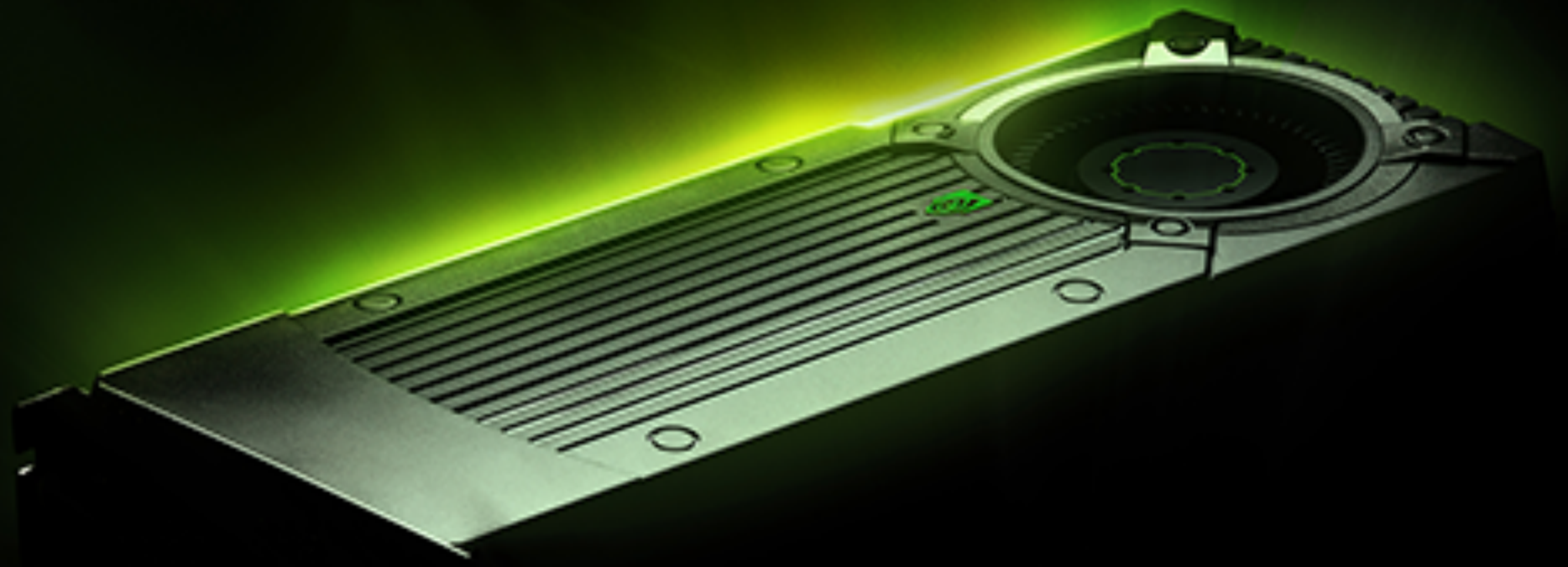


From Playstation to Supercomputers: Accelerating Physics with GPUs

Wes Gohn
University of Kentucky

6 January 2019

Introduction

- I am a particle physicist (and hence the talk will be mainly about that)
- I work on the Muon g-2 experiment (so more details about that)
- There are many great applications of GPUs in physics outside of HEP and g-2, which I am not an expert on, but I will mention some interesting use cases.

Introduction

- I am a particle physicist (and hence the talk will be mainly about that)
- I work on the Muon g-2 experiment (so more details about that)
- There are many great applications of GPUs in physics outside of HEP and g-2, which I am not an expert on, but I will mention some interesting use cases.

But actually today is my last day, so I will also talk a bit about what I'm doing next…

Improvements in Computing Performance

Apollo 11 (1969)
1 MHz

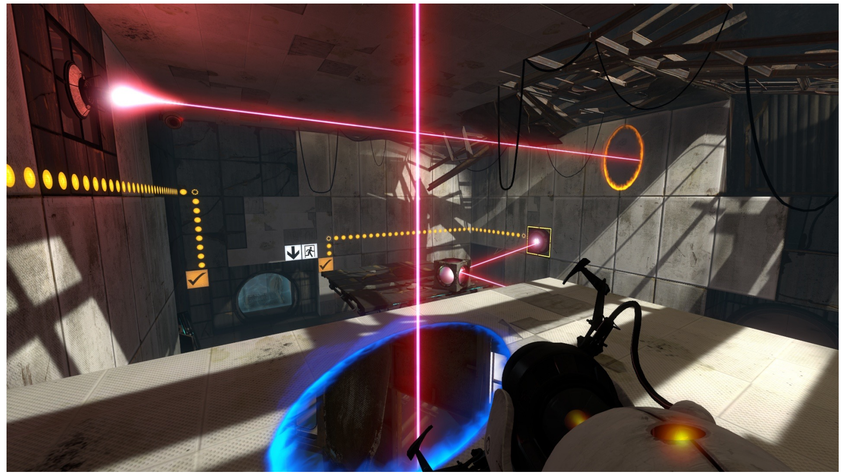
Macintosh Centris (1992)
40 MHz

Macbook Pro (2004)
1000 MHz

Macbook Pro (2018)
4600 MHz (x6)

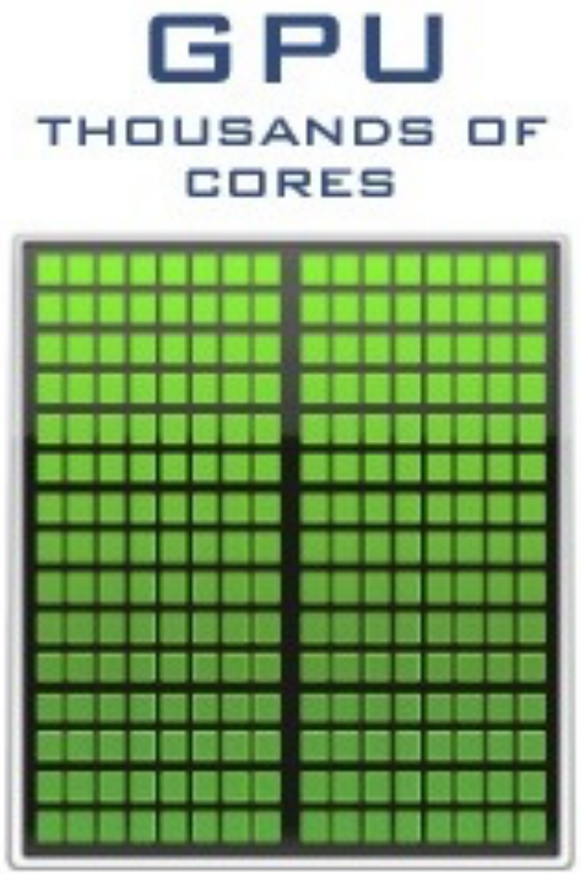
What is a GPU?

- GPU = Graphics Processing Unit
- Designed to make pretty graphics for video games
- Uses massive parallelization to process thousands of threads simultaneously
- Technical definition: "*A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creating of images in a frame buffer*" [Wikipedia]



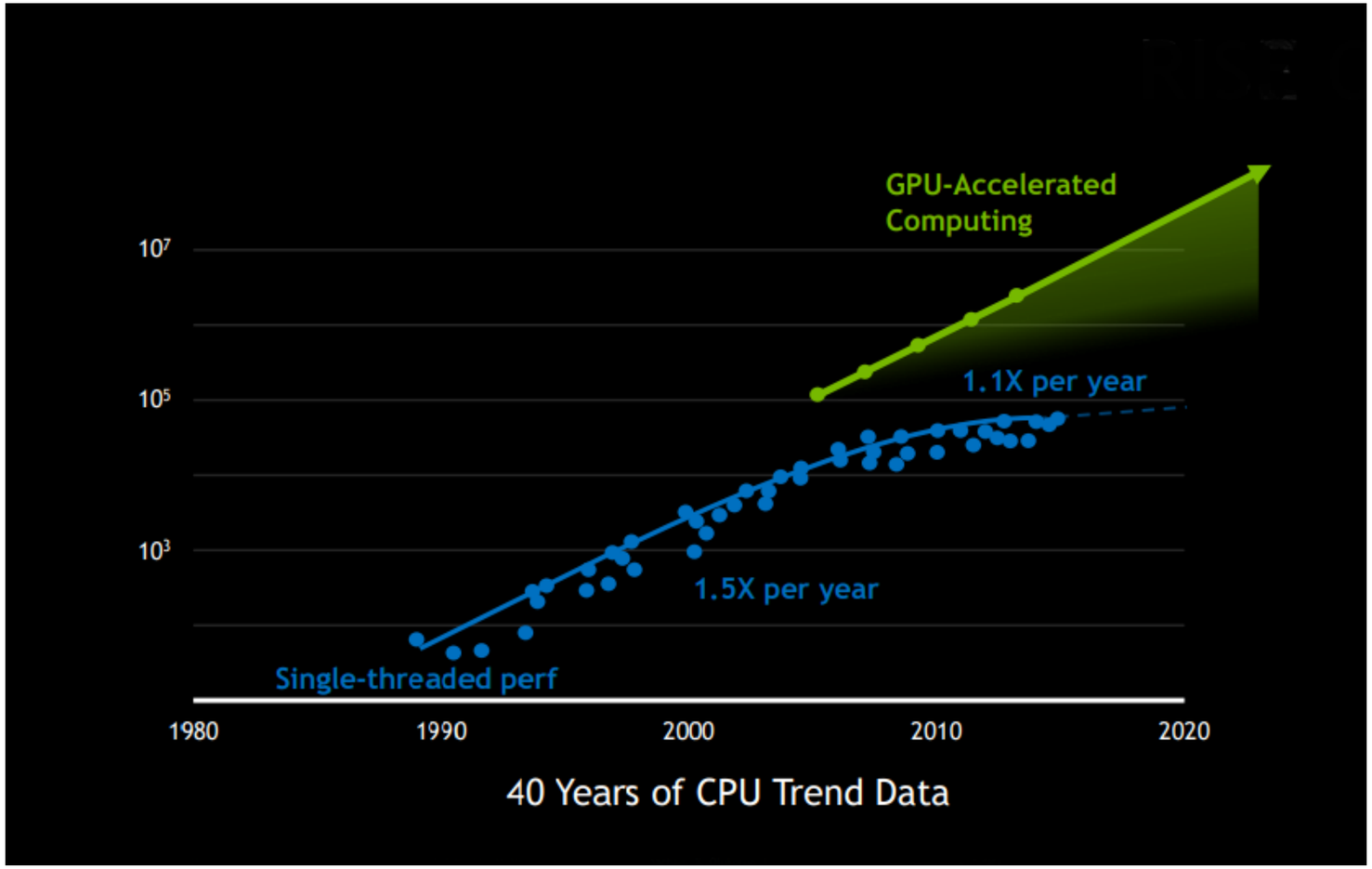
What is a GPU?

- GPU = Graphics Processing Unit
- Designed to make pretty graphics for video games
- Uses massive parallelization to process thousands of threads simultaneously
- Technical definition: "*A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creating of images in a frame buffer*" [Wikipedia]



Moore's Law is Dead without GPUs

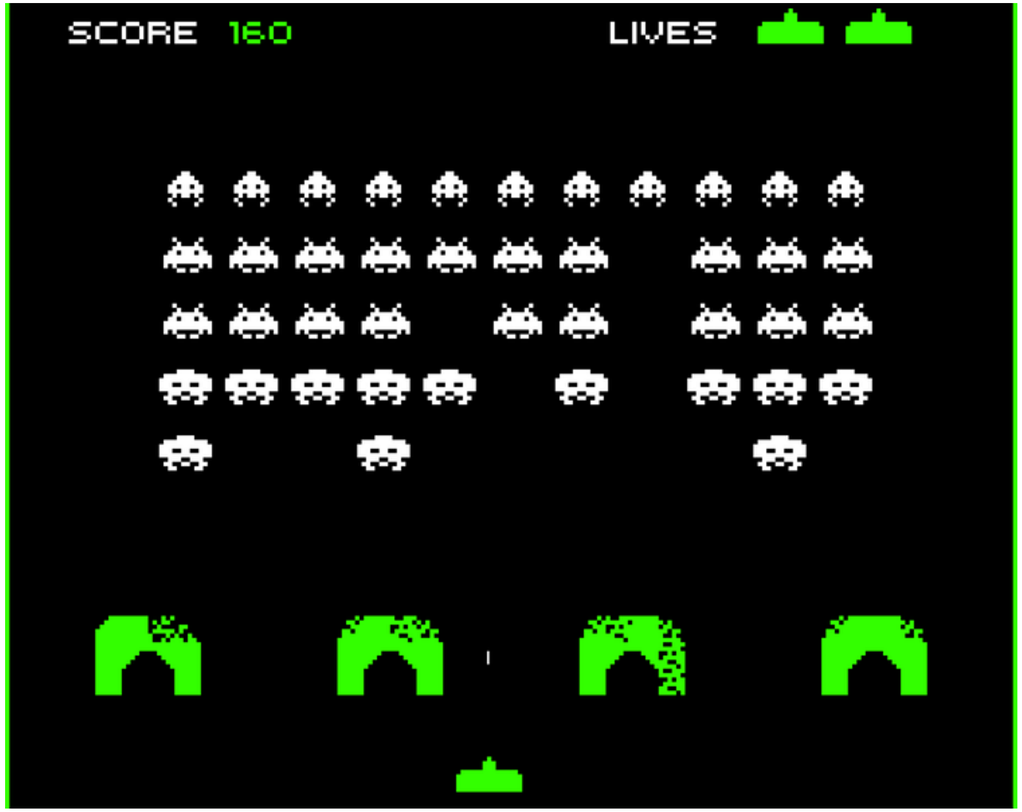
- Moore's Law: the number of transistors in a dense integrated circuit doubles about every two years



GPUs are the future of high performance computing!

History of GPUs

- 1970s: Arcade System Boards.
- 1980s: First video graphics cards with 16 bit color bitmaps
- 1990s: Real-time 3D graphics emerge
- 2000s: Nvidia GeForce3 capable of programmable shading
- 2010s: Virtual Reality, HPC, AI



History of GPUs

- 1970s: Arcade System Boards.
- 1980s: First video graphics cards with 16 bit color bitmaps
- 1990s: Real-time 3D graphics emerge
- 2000s: Nvidia GeForce3 capable of programmable shading
- 2010s: Virtual Reality, HPC, AI

History of GPUs

- 1970s: Arcade System Boards.
- 1980s: First video graphics cards with 16 bit color bitmaps
- 1990s: Real-time 3D graphics emerge
- 2000s: Nvidia GeForce3 capable of programmable shading
- 2010s: Virtual Reality, HPC, AI

History of GPUs

- 1970s: Arcade System Boards.
- 1980s: First video graphics cards with 16 bit color bitmaps
- 1990s: Real-time 3D graphics emerge
- 2000s: Nvidia GeForce3 capable of programmable shading
- 2010s: Virtual Reality, HPC, AI

History of GPUs

- 1970s: Arcade System Boards.
- 1980s: First video graphics cards with 16 bit color bitmaps
- 1990s: Real-time 3D graphics emerge
- 2000s: Nvidia GeForce3 capable of programmable shading
- 2010s: Virtual Reality, HPC, AI

GPUs in high performance computing

- In 2003 UIUC built a 500 gflop supercomputer by linking the graphics cards of 65 playstation 2 gaming consoles — total cost < $50k.

- NVIDIA introduced CUDA in 2007 — enables relatively simple C++ programming for GPU applications.

- In 2012 NVIDIA introduced it's TESLA line of GPUs: The first designed specifically for HPC rather than graphics.

- In 2018 Summit came online — the worlds fastest Supercomputer is based on GPU

GPUs in high performance computing

- In 2003 UIUC built a 500 gflop supercomputer by linking the graphics cards of 65 playstation 2 gaming consoles — total cost < $50k.

- NVIDIA introduced CUDA in 2007 — enables relatively simple C++ programming for GPU applications.

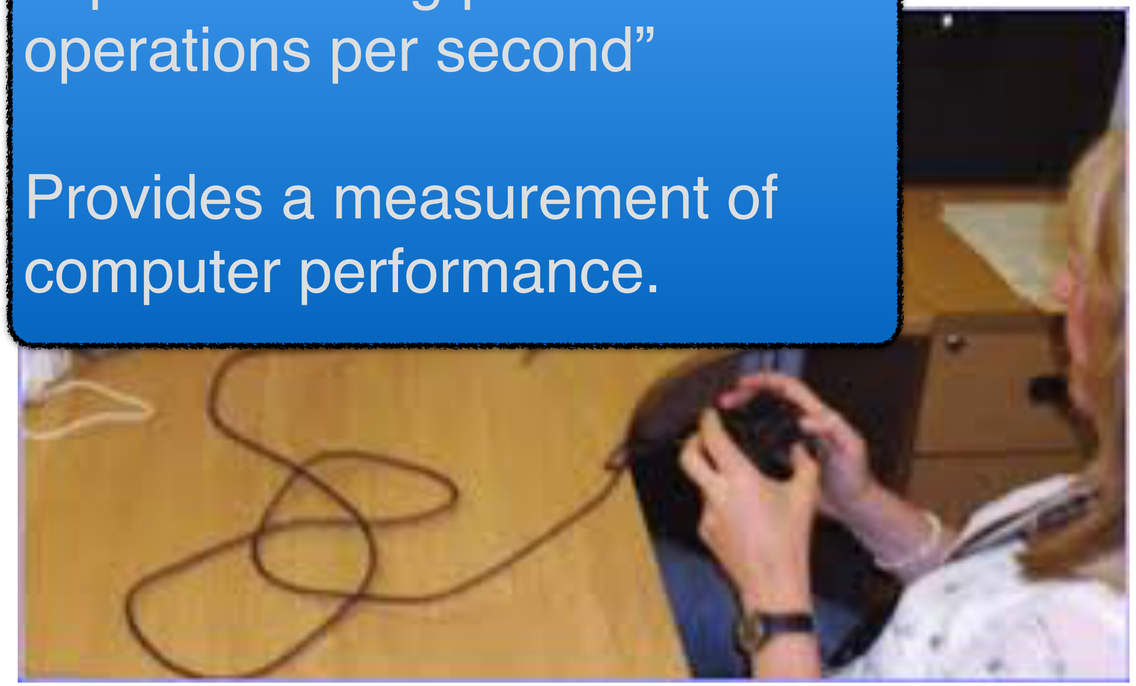
- In 2012 NVIDIA introduced it's TESLA line of GPUs: The first designed specifically for HPC rather than graphics.

- In 2018 Summit came online — the worlds fastest Supercomputer is based on GPU

What are "flops"?

flops = "floating point operations per second"

Provides a measurement of computer performance.



GPUs in high performance computing

- In 2003 UIUC built a 500 gflop supercomputer by linking the graphics cards of 65 playstation 2 gaming consoles — total cost < $50k.

- NVIDIA introduced CUDA in 2007 — enables relatively simple C++ programming for GPU applications.

- In 2012 NVIDIA introduced it's TESLA line of GPUs: The first designed specifically for HPC rather than graphics.

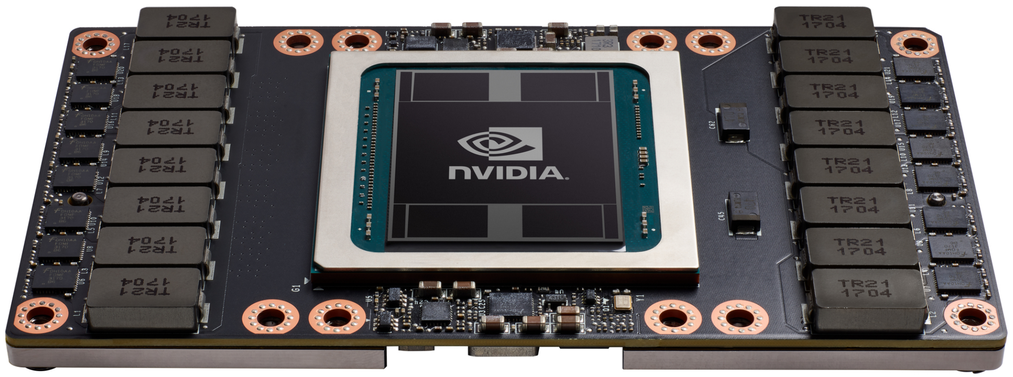
- In 2018 Summit came online — the worlds fastest Supercomputer is based on GPU

Current state-of-the-art hardware

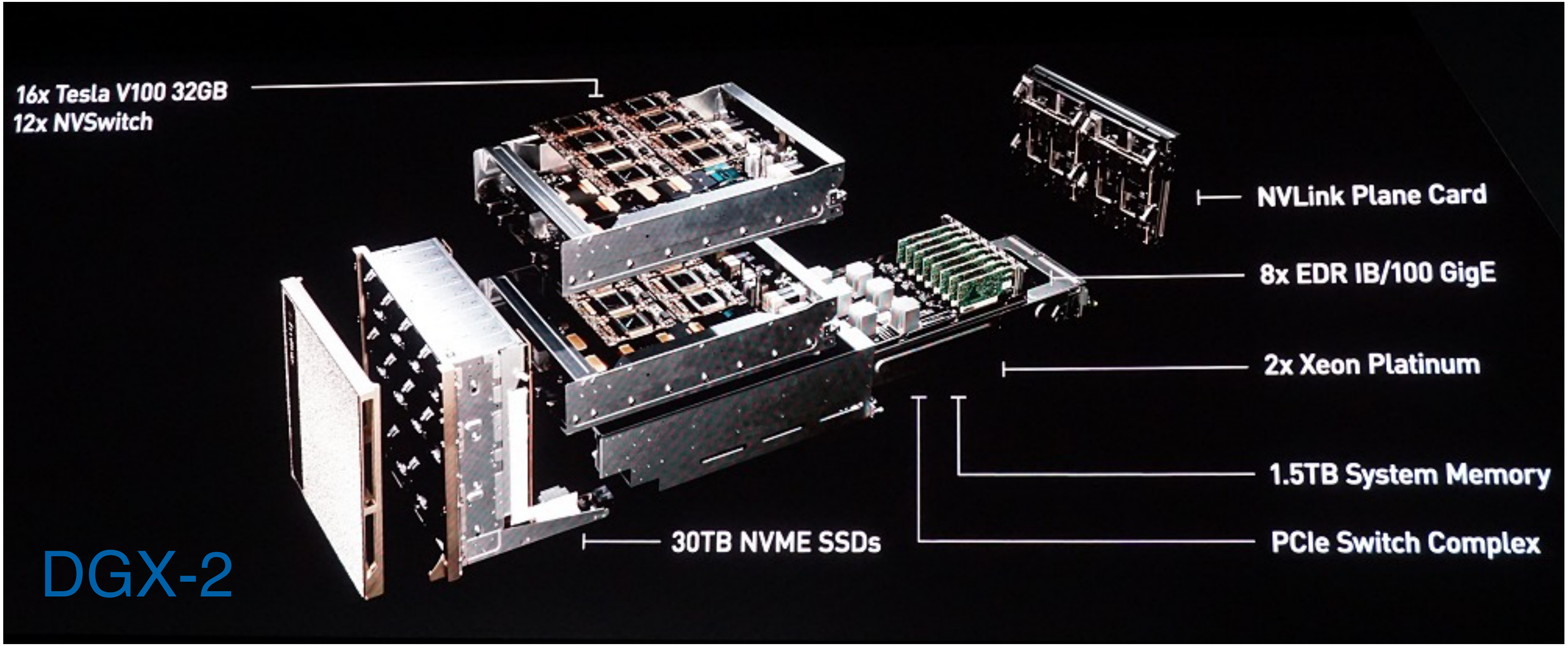
- NVIDIA Tesla V100
  - 7.8 tera-flops double precision
  - 5120 CUDA Cores
  - 15.7 tera-flops single precision
  - 125 tera-flops deep learning

- Can attach via NVLINK (300 GB/s) or PCIE (30 GB/s)

- Provides “high performance computing in your desktop”



Current state-of-the-art hardware



Summit

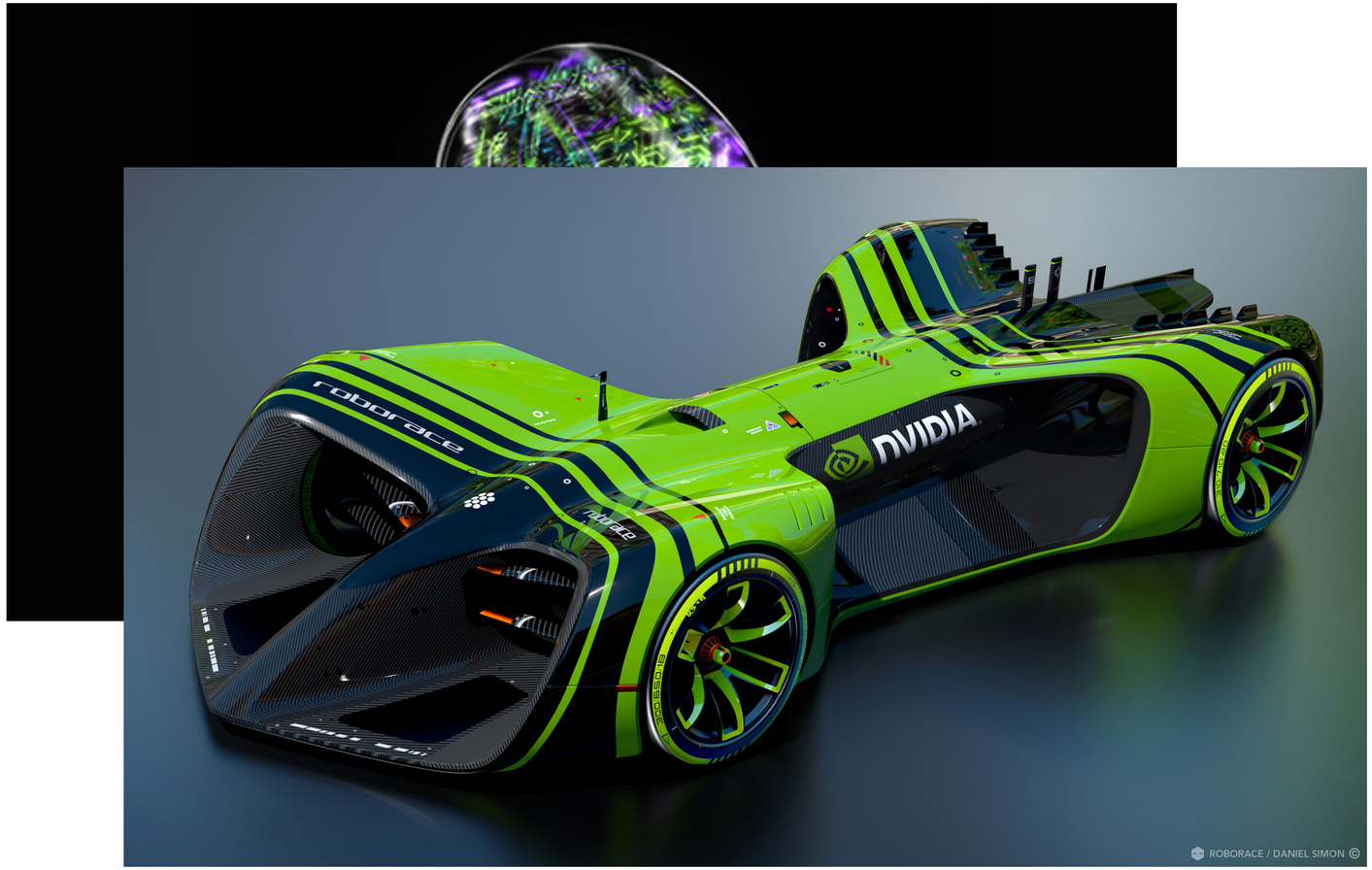
- Located at Oak Ridge National Lab
- 27,648 Nvidia Tesla V100 GPUs
- 200 petaflops
- 13 MW of power
- Currently the world's most powerful supercomputer.
  - (#2 in China and #3 at LBNL)

Applications of GPUs (non-physics)

- Artificial Intelligence (AI) / Deep learning
- Self-driving cars
- Medical Imaging
- Graphics!

Applications of GPUs (non-physics)

- Artificial Intelligence (AI) / Deep learning
- Self-driving cars
- Medical Imaging
- Graphics!

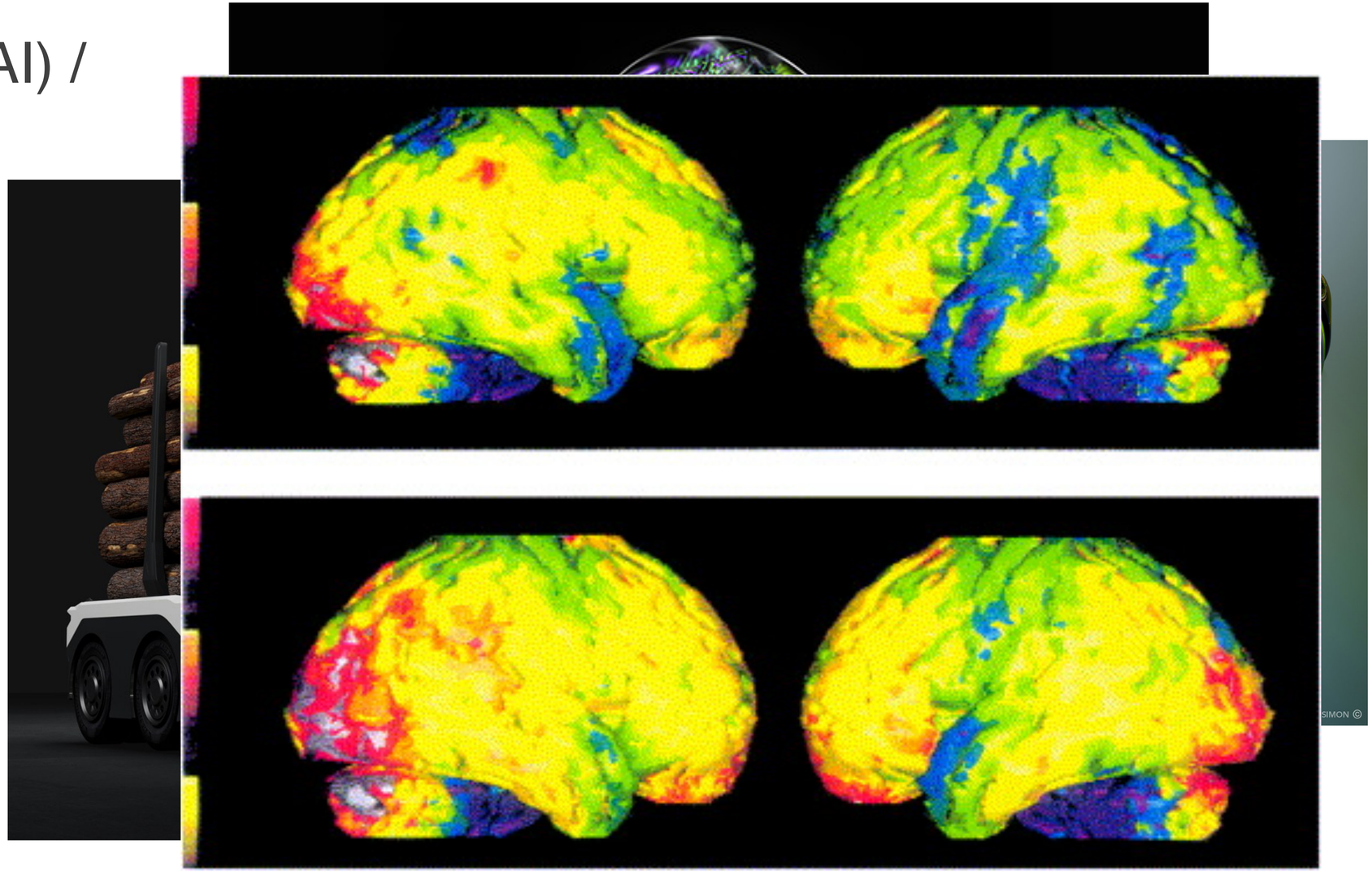


Applications of GPUs (non-physics)

- Artificial Intelligence (AI) / Deep learning
- Self-driving cars
- Medical Imaging
- Graphics!

Applications of GPUs (non-physics)

- Artificial Intelligence (AI) / Deep learning
- Self-driving cars
- Medical Imaging
- Graphics!

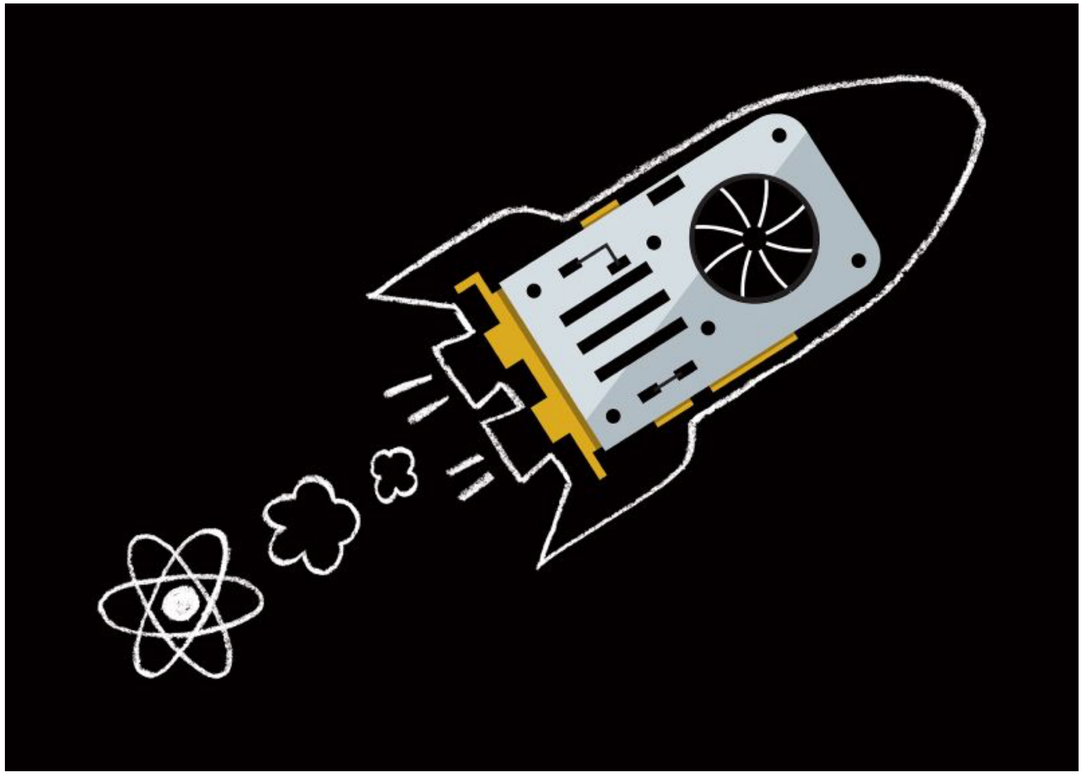


Applications of GPUs (non-physics)

- Artificial Intelligence (AI) / Deep learning
- Self-driving cars
- Medical Imaging
- Graphics!

Applications of GPUs (physics)

- Massively parallel simulations of black holes(M. Liska et al. Mon. Not. R. Astron. Soc. Lett. **474**, L81–L85; 2018)
- Multi-object Adaptive Optics for E-ELT (D. Gratadour & H. Ltaief)
- Molecular simulations to explore artificial photosynthesis with QBox
- Galaxy classification with machine learning (space.ml)
- Simulation of stellar explosions on Summit (E. Schneider, arXiv:1410.4194)



"Supercharge your data wrangling with a graphics card", Nature 562, Oct 2, 2018

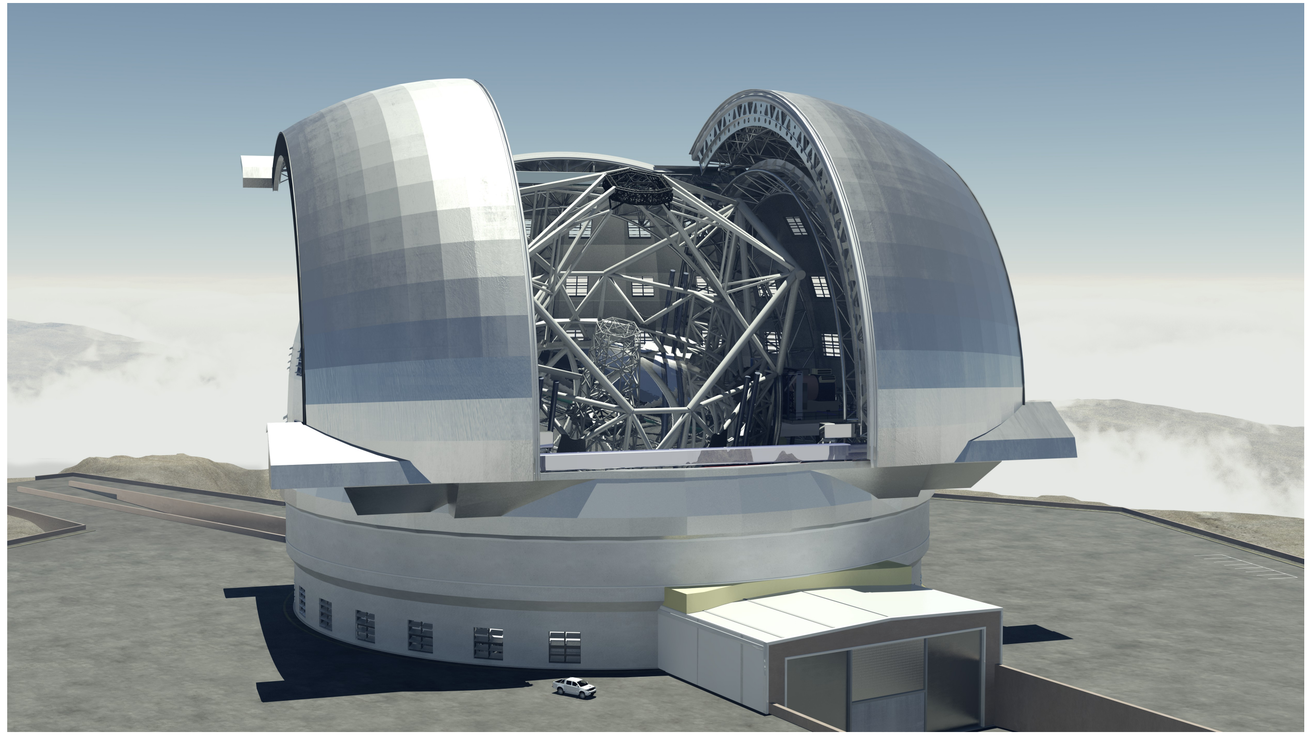
Applications of GPUs (physics)

- Massively parallel simulations of black holes(M. Liska et al. Mon. Not. R. Astron. Soc. Lett. **474**, L81–L85; 2018)
- Multi-object Adaptive Optics for E-ELT (D. Gratadour & H. Ltaief)
- Molecular simulations to explore artificial photosynthesis with QBox
- Galaxy classification with machine learning (space.ml)
- Simulation of stellar explosions on Summit (E. Schneider, arXiv:1410.4194)

"Supercharge your data wrangling with a graphics card", Nature 562, Oct 2, 2018

Applications of GPUs (physics)

- Massively parallel simulations of black holes(M. Liska et al. Mon. Not. R. Astron. Soc. Lett. **474**, L81–L85; 2018)
- Multi-object Adaptive Optics for E-ELT (D. Gratadour & H. Ltaief)
- Molecular simulations to explore artificial photosynthesis with QBox
- Galaxy classification with machine learning (space.ml)
- Simulation of stellar explosions on Summit (E. Schneider, arXiv:1410.4194)



"Supercharge your data wrangling with a graphics card", Nature 562, Oct 2, 2018

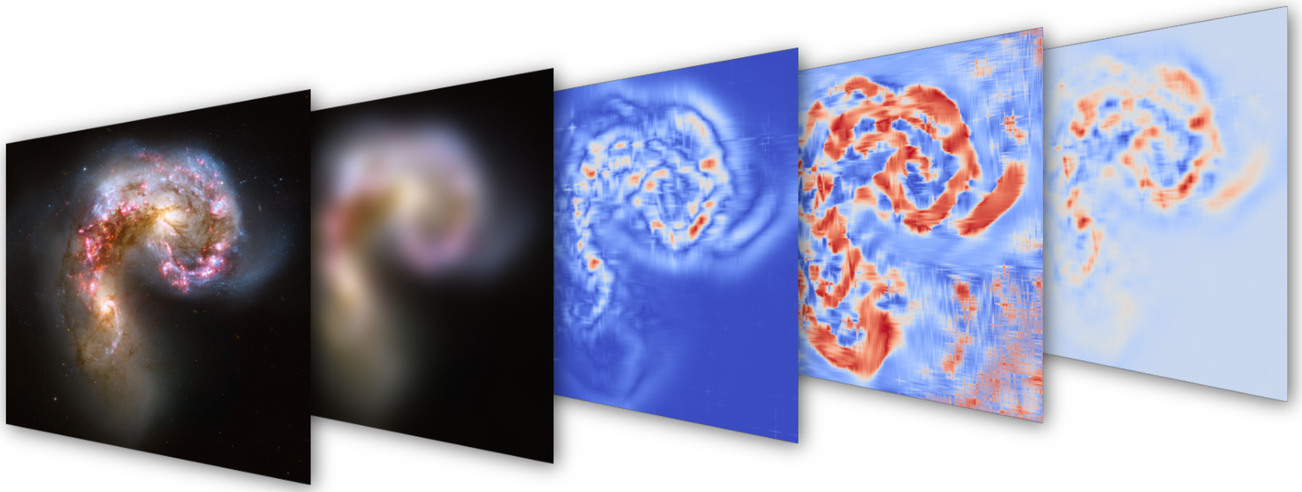
Applications of GPUs (physics)

- Massively parallel simulations of black holes(M. Liska et al. Mon. Not. R. Astron. Soc. Lett. **474**, L81–L85; 2018)
- Multi-object Adaptive Optics for E-ELT (D. Gratadour & H. Ltaief)
- Molecular simulations to explore artificial photosynthesis with QBox
- Galaxy classification with machine learning (space.ml)
- Simulation of stellar explosions on Summit (E. Schneider, arXiv:1410.4194)

a graphics card", Nature 562, Oct 2, 2018

Applications of GPUs (physics)

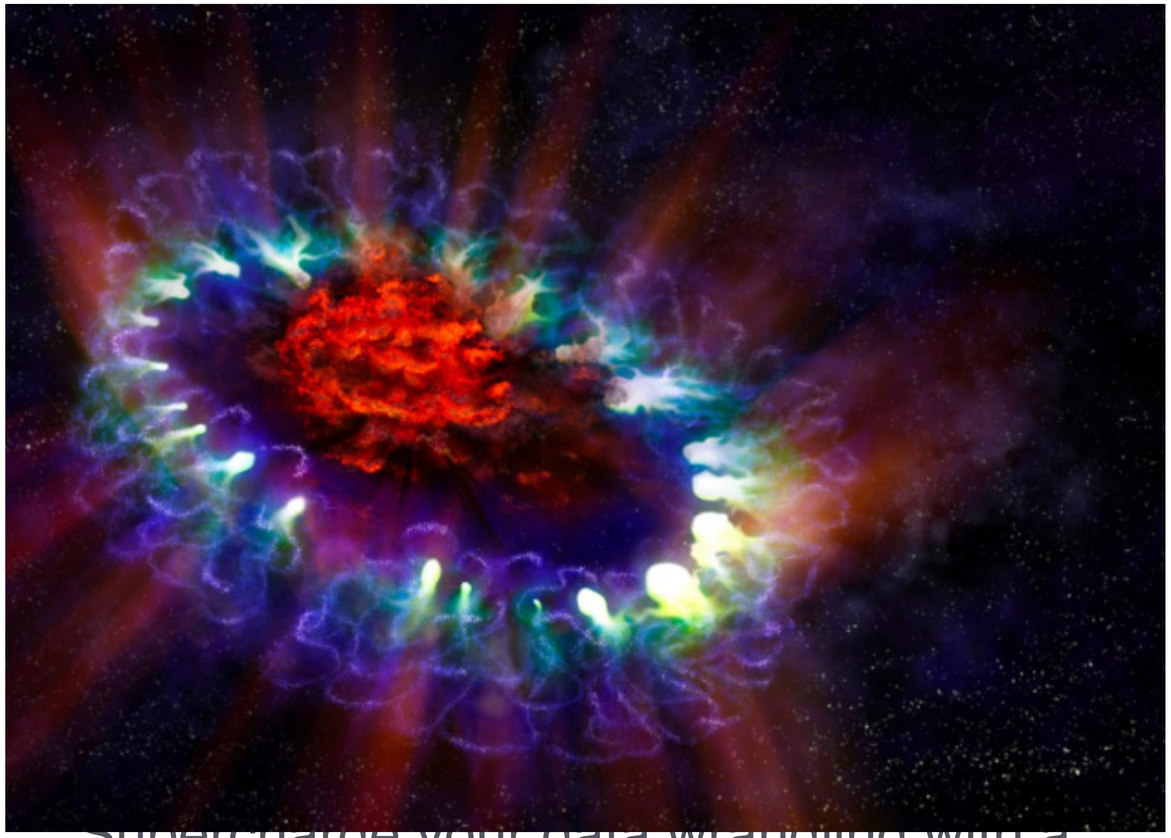
- Massively parallel simulations of black holes(M. Liska et al. Mon. Not. R. Astron. Soc. Lett. **474**, L81–L85; 2018)
- Multi-object Adaptive Optics for E-ELT (D. Gratadour & H. Ltaief)
- Molecular simulations to explore artificial photosynthesis with QBox
- Galaxy classification with machine learning (space.ml)
- Simulation of stellar explosions on Summit (E. Schneider, arXiv:1410.4194)



"Supercharge your data wrangling with a graphics card", Nature 562, Oct 2, 2018

Applications of GPUs (physics)

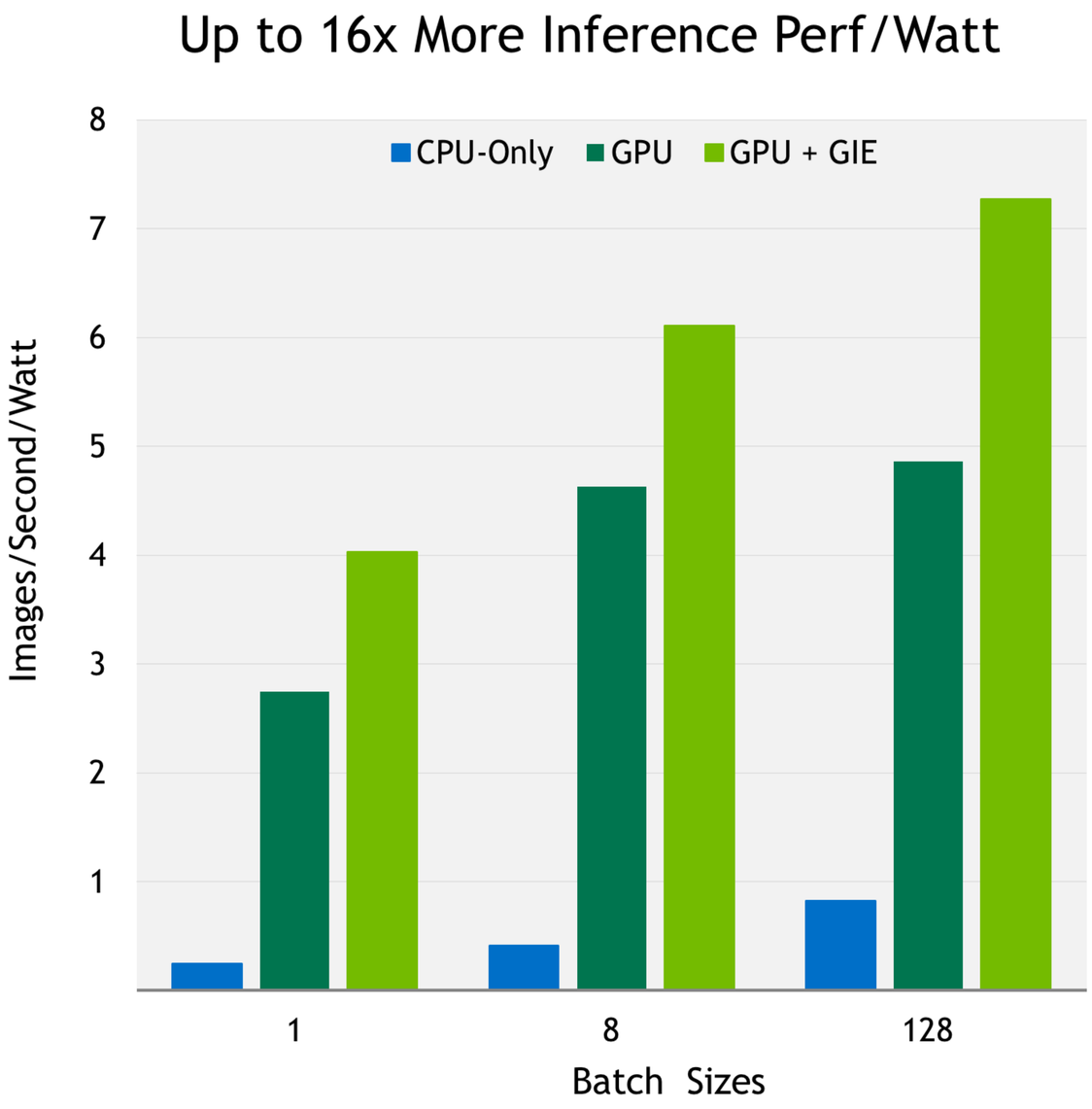
- Massively parallel simulations of black holes(M. Liska et al. Mon. Not. R. Astron. Soc. Lett. **474**, L81–L85; 2018)
- Multi-object Adaptive Optics for E-ELT (D. Gratadour & H. Ltaief)
- Molecular simulations to explore artificial photosynthesis with QBox
- Galaxy classification with machine learning (space.ml)
- Simulation of stellar explosions on Summit (E. Schneider, arXiv:1410.4194)



Supercharge your data wrangling with a graphics card", Nature 562, Oct 2, 2018

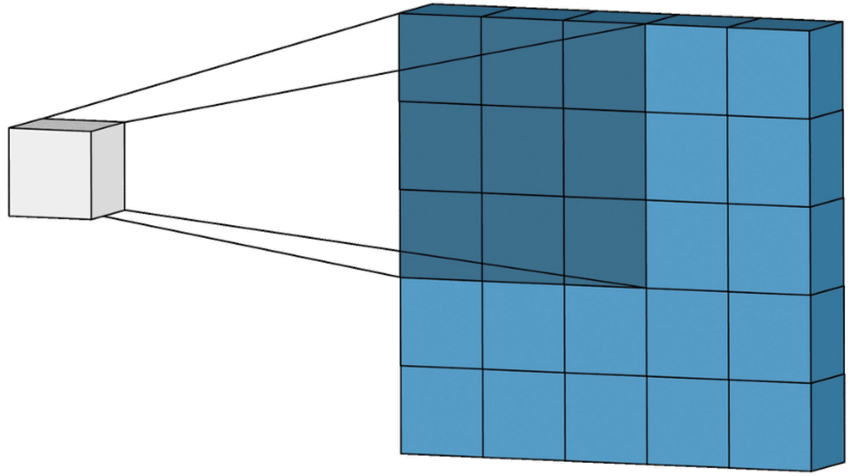
GPUs for machine learning

- Deep learning applications can be dramatically accelerated when run on a GPU.



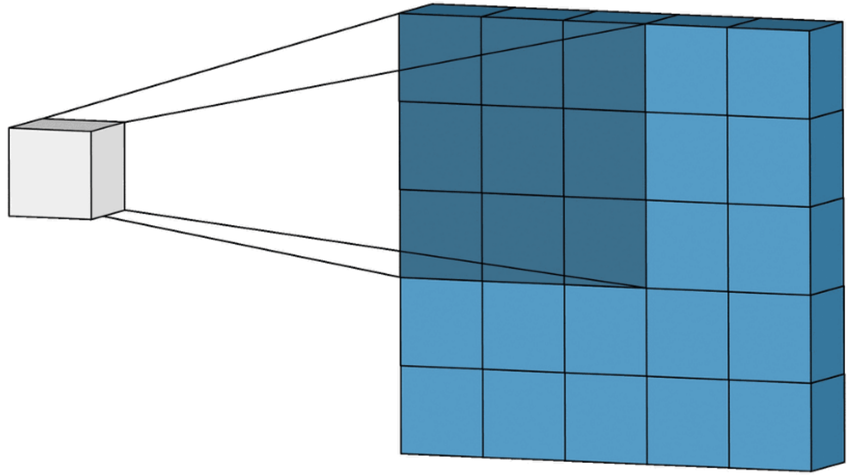
What is machine learning?

- A mechanism of using artificial intelligence to “teach” a computer to identify patterns using neural networks.
- Can be significantly accelerated by running on GPUs or other highly parallel architectures.
- In use by many other particle physics experiments.



What is machine learning?

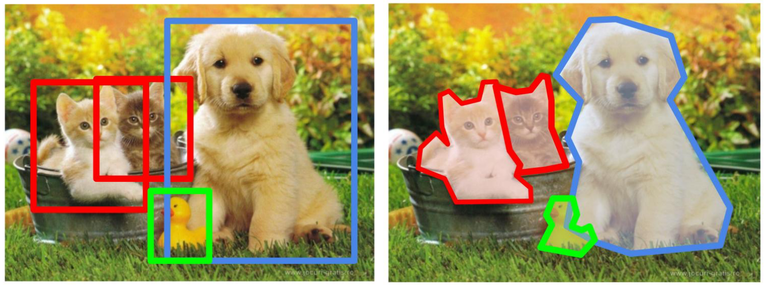
- A mechanism of using artificial intelligence to “teach” a computer to identify patterns using neural networks.
- Can be significantly accelerated by running on GPUs or other highly parallel architectures.
- In use by many other particle physics experiments.



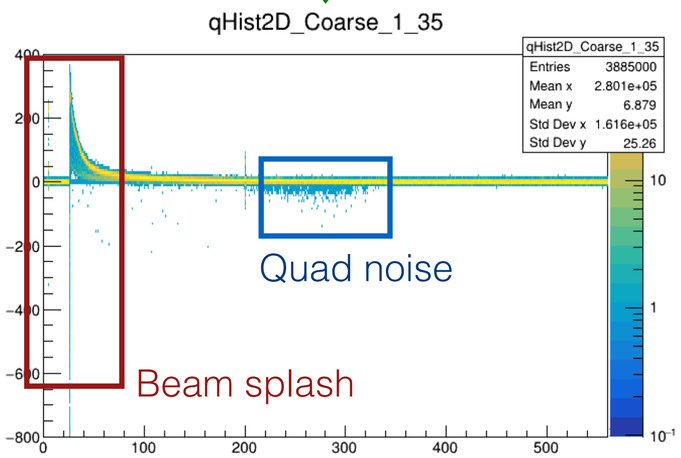
Deep Learning in Particle Physics

- Neural Networks invented to do image recognition (Facebook, google, etc)
- Convolutional neural networks have been successfully applied to improve data processing for many Fermilab neutrino experiments.
- Active machine learning community at Fermilab.
- Many opportunities to apply deep learning.

Object Detection | **Instance Segmentation**

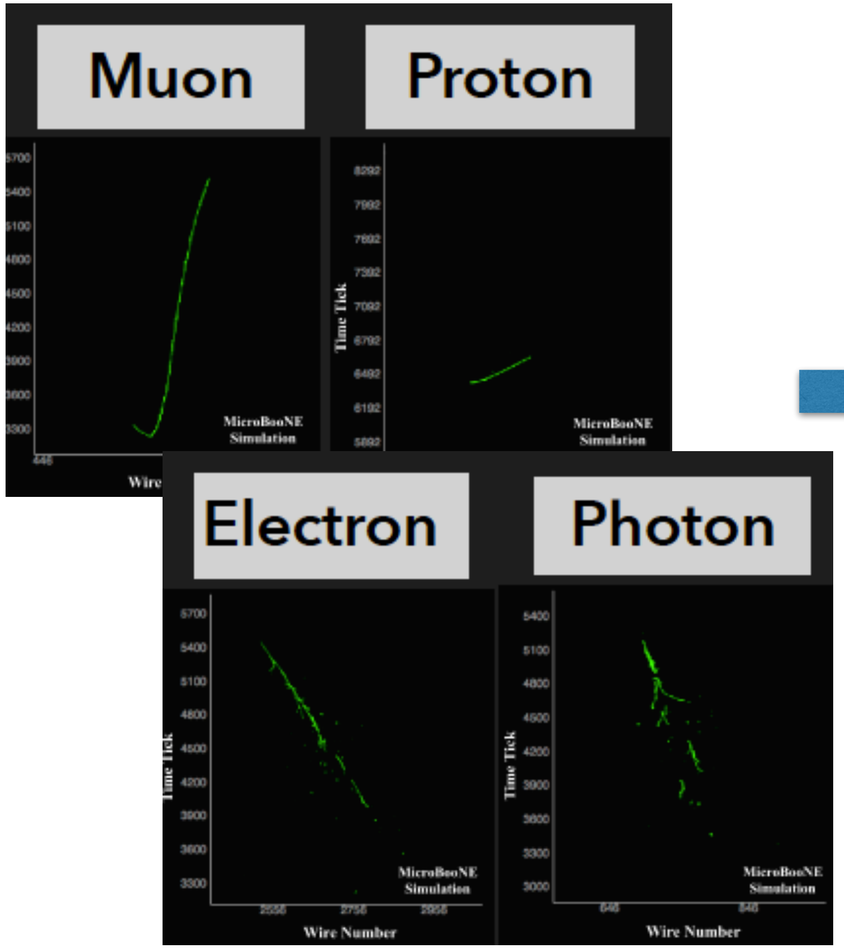


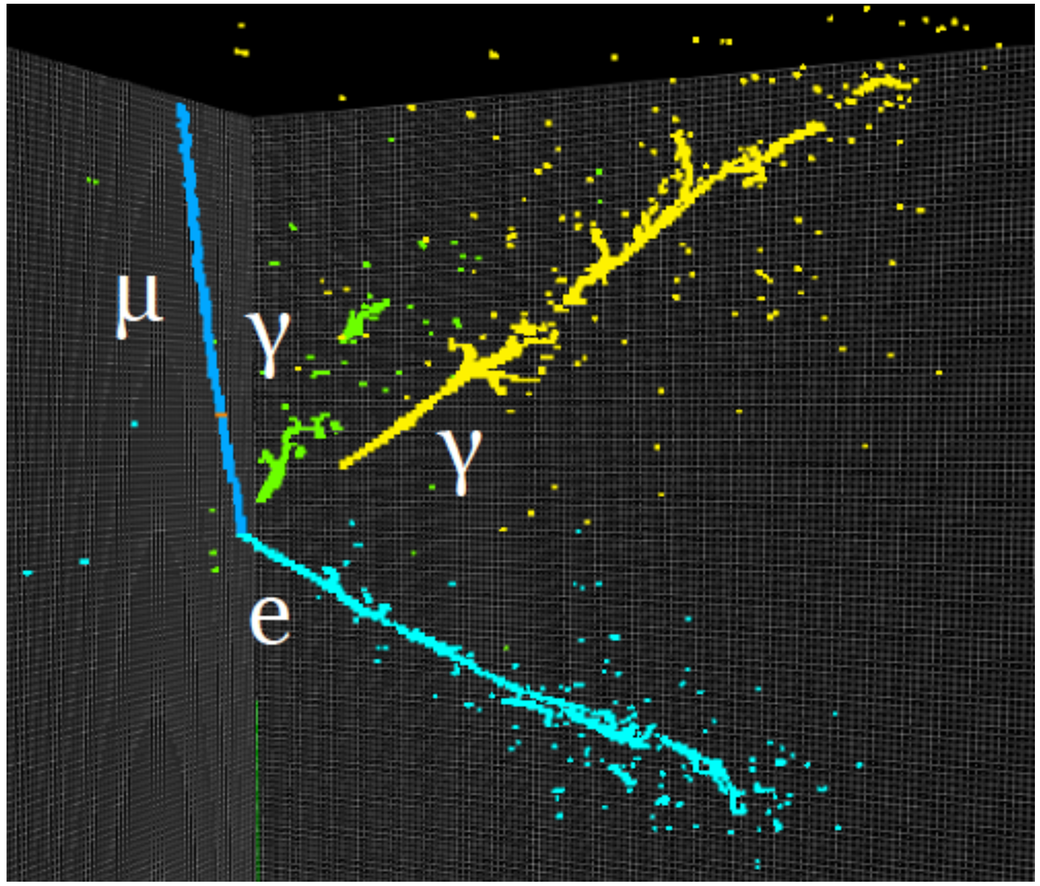
CAT, DOG, DUCK | CAT, DOG, DUCK



Using Deep Learning to look for Neutrinos

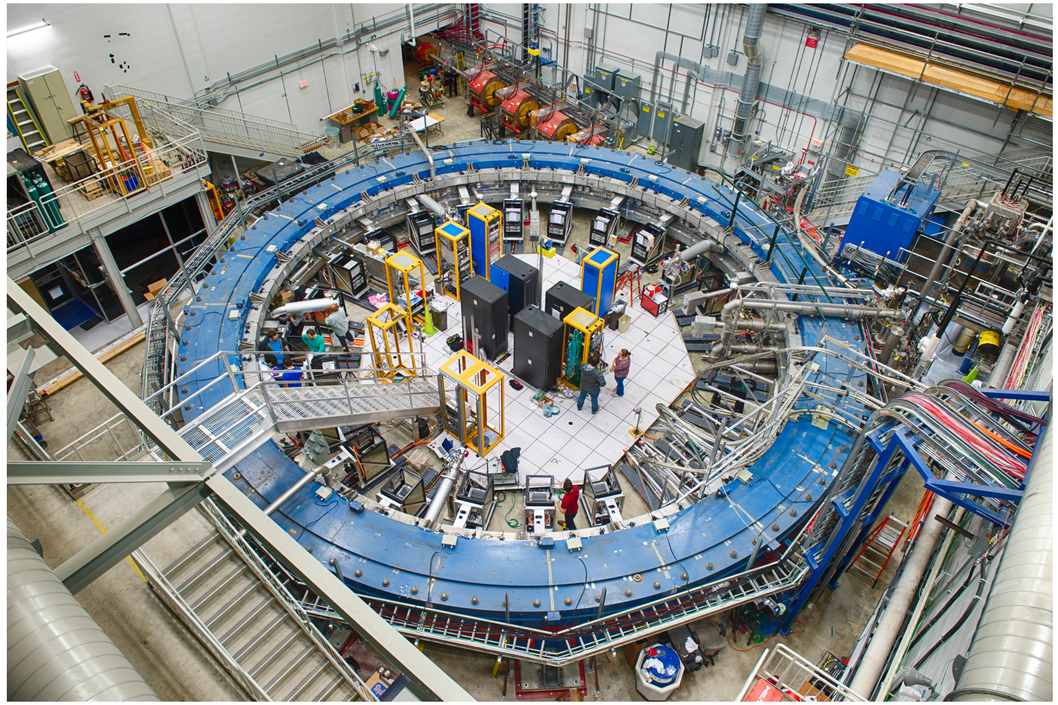
- CNNs identify particle tracks in the MicroBoone detector





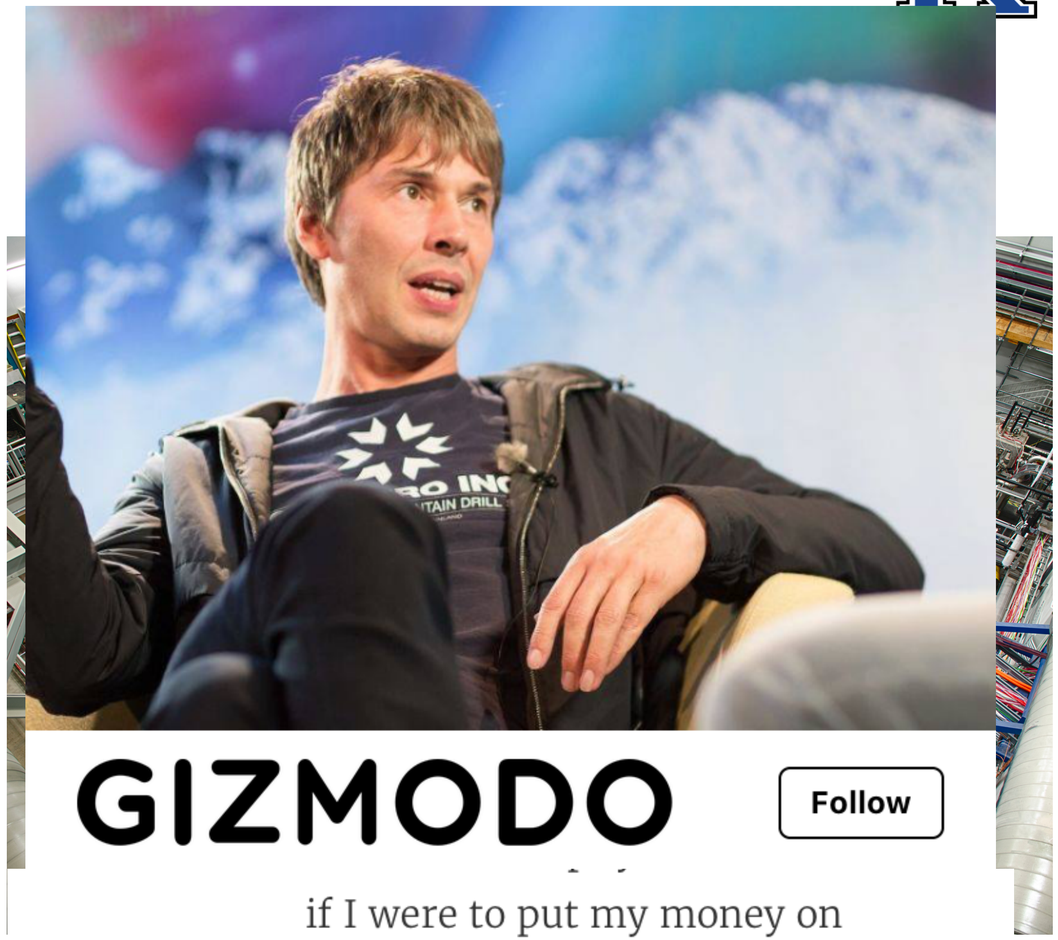
What is Muon g-2?

- Muon g-2 is an experiment to measure the magnetic moment of the muon.
- Muons (like big electrons) are injected into a 50’ diameter magnetic ring.
- As they travel around the ring, they precess like tiny tops.
- We measure this precession frequency to infer how muons interact with the quantum background and the rest of the universe.



What is Muon g-2?

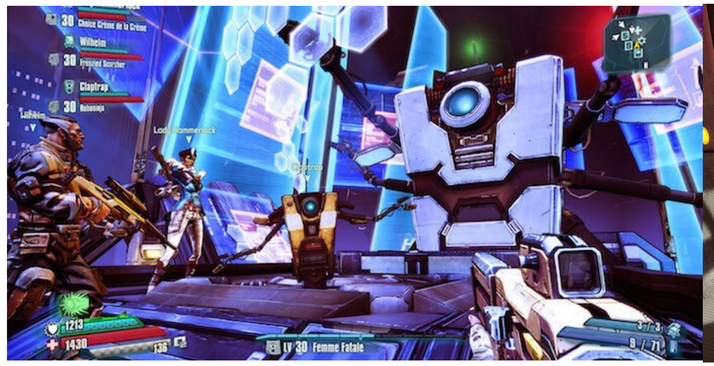
- Muon g-2 is an experiment to measure the magnetic moment of the muon.
- Muons (like big electrons) are injected into a 50’ diameter magnetic ring.
- As they travel around the ring, they precess like tiny tops.
- We measure this precession frequency to infer how muons interact with the quantum background and the rest of the universe.



if I were to put my money on something that would signal new physics, it’s the g-2 experiment at Fermilab. I think it’s really fascinating.

Why GPUs?

- The GPUs dramatically improve performance by parallelizing processing.
- Technology was developed for commercial applications, so it is well supported.
- Easier to code in C++ than to learn specialized language.
- Without GPUs, we could not keep up with our data rates.

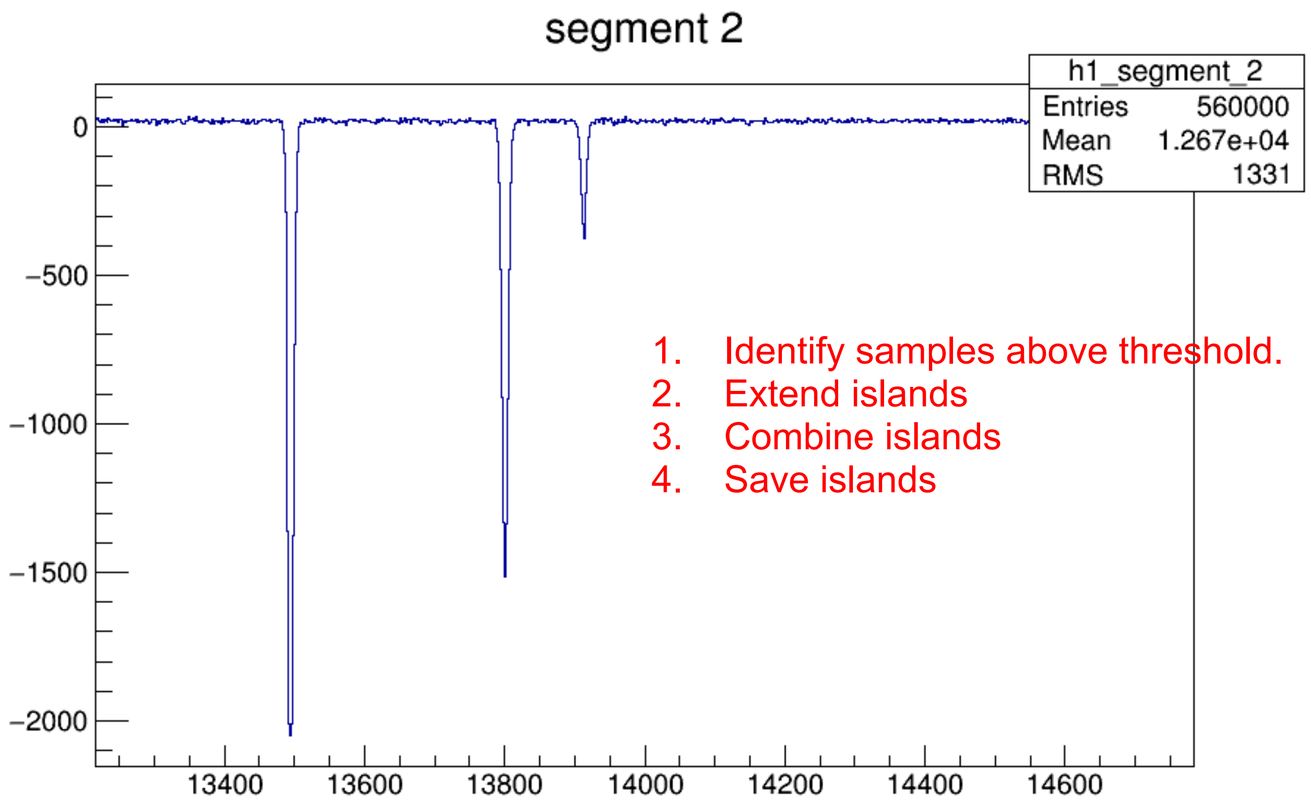


Muon g-2 Data

- The ring is filled with muons 12 times per second.
- Each time a fill enters the ring, we record our detectors for almost 1 millisecond to record all of the muon decays.
- Each decay puts a tiny spike in the data, which we need to keep. 99% of the data does not have this tiny spike, and can be thrown away.
- The total data rate is 20 GB/s. Imagine downloading 10 iTunes movies to your computer every second for the next two years — that's a lot of data!

How data is processed online in GPUs

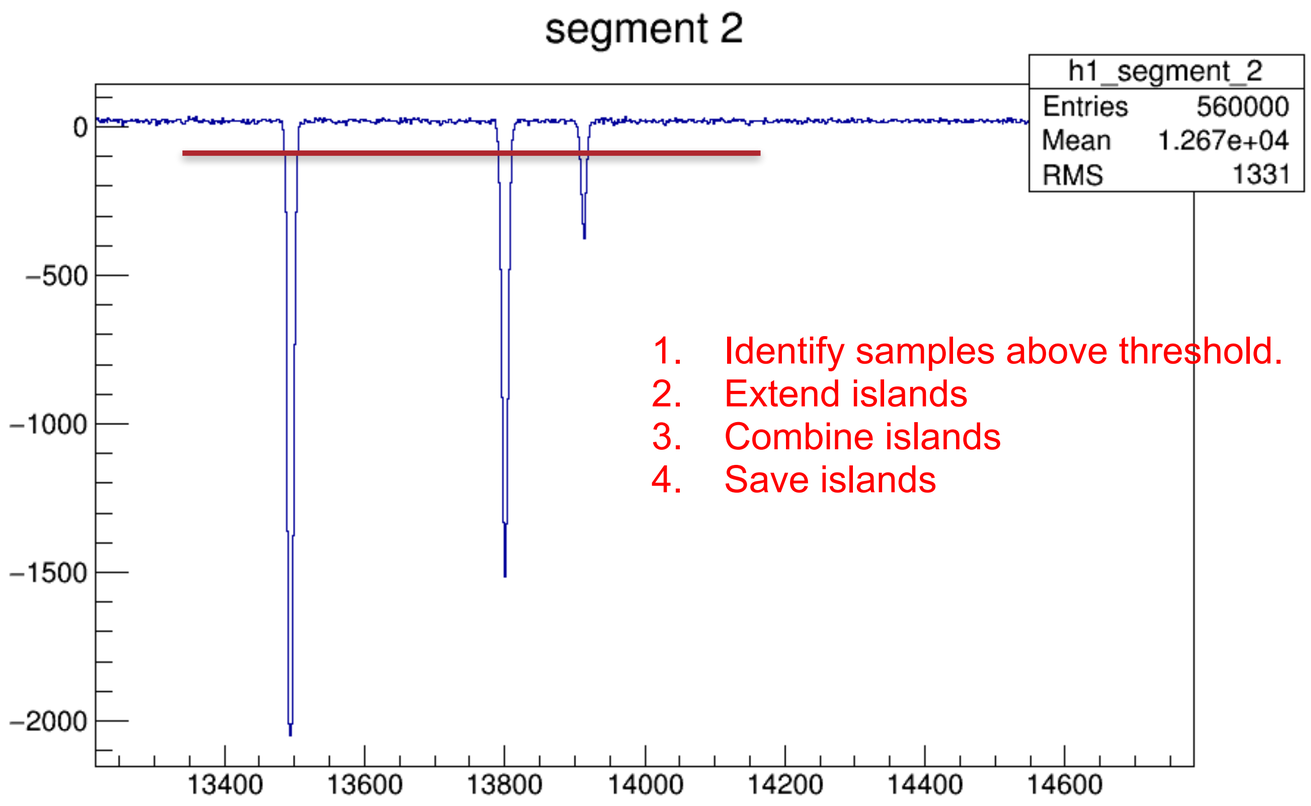
- Identify and save regions of the waveform containing positron hits.
- A typical waveform will have ~180 islands.



1. Identify samples above threshold.
2. Extend islands
3. Combine islands
4. Save islands

How data is processed online in GPUs

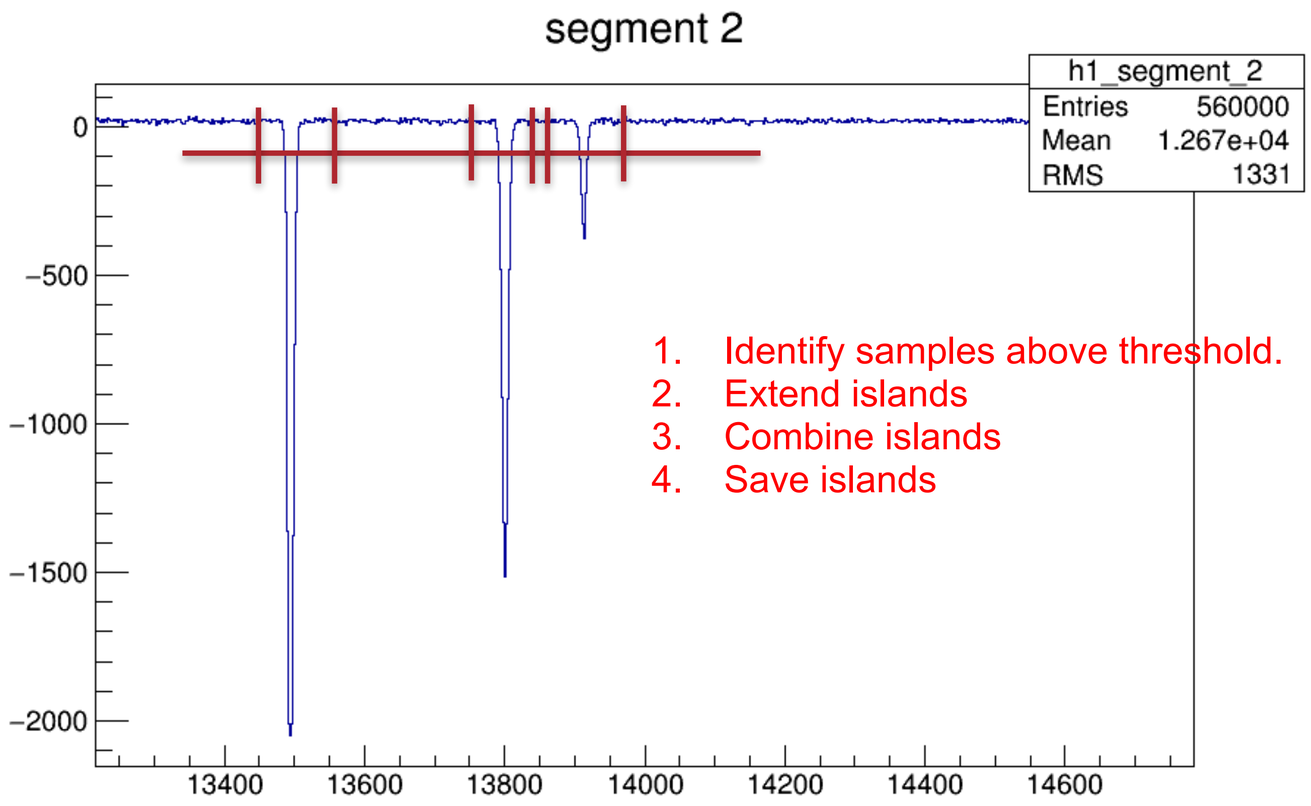
- Identify and save regions of the waveform containing positron hits.
- A typical waveform will have ~180 islands.



1. Identify samples above threshold.
2. Extend islands
3. Combine islands
4. Save islands

How data is processed online in GPUs

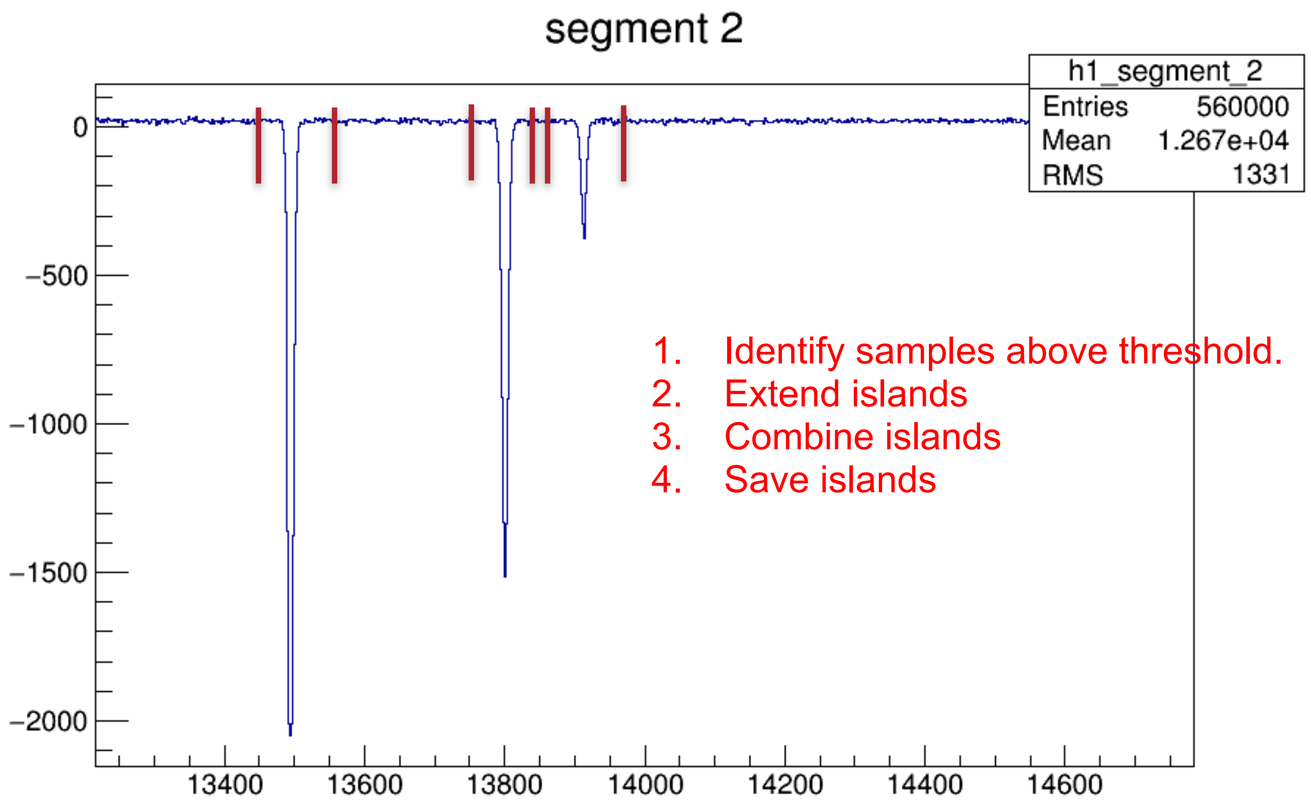
- Identify and save regions of the waveform containing positron hits.
- A typical waveform will have ~180 islands.



1. Identify samples above threshold.
2. Extend islands
3. Combine islands
4. Save islands

How data is processed online in GPUs

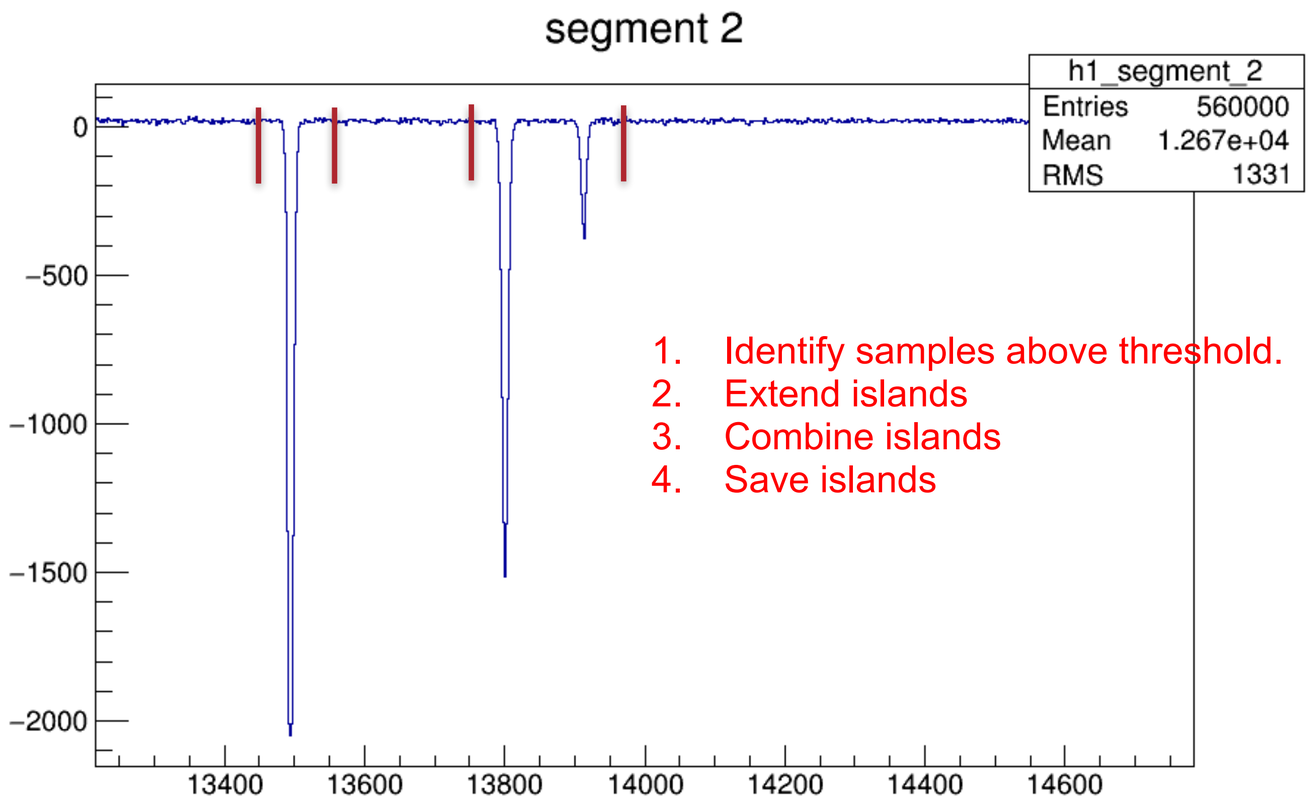
- Identify and save regions of the waveform containing positron hits.
- A typical waveform will have ~180 islands.



1. Identify samples above threshold.
2. Extend islands
3. Combine islands
4. Save islands

How data is processed online in GPUs

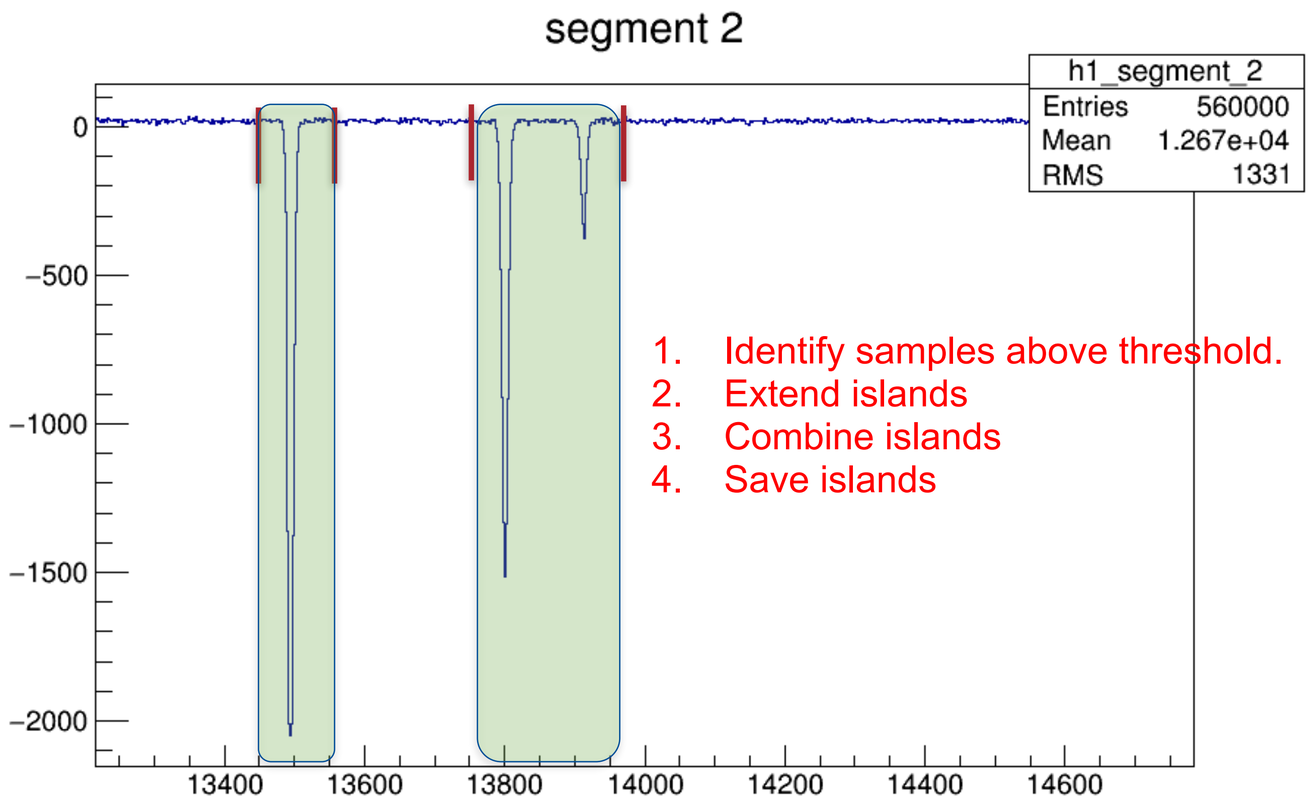
- Identify and save regions of the waveform containing positron hits.
- A typical waveform will have ~180 islands.



1. Identify samples above threshold.
2. Extend islands
3. Combine islands
4. Save islands

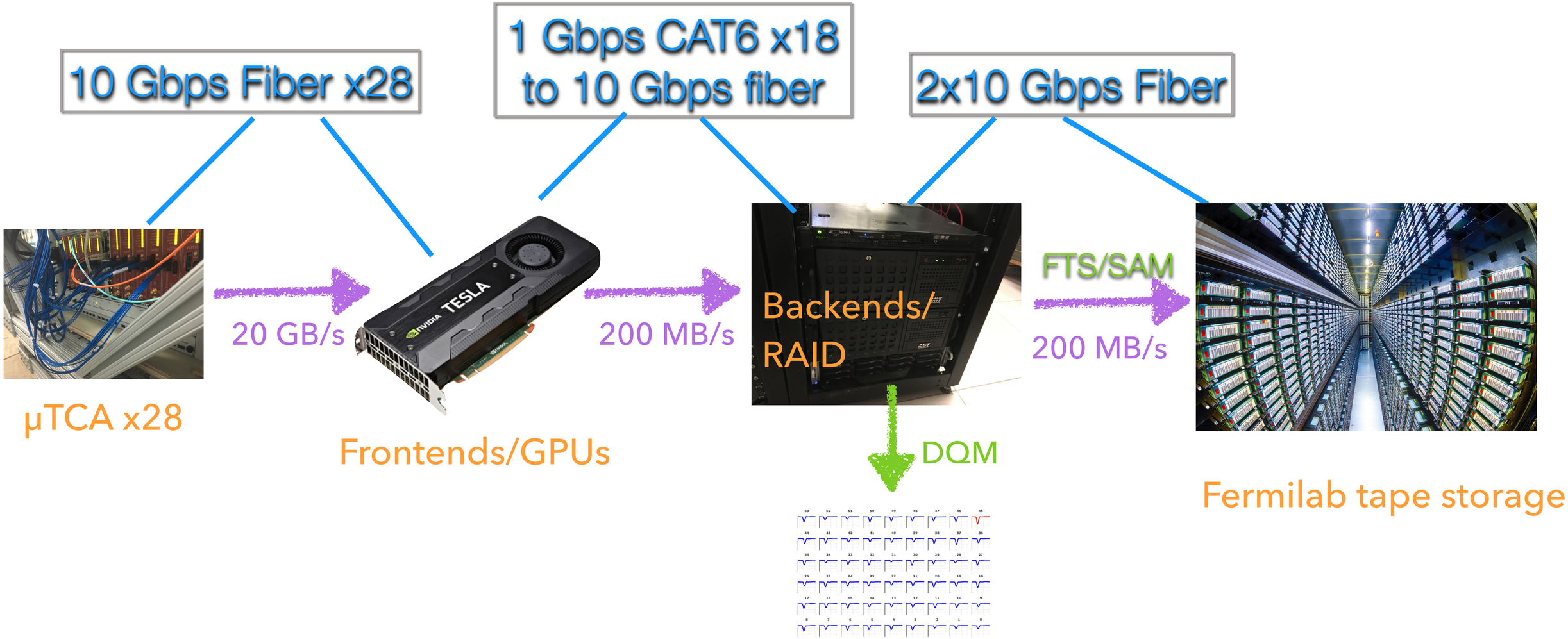
How data is processed online in GPUs

- Identify and save regions of the waveform containing positron hits.
- A typical waveform will have ~180 islands.



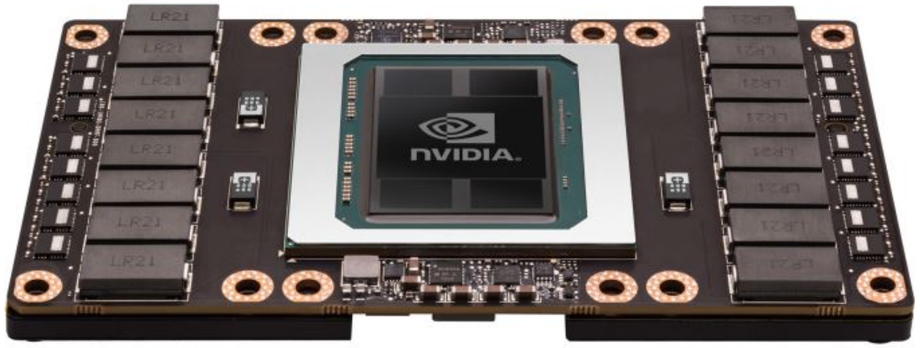
1. Identify samples above threshold.
2. Extend islands
3. Combine islands
4. Save islands

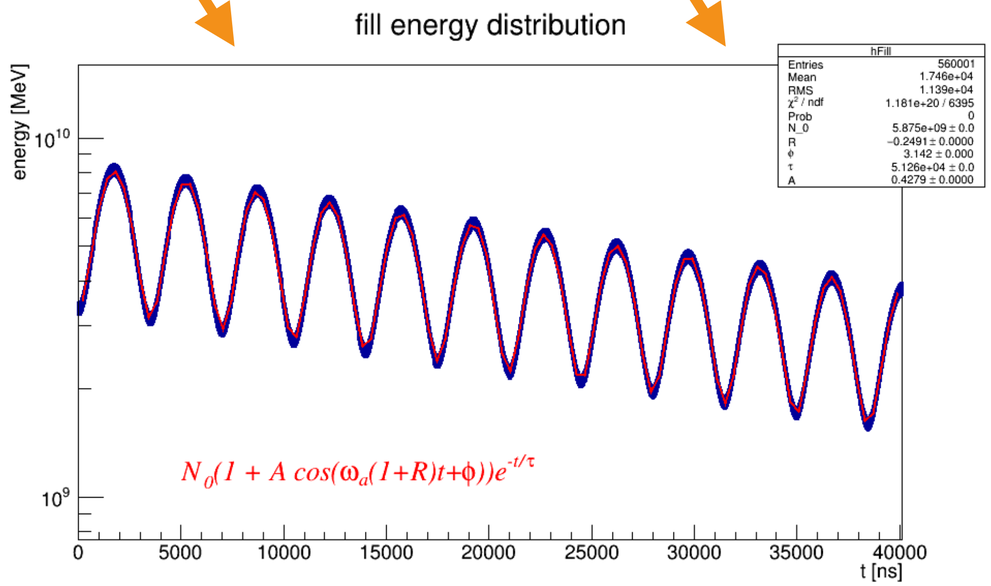
Data Flow


10 Gbps Fiber x28
1 Gbps CAT6 x18
to 10 Gbps fiber
2x10 Gbps Fiber
µTCA x28
20 GB/s
Frontends/GPUs
200 MB/s
Backends/
RAID
FTS/SAM
200 MB/s
Fermilab tape storage
DQM

Using GPUs for data simulation

- We also use GPUs to create simulations of our data.
- We use CUDA to multithread the simulation on Nvidia GPUs, so we can simulate thousands of particles at the same time.
- The simulation runs on the Open Science Grid, and can simulate 3 years of data in about 2.5 hours.

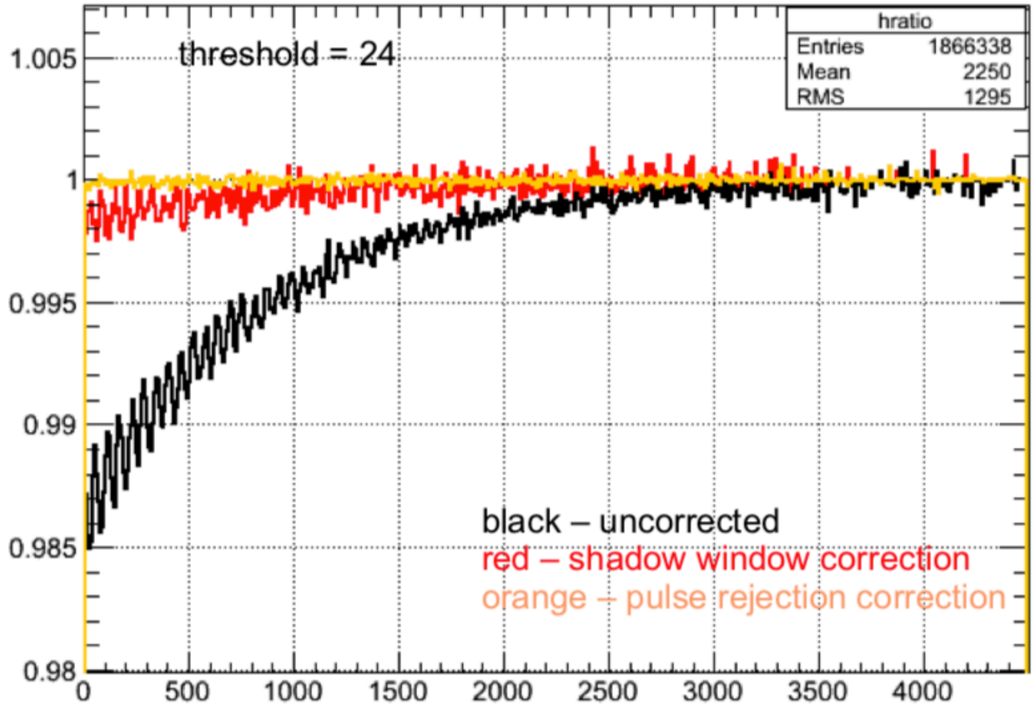




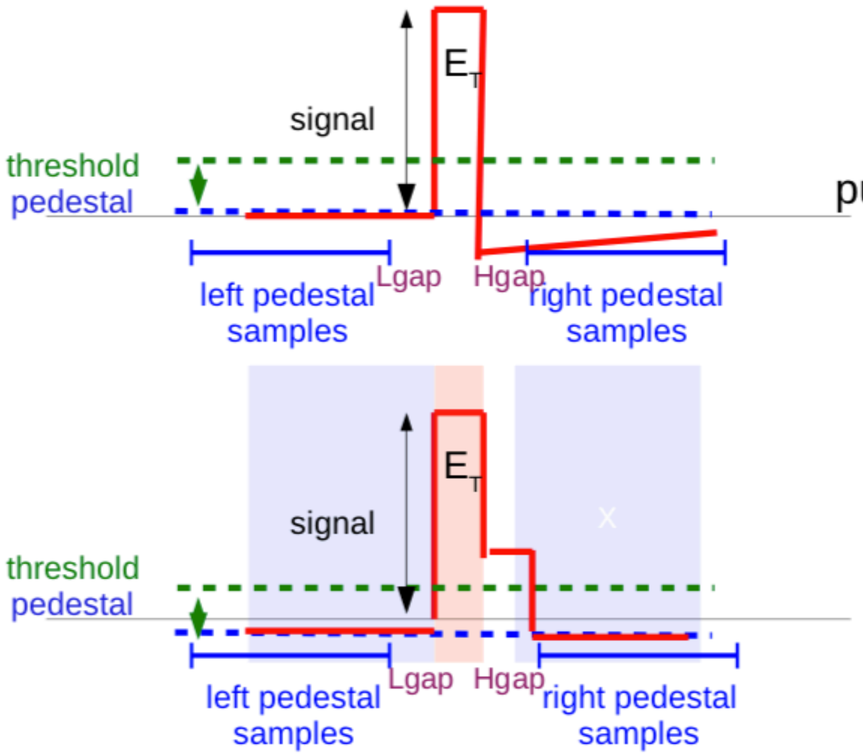
What we learn from the simulation

- Can simulate effects from pileup, pedestal variations, beam effects.

Ratio to truth



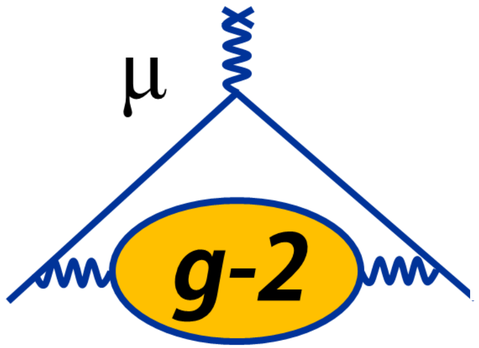
Comparison of different pileup correction methods.

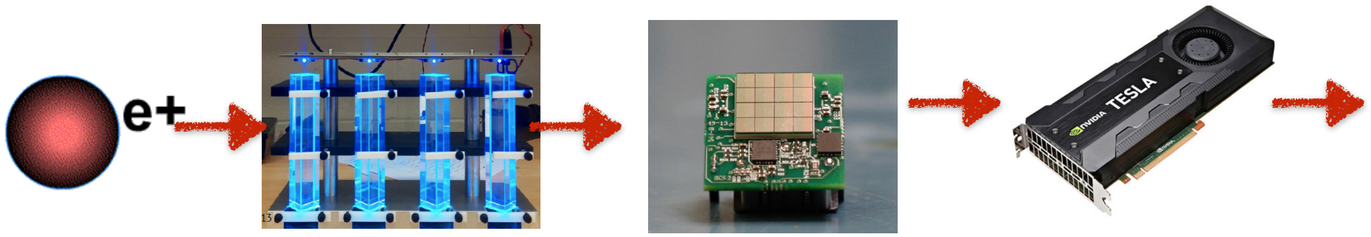


pulse undershoot affects w2,4,8 calc differently

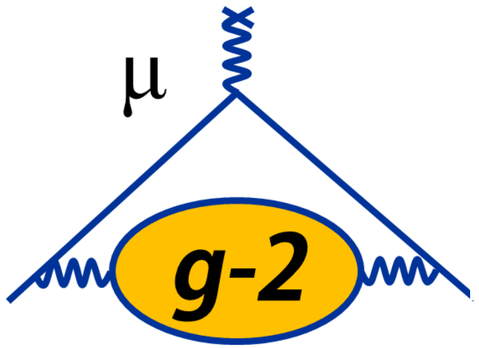
split pulse affects w2,4,8 calc differently

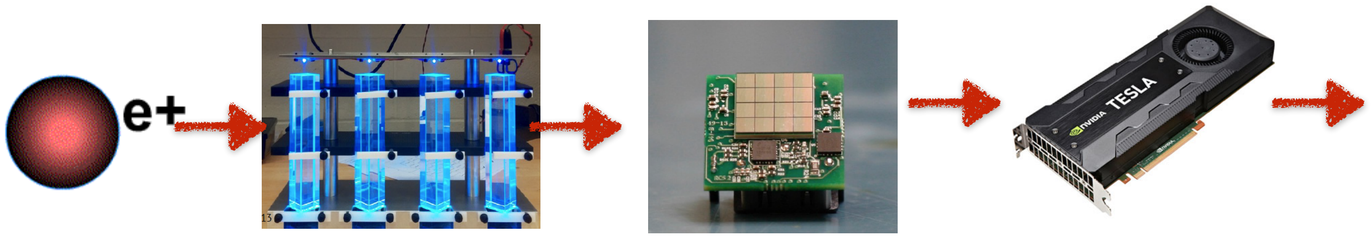
Aside: Medical Imaging

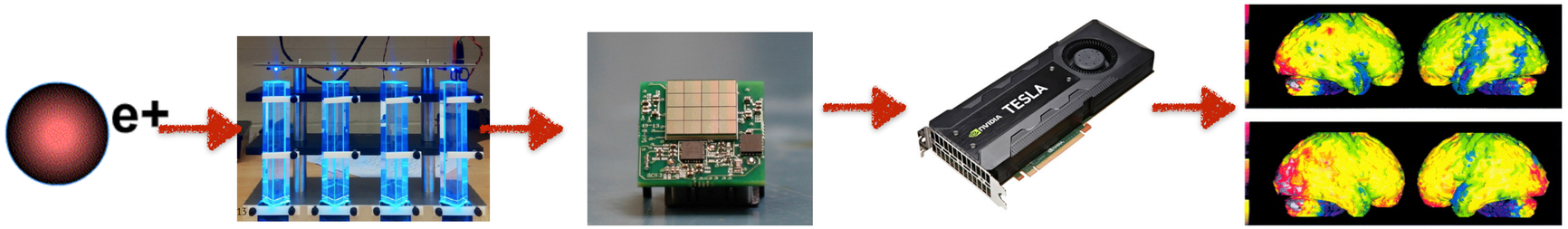




Aside: Medical Imaging





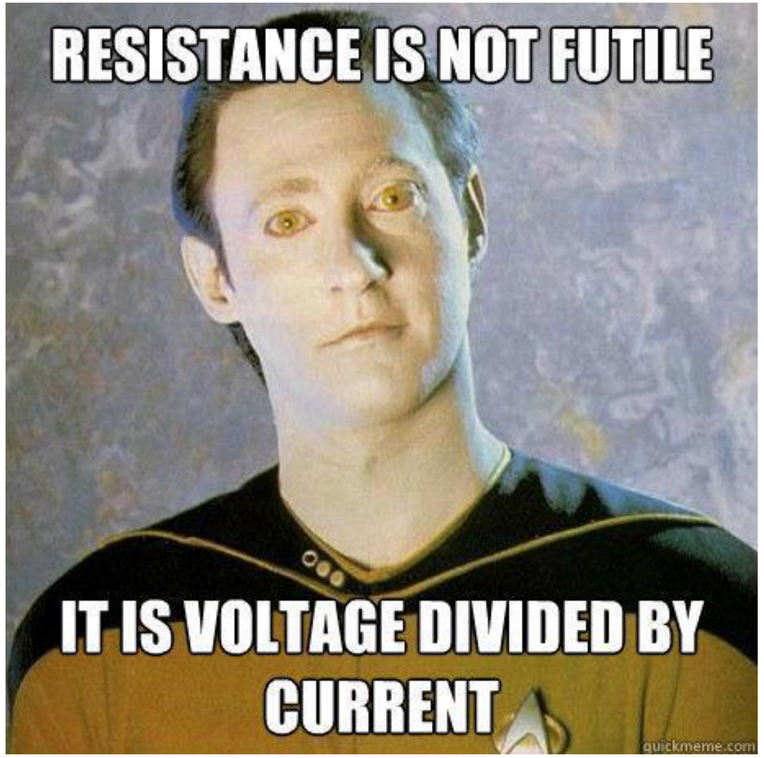


Conclusion

- GPUs are revolutionizing the world of High Performance Computing, and creating new possibilities for computing in physics.

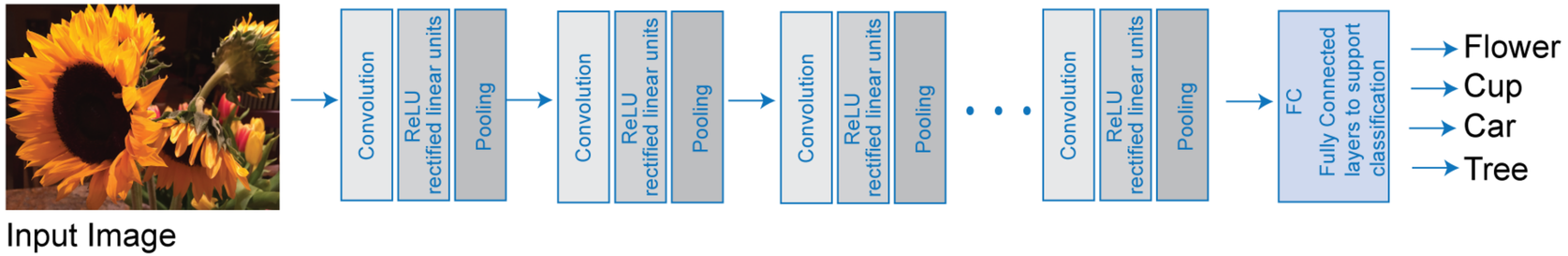
- The increased parallelization available in GPUs is creating a boom in Artificial Intelligence, which we are just learning how to apply both in industry and in physics.

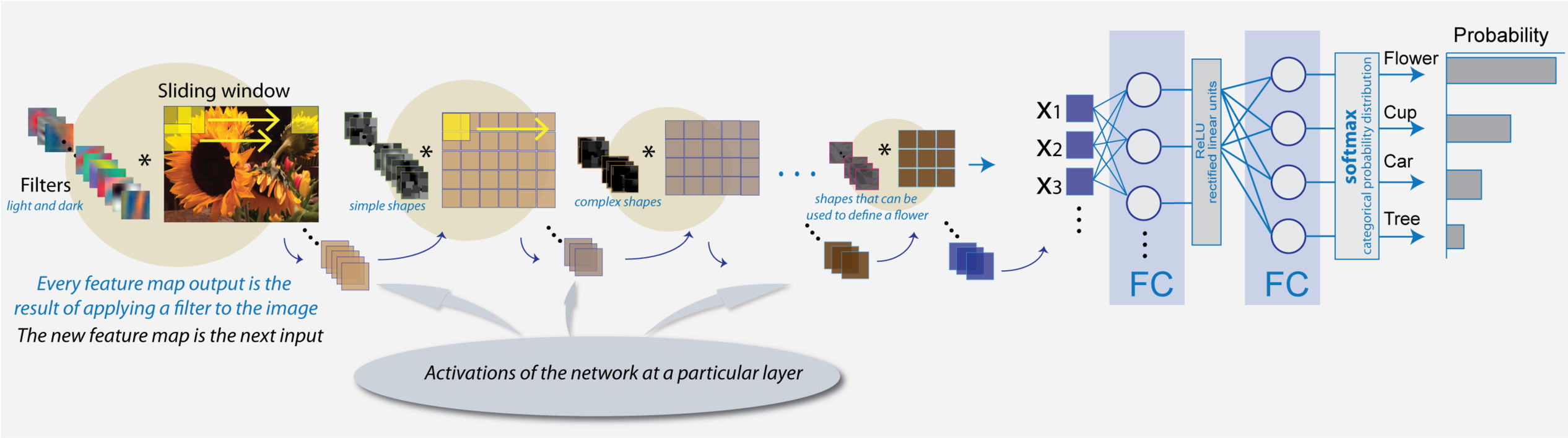
- Many particle physics experiments are using this technology to accelerate data processing and simulation.



backup

Convolutional Neural Networks

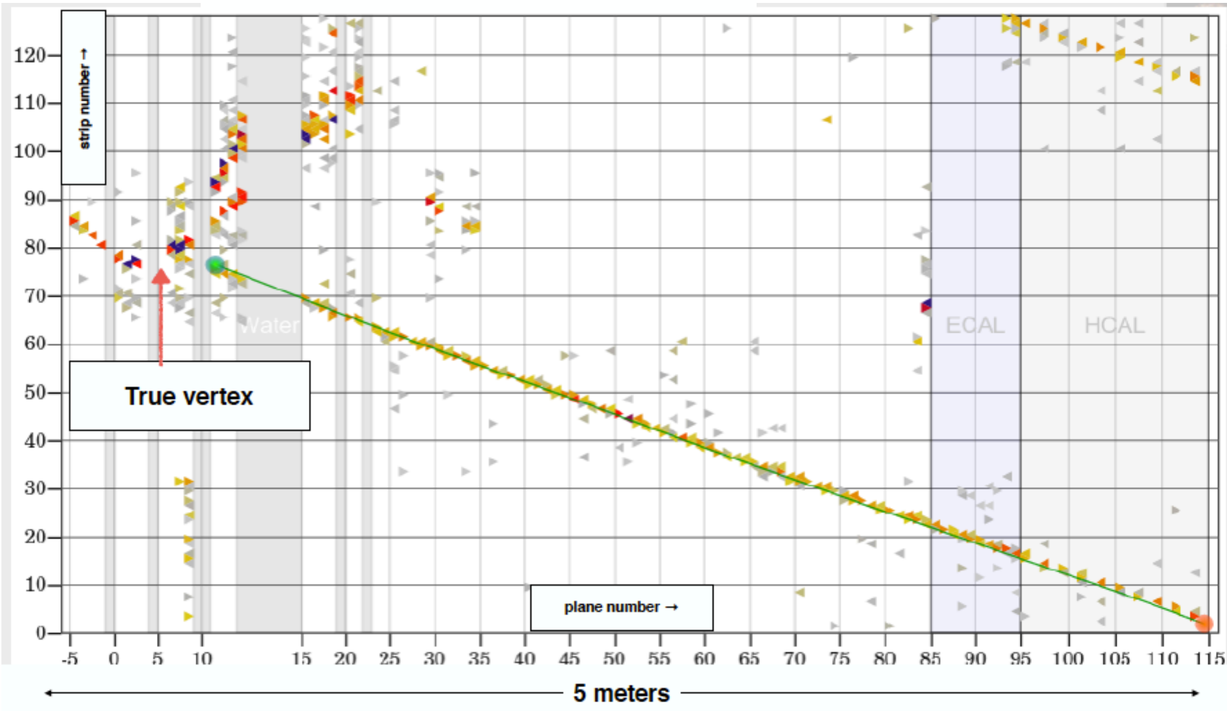




Using Machine Learning to look for Neutrinos

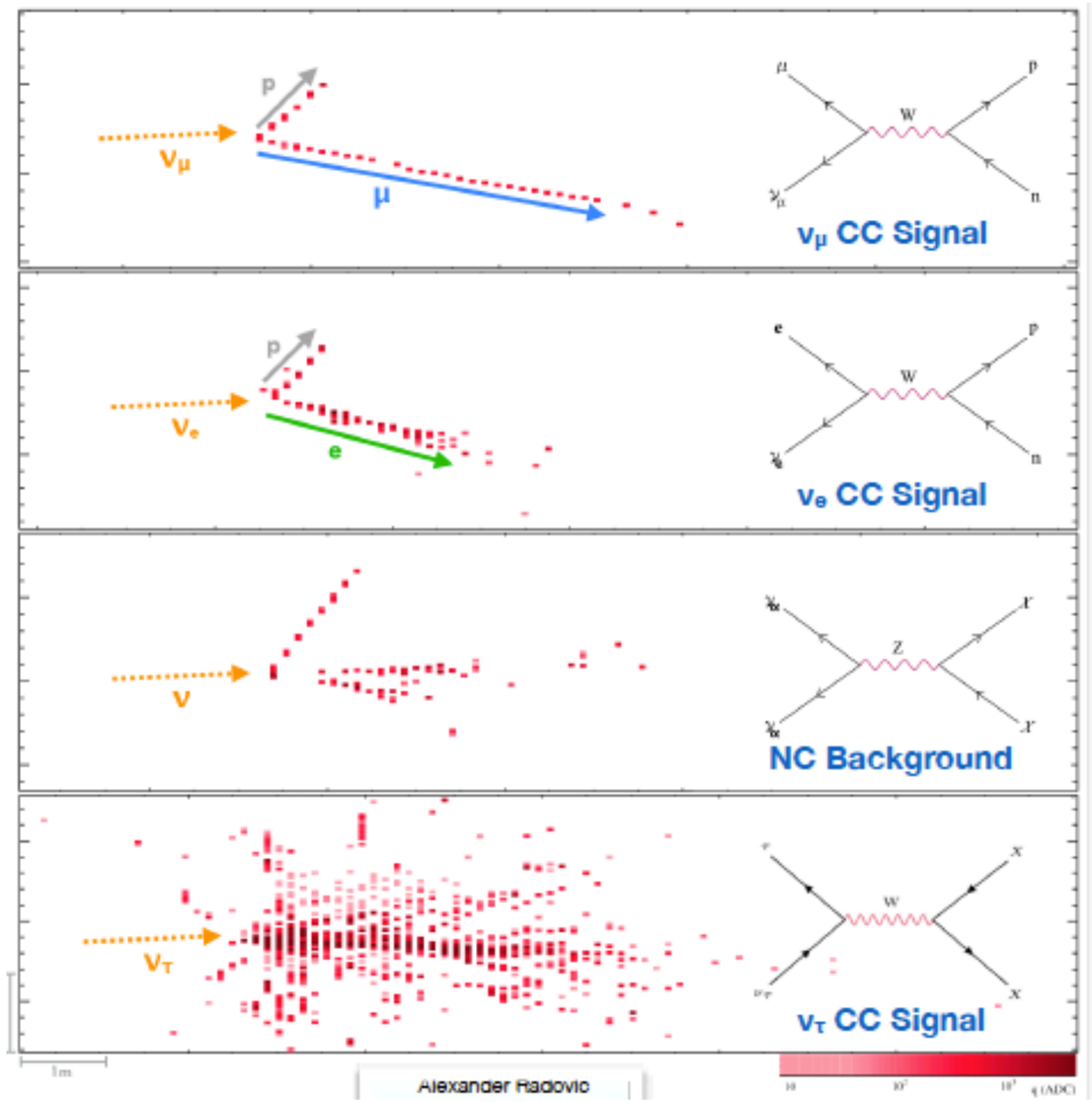
- Running large-scale machine learning applications on Summit (and previously TITAN)
- Network is trained using simulation data to identify vertex of particle decay from 2D vector input (E, position)
- Significantly outperforms track-based reconstruction methods.
- Utilizes “Domain Adversarial Neural Networks”, which minimize bias coming from the MC (arXiv:1505.07818)

$\nu_\mu + N \rightarrow \mu^- + X$



from G. Perdue

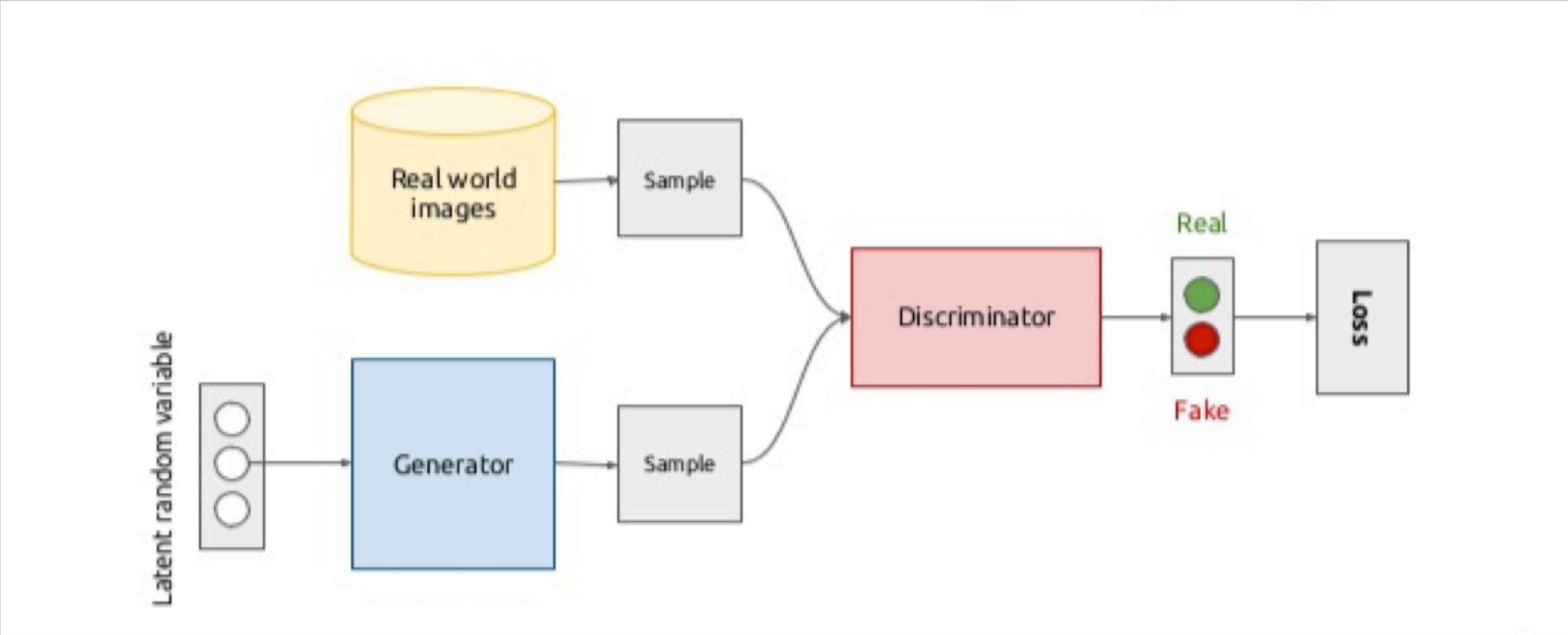
Using Machine Learning to look for Neutrinos



- To search for neutrino oscillations, must effectively identify electron, muon, and tau type neutrinos.
- Using CVN (Convolutional Visual Network) for event classification.
- Network is based on the first Googlenet, which is a 22 layer deep network that won ImageNet2014.

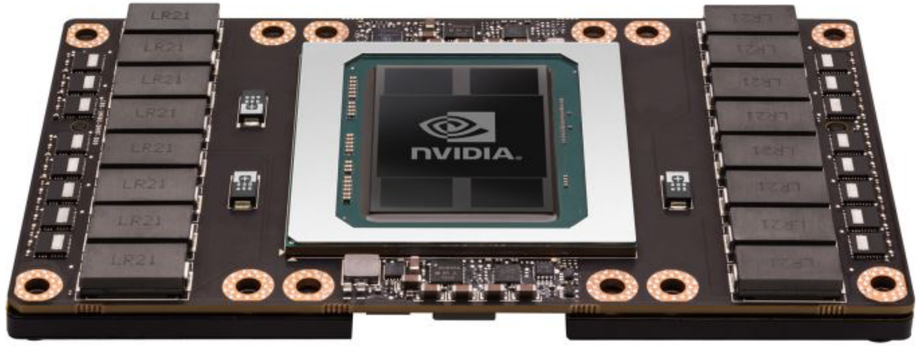
Generative Adversarial Networks (GANs)

- A GAN can be used to perform an incredibly fast particle physics simulation.
- One network is trained to learn what particle physics events look like using real or simulated data.
- A second network generates fake events, which are sorted by the first network to determine whether or not they look realistic.
- Currently being used in LHC experiments.



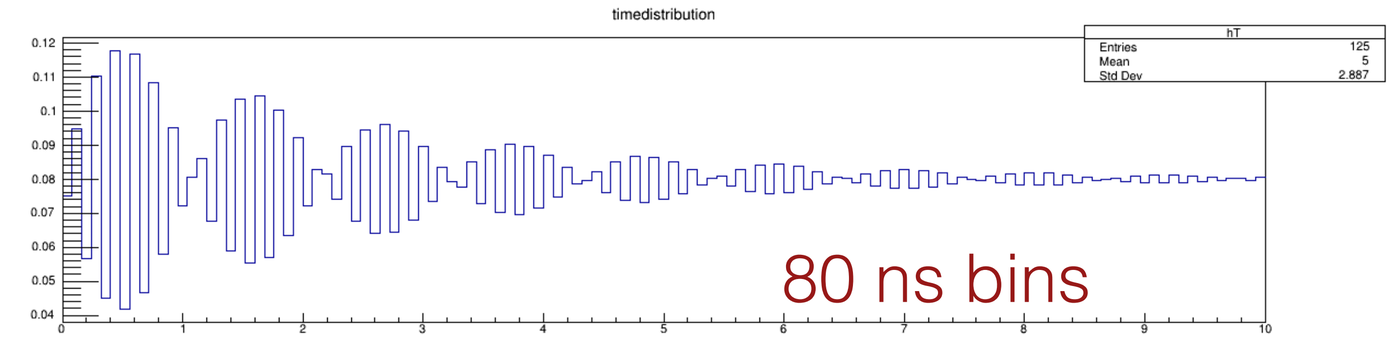
GPU Nodes on Open Science Grid

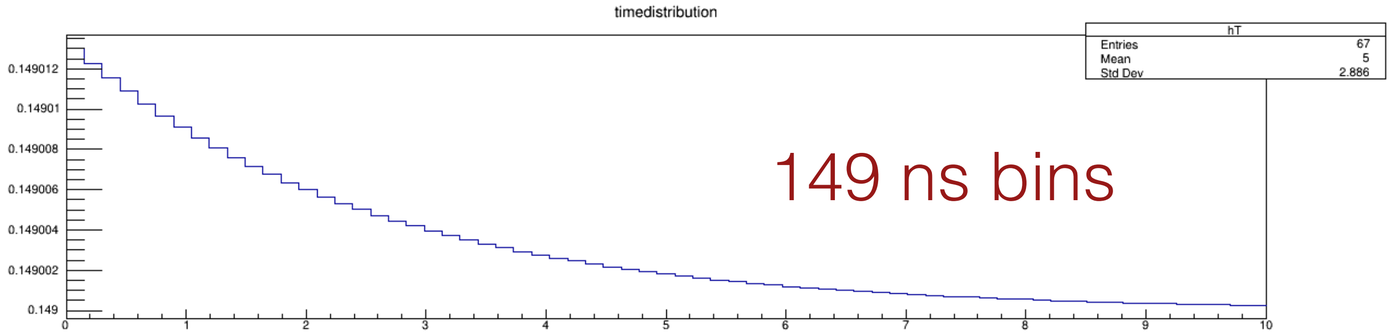
- Run on Open Science Grid GPU nodes to simulate 10^{11} positrons in 2.5 hours.
- Currently two OSG sites with GPU nodes available:
  - Omaha: Nvidia Tesla P100, 3584 Cuda cores
  - Syracuse Orange Grid: GeForce GTX 750 Ti, 640 Cuda cores
- Must insure that code is optimized for either architecture.



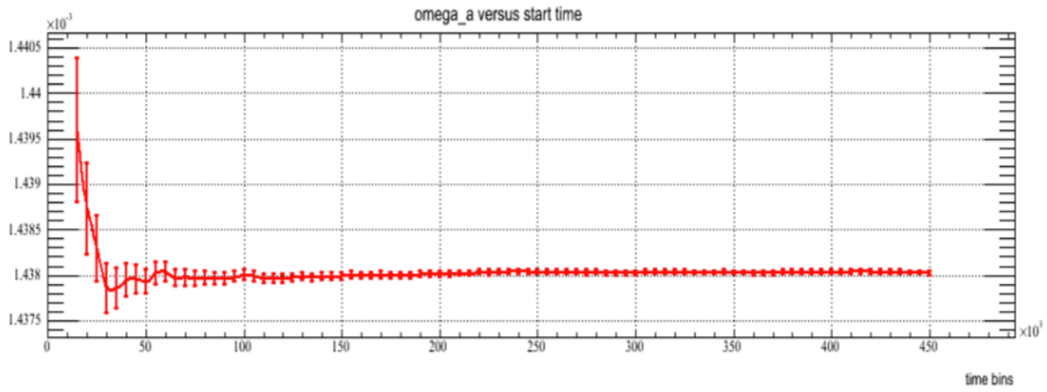
What we learn from the simulation

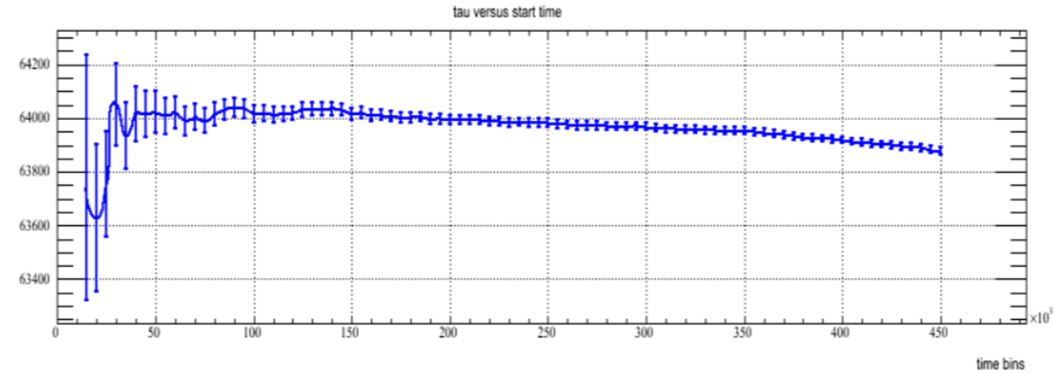
- Results of the simulation were used to set DAQ parameters.



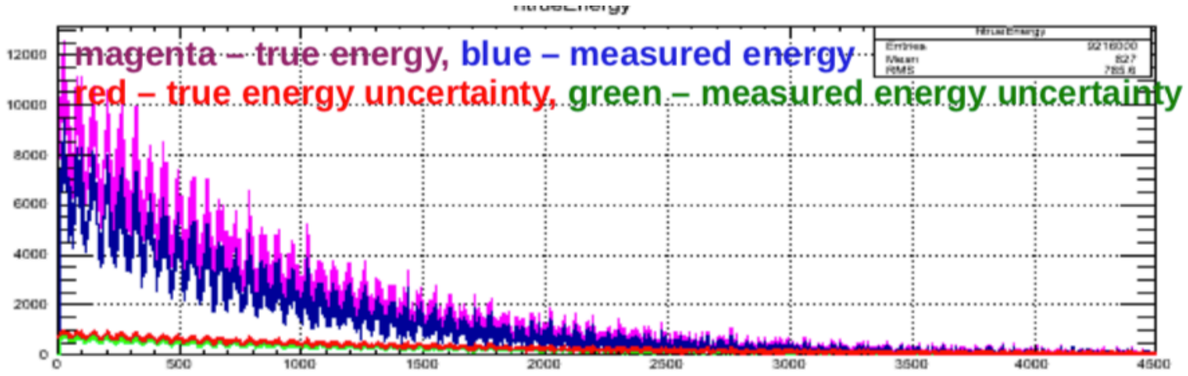
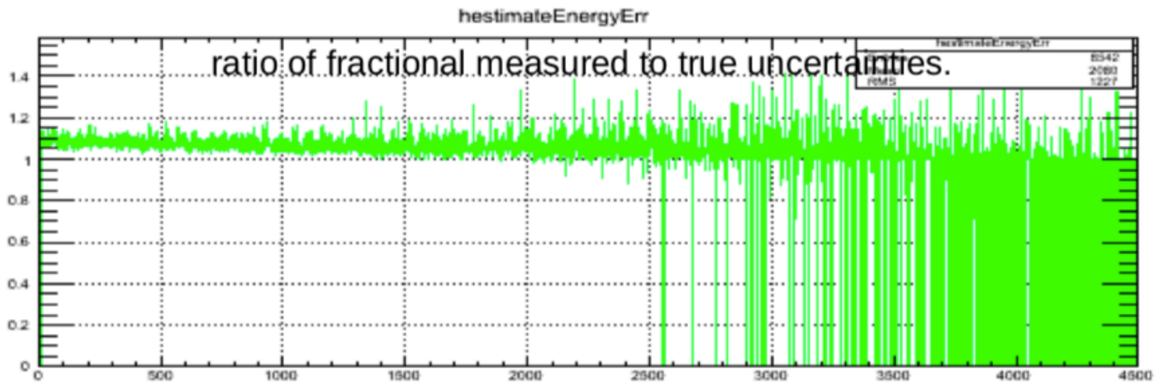


Beam dynamics effects on time spectrum





Fit end-time scans

Improved Estimation of statistical uncertainties

- Must calculate statistical error on *sum* of integrals from many bins of histogram.
- How do we handle lost energy?
  - Particle that deposits a portion of energy in an edge crystal
  - Energy that does not pass threshold in one crystal but does in another
- How does pileup affect the uncertainty?