Computing Report

IceCube Maintenance and Operations Review
Madison, WI
March 11th, 2019
Benedikt Riedel

Outline

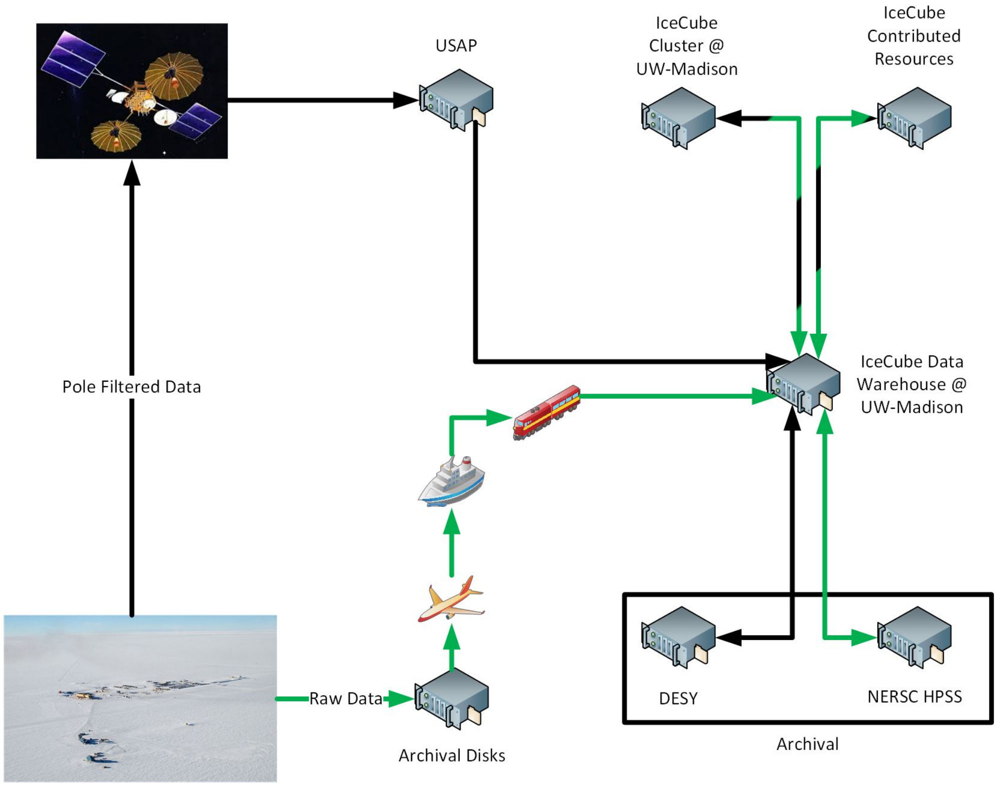
- Overview - Deliverables, Data Flow/Processing, Simulation, Challenges
- Computing Infrastructure
  - UW, Collaboration, National, Future Plans
- Production and Physics Software
  - Simulation Software
  - Dataset and Workflow Management - IceProd
  - Long Term Archive
- Data Processing - Continuous L2, Pass 2, and Pass 3
- Simulation Production
- IceCube Upgrade
- Summary

Deliverables

- Data Warehouse and Storage Infrastructure for experimental, simulation, and analysis data, including data retrieval from Pole
- High Performance Computing cluster for timely offline data analysis and simulation production, including GPU computing
- Data Center Infrastructure, i.e. infrastructure to maintain data warehouse and cluster
- Provide infrastructure and support to utilize collaboration computing resources
- Offline/analysis software support and maintenance, including distributing workloads across a global computing grid

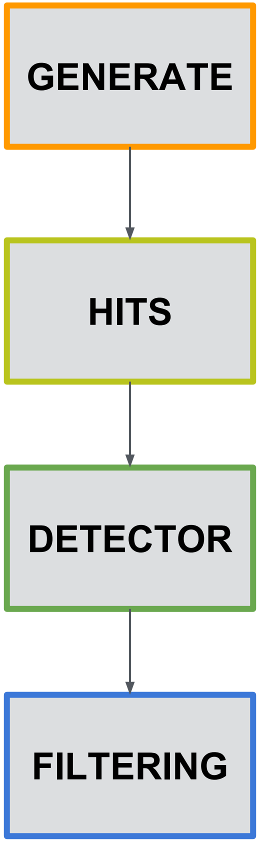
Data Flow and Processing

- Pole Filtered Data arrives via satellite - Arrives at UW-Madison and is reduced further to higher levels
- Raw data is written to archival disk at pole, retrieved once a year
- Raw data is archived at National Energy Research Scientific Computing Center (NERSC)
- Filtered data is archived at Deutsches Elektronen-Synchrotron (DESY)



Simulation Chain

Generate background (Cosmic Ray showers) and signal (neutrino interactions)
Propagate to detector



Photon propagation in ice - Requires GPUs

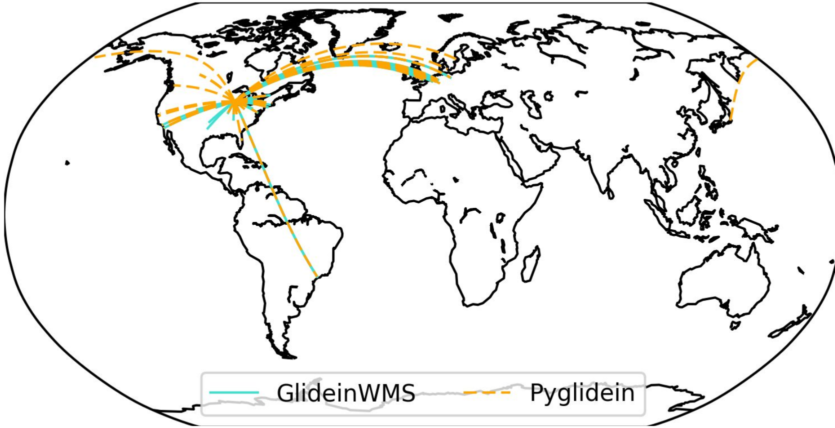
DOM hardware simulation
DAQ trigger emulation

Pole (Level 1) and offline (Level 2) reconstruction and filtering

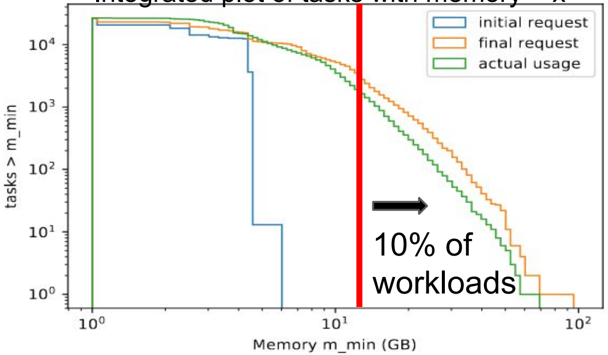
Challenges

- Global heterogeneous resources pool
- Mostly shared and opportunistic resources
- Atypical resources requirements and software stack
  - Accelerators (GPUs)
  - Broad physics reach - Lots of physics to simula
  - Data flow includes leg across satellite
  - “Analysis” software is produced in-house
    - “Standard” packages, e.g. GEANT4, don’t support everything or don’t exist
    - Niche dependencies, e.g. CORSIKA (air showers)
  - Detector up time at 99+% level
- Significant changes of requirements over the course of experiment - Accelerators, Multimessenger Astrophysics, alerting, etc.

Glidein Locations



Integrated plot of tasks with memory > x



* All IceProd 2 tasks as of April 2018

Computing Infrastructure

Computing Infrastructure — UW

- WIPAC and UW resources are the backbone of computing infrastructure for IceCube
- WIPAC hosts the central data warehouse for IceCube detector and simulation data, and central data analysis facility
  - 8.5 PB of storage available
  - ~6000 CPU cores (90+% usage), ~300 GPUs (90+% usage) dedicated to IceCube
  - Interactive analysis and support infrastructure
- Resources are split between 222 West Washington, UW Physics Department, and OneNeck facility in Madison
  - 222 West Washington - Core services, older storage, etc.
  - UW Physics Department - Compute cluster and storage
  - OneNeck - New storage infrastructure
  - OneNeck will replace 222 in the coming months - Aim is to have everything complete by H1 2019

Computing Infrastructure — UW

- Network infrastructure now maintained and provided by UW
- Upgraded and reconfigured storage infrastructure
  - Bought 8.5 PB storage for experimental and simulation data
  - New infrastructure now a single vendor and located at OneNeck facility
  - Remaining storage will be reconfigured to provide
    - More storage for users
    - R&D area to study feasibility of different storage technologies: Ceph, dCache, etc.
- Improved GPU capabilities
  - Continuous increase in GPU compute capacity - Both upgrades of older cards and new purchases
  - Growing GPU/accelerator resources through applying to outside resources, e.g. NSF's XSEDE program

Computing Infrastructure ─ Collaboration

- Introduced **computing pledge system** to incentivise investment in computing - Computing resources are in-kind contributions
- Continually expanding the IceCube processing grid using in-house developed `pyglidein` - Able to include campus clusters, regional computing centres, national supercomputers
- Direct investments in IceCube computing resources by other institutions, e.g. UAlberta, MSU, UMD, DESY, Mainz
- Established **long-term archive** at NERSC for IceCube raw data
- Working with LHC Tier 2 centers at collaboration institutions for access or higher priority
  - Already have access to DESY and Belgian Tier 2 site
  - Working on higher priority with US Tier 2 sites at MSU and UT-Arlington
- SCAP met in 2016 and 2018, see Kael's talk for details

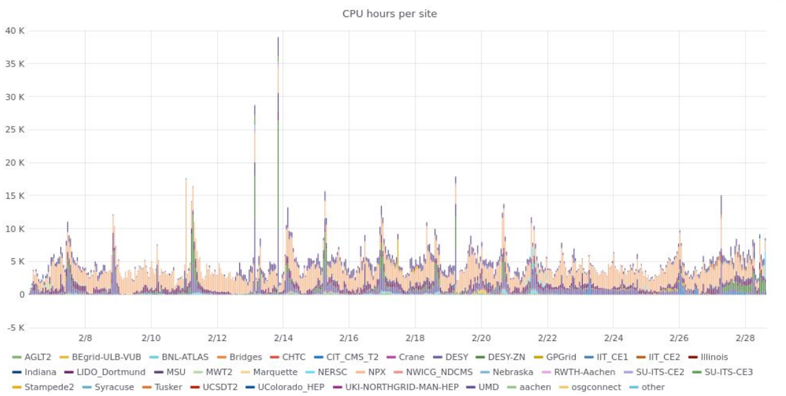
Computing Infrastructure — Collaboration

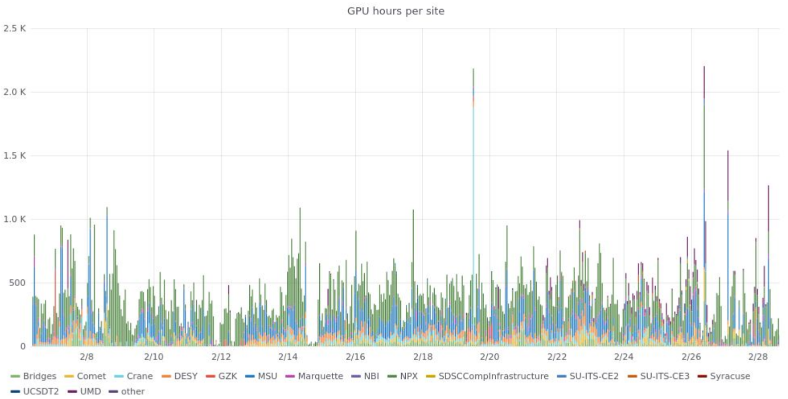
Collaboration-contributed CPU and GPU Compute

- Continuously using ~2k CPU-hours and ~150 GPU-hours
- Significant amount of “dark” computing - Users at local institutions

Storage

- Primary Data Warehouse: 8.5 PB of disk provisioned at UW-Madison
- Backups:
  - 4 PB of tape storage provisioned at NERSC for raw data backup
  - 4 PB of tape provisioned at DESY for offline processed data backup





Computing Infrastructure — National

Significant invest in GPU resources on national-scale HPC resources

- USA
  - Extensive use of NSF's **XSEDE** GPU resources - XStream, Comet, Bridges
  - **Open Science Grid** (OSG) infrastructure and resources are essential
  - Started exploiting DOE resources (Titan and NERSC) - Significant restrictions compared to most XSEDE resources
- EU
  - Significant number of possible resources targets, e.g. LHC facilities, supercomputers, etc. - Some come with significant restrictions similar to DOE
  - Non-local resources have not been exploited yet
- Japan
  - Small usage so far

Computing Infrastructure — National

NSF XSEDE - 2018 Allocation

- PSC Bridges: 287k SUs of GPUs - 53% used with 4 months remaining
- SDSC Comet: 180k SUs of GPUs - 55% used with 4 months remaining
- OSG: 4M SUs of CPU - 100% used
- SU = Service Unit

DOE - 2018 Allocation

- Titan: 1M node-hours - Used, 2019 allocation applied for
- NERSC - Cori: 1.25M NERSC-hours - Used jointly for production and user analysis of UC-Berkeley/LBL group

Computing Infrastructure — Future Plans

- Leverage existing and upcoming resources at collaboration institutions and national facilities
  - Focus on ability to use supercomputers with limited network connectivity - Similar issues faced by HL-LHC
  - XSEDE resources (Stampede2 and Frontera), DOE resources (NERSC 9/Perlmutter)
  - IRIS-HEP - Software institute funded by NSF for the HL-LHC area
  - SCiMMA - Conceptualization for computing in Multi-Messenger Astronomy
  - Morgridge Institute of Research has hired new Associate Scientist with experience in CMS, LIGO, OSG, and data management
- Additional resources through NSF programs and solicitations
  - Submitted proposal for Mid-scale Research Infrastructure-1 “Infrastructure” proposal with cryoEM group at UCSD and SDSC for a GPU cluster hosted at SDSC
  - Approved funding for NSF-sponsored Internet2’s Exploring Clouds for Acceleration of Science (E-CAS) for commercial cloud credits
- Modernization of Workflows
  - Deployment of software with containers
  - Continuous integration and testing solutions to improve production software and **reproducibility**
  - Analytics and traceability of production systems, including **improved monitoring**
- Data organization, management, and access will transition to software-driven era

Physics Software

Physics Software — Releases

Releases of production software around season changes - as needed

- Vernal Equinox - March 20th
- Summer Solstice - June 21st
- Autumnal Equinox - September 22nd
- Winter Solstice - December 21st

Quick incremental releases as needed

Code Sprints - Support release preparation

- Week before the scheduled release
- At most four per year

Physics Software —Workshops

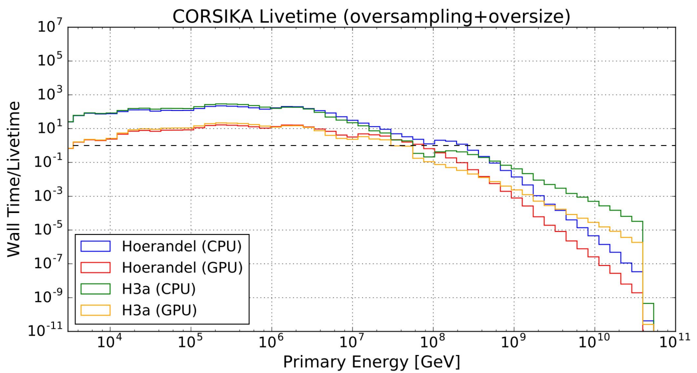
Workshops held pre-/post-collaboration meeting

- High level of productivity
- Code optimization
  - Memory, CPU profiling
  - Data structures
  - Optimization schemes
  - Simulation quality/improvements

Yearly Software Bootcamps - Introduce new students and postdocs to IceCube and IceCube software

Physics Software — CORSIKA

CORSIKA Livetime Issues



- *A priori* simulation doesn't know if a shower is "interesting" to IceCube - It can take over **500x** the compute time to get the desired livetime
- Even more problematic for generating air showers with <10^3 GeV primaries - Scientifically interesting, yet wasteful production wasteful - Products don't trigger detector, so resources are "wasted"
- Single muon simulation (MuonGun) is much faster, but introduces systematics (muon bundles)
- Analyzers would prefer CORSIKA, not possible by brute force simulation

Physics Software — CORSIKA

CORSIKA Dynamic Stack

- D. Baack (Dortmund), J.van Santen (DESY), K. Meagher (WIPAC)
- Better control shower generation
  - Kill showers as early as possible
  - Save CPU and GPU time
- Initial simple settings show factor of 2 reduction in CPU across all energy ranges.

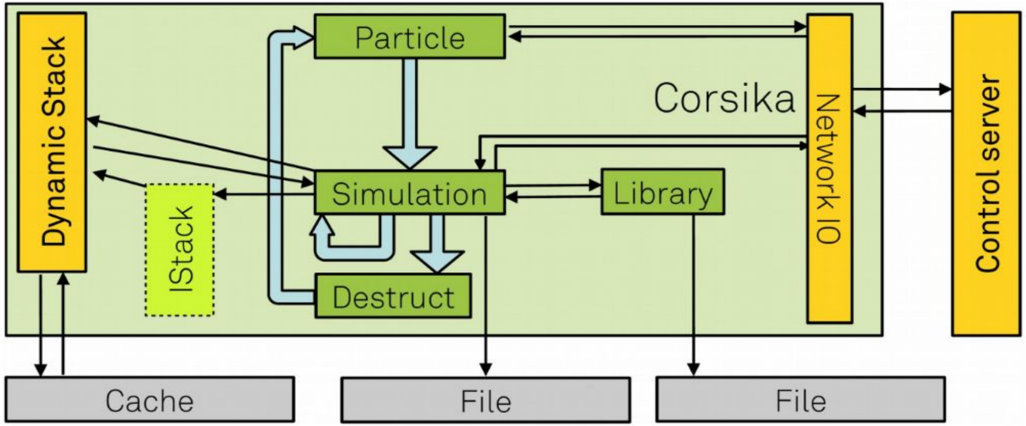
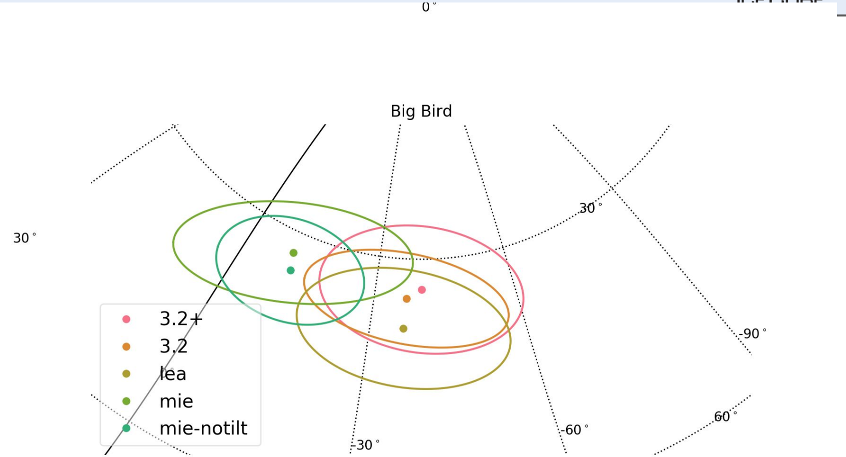


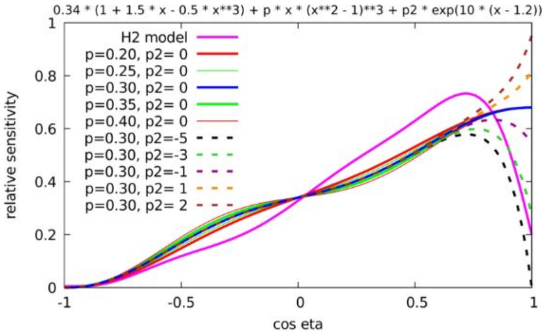
*Image from D. Baack (Dortmund)

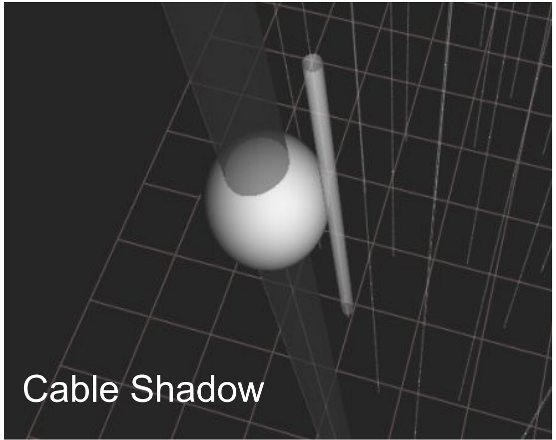
Physics Software — Photon Propagation

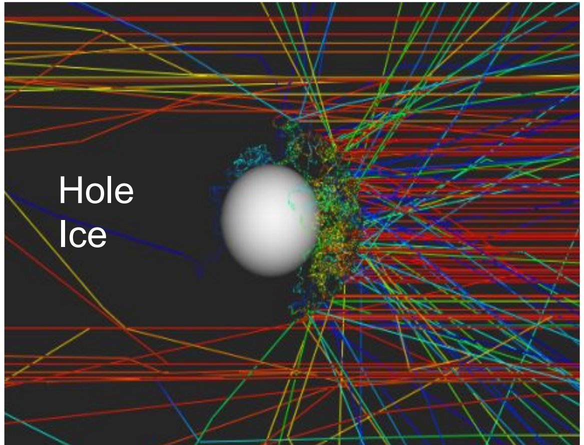
Ice model uncertainties

- Modeling the proper angular and overall acceptance of DOMs is an extremely hard problem *in situ*
- Important systematic effect





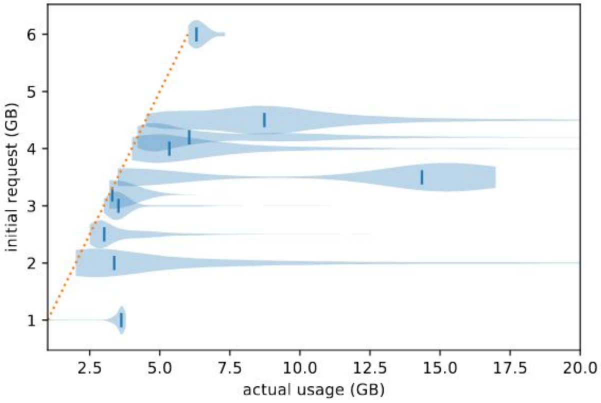




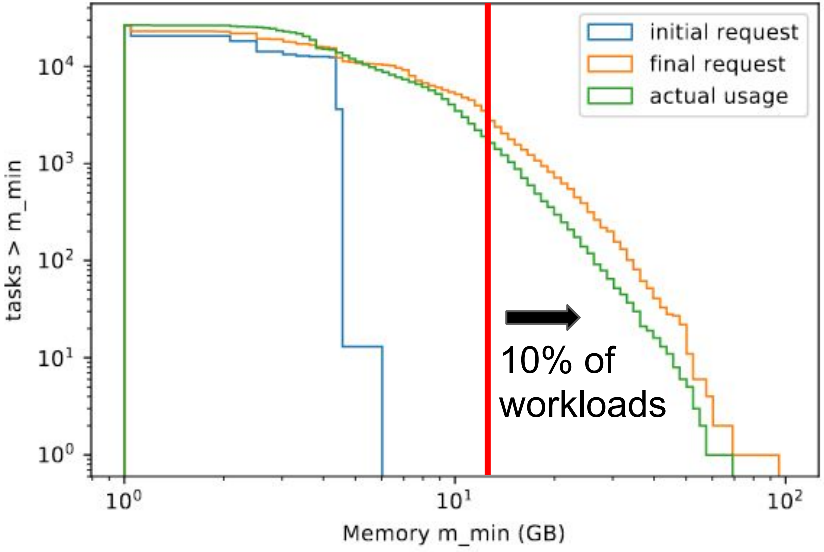
Physics Software — Photon Propagation

High memory usage

- A headache for scheduling
  - Initial request is a (hapless) guess
  - We continually retry with 1.5x higher requests
- Promising solution in testing



Integrated plot of tasks with memory > x

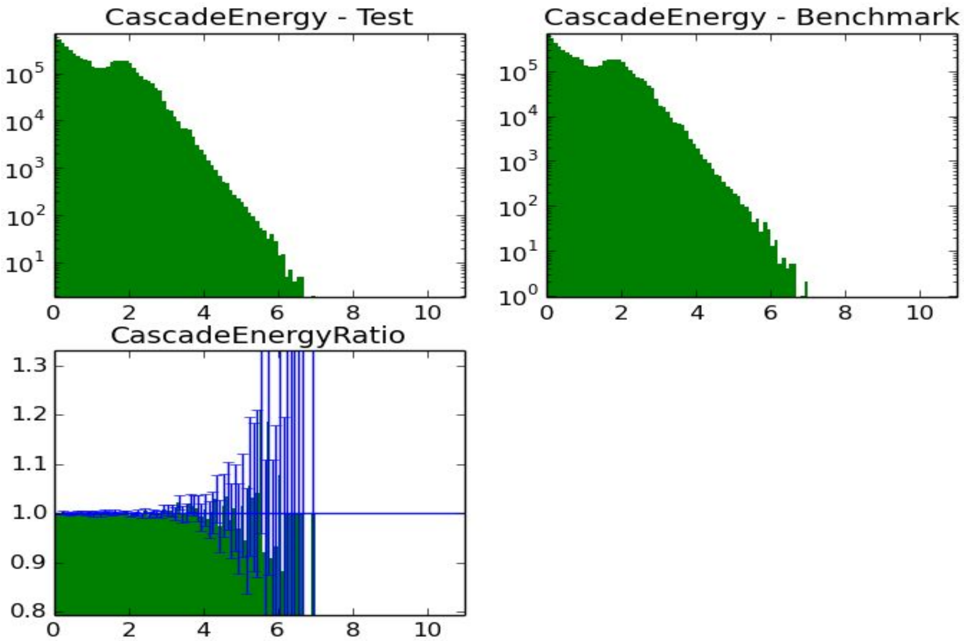


* All IceProd 2 tasks as of April 2018

Physics Software — Validation and Monitoring

Sanity Checkers - Data Quality

- Nightly comparisons of high-level physics
- Quick detection of software changes that might affect results
- Verify production datasets too



Production Software

Production Software — Overview

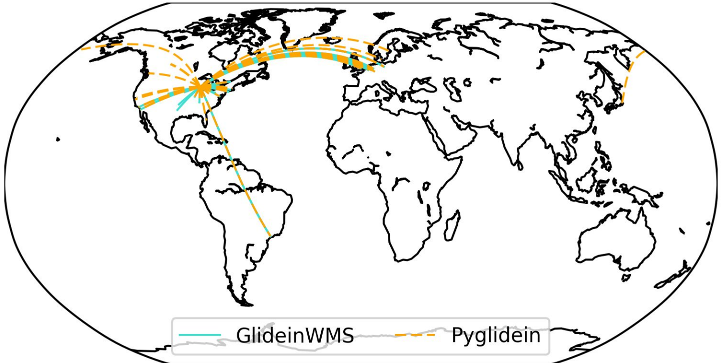
New Efforts | Needs Work | Working/ Stable

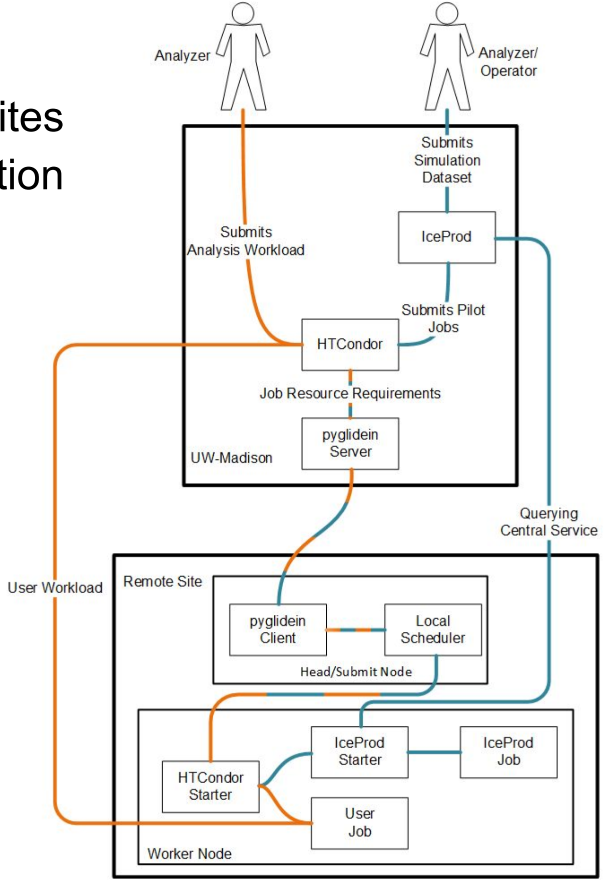
Production Software — pyglidein

pyglidein - IceCube Job Submission

- Lightweight python library that submits jobs at remote sites
- First developed to reduce need for site-specific information in IceProd
- Creates a **global HTCondor pool** for IceCube independent of OSG infrastructure
- Makes IceCube collaboration resources accessible to individual users and production alike

Glidein Locations



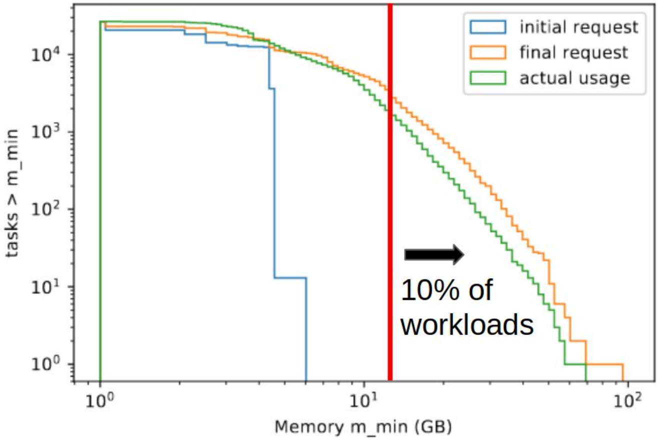


Production Software — IceProd Dataset Management

IceCube requires its own workflow management system - IceProd2

- Data provenance, dataset submission
- Diverse job requirements not experienced by similar experiments
  - Simulation requires GPUs
  - Large energy range
  - 10% of jobs require order of magnitude more memory
- Build with supercomputer support in mind
  - Demand for GPUs is increasing - Both from analyzers and production
  - Demand for Machine Learning focused environments increasing
  - Current and future supercomputers are GPU-equipped and built with machine learning in mind
  - Each supercomputer is an idiosyncratic system

Integrated plot of tasks with memory > x



Production Software — IceProd Dataset Management

Growing pains moving from IceProd1 to IceProd2+pyglidein+HTCondor:

- Database was not responsive enough
- Synchronization problem between distributed databases
- Scaling of storage servers
  - Issues with # connections for scratch, DESY gridftp servers
  - Bandwidth, storage limitations for scratch

IceProd2.4 release in October 2018

- Fix the scaling bottlenecks - Unified, more performant database
- Simple REST API - For services and users to connect to
- Multi-user + authentication - Having “normal” users operate the system

New scratch servers in Q1 2019

- Currently: single ZFS server
- Future: Ceph cluster (completed), multiple sites providing scratch disk (e.g. MSU)

Production Software — IceProd Dataset Management

Future goals:

- Distributed storage support
  - Intermediary file storage at more than one location
  - Spread load away from UW-Madison
  - Make queueing decisions based on location of input files
- Supercomputer support
  - Some clusters have limited external network
  - Still need to submit and monitor jobs with no external connections
  - Exploring this at a HTCondor, glidein, or IceProd level

Production Software — Long Term Archive — Current

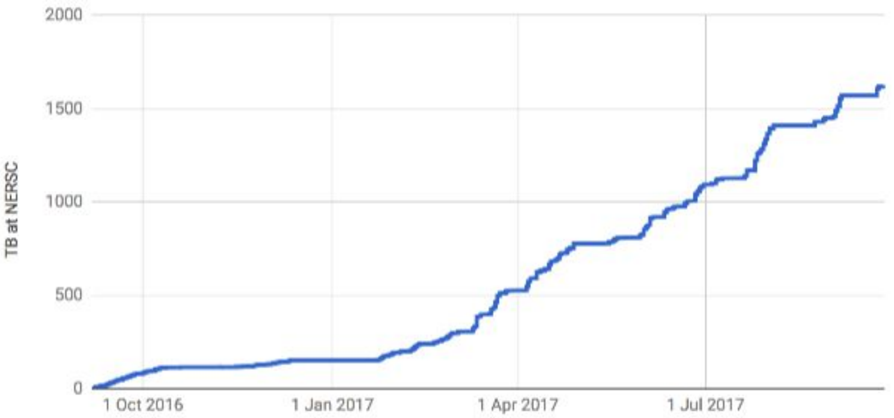
JADE extension (kanoite)

- This version archives data to tape at NERSC and DESY

How it works:

- JADE indexes data and prepares large bundle archives ~500GB
- The Globus transfer service manages transfers - Going closed/commercial soon; we are migrating away

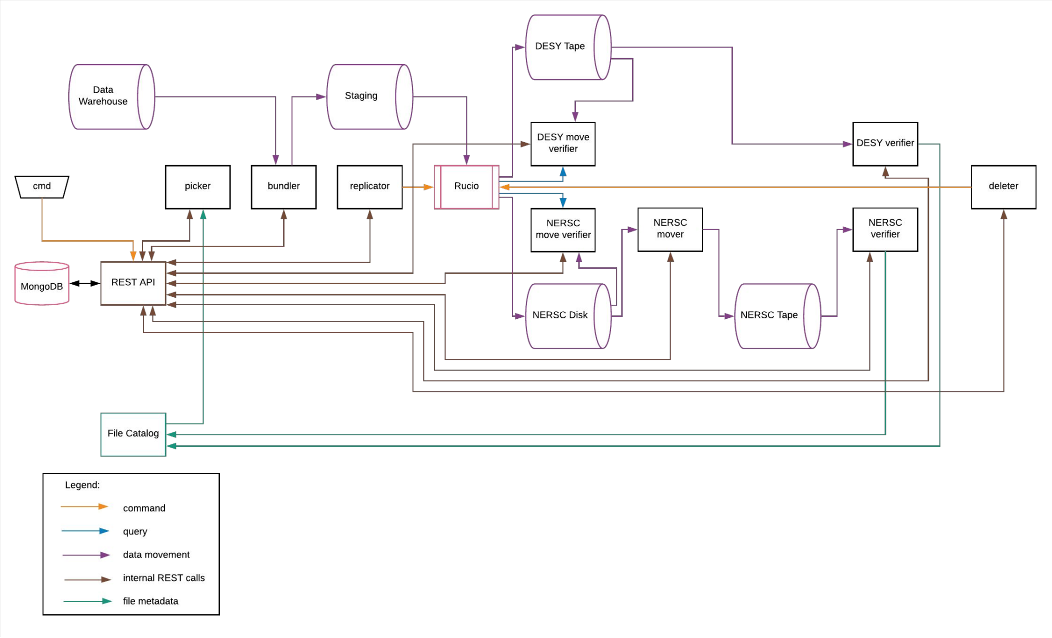
Pain point: Substantial operator effort



Production Software — Long Term Archive — Future

New software designated "Long Term Archive" (LTA)

- Written in Python
- Designed specifically for this purpose
- Integrates Rucio - ATLAS data transfer software
- NSF award 1841479 (CESER)
  Collaborative Research: Data Infrastructure for Open Science in Support of LIGO and IceCube



Data Processing – Ongoing L2, Pass 2 and Pass 3

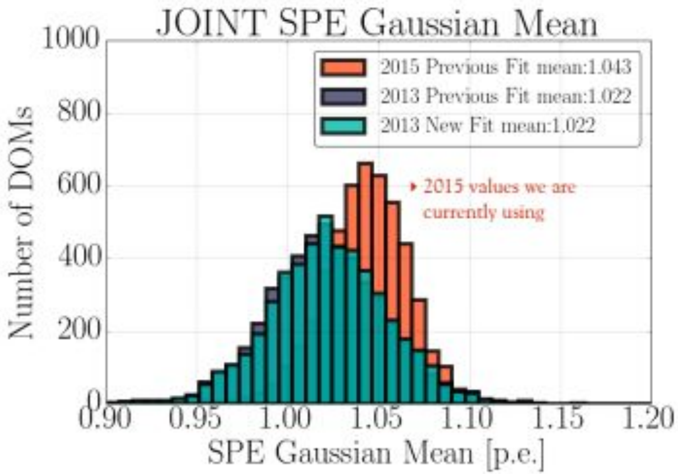
Data Processing ー Level 2

- The data taking for IC86-2018 began July 10, 2018
- Minimal differences with respect to IC86-2017
- Estimated resources required:
  - ~750 kCPU hours on NPX cluster at WIPAC
  - 100 TB of storage for both input and output data
- Production based on new database structure at pole and in Madison
- Level 2 data are typically available 1.5 weeks after data taking - Used to be 1 year
- Additional data validations have been added

Data Reprocessing ― Pass 2

In 2015, it was found that the SPE distribution peak obtained from the calibration chain is not centered around 1

- Correction of the SPE peak was introduced for the 2015 season
- The IC2015 24h test-run showed some changes when comparing data to the previous season
- Chance to make sure that all detector configurations (from IC79 to IC86-2014) are processed with the same L2 processing
  - Experimental data is more uniform across the science run years
  - Reduced impact on simulation requirements for individual years
  - Significantly less overhead for analysers to understand variations in seasons



Data Reprocessing ー Pass2

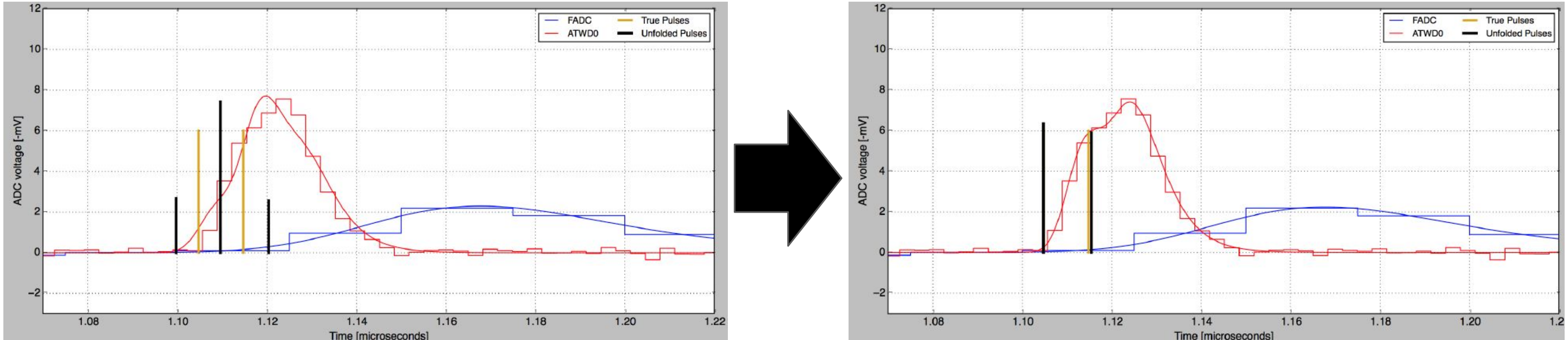
Completed:

- Reprocessing L2 and L3 of 7 years: 2010 (IC79) – 2016 (IC86-6)
  - Using software of season 2017 (IC86-7)
  - 8 years of data w/ same filters and reconstructions: 2010 – 2016 + 2017
- Total CPU hours:
  - 11M (L2) + 2M (L3)
  - About 15% more than anticipated
- Total storage:
  - 520 TB (L2) + 30 TB (L3)

Data Reprocessing — Pass 3

We recently discovered a mismatch between the first unfolded pulse and the first injected charge in feature extraction

- Checking impact on online filter and high level analyses to assess need for Pass 3 reprocessing - Appears to be subtle affect
- Opportunity to apply leap second correction at SDST level
- Reprocessing is large but we have the machinery in place and tested - Good exercise for processing needed with IceCube-Upgrade to utilize new information about systematics



Simulation Production

Simulation Production

Simulation Production is and has been transitioning

- Monte Carlo production has become individual analysis driven
- CORSIKA background generation still requires a unified plan — Too expensive
- Single Muon (MuonGun) simulations optimized for targeted volume and single muon backgrounds, e.g. oscillation analysis
- SimProd team provides production framework and technical assistance for running dedicated productions

Simulation Production — Analysis Drivers

Neutrino production

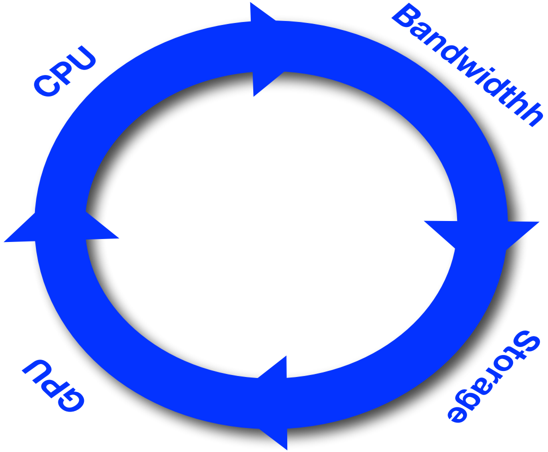
- Large matrix of systematic datasets
  - Photon-level production to accomodate - Large storage footprint
  - Systematic variations - ice model, DOM acceptance, hole ice
- Multiple generators
  - Low-Energy production - GENIE
  - Most other analyses - NeutrinoGenerator
  - High-Energy Sterile Neutrino - LeptonInjector (final state neutrino)
  - Moving to LeptonInjector as new neutrino event generator

Simulation Production — Optimization Cycle

Tackling one issue often exposes (or even introduces) a different challenge

Example: Speedup in individual steps (generation oversampling, GPU performance) can lead to alternatives:

- Larger files that are difficult to transfer
- Inefficient shorter jobs with large overheads



Simulation Production — Dynamics

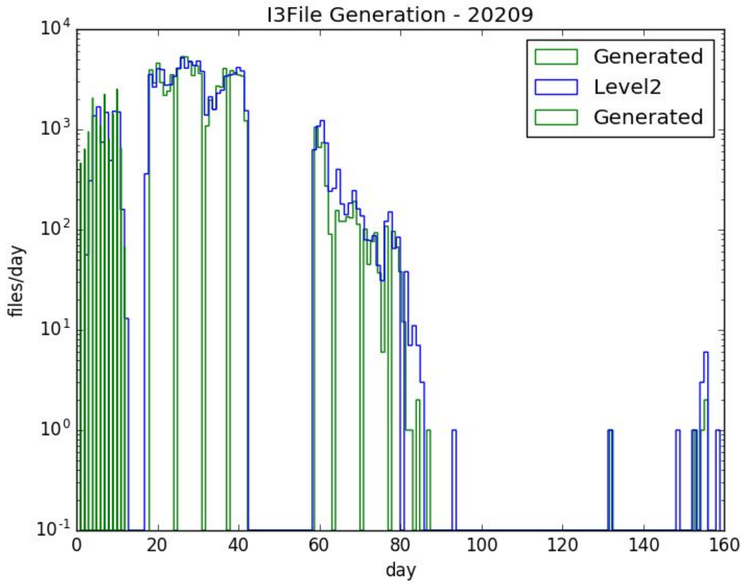
Issues with production dynamics

- 17 days to get to "full production"
- 2 suspensions due to disk issues
- ~25 day spin down? IceProd2 or IceSim?

Collaborators still don't know if this dataset is ready for use.

Publish at the 99% level

- Warn of potential bias due to failures
- Investigate further
- Roll fixes into the next release



IceCube Upgrade

IceCube Upgrade Considerations

IceCube Computing is a stable system that can be expanded for the needs of the Upgrade

Storage

- UW-Madison system can be expanded as needed - Will require negotiation with UW
- Need to negotiate new agreements with NERSC and DESY regarding backups

Compute

- Expand as needed - Greater focus on collaboration, in discussions with MSU to deploy hardware there
- Leverage national-level resources more, e.g. TACC’s upcoming Frontera supercomputer with GPUs, European supercomputers

Software

- Biggest area of work - Already being addressed
- Supercomputer integration with IceProd is essential
- Data organization, management, and access will be more software-driven

Summary

Summary

- IceCube Computing is providing the services as outlined in the M&O proposal
  - Data Warehouse and Storage infrastructure for the IceCube experiment
  - High performance computing cluster
  - Data Center support
  - Means to utilize collaboration resources
  - Offline software support and maintenance
- Expanded capabilities, availability, and use of IceCube computing grid
- Software capabilities and maintenance a focus
- Adoption of industry standards on the way
- Timely offline processing
- Proven the ability to (re)process current IceCube dataset in a timely fashion
- Facilitating transition to analysis-driven simulation production

Questions?

Backup

Personnel Changes

Significant personnel changes

- Management:
  - Gonzalo Merino returned to PIC as Deputy Director in Aug 2018
  - Benedikt Riedel took over as Computing Manager as of Dec 2018
  - David Schultz now manages the Production Software group
- Staff:
  - Heath Skarlupka (Operations Engineer) left for industry in March 2018 - Hiring replacement
  - Chad Sebranek (Web Developer) moved to another UW position in Aug 2018 - Hiring Replacement, important for public data releases
  - Paul Wisniewski (Network Engineer) moved to another UW position in 2017 - Services provided by UW-Madison
  - Alec Sheperd replaced Ben Stock as system administrator
- Overall, significant turnover for IceCube, but not atypical for industry.
- Small team, can and has lead to disruptions in service

Production Software — IceProd Dataset Management

What is IceProd?

Data provenance

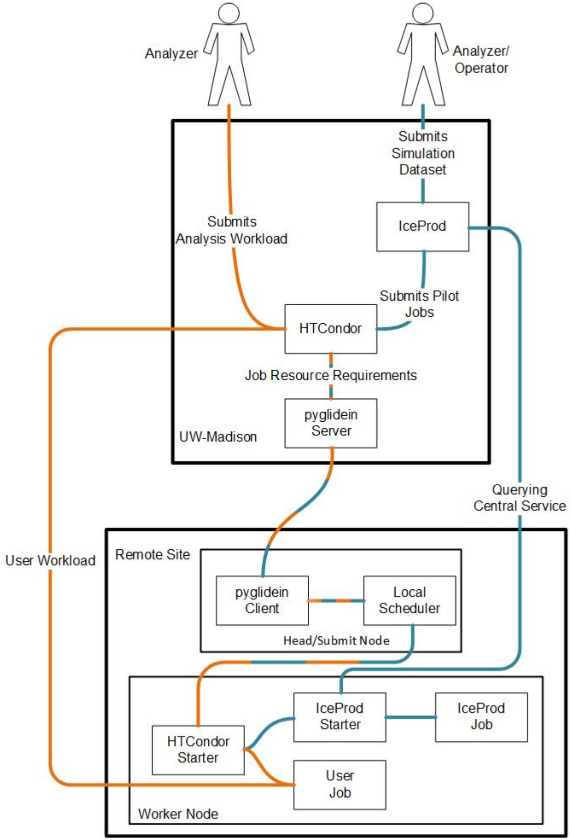
- Configuration for how a file was generated or processed
- Which software, what versions, when/where it ran, etc.

Dataset submission

- Monitor job status, resource usage
- Retry failed jobs - resubmit with different requirements

Use cases:

- Simulation production
- Experiment data processing
- Common analysis processing
- Other large-scale workloads



Production Software — IceProd Dataset Management

Switch from IceProd v1 to v2 in late 2016

- Moved from IceProdv1 to IceProd2+pyglidein+HTCondor

Software distribution using CVMFS

- /cvmfs/icecube.opensciencegrid.org
- Uniform software versions across all OS types
  - Simulation and reconstruction software
  - IceProd 2 software

Pilot job infrastructure

- Run multiple tasks sequentially and in parallel - Reduces startup overhead, connection costs with server
- Resource monitoring in real-time

Pledges as of Oct 2018

Site	Pledged CPUs	Pledged GPUs
Aachen	27700*	44*
Alabama		6
Alberta	1400	178
Brussels	1000	14
Chiba	196	6
Delaware	272	
DESY-ZN	1400	180
Dortmund	1300*	40*
LBNL	114	
Mainz	1000	300
Marquette	96	16
MSU	500	8
NBI		10
Penn State	3200*	101*
Queen’s		55
Uppsala	10	
UMD	350	112
UTA	50	
UW-Madison	7000	440
Wuppertal	300	
TOTAL (exclusive)	13688	1325
TOTAL (all)	45888	1510

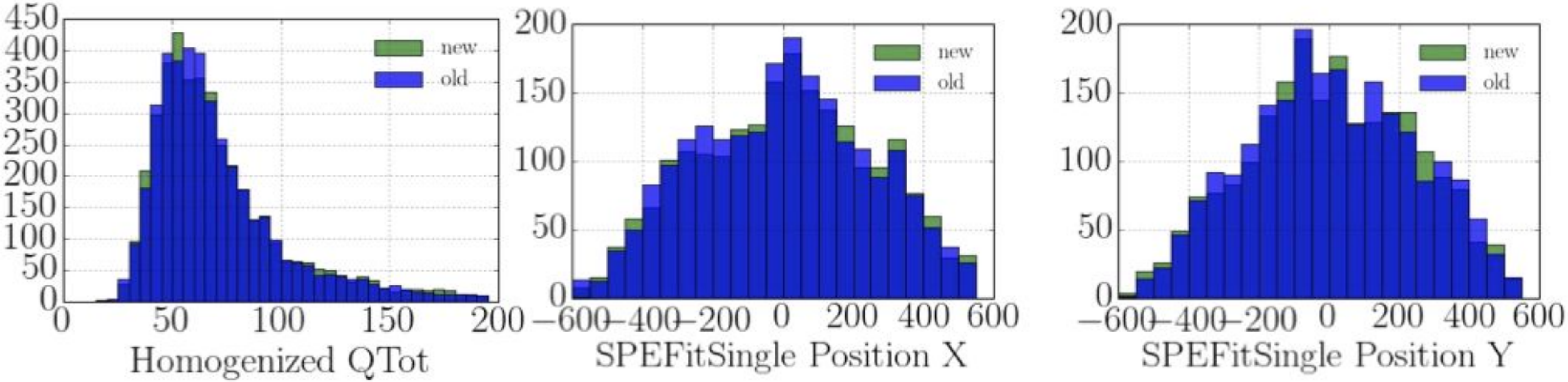
*Indicates maximum shared resources, not exclusively for IceCube

Pass2: Level2 validation

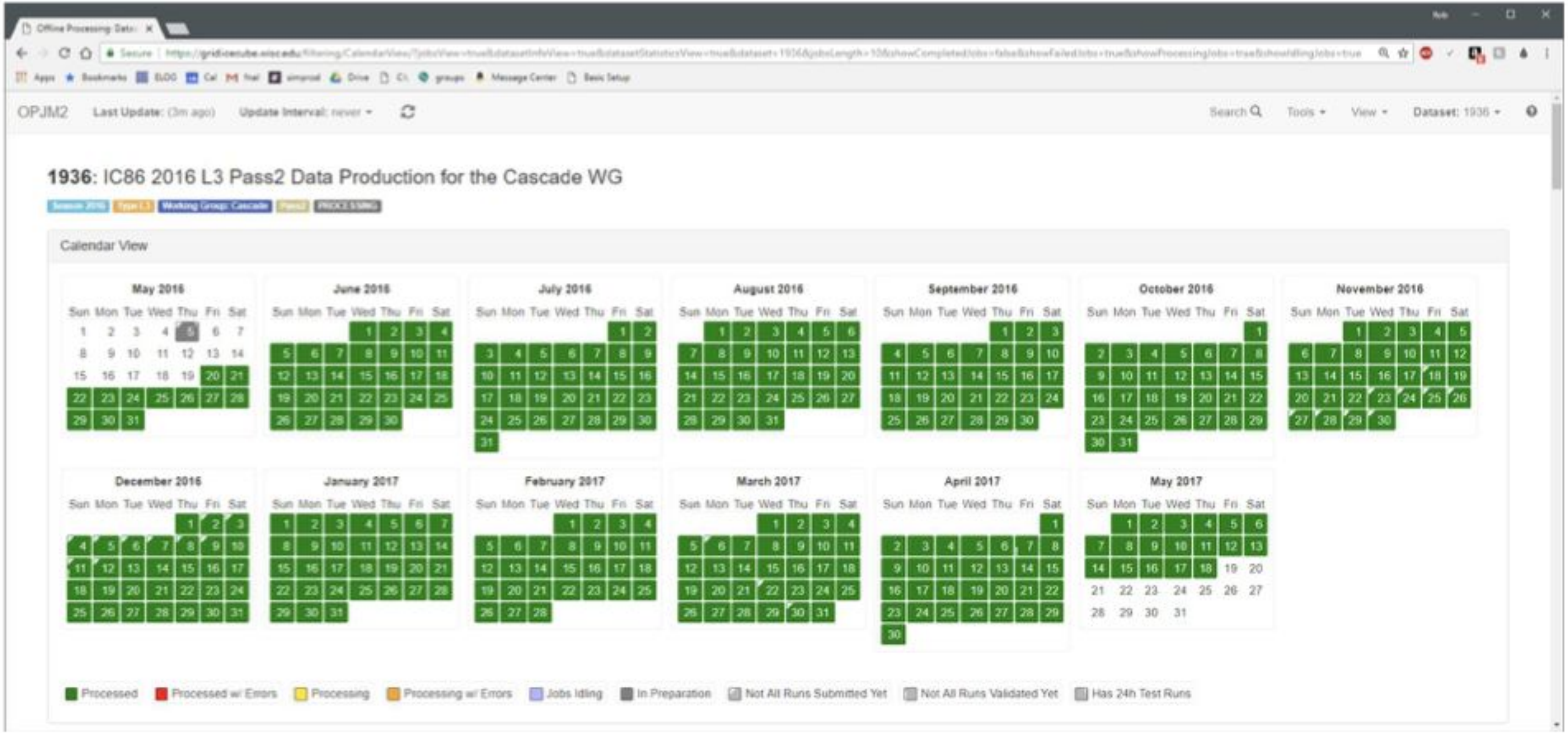
Fully processed 9 runs of interest for the HESE analysis:

128973 129112 129253 129281 129316 129402 129474 129497 129510

Spencer Axani compared 1 run from 12/26/2016 (new) to a run from 12/26/2015 (old)



Pass2 L3 production example: Cascade filter



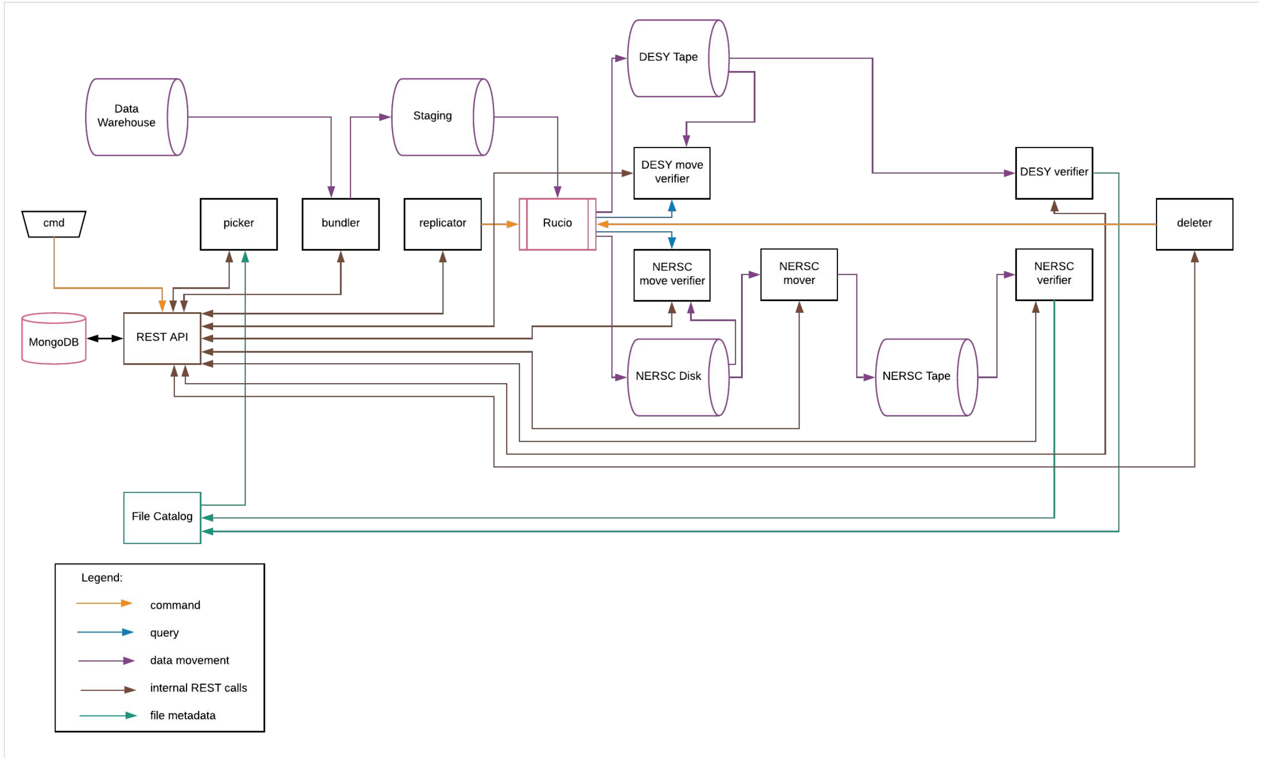
South Pole Data Transfer - JADE

JADE data transfer tool:

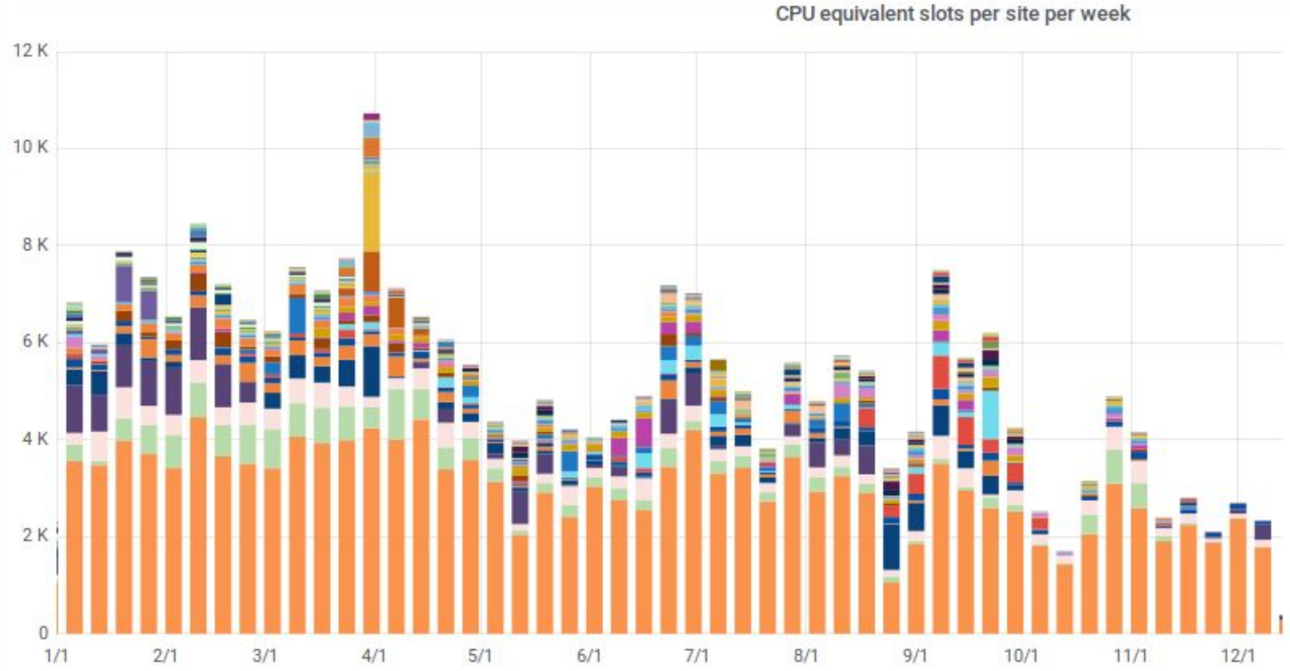
- Written in Java
- Transfers data from South Pole to Madison
  - Via satellite managed by ASC polar contractor
  - FTP input server at pole, output server in US

Long Term Archive - Future

New software designated “Long Term Archive” (LTA)

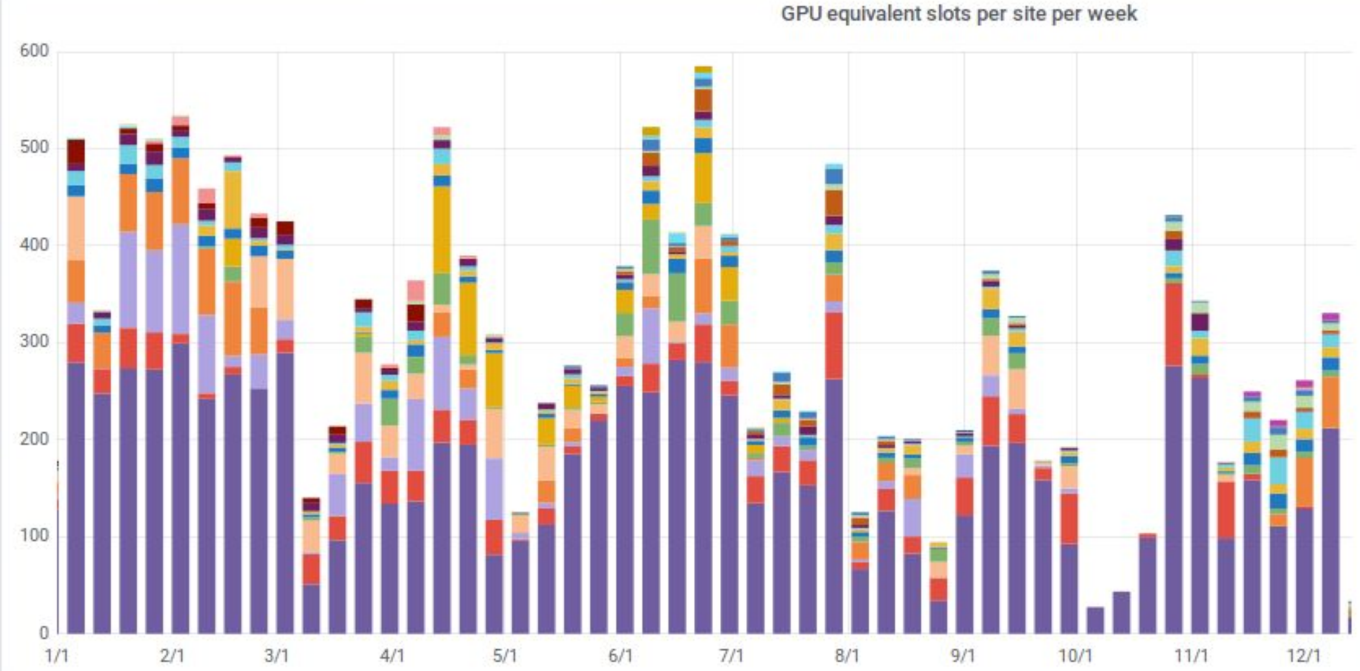


All Sources - CPU Usage 2018 and Site



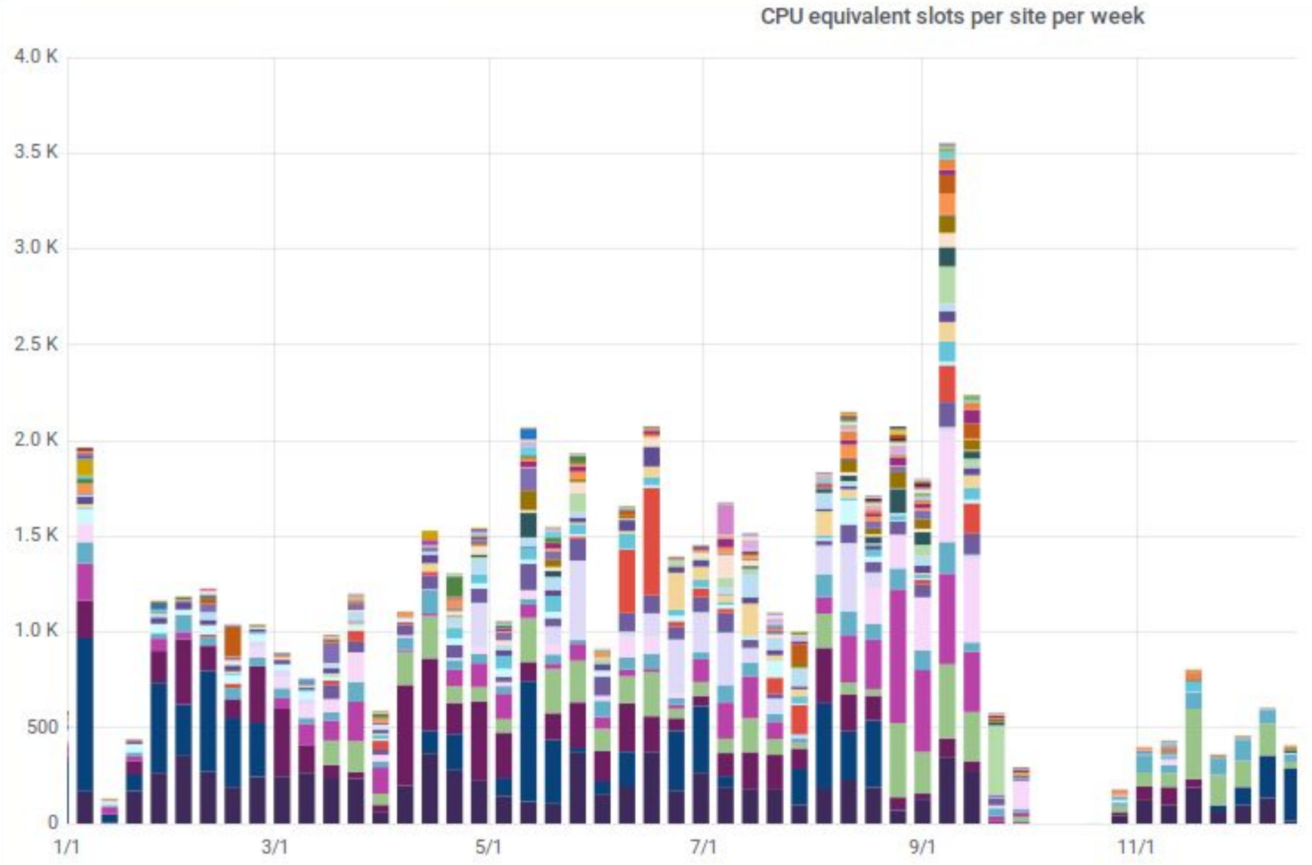
	max	avg	current
NPX	4.463 K	2.951 K	282
DESY-ZN	1.040 K	324	1
T2B_BE_IIHE	639	291	2
DESY	1.088 K	251	91
msu	1.044 K	189	0
UKI-NORTHGRID-MAN-HEP	420	107	0
LIDO_Dortmund	219	92	10
UMD	681	85	0
GPGrid	996	71	0
USCMS-FNAL-WC1	739	71	0
RWTH-Aachen	376	66	0
Bridges	596	59	0
CHTC	208	58	0
BEgrid-ULB-VUB	204	51	0
Nebraska	282	41	0

All Sources - GPU Usage 2018 and Site



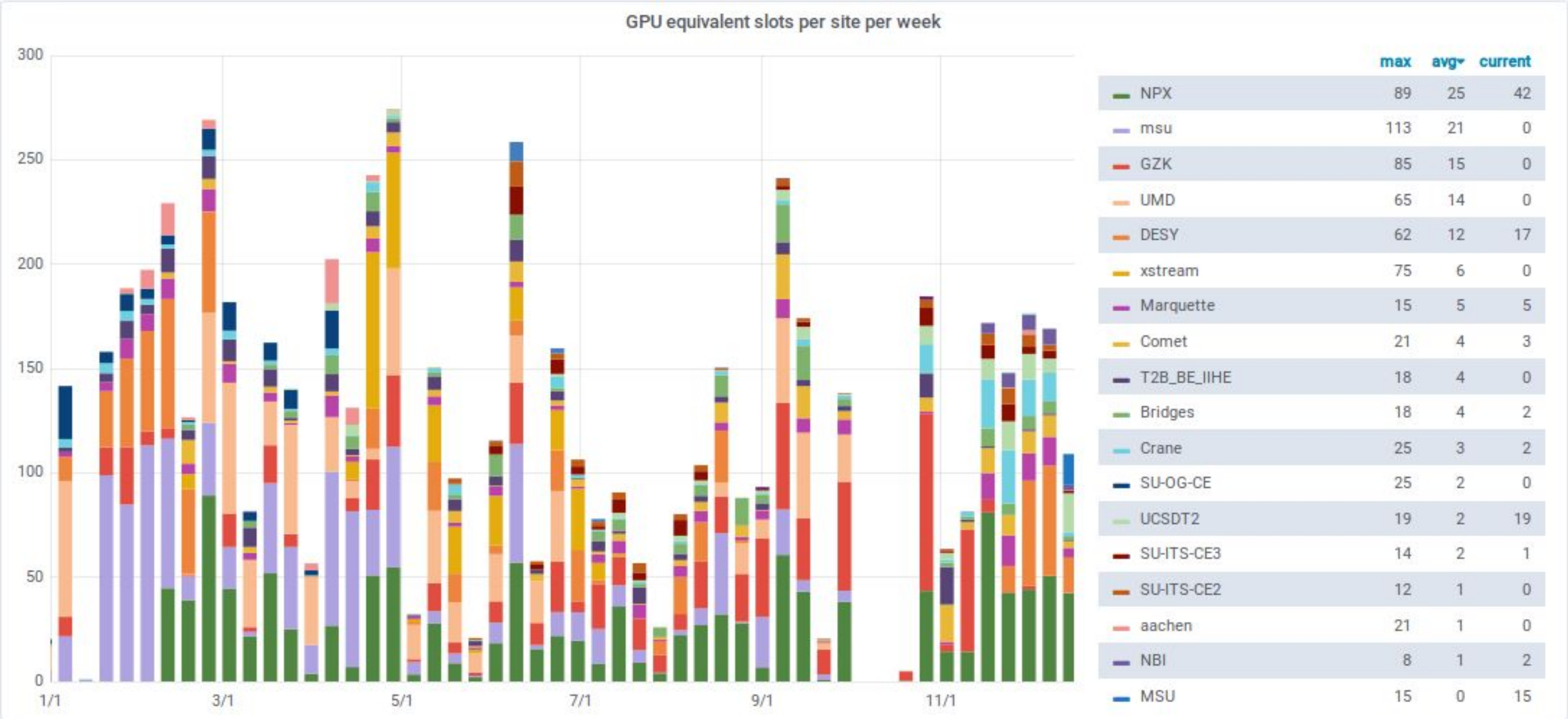
	max	avg	current
NPX	299	172	17
GZK	85	23	0
msu	113	21	0
DESY	76	18	7
UMD	65	15	0
Bridges	56	9	0
xstream	90	9	0
Marquette	16	7	2
Comet	59	6	1
Crane	27	6	0
T2B_BE_IIHE	18	5	0
SU-ITS-CE3	27	3	0
SU-OG-CE	25	2	0
UCSDT2	15	2	4
SU-ITS-CE2	16	2	0

Simulation Production - CPU Usage 2018 and Site

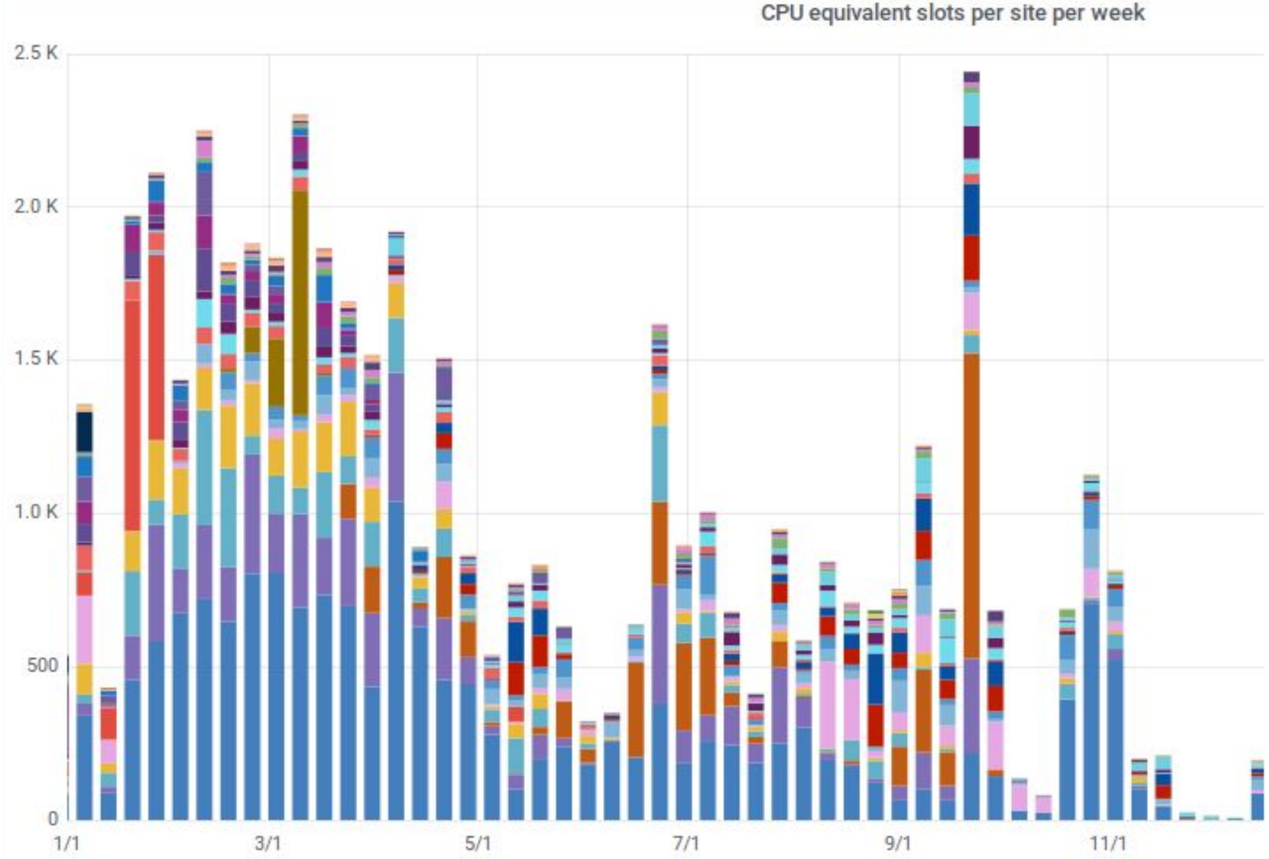


	max	avg	current
T2B_BE_IIHE	375	168	13
DESY	800	144	275
DESY-ZN	525	129	3
NPX	387	102	31
msu	699	96	0
LIDO_Dortmund	166	58	40
UMD	571	55	5
USCMS-FNAL-WC1	411	51	0
CHTC	140	39	0
Bridges	563	36	0
UCSDT2	126	22	0
NWICG_NDCMS	107	22	3
GZK	192	21	0
Marquette	103	20	0
UKI-NORTHGRID-MAN-HEP	117	19	0
GPGrid	361	19	0
SU-ITS-CE3	127	13	13
UColorado_HEP	121	13	0

Simulation Production - GPU Usage 2018 and Site



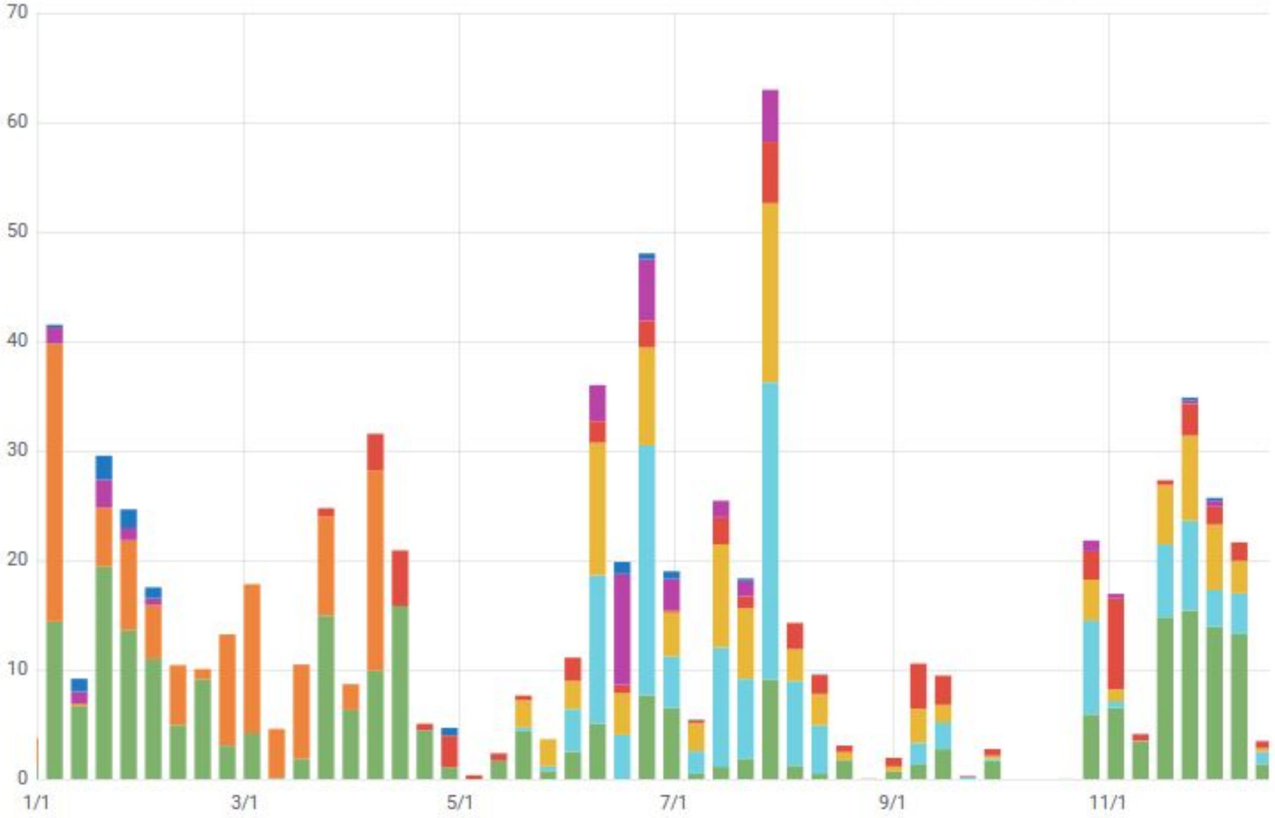
GlideinWMS - CPU Usage 2018 and Site



	max	avg	current
DESY-ZN	1.040 K	326	89
UKI-NORTHGRID-MAN-HEP	420	107	0
GPGrid	996	71	0
RWTH-Aachen	376	66	0
BEgrid-ULB-VUB	204	51	0
Nebraska	282	42	9
CA-MCGILL-CLUMEQ-T2	753	31	0
NWICG_NDCMS	127	31	35
UColorado_HEP	129	28	10
SU-ITS-CE2	148	25	11
SU-ITS-CE3	170	24	13
USCMS-FNAL-WC1	735	21	0
UCSDT2	79	20	4
MWT2	93	19	10
IIT_CE1	108	16	0
SU-OG-CE1	140	15	0
BNL-ATLAS	107	14	7
SU-OG-CE	110	14	0

GlideinWMS - GPU Usage 2018 and Site

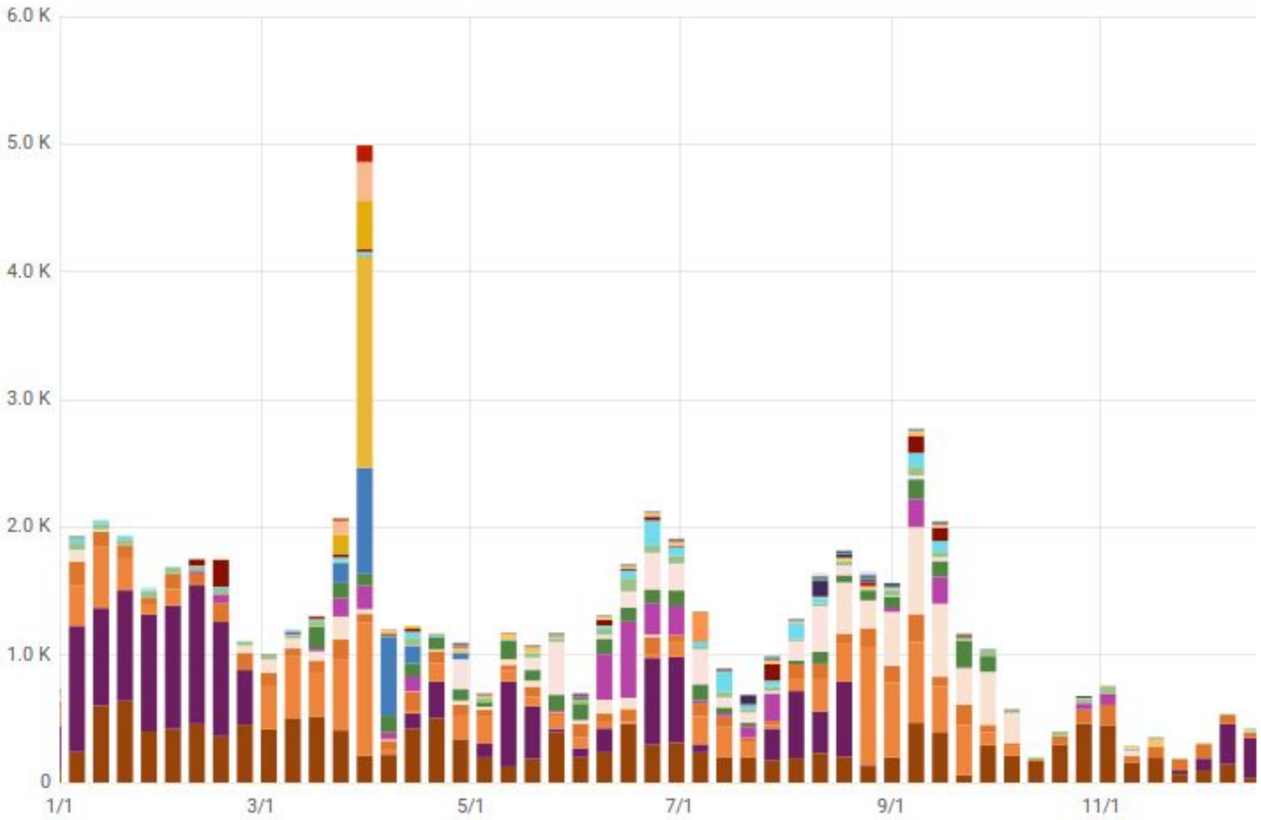
GPU equivalent slots per site per week



	max	avg	current
Crane	19.39	5.24	1.38
SU-ITS-CE3	27.14	2.87	1.11
SU-OG-CE	25.40	2.35	0
SU-ITS-CE2	16.37	2.17	0.33
UCSDT2	8.28	1.23	0.65
UKI-NORTHGRID-MAN-HEP	10.13	0.76	0
UKI-LT2-QMUL	2.18	0.20	0

Pyglidein - CPU Usage 2018 and Site

CPU equivalent slots per site per week



	max	avg	current
T2B_BE_IIHE	639	292	39
DESY	1.088 K	255	312
msu	1.044 K	189	0
LIDO_Dortmund	219	93	46
UMD	681	85	12
Bridges	596	59	0
CHTC	208	58	0
USCMS-FNAL-WC1	411	50	0
Cedar	830	35	0
AWS	1.643 K	32	0
Marquette	105	32	10
GZK	200	25	0
Comet	210	17	0
Illume	385	11	0
NWICG_NDCMS	54	11	3
Guillimin	300	8	0
UCSDT2	125	6	0
PSU	231	5	0

Pyglidein - GPU Usage 2018 and Site