CUP-ECS Center Overview

Annual Review Meeting

Prof. Patrick G. Bridges

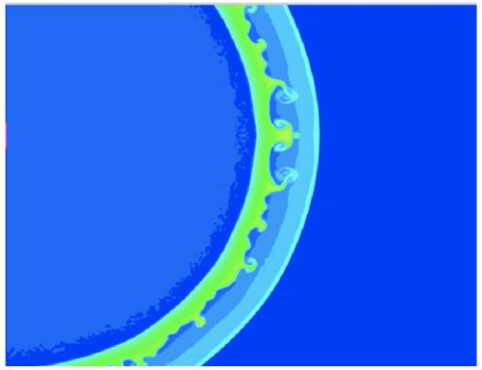
Lead PI and Center Director

August 23, 2021

Communication Challenges: Motivating Application Example

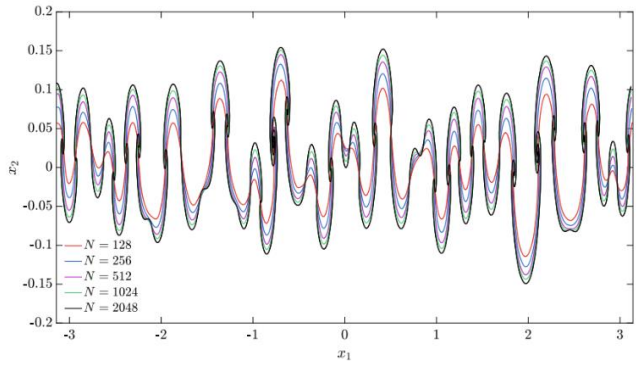
Communication overheads are key performance and design challenges for production exascale frameworks and applications

Eulerian Fluid Model (HIGRAD, xRage, etc.)



Mesh-mesh or mesh-particle remapping

Particle-based model of fluid interface



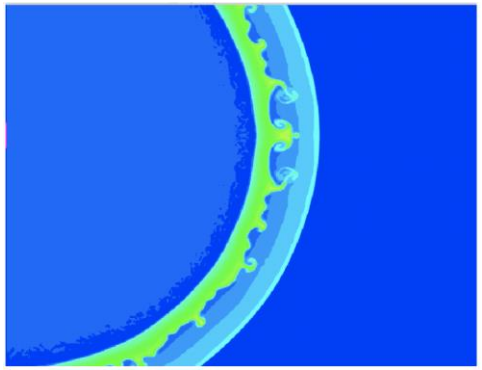
- **Explicit regular mesh CFD GPU solvers can spend 30-50% of their time in communication**
- **Low-order Z-model of interface growth spends 97% of runtime in FFT communication**

- Ramani, Raaghav, and Steve Shkoller. "A multiscale model for Rayleigh-Taylor and Richtmyer-Meshkov instabilities." *Journal of Computational Physics* 405 (2020): 109177.
- Robey, Robert W., Yuliana Yajaira Zamora, and Jenniffer Marie Estrada Lupianez. *Experience on New Architectures with the Higrad Code*. No. LA-UR-17-20718. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2017.

Communication Challenges: Motivating Application Example

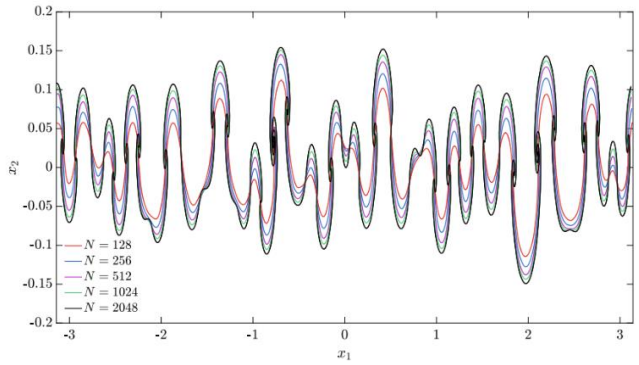
Communication overheads are key performance and design challenges for production exascale frameworks and applications

Eulerian Fluid Model (HIGRAD, xRage, etc.)



Mesh-mesh or mesh-particle remapping

Particle-based model of fluid interface



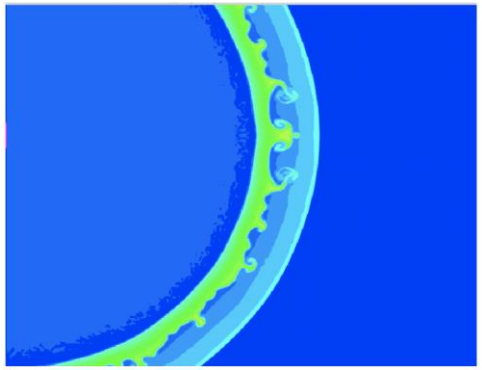
- **Irregular communication in adaptive meshes can cause anomalous 20% performance slowdowns**
- Low-order Z-model of interface growth spends 97% of runtime in FFT communication

- Ramani, Raaghav, and Steve Shkoller. "A multiscale model for Rayleigh-Taylor and Richtmyer-Meshkov instabilities." *Journal of Computational Physics* 405 (2020): 109177.
- Robey, Robert W., Yuliana Yajaira Zamora, and Jenniffer Marie Estrada Lupianez. *Experience on New Architectures with the Higrad Code*. No. LA-UR-17-20718. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2017.

Communication Challenges: Motivating Application Example

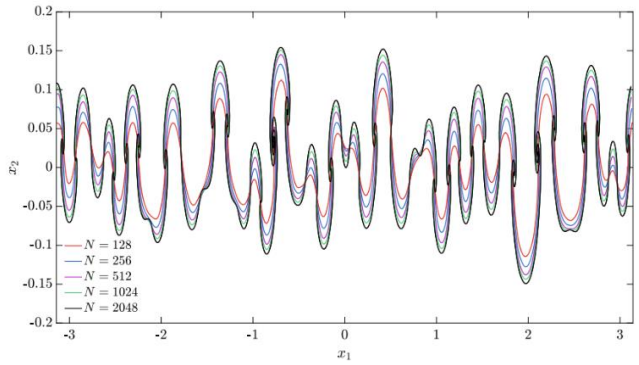
Communication overheads are key performance and design challenges for production exascale frameworks and applications

Eulerian Fluid Model (HIGRAD, xRage, etc.)



Mesh-mesh or mesh-particle remapping

Particle-based model of fluid interface



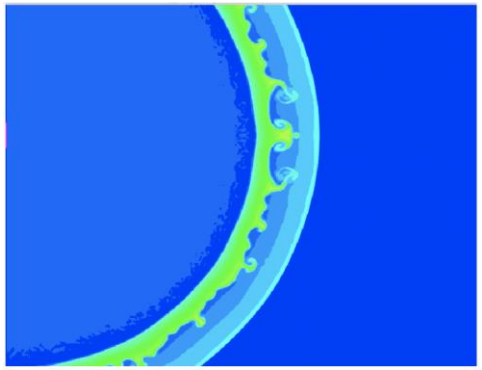
- **AMG methods used by implicit solves can spend 95% or more of their runtime in MPI communication**
- Low-order Z-model of interface growth spends 97% of runtime in FFT communication

- Ramani, Raaghav, and Steve Shkoller. "A multiscale model for Rayleigh-Taylor and Richtmyer-Meshkov instabilities." *Journal of Computational Physics* 405 (2020): 109177.
- Robey, Robert W., Yuliana Yajaira Zamora, and Jenniffer Marie Estrada Lupianez. *Experience on New Architectures with the Higrad Code*. No. LA-UR-17-20718. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2017.

Communication Challenges: Motivating Application Example

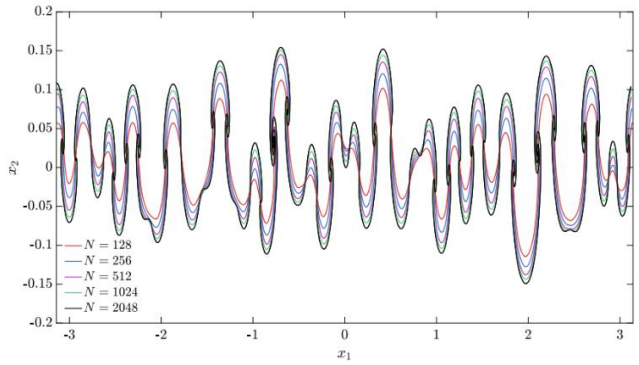
Communication overheads are key performance and design challenges for production exascale frameworks and applications

Eulerian Fluid Model (HIGRAD, xRage, etc.)



Mesh-mesh or mesh-particle remapping

Particle-based model of fluid interface



- AMG methods used by implicit solves can spend XX% of their runtime in MPI communication
- **High-order Z-model relies on fast multi-pole method with complex global communication**

- Ramani, Raaghav, and Steve Shkoller. "A multiscale model for Rayleigh-Taylor and Richtmyer-Meshkov instabilities." *Journal of Computational Physics* 405 (2020): 109177.
- Robey, Robert W., Yuliana Yajaira Zamora, and Jenniffer Marie Estrada Lupianez. *Experience on New Architectures with the Higrad Code*. No. LA-UR-17-20718. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2017.

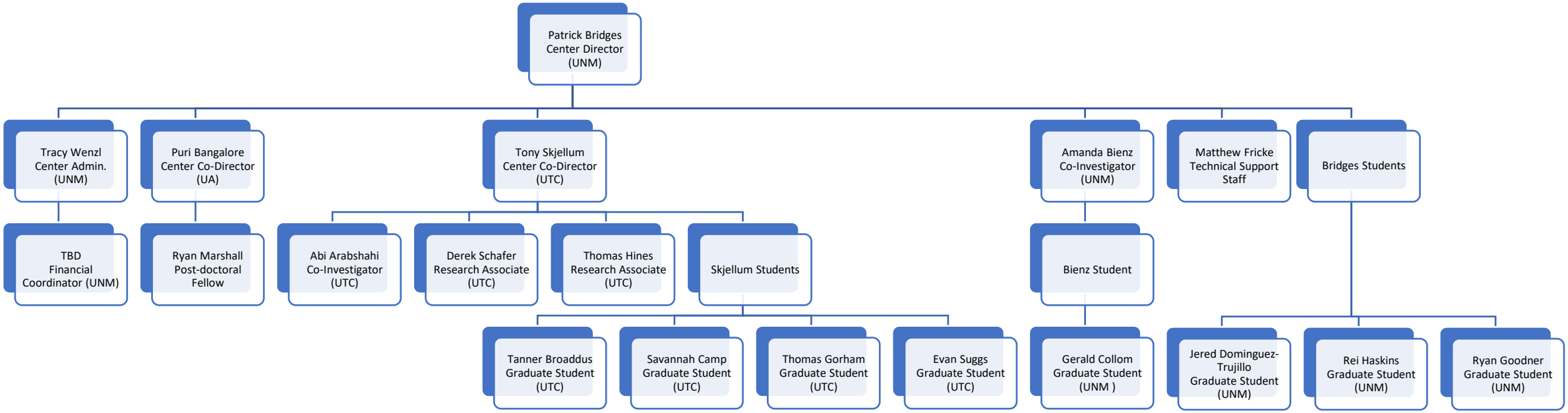
Challenge

- Communication tradeoffs in modern hardware systems and computational algorithms are very complex and hard to optimize
- **Communication abstractions** are the problem!
- There is a large **semantic gap** between:
  - **Application communication patterns** that need to be optimized
  - **Optimized communication primitives** that programmers actually use
- NNSA application and framework designers offered an impossible choice:
  - **High-level primitives** (e.g. MPI datatypes, neighbor collectives, etc.) may capture application semantics but are poorly optimized and can be hard to use
  - **Low-level primitives** (e.g. MPI_Win_XXX) yield brittle, hard-to-maintain code that may need to be re-optimized platform-to-platform
  - **“Normal” primitives** (e.g. MPI_Isend) can combine the worst features of both, especially with non-trivial communication patterns.

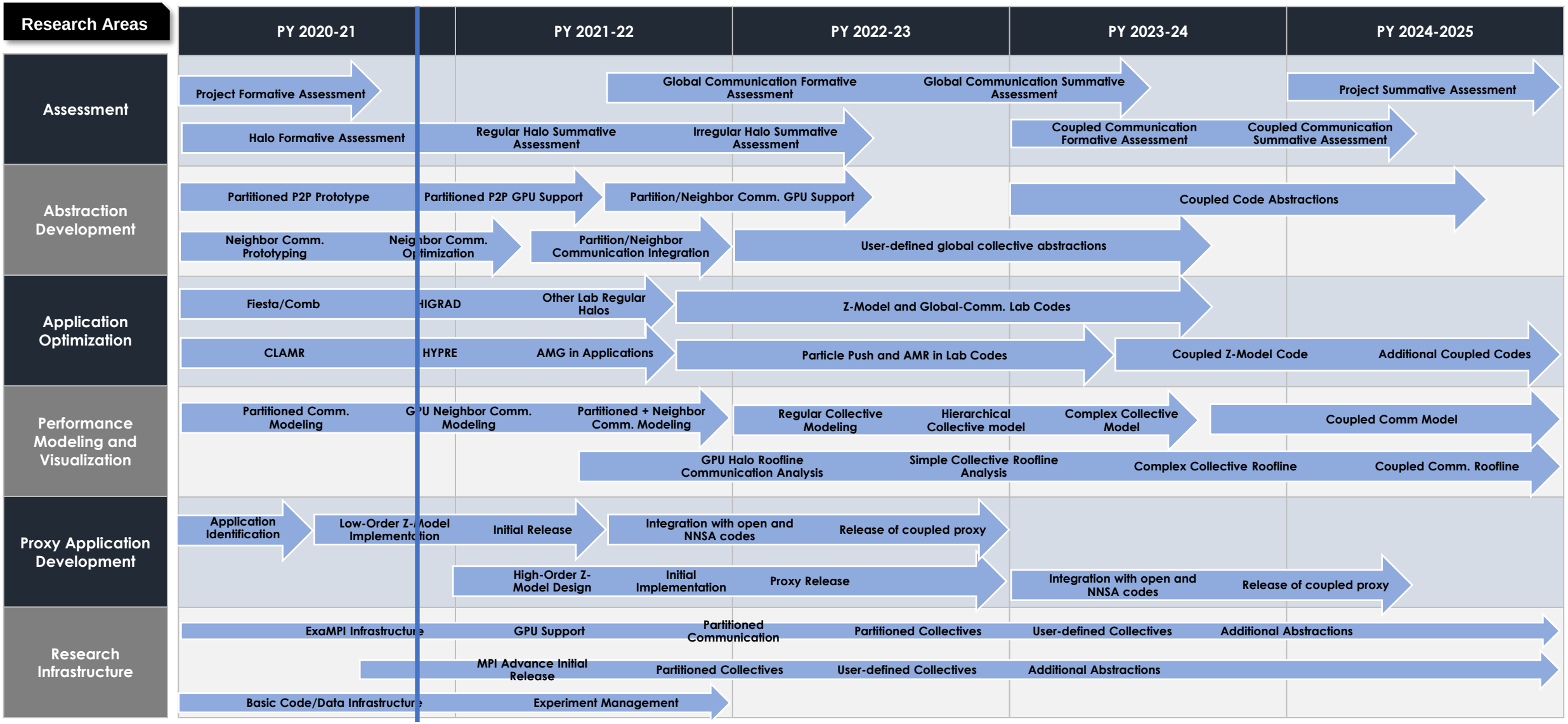
Center Goals

- 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.

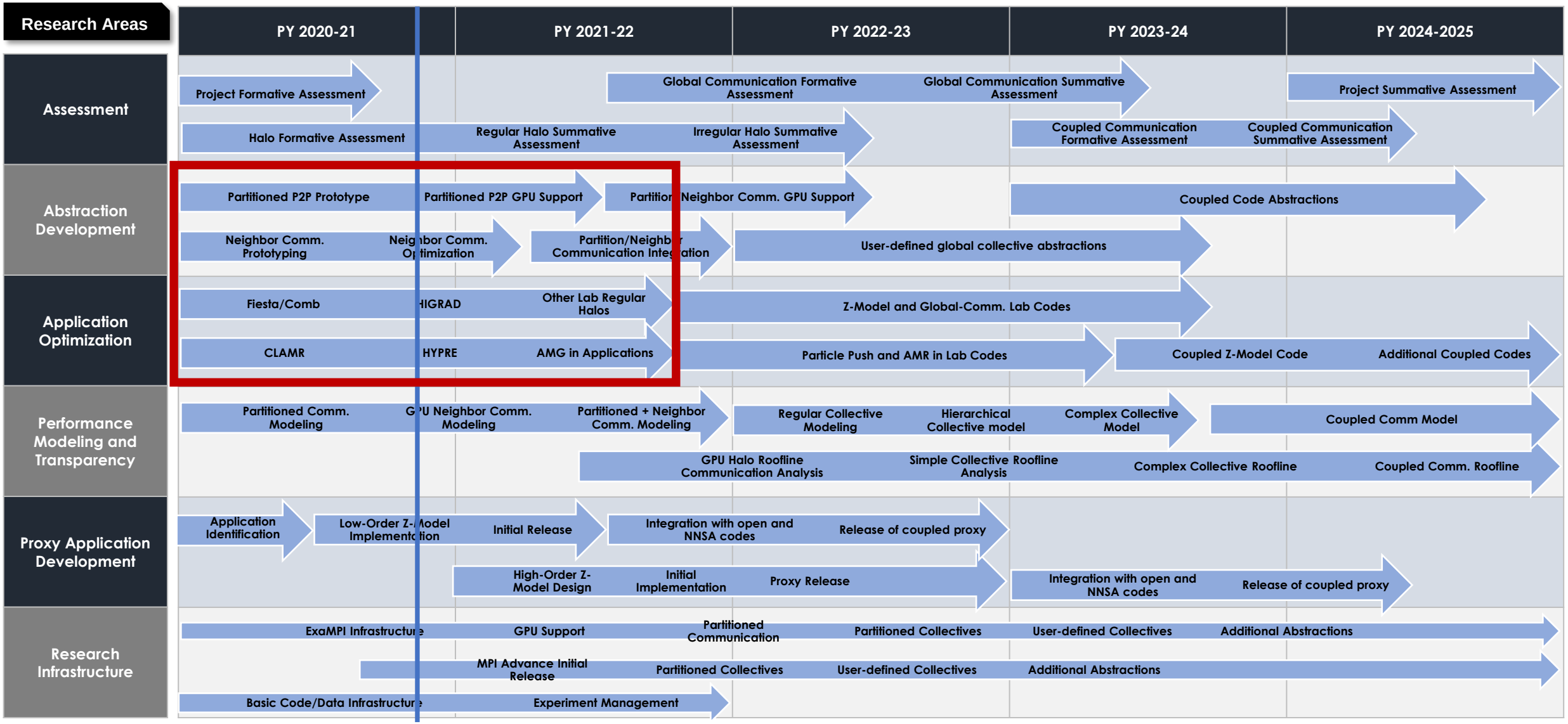
Center Personnel and Organizational Structure



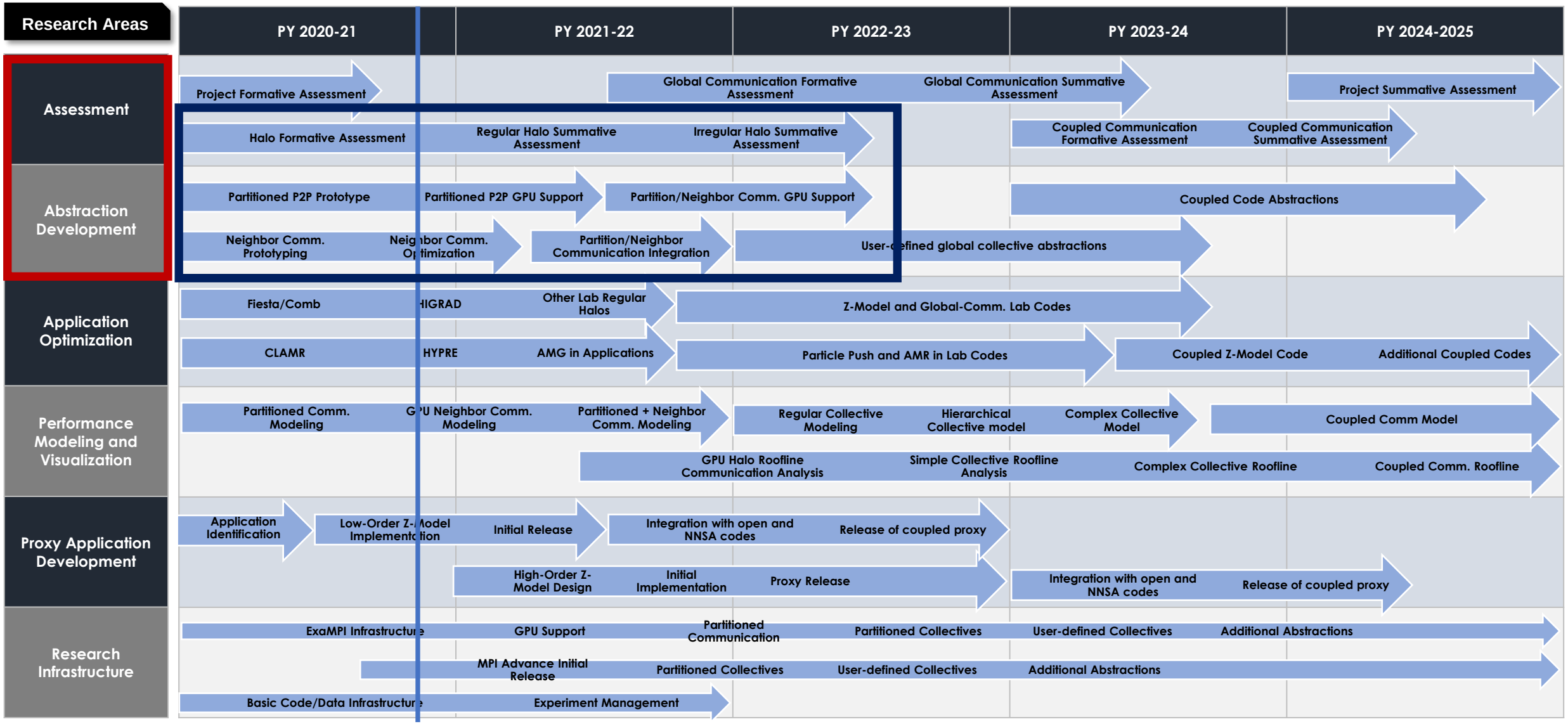
5-year Project Roadmap



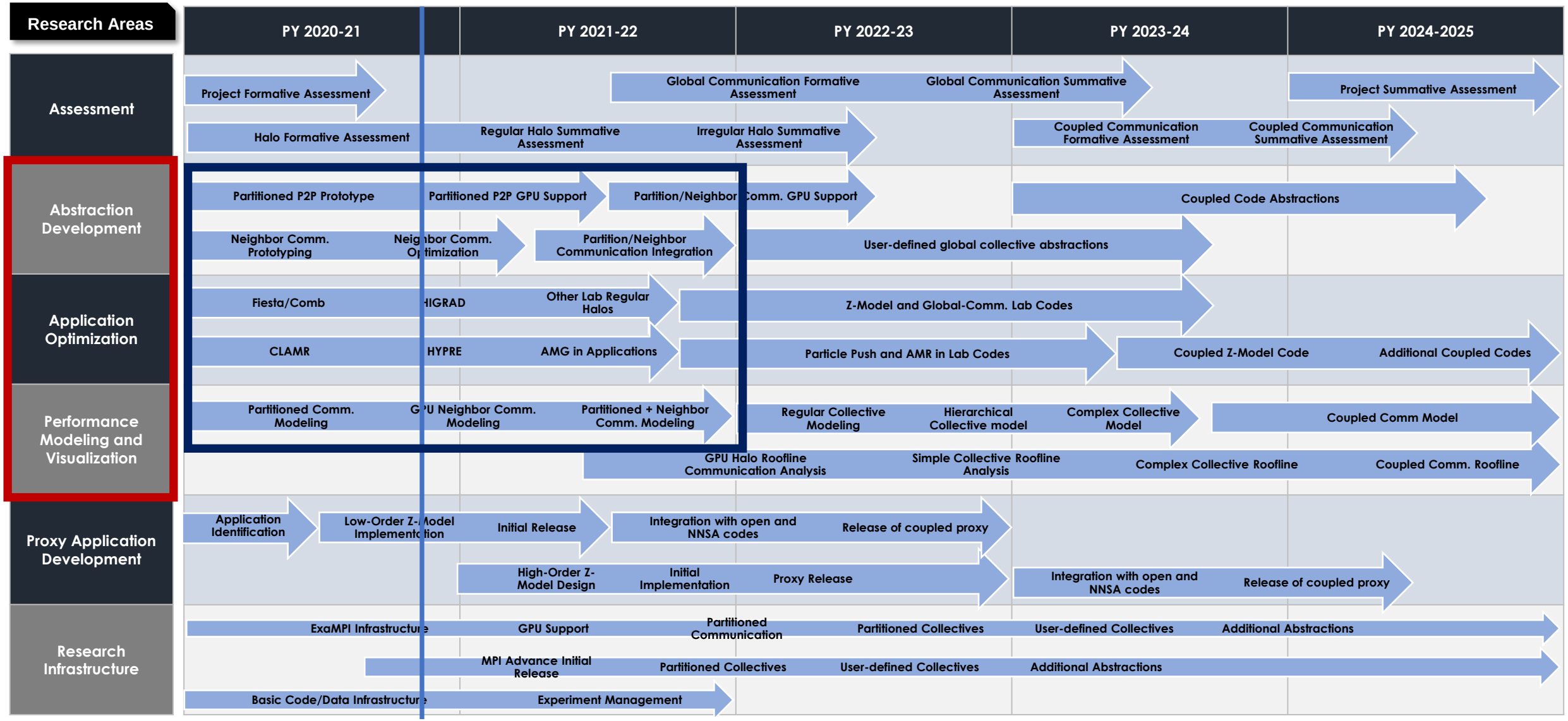
5-year Project Roadmap - Integrations



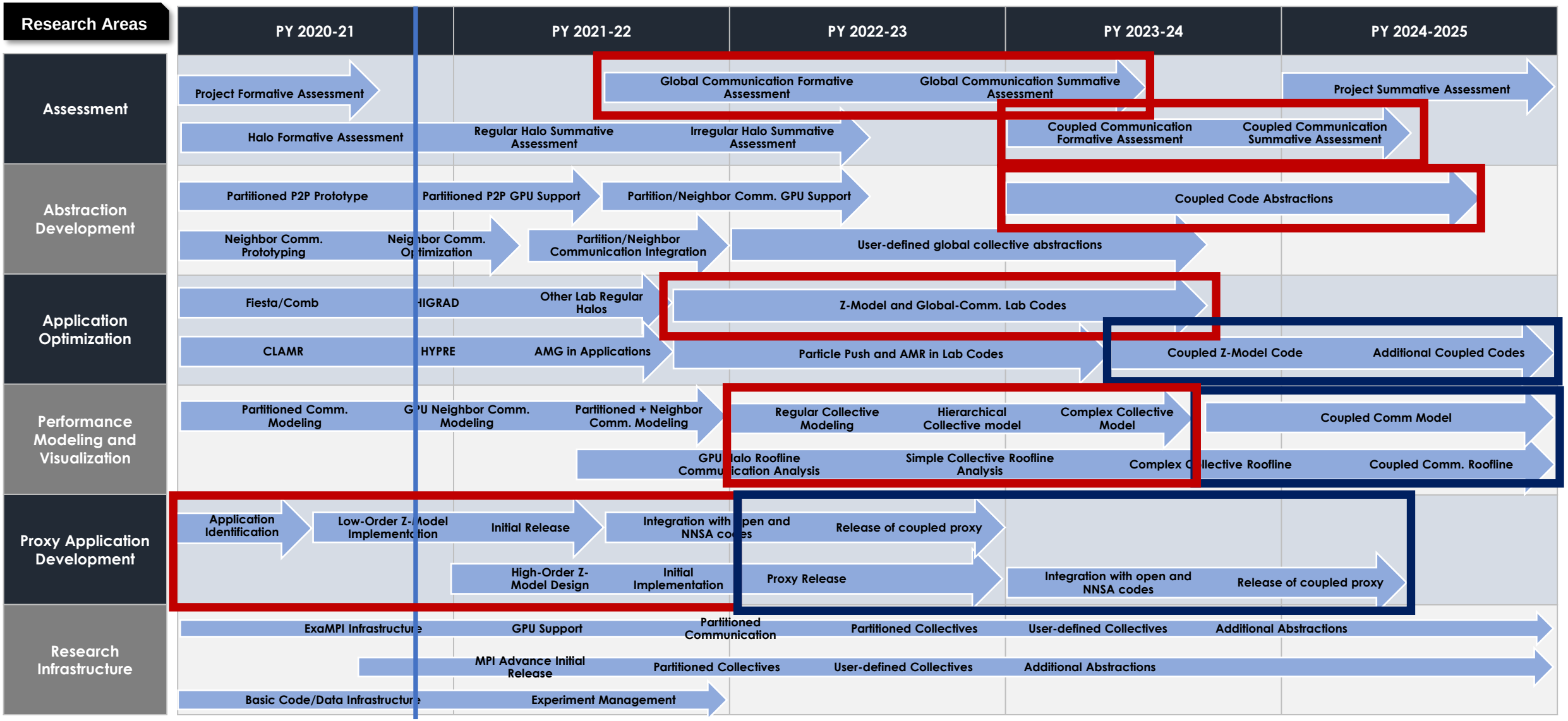
5-year Project Roadmap - Integrations



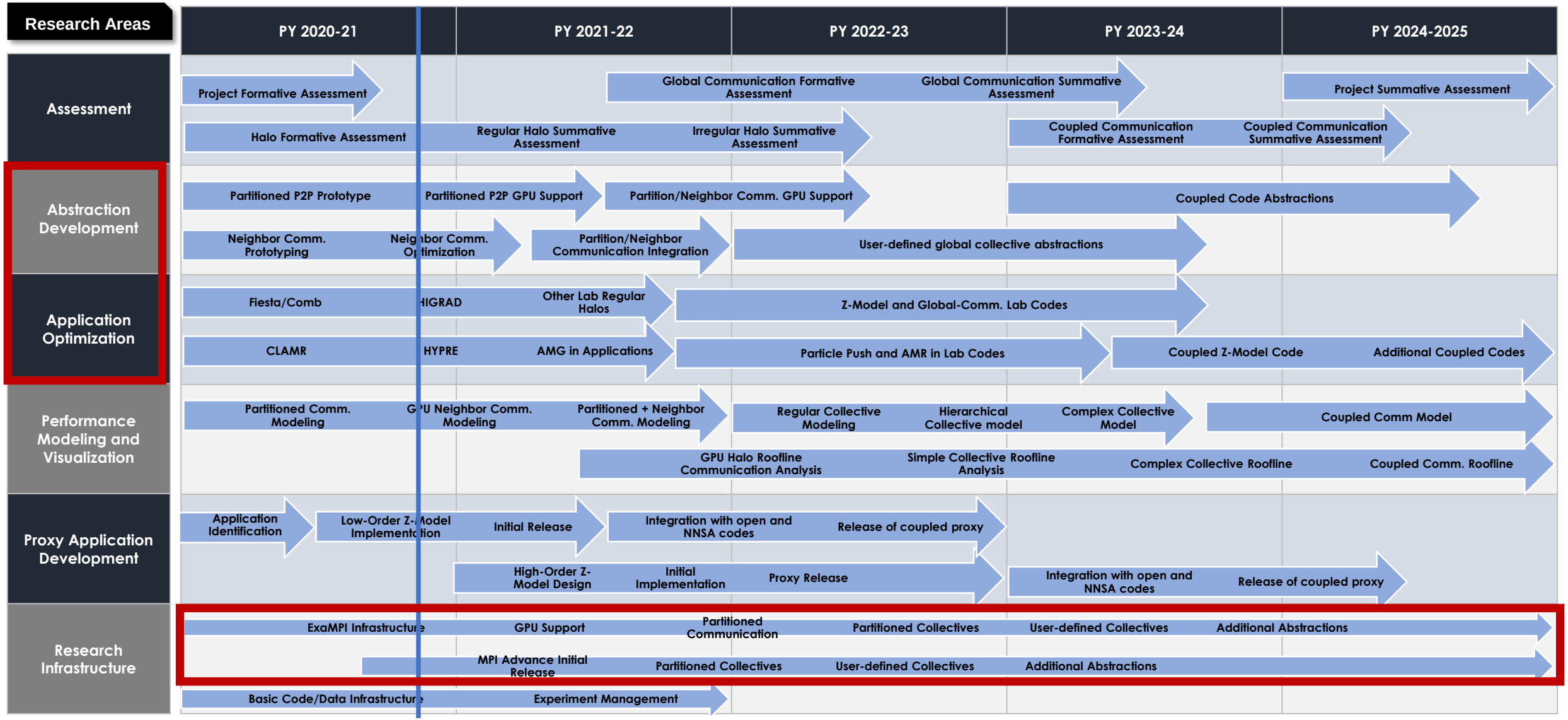
5-year Project Roadmap - Integrations



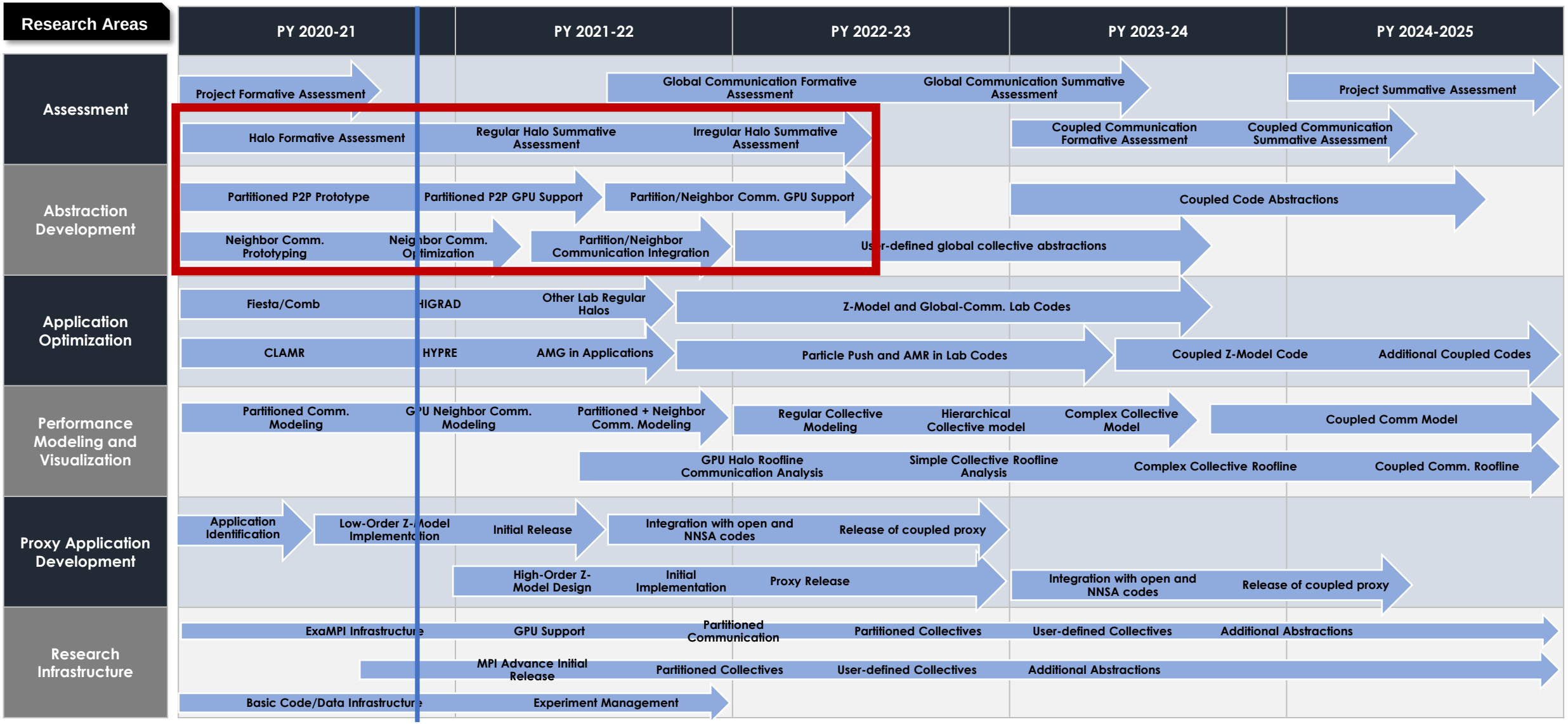
5-year Project Roadmap - Integrations



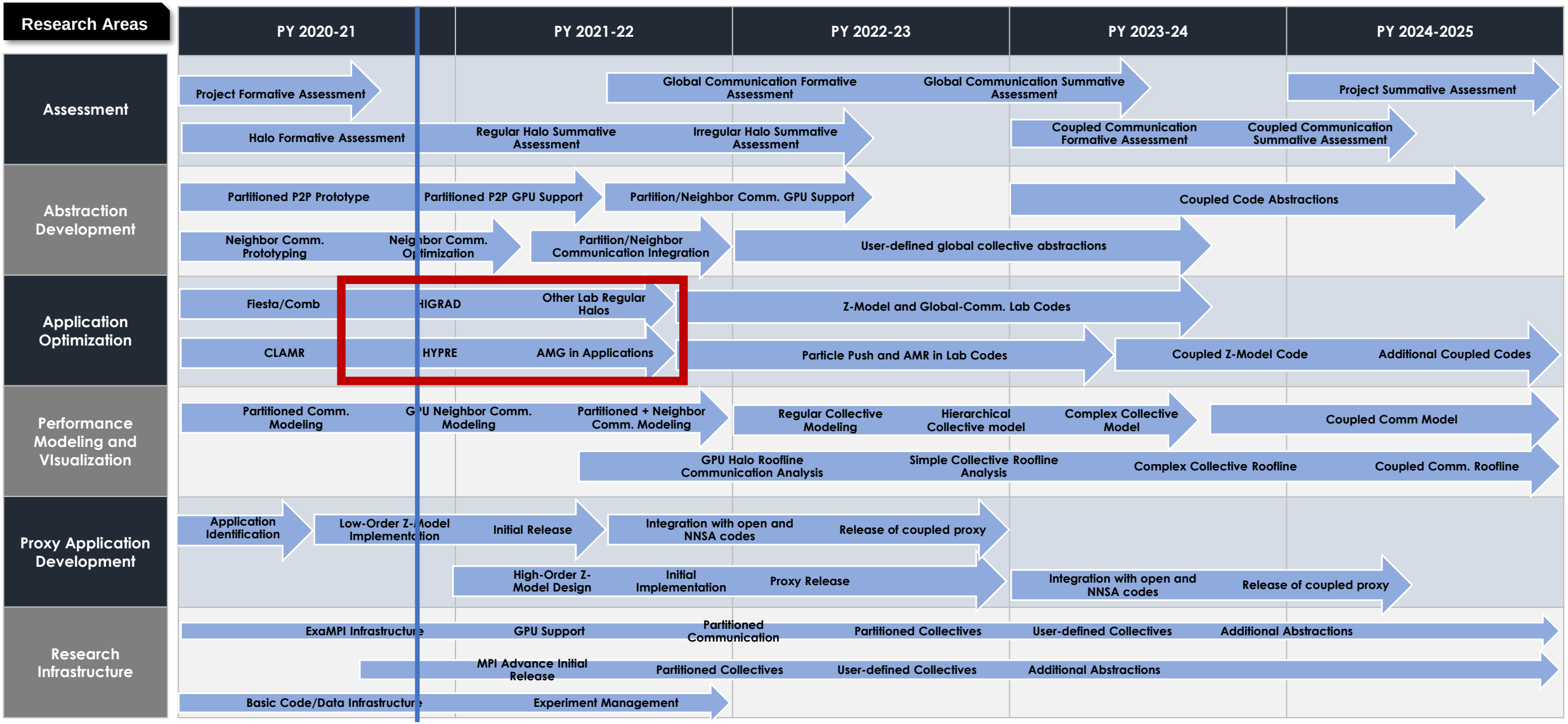
5-year Project Roadmap - Integrations



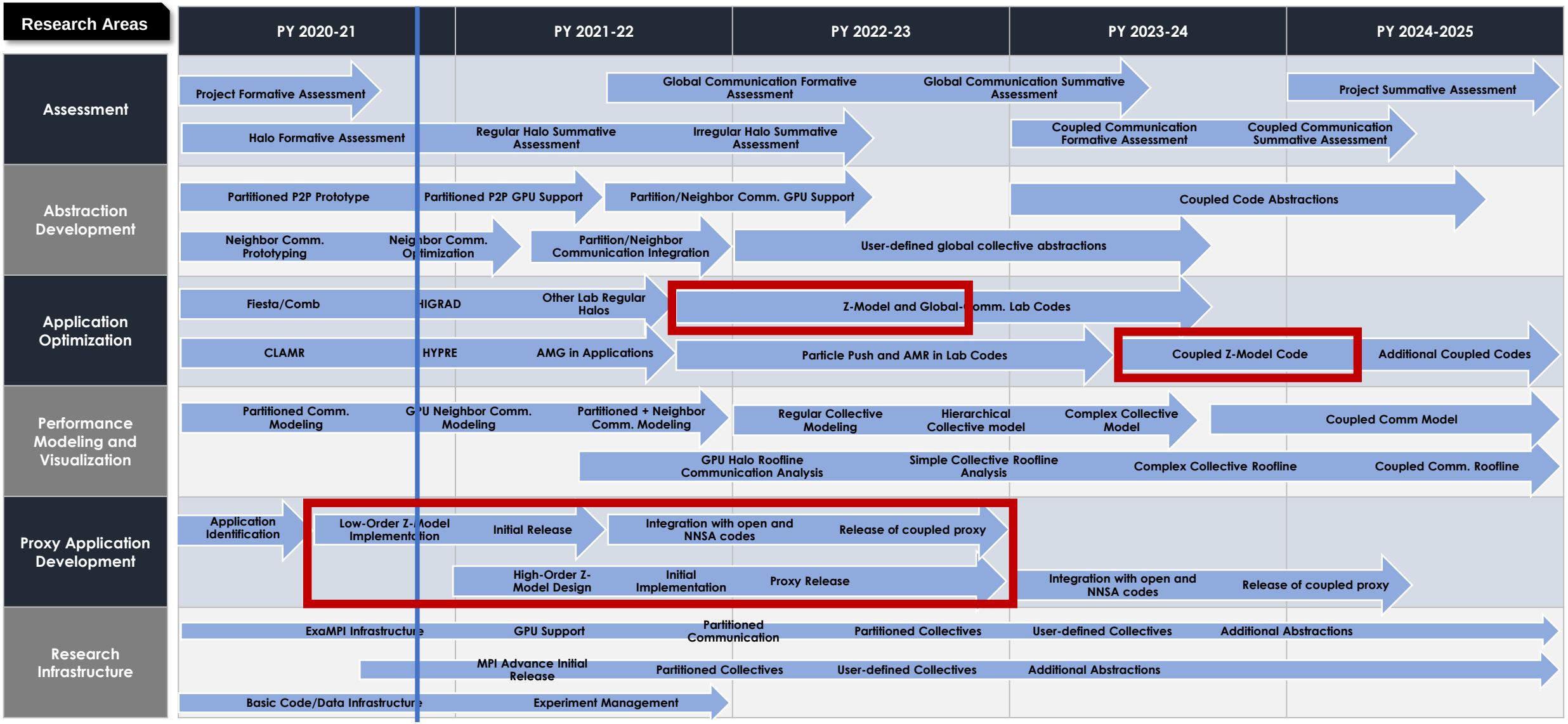
5-year Project Roadmap - Changes



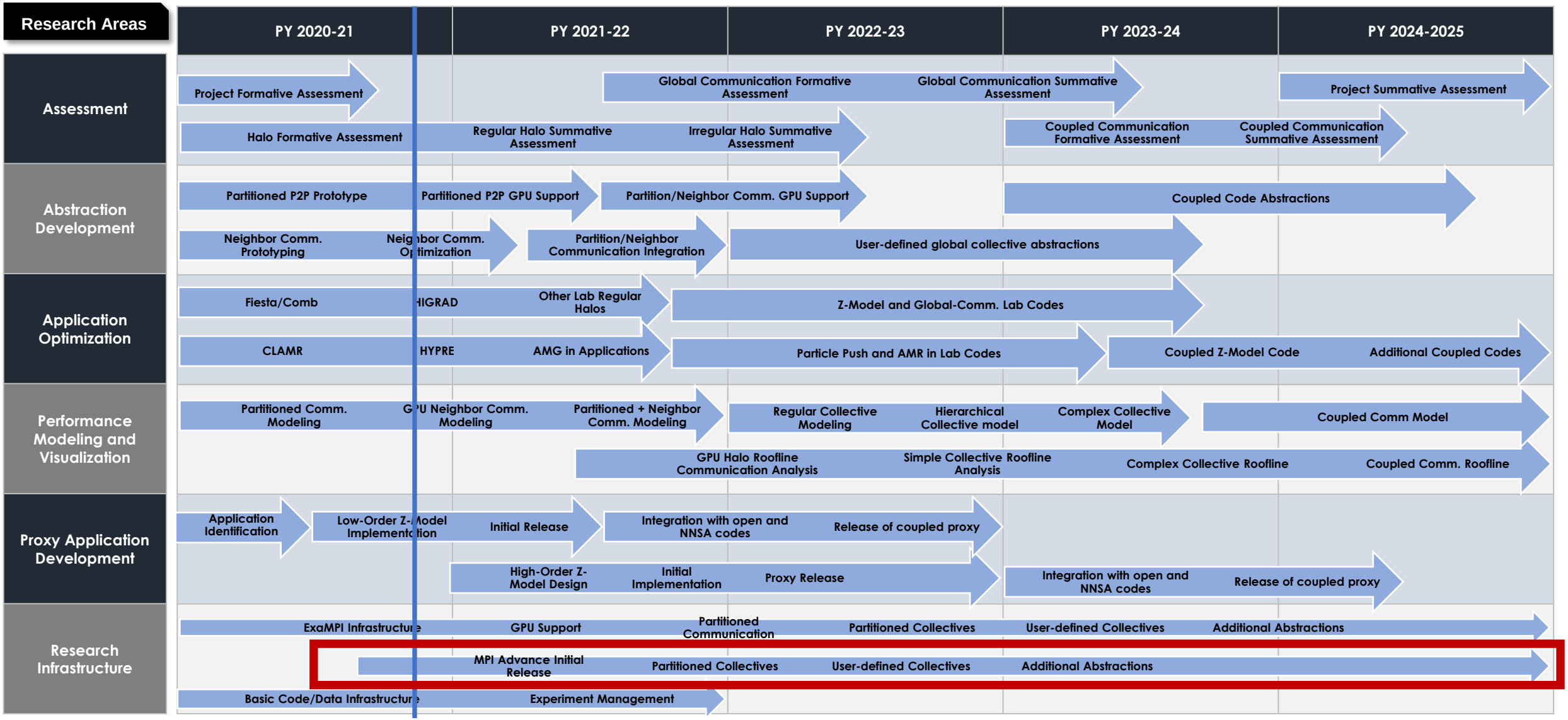
5-year Project Roadmap - Changes



5-year Project Roadmap - Changes



5-year Project Roadmap - Changes



Other Project Changes

- Addition of co-investigator Amanda Bienz (UNM) to project
  - Rebudget from a postdoc and grad student to faculty time and grad student
- Move of Prof. Puri Bangalore from UAB to UA
  - Requested move of subaward between institutions
- Departure of senior personnel Craig Tanis from UTC and project
  - Replaced with additional senior personnel (Thomas Hines)

Project Risks and Mitigation Strategy

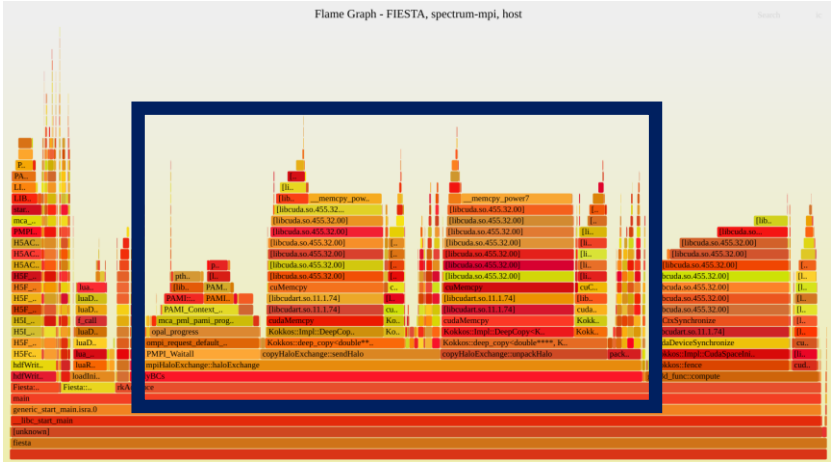
- Primary risk is in personnel loss and recruiting/retaining students
- Overall mitigation strategy
  - Aggressive student recruiting is required
  - Close collaboration/internships with national lab partners helps
  - Students enjoy and are excited by internship prospects!
- Experiences so far
  - Mitigated personnel loss in year 1 (see previous slide)
  - Student recruiting proceeding well
    - Successfully working with collaborators at other institutions to identify/recruit promising students (e.g. Tennessee Tech)
    - One student recruit (UNM) delayed due to USCIS I-20 processing delay
- Aggressive use of internships helps retain students, creates pipeline to labs

Year 1 Research Areas and Directions

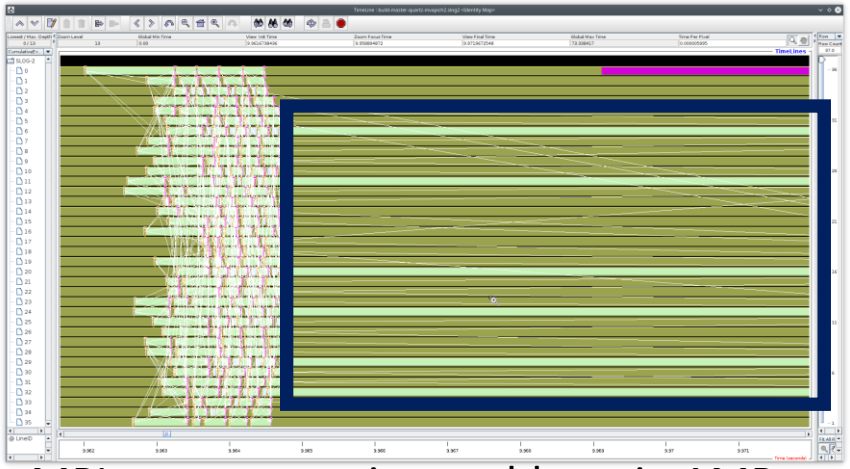
- Assessment
  - Qualitative formative assessment of DOE application communication abstraction usage
  - Quantitative formative assessment of regular and irregular halo communication challenges
- Abstraction Optimization and Development
  - Persistent and Partitioned Communication Abstractions
  - Neighbor Communication Abstractions
  - Optimization and abstraction deployment in Fiesta, Comb, and HYPRE
- Proxy Application Development
  - Identified fluid interface modelling application
  - Created parallel low-order implementation
- Performance Modeling and Transparency
  - Communication propagation of application variance
  - Partitioned communication effective bandwidth
- Research and Deployment Infrastructure
  - ExaMPI development and MPI Advance initiation for abstraction deployment
  - Code, data, and experiment management

Assessment Highlights

- Project Formative Assessment – Qualitative assessment of MPI use in DOE proxy codes, refined app focus from AST/TST feedback
- Halo Formative Assessment -
  - Fiesta – UNM Kokkos regular halo exchange similar to LANL HIGRAD production code
  - HIGRAD – LANL Fortran/OpenACC code (assessment just beginning)
  - CLAMR – Irregular AMR halo exchange similar to ones used in production AMR codes
  - HYPRE – coarse AMG exchanges dominated by comm. Costs
- Details in upcoming talks and posters
- Related Publication (prior to award)
  - Nawrin Sultana, et al. Understanding the use of message passing interface in exascale proxy applications. *CCPE*, 2020.



Original Fiesta (Kokkos) spends >50% of time spent in MPI running on Lassen



MPI progress engine problems in AMR halo exchanges can slow app by 20%+

Halo Exchange Abstraction Highlights

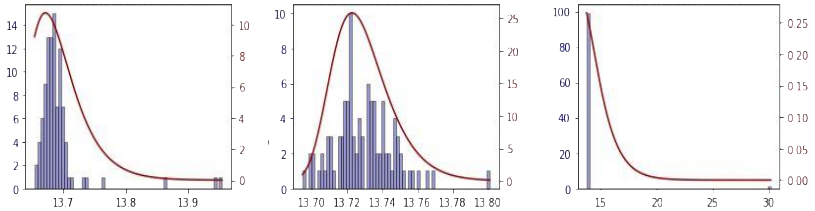
- Regular Halos
  - Optimization of Fiesta halo exchange improved Lassen performance significantly
  - Released open-source MPI partitioned communication implementation
  - More details on these steps in halo exchange talk (next) and posters this afternoon
- Irregular Halos
  - Converted HYPRE to use neighbor collectives
  - Designed extended persistent neighbor collective interface to improve optimizability
- Publications
  1. Andrew Worley et al. Design of a Portable Implementation of Partitioned Point-to-Point Communication Primitives. In *Proceedings of the Fourteenth International Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2)*, 2021.
  2. Matthew G. F. Dosanjh et al. Implementation and Evaluation of MPI 4.0 Partitioned Communication Libraries. *Parallel Computing*, 2021.

Proxy Application Development

- Identified fluid interface movement model as target proxy application
  - Low and high-order numerical models by Shkoller et al.
  - Diverse global communication demands (FFT and/or FMM)
  - Good target for coupling with open and lab CFD codes
- Implemented MPI version of 3-D low-order model
- Examining FlecSI as potential framework for high-order model impl. because of it's support for complex data structures
- Additional discussion of Z-model implementation, scaling, and next steps by Thomas Hines this afternoon

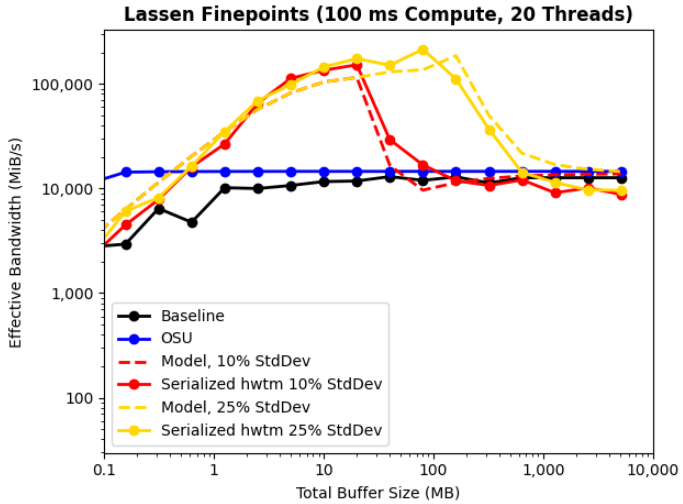
Performance Modeling and Research Infrastructure Highlights

- Performance Modeling
  - Modeled communication system propagation performance variation in HPC application codes
  - Modeled performance of partitioned communication primitives with varying communication/computation overlap
- Research Infrastructure
  - Created MPI Advance library for early application access to new MPI abstractions and optimizations
  - Enhanced ExaMPI to support additional applications and primitives for upcoming tests
- Publications
  3. J. Dominguez-Trujillo et al. Lightweight measurement and analysis of HPC performance variability. In *Proceedings of the 2020 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High-Performance Computer Systems (PMBS)*, 2020



b) HPCG workload variation and MOM fitting

Fit of statistical model to variation in HPCG runtime on Sandia Attaway cluster



Model vs. actual partitioned communication bandwidth on LLNL Lassen cluster

Other Contributions - Highlights

Additional studies and publications on many other MPI-related topics (e.g. FPGA support, language bindings, fault tolerance, etc.)

4. Daniel J Holmes et. al. Why is MPI (perceived to be) so complex? part 1—does strong progress simplify MPI? In *27th European MPI Users' Group Meeting*, pages 21–30, 2020.
5. Pouya Haghi et al. A reconfigurable compute-in-the-network FPGA assistant for high-level collective support with distributed matrix multiply case study. In *2020 International Conference on Field Programmable Technology (ICFPT)*, pages 159–164. IEEE, 2020.
6. Pouya Haghi et al. FPGAs in the network and novel communicator support accelerate MPI collectives. In *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–10. IEEE, 2020.
7. Qingqing Xiong et al. Accelerating MPI collectives with FPGAs in the network and novel communicator support. In *2020 IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM).* IEEE, 2020.
8. Derek Schafer et al. Extending the MPI stages model of fault tolerance. In *2020 Workshop on Exascale MPI (ExaMPI),* pages 52–61. IEEE, 2020.
9. Reed Milewicz et al. Negative perceptions about the applicability of source-to-source compilers in HPC: A literature review. In Proceeding of the Workshop on Compiler-assisted Correctness Checking and Performance Optimization for HPC,, 2021.
10. Tom Herschberg et al. Integrating MPI sessions with topological connection building and collective communication, September 2021. Accepted to EuroMPI 2021.
11. Anthony Skjellum et al. MPI sessions and fault tolerance: A position paper, September 2021. Accepted to EuroMPI 2021.
12. Martin Ruefenacht et al. MPI's language bindings are holding MPI back, 2021. ArXiv 2107.10566; cs.PL.

Year 2 Plan Highlights

- Abstractions and Summative Assessment
  - Complete integration and evaluation of partitioned communication primitives in Comb proxy and Fiesta/HIGRAD applications
  - Begin integration and evaluation of optimized irregular neighbor collectives in HYPRE
  - Global communication formative – particle codes (EMPIRE, etc.), fast Fourier and fast multipole methods (Z-Model, FlecSPH, etc.)
- New Abstraction Development
  - Evaluation and optimization of prototype GPU partitioned communication in Comb proxy application
  - Design of partitioned neighbor collective abstraction as a general optimized halo exchange communication mechanism
- Fluid Interface Proxy
  - Design and begin implementation of parallel version of higher-order fluid interface model for use as a stand-alone proxy
- Research Infrastructure
  - Initial release of MPI Advance library with example usage in DOE applications
  - Design of general communication performance assessment experiment management system

Acknowledgements

- Academic Collaborators and Colleagues
  - Other affiliated students – Pepper Marts, Carson Woods, Quincy Wofford
  - Steve Shkoller (UC-Davis), HeFFTe team (UTK)
- Lab Collaborators (more details later)
  - LANL – Jon Reisner, Bob Robey, Galen Shipman, Bob Bird
  - LLNL – Olga Pearce, Ignacio Laguna, Ruipeng Li, Kathryn Mohror
  - Sandia – Kevin Pedretti, Matthew Dosanjh, James Elliot, Kurt Ferreira, Ryan Grant, Scott Levy, Carl Pearson, Patrick Widener
- Funding
  - PSAAP-III Award DE-NA0003966
  - DoD SMART Scholarship to Jered Dominguez-Trujillo
  - NSF award OAC-1807583