Data Management for High Speed, Distributed Data Acquisition

Jeff Maggio

Principal Investigator

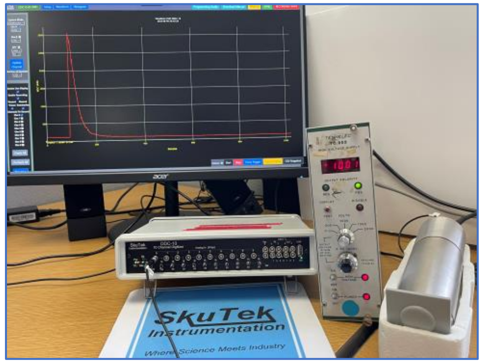
SBIR Exchange Aug 2023

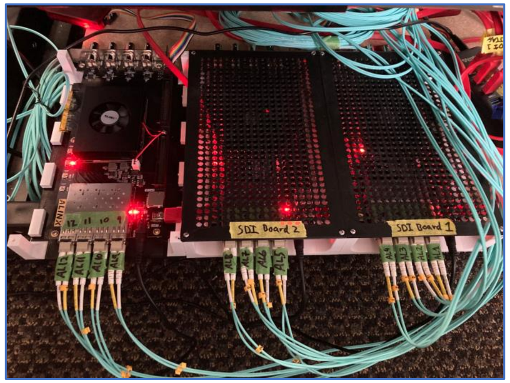
Outline

- Our Company and Capabilities
  - Team
- Our Current Product Line
  - Digitizers & Logic Modules
- Data Management Research
  - Performance networking
  - Data Storage
- Acknowledgements and Future Plans

SkuTek Instrumentation Makes… Instrumentation!

- The Team
  - Full time: 3 Research Engineers with Physics Backgrounds
  - Part time: 2 Other Senior Engineers, 1 Manager, 1 EE consultant
  - Interns rotating in and out constantly
- Our Focus
  - Instrumentation & Data Acquisition (DAQ) for High Energy Physics, Astrophysics, and Nuclear Physics.
- Our Capabilities: full end-end DAQ expertise
  - Electronic Design
  - Firmware Development
  - Digital Pulse Processing
  - Detector Assemblies
  - High Speed Networking
  - High Performance I/O





SkuTek Instrumentation
Where Science Meets Industry

We Serve National and International Customers

UNIVERSITY of ROCHESTER

Los Alamos National Laboratory

Brookhaven National Laboratory

HOUGHTON COLLEGE

Fermilab

AUGUSTANA UNIVERSITY

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Albert Einstein Center for Fundamental Physics

NSCL National Superconducting Cyclotron Laboratory

SFU SIMON FRASER UNIVERSITY ENGAGING THE WORLD

BROWN

UNIVERSITÄT BERN

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

WIENER Plein & Baus, Corp.

NORTHROP GRUMMAN

Raja Ramanna Centre for Advanced Technology

UNIVERSITÉ LAVAL

We Specialize in High Performance Digitizers

FemtoDAQ Family of Benchtop Digitizers

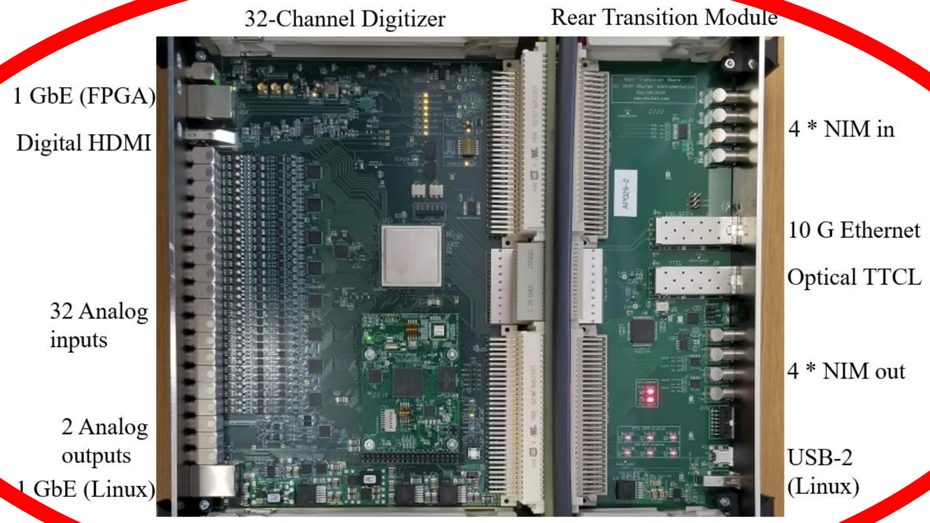
FemtoDAQ Kingfisher
(formally DDC10)

FemtoDAQ Merlin
(formally FemtoDAQ+)

FemtoDAQ Vireo
(prototype)

FemtoDAQ Classic

Chickadee-32 High Density Digitizer



What is Data Management?

Scientific Data Demands are getting Bigger

1) The DOE's Energy Science Network (ESNet) estimates data rates and volumes will increase by several orders of magnitude this decade. [6]
   - Driven by new instrumentation, larger channel counts, AI & machine learning, etc

2) There is growing adoption of the "Distributed Computing Infrastructure" (DCI) data model.
   - Data processing occurs at High Performance Computing (HPC) centers, often geographically separated from the experimental facility.
   - Examples: Square Kilometer Array, Cherenkov Telescope Array, Linac Coherent Light Source (LCLS), Gamma Ray Energy Tracking Array (GRETA), LuxZeplin, etc. [2,3,4,6]

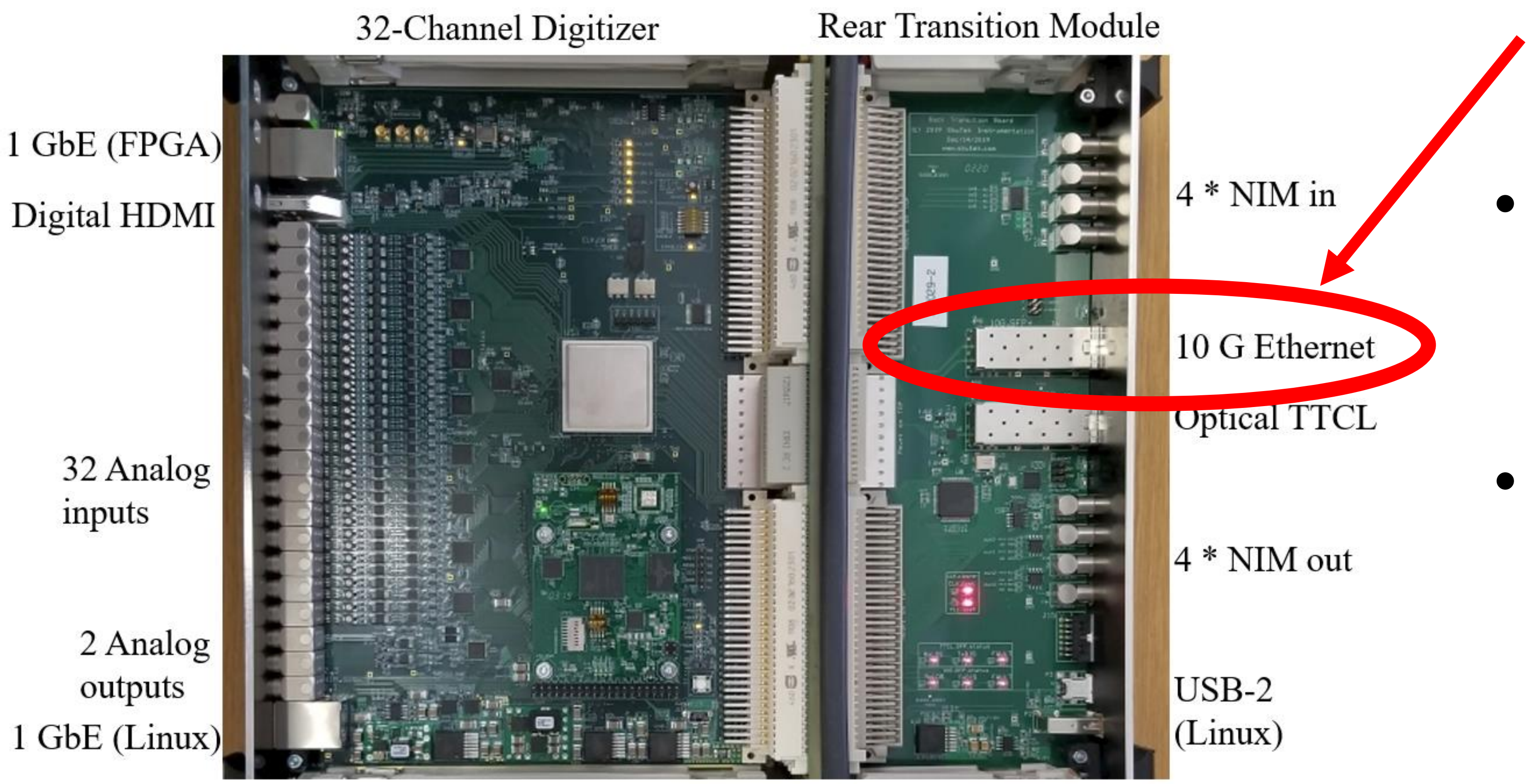
Takeaway: DAQ systems must account for this new paradigm.

The Goals:

1. To offer the DOE and our other customers end-end DAQ Solutions which can operate at 100 Gbps in bursts.
   - This means connecting our digitizers products to their eventual destination at an HPC facility.

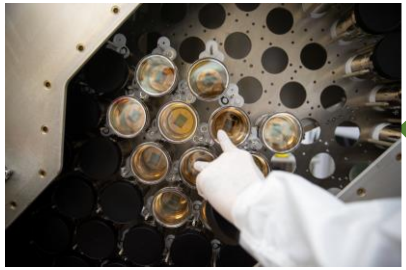
2. We want to collaborate to establish future DOE data management standards
   - ESNet’s 2022 report on readiness documents the urgent need for a standard API and reduction of adhoc custom systems developed by each lab. [6]

Modern digitizers can Produce a Lot of Data



- 10 Gbps readout from our Chickadee-32 digitizer
- Up to 1.2GB per 32 channels per second.
- Imagine 1000 channels…
  - >100TB / hour

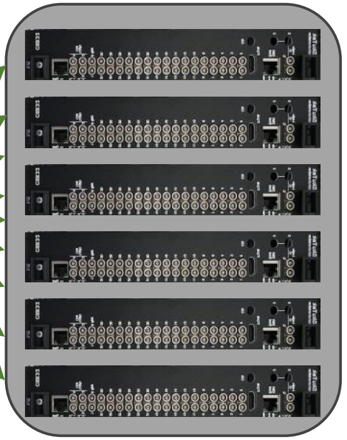
The Modern End-End Pipeline



Detector

(pictured: LuxZeplin)

Analog Signal



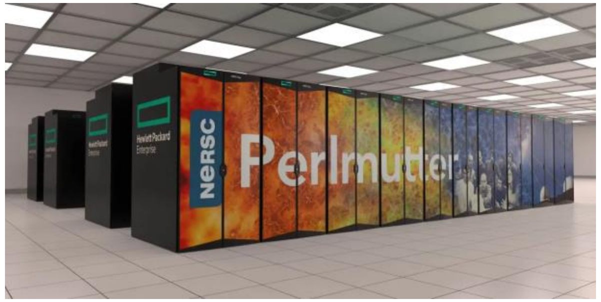
The DAQ

(digitizers)

Digitized UDP Streams

?

File Transfer



Data Analysis at HPC Center

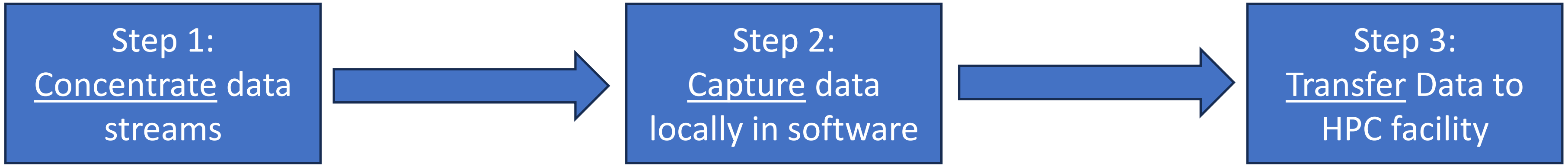
(pictured: NERSC)

Step 1: Concentrate data streams

Step 2: Capture data locally in software

Step 3: Transfer Data to HPC facility

Our Three Steps to Data Management



Step 1:
Concentrate data streams

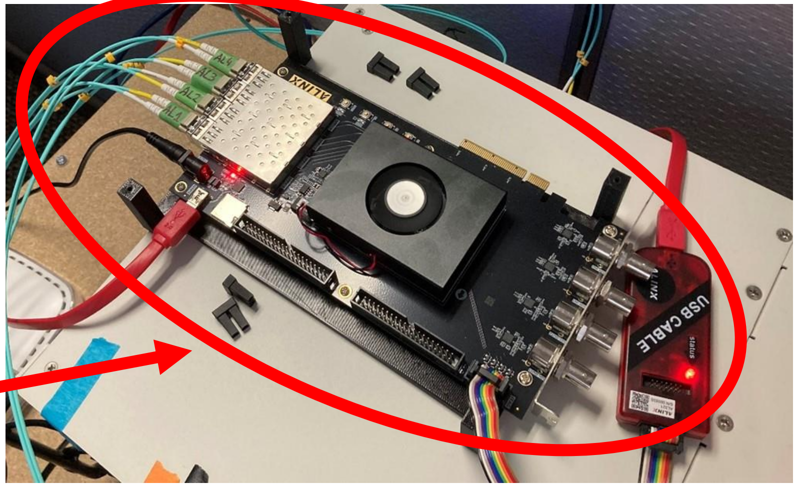
Step 2:
Capture data locally in software

Step 3:
Transfer Data to HPC facility

- 1) **Concentrate** multiple 10G digitizer outputs to fewer 100G cables.
- 2) **Capture** the consolidated data streams and save to files.
- 3) **Transfer** files to an HPC facility.

Step 0: We Made a 100G data source

- We needed to reliably generate 100G UDP data streams to test our system.
- Further, these streaming sources had to mimic digitizer streaming behavior precisely!



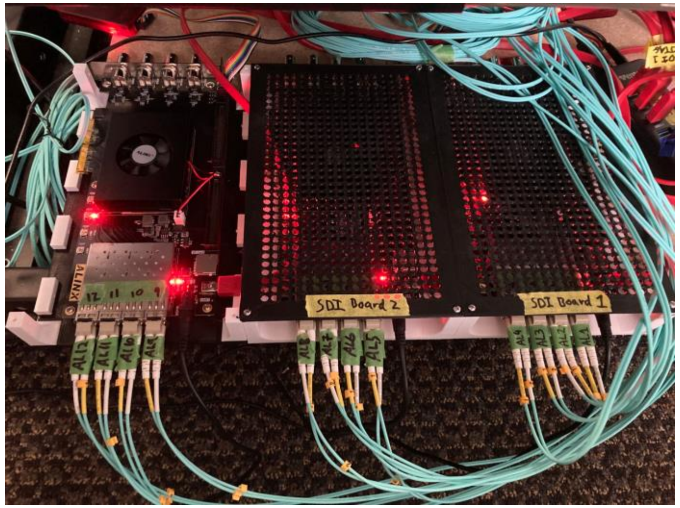
The Solution:

We developed an FPGA-based “**UDP Cannon**”

- Each port can stream randomized “bursts” or maintain steady-state between 0-10 Gbps.
- Packet content is deterministic to easily check for errors.

Next step: implement GRETA’s header format so this system will be GRETA compatible.

I’m so proud of my team for this accomplishment. They developed it so quickly and it’s still rock-solid reliable.



Tiled Design allows multiple boards operating in parallel. Above configuration can stream up to 120 Gbps

Stage 1: Concentration – Can Commercial Hardware do it?

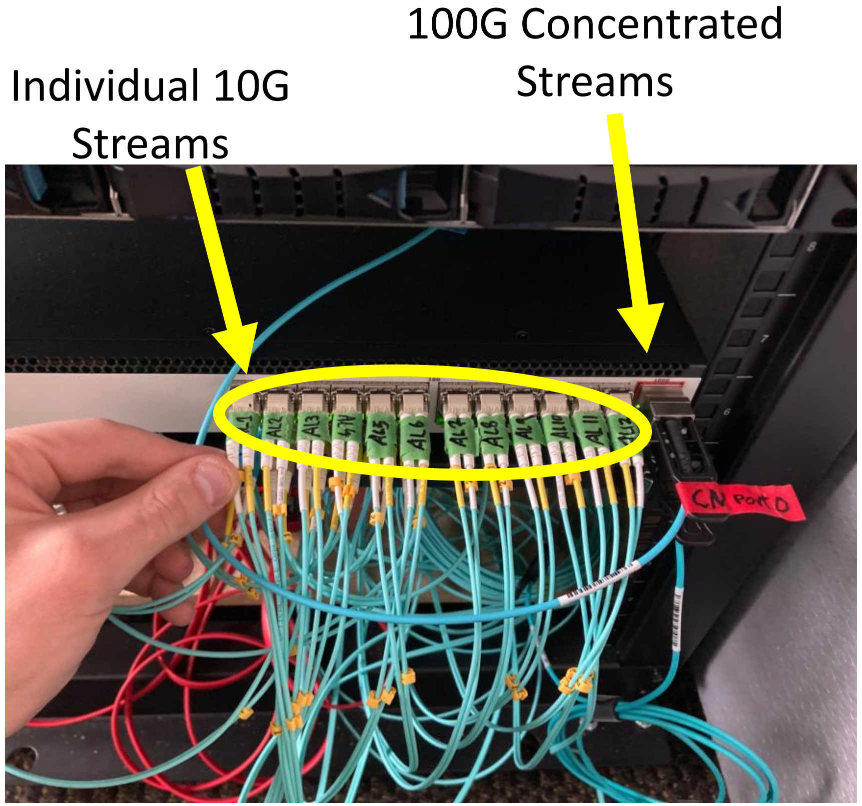
There was an open question whether commercial hardware could operate at 100G without losing data.

- ESNet has published accounts of cheaper “shallow buffered” switches being unable to handle 100G of traffic.

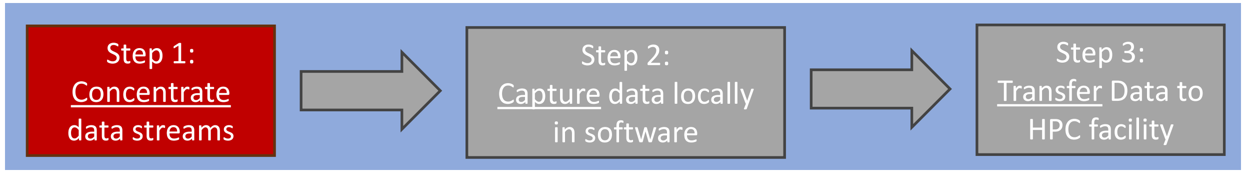
We wanted to avoid developing our own concentration hardware!

- Most of the aforementioned tests used the Transmission Control Protocol (TCP) or were in “over subscribed” applications. [5]

We hypothesized that these problems wouldn’t manifest with our User Datagram Protocol (UDP) streams and with appropriate configuration.



The Commercial Networking Switch we purchased. ~$2500 with shallow buffering



Step 1: Concentrate data streams

Step 2: Capture data locally in software

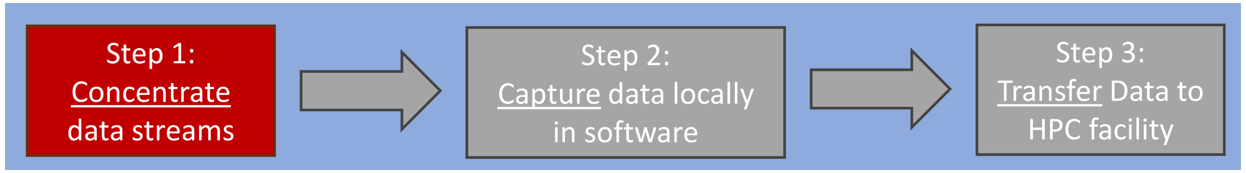
Step 3: Transfer Data to HPC facility

And we were right! Although it took awhile…

- Initially, we saw up to 5% packet loss in the switch concentrating ten 10G streams into a single 100G wire.
- We had to develop network analysis software first.
  - More on this in a second!
- We met with the vendor and over several weeks reconfigured the switch for optimal UDP streaming.

The switch now concentrates 100 Gbps losslessly!

Test Time Period (real)	Data Rate	Packet Size	Total Quantity Streamed	Dropped Packets
26.5 Hours	97.18 Gbps	8800 Bytes	1.15 PB	0



Step 1: Concentrate data streams

Step 2: Capture data locally in software

Step 3: Transfer Data to HPC facility

Stage 2: Onward to Linux Networking

100G capable hardware != 100G capable computer.

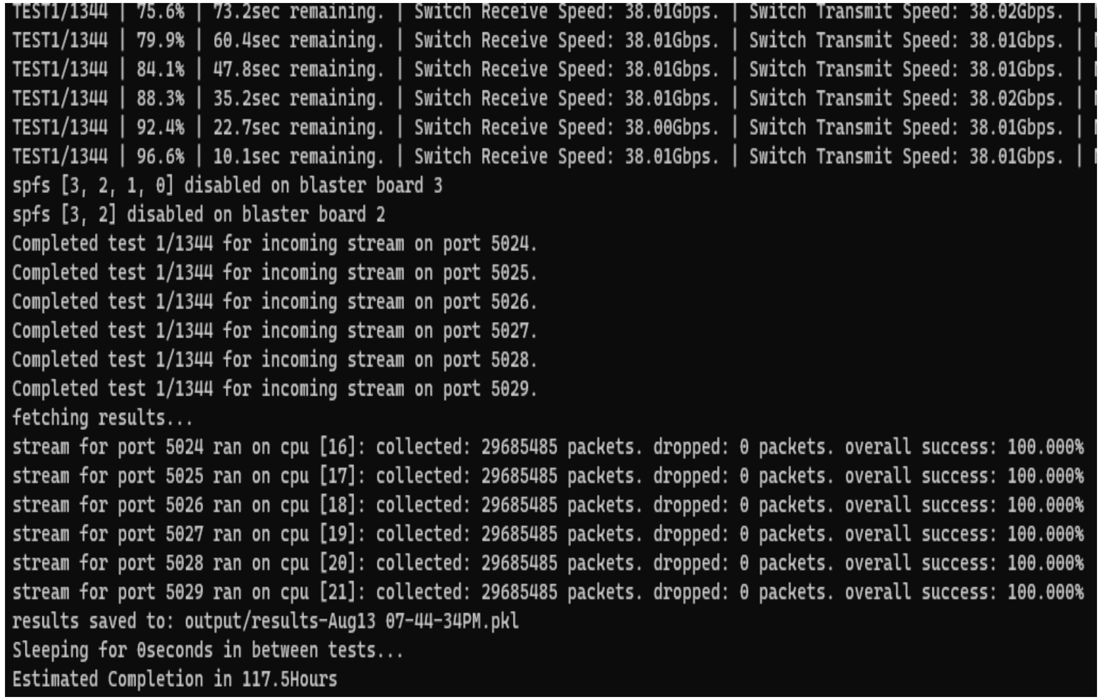
- Performant networking requires “tuning”.
- This is difficult task! We found at least 60 parameters that influence Linux networking speeds.

We developed software to automate testing different network configurations.

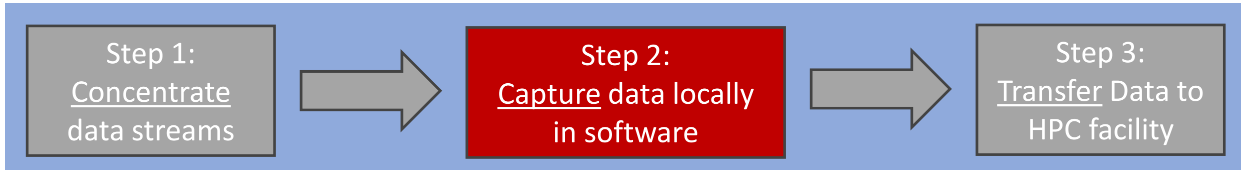
This software can:

1. Automatically modify CPU, NIC, kernel, & streaming parameters according to a config file.
2. Identify where packets were lost (switch, NIC, networking stack, etc)
3. Verify packet integrity and ordering
4. Monitors hardware utilization

- It operates similarly to *iperf*, but it uses our UDP Cannon



```
TEST1/1344 | 75.6% | 73.2sec remaining. | Switch Receive Speed: 38.01Gbps. | Switch Transmit Speed: 38.02Gbps. |
TEST1/1344 | 79.9% | 60.4sec remaining. | Switch Receive Speed: 38.01Gbps. | Switch Transmit Speed: 38.01Gbps. |
TEST1/1344 | 84.1% | 47.8sec remaining. | Switch Receive Speed: 38.01Gbps. | Switch Transmit Speed: 38.01Gbps. |
TEST1/1344 | 88.3% | 35.2sec remaining. | Switch Receive Speed: 38.01Gbps. | Switch Transmit Speed: 38.02Gbps. |
TEST1/1344 | 92.4% | 22.7sec remaining. | Switch Receive Speed: 38.00Gbps. | Switch Transmit Speed: 38.01Gbps. |
TEST1/1344 | 96.6% | 10.1sec remaining. | Switch Receive Speed: 38.01Gbps. | Switch Transmit Speed: 38.01Gbps. |
spfs [3, 2, 1, 0] disabled on blaster board 3
spfs [3, 2] disabled on blaster board 2
Completed test 1/1344 for incoming stream on port 5024.
Completed test 1/1344 for incoming stream on port 5025.
Completed test 1/1344 for incoming stream on port 5026.
Completed test 1/1344 for incoming stream on port 5027.
Completed test 1/1344 for incoming stream on port 5028.
Completed test 1/1344 for incoming stream on port 5029.
fetching results...
stream for port 5024 ran on cpu [16]: collected: 29685485 packets. dropped: 0 packets. overall success: 100.000%
stream for port 5025 ran on cpu [17]: collected: 29685485 packets. dropped: 0 packets. overall success: 100.000%
stream for port 5026 ran on cpu [18]: collected: 29685485 packets. dropped: 0 packets. overall success: 100.000%
stream for port 5027 ran on cpu [19]: collected: 29685485 packets. dropped: 0 packets. overall success: 100.000%
stream for port 5028 ran on cpu [20]: collected: 29685485 packets. dropped: 0 packets. overall success: 100.000%
stream for port 5029 ran on cpu [21]: collected: 29685485 packets. dropped: 0 packets. overall success: 100.000%
results saved to: output/results-Aug13 07-44-34PM.pkl
Sleeping for 0seconds in between tests...
Estimated Completion in 117.5Hours
```



Step 1:
Concentrate data streams
Step 2:
Capture data locally in software
Step 3:
Transfer Data to HPC facility

Our Tuning Software also has an analysis GUI

We routinely use it to analyze hundreds of network configurations

- We generate matrix plots identify correlations

We achieved high speed performant networking in Linux.

0-60 Gbps : 100% success

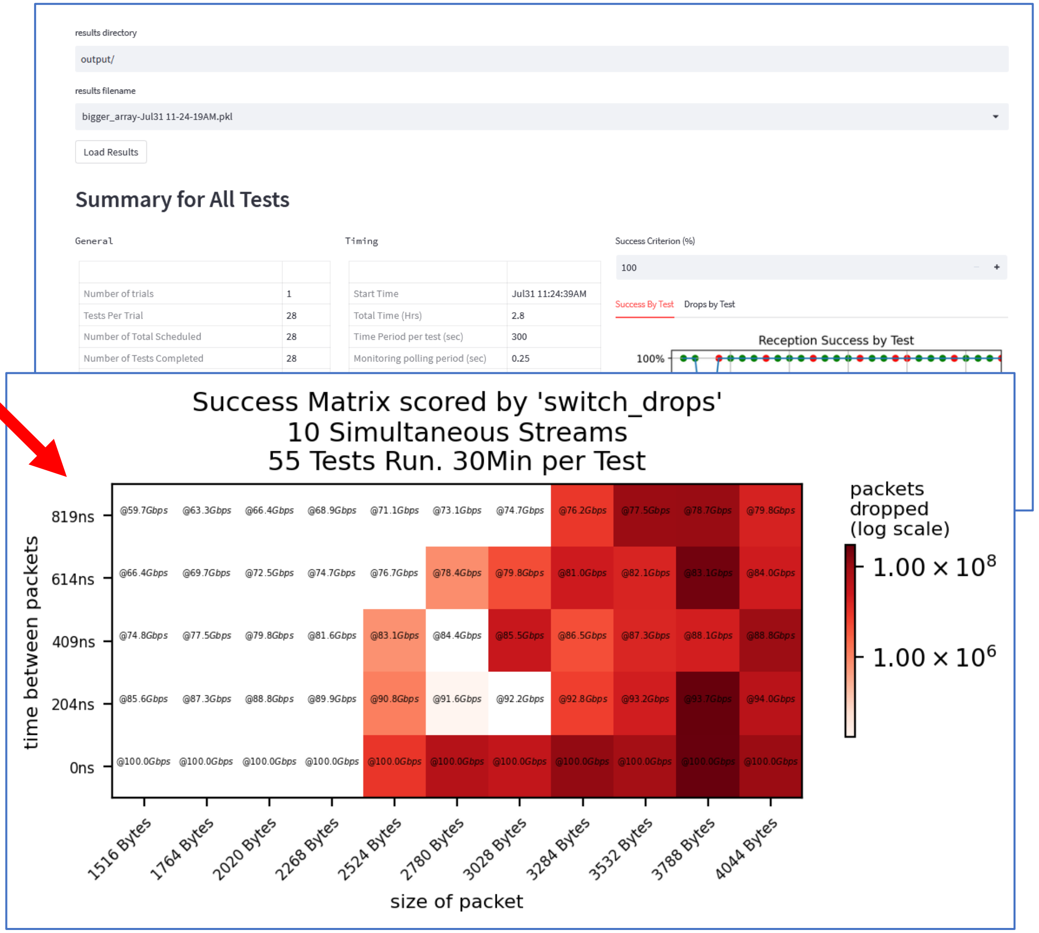
60-80 Gbps : >99.999% success

100 Gbps: 90% success

*Results for steady-state streaming with jumbo packets *(no bursts or variation)*

Future Work:

- We will begin testing “burst” streaming which is more typical in pulsed accelerators



results directory

output/

results filename

bigger_array-Jul31 11-24-19AM.pkl

Load Results

Summary for All Tests

General

Number of trials	1
Tests Per Trial	28
Number of Total Scheduled	28
Number of Tests Completed	28

Timing

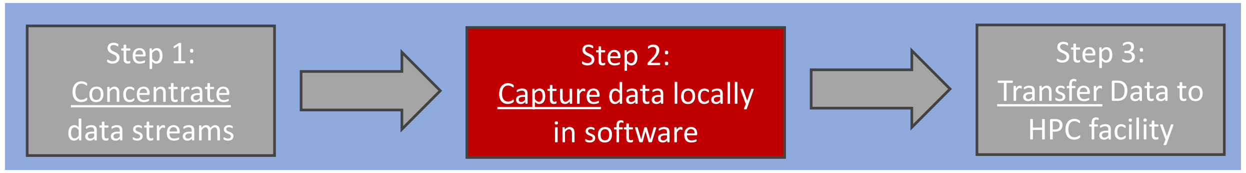
Start Time	Jul31 11:24:39AM
Total Time (Hrs)	2.8
Time Period per test (sec)	300
Monitoring polling period (sec)	0.25

Success Criterion (%)

100

Success Matrix scored by 'switch_drops'
10 Simultaneous Streams
55 Tests Run. 30Min per Test

time between packets	1516 Bytes	1764 Bytes	2020 Bytes	2268 Bytes	2524 Bytes	2780 Bytes	3028 Bytes	3284 Bytes	3532 Bytes	3788 Bytes	4044 Bytes
819ns	@59.7Gbps	@63.3Gbps	@66.4Gbps	@68.9Gbps	@71.1Gbps	@73.1Gbps	@74.7Gbps	@76.2Gbps	@77.5Gbps	@78.7Gbps	@79.8Gbps
614ns	@66.4Gbps	@69.7Gbps	@72.5Gbps	@74.7Gbps	@76.7Gbps	@78.4Gbps	@79.8Gbps	@81.0Gbps	@82.1Gbps	@83.1Gbps	@84.0Gbps
409ns	@74.8Gbps	@77.5Gbps	@79.8Gbps	@81.6Gbps	@83.1Gbps	@84.4Gbps	@85.5Gbps	@86.5Gbps	@87.3Gbps	@88.1Gbps	@88.8Gbps
204ns	@85.6Gbps	@87.3Gbps	@88.8Gbps	@89.9Gbps	@90.8Gbps	@91.6Gbps	@92.2Gbps	@92.8Gbps	@93.2Gbps	@93.7Gbps	@94.0Gbps
0ns	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps	@100.0Gbps



Step 1: Concentrate data streams

Step 2: Capture data locally in software

Step 3: Transfer Data to HPC facility

Once Received, We have to Write Data to Disk

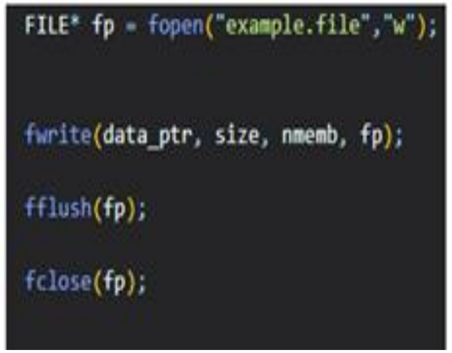
Writing to disk efficiently presents 2 major challenges:

1. Performant writing strategies vary widely depending on the storage configuration *(device contension, solid state vs hard disk, etc)*
2. Different experiments will have different storage requirements.
   - High speed solid state drives (SSD) can be expensive. Even today SSDs are still 6-10x more expensive per GB

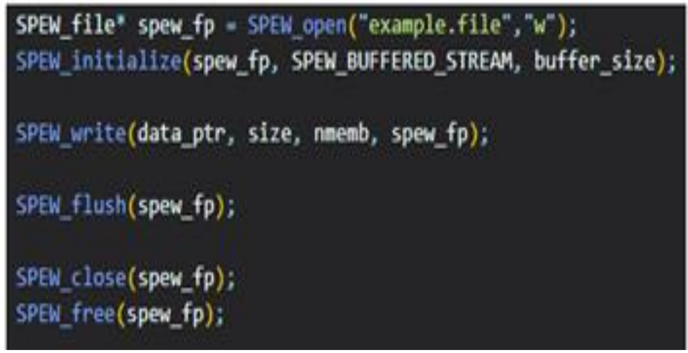
The solution:

Skutek PErformance library for Writing (SPEW)

- SPEW is a drop-in replacement C library for Linux filestream writing



```
FILE* fp = fopen("example.file","w");


fwrite(data_ptr, size, nmemb, fp);

fflush(fp);

fclose(fp);
```

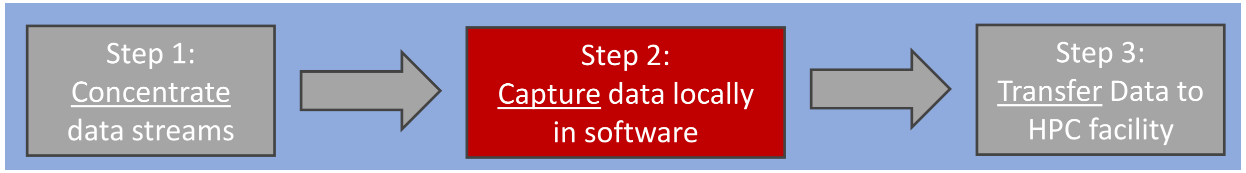


```
SPEW_file* spew_fp = SPEW_open("example.file","w");
SPEW_initialize(spew_fp, SPEW_BUFFERED_STREAM, buffer_size);

SPEW_write(data_ptr, size, nmemb, spew_fp);

SPEW_flush(spew_fp);

SPEW_close(spew_fp);
SPEW_free(spew_fp);
```

Behind the scenes, SPEW uses threading, buffering, and preallocation depending on the mode it is in.

Up to 20-40% performance improvements over unoptimized writing.

SPEW will be the bedrock of all future software for this project.



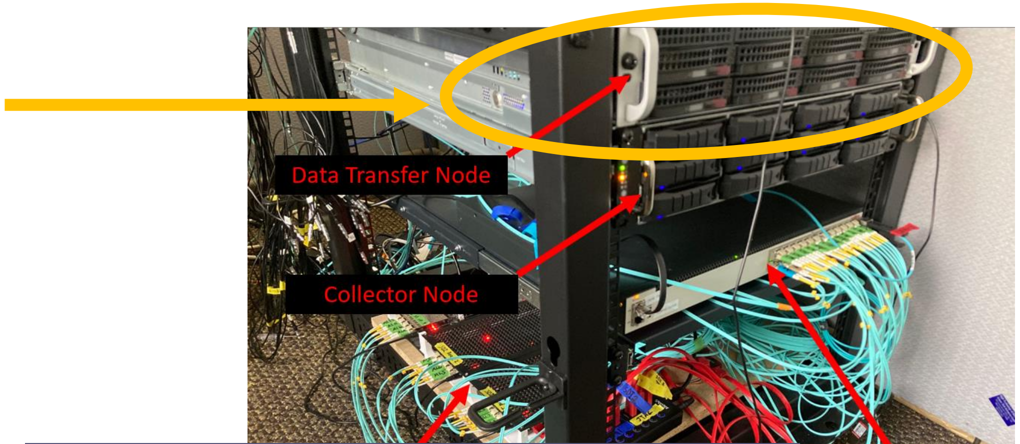
Step 1: Concentrate data streams

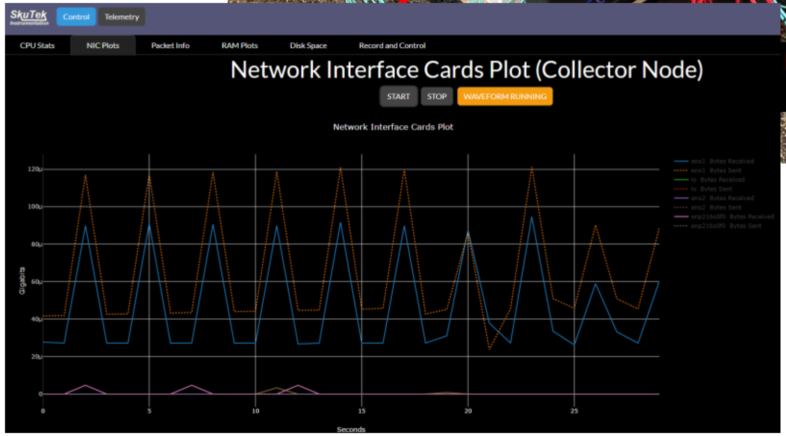
Step 2: Capture data locally in software

Step 3: Transfer Data to HPC facility

Stage 3: Getting Our Data to NERSC

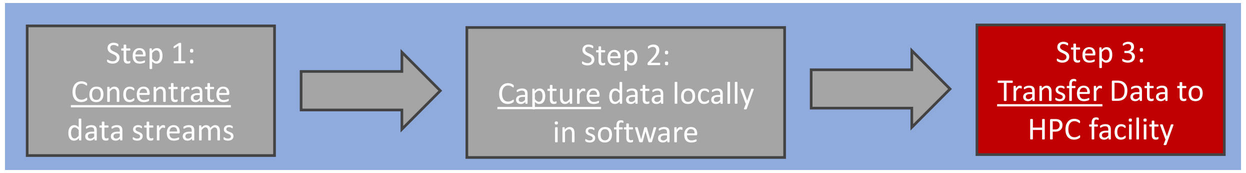
- We have hardware in house for our Data Transfer Node (DTN) and acquired access to NERSC.
- We developed a REST API to initiate file transfers using Globus to NERSC.
  - Transfer speeds are currently limited by our internet connection.
- We developed a web-based hardware monitor to monitor transfer speeds.
- We're seeking partnership to test our Data Transfer Node with access to ESNet.





Network Interface Cards Plot (Collector Node)

*software was being tested on Collector Node in this screenshot, but will be identical on DTN.


Step 1: Concentrate data streams

Step 2: Capture data locally in software

Step 3: Transfer Data to HPC facility

Thank you!

My Colleagues: Jim Vitkus, Jackson Hebel, Wojtek Skulski, David Miller

Our EE consultant: Eryk Druszkiewicz

Extra thanks to our interns!

(left) Joshua Rosenberg: Mechanical Engineering @RIT
(right) Edmond Tan: Computer Engineering Technology @RIT

Michelle Shinn and Manouchehr Farkhondeh

Mario Cromaz and the team at ESNet for being so approachable

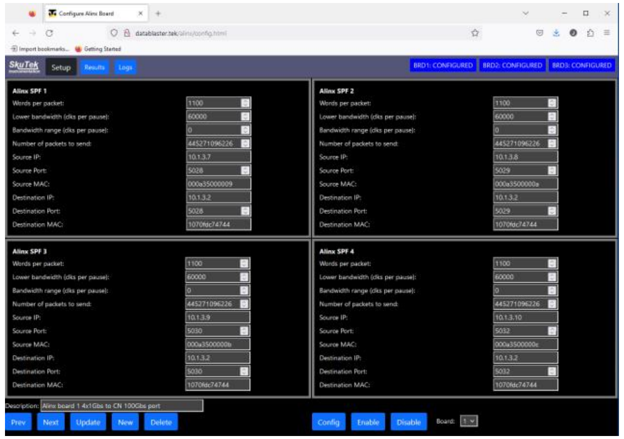
References

1. Sands, A. E. (2017). Managing Astronomy Research Data: Data Practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects. UCLA. Retrieved from https://escholarship.org/uc/item/80p1w0pm
2. Wang, Ruonan, Tobar, Rodrigo, Dolensky,… Processing Full-Scale Square Kilometre Array Data on the Summit Supercomputer. United States. https://doi.org/10.1109/SC41405.2020.00006
3. Lamanna, G., Antonelli, L. A., Contreras, J. L., Knödlseder, J., Kosack, K., Neyroud, N., … Zoli, A. (2015). Cherenkov Telescope Array Data Management. doi:10.48550/ARXIV.1509.01012
4. Cromaz, M., Dart, E., Pouyoul, E., & Jansen, G. (2021). Simple and Scalable Streaming: TheGRETA Data Pipeline*. EPJ Web Conf., 251, 04018
5. “FasterData.” Fasterdata.es.net, United States Department of Energy, https://fasterdata.es.net/.
6. “ESNet.” www.es.net, United States Department of Energy, https://www.es.net/science-engagement/science-requirements-reviews/esnet-network-requirements-reviews/bes-requirements-review-2022/

Backup Slides

Our UDP Cannon will become a spin-off product

- The UDP Cannon has been **invaluable** in our own networking stress tests.
- We believe it has commercial value in other industries: such as networking and datacenter engineering!
    - It's ideal to stress-test networking systems such as switches or routers.
    - Firmware solution means zero tuning or software optimization required!
- This will be the first commercial product produced by this grant!



Alinx SPF 1
Words per packet: 1100
Lower bandwidth (clks per pause): 60000
Bandwidth range (clks per pause): 0
Number of packets to send: 445271096226
Source IP: 10.1.3.7
Source Port: 5028
Source MAC: 000a35000009
Destination IP: 10.1.3.2
Destination Port: 5028
Destination MAC: 1070fdc74744

Alinx SPF 2
Words per packet: 1100
Lower bandwidth (clks per pause): 60000
Bandwidth range (clks per pause): 0
Number of packets to send: 445271096226
Source IP: 10.1.3.8
Source Port: 5029
Source MAC: 000a3500000a
Destination IP: 10.1.3.2
Destination Port: 5029
Destination MAC: 1070fdc74744

Alinx SPF 3
Words per packet: 1100
Lower bandwidth (clks per pause): 60000
Bandwidth range (clks per pause): 0
Number of packets to send: 445271096226
Source IP: 10.1.3.9
Source Port: 5030
Source MAC: 000a3500000b
Destination IP: 10.1.3.2
Destination Port: 5030
Destination MAC: 1070fdc74744

Alinx SPF 4
Words per packet: 1100
Lower bandwidth (clks per pause): 60000
Bandwidth range (clks per pause): 0
Number of packets to send: 445271096226
Source IP: 10.1.3.10
Source Port: 5032
Source MAC: 000a3500000c
Destination IP: 10.1.3.2
Destination Port: 5032
Destination MAC: 1070fdc74744

Description: Alinx board 1 4x1Gbs to CN 100Gbs port
Board: 1

Web-based Graphical Interface lets users control each stream individually
Also uses REST API for automated control

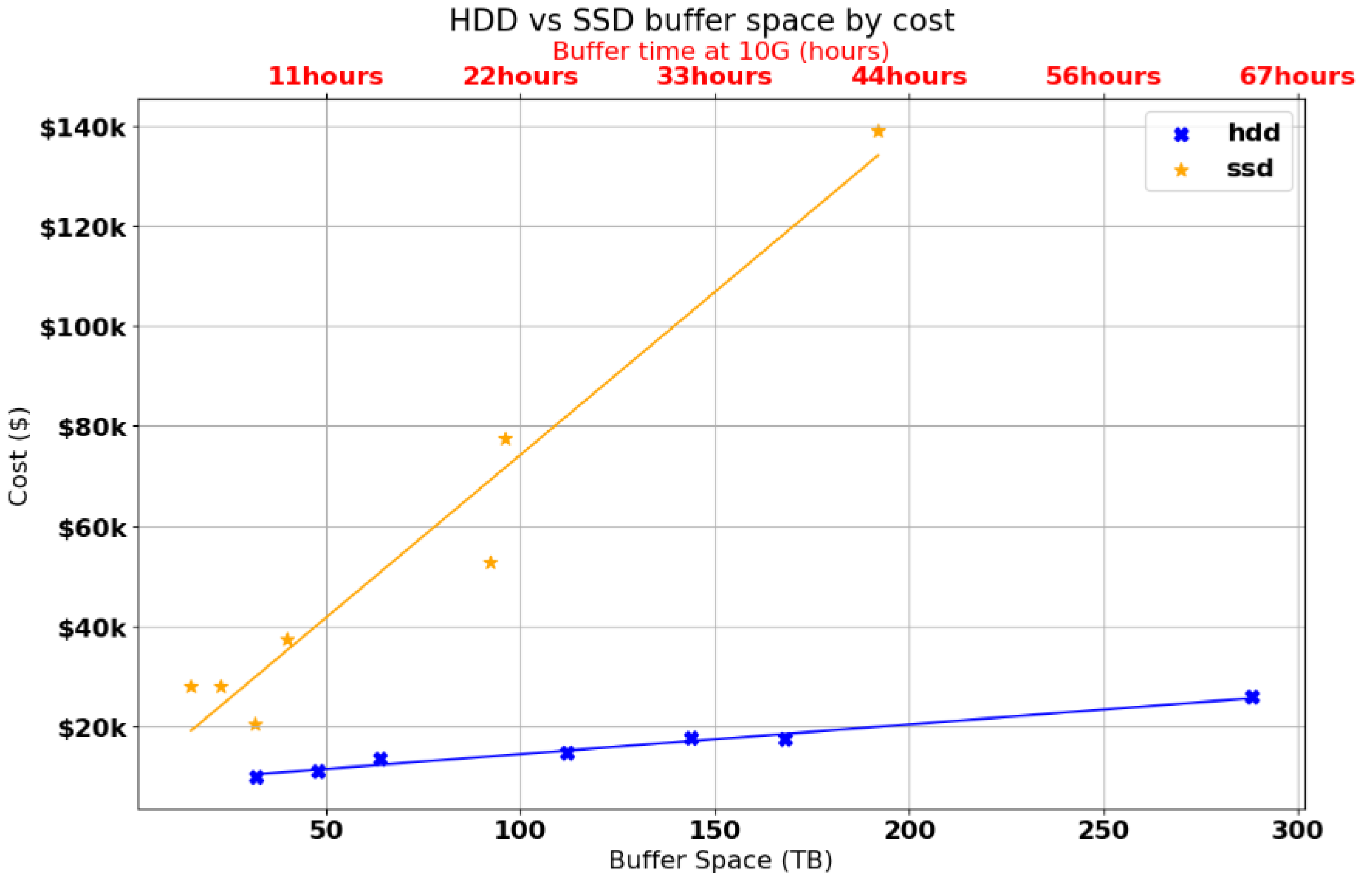
SkuTek Instrumentation

Prototype enclosure for a 4-board product
(designed by our intern!)

SSD vs HDD expense

- SSD storage servers are between 6-10x more expensive than harddrive ones.

- Harddrives are semi-competitive at storage speeds



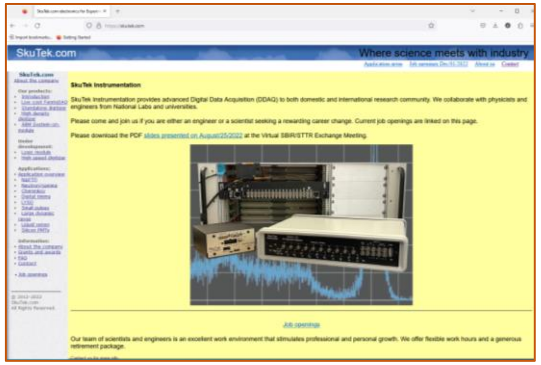
SkuTek Instrumentation
Where Science Meets Industry

TABA funding has been fantastic

- TABA funding has helped our business has develop considerably!

- We're building our brand by sponsoring conferences

- We have a new website focused on showcasing our products and providing support



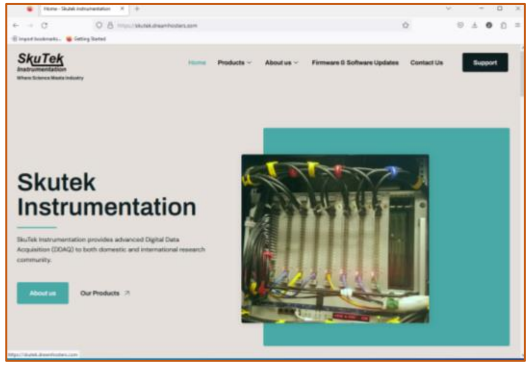
SkuTek Instrumentation

SkuTek Instrumentation provides advanced Digital Data Acquisition (DDAQ) to both domestic and international research community. We collaborate with physicists and engineers from National Labs and universities.

Please come and join us if you are either an engineer or a scientist seeking a rewarding career change. Current job openings are linked on this page.

Please download the PDF slides presented on August/25/2022 at the Virtual SBIR/STTR Exchange Meeting.

Our team of scientists and engineers is an excellent work environment that stimulates professional and personal growth. We offer flexible work hours and a generous retirement package.


Home - Skutek Instrumentation
SkuTek
Instrumentation
Where Science Meets Industry
Home
Products
About us
Firmware & Software Updates
Contact Us
Support
About us
Our Products

Skutek Instrumentation

SkuTek Instrumentation provides advanced Digital Data Acquisition (DDAQ) to both domestic and international research community.

Visit us at www.skutek.com!