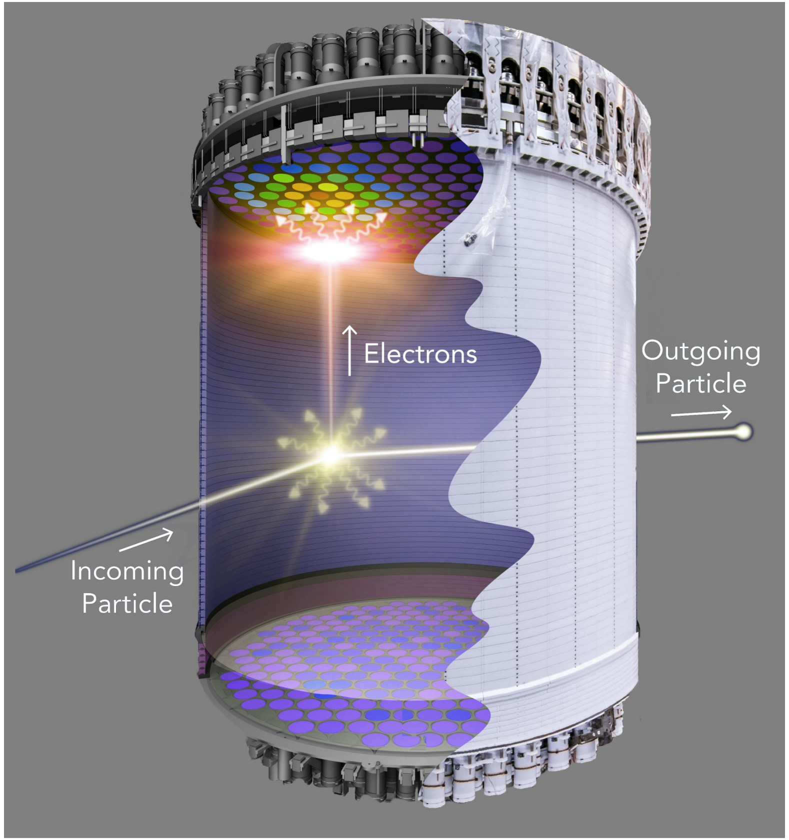
LUX-ZEPLIN: data-intensive search for Dark Matter

Maria Elena Monzani
(SLAC/KIPAC)

HEP-CCE AHM
June 22 2024

The LUX-ZEPLIN (LZ) Dark Matter Experiment



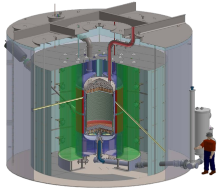
LZ is a 10-ton Liquid Xenon TPC

- Located underground at SURF, South Dakota
- Initial science run data in winter/spring 2022
- Set world-record WIMP sensitivity in July 2022
- (5 weeks turnaround between run and results)!
- LZ data is stored and processed at NERSC

Data Throughput (order of magnitude)

- Fermi-LAT (>2008): 0.3 PB/year
- LZ (2021-2028+): 3.5 PB/year, 7+ years
- ATLAS (>2010): 3.2 PB/year (raw)
- PS: extreme “needle in a haystack” problem!

Construction and Data Taking Timeline

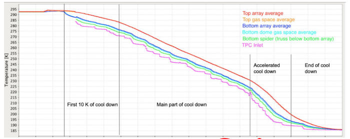


CD3 and TDR

TPC Complete

Underground

Cold GXe



Science Running!

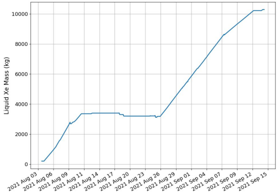
MDC1 | MDC2 | MDC3

2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023

Cryostat arrives

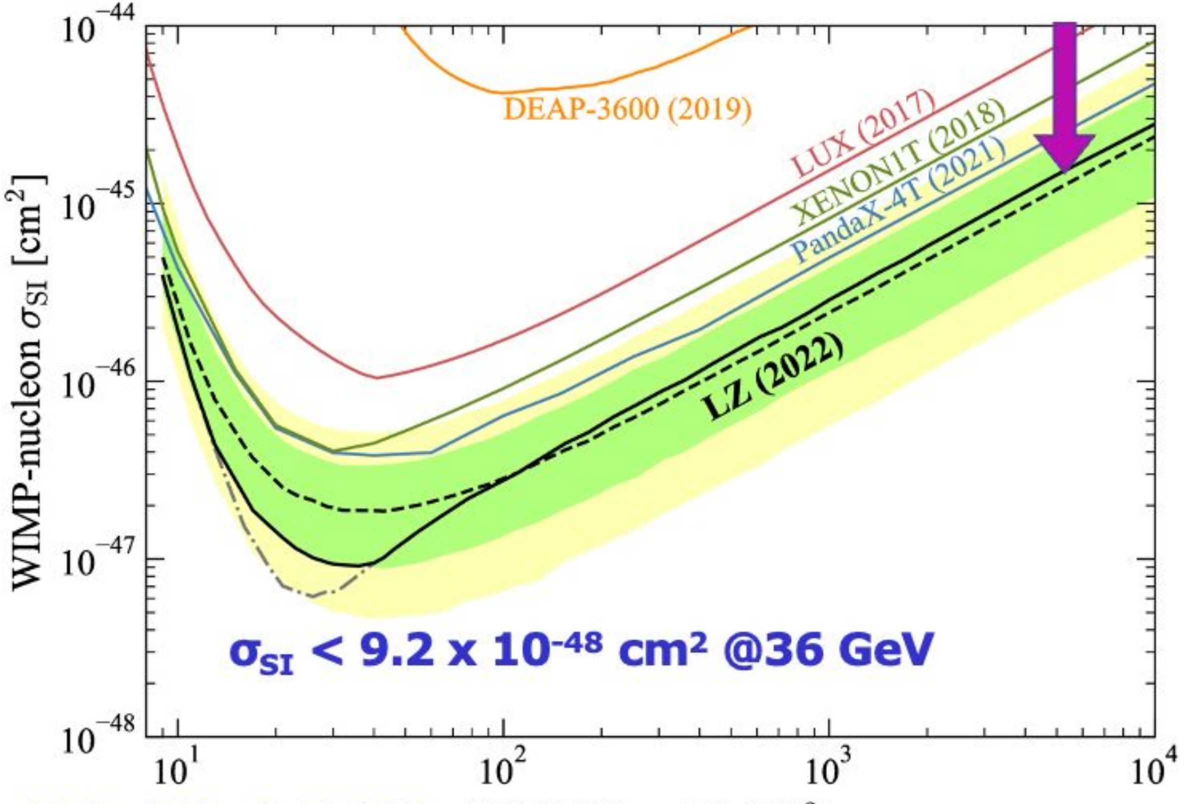
Sealed Up

LXe fill



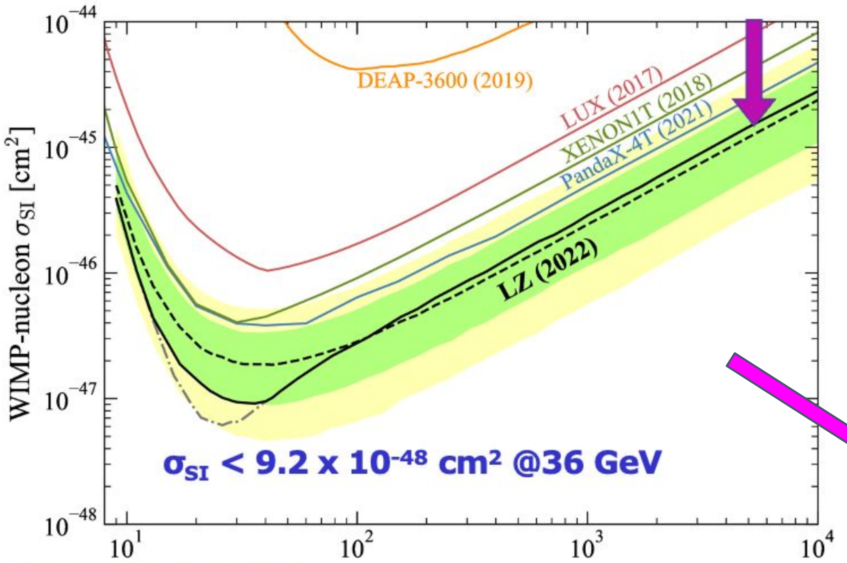
First results!

World-leading WIMP sensitivity (July 10, 2022)



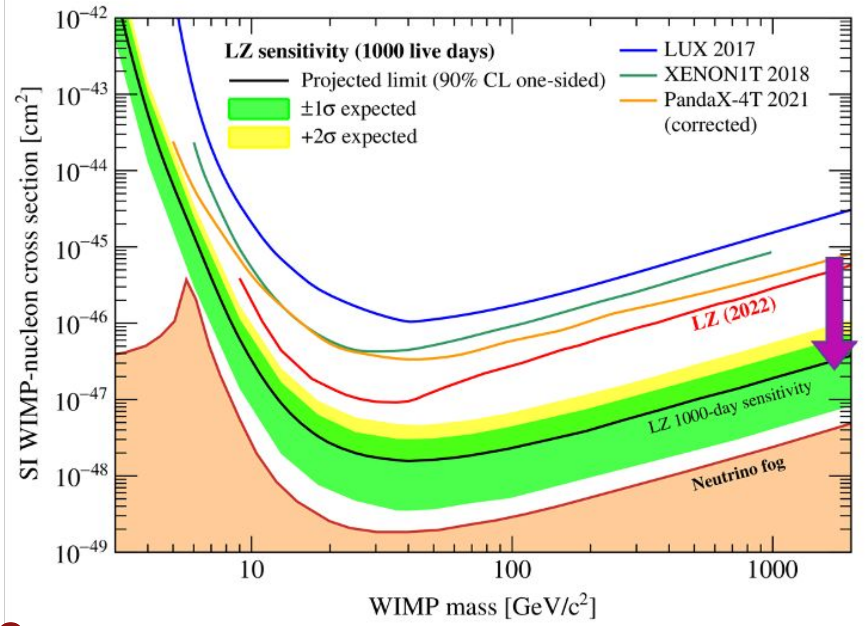
PRL 131, 041002

Extension of the LZ program reviewed/approved last month



PRL 131, 041002

x15 more data in the full run!



New plan! Data taking through CY2027
Decommissioning & data analysis in 2028

LZ: Offline Computing and Software

Data is staged at SURF and transferred to the remote data centers

- Fully redundant data center design (each site can run data processing and simulation production… and store a complete copy of all the data)!
- Data rate: ~3 PB/year, including raw, reconstructed, calibrations, etc.
- All detector data are processed automatically 24/7 at the USDC.
- Data can be reprocessed on-demand based on calibrations and analysis.
- Reconstructed and simulated data is then made available to all analyzers.

Temporary Storage:

SANFORD LAB HOMESTAKE

US Data Center (USDC):

- Prompt Processing
- Long-term Archiving
- Supercomputers!

NERSC

UK Data Center (UKDC):

- Data Reprocessing
- Sims Production
- Distributed CPUs!

GridPP
UK Computing for Particle Physics

Reconstructed & simulated data can be analyzed at either data center

- NERSC and GridPP have diverging CPU architectures. All LZ software & analysis tools can run seamlessly on either architecture.
- System choice is based on user preference, but several team members have become proficient at both supercomputers and distributed computing.

Offline Requirements and Design Principles

Store all raw & reconstructed data from LZ

- 2 “live” copies of all raw & reconstructed data at NERSC and UKDC
- 1 “tape” archive of all raw data at NERSC before bias mitigation
- At least 1 backup of all versions of reconstructed data at NERSC

Process detector data early and often

- Automatic prompt-processing at USDC upon data reception
- Redundant capabilities to reprocess/simulate multiple times based on calibration/analysis results (rerun 1 year of data in 1 month)

Time is of the essence! Rapid (<1 day) turnaround

- Very limited computing resources are available at SURF (RAID array for storage and “first look” online quality monitoring tool)
- Full-scale detector health assessment happens at NERSC. Quasi-real-time analysis feedback during commissioning

US Data Center (USDC):

- Prompt Processing
- Long-term Archiving
- Supercomputers!

NERSC

UK Data Center (UKDC):

- Data Reprocessing
- Sims Production
- Distributed CPUs!

GridPP
UK Computing for Particle Physics

The LZ Data Centers

UKDC Overview

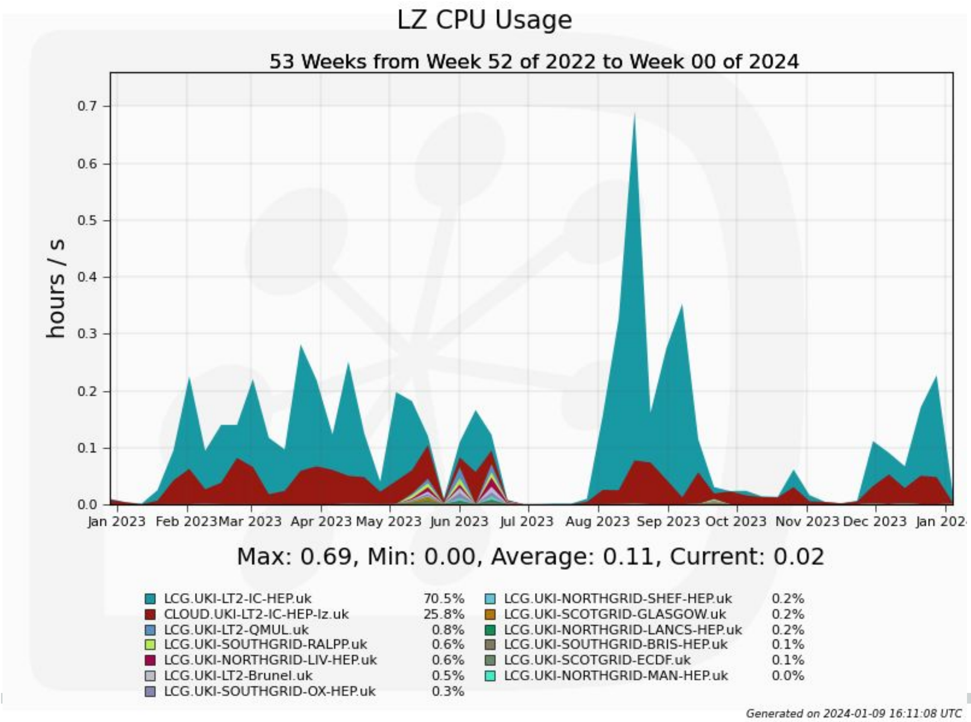
- **Follows LHC distributed Grid computing model**
  - Based on GridPP and IRIS resources (~70k job slots, >50PB)
  - Hardware buy in → access this pool of distributed resources
- **Data hosted by Imperial College London (ICL)**
  - Housed in the VIRTUS Data Centre in Slough
  - 7.2 PB currently available → ramps up as data collected
  - +3 PB agreed for 2025
- **CPU distributed across ICL and other GridPP sites**
  - Expect ~500 slots average, but with opportunistic use of more (achieved 2000-6000 in productions)
  - No central login node(s) for collaboration users (relying in institutional clusters at GridPP member institutions).

UKDC Role during LZ Operations

- **Host a complete copy of raw and processed data**
- **Official processing of data (asynchronous/reprocessing)**
  - Keep-up processing capability planned, not implemented yet
- **CPU, storage, staffing for MC production campaigns**
- **User analysis tools and support (75 active users):**
  - CPU and storage; Job submission tools & interface
  - Integration with core tools: ALPACA, Stats, LZLAMA, etc.
- **+ a number of hosted services:**
  - Data Movement Endpoint; Data catalog replica; Offline event viewer; UK instance of PREM (data quality tool)

UKDC: CPU usage in 2023

- 3.7 mln CPU-hrs in 2023: 92% production; 8% user analysis
- Mixture of of official sims production and LZAP reprocessing



- 3.7 mln CPU-hrs on GridPP: ~30k node hours on Perlmutter
- 2023 allocation at NERSC: used ~105k CPU raw node hours
- At NERSC: 60% user analysis/jobs, 40% prompt processing

USDC at NERSC Overview

Compute Platform	Node types	# Available	Comments	Doc.
Perlmutter CPU	AMD Milan Nodes 512 GB DRAM/node	3072 nodes 128 cores/node	"Standard" CPUs with good memory footprint. Prompt processing, simulations and user analysis are performed on this system.	link
Perlmutter GPU	AMD Milan Nodes 256 GB DRAM/node	1792 nodes 64 CPU + 4 GPU	Can be used as a "standard" CPU if necessary. Potential GPU applications: raytracing for simulations., ML modeling, etc.	link

File system	Performance	Available space	Comments
Community	>100 GB/s	10 PB for LZ	Large, permanent, medium-performance, shared across LZ
PM Scratch	5+ TB/s	35 PB total	Temporary, flexible, high-performance SSD
HPSS	>1 GB/s	6 PB for LZ	Tape archive for long-term storage

Other Resources	Function	Trained users	Comments
SPIN	Host Cloud Services	20+ from LZ	SPIN hosts all our web services and DataBases (+mirrors). Infrastructure for data movement and prompt processing.

USDC Role during LZ Operations

All LZ software/tools running on Perlmutter since early 2023

- This includes: prompt processing, simulations, inference, user analysis, etc.
- A mix of CPU and GPU allocation (GPU is currently underutilized in LZ)
- Allocation awarded yearly via ERCAP; multi-year plans requested since 2020
- Reliability of Perlmutter & its infrastructure are a top risk item for LZ operations

USDC Role during operations:

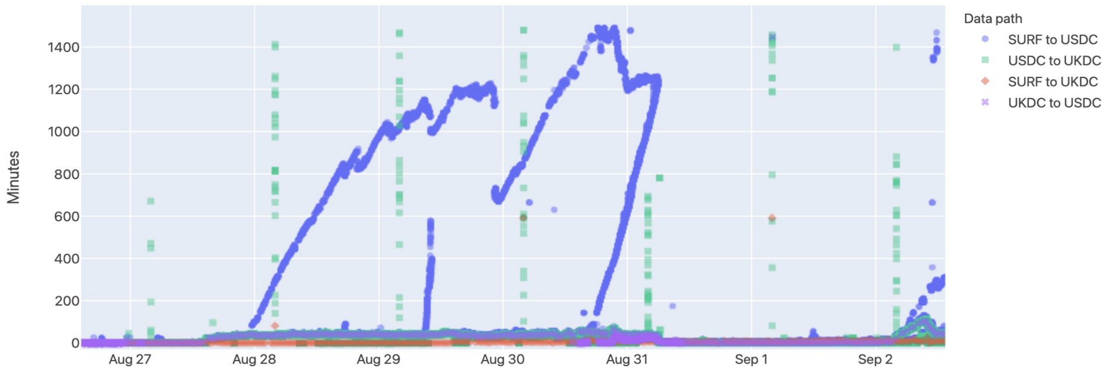
- Data Movement Endpoints (from SURF and to/from UKDC)
- Host 2 full copies of raw and processed data (on disk & tape, bias mitigation)
- Prompt Processing of all detector data (reprocessing planned, not available yet)
- MC production when necessary (halted in 2022, and restarted this year)
- User analysis: tools, resources, support, software, etc. (200 active users)
- Infrastructure software: web services, DBs, data catalog, bias mitigation, etc.

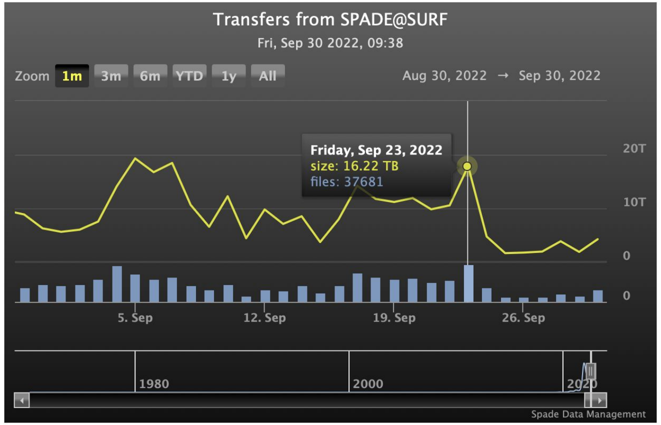
Data Movement

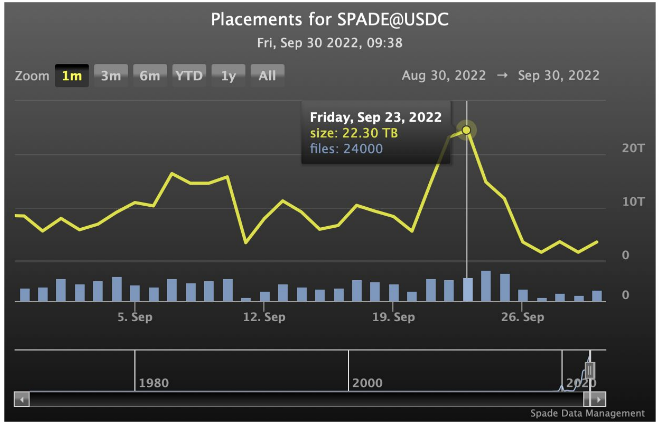
Our Data Movement framework is SPADE

- SPADE (South Pole Archival and Data Exchange) originally built for IceCube, then adopted by DayaBay and light-sources
- Modular application written in Java. It supports a variety of underlying transfer protocols (including GridFPT)
- LZ has SPADE endpoints at USDC, SURF, and UKDC. All data movement and warehousing operations are fully automated

Example: Data Transfer Latency



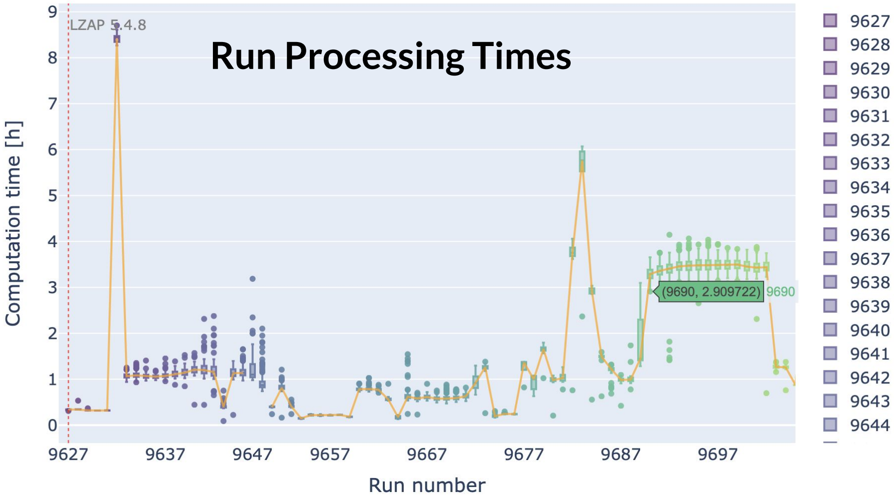


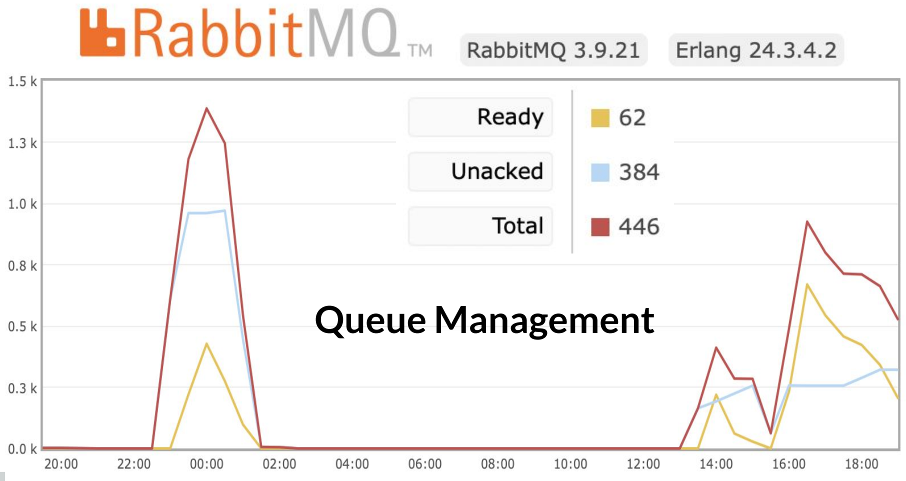


Prompt Processing at NERSC

We use P-Squared (P^2) for job management and submission at NERSC

- P-Squared is used to define, schedule, monitor and control large numbers of jobs. It's a custom framework, originally developed for DayaBay, built on top of RabbitMQ
- Prompt processing happens automatically as raw data files are received at NERSC and ingested in the data catalog. It is triggered by SPADE and managed by P-Squared. During science operations, raw data are typically processed within 30 minutes from submission (including queue wait times)

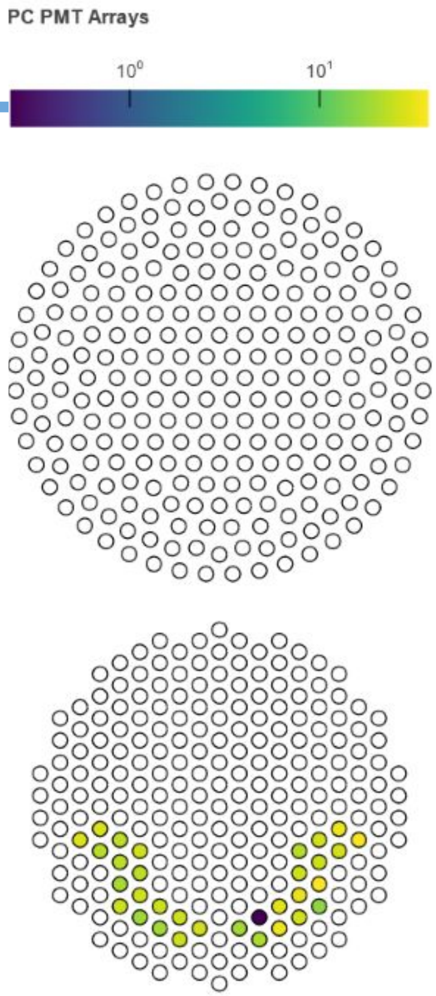




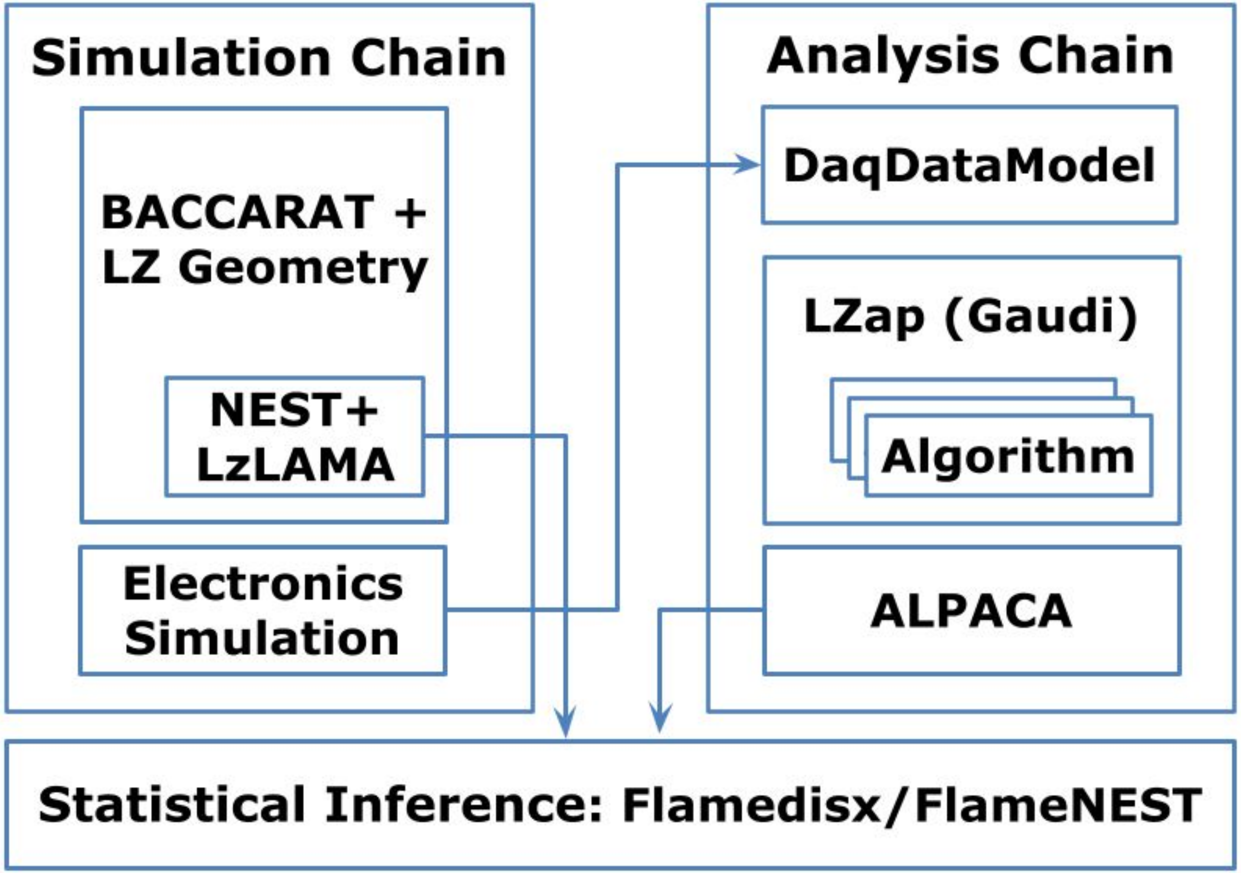
Extensive use of SPIN services

Supporting both production tools and user access!

- Data transfer (SPADE)
- Job submission engine (PSQUARED)
- Monitoring data movement and processing (SPADE/PSQUARED)
- Offline event viewer
- PREM (Offline Data Quality Monitor)
- Databases, database mirrors, and associated web service interfaces
- Data Catalog and its interfaces
- Code Quality and Software Release validation
- Web Services using SAML/NGINX authentication tools



LZ Software Elements



LZ simulation and reconstruction rely heavily on "standard" HEP frameworks

Crucial dependencies: Geant4, Gaudi, ROOT

Resiliency, robustness, and reliability of NERSC

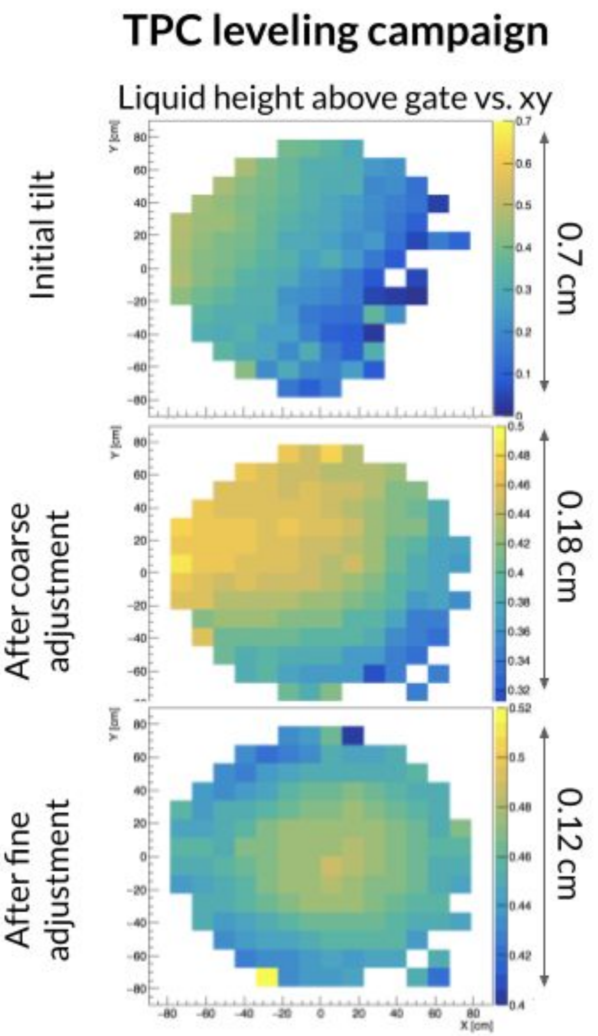
Quasi-real-time computing

Commissioning success: the leveling campaign of October 2021

- [Premise: SURF underground days are Mon-Thu or Tue-Fri]
- We performed the leveling of the detector on a Mon-Thu week
- However, there was a scheduled Cori outage that Wed
- We needed to be able to look at/analyze data every night
- (heroic effort from NERSC to keep us running on Gerty that week)

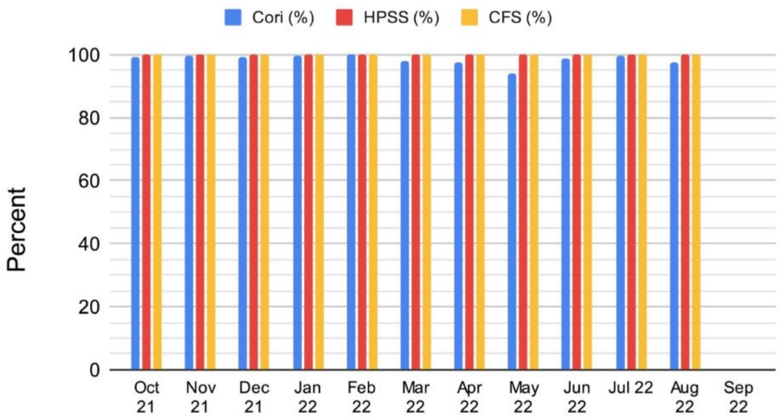
Superfacility uptime: uptime of LZ x uptime of computing services

- Downtime is expensive:
  - Defensive Engineering
  - Impact on Detector Operations
  - Reputation with Science Partners
- Our computing infrastructure is quite complex. Instabilities on a single subsystem (DB, disk, CPU) can impact the entire workflow.



Not all uptime is created equal: 2022

System availability excluding scheduled outages.



Cori: 98.6%
HPSS: 100%
CFS: 100%

Month	Cori (%)	HPSS (%)	CFS (%)
Oct 21	99.4	100	100
Nov 21	99.8	100	100
Dec 21	99.4	100	100
Jan 22	99.6	100	100
Feb 22	100	100	100
Mar 22	98.1	100	100
Apr 22	97.6	100	100
May 22	94.2	100	100
Jun 22	98.7	100	100
Jul 22	99.8	100	100
Aug 22	97.7	100	100
Sep 22	99.3	100	100

https://www.nersc.gov/assets/NUG-Metrics-2022.pdf

Not all uptime is created equal: 2023

System Availability Has Remained High

Month	Cori (%)	HPSS (%)	CFS (%)	Perlmutter (%)
Oct 22	100.0	100.0	100.0	
Nov 22	99.2	100.0	100.0	
Dec 22	99.6	99.9	100.0	
Jan 23	98.7	100.0	100.0	
Feb 23	100.0	100.0	99.8	
Mar 23	97.1	100.0	100.0	
Apr 23	97.9	100.0	100.0	
May 23	99.3	100.0	100.0	
Jun 23		100.0	100.0	98.5
Jul 23		100.0	100.0	99.6
Aug 23		100.0	100.0	97.7
Sep 23				

Rebecca Hartman-Baker, NUG presentation 2023
https://youtu.be/IXCW-YnYRAU?si=fUop6OkMtb1znghp&t=960

NERSC MOTD

- Live Status (MOTD)
- Scheduled outages for the rest of the calendar year
- One additional multi-day outage expected this summer for power inspection work

ASCR "uptime"

≠

Science uptime

Planned Outages:

Perlmutter:	06/26/24 6:00-18:00 PDT Scheduled Maintenance
HPSS Archive (User):	06/26/24 9:00-15:00 PDT Scheduled Maintenance HPSS Archive will remain available during scheduled maintenance. Some tape file retrievals may be delayed during the maintenance window.
Data Transfer Nodes:	06/26/24 9:00-12:00 PDT Scheduled Maintenance This will affect globus and interactive DTN nodes.
Globus:	06/26/24 9:00-12:00 PDT Scheduled Maintenance This will affect globus and interactive DTN nodes.
HPSS Archive (User):	07/14/24 19:00-07/19/24 17:00 PDT Scheduled Maintenance HPSS Archive upgrade from V7.4 to V9.3. the main HPSS environment, Archive, will be shutdown from 7 pm (Pacific Time) on July 14 till 5 pm on July 19, 2024, while a new version of the system is deployed. During this 5 days period, it will not be possible to store or retrieve data into or from the system. Users are strongly encouraged to plan accordingly.
Perlmutter:	07/17/24 6:00-07/18/24 18:00 PDT Scheduled Maintenance Perlmutter unavailable for major upgrades. Login nodes and Perlmutter scratch available starting 7/17 at 18:00. Job submission (e.g. sbatch) and querying (e.g. squeue) will be unavailable. No user jobs or scrontabs will run. There may be intermittent disruptions to login nodes or Perlmutter scratch access once they become accessible.
Perlmutter:	08/27/24 6:00-20:00 PDT Scheduled Maintenance
Perlmutter:	09/18/24 6:00-20:00 PDT Scheduled Maintenance
Perlmutter:	10/16/24 6:00-20:00 PDT Scheduled Maintenance
Perlmutter:	11/13/24 6:00-20:00 PST Scheduled Maintenance
Perlmutter:	12/18/24 6:00-20:00 PST Scheduled Maintenance

Not all uptime is created equal: 2024

Time range: Jan 1st - Jun 23 2024 (175 days). Data from: https://my.nersc.gov/outagelog-cs.php	Total downtime (scheduled + unscheduled + degraded)	
	Days	Fraction
Perlmutter (excluding GPU-only events)	8.7	4.9%

- Superfacility uptime: uptime of LZ x uptime of NERSC services
- Uptime fraction over scheduled has limited usefulness for LZ ops
- We use so many part of NERSC, that downtime or degradation anywhere (DTNs, CFS, Slurm, SPIN, etc.) impacts entire workflow

Not all uptime is created equal: 2024

Time range: Jan 1st - Jun 23 2024 (175 days). Data from: https://my.nersc.gov/outagelog-cs.php	Total downtime (scheduled + unscheduled + degraded)	
	Days	Fraction
Perlmutter (excluding GPU-only events)	8.7	4.9%
Perlmutter, SPIN, CFS, DTN, Globus, Superfacility API	12.8	7.3%

- Superfacility uptime: uptime of LZ x uptime of NERSC services
- Uptime fraction over scheduled has limited usefulness for LZ ops
- We use so many part of NERSC, that downtime or degradation anywhere (DTNs, CFS, Slurm, SPIN, etc.) impacts entire workflow

Downtime and impact on operations

Time range: Jan 1st - Jun 23 2024 (175 days). Data from: https://my.nersc.gov/outagelog-cs.php	Total downtime (scheduled + unscheduled + degraded)		# of days with outage or system degradation	
	Days	Fraction	Count*	Fraction
Perlmutter (excluding GPU-only events)	8.7	4.9%	32	18%
Perlmutter, SPIN, CFS, DTN, Globus, Superfacility API	12.8	7.3%	53	30%

- LZ experienced 53 “events” impacting computing operations at NERSC in 2024, affecting 30% of calendar days (or 44% of business days - excluding instabilities e.g. SPIN)
- Impact on LZ operations: we had to give up on our plan for data turnaround and detector data quality monitoring on the ~day scale (more in this session from David/Ibles)

*consistent with reports from DESI etc.

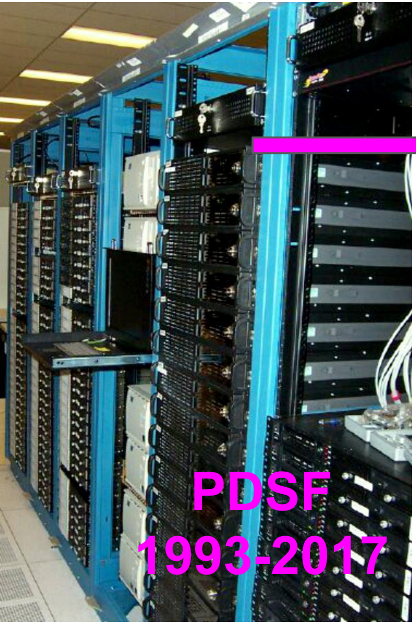
Downtime and impact on operations

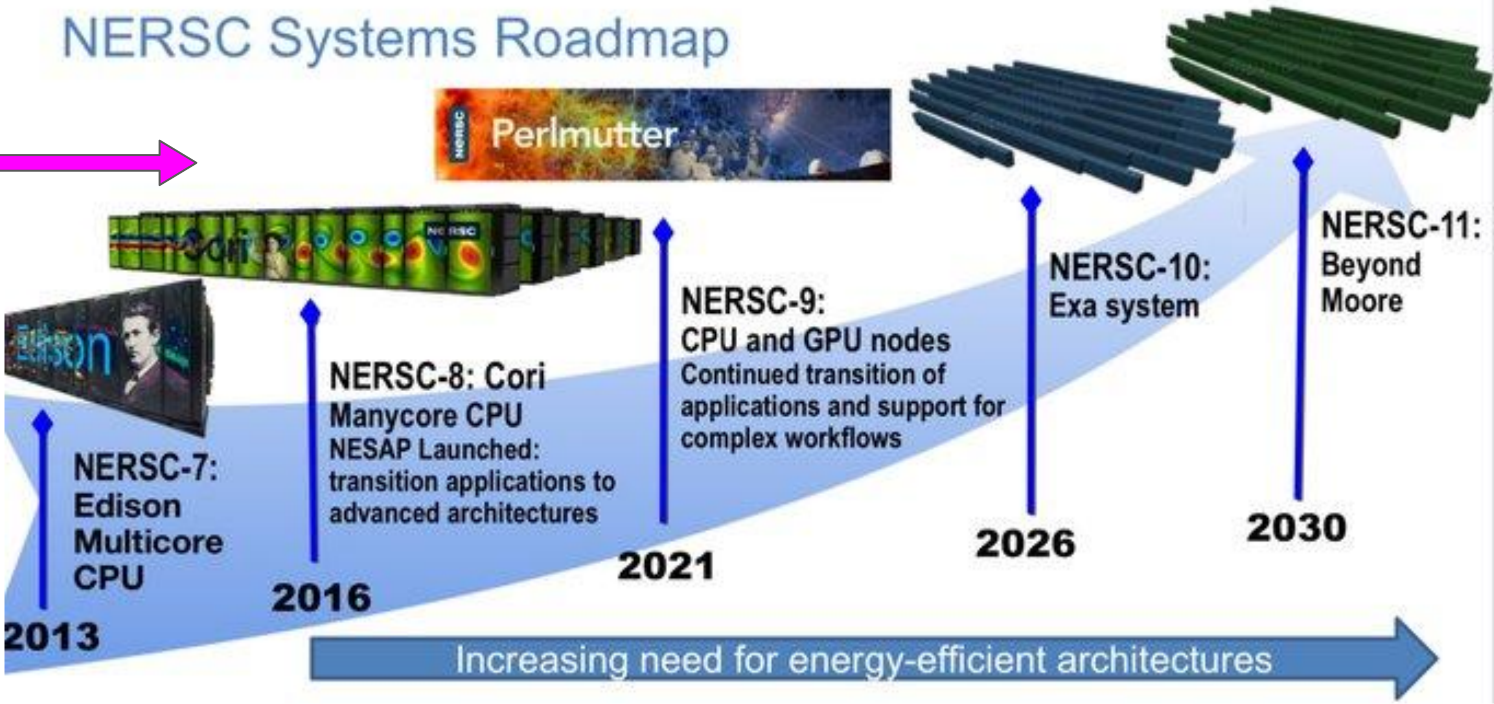
Time range: Jan 1st - Jun 23 2024 (175 days). Data from: https://my.nersc.gov/outagelog-cs.php	Total downtime (scheduled + unscheduled + degraded)		# of days with outage or system degradation	
	Days	Fraction	Count	Fraction
Perlmutter, SPIN, CFS, DTN, Globus, Superfacility API	12.8	7.3%	53	30%

- LZ experienced 53 “events” impacting computing operations at NERSC in 2024, affecting 30% of calendar days (or 44% of business days - excluding instabilities e.g. SPIN)
- Impact on LZ operations: we had to give up on our plan for data turnaround and detector data quality monitoring on the ~day scale (more in this session from David/Ibles)
- Consequences for staffing/retention: we keep recruiting people to cope with this rate of disruption, but people get quickly burned out and discouraged
- Cautionary tale in view of HPDF and of upcoming HEP experiments (Rubin, CMB-S4, etc.)

Roadmap through 2028

Computing model evolution at NERSC





Computing architectures evolve much faster than physics experiments!

- LZ is now on its 4th NERSC system. Adoption of NERSC-10 in 2026+ is likely

What else is on the horizon for ASCR?

ASCR Facilities Program Element	Major Project(s)
High Performance Networking	ESnet6 ✓
High Performance Production Computing	NERSC-10, HPDF
Leadership Computing	OLCF-6, ALCF-4

IRI Pathfinding Testbed

B. Brown, June 12. 2023

ESnet, OLCF, NERSC, and ALCF are finalizing a jointly-authored concept paper, quoted at right.

Each facility will contribute resources to creating this research environment.

"The proposed IRI Testbed is a progressive design-experiment and test-refine approach proposed to ***establish a shared environment for IRI developers and pilot application users to come together*** *and advance the IRI vision.*

"The goal is to build cyberinfrastructure enabling multiple user facilities to experiment with the design patterns and address the gaps identified in the IRI Architecture Blueprint Activity (ABA) report."

Prompt processing @ UKDC?

Staffing and operational constraints at the UKDC

- UKDC has ~1 FTE of engineering, across 5 different people
  - + 1 additional FTE for management and user support
  - + 1->2 FTEs for production management and operation
  - no bandwidth for a separate prompt processing chain

Why does it have to be a separate processing chain?

- Summary: NERSC is **not** a grid site / GridPP is **not** an HPC
  - diverging job submission interfaces (slurm vs ganga)
  - diverging data access interfaces (CFS vs xrootd)
  - diverging identity management (certificates vs MFA)
  - we don't "own" architecture or policies at either facility
- These challenges have a major impact on data movement
  - example: limited support for grid certificates at NERSC

US Data Center (USDC):

- Prompt Processing
- Long-term Archiving
- Supercomputers!

NERSC

UK Data Center (UKDC):

- Data Reprocessing
- Sims Production
- Distributed CPUs!

GridPP
UK Computing for Particle Physics

The workflow portability pilot

Approach: LZ is investigating a workflow portability pilot

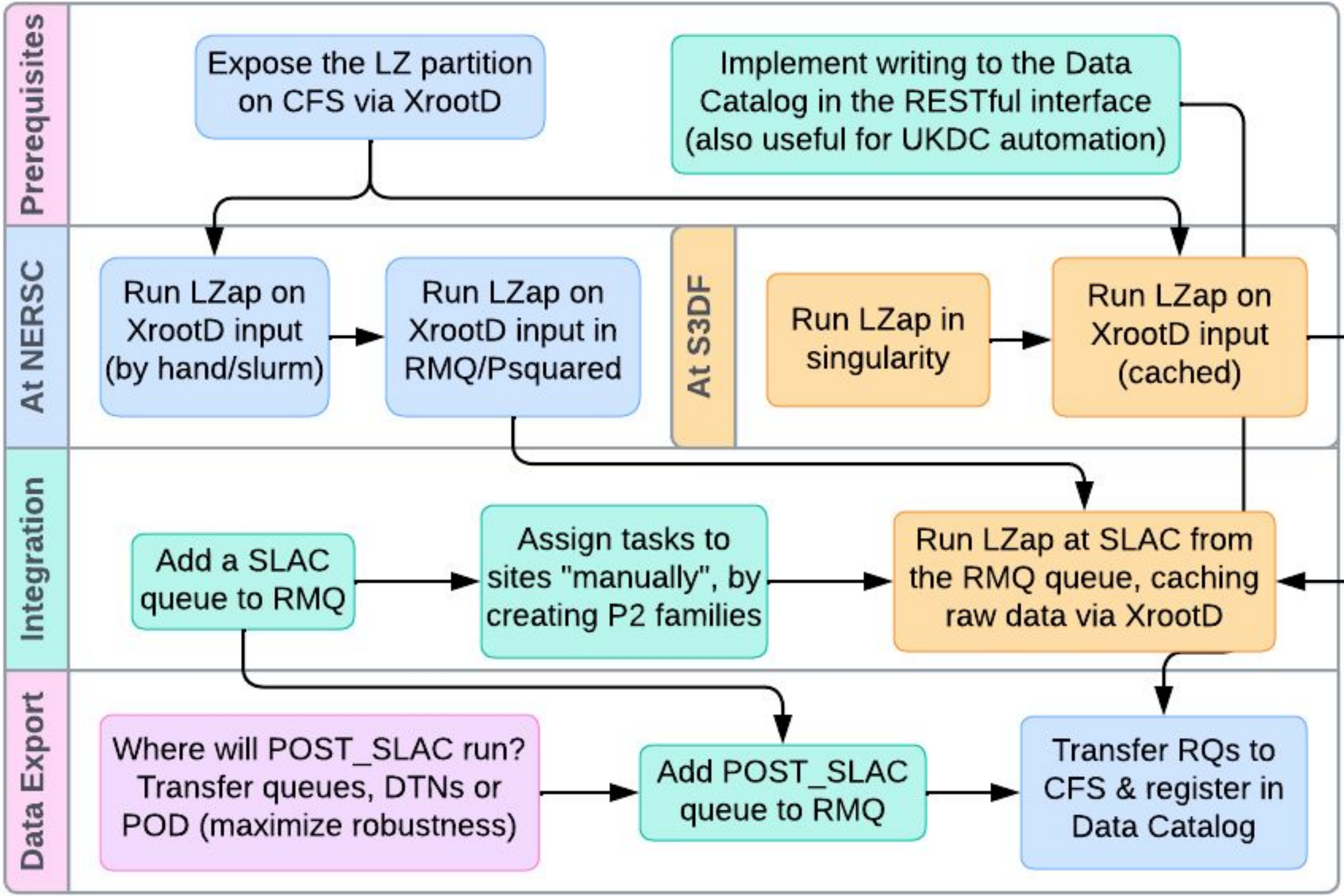
- Goal: maximize uptime, guarantee fast turnaround (<1 day)
- Plan: a “backup” system in the US to mitigate NERSC downtime
- Bonus: facilitate the transition to NERSC-10 if/when needed
- Resources: we have recruited additional staffing for this effort
- Support: work will be performed in collaboration with HEP-CCE

Multiple options for alternate data center(s) in the US

- ANL: hoping for similar interfaces and protocols to NERSC
- FermiGrid: simplify the data movement issues with GridPP
- SLAC S3DF: same architecture as Perlmutter (AMD Milan)
  - Stringent uptime constraints for Rubin operations
  - Additional benefit: synergies with DESC and LCLS-II
  - Future: will S3DF be a “spoke” in the HPDF ecosystem?

Workflow portability pilot

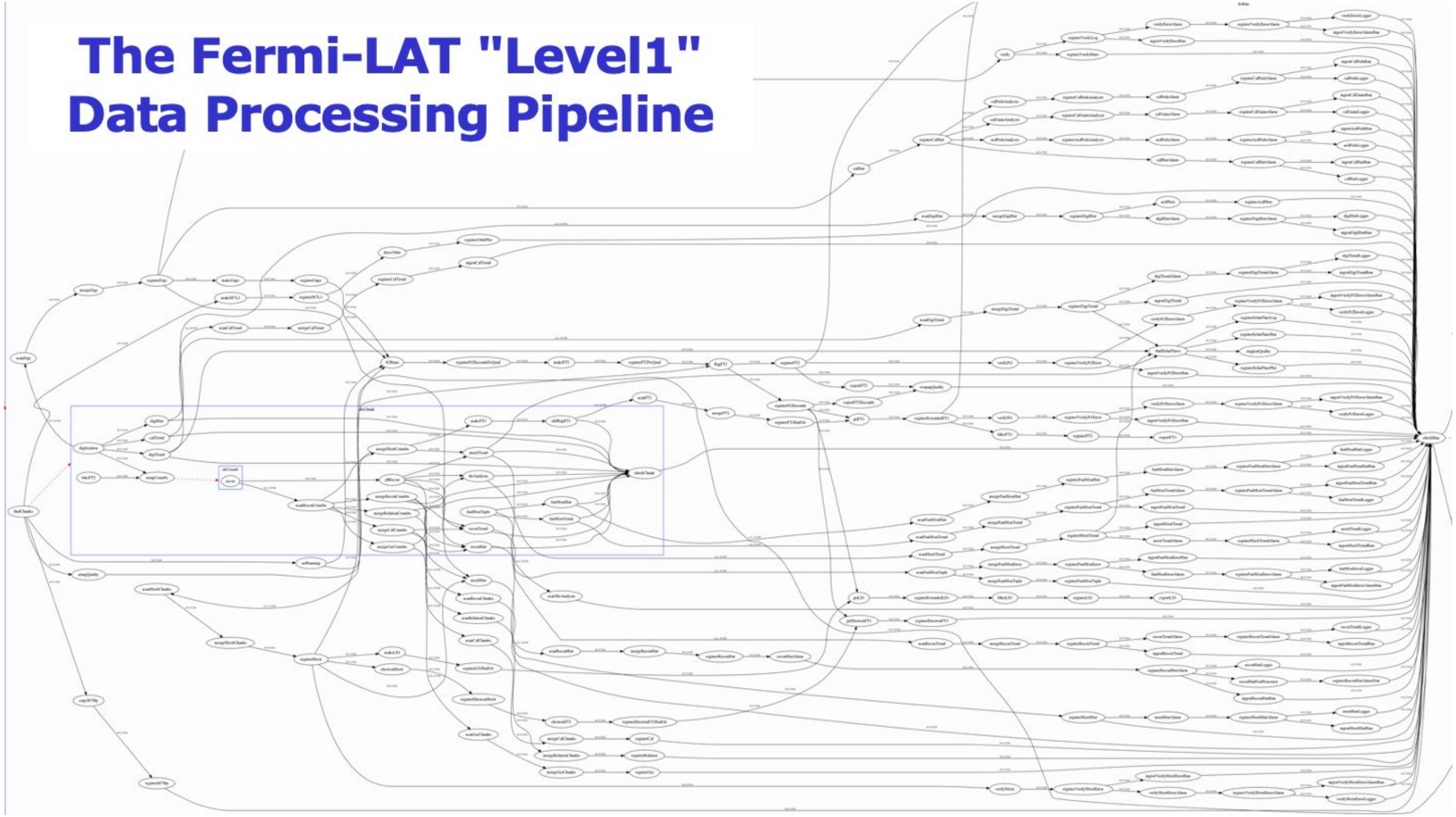
Maria Elena Monzani | October 20, 2023



Relocating the Fermi-LAT pipeline to S3DF

B50 -> S3DF: 1+ year

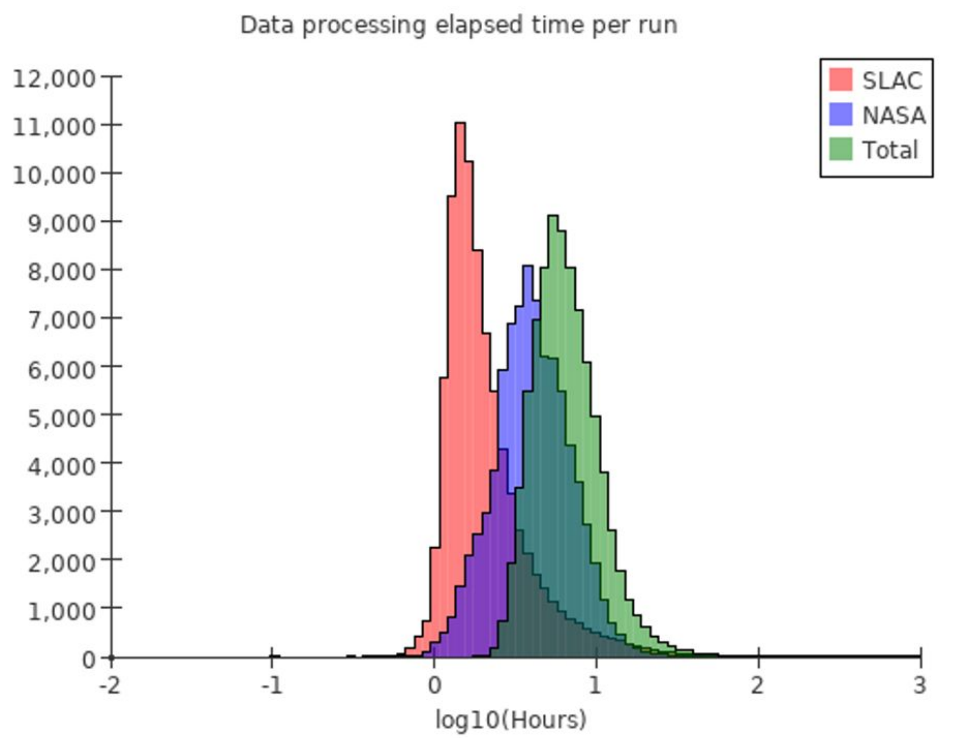
Relocating the Fermi-LAT pipeline to S3DF



Relocating the Fermi-LAT pipeline to S3DF

B50 -> S3DF: 1+ year

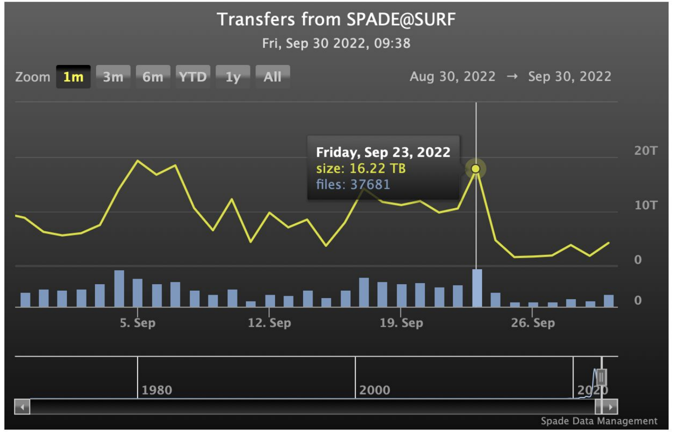
16-year Data Latency (pre-S3DF)

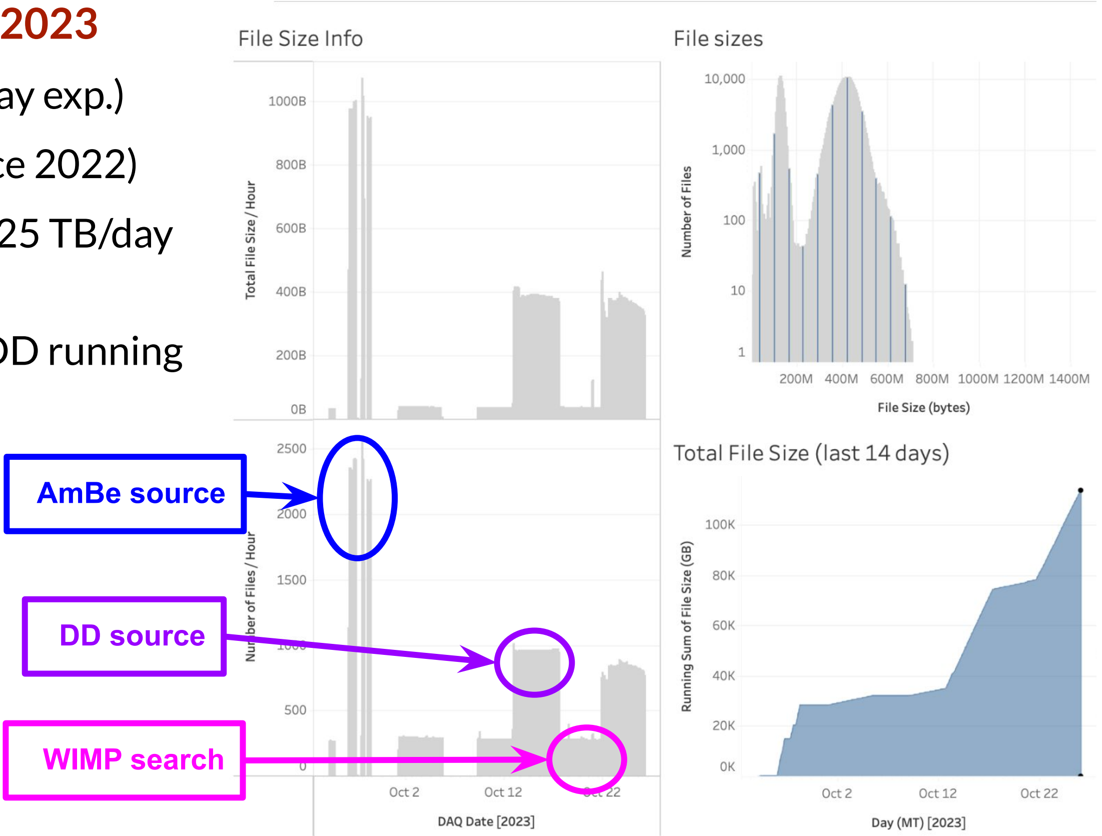


Data Movement Robustness

Neutron calibration campaign of Oct 2023

- WIMP search rate: 1 TB/day (3 TB/day exp.)
- Demonstrated: 15 TB/day (DD source 2022)
- How high is too high? AmBe source: 25 TB/day
  - that was definitely too high
  - also, didn't plan for continuous DD running





Data Movement Robustness

Data Movement is an infrastructure vulnerability

- Data rates are often higher (or much higher) than planned (calibrations - up to 30x higher than WS, skin emission, etc.)
- SPADE is an "ancient" tool, and is showing its limitations
- Integration with GridPP is challenging (diverging identity management protocols and interfaces: certificates vs MFA)

Plan: replace SPADE with a more modern tool

- Currently looking at RUCIO, which is being adopted widely
- Improved GridPP integration (designed for LHC experiments)
- Expose all datasets via xrootd, to support portable workflow
- Resources: we have recruited additional staffing to this effort

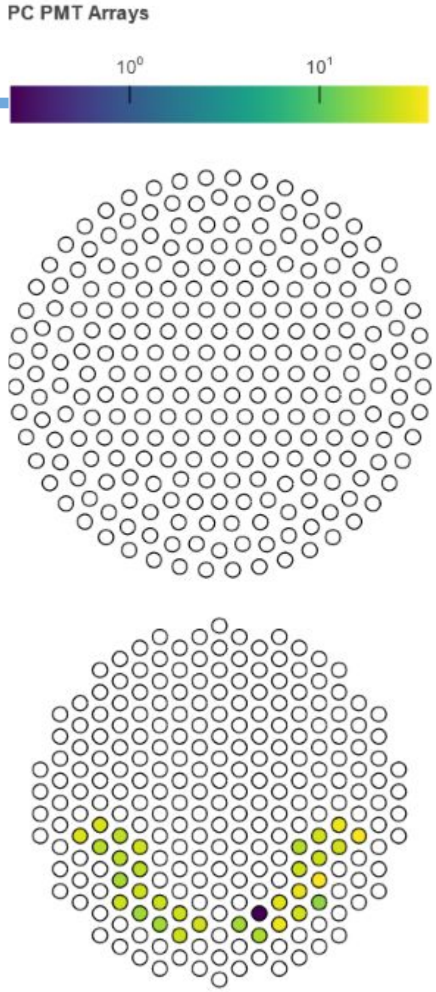
Reliability of SPIN services

LZ makes extensive use of SPIN services, for production and user access

- Reliability of database workloads on SPIN has been inadequate for almost two years. Recovery from frequent failures is very labor intensive
- Underlying cause: incomplete separation of development and production clusters. Storage designed for file I/O rather than block I/O
- A recent policy change (zero-trust architecture) required LZ to update its policies for DB access and revamp some of our interfaces

We may need to move our DB workloads to a commercial cloud provider

- We are working with NERSC to address some of these vulnerabilities, but solving the issue requires a refurbishment of the underlying hardware
- Timeline for this hardware upgrade is uncertain. Hopefully CY2024?
- Backup solution: pursue external avenues (google cloud, AWS) to host our DB workloads down the line. Keeping this as a risk item for the time being



"Zero-trust" architecture: DB access

Existing security exemption for database access on SPIN was revoked in December 2023

- This policy change was required to comply with "zero-trust architecture" mandate from DOE
- All DB connections routed through a proxy server with firewall (limited set of IPs are allowed)
- Starting this month, DB access is only permitted from IPs belonging to LZ institutional clusters
- LZ officially discontinued "analysis from your laptop" support (not widely used in recent years)
- Complying with both mandates is becoming increasingly more complex AKA more expensive
- We were forced to revamp some interfaces and tools. This transition took about 6 months

Potential tension between "zero-trust architecture" & "OSTP public access" initiative

2023 PAP: Data Management and Sharing Plan (DMSP) Requirements				
Validation and replication of results	Timely and equitable access	Data repository selection	Data management and sharing resources	Data sharing limitations

DOE Implementation of the OSTP "Nelson Memo", HEPAP meeting May 2024

What else is of interest?

Scaling up HEP AI/ML applications

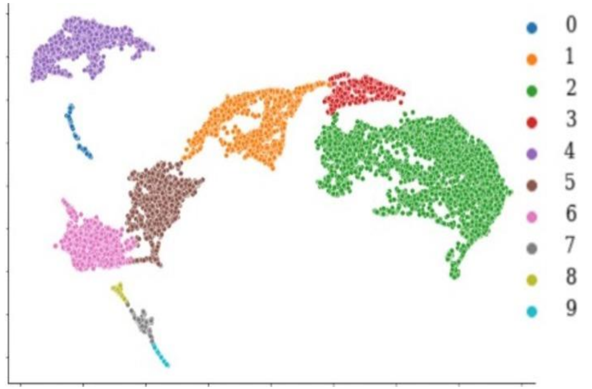
Extreme needle in a haystack problem:

- Identify a handful of DM events (if nature cooperates)
- Expected background is of order ~5-10 billion events
- Background rejection problem with a rarity of order 10^{-9}
- Ideal playground for the development of novel ML algorithms
- Rare/unmodeled backgrounds can spoil bias mitigation schema

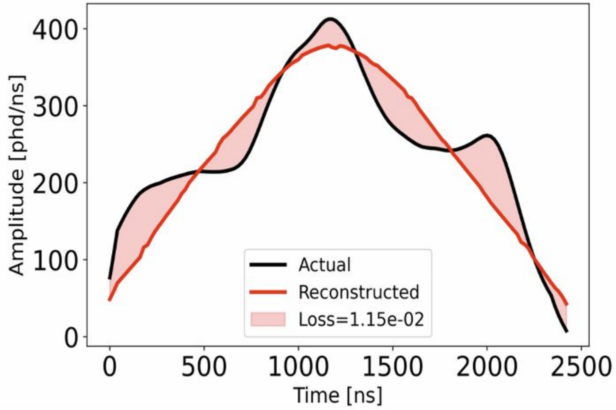
Approach: anomaly detection at the 10^{-9} sensitivity

- Collaboration with Stanford ICME (School of Engineering)
- Tools: event clustering and resilient-VAEs (in recursive mode)
- Challenge: train ML models on the waveform (multi-PB dataset)
- There are currently no machines with a multi-PB scale RAM

UMAP + DBSCAN (credit: Maris Arthurs)



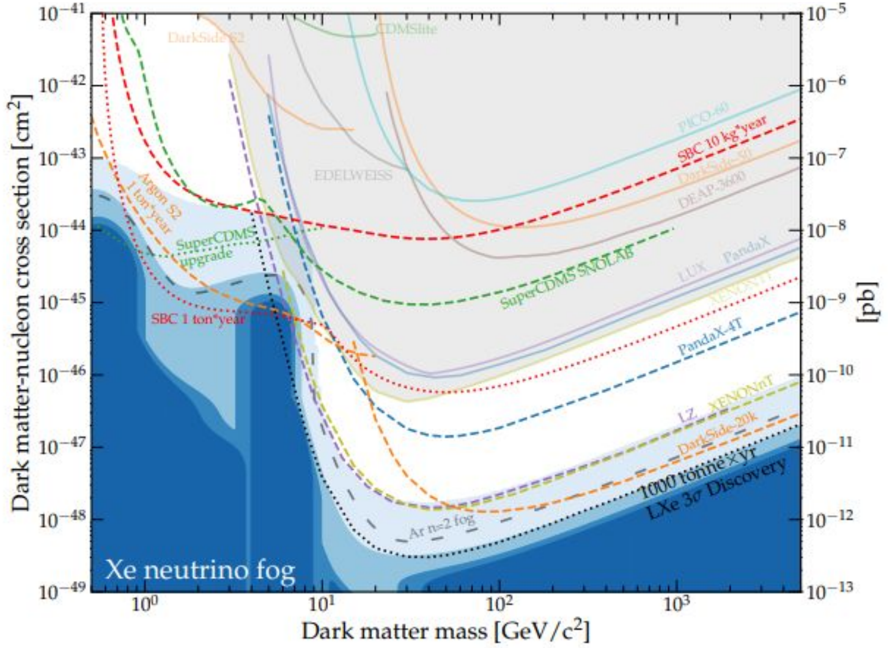
VAE on full WF (credit: Tyler Anderson)

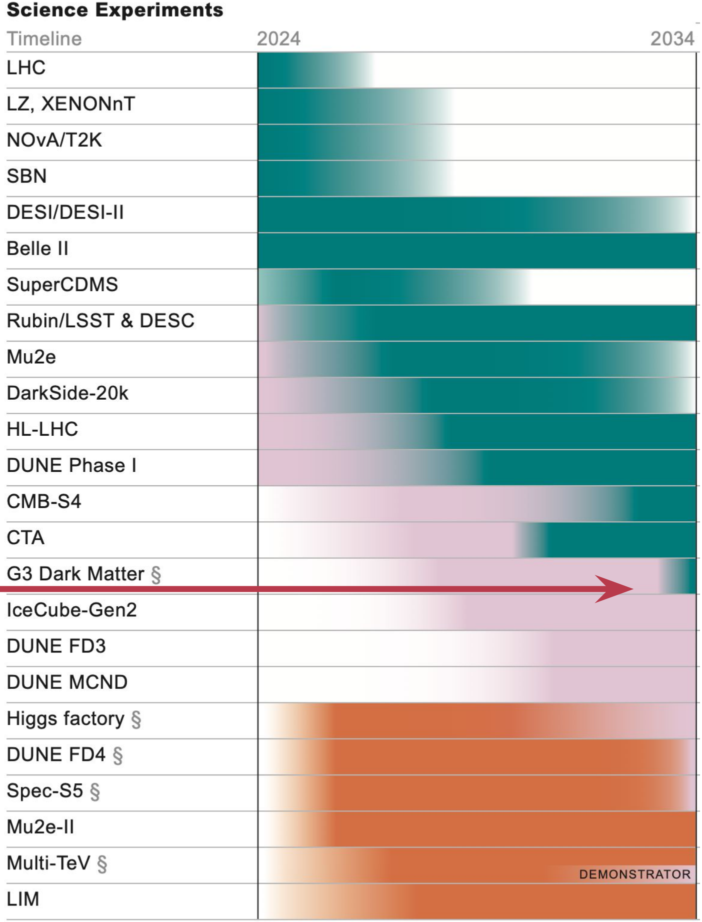


What will happen after LZ?

LZ taking data through 2027. Analysis through 2028+

P5 endorsed an "ultimate" Dark Matter experiment



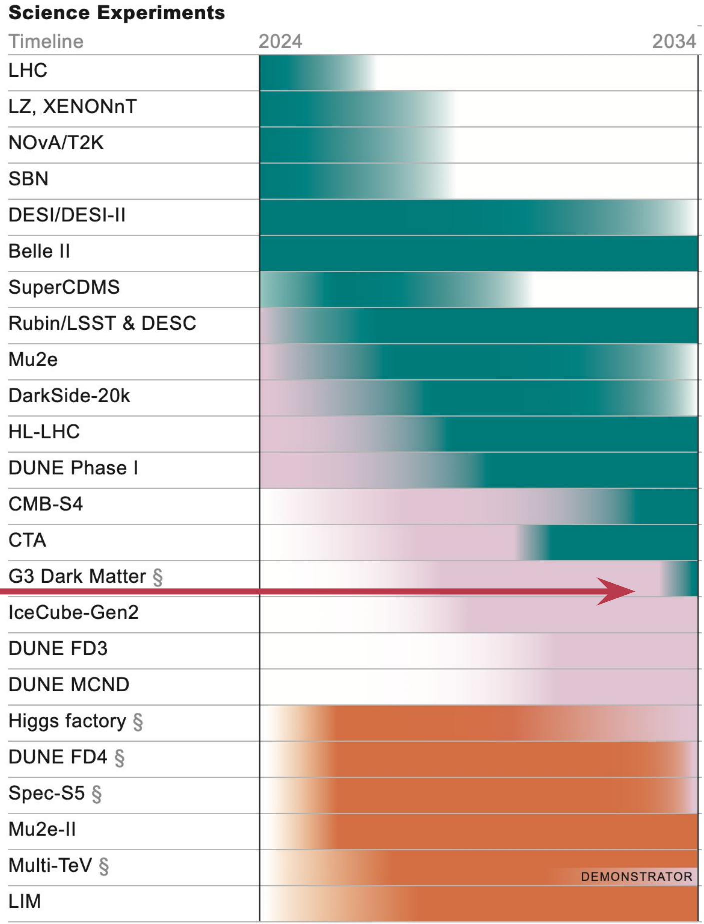


Simulation needs for the post-LZ

LZ taking data through 2027. Analysis through 2028+

P5 endorsed an "ultimate" Dark Matter experiment

- Multi-purpose observatory for a multitude of dark matter models, neutrinoless double beta decay, and astrophysical neutrinos
- Fully probe WIMP parameter space into the neutrino fog (50-100 tonne experiment)
- A x10 scale-up from LZ: will need accurate simulations to design the "ultimate" experiment
- This level of accuracy requires raytracing on the GPU, which is needed in the next ~few years



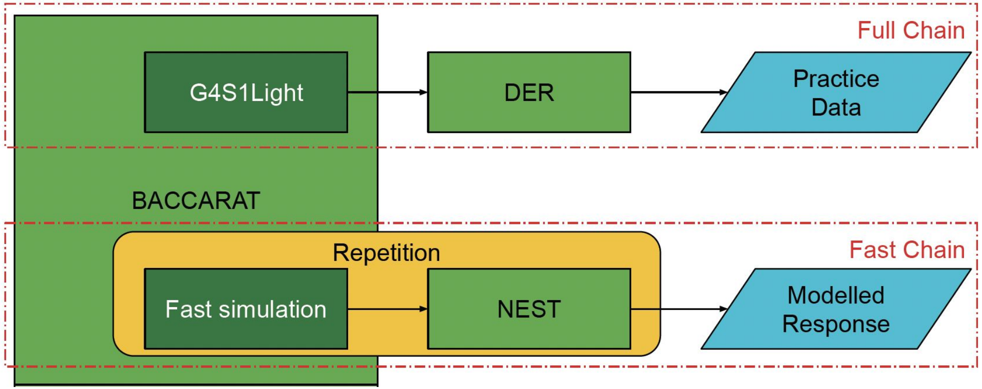
LZ Simulation on GPU (NESAP project, 2020-2023)

BACCARAT tracks particles using Geant4. Various features have been added to BACCARAT to better model the xenon and GdLS response from the LZ detector

DER is a software package designed to simulate the signal processing done by the analogue front-end electronics of LZ

Full chain simulation tracks photons and electrons generated by the interaction and record individual photon hits on the PMTs, optical tracking consume >95% of CPU time used in LZ simulations

Fast chain speed-up the simulation factor nearly 20x, however result do not contain information on the time of interactions or specific photon hits on PMTs (energy deposits are passed to the NEST module which uses detector averaged quantities to generate S1 and S2 signals)



A schematic of the current LZ simulation workflow

LZ Simulation Challenges

- Simulating particles requires navigation through geometry trees built for each solid in the geometry, a solid tree consists of simpler shapes or primitives
- Optical photon simulations may required >95% of the simulation time in BACCARAT, but they only interact at the boundary of the volume (don't interact within the volume)
- Treating optical photons separately can save a significant amount of simulation time
- To avoid optical photon simulations, the S2 Light Map was developed, but it is not optimal and differs from the full simulation approach
- Simulating events that involve a significant number of optical photons, like muons, is not possible due to their large quantity
- **GPUs** can be used to perform ray-tracing for physics rather than visualization, potentially accelerating the simulation process

Different Approaches for GPU Simulations

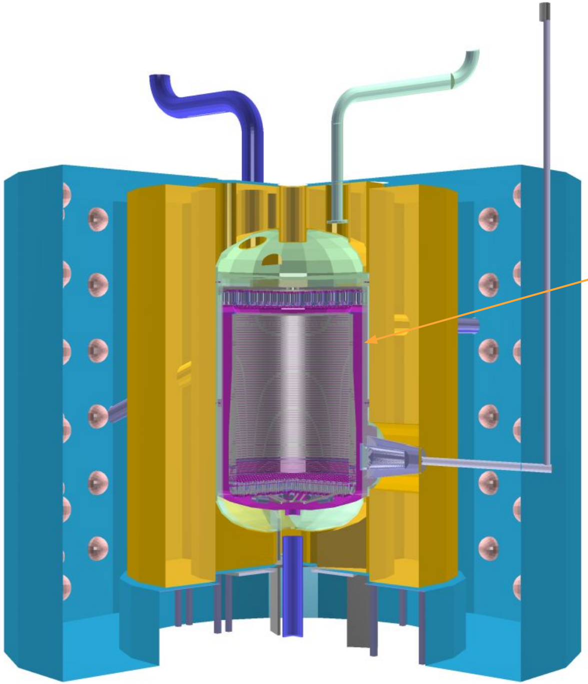
Some highlights:

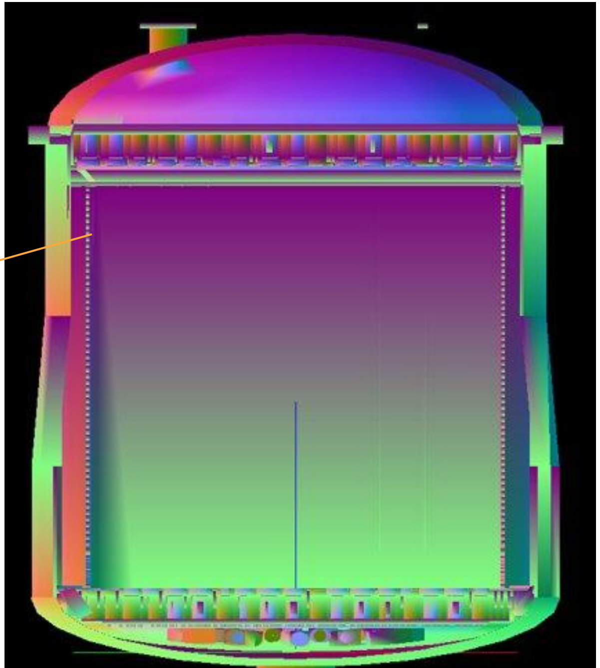
- Opticks/Optix
  - A system which maps Geant4 geometry and photon generation steps to NVIDIA's OptiX GPU ray-tracing framework
- Celeritas
  - GPU-accelerated particle transport for detector simulation
- Mitsuba-3
  - Industry open-source rendering software for optical photon simulation
- *Larnd-sim*
  - Highly-parallelized simulation of a pixelated LArTPC on a GPU (DUNE)

+ *many more:*

- ***26TH INTERNATIONAL CONFERENCE ON COMPUTING IN HIGH ENERGY & NUCLEAR PHYSICS (CHEP2023)***, **May 8 – 12, 2023, Norfolk Waterside Marriott, VA, USA**
- ***GridPP49 & SWIFT-HEP05***, **March 28-30, 2023, Rutherford Appleton Laboratory, UK**

LZ Geometry : GPU useable input



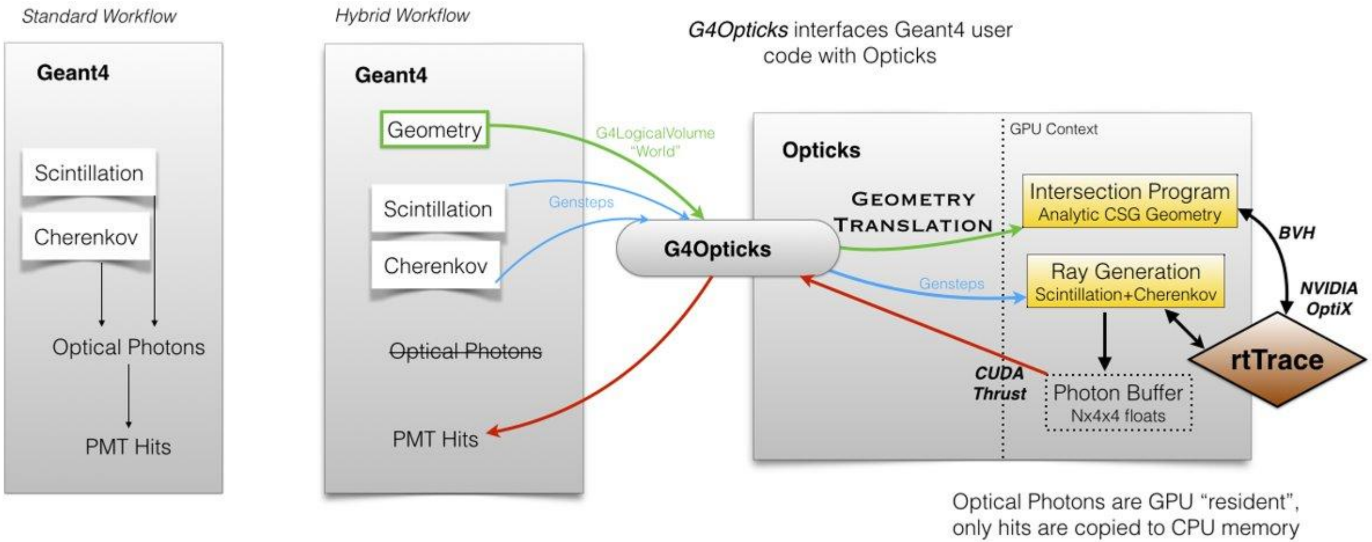


picture from Sam Eriksen

- Full LZ geometry has thousands of logical volumes
- LZ TPC itself contains >8000 different solids (tens of thousands of primitives)
- LZ geometry (BACCARAT) is converted to GDML and then OBJ as part of the CI --> Work by Sam Eriksen
- Most of the GPU simulations (approaches) would take either GDML/OBJ as a input files

Opticks: Overview

An illustration of how Opticks integrates OptiX into a particle physics workflow



More details:
[O. Creaner and et al.](#)

- Opticks translates Geant4 geometry and photon generation steps for OptiX
- Geant4 geometry is converted to GPU-compatible form and uploaded to GPU
- OptiX performs photon generation and propagation using ray tracing during event processing
- Only photon hits on PMTs are sent to CPU for further processing after OptiX ray tracing is complete

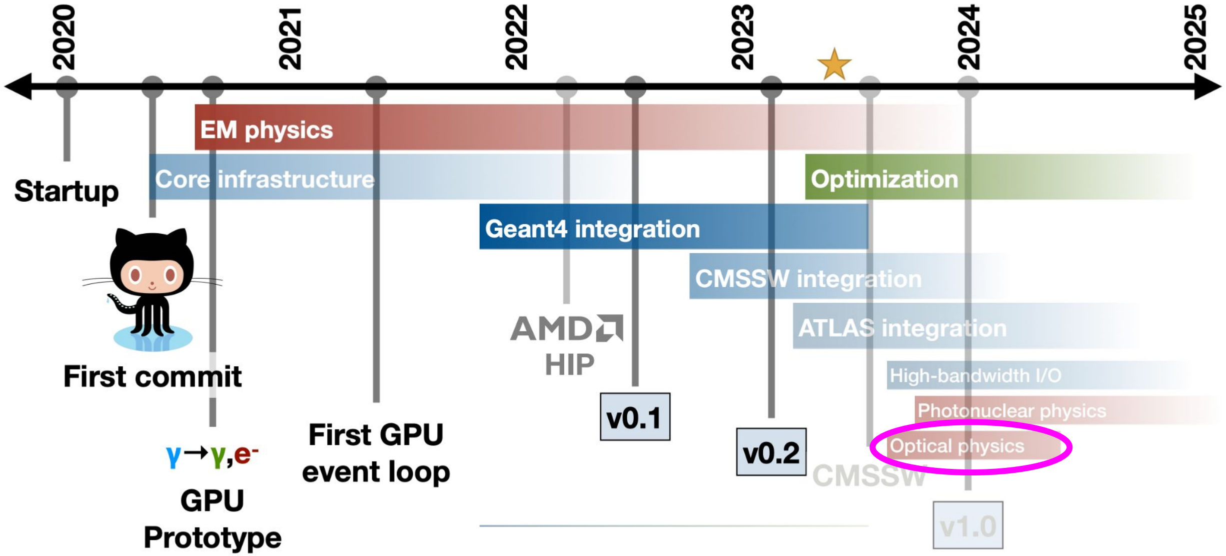
Opticks: What has been done so far?

- Containerize Opticks / Optix for LZ simulation, O. Creaner and et al.
    - Docker image was created a few years ago to run on Cori GPU, https://gitlab.com/luxzeplin/sim/opticks-on-shifter
    - Existing instructions are outdated and it took multiple steps to get the container running on Cori (did not spend much time since Cori is retiring soon)
    - Real physics or LZ examples for testing are not yet available

- Prior experience with using Opticks to simulate JUNO indicates the potential for speed-up factors over 1000x for LZ
    - From Sam Eriksen's thesis, a photonbomb in the TPC is 720x faster on a T4 GPU than Geant4

Celeritas: Overview

- **GPU**-focused implementation of **HEP** detector simulation
- Physics derived from Geant4 methods and implementation
- Tracking of EM interaction through particles and Geant4 (10.6-11.0) integration is ongoing
- Planning to implement the optical physics for the GPU simulation
- User+developer documentation, link



Picture from Celeritas' CHEP May 8, 2023 Presentation, link

Celeritas: What has been done so far?

- Waiting for Optical Physics implementation to be ready!
- Meanwhile, created a shifter image to run on Perlmutter with CUDA base image (cuda:11.8.0-devel-ubuntu22.04), installed spack and pre-requisites for celeritas (based on the instructions)
  - Dockerfile for this image can be found here, https://gitlab.com/luxzeplin/sim/gpu/lz-celeritas
  - Used *podman-hpc*, a *very useful tools to create, pull and push the images on the Perlmutter*
  - Shifter image can be found on DockerHub https://hub.docker.com/repository/docker/mtimalsina/lz-celeritas/general
- Executed the docker image on Perlmutter GPU. Attempted to build the existing celeritas on it but this aspect was never completed.

Appendix: environmental impact of LZ computing vs collaboration travel

NERSC computing: energy expenditure

HPC System	System Power (MW) Average	Standard Deviation	Maximum	TDP
Cori	3.18	0.36	4.21	5.72
Perlmutter	3.19	0.49	4.86	6.90

Peak Power
30 PFLOPS
70 PFLOPS

NERSC average: 4 MW, including cooling etc. (Table Source)

Order of magnitude for energy and transportation

- Peak power output for a standard GE wind turbine: 2 MW
- Total power output of Titanic's coal-fueled steam engines: 4.4 MW
- Average power consumption of a Boeing 747 passenger aircraft: 140 MW

Order of magnitude comparison: computing vs air travel

- NERSC consumes 1/35 of the power of a Boeing 747, in average
- Running Perlmutter for a year is equivalent to flying an airplane for ~~10 days~~
  - Or about 3 weeks for a modern aircraft, like the Boeing 787 Dreamliner
- [neglecting construction costs for both systems in this approximation]

LZ computing: energy expenditure

HPC System	System Power (MW) Average	Standard Deviation	Maximum	TDP
Cori	3.18	0.36	4.21	5.72
Perlmutter	3.19	0.49	4.86	6.90

Peak Power
30 PFLOPS
70 PFLOPS

NERSC average: 4 MW, including cooling etc. (Table Source)

How much power did LZ use in 2023 for computing?

- 100k node hours (total NERSC hours: ~40M), or 1/400 of NERSC
- In 787 Dreamliner units, that comes out to 70 minutes of flight
- We also ran the equivalent of ~30k NERSC hours on GridPP
- GridPP is closer to Cori than Perlmutter for energy efficiency (x2)
- Let's call this 2h of flight in a Boeing 787 Dreamliner (upper limit)
- [ignoring individual laptops, monitors, institutional clusters]

Computing vs collaboration meeting

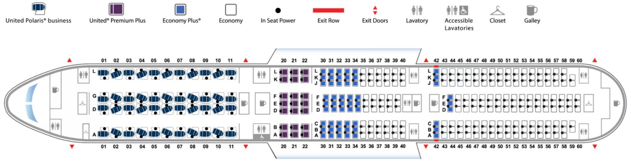
June collaboration meeting at Brown University

- 100 participants: 20% “local”, 20% traveled from EU/UK
- Average flight time: 11h/person round trip (guesstimate)
- 1100 flight hours/320 passengers: 3.5h in Dreamliner units

January collaboration meeting at University of Edinburgh

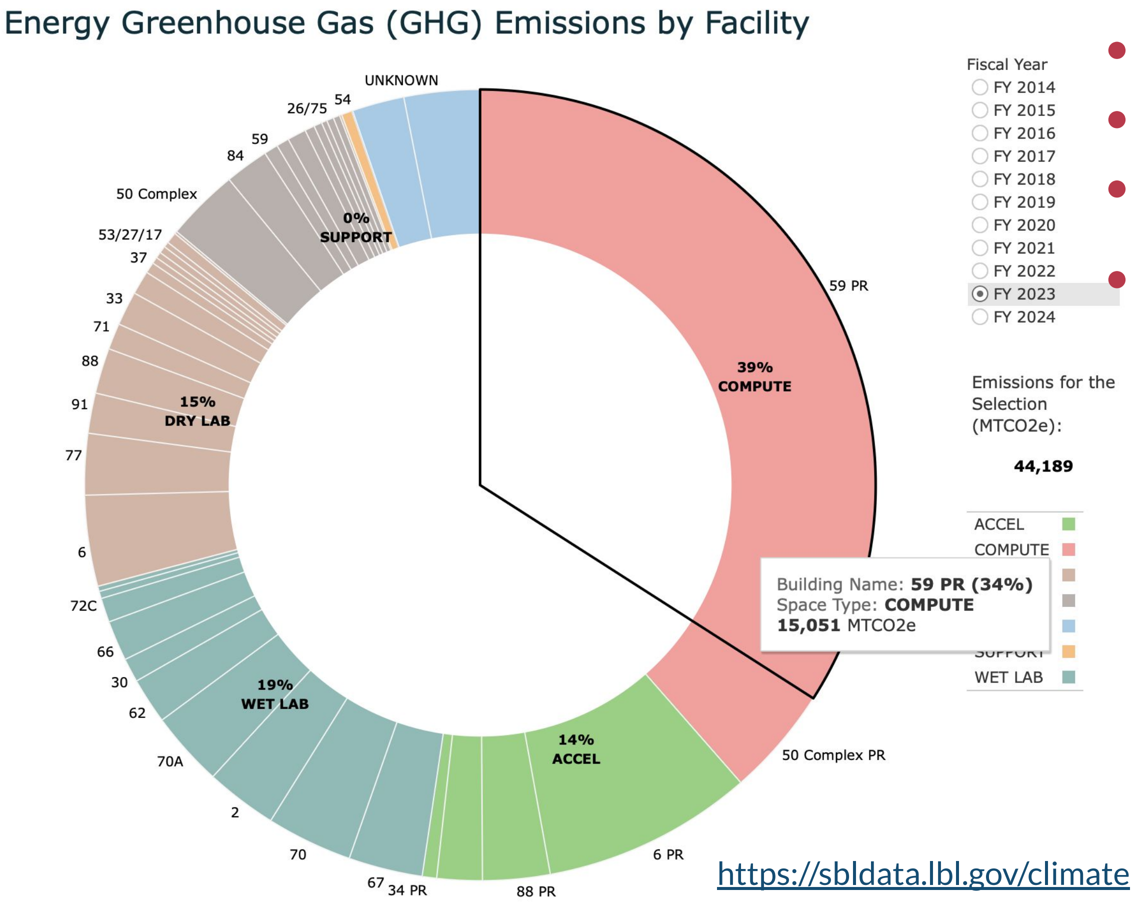
- 80 participants: 50% “local”, 50% traveled from overseas
- Average flight time: 12h/person round trip (guesstimate)
- 960 flight hours/320 passengers: 3h in Dreamliner units

Seat map (44/21/253)



In the average year, LZ computing consumes 30% of the energy expended for collaboration meeting travel (the number may be closer to 20% if we consider analysis workshops, etc.)

NERSC computing: carbon footprint



https://sbldata.lbl.gov/climate

- NERSC total compute: 15k MTCO2e
- LZ fraction: NERSC/400 = 37.5 MT
- Including the UKDC: ~60 MT CO2e
- Collaboration mtg travel from SFO:

13 hr 45 min SFO–EDI	1 stop 1 hr 38 min ORD	811 kg CO2e Avg emissions ⓘ
19 hr 29 min EDI–SFO	1 stop 6 hr 10 min ORD	799 kg CO2e Avg emissions ⓘ
5 hr 46 min SFO–BOS	Nonstop	412 kg CO2e Avg emissions ⓘ
6 hr 30 min BOS–SFO	Nonstop	422 kg CO2e Avg emissions ⓘ

NERSC vs meeting travel: carbon footprint

June meeting at Brown University

- 100 participants: 20% “local”, 20% from EU/UK
- Average flight emissions: 800 kg CO2e

January meeting at University of Edinburgh

- 80 participants: 50% “local”, 50% from US
- Average flight emissions: 800 kg CO2e

Total Collaboration meeting travel: 144 MT CO2e

→ At least 2.5x higher than total annual computing

[CO2 estimate for travel is 20% lower than power-only calculation because assumes economy seats only]

- NERSC total compute: 15k MTCO2e
- LZ fraction: NERSC/400 = 37.5 MT
- Including the UKDC: ~60 MT CO2e
- Collaboration mtg travel from SFO:

13 hr 45 min SFO–EDI	1 stop 1 hr 38 min ORD	811 kg CO2e Avg emissions ⓘ
19 hr 29 min EDI–SFO	1 stop 6 hr 10 min ORD	799 kg CO2e Avg emissions ⓘ
5 hr 46 min SFO–BOS	Nonstop	412 kg CO2e Avg emissions ⓘ
6 hr 30 min BOS–SFO	Nonstop	422 kg CO2e Avg emissions ⓘ