Fermilab | U.S. DEPARTMENT OF ENERGY | Office of Science

HEPCloud Operations At Fermilab—The First Five Years

Steve Timm, Nick Smith

CHEP 2024

Why HEPCloud Was Built:

- In 2014 Fermilab Computing Management began “Virtual Facility Project”
- Expand to commercial clouds and HPC Centers as well as Grid computing sites.
- Strong support from experimental stakeholders, developers, funding agencies.
- "HEPCloud" name given by Jim Siegrist, then-head of DOE OHEP
- Some proof-of-principle tests with Amazon Web Services and various private clouds had already been done, as well as some basic tests at NERSC.
- In 2016 two large-scale demonstrators were done:
  - 60K cores on Amazon Web Services
  - 160K cores on Google Compute Engine.
- Mandate from P. Spentzouris—Build system that could
  - Do all the activities of data-intensive high-throughput computing
  - Run on grid, cloud or HPC
  - Have Fermilab facility be in control of where they would run.
- Building on successful GlideinWMS development.

HEPCloud Decision Engine Development

- From demonstrators, command and control emerged as biggest challenge
  - Who was authorized to run on cloud/HPC
  - How much were they allowed to spend
  - How to detect unauthorized use
  - How to optimize the selection of resources for most economical running
  - How to protect all the valuable secrets needed to run jobs at these sites.
- This led to the development of the Decision Engine
  - All resources ranked by a “Figure of Merit” based on price/performance with units of money—lower is better.
  - Logic Engine for rule-based decision on which resources are allowed and which are disabled
    - Based on budget for cloud, or allocation for HPC resources.
  - 2 years, several full-time developers.

First Production Deployment—March 2019

- Initial configuration queried job queues and could send jobs to any of AWS, Google, NERSC, OSG and local simultaneously
- USCMS was first big user, they had been running jobs at NERSC on our integration instance for more than a year.
- Stakeholders wanted control of where their jobs would go, and when.
- Configuration that was quickly agreed to:
  - HEPCloud would only take jobs from CMS Global that explicitly requested cloud or HPC sites
  - Other stakeholders (NOvA, Muon g-2, Mu2e, and DUNE) did the same on the shared “FIFE batch pool”
  - DUNE now has its own Global Pool with its own decision engine.

Expansion to XSEDE (ACCESS) Resources

- USCMS received allocations on XSEDE (later ACCESS) resources under PI's Ken Bloom and later Tulika Bose.
  - These are National Science Foundation funded resources
  - Pittsburgh Supercomputing Center (PSC), first Bridges, now Bridges2
  - Texas Advanced Computing Center, first Stampede2, soon Stampede3
  - San Diego Supercomputing Center, first Comet, now Expanse
  - Purdue Rosen Center for Advanced Computing, Anvil
- Open Science Grid developed concept of "Hosted CE"
  - An HTCondor-based compute element which we could submit to with normal grid authentication methods.
  - Used HTCondor "batch" universe to ssh into the remote batch system.
- TACC Frontera not an XSEDE/ACCESS resource but also accessed via HOSTED-CE
- Also used 120 VMs on JetStream cloud-based resource.

Batch submission protocols

- Amazon Web Services uses "EC2" based protocol.
- Google Compute Engine (Platform) uses google API
- NERSC uses HTCondor "batch universe" (originally known as BOSCO but now part of HTCondor).
  - ssh into the remote batch host and run a set of shell scripts which produces a submit file for remote batch system (SLURM)
  - ssh periodically and poll the job ids to see if they have completed
  - Also forward updated credentials to the job as needed.
  - Once NERSC introduced multi-factor authentication we used what is known as a "ssh proxy" to have an object similar to an ssh key except with a 30 day expiry date and restricted hosts from which it could be used.

Inference Server testing

- The main recent use of commercial clouds in HEPCloud in recent years is for testing the Triton inference server
- Special purpose hardware, either small cluster of GPU, FPGA, or tensor processors, to which O(1000) running jobs call out across the network to do the short inference phase of the processing.
- In one DUNE test with worker node jobs running at Fermilab and the Triton server at Google's location in Iowa, we managed to use all then-available off-site bandwidth of Fermilab (100Gbps)
- Now beginning work of co-scheduling Triton inference servers on GPU and CPU jobs at NERSC.
- Some use of SuperFacility API now, planning towards IRI era

Data transfer

- CMS jobs are dominated by step-chain MC which needs pre-mix pileup files.
  - In early days we transferred a set of pileup to NERSC
  - Nowadays read via xrootd straight from Fermilab.
  - Output files staged straight back to Fermilab.
- DUNE and other neutrino experiments
  - Stage local flux files, databases, shower libraries on Community File System
  - Output to Perlmutter Scratch
  - Transport of output back with FTS3 / Rucio.
- Globus file transfer to ALCF via NERSC has been done
  - as proof of principle, not yet in regular use.

Code distribution

- Originally tried to bundle all versions of CMS code into custom “shifter” image at NERSC ~600GB.
  - Jobs failed to start due to timeout.
- NERSC then agreed first to rsynced version of CVMFS
- Now NERSC runs native CVMFS.
- On other sites, use “cvmfsexec” which fetches a cvmfs tree into user space and bind-mounts it into your singularity container.
- Problems at first with some sites not supporting user-defined namespaces but they all do now.

Provisioning challenges

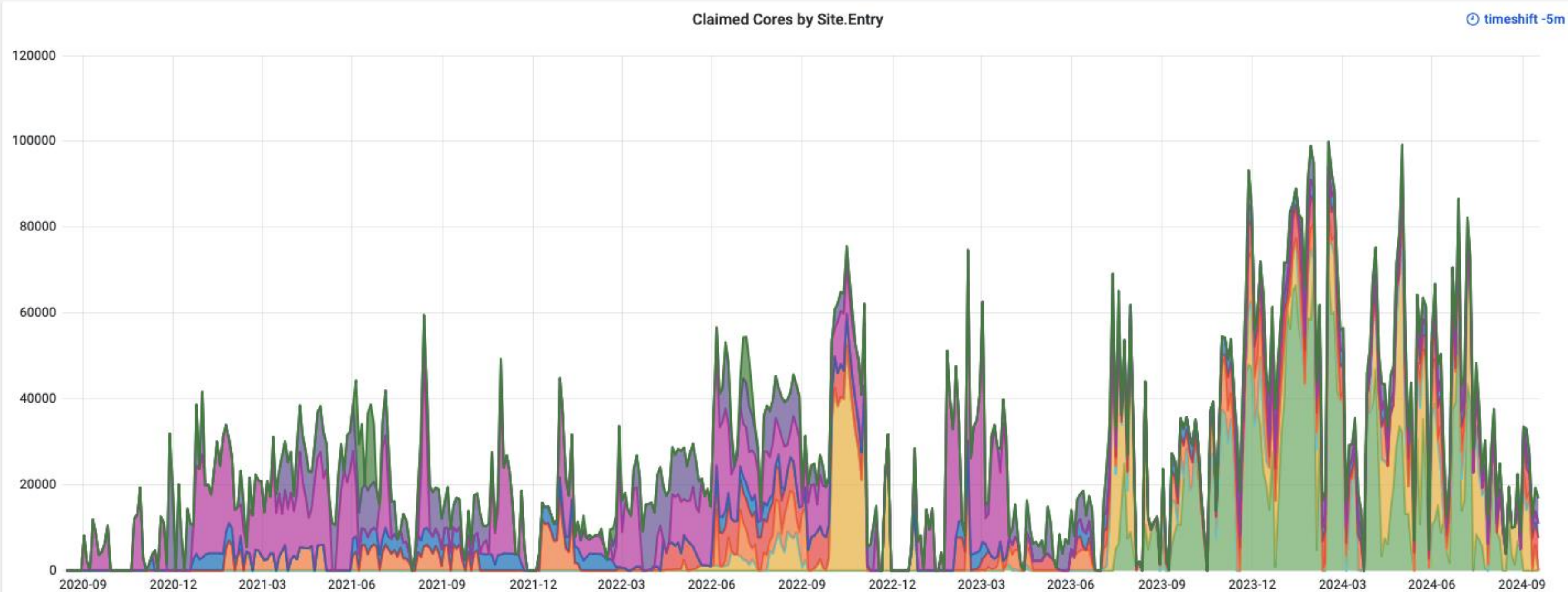
- Began running on NERSC just as “Cori” was deployed.
  - Originally went for the slower “Knights Landing” nodes because there were more of them and they were less requested
  - Shifted to “Haswell” nodes for better performance
  - Near the end of Cori run had to request 100 nodes at a time.
- Started on Perlmutter in early 2023 with 5 nodes at a time
  - Currently asking for 50 nodes at a time.
  - 24 hour pilot length, 48 hours just recently became allowed.
- TACC Frontera we are throttled to 4x28-node jobs due to limitations of their internal site firewalls.
- Other ACCESS nodes we run single-node jobs.

I/O Scalability Challenges

- HEPCloud workflows big challenge for shared file systems
  - Large number of inodes

    (Simultaneous untar of 250 Madgraph tarballs on same node)
  - Large number of IOPS
- Most shared file systems don’t handle that well
- At NERSC use the ”Node Cache” to export a whole block device to each running job from Perlmutter scratch
- Most other places use the limited local disk on each worker.
- Challenging transition from in-cluster HPC network to the internet
  - In early days saw lots of dropped connections and partially transferred files. That’s rare now.

USCMS Resource Usage 2020-present

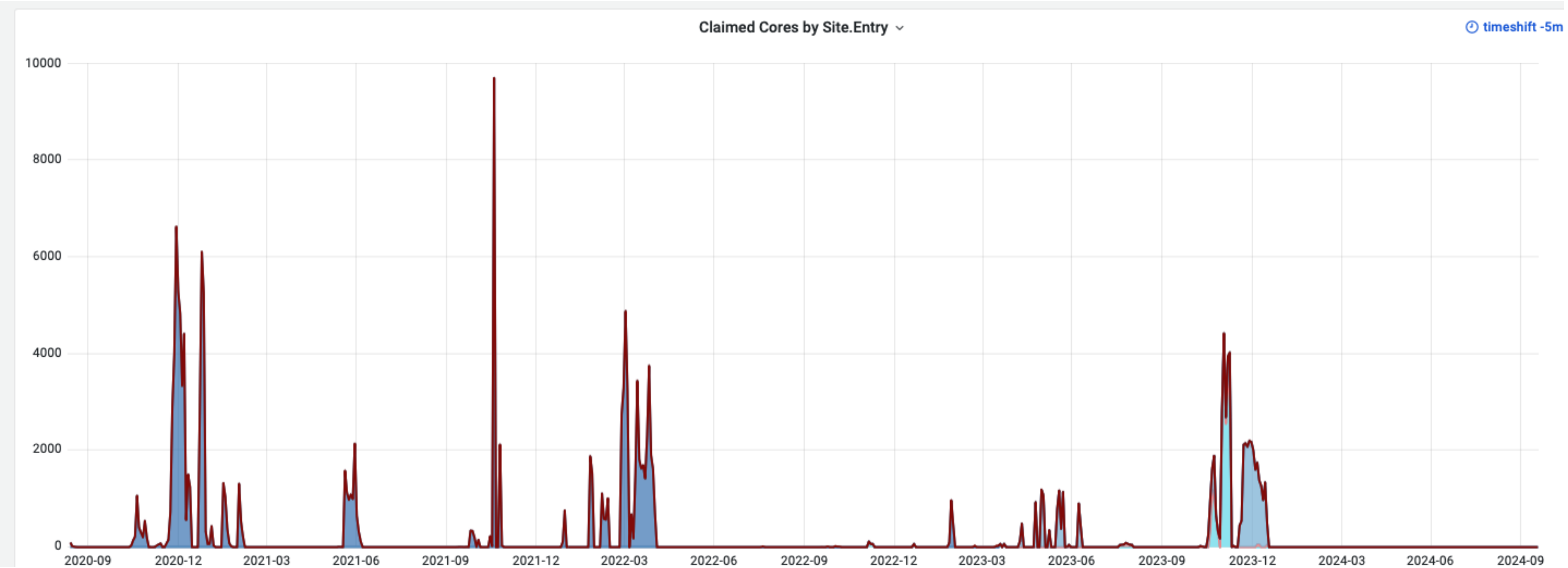


	min	max	avg	current
NERSC-Perlmutter-CPU.CMSHTPC_T3_US_NERSC_Perlmutter_CPU_RHEL8	0	70454	5825	0
NERSC-Perlmutter-CPU.CMSHTPC_T3_US_NERSC_Perlmutter_CPU_SL7	0	52263	4148	0
PSC-Bridges2.CMSHTPC_T3_US_Bridges2	0	22743	2311	0
Purdue-Anvil.CMSHTPC_T3_US_PURDUE-Anvil	0	7904	2015	7551
SDSC-Expanse.CMSHTPC_T3_US_SDSC-Expanse	0	4065	1781	3412
T3_US_NERSC.CMSHTPC_T3_US_NERSC_Cori_SL7	0	55912	6126	0
TACC-Frontera.CMSHTPC_T3_US_TACC_FRONTERA	0	31930	4186	5880
TACC-Stampede2.CMSHTPC_T3_US_TACC-Stampede2	0	18804	337	0

Average of 26729 cores DC over last 4 years

Peaks of >100K cores! 2024 ERCAP=1.13M NERSC-HOURS

DUNE/FIFE usage 2020 to present



Usage shown from NOvA, Mu2e, and DUNE experiments
Not shown is direct SLURM submission from other smaller projects.

Current integration/deployment challenges:

HEPCloud

- Strong interest from all stakeholders in leadership facilities at Argonne and Oak Ridge
  - These machines have no inbound or outbound network access allowed from their workers.
  - Not compatible with standard GlideinWMS pilot model.
- Several demos done thus far (both at OLCF and ALCF) using “split-starter” technique in HTCondor, using shared file system to communicate with a node on the edge.
- In conversation with both labs for best way to move into the Integrated Research Infrastructure (IRI) era.
- DUNE has part of its reconstruction “MLreco” which currently must be run on GPUs, which could easily take more than 100K GPU node-hours per year.
  - Key use case for mostly-GPU supercomputers.

Conclusions

- HEPCloud is a mature provisioning system which provides access to compute resources similar to the size of the US CMS Tier-1 facility at Fermilab.
- Over last six years we have been able to grow the capacity thanks to increasing allocations from DOE OHEP and NSF ACCESS program
- Important access point to heterogenous or experimental compute resources, such as GPU, FPGA, TPU, Quantum.
- Continue to onboard more users and more resources.
  - Submitting to leadership-class facilities
  - Learning how to provision native MPI workflows.

Thanks to External Organizations

- NERSC, ALCF, OLCF
- SDSC, TACC, PSC
- HTCondor developers
- OSG HTCondor-CE maintainers
- KISTI
- IIT Department of Computer Science
- DOE OMNI intern program
- INFN internship program
- DOE SULI/SIST internship program

Thanks to HEPCloud Team Members:

- Project Sponsors: P. Spentzouris, J. Amundson, S. Fuess, B. Holzman, A. Norman
- Project Managers: Rob Kennedy, Tanya Levshina, Eileen Berman, Krista Majewski, Gabriele Garzoglio, Parag Mhashilkar
- Architecture Consultants: Jim Kowalkowski, Marc Paterno
- Technical Leads: Anthony Tiradani, Marco Mambelli
- Development Leads: Dmitry Litvintsev, Kyle Knoepfel
- Developers: A. Moibenko, D. Dagenhart, Q. Lu, B. Coimbra, V. di Benedetto, S. Bhat, L. Goodenough, P. Riehecky, D. Box, N. Urs.
- CMS Team: D. Hufnagel, A. Mohapatra, I. Fisk, N. Magini, D. Mason, D. Dykstra, E. Vaandering.

HEPCloud Team Members Continued:

- CMS R+D: Hyunwoo Kim, Maria Acosta
- DUNE and FIFE Experimental Liaisons
  - Ken Herner, Andrew Norman, Alexander Booth, Alex Himmel, Rob Kutschke, Ray Culbertson, Eremey Valetov.
- Security: Mine Altunay
- HEPCloud Operations and Integration Testing:
  - Steven Timm (lead 2015-2024), Vito di Benedetto(lead 2024-onwards), Nicholas Peregonow, Arshad Ahmad, Merina Albert, Farrukh Khan, Joe Boyd, Gerard Bernabeu, Neha Sharma
- Current HEPCloud management: S. Lammel, facility head, M. Mambelli, project lead, V. DiBenedetto, operations lead.
- Students, interns, and contractors—Many!

BACKUP SLIDES

XSEDE/Access allocations

YEAR	PSC	SDSC	TACC	Purdue
2019	15M(B)	10M(C)	2.4M (S2)	
2020	10M(B)+1.9M(B2)	8.7M(C)+4M(E)	600K (S2)	
2021	17M(B2)	17M(E)	240K(S2)	7M
2022	14.7M(B2)	13.2M(E)	75K(S2)	18M
2023	30M(B2)	30M(E)		33M
2024	23M(B2)	23M(E)	1M(S3)	23M