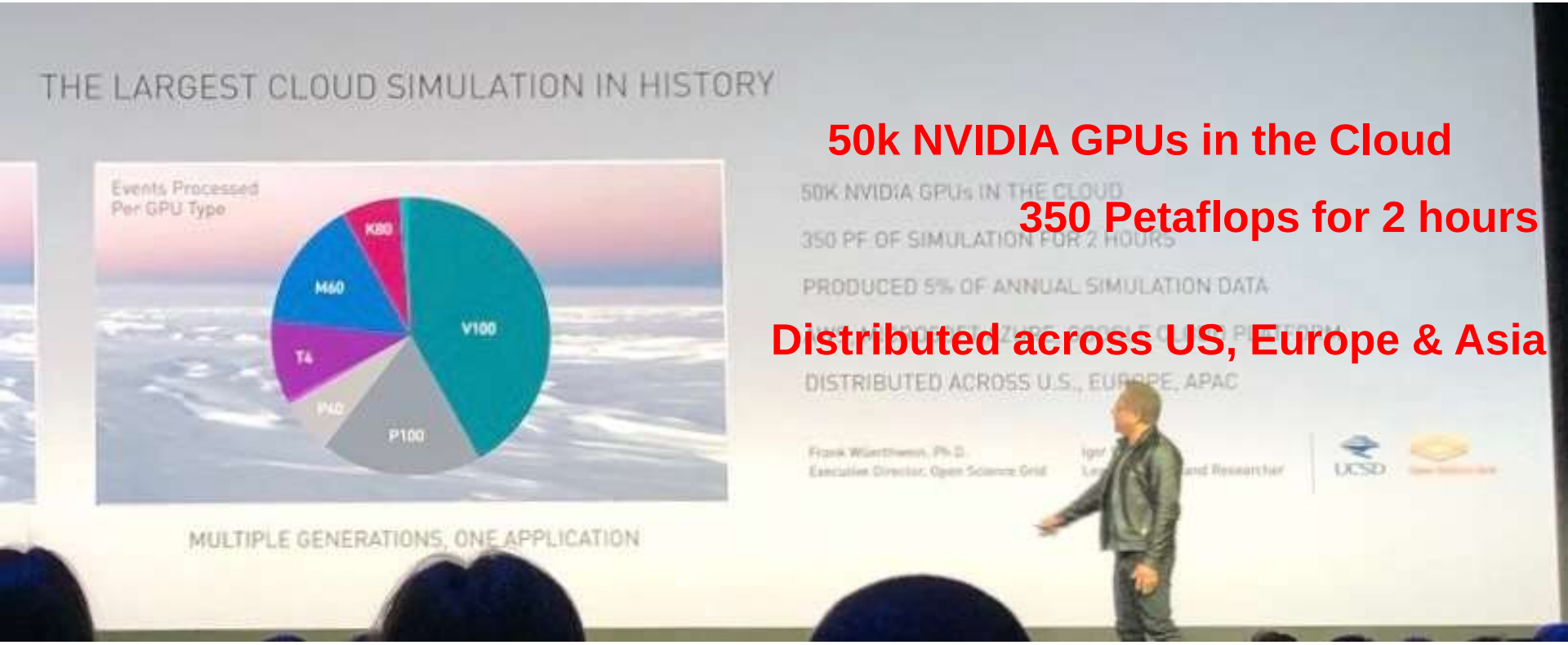
Running a GPU burst for Multi-Messenger Astrophysics with IceCube across all available GPUs in the Cloud

Frank Würthwein
OSG Executive Director
UCSD/SDSC

Jensen Huang keynote @ SC19

The Largest Cloud Simulation in History



50k NVIDIA GPUs in the Cloud

350 Petaflops for 2 hours

Distributed across US, Europe & Asia

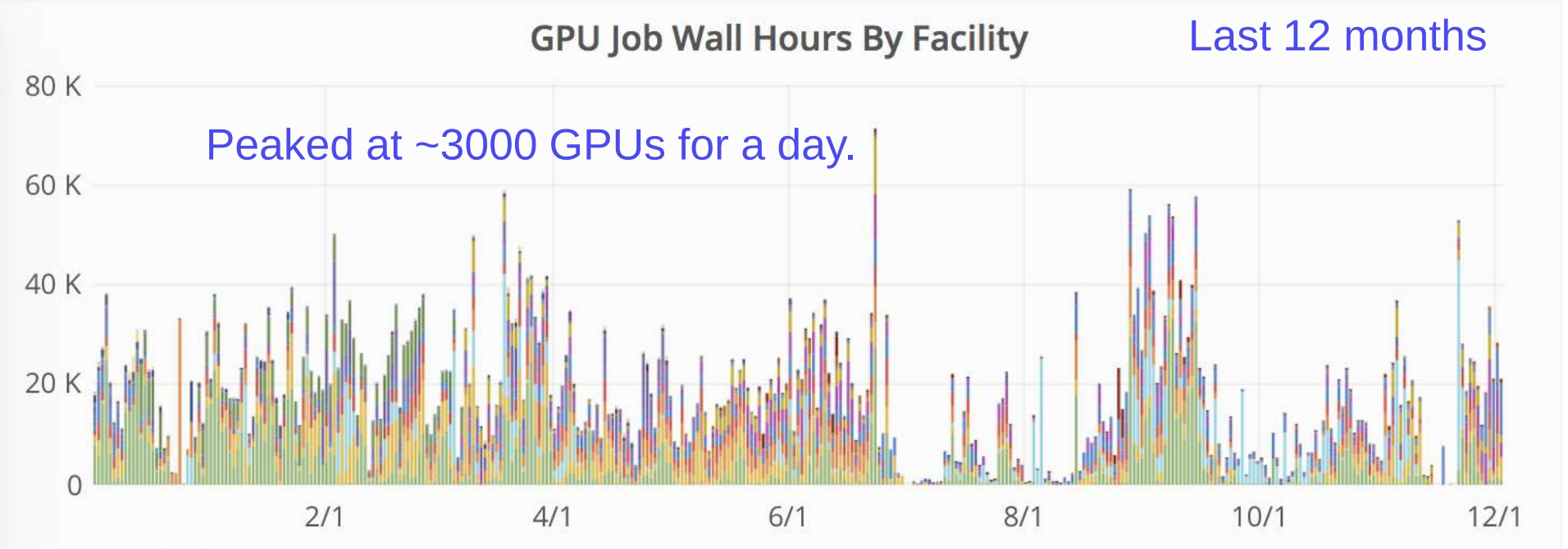
Saturday morning before SC19 we bought all GPU capacity that was for sale in Amazon Web Services, Microsoft Azure, and Google Cloud Platform worldwide

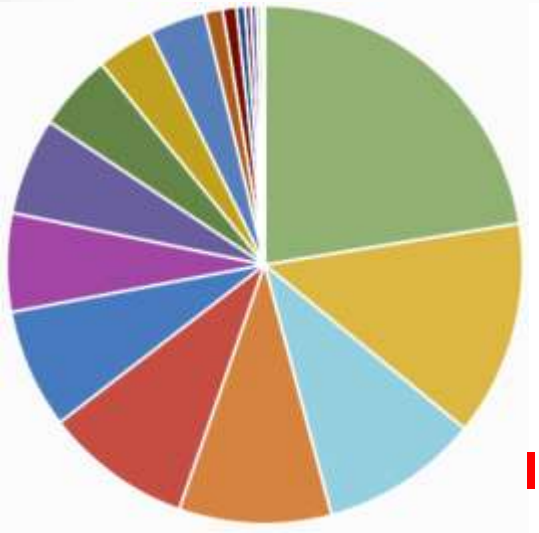
Open Science Grid

UCSD

How did we get here?

Annual IceCube GPU use via OSG





OSG supports global operations of IceCube. IceCube made long term investment into dHTC as their computing paradigm.

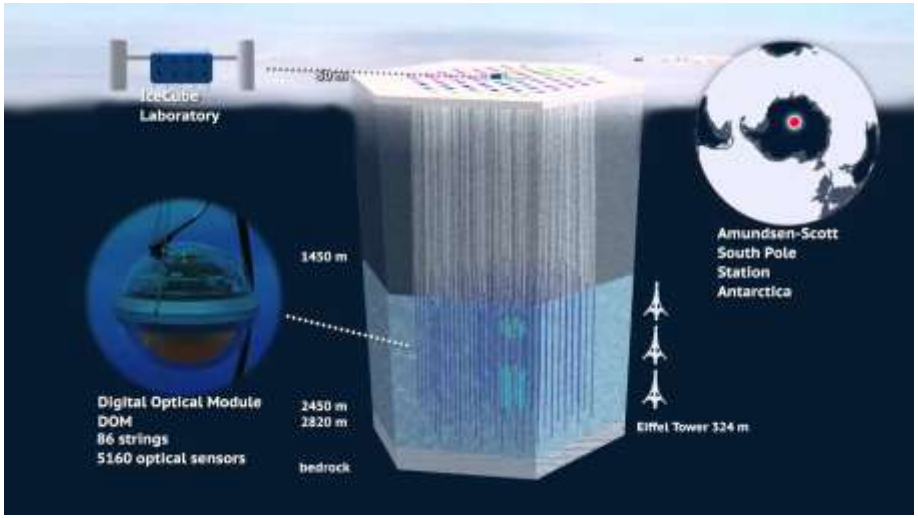
We produced ~3% of the annual photon propagation simulations in a ~2h cloud burst.

Longterm Partnership between IceCube, OSG, HTCondor, … lead to this cloud burst.

The Science Case

IceCube

A cubic kilometer of ice at the south pole is instrumented with 5160 optical sensors.



A facility with very diverse science goals

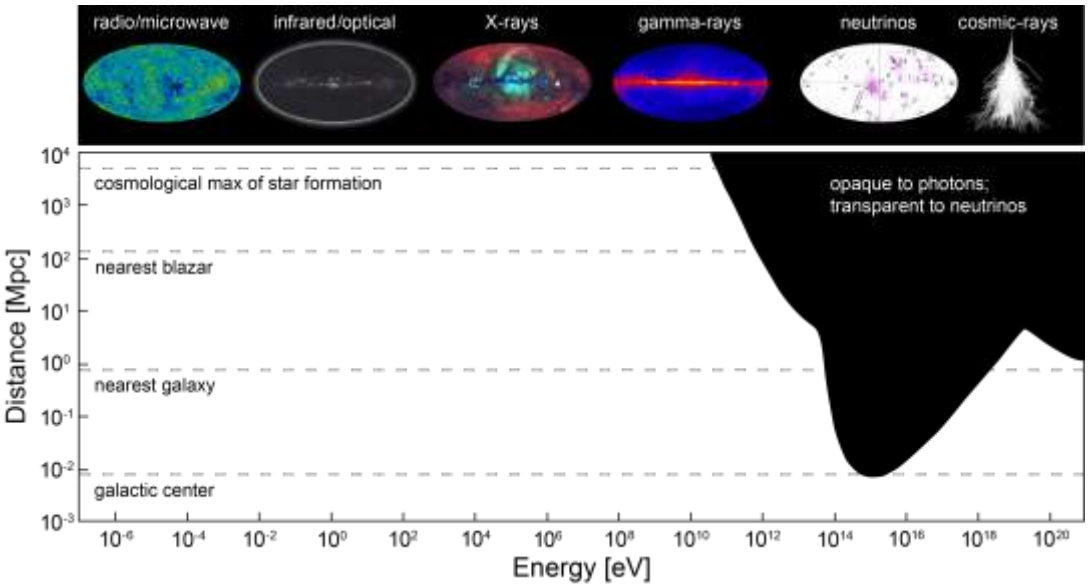
Astrophysics:
- Discovery of astrophysical neutrinos
- First evidence of neutrino point source (TXS)
- Cosmic rays with surface detector

Particle Physics:
- Atmospheric neutrino oscillation
- Neutrino cross sections at TeV scale
- New physics searches at highest energies

Earth Science:
- Glaciology
- Earth tomography

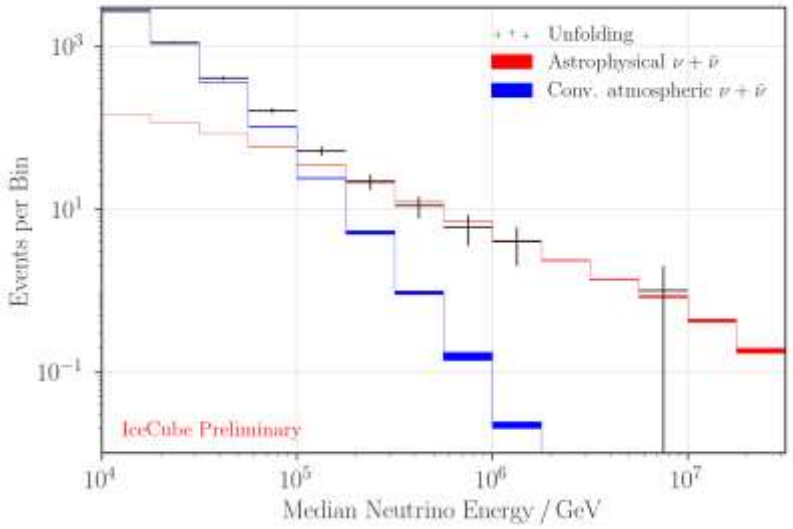
Restrict this talk to high energy Astrophysics

High Energy Astrophysics Science case for IceCube



Universe is opaque to light at highest energies and distances.

Only gravitational waves and neutrinos can pinpoint most violent events in universe.

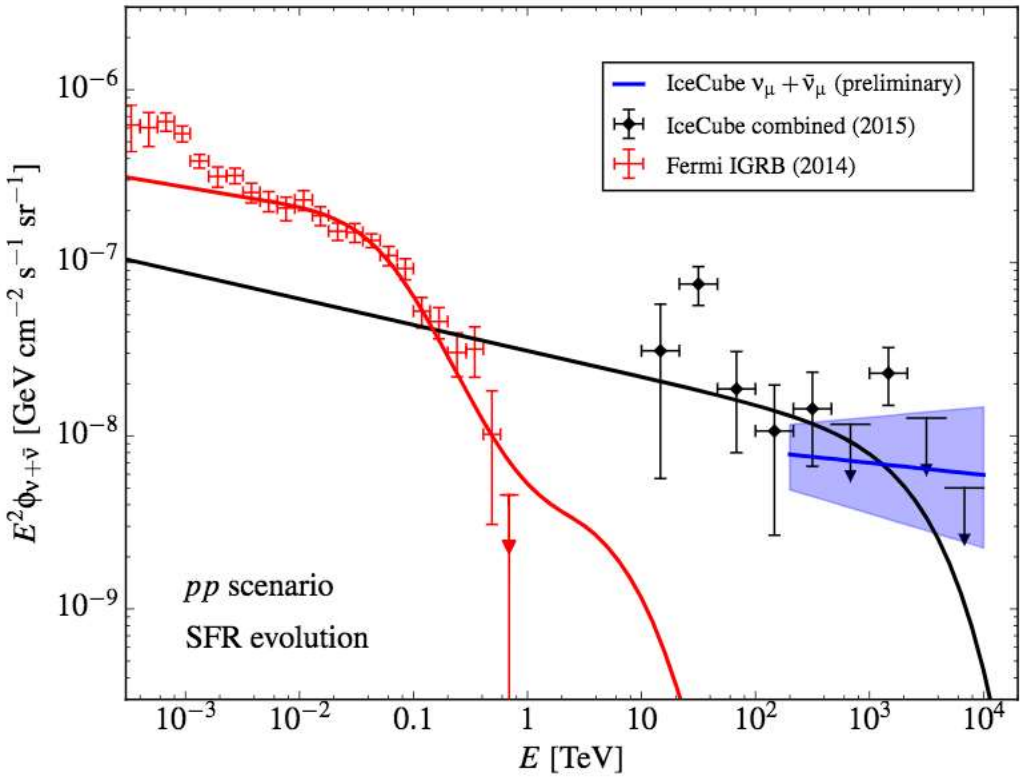


Fortunately, highest energy neutrinos are of cosmic origin.

Effectively "background free" as long as energy is measured correctly.

High energy neutrinos from outside the solar system

First 28 very high energy neutrinos from outside the solar system



Science 342 (2013). DOI: 10.1126/science.1242856

Red curve is the photon flux spectrum measured with the Fermi satellite.

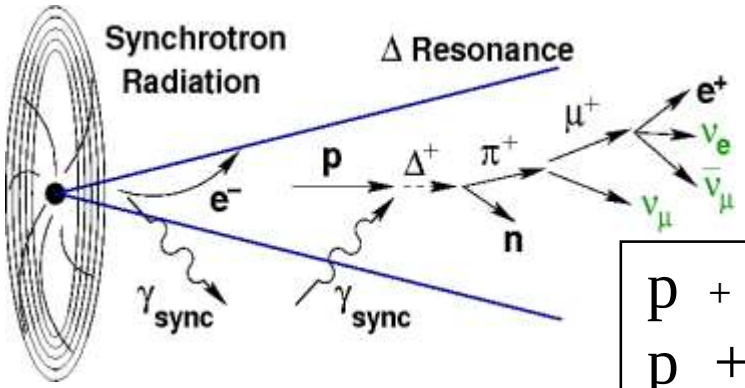
Black points show the corresponding high energy neutrino flux spectrum measured by IceCube.

This demonstrates both the opaqueness of the universe to high energy photons, and the ability of IceCube to detect neutrinos above the maximum energy we can see light due to this opaqueness.

Understanding the Origin

We now know high energy events happen in the universe. What are they?

The hypothesis:



The same cosmic events produce neutrinos and photons

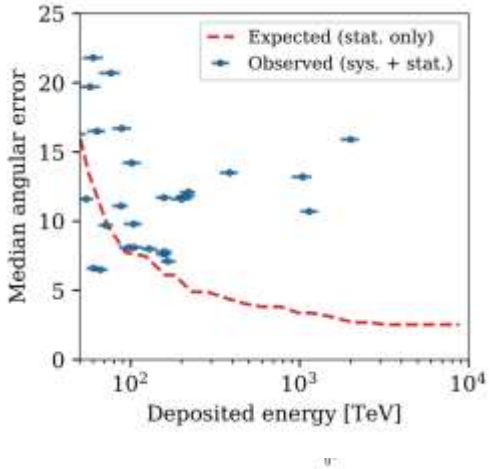
$$p + g \rightarrow D^{+} \rightarrow p + p^{0} \rightarrow p + gg$$
$$p + g \rightarrow D^{+} \rightarrow n + p^{+} \rightarrow n + m + n$$

We detect the electrons or muons from neutrino that interact in the ice.
Neutrino interact very weakly => **need a very large array of ice instrumented** to maximize chances that a cosmic neutrino interacts inside the detector.
Need pointing accuracy to point back to origin of neutrino.
Telescopes the world over then try to identify the source in the direction IceCube is pointing to for the neutrino.

Multi-messenger Astrophysics

The ν detection challenge

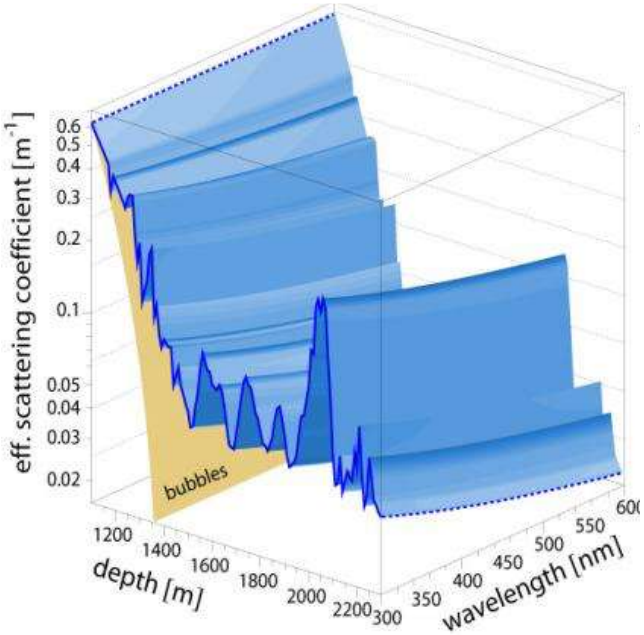
Observed pointing resolution at high energies is systematics limited.

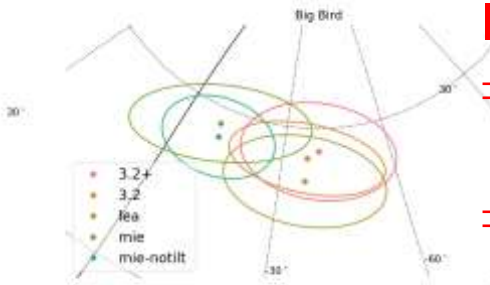


Photon propagation through ice runs efficiently on single precision GPU.

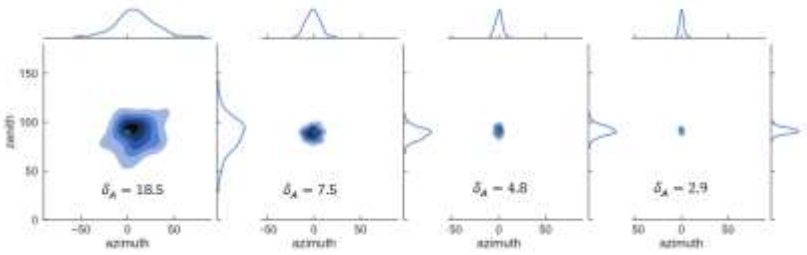
Detailed simulation campaigns to improve pointing resolution by improving ice model.

Ice properties change with depth and wavelength





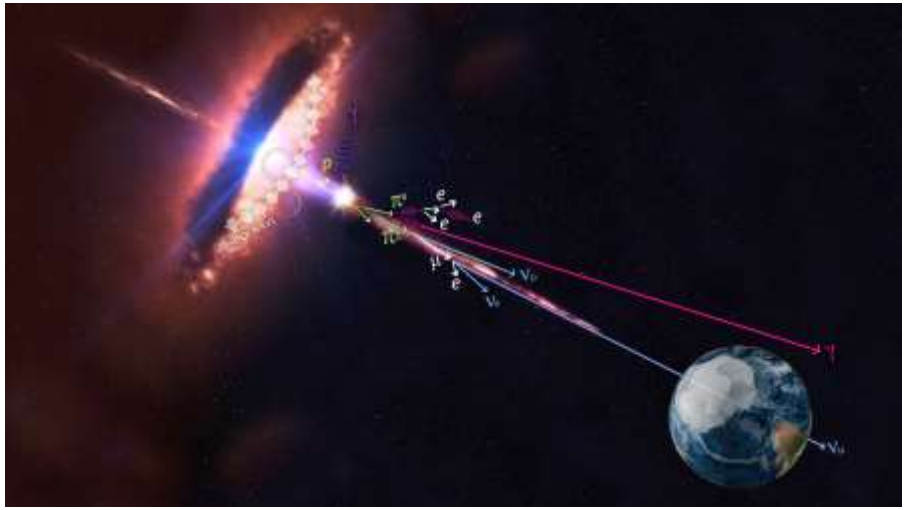
Improved e and τ reconstruction
⇒ increased neutrino flux detection
⇒ more observations



Central value moves for different ice models

Improvement in reconstruction with better ice model near the detectors

First evidence of an origin

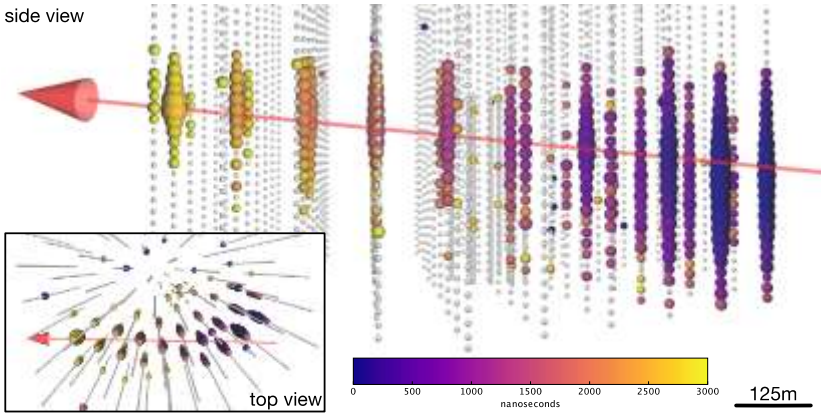


IceCube alerted the astronomy community of the observation of a single high energy neutrino on September 22 2017.

A blazar designated by astronomers as TXS 0506+056 was subsequently identified as most likely source in the direction IceCube was pointing. Multiple telescopes saw light from TXS at the same time IceCube saw the neutrino.

Science 361, 147-151 (2018). DOI:10.1126/science.aat2890

First location of a source of very high energy neutrinos.



Neutrino produced high energy muon near IceCube. Muon produced light as it traverses IceCube volume. Light is detected by array of phototubes of IceCube.

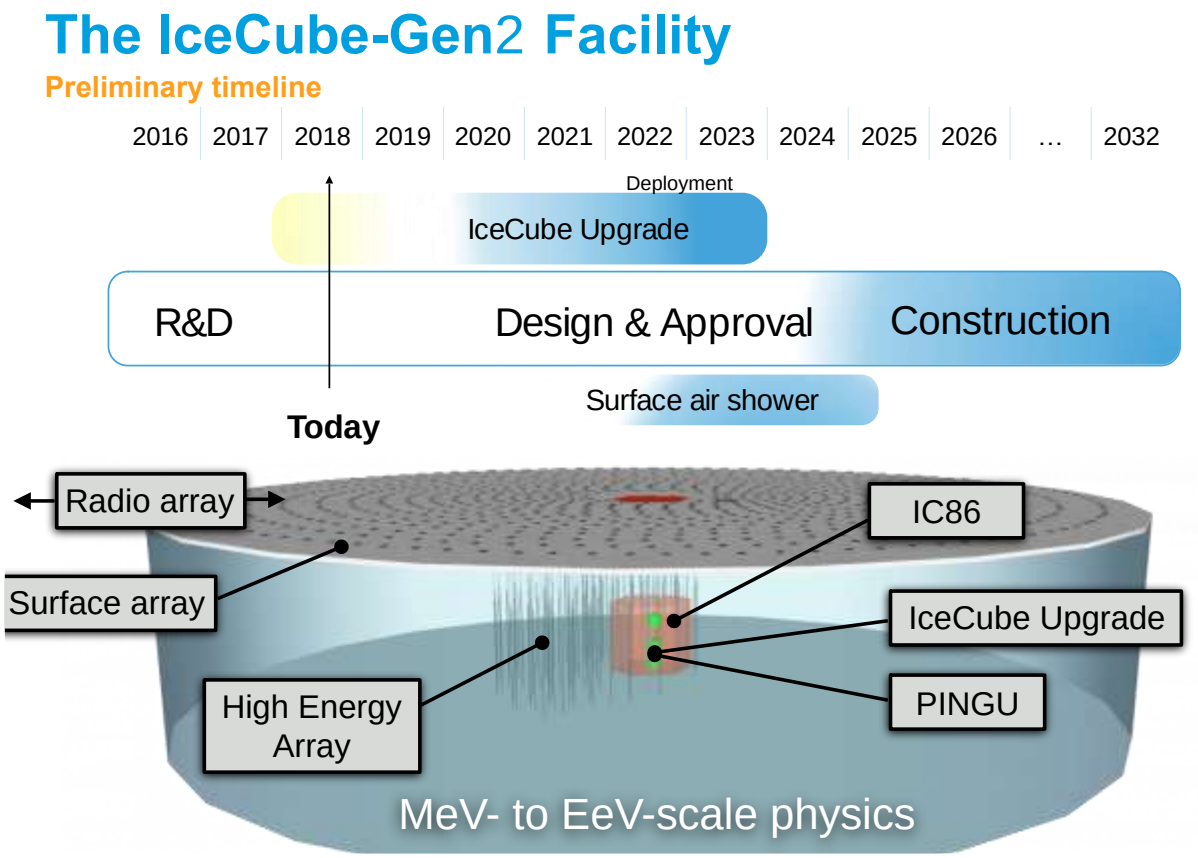
IceCube's Future Plans

Near term:

add more phototubes to deep core to increase granularity of measurements.

Longer term:

- Extend instrumented volume at smaller granularity.
- Extend even smaller granularity deep core volume.
- Add surface array.



Improve detector for low & high energy neutrinos

Details on the Cloud Burst

The Idea

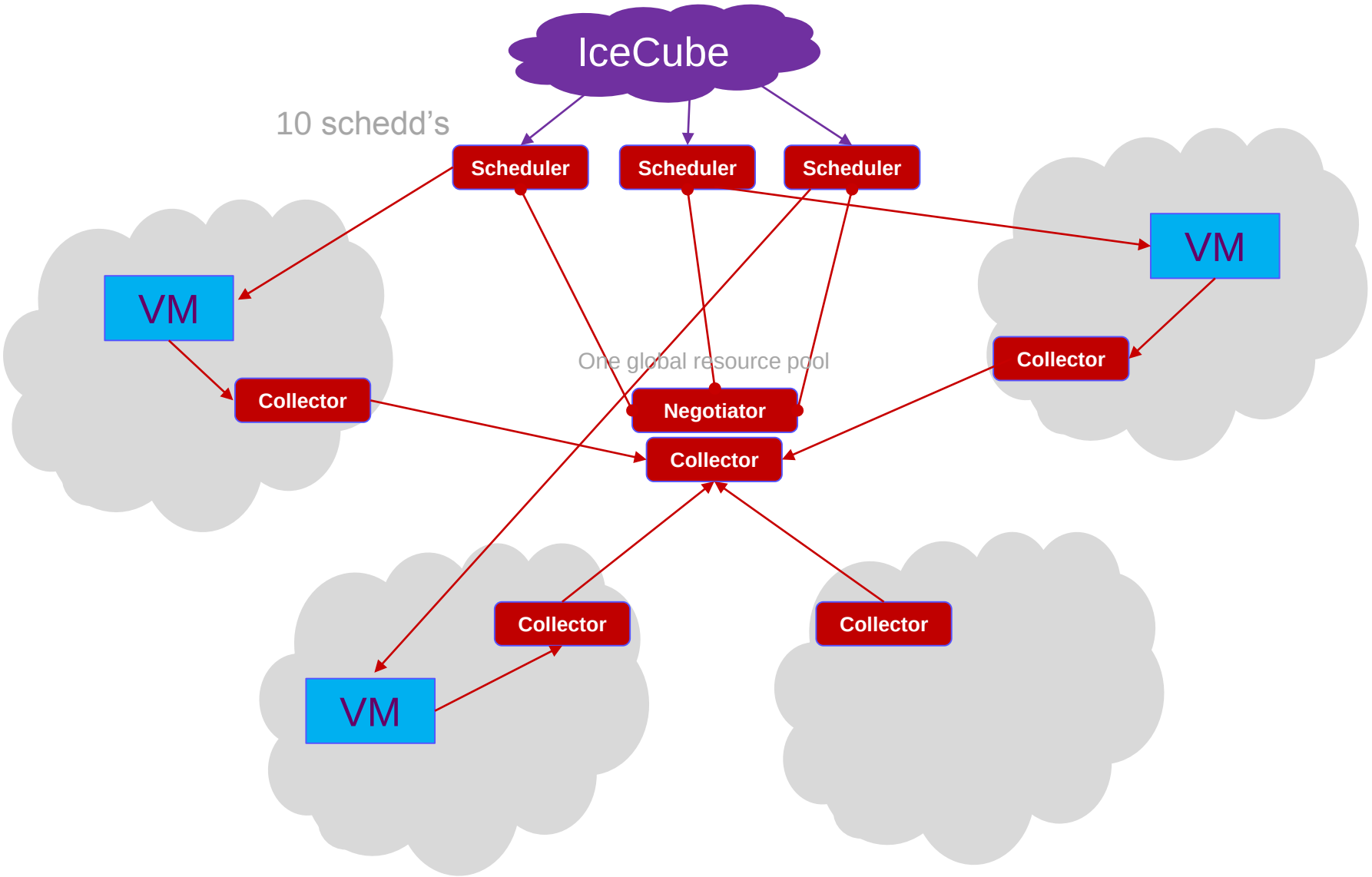
- Integrate all GPUs available for sale worldwide into a single HTCondor pool.
  - use 28 regions across AWS, Azure, and Google Cloud for a burst of a couple hours, or so.
- IceCube submits their photon propagation workflow to this HTCondor pool.
  - we handle the input, the jobs on the GPUs, and the output as a single globally distributed system.

Run a GPU burst relevant in scale for future Exascale HPC systems.

A global HTCondor pool

- IceCube, like all OSG user communities, relies on HTCondor for resource orchestration
  - This demo used the standard tools
- Dedicated HW setup
  - Avoid disruption of OSG production system
  - Optimize HTCondor setup for the spiky nature of the demo
    - multiple schedds for IceCube to submit to
    - collecting resources in each cloud region, then collecting from all regions into global pool

HTCondor Distributed CI

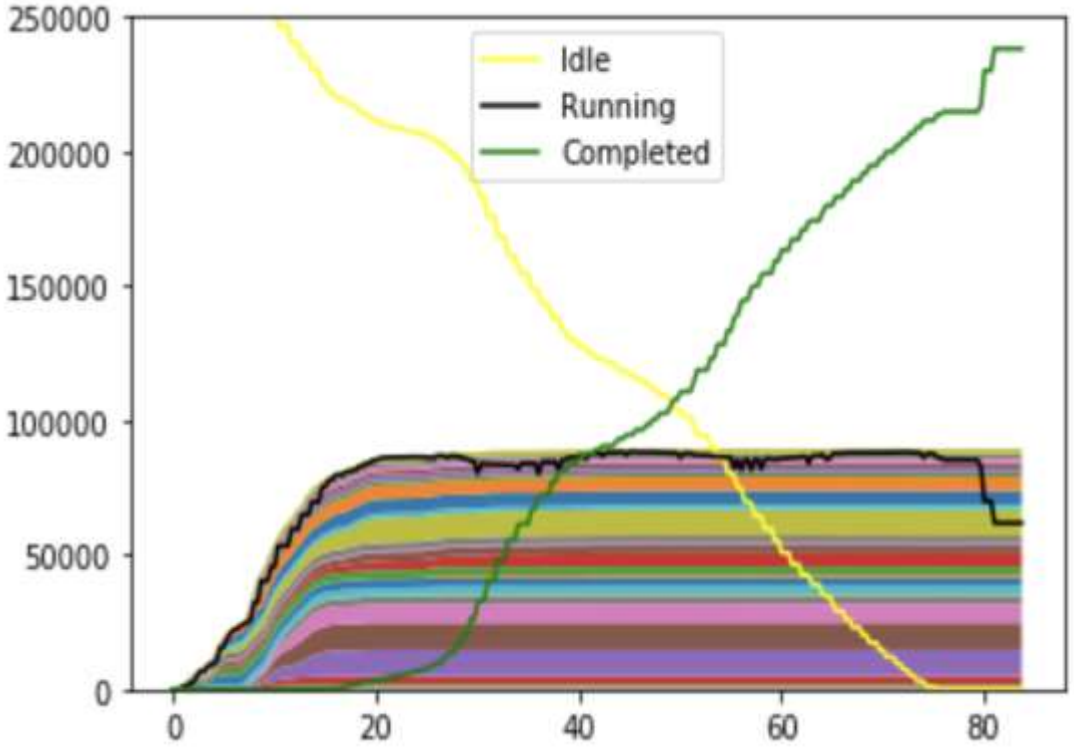


Using native Cloud storage

- Input data pre-staged into native Cloud storage
  - Each file in one-to-few Cloud regions
    - some replication to deal with limited predictability of resources per region
  - Local to Compute for large regions for maximum throughput
  - Reading from “close” region for smaller ones to minimize ops
- Output staged back to region-local Cloud storage

- Deployed simple wrappers around Cloud native file transfer tools
  - IceCube jobs do not need to customize for different Clouds
  - They just need to know where input data is available (pretty standard OSG operation mode)

The Testing Ahead of Time



~250,000 single threaded jobs run across 28 cloud regions during 80 minutes.

Peak at 90,000 jobs running.

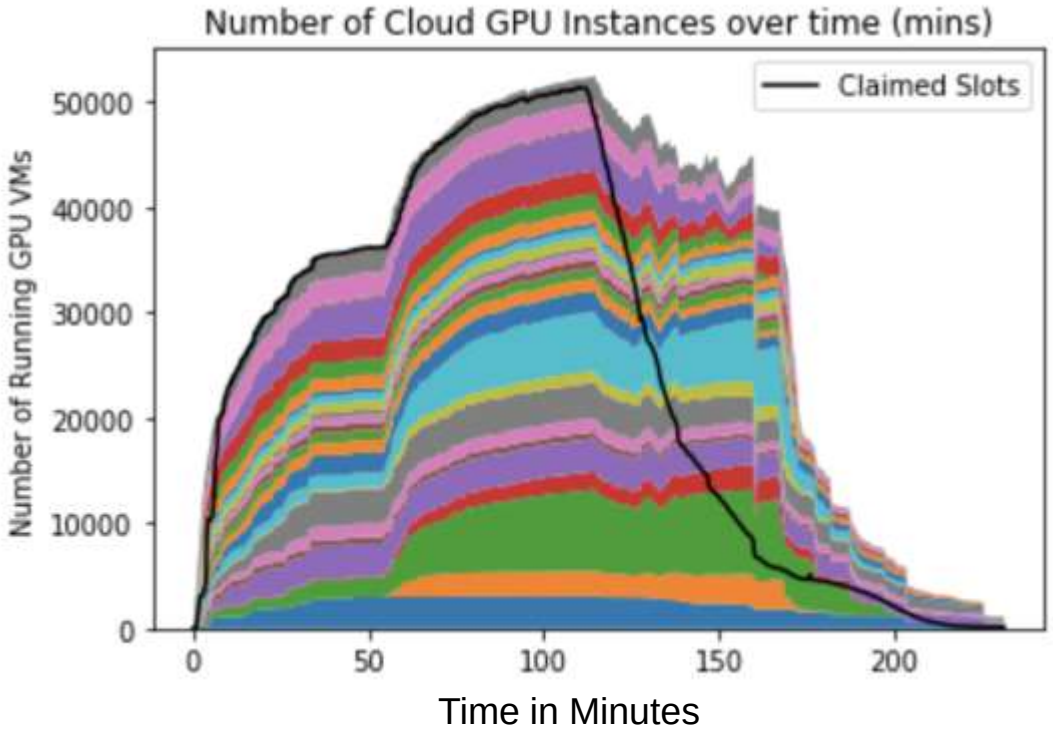
up to 60k jobs started in ~10min.

Regions across US, EU, and Asia were used in this test.

Demonstrated burst capability of our infrastructure on CPUs.

Want scale of GPU burst to be limited only by # of GPUs available for sale.

Science with 51,000 GPUs achieved as peak performance



Each color is a different cloud region in US, EU, or Asia.

Total of 28 Regions in use.

Peaked at 51,500 GPUs

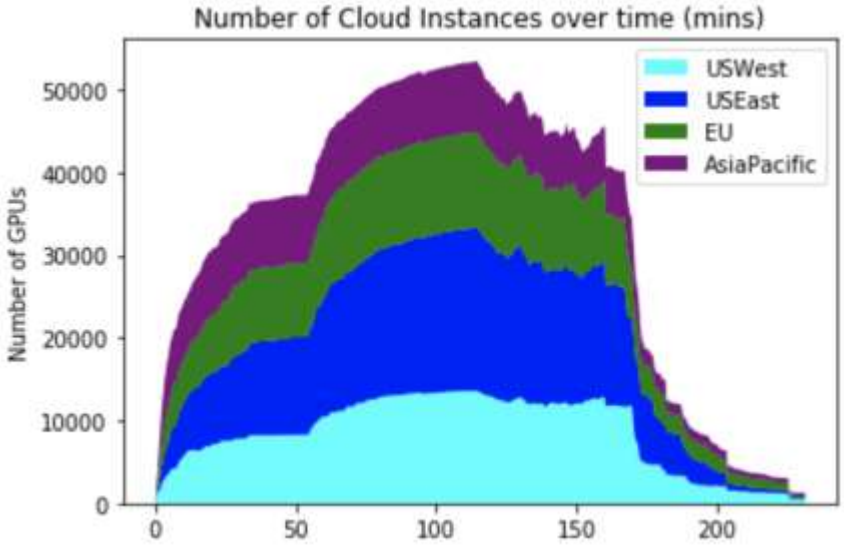
~380 Petaflops of fp32

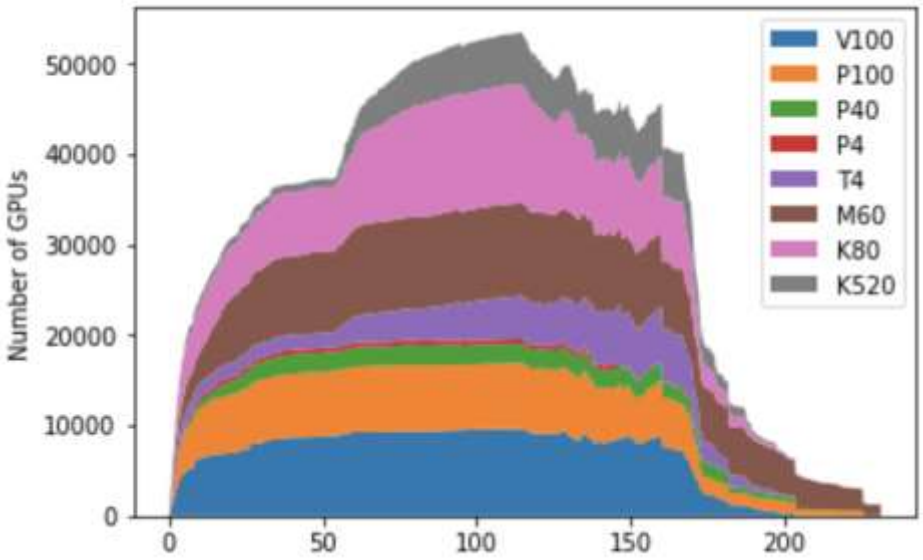
Summary of stats at peak

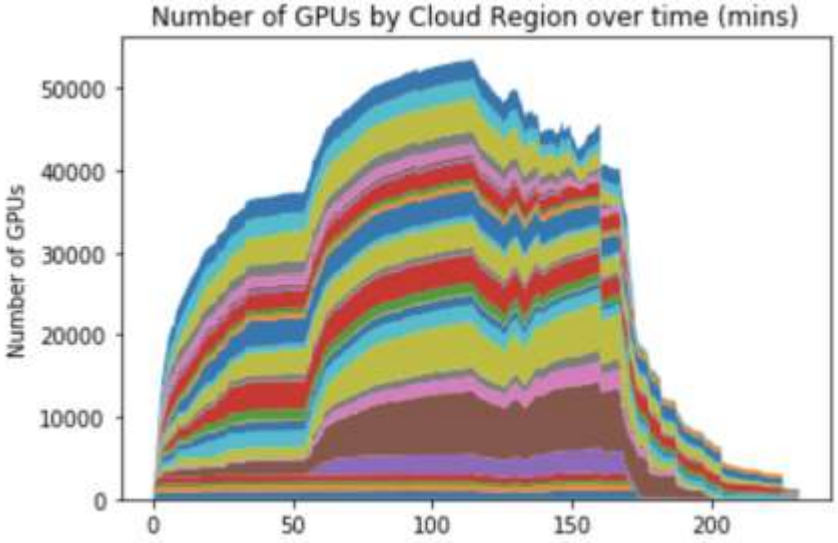
	V100	P100	P40	P4	T4	M60	K80	K520
Num GPUS	9.2k	7.2k	2.1k	0.5k	4.6k	10.1k	12.5k	5.4k
PFLOP32s	132.2	68.1	25.2	2.5	38.6	48.8	51.6	12.4

8 generations of NVIDIA GPUs used.

A Heterogenous Resource Pool



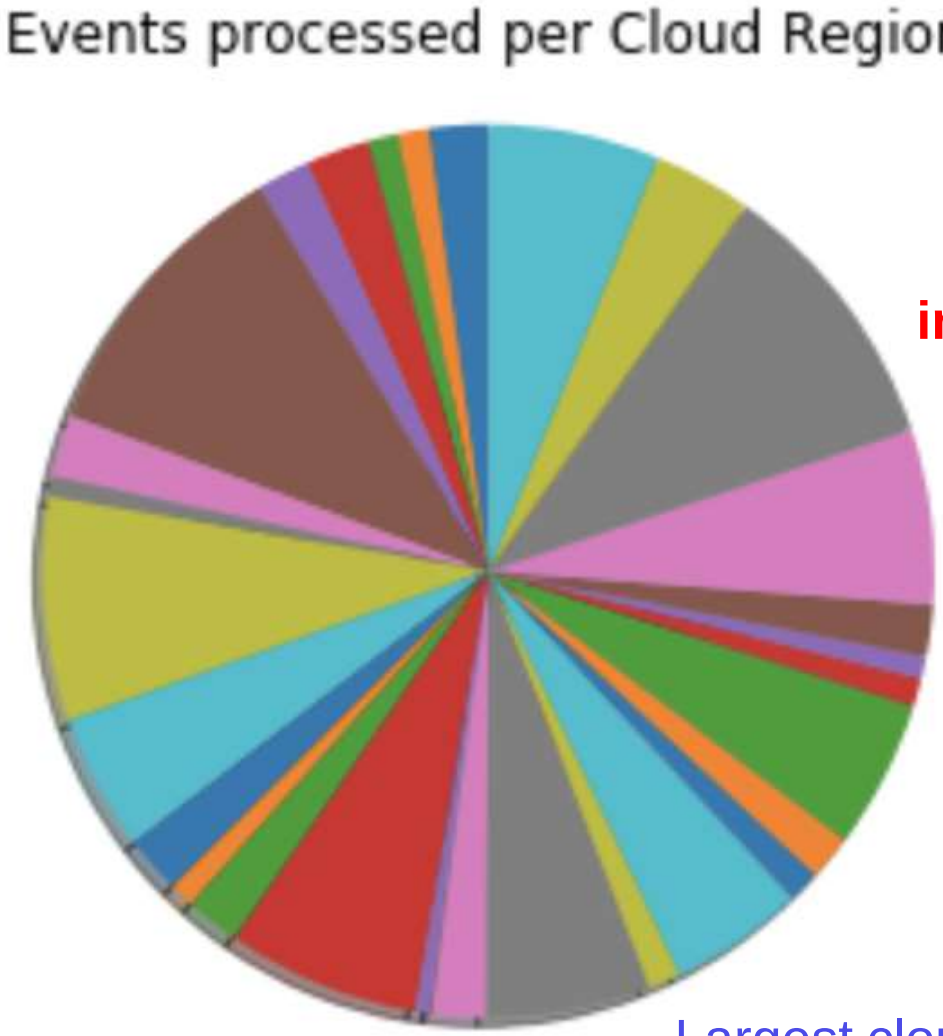




28 cloud Regions across 4 world regions providing us with 8 GPU generations.

No one region or GPU type dominates!

Science Produced



Distributed High-Throughput Computing (dHTC) paradigm implemented via HTCondor provides global resource aggregation.

dHTC paradigm can aggregate
on-prem anywhere
HPC at any scale
and multiple clouds

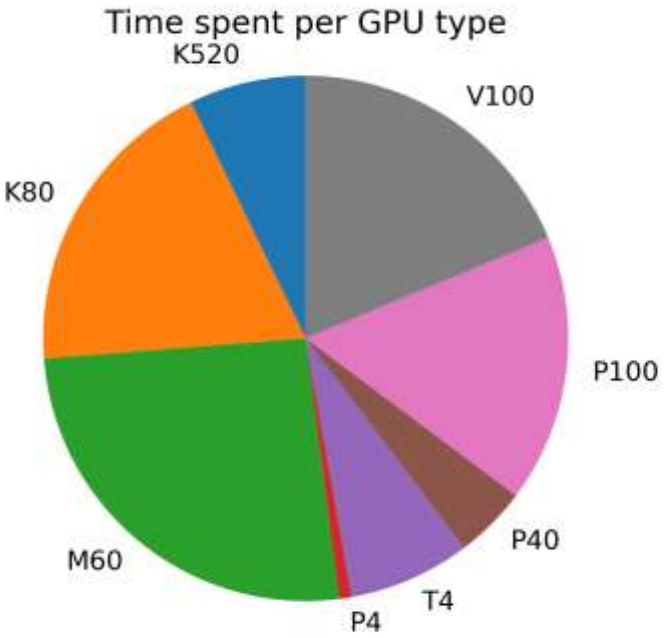
Largest cloud region provided 10.8% of the total

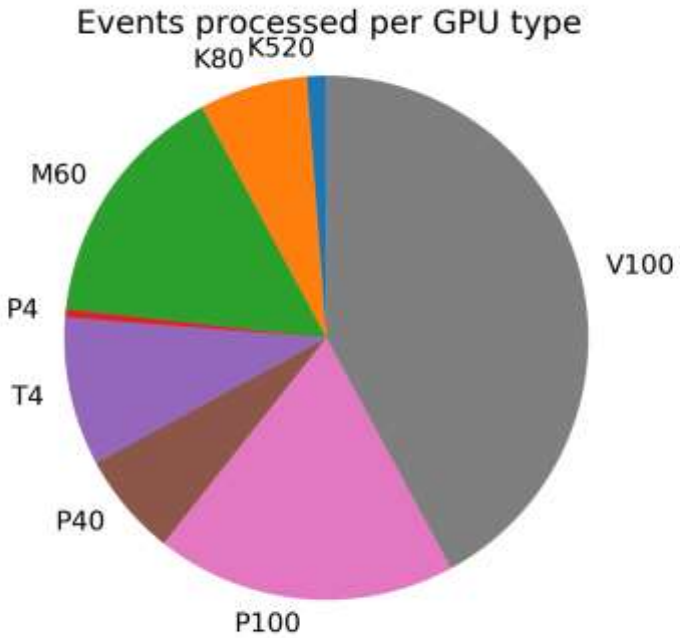
Open Science Grid

UCSD

Performance and Cost Analysis

Performance vs GPU type





42% of the science was done on V100 in 19% of the wall time.

IceCube Performance/$$$

GPU	V100	P100	T4	M60	RTX 2080 Ti	GTX 1080 Ti
Relative TFLOP32	100%	67%	57%	34%*	82%	62%
Relative IceCube Performance	100%	56%	48%	30%*	70%	48%
Relative Science/$**	1.1-1.4	1.1-1.3	1.7-2.1	1.0-1.3	N/A	N/A

*per CUDA device

**spot market prices, range indicates cost differences between cloud vendors

IceCube performance scales better than TFLOP32 for high end GPUs
Science/$$$ x1.5 better for T4 than V100
Price differential between vendors ~10-30%
Price differential on-demand vs spot ~x3

Aside: Science/$$$ for on-prem of 2080Ti ~x3 better than V100

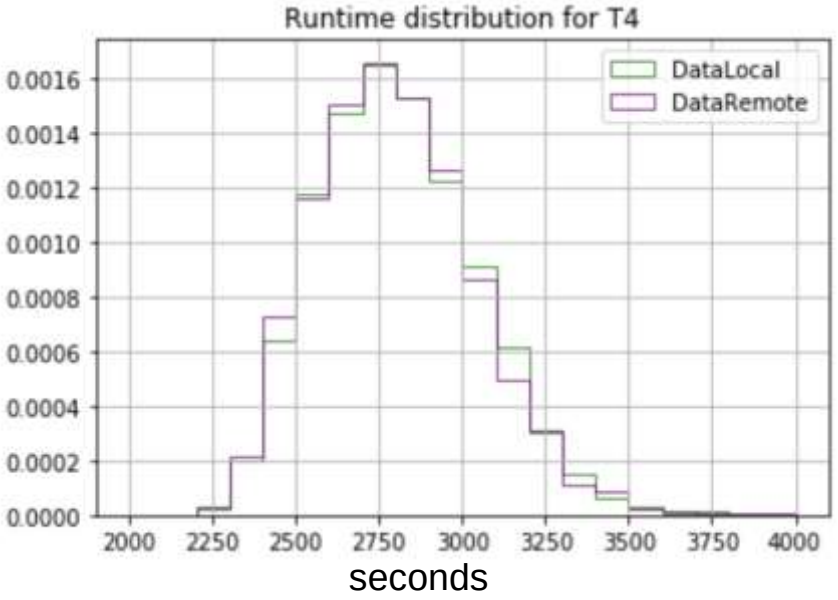
Open Science Grid

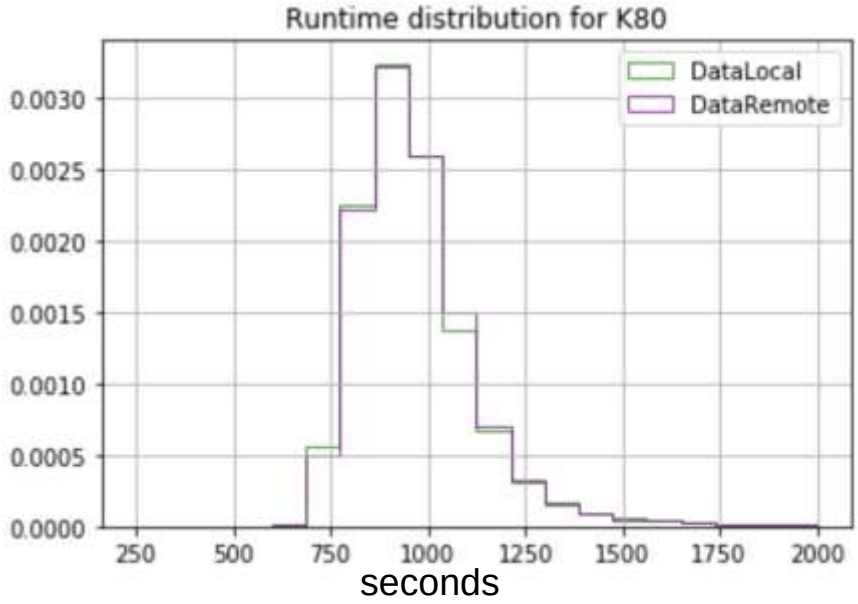
UCSD

IceCube and dHTC

dHTC = distributed High Throughput Computing

IceCube Input Segmentable





IceCube prepared two types of input files that differed in x10 in the number of input events per file.

Small files processed by K80 and K520, large files by all other GPU types.

A total of 10.2 Billion events were processed across ~175,000 GPU jobs.

Each job fetched a file from cloud storage to local storage, processed that file, and wrote the output to cloud storage. For ¼ of the regions cloud storage was not local to the region. => we could have probably avoided data replication across regions given the excellent networking between regions for each provider.

Applicability beyond IceCube

- All the large instruments we know off
  - LHC, LIGO, DUNE, LSST, …
- Any midscale instrument we can think off
  - XENON, GlueX, Clas12, Nova, DES, Cryo-EM, …
- A large fraction of Deep Learning
  - But not all of it …
- Basically, anything that has bundles of independently schedulable jobs that can be partitioned to adjust workloads to have 0.5 to few hour runtimes on modern GPUs.

Cost to support cloud as a “24x7” capability

- Today, roughly $15k per 300 PFLOP32 hour
- This burst was executed by 2 people
  - Igor Sfiligoi (SDSC) to support the infrastructure.
  - David Schultz (UW Madison) to create and submit the IceCube workflows.
    - “David” type person is needed also for on-prem science workflows.
- To make this a routine operations capability for any open science that is dHTC capable would require another 50% FTE “Cloud Budget Manager”.
  - There is substantial effort involved in just dealing with cost & budgets for a large community of scientists.

IceCube is ready for Exascale

- Humanity has built extraordinary instruments by pooling human and financial resources globally.
- The computing for these large collaborations fits perfectly to the cloud or scheduling holes in Exascale HPC systems due to its “ingeniously parallel” nature. => dHTC
- The dHTC computing paradigm applies to a wide range of problems across all of open science.
  - We are happy to repeat this with anybody willing to spend $50k in the clouds.

Demonstrated elastic burst at 51,500 GPUs

IceCube is ready for Exascale

Contact us at: help@opensciencegrid.org
Or me personally at: fkw@ucsd.edu

Acknowledgements

- This work was partially supported by the NSF grants OAC-1941481, MPS-1148698, OAC-1841530, and OAC-1826967

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov