

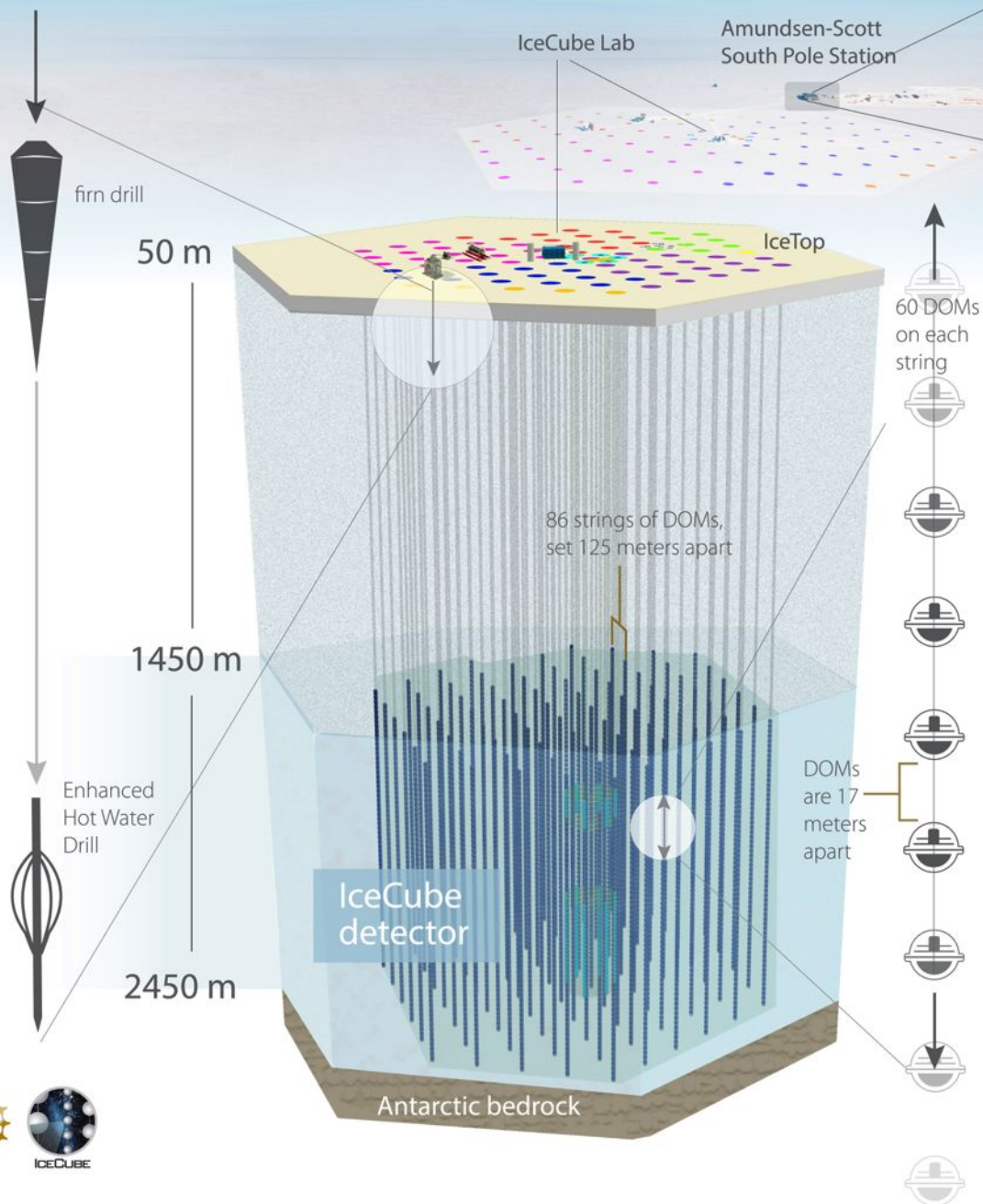


# IceCube Computing






OSG All Hands Meeting  
Mar 14, 2015

Gonzalo Merino and David Schultz  
UW-Madison



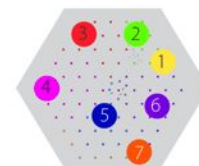


## Detector Design

-  1 gigaton of instrumented ice
-  5,160 light sensors, or digital optical modules (DOMs), digitize and time stamp signals
-  1 square kilometer surface array, IceTop, with 324 DOMs
-  2 nanosecond time resolution
-  IceCube Lab (ICL) houses data processing and storage and sends 100 GB of data north by satellite daily

## Detector Construction

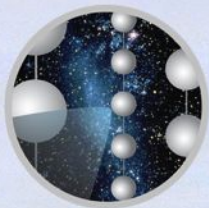
7 years of construction, 2004-2011



-  28,000 person-days to complete construction, or 77 years of continuous work
-  4.7 million pounds of cargo shipped, 1.2 million of which was the drill
-  48 hours to drill and 11 hours to deploy sensors per hole
-  4.7 megawatts of drill thermal power with 200 gallons of water per minute delivered at 88 °C and 1,000 psi







# The IceCube Collaboration



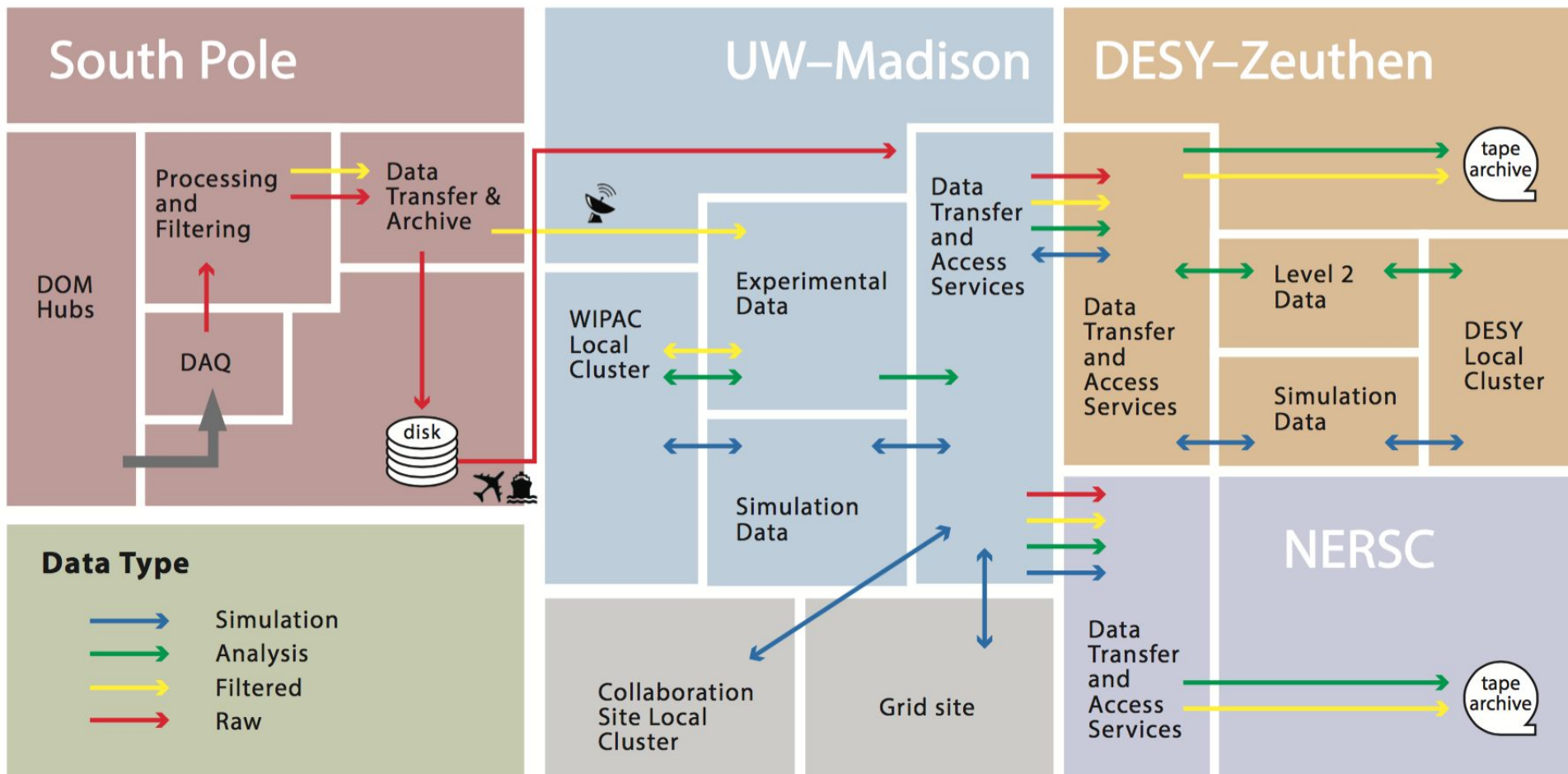
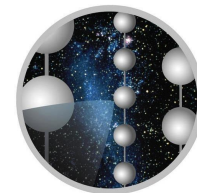
## Funding Agencies

Fonds de la Recherche Scientifique (FRS-FNRS)  
Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO-Vlaanderen)  
Federal Ministry of Education & Research (BMBF)  
German Research Foundation (DFG)

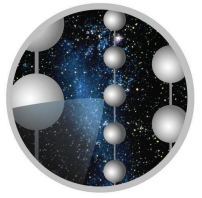
Deutsches Elektronen-Synchrotron (DESY)  
Japan Society for the Promotion of Science (JSPS)  
Knut and Alice Wallenberg Foundation  
Swedish Polar Research Secretariat  
The Swedish Research Council (VR)

University of Wisconsin Alumni Research Foundation (WARF)  
US National Science Foundation (NSF)

# IceCube Data Flow



# Data retention/archival policies



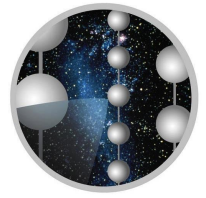
Data type	Subtype	Growth (TB/yr)	DESY-ZN tape	NERSC tape	Years on disk at WIPAC
Experimental	Raw	286		yes	2
	SNraw	31		yes	2
	Ancillary	5		yes	2
	SuperDST	64	yes	yes	2
	Filtered	36	yes	yes	2
	Level2	94	yes	yes	3
	Level3	90		yes	10
Simulation	Level2	393			3
	Level3	103		yes	10
	Photon tables	8			5

~700 TB/year to NERSC archive

~200 TB/year to DESY archive



# Long Term Archive



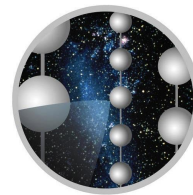
Large fraction of the data eventually becomes **archival data**. Needs to be preserved for the lifetime of the detector, and beyond.

- Managing a multi-PB near-line tape archive not an easy task - Large infrastructure and manpower costs.
- Decided to **outsource** the service to larger centers that can benefit from economies of scale.
- May 2015: Collaboration group at LBNL offered to provide tape storage service at NERSC (~6 PB in 5 years).

NERSC requires big files (100GB→1TB) ⇒ Need to bundle files. We are currently developing sw to handle this. Plan is:

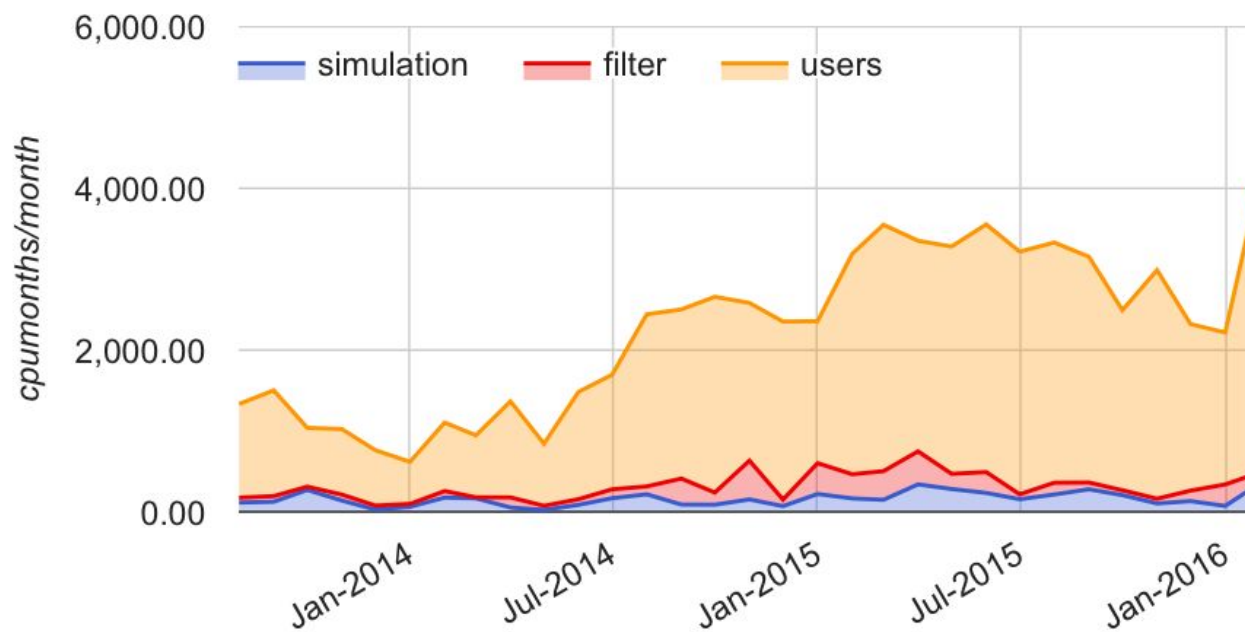
- Decouple archive from “live” data (no HSM).
- Bundling: re-use the in-house developed sw for transferring data from the South Pole.

# IceCube Computing Resources

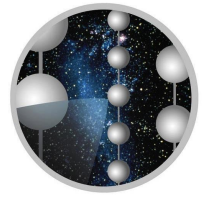


## UW-Madison data center (Tier-0)

- ~ 5000 CPU (HT)cores (recently upgraded to ~7000)
  - 2GB RAM per (HT)core
- ~ 350 GPUs
- ~ 4PB disk



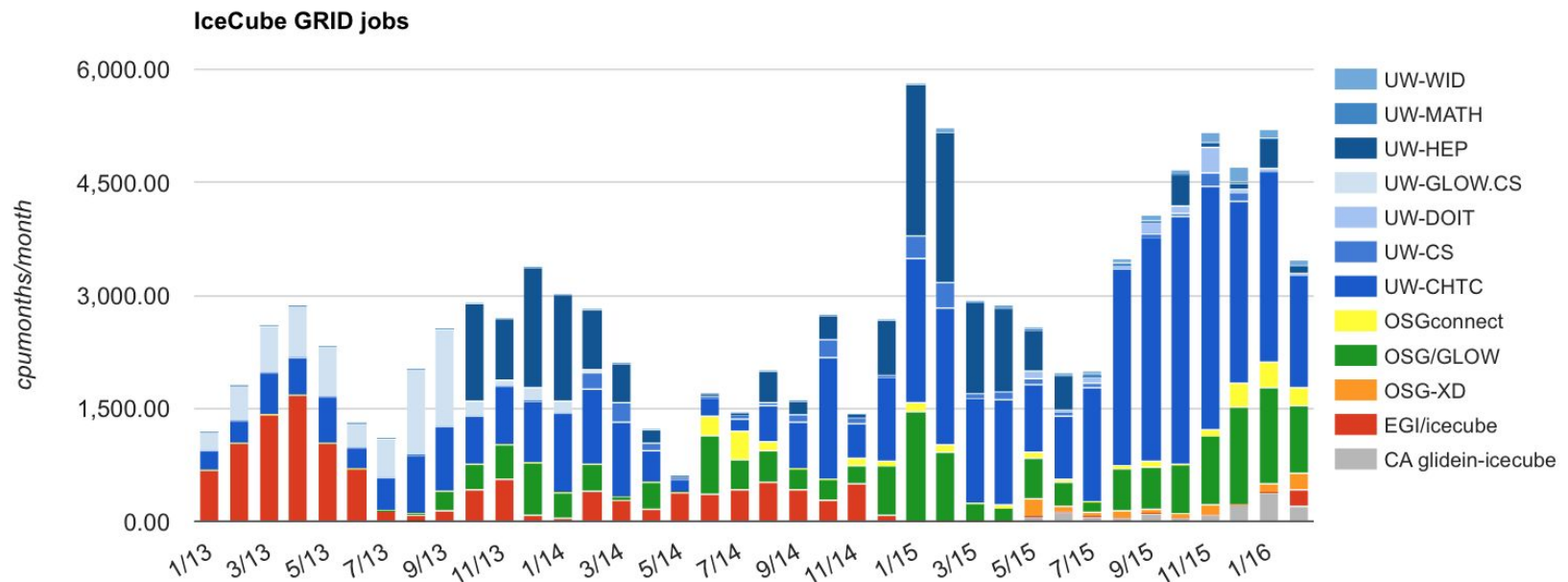
# Opportunistic Resources



IceCube makes extensive use of opportunistic shared resources.

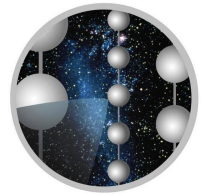
Currently about ~50% of our used CPU is opportunistic

- Largest chunk from UW clusters (HTCondor flocking)
- Substantial amount from OSG (GLOW & OSG VOs)





# Grid Tools

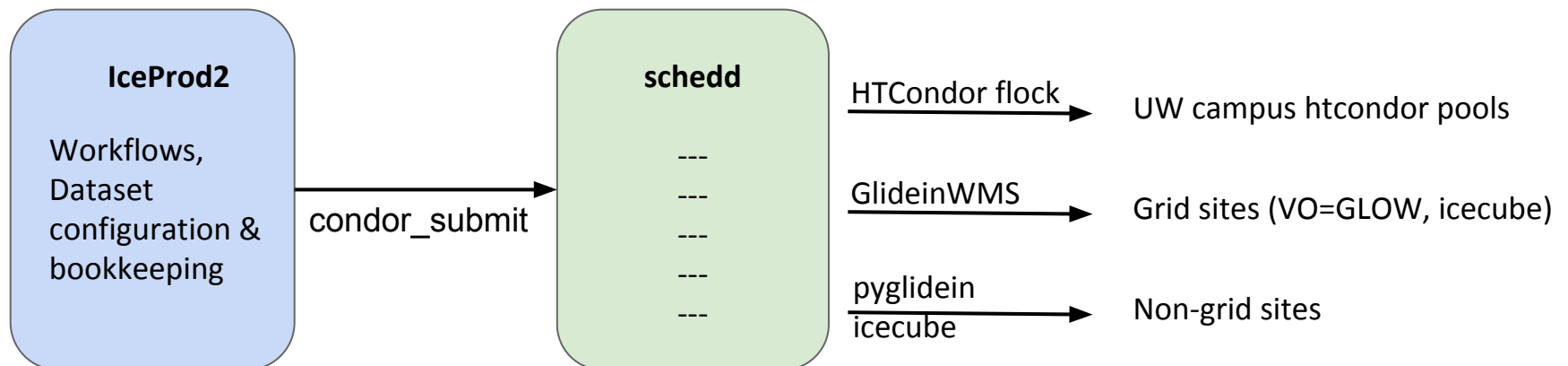


Evolving from a setup where the distributed infrastructure was managed end-to-end by our in-house Grid framework:

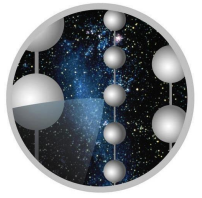
- IceProd (started in 2006, <http://arxiv.org/abs/1311.5904>)

Towards a model where the new framework (IceProd2) focuses more in the IceCube specifics (dataset configuration & bookkeeping, ...) and “delegates” the resources federation to 3rd party tools like HTCondor.

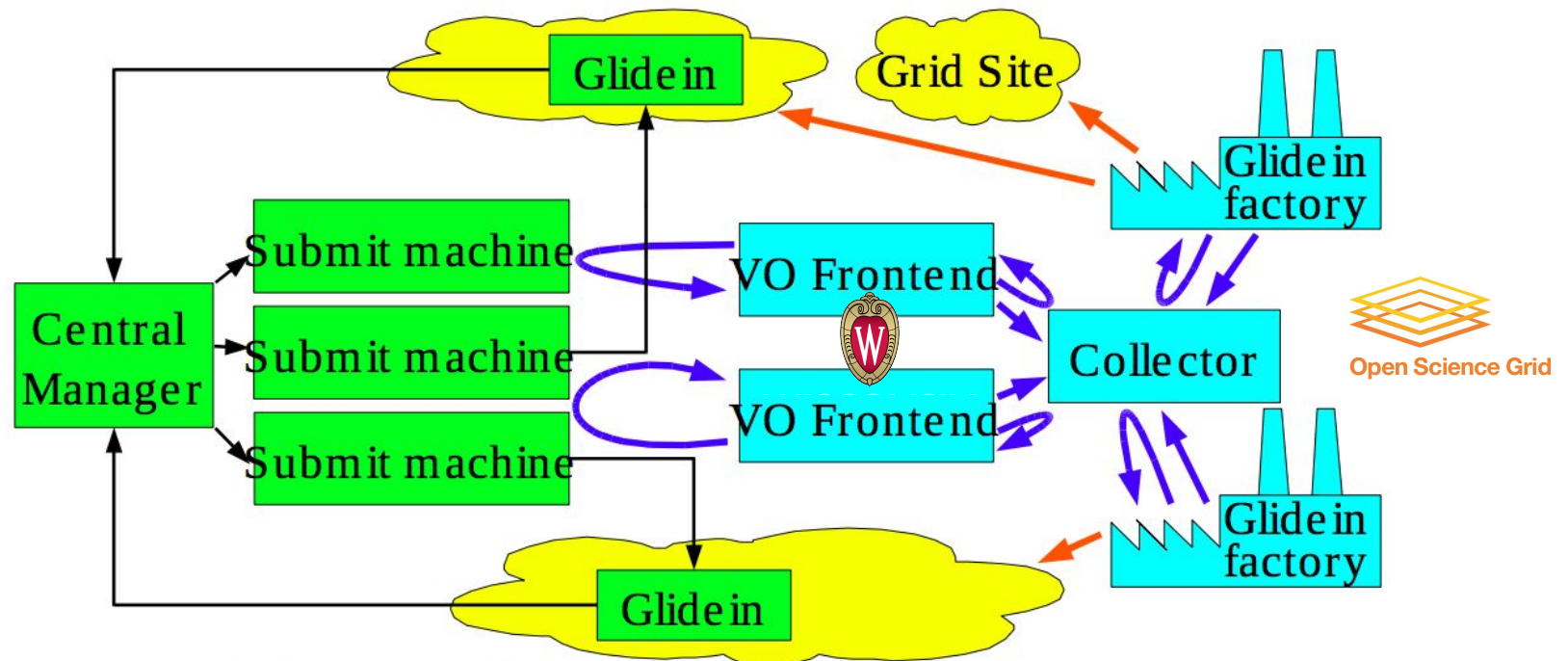
Current system:



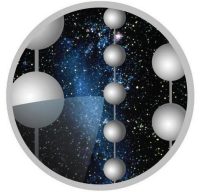
# GlideinWMS



Used since 2013 “as a service” via the GLOW VO (thanks!)



# GlideinWMS: GLOW



Some of the IceCube sites out there are Grid sites (shared w LHC). We try to use them with standard tools.

- Did this with DESY-ZN (Berlin) and SCINET (Toronto) in 2014/2015

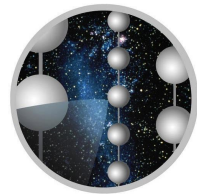


Downside:

- Requires VOFrontend configuration - sync with list of IceCube sites
- Not all sites will be ok with accepting GLOW VO for IceCube



# GlideinWMS: IceCube



Next → try and use VO=icecube for our pilot based Grid infrastructure

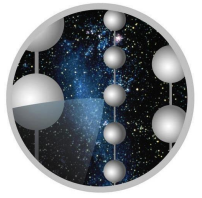
OSG/GlideinWMS proposed configuration:

- Configure UW/CHTC VOFrontend to manage 2 sets of credentials: GLOW, icecube.
- OK! We still get this “as a service” from UW/OSG. No need to run our own Frontend/Factory.

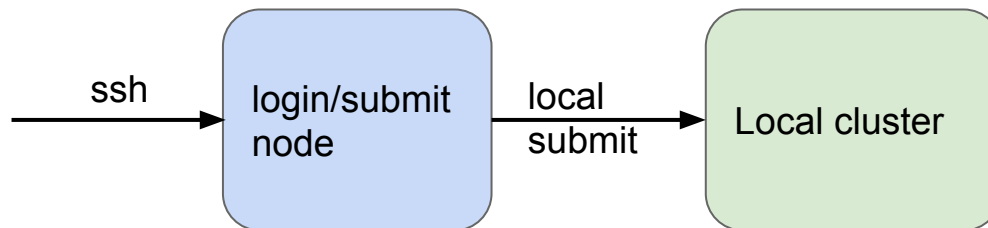
Got a lot of support from UW/CHTC & OSG (thanks @moate, @efajardo, @mkandes, @bbockelm!)

- Feb-8 : initial phone call to set requirements & goals.
- Feb-9: everyone in a slack team, active discussion.
- Feb-10: 1st icecube glideins running at DESY and SDSC.
- ...
- Today: icecube glideins from OSG factory running at 5 sites (3-5 more in the pipeline with open GGUS tickets, more to come ...)

# pyglidein icecube



Several IceCube sites are “non-Grid”



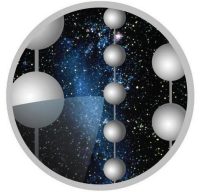
2015: started experimenting with BOSCO for this

- Our experience was that often lots of jobs ended up on “hold” because the ssh tunnel becoming flaky.

The BOSCO idea of a glidein factory “via ssh” is nice.

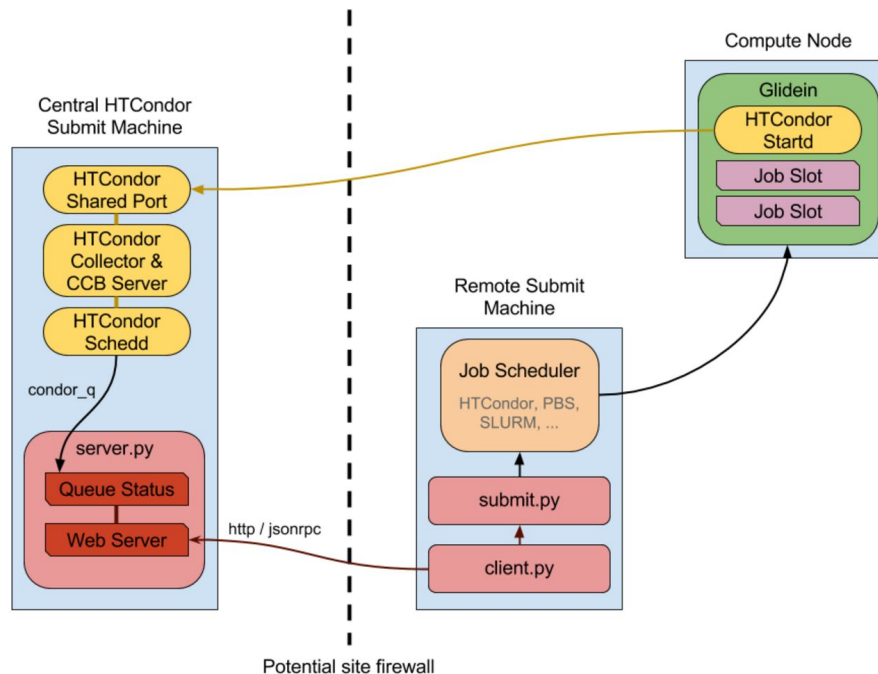
- Why do not try and move the factory to the other side of the ssh connection?

# pyglidein icecube



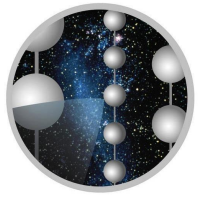
Decided to try and write our “minimalistic” factory - it might be worth as long as it is simple (currently ~1000 lines of python code)

- Developer: David Schultz
- Code: <https://github.com/dsschult/pyglidein>



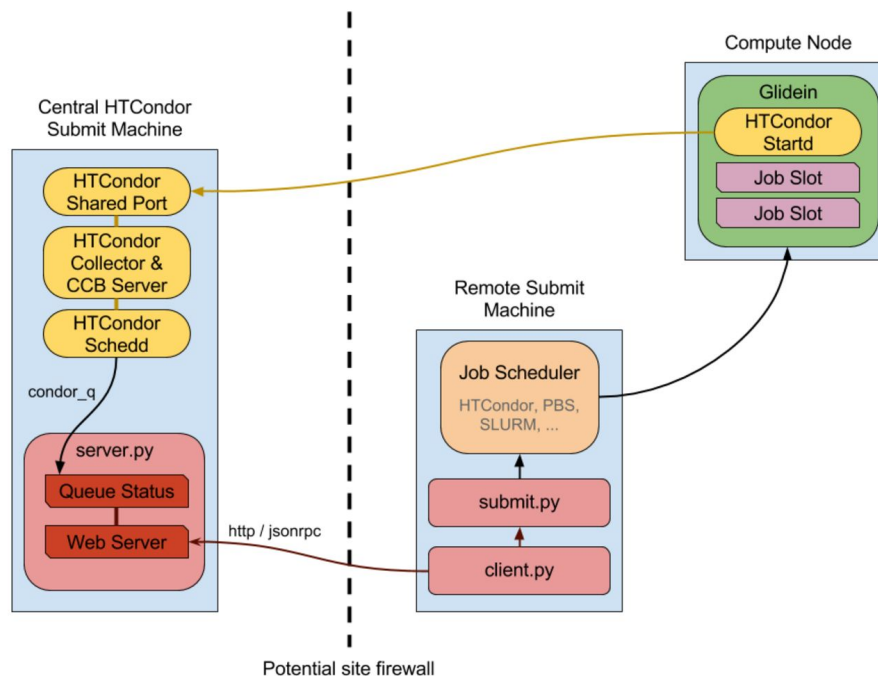


# pyglidein icecube



Decided to try and write our “minimalistic” factory - it might be worth as long as it is simple (currently ~1000 lines of python code)

- Developer: David Schultz
- Code: <https://github.com/dsschult/pyglidein>



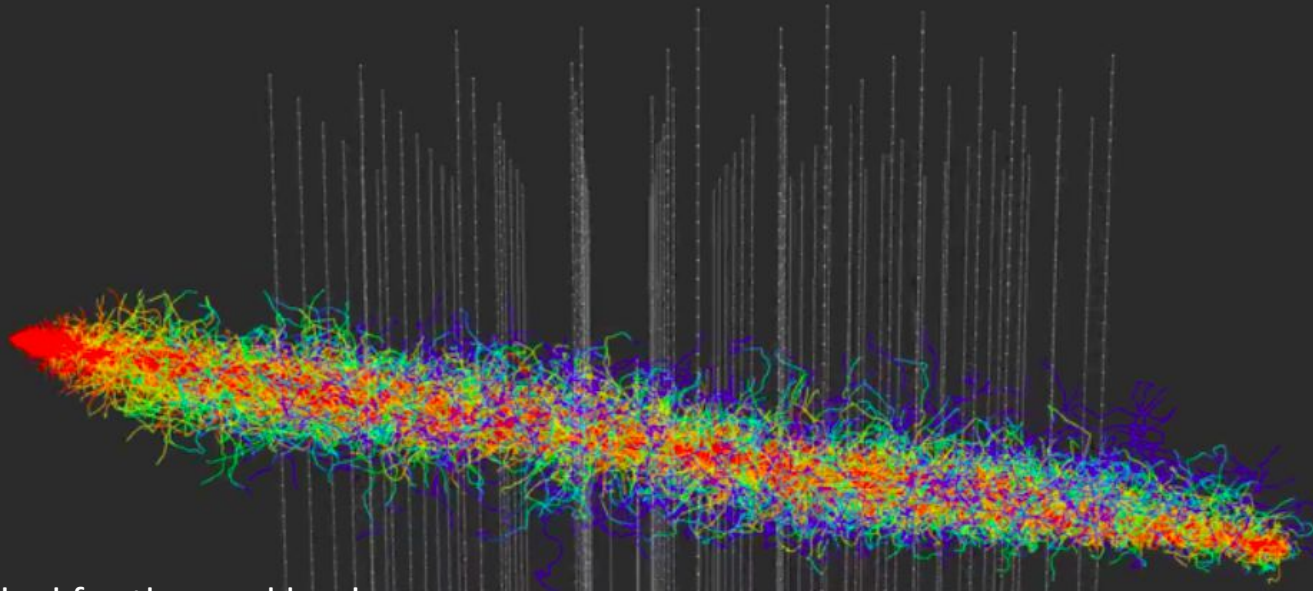
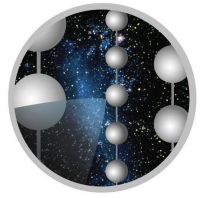
Running in production at 5 sites since mid-2015.

**Cons:** yet another factory, yet another glidein, ...

**Pros:** Useful to be able to customize our glidein quick, e.g.

- GPU discovery/assignment
- ClassAdd to advertise CVMFS/icecube
- Parrot

# GPUs: direct photon propagation



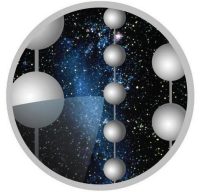
GPUs are ideal for the workload

- Many independent photons + scattering model is simple (scatter, absorb, change ice layer or hit a DOM)
- Simulate each photon with an independent thread
- Only interrupt parallelism when a photon hits a DOM and signal needs to be stored (very rare!)

GPUs are  $O(\sim 100)$  faster than CPUs for this workload

time delay  
vs. direct light  
"on time" → delayed

# IceCube GPU Cluster



Good news: code is ok with consumer-grade GPUs

Not so good: GPUs still a rare beast, not easy to find accessible GPU clusters out there.

⇒ needed to build an in-house sizeable cluster.



Current IceCube GPU cluster at UW-Madison:

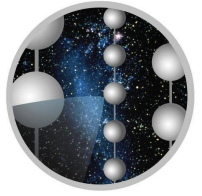
- 48 Nvidia Tesla M2070
- 32 AMD 7970
- 32 Nvidia GeForce GTX690
- 256 Nvidia GeForce GTX980

(~1.5 PFLOPs single precision  
... small gaming supercomputer)





# GPU Resources - XSEDE



We want to explore the possibility of expanding our GPU capacity by requesting time allocations in GPU-enabled supercomputers.

2015: requested a “startup” allocation to test running IceCube GPU jobs: 50,000 SU at TACC Stampede awarded

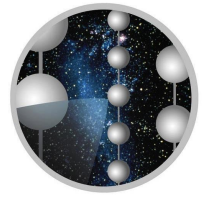
- CVMFS was there. David Lesny (ATLAS MWT2) got /cvmfs/icecube.opensciencegrid.org/ replicated in 1 day (thanks!)
- Successfully ran IceCube GPU jobs (glidein was not possible, due to firewall)

2016: XSEDE “research” allocation awarded in 2 GPU-enabled systems:

- **Comet** at SDSC: 5,543,895 SUs (36 nodes with 2x NVIDIA K80 GPUs each)
- **Bridges** at PSC: 512,665 SUs (16 nodes with 2x NVIDIA K80 GPUs each)
  - Fall 2016: +32 nodes with 2x NVIDIA Pascal GPUs each
- Requested ECSS support → working with Mats Rynge to integrate these resources in our workload (mostly: CVMFS + glidein-friendly network)
  - Good news: we are already running GPU glideins in Comet/SDSC!

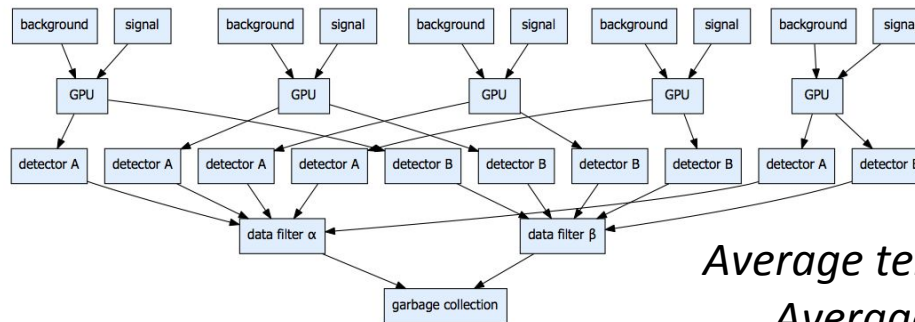
-

# Data Management



The IceProd framework orchestrates the simulation production workflows. Tasks write/read intermediate output/input from the UW-Madison GridFTP.

- Most IceCube sites that provide a CE, do not provide an SE.



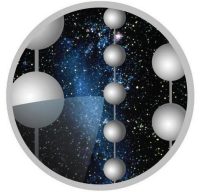
*Average temporary output ~10-200 MB*  
*Average task duration ~0.5-2 hours.*

We do not see big problems with the “central SE” model so far. However, we need to tackle it if we want to scale in the next years.

- Few IceCube sites might provide SE's → ~5 “regional” gridftp servers?
- Need to add some “locality awareness” to the scheduling

# Summary

---



IceCube benefits a lot from OSG. Big users of opportunistic CPU (thanks!)

- Plans for the UW-Madison site to become a fully functional OSG site (including sharing the CPU/GPU cluster)

GPU continues to be a critical resource in the simulation chain. Main facility is the UW-Madison cluster.

- Work with IceCube sites to integrate their GPU clusters seamlessly with simulation production framework using pyglidein icecube.
- Actively explore new opportunities for tapping on other GPU resources (XSEDE, opportunistic GPU at OSG sites ... )

Long Term Archive service using remote DESY and NERSC sites to be rolled out this year. Plan is to write software to handle data transfers to archive.

- Remote archive includes one ~400 TB bulk transfer UW→NERSC once a year. Plan is to leverage gridftp/globus.org services as much as possible.