



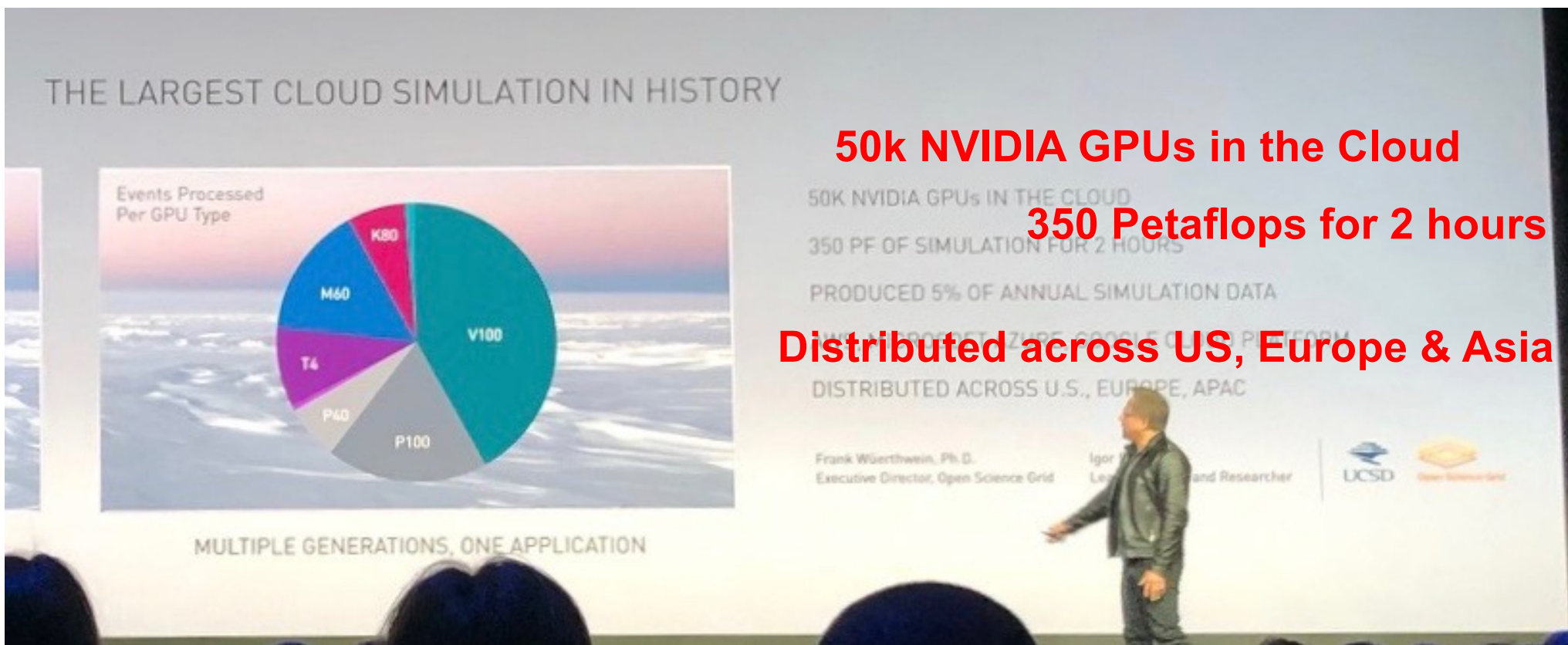
# Running a GPU burst for Multi-Messenger Astrophysics with IceCube across all available GPUs in the Cloud



**Frank Würthwein**  
**OSG Executive Director**  
**UCSD/SDSC**



## The Largest Cloud Simulation in History



Saturday morning before SC19 we bought all GPU capacity that was for sale worldwide across 3 commercial cloud providers, including AWS

# How did we get here?



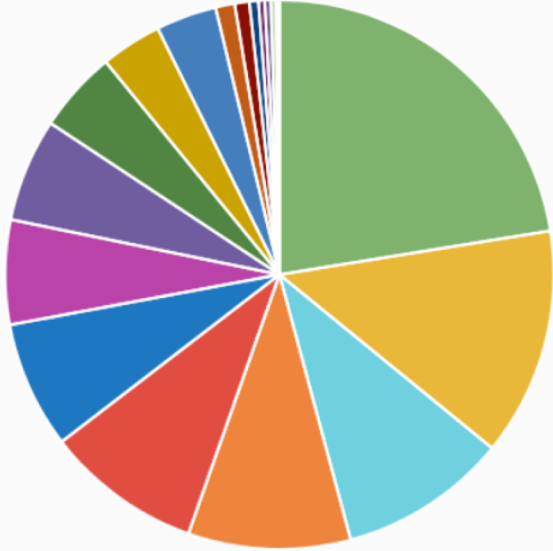
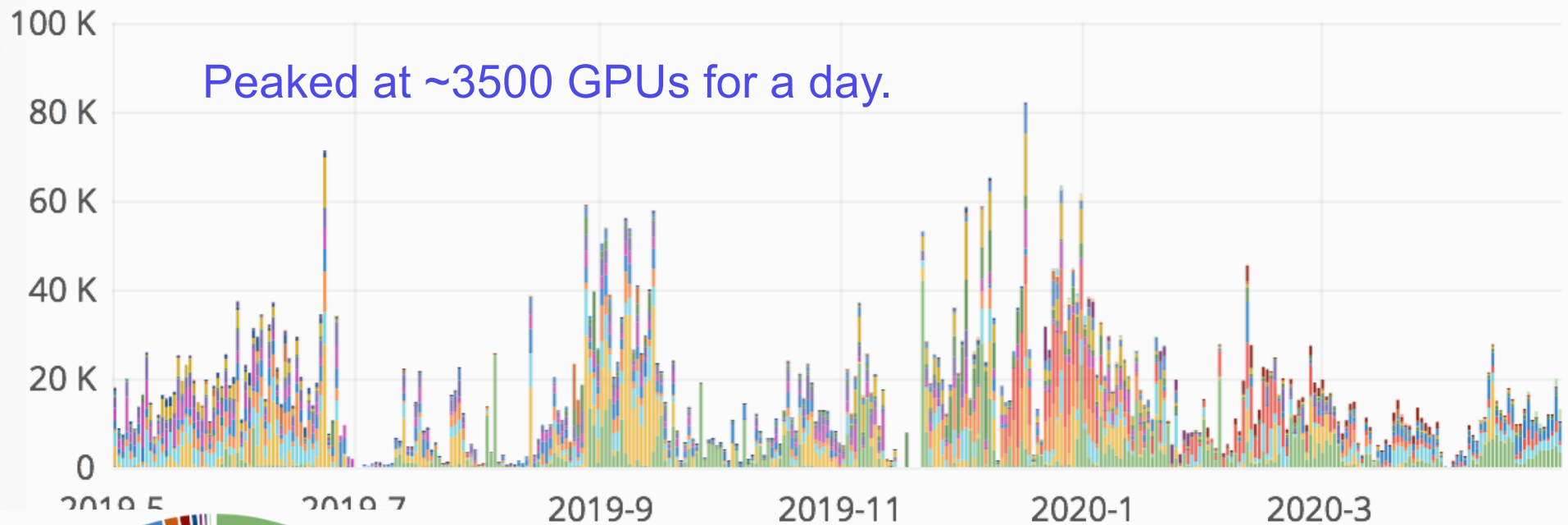
Open Science Grid

# Annual IceCube GPU use via OSG



GPU Job Wall Hours By Facility

Last Year



**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



Open Science Grid

# IceCube



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

**A facility with very diverse science goals**

Astrophysics:

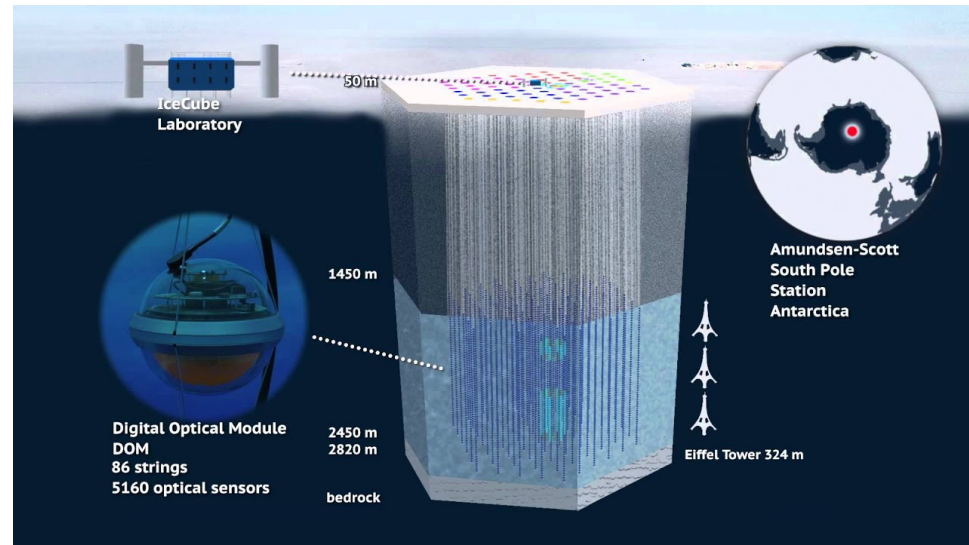
- Discovery of astrophysical neutrinos
- First evidence of neutrino point source (TXS)
- Cosmic rays with surface detector

Particle Physics:

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

Earth Science:

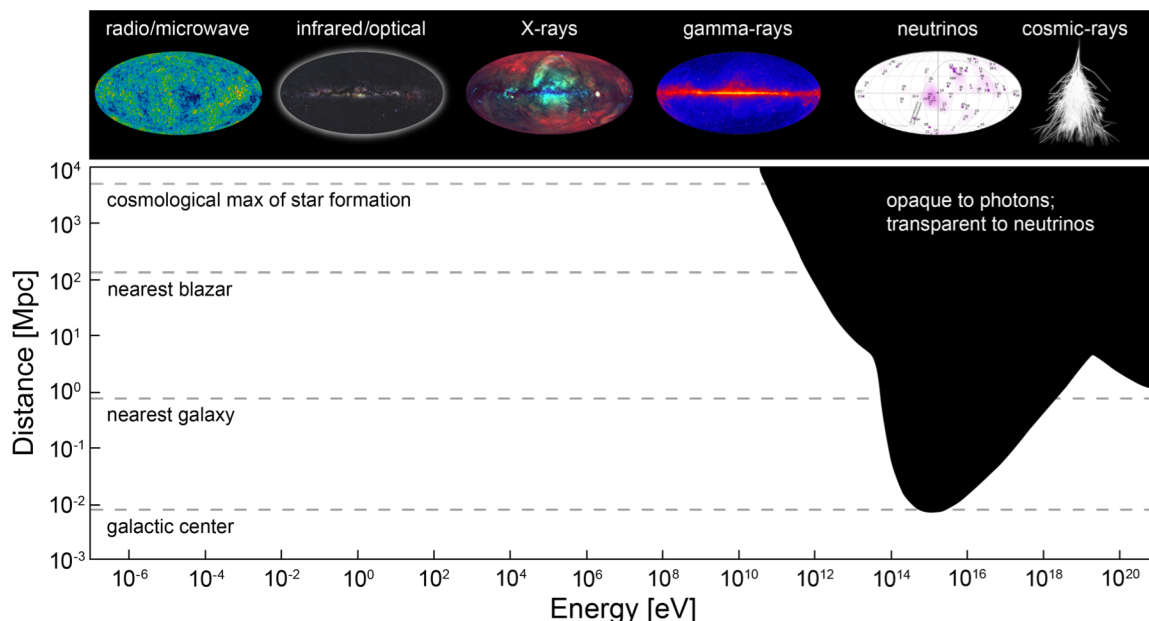
- Glaciology
- Earth tomography



**Restrict this talk to high energy Astrophysics**

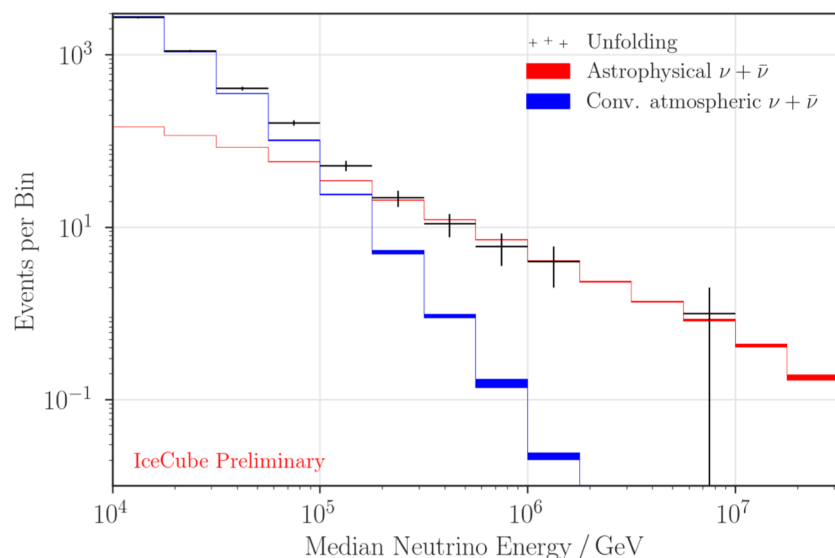


# High Energy Astrophysics Science case for IceCube



Universe is opaque to light at highest energies and distances.

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

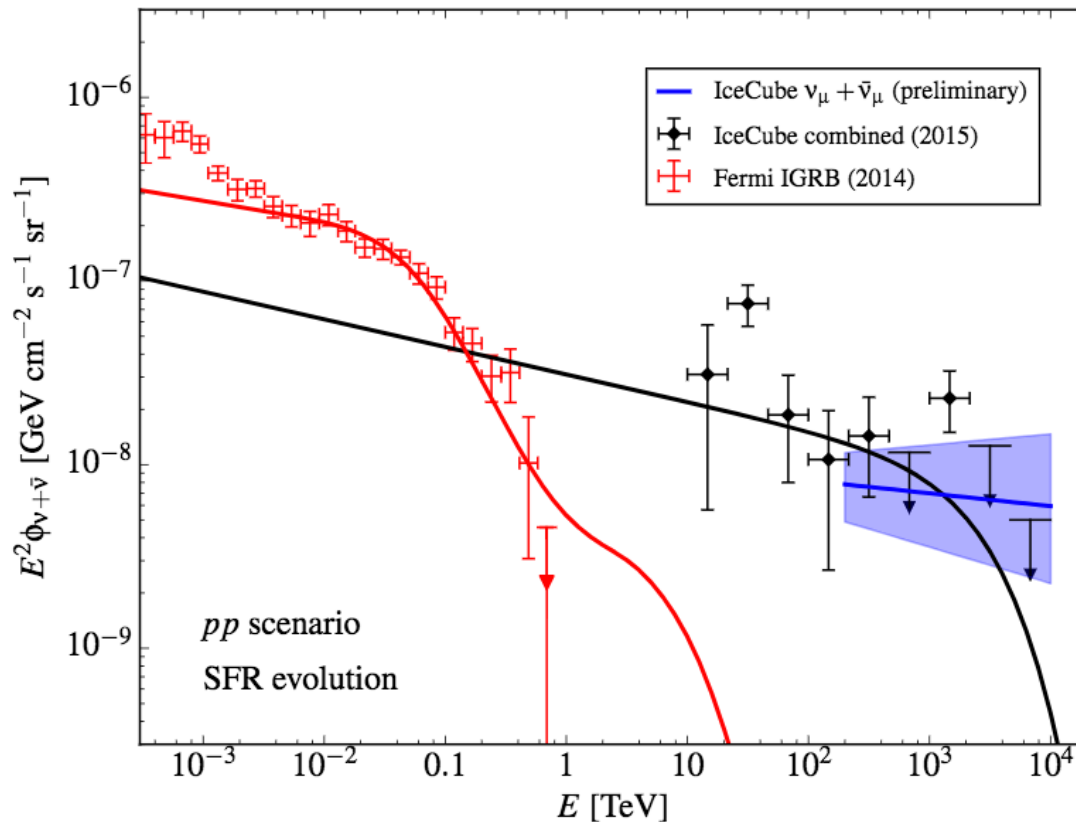


Fortunately, highest energy neutrinos are of cosmic origin.

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](https://doi.org/10.1126/science.1242856)

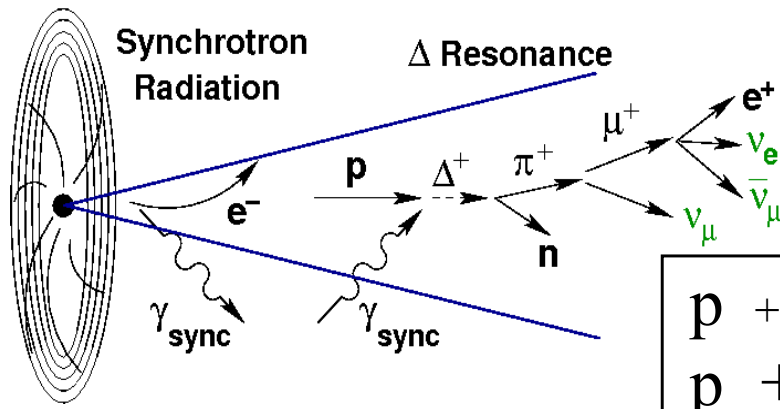
Red curve is the **photon** flux spectrum measured with the Fermi satellite.

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

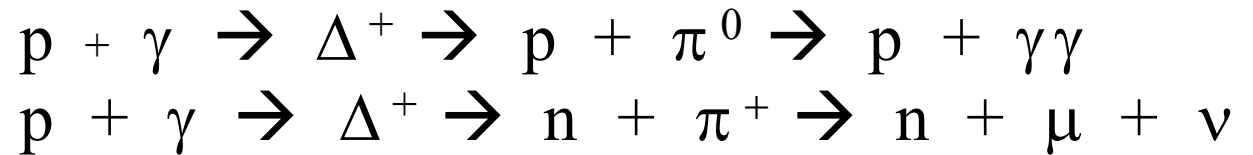
This **demonstrates** both the opaqueness of the universe to high energy photons, and **the ability of IceCube to detect neutrinos above the maximum energy we can see light** due to this opaqueness.

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

**The hypothesis:**



The same cosmic events produce neutrinos and photons



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**

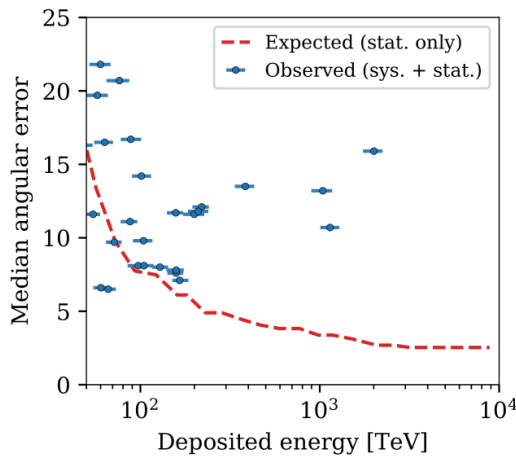


Open Science Grid

# The $\nu$ detection challenge

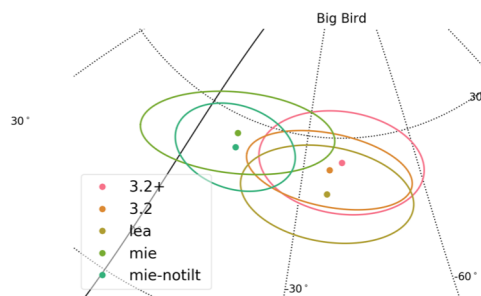


Observed pointing resolution at high energies is systematics limited.



Photon propagation through ice runs efficiently on single precision GPU.

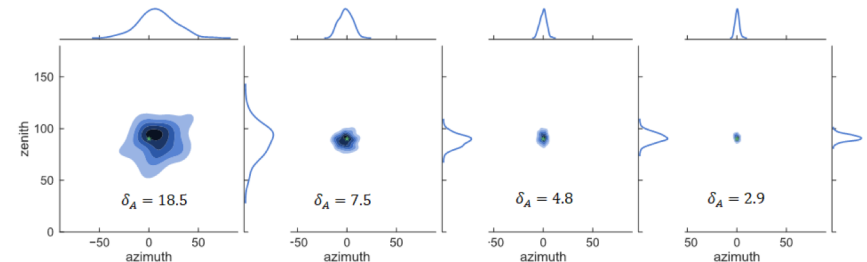
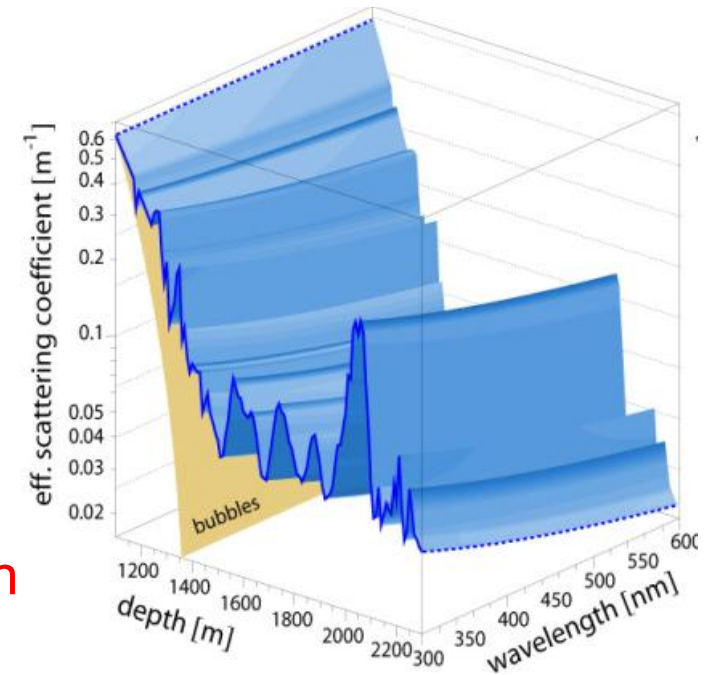
Detailed simulation campaigns to improve pointing resolution by improving ice model.



Improved  $e$  and  $\tau$  reconstruction  
 $\Rightarrow$  increased neutrino flux detection  
 $\Rightarrow$  more observations

Central value moves for different ice models

Ice properties change with depth and wavelength



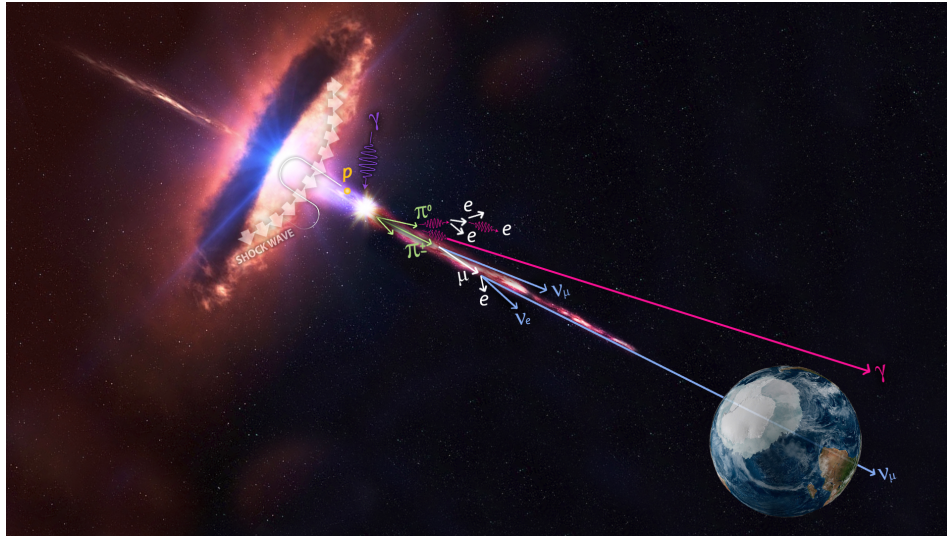
Improvement in reconstruction with better ice model near the detectors





Open Science Grid

# First evidence of an origin

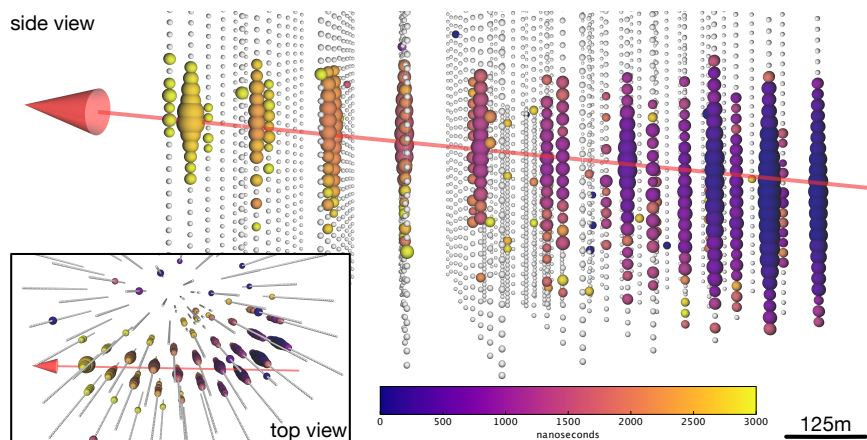


IceCube alerted the astronomy community of the observation of a single high energy neutrino on September 22 2017.

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](https://doi.org/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.





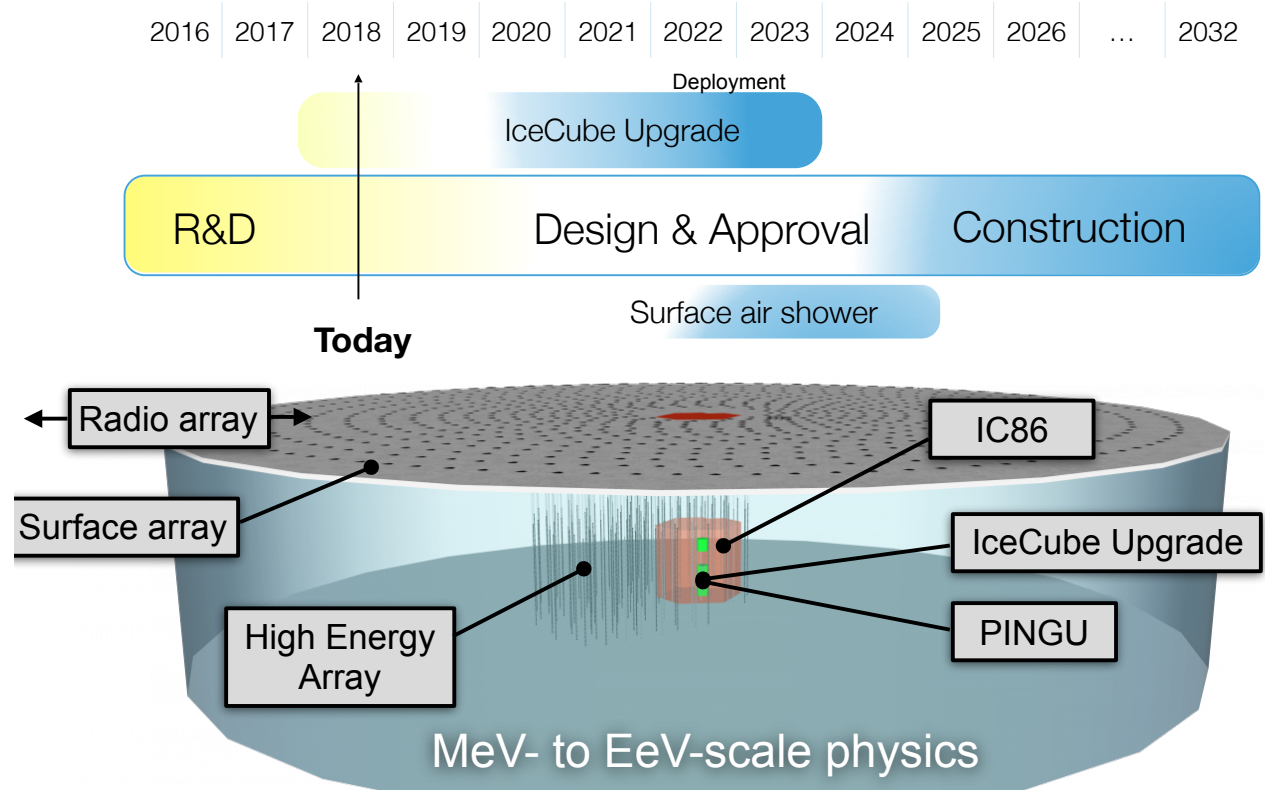
# IceCube's Future Plans

Near term:

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

## The IceCube-Gen2 Facility

Preliminary timeline



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

- Integrate **all GPUs available for sale worldwide** into a single HTCondor pool.
  - use 28 regions across 3 cloud providers, including AWS 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.

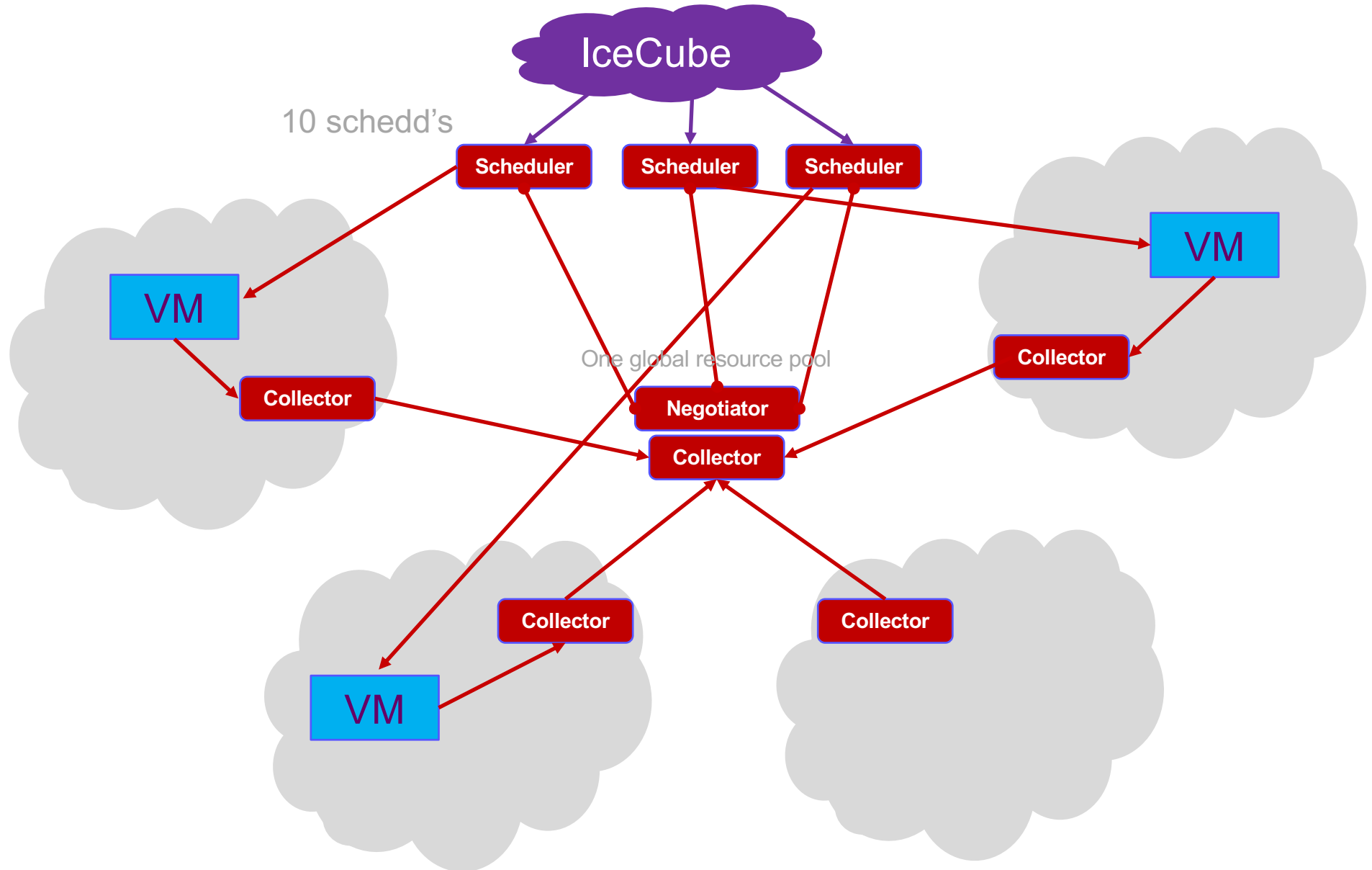
**Run a GPU burst relevant in scale for future Exascale HPC systems.**

# 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

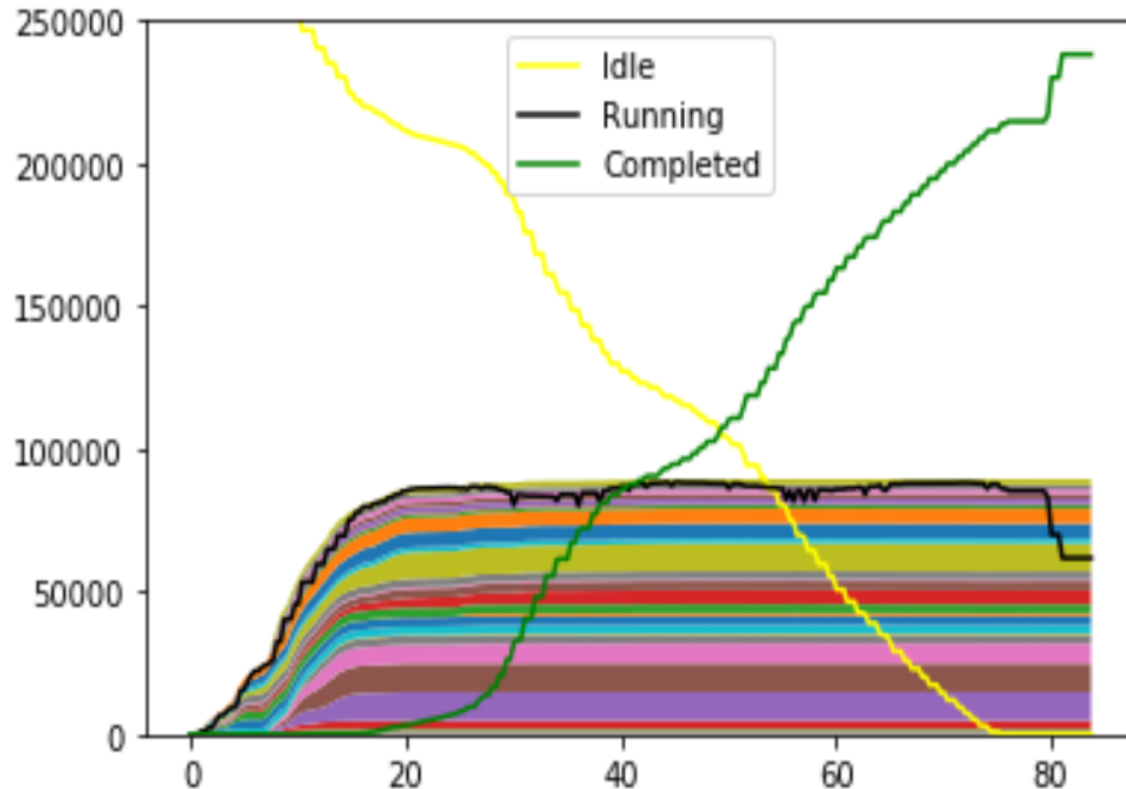


- Input data pre-staged into native Cloud storage
  - Each file in one-to-few Cloud regions
    - some replication to deal with limited predictability of resources per region
  - Local to Compute for large regions for maximum throughput
  - Reading from “close” region for smaller ones to minimize ops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



Open Science Grid

# 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.**

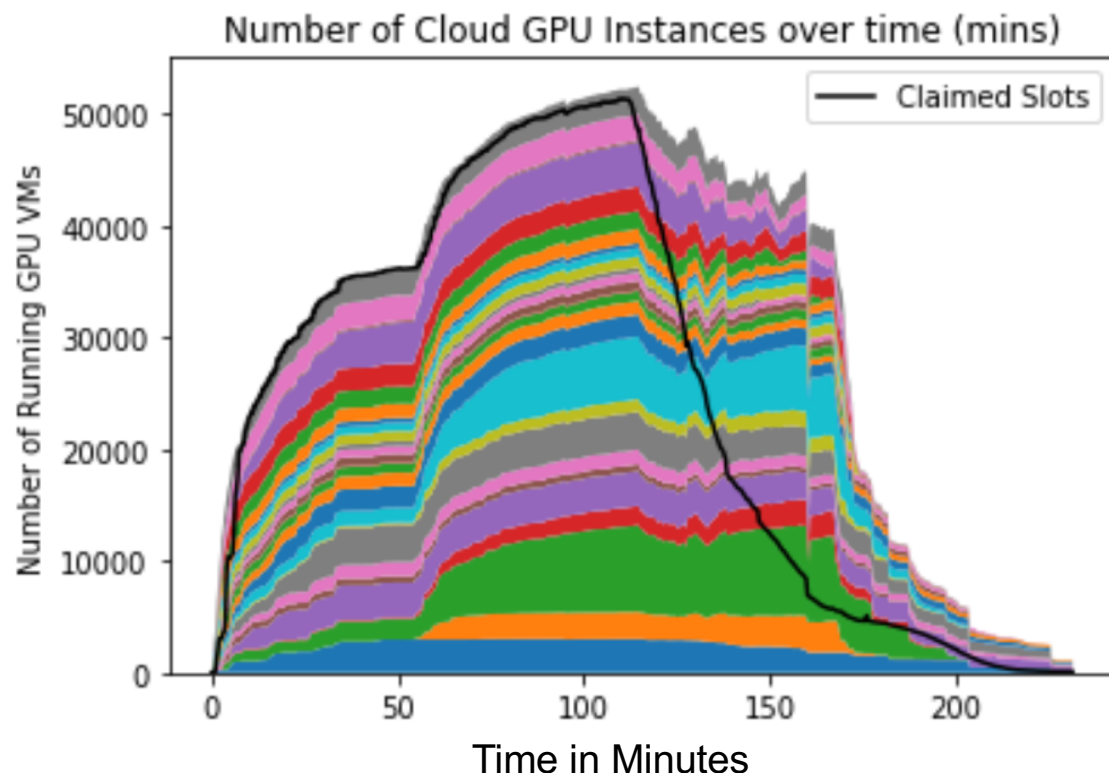
**Want scale of GPU burst to be limited  
only by # of GPUs available for sale.**





Open Science Grid

# 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

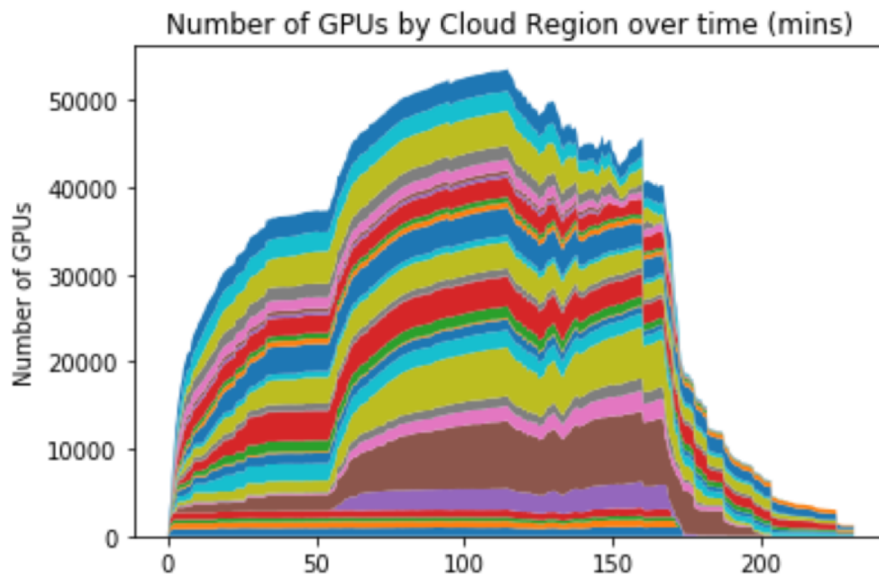
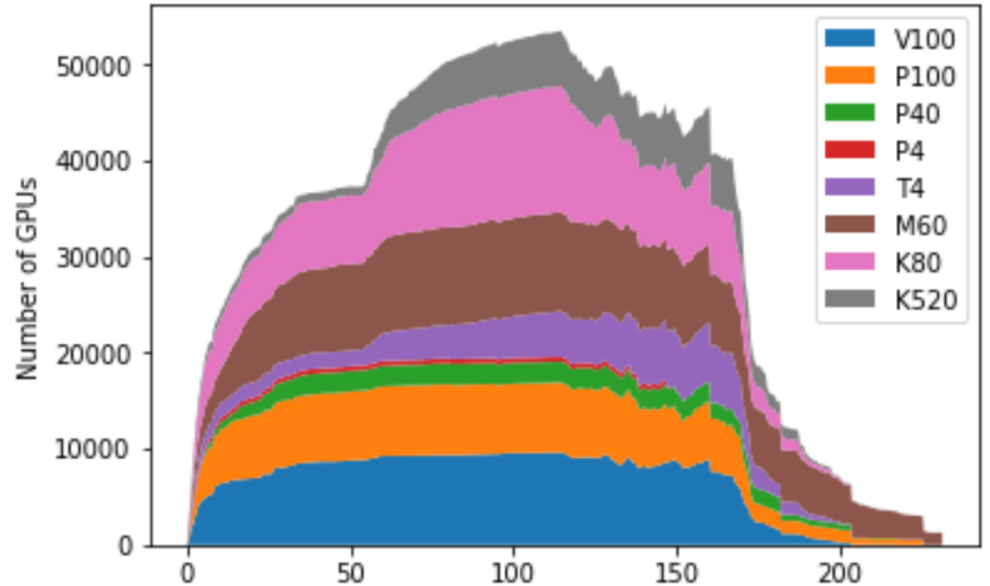
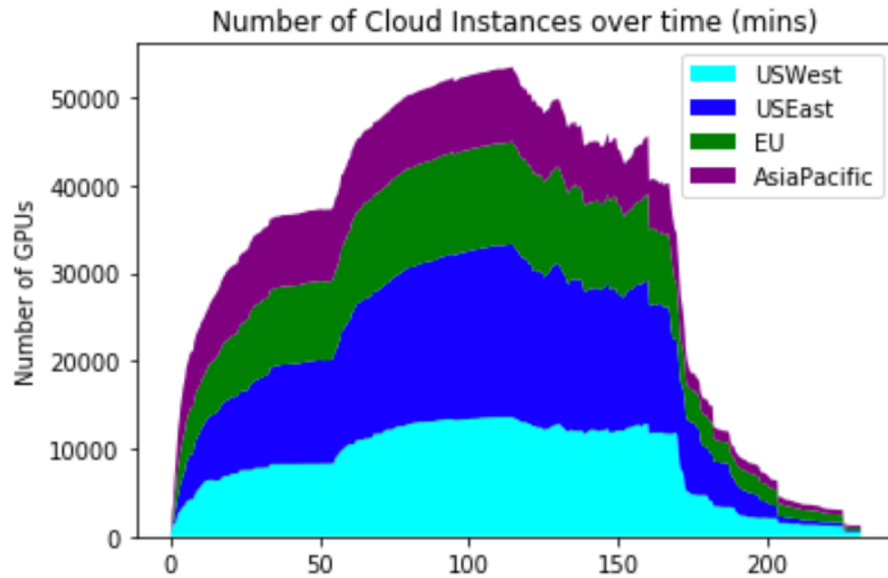
## Summary of stats at peak

	V100	P100	P40	P4	T4	M60	K80	K520
<b>Num GPUS</b>	9.2k	7.2k	2.1k	0.5k	4.6k	10.1k	12.5k	5.4k
<b>PFLOP32s</b>	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!**

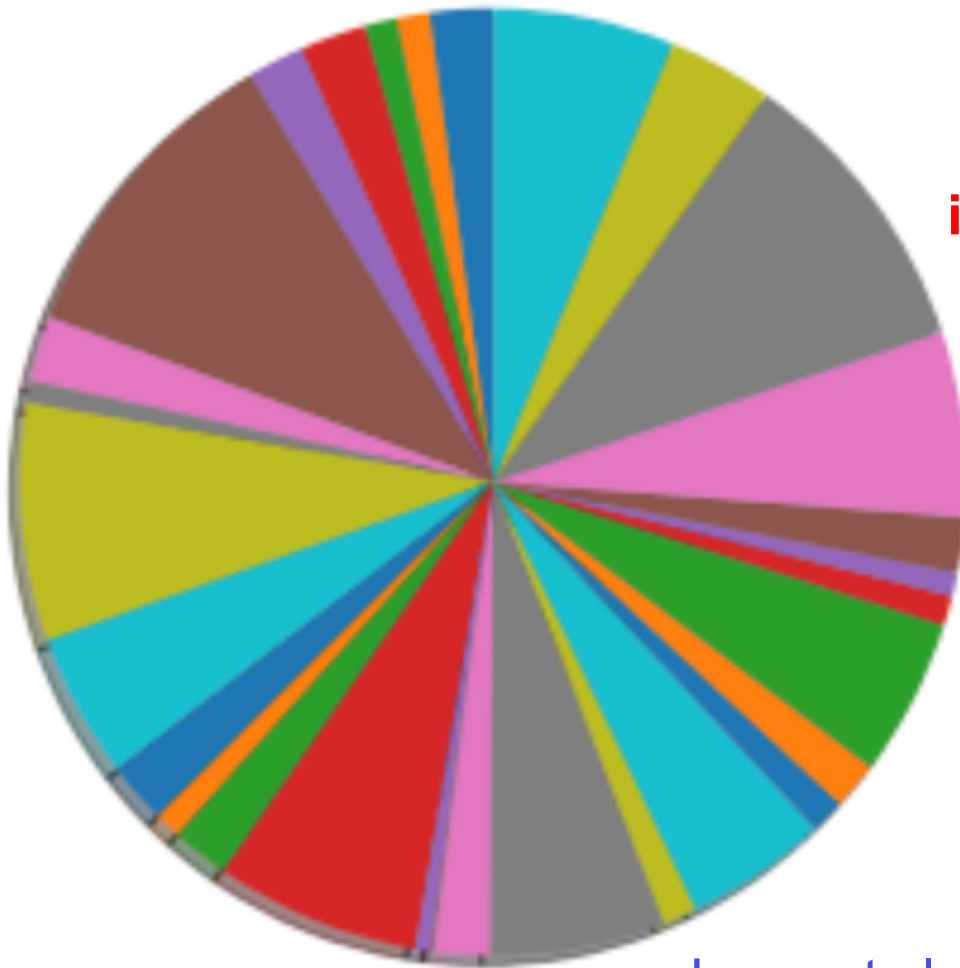


Open Science Grid

# Science Produced



Events processed per Cloud Region



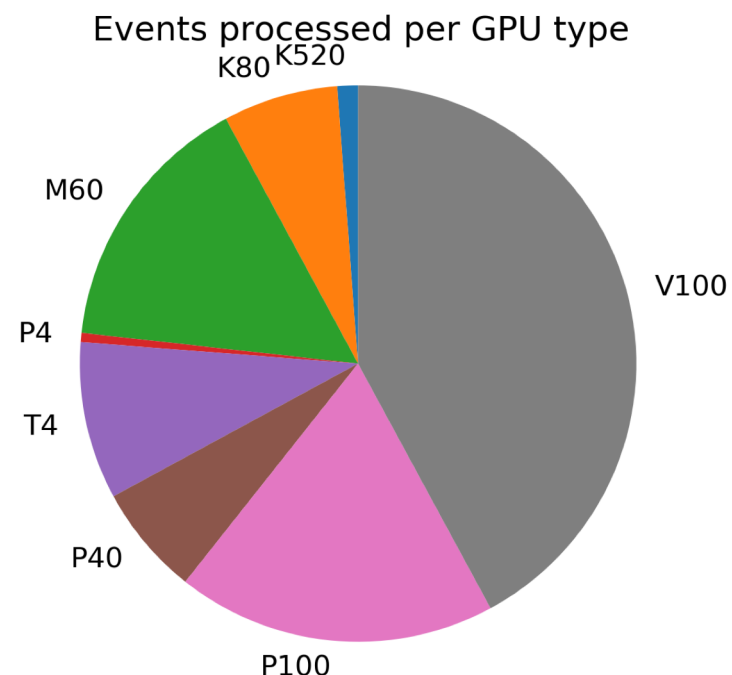
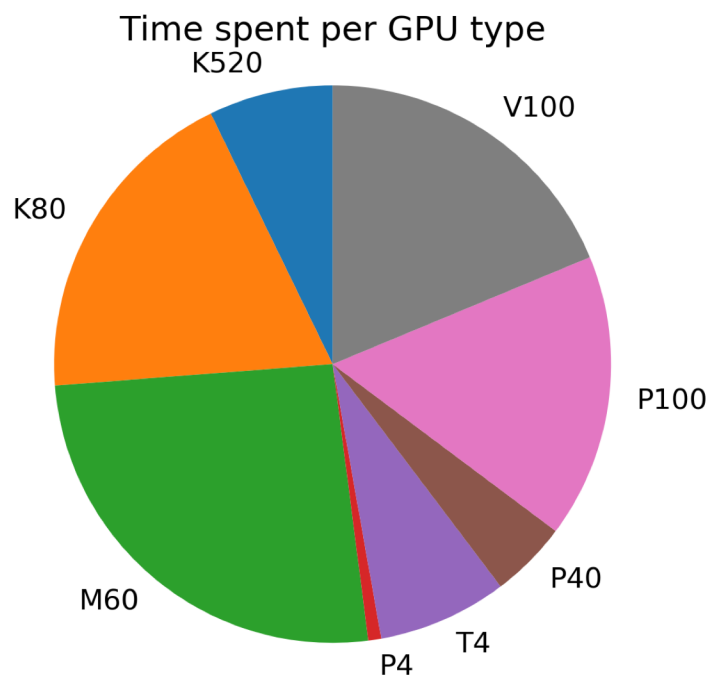
**Distributed High-Throughput Computing (dHTC) paradigm implemented via HTCondor provides global resource aggregation.**

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

Largest cloud region provided 10.8% of the total

# Performance and Cost

# Performance vs GPU type



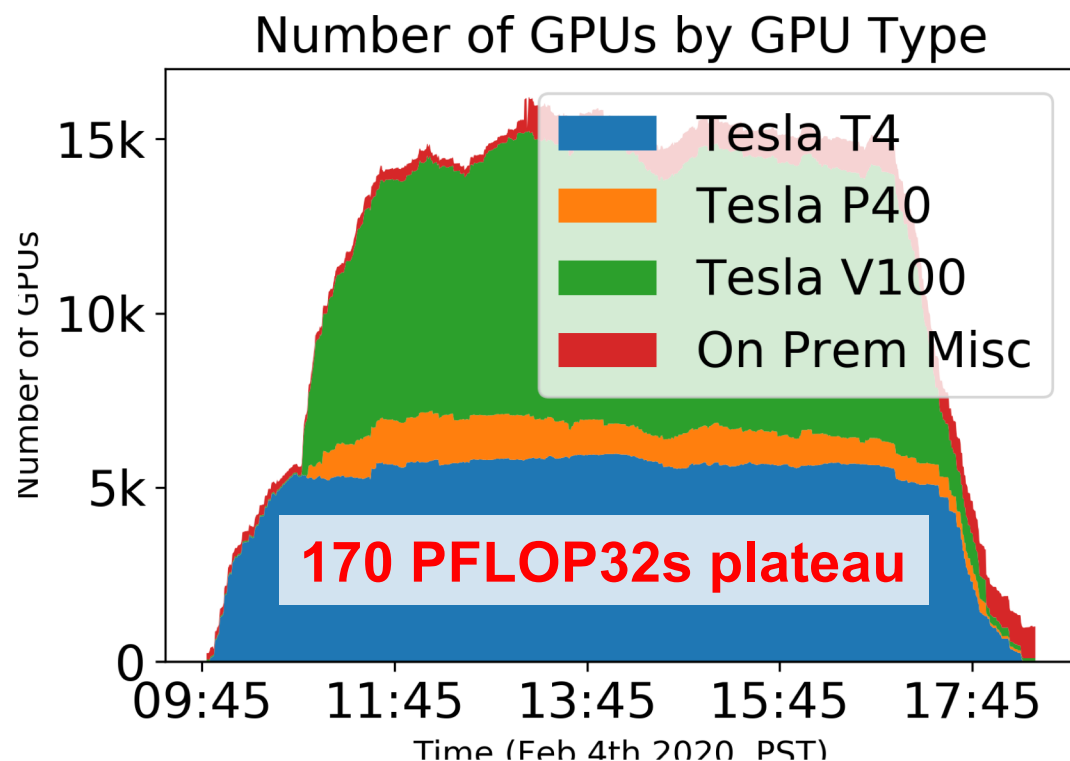
**42% of the science was done on V100 in 19% of the wall time.**

# **Second Cloud Burst focused on maximizing science/\$\$\$**

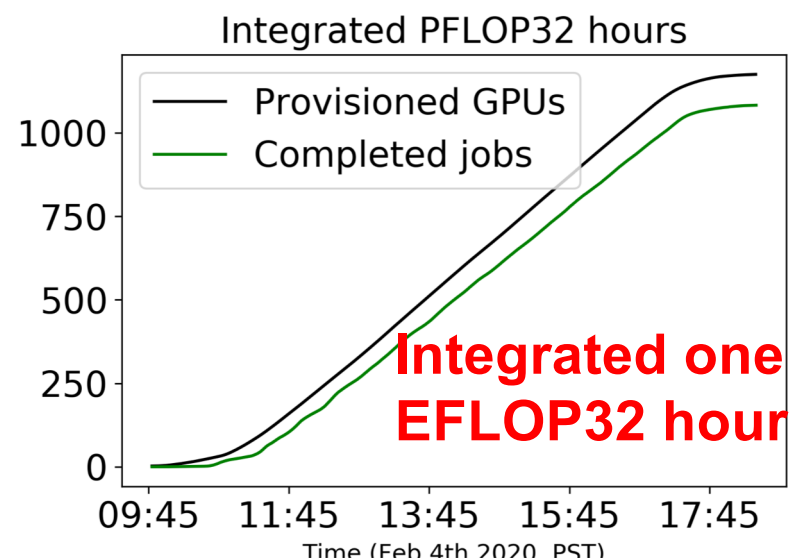
2<sup>nd</sup> burst was an 8h work day in pacific time zone on a  
random Tuesday in February

# A Day of Cloud Use

We did a 2<sup>nd</sup> run on February 4<sup>th</sup> 2020 to focus on a cost-effective 8h work day



Total bill: **~\$60k**,  
including networking and storage



We picked a “random” Tuesday during peak working hours (pacific).



# Cost to support cloud as a “24x7” capability

- February 2020: roughly \$60k per ExaFLOP32 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.

**To provide an aggregate ExaFlop32  
hour per day dHTC production capability  
in the commercial cloud for the sum of  
many sciences today would require:**

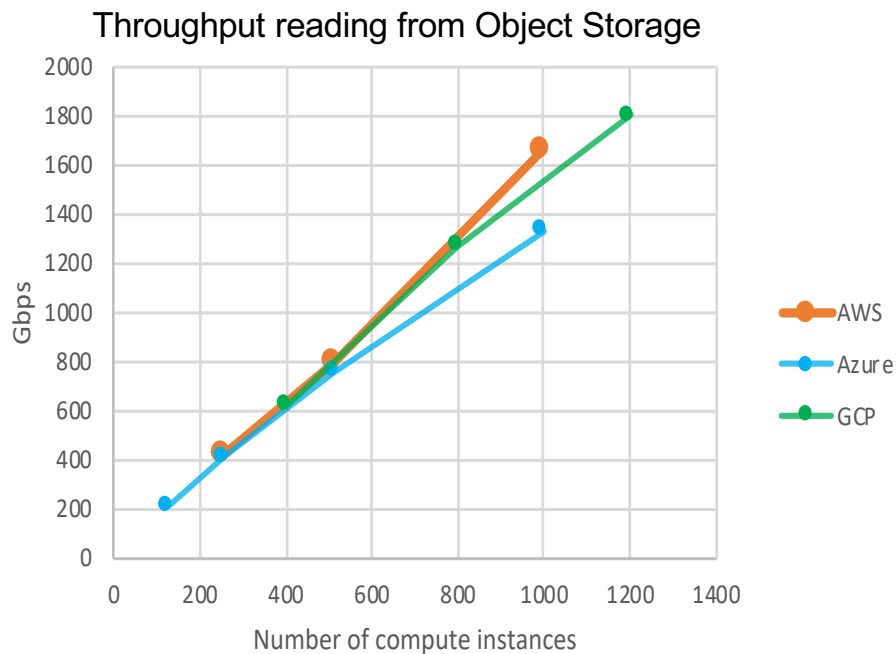
**1.5FTE of human effort  
\$60k of cloud costs per day**

This does not include the human effort to train the community,  
define the workflows, run the workflows, ... i.e. it does not include  
what the scientists themselves still have to do.

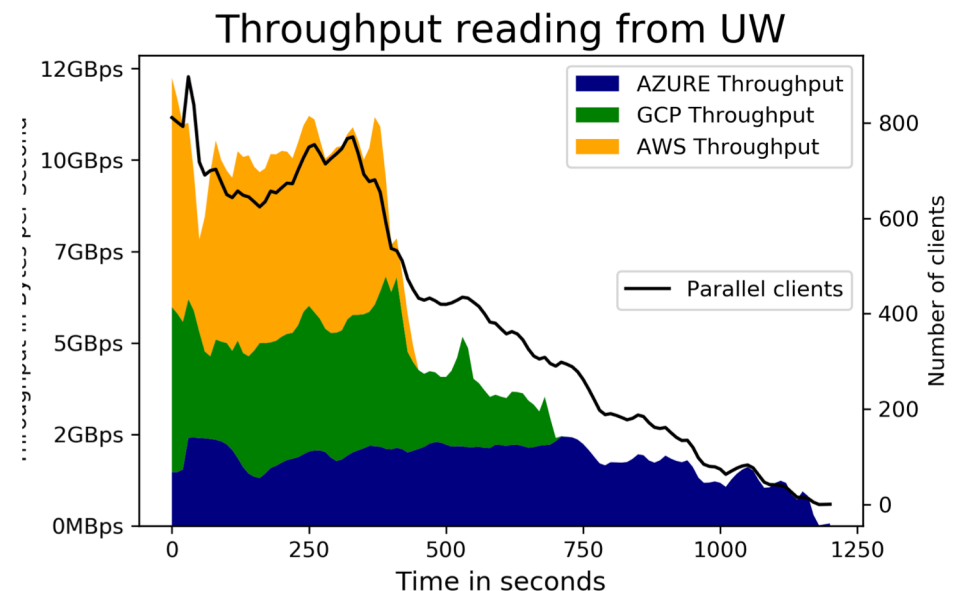


# Aside on Networking and Data

- Public cloud providers have excellent networking in place.
  - Observed close to 2Tbit/sec within a region
  - Measured up to 1Tbit/sec between regions
  - Transferred in and out of cloud and on-prem with ease.
    - But beware of cost of \$20-\$120 per TB egress
      - Waived if <15% of bill => 2GB/hour of V100 on spot is free



## Easily filling 100Gbps pipe to campus



- 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.



# 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](mailto:help@opensciencegrid.org)  
Or me personally at: [fkw@ucsd.edu](mailto:fkw@ucsd.edu)

# Acknowledgements

- This work was partially supported by the NSF grants OAC-1941481, MPS-1148698, OAC-1841530, OAC-1904444, and OAC-1826967

