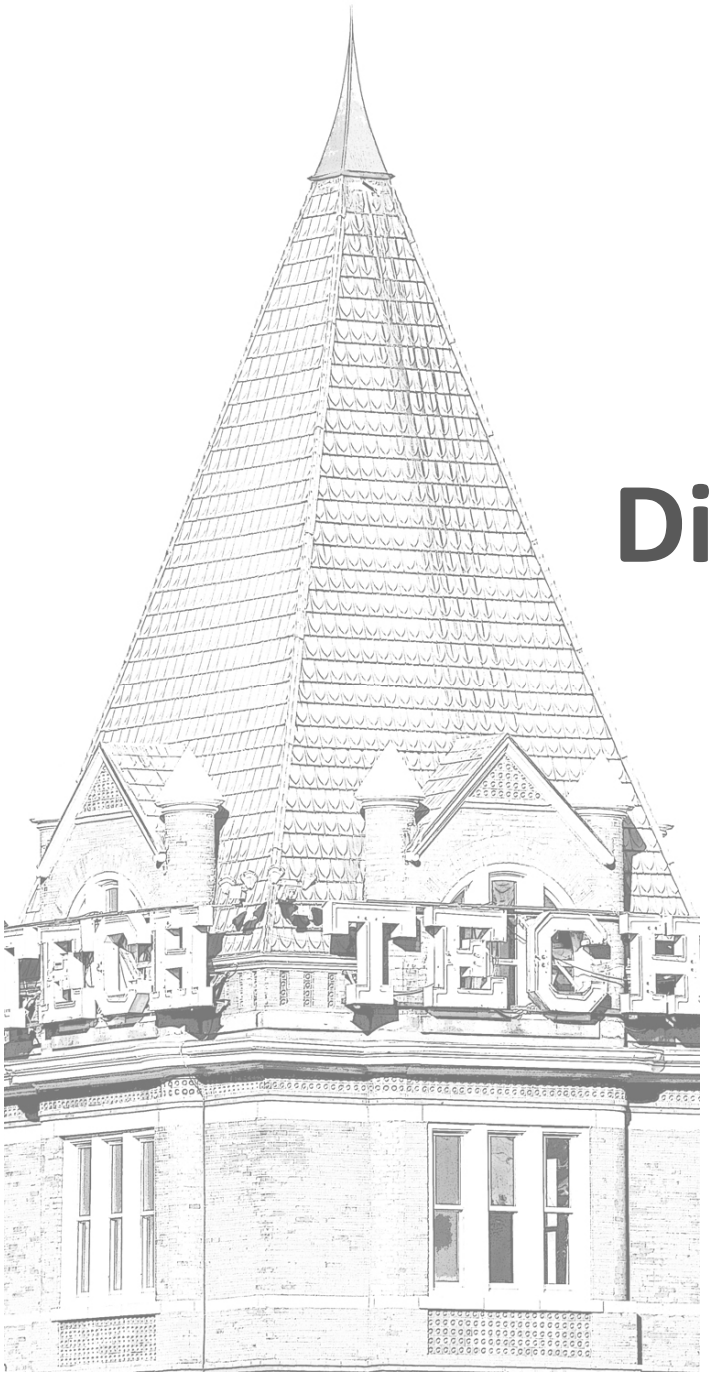


Georgia Institute of Technology

Enabling Science and Discovery at Georgia Tech With MVAPICH2

3rd Annual MVAPICH User Group (MUG) Meeting

August 19-21, 2015

Mehmet Belgin, Ph.D.
Research Scientist
PACE Team, OIT/ART

Georgia Tech

- #7 best public university (U.S. News & World report, 2014)
- College of Science consistently in top 5
- #1 Industrial Engineering Program for the past 2 decades
- 21,500 undergrad and grad students
- Colleges: Architecture, Computing, Engineering, Sciences, Business, Liberal Arts

PACE (PACE.GATECH.EDU)

Georgia Tech

what is:

A **P**artnership for an **A**dvanced **C**omputing **E**nvironment

provides:

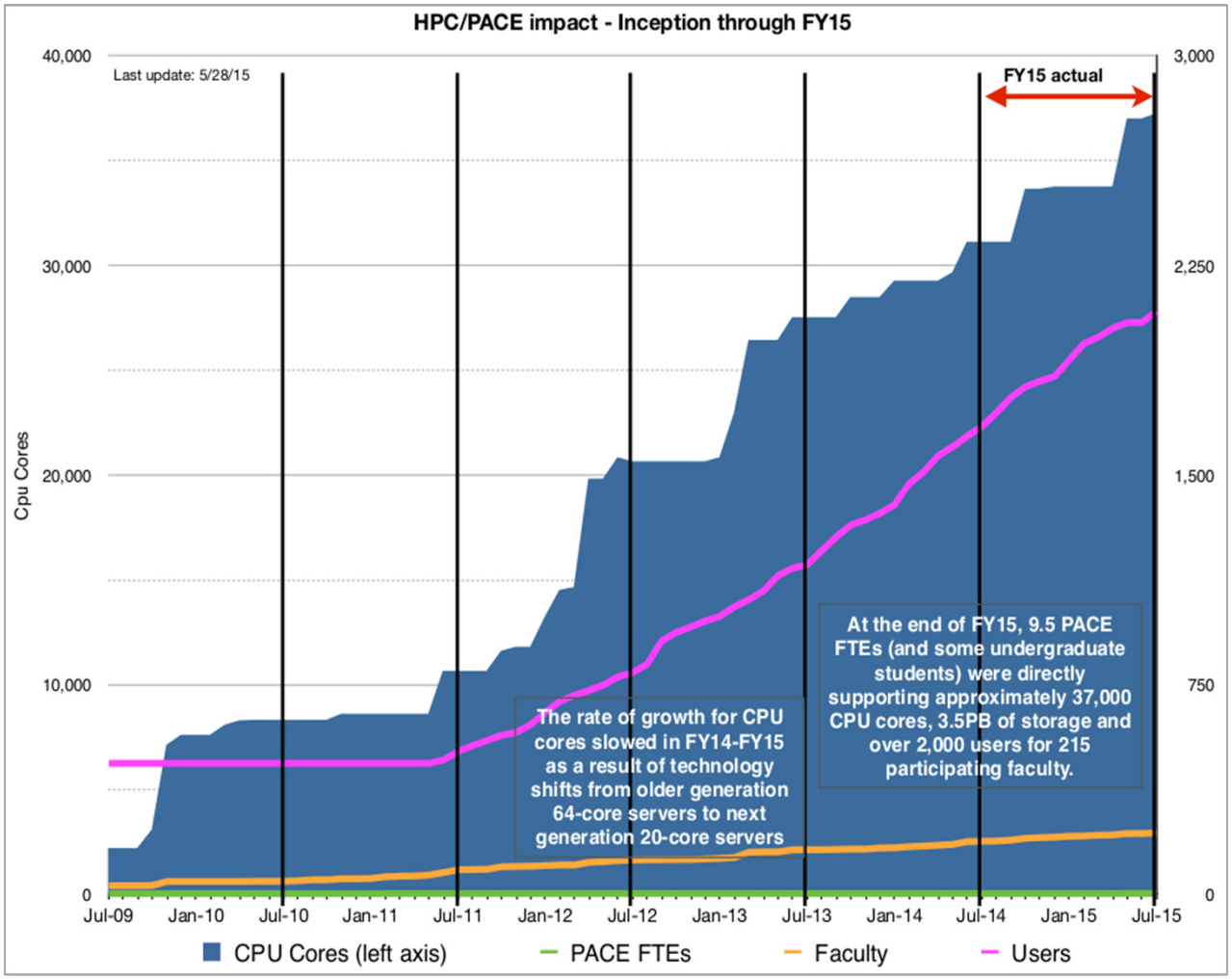
Centralized HPC services for federated clusters

consists of:

- 11 active members (incl. 3 research scientists)
- 3 student assistants

PACE Structure

PACE



- > 2000 users (~1700 active)
- 215 participating faculty (PIs)
- > 100 “queues”
- 37k cores, most with QDR IB, but not all
- 3.5 PB of storage
- Total 9000 ft sq datacenter(s)
- 100 Gb/sec to Internet2 AL2S

MVAPICH2 @ PACE

- First encounter: mvapich2/1.4.1, May 2010
  (end of mpich2 for us)

- PACE software repo (2011-2015)
  mvapich2/1.6, 1.7, 1.8, 1.9, 2.0

- First encounter with the MVAPICH2 Team (Sep 2011)
  - mvapich2/1.6 not working for > 64 cores (reg cache issue)
  - received a workaround the next day!

- Another crisis (June, 2013)
  - mvapich2/1.6 & 1.7 hanging for a user, critical simulations in danger
  - workaround in 3 days! (unset MALLOC_PERTURB_)
  - a patch in 2 weeks
  - official integration in mvapich2/1.9a

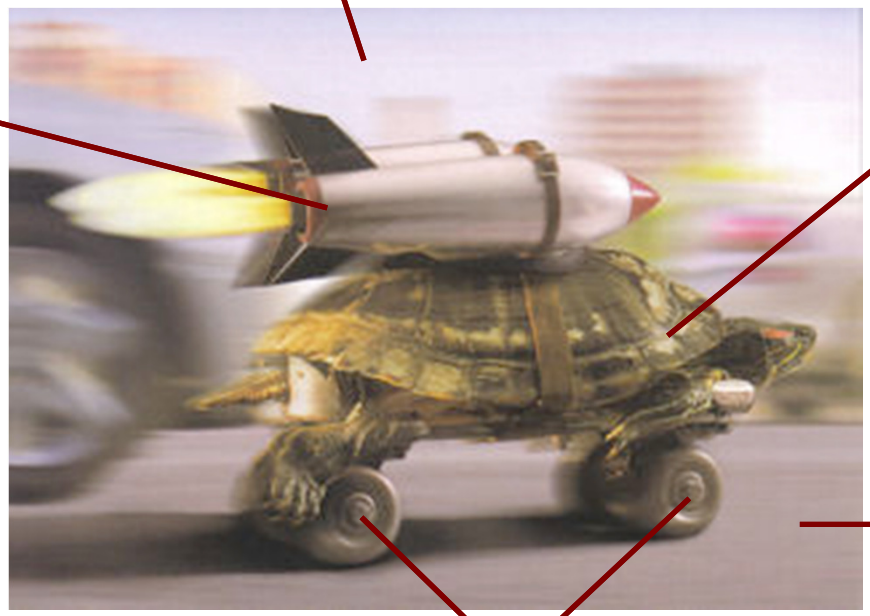
- New PACE software repo (2015-)
  mvapich2/1.9, 2.0, 2.1, …

MVAPICH2: powerful but familiar

Georgia Tech

Same world
(std OS, OFED, compilers)

Turbo Boost!
(MVAPICH2)



Same animal
(no code changes)

Existing
Infrastructure
(IB)

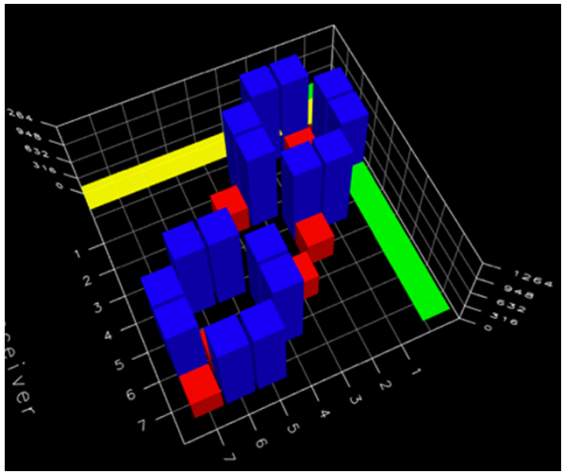
Familiar Technology
(MPICH)

MVAPICH2 provides superior performance without changing your world

MVAPICH2 for sysadmins

Georgia Tech

- Acceptance testing: 10-days of “uninterrupted” runs with mvapich2 compiled:
  - VASP (the “node killer” case!)
  - LAMMPS
  - HPL
  - SPEC2007 (will be added soon)



- High compilation success rate with MPI packages

- Node/IB fabric health analysis: p2p OSU benchmarks
  - Bandwidth and latency
  - A wrapper script to submit one-to-all jobs and analyze data
  - A summary to report slow paths with std deviations

- Excellent Compatibility with debuggers/profilers
  - Valgrind (compiled with MPI wrappers)
  - TAU
  - Allinea DDT (debugger) and MAP (profiler)

Tuning and Analysis Utilities

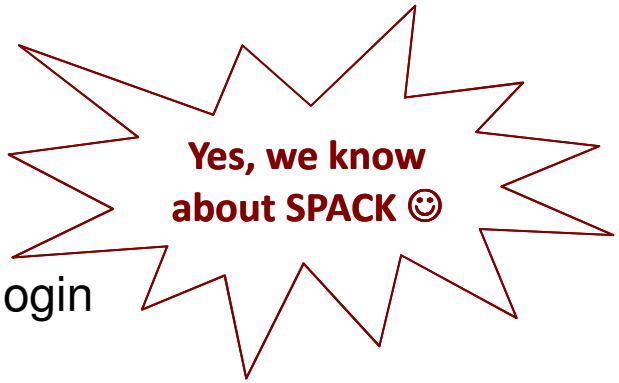
allinea

PAPI

Valgrind

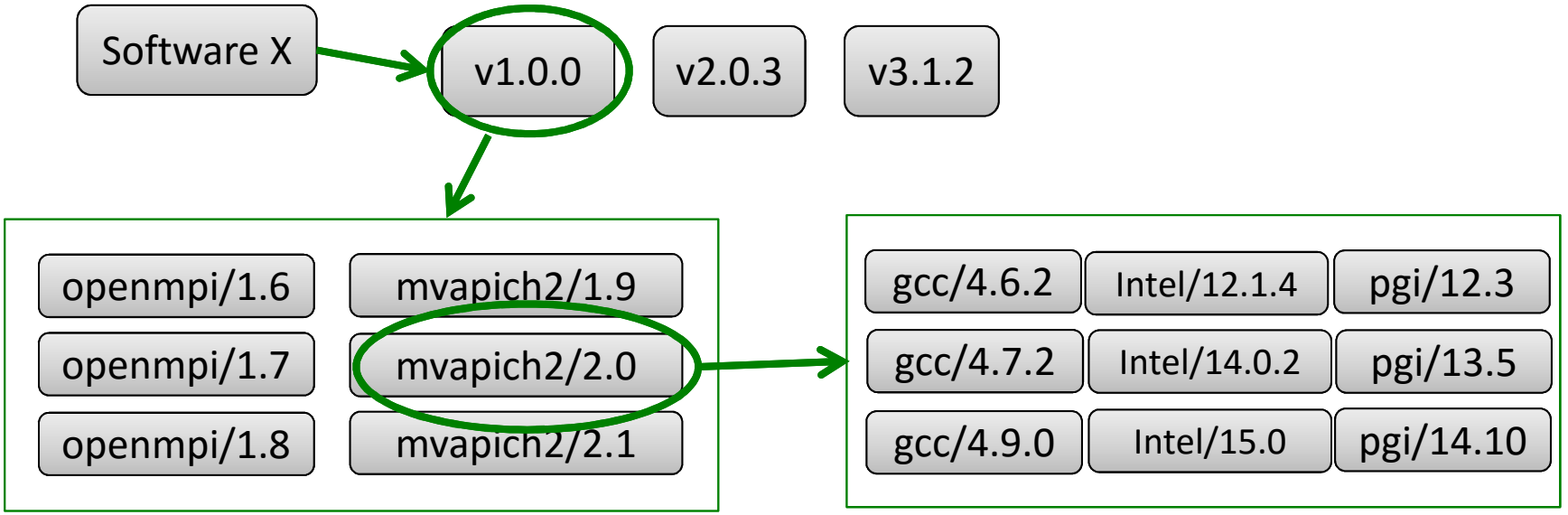
PACE software repository

- 420 packages, over 1TB
  - 54 MPI packages with mvapich2
  - 49 MPI packages with openmpi



- 576 of ~2000 users choose to load an MPI module on login
  - Mvapich2: 504
  - OpenMPI: 72 (mostly from a non-IB cluster)

- Hierarchical format for all version/MPI/compiler combinations (as possible)



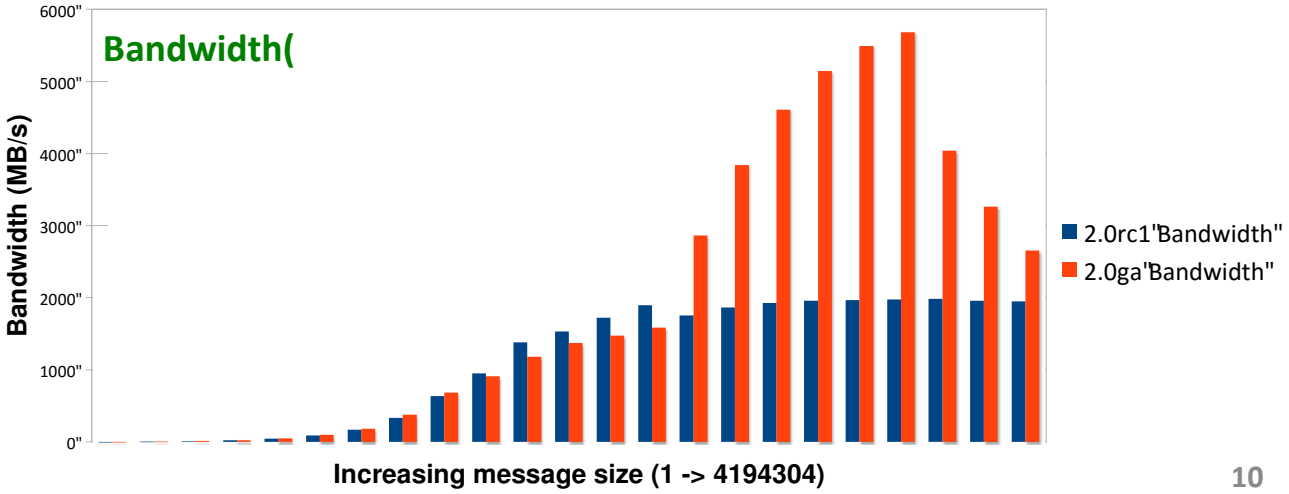
Getting better every day

- 2.0rc1 vs. 2.0ga (rc2?) (available in 2.0rc1 but not default)
- Improved intra-node communication performance using Shared memory and Cross Memory Attach (CMA)

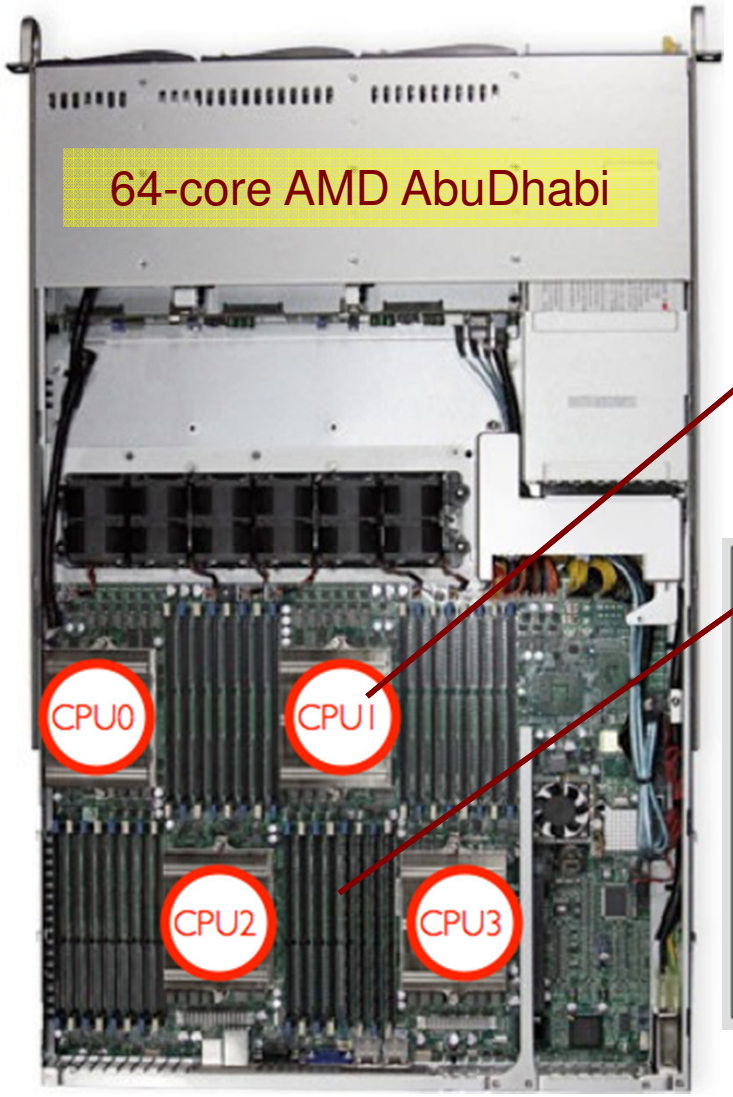
- p2p OSU benchmarks
- 64-core AMD node

XSEDE'14 article by Jerome Vienne

"Benefits of Cross Memory Attach for MPI libraries on HPC Clusters"

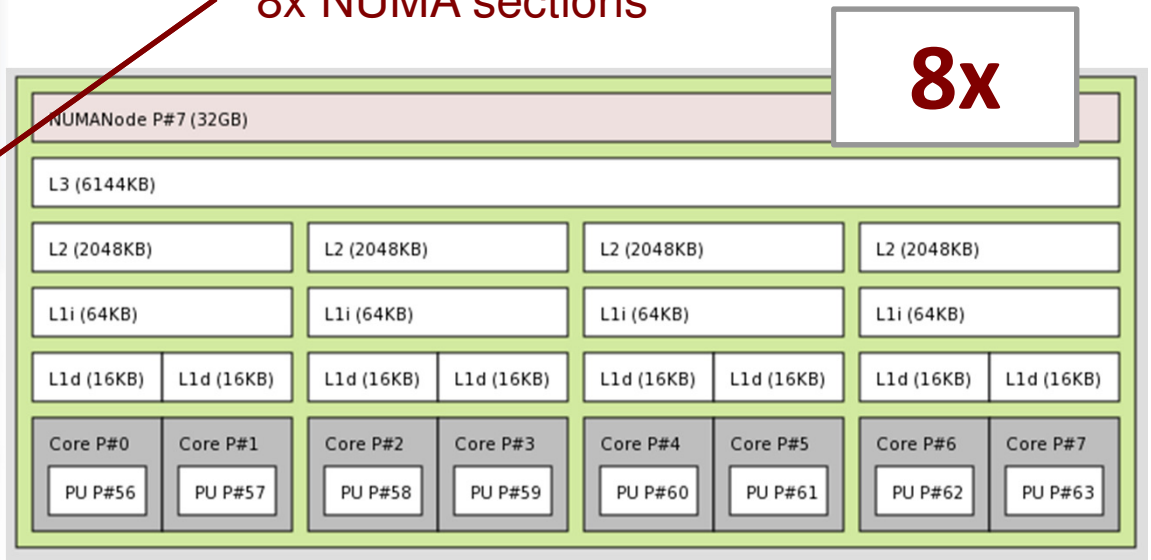


Challenges in multicore performance



Each Processor has
16 cores!

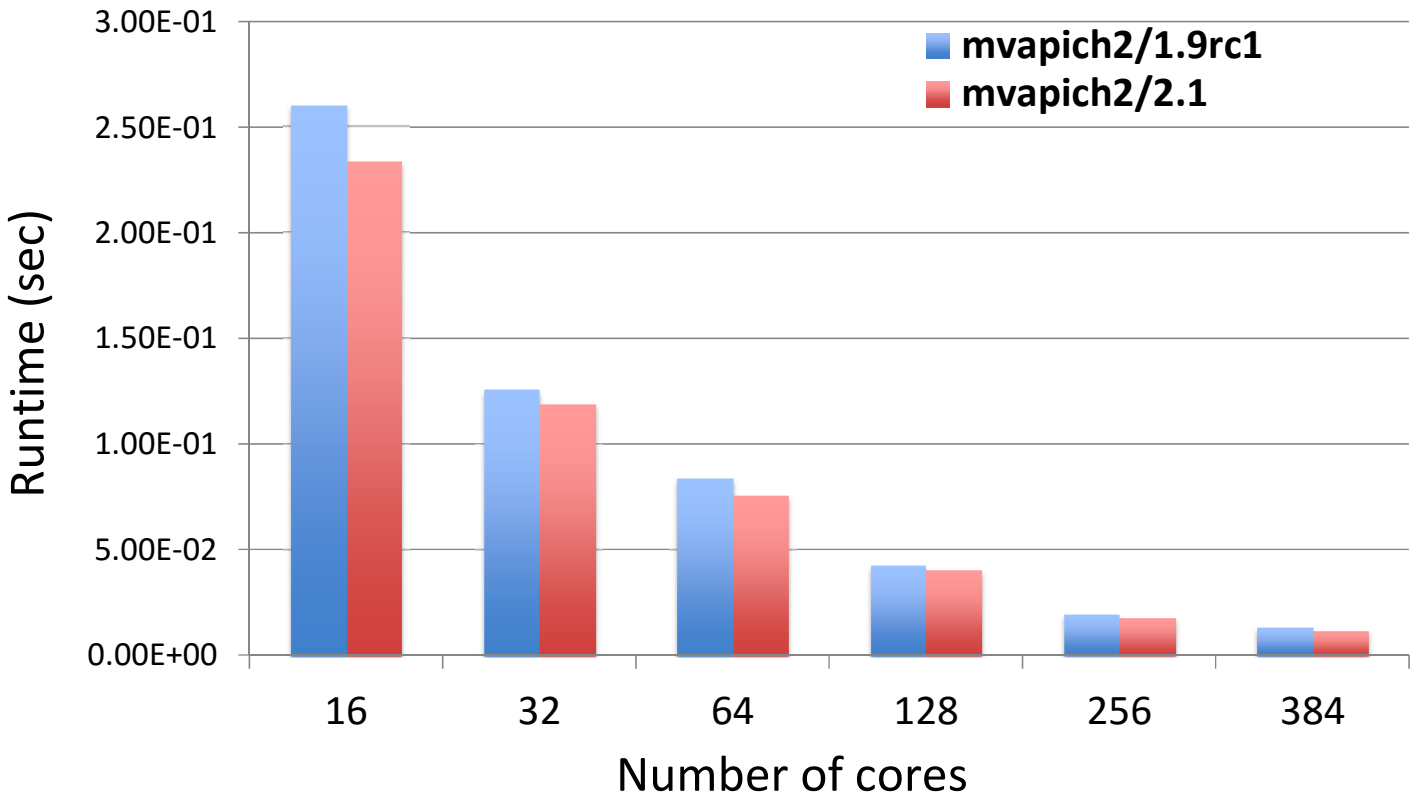
4x sockets,
8x NUMA sections



(hwloc 'lstopo')

Improved overall performance

- Leslie 3d from SPEC2007 benchmark, 128cube case (https://www.spec.org/mpi2007/)
- ~10% consistent performance improvement on average since 1.9rc1
- 195 QDR connected 16-core Intel sandybridge nodes, with 64GB memory
- 10% of a $1.2 million cluster is…

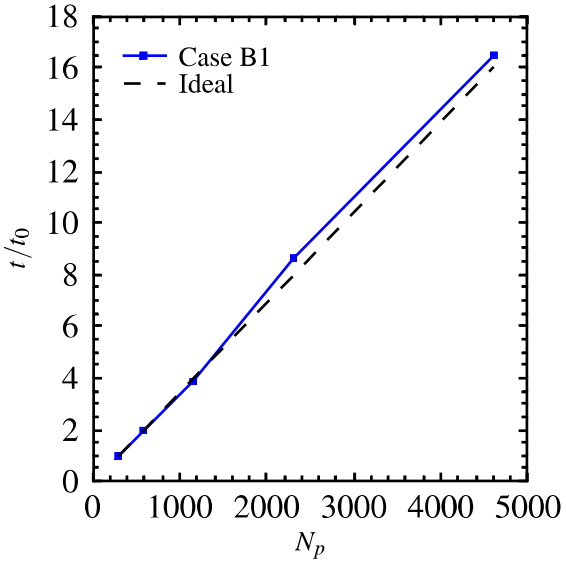


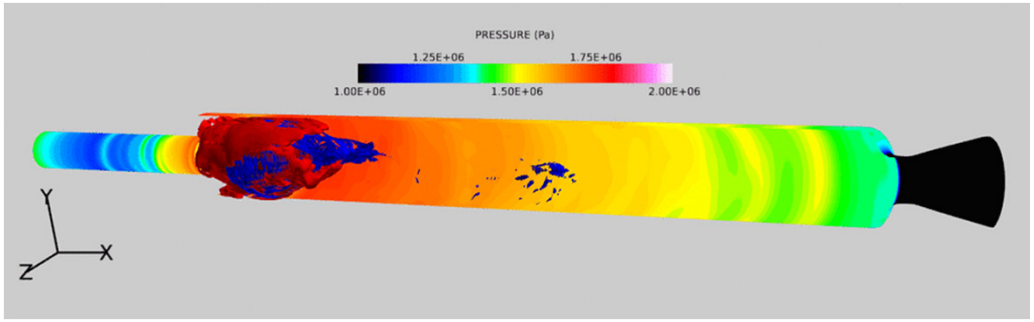
Impact on Research: Leslie

Prof. Suresh Menon's Computational Combustion Lab @ GT

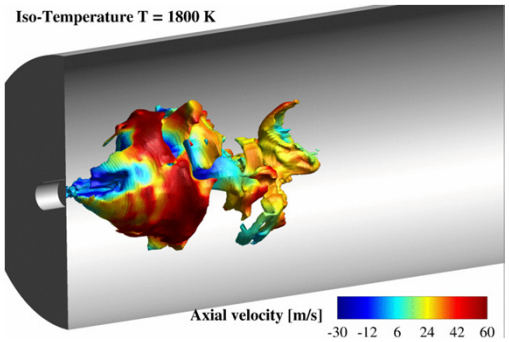
- **LESLIE** is a three-dimensional, parallel, multi-block, structured, finite-volume, compressible flow solver with multiphysics capability.

- It has been used to study wide variety of flow systems such as canonical turbulent flames, thermo-acoustic combustion instability, swirl spray combustion, real-gas systems, MHD flows etc.





Combustion instability in model high-pressure rocket combustor

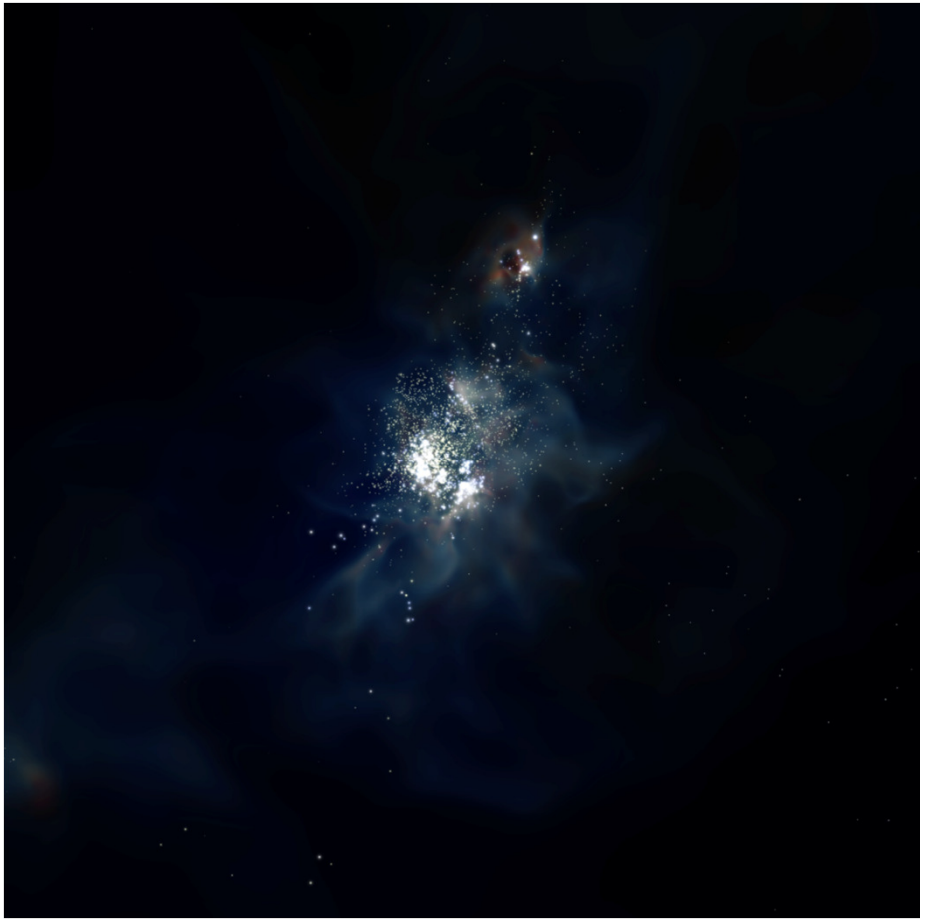


Swirl spray combusion:
Evolution of flame surface

Impact on Research: Enzo

The Enzo Project: Prof. John Wise, Center for Relativistic Astrophysics @ GT

- One of the lead developers of publicly-available and open-source Enzo (http://enzo-project.org/)
- Simulations of early star and galaxy formation that include hydrodynamics, gravity, chemical networks, magnetic fields, and radiation transport.
- Interpreting observations of the farthest galaxies and to understand how galaxies form over cosmic time.

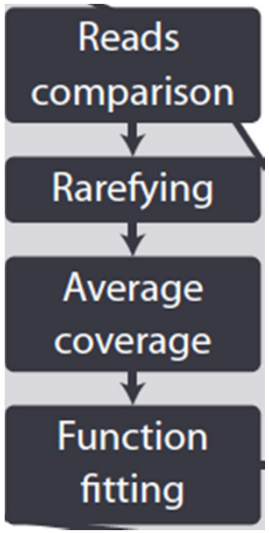


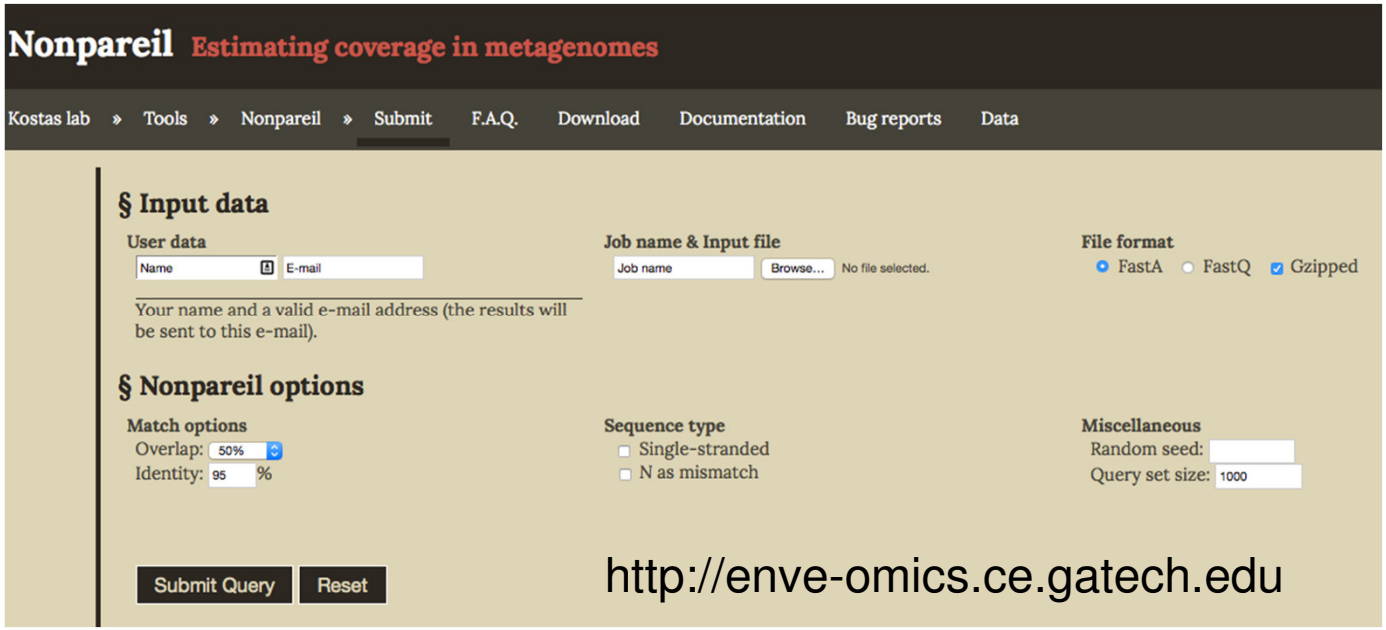
Close up of a young "dwarf" galaxy produced as part of simulation (SDSC)*

*** Also killer of black toners, do not print out this slide ☺**

Impact on Research: Nonpareil

Prof. Kostas Konstantinidis: Environmental Microbial Genomics Lab @ GT



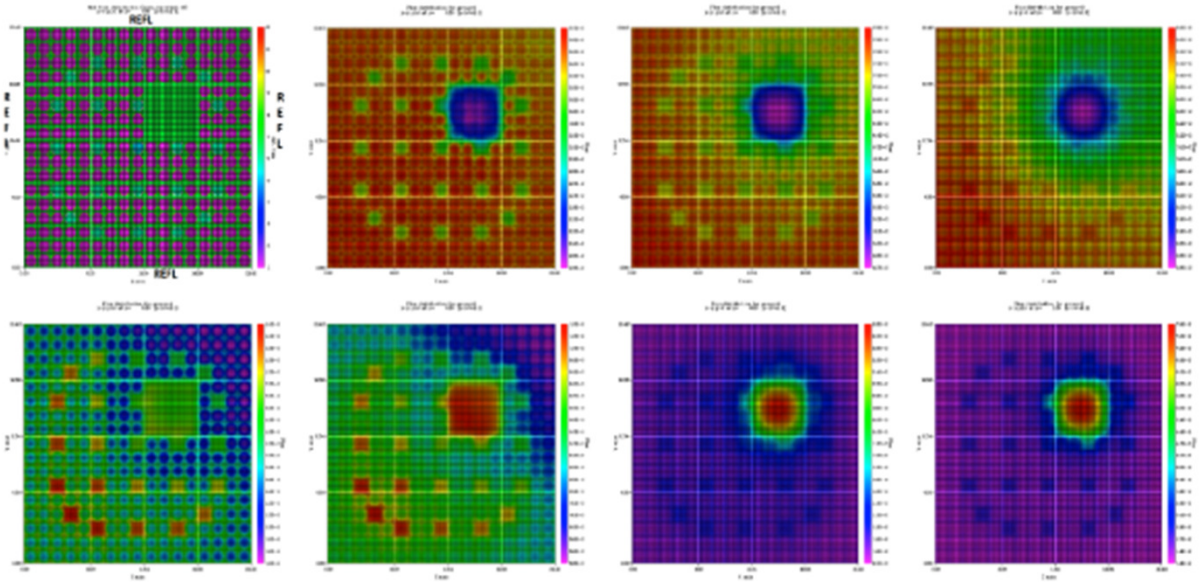


- Developing bioinformatics algorithms and tools to analyze genomic and metagenomic data from microbiome project. For instance, our tools are applied to the Human Microbiome Project to identify how the gut microbial community cause disease vs. healthy state.

- Nonpareil uses the redundancy of the reads in a metagenomic dataset to estimate the average coverage and predict the amount of sequences that will be required to achieve "nearly complete coverage", defined as ≥95% or ≥99% average coverage.

Impact on Research: Pentran

Prof. Glenn Sjoden: Chief Scientist, Air Force Technical Applications Center
Former Director, Radiological Science and Engineering Laboratory @GT



Top left: Water Hole pressurized water reactor model.

Others: Flux from high energy (red) to low energy (purple)

- Pentran: 3D Parallel deterministic radiation transport code
- Phase space decomposition with 3D topology in MPI in angle/direction, energy, and space, with further angular refinement inside each MPI task with OpenMP threading.

Today

Busted Myths

- MPI will have no place in Exascale world
- Mvapich2 is IB dependent (not-so-good for cloud)

Known issues

- Affinity problems with cpusets
- Mpi4py incompatibility

Wishlist

- Ability to run seamlessly on non-IB networks
- A framework to analyze and publish OSU benchmark results

  => INAM!! ☺
- Download links for old versions

Thank You!