LSSw Meeting 4

January 20, 2022

Announcements

Preview for LSSw Meeting 5: Feb 17, 2022

- Topic: Other Technical Software Ecosystems: A panel discussion
- Description: This month we have panelists representing other technical software ecosystems:
  - Panelists, TBD
- Prompts:
  - What is the value proposition of your ecosystem to it developer and user community?
  - What is the business model of your ecosystem (how do people fund their efforts)?
  - What are some of the challenges you face in providing value?
  - What are your sustainability challenges?

LSSw Meeting 4

- Topic: Expanding the Leadership Scientific Software Developer and User Communities: A panel discussion
- Description: This month we have panelists representing the broader scientific software developer communities:

  - Deb Agarwal, Berkely Lab
  - Anshu Dubey, Argonne National Laboratory
  - Bill Hart, Sandia National Labs
  - Addi Malviya-Thakur, Oak Ridge National Laboratory
  - Katherine Riley, Argonne National Laboratory

  Prompts:

  - How has the traditional definition of leadership (HPC) scientific software developers limited who can be involved?
  - What is required to make the traditional definition more inclusive?
  - What do you see as the most urgent priority activities in planning for a holistic leadership software ecosystem over the next few years?
  - What is missing from the conversation about sustainable leadership scientific software?

Deb Agarwal

Background:

Berkeley Lab - 27 years

- Data Science and Distributed Systems Researcher
- Computational Science
- Manager (currently Division Head)

Types of Software Projects:

- Research prototypes
- Research output that results in releasable library
- Scientific data processing/workflows/pipelines
- HPC performance software
- Scientific user interfaces

What are some important similarities and differences between the software development and use in your community relative to large-scale HPC environments such as the DOE Leadership Computing Facilities (LLNL, OLCF, ALCF, NERSC)?

- Similarities
    - Software needs to be reliable, maintained, and dependable
    - Not all software is heavily used
    - Science depends on the software
- Differences
    - Scalability and robustness to availability/scheduling required
    - Frequency of significant changes in computing environment

Do you think it makes sense to include your scientific software community as part of what we mean by Leadership Scientific Software?

- Our mission is supporting DOE science
- Both HPC and non-HPC software is needed to accomplish science
- Where is the boundary defined? Many pipelines/workflows involve HPC and non-HPC elements

If so, what is most important to consider if we attempt to expand the definition of leadership computing to include the community you represent? If not, why?

- It would be valuable to be addressing sustainability of DOE Science software
- HPC, large-scale experiments, data repositories, processing pipelines, etc

What are some issues in the HPC software community that are not being sufficiently addressed right now that need to be considered to better address your important requirements?

- Lack of knowledge of good software engineering practices
- Complex aggregations of software combined into a workflow
- Best practices, consultants, training, site licenses for software engineering tools

Anshu Dubey

- Computational Scientist, Mathematics and Computer Science Division, Argonne National Laboratory
- Chief software architect for FLASH, a multiphysics multidomain application code from version 3 on.
  - FLASH was first released in 2000 for astrophysical thermonuclear flashes simulation
  - It went on to become used by several other science communities
  - Papers have been published about its architecture, software process and evolution, and impact on scientific communities
- Leading development of a new multuphysics code derived from FLASH, designed for heterogeneous platforms
  - Fundamentally rearchitected with several solver upgrades
  - Shifting over to community development and governance model
  - Watch for announcement in the next 2-3 weeks
- Deeply engaged with the IDEAS-ECP project
  - Leading the development and documentation of methodologies

Prompts

- What are some important similarities and differences between the software development and use in your community relative to large-scale HPC environments such as the DOE Leadership Computing Facilities (LLNL, OLCF, ALCF, NERSC)?
  - I work with software that is used on all types of platforms ranging from workstations to LCFs because it is an application software. Therefore, it has the constraint that it needs to be robust but also performance portable across a wide variety of platforms. And then there is SSWG that is looking at even greater diversity in software.
- Do you think it makes sense to include your scientific software community as part of what we mean by Leadership Scientific Software?
  - Absolutely. You could argue that LSS is meant to help communities like ours
- If so, what is most important to consider if we attempt to expand the definition of leadership computing to include the community you represent? If not, why?
  - Leadership computing often takes the perspective of libraries and system software, and their concerns. It is becoming apparent that more varied software will be needing the use of HPC platforms, and new science communities will be moving their analysis and operations to HPC. They need to have a place at the table.
- What are some issues in the HPC software community that are not being sufficiently addressed right now that need to be considered to better address your important requirements?
  - Now I will put on the hat of SSWG which is also championing the cause of experimental or observational facilities that are fast moving towards needing the use of LCFs in non-traditional ways. Their biggest challenge is that their requirements are diverse and may not neatly fit into the model that has worked well for the more traditional simulation and analysis use of LCFs

William Hart

Background

- Ph.D. Computer Science, B.A. Mathematics
- Sandia Labs
  - Research Staff (19 years)
  - Manager (8 years)
- Research focus: optimization methods and applications

ECP Lead – Data Analytics and Optimization Applications

- CANDLE (ML)
- ExaBiome (Graph Algs)
- ExaFEL (ML/Opt)
- ExaSGD (Opt)

Major Scientific Software Activities

- Autodock – Flexible drug docking
- DAKOTA – Black/Grey-box optimization
- Acro/PICO/PEBBL – Parallel branch-and-bound
- Pyomo – Optimization modeling
- Gcovr – Code coverage analysis

Prompts

- What are some important similarities and differences between the software development and use in your community relative to large-scale HPC environments such as the DOE Leadership Computing Facilities (LLNL, OLCF, ALCF, NERSC)?
    - Key software components might be commercial. How can we manage licenses?
    - End-users more focused on application. Need to prioritize ease of use. Cannot assume deep SW skills.
    - Different funding models for SW development. Big govt investments vs commercial open source vs academic vs
- Do you think it makes sense to include your scientific software community as part of what we mean by Leadership Scientific Software?
    - Yes. ECP has demonstrated synergies that can be exploited.
- If so, what is most important to consider if we attempt to expand the definition of leadership computing to include the community you represent? If not, why?
    - Consider how HPC trends create an implicit bias in our scope of scientific computing. What does that mean for LC?
- What are some issues in the HPC software community that are not being sufficiently addressed right now that need to be considered to better address your important requirements?
    - Data streaming from experimental facilities to HPC facilities requires further maturation and investment
    - Cross-domain tool integration remains a challenge (e.g. modeling with machine learning & operations research)

Addi Malviya Thakur

- Group Leader, Software Engineering Group, ORNL
- Experience related to scientific software development and use
  - *PI for the Software Framework for ORNL's Initiative on Self-Driven Experiments for Science/Interconnected Science Ecosystem (INTERSECT)*
  - *Co-PI for the Data Interpretation Platform for ORNL Initiative on Neutrons Data Interpretation Ecosystem*
  - *Workflow and Integration Lead: ExaAM-ECP (Transforming Additive Manufacturing through Exascale Simulation)*
  - *Collaborator/Member of the IDEAS-ECP Project*
  - *Collaborator/Member for the Data Reduction Software Project at the ORNL SNS/HFIR facilities*
  - *Collaborator/Member for several other scientific software projects*
  - *Worked on Biomedical Imaging Research software before ORNL*
  - *Better Scientific Software (BSSw) Fellow - honorable mention*

Prompts

- What are some important similarities and differences between the software development and use in your community relative to large-scale HPC environments such as the DOE Leadership Computing Facilities (LLNL, OLCF, ALCF, NERSC)?
  - Similarities
    - Both benefit from Software Engineering best practices like testing, documentation, code versioning, continuous integration/continuous deployment (CI/CD) and from a reliable software ecosystem (e.g., e4s, spack)
    - All Research software may not start with a set of defined requirements, unlike other non-research software
    - Both depend on vendor stacks and third-party libraries
  - Differences
    - HPC software can be highly specialized with a high initial investment in expertise
    - The scales and heterogeneity are unique
    - Performance and portability are crucial

With contributions from William Godoy and Jon Fortney, ORNL

Prompts

- Do you think it makes sense to include your scientific software community as part of what we mean by Leadership Scientific Software?

    - Yes!
    - Integrated Research workflows and interconnected facilities will bring more application areas closer to LCFs
    - Research software and application developers from the non-HPC community will work with LCFs to run a part or all of their science workflows in this interconnected ecosystem
    - Including this scientific software community within the Leadership Scientific Software would be very timely

Prompts

- If so, what is most important to consider if we attempt to expand the definition of leadership computing to include the community you represent? If not, why?
  - Consider other ways LCFs can help push science boundaries besides the traditional ways
  - Consider how to bring LCFs to the edge effectively
  - Determine steps to facilitate integration and interconnection with other scientific facilities across DOE

Prompts

- What are some issues in the HPC software community that are not being sufficiently addressed right now that need to be considered to better address your important requirements?

  - Interconnected research workflows; connecting scientific instruments/facilities to the LCF(s)
  - Workflow management
  - Reproducibility
  - A security model that makes the interaction with LCF(s) seamless and standardized.
  - Availability for some flavor of Continuous Integration and Continuous Deployment at the leadership facility that is easily accessible for projects

With contributions from William Godoy and Jon Fortney, ORNL

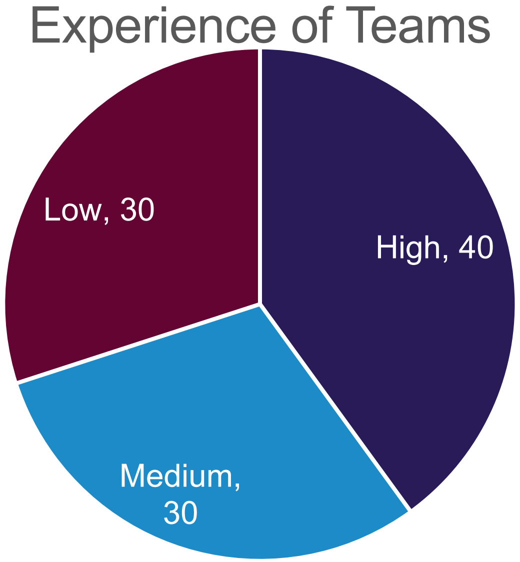
Katherine Riley

- Director of Science, Argonne Leadership Computing Facility
- Program Manager for DOE LCF’s INCITE Program
- At the ALCF since the start overseeing applications engagements, porting, performance, and science missions
- Code development and engagement in
  - —Multiphysics applications
  - —Scientific software engineering and practices
    - ▪ Transitioning applications into the HPC space

Similarities and Differences by Experience

What are some important similarities and differences between the software development and use in your community relative to large-scale HPC environments such as the DOE Leadership Computing Facilities (LLNL, OLCF, ALCF, NERSC)?

- Cost -
  - Under-discussed. Researchers might not fully understand the cost of a campaign or even a run. Systems are expensive, software investment should match this.
  - Raises the bar on development, validation, verification, uncertainty, reproducibility. Data.
- For the all levels of experience
  - Limited environment can slow them down
  - Other aspects of their scientific workflow are challenges – not just the software they have on LCFs. Attention is needed on the full pipeline to ensure good use of resources
  - Struggle with people/resources to tackle the scientific software challenges
- For the Medium and Low Experience teams
  - Software environment can stop them in their tracks – if they have a background fixing their environment, don't always have the ability
  - Software engineering focus is often different and, therefore, harder to align with LCFs
  - Security issues – at every level of workflows and environment
- For the Low
  - Unprepared for the complexity of the software challenges, debugging, performance
  - If there is a design to software, often missing needs of a supercomputer
    - (don't know limitations of stack, compiler differences, etc)



Prompts (2)

Do you think it makes sense to include your scientific software community as part of what we mean by Leadership Scientific Software?

- Secure the huge scientific potential of computing facilities by matching that level of investment in software
  - —Challenge is bigger than just LSS, but LSS type activities can play a vital role
- The science comes from the scientific software and not all teams will have expertise at hand
- Even as LCF software environment improves, they will always be cutting edge machines and be challenging
- Overall, HPC should be more accessible to science

If so, what is most important to consider if we attempt to expand the definition of leadership computing to include the community you represent? If not, why?

- Is this expanding the definition of leadership computing or expanding what constitutes important areas of investment?
- Reasonable software engineering and support that will maximize investments in hardware
- Reducing the barrier to effective use of HPC while *not compromising integrity of the science*
- Remember, only a small slice of those who might use LCFs have been/are in ECP

Prompts (3)

What are some issues in the HPC software community that are not being sufficiently addressed right now that need to be considered to better address your important requirements?

- Designing applications to deal with the challenging, very heterogeneous, changing environment without collapsing into single solutions or too many solutions
  - Ensuring that scientific software tools are used properly
- Adoption and integration of practices that will improve the science capabilities and flexibility of applications
- With rapidly increasing heterogeneity, help ensure good science is delivered by scientific software
  - Reproducibility, Uncertainty, Validation, etc