CUP-ECS Center Overview

PSAAP-III Annual Review

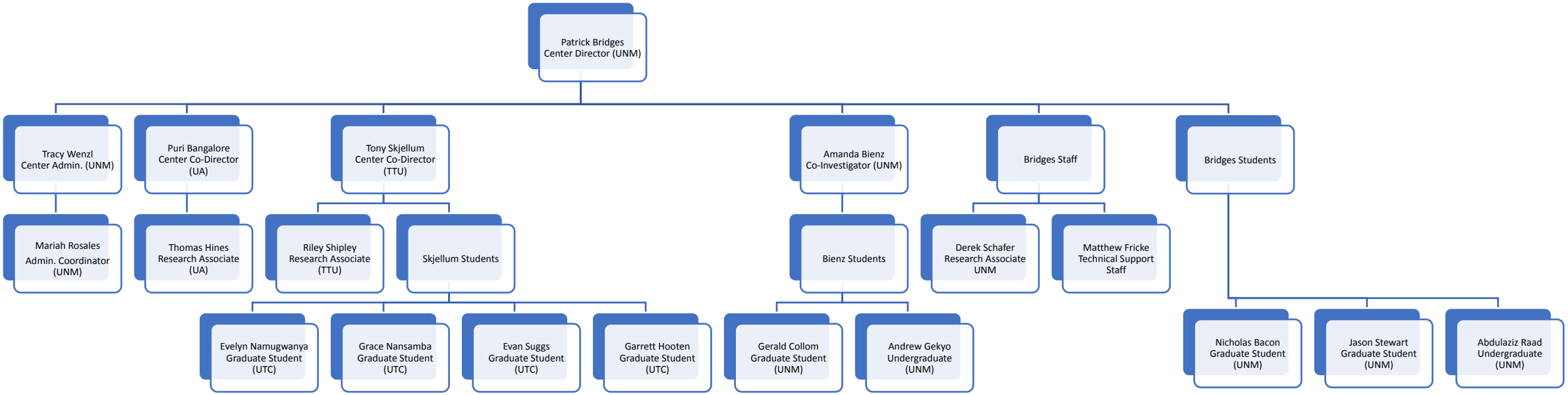
Prof. Patrick Bridges

September 28, 2023

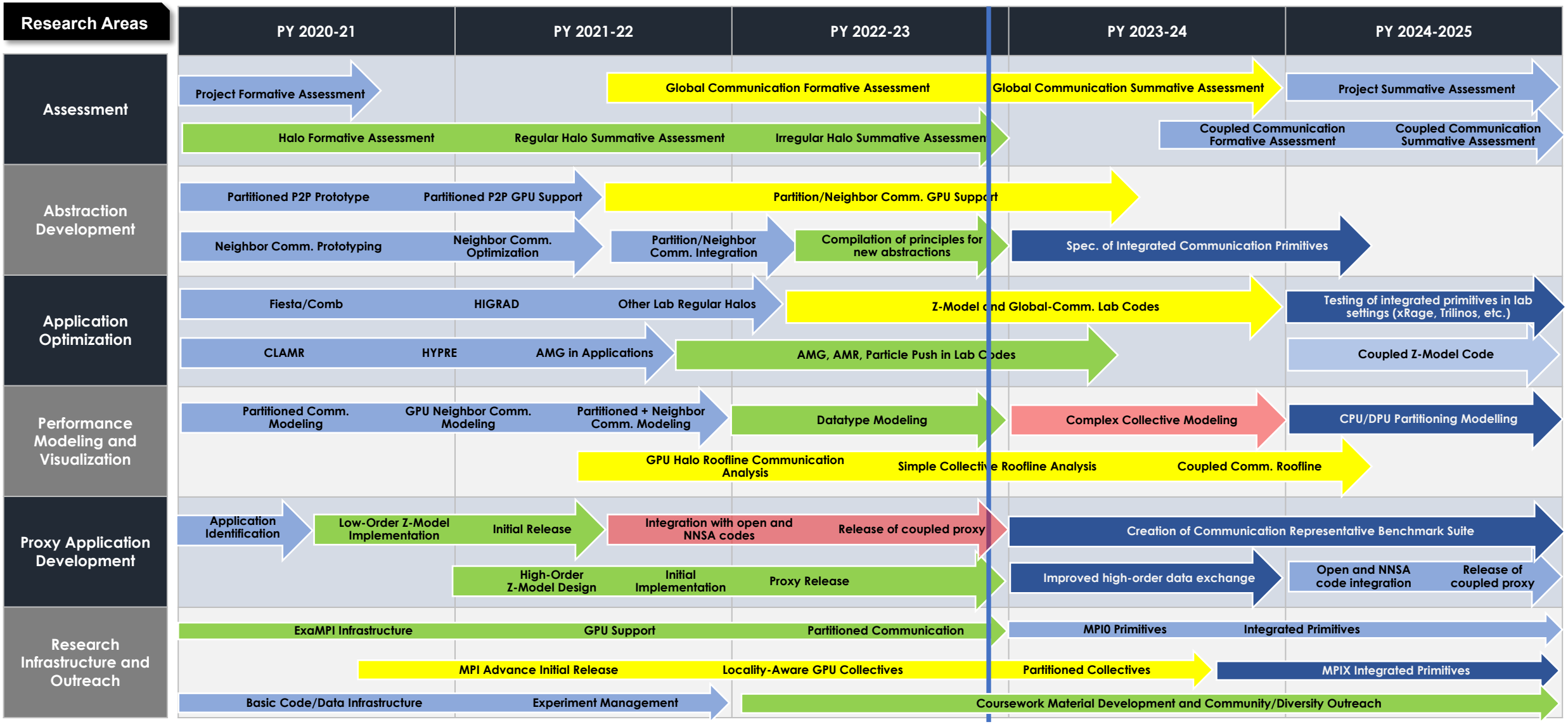
Center Goals

- Mission: “Provide optimized, performance-transparent communication systems for NNSA exascale applications.”
- **Goal: Research, demonstrate and deploy better communication abstractions that make NNSA mission applications faster, more predictable, and easier to write**
- Approach
  1. Revisit and re-architect the relationship between exascale communication systems, applications, and hardware to support transformative scientific insights
  2. Research communication system innovations that accurately quantify, predict, abstract, and optimize exascale communication systems
  3. Develop and integrate enabling technologies and leverage these fundamental research advances in support of NNSA applications and systems
  4. Continuously refine research, development, and system integration based on feedback from NNSA collaborators and stakeholders.

Center Personnel and Organizational Structure

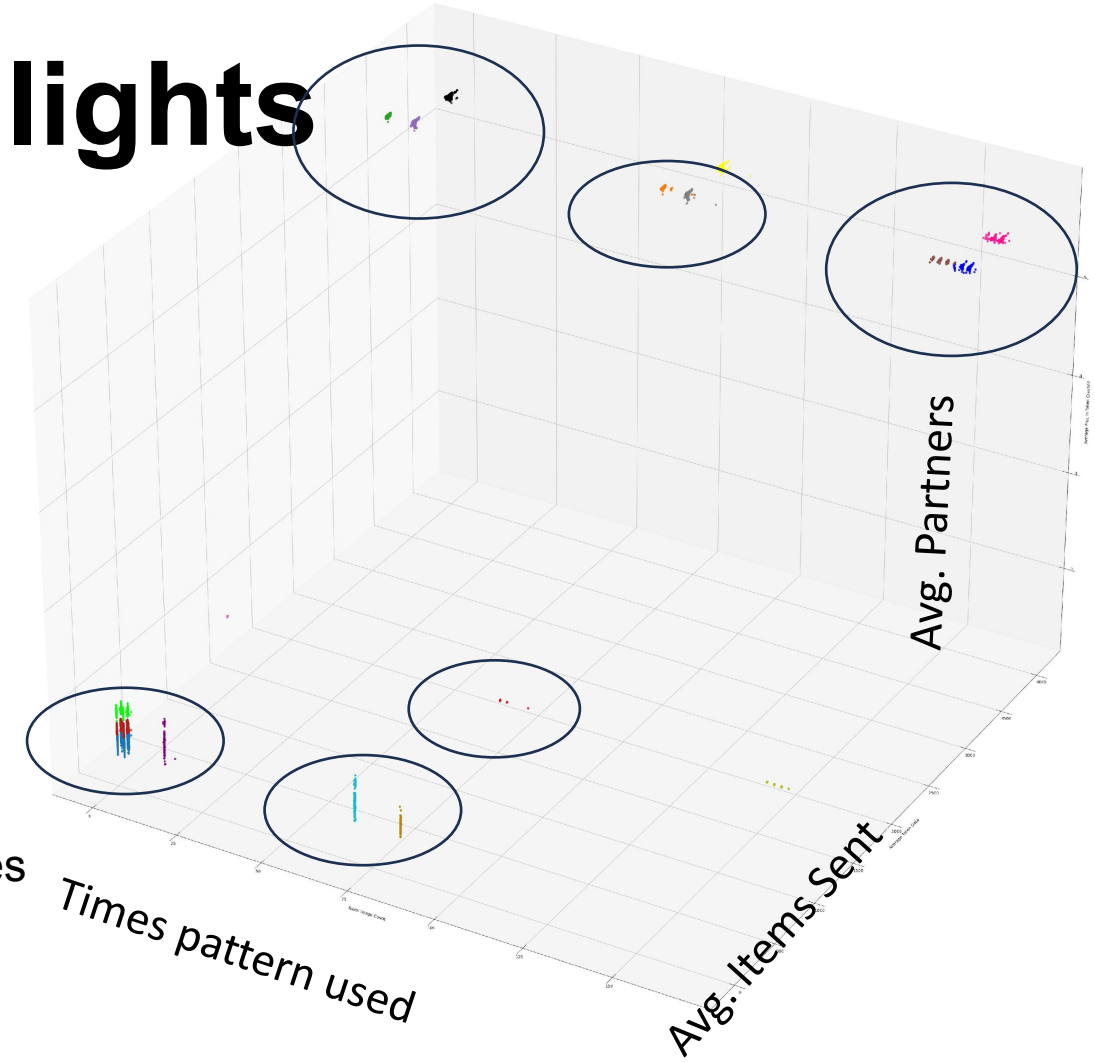


Updated 5-year Project Roadmap



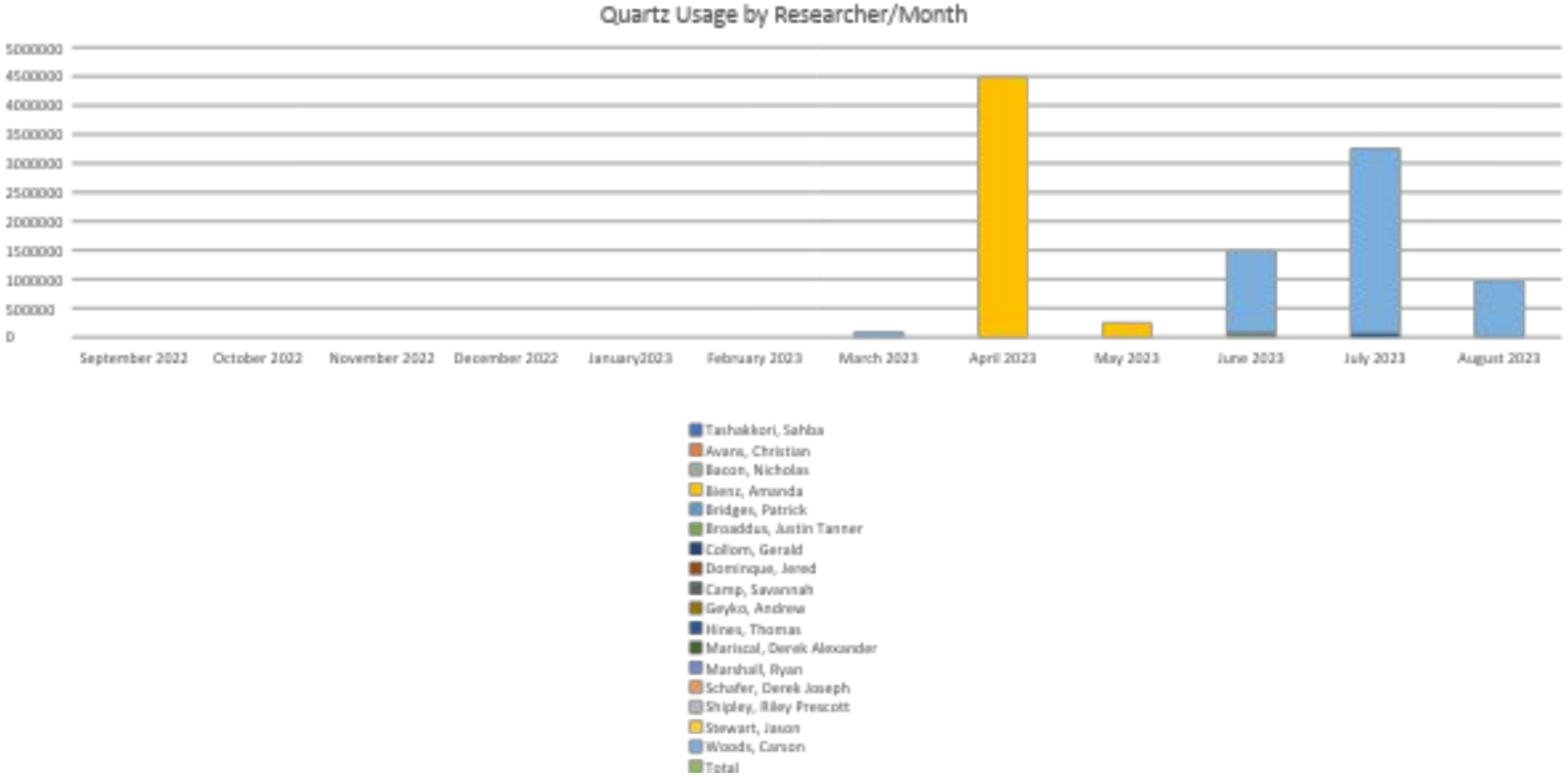
Selected Research Highlights

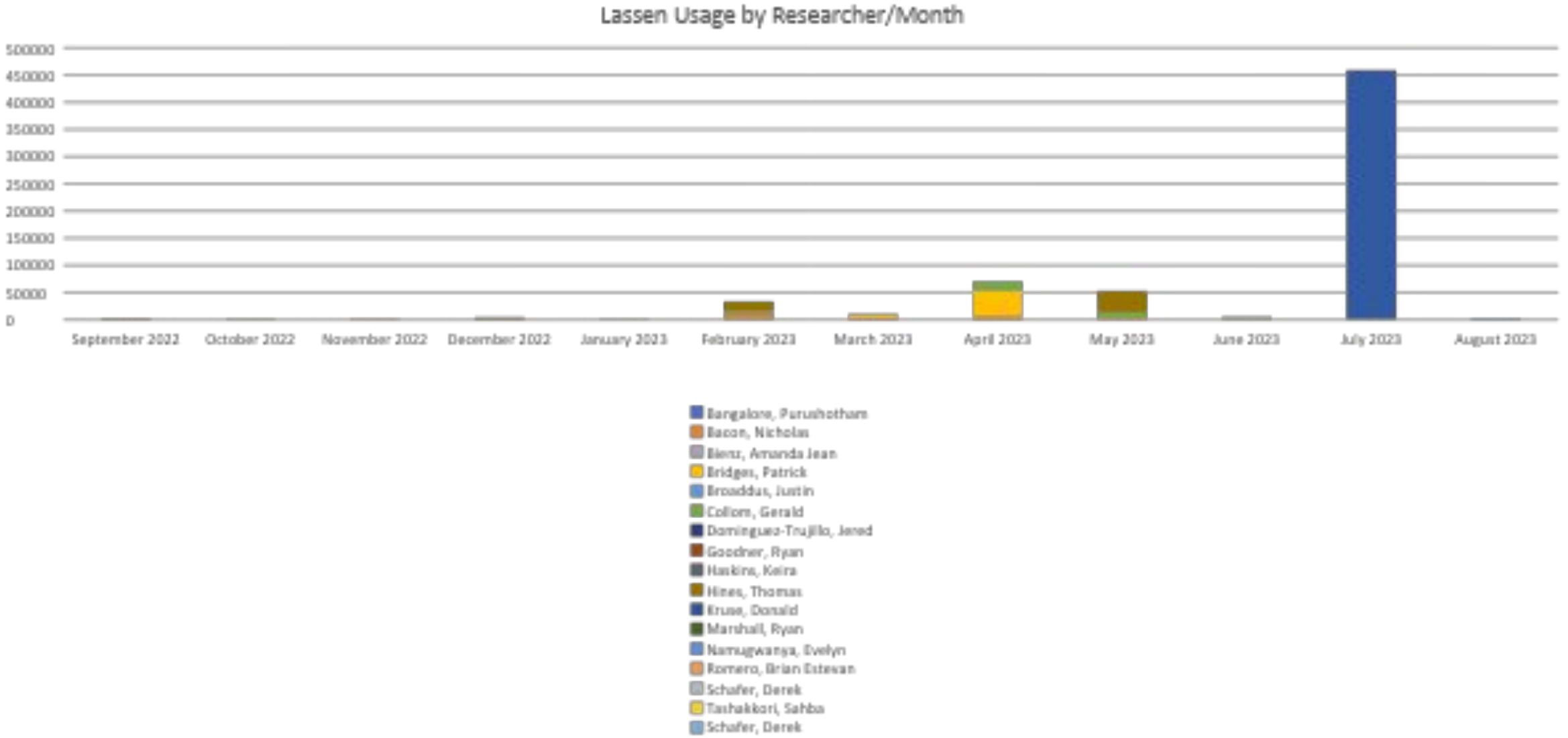
- Multiple high quality assessments and benchmarks
  - MPI implementation modeling and assessment
  - Irregular communication assessment
  - Fluids interface proxy application release
  - Irregular communication benchmark creation
  - Foundations for a new suite of application-informed communication benchmarks and proxies
- New insights for new primitives
  - Aggregating neighbor collectives
  - Partitioned communication for threaded communication
  - Topology creation and discovery
  - Datatypes replacements
  - Collecting principles for reimagined communication primitives
- Deployment into Trilinos and HYPRE via MPI Advance



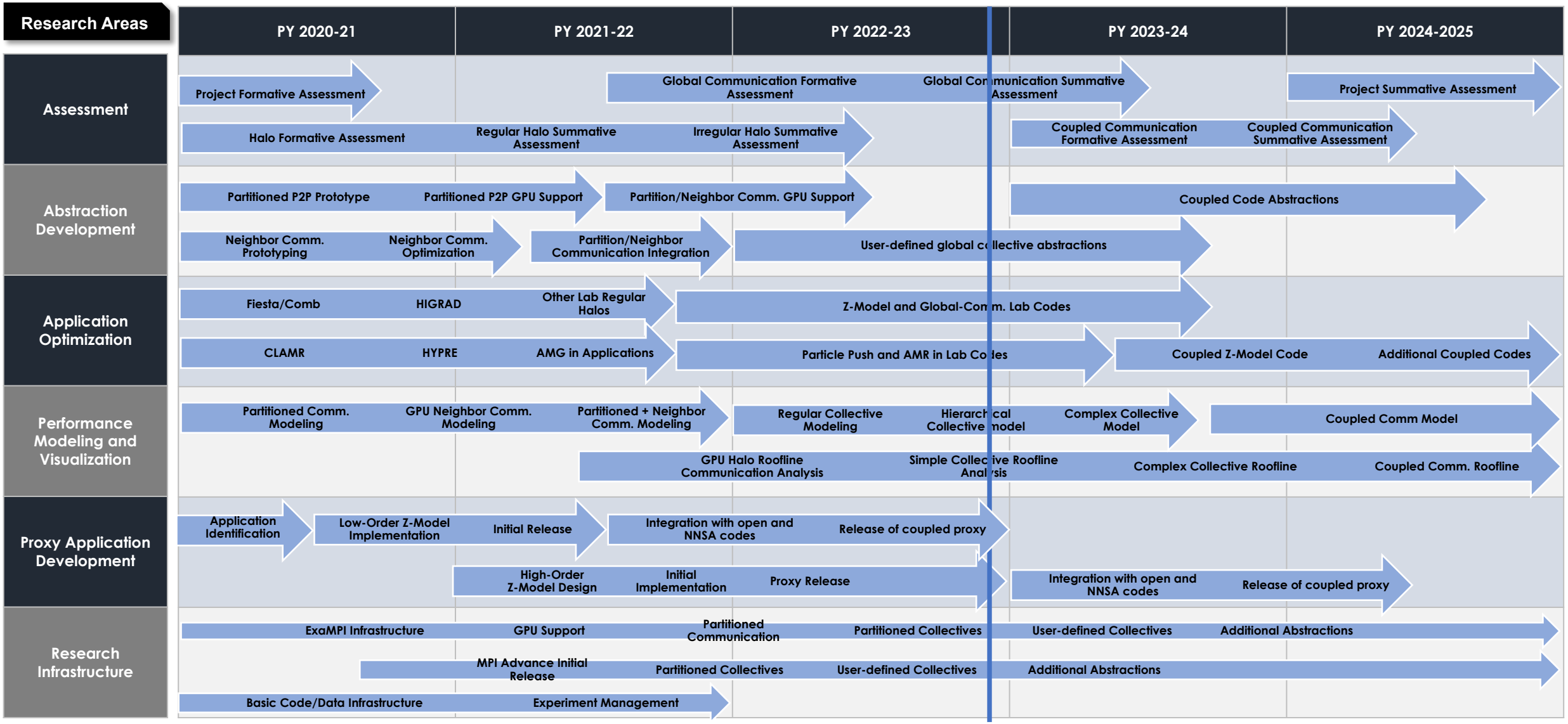
Testbed Usage

- LLNL Systems
  - Frequent performance testing on Lassen, particularly for GPU datatype testing and low-level communication primitive kernel-triggered communication
  - Scaling tests of performance models (machine learning and roofline) studies on Quartz
  - Started benchmarking runs on Tioga – initial runs have been low-level benchmarking (see talks today)
  - Do not yet have access to Cray stream triggering implementation
- Other systems
  - Chicoma has been occasionally challenging to build and run on (cray mpich + gpu aware MPI + spack is tricky to get right), but LANL support has provided the support needed
  - Using LANL Darwin for xRage under separate LANL support;
  - UNM PSAAP A100 systems useful for fast turnaround but most testing transitioned to lab platforms
  - Planning to rebudget Y3+ equipment funds to personnel
- Other potential needs
  - Access to DPU systems for testing performance and performance models (will discuss with Sandia contacts)


Quartz Usage by Researcher/Month
5000000
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
September 2022
October 2022
November 2022
December 2022
January2023
February 2023
March 2023
April 2023
May 2023
June 2023
July 2023
August 2023
Tashakkori, Sahba
Avara, Christian
Bacon, Nicholas
Bienz, Amanda
Bridges, Patrick
Broaddus, Justin Tanner
Collom, Gerald
Dominique, Jared
Camp, Savannah
Geyko, Andrew
Hines, Thomas
Mariscal, Derek Alexander
Marshall, Ryan
Schafer, Derek Joseph
Shipley, Riley Prescott
Stewart, Jason
Woods, Carson
Total


Lassen Usage by Researcher/Month
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
September 2022
October 2022
November 2022
December 2022
January 2023
February 2023
March 2023
April 2023
May 2023
June 2023
July 2023
August 2023
Bangalore, Purushotham
Bacon, Nicholas
Benz, Amanda Jean
Bridges, Patrick
Broaddus, Justin
Collom, Gerald
Dominguez-Trujillo, Jared
Goodner, Ryan
Haskins, Keira
Hines, Thomas
Kruse, Donald
Marshall, Ryan
Namugwanya, Evelyn
Romero, Brian Estevan
Schafer, Derek
Tashakkori, Sahba
Schafer, Derek

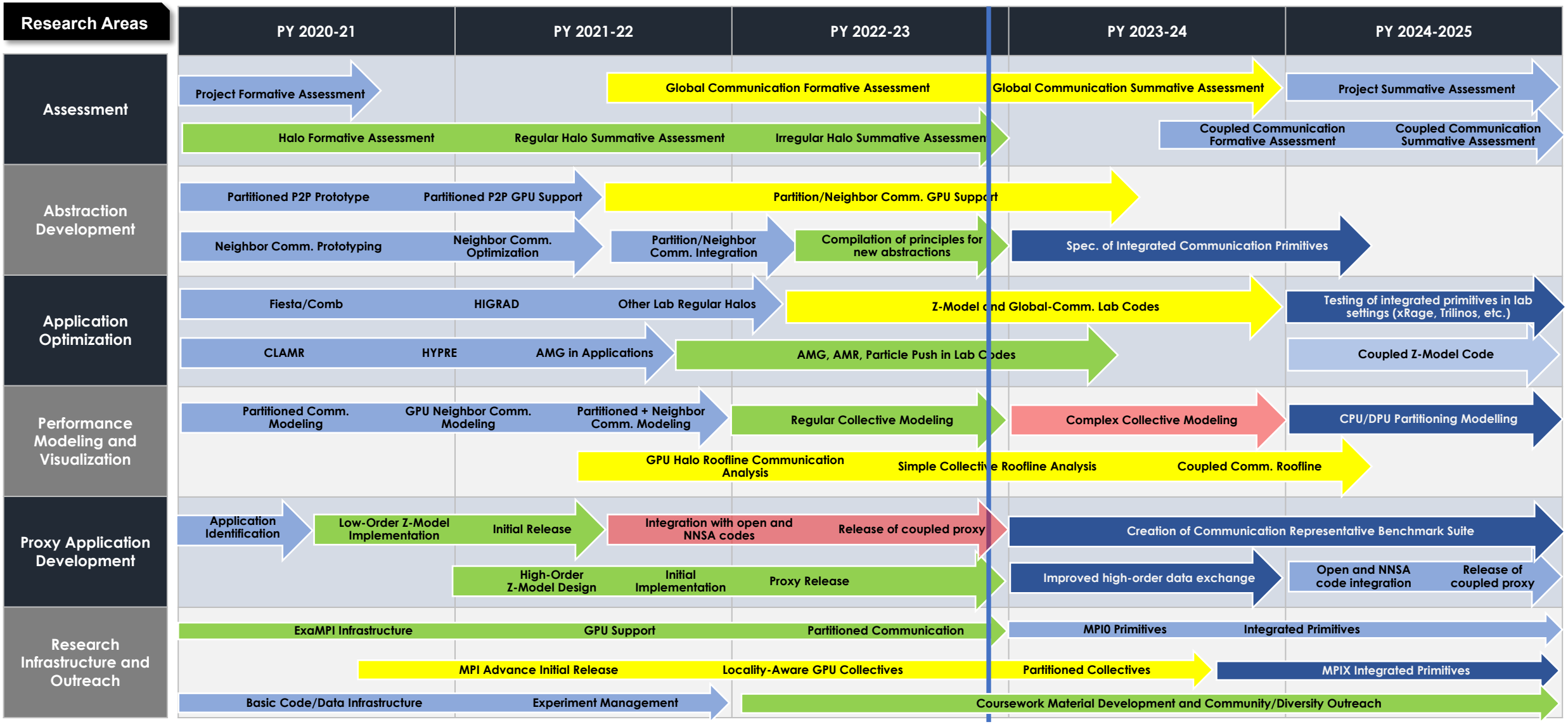
Original 5-year Project Roadmap



Proposed Focus of Remaining 2 years

- Specification and initial implementation of comprehensive benchmark suite that exposes communication behaviors relevant to DOE applications
  - Existing apps, proxies, and mini-applications when appropriate
  - Inputs for proxies, mini-apps, and macro- and micro- benchmarks derived from lab applications
  - New coupled application proxy
- Modeling of coupled code tradeoffs on CPU/GPU/DPU hardware – already begun discussion of the challenges with HPC use of DPUs with NVIDIA
- Specification and initial implementation of new primitives *beyond MPI*
  - Integrate lessons learned from 30 years of MPI, and datatype, collective, and partitioning studies
  - Low-level primitives for frameworks and high-level APIs for app programmers
- Demonstration and evaluation in lab applications and frameworks
  - Beyond MPI abstractions in lab frameworks (e.g. Kokkos, Trilinos, Cabana, etc.)
  - MPIX abstractions (including limited backports from beyond MPI) in lab applications

Updated 5-year Project Roadmap



Summary of changes

- Global communication assessment and optimizations partly delayed
  - FFTs under close study Discussing global sort testing with lab personnel
  - Need to identify specific particle code tests to run - VPIC-Kokkos?
- GPU-triggered partitioned communication delayed
  - Challenges with lack of access to quality GPU triggering abstractions remain
- Coupled benchmark deferred to Spring 2024 after discussions with Jon Reisner
  - Dropped low-order coupled benchmark
  - Implementing global binning in Beatnik as another global communication benchmark
- Modeling studies exposing limits of roofline-based approach - refocusing
- Identifying specific challenge questions/problems to maximize impact

Lab Internships, Placements, and Interactions

- Internship placements for current students (many new)
  - Nick Bacon (UNM Ph.D.): Weekly meetings with Sandia staff (Ferreria, Levy). Internship to be scheduled
  - Jason Stewart (UNM M.S.): Periodic meetings with ECP CoPA team (LANL, ORNL). Internship under discussion
  - Gerald Collom (UNM Ph.D.): Additional summer internship with LLNL
  - Nicole Avans (TNTech PhD): Worked on internship-type experience with LLNL at UTC in Summer 2023
- Lab personnel placements in the past year
  - Ryan Marshall (UTC/UA) as LANL postdoc
  - Keira Haskins (UNM) as SNL staff
  - Garrett Hooten (UTC) and Tommy Gorham (UTC) as LLNL staff
- Additional lab interactions
  - Discussions with Shipman and Junghans (LANL), Dosanjh (SNL) on benchmark development
  - Discussions with Kokkos team on Kokkos communication abstraction options
  - Work with Trilinos and HYPRE team on integration of MPIAdvance into these libraries
  - Discussions with Dosanjh (SNL) and Schonbeim (SNL) on MPI0, Portals, and MPI Standardization
  - Discussions and collaboration with Boehme (LLNL), O. Pearce (LLNL) on tools and instrumentation integration with MPI
  - Discussions and collaboration with R. Pearce on applying MPI0 to YGM acceleration in MPI environments
  - Discussions and collaboration with M. Curry (SNL) on software engineering infrastructure for HPC software stacks (e.g., ATSE)

Other Project Changes

- Anthony Skjellum moved from UT-Chattanooga to Tennessee Tech
- Thomas Hines moved from UTC to UA
- Switched UNM funding from a postdoc to two undergraduates (Raad, Gekyo) based on personnel availability
- Planning to rebudget equipment funds (hardware, additional nodes) to additional personnel

Education Activities

- Hackathons
  - Weekly online mini-hackathons with students to help with their research
  - In-person hackathons for students focused on specific research problems
    - February 2023 – Partitioned Communication
    - September 2023 – Topology creation and management
- Colloquiua by DOE lab staff focused on MPI, threading and GPU issues
- Course Development
  - Homework assignments created and implemented by center students
  - Book materials outlined and under development

Diversity and Outreach Activities

- Student recruiting – continued focus on training and recruiting diverse students
  - UNM PSAAP staff member Fricke and grad student Bacon supporting and training UNM's cluster competition team
  - Supports efforts on diversity in student recruiting
- Grace Hopper
  - Decided after feedback to look into hosting a tutorial on MPI/HPC
  - Dates for Grace Hopper 2023 overlapped with this meeting
  - Planning and developing materials for next year's meeting
- Presentations by center researchers at SIAM CSE 2023
- SIAM PP24 Mini-symposium proposal submitted on MPI benchmarking

UNM HPC Competition Team

Photo courtesy of Stewart Copeland and Graphic Design by Carter Frost

Year 3 Milestones

- **Milestone 1: Formative assessment of irregular communication demands in DOE applications, including but not limited to the LANL HOSS application.** Assessment primarily on the LANL XRage application on LANL systems, as well as the usage of the HYPRE library by LANL benchmarks (e.g. AMG2023). Discussions of HOSS assessment underway with Marshall (LANL)
- **Milestone 2: Submission of partitioned collective abstraction specification to MPI forum for future inclusion in MPI standard and revision based on community feedback.** Preliminary partitioned collective abstractions successfully prototyped and published. Presented to MPI Forum WG; Forum is prioritizing MPI 4.1, delaying standardization.
- **Milestone 3: Design and initial implementation of GPU-triggered neighbor collective abstractions in MPI advance.** Designed Pulse benchmark to explore various mechanisms to perform GPU-triggered communication for halo exchanges. Plan to release MPI Advance that uses these GPU-triggered abstractions early in next calendar year.
- **Milestone 4: Release of higher-order fluid interface model benchmark specification, implementation, and initial performance results.** Release 1.0 with distributed high order support and support for Quartz, Lassen, and Tioga this week.
- **Milestone 5: Summative assessment of optimized performance of different GPU halo communication approaches in DOE benchmarks and applications.** Used Pulse benchmarks to fully quantify the varied tradeoffs for stencil-grid halo exchanges between GPUs on heterogeneous architectures. Optimization withing DOE HIGRAD application planned for early year 4

Year 2 Review Recommendations

Main Recommendations

1. Technical details on the Center's new MPI abstractions – See talks/posters later today
2. Models of communication that can be used to understand scaling of HPC architectures – under consideration
3. Please continue to establish a pipeline of students who can work at the labs - continuing
4. Hold a student poster session instead of student presentations – see poster session ☺️
5. Consider communication primitives for large scale AMD Accelerated Processor Units (APU) – under consideration

Additional Recommendations

1. NDAs/etc. with AMD/HPE/etc. – AMD complete, still waiting for HPE Legal
2. Create more tutorials and materials - currently in progress (hackathons, student seminars)

Year 4 Milestones

- Milestone 1: Specify initial benchmarks and codes for a curated MPI communication benchmark suite informed by assessment results from DOE applications.
- Milestone 2: Model and measure impact of improved GPU communication latency on strong scaling of key DOE benchmarks and applications
- Milestone 3: Prototype low-level abstractions and APIs to provide highest levels of performance on current and emerging hardware (MPI_0)
- Milestone 4: Evaluation of performance of locality-aware global communication primitives for improving performance of Fast Fourier Transforms (FFTs) on heterogeneous architectures.
- Milestone 5: Release of example assignments, hackathon materials, and emerging book contents, and incorporate these educational materials into HPC courses at respective institutions