Future of Computational Infrastructures: Exascale Computing and an Integrated Research Infrastructure

Barbara Helland, Associate Director

Advanced Scientific Computing Research

U.S. DEPARTMENT OF ENERGY | Office of Science

SC's Computational Infrastructure is changing

New systems are designed to support simulations, AI/ML and Data analysis

System attributes	ALCF Now	NERSC Now	OLCF Now	NERSC Pre-Exascale	ALCF Pre-Exascale	OLCF Exascale	ALCF Exascale
Name (Planned) Installation	**Theta 2016**	**Cori 2016**	**Summit 2017-2018**	**Perlmutter (2020-2021)**	**Polaris (2021)**	**Frontier (2021-2022)**	**Aurora (2022-2023)**
System peak	> 15.6 PF	> 30 PF	200 PF	> 120PF	35 – 45PF	>1.5 EF	≥ 1 EF DP sustained
Peak Power (MW)	< 2.1	< 3.7	10	6	< 2	29	≤ 60
Total system memory	847 TB DDR4 + 70 TB HBM + 7.5 TB GPU memory	~1 PB DDR4 + High Bandwidth Memory (HBM) + 1.5PB persistent memory	2.4 PB DDR4 + 0.4 PB HBM + 7.4 PB persistent memory	1.92 PB DDR4 + 240TB HBM	> 250 TB	4.6 PB DDR4 +4.6 PB HBM2e + 36 PB persistent memory	> 10 PB
Node performance (TF)	2.7 TF (KNL node) and 166.4 TF (GPU node)	> 3	43	> 70 (GPU) > 4 (CPU)	> 70 TF	TBD	> 130
Node processors	Intel Xeon Phi 7320 64-core CPUs (KNL) and GPU nodes with 8 NVIDIA A100 GPUs coupled with 2 AMD EPYC 64-core CPUs	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	2 IBM Power9 CPUs + 6 Nvidia Volta GPUs	CPU only nodes: AMD EPYC Milan CPUS; CPU-GPU nodes: AMD EPYC Milan with NVIDIA A100 GPUs	1 CPU; 4 GPUs	1 HPC and AI optimized AMD EPYC CPU and 4 AMD Radeon Instinct GPUs	2 Intel Xeon Sapphire Rapids and 6 Xe Ponte Vecchio GPUs
System size (nodes)	4,392 KNL nodes and 24 DGX-A100 nodes	9,300 nodes 1,900 nodes in data partition	4608 nodes	> 1,500(GPU) > 3,000 (CPU)	> 500	> 9,000 nodes	> 9,000 nodes
CPU-GPU Interconnect	NVLINK on GPU nodes	N/A	NVLINK Coherent memory across node	PCIe		AMD Infinity Fabric Coherent memory across the node	Unified memory architecture, RAMBO
Node-to-node interconnect	Aries (KNL nodes) and HDR200 (GPU nodes)	Aries	Dual Rail EDR-IB	HPE Slingshot NIC	HPE Slingshot NIC	HPE Slingshot	HPE Slingshot
File System	200 PB, 1.3 TB/s Lustre 10 PB, 210 GB/s Lustre	28 PB, 744 GB/s Lustre	250 PB, 2.5 TB/s GPFS	35 PB All Flash, Lustre	N/A	695 PB + 10 PB Flash performance tier, Lustre	≥ 230 PB, ≥ 25 TB/s DAOS

U.S. DEPARTMENT OF ENERGY | Office of Science

Frontier: https://www.olcf.ornl.gov/frontier/
Aurora: https://www.alcf.anl.gov/aurora

ASCR Computing Upgrades At-a-Glance
November 24, 2020

The Exascale Computing Project (ECP) enables US revolutions in technology development; scientific discovery; healthcare; energy, economic, and national security

ECP Mission

Develop exascale-ready applications and solutions that address currently intractable problems of strategic importance and national interest.

Create and deploy an expanded and vertically integrated software stack on DOE HPC exascale and pre-exascale systems, defining the enduring US exascale ecosystem.

Deliver **US HPC vendor technology advances and deploy ECP products** to DOE HPC pre-exascale and exascale systems.

ECP Vision

Deliver **exascale simulation and data science innovations and solutions to national problems** that enhance US economic competitiveness, change our quality of life, and strengthen our national security.

- Co-Funded by SC/ASCR and NNSA/ASC
- 7 year project – $1.8B
- 6 lead labs: ORNL, ANL, LBNL, LLNL, SNL, LANL
- More than 80 research teams
  - >1000 researchers
  - Drawn heavily from 17 DOE labs plus national universities and US companies (100+ each)

Efficiently utilizing GPUs goes far beyond typical code porting

Port Code

- Rewrite, profile, and optimize
- Memory coalescing
- Loop ordering
- Kernel flattening

Adapt Numerics

- Reduced synchronization
- Reduced precision
- Communication avoiding

Adapt Models

- Mathematical representation
- “On the fly” recomputing vs. lookup tables
- Prioritization of new physical models

ECP is delivering a curated software ecosystem: Extreme-scale Scientific Software Stack (E4S)

- E4S: HPC software ecosystem – a curated software portfolio
- A **Spack-based** distribution of software tested for interoperability and portability to multiple architectures
- Available from **source**, **containers**, **cloud, binary caches**
- Leverages and enhances SDK interoperability thrust
- Not a commercial product – an open resource for all
- Growing functionality: Aug 2022: E4S 22.08 – 100+ full release products

https://spack.io

Spack lead: Todd Gamblin (LLNL)

https://e4s.io

E4S lead: Sameer Shende (U Oregon)

Community Policies Commitment to software quality	DocPortal Single portal to all E4S product info	Portfolio testing Especially leadership platforms
Curated collection The end of dependency hell	Quarterly releases Release 1.2 – November	Build caches 10X build time improvement
Turnkey stack A new user experience	https://e4s.io	E4S Strategy Group US agencies, industry, international

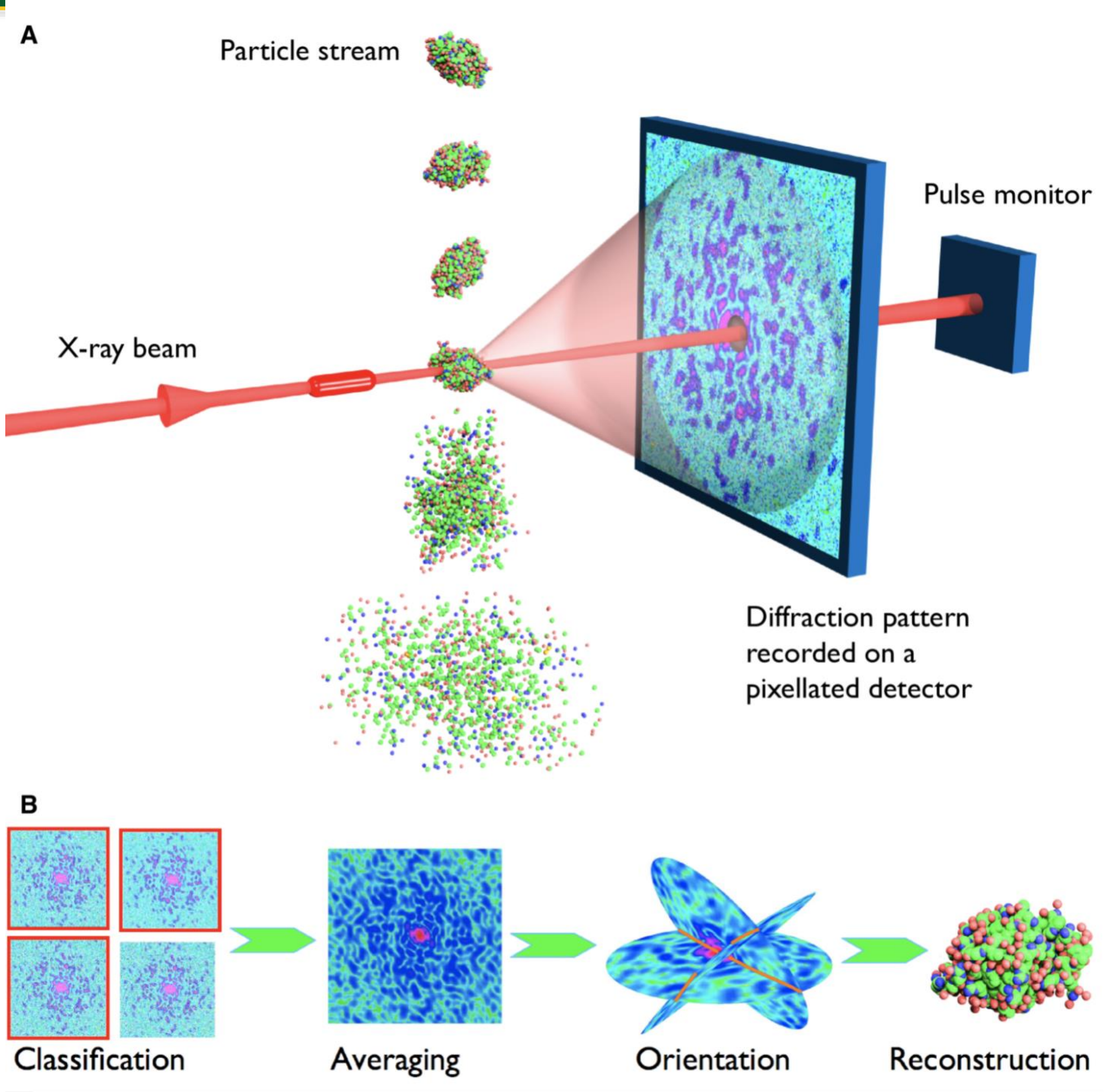
Also includes other products, e.g.,
AI: PyTorch, TensorFlow, Horovod
Co-Design: AMReX, Cabana, MFEM

ExaFEL: An ECP Case Study

Goal: LCLS **free electron laser** will increase its data throughput by three orders of magnitude by 2025. Near real-time analysis (~10 min) of data bursts, requiring burst computational intensities exceeding an exaflop

Computational challenges

- Complex multi-component workflow, integration of DOE HPC and experimental facilities
- Non-uniform FFTs on GPUs
- Maximum likelihood estimation non-linear, sparse optimization loop

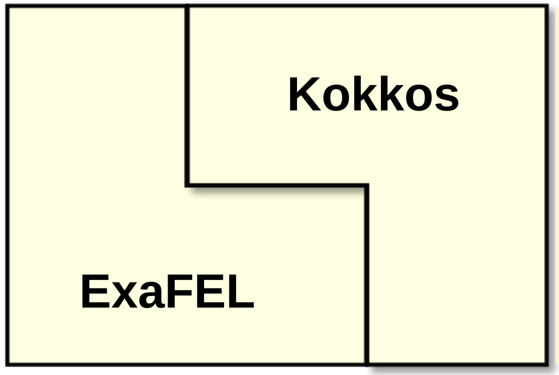


PI: Amedeo Perazzo, SLAC

Single Particle Imagine (SPI) Acceleration on GPUs

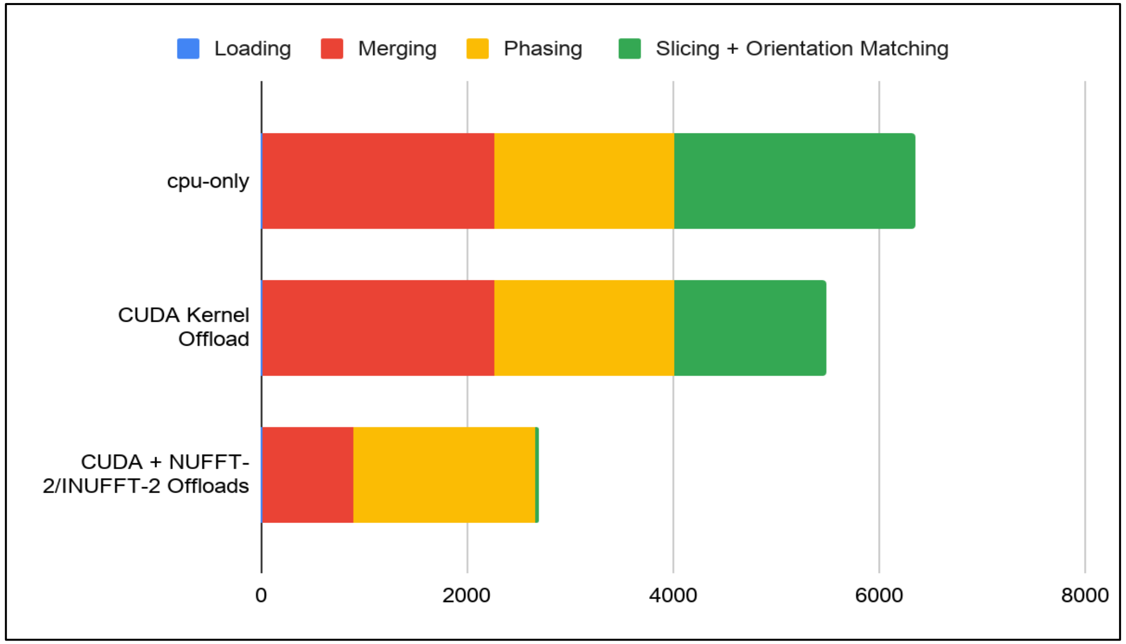
Single-node analysis: 1,500 images

- 1 CPU vs 1 GPU
- GPU-optimized slicing
- Uses ***spinifel*** proxy app



nanoBragg code ported from Nvidia to AMD GPUs with minimal effort

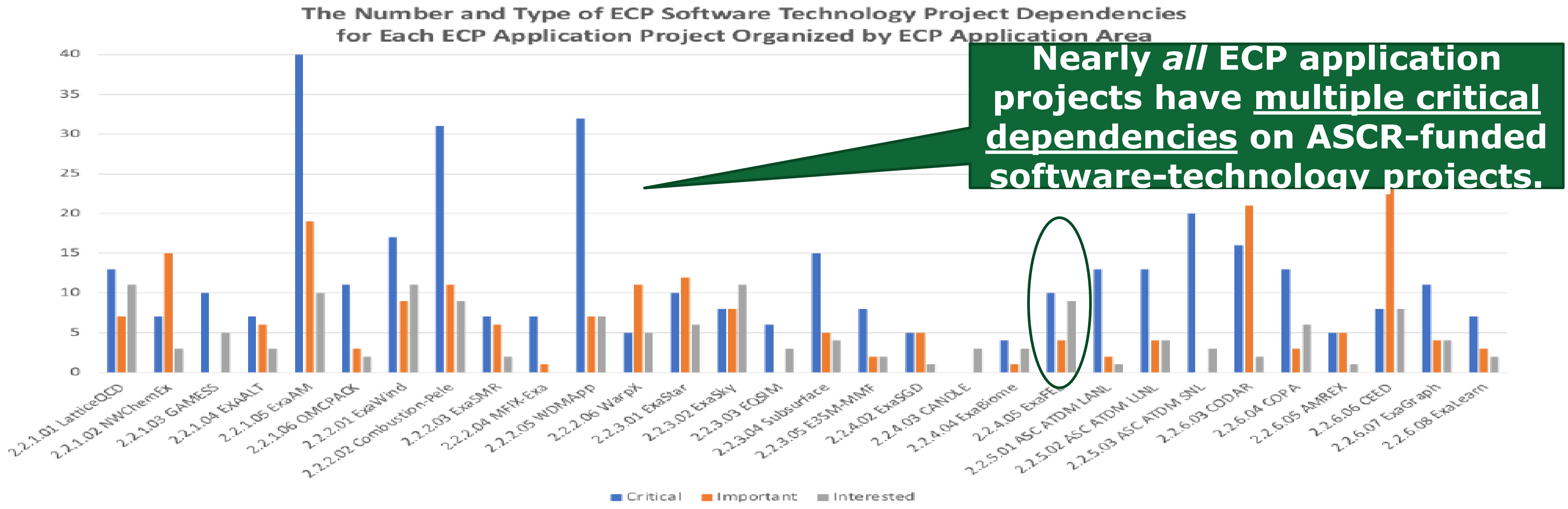
Time (s) spent in different modules



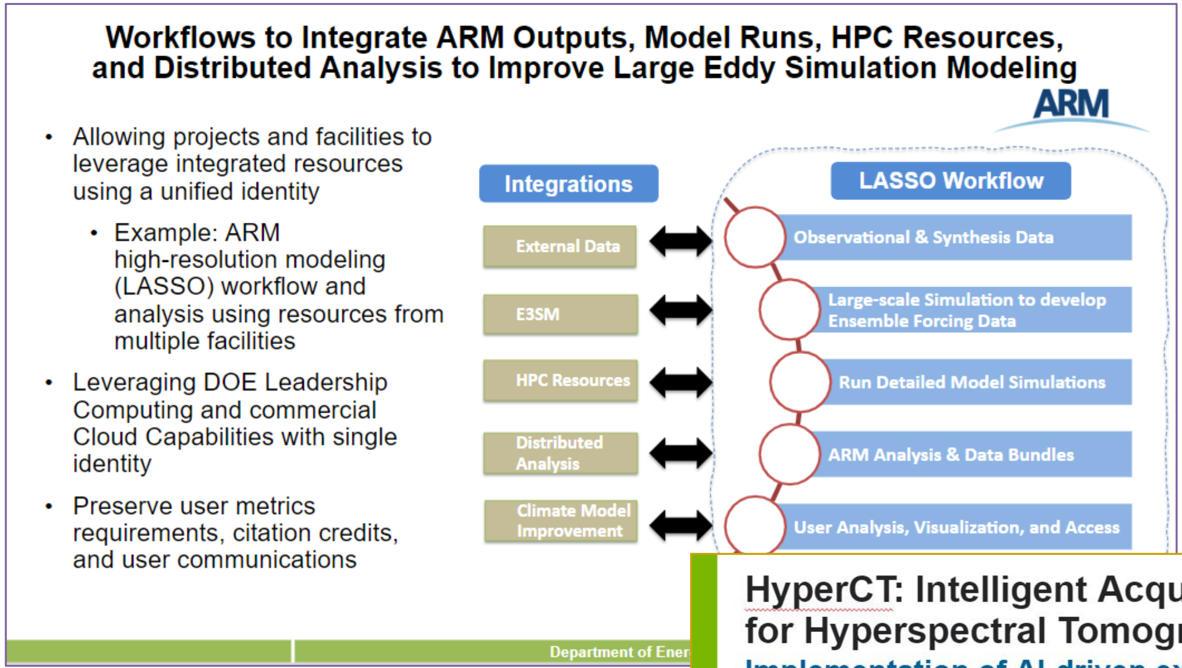
Optimization level	Wall Time (s)	Speed Up
CPU only	6345	-
CUDA kernels offload	5495	13%
CUDA kernel + NUFFT-2/INUFFT-2 offloads	2697	57%

ASCR Software Sustainability

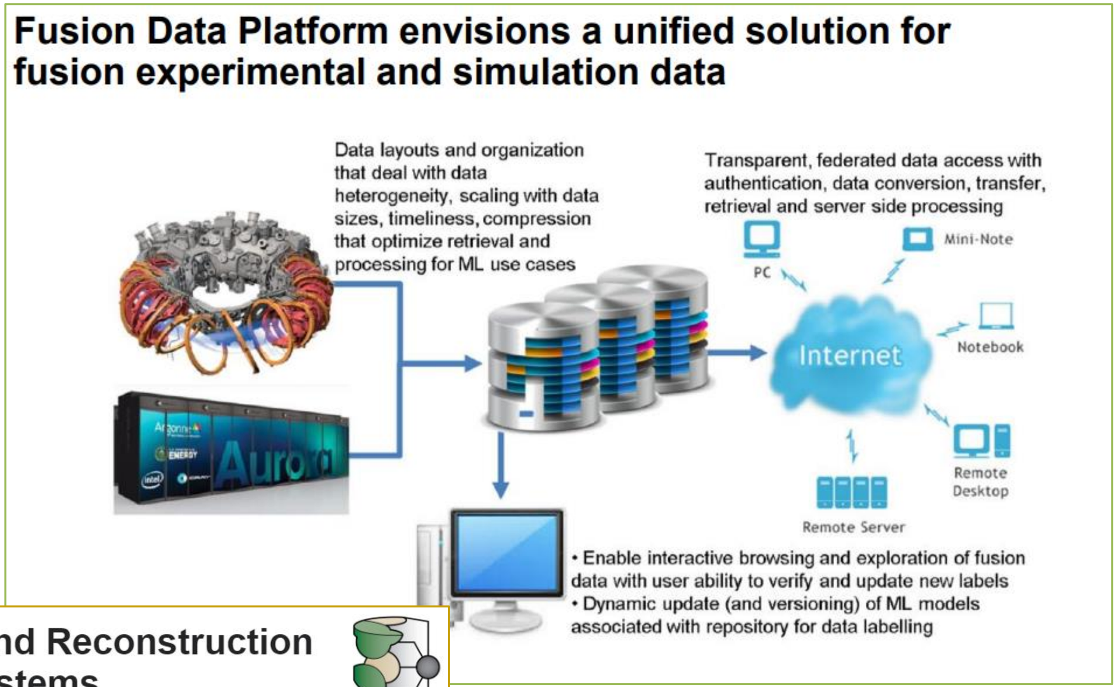
- ASCR-supported software technologies are critical to DOE scientific software on *all* platforms.
- In the CHIPS and Science Act, congress explicitly added *Exascale Ecosystem Sustainment* to ASCR's portfolio balance, "**<u>It is the sense of Congress that</u>** the Exascale Computing Project has successfully created a broad ecosystem that provides shared software packages, novel evaluation systems, and applications relevant to the science and engineering requirements of the Department, and that **<u>such products must be maintained and improved in order that the full potential of the deployed systems can be continuously realized.</u>**"



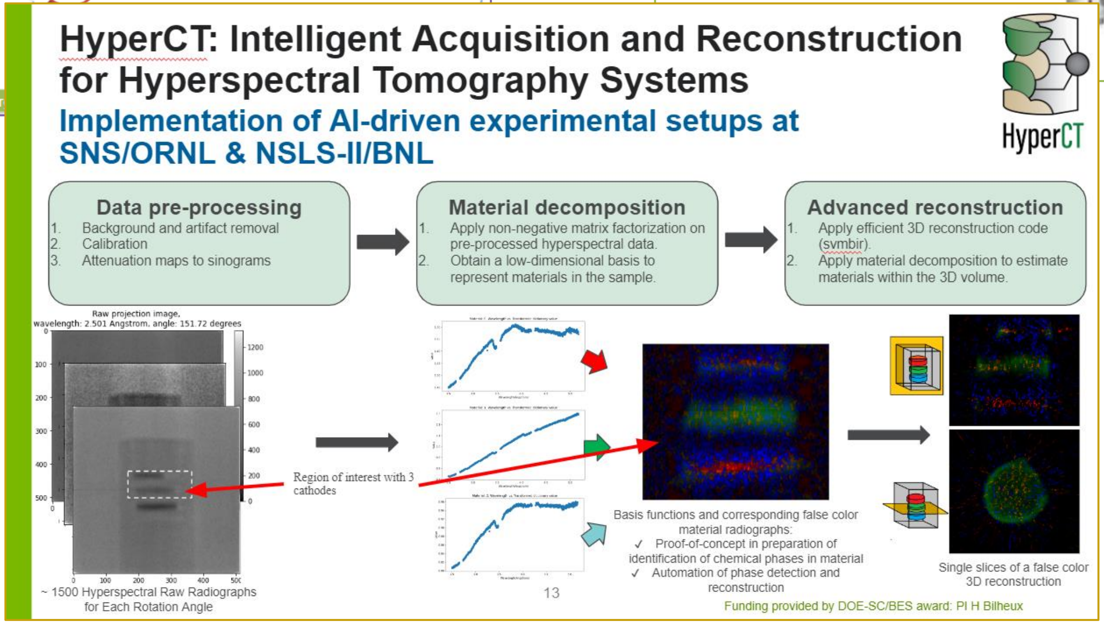
Data Explosion: SC Programs are grappling with integration across facilities and resources



Courtesy Paul Bayer, BER



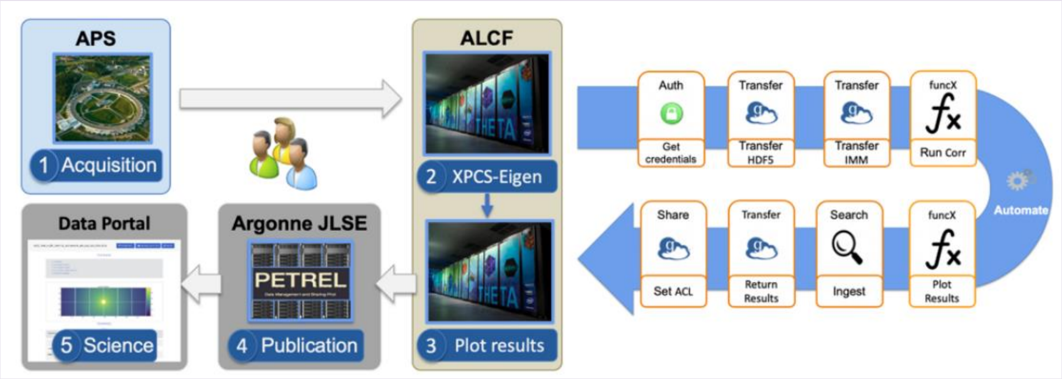
Courtesy Matt Lanctot, FES

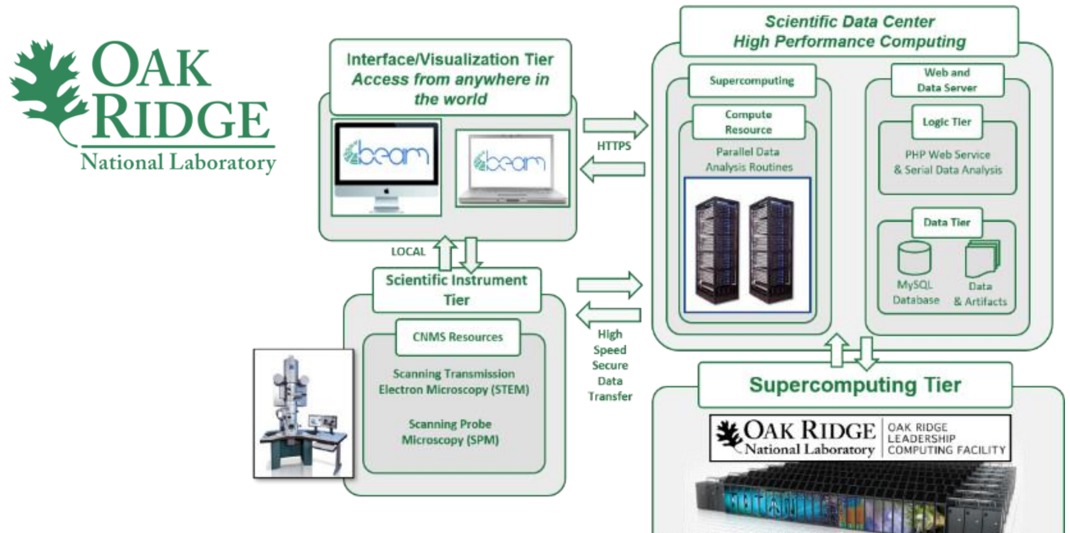


Courtesy Tom Russell, BES

Across DOE, innovators have been taking concerted steps towards integration through research, partnerships, and lab-level projects

LBNL's Superfacility project
ORNL's INTERSECT initiative
ANL's ALCF-APS Balsam software project
NERSC-LCLS LLANA software project
ECP ExaWorks & ExaFEL projects
BES DISCUS Light Source Data Working Group project
BES-ASCR CAMERA applied math center
BER joint EMSL-JGI FICUS joint-allocation program
… and more





Lingerfelt et al., *Procedia Computer Science* **80**, 2276 (2016).

➢ These are all separate initiatives with similar integration goals.

DOE is positioned to lead the new era of integrated science within the USG and the world.

Linking distributed resources is becoming paramount to modern collaborative science.

The next era of the DOE laboratories is as an open innovation ecosystem:

Accelerating discovery & innovation

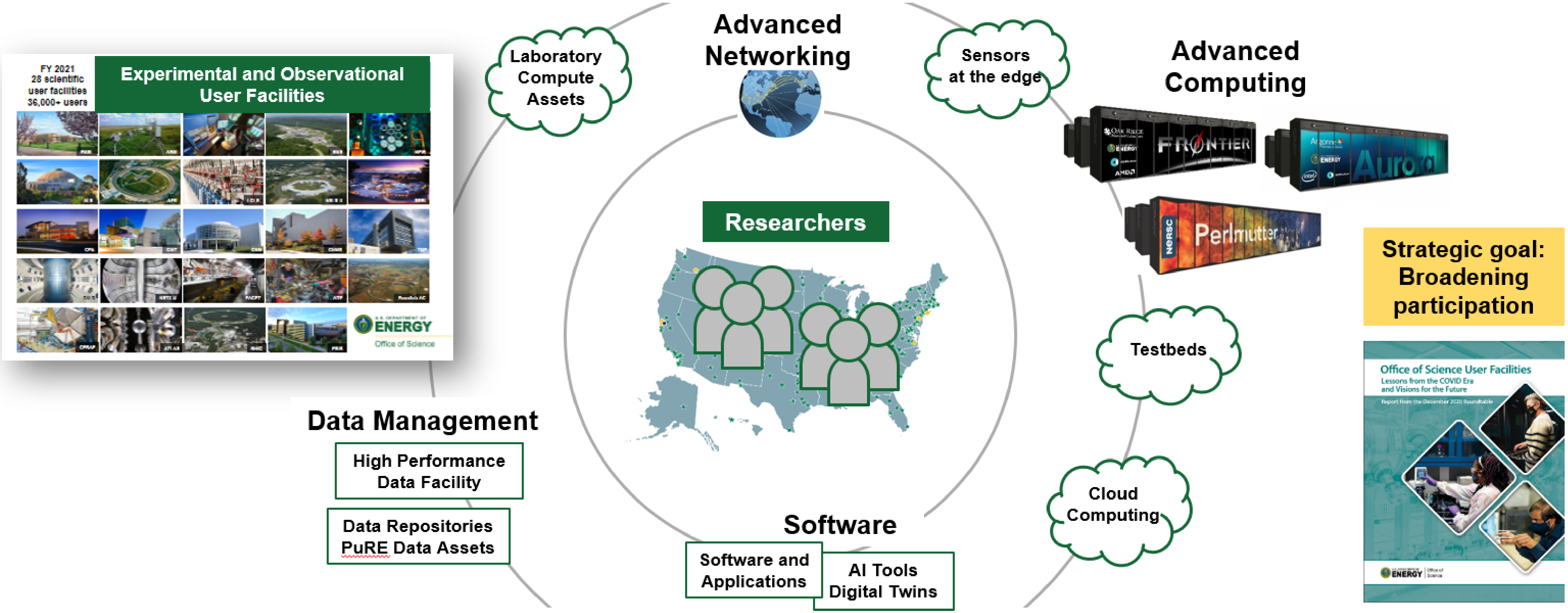
Democratizing access

Drawing new talent

Advancing open science

Integrated Research Infrastructure (IRI)

The IRI vision: A DOE/SC integrated research ecosystem that transforms science via seamless interoperability



New modes of integrated science

Rapid data analysis and steering of experiments	Novel workflows using multiple user facilities	AI-enabled insight from integrating vast data sources

FY 2022 Integrated Research Infrastructure Architecture Blueprint Activity

Aim: Produce the **reference conceptual foundations** to inform a coordinated "whole-of-SC" strategy for an integrative research ecosystem.

Gathering Insight phase

- **IRI Science Cases and IRI Modes/Patterns**
- Existing IRI projects' lessons learned
- Urgencies
- Common requirements
- Gaps and challenges

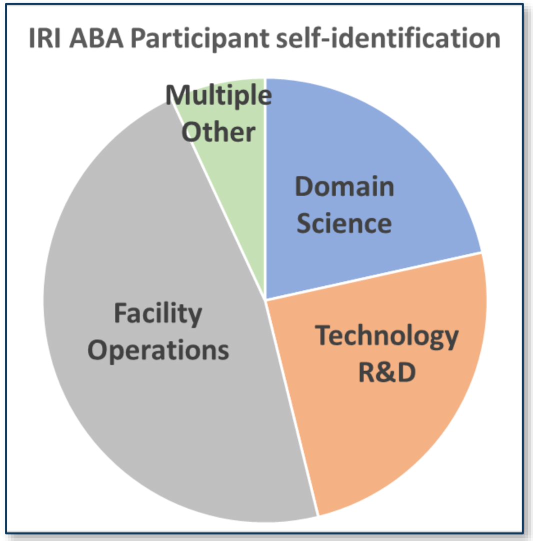
Synthesis & sharing

Design phase

- **Develop one or more IRI Architecture Blueprints** that address ensembles of design patterns.
- Develop operational principles for achieving an SC IRI ecosystem.

Final synthesis

Feb 2022 | Mar | Apr | May | Jun | Jul | Aug | Sep



Engaged 160+ DOE experts across the labs, all SC User Facilities, research programs, and DOE HQ.

Key Preliminary Conclusions:

- IRI requires a **distributed and interoperable approach** to computational and data infrastructure.
- IRI high performance data infrastructure will be an **orchestrated system of systems**.

IRI Blueprint Activity Governance

SC HQ Executive Leadership

Ben Brown
Director, ASCR Facilities Division

SC HQ Coordination Group

BER	Paul Bayer, Jay Hnilo, Resham Kulkarni
BES	Tom Russell
FES	Josh King, Matt Lanctot
HEP	Jeremy Love, Eric Church
IP	Kristian Myhre
NP	Xiaofeng Guo, Jim Sowinski

Field Leadership Group

Debbie Bard, NERSC, LBNL	Eric Lancon, SCC, BNL
Amber Boehnlein, CIO, JLab	Jini Ramprakash, ALCF, ANL
Kjiersten Fagnan, JGI, LBNL	Arjun Shankar, OLCF, ORNL
Chin Guok, ESnet, LBNL	Nicholas Schwarz, APS, ANL