Sensors-to-Edge-to-HPC ecosystem

In flight data analysis for autonomous control systems

Ryan Coffee / Sr. Research Scientist / LCLS+TID+PULSE

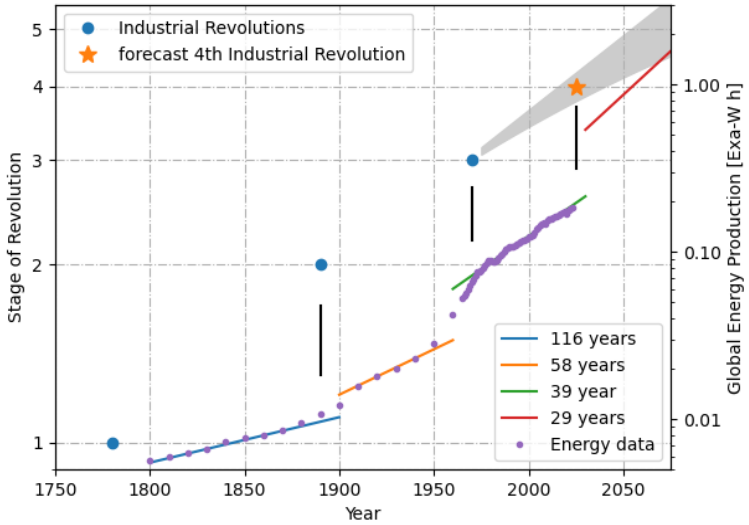
October 21, 2025

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University

U.S. DEPARTMENT of ENERGY

The Revolution is Now… Ubiquitous Distributed Compute



In the midst of a technology jump

- Greater than 2x jump in global energy consumption, borne by the technically advanced nations
- I bet **$20 on an 8x jump in the US**
- Doubling time drops from 39 to **29 years**.

Deployable Adaptivity

Low-power context-aware inference = **on-device autonomy**

We're not saving energy, just reducing energy per inference and moving to **ubiquitous, scaled, deployed, and continuous training and inference**

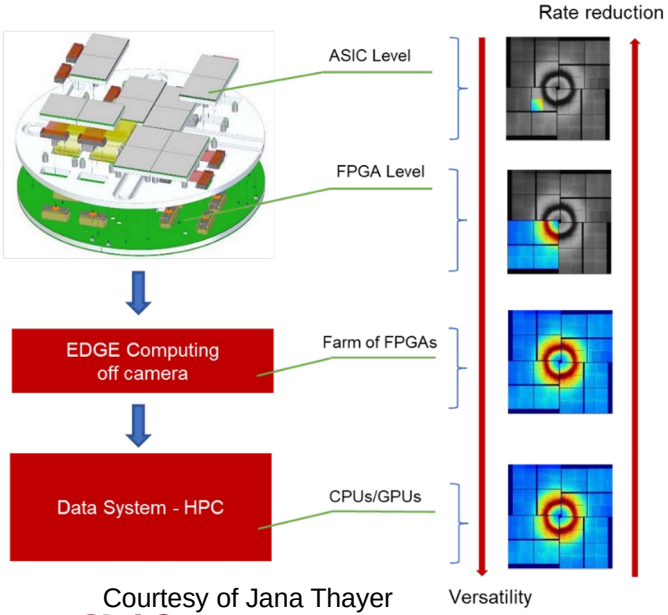
The Revolution is Now… Computational Ecosystem

Exemplar Systems of Systems

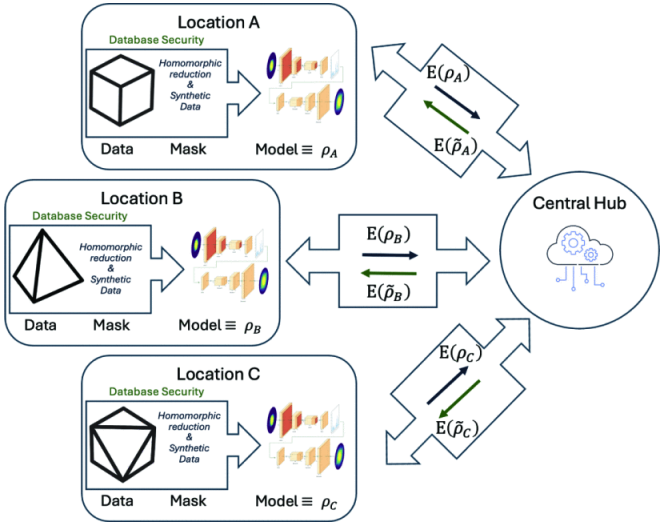
- Interconnected Light Sources
- Interconnected Fission and Fusion Reactors

Inherently Multi-scale

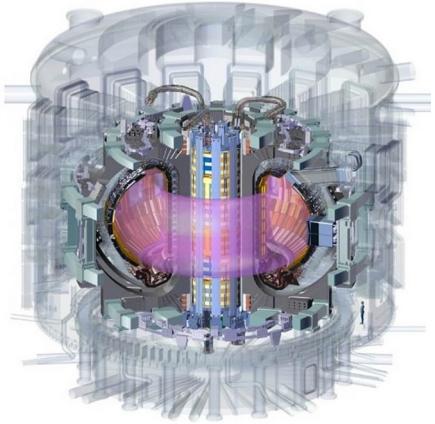
- µs-ms latency decisions
- Minutes-hours scale “run” evolution
- Days-months scale “campaign” forecasting



Courtesy of Jana Thayer



R. Archibald et al., "Privacy Preserving Federated Learning … ," 2024 IEEE BigData 2024, p. 4132, doi: 10.1109/BigData62323.2024.10825977.



General Atomics – DIII-D

SLAC

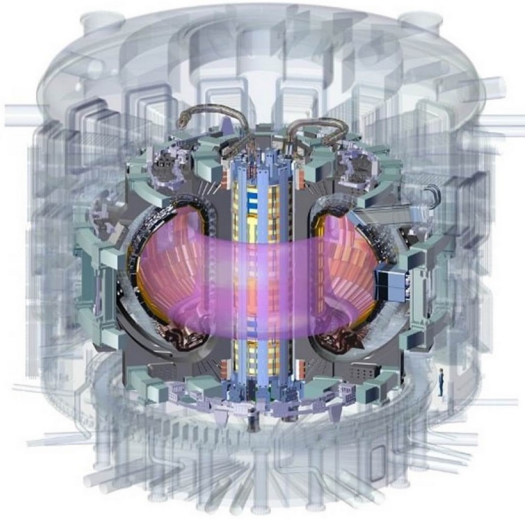
Autonomous Systems of Systems

At the light source

- 1MPix at 1MSps rates
- 1Byte/sample = 1TBps

At the tokamak

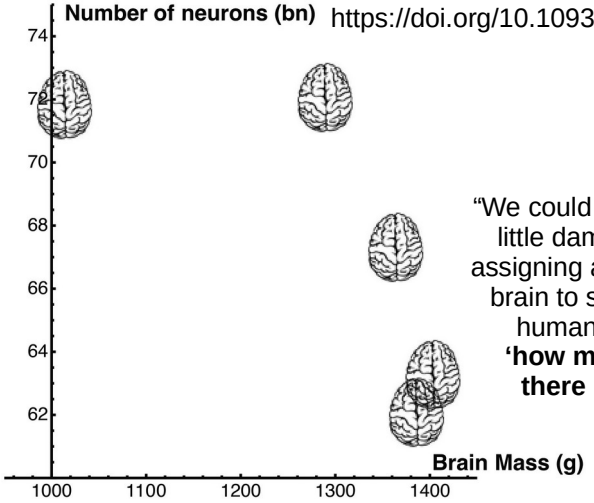
- 1k channels at GSps rates
- 1Byte/sample = 1TBps



Is AGI actually our goal?

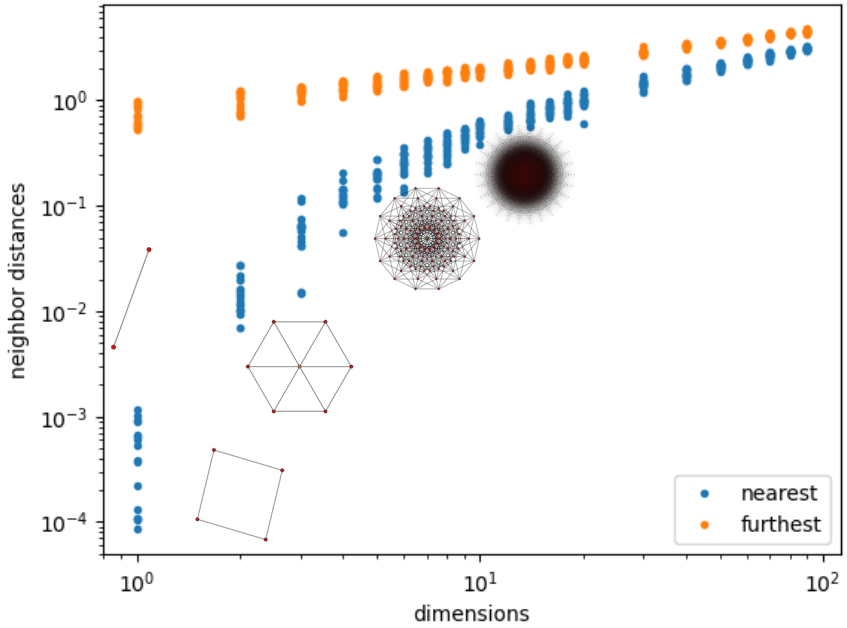
A word for dense information

- Three **orthogonal shadows** are likely better than 13 coupled
- **Interpolation evaporates** in high dimensions
- Reducing the chatter among information channels **reduces extraneous network traffic**
- Correlation dilution across channels increases the cross-section for **information leakage**.



https://doi.org/10.1093/brain/awae390

"We could argue that very little damage is done by assigning a number to the brain to satisfy our basic human curiosity about **'how many things are there in this thing?'**"
– A. Gorieli,



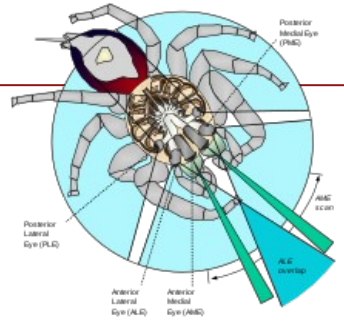
All you need is Edge

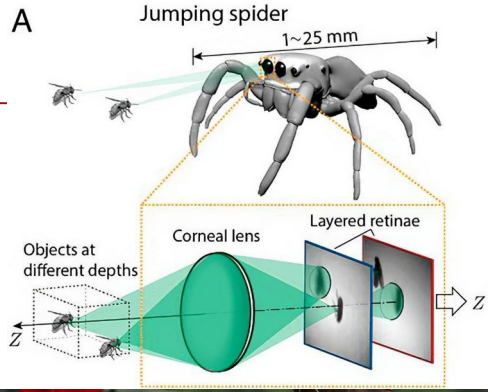
Hardware and wetware work in unison

ML in Science is predominantly acceleration of known interpretation

Let’s design for Jumping Spider Specificity and efficiency…

… We need an AI Pit Crew



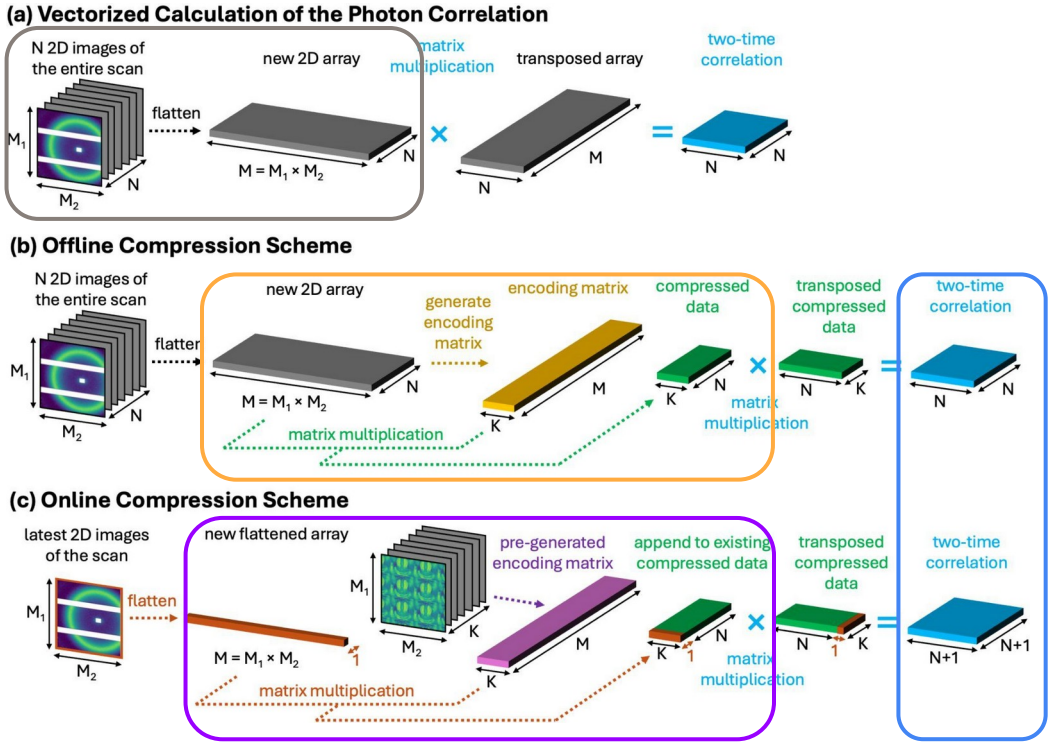


Domain and Detector specific

Generalizabe algorithms – what users **need**, not what they **want**

Recast offline for streaming – **offline encoding** explores **information sufficiency**

FPGA/ASIC encoding opens a new can of worms, e.g. **firmware is meta-data**



Strempfer *et al.*, "Homomorphic data compression for real time photon correlation analysis," Opt. Express 33, 12059-12070 (2025)

A case study – The Cookie Box

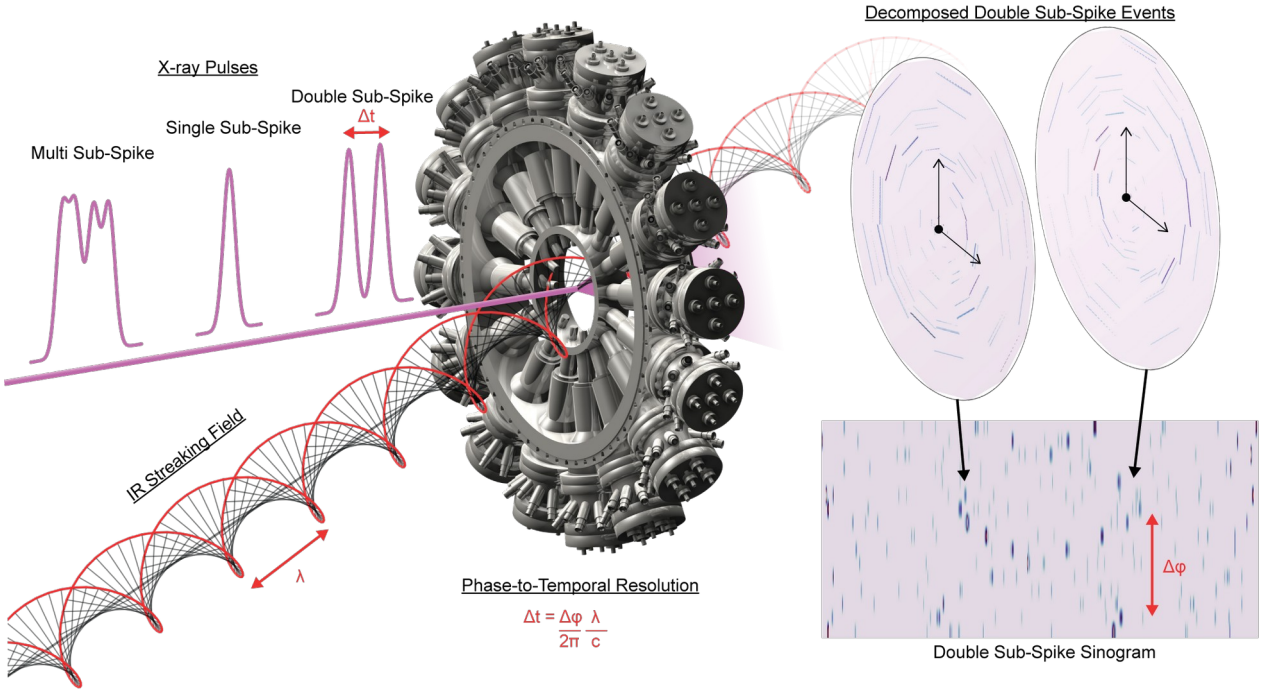
Caught with our hands in the…

X-ray Laser is **a wild ride**, just like the big wide world

Nature uses **context-dependent reasoning/responses**

She even pricks ears and shifts gaze as context evolves

Our detectors should do the same



Hirschman *et al., accepted* APL-ML

Autonomous Streaming Decisions

Modular parallel streams

- Denoiser takes longer than 23M parameter Regressor... **Trix are for FPGAs.**
- Motivates **direct coupling of different hardware**
- S3AI working with vendors (Groq and Cornami for now) to **open the communication** to their chips

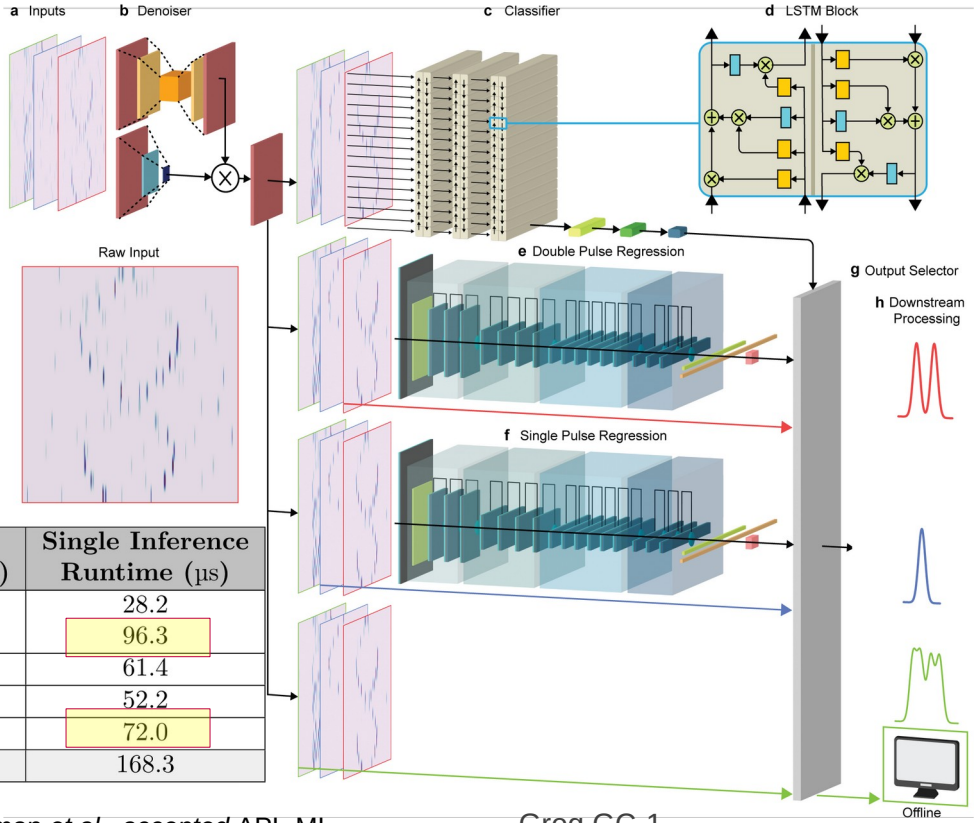


Table 2 DCIPHR parameters and runtime

Model (Identifier)		# Parameters	Parameter Memory (MB)	Single Inference 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

Hirschman *et al., accepted* APL-ML

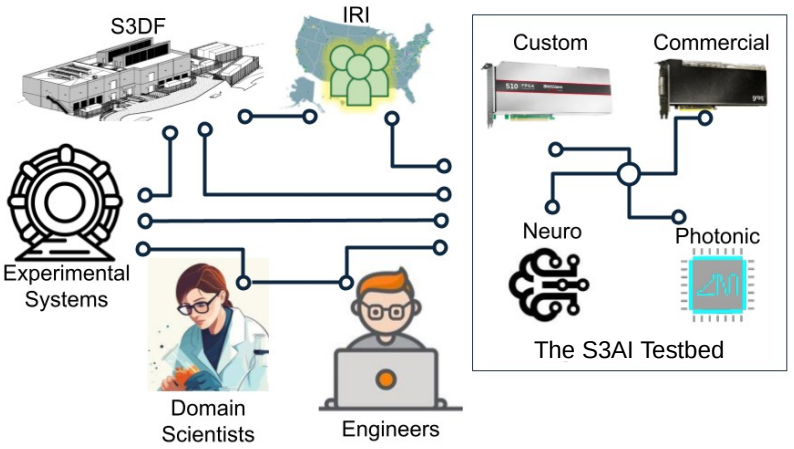
Groq GC-1

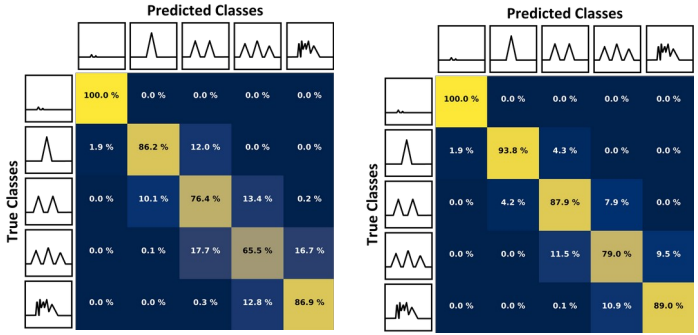
A Testbed of our Own

S3AI for heterogeneous exploration

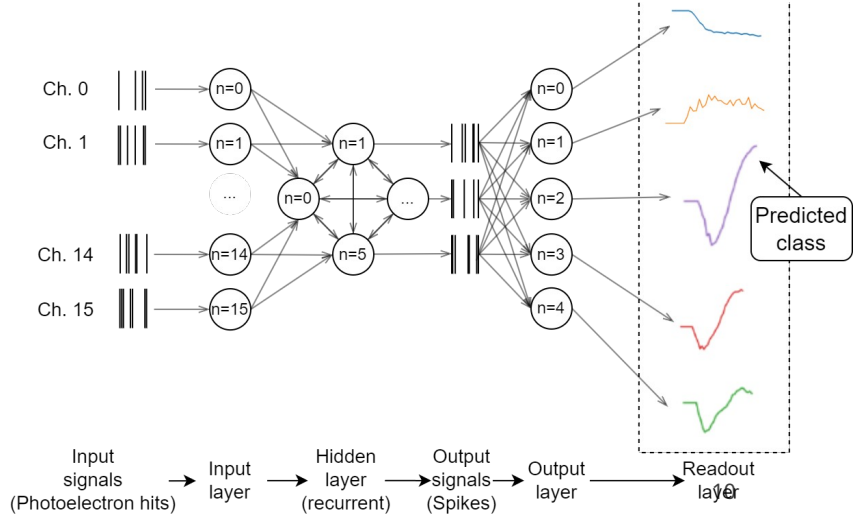
- Simulating direct conversion of time-series into spiking binary signals
- Comparing traditional approaches to **Spiking Neural Networks** ... just as Nature intended!

Courtesy of Ryan Herbst





SNN CNN



Gouin-Ferland *et al., submitted*

Taking Control of Our Destiny

Autonomous Control Signals

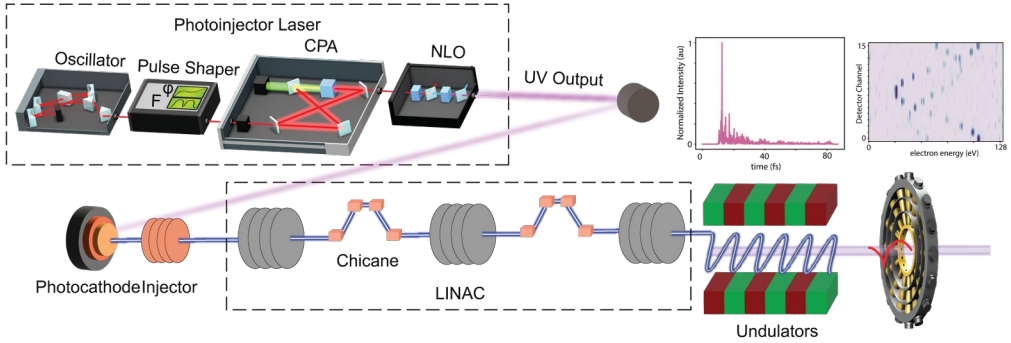
Real-time diagnostics **do not *need* to be human-readable**

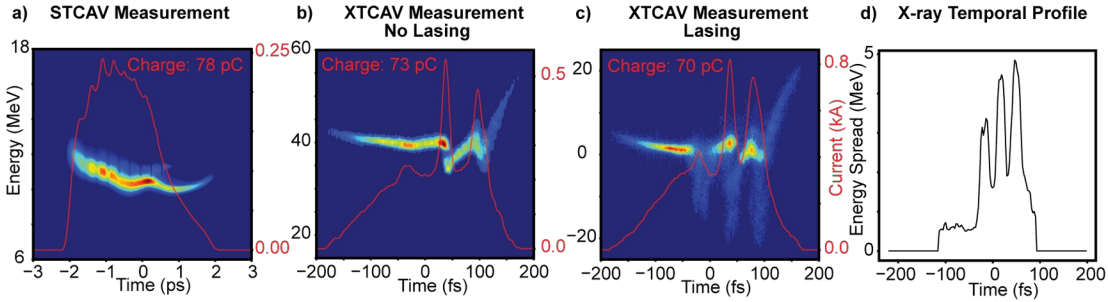
But **humans know *how*** to read the important minor features

Compression = reduction to those features

Multi-modal signal interpretation

- CookieBox for fast classification
- Xtcav/Passive Streaker for constraint
- Mfps distributed mutual information
- **Control feedback to injector laser shaper**





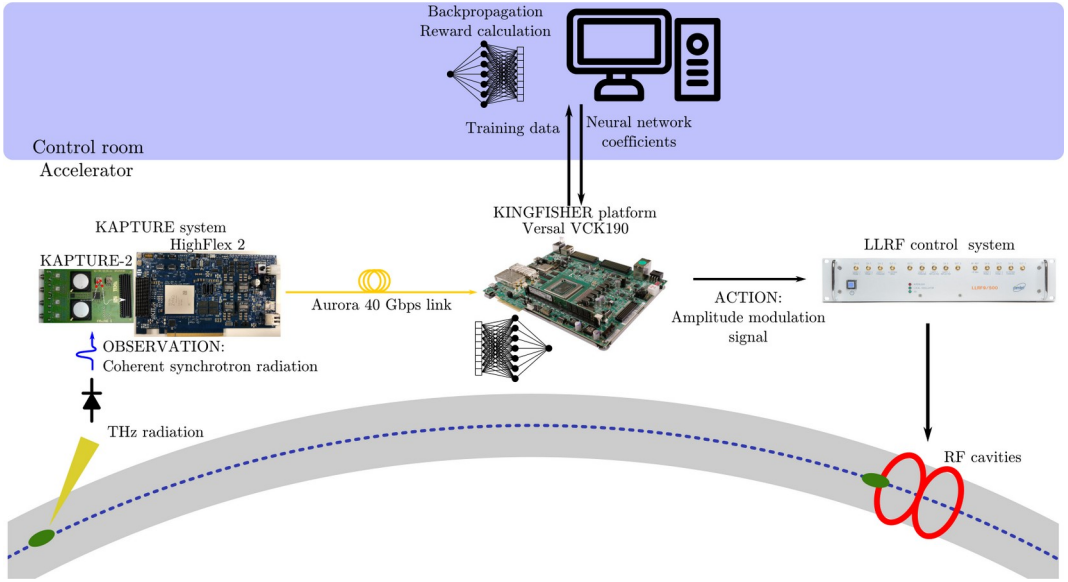
Hirschman *et al.*, submitted IPAC25

Actual Self-Adaptation – Even on FPGA

Real-time learning +
Ultrafast inference =
Autonomous Revolution

Context control = Agency for AI

- SNL allows for weight updates in real-time on FPGA
- Weights are trained on remote LCF e.g. Frontier ALCC project LRN045 (DIII-D forecasting)
- But honest autonomy requires bi-directional **real-time exchange** between the sensors, the controls, and the LCF.



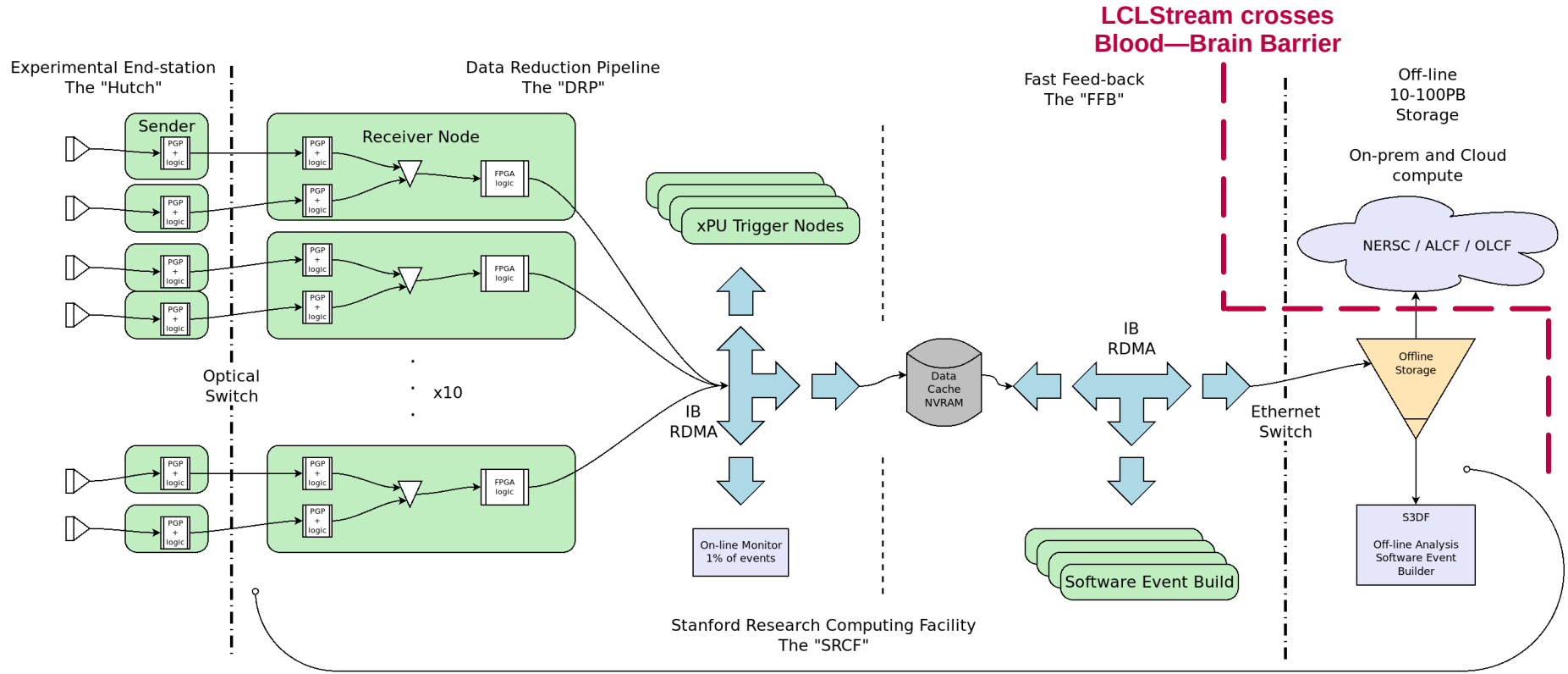
L. Scomparin *et al.*, "Preliminary results on the reinforcement learning-based control of the microbunching instability" (2024) 15th IPAC doi:10.18429.JACoW-IPAC2024-TUPS61

Autonomous Decisions – The Light Source Ecosystem

The Data Flow Reality

100s of meters of optical fiber between The Hutch and The DRP.

Event Building is a luxury that users take for granted



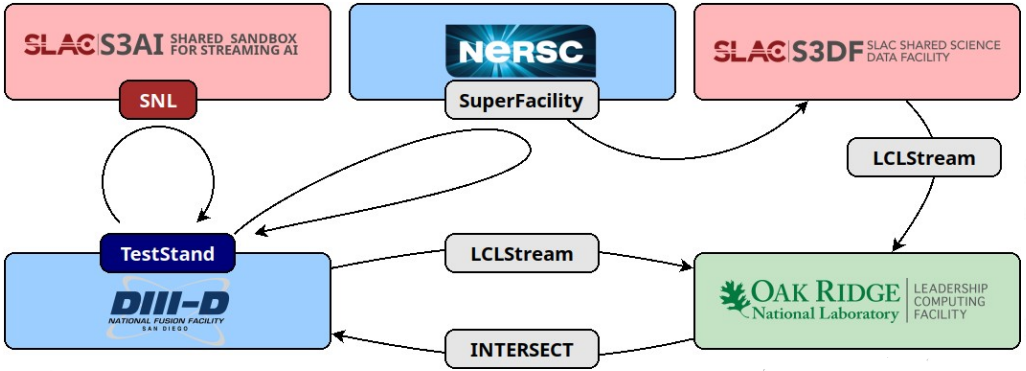
Autonomous Decisions – The Fusion Ecosystem

At the tokamak (eventually)

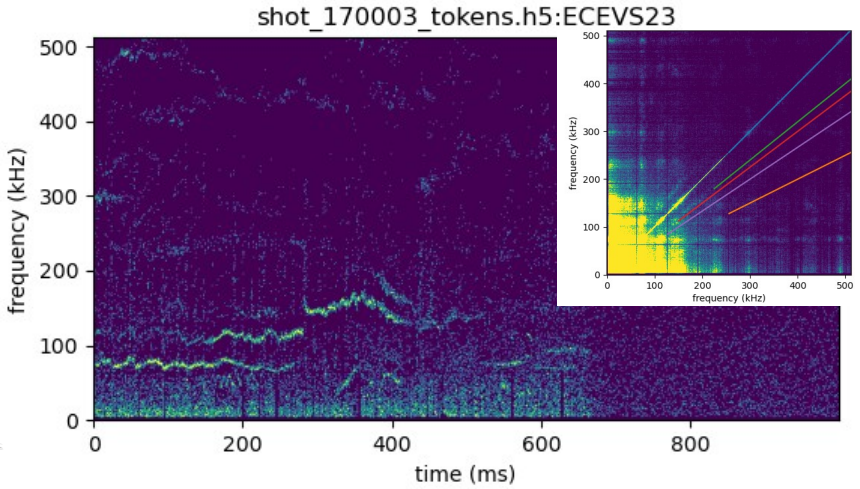
- 1k channels at GSps rates
- 2Byte/sample = 2TBps
- Get smart about mutual information

U.S. DEPARTMENT of ENERGY | Office of Science

Title:	**Real-Time Adaptive Disruption Forecasting at DIII-D**
Principal Investigator:	Ryan Coffee (SLAC National Accelerator Laboratory)
Co-investigators:	David Rogers (ORNL-NCCS)
ALCC Allocation Site(s):	Oak Ridge Leadership Computing Facility (OLCF)



FEDER
Fusion Energy Data Ecosystem & Repository
SLAC



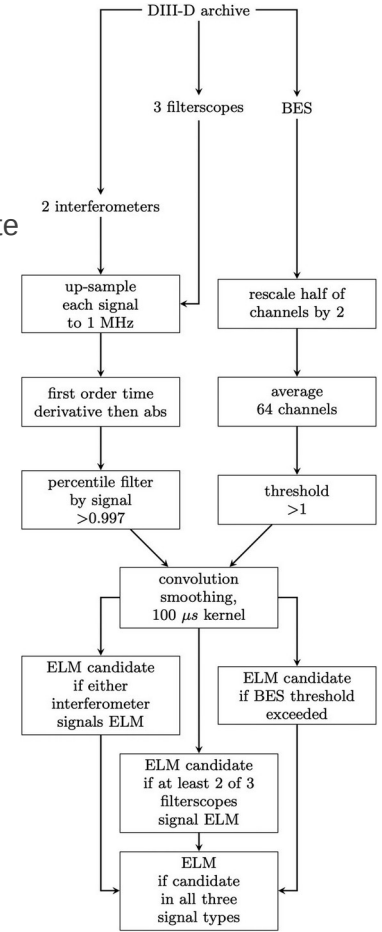
Triumvirate for the Win

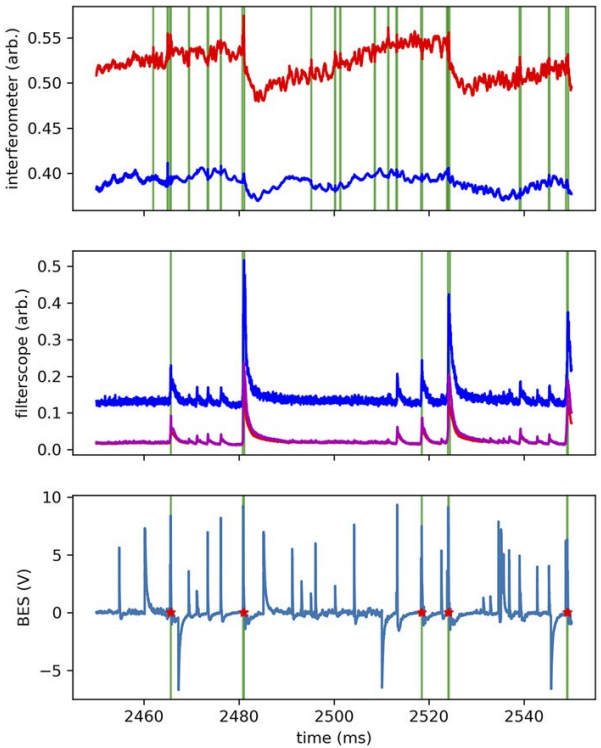
Decision Support Diagnostics

The DIII-D ELM identifier is itself **multi-modal**

- BES is an imaging modality highest sample rate
- Filterscopes are point detection but uniquely see ELMs
- Interferometers are ultrasensitive belt & suspenders

Incarnation of a mixture of **different** experts





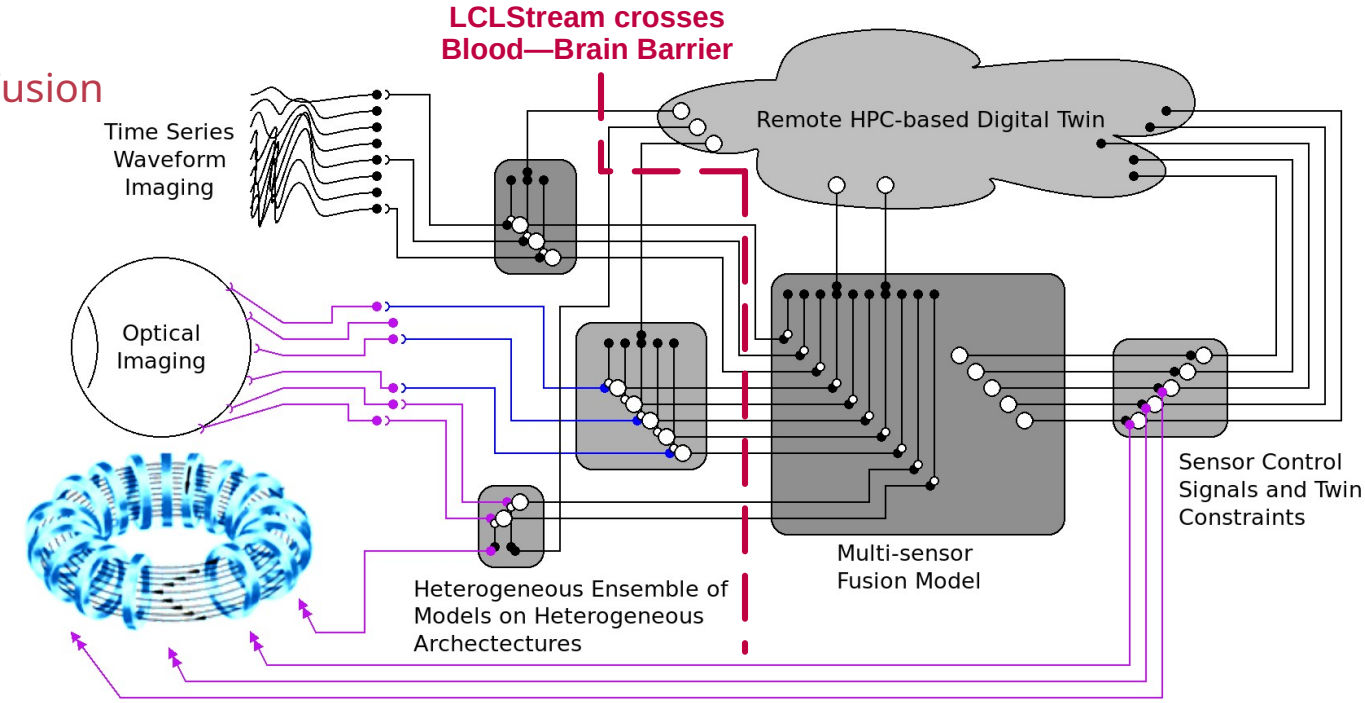
Finn H. O'Shea, *et al.,* "Automatic identification of edge localized modes in the DIII-D tokamak" APL Mach. Learn. (2023) 1, (2) 026102 https://doi.org/10.1063/5.0134001

An Important Destiny to Control

Orthogonal Views on Fusion

Breaking through silos, linking across Labs, and partnering with industry

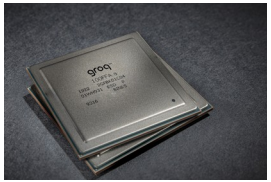
- LCLStream model for **Edge-to-Exa streaming**
- INTERSECT enables geographically **remote autonomous control**
- Embrace industry collaboration for **direct coupling of sensing to inference acceleration**



XILINX

CORNAMI
Intelligent Computing

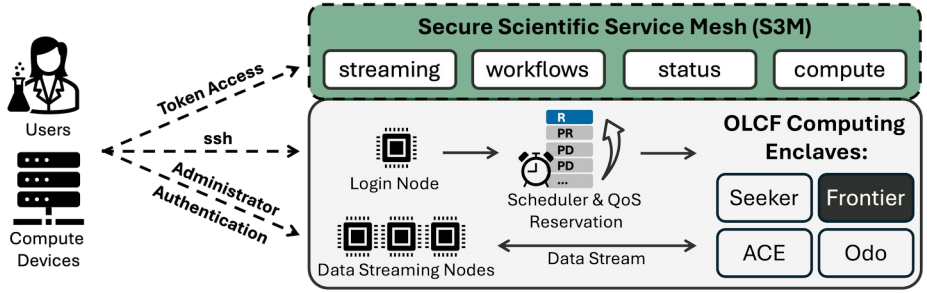
NVIDIA



Fitting the Pieces Together for the Ecosystem

Shovel-ready Components

- Focus on what we **have** rather than what we **want**
- **Interconnect modules** rather than boil the ocean
- Must **break down rivalries**
- This is the time to **burrow under silo walls**

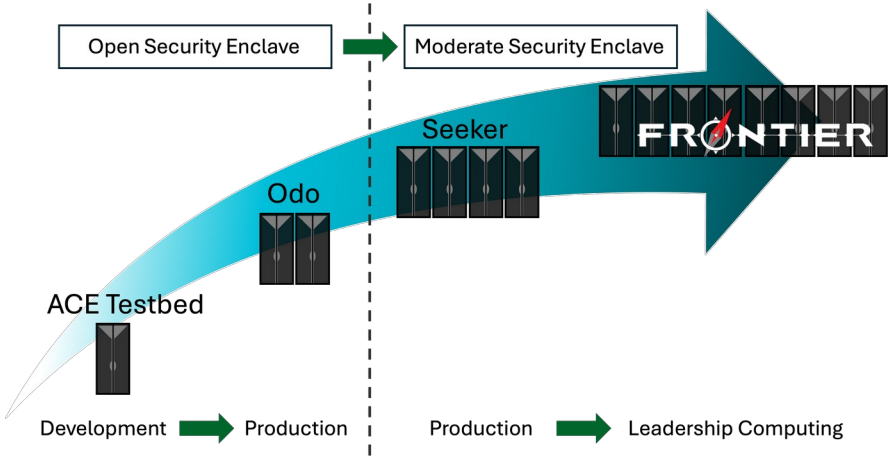


U.S. DEPARTMENT of ENERGY | Office of Science

Title: **Real-Time Adaptive Disruption Forecasting at DIII-D**
Principal Investigator: Ryan Coffee (SLAC National Accelerator Laboratory)
Co-investigators: David Rogers (ORNL-NCCS)
ALCC Allocation:
Site(s): Oak Ridge Leadership Computing Facility (OLCF)
Allocation(s): 350,000 on Frontier
Research Summary:

This project will extend an existing statically trained machine-learning based disruption prediction model for tokamak fusion reactors by leveraging a meta-learning method for fast optimization of the plasma state forecasting model and also leveraging an encoder/decoder model that accommodates a dynamic quantization scheme. The quantization will be optimized to explore the space of model and feature encoding. This exploration will inform decisions about reduced fidelity diagnostics that would remain sufficient for a reactor regime tokamak state prediction consistent for energy on the grid.

2025 ASCR Leadership Computing Challenge Award

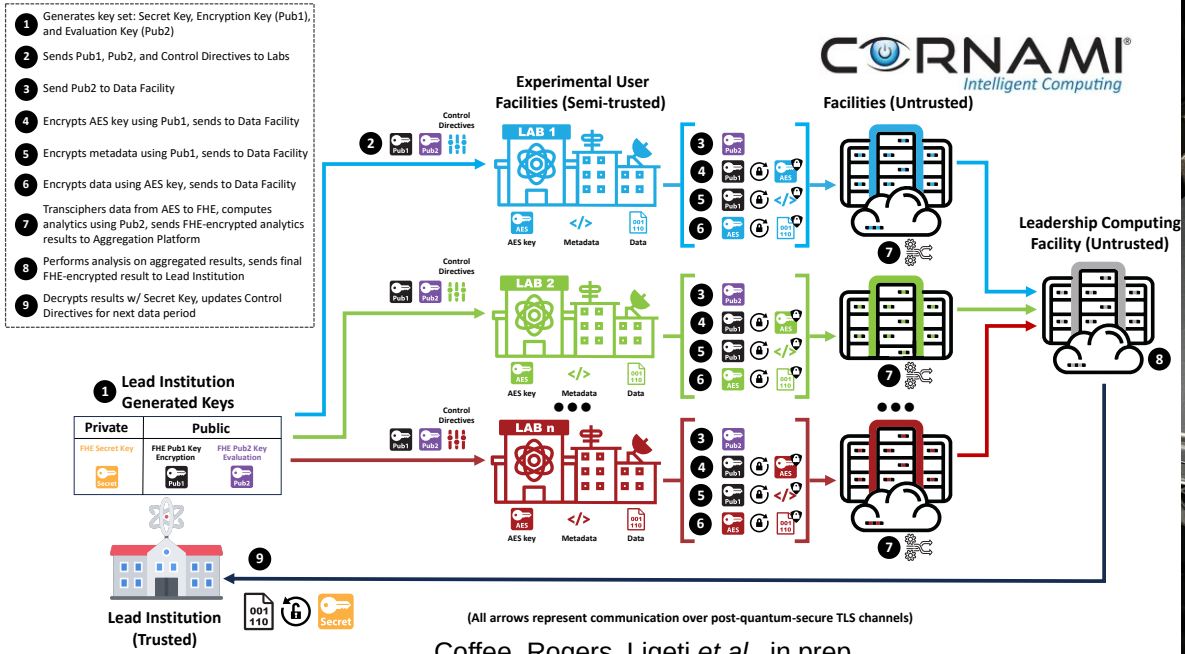


Etz, Rogers, Brim, *et al.*, "Enabling Seamless Transitions from Experimental to Production HPC for Interactive Workflows" arXiv:2506.01744v1 [cs.DC]

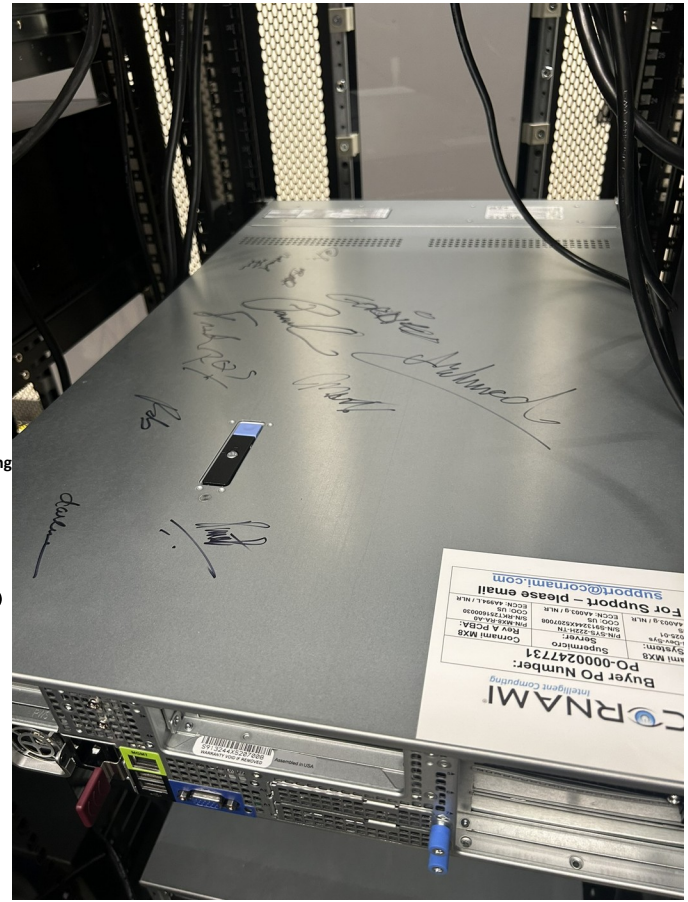
Future-proofing the Ecosystem

Inherently Multi-Party

- Security with **Shared Resources/Orthogonal views**
- Interconnecting **secure enclaves with open cloud**
- Design for **“Collaborating Competitors”**



Coffee, Rogers, Ligeti *et al.,* in prep



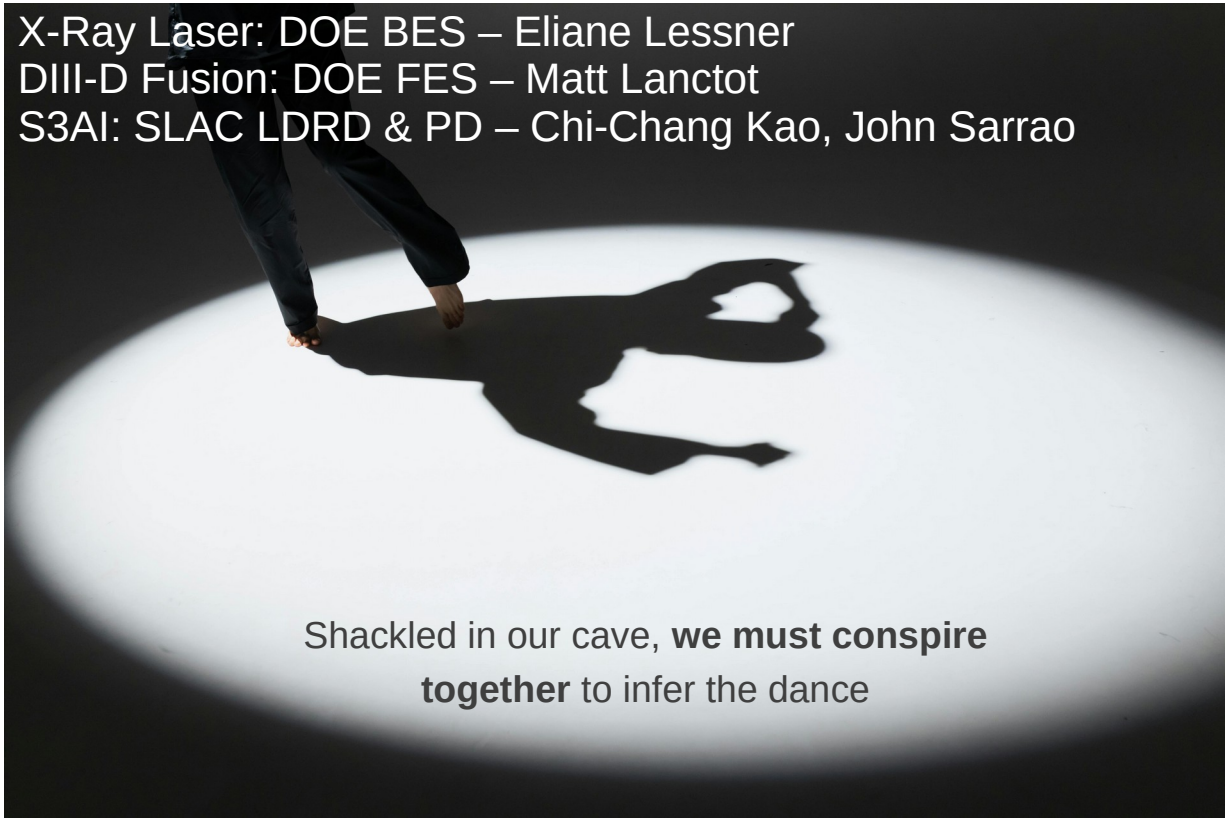
How I learned to stop worrying…

Dance of Shadows

Although beautiful, a single shadow does not well represent the dancer

Multi-modal signal interpretation

- Every sensor is responsible for its own **brutal parsimony**
- Nature abhors redundant arrays
- Autonomy requires a rich, diverse, and efficient sensor environments
- It's all for naught if you have to deliberate every move **… latency really matters**



Thank You!

Now let’s get crackin’

coffee@slac.stanford.edu

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University

U.S. DEPARTMENT of ENERGY