FERMILAB-SLIDES-24-0090-CSAID

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

An Intermediate Level Supernova Pointing Trigger for DUNE Using In-storage AI

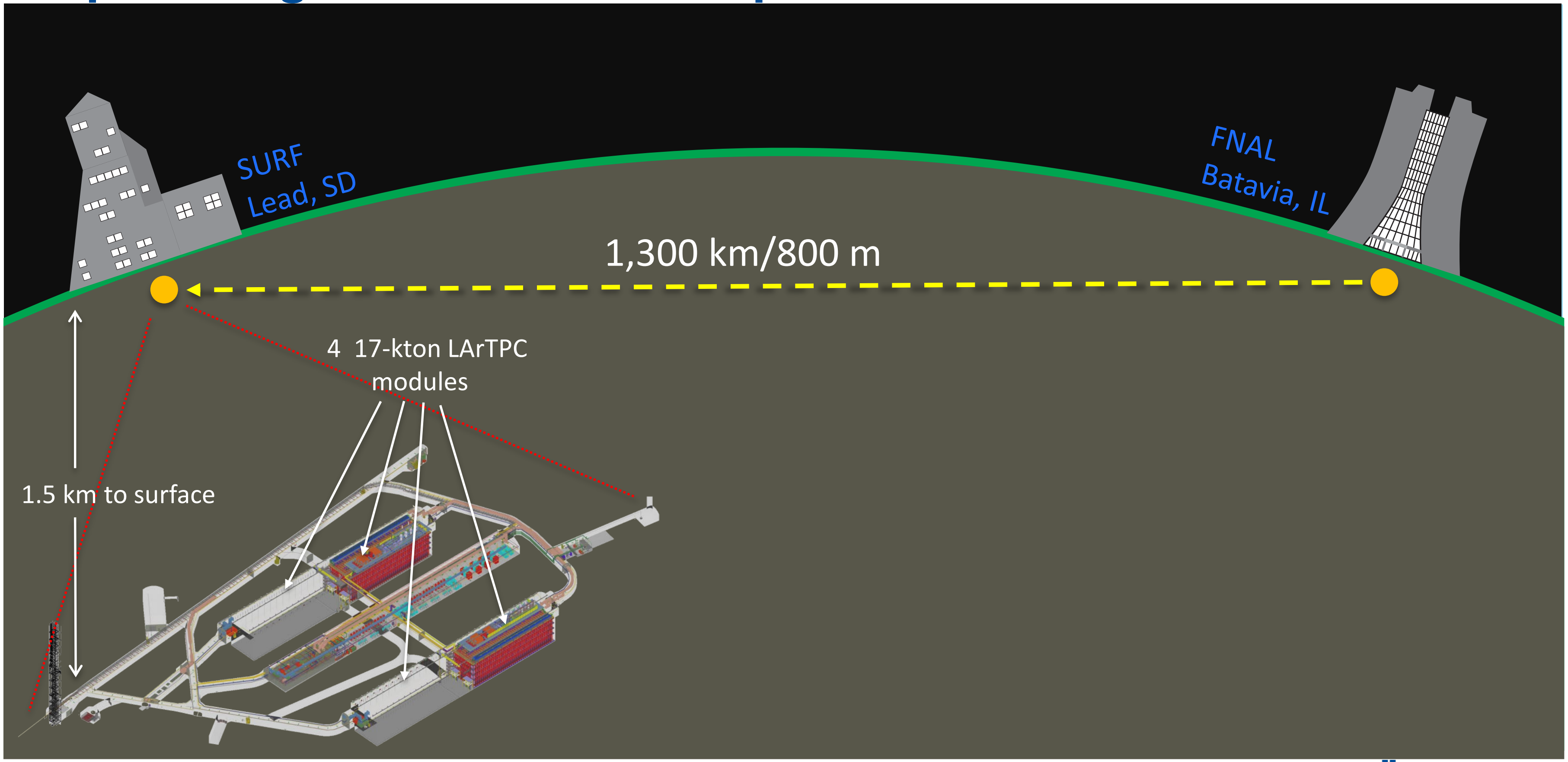
Michael Wang (Fermilab) for the DUNE collaboration

24th IEEE Real Time Conference - ICISE, Quy Nhon, Vietnam

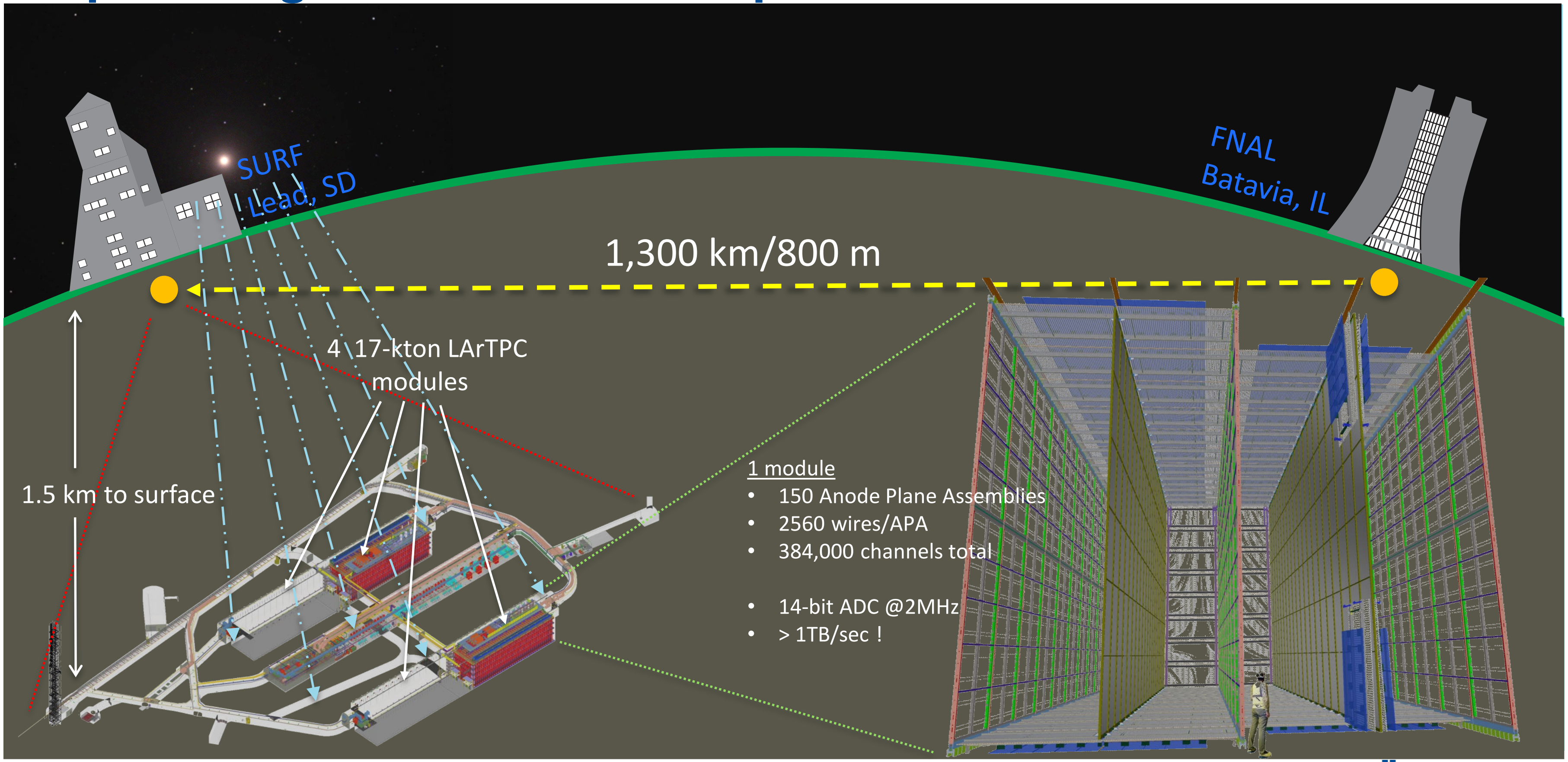
April 22-26, 2024

This document was prepared by the DUNE Collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, Office of High Energy Physics HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

Deep Underground Neutrino Experiment



Deep Underground Neutrino Experiment

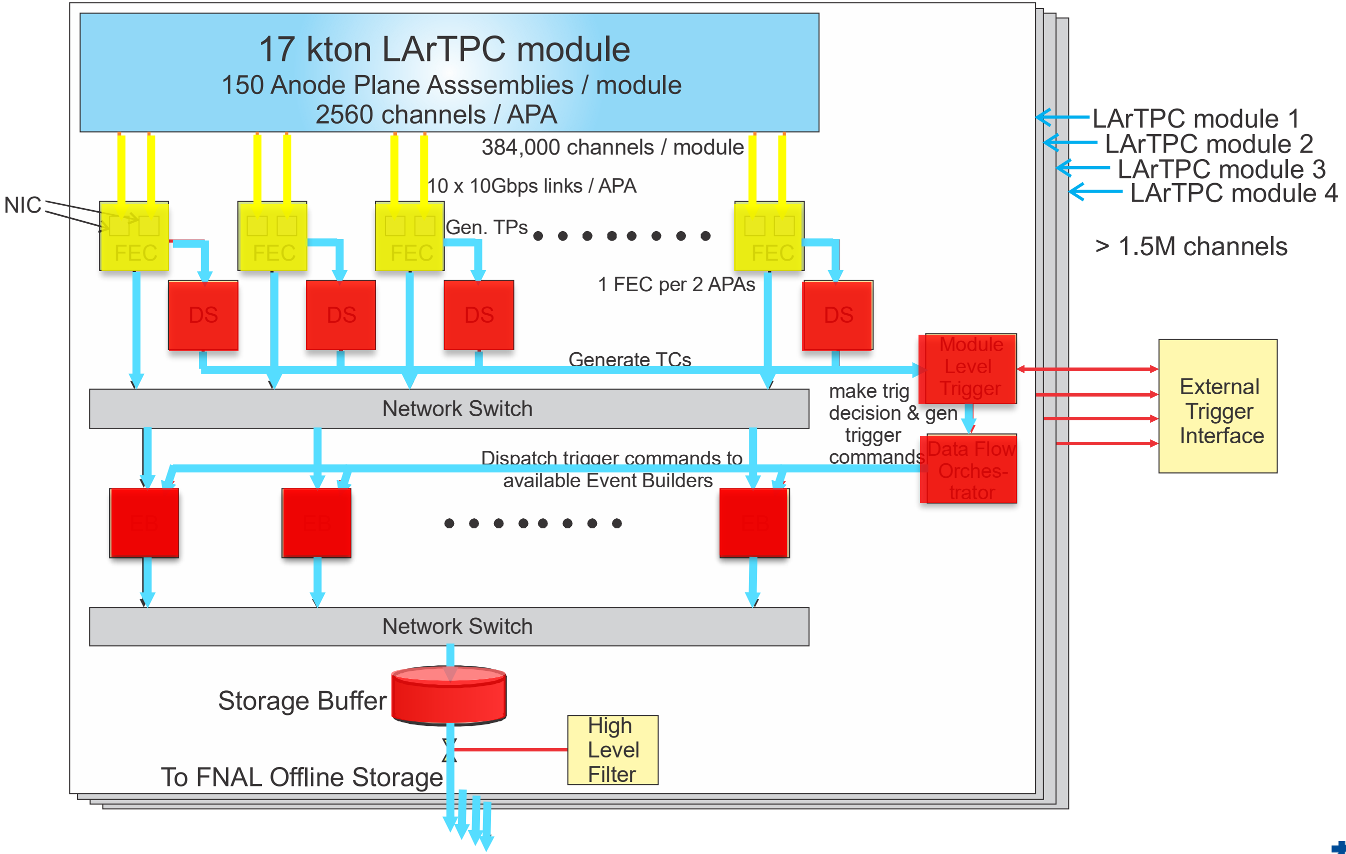


<u>1 module</u>

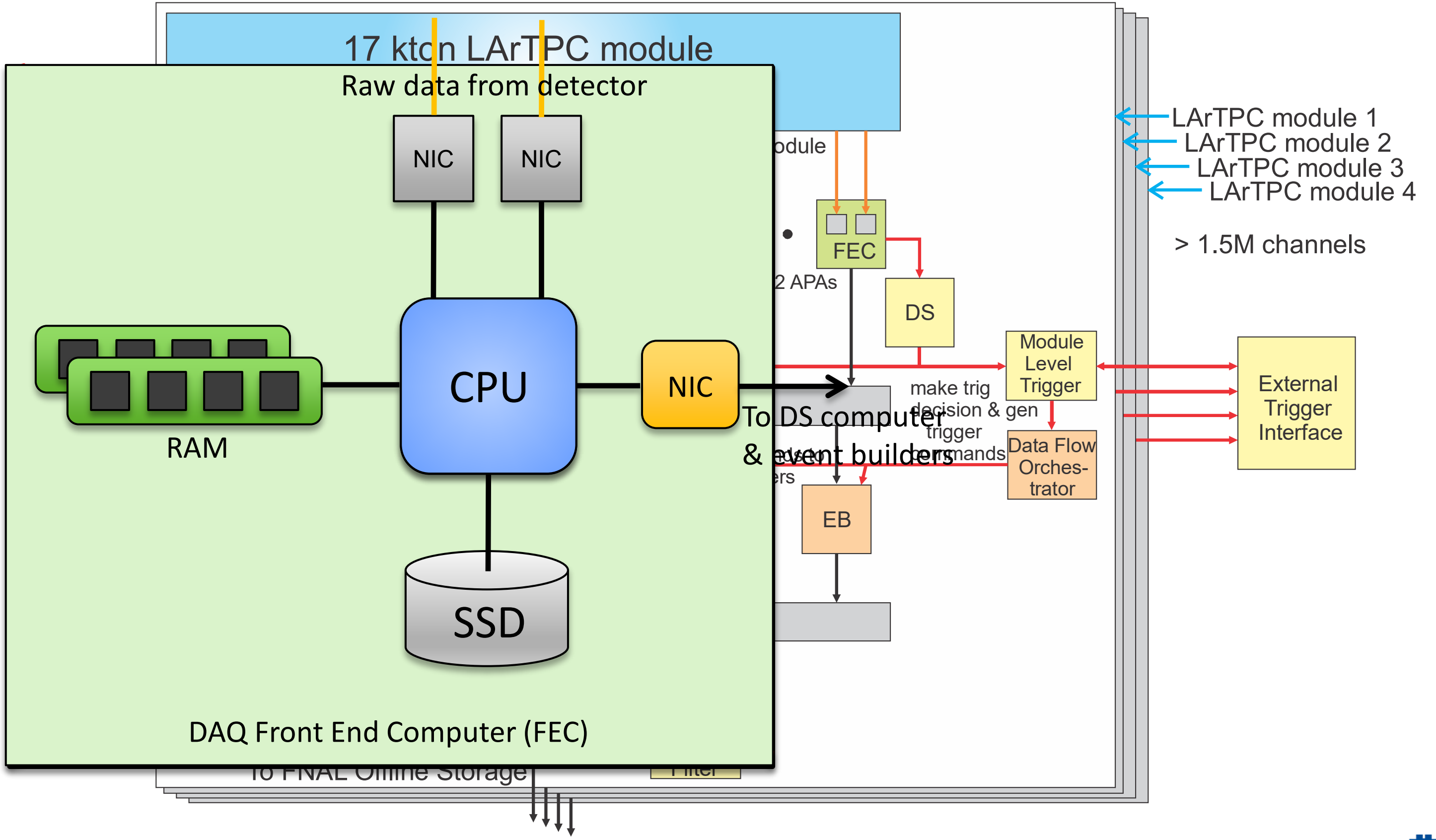
- 150 Anode Plane Assemblies
- 2560 wires/APA
- 384,000 channels total

- 14-bit ADC @2MHz
- > 1TB/sec !

DUNE DAQ and Trigger System

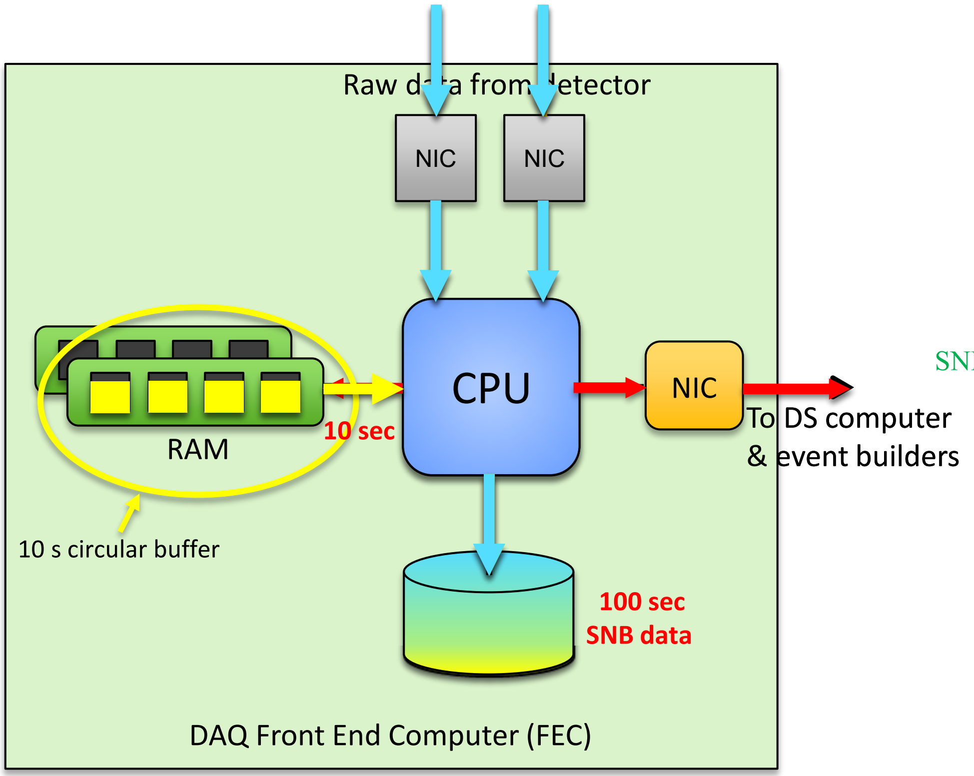


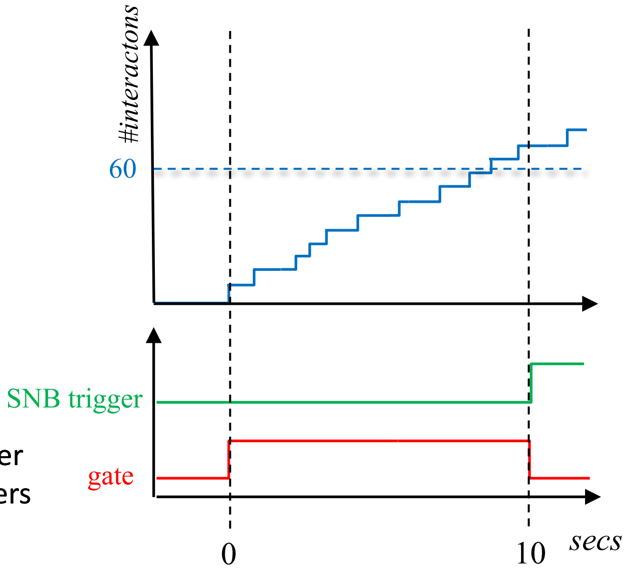
DUNE DAQ and Trigger System



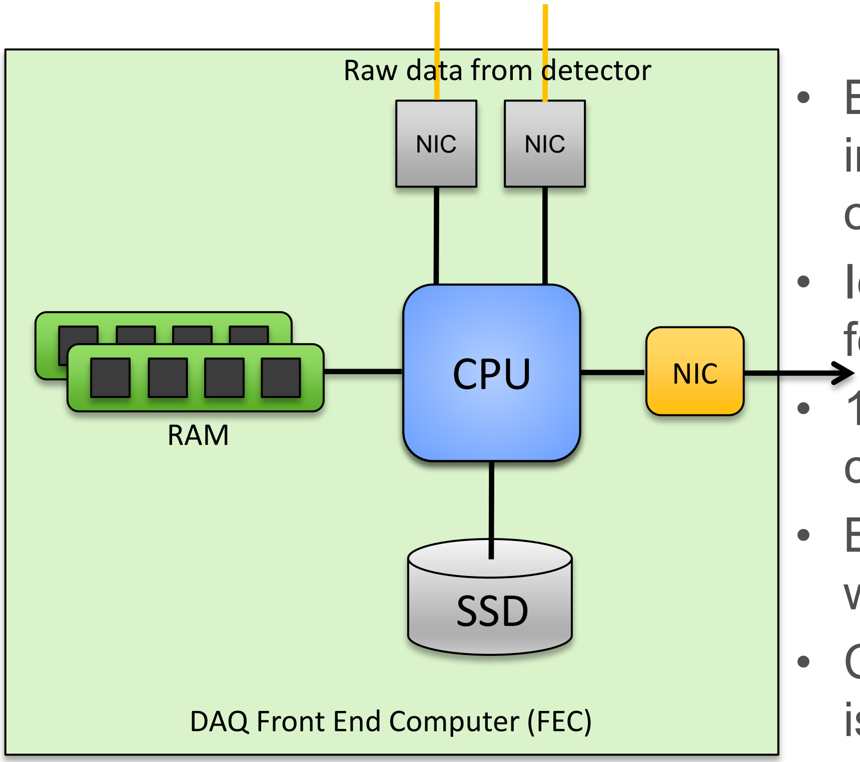
Fermilab

DUNE SNB Trigger





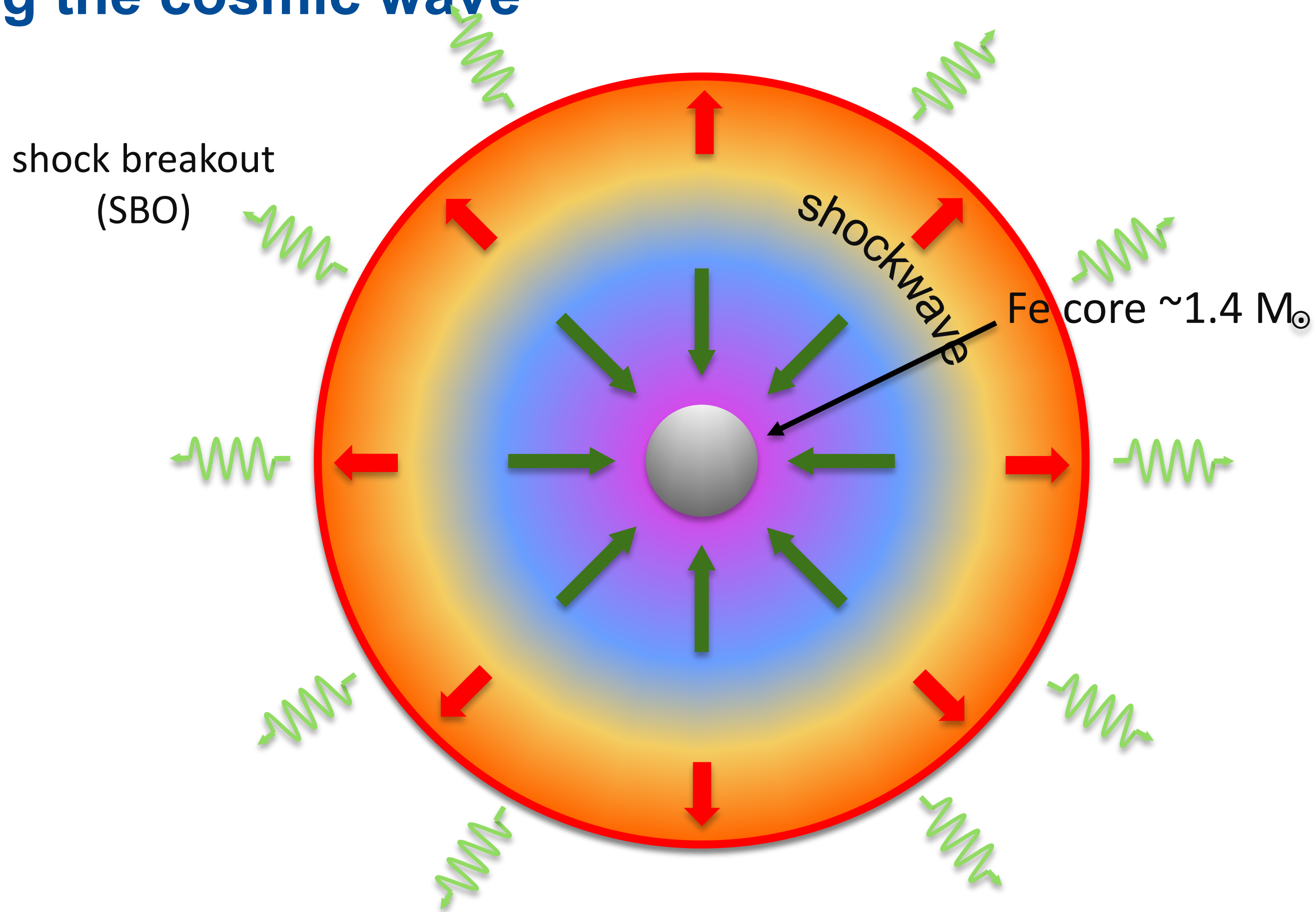
DUNE SNB Trigger – baseline requirements



- Baseline trigger provides no pointing information, which is needed for optical follow-ups
- Idea is to send data back to FNAL for more processing
- 100 Gbps links from underground caverns to surface & back to FNAL
- Best case scenario: 120 TB/module would take ~3 hrs to transfer
- Could be worse: DUNE requirement is to copy data back within 24 h

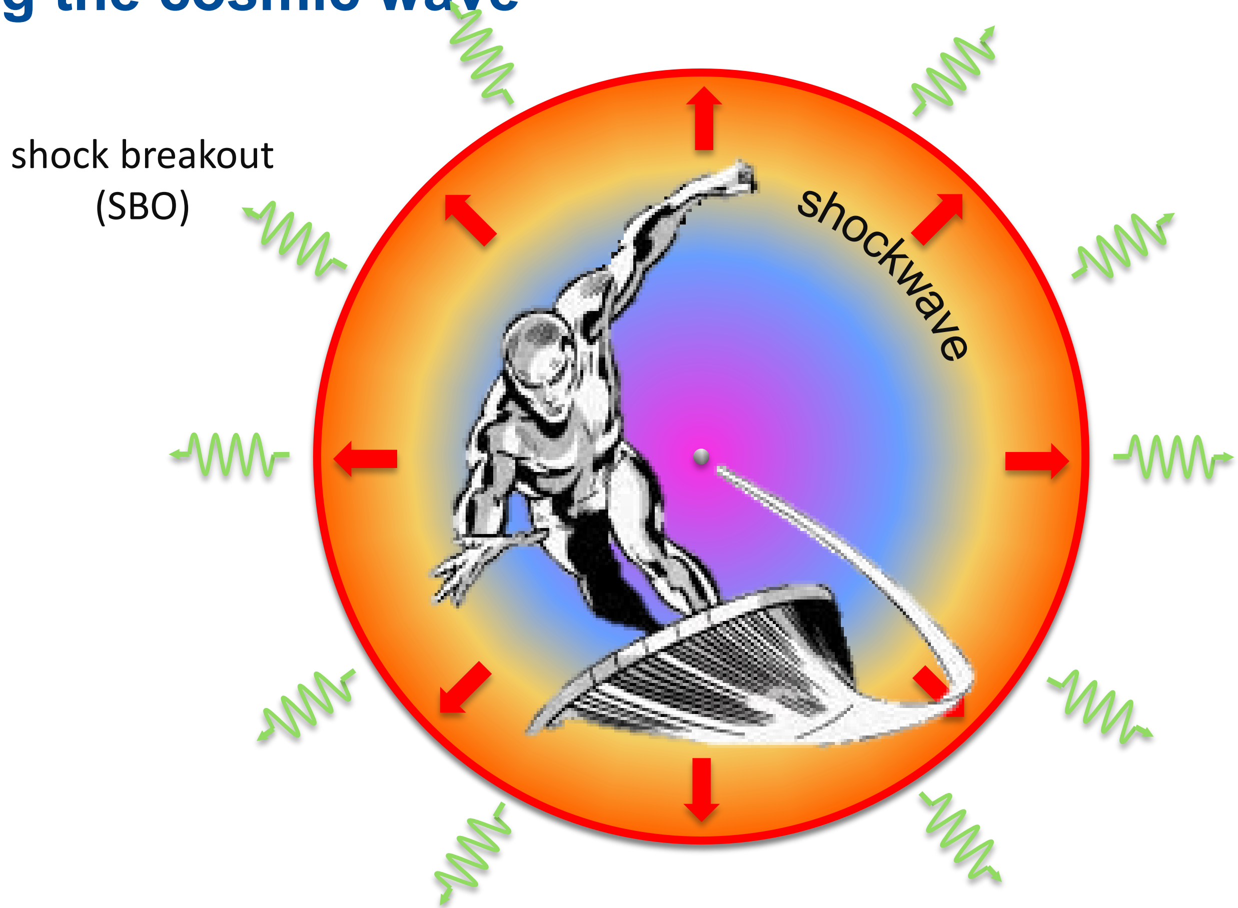
Fermilab

SURFing the cosmic wave



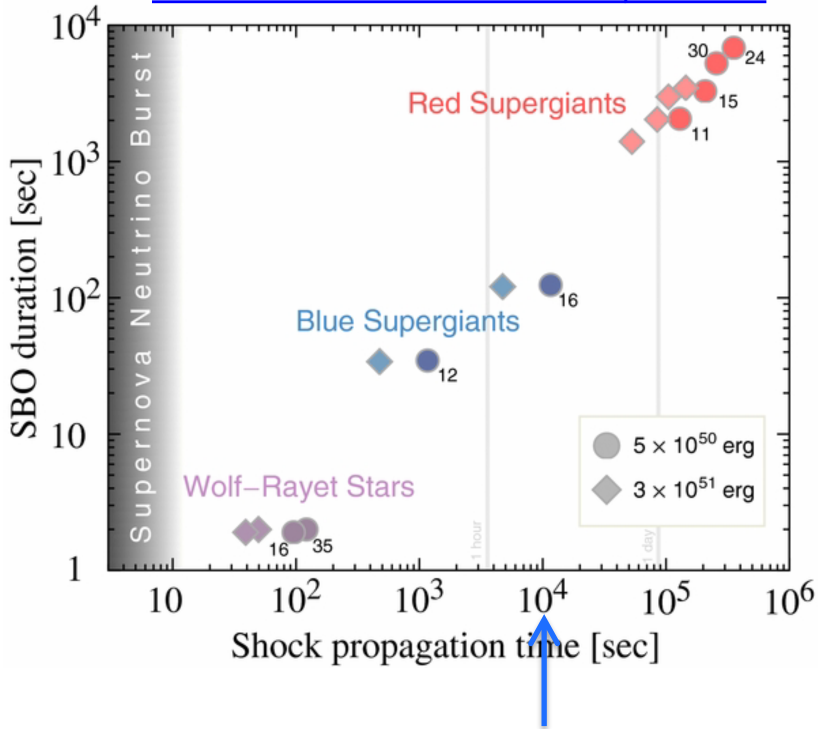
Fermilab

SURFing the cosmic wave



Shock propagation time

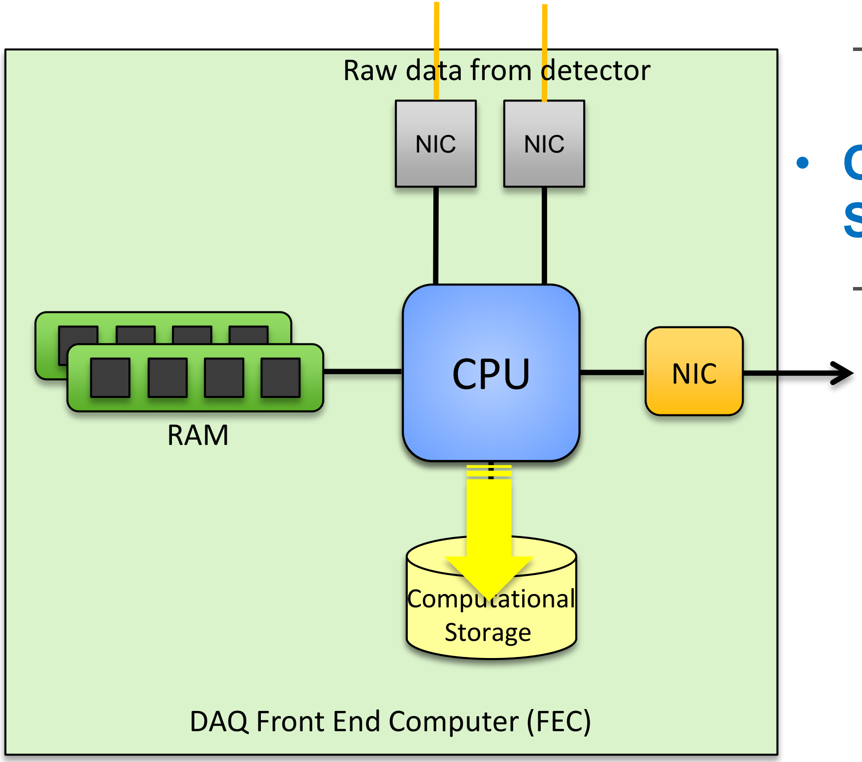
[Matthew D. Kistler *et al* 2013 *ApJ* **778** 81]



SN 1987A ~3 hrs

- Delay between arrival of neutrinos and optical light: ~shock propagation time
- Range: ~1 min to several days
- Unless progenitor is red supergiant, network transfer time back to FNAL alone already exceeds available window of opportunity

Computing paradigm shift



- **A solution: Near-data computing**
  - Instead of moving data to processors, move processing to the data
- **One example is Computational Storage Technology**
  - COTS products available now

Samsung SmartSSD CSD

Standard 2.5" form factor

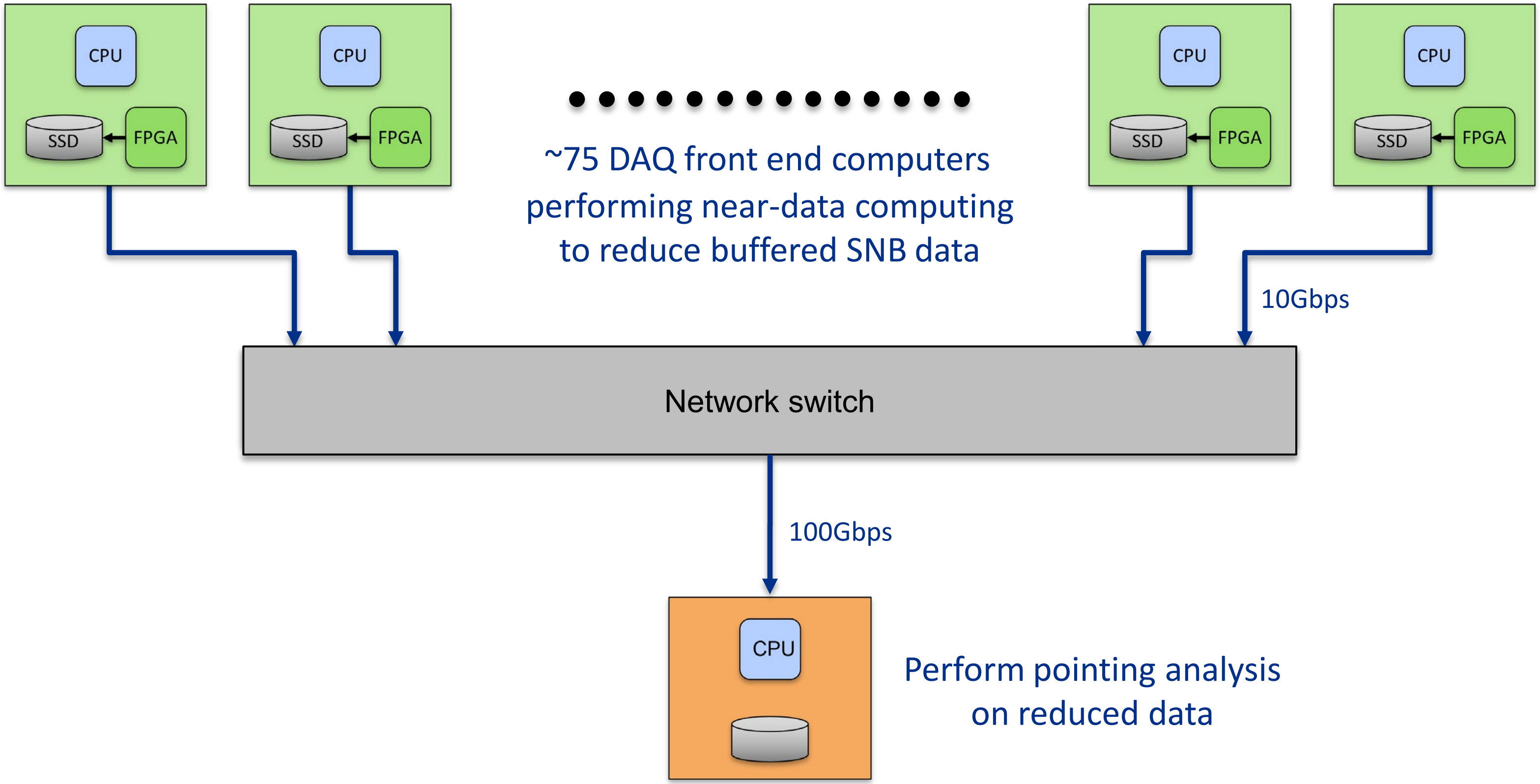
BittWare 250-U2 CSP

Xilinx Alveo card

Strategy for fast pointing determination

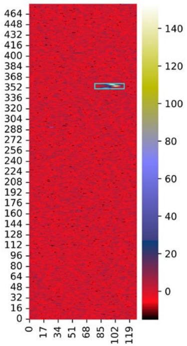
- Two step approach:
  - Data reduction: perform *in situ* on data, using FPGAs or GPUs to reduce buffered SNB data to a point where it can be transported quickly over conventional network to destination server
  - Pointing determination: execute optimized “offline-like” pointing analysis on the reduced data set on the server
- Use AI/ML methods for the in situ data reduction step:
  - Run ML models on accelerators like FPGAs, GPUs, etc.
  - Fast inference times for low latencies
  - Apply ML methods on both:
    - Raw 2D LArTPC wire plane images
    - Raw 1D LArTPC wire waveforms

Strategy for fast pointing determination: hardware



Fermilab

Strategy for fast pointing determination: algorithms



Raw wire plane data

ML-based *in situ* data reduction

Radiological Bkg rejection with **2DCNN** → Raw waveform ROI finding with **1DCNN** → Denoising raw waveforms in ROIs with **1D AutoEncoder**

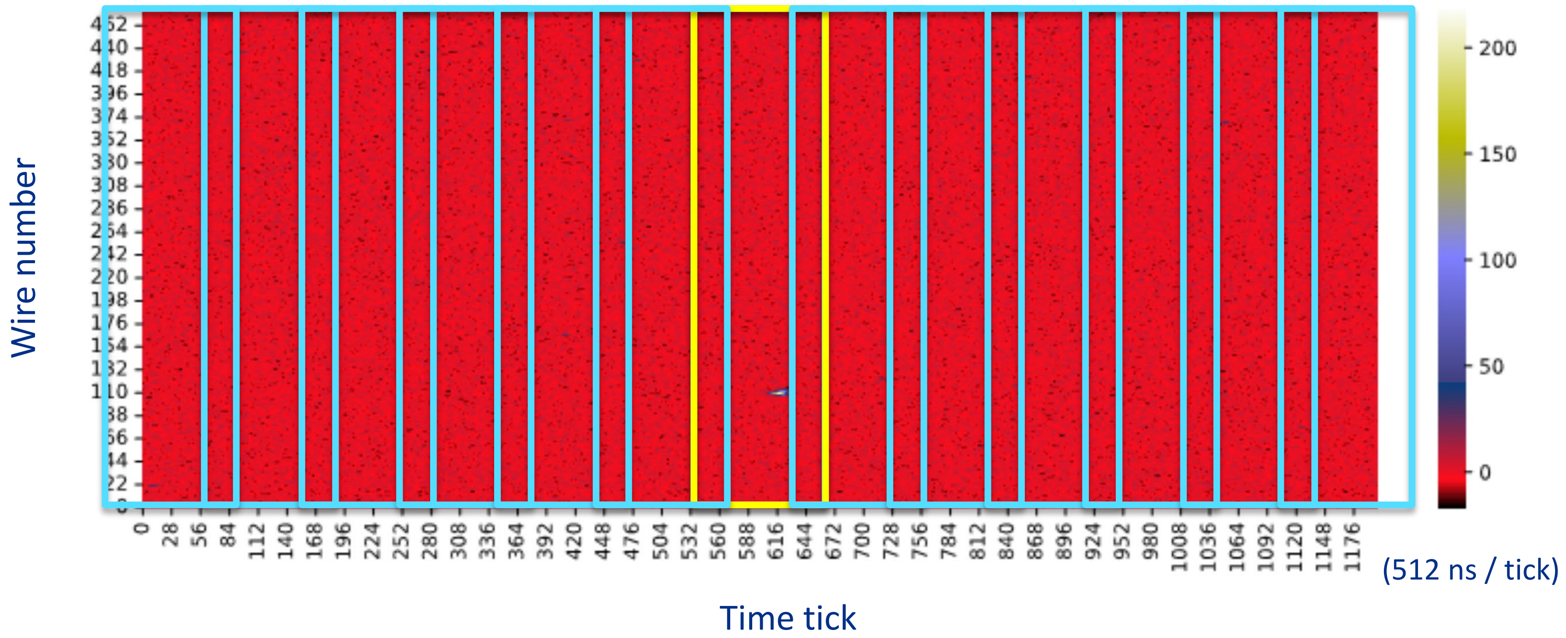
network

Dataprep (sig-proc) → Gaushit → SP Solver → DisAmbig → TrajCluster → PmTrack → PointTreeRes → Likelihood fit for direction

"Offline-like" reconstruction and pointing analysis

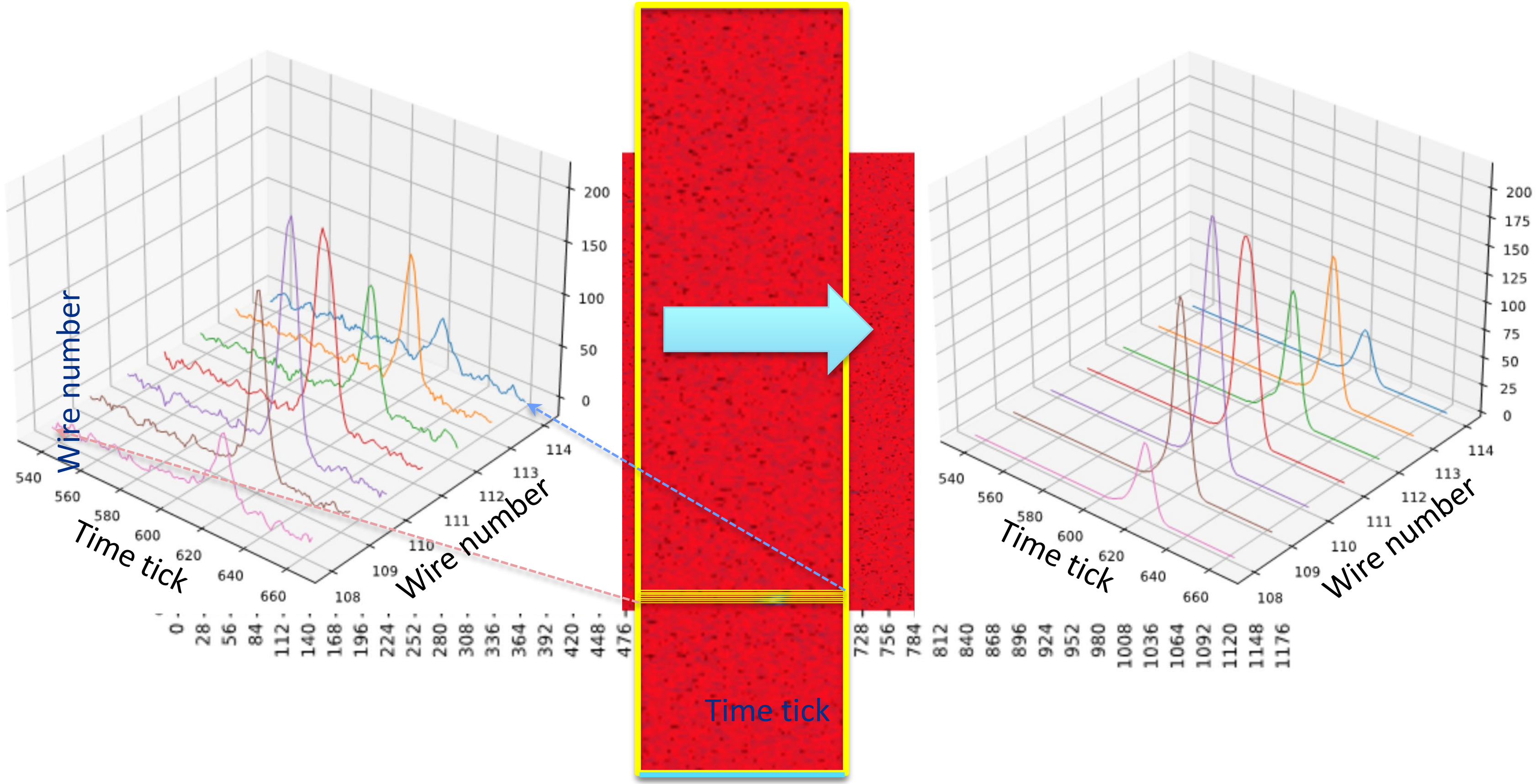
Fermilab

Radiological background rejection with 2D-CNN



Fermilab

ROI finding with 1D-CNN and denoising with 1D-AE



ML-based data size reduction estimates

- Data rate from 1 far detector module (horizontal drift):
  - 150 APAs × 2560 ch/APA × 14 bits/sample × 2 Msamples/sec ~ 1.2 TB/s
- Buffered supernova data per detector module:
  - 100 seconds: 120 TB
  - For pointing determination, focus on first 10 seconds: 12 TB
- Estimated size of data per SN neutrino candidate from ML-based reduction pipeline:
  - CC: 47,306 bytes
  - ES: 29,680 bytes
- Assuming 2DCNN rejects 100% radiologicals, assume we retain all CC + ES neutrinos interactions:
  - 3,300 CC × 47,306 bytes + 326 ES × 29,680 bytes ~ 158 MB for all 4 modules
  - 48 TB → 0.000151 TB ~ over 5 orders of magnitude reduction !

Execution time for track reconstruction on reduced sample

- Track reconstruction pipeline:

  → Dataprep → Gaushit → SP Solver → DisAmbig → TrajCluster → PmTrack →

- Use of 1D denoising AutoEncoder to clean up electronics noise of raw waveforms in ROIs from 1DCNN allows us to use “legacy” 1D FFT deconvolution in the “Dataprep” stage to speed up things.

Interaction	Reco time (sec/event)
CC	0.061
ES	0.026

- Assume we reject all radiologicals with 2DCNN and retain all CC+ES events:
  - 3300 × .061 + 326 × .026 ~ 210 seconds = 3.5 min
  - for full DUNE detector executed using one CPU core !

Compare with no reduction case

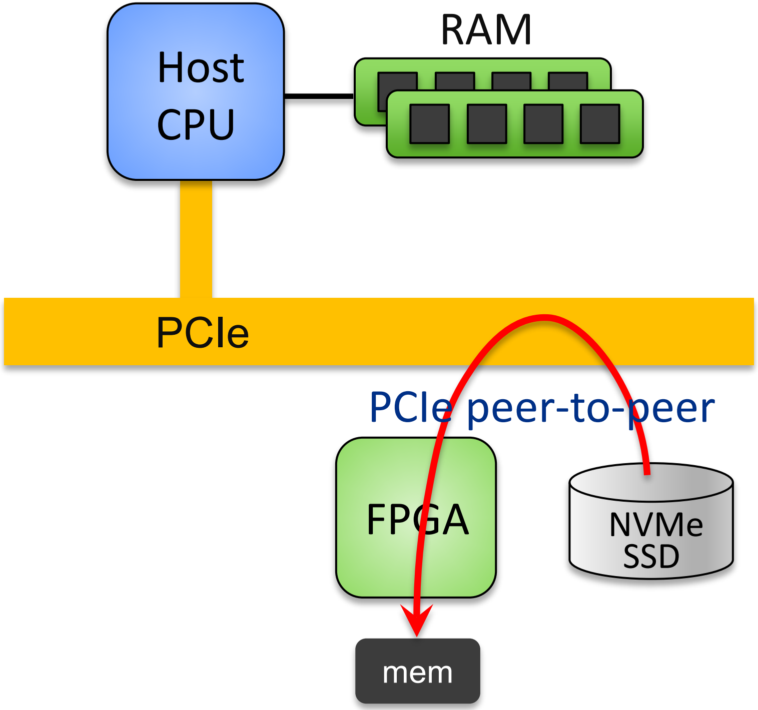
- How long would it take to process the SN data using the reco pipeline in the previous slide on a full 10 second raw dataset
- It takes ~9.5 seconds to process 1 APA worth of data in a 6000 time tick (500 ns/tick) readout window using one CPU core:
  - Assuming we dedicated one CPU core to 1 APA on each DAQ readout computer:
    - 100 seconds of SN data would take 88 hrs to complete
    - 1st 10 seconds of SN data would take ~9 hrs to complete
- For simplicity assume all 4 detector modules identical to first horizontal drift module, i.e. 150 APAs per module:
  - 9 hrs using $4 \times 150 = 600$ CPU cores versus 3.5 min using 1 CPU core !

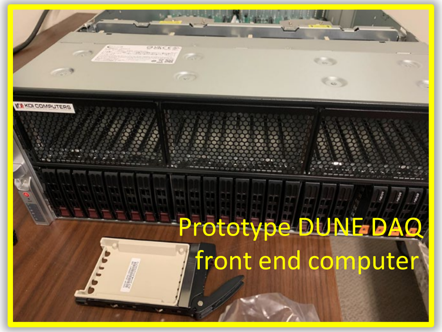
Execution time for ML-based data reduction pipeline

- ML-based data reduction algorithm is meant to run on FPGAs that access buffered SN data on SSDs directly. FPGA's main advantage over GPU is lower power consumption – important because of limited power budget in SURF underground caverns.
- However, since we were not able to get results in time, we benchmarked the algorithm on a typical GPU to get an idea of what is achievable
- Using half of an Nvidia A100 GPU*:
  - Since each front end DAQ computer serves 2 APAs
  - 15 minutes to perform the ML-based data reduction for 1 APA

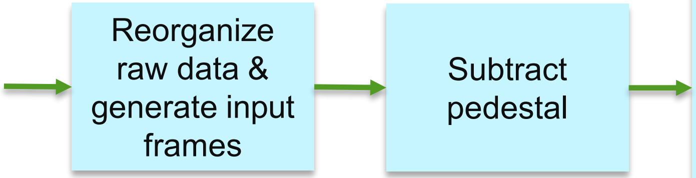
  * A100 is a datacenter GPU and not the ideal for this application due to power consumption. Only used here to get an idea of inference times possible. There are other GPUs targeted at low latency inference combined with low power consumption
- Total of ~20 minutes to do ML-based data reduction + track reconstruction
  - Still considerably less time than it takes just to transfer the SN data back to Fermilab:
    - 10 seconds of SN data for all 4 modules: 48 TB takes ~1 hr to transfer over 100 Gbps ethernet

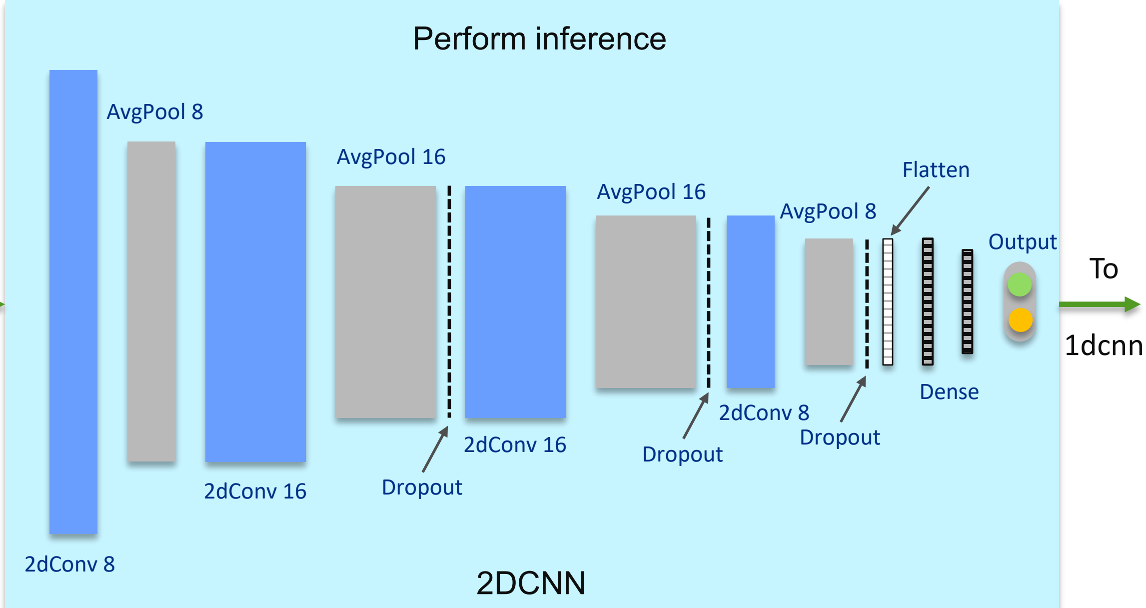
In-storage ML-based data reduction





Hardware implementation 1





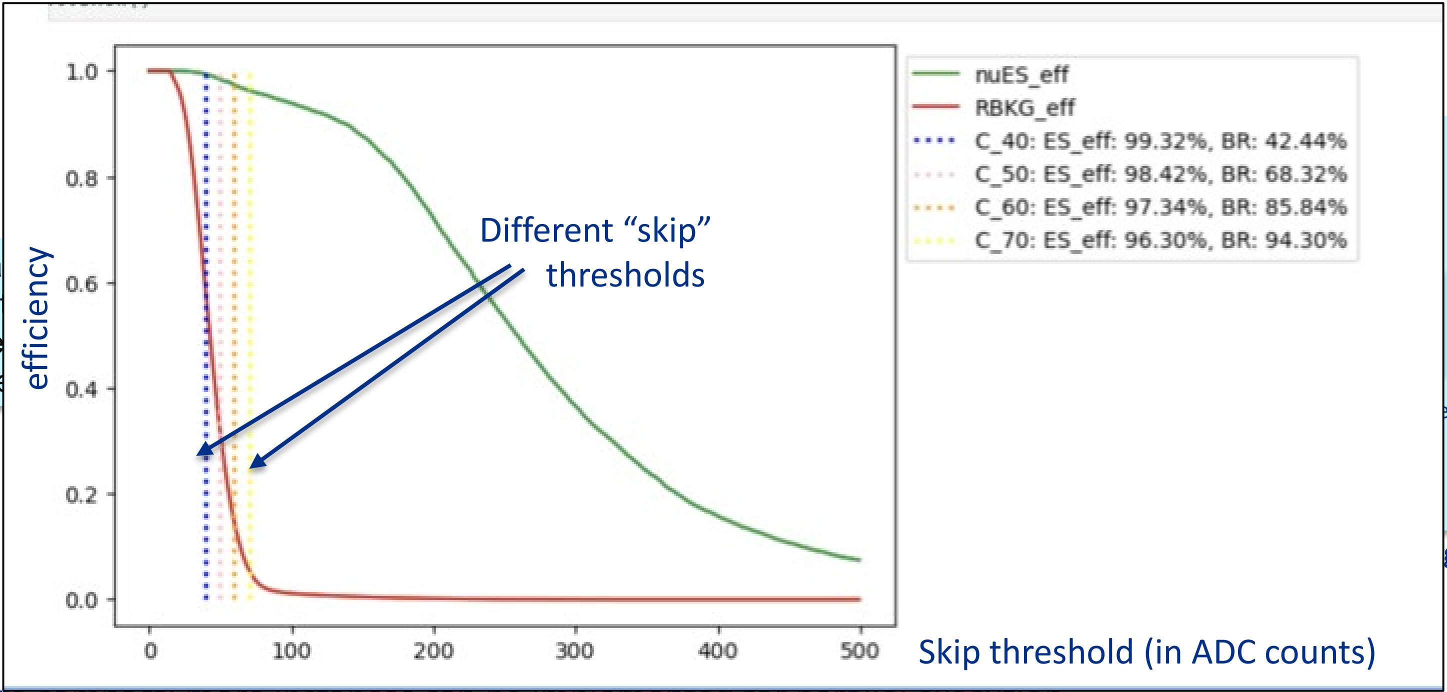
Full-size frames sent to 2dcnn for inference

Plane	Input frame size
induction	200 x 1,148(9)
collection	200 x 480

Inference time/frame ~3.2 ms
Compared with ~0.5 ms for the GPU

Hardware implementation 2

Reorga
raw dat
generate
frame



Output
To 1dcnn

ge

- Loo
- More compact: more instances can be implemented for parallel execution
- Requires more testing with wider input frames and inclusion of all pre-processing steps

Signal efficiencies when using ML-based data reduction

- Fast execution times are useless unless we retain a significant amount of the SN neutrino interactions. Here we compare signal efficiencies after the GausHit finder stage between a standard full dataset reconstruction and our ML-based reduced dataset reconstruction:

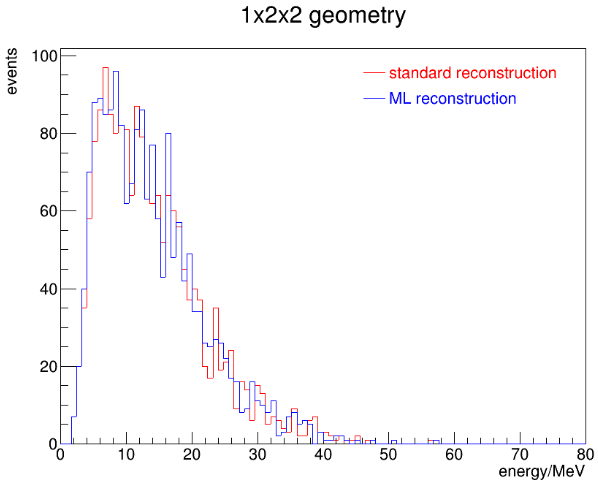
Reconstruction chain	Sig eff for primary trk hits			Sig eff for daughter trk hits		
	U	V	Z	U	V	Z
ML-reduced (1D deconvolution)	0.69	0.71	0.66	0.16	0.18	0.073
Standard full dataset (2D deconvolution)	0.68	0.67	0.62	0.14	0.17	0.062

GausHit finder hit efficiencies for fully simulated nuES events

- Able to achieve same signal efficiencies but with tremendous reduction in execution time !

Energy distributions

- Here we compare the reconstructed energy of the scattered electron in the SN ES interactions between the standard full-dataset reconstruction and the ML-based reduced dataset reconstruction:



- Basically, the ML-based reconstruction pipeline produces identical results as a standard offline reconstruction

More info on DUNE SN pointing capability & ML models used

- For more details on DUNE's supernova pointing capabilities, please refer to:
  - Abi, B., Acciarri, R., Acero, M.A. *et al.* Supernova neutrino burst detection with the Deep Underground Neutrino Experiment. [*Eur. Phys. J. C* **81**, 423 (2021).]
  - Updated version to be published:
    Abed Abud, A., Abi, B., Acciarri, R. *et al.* Supernova pointing capabilities of DUNE.
    (DUNE-doc-27538-v13, primary authors: Shen, J., Roeth, A. J., Hakenmueller, J., Queen, J., Pershey, D., Scholberg, K.)
- For more details on ML models used in this work:
  - Clair, J. Real-Time Detection of Low-Energy Events for the DUNE Data Selection System. DUNE-doc-27333-v1. (2DCNN)
  - Uboldi, L., Ruth, D., Andrews M., Wang, M.H.L.S. *et al.* Extracting low energy signals from raw LArTPC waveforms using deep learning techniques — A proof of concept. *Nucl. Instrum. Methods Phys. Res. A* **1028** 166371 (2022). (1DCNN)
  - Mitrevski, J. Low Energy LArTPC Signal Detection using Anomaly Detection. Fast Machine Learning for Science, Imperial College London, 25-28 Sept 2023. (1D denoising autoencoder)

Conclusions

- Common assumption: DUNE SN processing requires significant computing resources: processing, network, and storage
  - SN data needs to be transferred back to FNAL for more processing to determine direction
  - HPC sites were also being considered for this purpose
- What we have demonstrated: applying ML-based data reduction as early as possible can reduce data to such a degree that makes it possible to perform offline-like analysis on-site with minimal computing resources – i.e. on a single server:
  - total processing times less than network transfer time back to FNAL appear achievable
  - no loss in quality of results wrt full reconstruction/analysis
- This has positive implications for previous assumptions about DUNE's computing requirements, perhaps warranting a re-examination of these assumptions

Conclusions (continued)

- While results look promising, this is a work in progress, and more work needs to be done:
  - Single largest contribution to the total processing time is the 2DCNN-based radiological background rejection:
    - Focusing on reducing 2DCNN inference time will reap the largest benefits
  - Continue exploring hardware accelerator options:
    - Continue with FPGA implementation and optimization
    - Benchmark GPUs geared towards low-latency inference applications & low power
    - Explore dedicated AI inference chips
  - Include algorithms for discriminating ES vs CC and optimize these algorithms
  - Implement end-to-end demonstrator within dune-daq framework

People directly involved in this effort

- ***Fermilab***
  - Maira Khan, Jovan Mitrevski, Ben Hawks,
    Tom Junk, Tingjun Yang, Jennifer Ngadiuba, Mike Wang,
    Pengfei Ding (now at LBNL)
- ***Duke University***
  - Kate Scholberg, Janina Hakenmueller, Van Tha Bik Lian
- ***Columbia University***
  - Georgia Karagiorgi, Judicael Claire, Guanqun Ge, Akshay Malige
- ***York University***
  - Tejin Cai (now at Synopsys Inc.)
- ***Iowa State University***
  - Amanda Weinstein, Avik Ghosh

Thank you!

How a LArTPC works