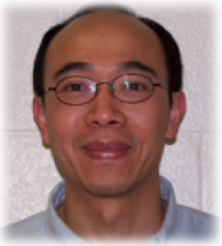
Co-Design of Hardware/Software for Predicting MAV Aerodynamics





E. de Sturler, **W. Feng**, C. Roy, A. Sandu, D. Tafti

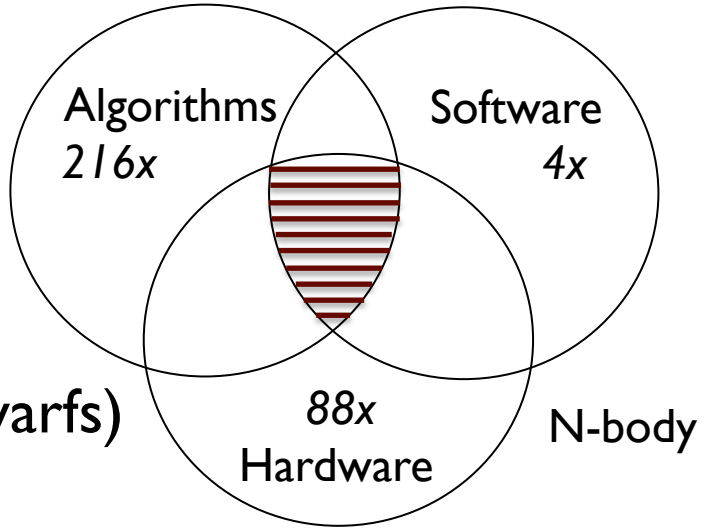




J. Edwards, H. Luo, F. Mueller

Fariba Fahroo, Computational Mathematics

2013 AFOSR Computational Mathematics Review
W. Feng, wfeng@vt.edu, 540.315.1545

Vision

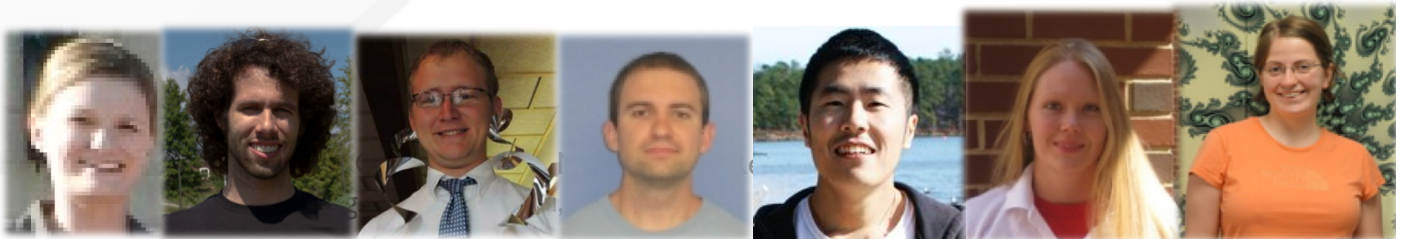


- Synergistic co-design process for the *structured/unstructured grid motifs* (or dwarfs) in computational fluid dynamics (CFD) to support aerodynamic predictions for micro-air vehicles (MAVs).
  - Malleable **algorithms**
    - … that can be mapped and optimized in **software**
    - … onto the right type of processing core in **hardware**
    - … at the right time
  - Co-design feedback to vendors to assist in guiding future hardware design

Team @ VT & NC State

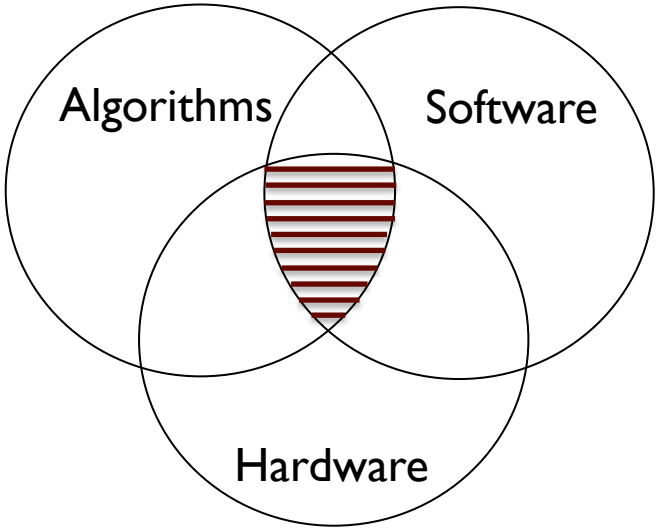
- Virginia Tech (14)
  - Eric de Sturler — Numerical Methods (solvers & preconditioners)
  - Wu Feng — Parallel Computing (performance, power, portability)
  - Chris Roy — CFD (structured grid and ALE mesh movement)
  - Adrian Sandu — Numerical Methods (time stepping & discretization)
  - Danesh Tafti — CFD (pressure-based multiblock structured)
  - Research scientist (1), postdocs (2), and graduate students (6)
- North Carolina State University (7)
  - Jack Edwards — CFD (multiblock structured w/ implicit solvers)
  - Hong Luo — CFD (unstructured grid / compressible)
  - Frank Mueller — Parallel Computing (languages, compilers, scalability)
  - Postdocs (2), graduate students (2)

VirginiaTech
Invent the Future

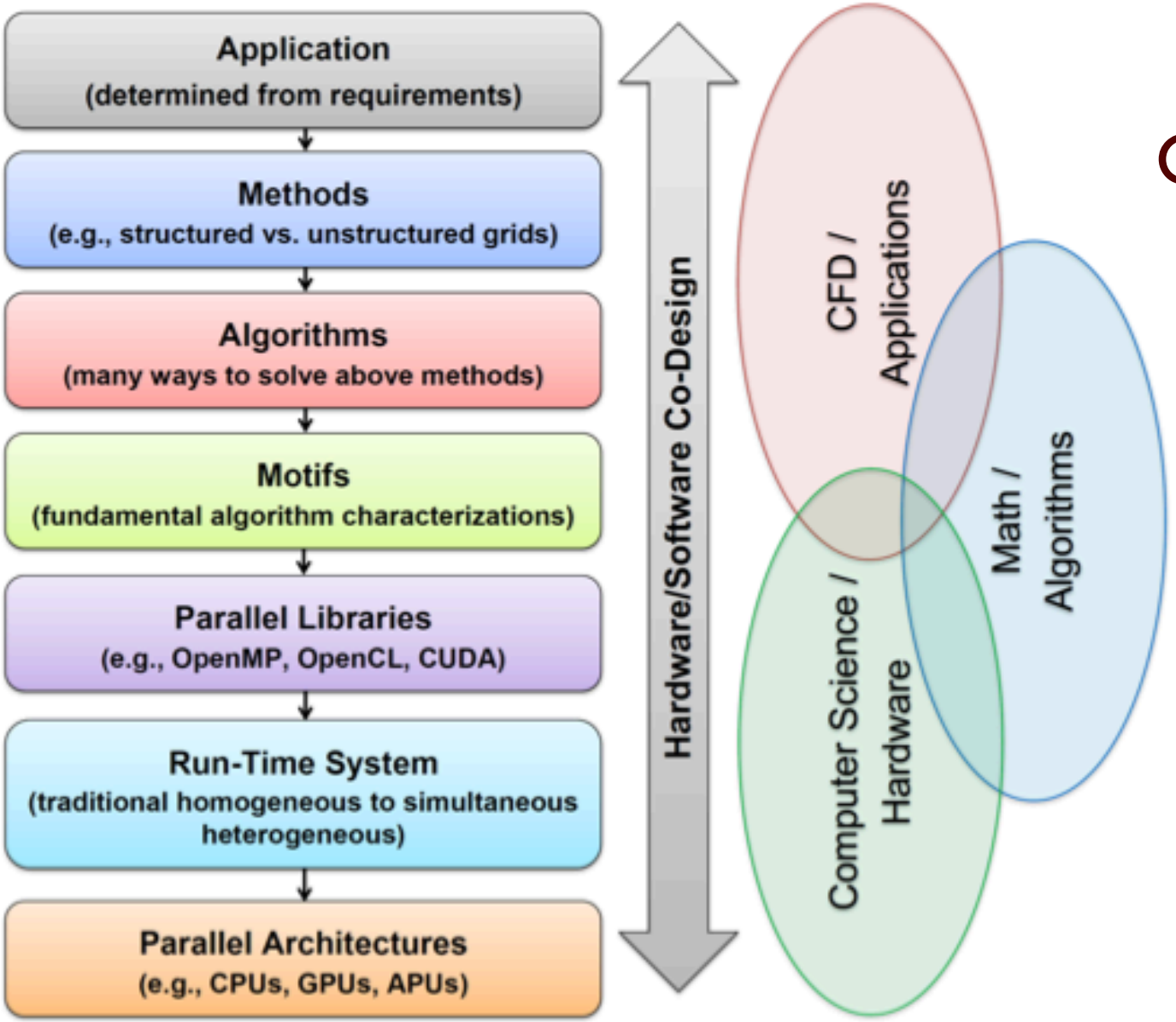


Roadmap

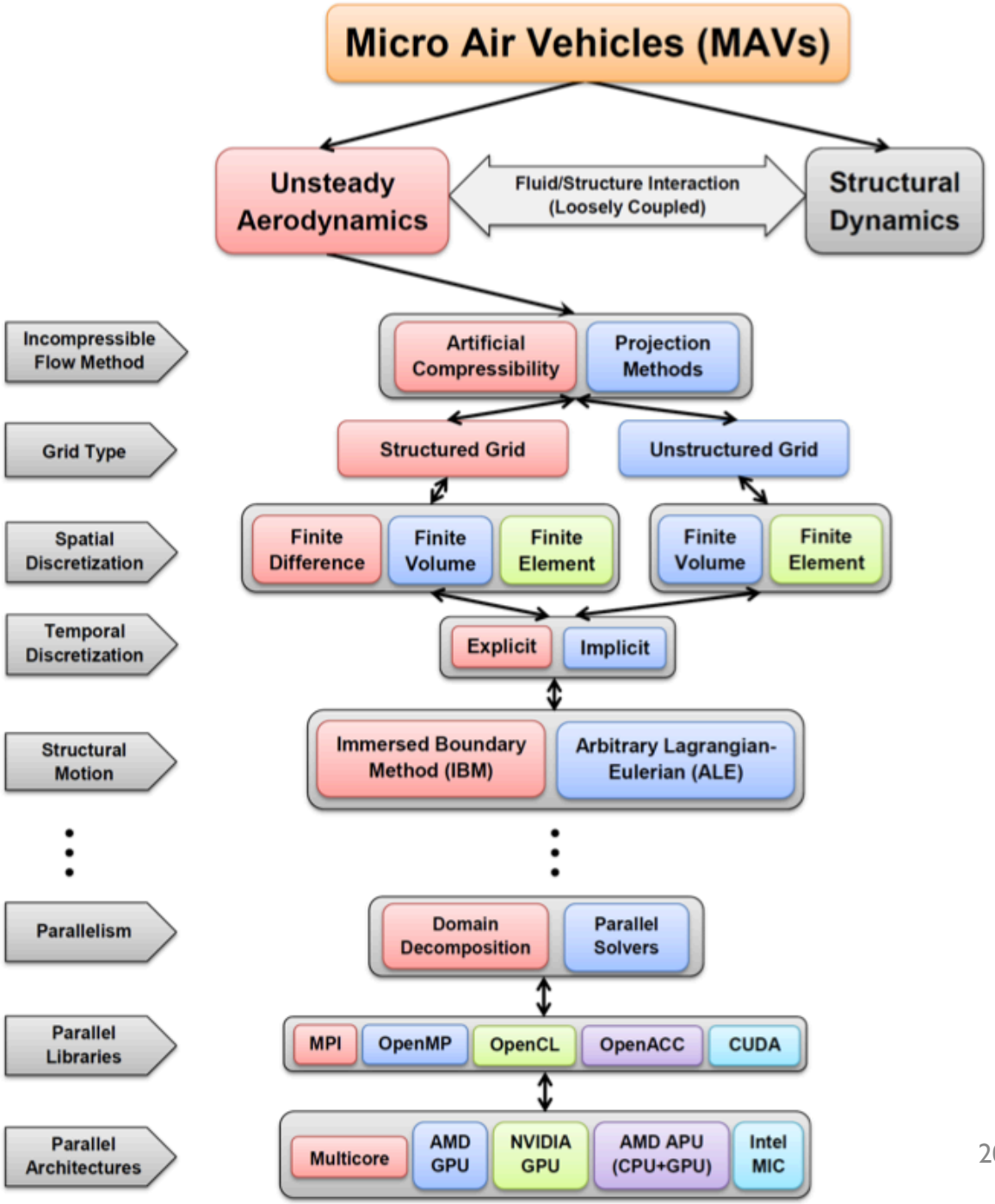
- Vision
- Team
- Approach
- Infrastructure
- Co-Design Research
  - Computer Science
  - CFD Codes (4)
  - Mathematics
- Achievements & Publications
- Next Steps



Layered Co-Design



W. Feng, wfeng@vt.edu, 540.315.1545



Co-Design for Micro Air Vehicles (MAVs)

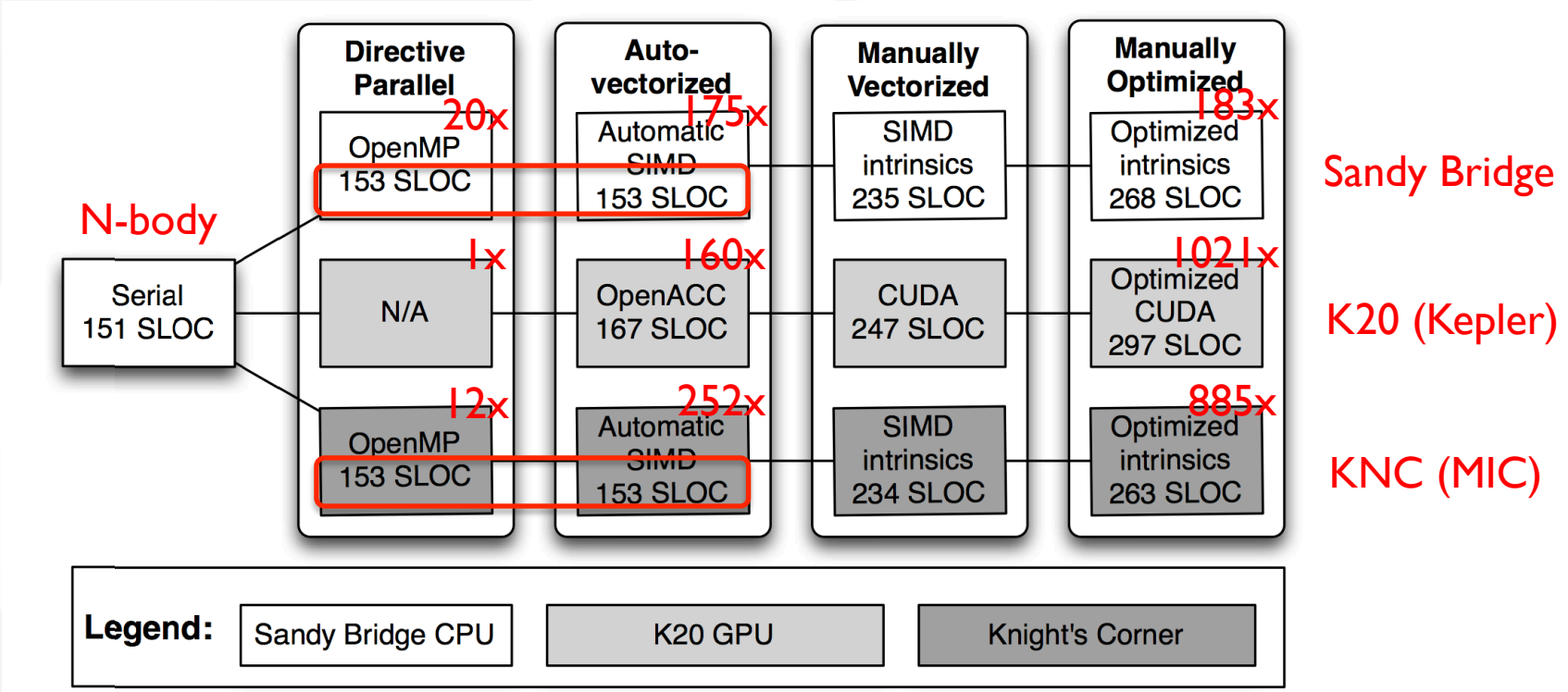
2013 AFOSR Computational Mathematics Review
W. Feng, wfeng@vt.edu, 540.315.1545

Co-Design Around Three P's: Performance, Programmability, Portability

"Productivity = Performance + Programmability + Portability"

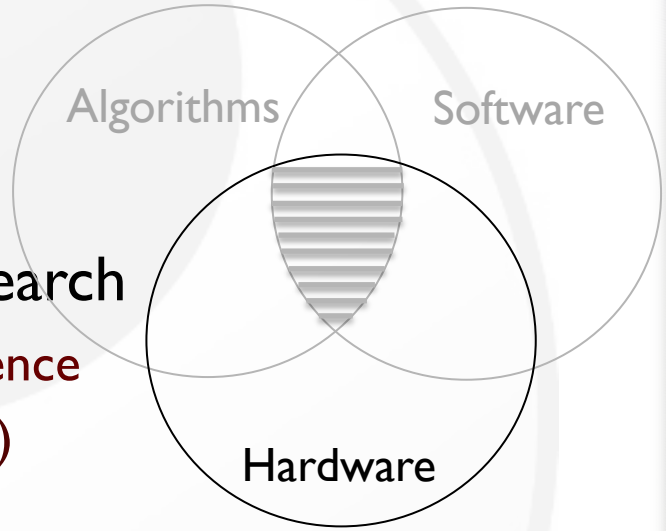
- Multi-dimensional optimization across two or more P's

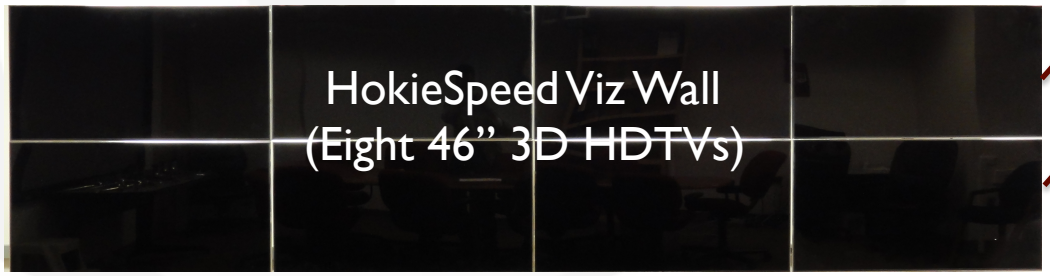
  … first *manual co-design* … then *automated co-design*



Roadmap

- Vision
- Team
- Approach
- Infrastructure
- Co-Design Research
  - Computer Science
  - CFD Codes (4)
  - Mathematics
- Achievements & Publications
- Next Steps





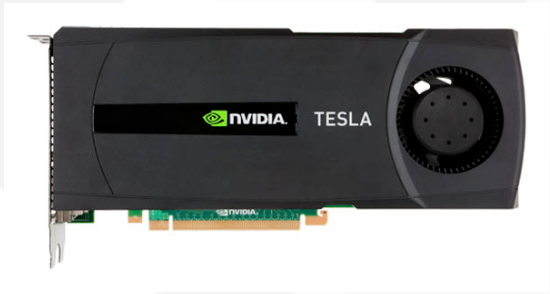
HOKIESPEED

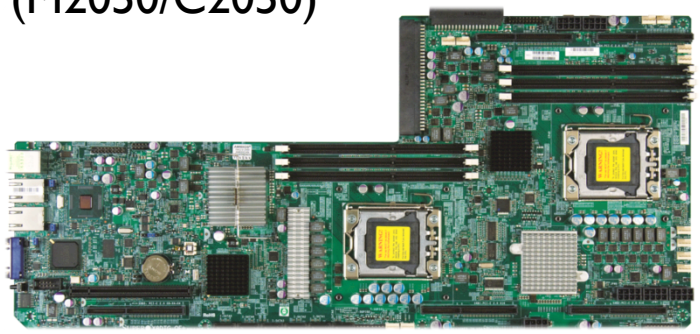
VT

NSF

A GPU-Accelerated Supercomputer for the Masses

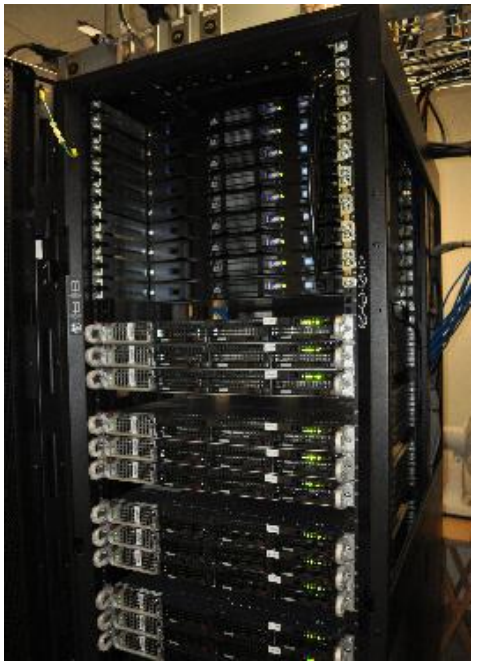
- Purpose
  - To catalyze new approaches for conducting research via the synergistic amalgamation of heterogeneous CPU-GPU hardware and software
- Profile
  - Total Nodes: 209, where each compute node consists of
    - Motherboard: Supermicro 2026GT-TRF Dual Intel Xeon
    - CPUs: Two 2.4-GHz Intel Xeon E5645 6-core (12 CPU cores per node)
    - GPUs: Two NVIDIA Tesla Fermi GPUs (M2050/C2050)





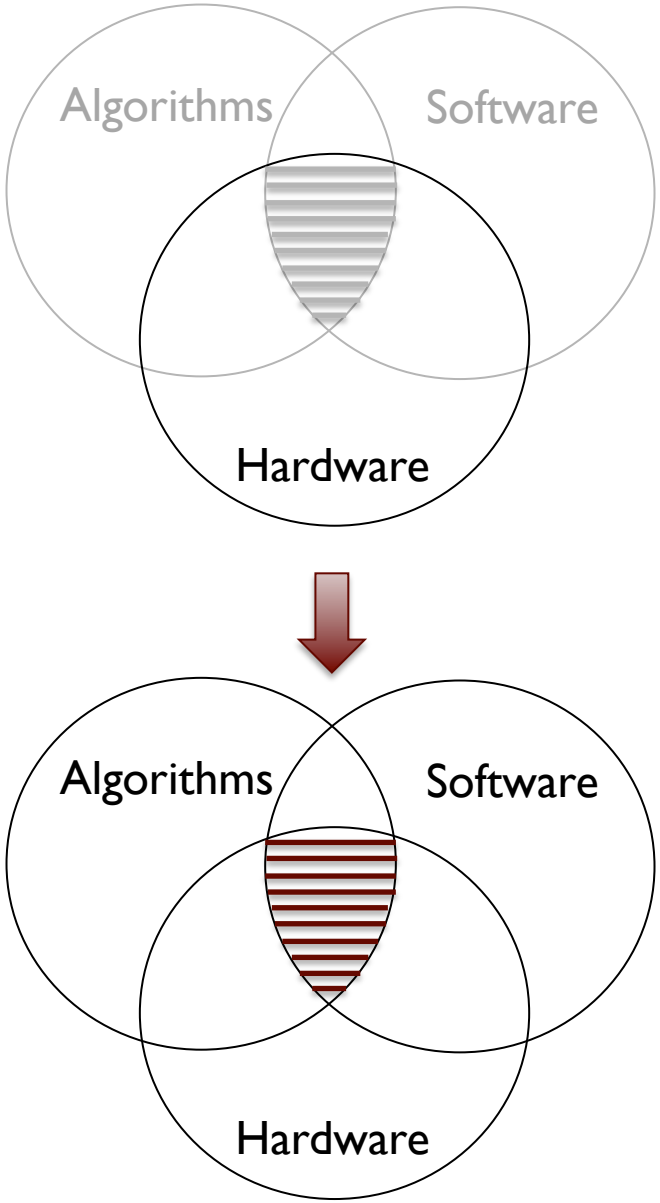
ARC Cluster

- Hardware
  - 2x AMD Opteron 6128 (8 cores each)
    - 108 nodes = 1728 CPU cores
  - NVIDIA GTX480, GTX680 , C2050, K20c: 108 GPUs
  - Mellanox QDR InfiniBand: 40Gbit/s
- Software
  - CUDA 5.0
  - PGI Compilers V13.4 w/ CUDA Fortran & OpenACC support
  - OpenMPI & MVAPICH2-1.9 w/ GPUDirect V2 capability
  - Torque/Maui job management system



Roadmap

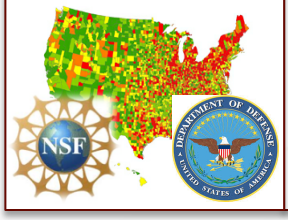
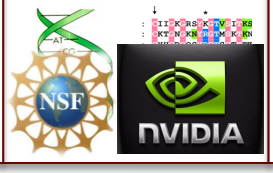
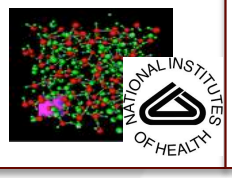
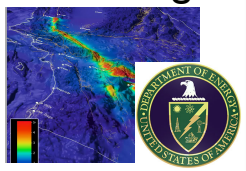
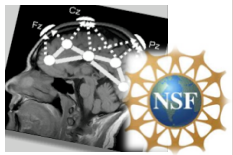
- Vision
- Team
- Approach
- Infrastructure
- Co-Design Research
  - Computer Science (Feng, Mueller)
  - CFD Codes (Edwards, Luo, Roy, Tafti)
  - Mathematics (de Sturler, Sandu)
- Achievements & Publications
- Next Steps



Intra-Node

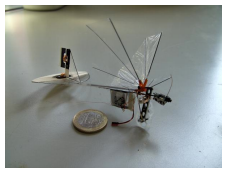
Ecosystem for Heterogeneous Parallel Computing

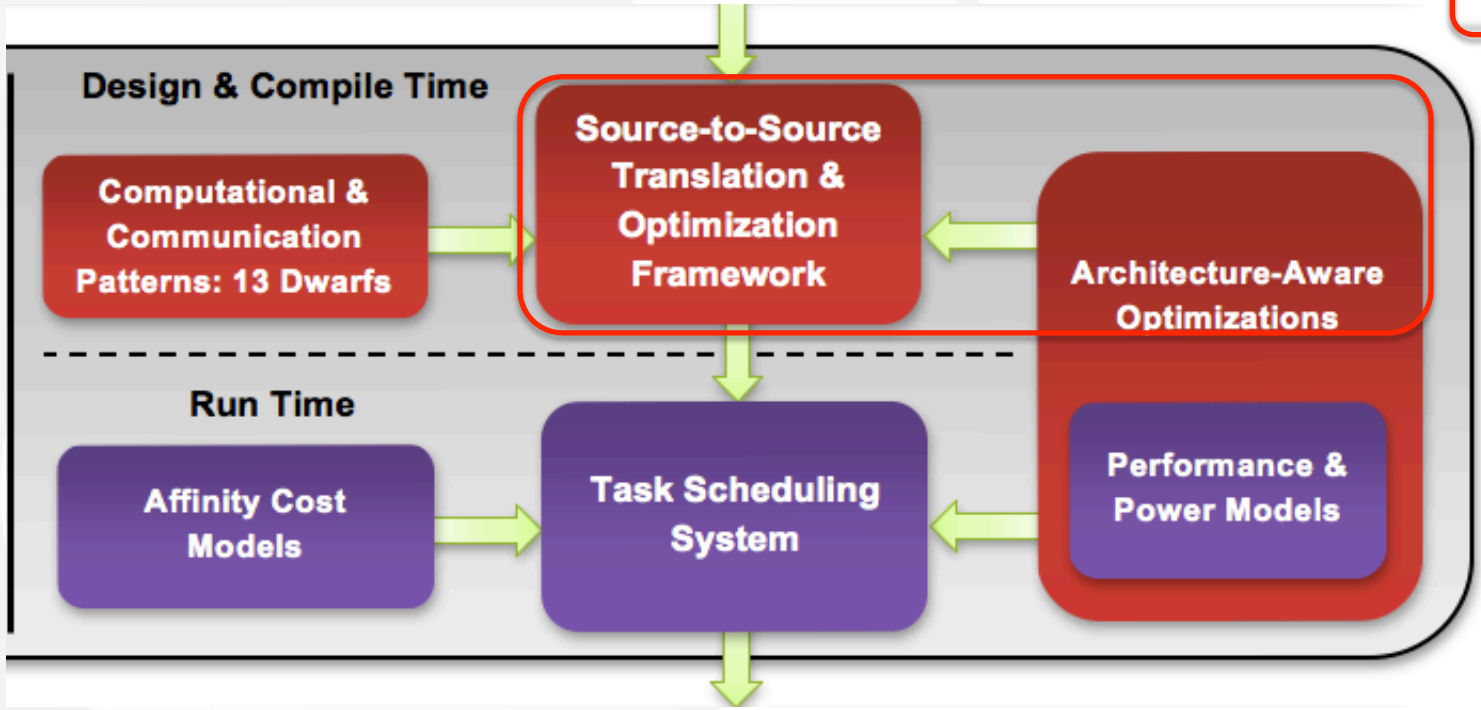
MANUAL CO-DESIGN

Epidemiology	Sequence Alignment	Molecular Dynamics	Earthquake Modeling	Neuro-informatics
				

Cybersecurity

MAVs





Manual Co-Design → Automated Co-Design

Heterogeneous Parallel Computing Platform

W. Feng, wfeng@vt.edu, 540.315.1545

Acceleration Potential of Kernels

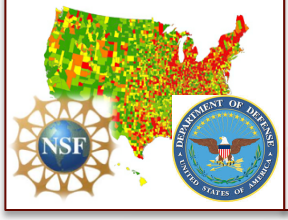
- Parallelism: Multicores and Accelerators (GPUs, Intel MIC)
- Problem: Existing Codes Sequential or Coarse Parallelism
  - Ad-hoc approach parallelization → unknown results
- Vision: Infer speedup potential before refactoring code
  1. Determine variable reuse for given architecture
  2. Estimate speedup for fine-grained parallelization
  3. Assess effects of manual code and data transformations
  4. Suggest (or auto-generate) code and data transformations

Intra-Node

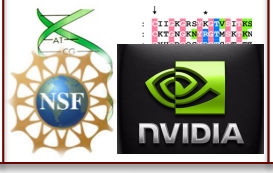
Ecosystem for Heterogeneous Parallel Computing

MANUAL CO-DESIGN

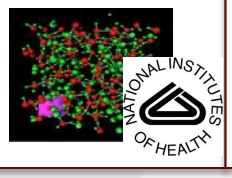
Epidemiology



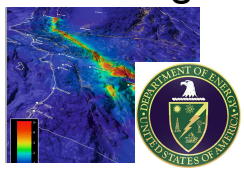
Sequence Alignment



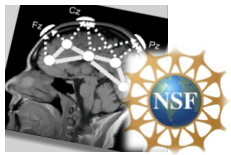
Molecular Dynamics



Earthquake Modeling

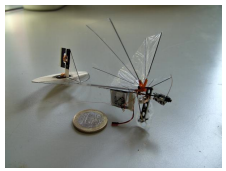


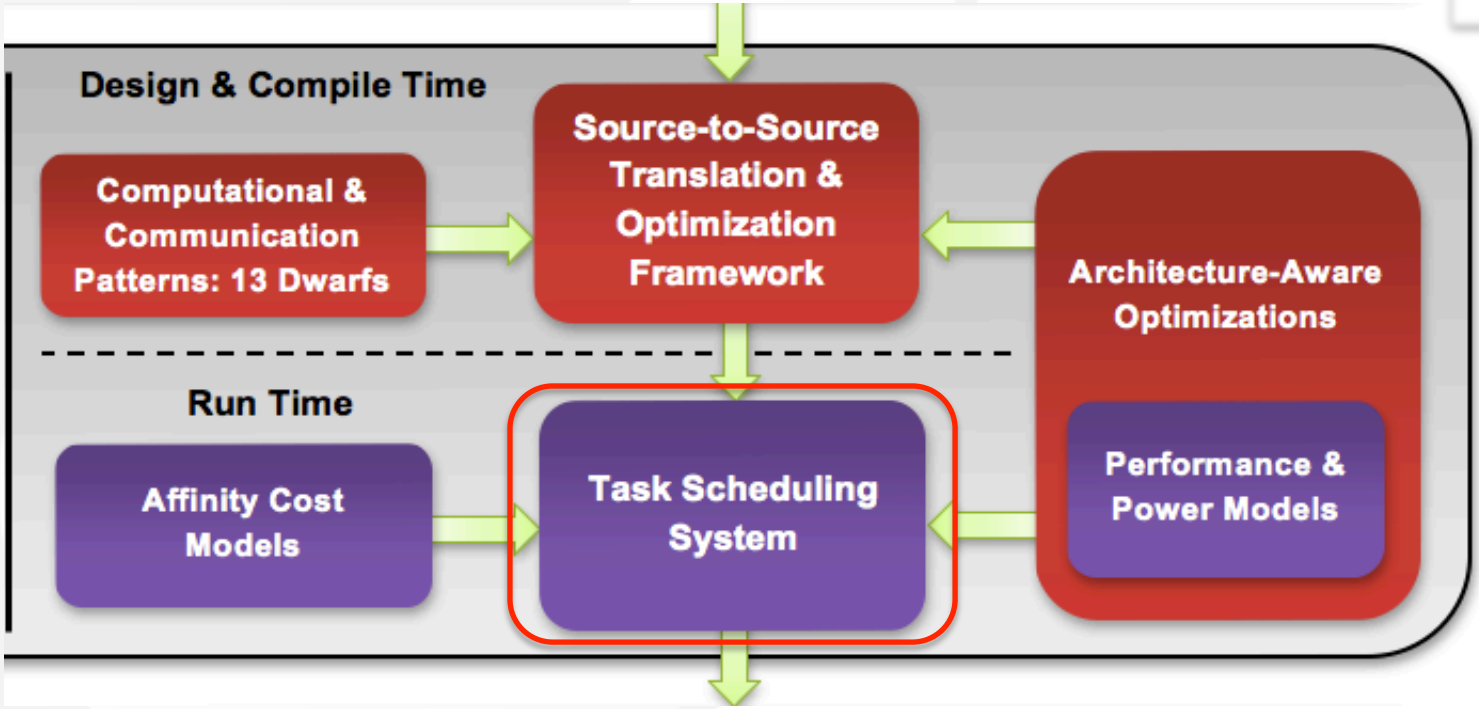
Neuro-informatics



Cybersecurity

MAVs

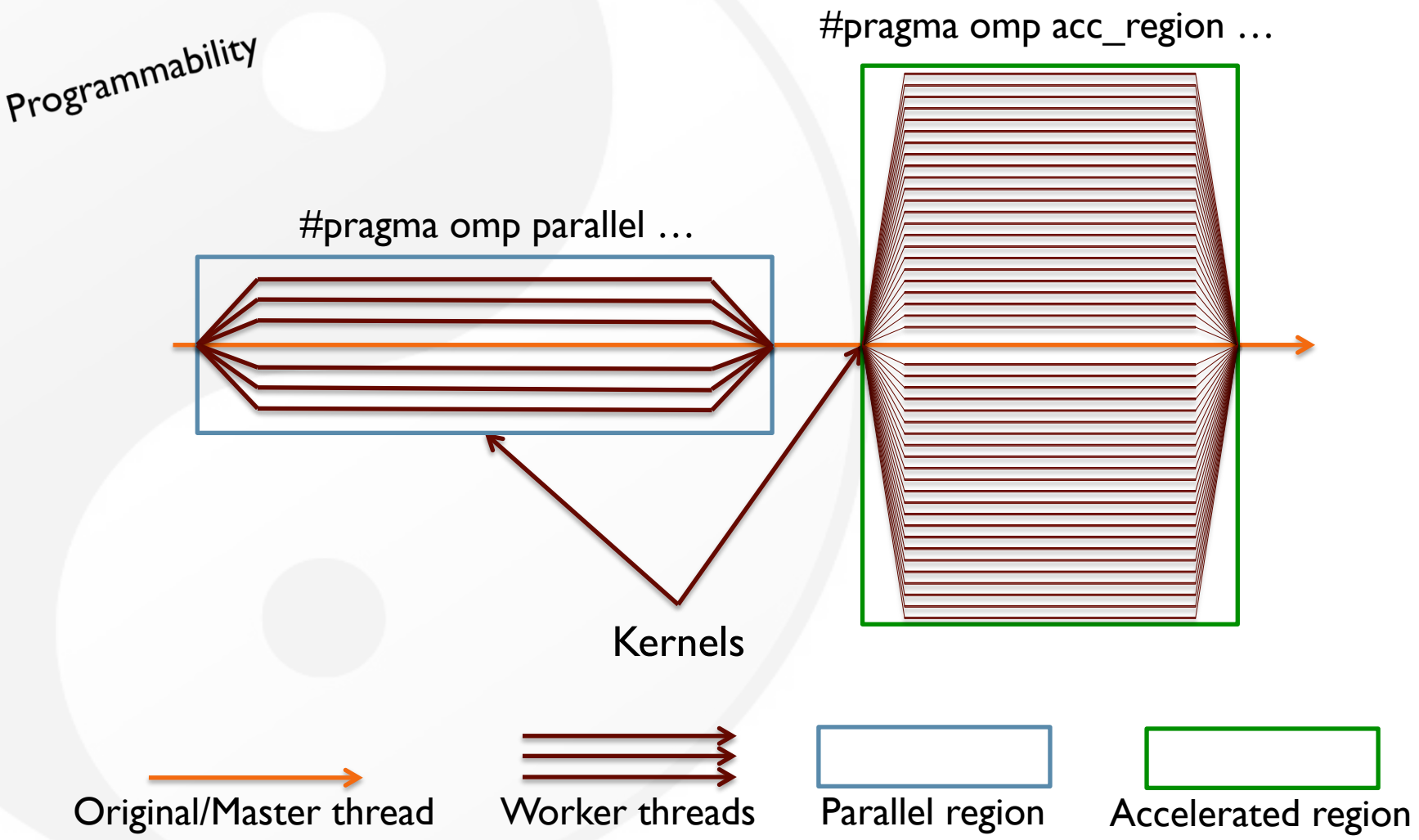




Manual Co-Design → Automated Co-Design

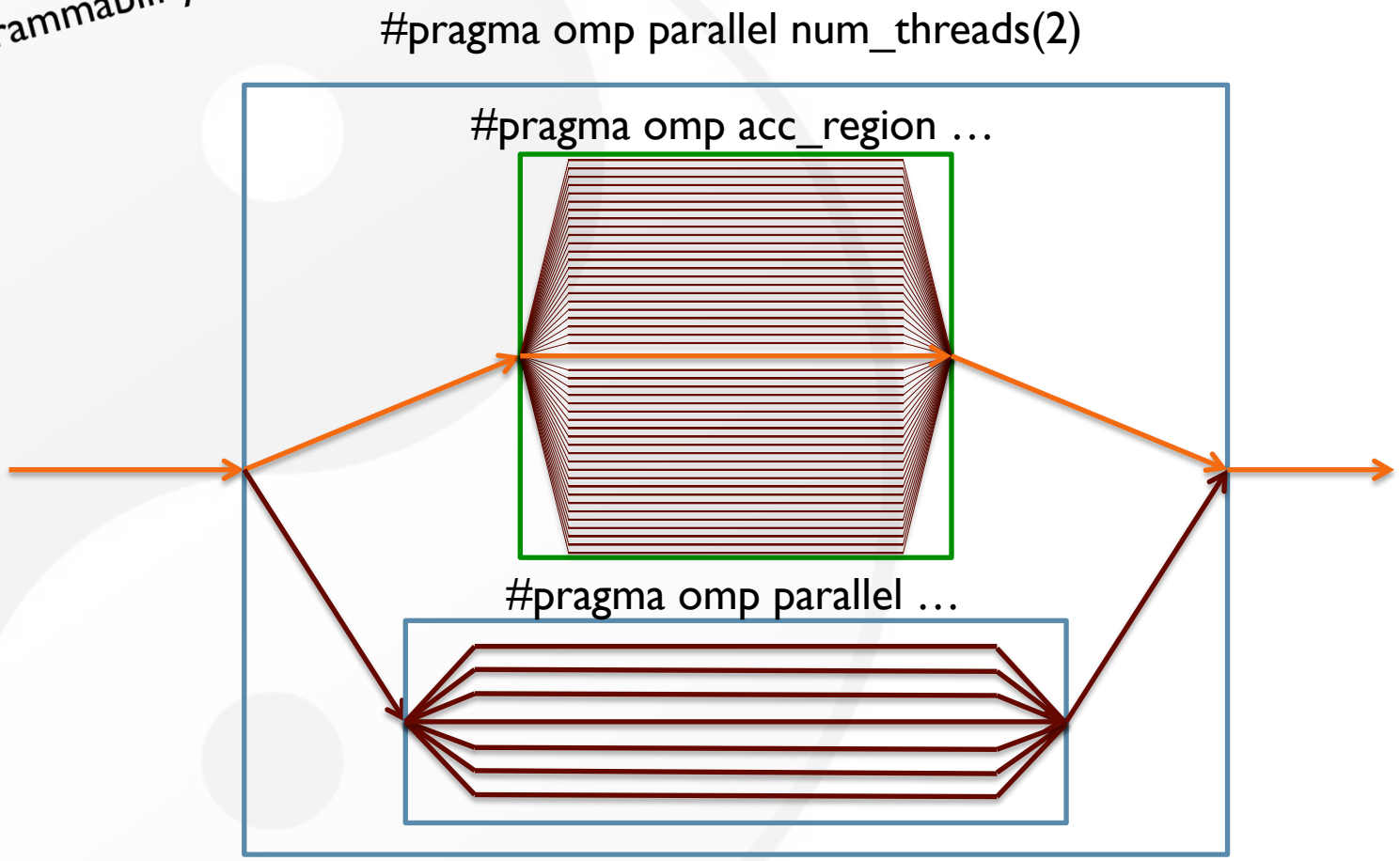
Heterogeneous Parallel Computing Platform

OpenMP Accelerator Behavior



Our Version

Programmability



Original/Master thread

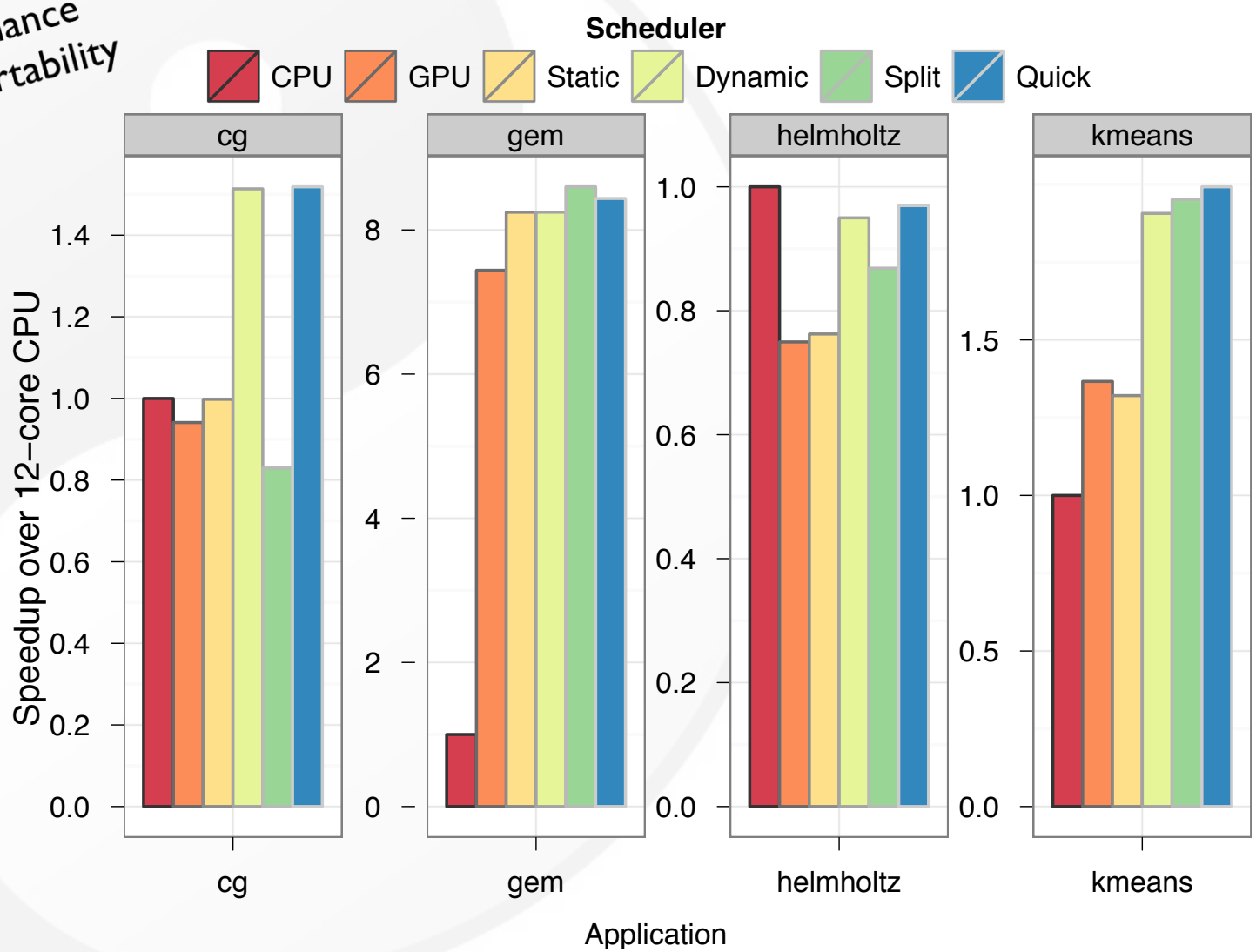
Worker threads

Parallel region

Accelerated region

Preliminary Results of Automated Task Schedulers

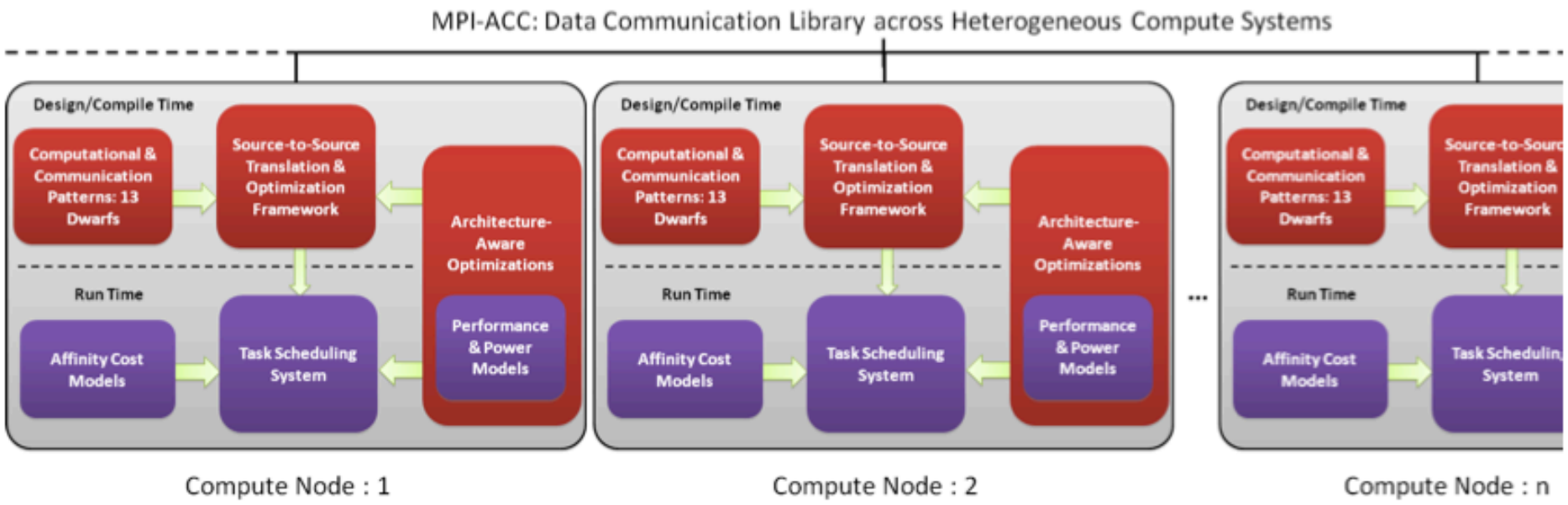
Performance and Portability



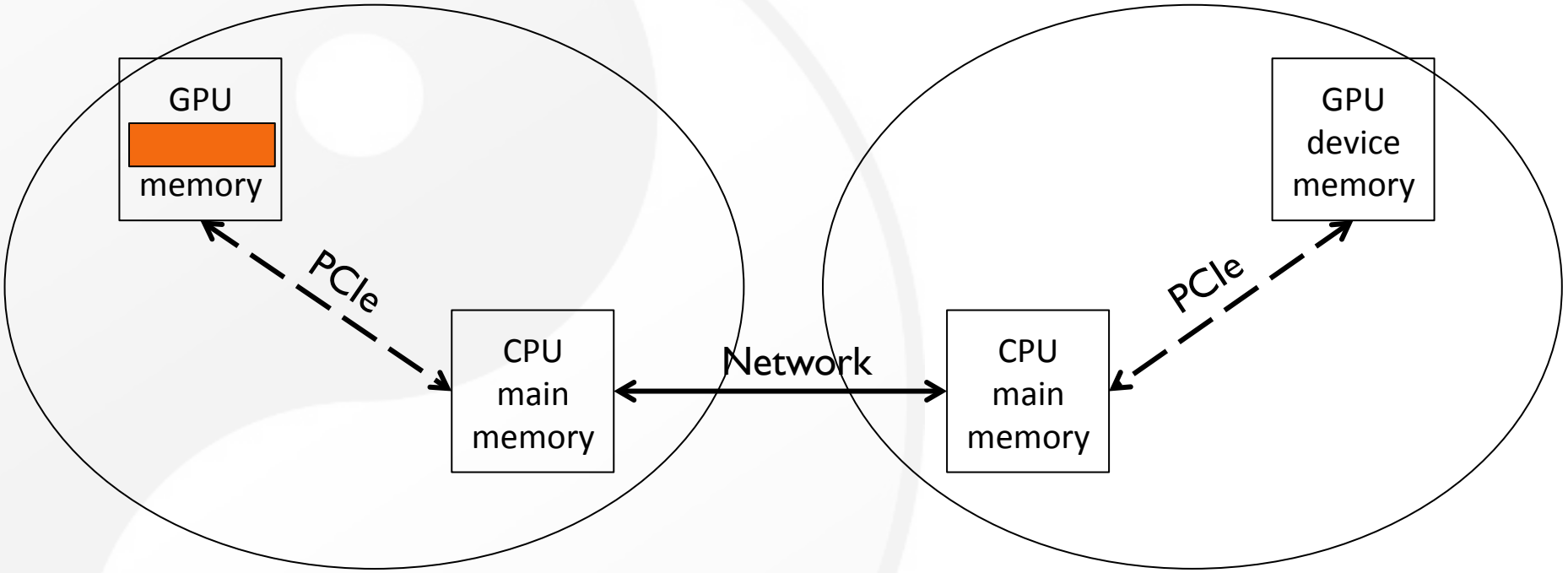
Inter-Node

Ecosystem for Heterogeneous Parallel Computing

- Goal
  - Unified data movement library that hides all the hardware and system software details from the algorithm developer while supporting a multitude of environments



Data Movement in CPU-GPU Clusters



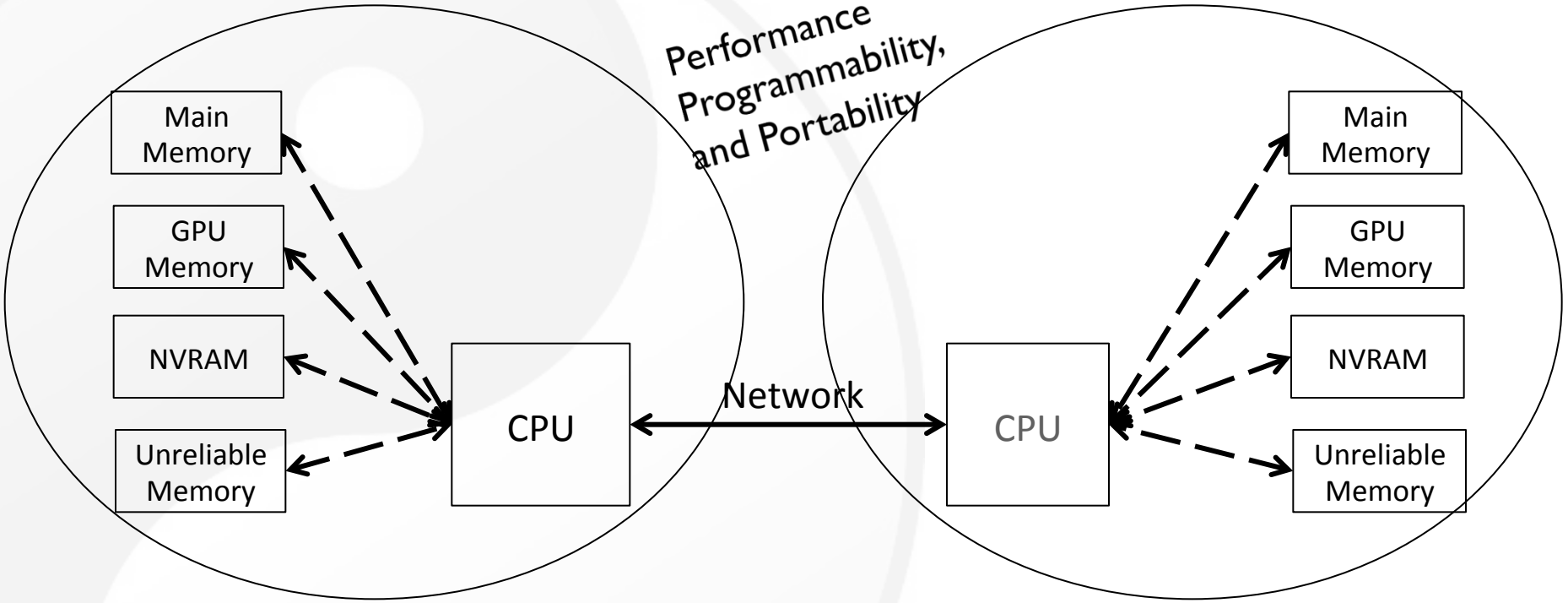
MPI Rank = 0

```
if(rank == 0)
{
  GPUMemcpy(host_buf, dev_buf, D2H)
  MPI_Send(host_buf, .. ..)
}
```

MPI Rank = 1

```
if(rank == 1)
{
  MPI_Recv(host_buf, .. ..)
  GPUMemcpy(dev_buf, host_buf, H2D)
}
```

MPI-ACC: Generalized Runtime for Accelerator Systems



Rank = 0

```
if (rank == 0)
{
    MPI_Send(s_buf, .. ..);
}
```

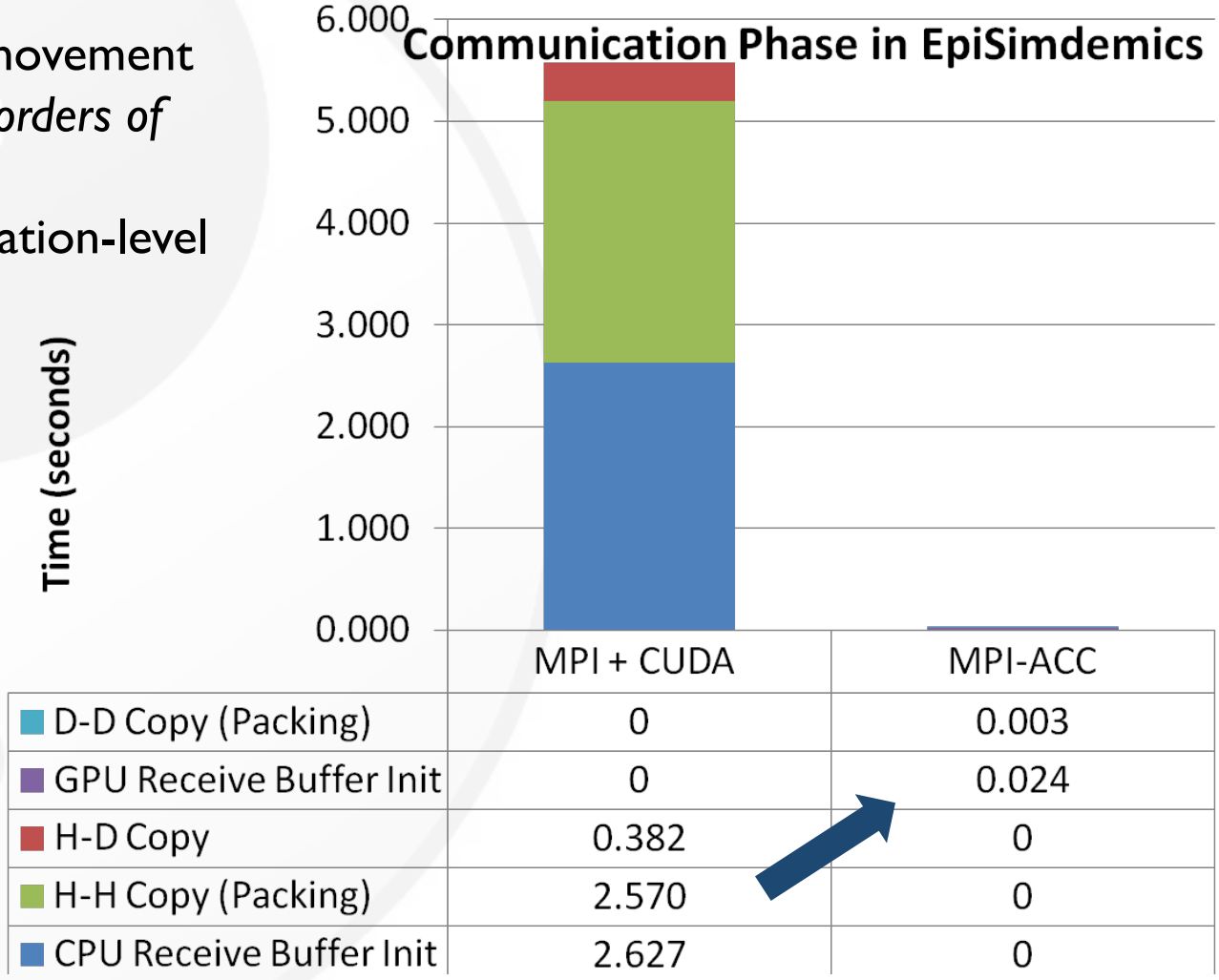
Rank = 1

```
if (rank == 1)
{
    MPI_Recv(r_buf, .. ..);
}
```

MPI-ACC Performance Comparison

- Accelerates data movement operations *by two orders of magnitude*
- *Enables* new application-level optimizations

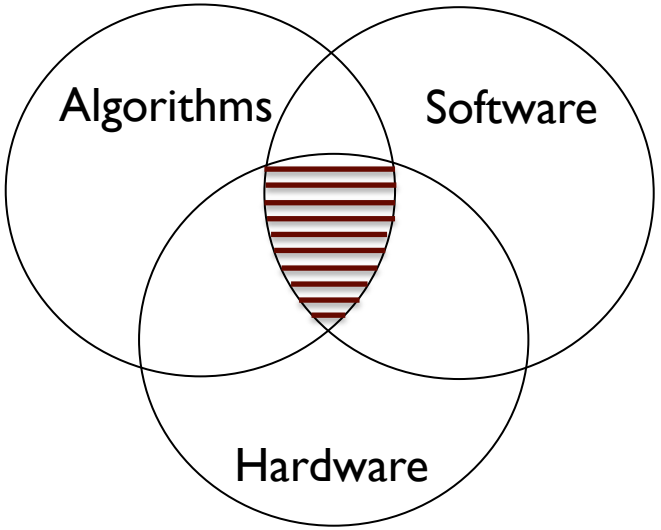
Performance Programmability, and Portability



MPI-ACC runtime optimized

Roadmap

- Vision
- Team
- Approach
- Infrastructure
- Co-Design Research
  - Computer Science
  - CFD Codes (4)
  - Mathematics
- Achievements & Publications
- Next Steps



Overall Objective

- Extend and port our (four) CPU-based CFD methods to modern heterogeneous computing systems, particularly those with GPUs, via a synergistic hardware/software co-design approach that seeks to significantly improve performance, thus significantly enhancing the computational capabilities of current CFD tools.
  - Characterize and implement the following methods:
    - Projection and artificial compressibility methods for incompressible flows
    - Structured, unstructured, and Cartesian grids
    - Arbitrary Lagrangian-Eulerian (ALE) and immersed boundary methods (IBM) for boundary deformation
    - Finite volume and finite element methods based on the reconstructed discontinuous Galerkin (RDG) technique.
  - Co-design, test, verify, and assess the above methods for solving a variety of low Reynolds number incompressible flow problems of interest to the U.S. Air Force in general and for predicting MAV aerodynamics in particular.

Targeted CFD Codes

SENSEI (C. Roy, Virginia Tech)

- Structured, multiblock, 2nd order, finite-volume code
- Artificial compressibility method
- Arbitrary Lagrangian/Eulerian (ALE) 2nd or higher order spatial accuracy
- Artificial compressibility (AC) and immersed boundary (IB) methods

GenIDLEST (D. Tafti, Virginia Tech)

- Structured, multiblock, 2nd order, finite-volume code
- Pressure projection method
- ALE and immersed boundary methods (IBM)

RDGFLO (H. Luo, NCSU)

- Unstructured, discontinuous Galerkin (DG) method
- High-order solution of compressible flows

INCOMP3D (J. Edwards, NCSU)

- Structured, multiblock finite-volume code
- Second or higher order spatial accuracy
- ALE and IBM

SENSEI: Overview

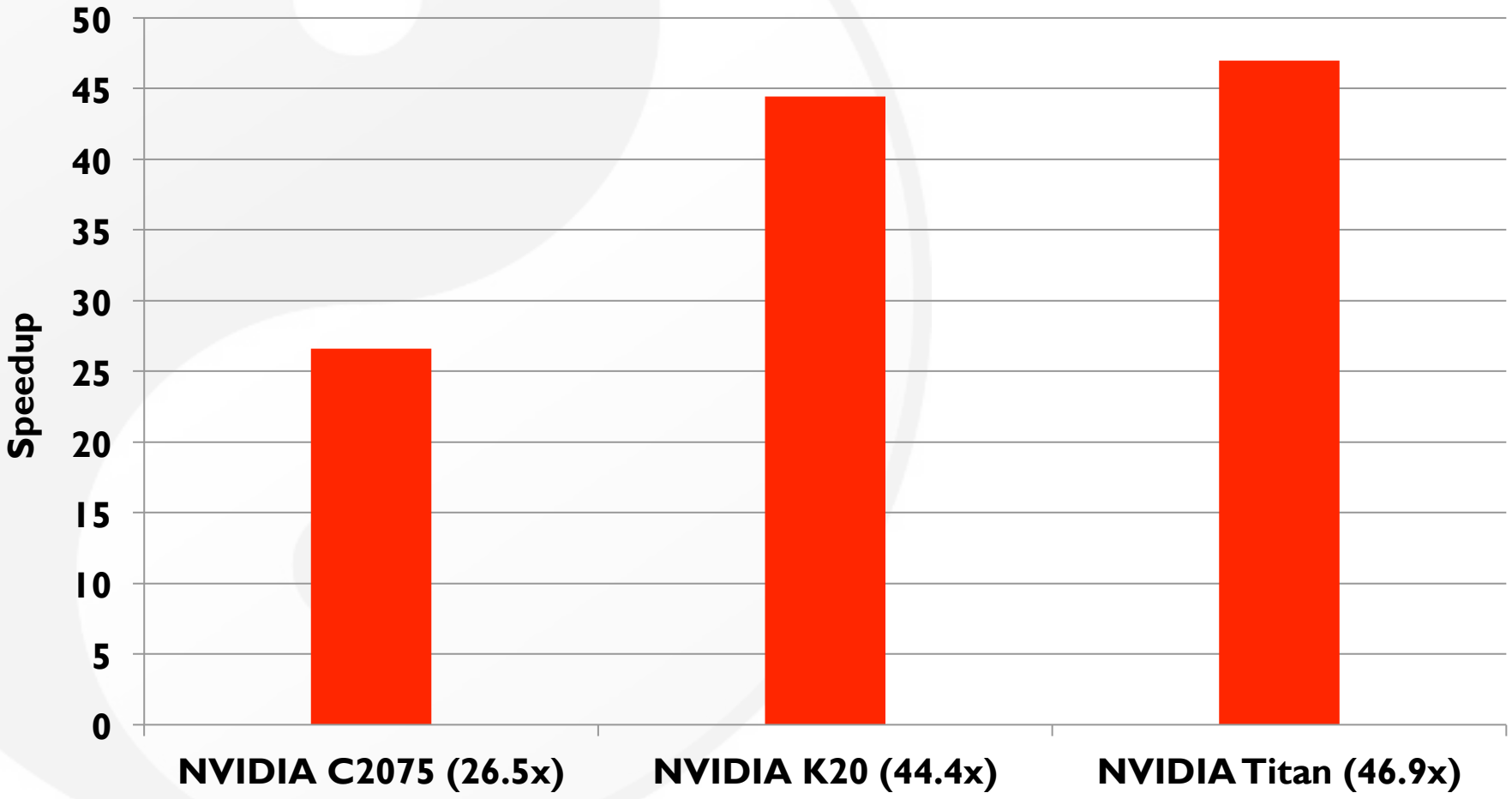
- SENSEI Lite (2D)
  - 2D, Cartesian, finite-difference code
  - Artificial compressibility
  - Explicit solver
  - CPU and GPU Optimization completed
- SENSEI (3D)
  - 3D, multiblock, finite-volume code
  - Artificial compressibility
  - Explicit and implicit solvers
  - Modern Fortran 2003/08 implementation
  - Challenge: Dynamically allocated arrays inside Fortran-derived types

Tradeoff between programmability, portability, and performance

Performance of SENSEI Lite

GPU speedup over *non-vectorized* single-thread CPU version

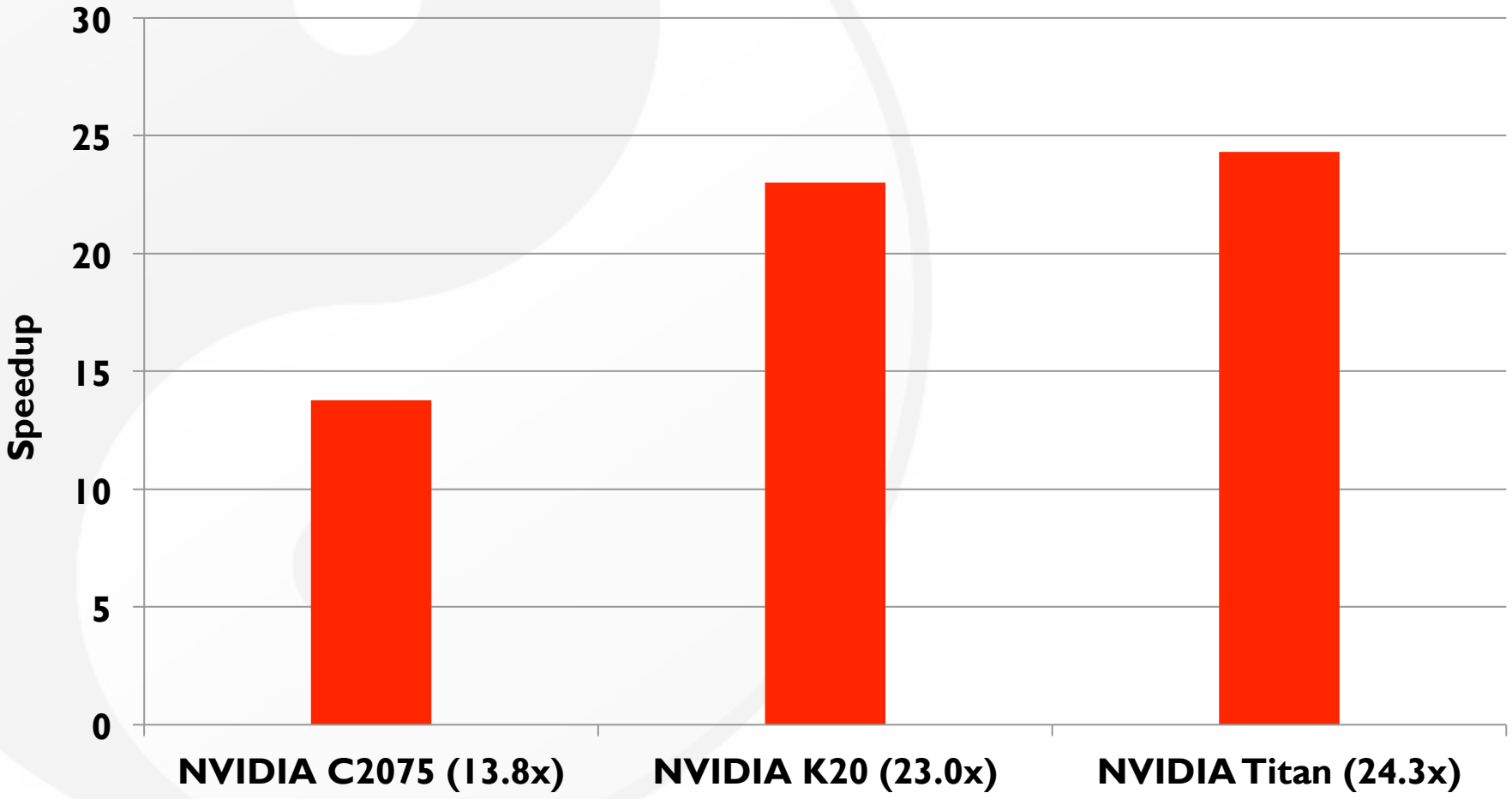
CPU = Xeon X5560 @2.8 GHz



Performance of SENSEI Lite

GPU speedup over *fastest (SSE vectorized)* single-thread CPU version

CPU = Xeon X5560 @2.8 GHz



Targeted CFD Codes

SENSEI (C. Roy, Virginia Tech)

- Structured, multiblock, 2nd order, finite volume code
- Artificial compressibility method
- Arbitrary Lagrangian/Eulerian (ALE) 2nd or higher order spatial accuracy
- Artificial compressibility (AC) and immersed boundary (IB) methods

GenIDLEST (D. Tafti, Virginia Tech)

- Structured, multiblock, 2nd order, finite volume code
- Pressure projection method
- ALE and immersed boundary methods (IBM)

RDGFLO (H. Luo, NCSU)

- Unstructured, discontinuous Galerkin (DG) method
- High-order solution of compressible flows

INCOMP3D (J. Edwards, NCSU)

- Structured, multiblock finite volume code
- Second or higher order spatial accuracy
- ALE and IBM

GPU GenIDLEST: Preliminary Profiling

- Performance Breakdown

Time %	Component
44%	kernel_pc_jac_blk2_pc_ortho
19.3%	***GPU-CPU Data Transfers***
7.32%	kernel_pc_jac_glb2_ortho
7.13%	kernel_matxvec2_ortho
….	….

- Data marshaling performed on GPU, but …
  - Data transfers still ~ 20% of execution time
  - Maximum speedup achievable: 5x! (Amdahl's law)
    - Better overlapping of computations with transfers → MPI-ACC

GPU GenIDLEST: Issues

- Kernels operating at low occupancy
  - 52% @ 75% occupancy, 33% @ 25% occupancy, 15% @ < 25% occupancy
  - Manual optimization techniques: (1) Improved register usage, (2) optimize use of in-kernel resources, and (3) concurrent kernel execution
- Reductions performed on CPU
  - Entails transferring data back to CPU for reduction
  - Efficient reduction algorithms available for GPU → port reduction to GPU
- Linear algebra kernels – saxpy, daxpy, matrix-vector multiply
  - 15% of execution time
  - Benchmark performance against GPU BLAS libraries
- Sliced array data transfers in CUDA Fortran
  - Results into multiple cudaMemcpy calls for a single data transfer. OUCH!

Targeted CFD Codes

SENSEI (C. Roy, Virginia Tech)

- Structured, multiblock, 2nd order, finite volume code
- Artificial compressibility method
- Arbitrary Lagrangian/Eulerian (ALE) 2nd or higher order spatial accuracy
- Artificial compressibility (AC) and immersed boundary (IB) methods

GenIDLEST (D. Tafti, Virginia Tech)

- Structured, multiblock, 2nd order, finite volume code
- Pressure projection method
- ALE and immersed boundary methods (IBM)

RDGFLO (H. Luo, NCSU)

- Unstructured, discontinuous Galerkin (DG) method
- High-order solution of compressible flows

INCOMP3D (J. Edwards, NCSU)

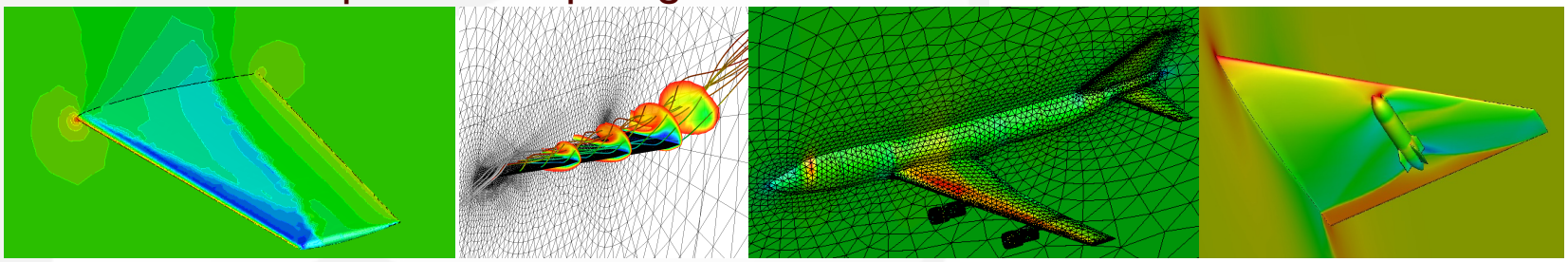
- Structured, multiblock finite volume code
- Second or higher order spatial accuracy
- ALE and IBM

RDGFLO: Overview

Reconstructed Discontinuous Galerkin Flow Solver

Key Features

- Compressible Navier-Stokes / Euler equations.
- Third-order reconstructed discontinuous Galerkin (DG) finite element method.
- Unstructured hybrid grids, i.e., tetrahedron, prism, pyramid, hexahedron.
- Time-accurate and steady-state solution schemes.
- MPI-based parallel computing on CPU clusters.



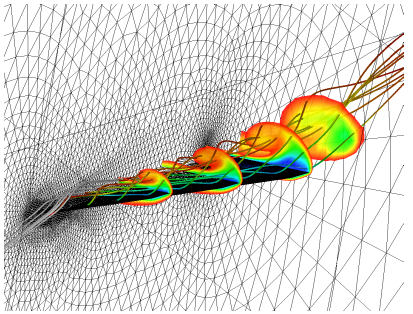
Need GPU acceleration for RDGFLO because …

High-order methods are expensive for large-scale problems in terms of computing time!

Rewriting a huge legacy code using CUDA is too costly. Alternatively, …

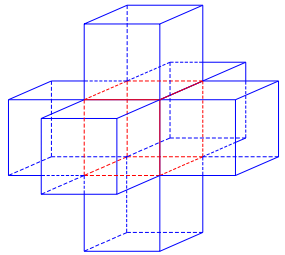
OpenACC does not require much change of data structures and algorithms in a legacy code.

RDGFLO: GPU Parallelization

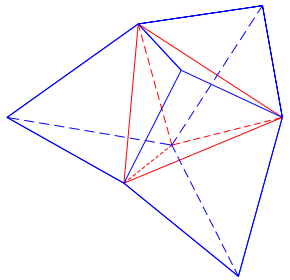


Race condition occurs in GPU parallelization in loops over faces when data writes to their left and right cell arrays.

Example: the elemental residual array for the cell (red) can be overwritten simultaneously in multiple GPU cores in loops over its faces by its face-neighboring cells (blue).



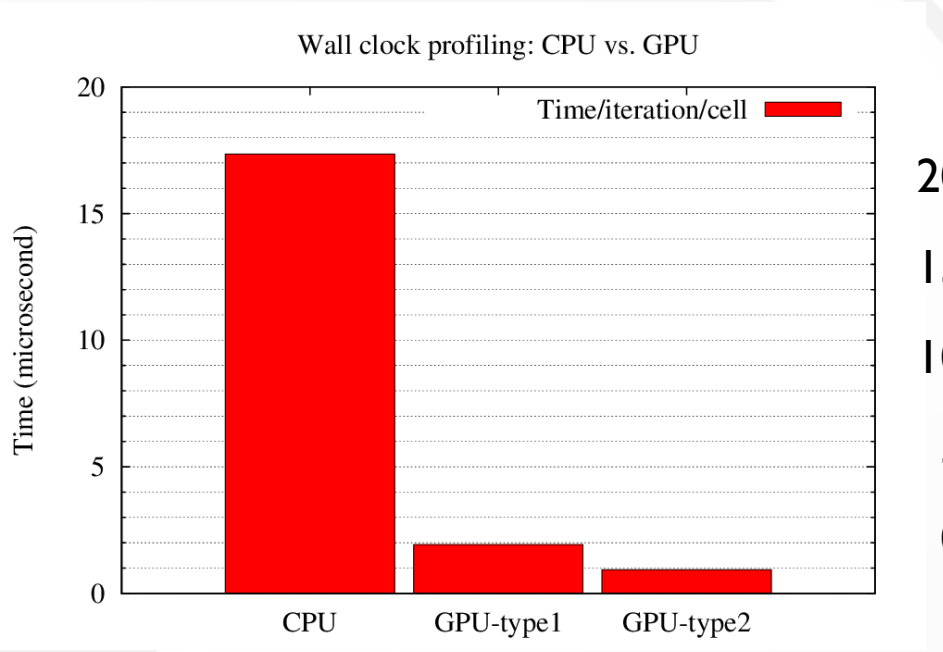
Coloring algorithm: reorder face indices and pack them in groups; Criterion: faces that share common elements do not reside in the same group (the same strategy in the case of OpenMP).

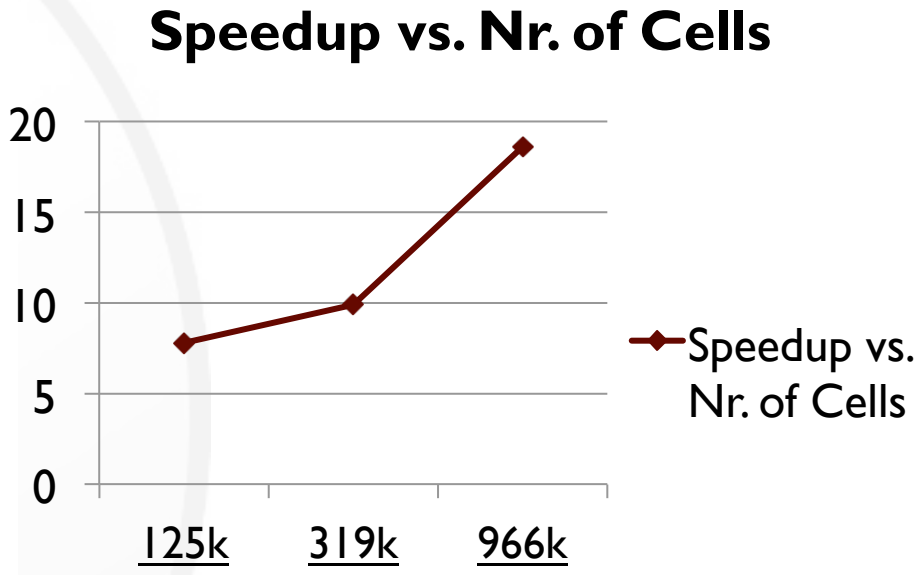


Example: an ACC sequential loop over the groups is nested outside the ACC parallel loop over the faces (the original loop is untouched).

All geometric and solution arrays are copied from host to device only once. For steady-state problems, solution arrays are only copied back to host memory at the end of time iterations and dumped in files.

RDGFLO: Weak Scaling Tests





Unit timings on the grid of 966k cells
-- CPU: AMD Opteron Processor 6128
-- GPU-type1: nVidia Tesla C2050
-- GPU-type2: nVidia Tesla K20c

Preliminary Result: A speedup factor of 20x (or more) is achievable.

Targeted CFD Codes

SENSEI (C. Roy, Virginia Tech)

- Structured, multiblock, 2nd order, finite volume code
- Artificial compressibility method
- Arbitrary Lagrangian/Eulerian (ALE) 2nd or higher order spatial accuracy
- Artificial compressibility (AC) and immersed boundary (IB) methods

GenIDLEST (D. Tafti, Virginia Tech)

- Structured, multiblock, 2nd order, finite volume code
- Pressure projection method
- ALE and immersed boundary methods (IBM)

RDGFLO (H. Luo, NCSU)

- Unstructured, discontinuous Galerkin (DG) method
- High-order solution of compressible flows

INCOMP3D (J. Edwards, NCSU)

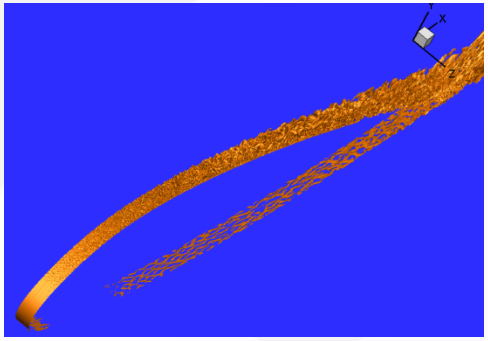
- Structured, multiblock finite volume code
- Second or higher order spatial accuracy
- ALE and IBM

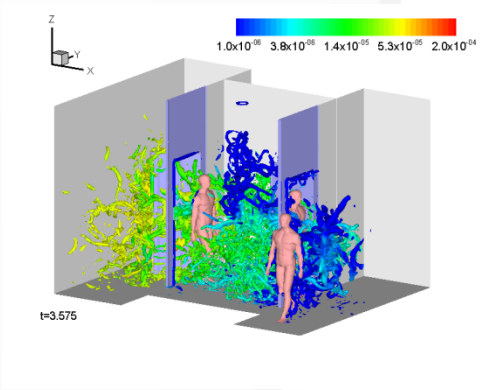
INCOMP3D: Overview

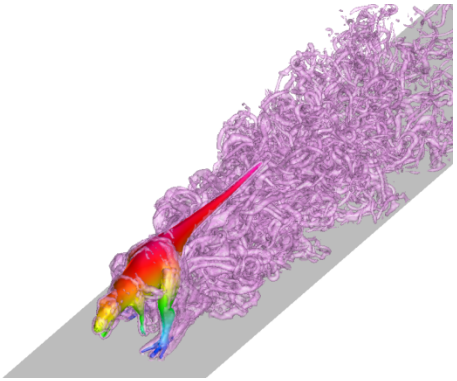
Multi-block Incompressible Navier-Stokes Solver for Large Eddy Simulation

Key Features

- Higher order PPM / central-difference schemes
- Fully implicit time evolution using dual-time stepping methodology
- Multi-block structured meshes (MPI parallelism)
- Immersed boundary methods for complex motion events

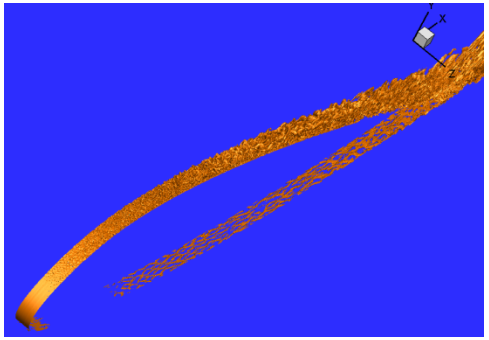






Need GPU acceleration for INCOMP3D to reduce costs associated with large-eddy simulation at high Reynolds numbers

INCOMP3D: GPU Parallelization

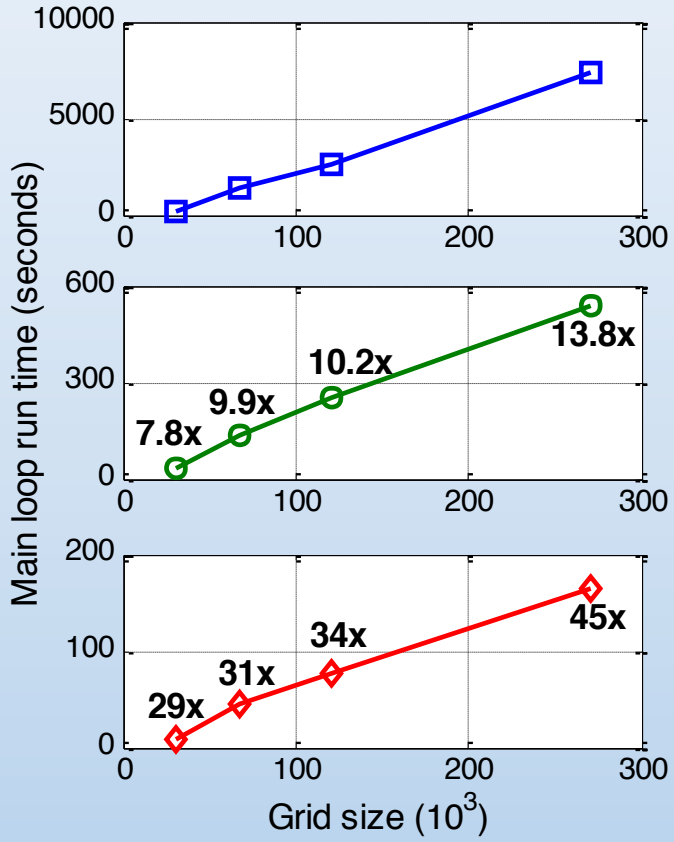


- Two scaled-down versions realized
- OpenACC (ACC) port
  - Main loop is carried out on GPU only. All essential arrays remain on GPU main memory. Temporary data arrays are created directly on GPU.
  - One code can be compiled into pure CPU or GPU-accelerated versions, using different compiler options. This facilitates long-term maintenance.
- CUDA Fortran (CUF) port
  - Each time step are carried out by one monolithic kernel, which includes residual calculation and time marching.
  - Residual array is directly created in shared memory; time step is local to each thread. Memory requirement is greatly reduced.
  - Overlapping blocks are used, due to inter-thread data dependency.
  - Residual calculation involves flux grouping by direction (i and j directions), to avoid memory contingency.
  - CUDA Fortran array overhead identified and solved using global variables.

INCOMP3D: Initial Findings

- ACC and CUF both achieved significant speedup over CPU
- CUF achieved better performance, but requires much more effort to port and maintain
  - Direct access of shared memory allows greater flexibility on algorithms.
  - Easier to make mistakes; harder to debug.
- ACC provides a good compromise between CPU and CUDA
  - Good speedup (~10x), with minimal to moderate effort on porting.
  - Easier to debug and maintain.

Example: steady-state flow inside a channel with 3 circular obstacles, Re=200. All results (in seconds) are obtained using nVidia c2050.

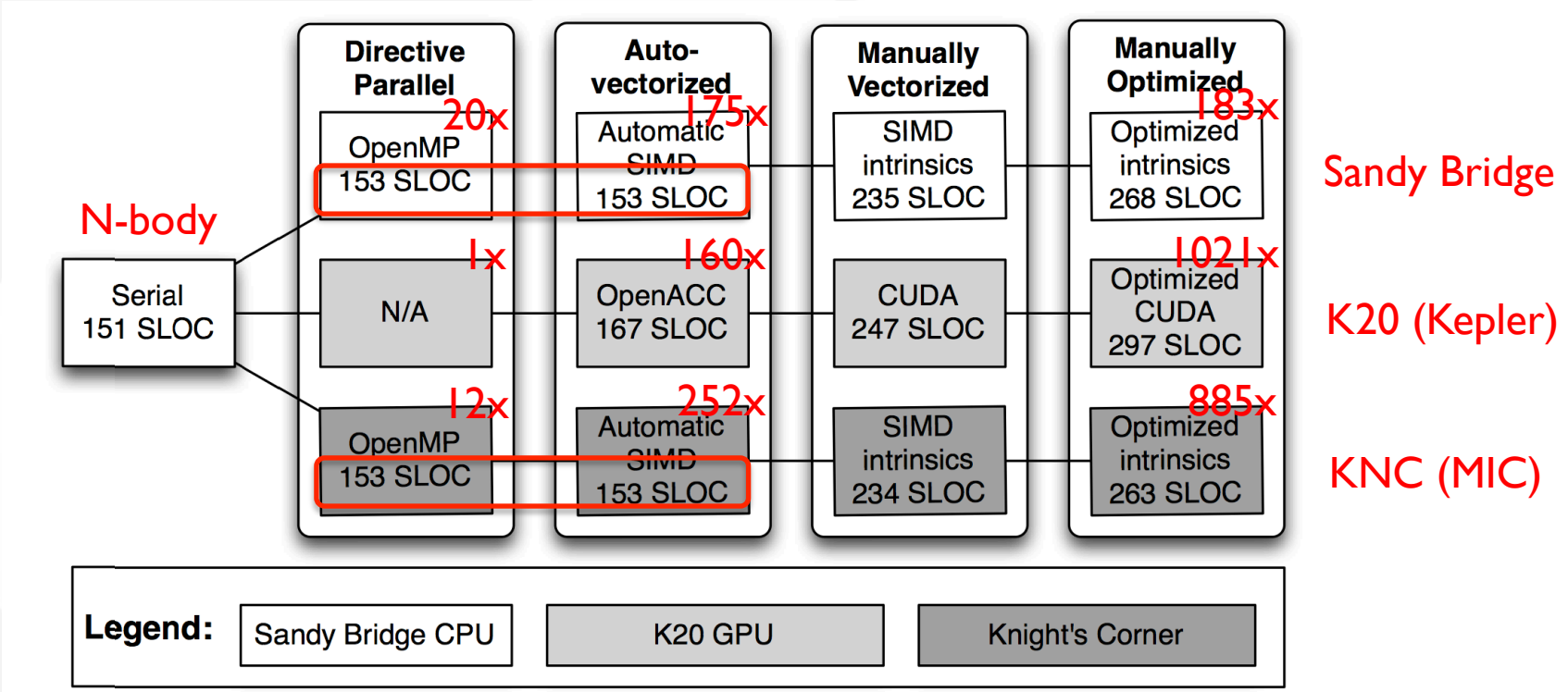


Co-Design Around Three P's: Performance, Programmability, Portability

"Productivity = Performance + Programmability + Portability"

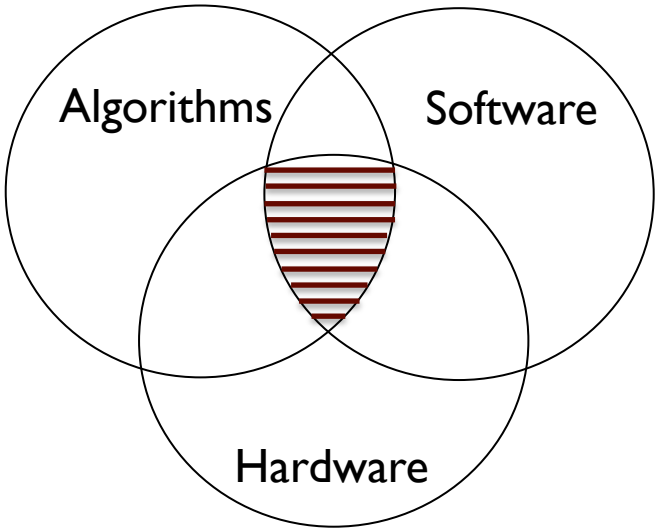
- Multi-dimensional optimization across two or more P's

  … first *manual co-design* … then *automated co-design*



Roadmap

- Vision
- Team
- Approach
- Infrastructure
- Co-Design Research
  - Computer Science
  - CFD Codes (4)
  - Mathematics (de Sturler, Sandu)
- Achievements & Publications
- Next Steps

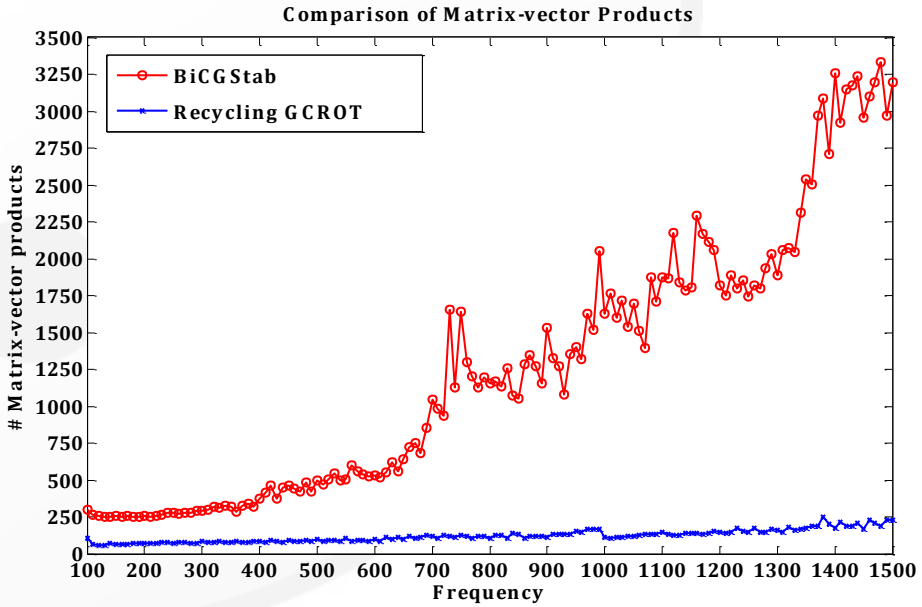


Solvers on GPUs and Multicore Processors

- Goal: Minimum Total Execution Time (vs Accuracy)
  - Fewer *expensive* iterations vs more *cheap* iterations
- Current General Effort
  - Efficient implementation of each kernel (data layout)
  - Efficient implementation of standard preconditioners
  - Better performance by varying parameters/preconditioners, memory-intensive algorithms have poor cache performance
- Future Efforts
  - Combine kernels/iterations, latency hiding, more arithmetic vs. data movement
  - Vary precision, analyze accuracy, and convergence
  - Alternative preconditioners
    - Faster convergence, better hardware utilization, data storage/access
    - Modify algorithms (solvers/preconditioners) – need to maintain good convergence

New(er) Solvers with Better Trade-offs

- Faster convergence (fewer iterations) for additional orthogonalizations and upfront matvecs, but all at once
  - Higher level BLAS/more work per data movement – trade off with sparse matvec
  - Better convergence allows cheaper preconditioner
  - Only one sync per extra vector in space
  - Especially useful for sequences of linear systems



Acoustics problem: Tire noise

Collaboration Jan Bierman, BMW

Preconditioner Effort

- Develop preconditioners with good GPU performance and fast convergence
- Avoid triangular solves (in ILU type prec.s)? Or improve performance by alternative storage schemes?
- Prec.s that use SpMV - Sparse Approximate Inverses
  - Improve convergence by multilevel correction
  - Multigrid-like preconditioners (AMG)
  - Domain decomposition preconditioners? (but without local direct solve)
- Computing preconditioners is expensive
  - Update preconditioners for sequence of problems

New Rosenbrock-Krylov Matrix-Free Integrators Using Minimal Implicitness

- New implicit matrix-free Rosenbrock-Krylov time integration methods avoid solution "to convergence" and guarantee order of consistency with small Krylov spaces (e.g., four-dimensional)
- Full-space linear algebra parallelized using cuBLAS. Arnoldi iteration parallelized with a custom CUDA kernel for the orthogonalization step.
- Tested on a parallel implementation of the shallow water equations using different grid sizes and tolerances.

Rosenbrock-Krylov Methods
Suited for Parallelization and Acceleration

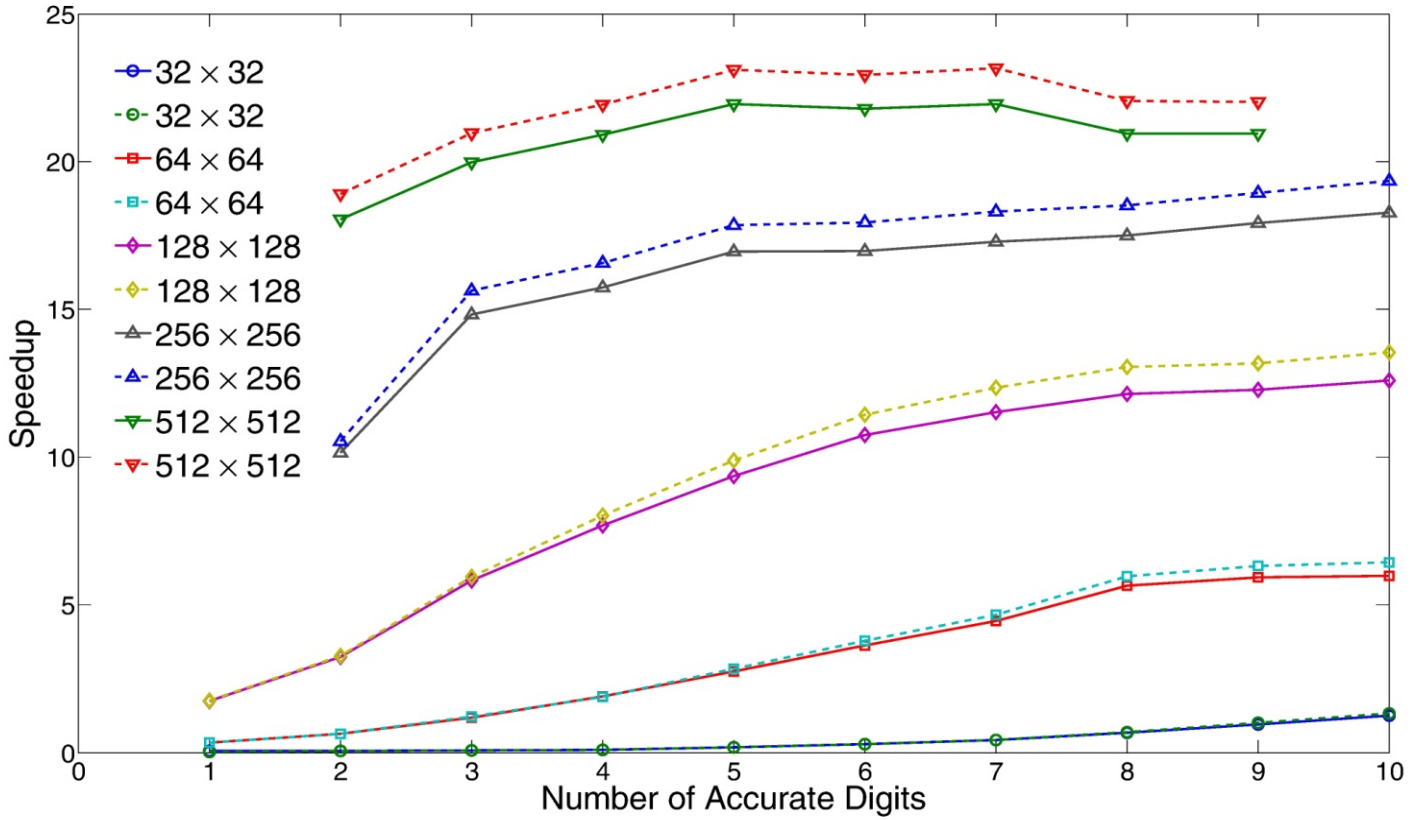


Figure: Speedup of GPU accelerated ROK-SWE over serial.
Dashed lines use a custom kernel for the Arnoldi, while solid use cuBLAS.
(GPU=NVIDIA Quadro 4000, CPU=Intel Xeon 2.5GHz).

Achievements

- CS
  - Compile/Design Time: Infer speedup potential before refactoring code
    - Memory trace compression algorithms to estimate parallelization benefits
  - At Run Time: Initial evaluation of our automated run-time system prototype (for Accelerated OpenMP and OpenACC) → vendors
  - Identified commonality for library for CFD codes: generalized GPU-to-GPU communication (via MPI), ghost cell exchange between GPUs, …
- CFD
  - Preliminary GPU parallelization of prototype CFD codes with OpenACC and CUDA
  - Up to 40x speed-up over a single CPU core (25x over SSE vectorized)
- Math
  - Initial GPU-parallelized Rosenbrock-Krylov method
  - Integrated initial CUDA-parallelized solvers with CFD

Publications

- P. Tranquilli, A. Sandu, "Rosenbrock-Krylov Methods for Large Systems of Differential Equations" http://arxiv.org/abs/1305.5481, May 2013.
- J. M. Derlaga, T. S. Phillips, C. J. Roy, "SENSEI Computational Fluid Dynamics Code: A Case Study in Modern Fortran Software Development," AIAA Paper 2013-2450, *21st AIAA Computational Fluid Dynamics Conf.*, June 2013.
- S. R. Glandon, P. Tranquilli, A. Sandu, " Acceleration of matrix-free time integration methods", *Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA) at SC13*, November 2013.
- B. P. Pickering, C. J. Roy, T. R. W. Scogland, W. Feng, "Directive-Based GPU Programming for Computational Fluid Dynamics," *52nd AIAA Aerospace Sciences Meeting*, January 2014.
- Y. Xia, L. Luo, H. Luo, J. Edwards, F. Mueller, 'GPU Acceleration of a Reconstructed Discontinuous Galerkin Method for Compressible Flows on Unstructured Grids", *52nd AIAA Aerospace Sciences Meeting*, January 2014.
- L. Luo, J. R. Edwards, H. Luo, F. Mueller, "Performance Assessment of Multi-block LES Simulations using Directive-based GPU Computation in a Cluster Environment," *AIAA SciTech 2014*, January 2014.

What's Next?

- Platforms
  - AMD & Intel CPU, AMD APU, AMD & NVIDIA GPUs, Intel MIC
- Towards Ease of Use and Automation (for Performance, Programmability, and Portability)
  - Web resource for *tenets of synergistic co-design* … between algorithms, software, and hardware → automation (long term)
  - Towards a CFD library for heterogeneous computing systems
    - GPU-integrated MPI vs. GPUDirect, ghost cell exchange, bounds checking, …
  - Code repositories for production codes
- GPU-Integrated MPI Evaluation
  - Experimental platforms (MIC and next-generation APU w/ "infinite memory")
- GPU mixed-precision solvers, GPU-efficient preconditioners
- GPU-efficient accurate and stable high-order time stepping

Acknowledgements

- This work was funded by the Air Force Office of Scientific Research (AFOSR) Computational Mathematics Program
  - Program Manager: Fariba Fahroo
  - Grant No. FA9550-12-1-0442