



# **The Instrumental Edge: Enabling Real Time AI Scientific Discovery**

Adam Thompson | Principal Technical Product Manager | NVIDIA

Abhik Sarkar | Staff Computer Scientist | Lawrence Livermore National Laboratory

Luigi Cruz | Staff Engineer | The SETI Institute

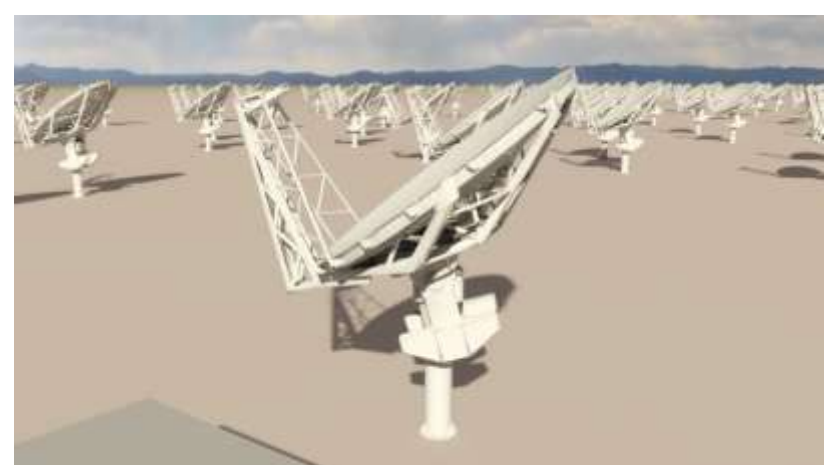


# Next Generation Instruments are Producing Increasing Amounts of Data

Complexity of Experiments is Booming and Human Insight is Now the Bottleneck

## Radio Astronomy

ngVLA – 244 Dishes  
100 Petabytes per Year



SKA – 200 Dishes  
1 Exabyte per Year

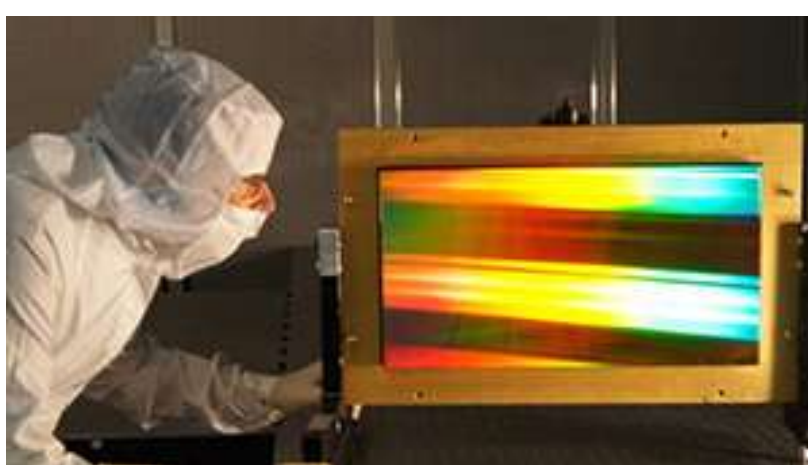


## Particle Physics

High Luminosity LHC  
100x Data than Higgs Boson Discovery



Advanced Laser Systems  
100x Increase in Repetition Rate



## Light Sources

APS-U – >60 beamlines  
100-200 Petabytes per Year



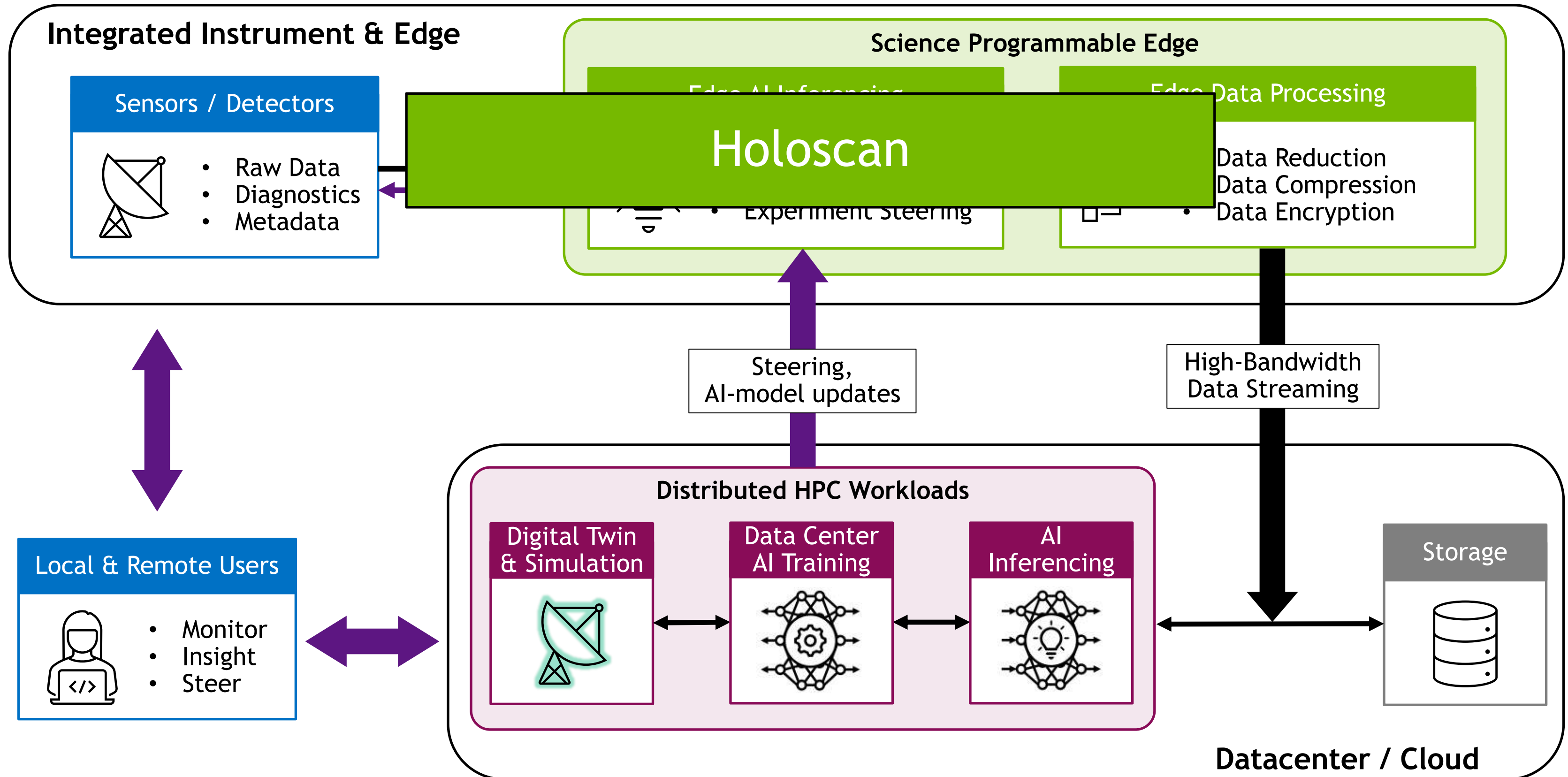
Free Electron Laser LCLS-II  
1 MHz Rep-Rate Upgrade





# Future Instruments Will Combine Edge and Distributed HPC Workloads

A Vision of Towards Integrated Ecosystem for Self-Driving Experiments and Laboratories



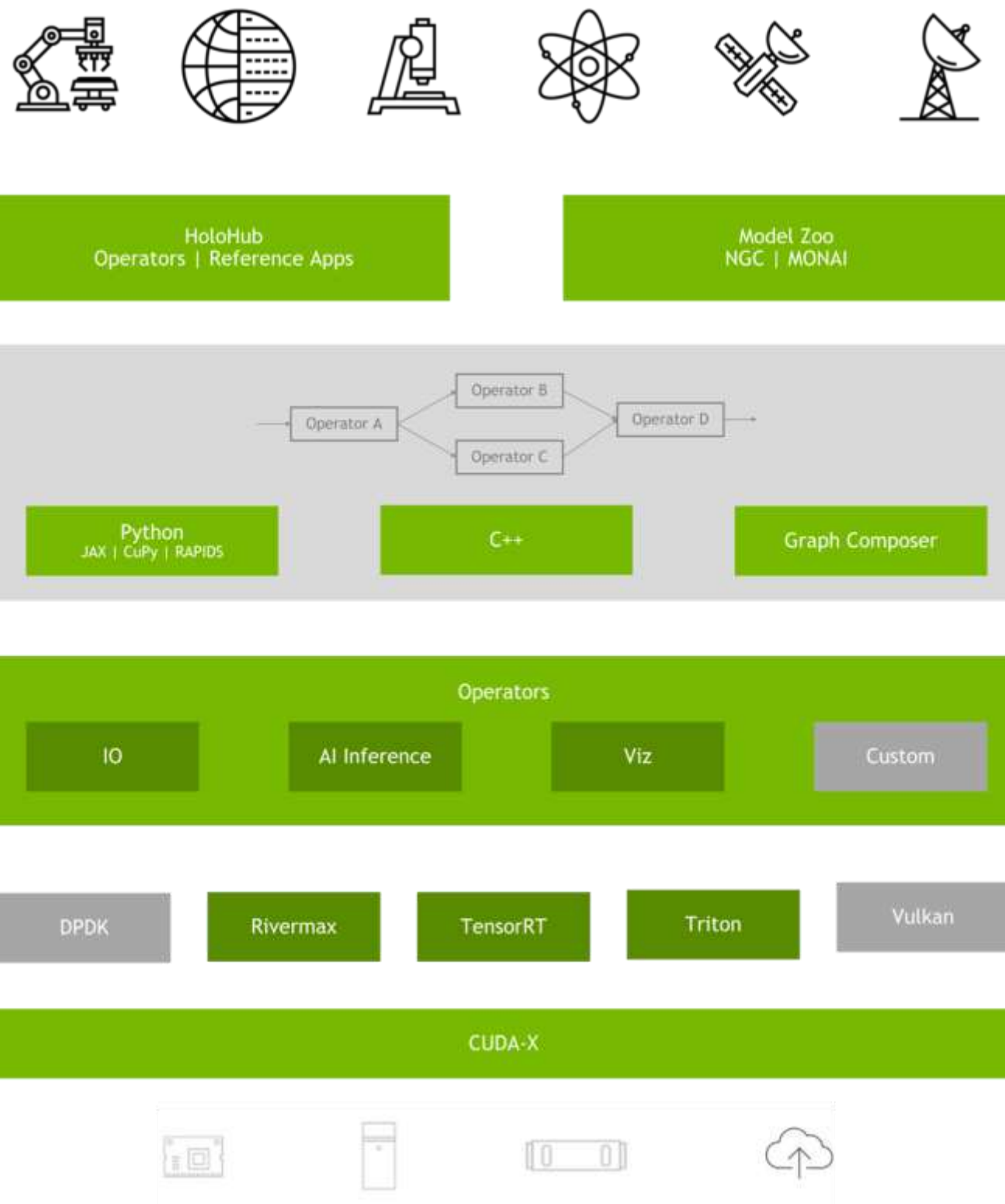




GitHub

# NVIDIA Holoscan

SDK for Building AI-Enabled Sensor Processing Applications



## Features

- C++ and Python APIs for **domain agnostic** sensor data processing workflows
- Multi-Node and Multi-GPU support with advanced pipeline scheduling options and network-aware data movement
- AI Inference with pluggable backends such as ONNX, Torchscript, and TensorRT
- Scalable from Jetson Orin Nano (ARM + GPU) to DGX (x86 + H100)
- Apache 2 Licensed and Available on [GitHub](#)

## Benefits

- Simplifies sensor I/O to GPU
- Simplifies the performant deployment of an AI model in a streaming pipeline
- Provides customizable, reusable, and flexible components to build and deploy GPU-accelerated algorithms
- Scale workloads with Holoscan's distributed computing features
- Deploy to the Cloud with Holoscan App Packager and Kubernetes



# Sidekick-Systems for High repetition rate laser experiments

---

Abhik Sarkar, Lawrence Livermore National Lab

NVIDIA GTC

March 17, 2025

LLNL-PRES-872790

Work performed under the auspices of the U.S. Department of Energy (DOE) by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and funded by the LLNL LDRD program under tracking code 23-ERD-035.





The National Ignition Facility is the world's largest and most energetic laser



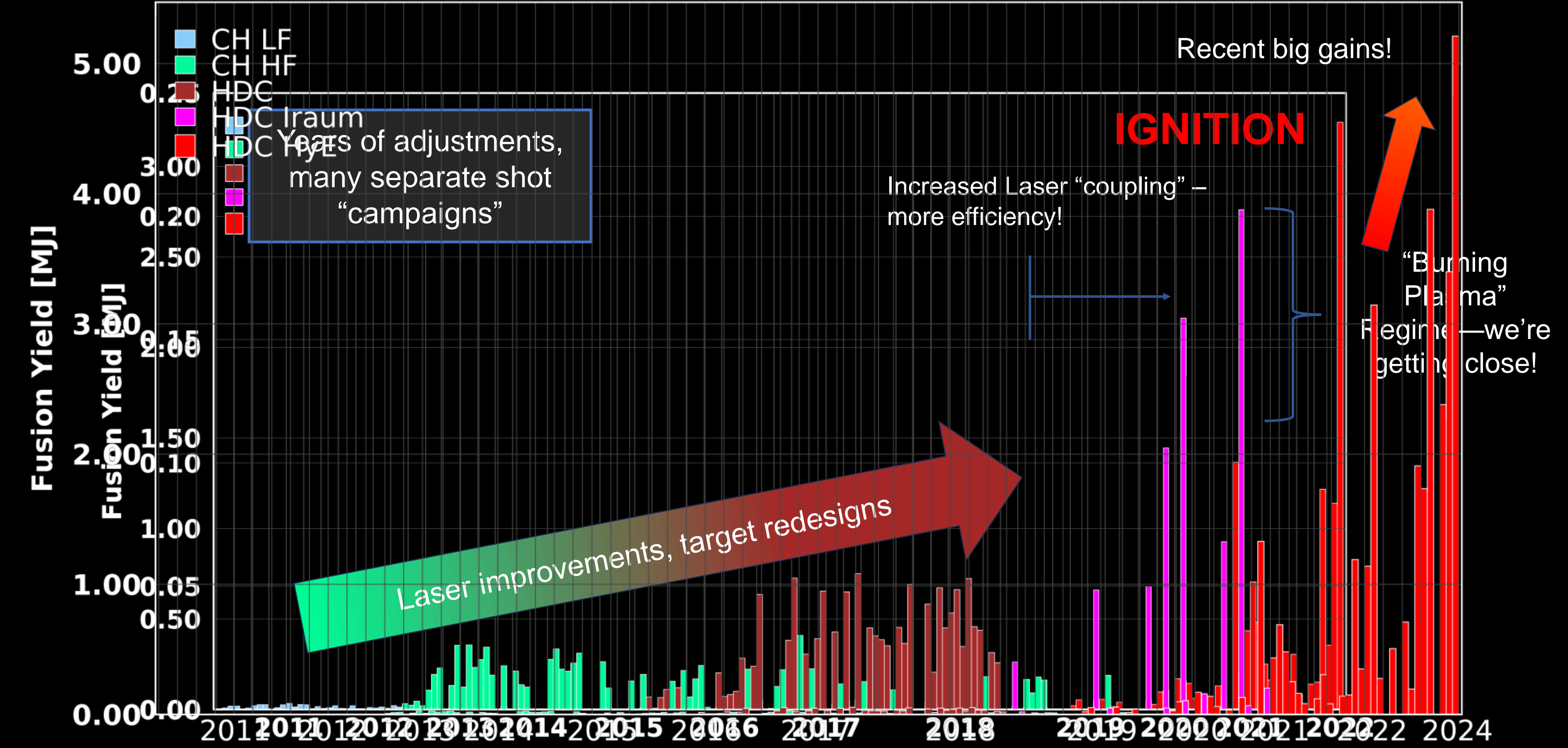


2 Mega-Joules of laser energy from 192 beams is delivered to centimeter-scale targets in a few nanoseconds for fusion experiments



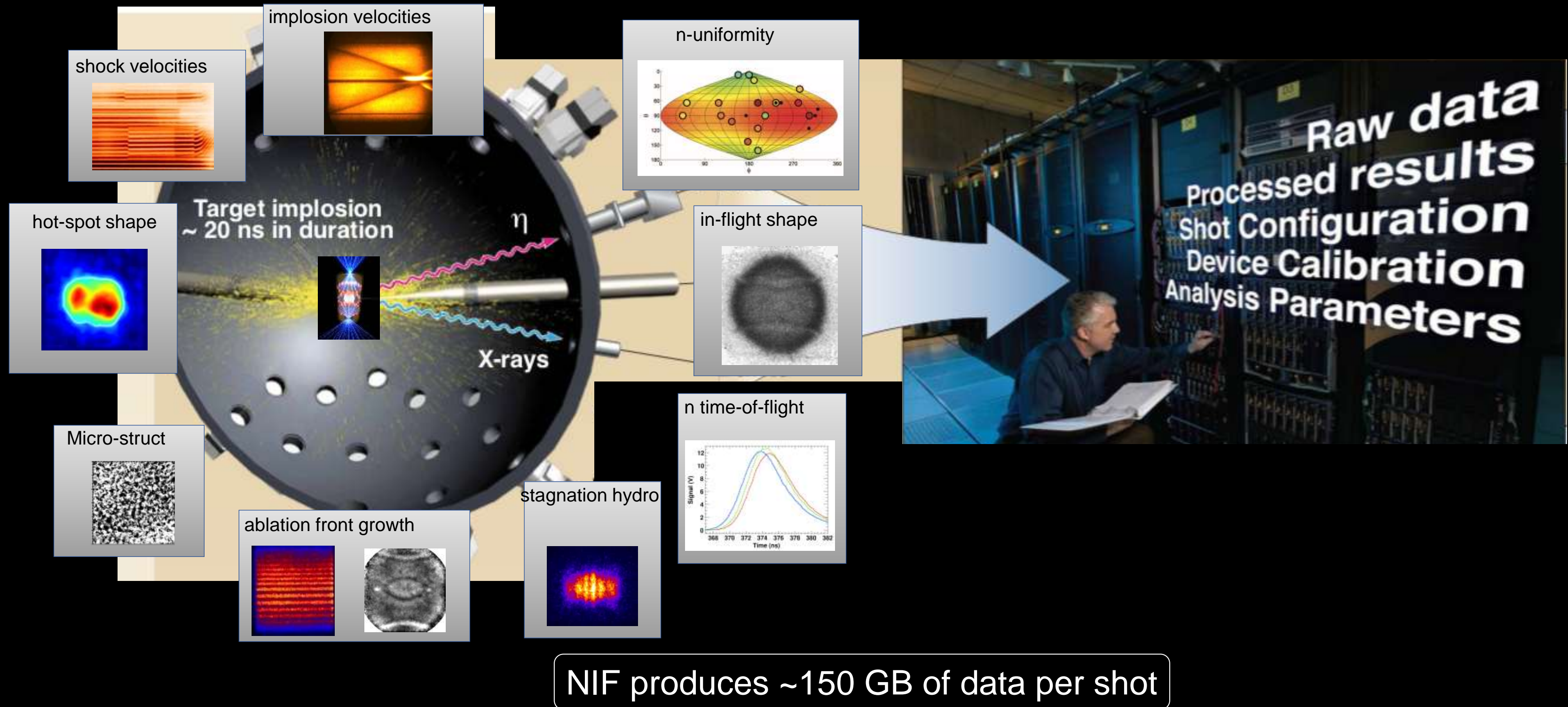


# It took around a decade of focused science to achieve ignition on NIF





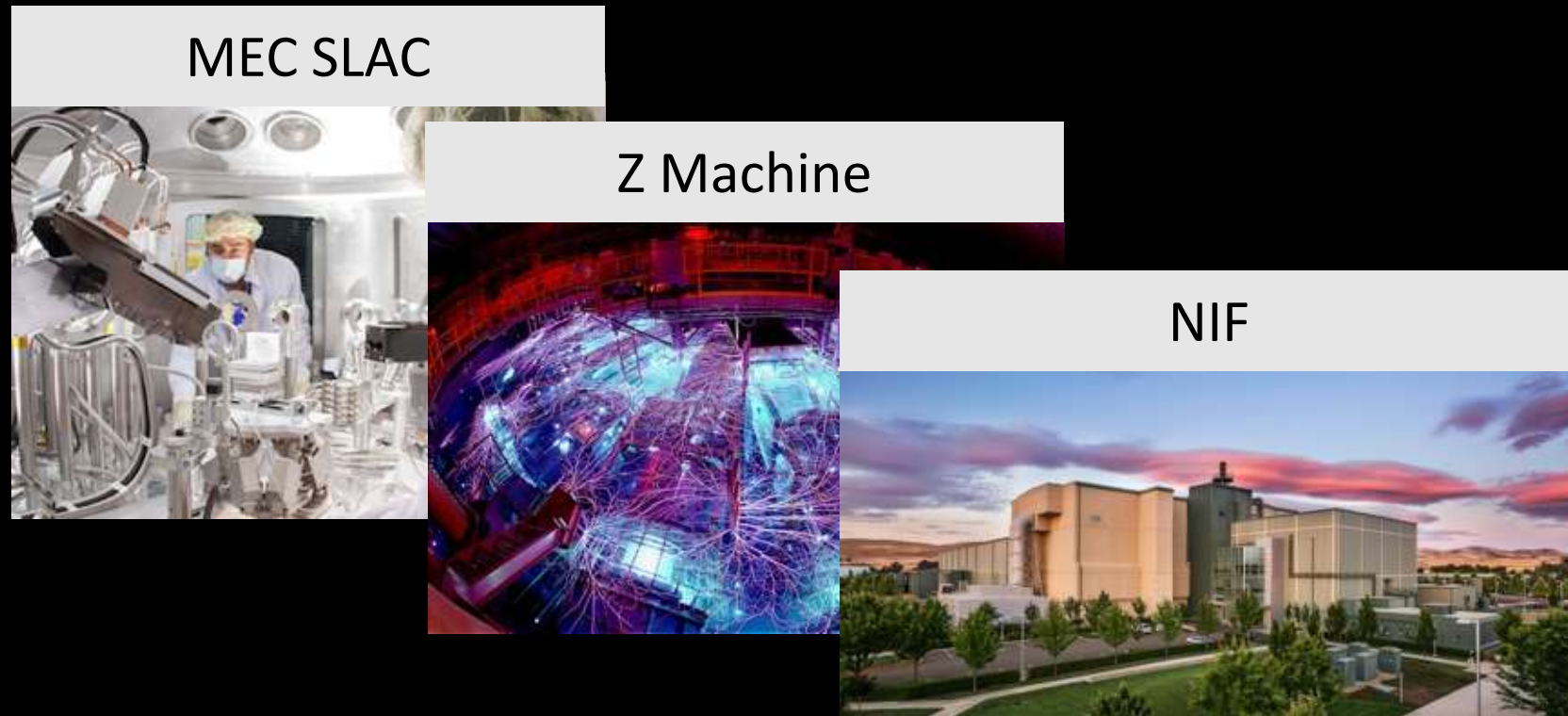
# Processing of large volumes of complex laser and resulting physics data will be crucial to autonomous laser-driven science





# Coupling high repetition rate laser with AI will enable a new regime of HED physics

Currently we make use of some of the premier laser scientific facilities around the US and the world to conduct forefront HED science



Until recently, much of HED has focused on large, energetic drivers that are mostly single-shot (~shot per hour)

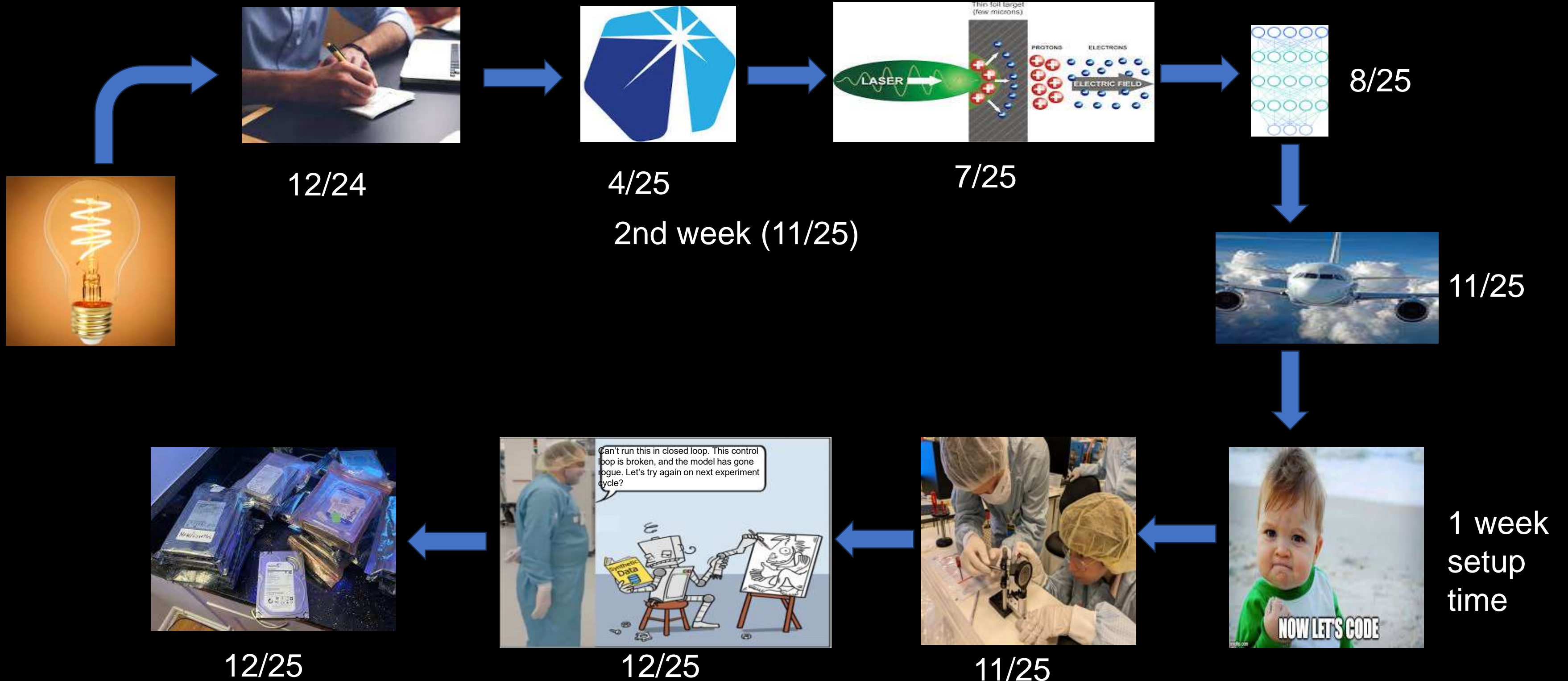
However, next-generation lasers coming online are already rep-rated ( $>10$  Hz)



We now have to shift paradigms, combining multiple emerging technologies with cognitive simulation to harness the possibilities of autonomous discovery

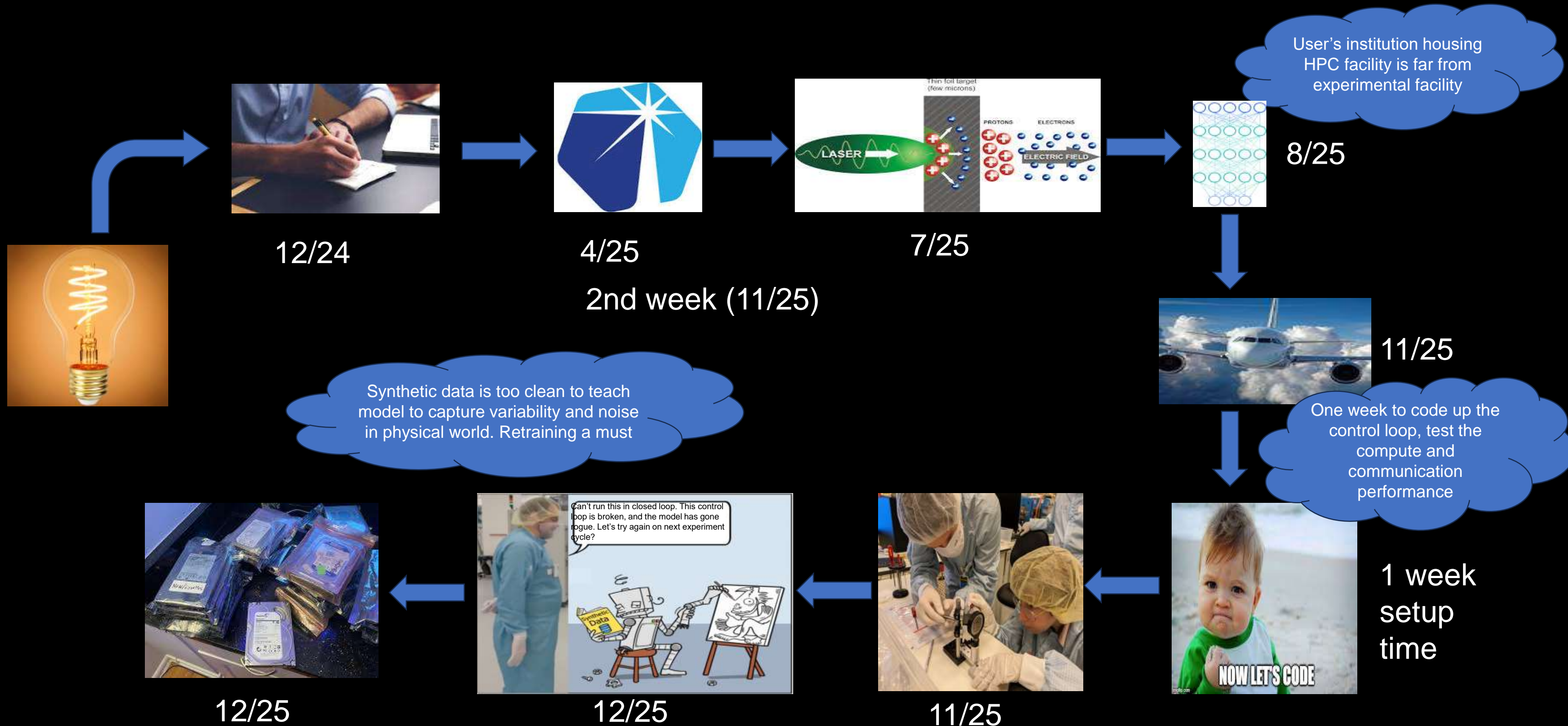


# Current AI-Driven High Repetition Rate HED Laser Experiment Workflow (rough timeline)



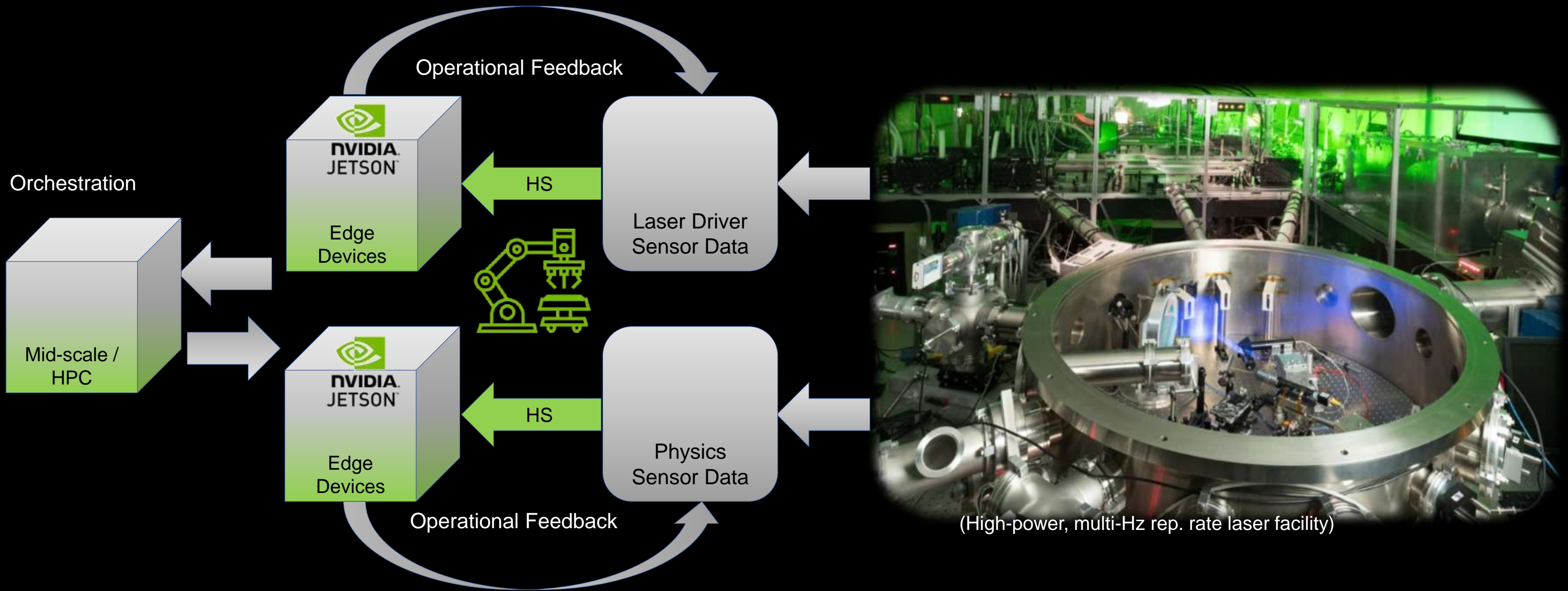


# Slowness of Scientific Discovery, why?





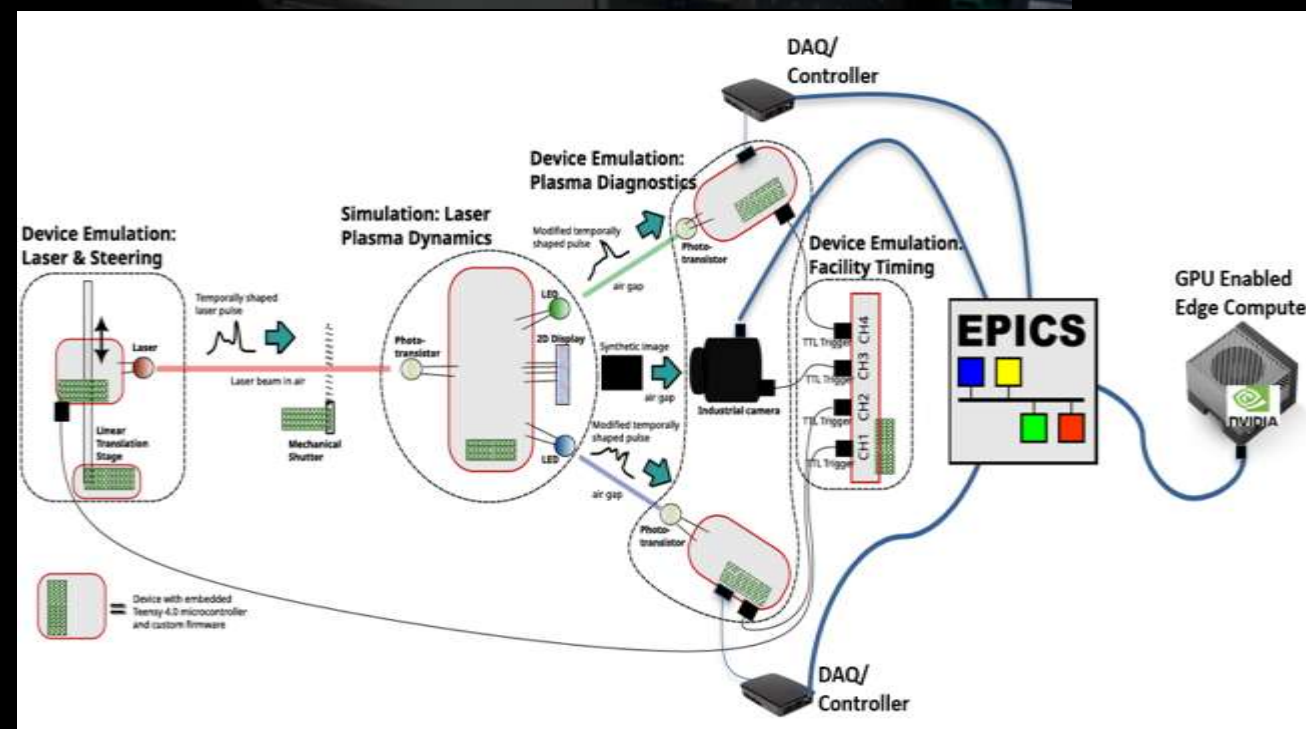
# We are working to integrate edge hardware throughout facilities to meet the low-latency demands of intelligent multi-Hz operations





# Sidekick Systems: Hardware-in-loop

Sidekicks mirror world-class facilities  
enabling development of  
robust AI self-driving playbooks





# Sidekick Systems: Hardware-in-loop

Sidekicks mirror world-class facilities enabling development of robust AI self-driving playbooks



Robust AI playbooks are deployed at world-class facilities to unlock self-driving mode

## Sidekick delivers ...

- The identical EPICS control system and environment
- Real experimental detection and high-speed timing *beyond* facilities
- Real-world noise and variability
- Live physics models adjustable from analytical to powerful AI surrogates
- Edge-based computing with microprocessors or GPUs
- LIVE testbed that mirrors its world-class sibling for self-driving dev
- 24/7 remote operation without occupancy of \$10M-\$1B machines

EPICS

THE EXPERIMENTAL PHYSICS AND INDUSTRIAL CONTROL SYSTEM

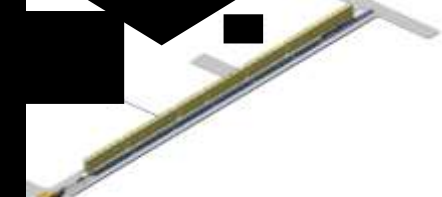
## ELI Laser Facility



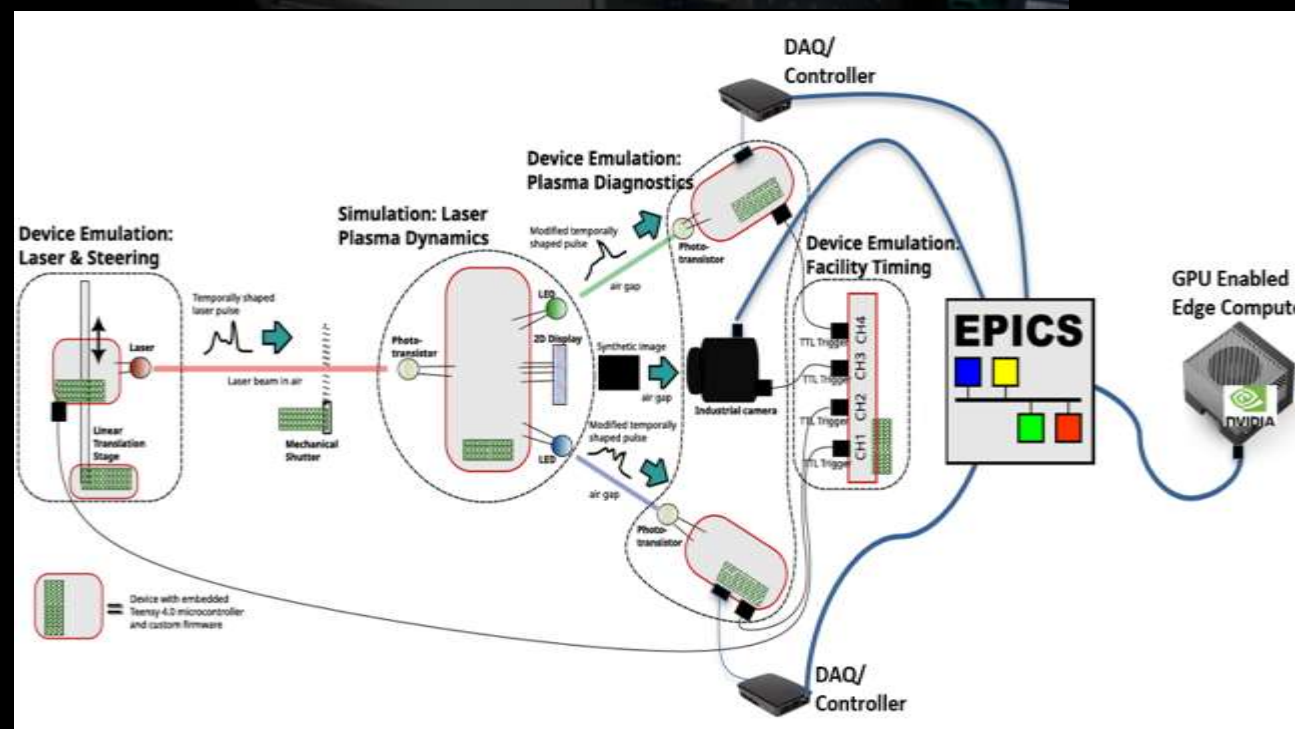
## Aleph Laser Facility



## Scorpius Accelerator

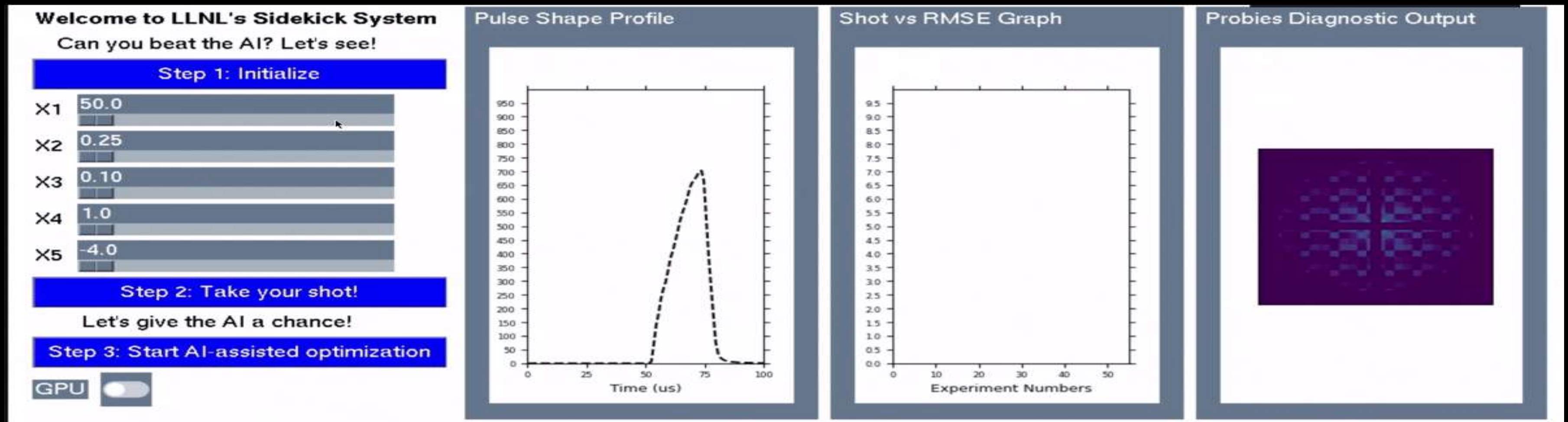


## Advanced Manufacturing





# The Sidekick can perform closed-loop experimental optimization



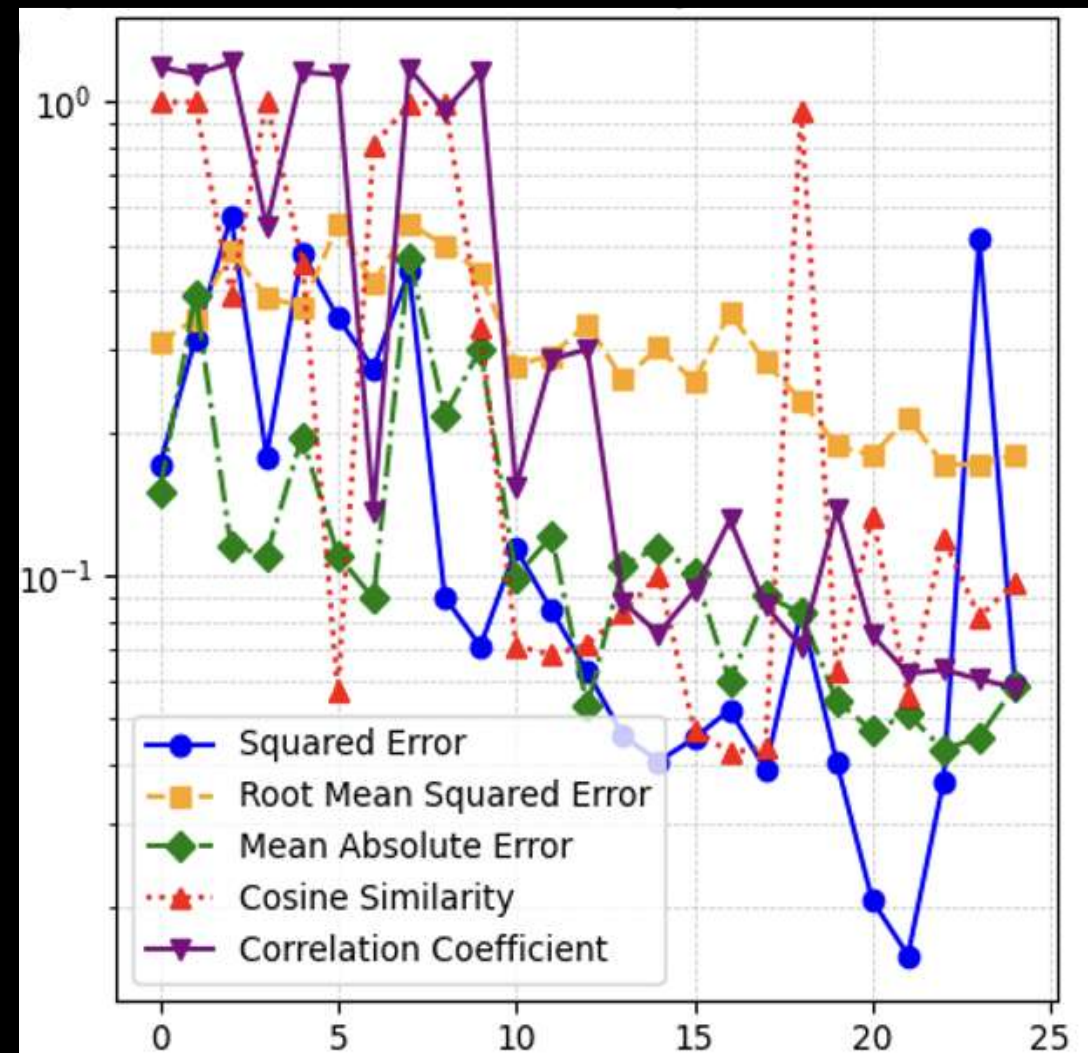
Analogous experiments generate synthetic diagnostic data and quickly optimize



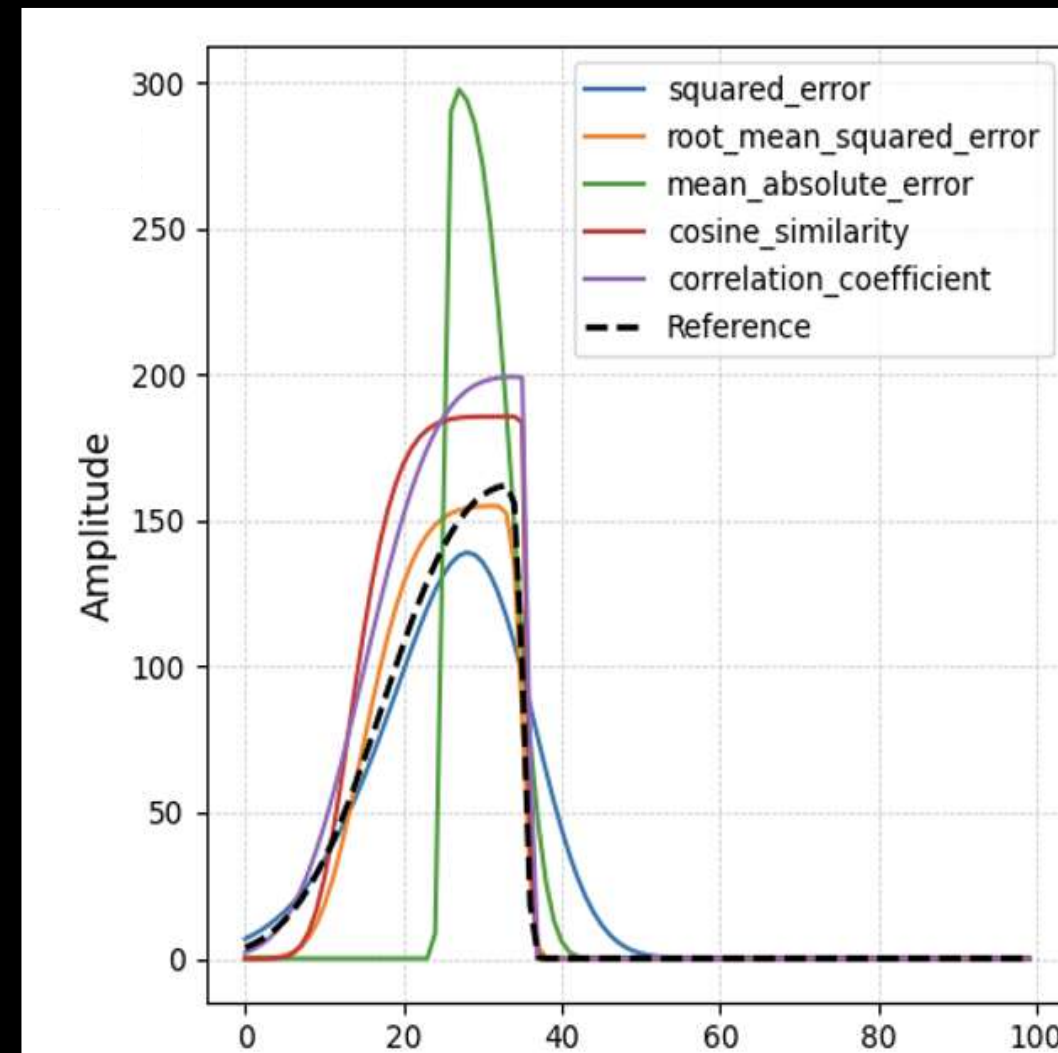
# Choosing optimization hyperparameters through in-silica experiments can be misleading

## Choosing Optimization Objective Function

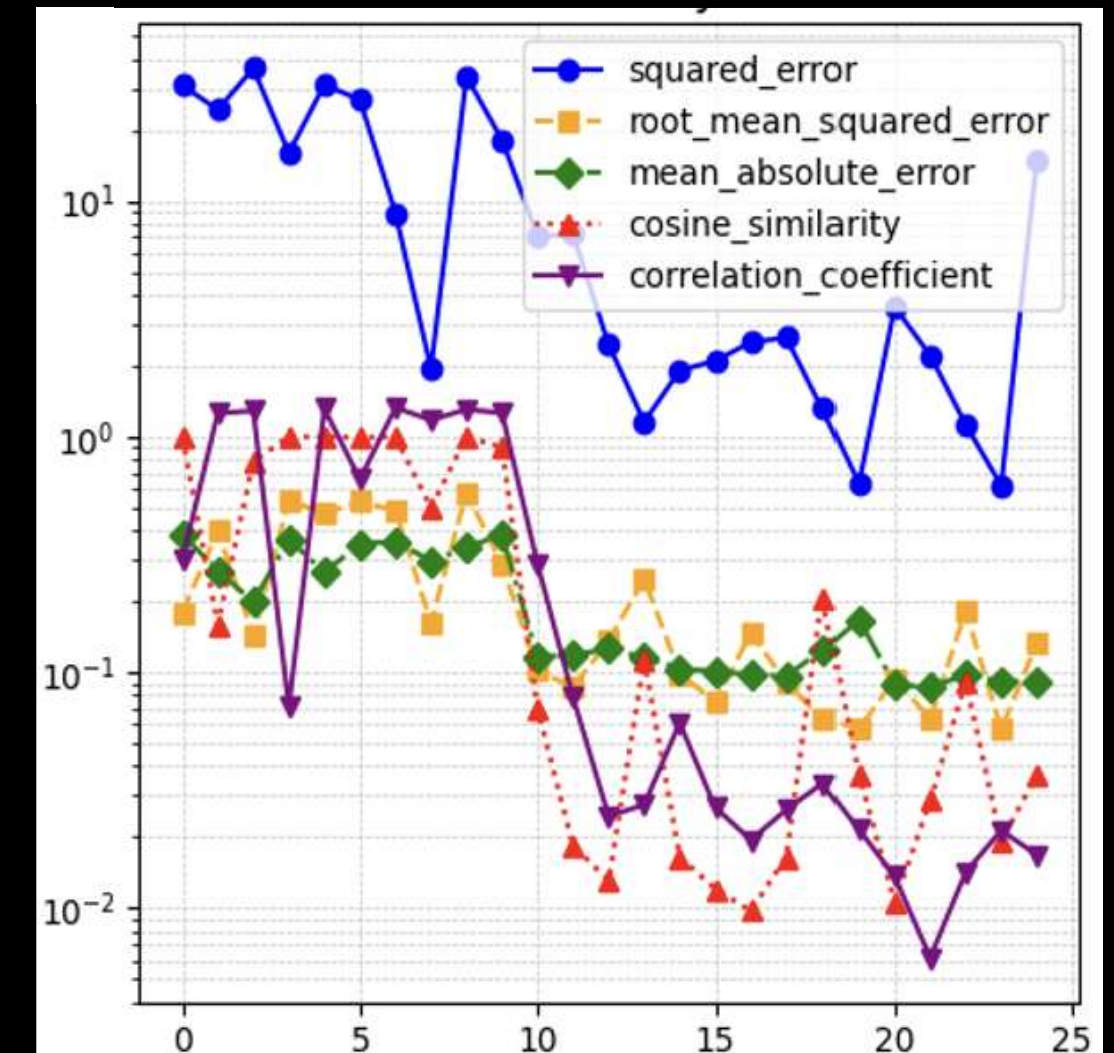
In-silica



Best Solution (Hardware-in-loop)



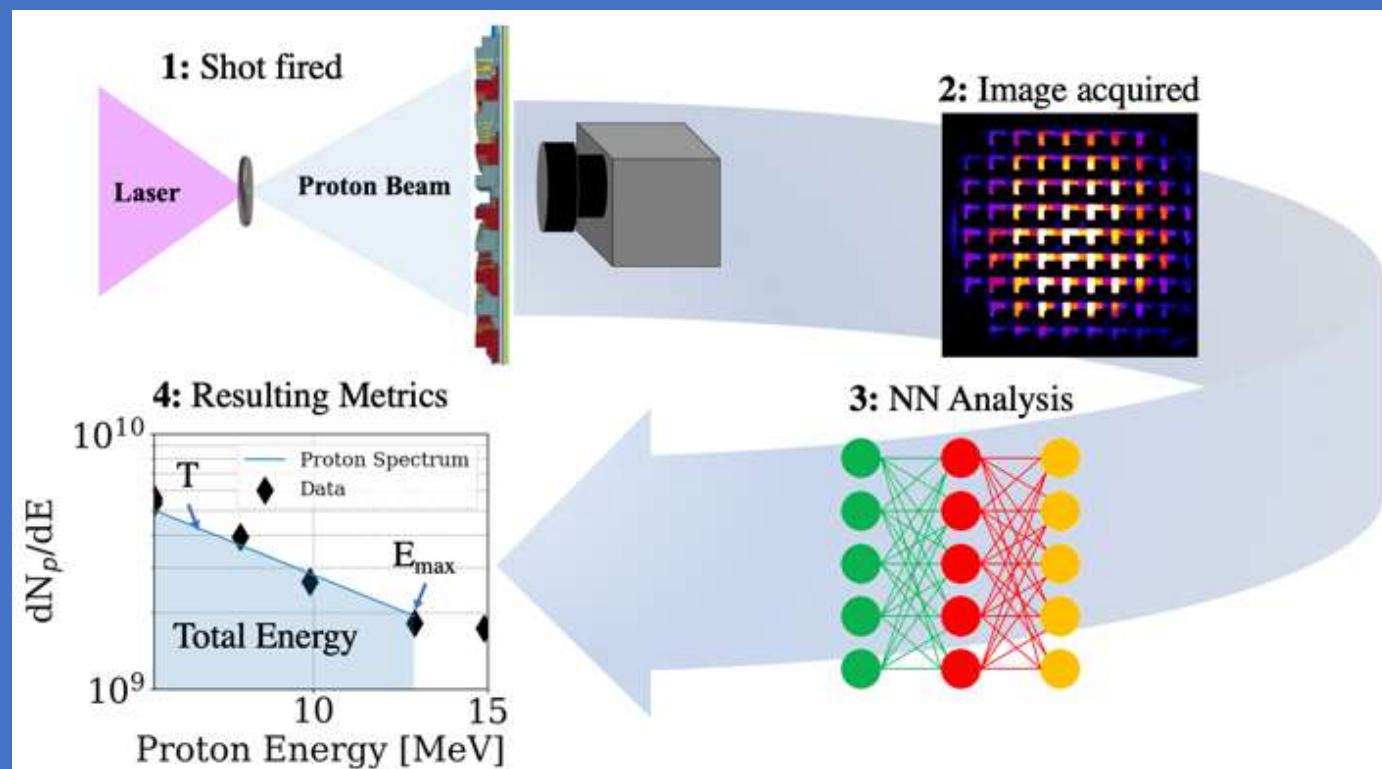
Hardware-in-loop



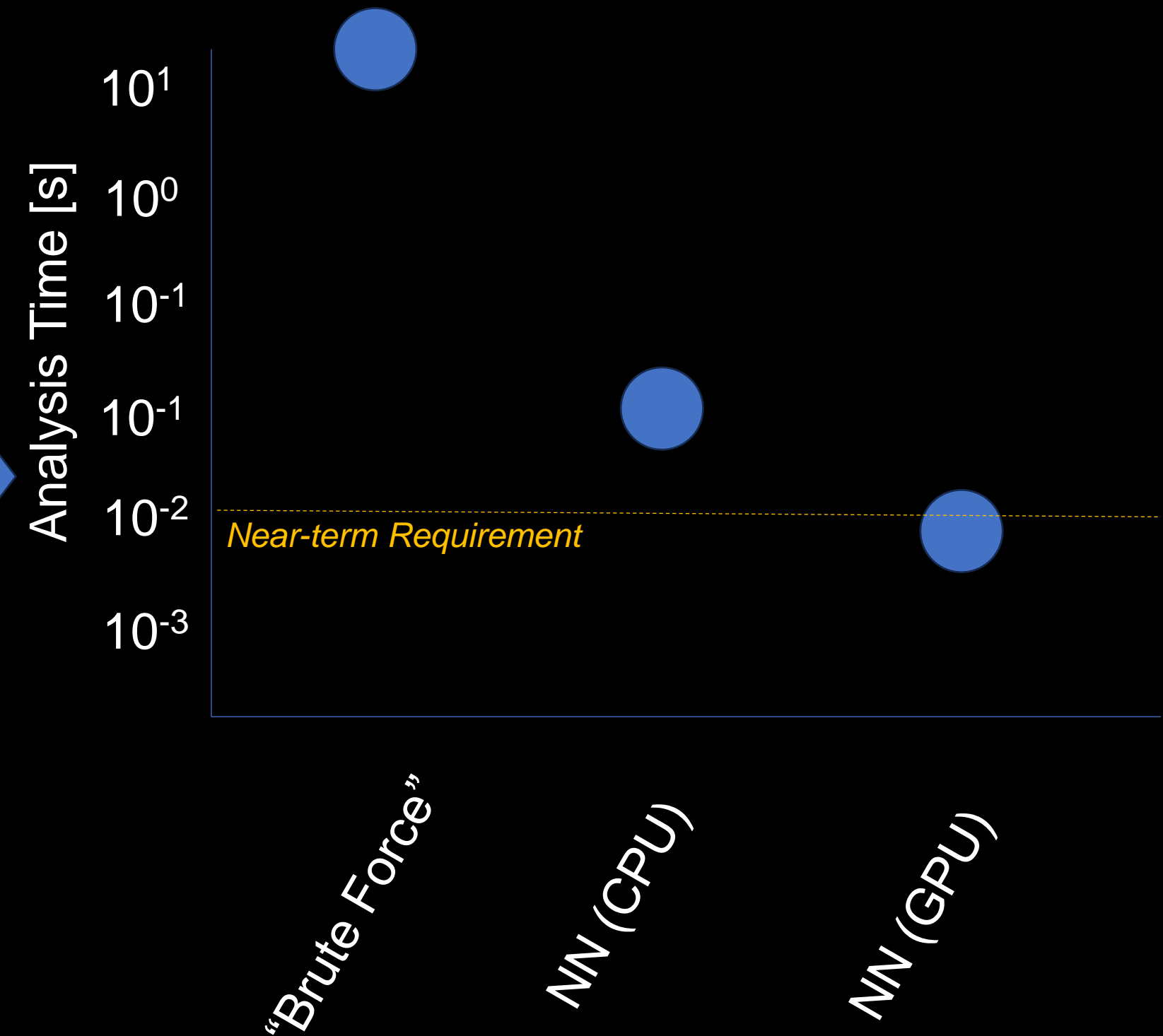
Hardware-in-loop simulations enable us to develop ML-optimization routines that can handle the types of noise that comes with real experiments



# Leveraging edge technologies is necessary to handle the extreme data processing rates



Data reduced to a handful of scalars





# Next-Gen Signal Processing Pipeline

Searching for Extraterrestrial Intelligence and the Origins of the Universe at the Allen Telescope Array

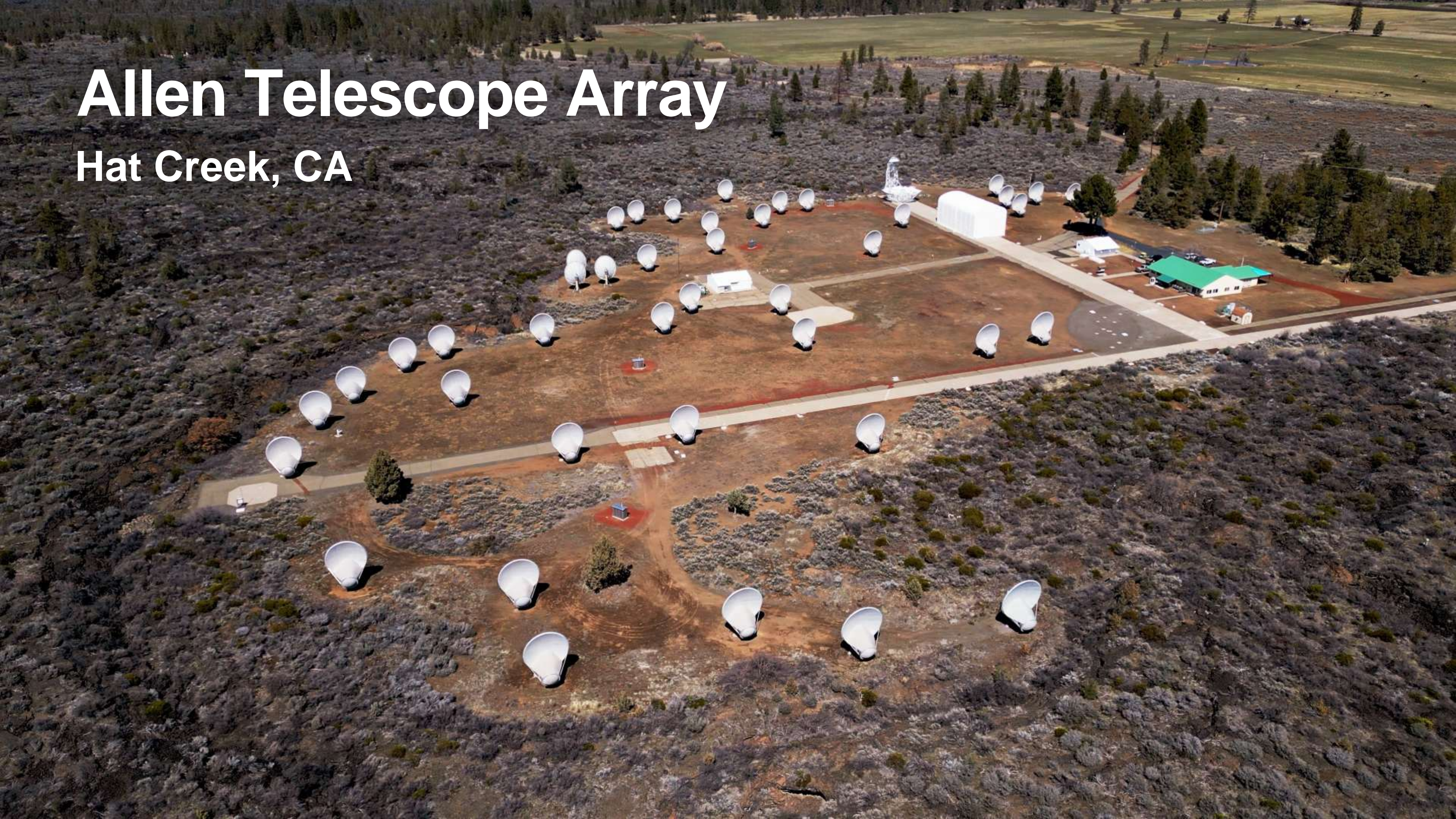
Luigi Cruz, Staff Engineer, SETI Institute

NVIDIA GTC 2025



# Allen Telescope Array

Hat Creek, CA





# AI For Real Time Instruments

Supercomputing '24 Demonstration with the SETI Institute

## AI for Real-Time Instruments

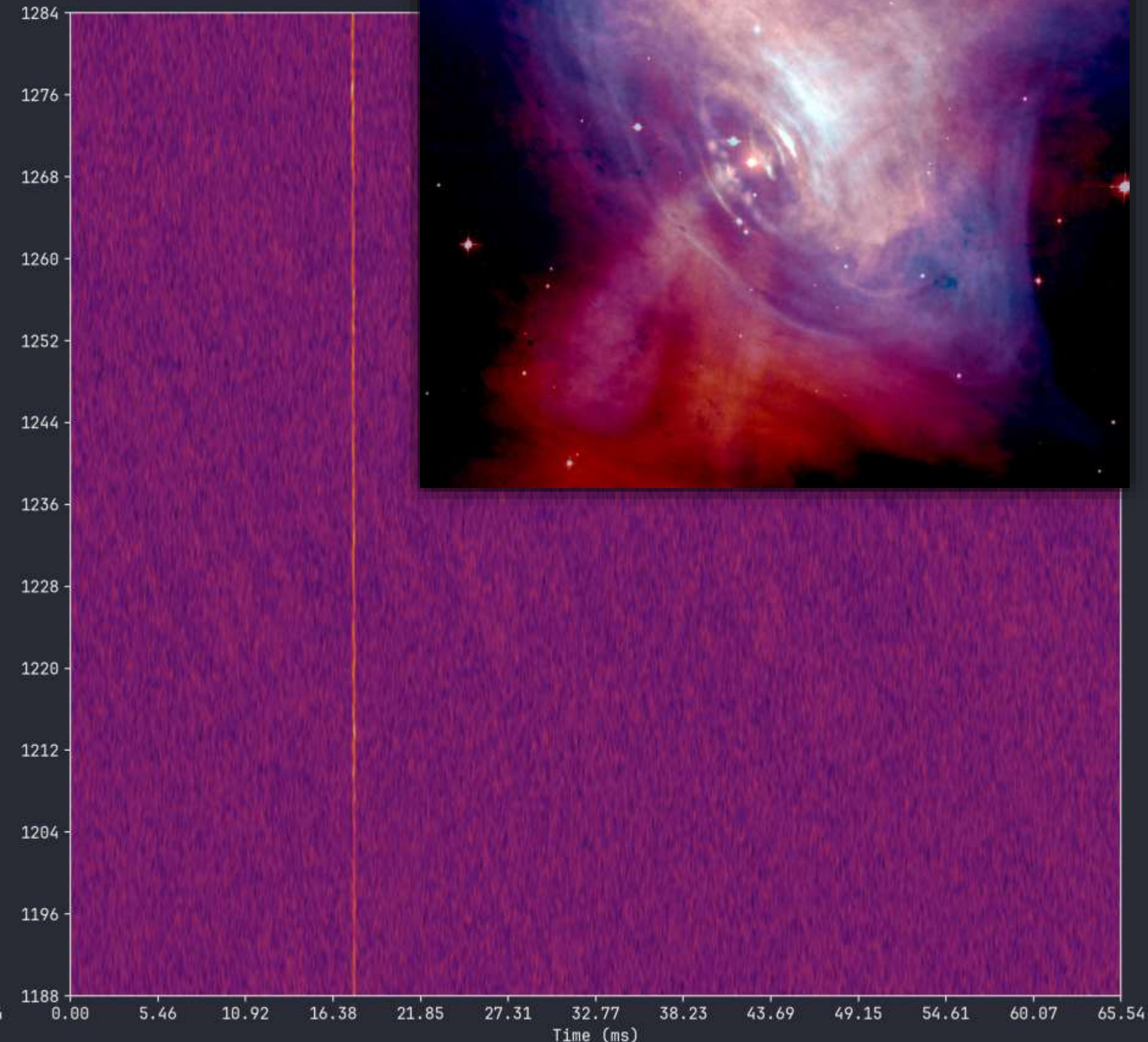
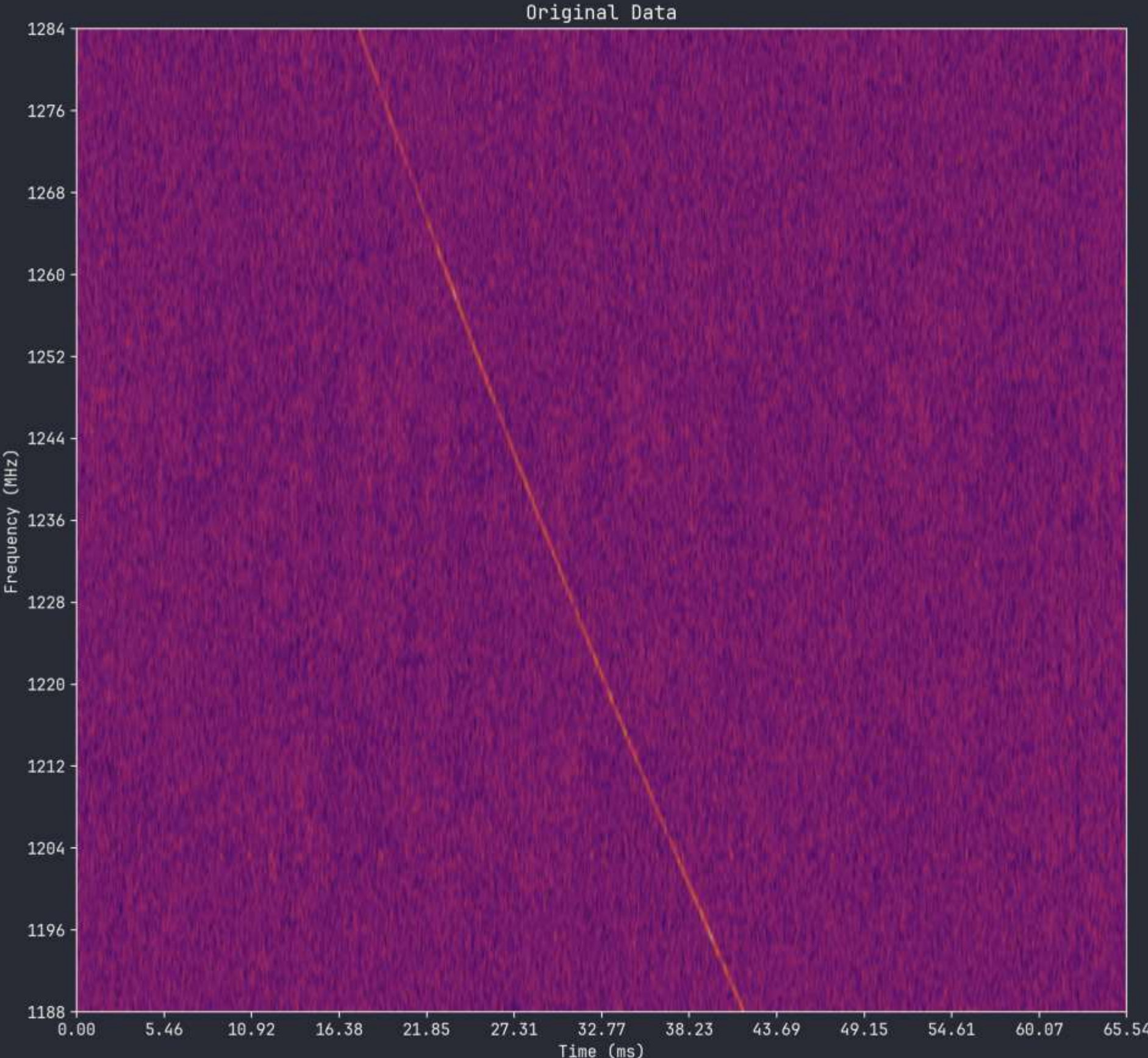
NVIDIA Holoscan and IGX™ in world's first streaming radio astronomy AI experiment at SETI Institute/Breakthrough Listen.





# Crab Nebula Detection

Crab Nebula Hit @ 1236.0 MHz  
(2025-02-10T17:04:01.786432Z) [100.000000%]

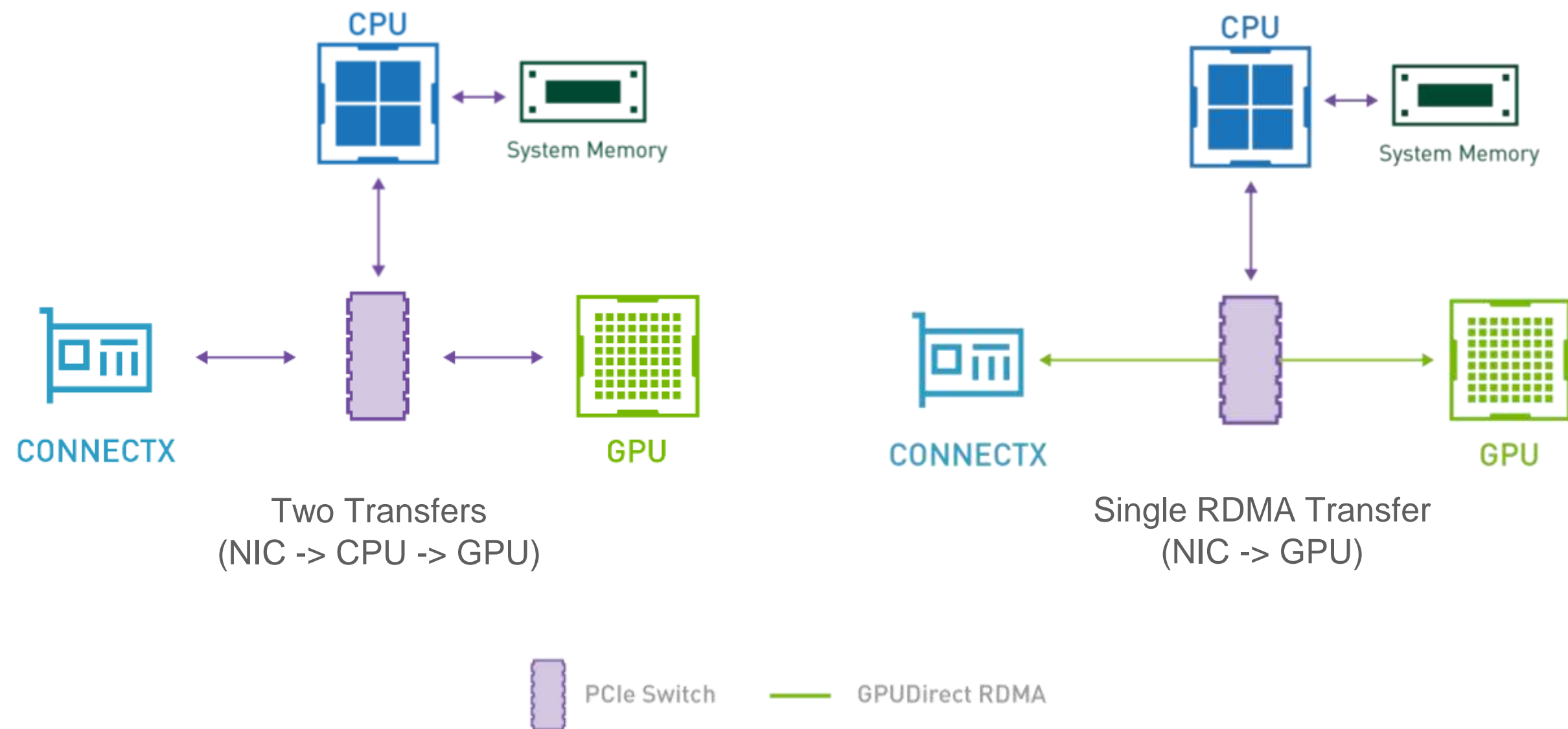




# Next Generation Pipeline with Holoscan

## Three rules for GPU-based scientific data processing:

- Acquisition: Packet ingestion handled with RDMA by Holoscan's Advanced Network Operator.
- Processing: Signal processing handled by BLADE modules via Holoscan Scheduler.
- Storage: Data product stored on NVMe with GPUDirect Storage.

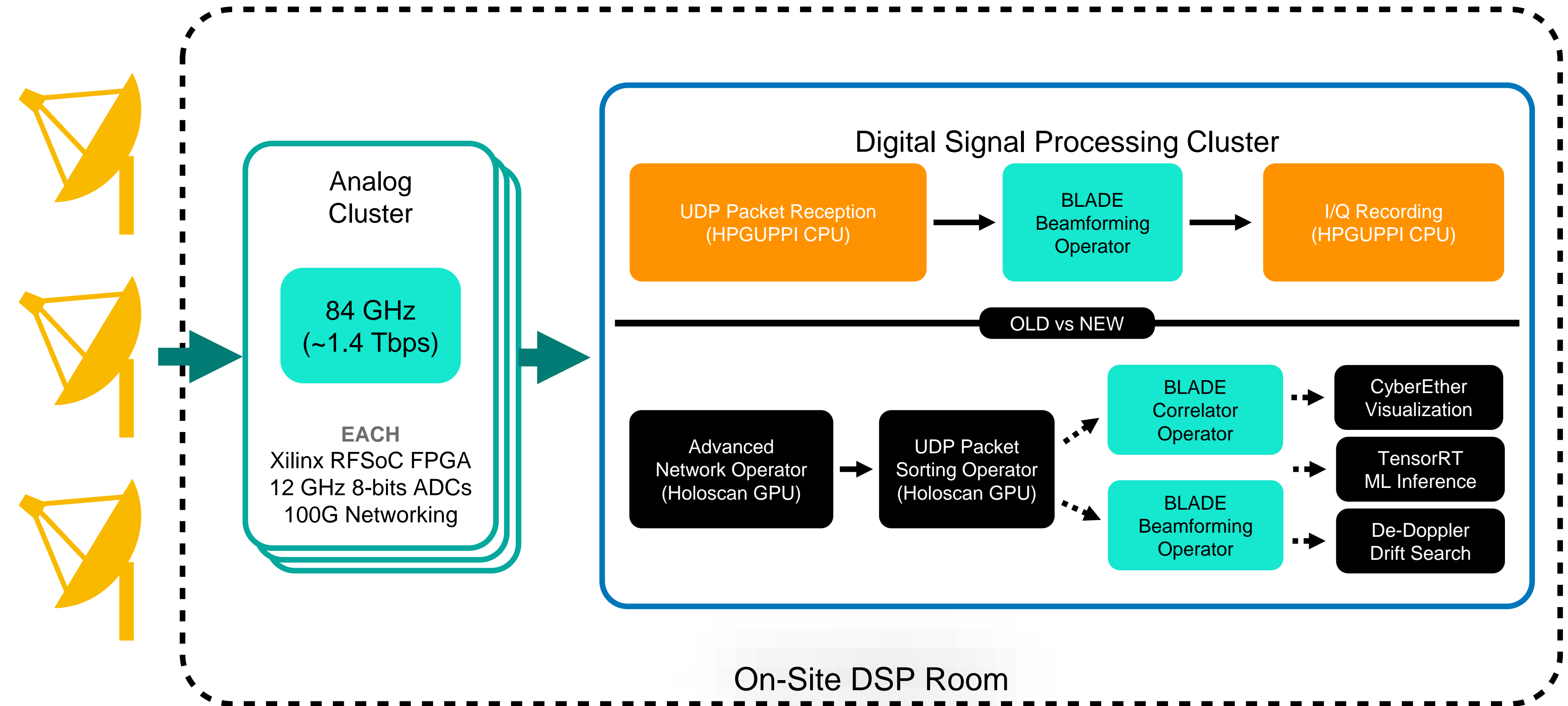


For more information, check out our GTC 2024 presentation:

["Are We Alone? Searching for Extraterrestrial Intelligence and the Origins of the Universe at the Allen Telescope Array"](#)

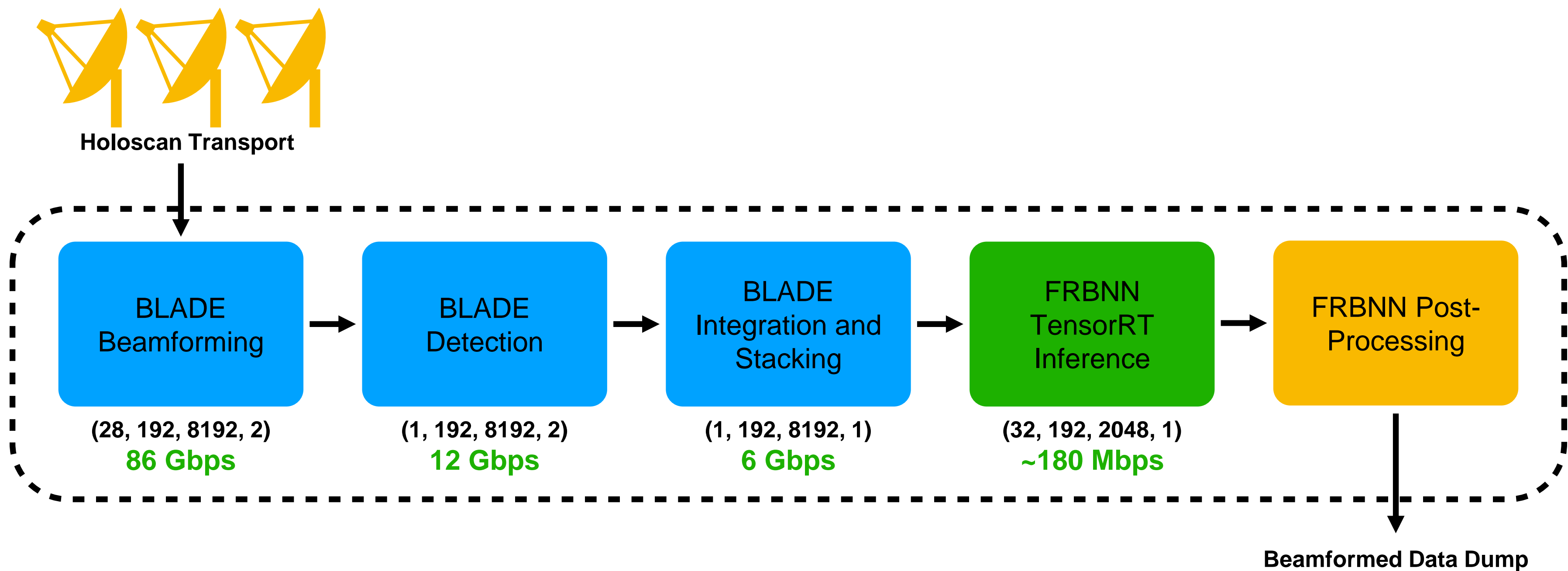


# Next Generation Pipeline





# Radio Burst Detection Pipeline



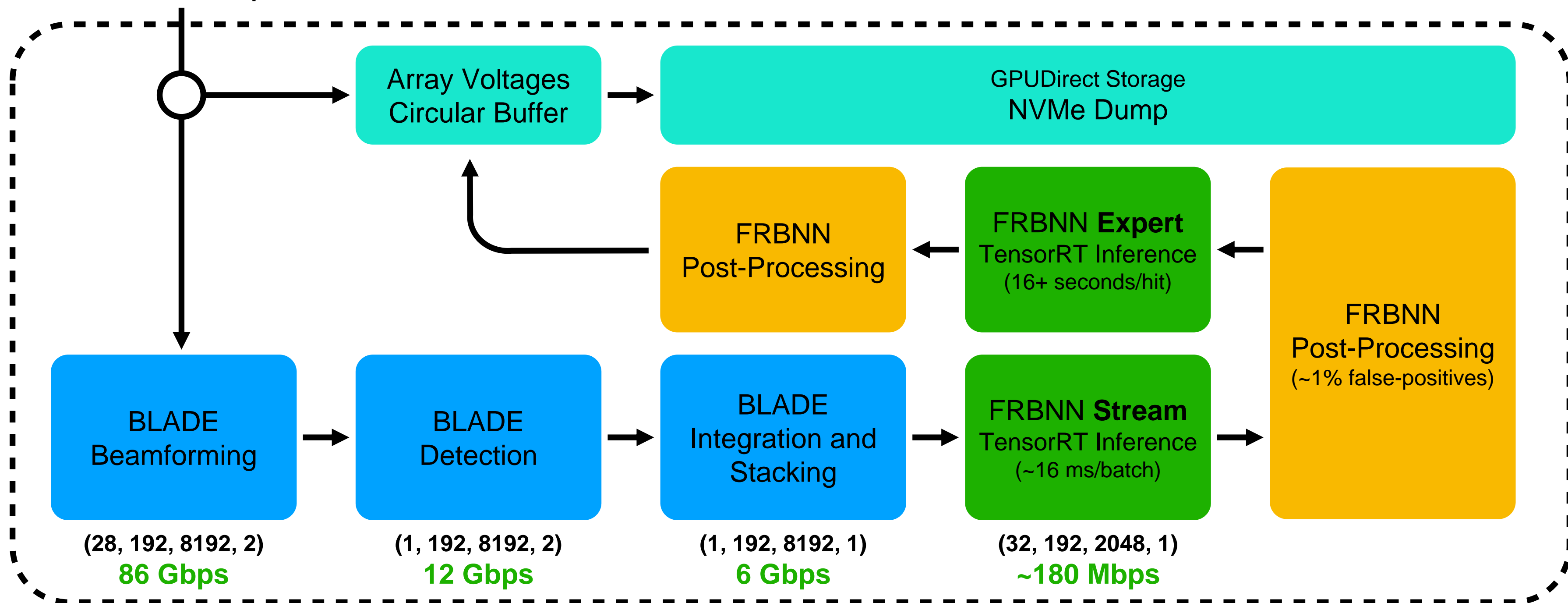
(A, F, T, P) = A - Antennas, F - Channels, T - Time, P - Polarization



# Enhanced Radio Burst Detection Pipeline



Holoscan Transport



(A, F, T, P) = A - Antennas, F - Channels, T - Time, P - Polarization



# Next Generation Digitizers

## Ultra-Wideband Direct Sampling at the Allen Telescope Array


	Current	Planned
RF Bandwidth	~1.5 GHz	~16 GHz
Number of Antennas	28	42
Antenna Data Rate	~46 Gbps	~490 Gbps
Aggregated Data Rate	~1.4 Tbps	~21 Tbps



**HITEK Systems**  
Agilex eSOM7C + AD9081

600 Gbps/antenna





**NVIDIA**  
Holoscan-based Next Generation Pipeline



# Getting Started with Holoscan

## Holoscan References



<https://github.com/nvidia-holoscan/holoscan-sdk>



```
docker pull nvcr.io/nvidia/clara-holoscan/holoscan:v3.0.0-dgpu
```



```
pip install holoscan  
conda install -c conda-forge holoscan
```



Debian Packages available on [NGC](#)



<https://docs.nvidia.com/clara-holoscan/sdk-user-guide/index.html>



