LEVERAGING LEADERSHIP COMPUTING FOR EXPERIMENTAL SCIENCE

TOM URAM, MISHA SALIM, TAYLOR CHILDERS, MICHAEL PAPKA
ARGONNE LEADERSHIP COMPUTING FACILITY
turam@anl.gov

ASCAC - JANUARY 2020

Man's capacities have never been measured. Nor are we to judge of what he can do by precedents, so little has been tried.

Henry David Thoreau

PROGRAMMING THE DOE COMPUTING COMPLEX FOR EXPERIMENTAL AND OBSERVATIONAL SCIENTIFIC COMPUTING

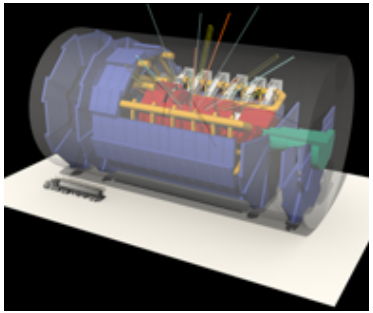
- DOE has dedicated billions of dollars in funding to building next-generation experimental and observational facilities
- Experimental and observational computing needs will increase dramatically over this period and will not be met with facility-local compute resources
- DOE is building exascale computing systems primarily aimed at simulation science
- The combined national resources dedicated to experimental science, observational science, and large-scale computing should function as a cohesive, programmable infrastructure for achieving state of the art scientific results for DOE (delivering significant science ROI on those billions)
- Can we *program* this national infrastructure?
  - We have been working to establish this new reality

ALCF SUPPORT FOR EXPERIMENTAL SCIENTIFIC COMPUTING

Allocation programs: Director's Discretionary, Early Science Program, ALCF Data Science Program

Technologies: Balsam Workflows, Cobalt Scheduler, Globus Transfer

HEP - ATLAS



2015-16

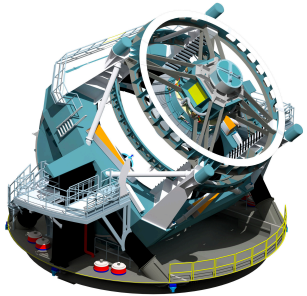
Hundreds of millions of core hours of simulation and analysis for ATLAS (DD, ADSP, ESP)

FES - DIII-D

2017

Near real-time analysis of DIII-D fusion experiment data, powered by Balsam workflows (DD); deep learning for fusion (ESP)

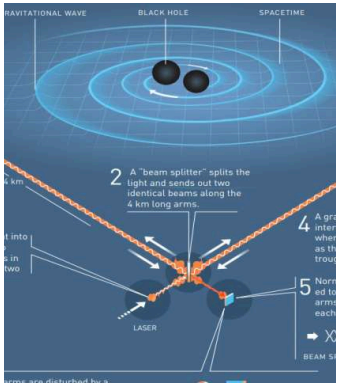
HEP/NP



2017-18

Simulation and analysis of telescope images for LSST-DESC (ADSP)

HEP - LIGO



2018-19

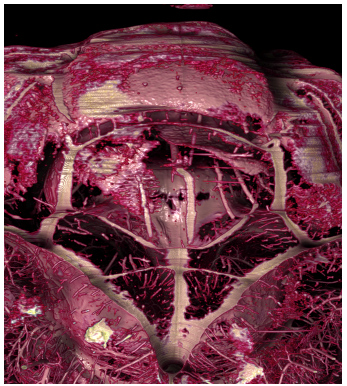
Deep learning for gravitational wave detection with LIGO (ADSP)

ALCF SUPPORT FOR EXPERIMENTAL SCIENTIFIC COMPUTING

Allocation programs: Director's Discretionary, Early Science Program, ALCF Data Science Program

Technologies: Balsam Workflows, Cobalt Scheduler, Globus Transfer

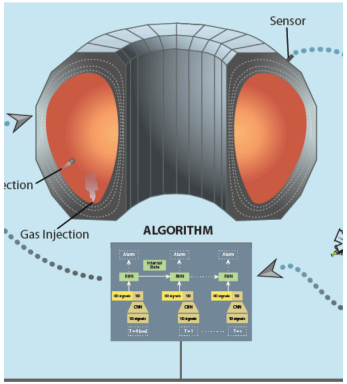
BES/OTHER



2018-

3D Reconstruction of mouse brain from electron microscopy images at Argonne (DD, ADSP, ESP)

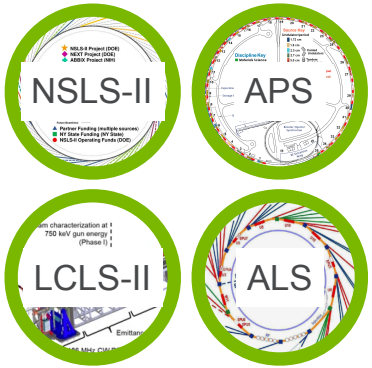
FES



2018-

Hyperparameter tuning at exascale to predict and mitigate disruption events (ESP)

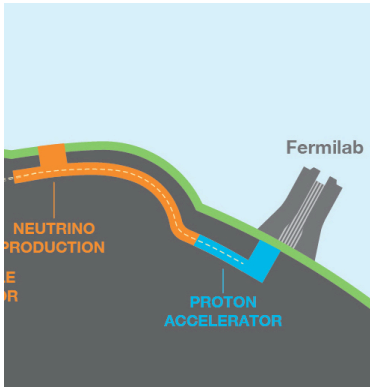
BES - LS



2019-

Online analysis of experiment data from multiple light sources using near-real-time queue on Theta (ADSP)

HEP - SBND



2019-

Large-scale simulation and deep learning for SBND/ DUNE (DD, ADSP)

ATLAS: ACCELERATING LHC SIMULATION WORKFLOWS THROUGH ADAPTATION FOR LEADERSHIP-CLASS SYSTEMS AT ALCF

ADSP PROJECT (PI: Childers, Argonne)

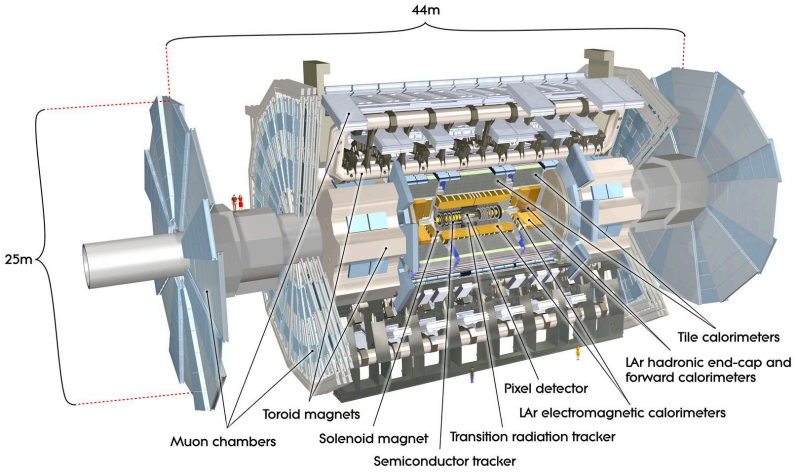
Objective: An end-to-end workflow to manage data movement and job management to significantly accelerate event generation and simulation on next-generation leadership systems

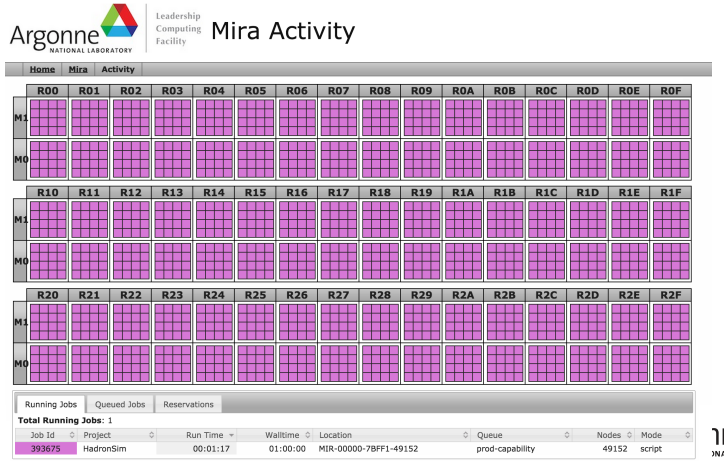
Impact: Leadership resources increases the analysis reach of LHC scientists, enabling the discovery of new particle physics

Results:

- ATLAS event generation scaled to the full Mira system (49152 nodes)
- Largest event generation jobs run by ATLAS

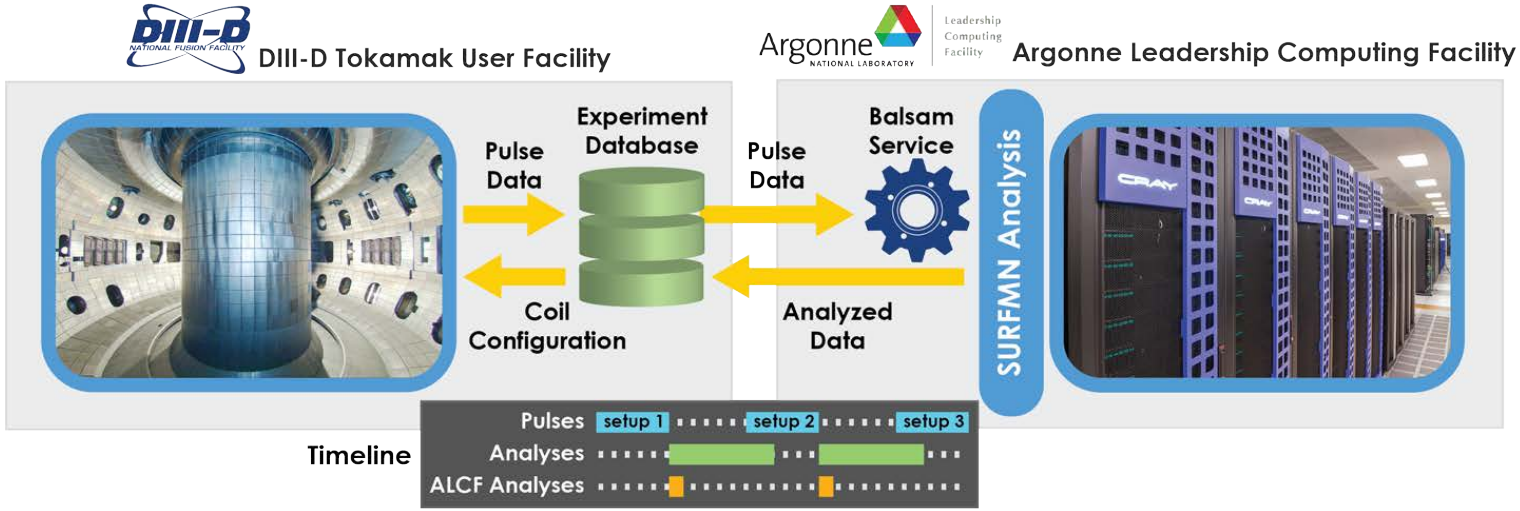
ALCF contribution to ATLAS computing ranked 6th on list of contributions *per country*





FUSION: AUTOMATIC BETWEEN-PULSE ANALYSIS OF DIII-D EXPERIMENTAL DATA

DD PROJECT (PI: Schissel, General Atomics)

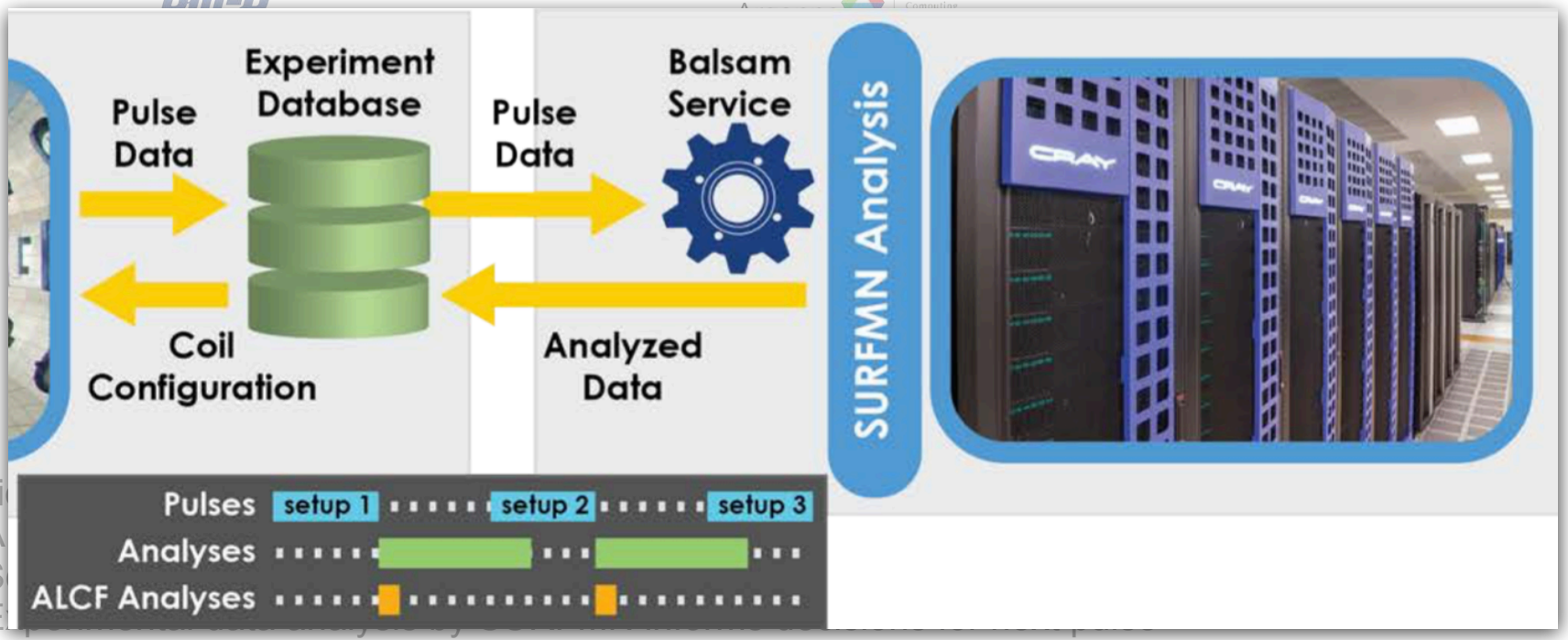


- Scientists set up DIII-D's experimental pulses every 15-20 minutes
  - A pulse is a confined plasma, hotter than the sun, lasting ~10s
  - Scientists configure the timing and current of control coils prior to each pulse
  - Experimental data analysis by SURFMN informs decisions for next pulse
- DIII-D pulses automatically trigger SURFMN analysis code to run at ALCF
- ALCF computation enables higher resolution analysis to be completed faster, improving accuracy of results for a more informed decision on control coil configuration for the next pulse

M. Kostuk, T. Uram, et al, 2nd IAEA Technical Meeting on Fusion Data Processing, Validation, and Analysis

FUSION: AUTOMATIC BETWEEN-PULSE ANALYSIS OF DIII-D EXPERIMENTAL DATA

DD PROJECT (PI: Schissel, General Atomics)



- Sci
  - A
  - S
  - E
- DIII-D pulses automatically trigger SURFMN analysis code to run at ALCF
- ALCF computation enables higher resolution analysis to be completed faster, improving accuracy of results for a more informed decision on control coil configuration for the next pulse

M. Kostuk, T. Uram, et al, 2nd IAEA Technical Meeting on Fusion Data Processing, Validation, and Analysis

REALISTIC SIMULATIONS OF THE LSST SURVEY AT SCALE

ADSP PROJECT (PI: Heitmann, Argonne)

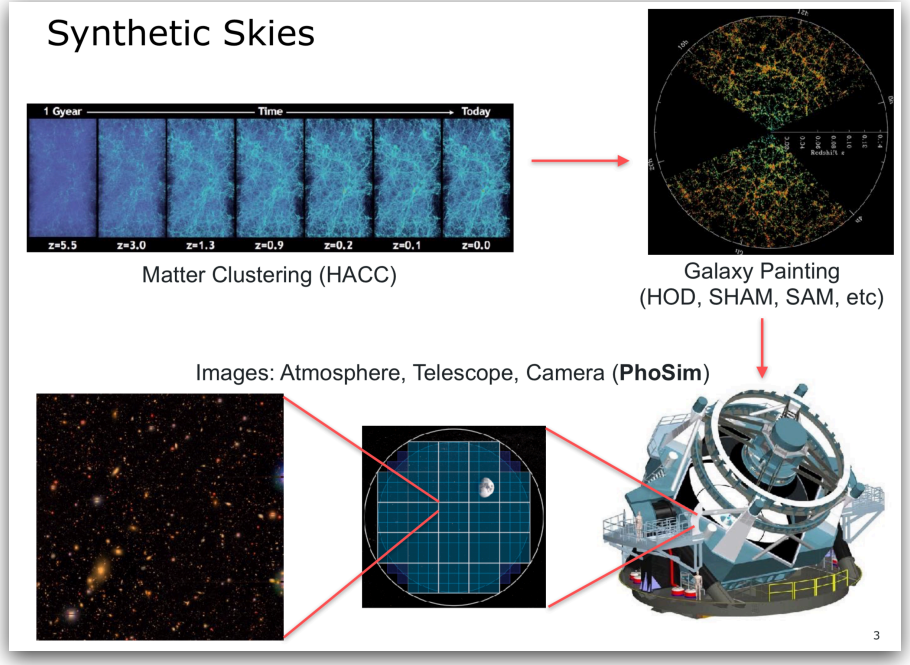
Objective: Development and execution of end-to-end workflow from simulation to the creation of sky maps with realistic galaxies

Impact: Deliver largest & most detailed synthetic sky maps ever created ready for the first data from LSST.

Approach

- Create a virtual survey with images almost indistinguishable from real LSST observations
- Develop an end-to-end pipeline for LSST data processing and analysis on ALCF supercomputers

Has run hundreds of millions of core hours at ALCF and NERSC to produce multiple years of simulated survey images

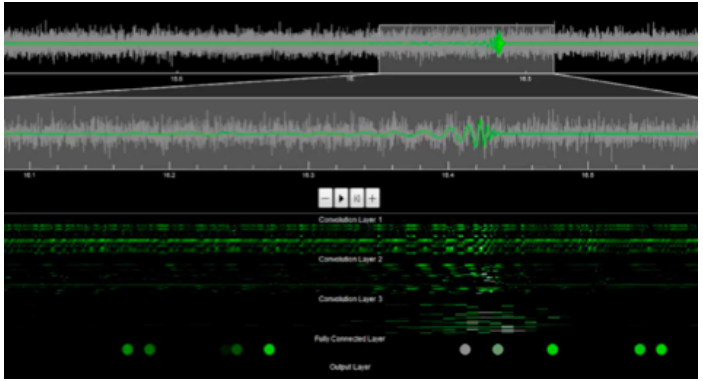


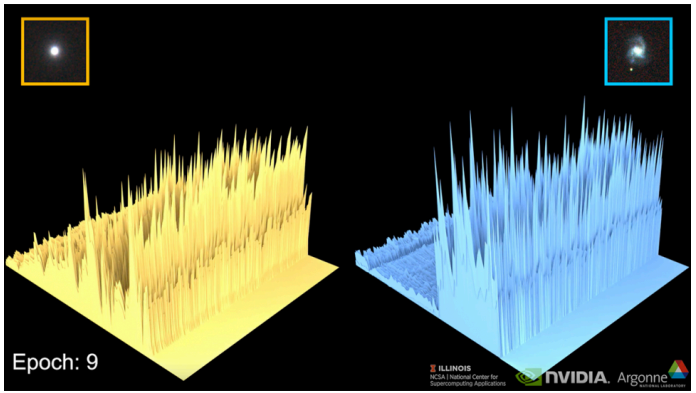
DEEP LEARNING AT SCALE FOR MULTIMESSENGER ASTROPHYSICS THROUGH THE NCSA-ARGONNE COLLABORATION

ADSP PROJECT (PI: Huerta, NCSA)

Objective: Novel data-parallel deep learning to fuse HPC and AI for multi-messenger astrophysics
Impact: Improve detection, and update methods in multiple domains, including gravitational waves (LIGO) and optical transients (LSST)
Results:

- Deep learning models running on Theta have been successful in identifying gravitational wave detection events; these models continue to be refined and extended
- Developing neural network visualization techniques toward interpreting what the trained models have learned





HIGH ENERGY PHYSICS: DEVELOPING HIGH PERFORMANCE COMPUTING APPLICATIONS FOR LIQUID ARGON NEUTRINO DETECTORS

ADSP PROJECT (PI: Szelc, UManchester)

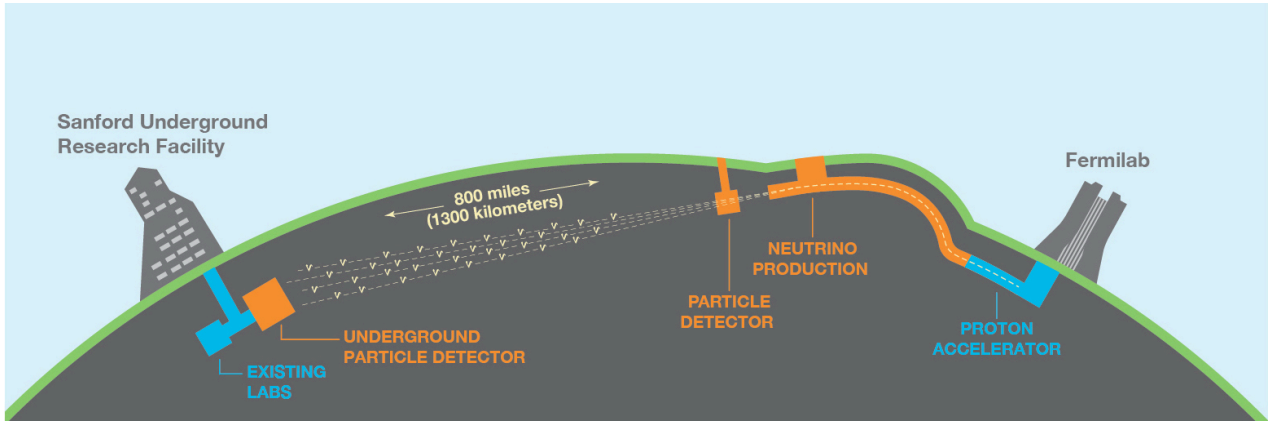
Objective: ML-accelerated simulation for next-generation neutrino physics experiments (SBND, DUNE)

Impact:

- Prepares neutrino codes to use HPC systems, delivering much-needed resources to future experiments
- Accelerates simulation through integrated ML, development of surrogate models

Results:

- Simulation and ML codes running on Theta, using Balsam



ENABLING CONNECTOMICS AT EXASCALE TO FACILITATE DISCOVERIES IN NEUROSCIENCE

ESP PROJECT (PI: Ferrier, Argonne)

Objective: Establish a pipeline to reconstruct a connectome from electron microscopy images (*connectome*: comprehensive map of neural connections in a brain; the wiring diagram of neurons)

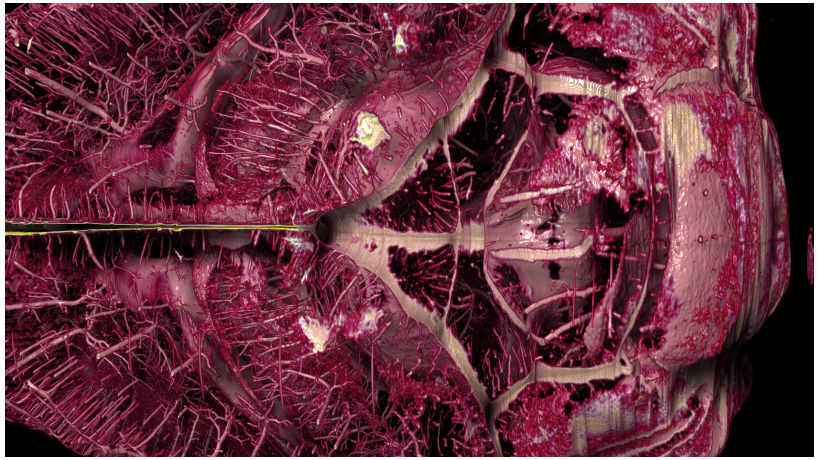
Impact:

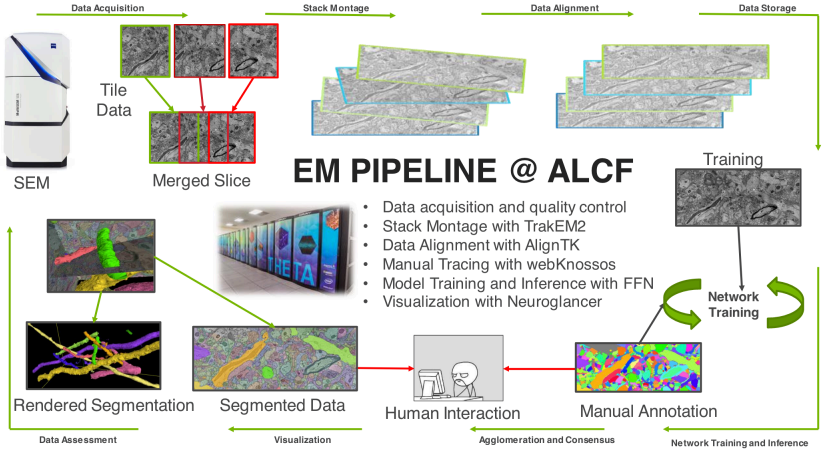
- Understand neurodegenerative diseases
- Study learning and aging and consciousness
- Improve neural network design
- Inform the design of neuromorphic chips

Results:

- Software running on Theta for stitching EM tiles, aligning resulting sections, training neural network on human annotations of neuron structure, large-scale inference for segmentation
- Initial pipeline operational with increasingly larger samples (fly, mouse, octopus, primate)

Mouse brain vasculature imaged at APS





ACCELERATED DEEP LEARNING DISCOVERY IN FUSION ENERGY SCIENCE

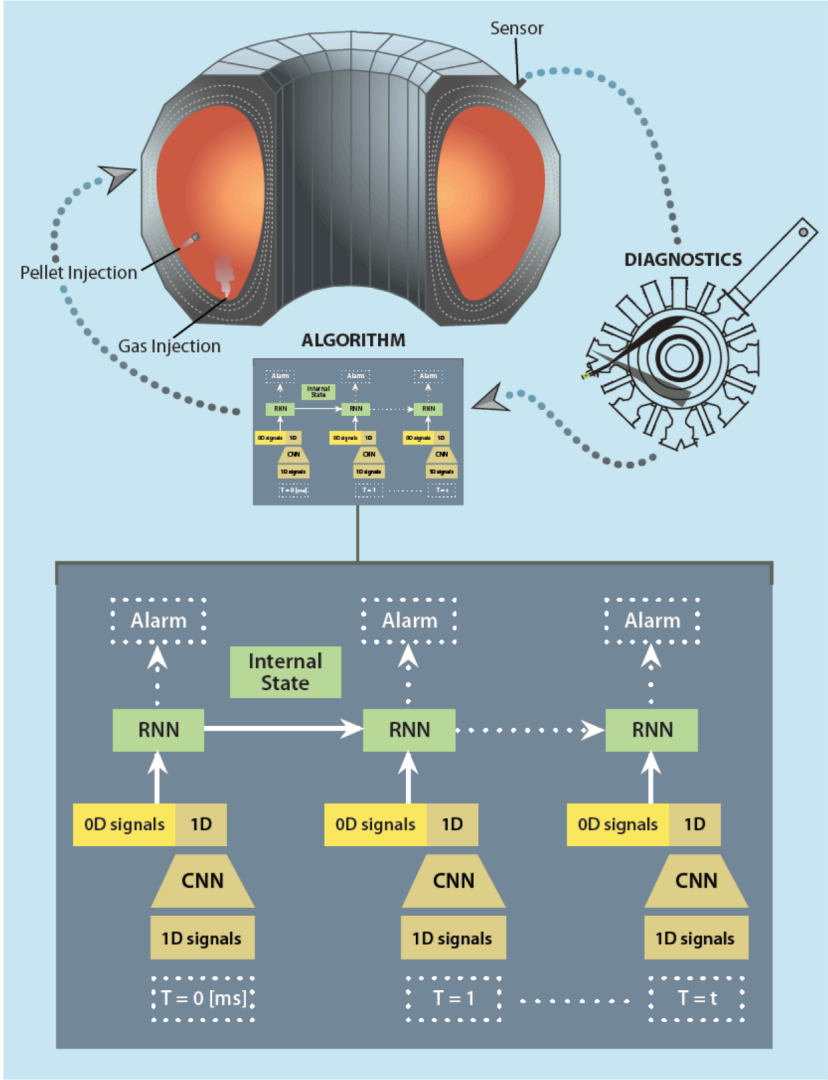
AURORA ESP PROJECT (PI: Tang, PPPL)

Objective: Reliably predict and avoid large-scale major disruptions in magnetically-confined tokamak systems such as the EUROfusion Joint European Torus (JET) today and the burning plasma ITER device in the near future

Impact: Dramatically accelerate progress in predictive modeling of clean energy/fusion R&D. Scaling this software to the largest computational scale will help enable impactful new science advances

Results:

- Measuring performance of network training on Theta
- Adding new training datasets to improve results
- Working to characterize failure modes of the model
- Configuring hyperparameter optimization with model (DeepHyper)



BALSAM WORKFLOW FRAMEWORK

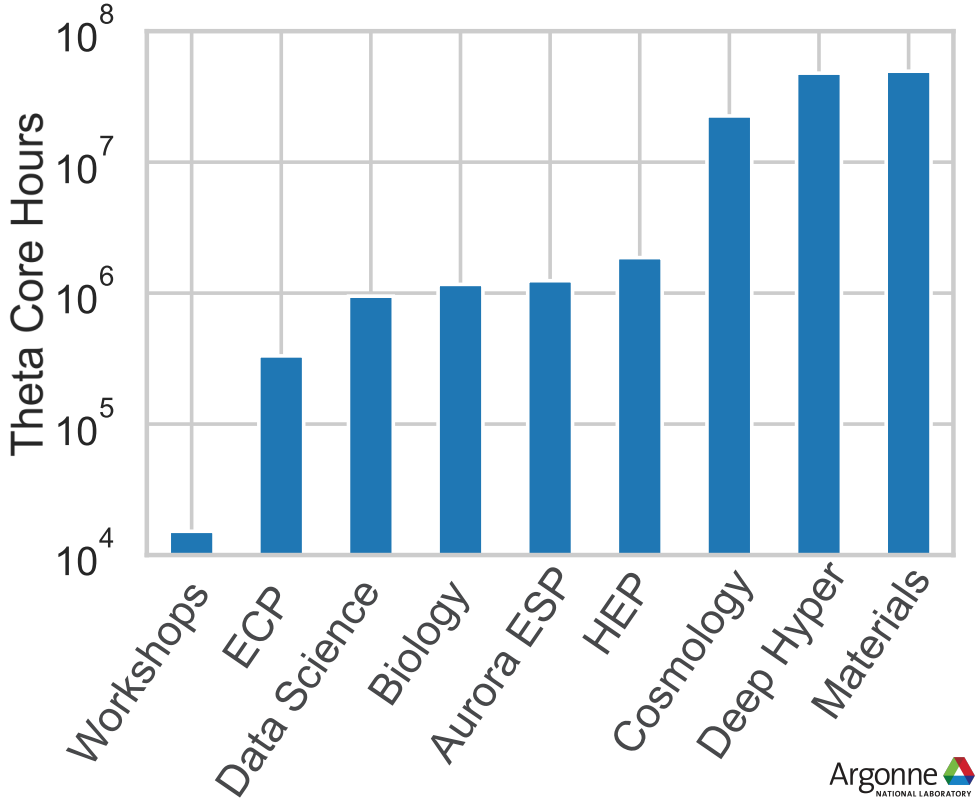
- ALCF is developing **workflow software** to help users manage the inherent complexity of large job campaigns on ALCF systems
- Frees users to concentrate on their science instead of computing
- Provides **unlimited queue depth**
- Automates and optimizes execution of job ensembles
- Python API and command-line interfaces require no modifications to applications
- Balsam has been used by multiple projects to run **hundreds of millions of compute hours on ALCF systems**

I will now make the precious balsam with which we shall cure ourselves in the twinkling of an eye.
-- Don Quixote

BALSAM USAGE AT ALCF

125M Theta core-hours between Sept 2018 - Sept 2019

- Users: 48
- Projects: 28
- Top usage categories:
  - Materials Science (39%)
  - DeepHyper (38%)
  - Cosmology (18%)



LEVERAGING LEADERSHIP COMPUTING FOR LIGHT SOURCE COMPUTING

INTEGRATING COMPUTING FACILITIES AND LIGHT SOURCES

a brief history

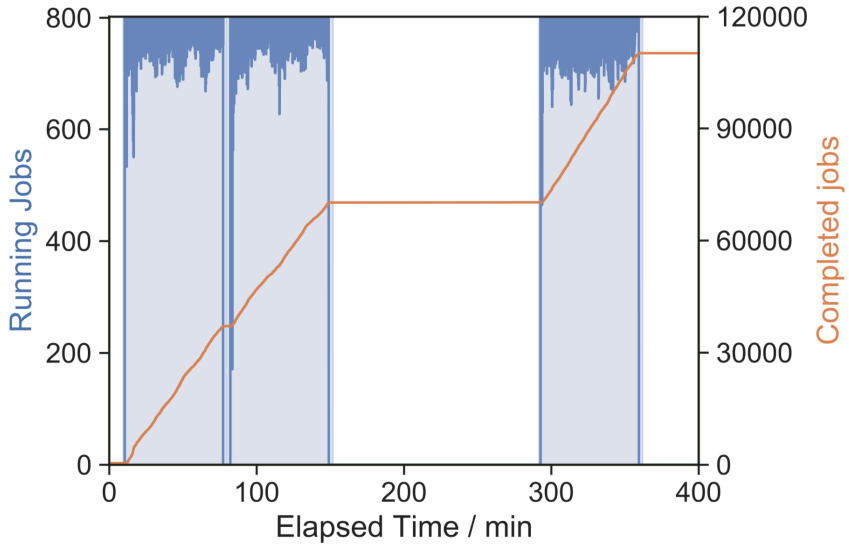
- Oct 2018: Ben Brown formed group across computing facilities to discuss support for data intensive experimental and observational scientific computing
- Nov 2018: Directors Meeting at SC18
- Dec 2018: Follow-up meeting
- Jan 2019: Group meeting
- Jun 2019: Light Source Directors and Computing Directors Meeting
- Jun 2019: Light Source and Computing Facilities Data Working Group
- Oct 2019: ALS User Meeting: Computational Challenges Across Light Sources
- Nov 2019: Facilities Directors meeting at SC19

NEAR REAL-TIME PROCESSING OF LIGHT-SOURCE WORKLOADS AT ALCF WITH BALSAM WORKFLOWS

- In an effort to simulate online processing of APS data at ALCF, APS provided 200,000+ datasets
- ALCF established a near-real-time queue fronted by a Balsam service; compute nodes are obtained from this queue as needed to accommodate incoming jobs
- Jobs were run in two scenarios:
  - **Bulk execution:** Analysis tasks were defined *en masse* in the Balsam database and run within scheduled jobs
  - **Streaming execution:** Datasets were drawn from this pool at random, transferred to ALCF, analyzed on Theta compute nodes from the near-real-time queue, and the results were transferred back to APS.

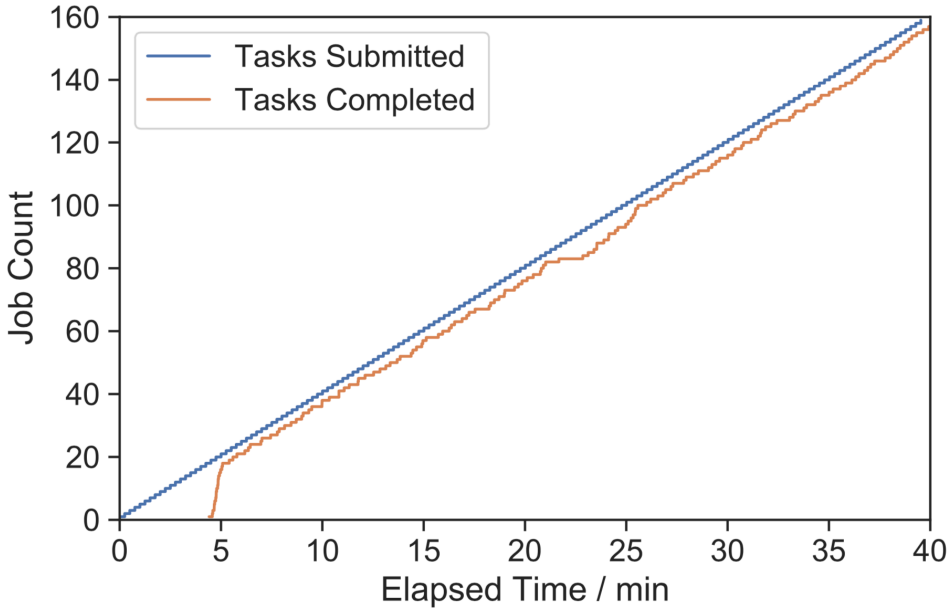
APS XPCS-EIGEN: BULK EXECUTION

- Selected 115,000 datasets for transfer to ALCF
- Programmatically defined Balsam jobs for the XPCS analyses
- Launched dedicated 802-node jobs within which Balsam executed the individual jobs
- Finished all jobs in 4 hours wall-clock time



APS XPCS-EIGEN: STREAMING EXECUTION

- Simulated online XPCS processing by repeatedly submitting a single Balsam job with a 40GB† dataset
- Balsam manages transfer of input data from APS via Globus, submits job to Theta, and, upon completion, transfers the output file back to APS
- As the number of ready jobs increases, Balsam grows its pool of Theta nodes and schedules more jobs
- High responsiveness is achieved due to near-real-time queue defined by ALCF



† 40GB dataset was the largest of the 200K datasets obtained

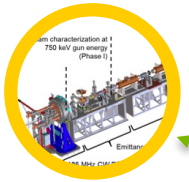
Balsam: Near Real-Time Experimental Data Analysis on Supercomputers, M. Salim, T. Uram, J. T. Childers, V. Vishwanath and M. Papka, SC2019

ALCF SUPPORT OF LIGHT SOURCE COMPUTING

Balsam enables **transparent access** to remote ALCF resources
- identity management
- simplified scheduler API
- high-speed transfers via Globus

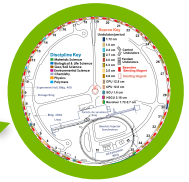
ALCF Theta (11.7 PetaFLOPs)

LCLS-II



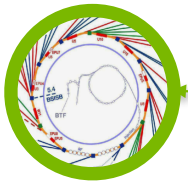
SPI workloads reaching 1-100PF by 2021

APS



XPCS workloads reaching 0.1PF by 2021

XPCS workloads reaching 0.1PF by 2021



ALS

XPCS workloads reaching 2.5PF by 2021

NSLS-II

NEAR REAL-TIME PROCESSING OF LIGHT-SOURCE WORKLOADS AT ALCF WITH BALSAM WORKFLOWS

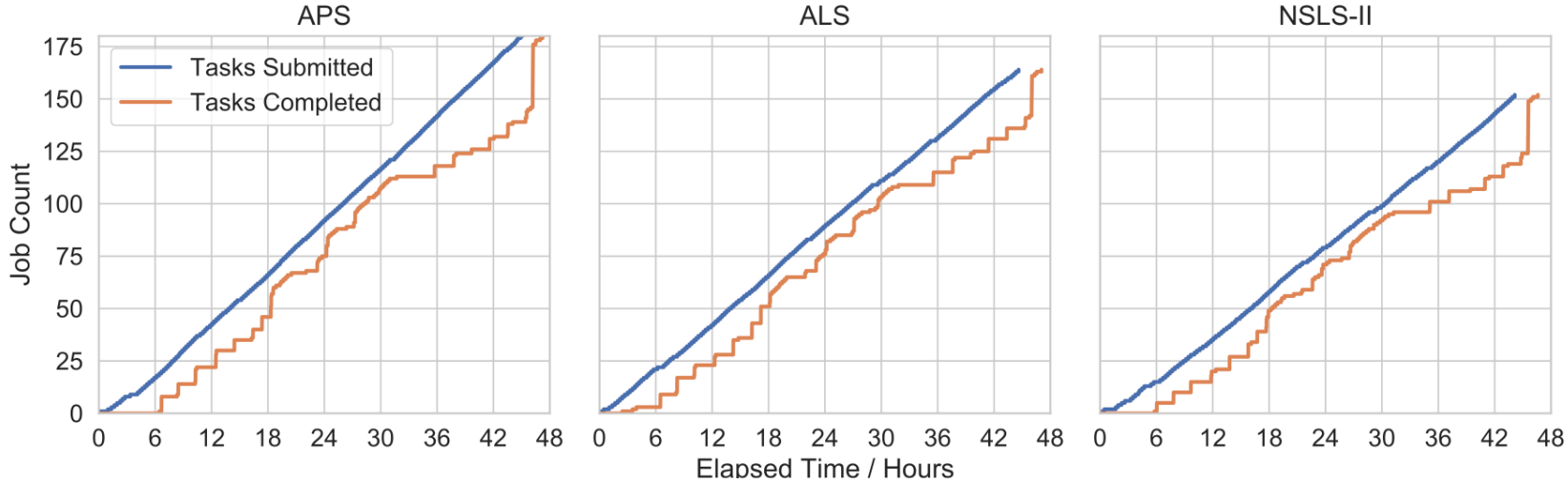
Experiment

1. Transfer 40GB† input dataset from APS, ALS, and NSLS-II
2. Process data with XPCS-Eigen using **near-real-time queue** on Theta
3. Transfer results to originating site

Results

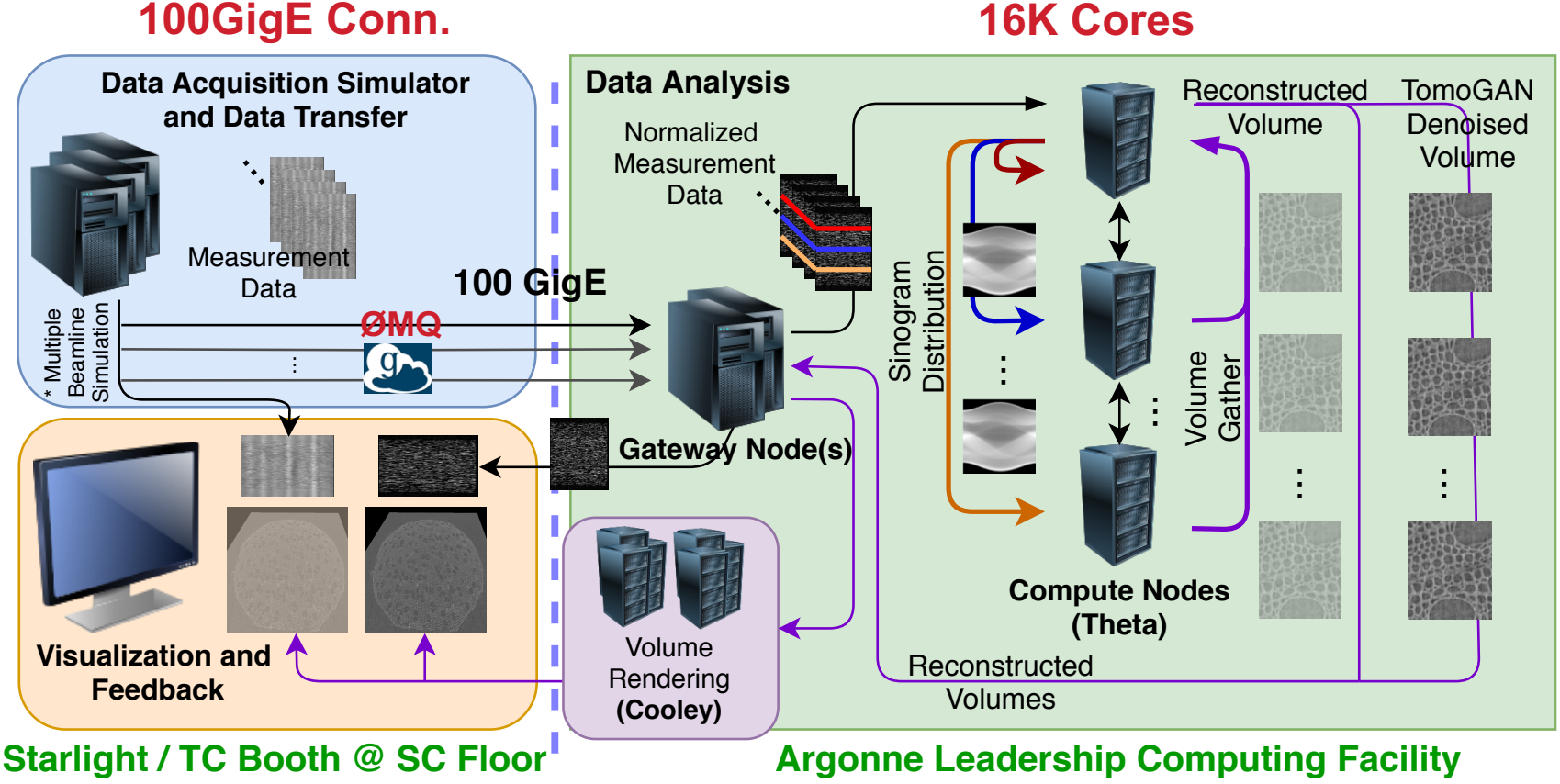
and simultaneously

- Continuously executed for **48+ hours**
- Transferred 23TB input data from APS/ALS/NSLS-II to ALCF (average bandwidth: 376Mbps)
- Analyzed 500+ datasets
- Transferred 179GB output data from ALCF to APS/ALS/NSLS-II (average bandwidth: 38Mbps)



† 40GB dataset was the largest of the 200K datasets obtained

SC19 TECHNOLOGY CHALLENGE DEMONSTRATION (winner)



* 100GigE network enables simulation of 10 beamlines each with 10GigE detector

NEXT STEPS WITH LIGHT SOURCES

- **XPCS Production:** We are working to deploy an XPCS service on Theta, to which APS users can easily submit jobs, based on the established Balsam setup
- **Engage more applications+beamlines:** Meetings with light source collaborators continue toward enabling more applications and beamlines at each light source
- **Maintain reliability** of infrastructure while improving performance
  - Improve networking in collaboration with ESNet
  - Improve analysis code performance in collaboration with light source staff (already working to improve XPCS-Eigen)
- **Meet the computational demand** of multiple light sources simultaneously (might exceed availability)
  - At present, we have nominally more computing available than the light sources can use
  - Near term, we plan to integrate their jobs as much as possible, enabling science and discovering where this model might break
  - Longer term (Aurora timeframe), we will have much more availability, and will need to adapt

LIGHT SOURCE COMPUTING AND DATA OUTLOOK

*It is estimated that the BES Light Sources will generate in the **exabyte (EB) range of data (per year)**, require **tens to 1,000 PFLOPS of peak on-demand computing resources**, which will only be available at DOE High-End Computing (HEC) facilities, and utilize **billions of core hours per year by 2028**. Unified solutions across the facilities are required in order to leverage efficiencies of scale, and to provide facility users with the ability to easily and transparently manipulate data across the complex. **The [light source] facilities do not have the operations resources to address these needs.***

Source: BES User Facilities Data Management and Analysis Resource Needs BES/ASCR Data Call Combined Response from the ALS, APS, LCLS, NSLS-II, and SSRL, May 2019

ASCR SUPPORT OF LIGHT SOURCE COMPUTING

Balsam will enable transparent switching between light-source-local computing resources and remote **ASCR resources**

- **intelligent compute resource selection**
- **portable, containerized execution**
- identity management
- high-speed transfers via Globus

ALCF Aurora

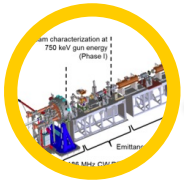
OLCF Summit

NERSC Saul

OLCF and NERSC via discretionary allocations

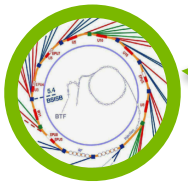
ASCR Computing

LCLS-II



SPI workloads reaching 1EF by 2028

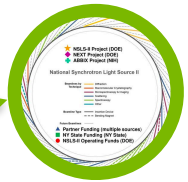
ALS



XPCS workloads reaching 10PF by 2028

XPCS workloads reaching 45PF by 2028

NSLS-II



XPCS workloads reaching 50PF by 2028

APS

SUMMARY

- ALCF has a history of enabling experimental and observational scientific computing, alongside its traditional simulation workloads
- Recent efforts with light sources have shown that we can leverage the national computing infrastructure to meet the needs of the experimental community
- **Enable transparent online processing of data from diverse experimental and observational facilities at ALCF and DOE compute facilities**
  - Continue to engage with light source and non-light source DOE experimental and observational facilities
  - Work with colleagues at other compute facilities to enable experimental sciences (ALCC proposal submitted together with NERSC/OLCF)
  - Continue building workflow software to close the gap and integrate large-scale computing with experimental and observational facilities **with low effort**