Edge ML & Heterogeneous Computing

Heterogeneous Edge for Ultra Low Latency

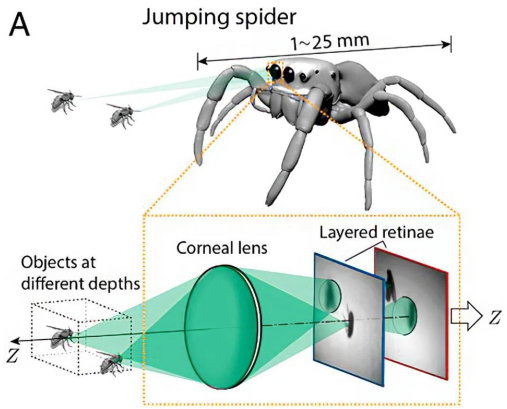
Ryan N Coffee / Sr. Research Scientist / LCLS-PULSE-TID

December 19, 2024

SLAC NATIONAL ACCELERATOR LABORATORY

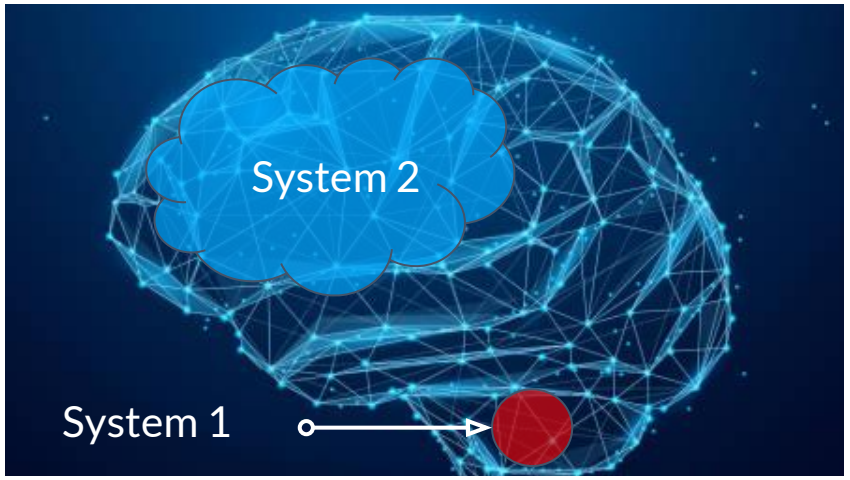
Stanford University | U.S. DEPARTMENT OF ENERGY

The Parsimonious Jumping Spider (100k neurons)



Eons of co-design

- Hardware and wetware work in unison
- Retinal cells ARE neurons, so are base of each hair on her body, they are acoustic sensors
- Not just computationally efficient… **energy efficient by minimizing bit flow**
- Only outliers are promoted (in humans) to prefrontal cortex (and late)
    - Why **waste** so much computation only for **rationalization**



GenAI aims (and misses) reasoning

- Aims to learn interpolative “logic”
- But our critical use cases need a formula one pit crew
    - **Performance (System 1) vs. Rationalization (System 2)**

Common function, different-domain

X-ray laser spectrometer

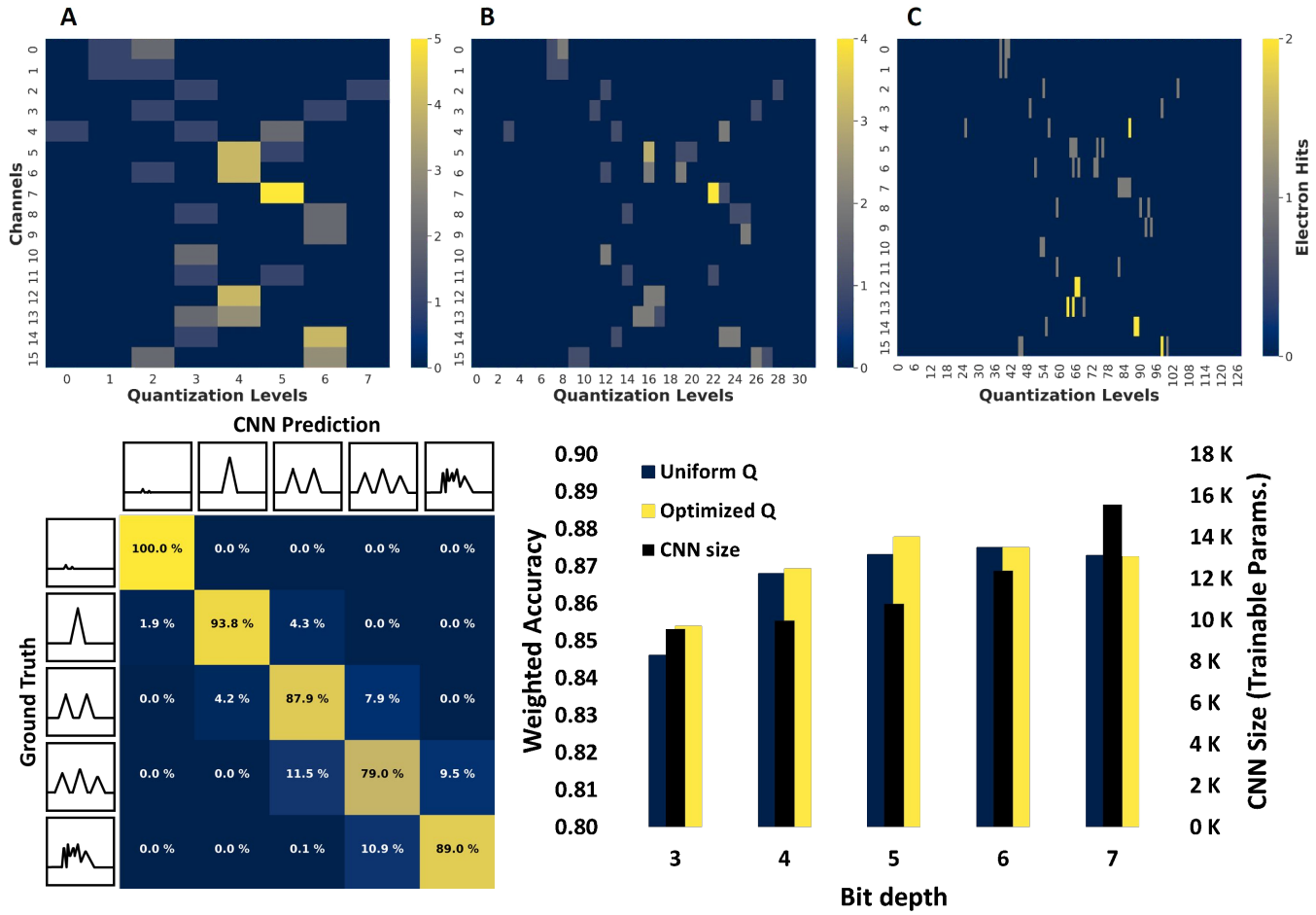
- Different domain, but similar signal interpretation
- Phase, amplitude, and number of “tracks” per microsecond
- Waveform to information FPGA prior to any system memory or NIC

Channel Information is Quasi-Static

- Prior distribution informs quantizaion
- FPGA, ASIC, or Analog implementation
- Stochasticity of output spectrum is a metric of “concept drift”

Maximize information/bit

- Far fewer, information dense, features
- Dense LinAlg Ops for encoding/tokenizing

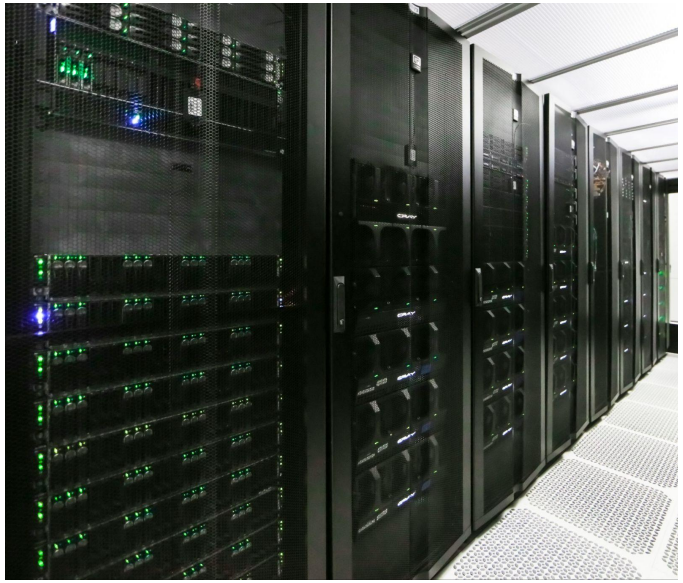


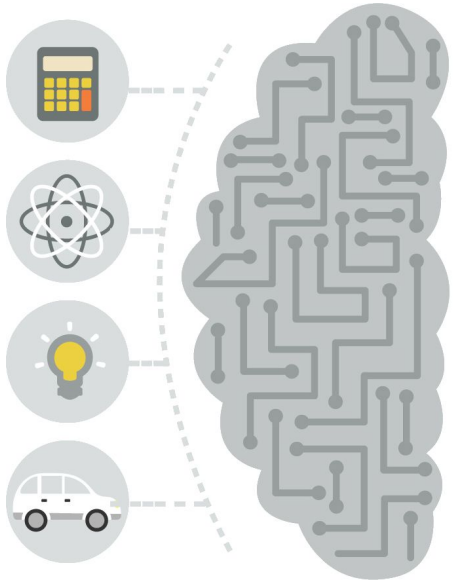
Gouin-Ferland, Coffee and Therrien, Front. Phys. 10 (2022)

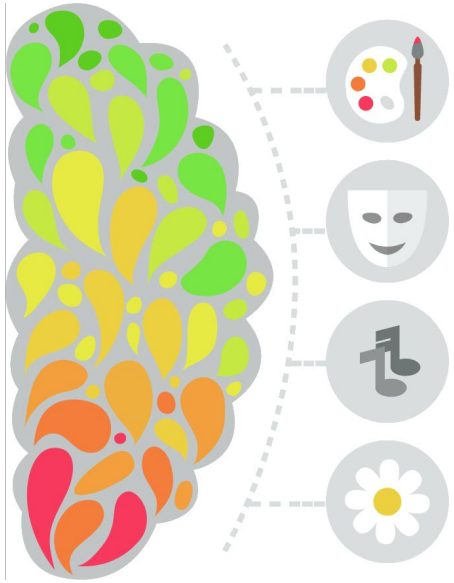
Edge-to-Exascale and back!

HPC testbeds linked to Edge Streaming Sensors and Early Access Hardware

- Testbeds that design for **Edge Integration with LCF**
- Real-world streaming tests to work out bugs and security
- Prototype domestic inter-lab federation, then international
- IRI Orchestration should align with future HEP international ecosystem
- Reconfigurable hardware and racks for **design exploration**
- Streaming imaging **(photonics)** and digitizers **(analog)**
- Early access for **inference hardware** and **custom ASICs** and HEP sensor prototypes
- Long DOE history in FPGA and leading **eFPGA** into age of chiplets for trigger, stream, and control systems



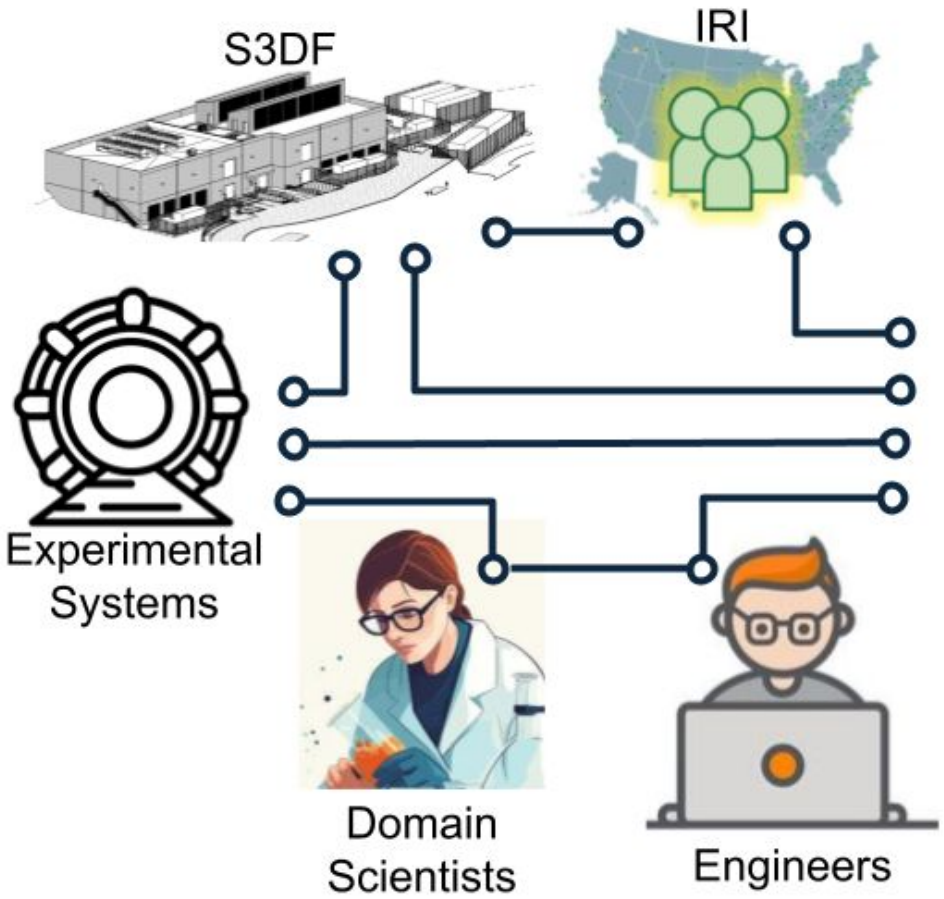


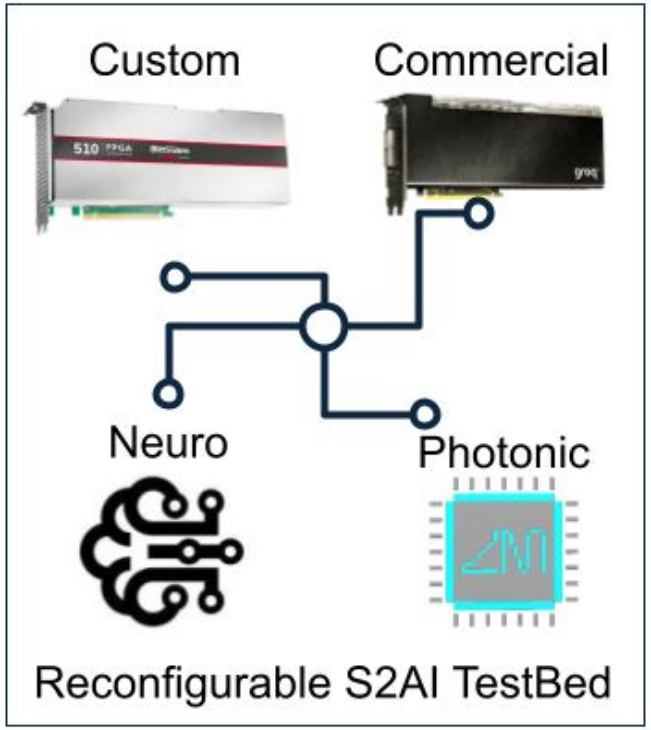


Edge-to-Exascale and back!

Domain Scientists and HPC and ASIC Engineers and Researchers

- Tiered Facilities
  - Experimental sensors
  - Mid-scale HPC – also archival storage
  - LCF
- Community Collaboration
  - Workforce Development
  - Open the hood on weird hardware
  - HEP science drives global technology mission





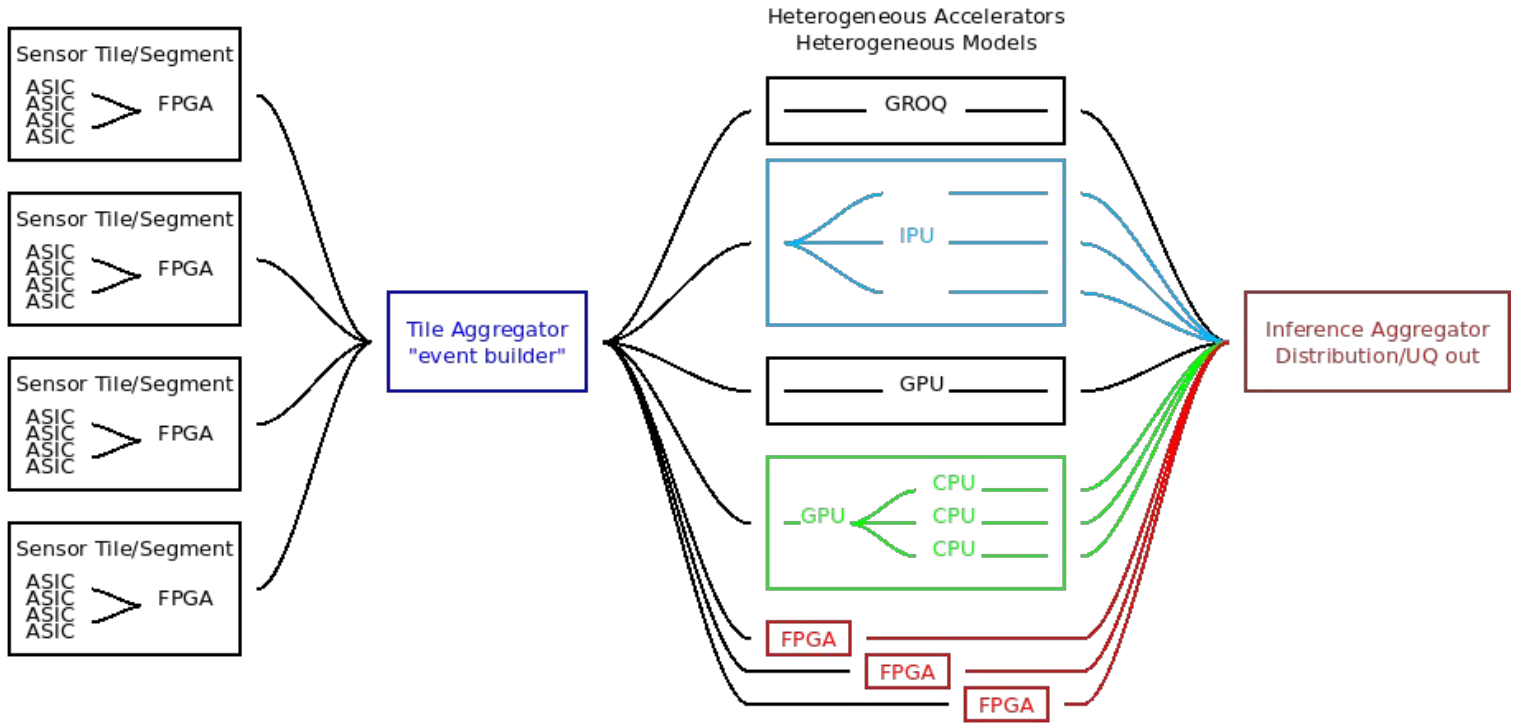
Heterogeneous Flow

Orthogonal models are like orthogonal minds

- Each architecture supports a different algorithm module or **Neural Layer**
- **Composability** of modules/layers allows flexibility
- **Orchestration** based on hardware simulators and then on real-time module metrics
- ASICs + eFPGAs at the sensor edge

  … or analog, photonic,

  … neuromorphic?



Model (Identifier)		# Parameters	Parameter Memory (MB)	Single Batch Runtime (μs)
Denoiser (1)	Zero Classifier (1a)	70,345	0.28	28.2
	Autoencoder (1b)	46,529	0.19	96.3
Classifier (2)		1,458,597	5.83	61.4
Single Pulse ϕ Regression (3)		12,196,240	48.78	52.2
Double Pulse $\Delta\phi$ Regression (4)		23,330,400	93.32	72.0
Totals		37,102,111	148.40	168.3

J Hirschman *in prep.*

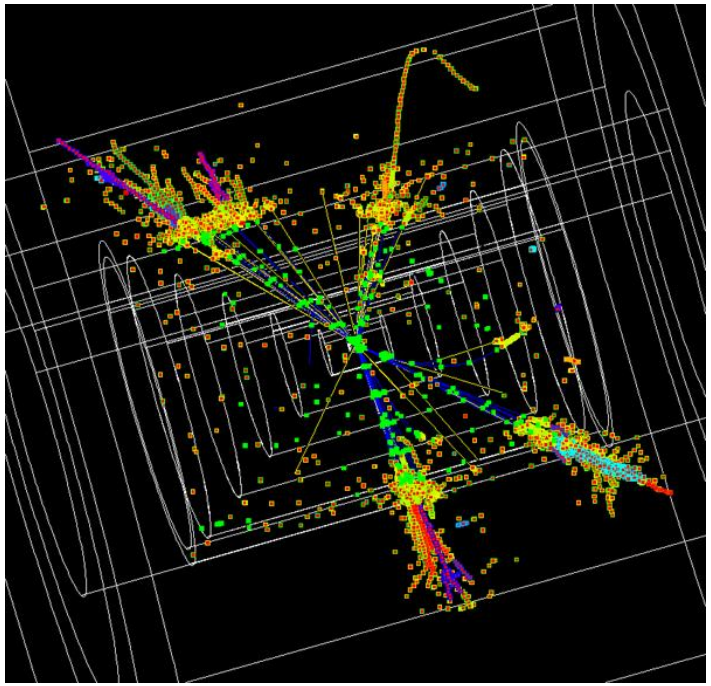
Toward Higgs Factory Heterogeneous Autonomous DAQ

Motivation

- Incorporating ML-based intelligence across the data pipeline is an R&D priority across exercises (DOE BRN, ECFA Detector R&D)
  - **Extreme data compression**, storage, efficiency, and performance, reducing costs and **increasing performance**
- Teams/institutions: SLAC, BNL, MPI, Uni Geneva, & more

Path Forward & Areas Of Focus

- Incorporation of front-end intelligence; see eFPGA talk from Kenny Jia (AIM)
- Triggerless readout: handling off-detector bandwidth, structure of off-detector compute stages (TDAQ)
- Full data pipeline and offline computing needs/optimization (S&C)
- Resources needed: engineering/physicist hours



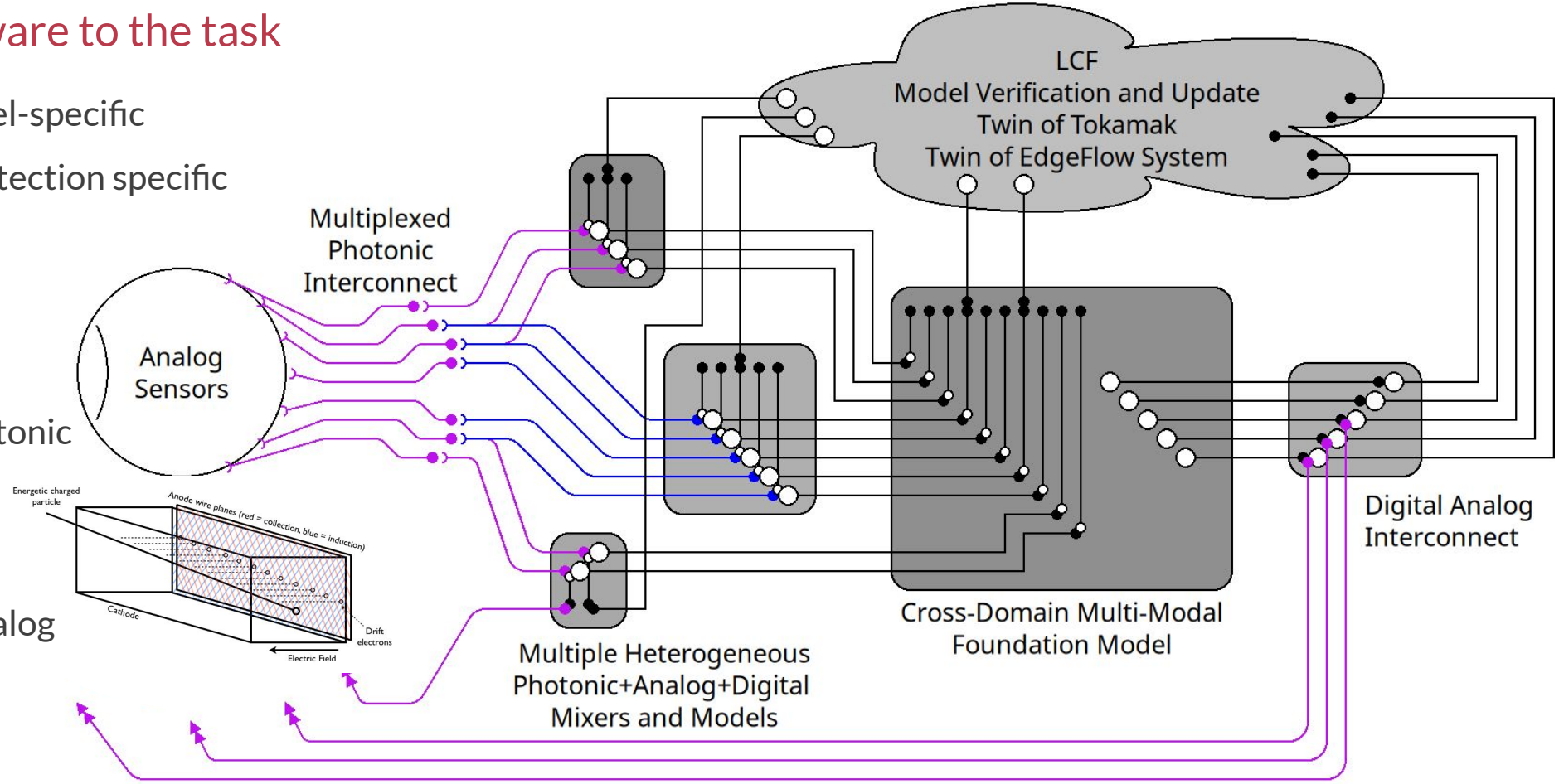
SLAC Synergies

- Phantom camera on-board processing is similar to the **“every event” readout** bandwidth problem.
- Photonic for **scintillation inference** (PET effort)
- analog for **charge cloud inference** (CookieBox)

Toward Higgs Factory Heterogeneous Autonomous DAQ

Optimize the hardware to the task

- Perform the channel-specific embedding with detection specific computing
- Scintillation → photonic
- Charge pulse → analog



TPC scheme by Rlinehan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45798181

Heterogeneous Computing Ecosystem … as it will be

Opportunity and Direction

- **International effort** for real-time Edge-HPC with early access and custom **streaming hardware**
- Nation Scale computing efforts linked internationally with global impact
- Plan for the bleeding edge of computing… in 2035!

Execution and Timeline

- Support Edge+HPC **linked testbeds** with **crisp HEP use case** as benchmark
- 5 years: Extending **IRI for International** HEP, Ultrawide Band Gap for **RadHard AI ASICs**
- 10 years: **Orchestration** of Heterogeneous flow informed/constrained by HPC resources and radiation environments
- 15 years: **Higgs Factory Autonomous Operation**

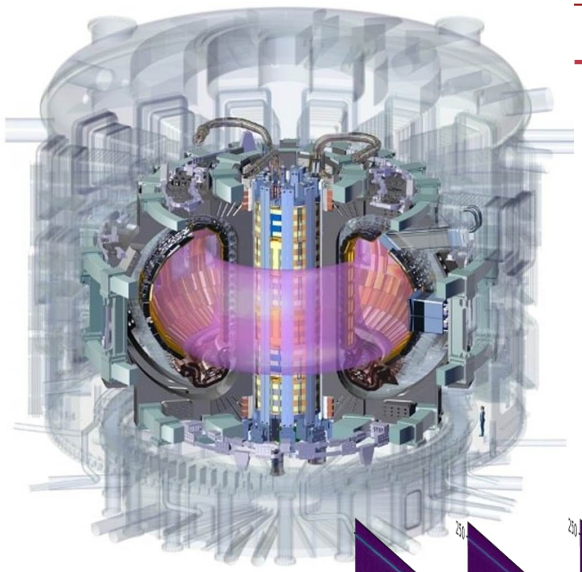
State of the Art and Challenges

- **Bulk data movement** to HPC is current tactic for conventional experiments
- Edge processing relegated to **isolated test stands**
  - repetition of effort
  - no economy of scale
- **Challenge**: Funding of Edge is siloed under each of HEP/BES/NP/FES/BER while for HPC it is ASCR

Potential Impact

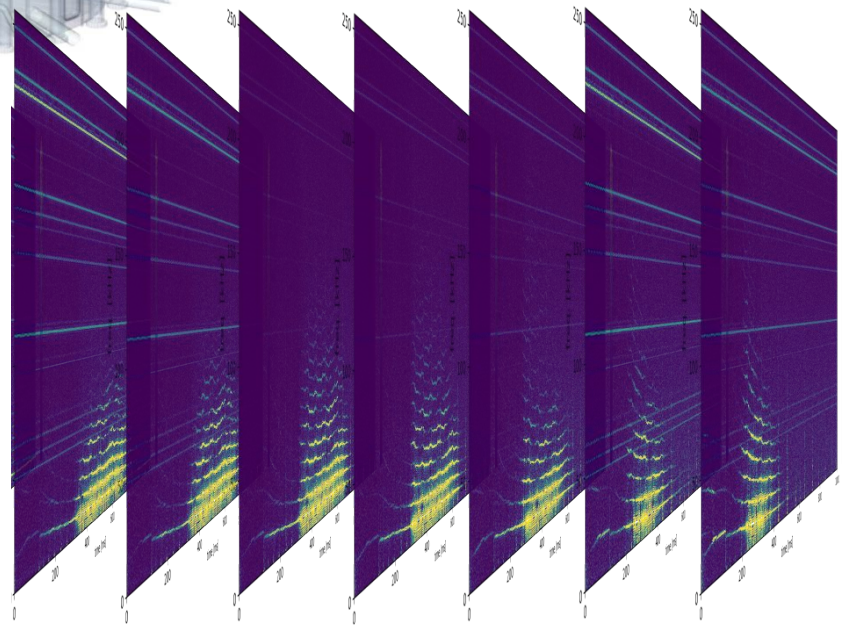
- Leverage **international network** of Nation Scale Computing from Cloud to Edge and back
- Computing infrastructure **as ubiquitous and essential as the interstate highway system**
- Coherent computing ecosystem from small to giant experiments via **Edge-to-HPC.**

Distributed sensors – Distributed computing



Tokamak magnetics

- Disruption **forecasting**
- Need **microsecond latency**
- Real-time controls fed by both **live and local** signal streams and **LCF twins**



Honeybee Acoustics

- **Natural** environmental sensors
- Signals functionally similar across **FES/BES/BER cases**
- ASCR build the tools to pull **all communities** into a Nation Scale computing ecosystem