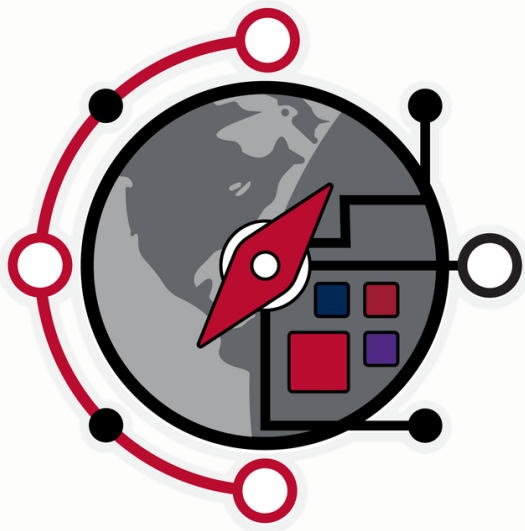




Center for
Optimized Modern
Parallel Adaptive
System Software

COMPASS PSAAP IV Annual Review: Center Overview

PROF. PATRICK BRIDGES · UNIVERSITY OF NEW MEXICO

COMPASS 2026 ANNUAL REVIEW · OCTOBER 29–30, 2026 · ALBUQUERQUE, NEW MEXICO

Schedule for the Review

UNM Student Union · Lobo A/B (Thu) · Fiesta A/B (Fri)

Thursday, October 29 · Morning

9:00–9:10	Introductions · Patrick
9:10–10:00	Center Overview, Roadmap, and Y1 Milestones · Patrick
10:00–10:30	Discussion and break
10:30–11:10	Project Foundations: Training Materials, Initial Sensitivity Studies, and System Monitoring · Derek & Puri
11:10–11:45	System Optimization: Workflow, Experiment/Data Collection, and Performance Tuning · Trilce & Patrick
11:45–12:00	Discussion
12:00–1:00	Lunch

Thursday, October 29 · Afternoon

1:00–1:40	System Software Primitives: Communication Optimization · Tony
1:40–2:00	Integrated Project: Sparse Solver Performance Optimization · Amanda
2:00–2:45	Integrated Project: Hydrodynamic Solver Methods · Steve & Jason
2:45–3:00	Plans for the Upcoming Year and Beyond · Patrick
3:00–3:30	Discussion and break
3:30–5:00	Poster Session: Student Introductions and Poster-side Discussions (17 posters)
6:00–8:30	Dinner at Mission Winery

Friday, October 30 · 8:45 Answers to overnight questions (Patrick) · 9:30 Discussion · 11:30 RT summary of findings · **Remote:** Zoom link in the agenda

SECTION 01

Center Context

Why optimizing modern HPC systems is hard, what COMPASS studies, and how the center is organized.



COMPASS Research Focus

Problem: Modern HPC systems are increasingly difficult to optimize

Focus: Novel system software that enables applications and systems to optimize and tune for post-exascale software and hardware

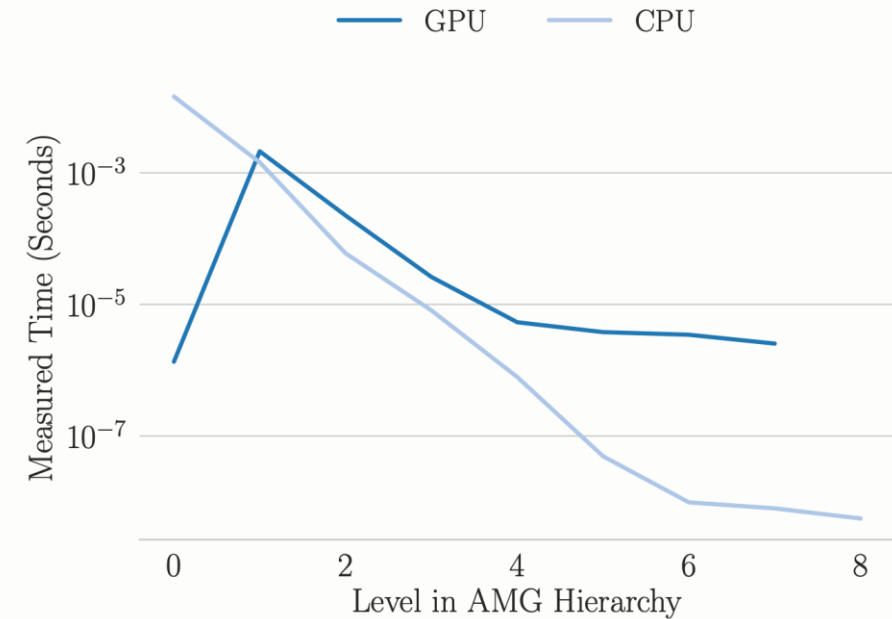
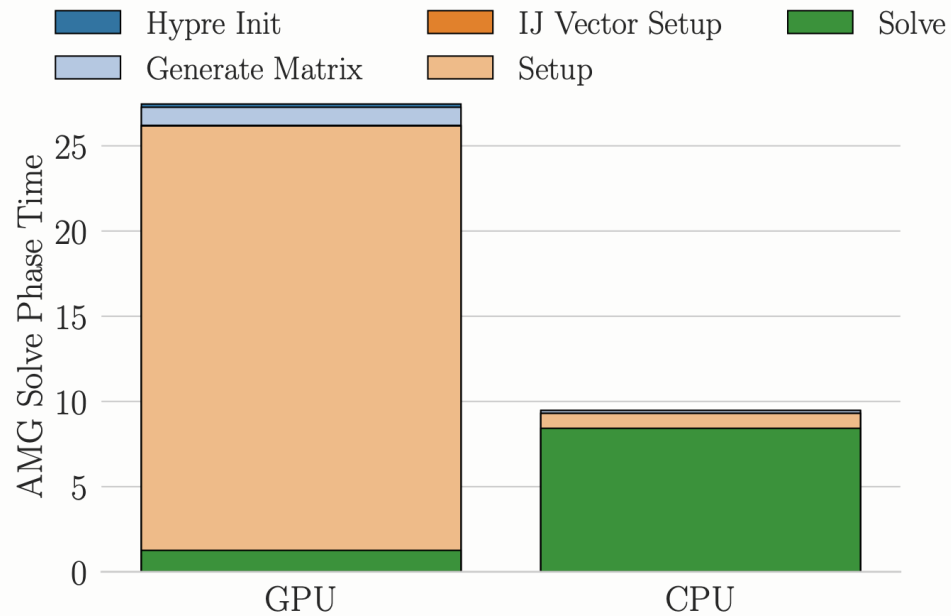
Goal: Create software systems that leverage AI techniques to enable scientific applications to exploit emerging hardware and algorithms

Approach: Integrate innovations in system software, numerical methods, artificial intelligence, and HPC performance modeling to *jointly optimize* every element of system behavior



Optimizing Solvers for Modern Hardware Is Challenging

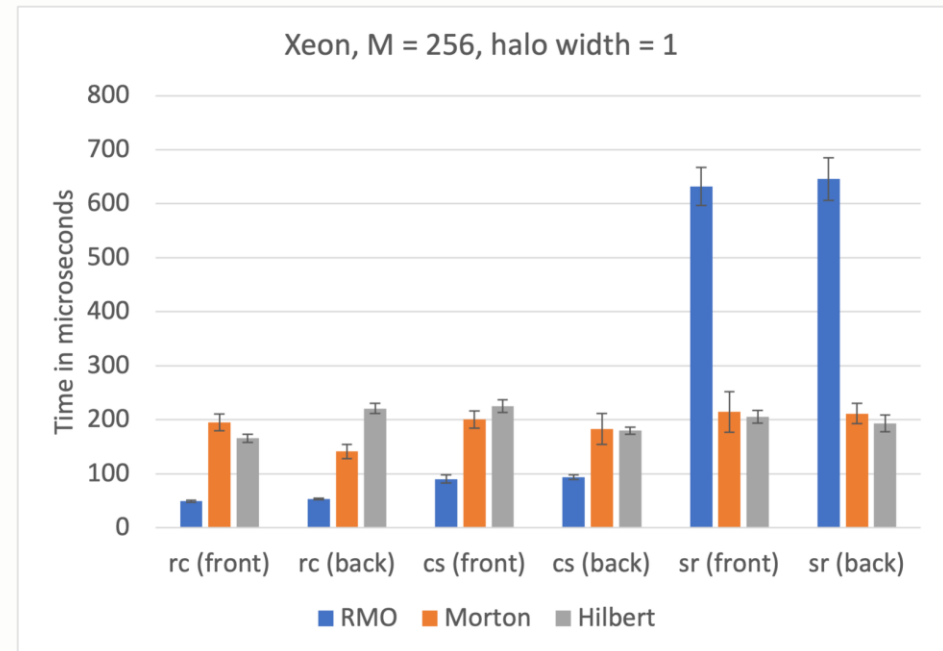
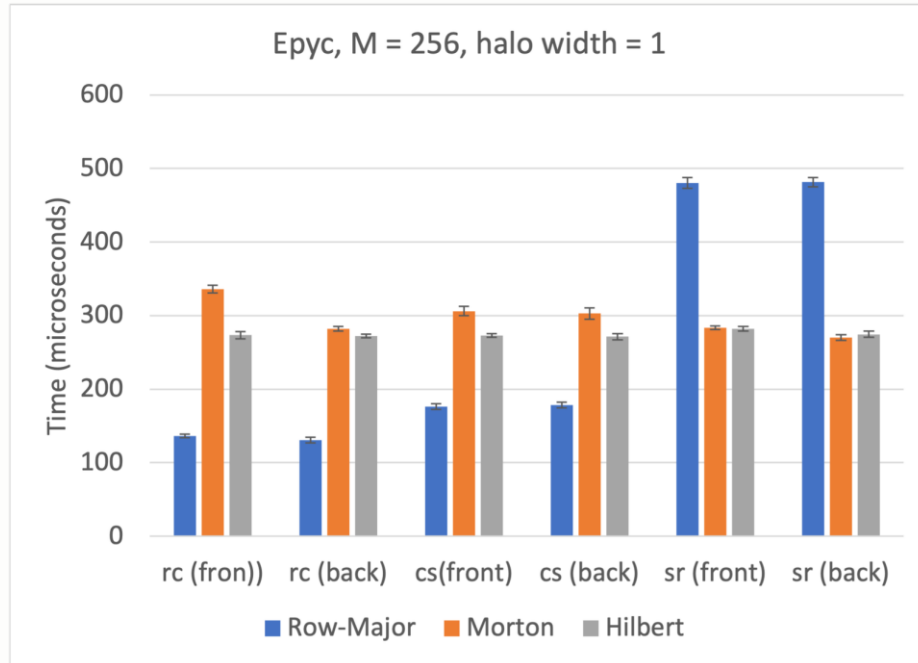
How should AMG use Grace Hopper APUs (288 CPU cores + 4 H100 GPUs) over the course of a solve?



And how should that change for new hardware (e.g., sparse memory acceleration) and new numerical methods?

Optimizing Runtime Data Layouts Is Challenging

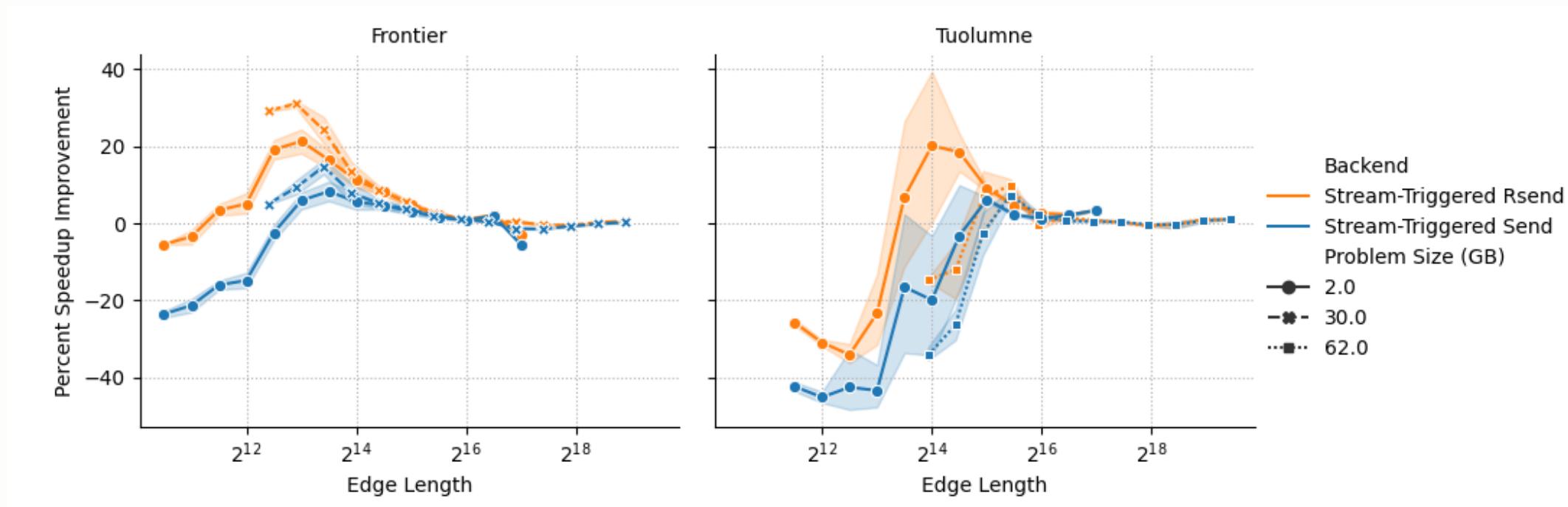
How should a multidimensional data structure be laid out in memory so it can be both computed on and communicated efficiently?



Halo packing performance of row-major vs. space-filling-curve memory layouts under different access patterns

Optimizing Communication for Modern Hardware Is Challenging

When should a solver use stream-triggered communication primitives?



Halo exchange speedup from stream-triggered communication on Frontier and Tuolumne

System and Application Complexity Keeps Growing

Novel systems, runtimes, numerical methods, and problems keep proliferating: more problems to solve, more interactions to chase down, more knobs to tune. **Manual intervention cannot keep up.**

COMPASS optimizes overall application and system performance through four research areas

1

AI-enhanced modeling

Characterize the performance of hardware/software systems

2

Cross-layer optimization

Search for near-optimal system and application configurations

3

Runtime and system software

Enhancements that adapt data layout and data movement

4

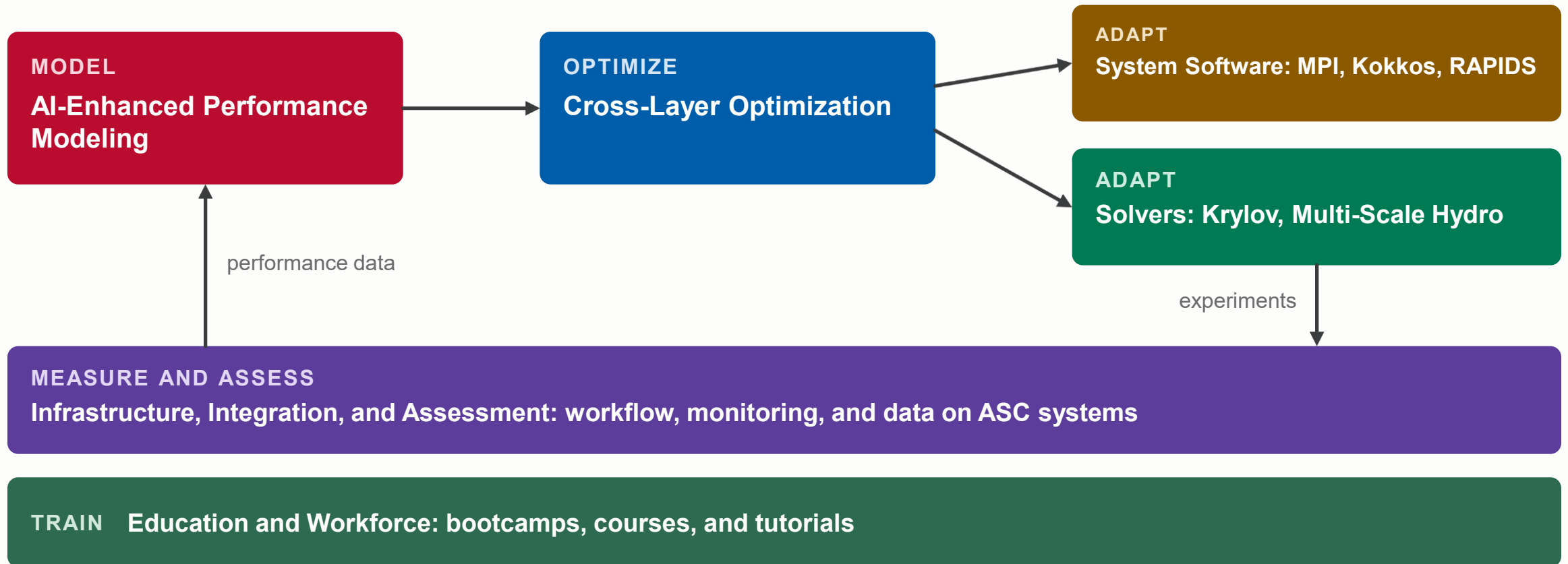
Novel numerical methods

Drive and demonstrate the benefits of the system

Tightly integrated with **integration and assessment**, **education and workforce**, and NNSA lab interactions.

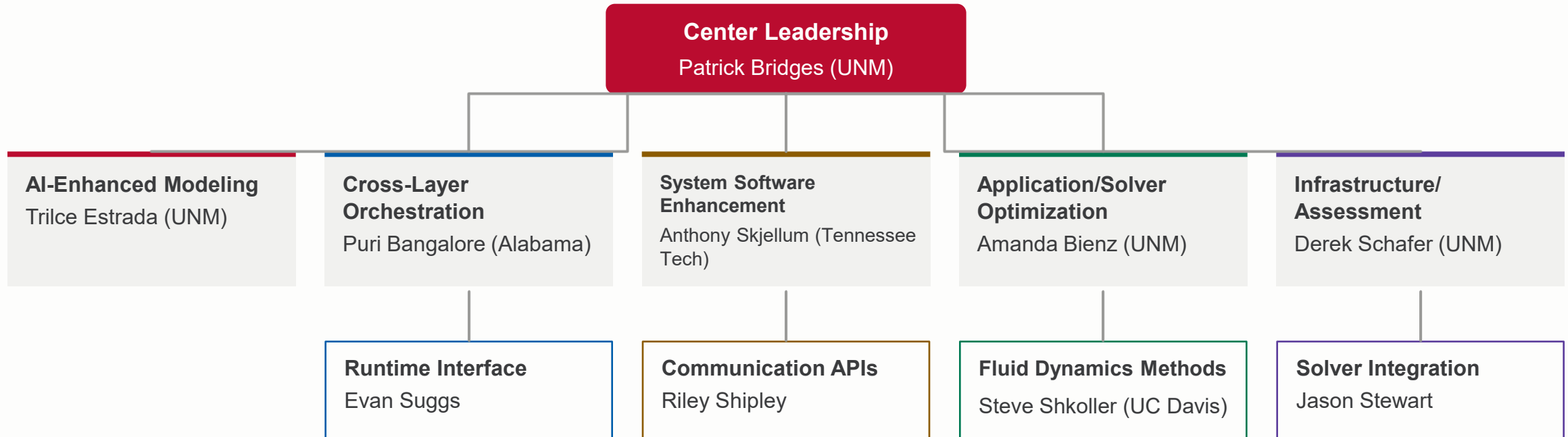
How COMPASS Fits Together

A **measure** → **model** → **optimize** → **adapt** loop, applied to NNSA-relevant solvers. Every talk and poster today is tagged with these areas.



Center Organization

Led by UNM, with Tennessee Tech, Alabama, and UC Davis



Collaboration: weekly and monthly online hackathons, semi-annual in-person hackathons; GitHub, Spack, Benchmark, Slack, Asana

Changes since the proposal and site visit: [leadership or area-lead changes, if any]

Team and Staffing

6

faculty

5

research staff

6

Paid graduate students

2

Paid undergraduates

Presenting posters today

UNM: Akhil Reddy Alasandagutti, Mike Adams, Nicholas Bacon, Xavier Barr, Shannon Kinkead, Nicholas Livingstone, Derek Schafer, Jason Stewart, Jackson Wesley

Tennessee Tech: Amr Akmal Abouelmagd, C. Nicole Avans, Thomas Hines, Javontae Martin, Evan Drake Suggs

Alabama: Purushotham Bangalore, Andrew Worley

Changes since the proposal: Andrew Worley, Abdulaziz Raad left project

Complete org chart with every faculty member, staff member, postdoc, and student (both funded and unfunded): Handout provided

SECTION 02

Roadmap and Year 1 Progress

The five-year roadmap, thrust by thrust: Year 1 milestones, where we stand, and what changed after the mid-year review.



COMPASS Project Roadmap

|| We are here: start of Year 2

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations		Assess Beatnik optimization				Layer-by-layer optimization		Search-based optimization				Integrate model uncertainty				Fully integrate models			
			Scheduler optimization sensitivity				Assess NNSA app optimization opportunities													
Solver Design	Beatnik fast multipole support		Z-Model adaptive meshing		Compare Krylov solver variants				Parallelize topological transitions		Specialize models for physics problems				Integrate models into production codes					
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions						Time-based communication primitive optimization									
	Create initial RAPIDS primitives				Additional RAPIDS extensions						Cross-layer control of system software APIs									
			Kokkos memory layout optimizations				Develop System Tuning APIs													
Integration & / Assessment	Design app/system workflow						Assess Z-Model/FIESTA				Assess Krylov Variants									
	Design/deploy system infrastructure						Assess NNSA application optimization				Assess integrated hydro codes									
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials				Create AI model course materials				Offer courses with integrated materials							
			Deploy bootcamp modules				Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials									

Year 5 goal: AI models, cross-layer optimization, and tunable system software fully integrated and assessed in Krylov and multi-scale hydrodynamics production codes

Roadmap: AI-Enhanced Performance Modeling

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Beatnik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Schedule optimization sensitivity				Assess NNSA app optimization opportunities								Fully integrate models			
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization								Cross-layer control of system software APIs			
	Create initial RAPIDS primitives				Additional RAPIDS extensions															
					Kokkos memory layout optimizations				Develop System Tuning APIs											
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure								Assess NNSA application optimization				Assess integrated hydro codes							
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials								Create AI model course materials				Offer courses with integrated materials			
	Deploy bootcamp modules								Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials							

YEAR 1 MILESTONE

Assess accuracy, precision, uncertainty, cost tradeoffs, and risks of grey-box models for performance forecasting

STATUS

Grey-box scaling models published (HiPC 2025), quantifying accuracy and exposing new limitations; modern ML alternatives under study; UQ work progressing, formally moved to Year 2

YEAR 2 MILESTONE

Non-intrusive data collection and data augmentation; incorporate uncertainty in model inputs and outputs

Scientific question: How can HPC performance models use machine learning to predict performance, capture uncertainty, and effectively inform system optimization decisions?

Roadmap: Cross-Layer Optimization

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation				Enhanced model for optimization integration			
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Beatrik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Scheduler optimization sensitivity				Assess NNSA app optimization opportunities											
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization								Cross-layer control of system software APIs			
	Create initial RAPIDS primitives				Additional RAPIDS extensions															
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure								Assess NNSA application optimization				Assess integrated hydro codes							
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials								Create AI model course materials				Offer courses with integrated materials			
	Deploy bootcamp modules								Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials							

YEAR 1 MILESTONE

Simulation-based study of the costs and benefits of model-driven, optimization-based scheduling on large-scale HPC systems

STATUS

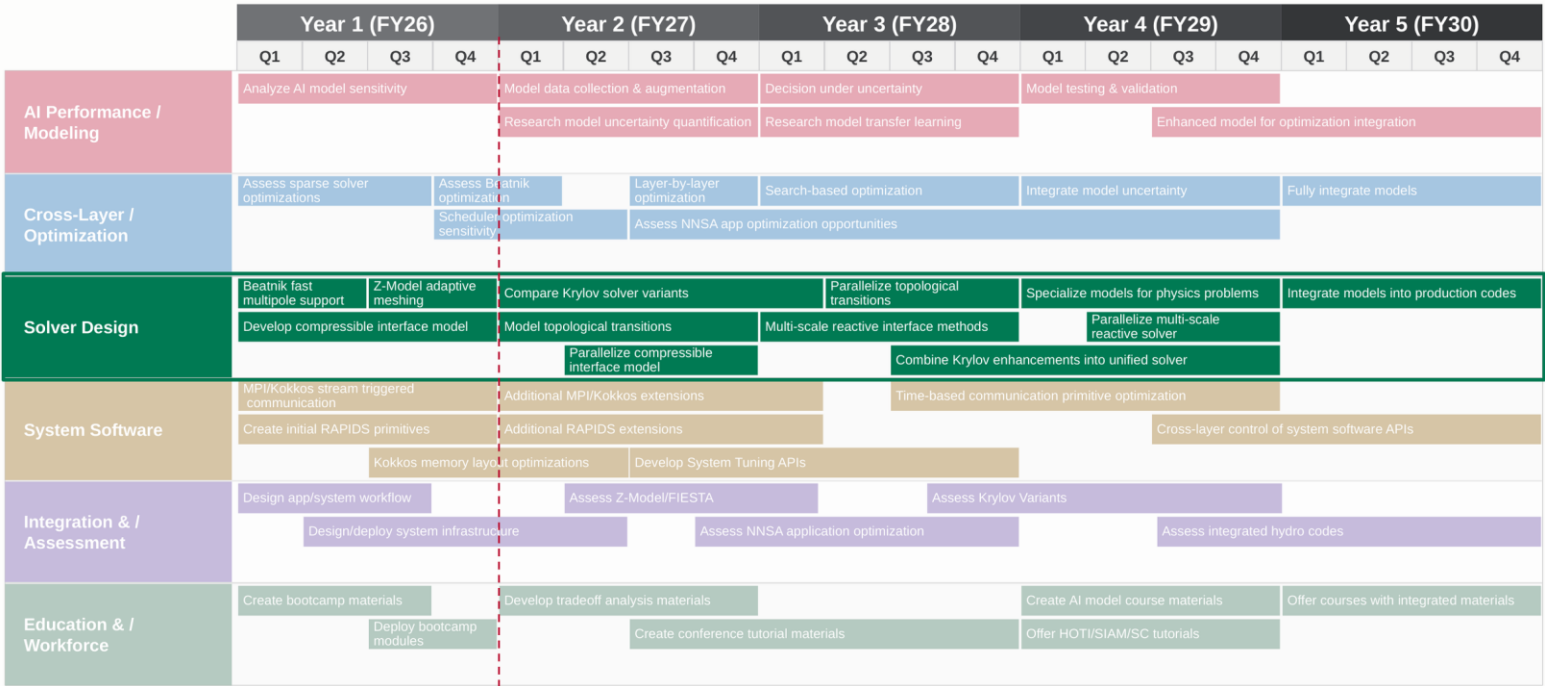
Re-scoped with TST input: optimization opportunities identified in hypre; cross-layer observability framework (scheduler, DCGM, Darshan) published at Cluster 2026; scheduler sensitivity study runs through Y2 Q2

YEAR 2 MILESTONE

Top-down and/or bottom-up layer-by-layer optimization of application and system software hyperparameters

Scientific question: What algorithms, techniques, and abstractions are needed to optimize and adapt cross-layer system configurations and resource allocations?

Roadmap: Application and Solver Design



YEAR 1 MILESTONE

Devise an adaptive mesh refinement (AMR) scheme for the 3D Z-model

STATUS

Delivered: 3D z-model with adaptive meshing, UNM working on parallelization; 3D merger and pinch-off started a year early; scalable GPU/MPI fast multipole solver built for Beatnik

YEAR 2 MILESTONE

Benchmark standard, communication-avoiding, and related Krylov methods on emerging architectures, pinpointing bottlenecks

Scientific question: How should we design solver methods to leverage novel system software and AI performance prediction?

Roadmap: System Software Enhancement

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Bratnik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Schedule optimization sensitivity				Assess NNSA app optimization opportunities								Fully integrate models			
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization											
	Create initial RAPIDS primitives				Additional RAPIDS extensions								Cross-layer control of system software APIs							
					Kokkos memory layout optimizations				Develop System Tuning APIs											
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure								Assess NNSA application optimization				Assess integrated hydro codes							
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials								Create AI model course materials				Offer courses with integrated materials			
	Deploy bootcamp modules								Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials							

YEAR 1 MILESTONE

Prototype, test, and demonstrate low-level RAPIDS API families

STATUS

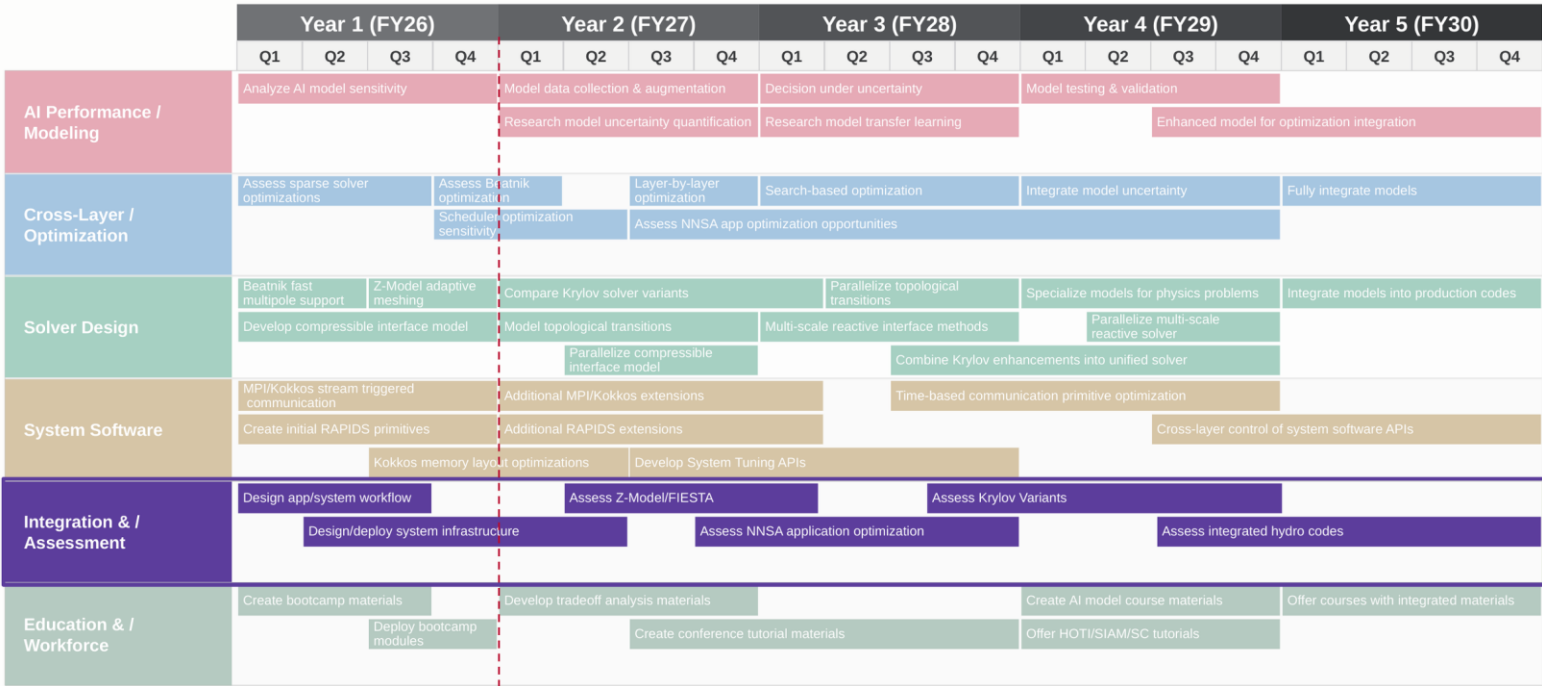
RAPIDS Mobius benchmarked in hypre/AMG2023; MailBox API for irregular communication prototyped (continues into Y2); MPI stream triggering published at Cluster 2026 and under MPI Forum discussion; KokkosComm stream-triggered primitives built

YEAR 2 MILESTONE

Enhance and demonstrate high-level (MPI, Kokkos) and low-level (RAPIDS) API extensions in an NNSA-relevant benchmark or application

Scientific question: What communication and data abstractions have the largest impact on application performance in dynamic, heterogeneous HPC architectures?

Roadmap: Infrastructure, Integration, and Assessment



Enabling technology: integrated system software infrastructure that supports COMPASS research on applications, solvers, and system software, and lets us assess its impact on NNSA-relevant codes

YEAR 1 MILESTONE

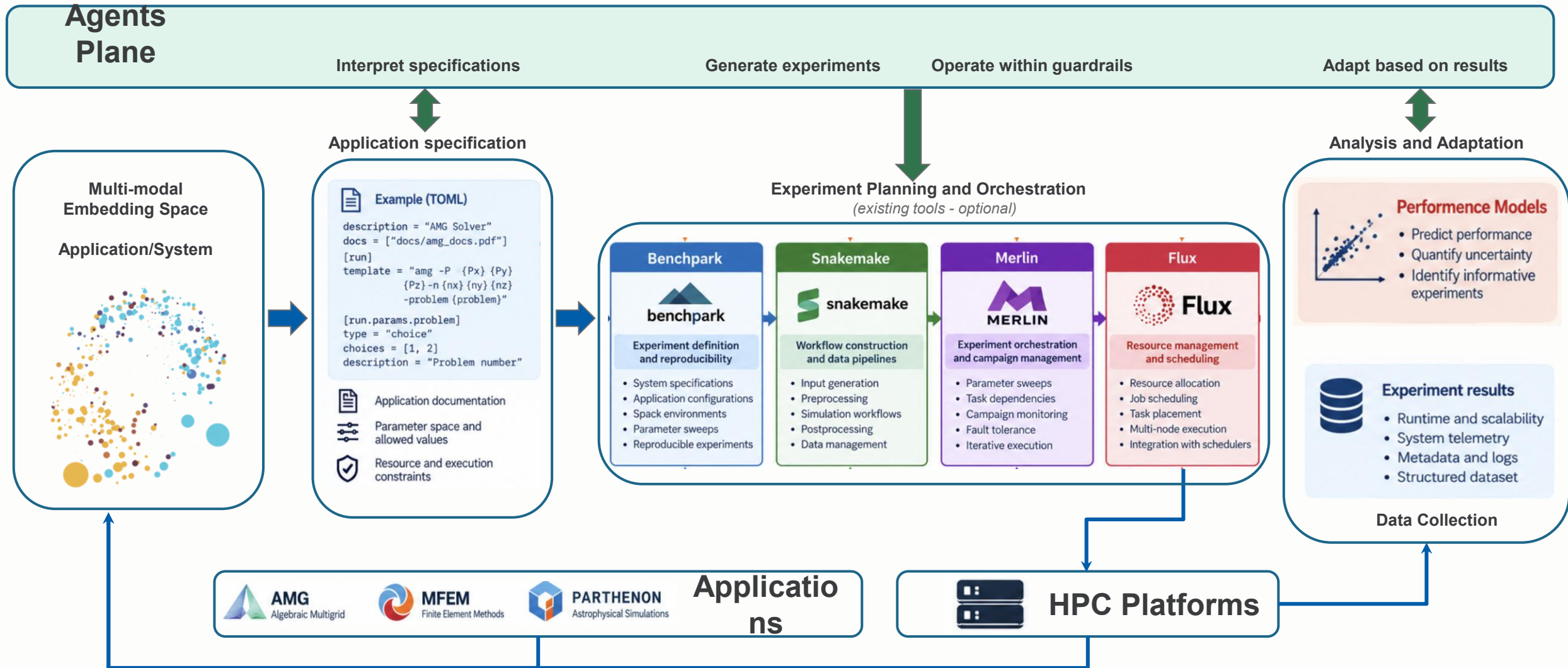
Create core COMPASS software infrastructure by enhancing and integrating existing HPC system software and middleware

STATUS

Prototype infrastructure for workflow launch, data acquisition, and storage (Flux, Merlin, Caliper, Vernier, Benchpark) under evaluation; Vernier paper at SC'25; runtime adaptation deferred to NNSA problems chosen with TST

YEAR 2 MILESTONE

Assess the 3D Z-model against LANL Fury on a rising bubble/plume with background shear and variable atmospheric stratification



Roadmap: Education and Workforce

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Bratnik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Schedule optimization sensitivity				Assess NNSA app optimization opportunities								Fully integrate models			
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization								Cross-layer control of system software APIs			
	Create initial RAPIDS primitives				Additional RAPIDS extensions															
					Kokkos memory layout optimizations				Develop System Tuning APIs											
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure								Assess NNSA application optimization				Assess integrated hydro codes							
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials								Create AI model course materials				Offer courses with integrated materials			
	Deploy bootcamp modules								Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials							

YEAR 1 MILESTONE

Create online training modules for core COMPASS techniques, workflows, and applications; deploy them as bootcamp trainings for new personnel

STATUS

Initial modules created and used to onboard new personnel, improving with feedback; documentation on the public COMPASS Bootcamp web page

YEAR 2 MILESTONE

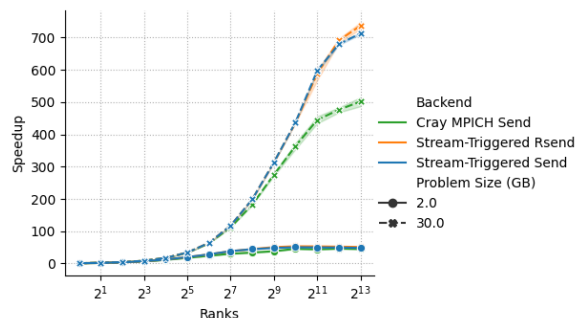
Course materials and tutorials on trade-offs between numerical and system-level decisions, for parallel systems and numerical algorithms courses

Goal: Leverage and broaden research impact through educational programs that train students and the community in full-stack parallel software development

Year 1 Highlights: What We're Most Excited About

SYSTEM SOFTWARE

CPU-free GPU communication

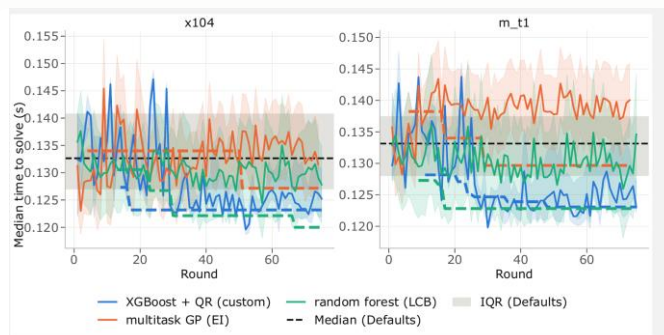


Co-designed MPI GPU abstraction (Cluster 2026) and Kokkos Comm (PASC 2026). **Stream-triggered MPI communication on Tuolumne improves halo exchange strong scaling by 30%+.** Cascade kernel-triggered allreduce beats RCCL and Cray MPICH from 256 B to 32 KiB on 128 GPUs.

Details: 1:00 talk · Schafer, Suggs, Avans, Hines posters

AI MODELING

Grey-box performance prediction

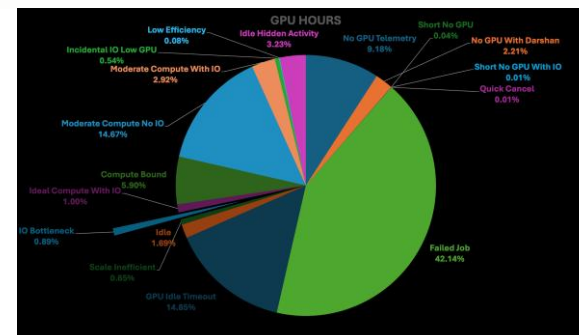


ML models of application scaling (HiPC 2025), now driving MPI autotuning: **quantile-regression Bayesian optimization finds better aCG configurations than a Gaussian process.**

Details: 11:10 talk · Barr poster

CROSS-LAYER OPTIMIZATION

Cross-layer characterization



Scheduler, GPU, and I/O telemetry in one job-level view (Cluster 2026): **262,634 Polaris jobs and 16.53M GPU-hours** classified to find what cross-layer tuning can fix.

Details: 10:30 talk · Bangalore poster

Year 1 Accomplishments: Publications

11 publications in Year 1: 6 conference and journal papers, 4 workshop papers, 1 SC'25 poster

Conferences and journals

Co-Design and Evaluation of a CPU-Free MPI GPU Communication Abstraction and Implementation · IEEE Cluster 2026

Cross-Layer Characterization Framework for HPC System Utilization · IEEE Cluster 2026

Kokkos Comm: Performance Portable Communication for Distributed Kokkos Applications · PASC 2026

Grey-Box Machine Learning Prediction of Parallel Application Scaling · HiPC 2025

Concepts for Designing Modern C++ Interfaces for MPI · LNCS, 2025

The Case of the Mysterious Citations · Computing in Science & Engineering, 2026

Workshops and posters

Scalable, High-Fidelity Monitoring of Application Communication Patterns in Vernier · SC'25 Workshops

Scaling All-to-All Operations Across Emerging Many-Core Supercomputers · SC'25 Workshops

Assessing the Impact of MPI Process Placement on Communication in Many-Core Systems · IPDPS Workshops 2026

Optimizing Allreduce Operations for Modern Heterogeneous Architectures with Multiple Processes per GPU · IPDPS Workshops 2026

Optimizing the GPU All-Reduce Using Multiple Processes per GPU · SC'25 research poster

Full list: compass.unm.edu/research · Code: github.com/COMPASS-PSAAP · MPI stream-triggering text under MPI Forum discussion

Year 1 Accomplishments: Software and Methods

Irregular-communication benchmark

Reproduces and generates models of irregular communication in HPC applications

Cross-layer observability framework

Scheduler, DCGM, and Darshan data in one job-level feature matrix with a workload taxonomy

KokkosComm stream triggering

New stream-triggered Kokkos communication primitives; paper testing under way

GPU-aware collectives

NUMA-aware and high-radix allreduce; aggregated neighbor collectives used in CG solves

Space-filling-curve layouts

CPU Hilbert, Gilbert, and Morton implementations and tests, built with an undergraduate

3D adaptive mesh refinement

For closed triangular meshes, to be integrated into Beatnik and production codes

GPU/MPI fast multipole solver

Scalable far-field solver for the expanded Beatnik interface benchmark

Bootcamp materials

Technical orientation published and in informal assessment with new personnel

Changes in Research Plan and Personnel

Changes made in response to TST discussion in March; **none affect long-term project goals**

Cross-layer study re-scoped and shifted

Opportunity assessment now focuses on scheduler optimization sensitivity, completing Y2 Q2; layer-by-layer optimization starts mid-Year 2. Uses public ALCF Polaris telemetry until Tuolumne data is available.

Sensitivity studies narrowed

Initial studies target scheduler and communication parameters; algorithmic choice and parameter selection will be studied later as needed.

Model UQ moved to Year 2

Based on the initial examination of model capabilities in the grey-box modeling work.

Krylov focus on preconditioned methods

Year 2 studies target existing preconditioned Krylov methods (communication-avoiding, pipelined), since Enlarged Krylov lacks a preconditioning method.

Personnel: [students and staff who joined or left; any changes in area leads]

Center Progression: Year 1

How has the roadmap changed?

First annual review: the roadmap is the proposal's, adjusted after the March TST meeting.

- Cross-layer study re-scoped; completes Y2 Q2
- Model UQ moved to Year 2
- Krylov work focused on preconditioned methods

Year 5 goals unchanged

What compute did we use?

163K

node-hours on LLNL systems in FY26: **134,262** on Dane (CPU) and **28,752** on Tuolumne (GPU)

First PSAAP IV year, so this is the baseline for future comparisons

Who is running on it?

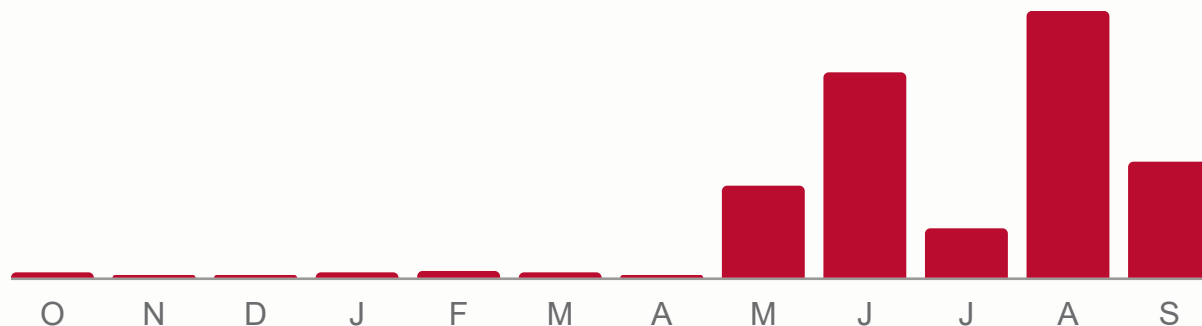
17

students and staff ran jobs on Dane or Tuolumne (19 people including faculty)

[__ graduate students · __ postdocs · __ staff]

ASC Computing Resources: FY26 Usage

Dane (CPU) · 134,262 node-hours · peak Aug: 47,377



Tuolumne (GPU) · 28,752 node-hours · peak Apr: 11,374



Job mix

- **Performance-data collection and parameter sweeps** for model training and tuning: most Dane use (Bacon, Alasandagutti) and the largest Tuolumne user (Barr)
- **Communication-primitive development and benchmarking** on GPUs (Schafer, Adams, Hines, Suggs)
- **Application scaling studies**: Beatnik (Stewart)
- Small- to mid-scale experiments and code development; [no full-scale production runs]

Issues

- Usage ramped up from May as workflow infrastructure came online
- Job-level telemetry not available on Tuolumne; cross-layer work uses the public ALCF Polaris corpus
- [Other issues: queue waits, accounts, allocations]

Lab Internships and Collaborations

Student Lab internships

Student	Lab	When	Status
[Name]	[LANL]	[Summer 2026]	[Completed]
[Name]	[LLNL]	[Summer 2026]	[Completed]
[Name]	[SNL]	[Summer 2027]	[Planned]
[Name]	[TBD]	[TBD]	[Needs plan]

[Students without a firm internship plan, to discuss with the AST]

Collaborations and interactions

- Year 1 papers co-authored with Sandia and LANL staff on process placement, all-to-all, Vernier monitoring, and CPU-free GPU communication [verify list]
- MPI stream-triggering text under discussion at the MPI Forum
- Kokkos Comm developed with Sandia and international Kokkos developers (PASC 2026)
- March 2026 TST meeting shaped the Year 1 roadmap changes
- Lab codes and systems: hypre (LLNL), Tuolumne and Dane; Fury (LANL) comparison planned for Year 2
- [Other visits, talks, and joint projects]

SECTION 03

Today's Program

The rest of the day's talks and posters, and where each one sits in the roadmap.



How Today's Talks Map onto the Roadmap

Talk	AI Modeling	Cross-Layer	Solvers	System Software	Integration	Education
10:30 Project Foundations · Derek & Puri	•	•			•	•
11:10 System Optimization · Trilce & Patrick	•	•			•	
1:00 Communication Optimization · Tony				•		
1:40 Sparse Solver Optimization · Amanda		•	•	•		
2:00 Hydrodynamic Solver Methods · Steve & Jason		•	•		•	
2:45 Plans for the Upcoming Year · Patrick	•	•	•	•	•	•
3:30 Posters (number by area)	—	5	2	8	1	—

Two foundation talks, then three integrated talks that each cut across several areas. Posters: 16 by area (Integration column = dynamic resource allocation), plus 1 cross-cutting.

10:30 · Project Foundations

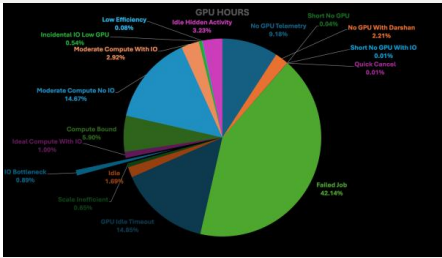
	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations		Assess Beatnik optimization		Layer-by-layer optimization		Search-based optimization		Integrate model uncertainty				Fully integrate models							
	Scheduler optimization sensitivity				Assess NNSA app optimization opportunities															
Solver Design	Beatnik fast multipole support		Z-Model adaptive meshing		Compare Krylov solver variants				Parallelize topological transitions		Specialize models for physics problems				Integrate models into production codes					
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization											
	Create initial RAPIDS primitives				Additional RAPIDS extensions								Cross-layer control of system software APIs							
	Kokkos memory layout optimizations				Develop System Tuning APIs															
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure				Assess NNSA application optimization				Assess integrated hydro codes											
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials				Create AI model course materials				Offer courses with integrated materials							
			Deploy bootcamp modules		Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials											

Related posters (3:30): Cross-Layer Characterization Framework for HPC Applications (Bangalore) · Guided Analysis for Understanding Global Communication Efficiency (Bacon) · Do AMD and NVIDIA Count the Same Thing? (Abouelmagd)

Derek & Puri · 10:30–11:10

Training Materials, Initial Sensitivity Studies, and System Monitoring

- Bootcamp and training materials
- Network and latency sensitivity on Tuolumne and Dane
- Cross-layer telemetry: scheduler, GPU (DCGM), I/O (Darshan)
- Applications chosen with TST: hypre, Trilinos, Parthenon



262,634 Polaris jobs, 16.53M GPU-hours
characterized

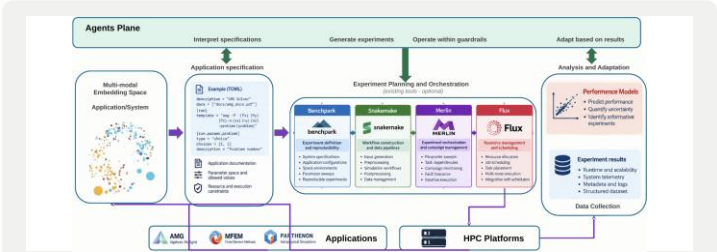
11:10 · System Optimization

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Beatnik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Schedule optimization sensitivity				Assess NNSA app optimization opportunities								Fully integrate models			
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver				Integrate models into production codes			
					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization								Cross-layer control of system software APIs			
	Create initial RAPIDS primitives				Additional RAPIDS extensions															
					Kokkos memory layout optimizations				Develop System Tuning APIs											
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure				Assess NNSA application optimization				Assess integrated hydro codes											
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials				Create AI model course materials				Offer courses with integrated materials							
	Deploy bootcamp modules				Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials											

Trilce & Patrick · 11:10–11:45

Workflow, Experiment/Data Collection, and Performance Tuning

- Center software infrastructure roadmap
- SparseAtlas and few-shot slowdown prediction
- AgHPC: safe agentic experiments
- MPI autotuning under uncertainty; when to deploy an optimization
- AI practice, and an integrated agentic adaptation framework



One infrastructure where agents plan experiments, models predict and adapt, and every run adds data

Related posters (3:30): Optimizing MPI Parameters under Uncertainty (Barr) · AgHPC: A Safe Agentic Framework for HPC Experiments (Livingstone) · Workflows Using Artificial Intelligence (Stewart & Hines) · Multi-Strategy Adaptive Optimization of HPC Communication (Kinkead)

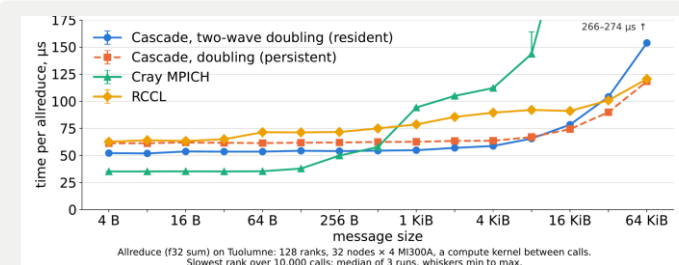
1:00 · System Software Primitives

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Beatnik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Schedule optimization sensitivity				Assess NNSA app optimization opportunities								Fully integrate models			
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver				Integrate models into production codes			
					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization								Cross-layer control of system software APIs			
	Create initial RAPIDS primitives				Additional RAPIDS extensions															
					Kokkos memory layout optimizations				Develop System Tuning APIs											
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure				Assess NNSA application optimization				Assess integrated hydro codes											
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials				Create AI model course materials				Offer courses with integrated materials							
	Deploy bootcamp modules				Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials											

Tony · 1:00–1:40

Communication Optimization

- GPU-driven: stream triggering (MPI Advance, Cabana, Kokkos Comm)
- Kernel-initiated transfers: OpenRAPIDS Cascade
- CPU-driven: ExaMPI strong progress; Earlybird collectives
- Locality discovery for tuning APIs



Cascade allreduce beats RCCL and Cray MPICH, 256 B to 32 KiB

Related posters (3:30): Schafer, Suggs, Avans (stream triggering) · Hines (Cascade) · Martin (ExaMPI RMA) · Wesley (Earlybird collectives) · Kinkead (locality discovery) · also Alasandagutti (CXL in an AMT runtime) and Abouelmagd (MI300A partitioning)

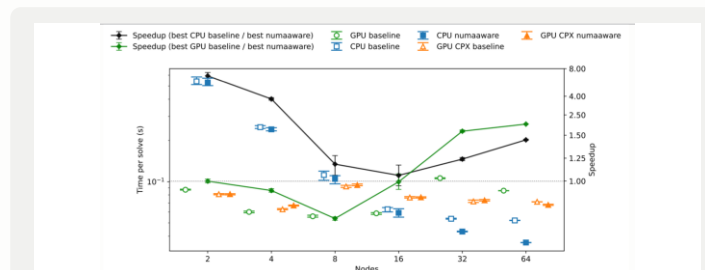
1:40 · Integrated Project: Sparse Solvers

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Beatnik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Schedule optimization sensitivity				Assess NNSA app optimization opportunities											
Solver Design	Beatnik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver				Integrate models into production codes			
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization											
	Create initial RAPIDS primitives				Additional RAPIDS extensions								Cross-layer control of system software APIs							
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure				Assess NNSA application optimization								Assess integrated hydro codes							
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials				Create AI model course materials				Offer courses with integrated materials							
	Deploy bootcamp modules				Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials											

Amanda · 1:40–2:00

Sparse Solver Performance Optimization

- ABCG research CG solver: CPU, GPU-aware, stream triggering
- Pipelined PCG: 1.87 s → 1.31 s on 64 processes
- Locality-aware collectives in hypre CG
- Next: hybrid CPU/GPU AMG; preconditioned Enlarged Krylov



NUMA-aware collectives speed up hypre CG at scale (Adams)

Related posters (3:30): Improving the Scalability of GPU-Aware Iterative Solvers Using Locality-Aware Collectives (Adams) · see also Kinkead and Wesley, covered at 1:00

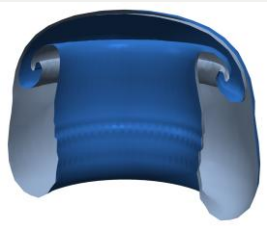
2:00 · Integrated Project: Hydrodynamics

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations			Assess Beatnik optimization		Layer-by-layer optimization		Search-based optimization				Integrate model uncertainty				Fully integrate models				
				Scheduler optimization sensitivity		Assess NNSA app optimization opportunities														
Solver Design	Beatnik fast multipole support		Z-Model adaptive meshing		Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems				Integrate models into production codes			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver							
System Software					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization								Cross-layer control of system software APIs			
	Create initial RAPIDS primitives				Additional RAPIDS extensions															
Integration & / Assessment					Kokkos memory layout optimizations				Develop System Tuning APIs											
	Design app/system workflow			Assess Z-Model/FIESTA				Assess Krylov Variants												
	Design/deploy system infrastructure								Assess NNSA application optimization				Assess integrated hydro codes							
Education & / Workforce	Create bootcamp materials			Develop tradeoff analysis materials								Create AI model course materials				Offer courses with integrated materials				
				Deploy bootcamp modules		Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials										

Steve & Jason · 2:00–2:45

Adaptive Interface Models for Massively Parallel Hydrodynamics

- **Delivered:** 3D z-model with adaptive meshing
- **Started early:** 3D merger and pinch-off
- **New:** variable-density plume; 1D shock model
- **Under way:** parallel C++ port with fast multipole solver



3,900× fewer unknowns
than a grid; full 3D roll-up in
1.4 h on one laptop core

Related posters (3:30): Engineering the Distributed Rising Bubble Problem in Beatnik (Stewart) · [Other hydrodynamics posters, if any]

Poster Session · 3:30–5:00

17 posters, each opening with a one-minute introduction, grouped by schematic area

Cross-Layer Optimization (5)

Cross-Layer Characterization Framework for HPC Applications · Bangalore

Guided Analysis for Understanding Global Communication Efficiency · Bacon

Optimizing MPI Parameters under Uncertainty · Barr

Multi-Strategy Adaptive Optimization of HPC Communication · Kinkead

Do AMD and NVIDIA Count the Same Thing? · Abouelmagd

Dynamic Resource Allocation (1) · Other (1)

AgHPC: A Safe Agentic Framework for HPC Experiments · Livingstone

Workflows Using Artificial Intelligence · Stewart & Hines

System Software (8)

Optimizations for a CPU-Free Stream-Triggered Communication Interface · Schafer

Streamlining Cabana + MPI Using CPU-Free Communication · Suggs

Accelerated Data Simulations Using Stream-Triggered MPI + Kokkos Comm · Avans

Implementing Software Strong Progress RMA in ExaMPI · Martin

OpenRAPIDS: Cascade · Hines

System Software (cont.)

Optimizing Persistent Collectives Using RMA and Earlybird Communication · Wesley

Advancing the State of the ARTS: CXL Memory in an AMT Runtime · Alasandagutti

Partitioning the AMD MI300A: Locality Rules, Power Throttles · Abouelmagd

Solver Design (2)

Engineering the Distributed Rising Bubble Problem in Beatnik · Stewart

Improving the Scalability of GPU-Aware Iterative Solvers Using Locality-Aware Collectives · Adams

Looking Ahead: Year 2 (FY27) Milestones

	Year 1 (FY26)				Year 2 (FY27)				Year 3 (FY28)				Year 4 (FY29)				Year 5 (FY30)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
AI Performance / Modeling	Analyze AI model sensitivity				Model data collection & augmentation				Decision under uncertainty				Model testing & validation							
					Research model uncertainty quantification				Research model transfer learning				Enhanced model for optimization integration							
Cross-Layer / Optimization	Assess sparse solver optimizations				Assess Beatrik optimization				Layer-by-layer optimization				Search-based optimization				Integrate model uncertainty			
					Scheduler optimization sensitivity				Assess NNSA app optimization opportunities											
Solver Design	Beatrik fast multipole support				Z-Model adaptive meshing				Compare Krylov solver variants				Parallelize topological transitions				Specialize models for physics problems			
	Develop compressible interface model				Model topological transitions				Multi-scale reactive interface methods				Parallelize multi-scale reactive solver				Integrate models into production codes			
					Parallelize compressible interface model				Combine Krylov enhancements into unified solver											
System Software	MPI/Kokkos stream triggered communication				Additional MPI/Kokkos extensions				Time-based communication primitive optimization											
	Create initial RAPIDS primitives				Additional RAPIDS extensions								Cross-layer control of system software APIs							
					Kokkos memory layout optimizations				Develop System Tuning APIs											
Integration & / Assessment	Design app/system workflow				Assess Z-Model/FIESTA				Assess Krylov Variants											
	Design/deploy system infrastructure				Assess NNSA application optimization				Assess integrated hydro codes											
Education & / Workforce	Create bootcamp materials				Develop tradeoff analysis materials				Create AI model course materials				Offer courses with integrated materials							
	Deploy bootcamp modules				Create conference tutorial materials				Offer HOTI/SIAM/SC tutorials											

Details at 2:45: Plans for the Upcoming Year and Beyond

AI Modeling: non-intrusive data collection and augmentation; uncertainty in model inputs and outputs

Cross-Layer: top-down and/or bottom-up layer-by-layer hyperparameter optimization

System Software: demonstrate MPI, Kokkos, and RAPIDS API extensions in an NNSA-relevant code

Solvers: benchmark standard vs. communication-avoiding Krylov methods on emerging architectures

Integration: 3D Z-model vs. LANL Fury on a rising bubble/plume with shear and stratification

Education: course materials and tutorials on numerical vs. system-level trade-offs

Discussion

- | Year 1 work is organized into cross-cutting projects that map directly onto the roadmap milestones
- | Today's talks and posters go into depth on each project
- | We are looking for RT input to refine the roadmap and to find new collaboration opportunities with the labs