Computing in LZ

Elena Korolkova and Brais López Paredes

on behalf of the UKDC team:

Elena Korolkova, Vitaly Kudryavtsev
University of Sheffield

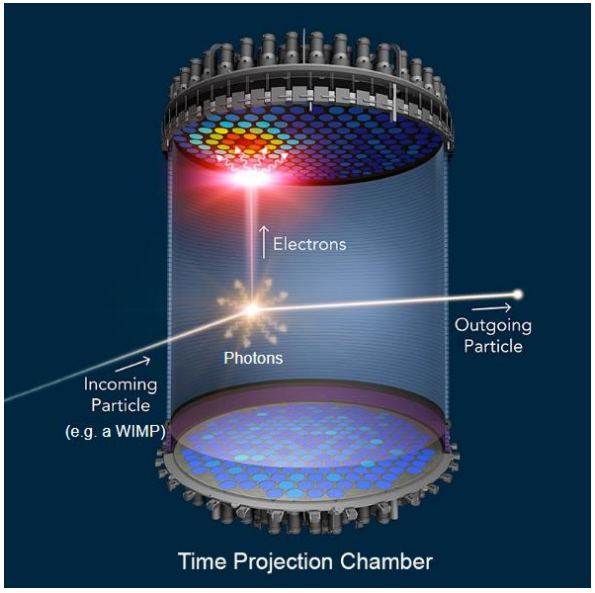
Daniela Bauer, Simon Fayer, Brais López Paredes, Alex Richards, Rob Taylor, Antonin Vacheret
Imperial College London

Luke Kreczko, Ben Krikler
University of Bristol

Chris Brew
Rutherford Appleton Laboratory, STFC

LUX-ZEPLIN (LZ) experiment

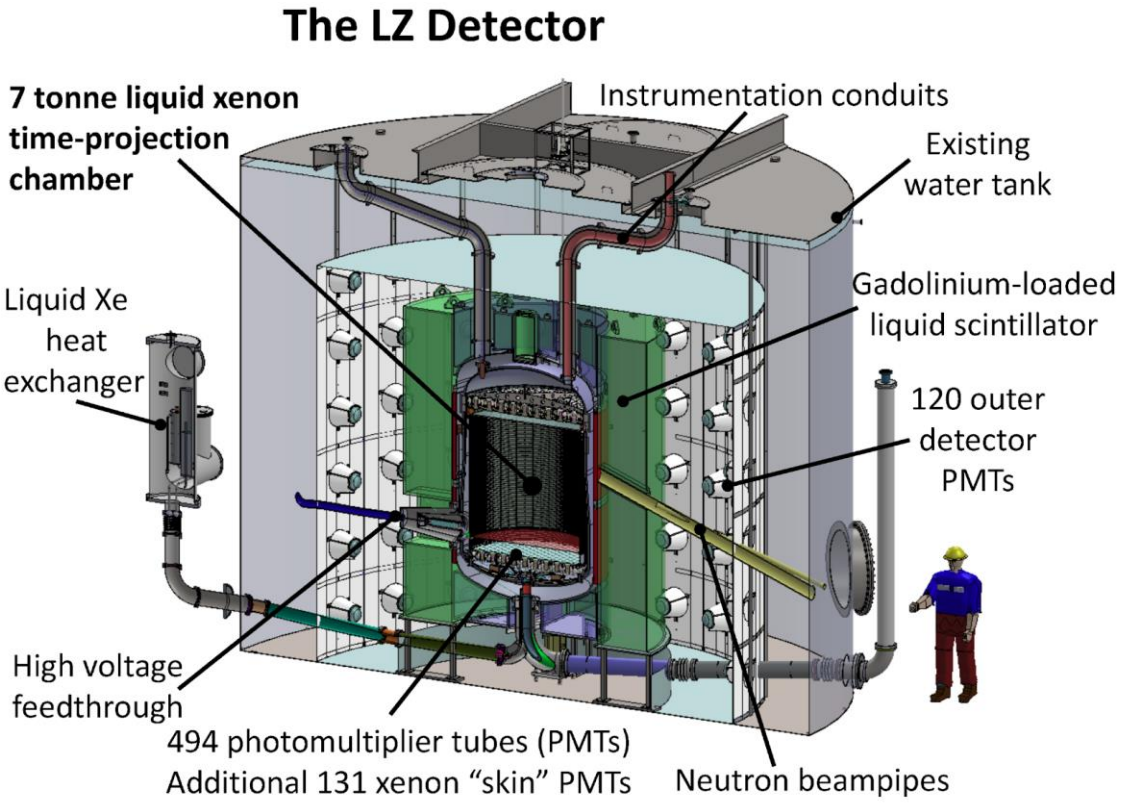
- 7 t of active liquid xenon (LXe) to search for dark matter interactions – Weakly Interacting Massive Particles (WIMPs).
- ~1 mile underground in the Sanford Underground Research Facility (SURF) in Lead, South Dakota.



- Active LXe in a 1.5m x 1.5m cylinder with an electric field – Time Projection Chamber (TPC).
- Recoling Xe nucleus causes prompt scintillation flash, followed by delayed electroluminescence light. Light signals are detected by 494 photomultiplier tubes above and below LXe.

The LZ detector

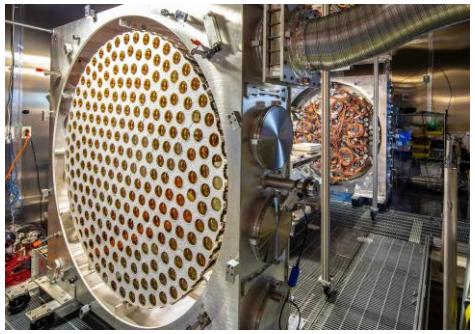
- LXe TPC:
  - 7 t active Xe
  - 5.6 t fiducial mass
- Two veto systems
  - LXe skin
  - Outer detector (20 t Gd-LS)
- Internal and external calibrations (ER and NR)
- 494 TPC, 131 Skin and 120 OD PMTs, >100 in-xenon sensors, monitor up to 2500 sensors



The LZ detector: cryostat - one of the UK deliverables.

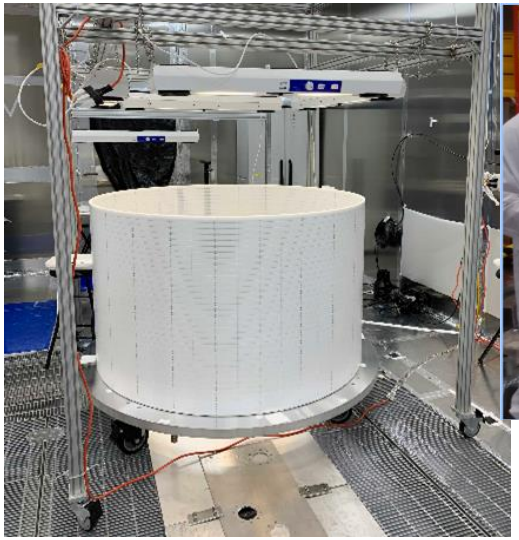
PTFE-tiled ICV

LZ is under construction now



PMT arrays

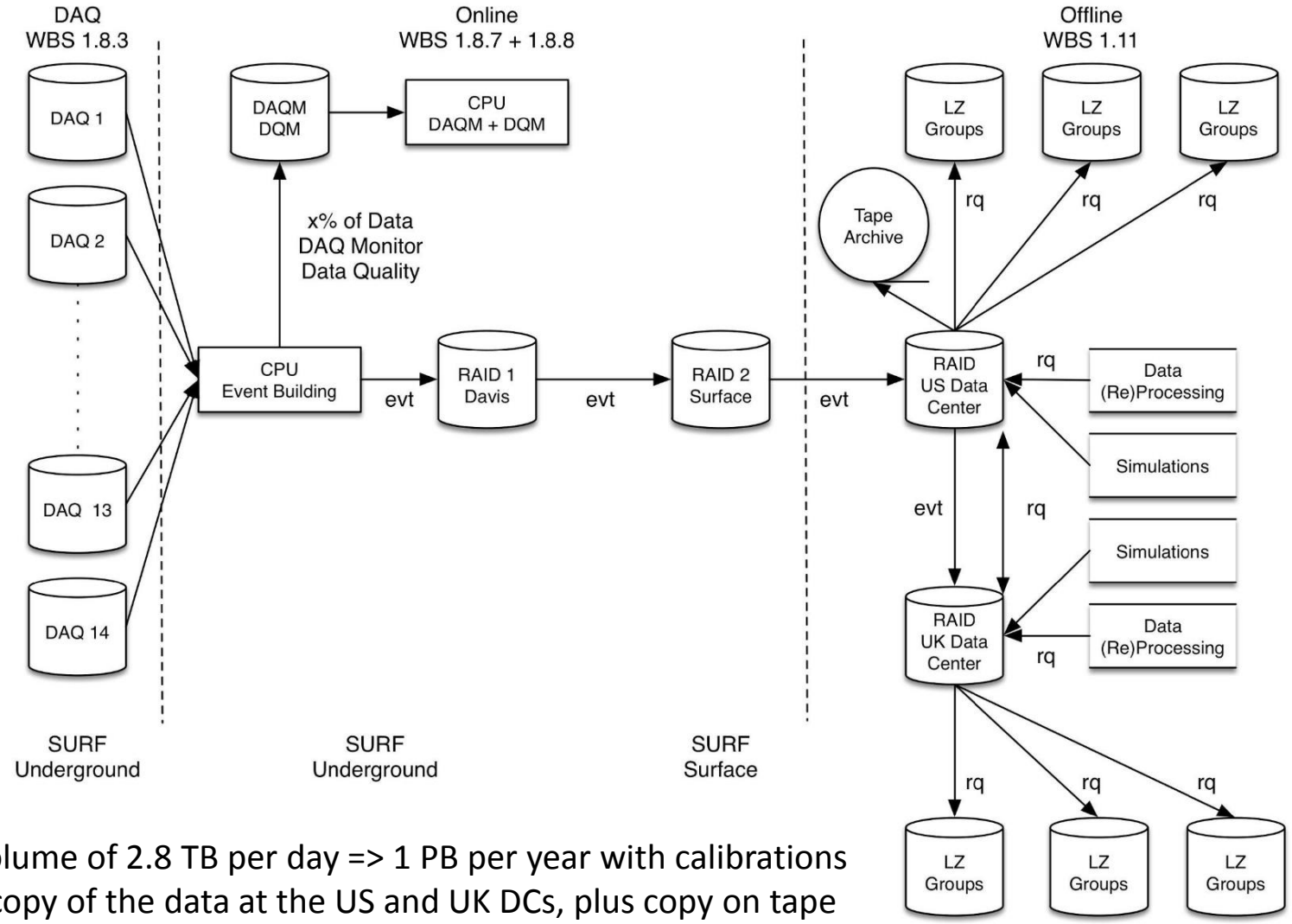
- Technical commissioning: Jan-Mar 2020
- Physics commissioning: Apr-Jun 2020
- Physics ready: Mid 2020
- Highly competitive science landscape



TPC stacking

4 large side acrylic tanks in water tank

LZ Data flow



Data volume of 2.8 TB per day => 1 PB per year with calibrations
“Live” copy of the data at the US and UK DCs, plus copy on tape

Data transfer strategy

- LZ data will be buffered locally at SURF, transmitted to primary data storage at the USDC, and mirrored to the UKDC
- Fully redundant storage of all raw and reconstructed data
- Underground and surface storage at SURF:
  - Dual staging (underground/surface), 192 TB each
  - Can accommodate 68 days of DM search data each
  - Large buffer in case of extended network outage
- Data Transfer Software (SURF-USDC-UKDC):
  - SPADE (South Pole Archival and Data Exchange) software
    - Interface with Data Catalog and job submission interface(s)
    - Can run on multiple protocols (gridftp, xrootd, globus)
    - Automatic checksum validations for all data transfers
  - Installed in NERSC and IC
    - Was tested in RALPP during MDC2 (slow speed transfers)
    - Current speed is 60 MB/s.
- Transfers can be done using dCache GridFTP Server in IC and data transfer nodes in NERSC
  - Used during Mock data challenges and for simulation transfers

LZ Data Centres

- US&UK DCs intended to provide comparable capability
- Each will host 100% of data and provide similar processing power
- UK Data Centre passed Acceptance Review in December 2017
- The US Data Centre passed Final Design Review in October 2018
- Two voms servers: Wisconsin and Imperial
- LZ cvmfs (/cvmfs/lz.opensciencegrid.org/):
  - deliver identical software builds to both data centres and users.

The UK Data Centre

- Hosted at Imperial College
- The UKDC uses the DIRAC framework for distributed computing:
  - Provides pilot-based job submission framework and data management system, has its own file catalog
  - Imperial hosts resource broker (DIRAC) for the GridPP resources (CPU & Storage)
- The UKDC delivered simulations for
  - Formal DOE “Critical Decision” reviews
  - Background reviews
  - Mock Data Challenges (MDC) 1 and 2, currently preparing for the MDC3
  - Recent sensitivity studies (arXiv:1802.06039)
  - requests from LZ groups
  - Simulated data available on UKDC storage and transferred to USDC

The UK Data Centre

- The UK data centre uses GridPP resources
  - Storage at Imperial: currently 960 TB (9.9 PB in 2025)
  - GridPP CPU resources
    - MDC2 2018 showed certain types of events (with abundant optical photons) intrinsically require 8-12 GB/core.
    - current resources are sparse, workaround by reserving multiple cores and memory for single job. Also expect a burst in reconstruction jobs during the commissioning phase in 2020.
    - IRIS resources to burst capacity during the MC production for MDC3 in 2019 early data taking in 2020
    - LZ has been granted 300 cores with 12 GB on IRIS
  - Access to the UKDC resources with grid certificates + LZ VO membership

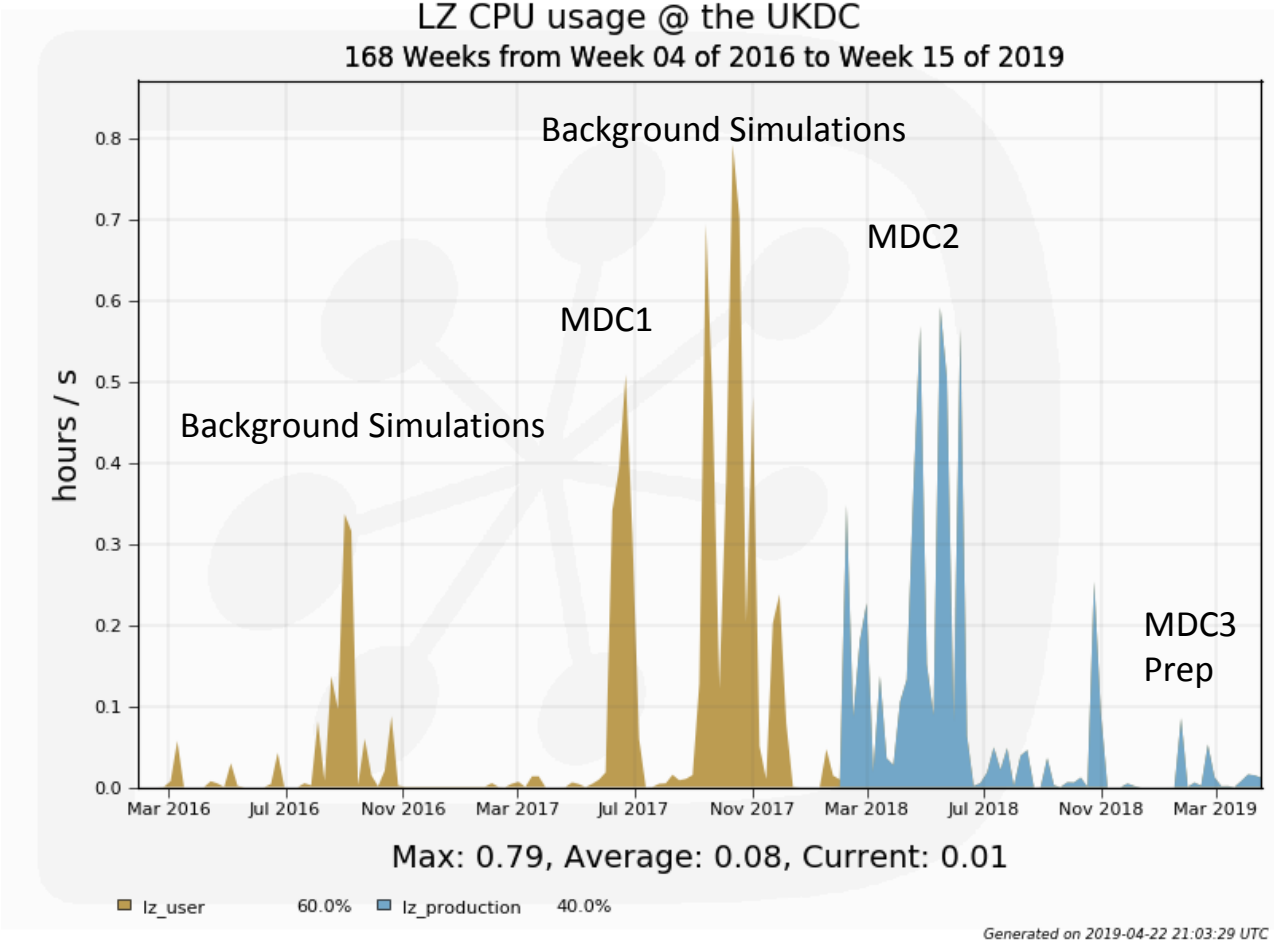
RESOURCE REQUIREMENTS

Table 11.2.2: Planned storage (in TB) and processing power by U.S. fiscal year at the U.S. and U.K. data centers.

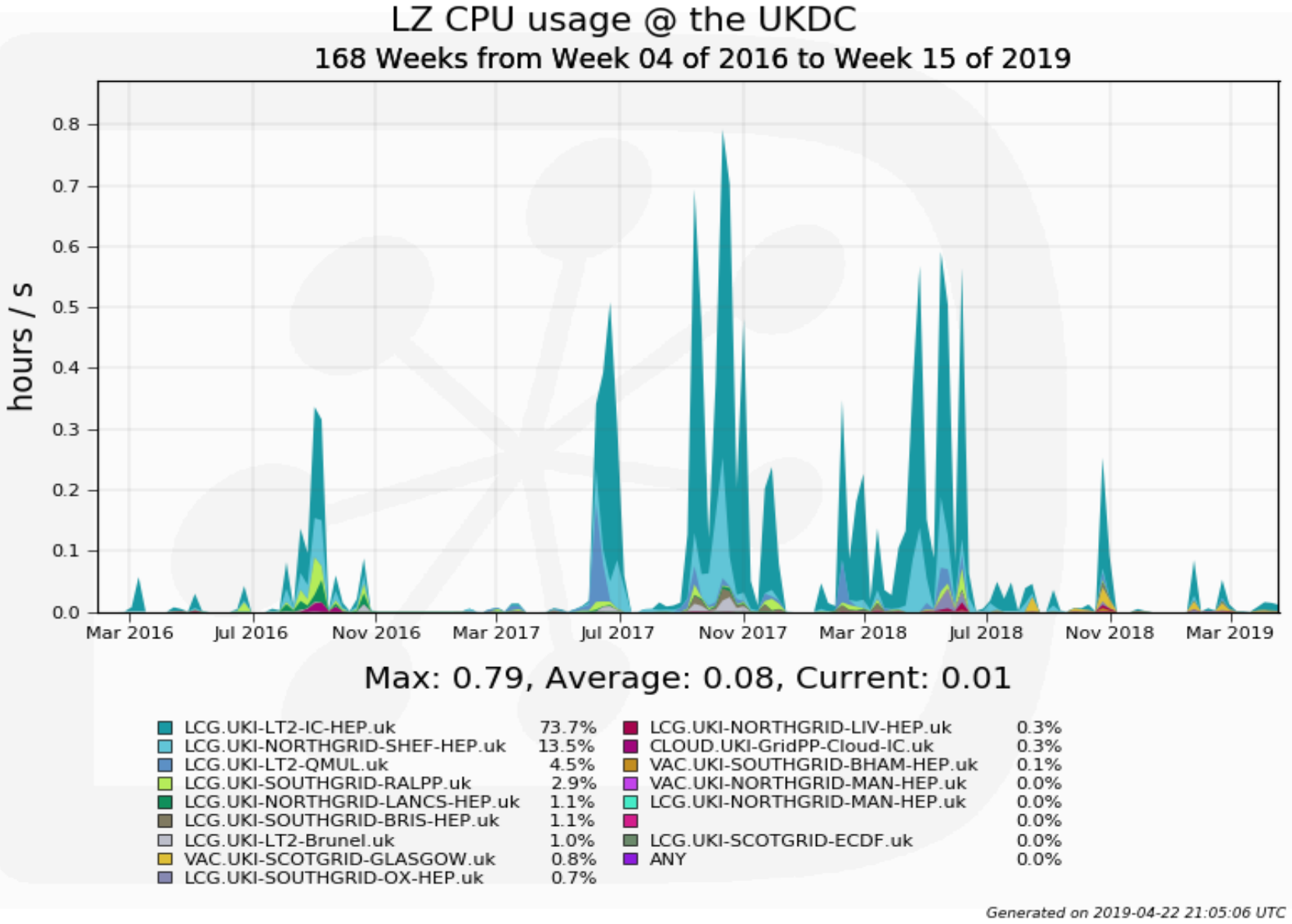
FY	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Raw data	–	–	–	–	–	560	1680	2800	3920	5040	6160
Calibration data	–	–	–	–	–	160	480	800	1120	1440	1760
Simulation data	40	80	80	100	100	200	200	200	200	200	200
Processed data	20	40	40	50	50	172	316	460	604	748	892
User data	20	40	40	50	50	55	134	213	292	371	451
Total data	80	160	160	200	200	1147	2810	4473	6136	7799	9463
USDC: Disk space	40	220	220	220	220	1360	3360	5360	7360	9360	11360
USDC: CPU cores	–	–	175	350	350	390	830	1270	1710	2150	2590
UKDC: Disk space	150	220	220	270	650	1597	3260	4923	6586	8249	9913
UKDC: CPU cores	150	175	350	350	350	390	830	1270	1710	2150	2590

960TB already provided

UKDC resource allocation agreed in principle with GridPP



Total CPU usage from January 2016 till April 2019 – 2.84 MHS06


LZ CPU usage @ the UKDC
168 Weeks from Week 04 of 2016 to Week 15 of 2019
hours / s
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Mar 2016
Jul 2016
Nov 2016
Mar 2017
Jul 2017
Nov 2017
Mar 2018
Jul 2018
Nov 2018
Mar 2019
Max: 0.79, Average: 0.08, Current: 0.01
LCG.UKI-LT2-IC-HEP.uk 73.7%
LCG.UKI-NORTHGRID-SHEF-HEP.uk 13.5%
LCG.UKI-LT2-QMUL.uk 4.5%
LCG.UKI-SOUTHGRID-RALPP.uk 2.9%
LCG.UKI-NORTHGRID-LANCS-HEP.uk 1.1%
LCG.UKI-SOUTHGRID-BRIS-HEP.uk 1.1%
LCG.UKI-LT2-Brunel.uk 1.0%
VAC.UKI-SCOTGRID-GLASGOW.uk 0.8%
LCG.UKI-SOUTHGRID-OX-HEP.uk 0.7%
LCG.UKI-NORTHGRID-LIV-HEP.uk 0.3%
CLOUD.UKI-GridPP-Cloud-IC.uk 0.3%
VAC.UKI-SOUTHGRID-BHAM-HEP.uk 0.1%
VAC.UKI-NORTHGRID-MAN-HEP.uk 0.0%
LCG.UKI-NORTHGRID-MAN-HEP.uk 0.0%
0.0%
LCG.UKI-SCOTGRID-ECDF.uk 0.0%
ANY 0.0%
Generated on 2019-04-22 21:05:06 UTC

The US Data Centre

- Uses NERSC resources:
  - CPU: 15 → 150 Mh via yearly HPC allocation
  - Storage: currently 290 TB (11 PB in 2025)
  - Tape (HPSS) space for backup ~effectively unlimited
- Production in 2018 (MDC2) performed across all available NERSC resources: pdsf, Edison, Cori Haswell, Cori KNL
  - pdsf, Edison: retired in April 2019
  - Cori Haswell (2000 nodes) oversubscribed
    - Each node has two sockets, each socket is populated with a 16-core Intel® Xeon™ Processor E5-2698 v3 ("Haswell") at 2.3 GHz , 32 cores/node; 2 hyper-threads/core, 128 GB DDR4 2133 MHz/node; 298.5 TB total aggregate memory.
  - Cori-KNL (9300 nodes) performs poorly per core
    - ~5-6x slower than on pdsf/Cori-Haswell (common issue for many HEP experiments):
    - Each node is a single-socket Intel® Xeon Phi™ Processor 7250 ("Knights Landing") processor with 68 cores/node @ 1.4 GHz . 4 hardware threads/core (272 threads total). 96 GB of memory/node, six 16 GB DIMMs (102 GiB/s peak bandwidth). Total aggregate memory (combined with MCDRAM) is 1.09 PB.
    - LZ lacks resources to optimize for KNL
- NERSC-9 is coming in 2020: will address these problems

Data and job processing software

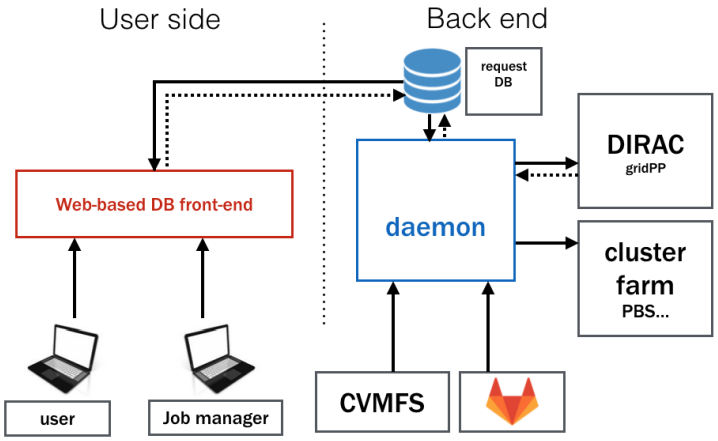
- LZ MC software based on Geant4 (Simulation) and Gaudi (Analysis) packages
- Conditions database – located at USDC with replica at Imperial (in progress)
- Data catalogue:
  - integrated SPADE catalogue w/ DIRAC data management system (Dirac catalogue)
- @ the UKDC:
  - DIRAC middleware for job submission and monitoring, meta data, grid storage
  - DIRAC API for scripting, splitting & merging jobs
  - Web-based tools to facilitate management and access to processed data
  - Job Submission Interface (JSI)
  - Event Viewer

Networking

- Transfer rates between UK and US DCs tested at ~30 Gb/s
  - above the requirement for LZ
  - Imperial already upgraded to 100Gb/s
- Average rate of compressed raw data acquisition ~ 0.3 Gb/s
- The peak rate during LED PMT calibration ~ 0.7 Gb/s
- 2 Gb/s transfer between data centres would be sufficient for LZ needs
- Long-term average < 0.5 Gb/s
- SPADE current speed is 480 Mb/sec
  - Improvements in short term roadmap
    - Third party copies
    - Multiple transfer endpoints at USDC
    - Integration with storage system checksumming

Job submission interface (JSI) @ UKDC

- The UKDC team has developed a web-based large scale production JSI
  - to simplify handling of requests for large scale submission of MC and data processing on the grid
  - reduce manual and time consuming operations to complete jobs
  - continuous development of functionalities to give more flexibility in writing and operating scripts
  - successfully tested and used during MDC1, MDC2 and in Background simulation campaigns



- Used by LZ and SOLID experiments
- The USDC uses job submission system Psquared and interface adapted from the UKDC

Infrastructure software

- Build environment is in constant evolution (moving from SL6 to CentOS7)
- Ongoing work on containers to ramp up after MDC2 (@USDC)
- LZ uses lz-git repository: http://lz-git.ua.edu/
- Software builds distributed via CVMFS to both data centres
- Continuous integration suite: testing OS, compilers, etc.
- Services at the USDC (e.g. DB, SPADE) deployed on containers
  - Possibility to deploy core software via containers as well
- Leverage tools and expertise from larger experiments

Mock Data Challenges (MDC)

- Overall goal: test functionality of the full chain from simulation all the way to physics analysis results.
- Scope: simulate a science run of LZ (including commissioning), producing science plots.
- Collaborators practicing full data analysis.
- Excellent tool to ensure physics readiness on Day 1.
- Two successful MDCs to date, preparing third
- UKDC essential in delivering tests, simulations and processing for MDC1,2&3

Summary

- The UKDC has been operational since 2016 and has been delivering (most) simulations to LZ
- LZ already has a fully functioning computing model
- Robust software infrastructure already in place
  - 100% redundancy of storage and CPU (UK + US DCs)
  - Can accommodate long network interruptions at SURF
- LZ software in advanced state of development and validation
- Two successful MDCs, upcoming MDC3
- Stress test of data centers and simulation/analysis software
- Collaborators practicing analysis well ahead of commissioning
- UK and GridPP researchers are currently and will continue to be integral in delivering LZ Computing
- **Critical need to continue to support resource allocation for LZ (as so far!); data start flowing early in 2020**

Thank you
for supporting LZ

Backup

- **Comments (continued)**
  - LZ presented (in discussions) a complete plan for operating software and computing for MDC3, CY19 simulation, LZ commissioning and Run 1, and Run 2+3. Our assessment is that the plan has no contingency on the US side, both in computing resources and in effort. The only contingency for LZ is that the UK carries a lot of the processing weight.
  - With its current computing infrastructure and software, LZ could succeed in publishing first results promptly after Run 1. However, because of the computing hardware changes at NERSC, there are also significant risks that the first publication after Run 1 could be delayed.
  - The current NERSC resource allocation is sufficient for MDC3 to succeed if exclusively CORI-I (Haswell) can be used. This would require reservations on the Haswell (traditional CPU) part. It is not sufficient, if the resources are a mix of CORI-I (Haswell) and CORI-II (KNL) resources. To increase the efficiency of running on CORI-II (KNL), changes to software and computing infrastructure would be needed.