Computing Report

IceCube Maintenance and Operations Review
Madison, WI
January 8th, 2019
Benedikt Riedel

Outline

- Deliverables
- 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

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
- 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 10 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 at UW
  - Both upgrades of older cards and new purchases
  - Growing GPU/accelerator resources through applying to outside resources, e.g. XSEDE

Computing Infrastructure — Collaboration

- Introduced **computing pledge system** to incentivise investment in computing - Computing resources are in-kind contributions
- Continually expanding the IceCube processing grid by using in-house developed `pyglidein` - Works on campus clusters, regional computing centres, national supercomputers
- Direct investments in IceCube computing resources by other institutions - UAlberta, MSU, UMD, DESY, Mainz
- Established **long-term archive** at NERSC for IceCube raw and processed 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

CPU and GPU Compute - Pledges as of Oct 2018

Storage

- Primary Data Warehouse:
  - 10 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

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

Computing Infrastructure — National

Significant invest in GPU resources on national-scale HPC resources

- USA
  - Extensive use of **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 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, but needs to be expanded

Computing Infrastructure — National

XSEDE - 2018 Allocation

- PSC Bridges: 287 kSUs of GPUs - 70% used with 6 months remaining
- SDSC Comet: 180 kSUs of GPUs - 65% used with 6 months remaining
- OSG: 4 MSUs of CPU - 100% used with 6 months remaining
- SU = Service Unit

DOE - 2018 Allocation

- Titan: 1 Mnode-hours - Used, 2019 allocation applied for
- NERSC - Cori: 1.25 MNERSC-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
- Resource sharing across Multi-Messenger Astronomy - WLCG as model
- Modernization of Workflows
  - Deployment of software with containers and kubernetes
  - 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
- Reorganization effort, details to follow in this talk

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

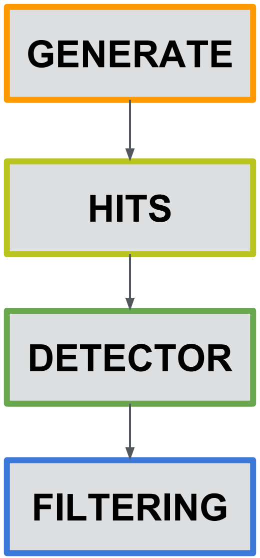
- High level of productivity
- Code optimization
  - Memory, CPU profiling
  - Data structures
  - Optimization schemes
  - Simulation quality/improvements

Yearly Software Bootcamps

- Introduce new students and postdocs to IceCube and IceCube software

Physics Software — Simulation Chain

Generate Cosmic Ray showers
and neutrino interactions
Propagate to detector



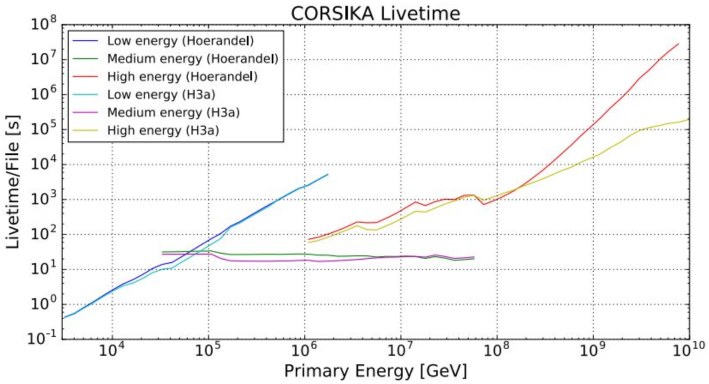
Photon propagation in ice

DOM hardware simulation
DAQ trigger emulation

L1 (pole) and L2 (offline)
reconstruction and filtering

Physics Software — CORSIKA

Low Energy CORSIKA Problem



- Generating CORSIKA with “low energy” primaries would be scientifically interesting
- Production wasteful - products don’t trigger detector, so resources are “wasted”
- MuonGun is much faster, but introduces systematics
- Analyzers would prefer CORSIKA, not possible with current resources by brute force

Physics Software — CORSIKA

CORSIKA Dynamic Stack

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

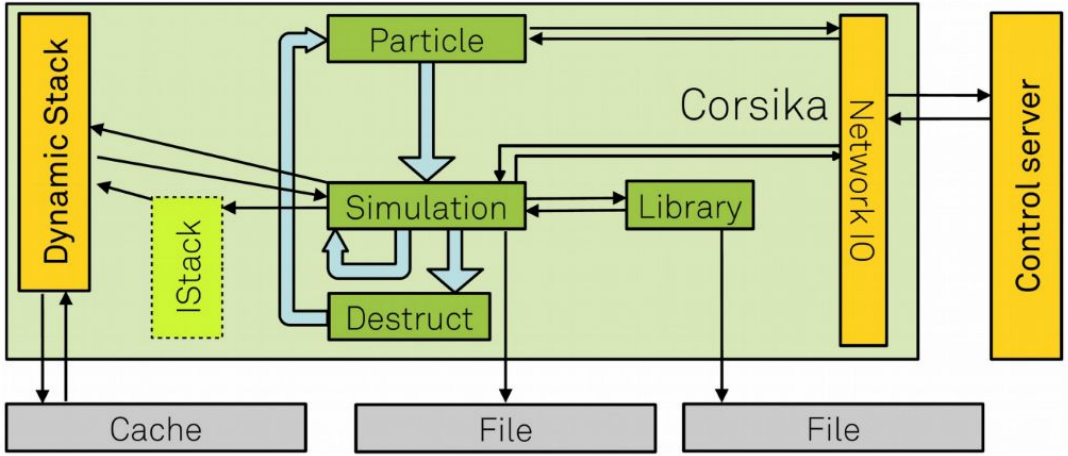
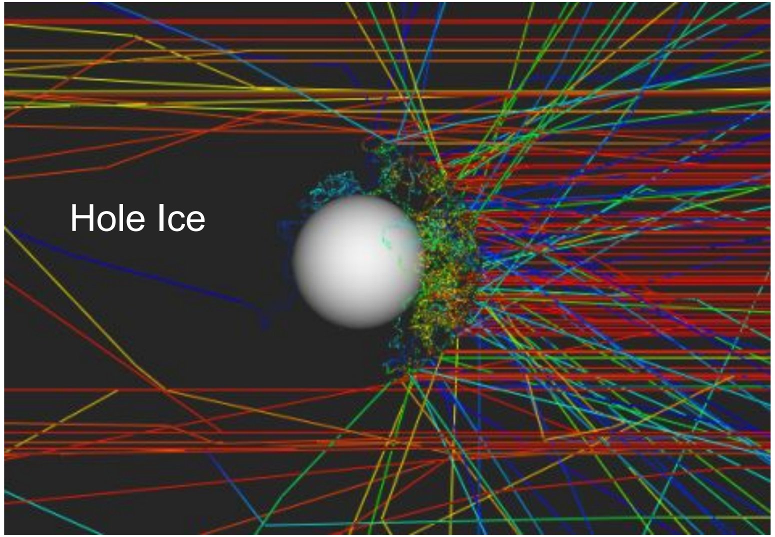


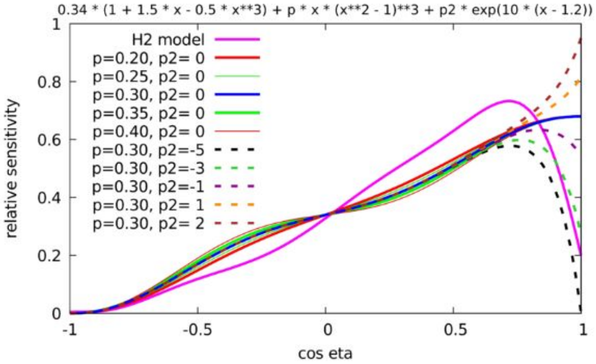
*Image from D. Baack (Dortmund)

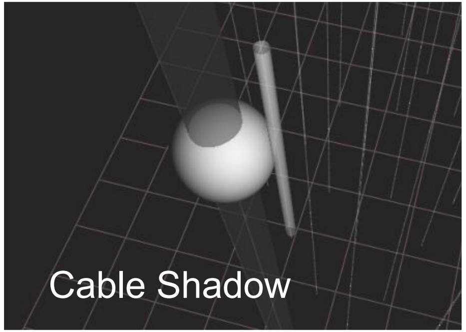
Physics Software — Photon Propagation

Ice model uncertainties

- Modeling the proper angular and overall acceptance of DOMs is an extremely hard problem *in situ*
- Important systematic effect, esp. in low-energy analyses



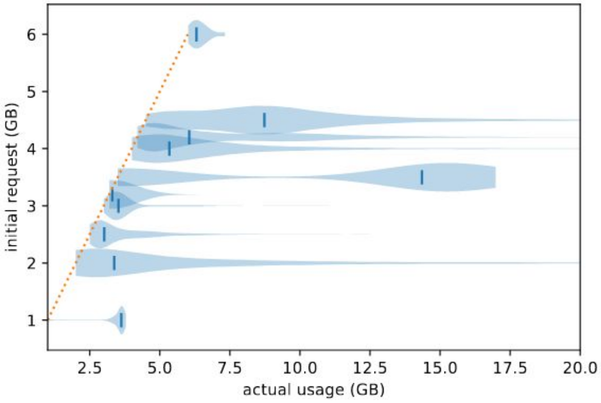




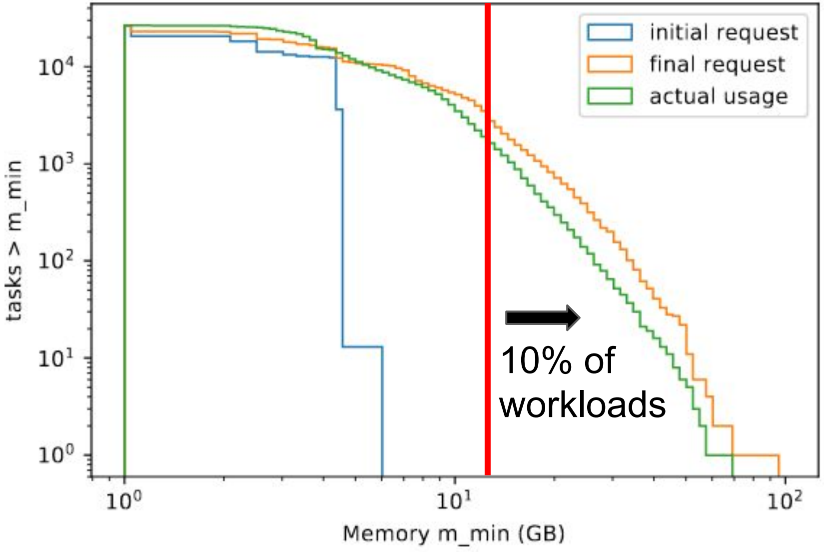
Physics Software — Photon Propagation

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

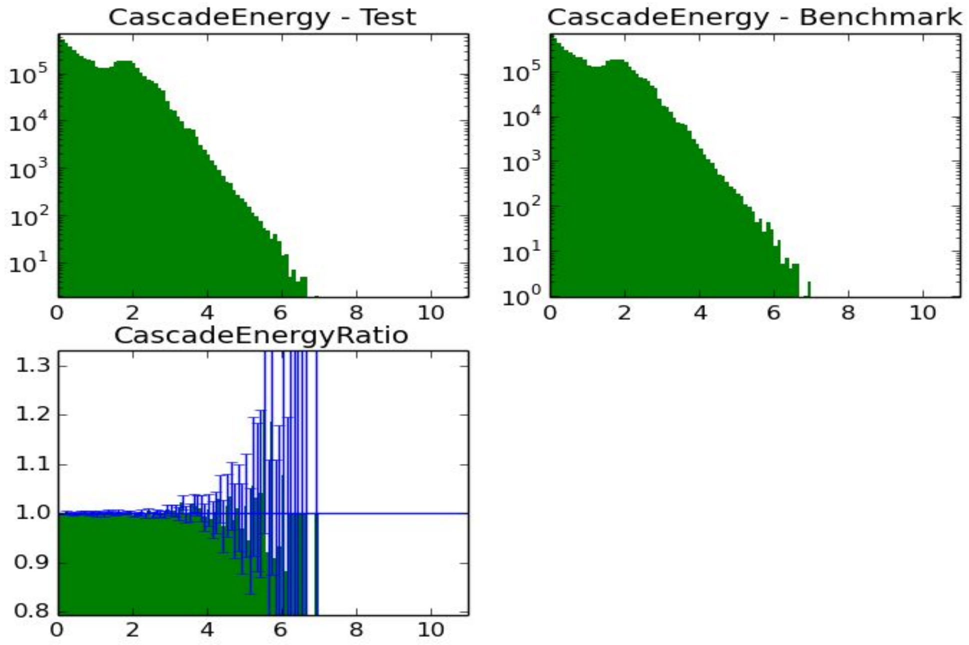


* All IceProd 2 tasks as of April 2018

Physics Software — Validation and Monitoring

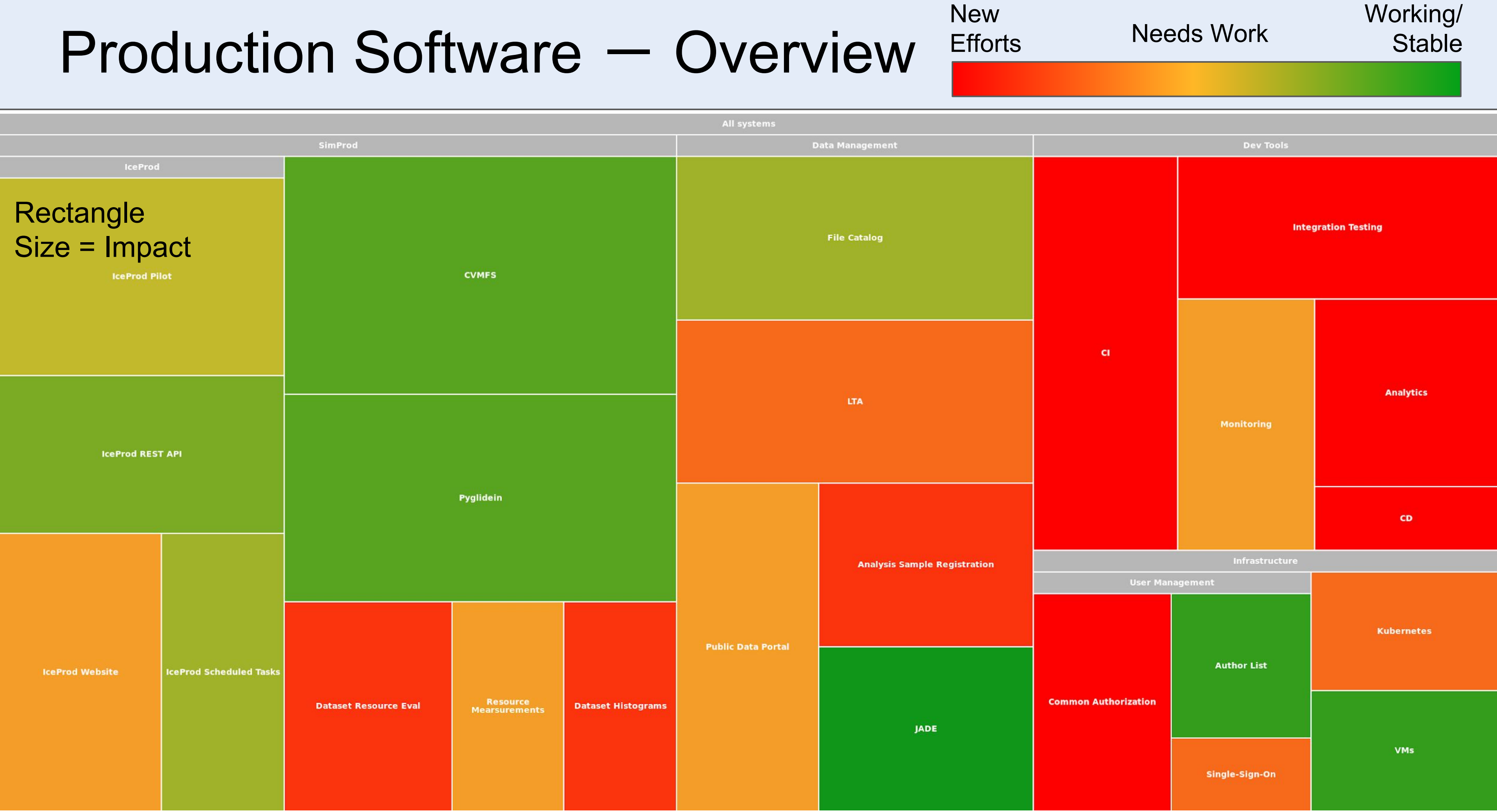
Sanity Checkers - Data Quality

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



Production Software

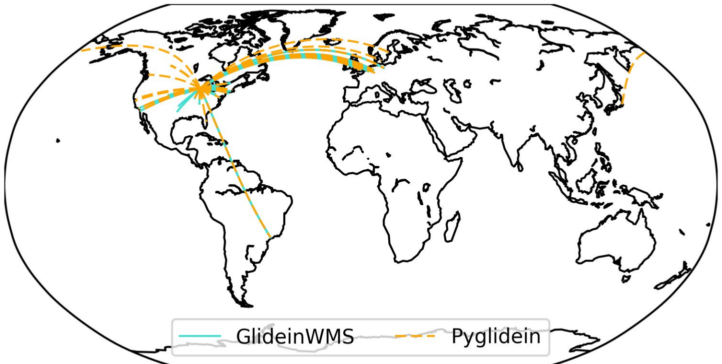
Production Software — Overview

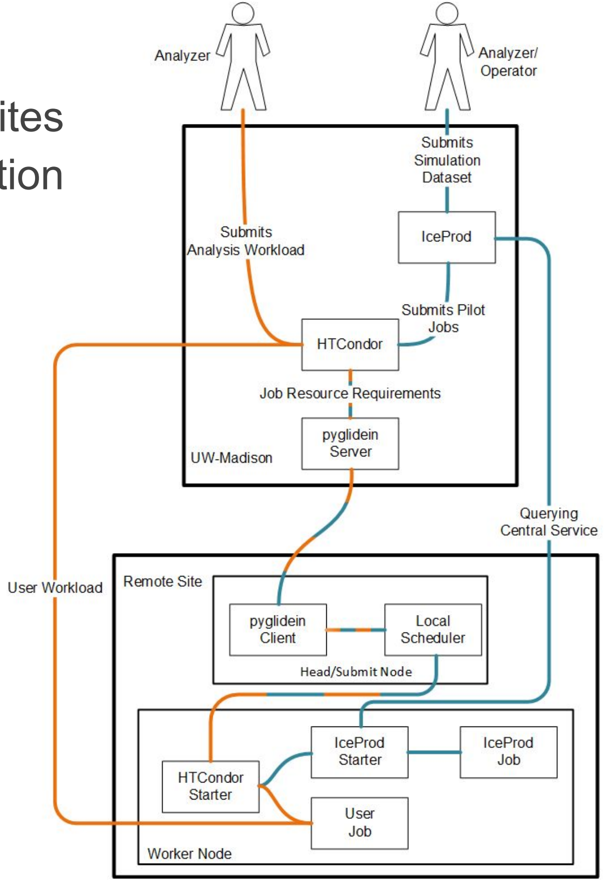


Production Software ─ pyglidein

pyglidein - IceCube Job Submission

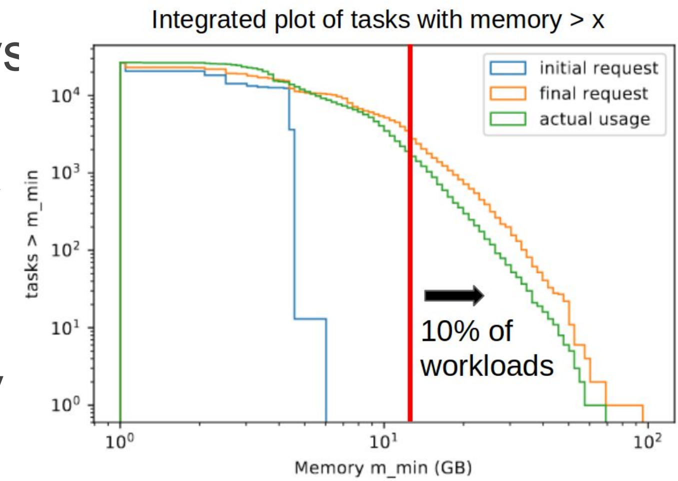
- Lightweight python library that submits jobs at remote sites
- First developed to reduce need for site-specific information in IceProd
- Creates a **global HTCondor pool** for IceCube independent of OSG infrastructure
- Makes IceCube collaboration resources accessible to individual users and production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





Production Software — IceProd Dataset Management

IceCube requires its own workflow management sys
- IceProd
  - Diverse job requirements not experienced by similar experiments
    - Bulk of simulation requires GPUs
    - 10% of jobs require order of magnitude more memory
  - Ability to run on supercomputers
    - 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



Production Software — IceProd Dataset Management

What is IceProd?

Data provenance

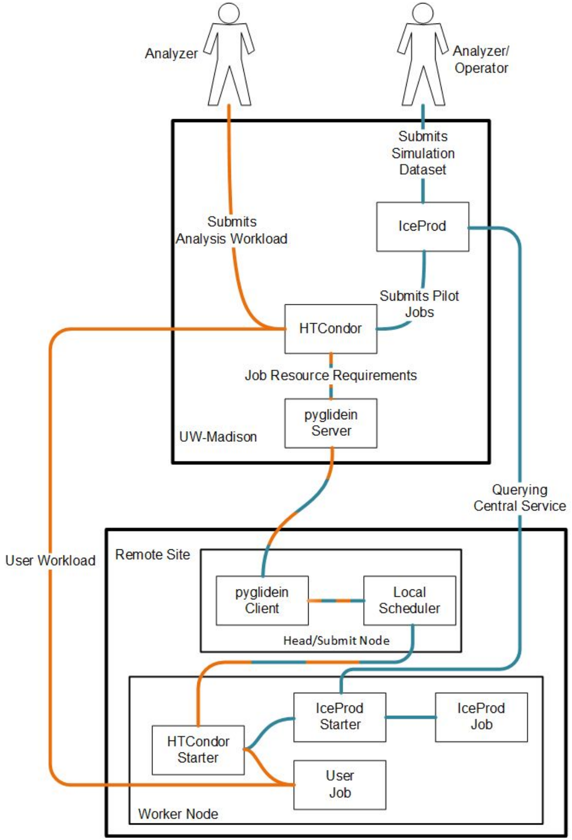
- Configuration for how a file was generated or processed
- Which software, what versions, when/where it ran, etc.

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

Production Software — IceProd Dataset Management

Growing pains:

- 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

IceProd 2.4 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

New scratch servers in Q1 2019

- Currently: single ZFS server
- Future: Ceph cluster

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

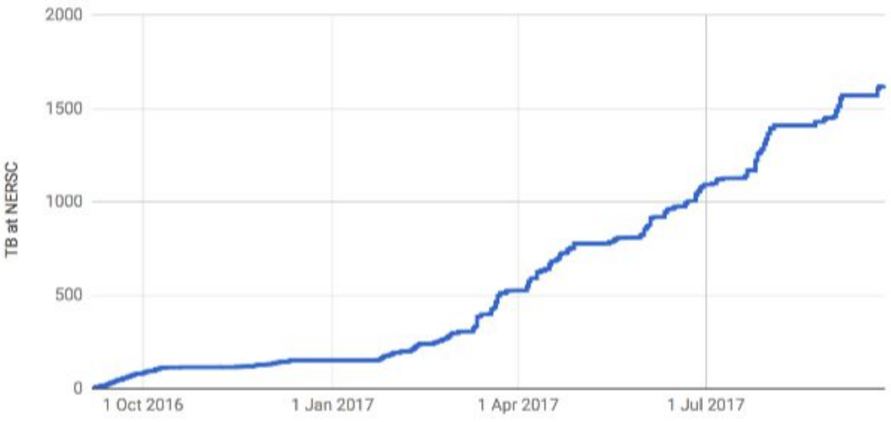
JADE extension (kanoite)

- 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

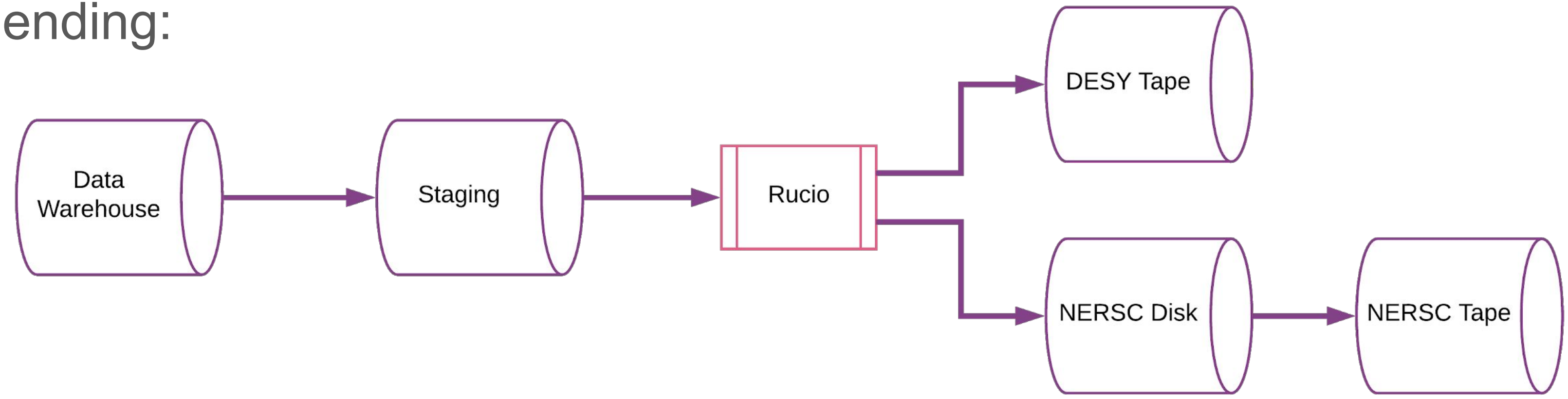


Production Software — Long Term Archive — Future

New software designated “Long Term Archive” (LTA)

- Written in Python
- Designed specifically for this purpose

Sending:



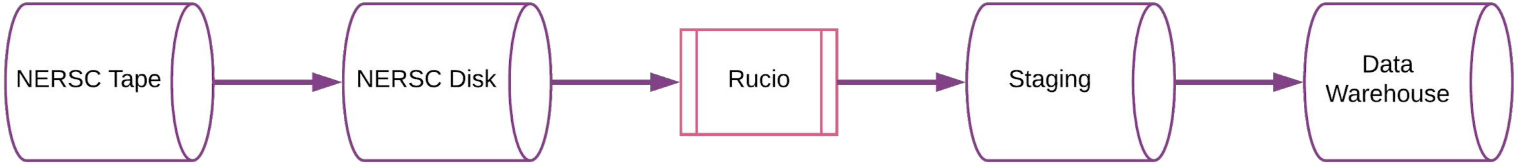
Production Software — Long Term Archive — Future

New software designated “Long Term Archive” (LTA)

- Integrates Rucio -- ATLAS data transfer software
- NSF award 1841479 (CESER)

  Collaborative Research: Data Infrastructure for Open Science in Support of LIGO and IceCube

Retrieving:



Data Processing – Ongoing L2, Pass 2 and Pass 3

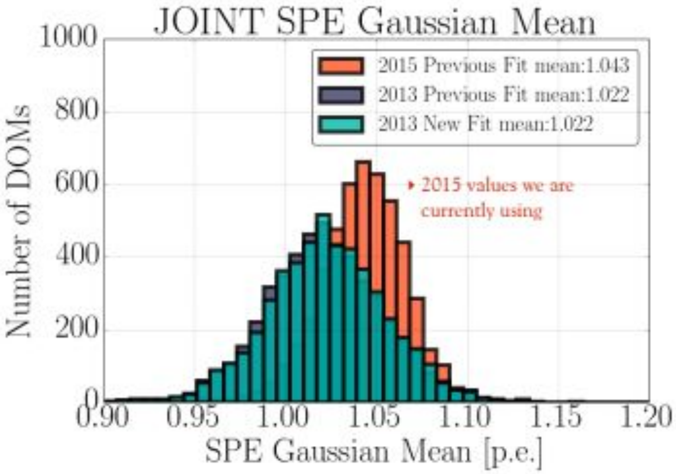
Data Processing ー Level 2

- 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
- Level2 data are typically available 1.5 weeks after data taking
- 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 exp. data to the previous season
- Needed correction



Data Reprocessing — Pass 2

Determination:

- Pass2: Apply the SPE corrections to all pre-2015 experimental data (back to IC79)
- Start from SuperDST data, apply the SPE correction, re-run L1 & L2
- Provided an opportunity to also 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

Data Reprocessing ― Pass2

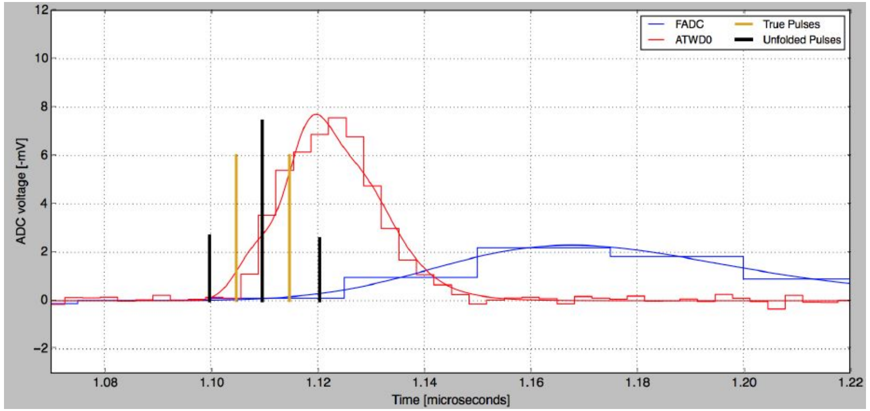
Now complete:

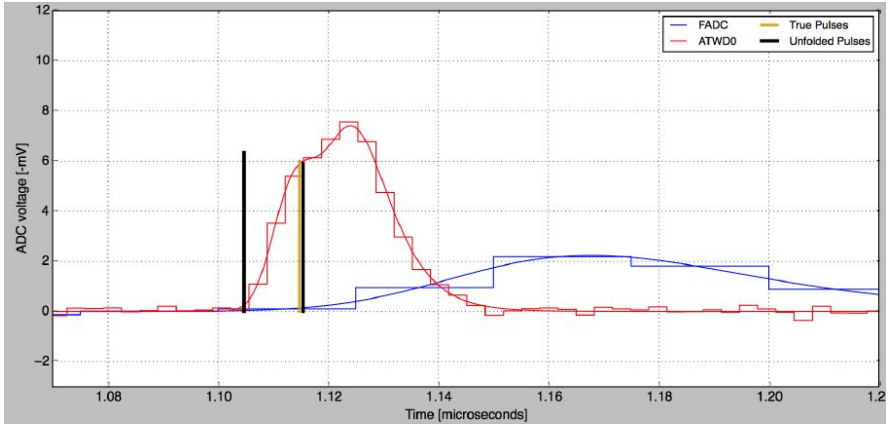
- Reprocessing L2 and L3 of 7 years: 2010 (IC79) – 2016 (IC86-6)
  - Using software of season 2017 (IC86-7)
  - 8 years of data w/ same filters and reconstructions: 2010 – 2016 + 2017
- Total CPU hours:
  - 11M (L2) + 2M (L3)
  - About 15% more than anticipated
- Total storage:
  - 520 TB (L2) + 30 TB (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 whether filter cuts need to be re-optimized
- Check impact on high level analyses to assess urgency for Pass 3 data reprocessing
- Opportunity to apply leap second correction at SDST level
- Reevaluate online filters for 2019 and apply to all years
- Reprocessing is large but we have the machinery in place and tested





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
- 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

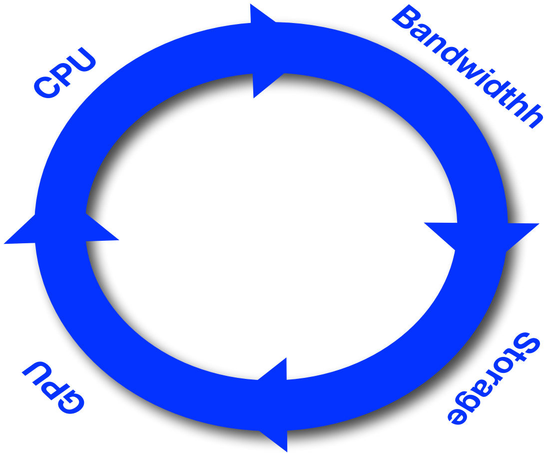
- Large matrix of systematic datasets
  - photons-level production
  - systematic variations - ice model, DOM acceptance, hole ice
- Multiple generators
  - Low-Energy production - GENIE
  - Most other analyses - NeutrinoGenerator
  - High-Energy Sterile Neutrino - LeptonInjector (final state neutrino)
  - 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

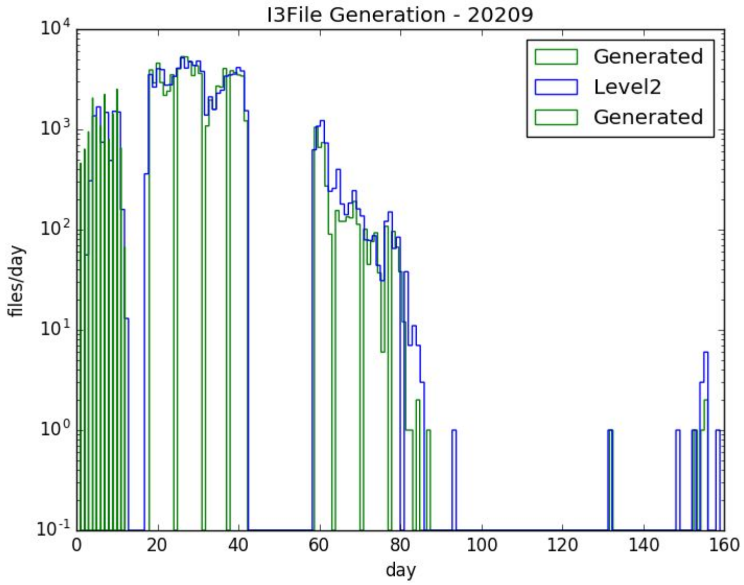
Issues with production dynamics

- 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

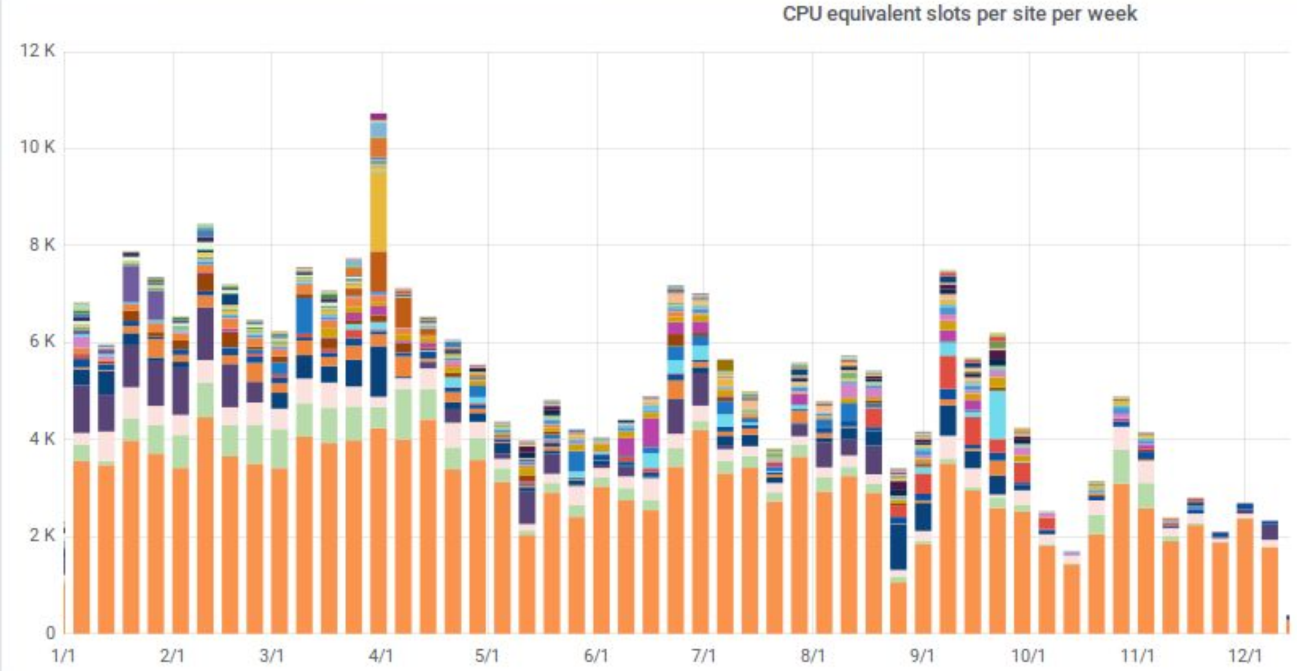
- Warn of potential bias due to failures
- Investigate further
- Roll fixes into the next release



Simulation Production — Dynamics

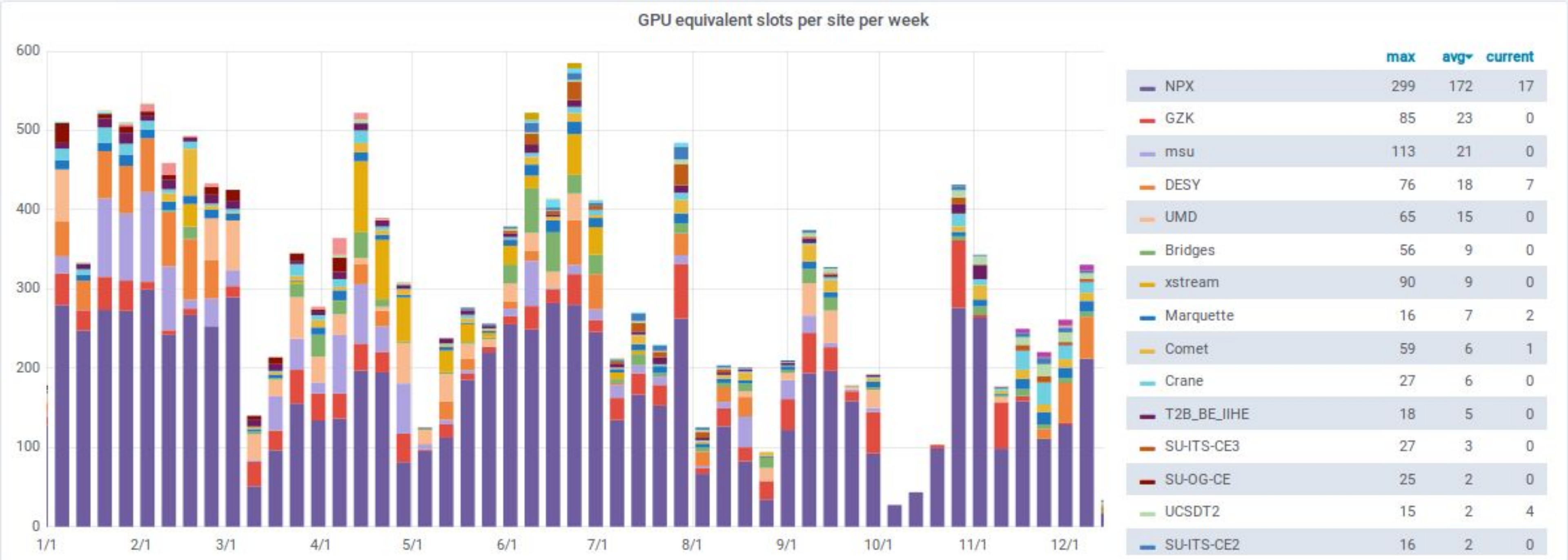
1. Resource checklist: estimate disk, CPU, GPU, running time
2. Short initial configuration and test period
3. Long, steady, and stable production run, with continuous monitoring
4. Short post-production validation period
5. Publish dataset - Send email to collaboration announcing dataset is ready for analysis

All Sources - CPU Usage 2018 and Site

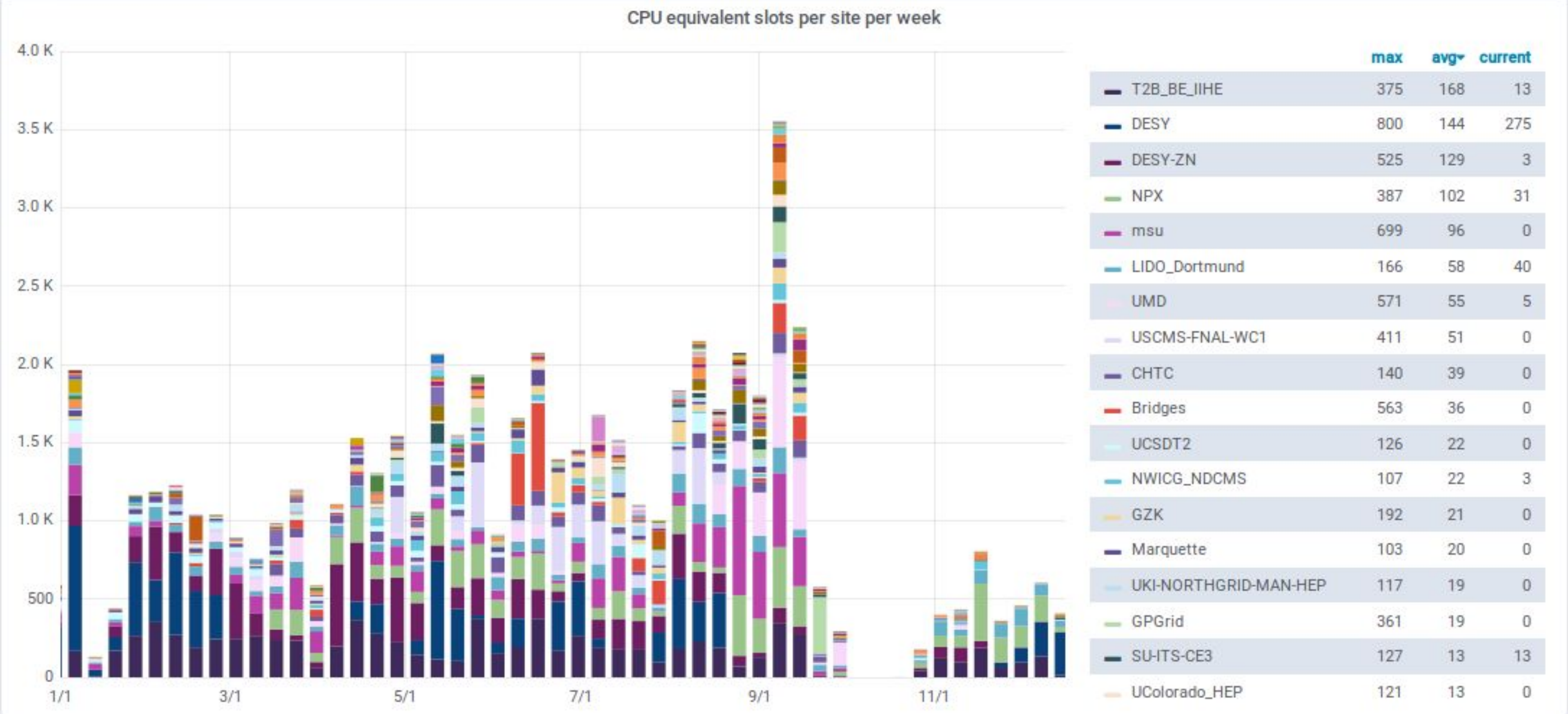


	max	avg	current
NPX	4.463 K	2.951 K	282
DESY-ZN	1.040 K	324	1
T2B_BE_IIHE	639	291	2
DESY	1.088 K	251	91
msu	1.044 K	189	0
UKI-NORTHGRID-MAN-HEP	420	107	0
LIDO_Dortmund	219	92	10
UMD	681	85	0
GPGrid	996	71	0
USCMS-FNAL-WC1	739	71	0
RWTH-Aachen	376	66	0
Bridges	596	59	0
CHTC	208	58	0
BEgrid-ULB-VUB	204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

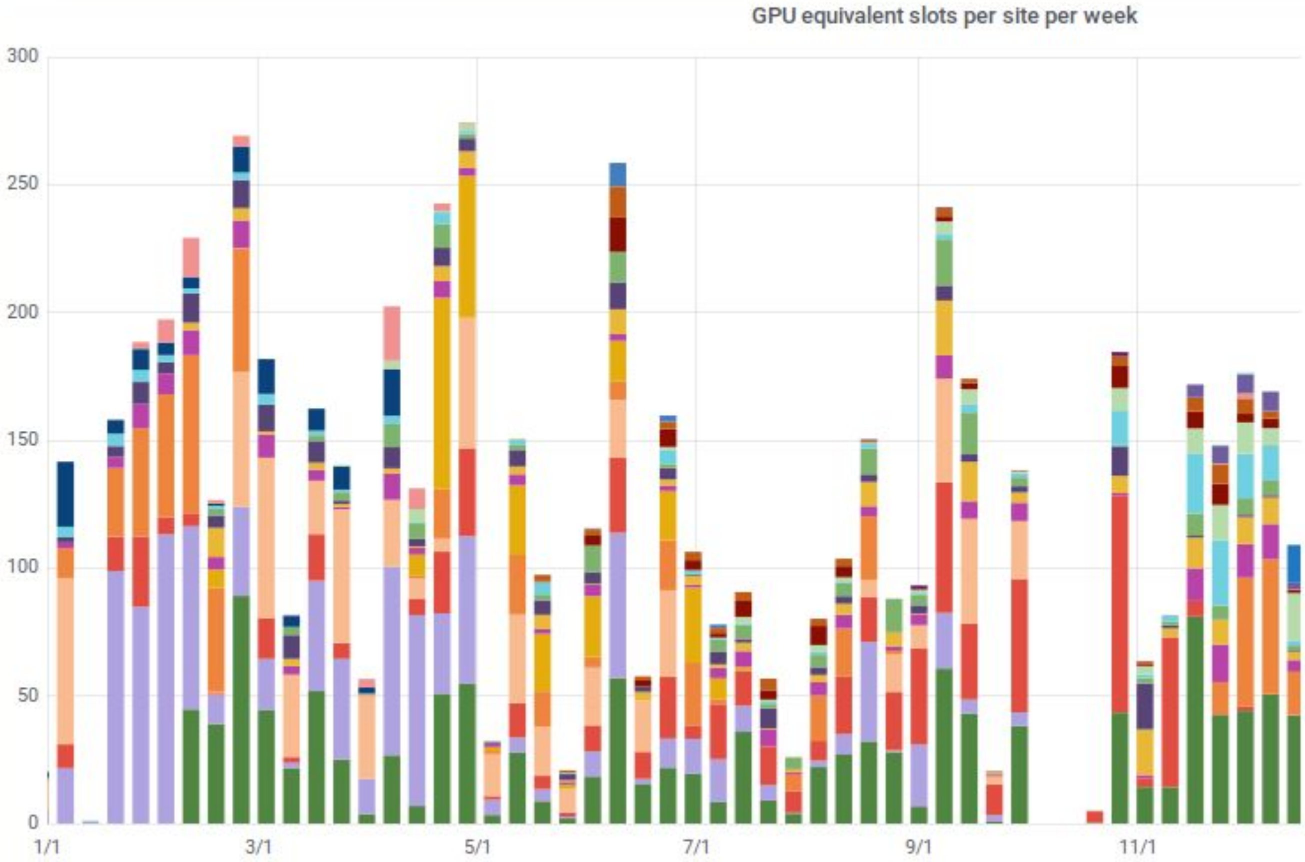
All Sources - GPU Usage 2018 and Site



Simulation Production - CPU Usage 2018 and Site

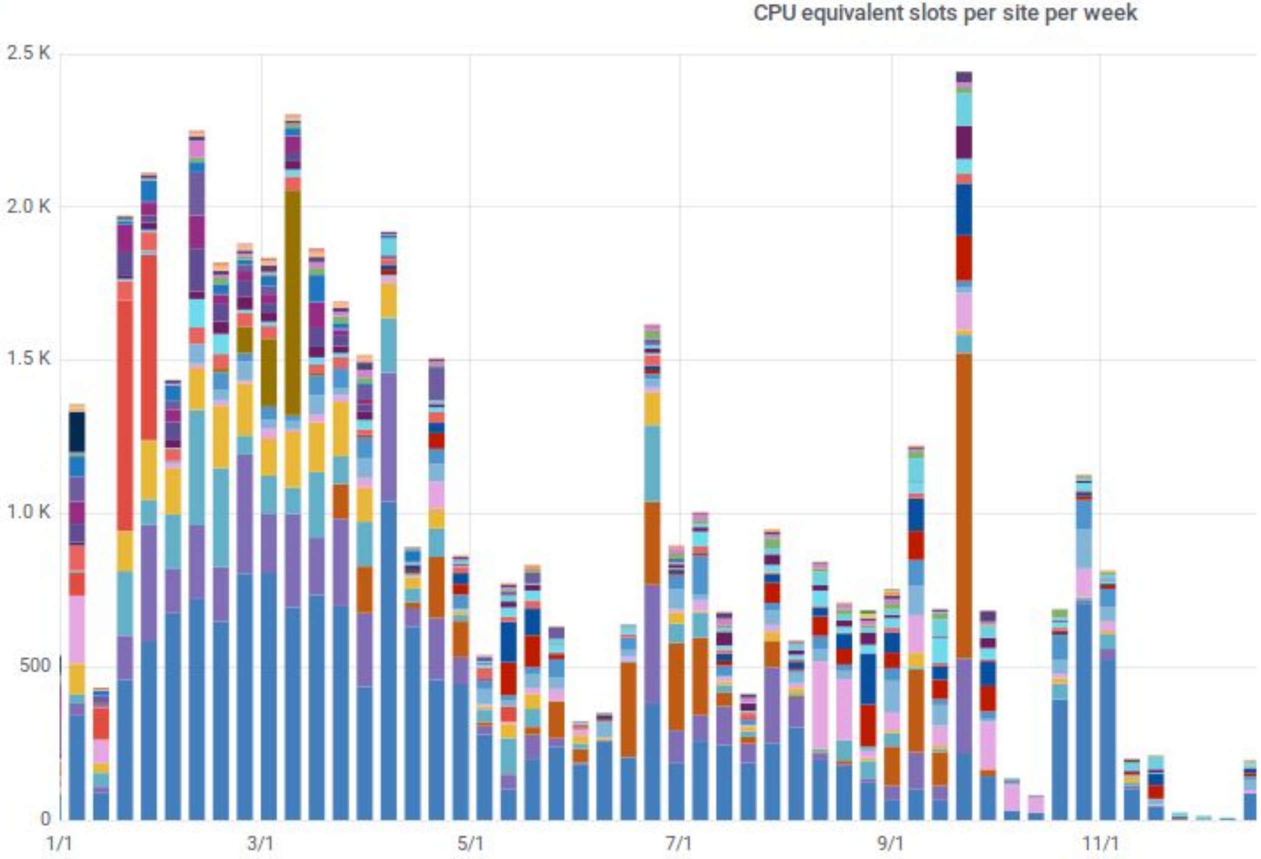


Simulation Production - GPU Usage 2018 and Site



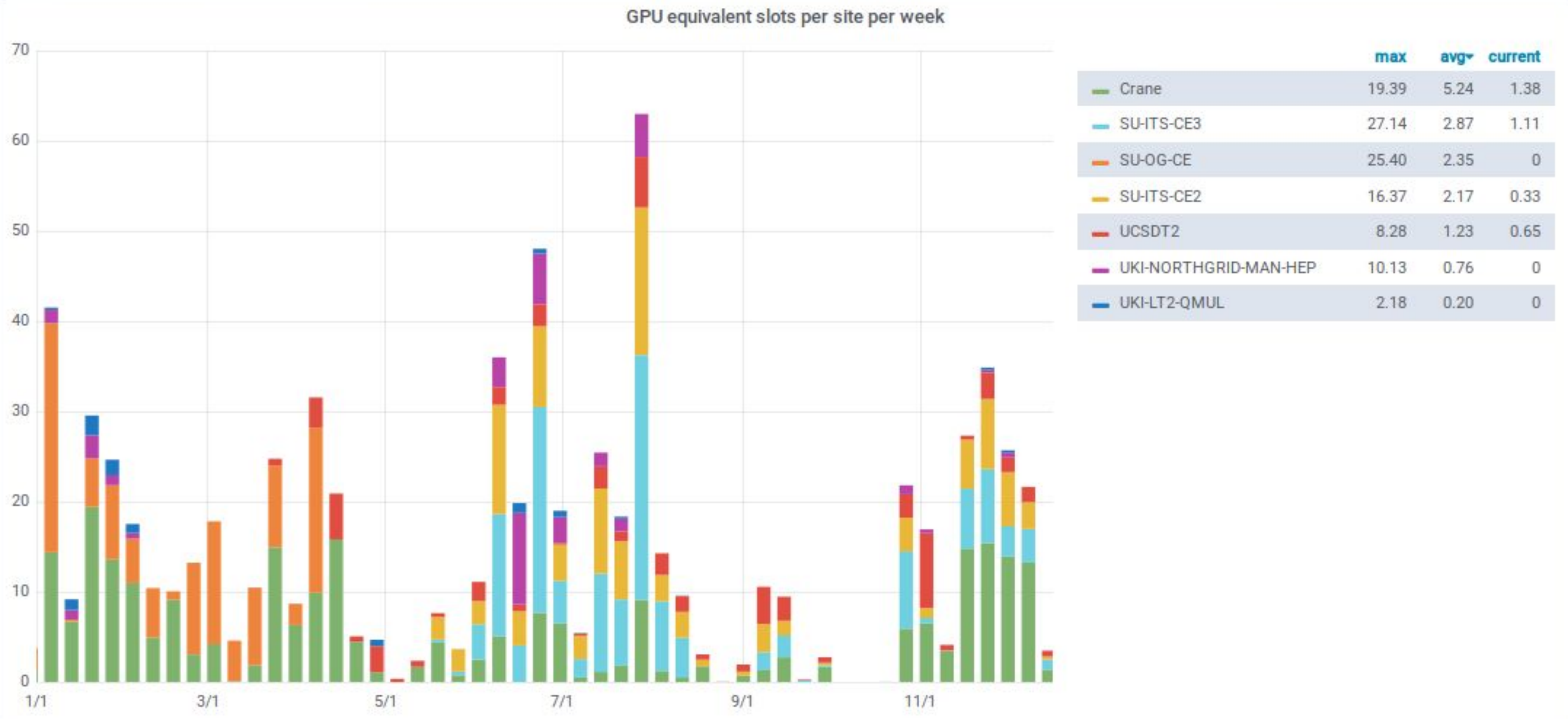
	max	avg	current
NPX	89	25	42
msu	113	21	0
GZK	85	15	0
UMD	65	14	0
DESY	62	12	17
xstream	75	6	0
Marquette	15	5	5
Comet	21	4	3
T2B_BE_IIHE	18	4	0
Bridges	18	4	2
Crane	25	3	2
SU-OG-CE	25	2	0
UCSDT2	19	2	19
SU-ITS-CE3	14	2	1
SU-ITS-CE2	12	1	0
aachen	21	1	0
NBI	8	1	2
MSU	15	0	15

GlideinWMS - CPU Usage 2018 and Site



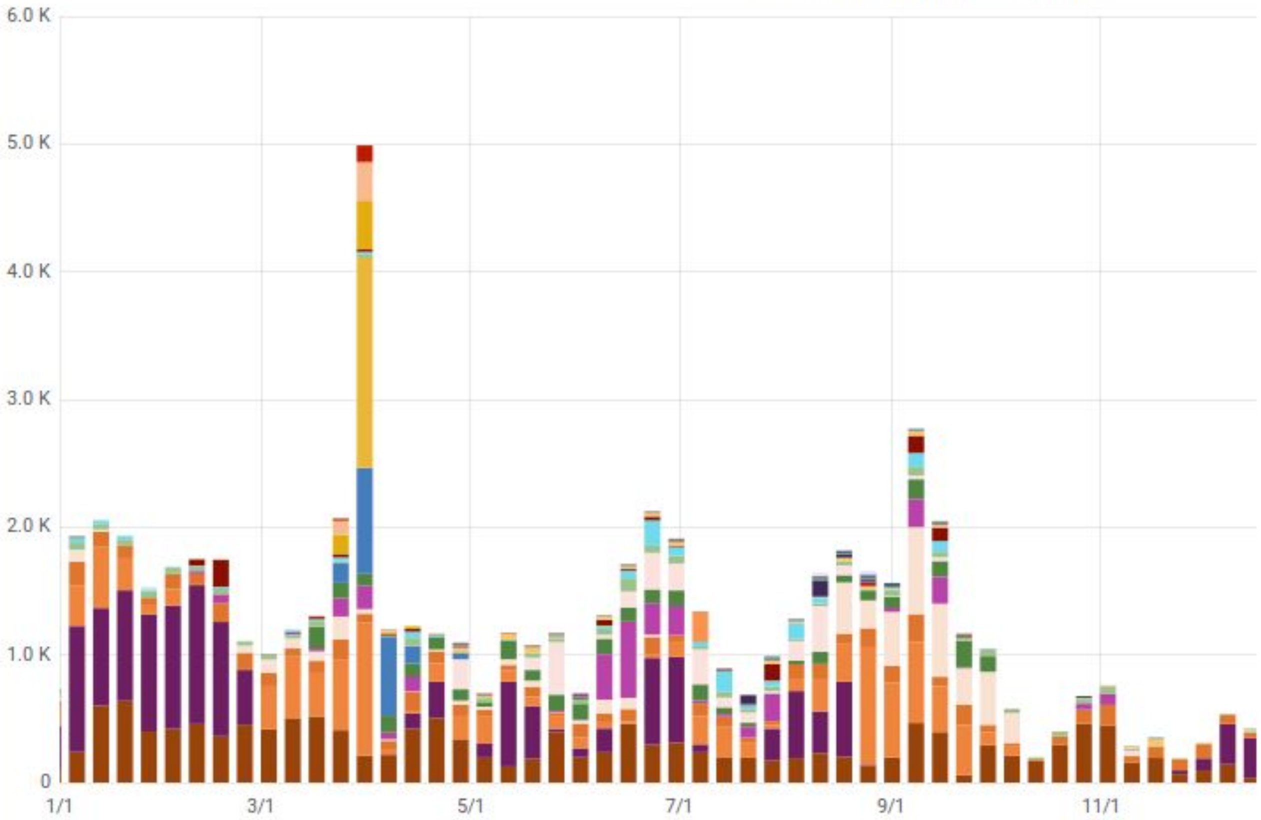
	max	avg	current
DESY-ZN	1.040 K	326	89
UKI-NORTHGRID-MAN-HEP	420	107	0
GPGrid	996	71	0
RWTH-Aachen	376	66	0
BEgrid-ULB-VUB	204	51	0
Nebraska	282	42	9
CA-MCGILL-CLUMEQ-T2	753	31	0
NWICG_NDCMS	127	31	35
UColorado_HEP	129	28	10
SU-ITS-CE2	148	25	11
SU-ITS-CE3	170	24	13
USCMS-FNAL-WC1	735	21	0
UCSDT2	79	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



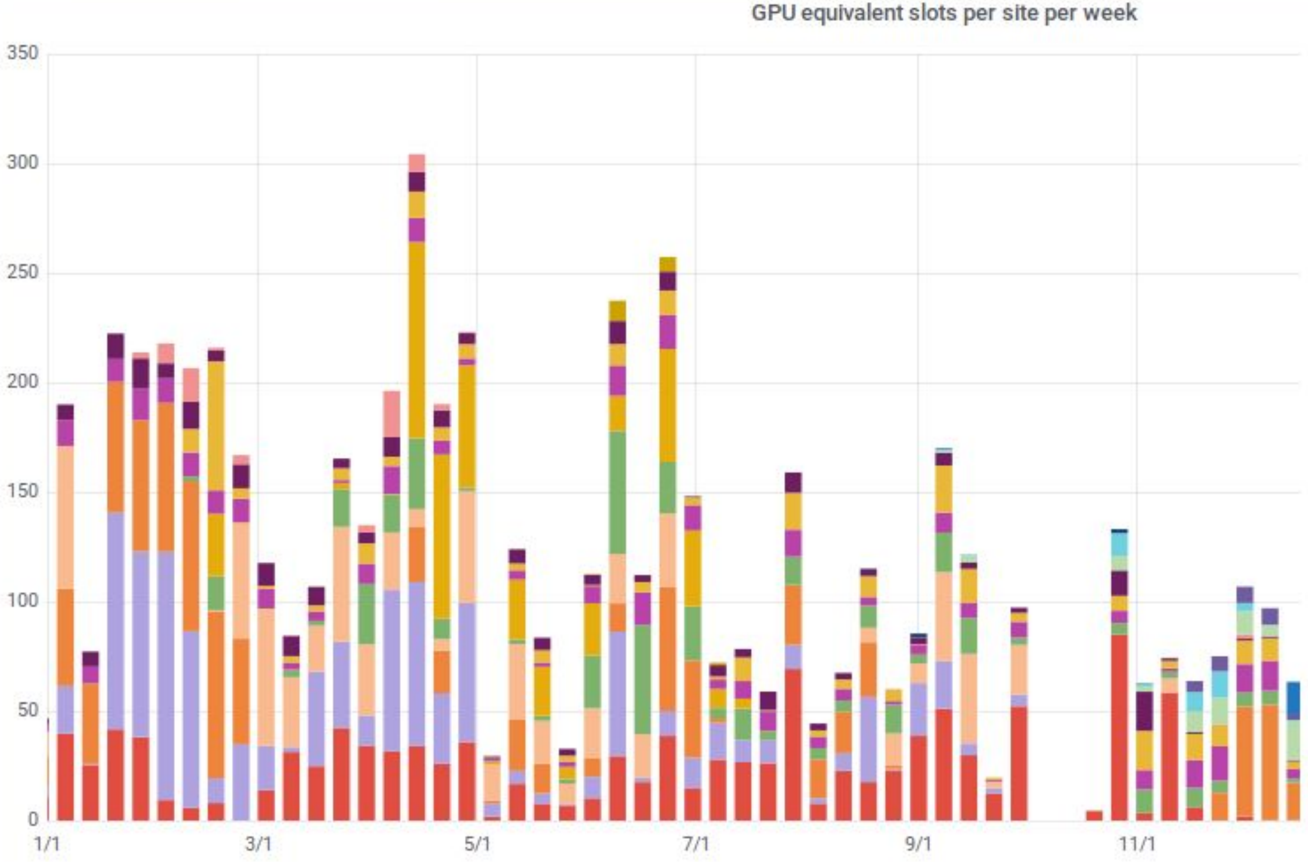
Pyglidein - CPU Usage 2018 and Site

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



	max	avg	current
GZK	85	23	0
msu	113	21	0
DESY	76	18	17
UMD	65	15	0
Bridges	56	9	2
xstream	90	9	0
Marquette	16	7	5
Comet	59	7	3
T2B_BE_IIHE	18	5	0
aachen	21	1	0
UCSDT2	18	1	18
Crane	12	1	1
NBI	8	1	2
PSU	9	0	0
MSU	15	0	15
Syracuse	2	0	0
illume-new	0	0	0
other	0	0	0

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
- Leverage national-level resources more, e.g. TACC's upcoming Frontera supercomputer with GPUs

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

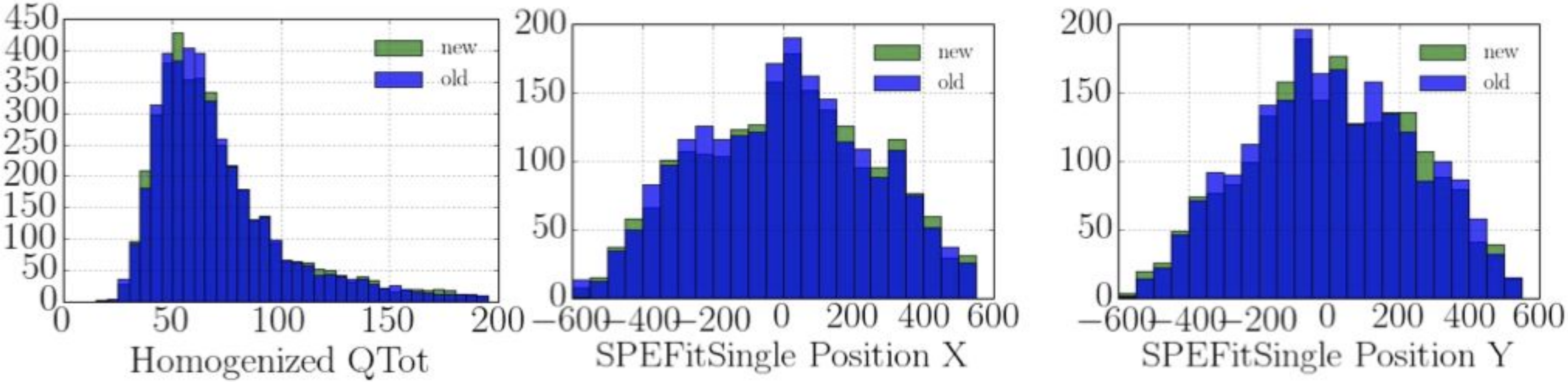
- Management:
  - Gonzalo Merino returned to PIC as Deputy Director in Aug 2018
  - Benedikt Riedel took over as Computing Manager as of Dec 2018
  - David Schultz now manages the Production Software group
- Staff:
  - Heath Skarlupka (Operations Engineer) left for industry in March 2018 - Hiring replacement
  - 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

Pass2: Level2 validation

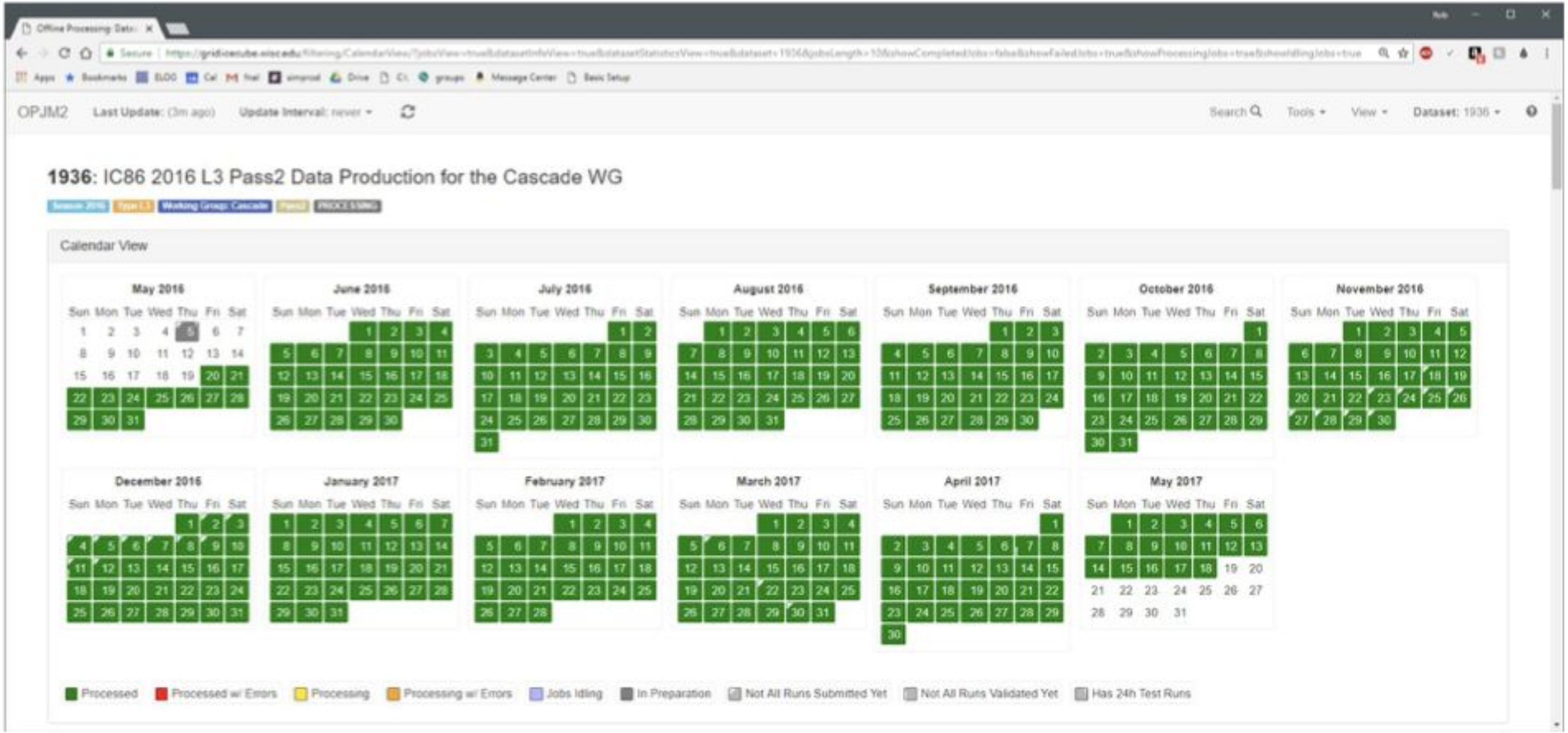
Fully processed 9 runs of interest for the HESE analysis:

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)



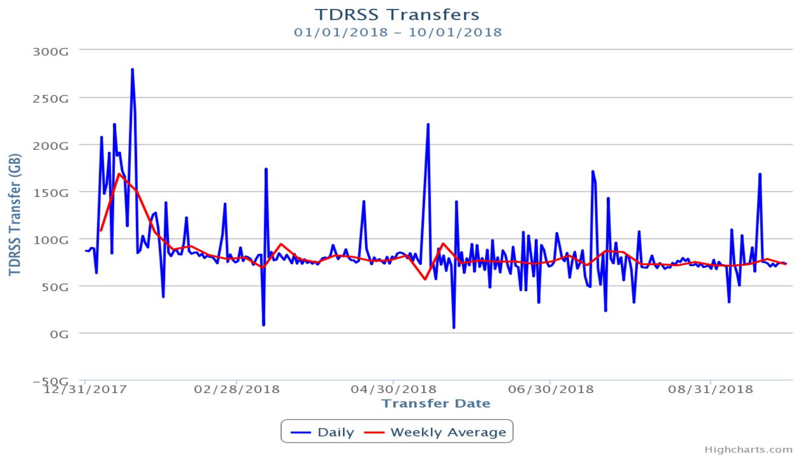
Pass2 L3 production example: Cascade filter



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)