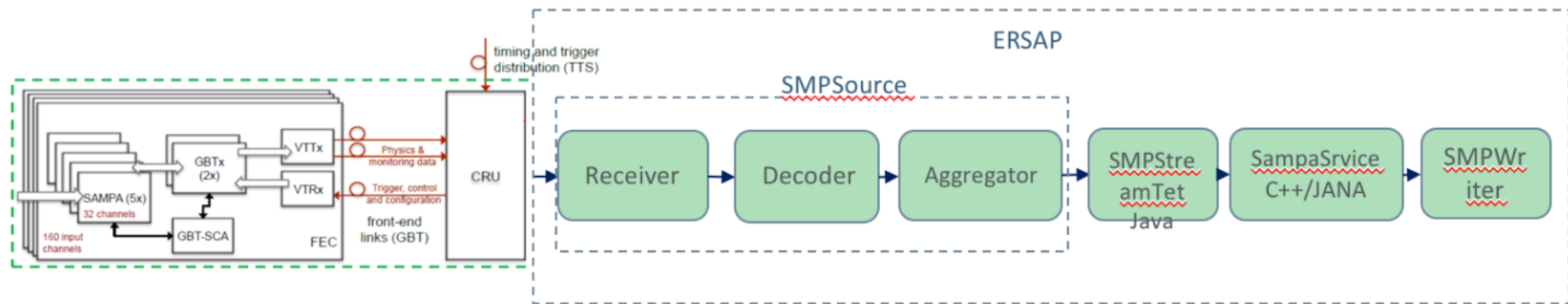
TriDAS ERSAP integration



Hall-D prototype calorimeter SRO



GEM prototype detector SRO (INDRA-ASTRA project)



- 5 SAMPAs based front end card (FEC) was fabricated at the JLAB for GEM (prototype) detector readout.
 - FEC has 5 SAMPAs chips: 160 ADC channels per FEC

- CRU (common readout unit. Provided by the ALICE collaboration)
 - Multiplexes data from front-end links
 - Control and configuration data for FECs
 - Xilinx Virtex-6 FPGA, can be used for data processing and formatting
 - 12 fast serial links



Streaming Readout – Modular Approach is Key

ERSAP also in use for DESY test run

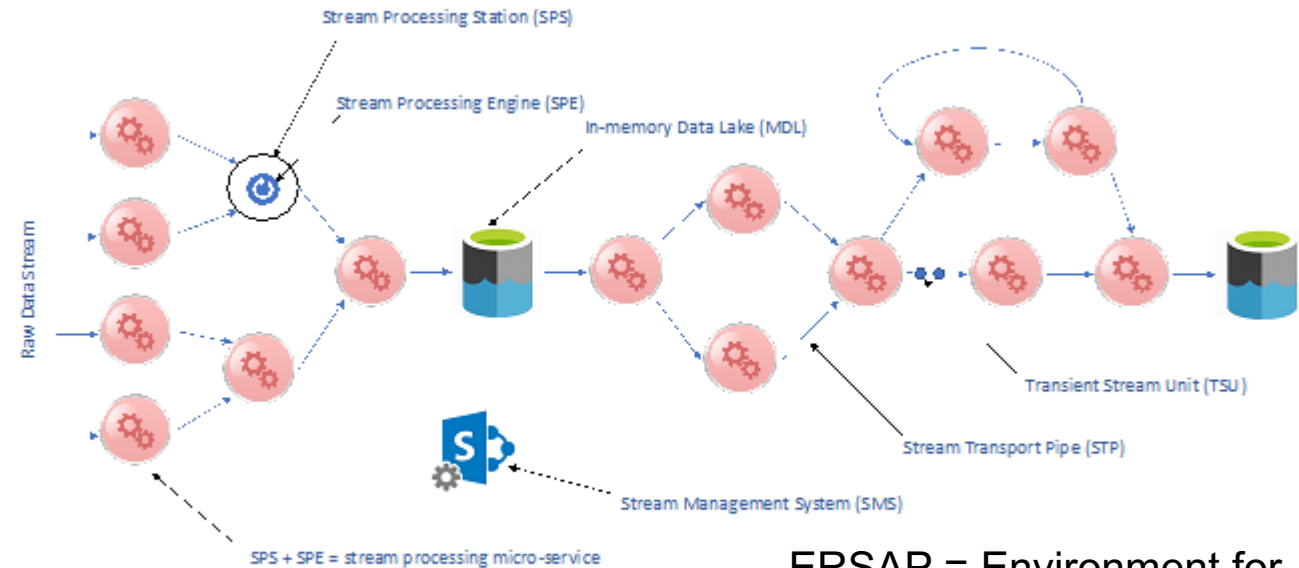
DESY Test Beam

May 2-15, 2022

Douglas Hasell, May 13, 2022

Goals

- Streaming Readout Comparison
 - JLab FADC250 - streaming and trigger mode operated stand-alone
 - CAEN Digitizer (streaming) and QDC (triggered) operated in parallel
- TPEX prototype calorimeter
 - 5x5 lead tungstate calorimeter (PbWO4 courtesy of Tanja Horn, CUA)
- EIC EMCal - Lead tungstate and scintillating glass
 - five 2x2x20 cm³ “crystals” (courtesy of Tanja Horn CUA)



ERSAP = Environment for Real-Time Streaming and Processing

Early Impressions

- JLab FADC250 system worked very well in both triggered and streaming modes but does not play well with others
- CAEN QDC (triggered) and Digitizer (streaming) worked together in parallel very well
 - Digitizer saw more events

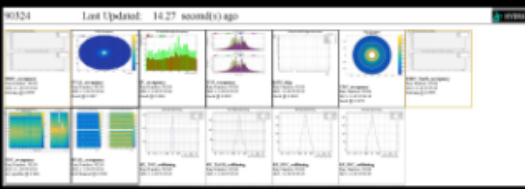
AI/ML for Streaming Readout at the EIC

Automated Data Quality Monitoring

Online Monitoring Tasks: Hydra

T. Britton, D. Lawrence, K. Rajan, arXiv:2105.07549v1 [cs.LG]

- Take off-the-shelf ML technologies and deploy in near real-time monitoring tasks for GlueX in Hall D.
- It was the online monitoring coordinator's job to sift through hundreds of images produced in the previous 24 hours, looking for missed anomalies. This "human-in-the-loop" method was prone to errors.
- Hydra was created to tackle these challenges. Hydra is an AI system that leverages Google's Inception v3 for image classification.



It uses for training the collection of monitoring plots that GlueX had previously recorded. A webpage was created to label the collected images and the entire system is driven by a detector. Hydra is able to spot problems missed by humans and has been shown to perform better than humans at diagnosing problems.

- Large network, ~70% of processing time spent on inference. Techniques are being tested to make Hydra models interpretable (e.g., Layerwise Relevance Propagation). Plans to deploy Hydra in other experimental halls.

See M. Bo and D. Lawrence talk

Event Reconstruction

AI-based Tracking

Keras

G. Giacalone, et al., arXiv preprint arXiv:2008.12885 (2020), G. Giacalone, arXiv preprint arXiv:2008.09144 (2020).

Different Network types were evaluated for accuracy and speed. MLP is chosen to be the best fit, due to implementation simplicity, accuracy and inference speed.

Features	TP	FP	FN	TA	Time (s)	
ERT	0	100%	0.14%	100%	0.30	
MLP	0	99.98%	10.77%	99.98%	0.12	
CNN	38x112	96.11%	28.11%	94.28%	94.28%	1.2
RNN	36	88.40%	11.60%	-	-	-

Autoencoders are typically used for denoising, but can be used for finding glitches.

AI track classification and segment recovery network was implemented as a CLARA service. Tracking code was modified to separate clustering from track finding.

- The implementation of AI assisted tracking into the CLARA12 reconstruction workflow and provided a 8 times code speedup.
- Implemented neural network was able to reliably reconstruct missing segment positions with accuracy of ± 0.35 wires, and lead to recovery of missing tracks with accuracy of $>99.8\%$.

See H. Bellard talk

Using Examples From Cristiano Fanelli's Presentation
Critical Path for Compute-Detector Model for the EIC

Automated Alignment and Calibrations

Autonomous Control and Experimentation

See M. Diefenthaler's talk
INDRAASTRA

Approach:

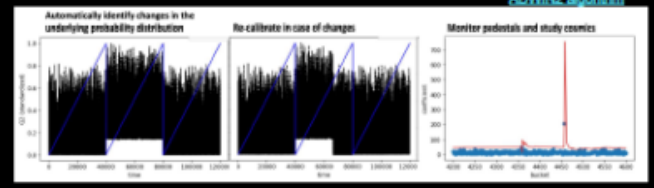
- Identify different data-taking periods Use ML for a) online change detection and b) online data-quality monitoring
- Calibrate different data-taking periods to a baseline

Learning how constant the data is within online adjustable thresholds

Developed Multi Scale Method:

- Represent data in multiscale basis: Increase of base coefficients \rightarrow Change.
- Transform to coefficient space: Outliers in the distribution \rightarrow Change.
- Detect Changes \rightarrow Detect outliers using IQR

ADWIN2 algorithm

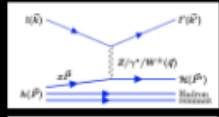


Automatically identify changes in the underlying probability distribution. No calibrate in case of changes. Monitor pedestals and study cosmic.

Reconstruction of DIS Events

Deeply Learning Deep Inelastic Scattering

M. Diefenthaler, et al., "Deeply Learning Deep Inelastic Scattering Kinematics," arXiv:2105.11638(2021).

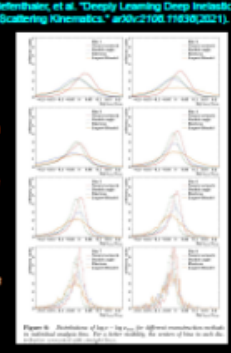


Use of DNN to reconstruct the kinematic observables Q^2 and x in the study of neutral current DIS events at the ZEUS experiment at HERA.

The performance of DNN-based reconstruction of DIS kinematics is compared to the performance of the electron method, the Jacquet-Blondelet method, and the double-angle methods using data-sets independent from those used for the training.

Compared to the classical reconstruction methods, the DNN-based approach enables significant improvements in the resolution of Q^2 and x .

DIS measurements at upcoming EIC



Slide courtesy Markus Diefenthaler (JLab)

Musings from a Recent Review

At a recent meeting I was amused to hear nearly all experiments, including the large LHC ones, proudly mention they were including GPU servers...

Some Highlights of Heterogeneous Computing

First use of CPUs and GPUs together, inspired by Lattice QCD needs



Number 310 | April 26, 2010

Hot Graphics Cards Fuel Supercomputing



GPU cluster

The hottest video games on the market often have the most realistic graphics. And the key to such remarkable video is a device called a graphics processing unit, or GPU. Now, scientists at DOE's [Jefferson Lab](#) are using the power of GPUs to study some of the most fundamental problems in the universe.

"The reason graphics processors are so powerful is so that they make your game look realistic. They need to be able to compute and draw lots of things—at least thirty times a second," said Chip Watson, manager of Jefferson Lab's High-Performance Computing group in the IT Division.

This fast computation can also be applied to "drawing things" that are too small to see directly, such as the sub-atomic particles studied by nuclear physicists at Jefferson Lab. Interestingly, GPUs owe their ability to render realistic graphics to physics.

Some Highlights of Heterogeneous Computing – cont.



Jefferson Lab's K20 supercomputer was listed on the TOP500 list of fastest computers in June 2013. This single rack system houses two supercomputers: one built of NVIDIA K20 GPUs and one built with Intel Xeon Phis.

<https://science.osti.gov/np/Highlights/2013/NP-2013-06-a>

Supercomputing on a Budget

The optimization of commercial hardware and specialized software enables cost-effective supercomputing.

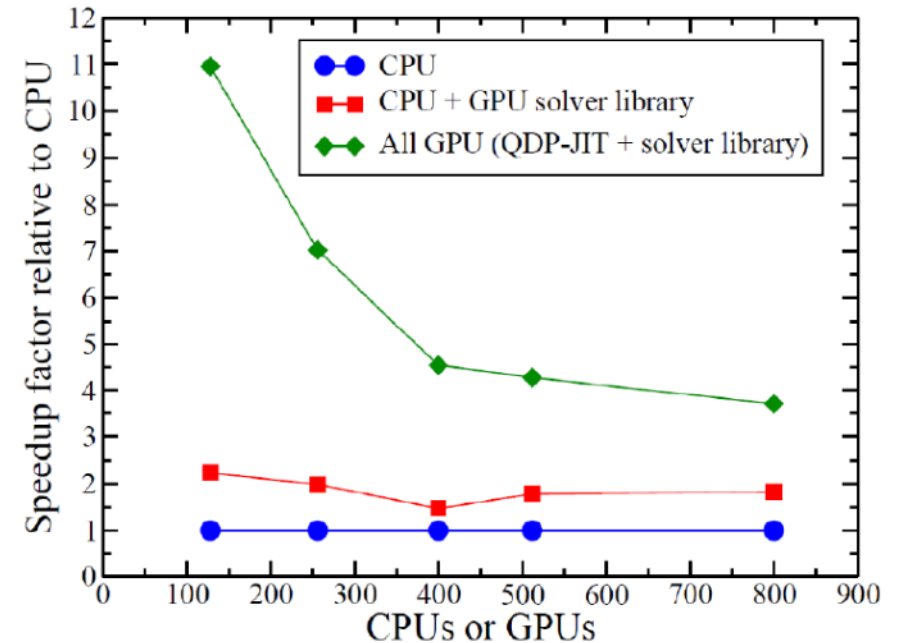


Figure 7.2: The strong reduction of time-to-solution as a result of software development is shown as the speedup gained as compared to the computing time with CPUs or GPUs only as the number of units increases. Shown are speedup factors only utilizing a solver library, and also including the QDP-JIT compilation framework on GPUs. [Image credit: B. Joo].

2015 NSAC Long-Range Plan,
Figure courtesy Balint Joo (ORNL)

Heterogeneous Architectures – Advantage for Streaming/AI

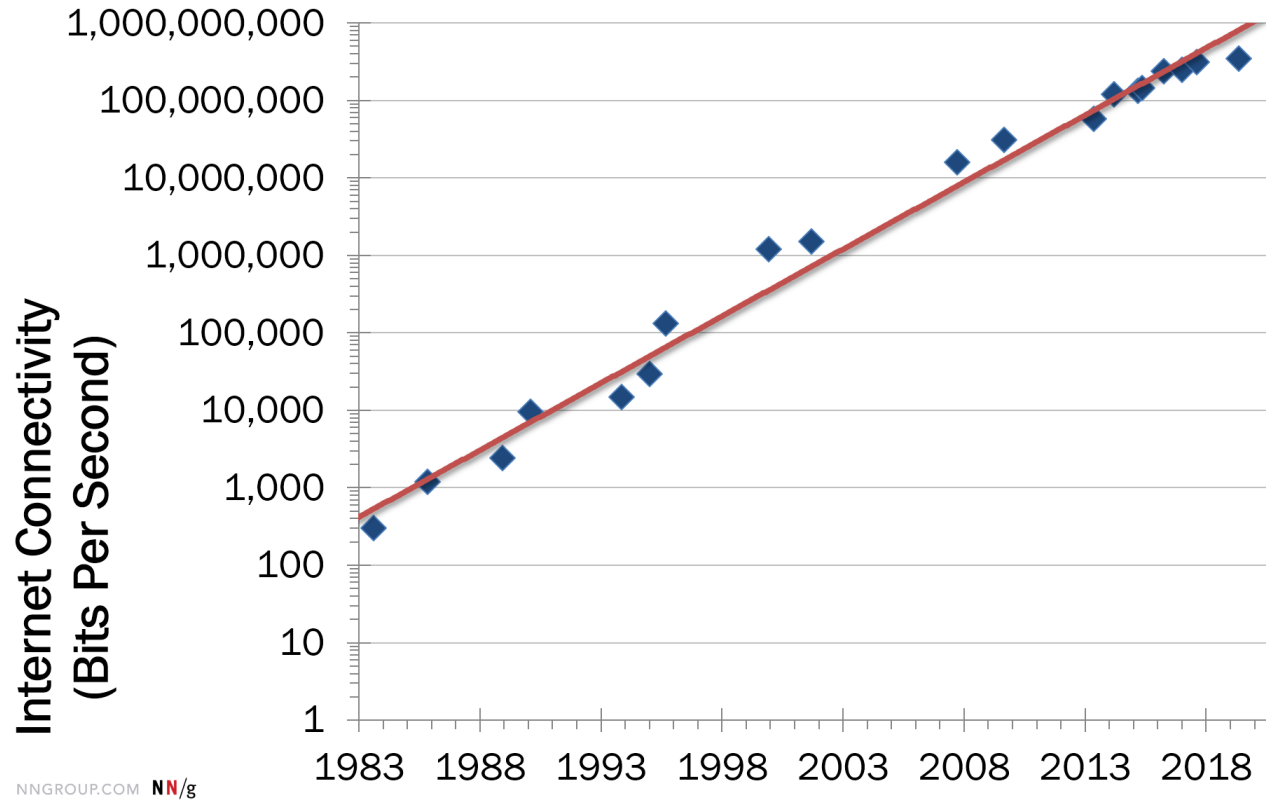
Based on slide of Nathan Brei (JLab)

- Heterogeneous hardware:
 - Commercial off-the-shelf
 - Can be used for general computing workloads but not a full CPU
 - Excluding: fast electronics, SAMPA, FELIX
 - In practice: GPUs, FPGAs, coprocessors (e.g. Xeon Phi), (TPUs)
- Themes:
 - AI/ML
 - Streaming readout
 - High-bandwidth real-time data processing

These technologies reinforce and enable each other.

Imagine a full end-to-end real-time, streaming system. What would it look like?

The Network Bandwidth will grow!



Eli Dart (leader of Science Engagement group of ESnet) e-mail:

"As far as network backbone capacity in the 2030s goes, it's too early to say. ESnet's optical system is highly capable (we just deployed it as Jason said), but we don't yet know what networking technologies will be available in 2032 - ten years is a long time. That said, we can make some guesses. It is likely that 1.6Tbps Ethernet will be available by then."

Jason Zurawski (Esnet, in group of above) e-mail:

"I think BNL plans to support multipole 400G connections, JLab is limited due to a lack of fiber recourses in the area, but will at a minimum upgrade some of the 10G connections that exist today to 100G soon (and move to 400G sometime in the future I would imagine)."

		Annualized Growth Rate	Compound Growth Over 10 Years
Nielsen's law	Internet bandwidth	50%	57×
Moore's law	Computer power	60%	100×

EIC Computing Butterfly Model

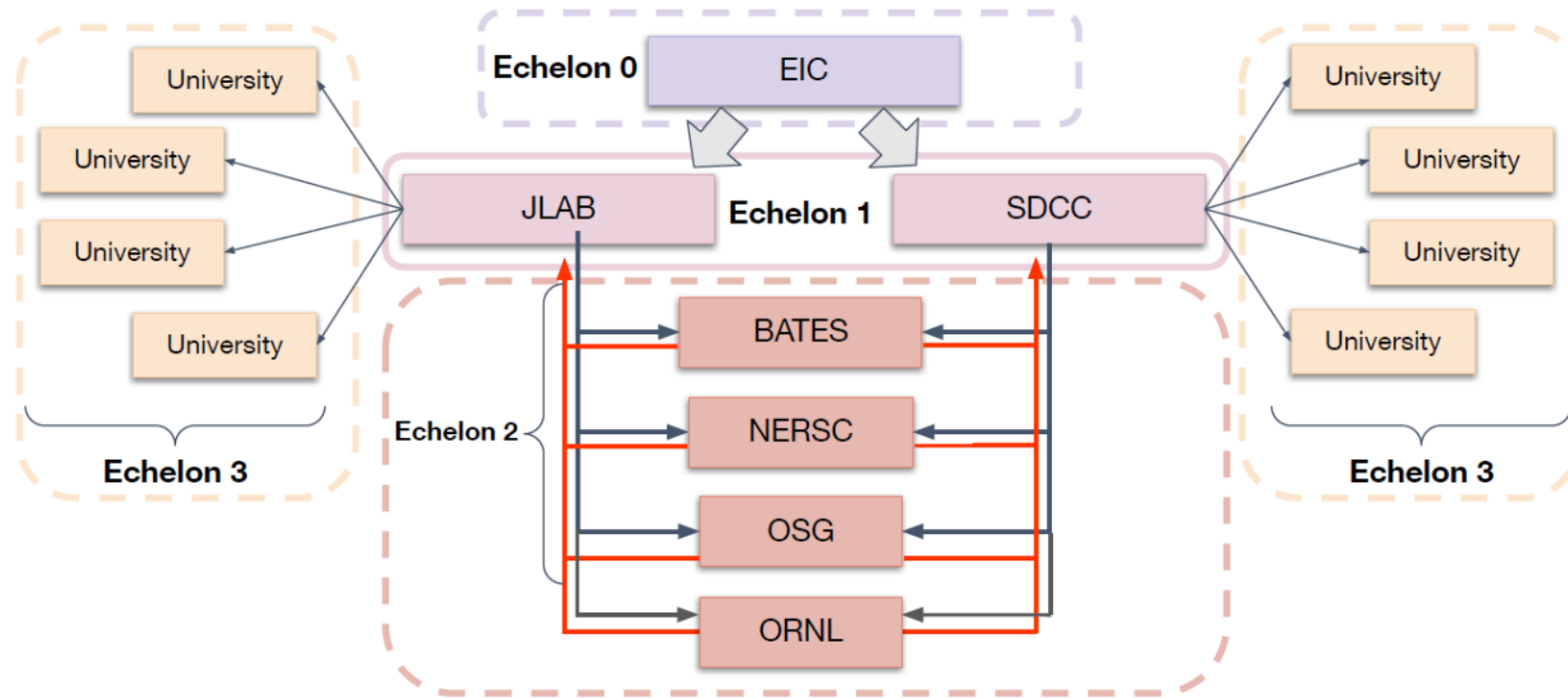


Figure 4: Butterfly model of federated offsite computing. In this model, nearly all storage is contained in echelon 1 while large portions of the raw data processing is delegated to multiple HTC/HPC facilities. The named facilities in this graphic are merely examples and do not represent commitments or final plans.

**Proof of Concept of
EIC Dual Echelon 1
data center approach
during EIC detector
proposal phase in
2021**

Benefits of Heterogeneous Hardware to Nuclear Physics

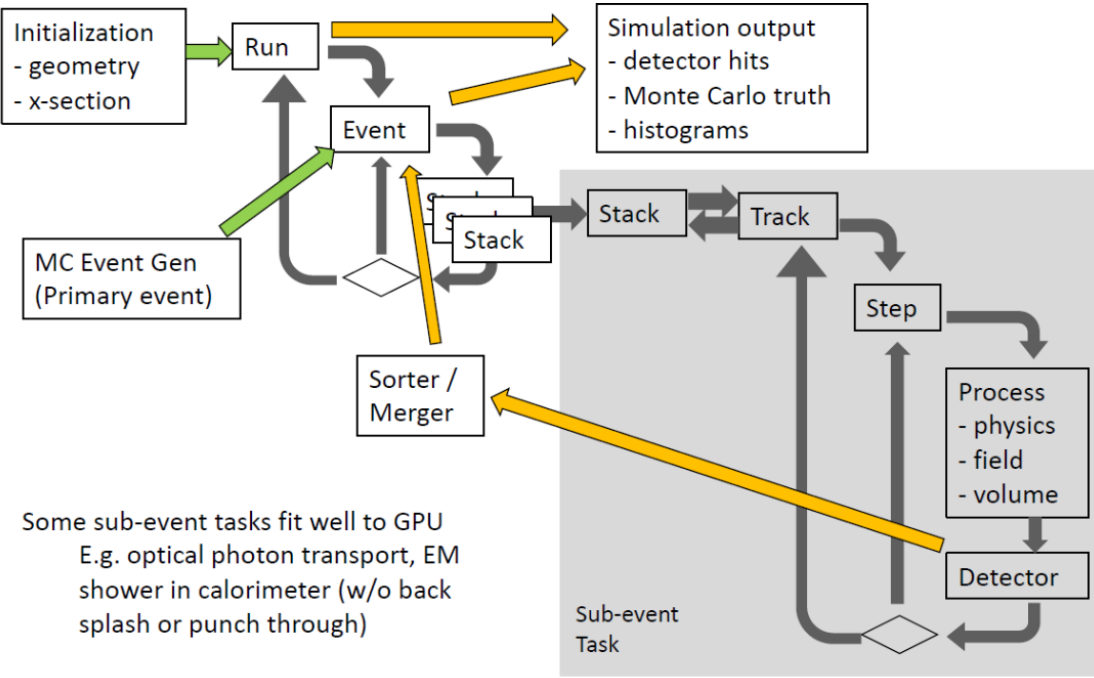
- GFlop/Watt is significantly lower for GPUs than CPUs
 - e.g. <https://www.karlsruhp.net/2013/06/cpu-gpu-and-mic-hardware-characteristics-over-time/>
- Price per flop is lower for GPUs than CPUs
 - *n.b. can be hard though to keep GPU fully busy*
- Large HPC/HTC systems have significant compute capability tied up in heterogeneous hardware (*including cloud services*)
- Higher memory bandwidth (*good for streaming*)
 - See LHCb Allen project: <https://arxiv.org/abs/1912.09161>
- Well-suited for AI/ML models
 - *not all models are efficient on GPUs, but some (e.g. CNNs) are extremely efficient*
 - *tools like HLS4ML making FPGAs more accessible*
(<https://fastmachinelearning.org/hls4ml/>)
- Faster simulation via GANs

Slide courtesy David Lawrence (JLab)

Benefits of Heterogeneous Hardware to GEANT4 Simulations



Turnkey Geant4 based simulation for EIC detector design (and more: detector R&D, Medical Physics applications)



<p>Detector Simulation</p>	<ul style="list-style-type: none"> • Turn-key application • Built on top of Geant4 for full and fast simulations • With library of potential detector options
<p>Requirements</p>	<ul style="list-style-type: none"> • Ease of leveraging new and rapidly evolving technologies: <ul style="list-style-type: none"> • AI/ML to accelerate simulations • Heterogeneous architectures: <ul style="list-style-type: none"> • AI/ML is the best near term prospect for using LCF/Exascale effectively. • Ability to reuse existing simulation work • Ease of switching detector options • Ease of switching between detailed and coarse detector descriptions
<p>Project</p>	<ul style="list-style-type: none"> • Support for high concurrency heterogeneous architectures and fast simulations integrated with full detector simulations allows to leverage AI/ML in Geant4. • Makoto Asai, who led Geant4's multi-threaded reengineering to support high concurrency heterogeneous architectures, is now at Jefferson Lab and leading the next phase in concurrent Geant4, sub-event parallelism. • We are building up team at Jefferson Lab on next-generation detector simulations with strong support from wider EIC community, in particular from BNL.

Slide courtesy Nathan Brei and Markus Diefenthaler (JLab)

Summary – EIC in the Streaming/AI Era

- Community efforts towards streaming at the EIC started in earnest in ~2018
 - Now, a mere four years later, streaming readout is the default for the envisioned EIC detector
 - The advances in microelectronics and commercial data handling hardware are our friends 😊
 - Many efforts are ongoing withing the EIC community
 - And we are only busy ~4 years with a decade to go before EIC detector operations start
- Community efforts towards AI at the EIC started in earnest in ~2020
 - The AI4NP workshop and white paper, AI4NP winter school (369 registered participants), AI4EIC workshop series (1st workshop with 243 registered participants) all were a huge success
 - AI is our friend and a perfect fit for the nuclear science we do
 - AI is being integrated in all aspects of the EIC detector (design, calibration, simulation, reconstruction, analysis)
 - Here also amazing momentum has been gathered
- To take full benefit of streaming and AI implies use of heterogeneous computing
 - AI requires to integrate the power of GPUs in our workflow
 - Our colleagues in Lattice QCD have illustrated the power of combining CPUs, GPUs and modern software
 - The increase in network bandwidth is our friend 😊 - just imagine 1.6 Tbps by then...
 - Similar, the developments in statistical methods and data science are our friends 😊

Summary – EIC in the Streaming/AI Era

Combining Streaming, AI, heterogeneous computing and modern software in our physics detector, data handling and analysis (from calibration to high-level physics analysis) is a **no-brainer**.

We “just” have to make it work.

*Heterogenous architectures, AI/ML, and the other technologies are rapidly evolving...
On the timescale of EIC data taking the landscape is likely to be completely different.
We must ensure an agile framework that can evolve rapidly over time!*

Backup

A Possible Fork in the Road

With GPUs computer and data sciences are taking two complementary approaches:

- Data science, via AI/ML, uses training data to create a model of a complex algorithm so that results are inferred from the input data without executing the algorithm. This allows use of GPUs without porting the code of the algorithm itself.
- Computer science is taking the approach of extending computing languages so that the same code can run on CPUs or GPUs without modification. For example, Intel's OneAPI:

<https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html>