Kicking our veto addiction:
Accelerating Edge Computing for Tailored Lossy Compression

Ryan N Coffee / Sr. Staff Scientist / PULSE, LCLS

November 8, 2023

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University | U.S. DEPARTMENT OF ENERGY

OUTLINE

The View from LCLS-II

Smart Triggers *vs* Smart Compression

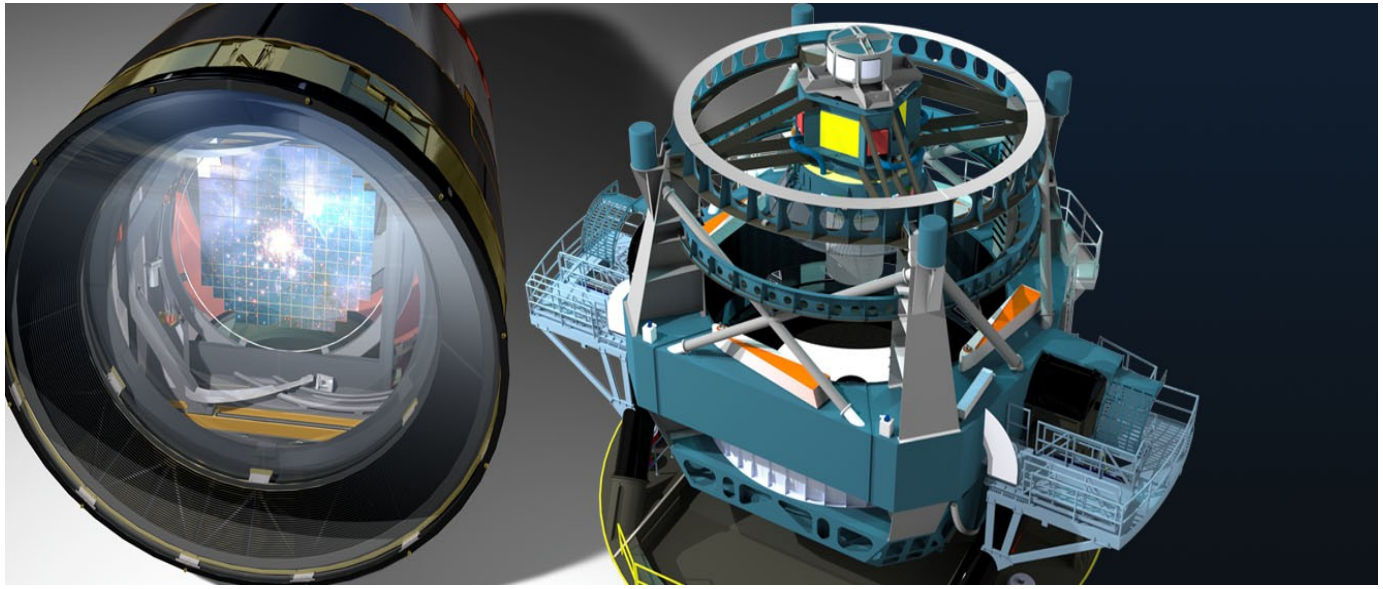
Data Flow pattern

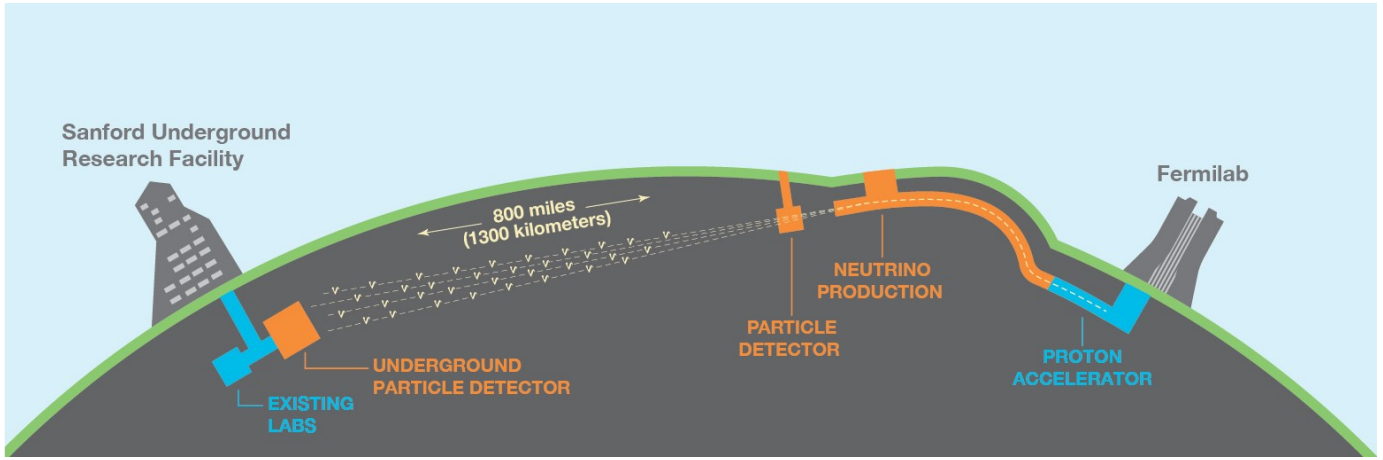
Distributed Federation

SLAC NATIONAL ACCELERATOR LABORATORY

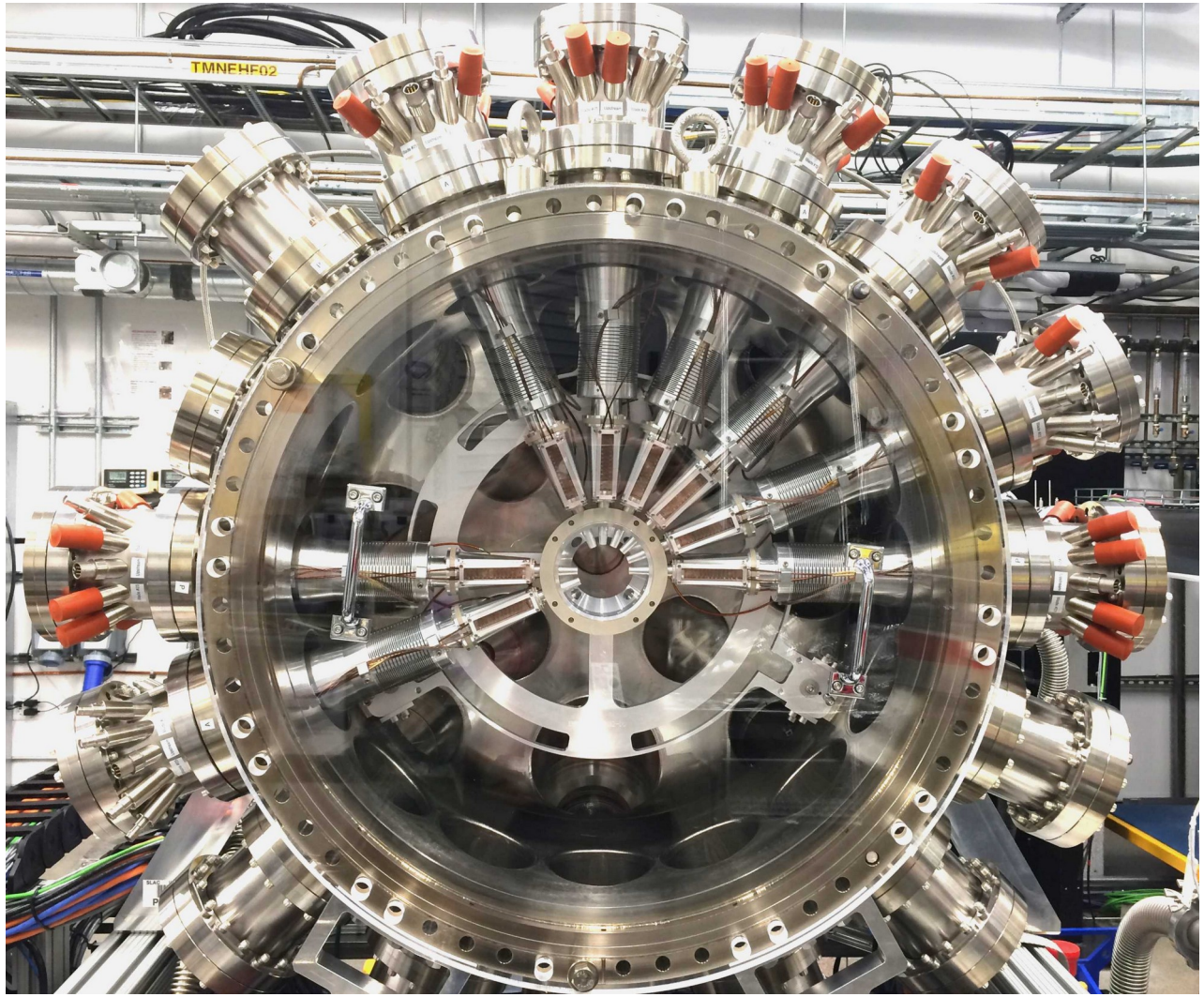
Stanford University | U.S. DEPARTMENT OF ENERGY

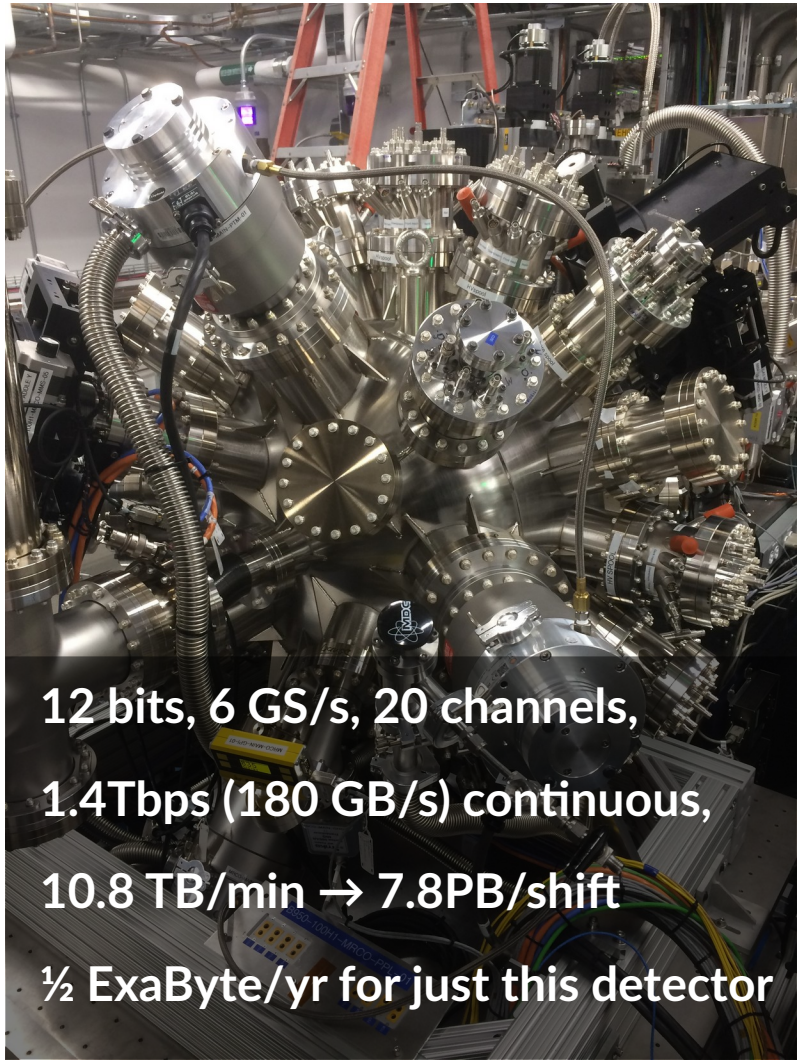
From HEP and Cosmology…





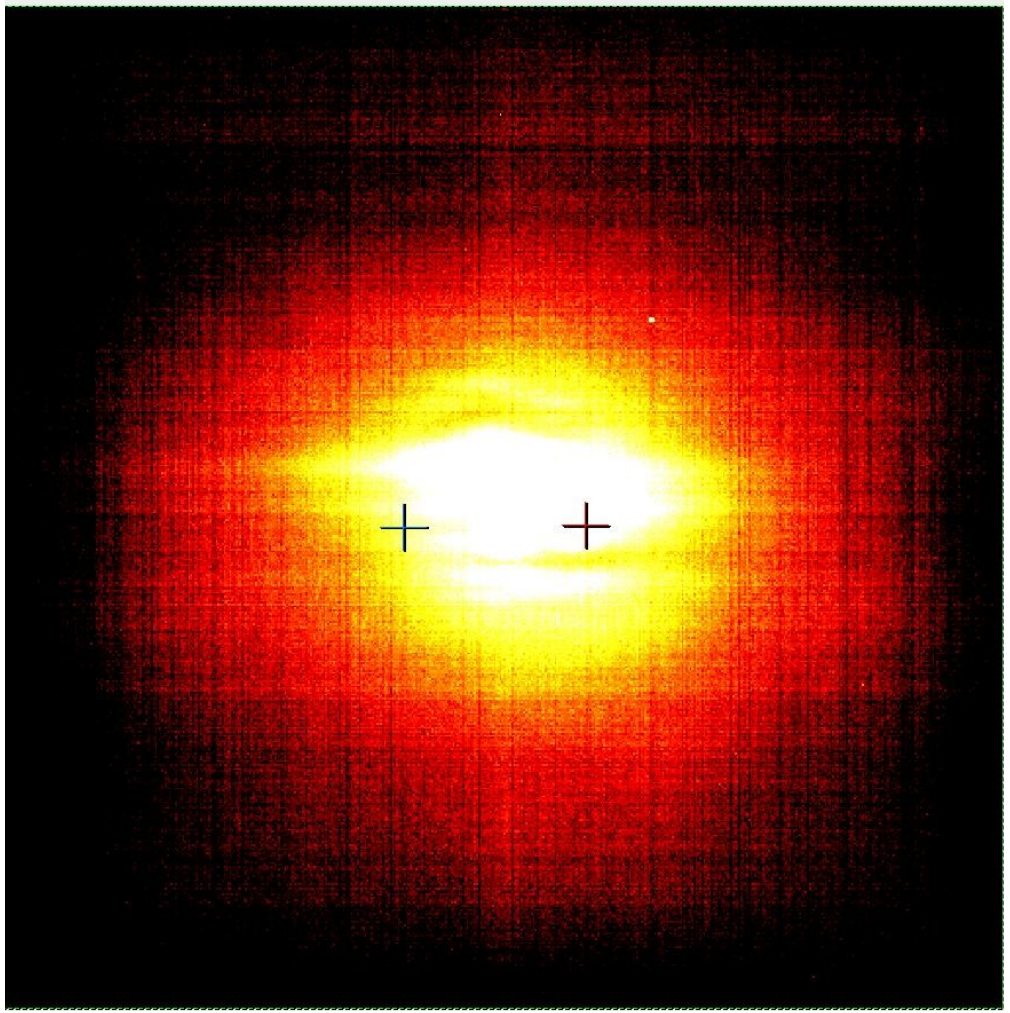
… to a view from LCLS-II





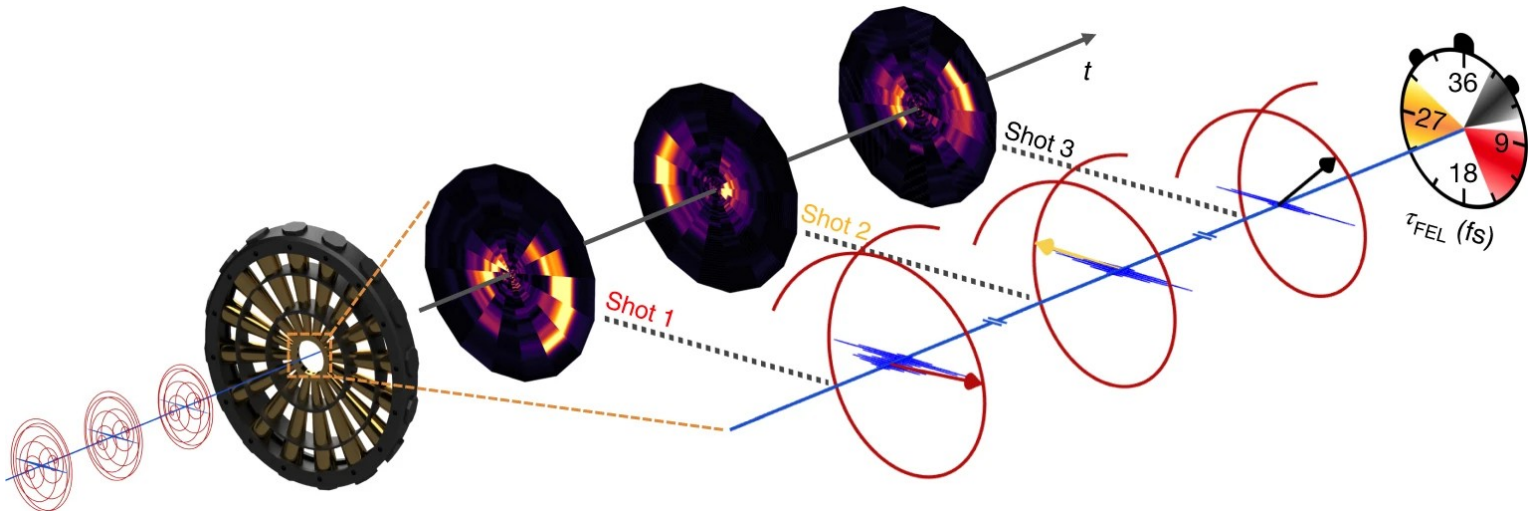
LCLS-II First Light

The Data Deluge has begun



SLAC

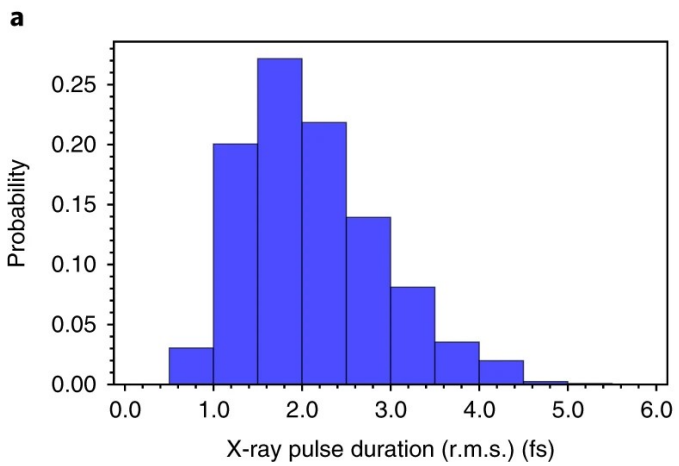
Smart Triggers at 1 MHz: The Attosecond Example

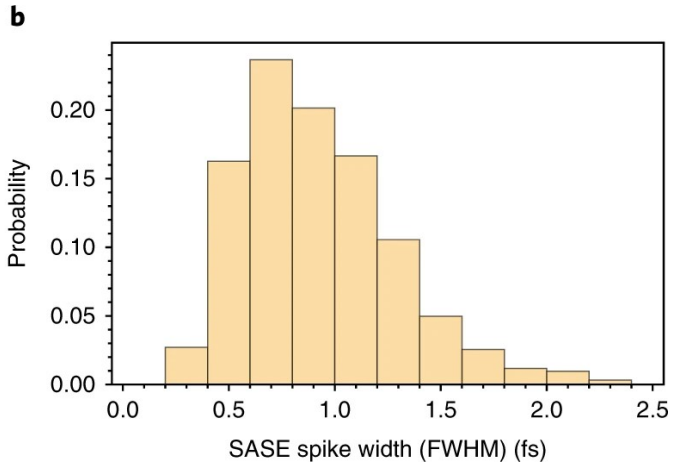


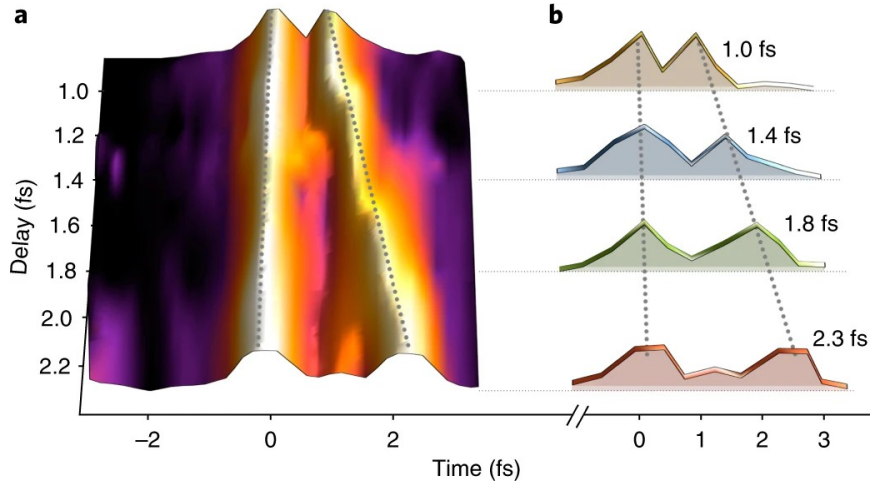
Angle resolving high-multi-resolution electron time-of-flight spectrometer

Aimed at multi-**everything** SASE pulse reconstruction

Creative use is correlation spectroscopy for non-linear x-ray interactions

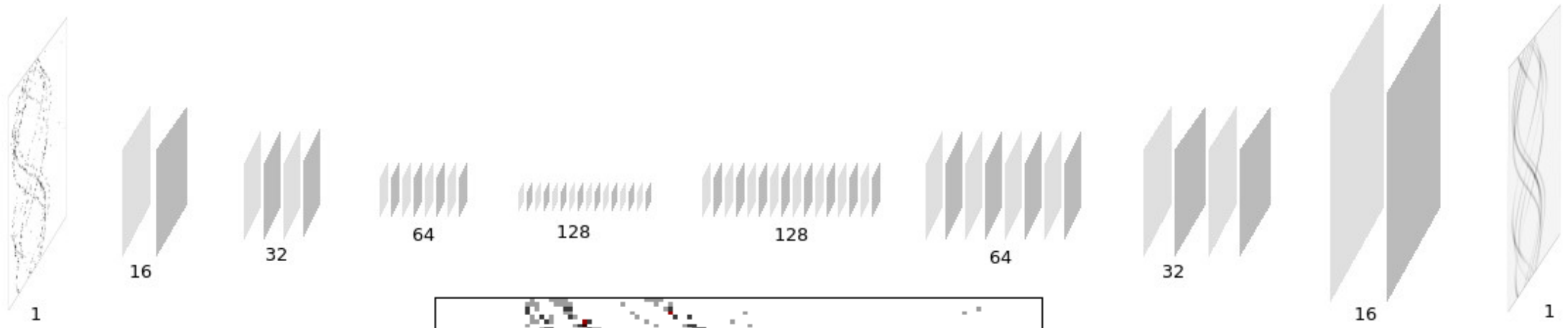






Hartmann, N. *et al.,* Nature Photonics **12** 215–220 (2018)

Smart Triggers at 1 MHz: The Attosecond Example



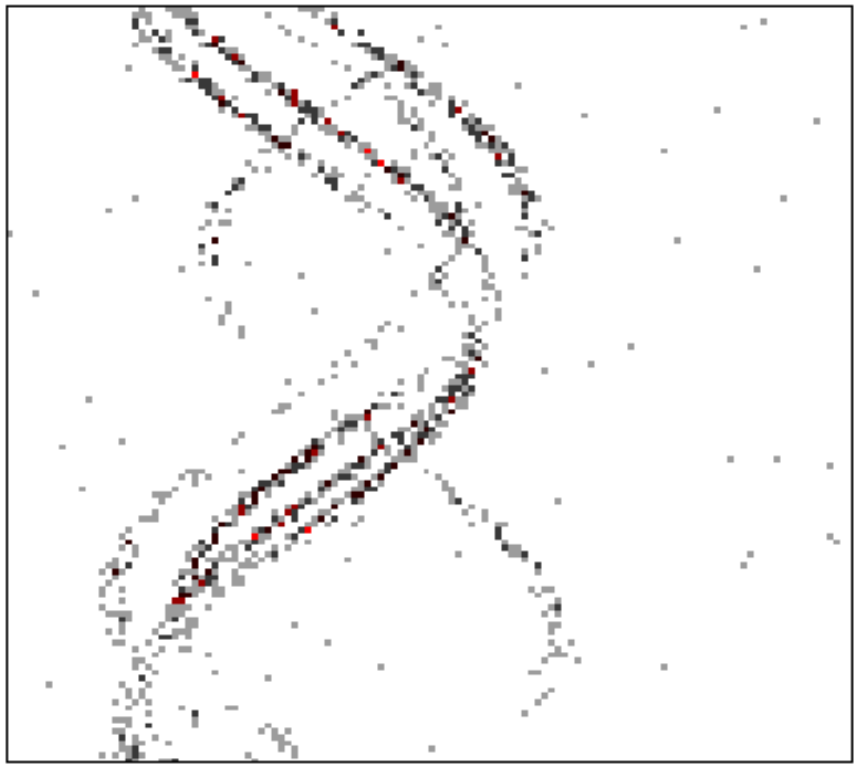
Simulation *vs* Data

Simulation gives ground truth

Data is “honest”

Transfer Learning is fraught

Both share the “structure” of the relevant information *vs* noise/stochasticity



Images *vs* Channels

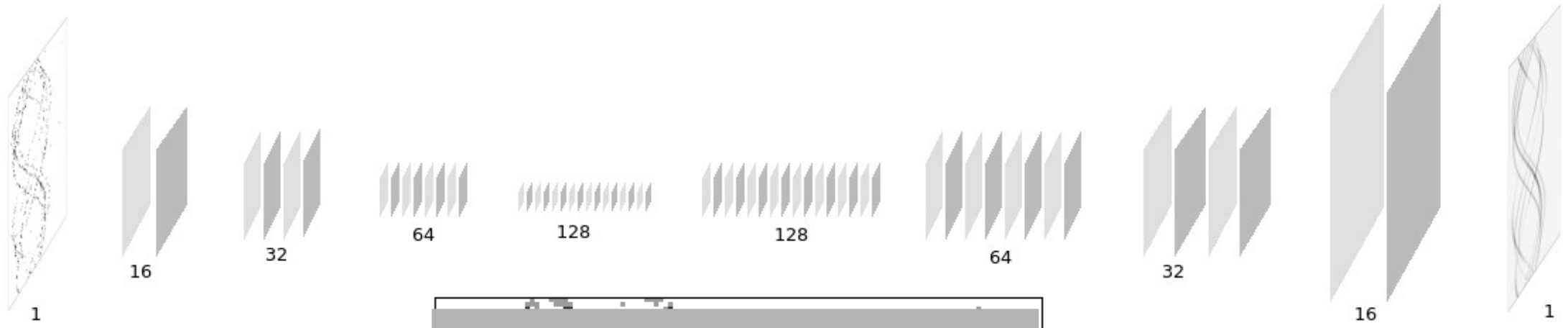
Images are grainy

Information is smooth

Channels can be very heavily compressed

Sparse information is bad for business (co-processor acceleration)

Smart Triggers at 1 MHz: The Attosecond Example



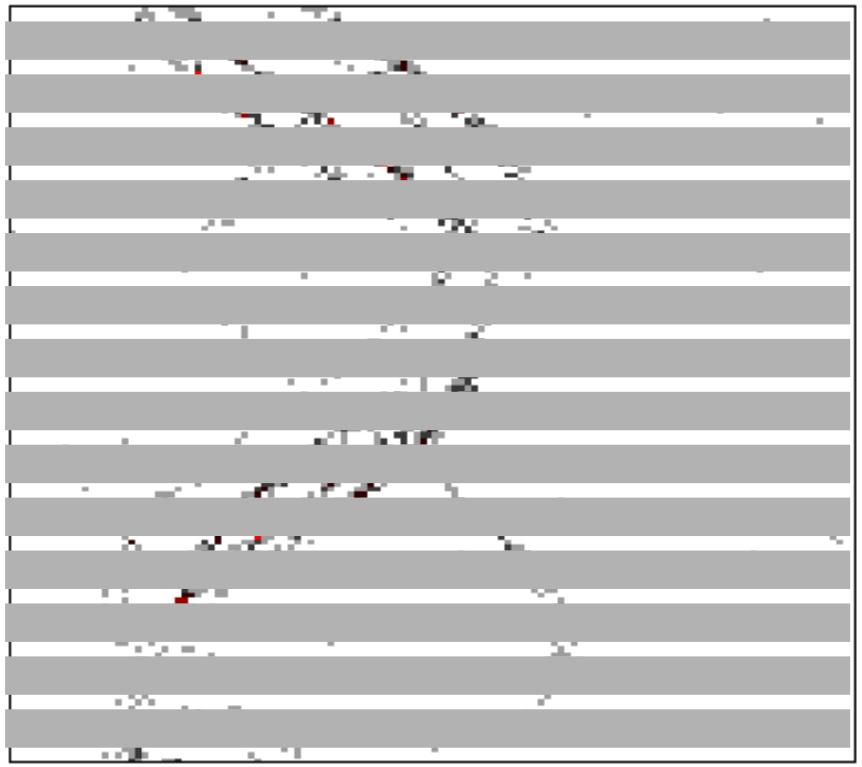
Simulation *vs* Data

Simulation gives ground truth

Data is “honest”

Transfer Learning is fraught

Both share the “structure” of the relevant information *vs* noise/stochasticity



Images *vs* Channels

Images are grainy

Information is smooth

Channels can be very heavily compressed

Sparse information is bad for business (co-processor acceleration)

CookieNetAE training on SambaNova and NVidia

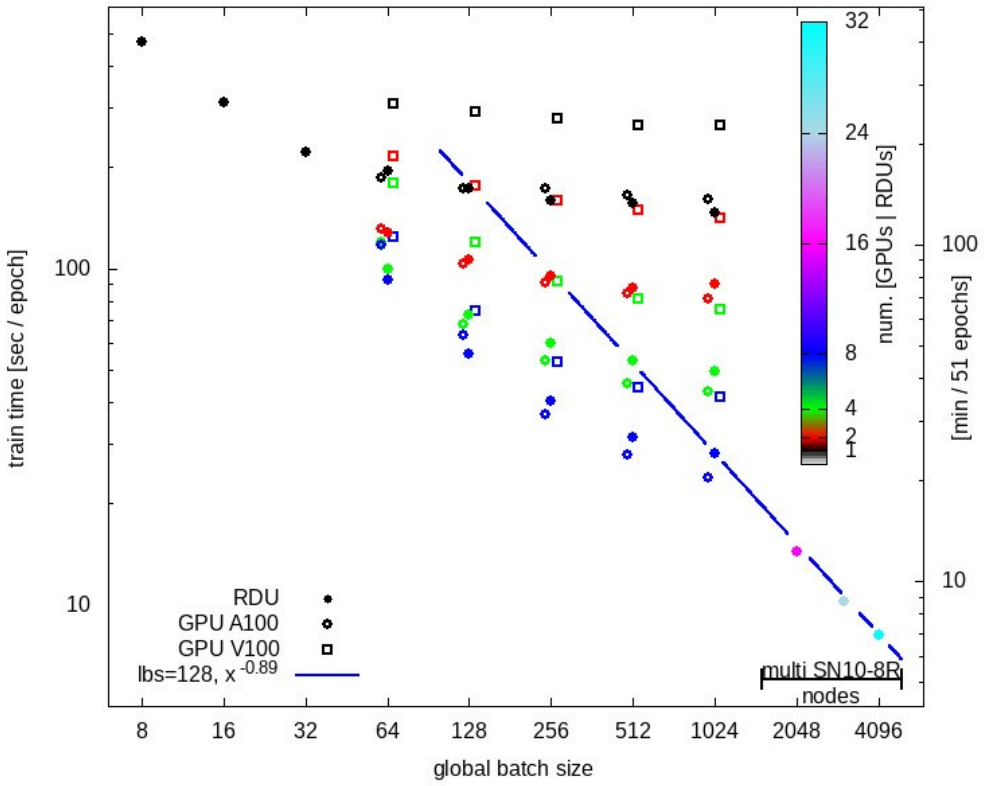
Encoder-Decoder architecture on SambaNova

Data parallel training for low to moderate batch size.

Under 10 min scratch retraining @ 1/3 M parameters

Scales well to multi-SN10-8chip nodes.

P Milan *et al.,* Front. Phys. **10** (2022)

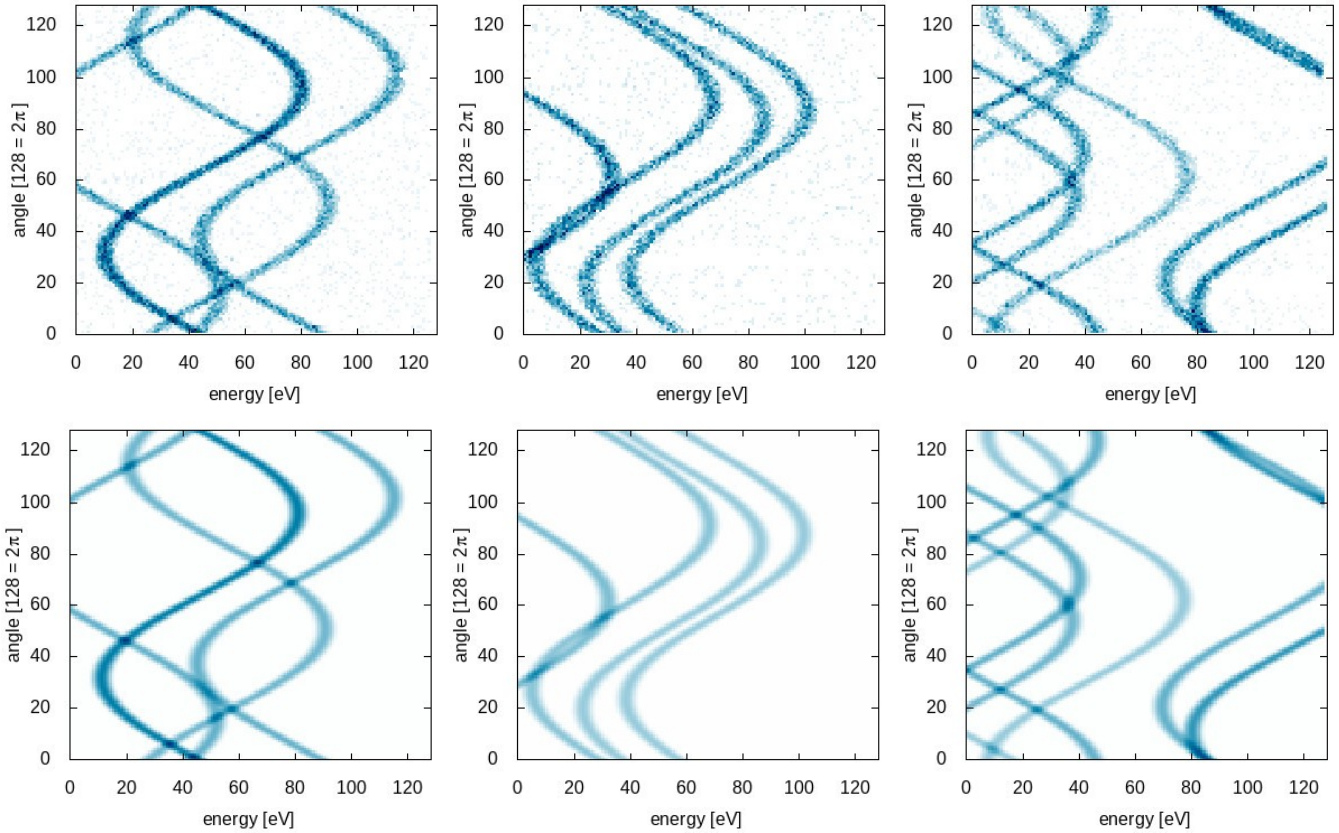


Testing across the heterogeneous ecosystem

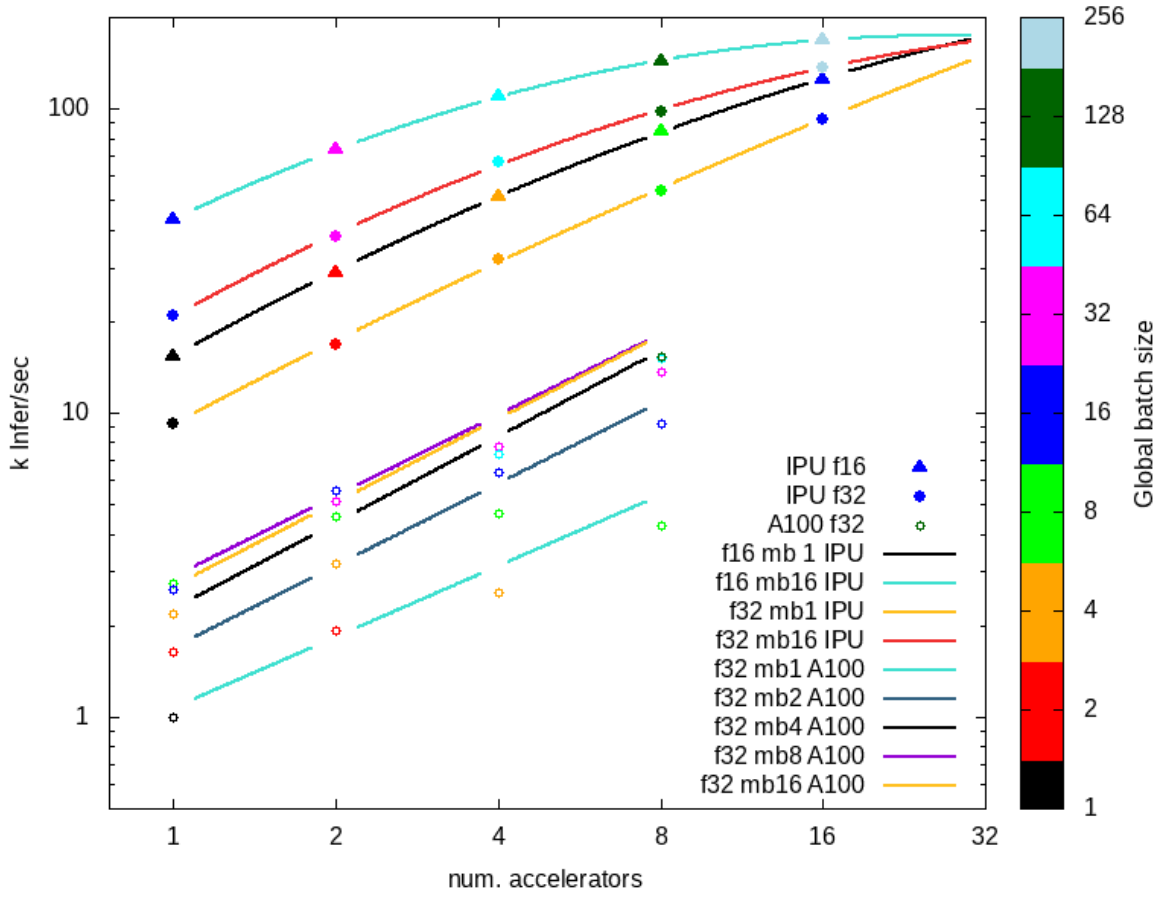
Cerebras CS2,

Graphcore POD-16,

ANL's ThetaGPU



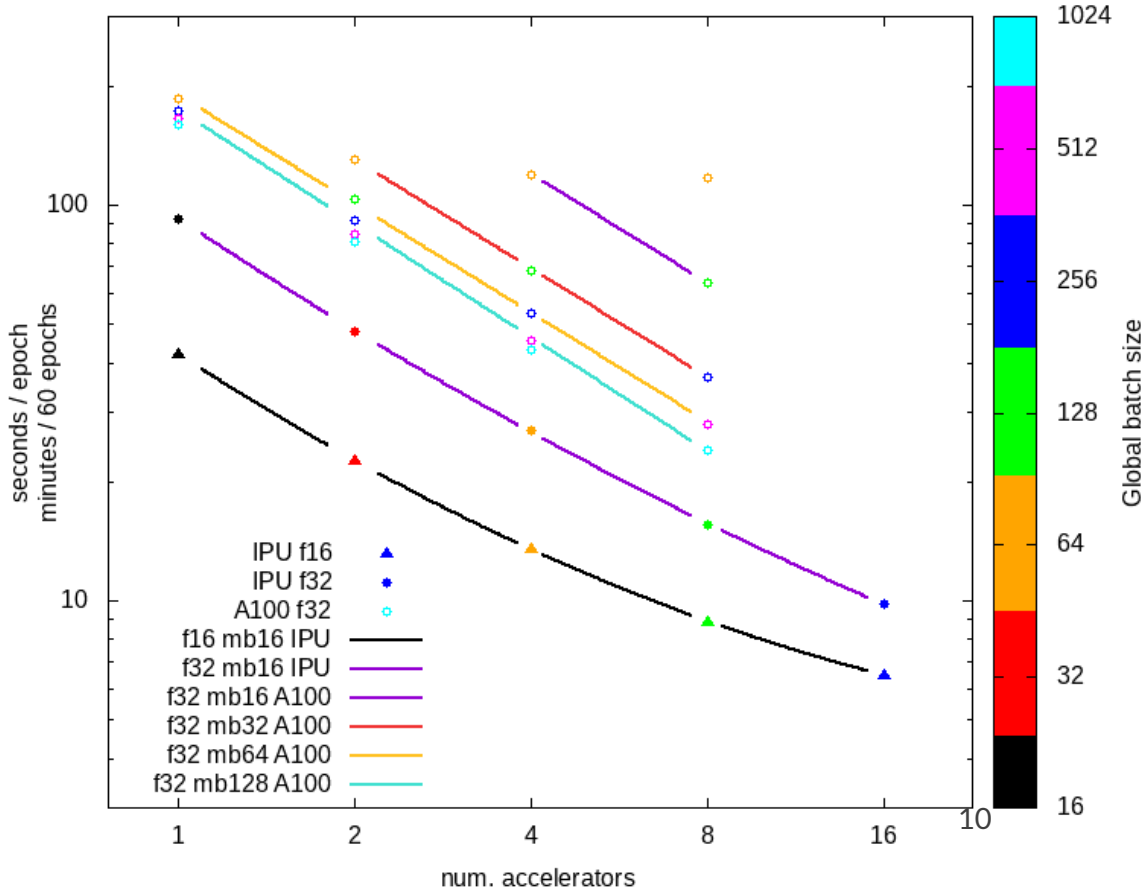
CookieNetAE on Graphcore and Nvidia



Data Parallel Patterns and Direct-Attached

Parallel inference is key for Batch Size = 1 streaming

Small Batch re-training is also expected to be highly used



SLAC

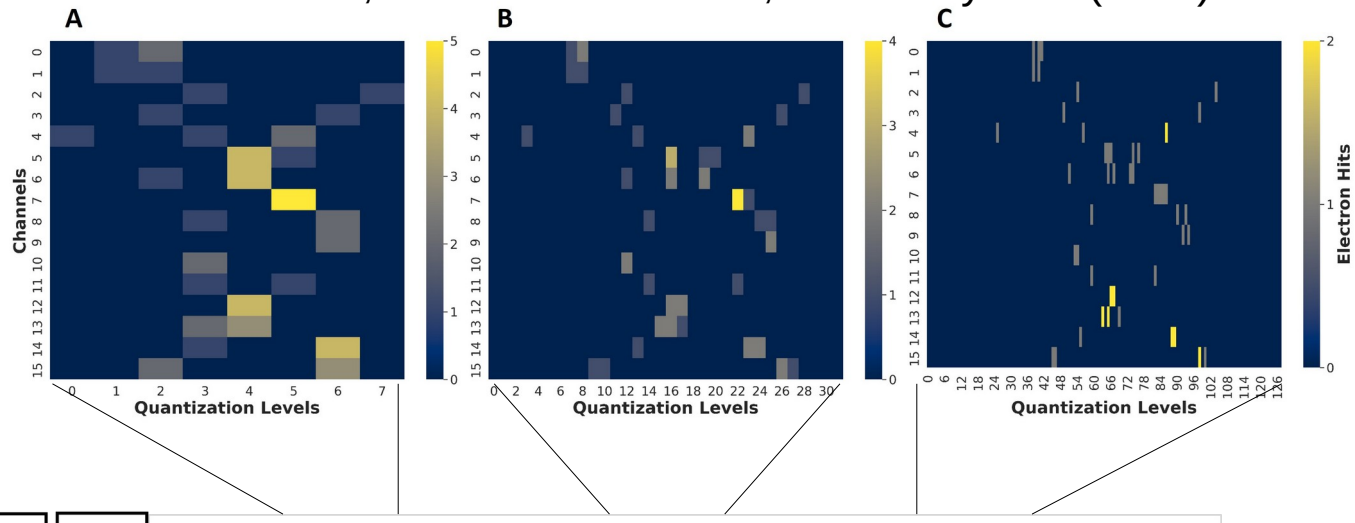
M Kraus *et al.,* Front. Phys., **10** (2022)

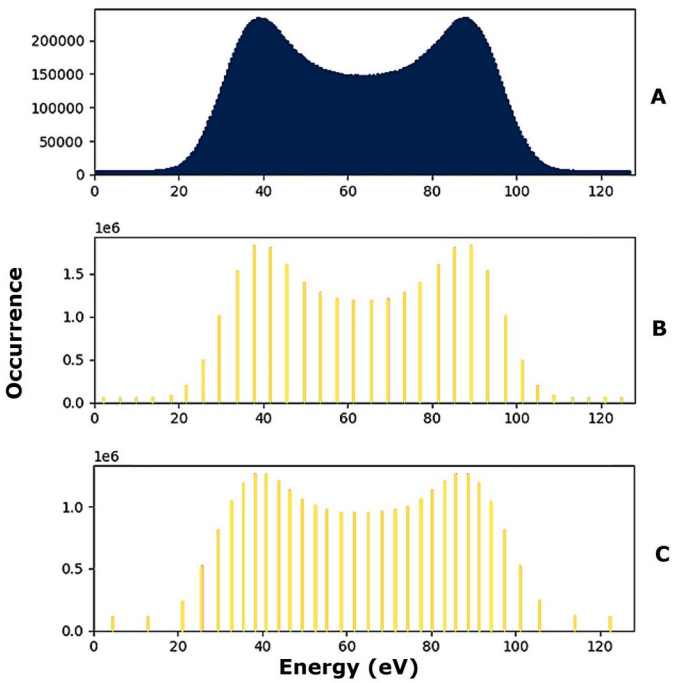
Smart Compression = Optimizing information per bit

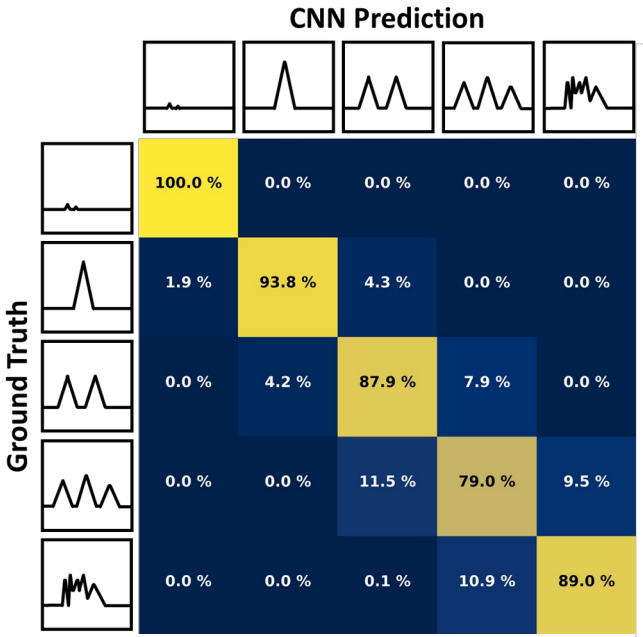
Quantization encourages parsimony

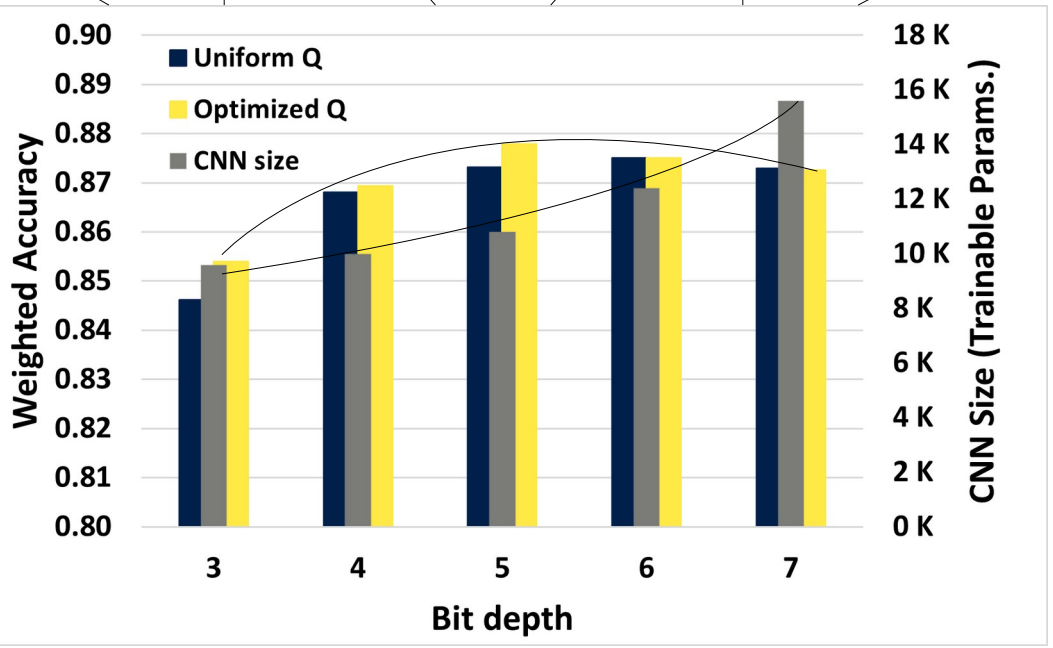
- Uniform quantization wastes token bits
- Reducing bits reduces input dimensionality
- Mitigates a universe of only corners

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









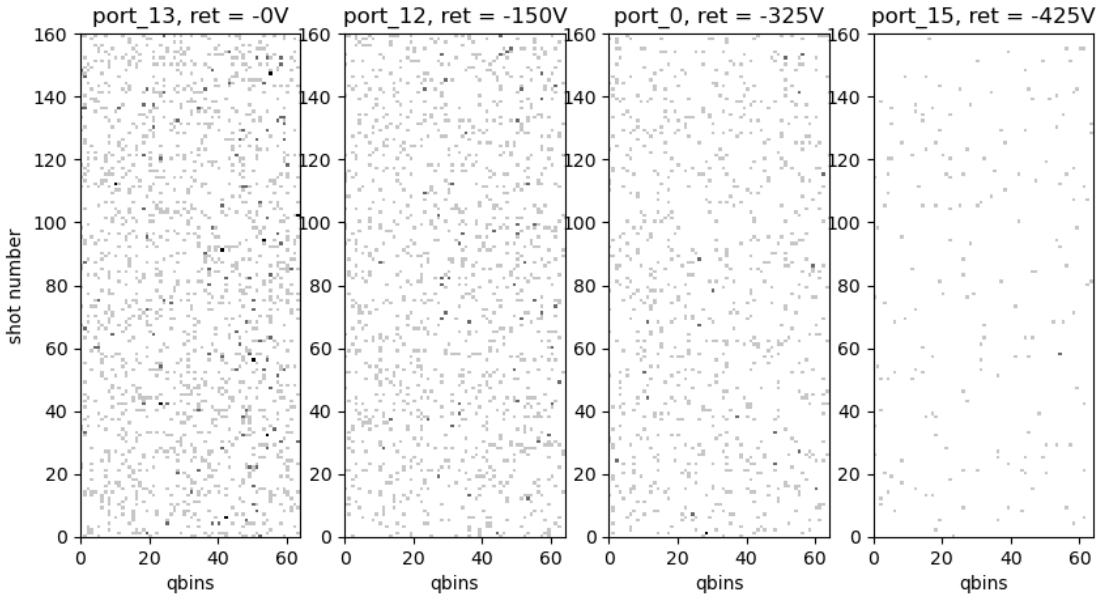
Smart Compression

Quantization

Think of quantization as a form of “tokenization”

Information is rotated into the Quantization Scheme, out of the raw data

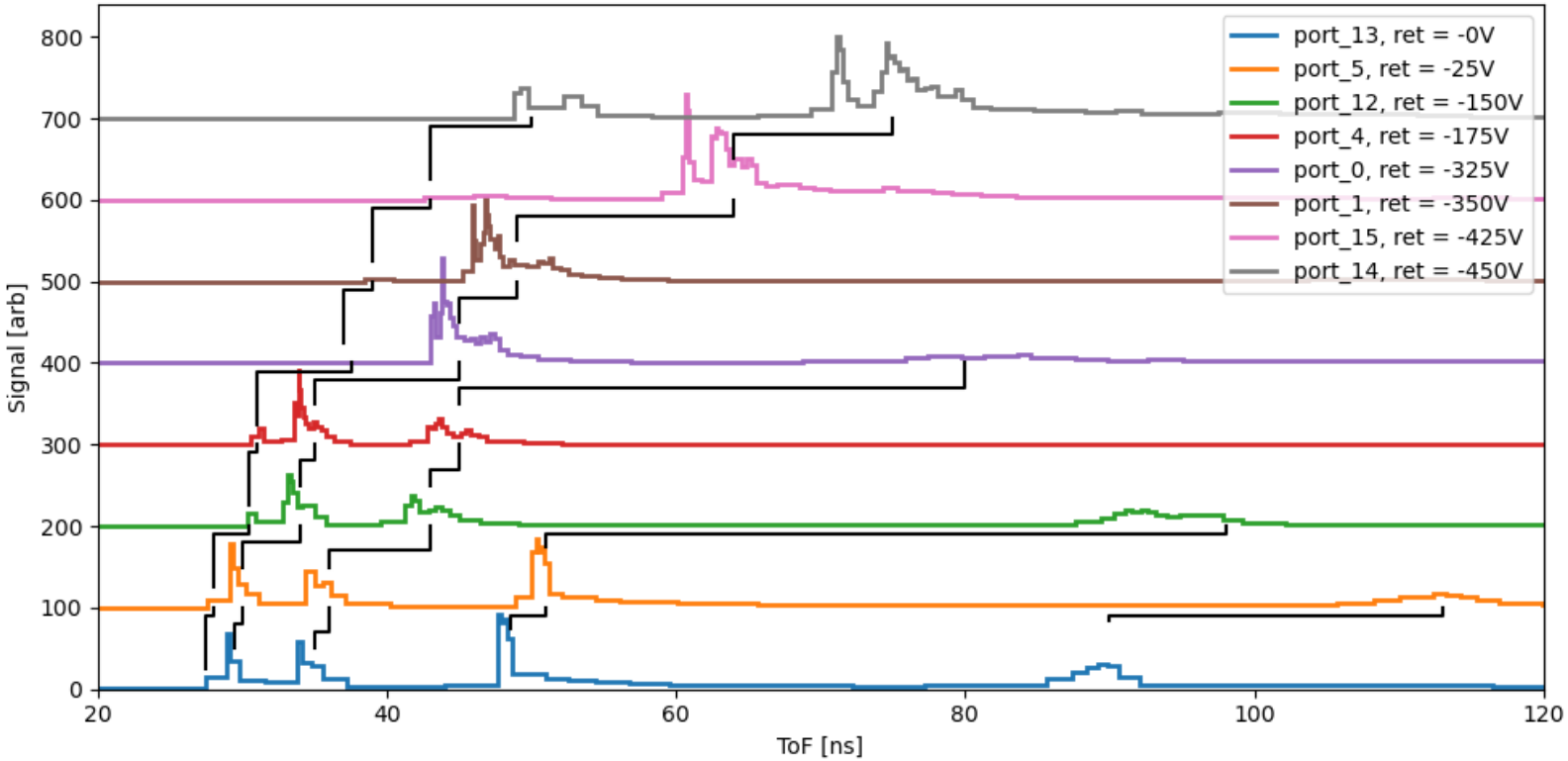
Easily implemented in EdgeAI/FPGA



Dimension Reduction

Mitigates the “Memory Wall” in getting data to ML training

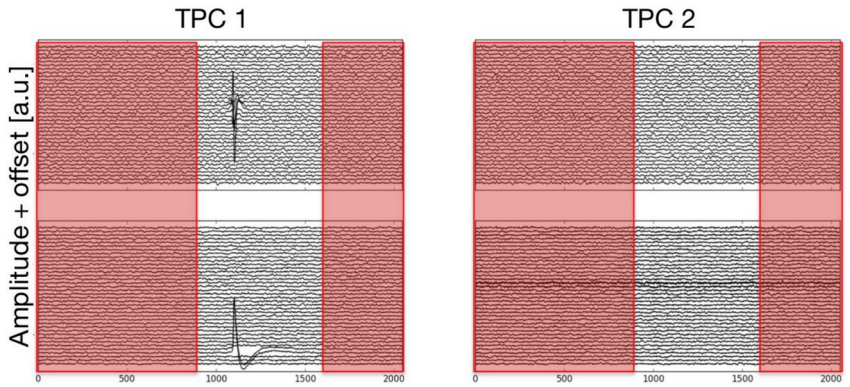
Alleviates the need for voluminous labeled datasets

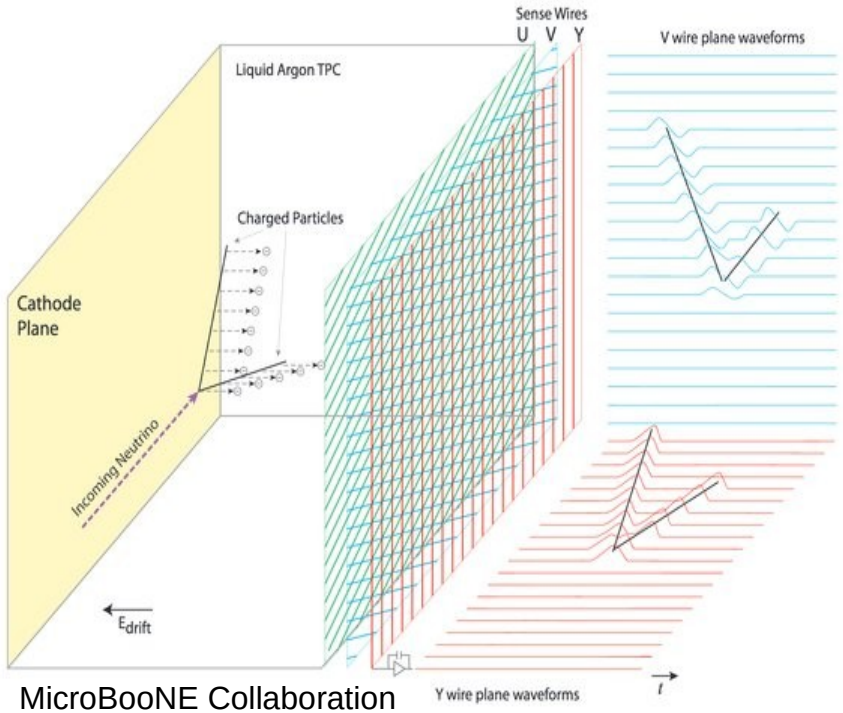


SLAC

Smart Compression

Reduction to basis-vector projections





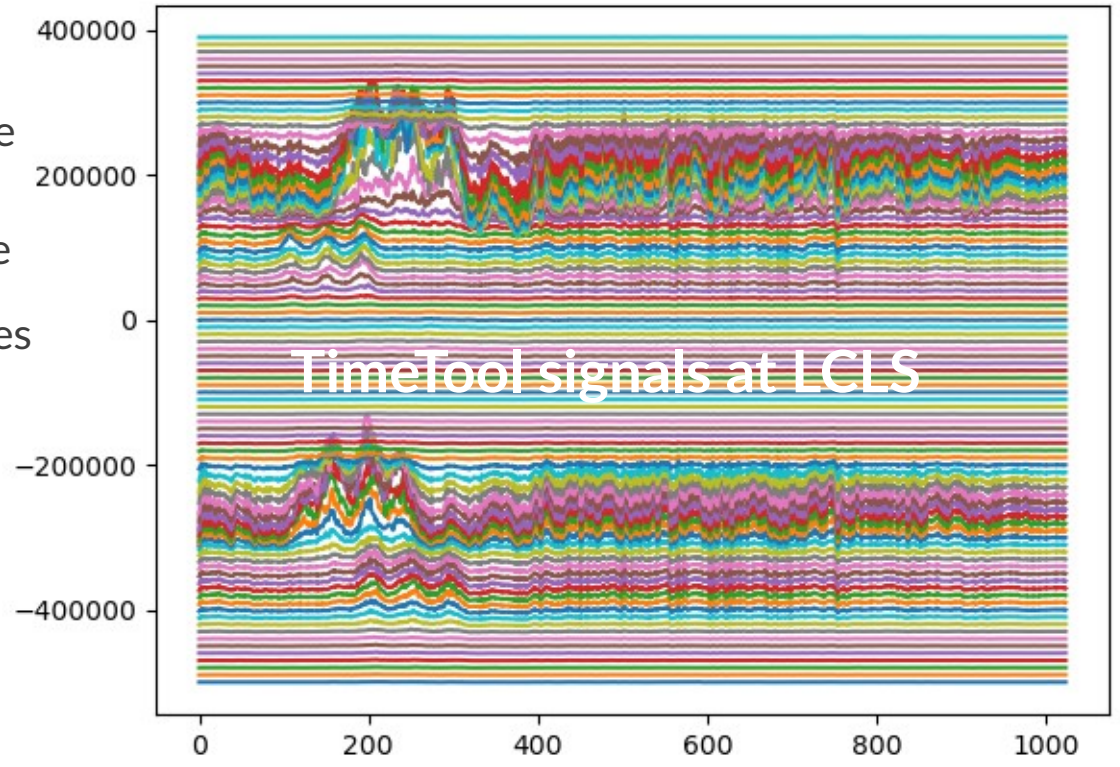
MicroBooNE Collaboration

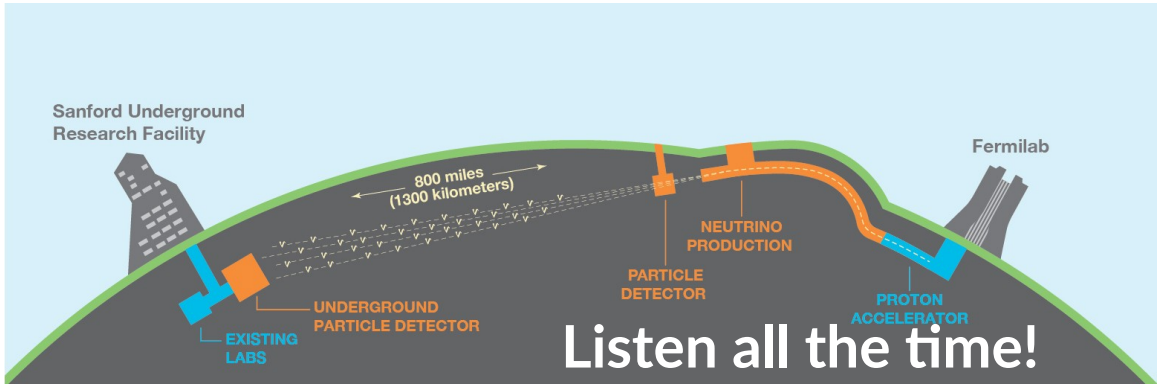
3D-TimeTool

Horizontal location of the feature gives fs-scale arrival time of x-ray pulse

Inter-row correlation gives x-ray spatial mode

Results inform real-time sorting/veto for downstream detectors





Smart Compression

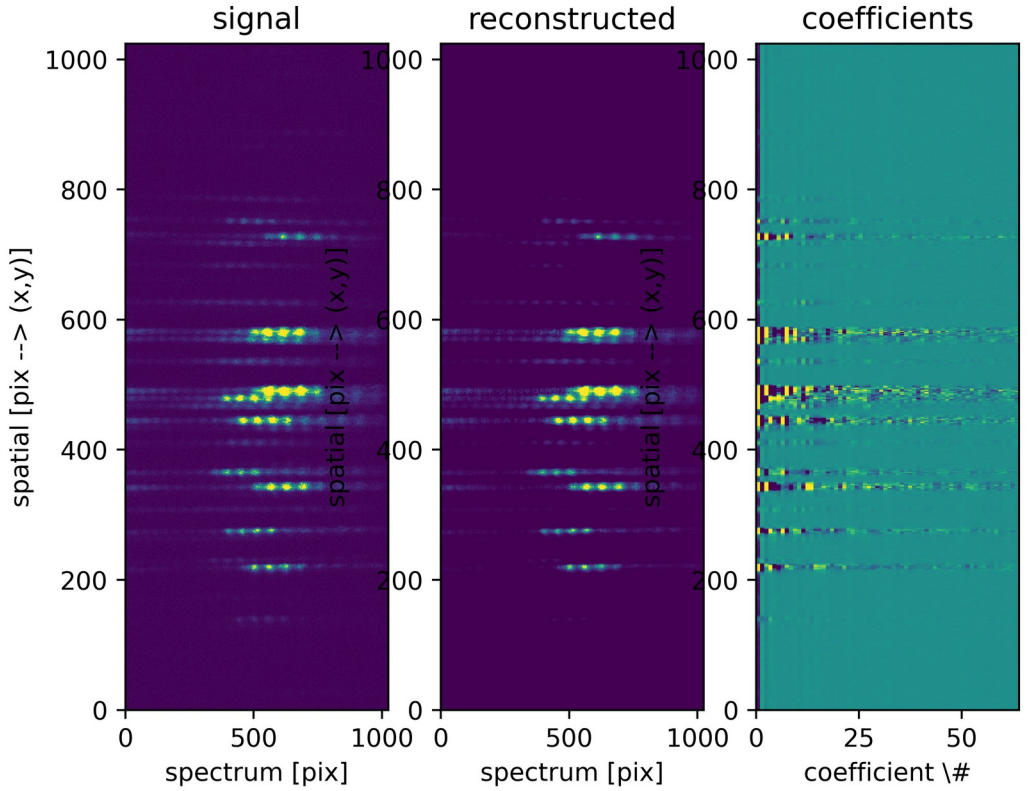
Reduction to basis-vector projections

ToF-PET

PET bi-photons are stochastic

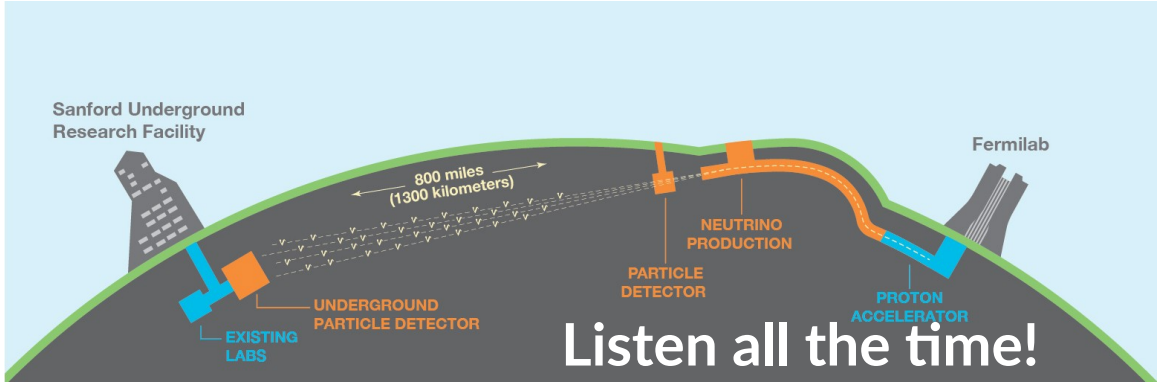
Laser and camera must be “always watching”

Event-rate in the few to tens kHz range.



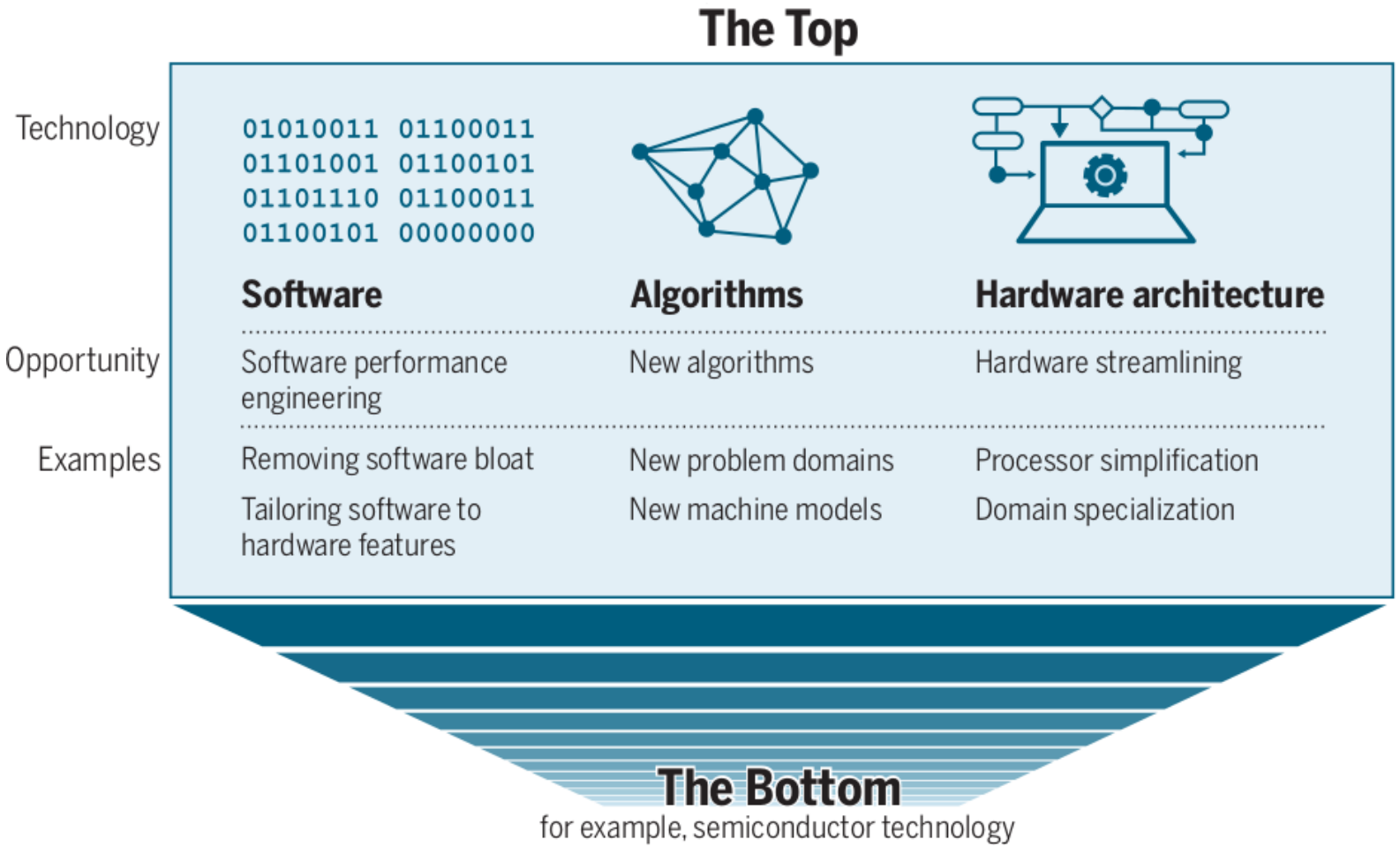
FORZA VISION RESEARCH AMETEK

SLAC



Edge Inference with Heterogeneity

By “inference” I mean “streaming data processing along the pipeline”



Leiserson *et al.*, “There’s plenty of room at the top:...” Science **368** 6495 (2020) 10.1126/science.aam9744

Edge Inference with Heterogeneity… Federation?

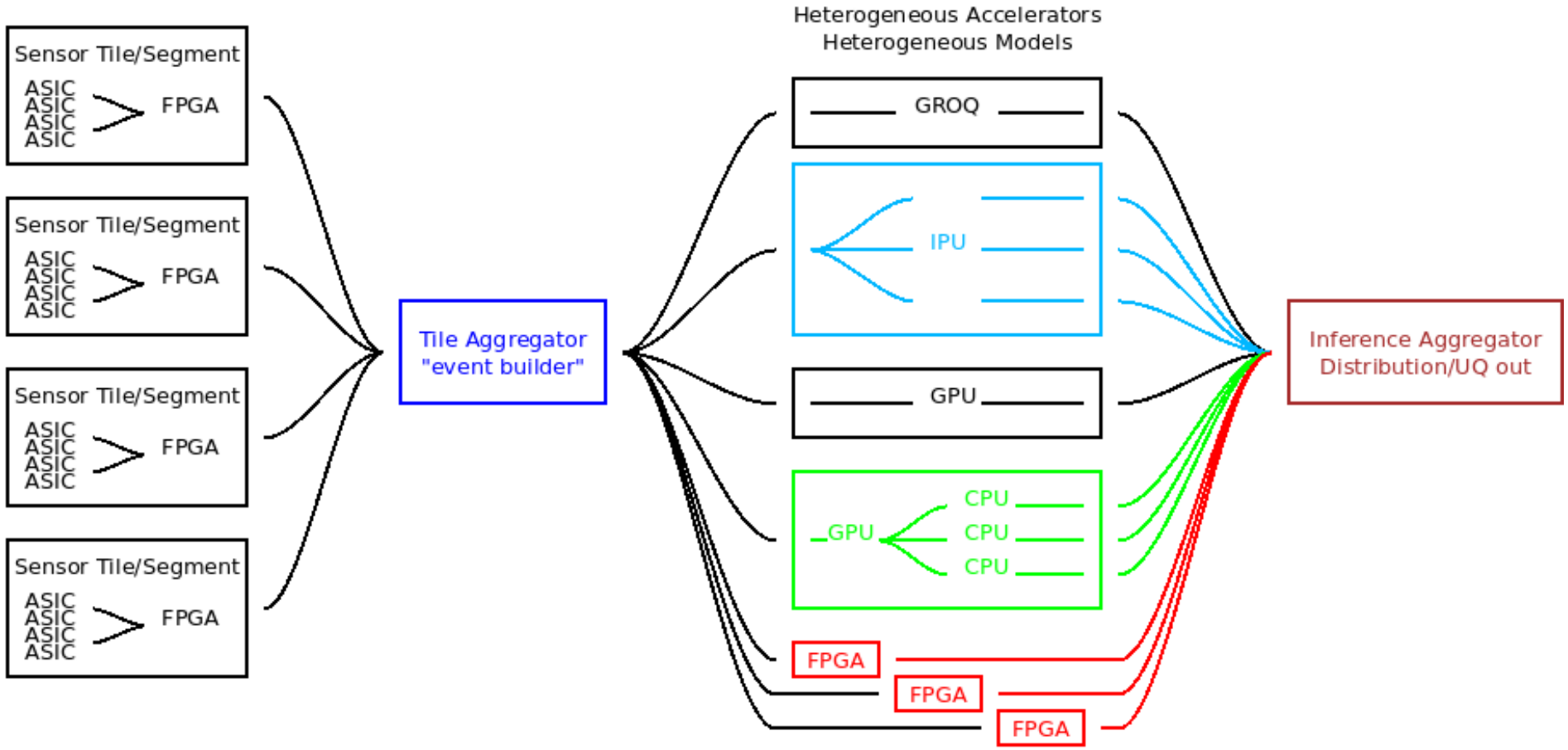
Orthogonal models need heterogeneous accelerators

Reduce multi-channel sparse to binary

Mixture-of-**ORTHOGONAL**-experts for honest UQ

Leverage different accelerators for a zoo of real-time models

Starting to look like Federated ML architectures



Edge Inference

Tokenization

Reduce multi-channel sparse vectors to binary arrays

Bit arrays look like ASCII characters

Pathological tokenization → common format for a Trillion Parameter Model Worth Wanting

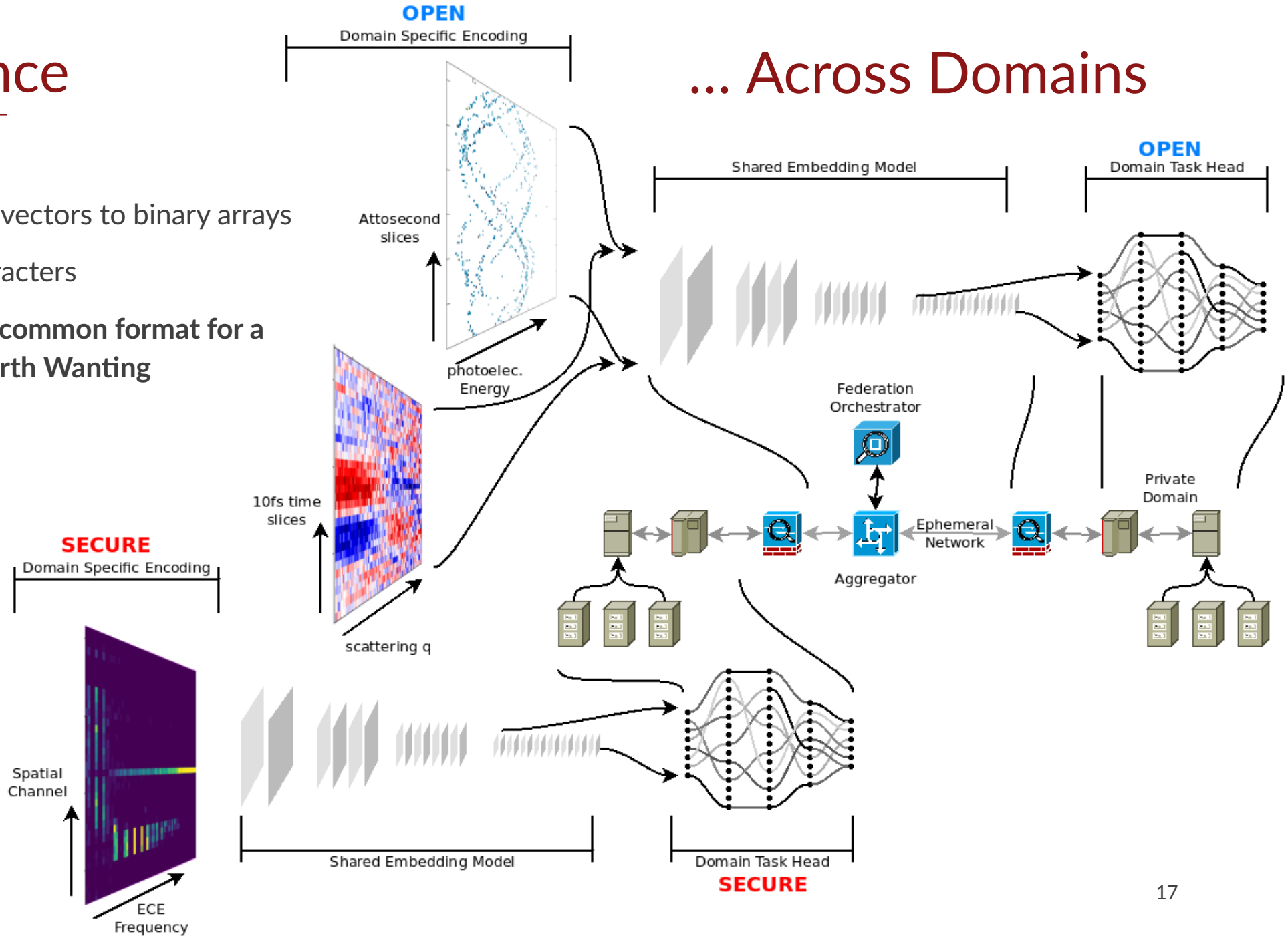
Sharing

Far fewer parameters for tuning with minimal labeled domain data

“Multi-lingual” sharing for improved generalizability

Great paradigm for Federated Resources

… Across Domains



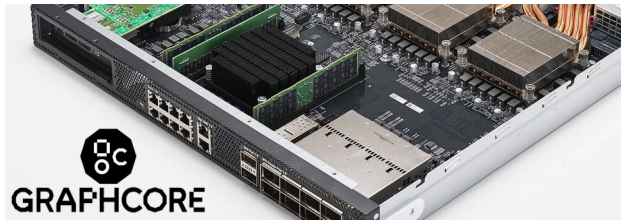
Federated Edge-to-HPC resources

Courtesy Greg Stewart
doi:10.1109/XLOOP54565.2021.00008

The heart of IRI might be in HPC,

… but its hands and feet are at the Edge

Let's Build… Let's Run!



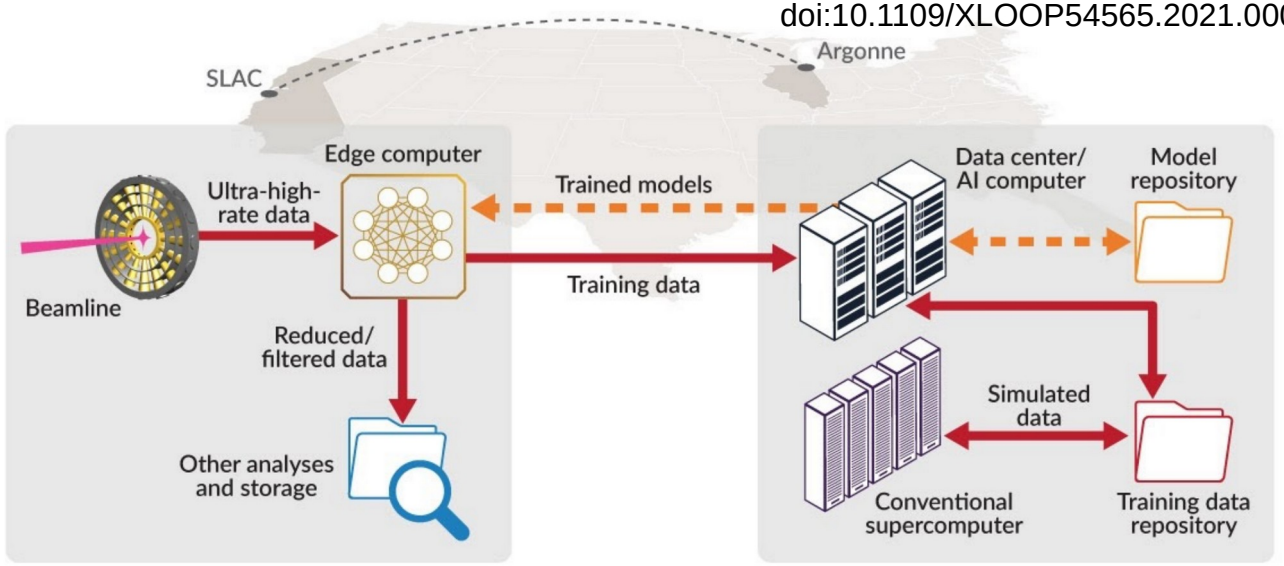
XILINX

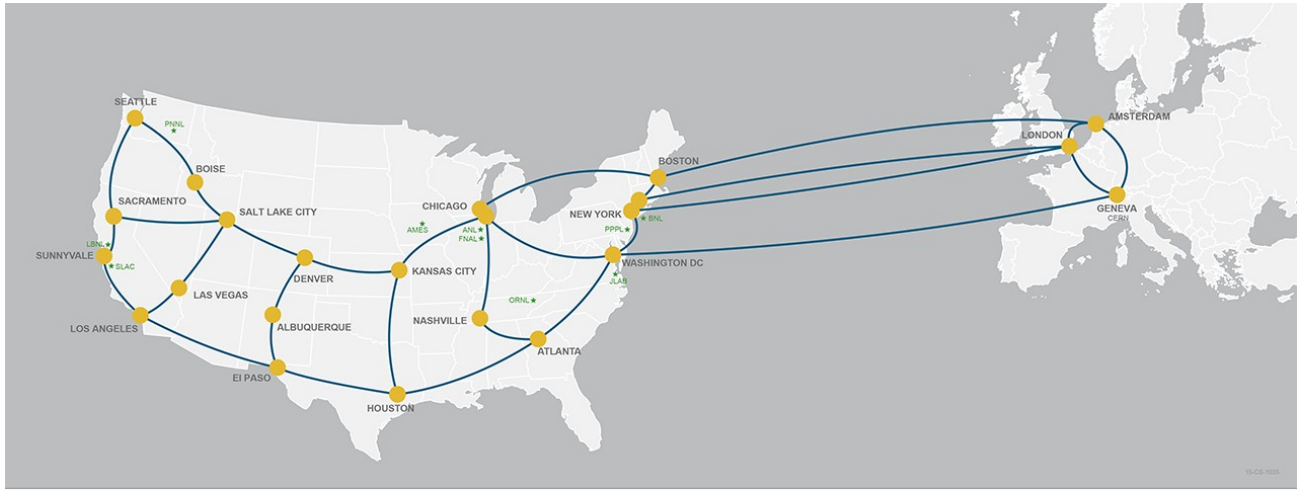
CORNAMI
Intelligent Computing

NVIDIA

SambaNova SYSTEMS

globus





ESnet
ENERGY SCIENCES NETWORK

★ Department of Energy Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)	LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ANL	Argonne National Laboratory (Argonne, IL)	ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
BNL	Brookhaven National Laboratory (Upton, NY)	PNNL	Pacific Northwest National Laboratory (Richland, WA)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)	PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)	SLAC	SLAC National Accelerator Laboratory (Menlo Park, CA)

Thanks

DOE BES – Eliane Lessner

CookieBox FWP-100498 PI Coffee
AISDC FWP-100643 PI Thayer

DOE FES – Matthew Lanctot

EdgeML Fusion FWP 100636 PI Kolemen
IFE Tammy Ma, *et al.*

SLAC

Omar Quijano, Abhilasha Dave, Ryan Herbst, Larry Ruckman, Jana Thayer, Amedeo Perazzo, ...

Enterprise Neurosystem

Bill Wright, Leo Hoarty, Sanjay Aiyagari, John Overton, Dinesh Verma, Erik Erlandson, Christine Payne, Dennis O'Connell, ...

Argonne

Zhengchun Liu, Ahsan Ali, Dennis Trujillo, Ian Foster, and the AISDC team

Oak Ridge

Tom Beck, Ben Mintz, Rob Moore, Rick Archibald, Paul Laiu, ...

Private Sector

Cornami Inc.: Mache Creeger, Deepika Natarajan; **APN Inc.:** Antonio Ransom, Jim Herbert, Harvey Rubin; **IsAdvice Inc.:** Pamela Isom

All of us who are actively building an Integrated National AI Infrastructure that advances rapid injection of human creativity into a securely and maturely shared environment. The boots on the ground in computing will make or break how we adapt to our evolving world.

We acknowledge: Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.
DIII-D, the largest magnetic fusion user facility in the U.S., is a tokamak confinement device with significant engineering flexibility to explore the optimization of the advanced tokamak approach to fusion energy production.

Thank You

We will only solve our challenges together…

energy – security – agriculture

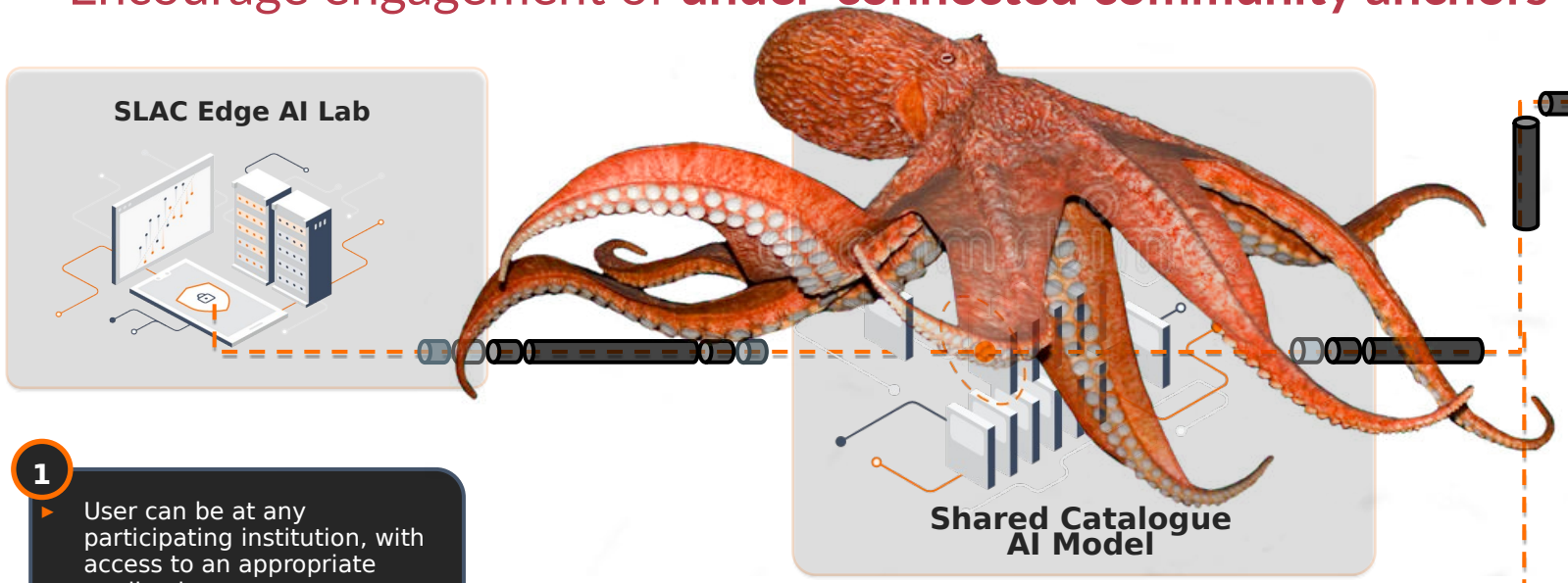
coffee@slac.stanford.edu

A “OneDOE” Octopus – we (DOE) are the trusted party

Secure (ephemeral) federation of data and models

Make highest use of national data and computing resources

Encourage engagement of **under-connected community anchors**



1

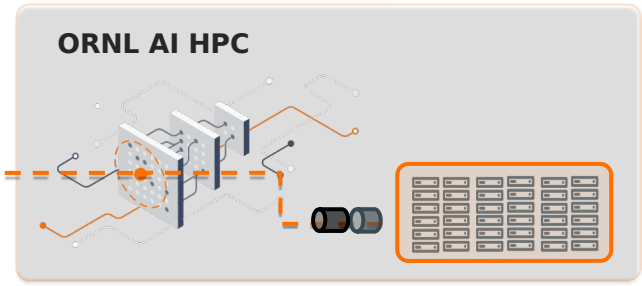
- User can be at any participating institution, with access to an appropriate application
- Since the system connects applications, not hosts, the application determines what runs, not the host
- In this way, expensive hardware is better leveraged for critical use

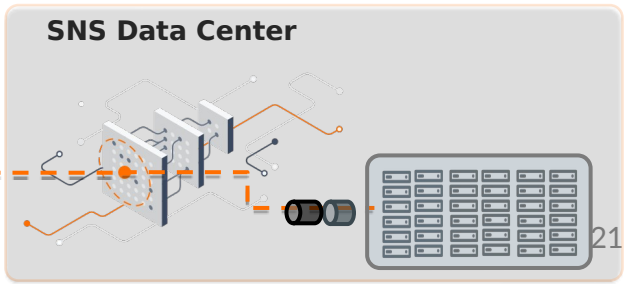
2

- Virtual catalog of all the models from participating institutions
- Any model can be executed on the data of any participating institution

3

- Model is pushed to the hardware of the institution that has the data, where it is executed
- Data is not shared to any central point, maintaining ownership and security
- The same model can be executed on multiple infrastructures, should it be necessary (i.e. different data sets)





SLAC

Courtesy Sanjay Aiyagari – Red Hat