SAND2021-5881PE

Sandia National Laboratories

Exceptional service in the national interest

Advanced Tri-lab Software Environment (ATSE)

MAY 19, 2021

PRESENTED BY

Kevin Pedretti
ktpedre@sandia.gov

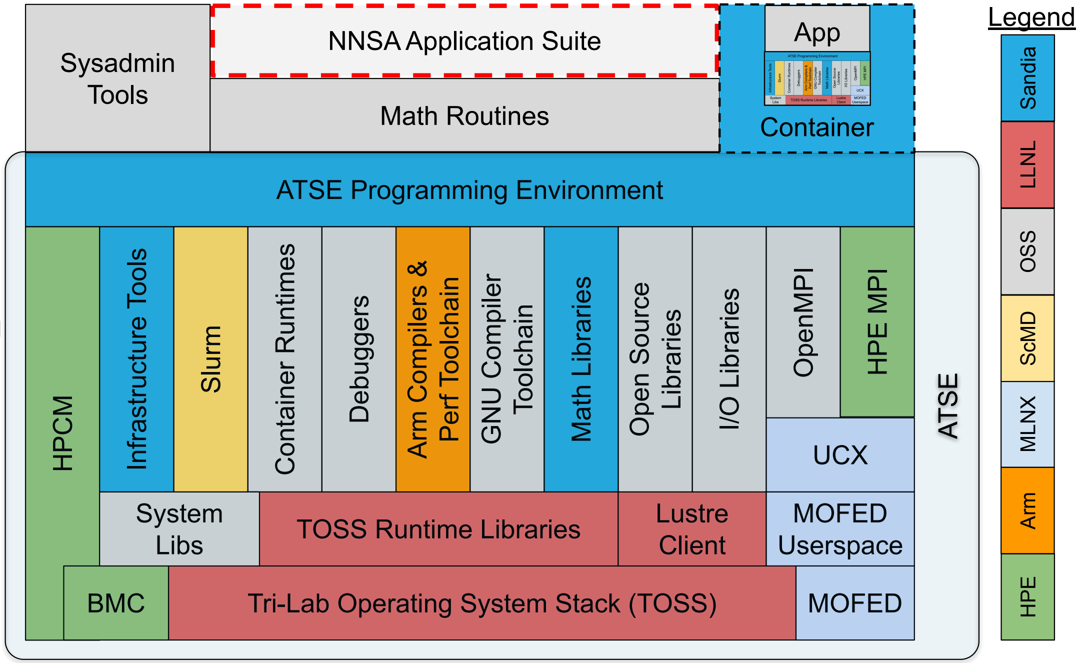
2021 ASC PI Meeting – Tri-lab Computing Session

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2021-5040 PE

What is ATSE?

- Modular, extensible, and open HPC software stack
  - Provide operational independence from any single vendor, encourage vendors to add value
  - Focal point for collaboration activities to mature new technologies (HW + SW)
- Prototype software stack for prototype systems: *Adv Arch Prototype Systems (AAPS), Vanguard2, Testbeds, Arm+GPU, A64FX, etc.*

ATSE:
Advanced Tri-lab Software Environment



Sysadmin Tools

NNSA Application Suite

Math Routines

App

Container

ATSE Programming Environment

HPCM

Infrastructure Tools

Slurm

Container Runtimes

Debuggers

Arm Compilers & Perf Toolchain

GNU Compiler Toolchain

Math Libraries

Open Source Libraries

I/O Libraries

OpenMPI

HPE MPI

UCX

System Libs

TOSS Runtime Libraries

Lustre Client

MOFED Userspace

BMC

Tri-Lab Operating System Stack (TOSS)

MOFED

ATSE

Legend

Sandia

LLNL

OSS

ScMD

MLNX

Arm

HPE

NNSA/ASC Astra Petascale Arm Supercomputer

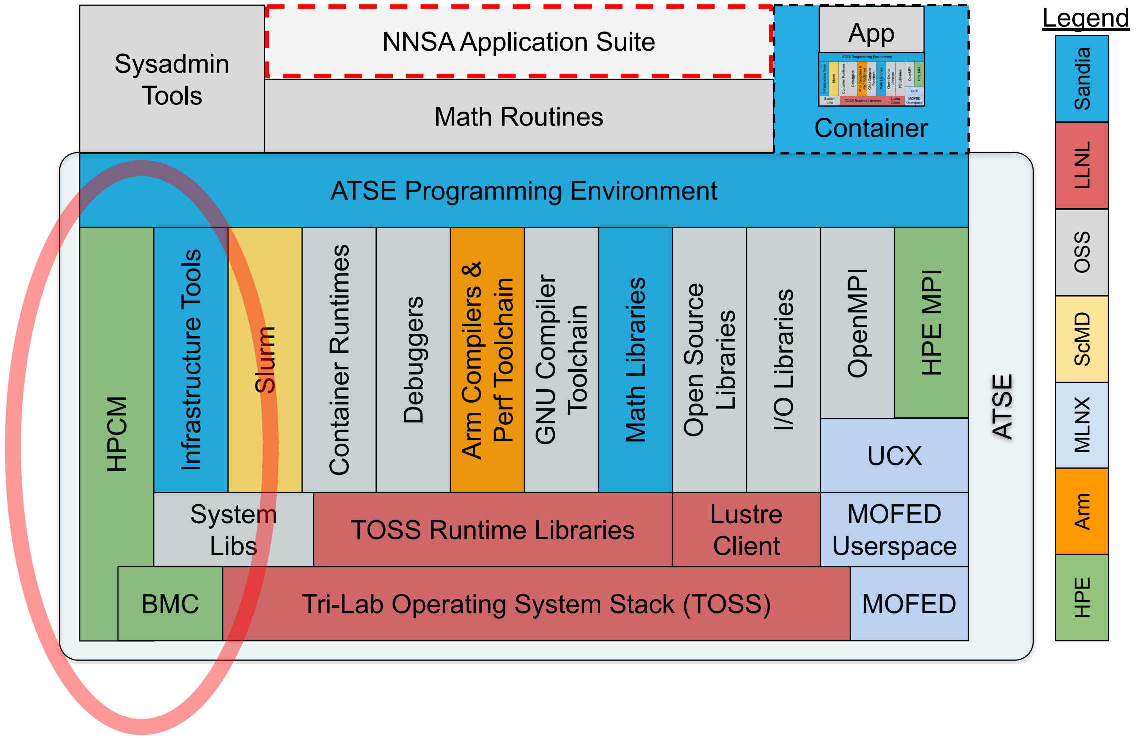
ASTRA
"Per aspera ad astra"
Sandia National Laboratories
NNSA
Advanced Simulation & Computing
Hewlett Packard Enterprise
ENERGY
ATSE
stack
VANGUARD

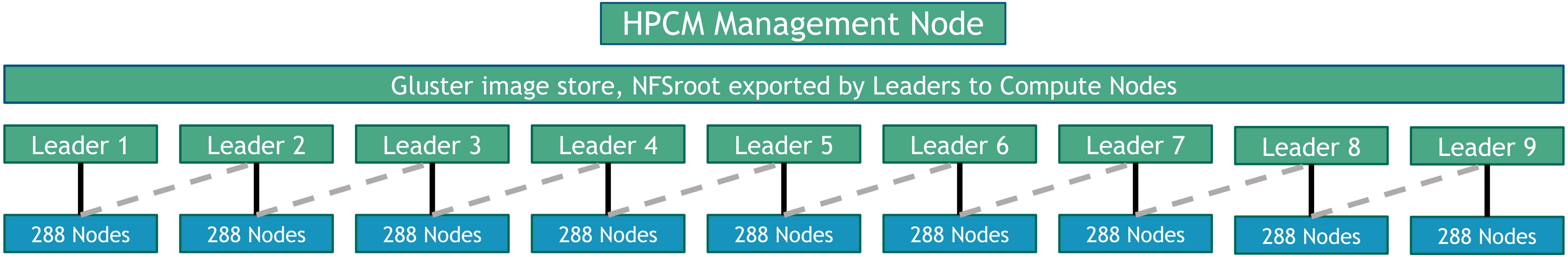
More details in SC20 paper – Chronicles of Astra: Challenges and Lessons

Scalable System Management

- HPCM: HPE Performance Cluster Manager
  - New and unproven at time of Astra deployment
  - Formed tight collaboration with HPE to mature
  - Now a technology option for large HPE/Cray systems
- Collaboration resulted in new capabilities:
  - Support for hierarchical leader nodes, Demonstrated boot of 2592 nodes in < 10 min
  - Scalable BIOS upgrades and configuration
  - Ability to build and deploy TOSS images (Tri-lab Operating System Stack)

ATSE:
Advanced Tri-lab Software Environment

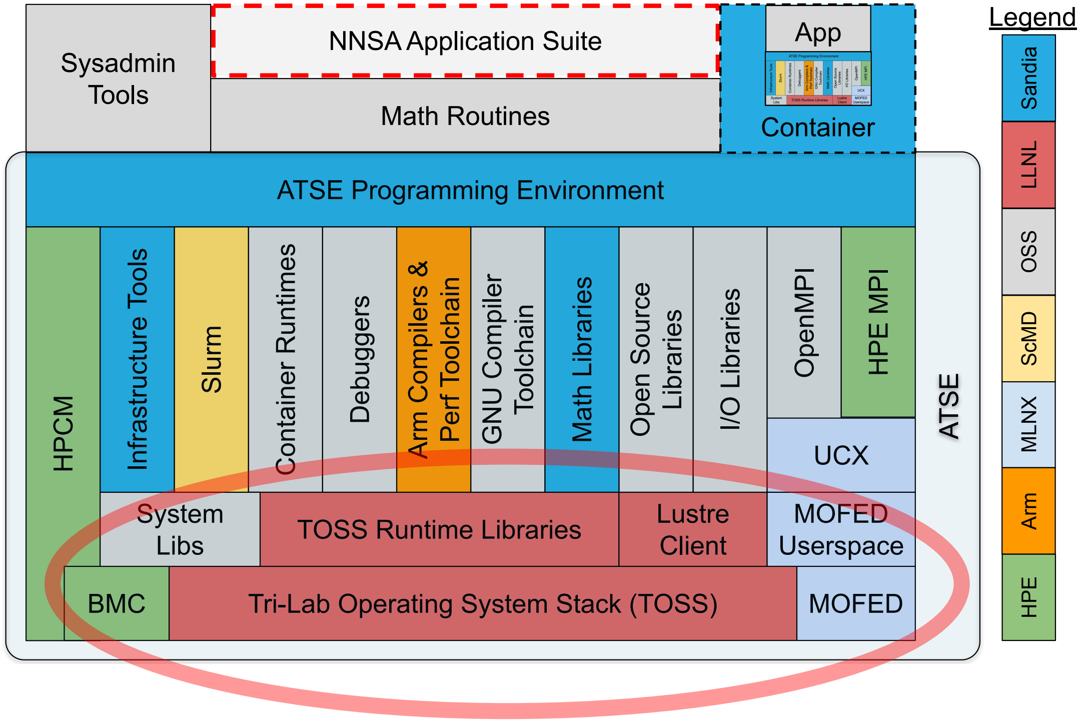




Scalable Base Operating System

- TOSS: Tri-lab Operating System Stack (Lead: LLNL, LANL, SNL)
  - Targets commodity technology systems (model: vendors provide HW, labs provide SW)
  - Red Hat based; x86_64, ppc64le, and aarch64
  - ~4K packages on all archs, 200+ specific to TOSS
  - Partnership with RedHat; direct engineering support
- Astra-related activities
  - Added packages needed for integration with HPCM
  - Added support for Mellanox OFED InfiniBand stack
  - Resolved Linux kernel bug(s) on Arm that were preventing large scale runs

ATSE:
Advanced Tri-lab Software Environment



Sysadmin Tools

NNSA Application Suite

Math Routines

App

Container

ATSE Programming Environment

HPCM

Infrastructure Tools

Slurm

Container Runtimes

Debuggers

Arm Compilers & Perf Toolchain

GNU Compiler Toolchain

Math Libraries

Open Source Libraries

I/O Libraries

OpenMPI

HPE MPI

UCX

System Libs

TOSS Runtime Libraries

Lustre Client

MOFED Userspace

BMC

Tri-Lab Operating System Stack (TOSS)

MOFED

ATSE

Legend

Sandia

LLNL

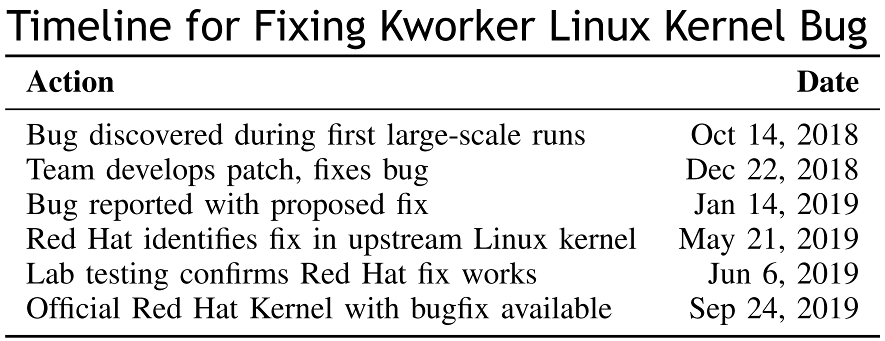
OSS

ScMD

MLNX

Arm

HPE



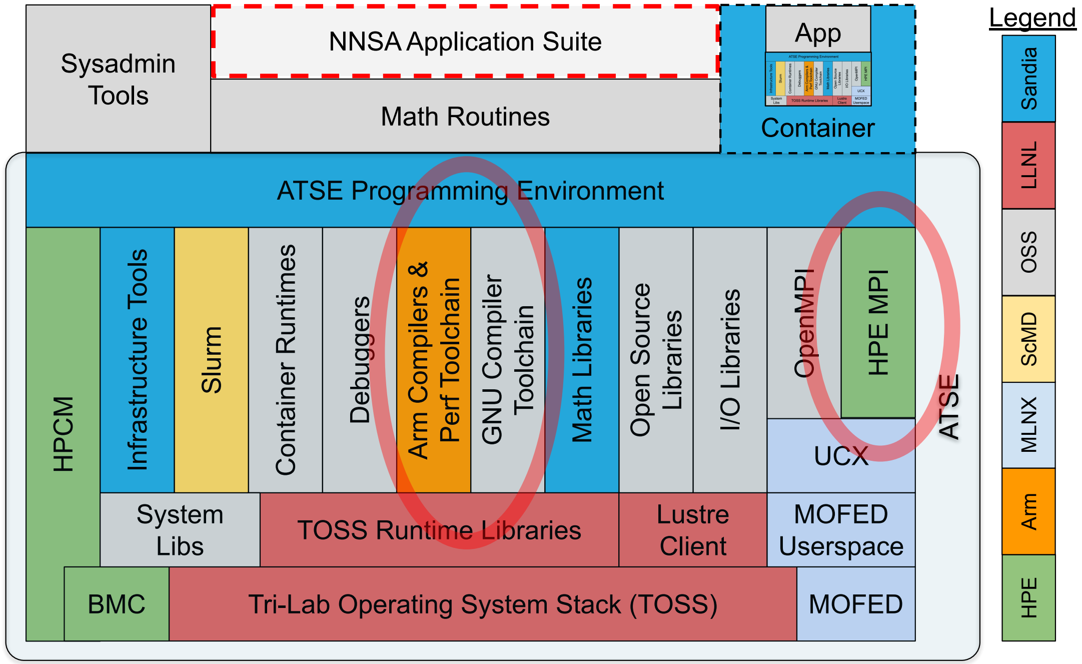
Timeline for Fixing Kworker Linux Kernel Bug

Action	Date
Bug discovered during first large-scale runs	Oct 14, 2018
Team develops patch, fixes bug	Dec 22, 2018
Bug reported with proposed fix	Jan 14, 2019
Red Hat identifies fix in upstream Linux kernel	May 21, 2019
Lab testing confirms Red Hat fix works	Jun 6, 2019
Official Red Hat Kernel with bugfix available	Sep 24, 2019

TOSS Provides a Scalable and Stable Base Operating System for ATSE / Astra

Vendor Components Add Value

ATSE:
Advanced Tri-lab Software Environment



- Encourage vendors to provide easily integrated components rather than big monolithic stacks
- ARM Allinea Studio
  - ARM Compiler for HPC (armclang, armflang, …)
  - ARM Performance Libraries (BLAS, LAPACK, FFT)
  - ARM Forge (Map and DDT) + Performance Reports
- HPE MPI
  - Proprietary MPI dates back to mid 90's, SGI MPI
  - Provides alternative MPI option for comparison / debug
- Placed contracts with Arm & Marvell to work on compiler and math lib optimizations
  - Improve threading for various matrix sizes (typically smaller sizes needed more optimization) on TX2/SVE
  - Vectorized (TX2/SVE) batched operations for multiple small operations – used in block-solver schemes
  - Performance results of 2.5X – 10X seen on multiple important BLAS and LAPACK routines

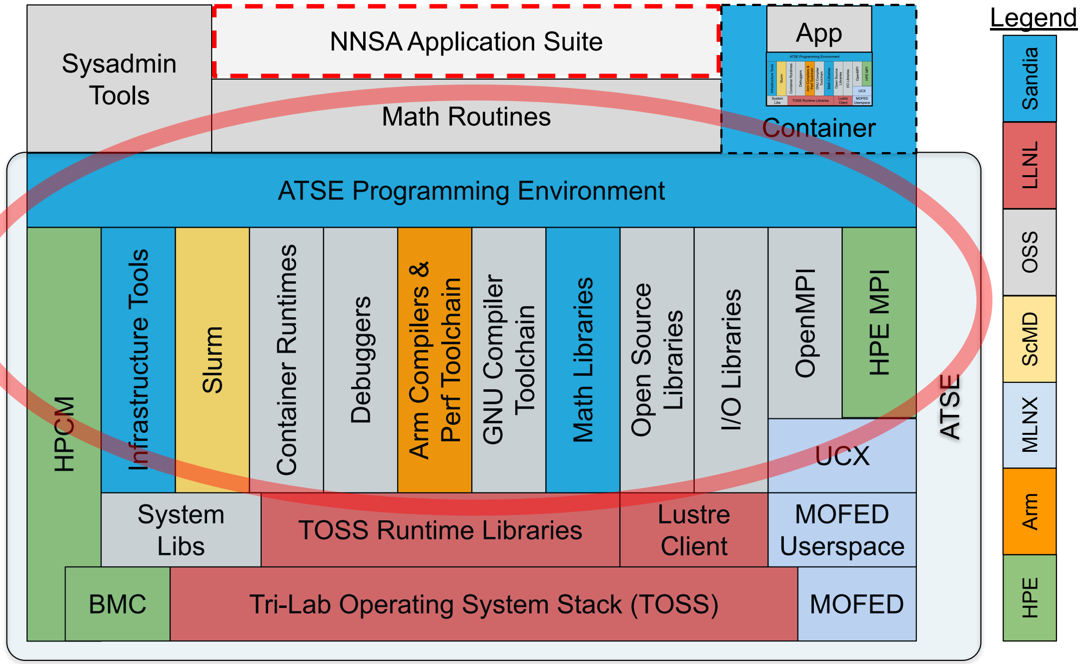
    Available in Arm 21.0 (latest release)

Vendor Components Integrate with ATSE to Provide Key Capabilities and Added Value

Scalable Programming Environment

- Curated HPC software stack
  - Provides base set of compilers, MPI implementations, third-party libraries, tools, and other components known to work well together
  - Focused on needs of ASC codes / ATDM L1 milestone
- Especially important for immature technologies
  - Many bugs, broken packages, and missing functionality
  - Need to do more to help users, avoid duplicated work
- Look and feel similar to OpenHPC, adapted for ASC:
  - Pin packages at specific versions, per code team requirements
  - Add missing packages (e.g., ParMETIS, CGNS)
  - Add microarchitecture and compiler optimizations
  - Add static library support, simplifies moving binaries between networks

ATSE:
Advanced Tri-lab Software Environment



ATSE Recipes Available @
https://doi.org/10.5281/zenodo.4006668

ATSE Provides "Ready to Go" Programming Environment for ASC Codes

Building ATSE with Spack

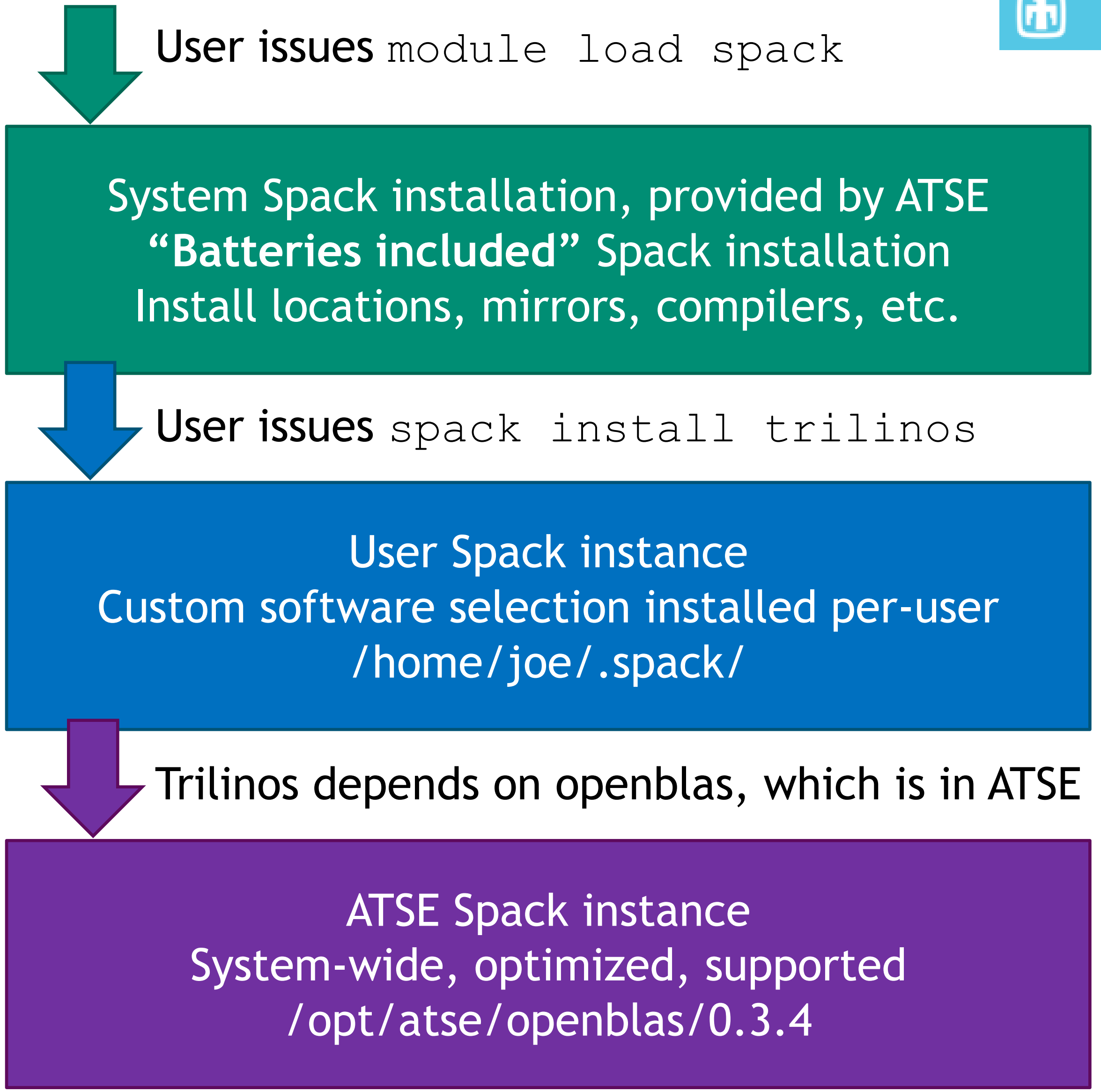
- Now using Spack to build ATSE Prg. Env.
  - Developed automated workflow for generating reproducible builds with same look and feel
  - Combines curation of ATSE with power of Spack
  - Build time reduced to 3 hours (was > 24 hours)
  - Used successfully on Arm+GPU and A64FX

- ATSE contributions to Spack

Package version bumps	12
Variant additions	17
Package additions	2
Core bug fixes	1
Major feature additions	2 (pending)

```
Package install metrics (#14705)
Shared spack instances  (#11871)
```

ATSE Shared Spack Instance Workflow



User issues `module load spack`

System Spack installation, provided by ATSE
“Batteries included” Spack installation
Install locations, mirrors, compilers, etc.

User issues `spack install trilinos`

User Spack instance
Custom software selection installed per-user
/home/joe/.spack/

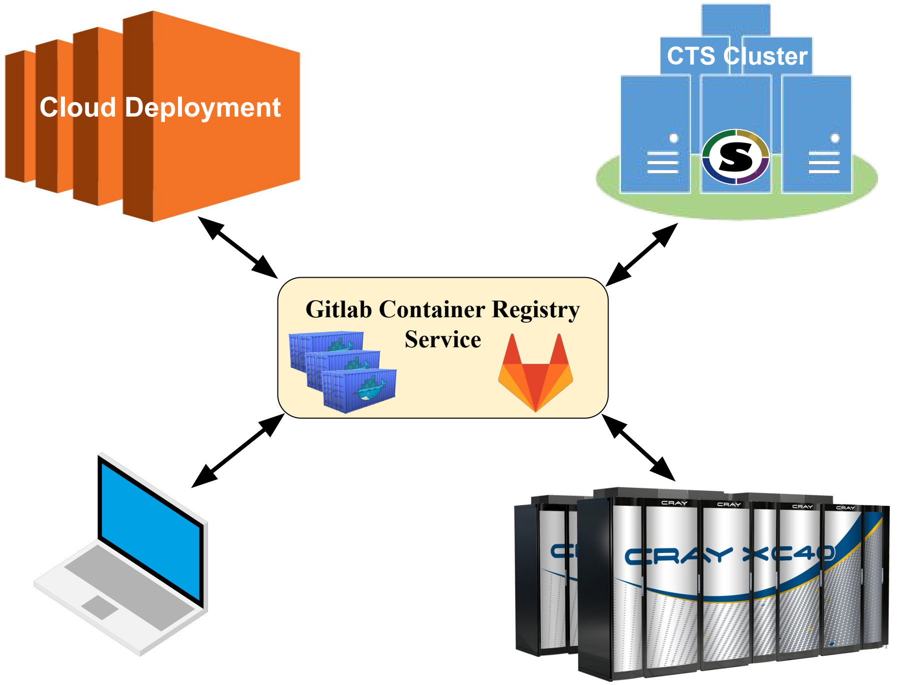
Trilinos depends on openblas, which is in ATSE

ATSE Spack instance
System-wide, optimized, supported
/opt/atse/openblas/0.3.4

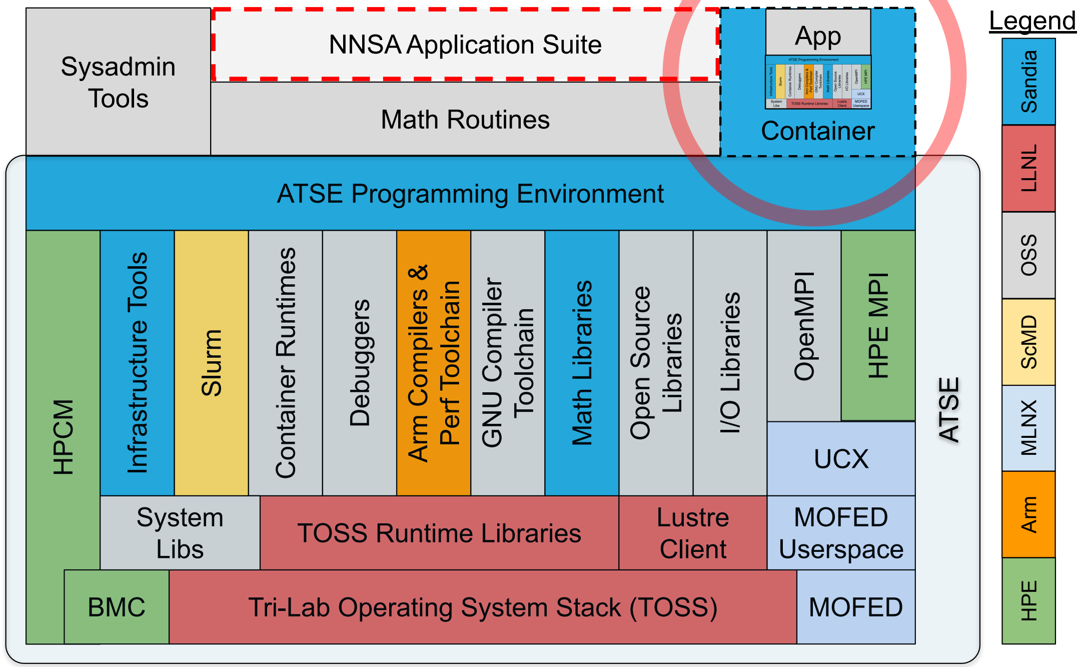
ATSE is Leveraging and Contributing New Capabilities to Spack

Advanced Container Workflows

- Release testing, Rollback, and Off-platform Test
  - Enabling Sierra / IC teams to do off-platform continuous integration build and test, freeing up Astra resources
- Pioneering in-platform unprivileged container builds
  - Podman, Charliecloud, Singularity, Docker, …



ATSE:
Advanced Tri-lab Software Environment



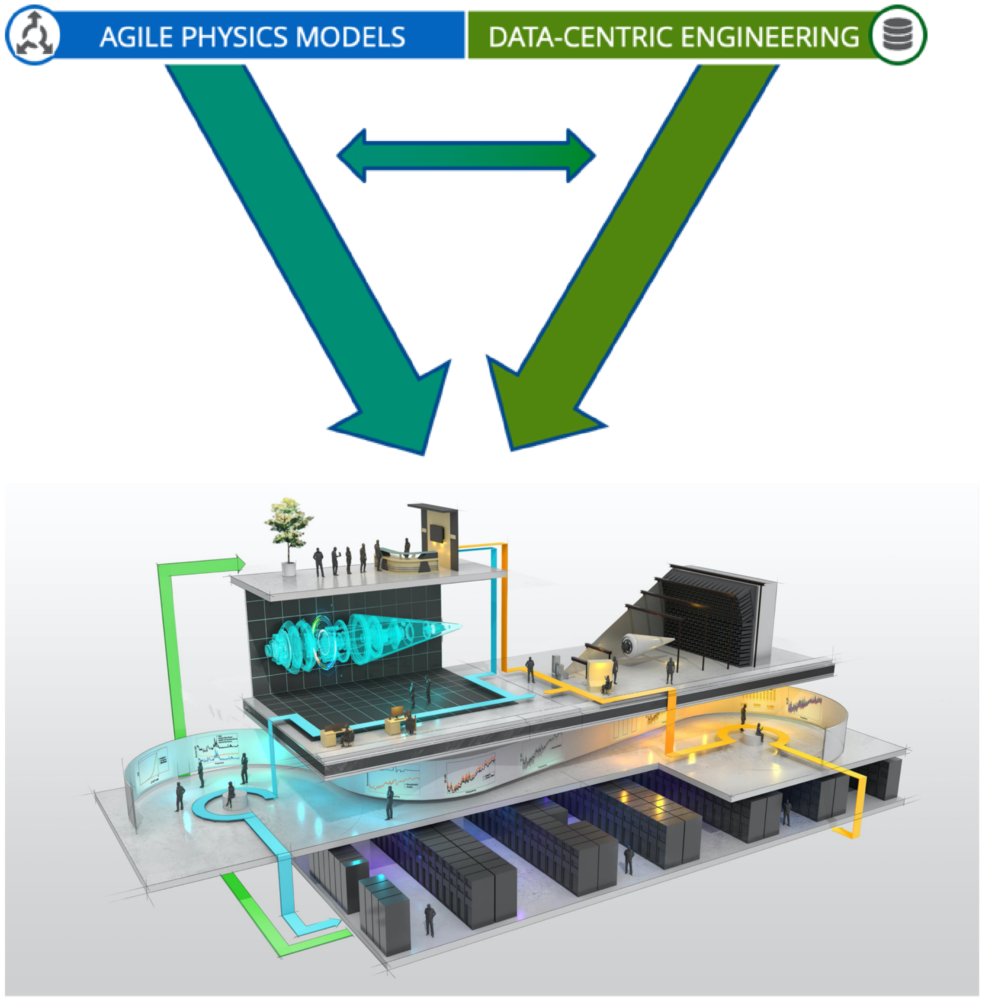
Collaborating with ECP Supercontainers Project
(SNL lead, LLNL, LANL, NERSC, U. Oregon)

Simplify Deployment of ASC Codes, Seamlessly Move Between Laptops / HPC / Cloud

Path Forward

- Develop ATSE for VANGUARD2
  - Build on foundational R&D from Astra / Vanguard-1
  - Close vendor collaboration and expect lots of iteration
- Extreme heterogeneity at node and system levels
  - Tightly-coupled CPU+X accelerators
  - Specialized node types for CPU, GPU, ML, Dataflow, …
- Dynamic resource management
  - Container orchestration for long-lived services + HPC apps
  - "Cloud-like" environments and usage models
- Enable cross-lab containerized workflows
  - Interface with TCE2, Charliecloud, and RCE efforts

Enable Converged HPC/HPDA
<u>Advanced Digital Engineering Workflows</u>



Conclusion

- ATSE is a modular, extensible, and open HPC software stack for Advanced Architecture Prototype Systems (AAPS) and Advanced Arch Testbeds
- Focal point for collaboration activities to mature new technologies (Hardware + Software) with potential to improve the ASC computing environment
- Approach has been successful on Astra
  - Facilitated collaboration activities with code teams and external system vendors
  - Supported FY21 ATDM L1 milestone successful completion
  - Deployed as a production system to support current LEP work
- Path forward enabling advanced containerized workflows on extremely heterogeneous systems, combining HPC and Cloud-like usage modules

Thanks to Matthew Curry, Andrew Younge, Si Hammond, Rob Hoekstra, Jim Laros, Ron Brightwell, Carson Woods, and the entire ATSE team for slide content