SAND2008-7395C

HPC Application Performance Analysis and Prediction

(Mantevo Project: software.sandia.gov/mantevo**)**

NNSA/ASC @ Supercomputing'08

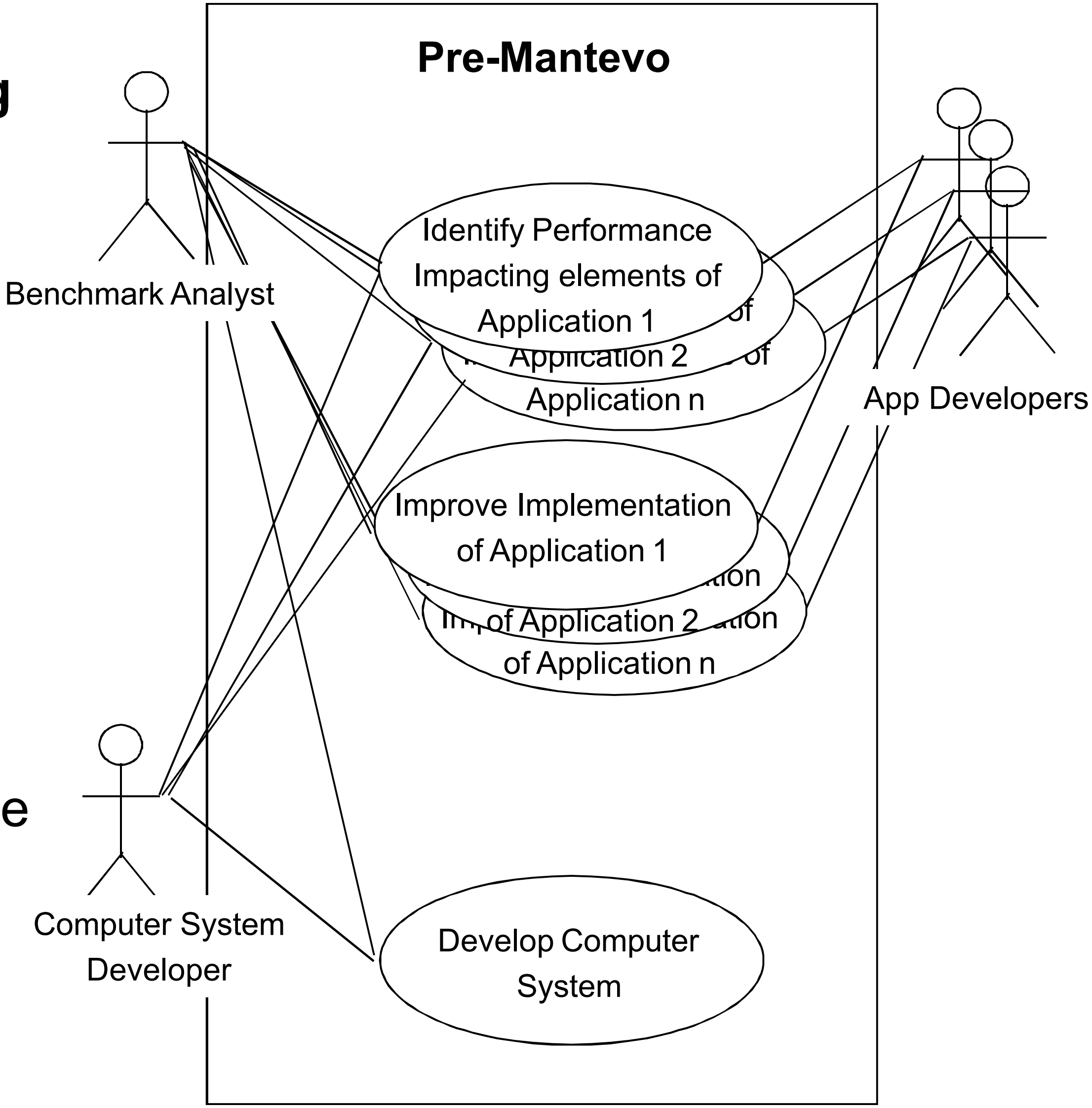
**Michael A. Heroux (project lead),
H. Carter Edwards (presenter),
Paul S. Crozier, and Alan Williams**

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

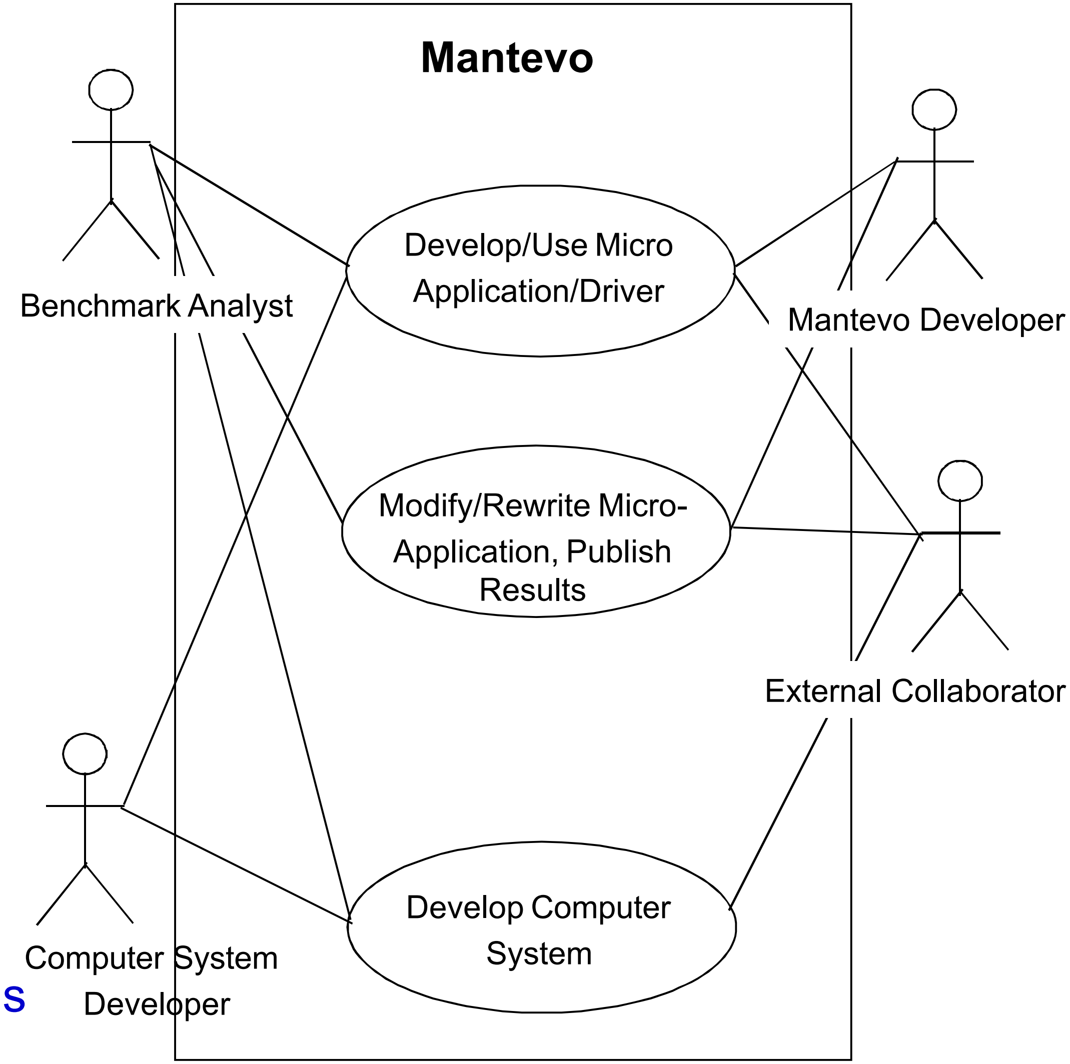
Background

- **Develop scalable computing capabilities via:**
  - Application analysis
  - Application improvement
  - Computer system design
- **Schedule driven**
- **Countless design decisions**
- **Collaborative efforts**
- **Pre-Mantevo:**
  - Work with each, large application
  - Application developers have competing needs:
    - Features
    - Performance
  - Application performance profiles have similarities



Mantevo* Project

- **Develop micro apps and micro drivers**
- **Aid system design decisions**
  - Proxies for real apps
  - Easy to use, modify, or rewrite
  - e.g., multicore studies
- **Guide application and library developers**
  - Early results in new situations: apps/libs know what to expect
  - Explore new programming models and algorithms
  - Predict performance of real applications in new situations
  - New collaborations
- **Results:**
  - Better-informed design decisions
  - Broad dissemination of optimization techniques
  - Incorporation of R&D results

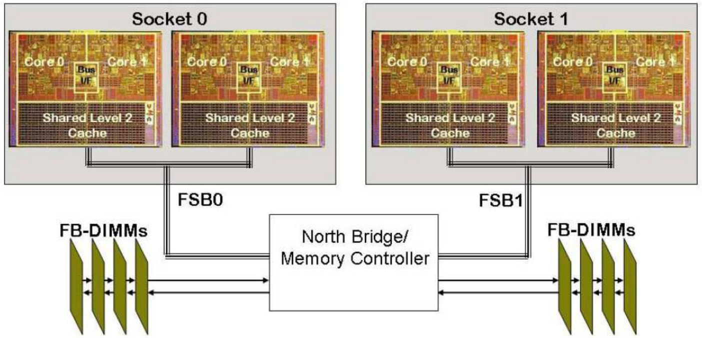

Mantevo
Benchmark Analyst
Develop/Use Micro Application/Driver
Mantevo Developer
Modify/Rewrite Micro-Application, Publish Results
External Collaborator
Computer System Developer
Develop Computer System

*** Greek: augur, guess, predict, presage**

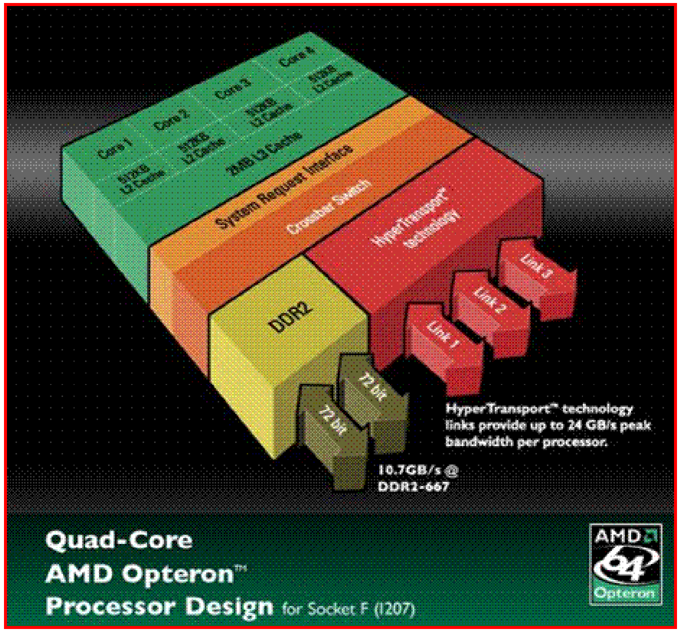
Key Focus Area:
Multicore Node Architectures

- **Multicore:**
  - New HPC systems axis
  - First Mantevo analysis focus
- **Quantitative results:**
  - Confirm, sharpen intuitive sense
  - Sometime counter intuition

Intel Clovertown



AMD Barcelona

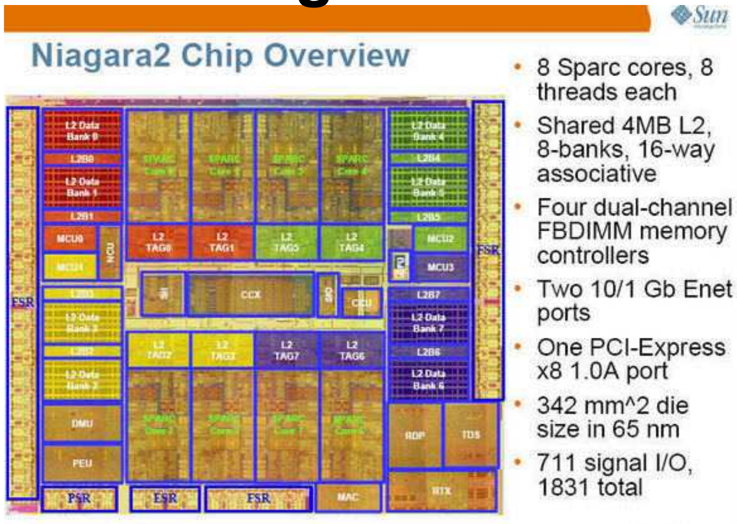


HyperTransport™ technology links provide up to 24 GB/s peak bandwidth per processor.

10.7GB/s @ DDR2-667

Quad-Core AMD Opteron™ Processor Design for Socket F (1207)

Sun Niagara2



Niagara2 Chip Overview

- 8 Sparc cores, 8 threads each
- Shared 4MB L2, 8-banks, 16-way associative
- Four dual-channel FBDIMM memory controllers
- Two 10/1 Gb Enet ports
- One PCI-Express x8 1.0A port
- 342 mm^2 die size in 65 nm
- 711 signal I/O, 1831 total

Sandia National Laboratories

Mantevo Microapps / Microdrivers

- Three types of packages:
  - **Microapps**: Small, self-contained programs
    - **HPCCG**: Implicit solution of unstructured FEM/FVM
    - **pHPCCG:** HPCCG with parameterized scalar/int, replaceable SpMV kernel
    - **miniMD**: molecular dynamics parameterized from simple to bio molecules
    - **phdMesh**: explicit FEM with contact detection
  - **Microdrivers**: Wrappers around Trilinos packages
    - **Beam**: Intrepid+FEI+Trilinos solvers
    - **Epetra Benchmark Tests**: Core Epetra kernels
  - **Motif framework**: Collection of “dwarves”
    - **Prolego**: Parameterized, composable fragment collection to mimic real apps
- Developed by application and library developers
- Open Source: software.sandia.gov/mantevo

Microapp: HPCCG/pHPCCG

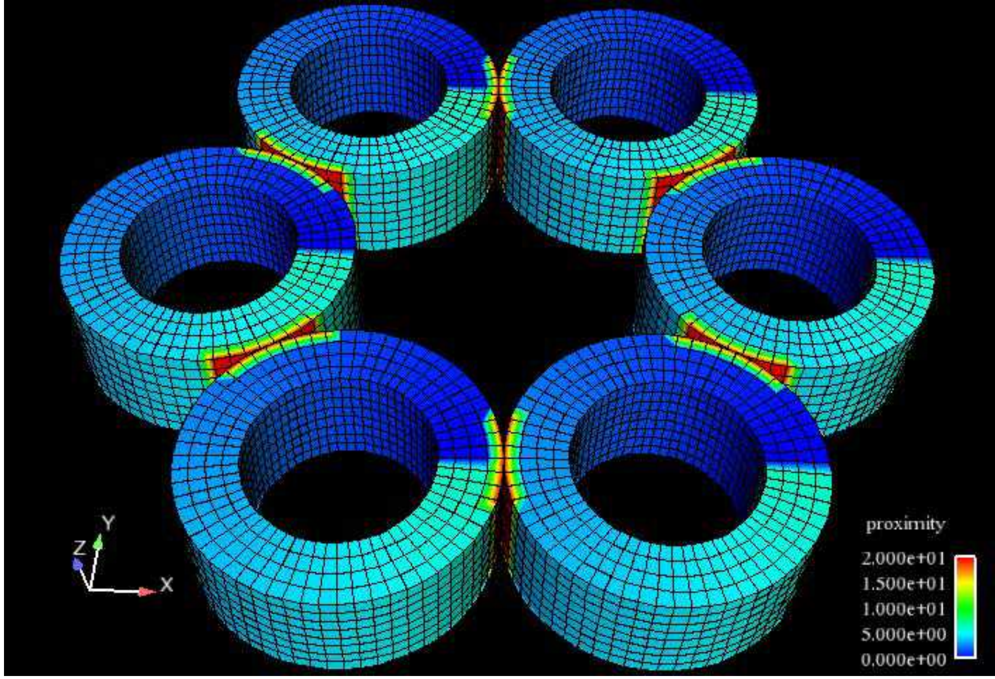
- **HPCCG: "Closest thing to an implicit unstructured FEM/FVM code in 500 semi-colons or less."**
  - **Simple application-like sparse matrix fill and solve**
  - **Compact, highly portable, and scalable**
  - **Baselined for MPI parallelism**
  - **Available as Open Source (LGPL License)**
- **Used in many early scalability and performance studies**
  - **ASC RedStorm, ASC Purple, SNL Thunderbird scalability**
  - **MPI-on-multicore studies**
  - **Several re-writes for parallelism comparisons:**
    - **Q-threads (massively threaded)**
    - **Bundle-Exchange-Compute (BEC)**
  - **Planned advanced-node studies**
    - **Cell Broadband Engine, Intel SSE, Woodcrest 128-bit architecture**
- **pHPCCG:** parameterized scalar/int types and replaceable SpMV

Microapp: miniMD

- Extracted computational core of LAMMPS, a scalable molecular dynamics simulation code
- Simulate O(10) to O(100) of atomic interactions
- Extreme scalability (10K atoms on 10K processors) is especially interesting (important science problem)
- Single precision
- Investigate novel architectures of interest
  - nVidia Tesla

Miniapp: phdMesh

- Parallel Heterogeneous Dynamic unstructured Mesh
- Explicit unstructured FEM/FVM with dynamic load balancing and parallel geometric search

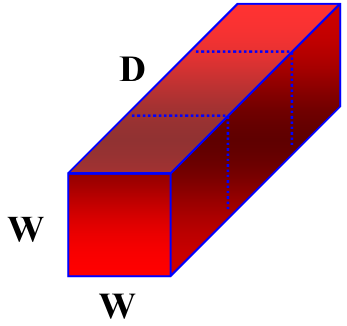


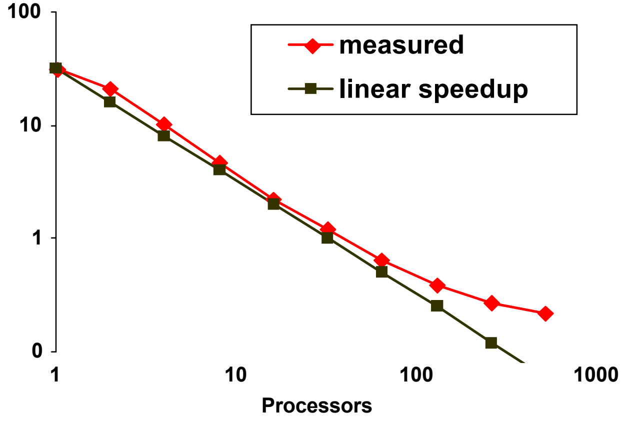
- Parallel geometric proximity search: a performance constraining algorithm for contact detection and multiphysics loose-coupling
- Dynamic mesh modification: a performance constraining capability for adaptive applications (e.g., load balancing, mesh refinement)
- Representative mini-application: Multiple 3D counter rotating “gears” with continually changing contact surfaces. Internal parallel generation of and domain decomposition of the meshed gears.
- phdMesh library: provides parallel, heterogeneous, and dynamic unstructured mesh and field data management

Sandia National Laboratories

Microdriver: Beam

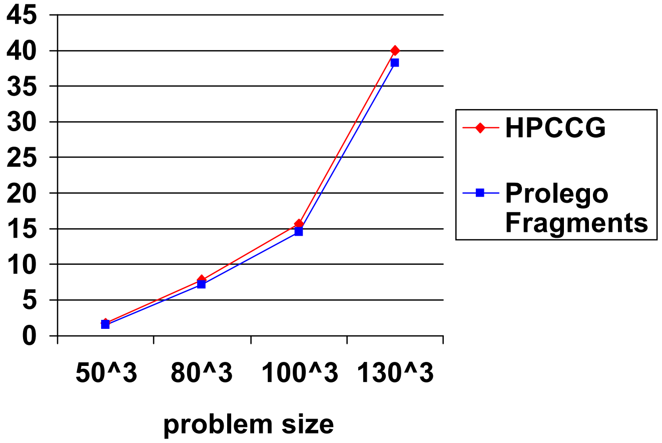
- Mimics important computational characteristics of implicit finite-element applications
- Heavily exercises Trilinos' (trilinos.sandia.gov) packages for filling and solving sparse linear systems of equations
- Scaled to 2 billion equations and 10k processors on ASC Red Storm
- Portable, scalable, and open source
- Representative mini-application: 3D beam of hexahedron elements with variable problem size/shape





Microdriver: Prolego

- **Configure a collection of computational kernels to model application performance**
- **Calibrate kernels to exhibit the performance characteristics of "real" application kernels**



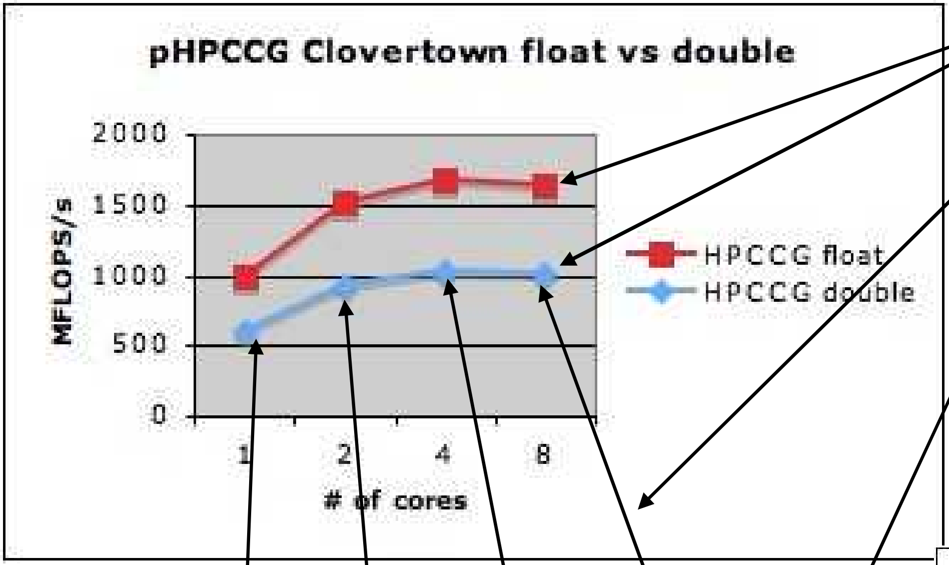
- **Current Prolego driver**
  - Run-time selection and calibration of kernels via XML input file
  - Initial kernels:
    - BLAS operations (vector axpy and dot, matrix-vector, matrix-matrix)
    - Sparse matrix-vector multiply
    - Binary-search operation, MPI operations (Allreduce, Barrier, Send/Irecv)

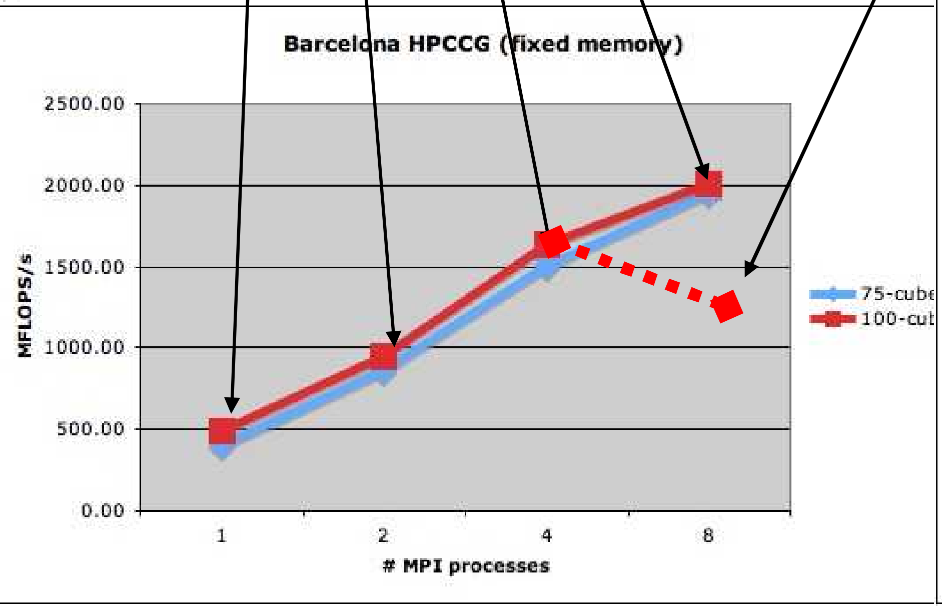
- **Planned:**
  - Finite-element oriented kernels
  - Input files calibrated to model Sandia applications
  - Comparison of modeled vs. actual performance

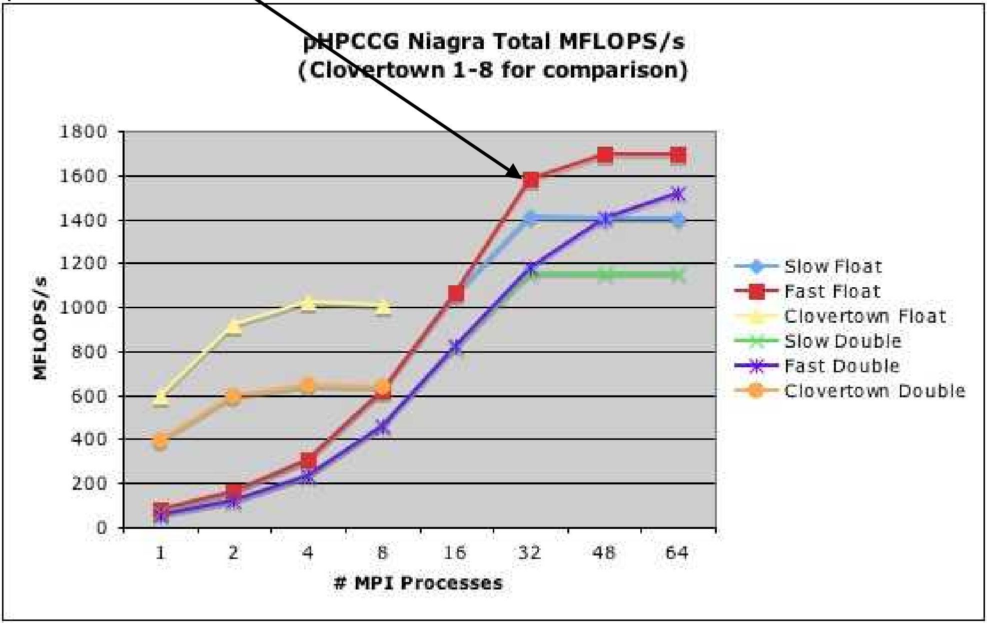
Some Performance Studies Using Mantevo Microapps

HPCCG Using MPI on Multicore Systems

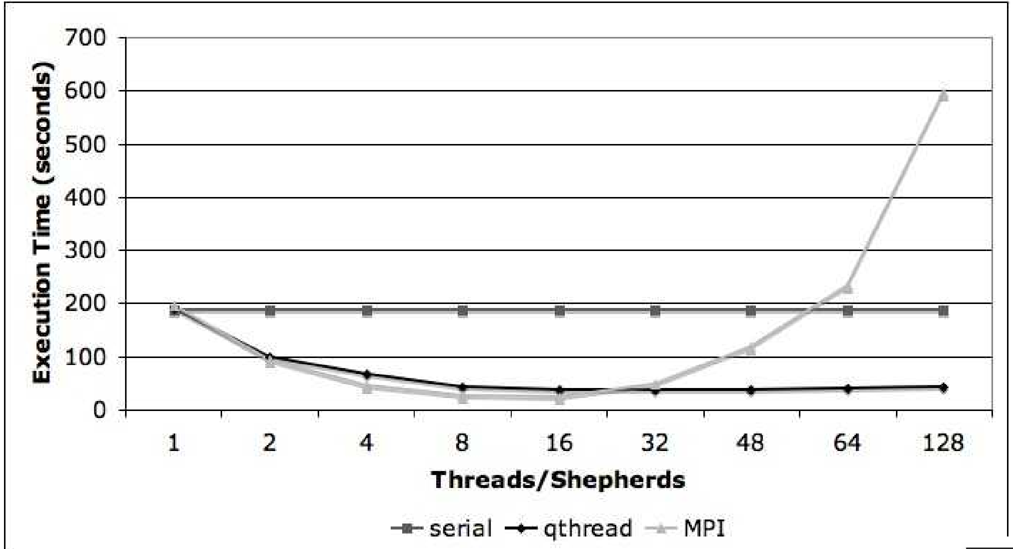
- **Float useful:**
  - Mixed precision algorithms.
- **Bandwidth even more important:**
  - Saturation means loss of cores.
- **Memory placement a concern:**
  - Shared memory allows remote placement.
- **NiagaraT2 threads hide latency:**
  - Easiest node to program.



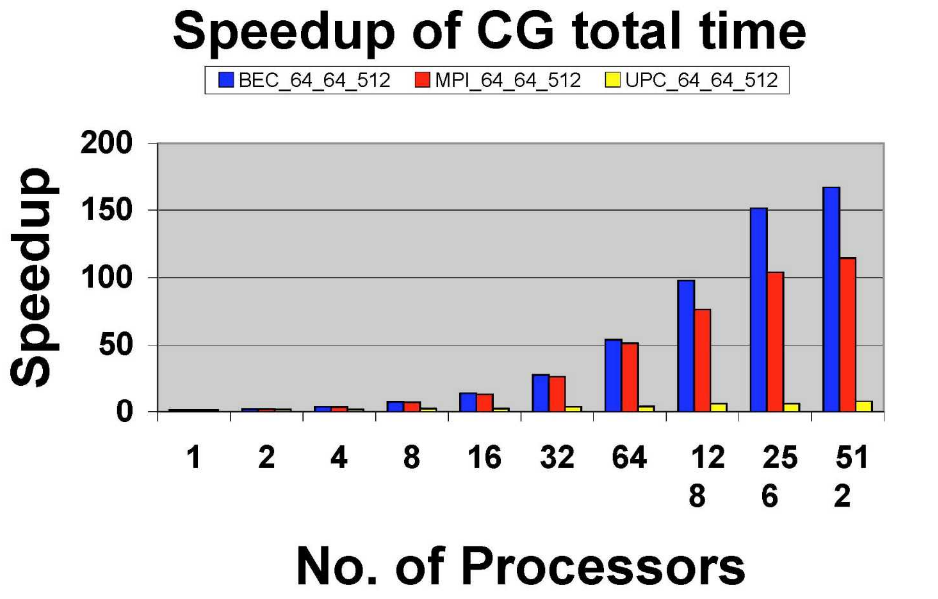


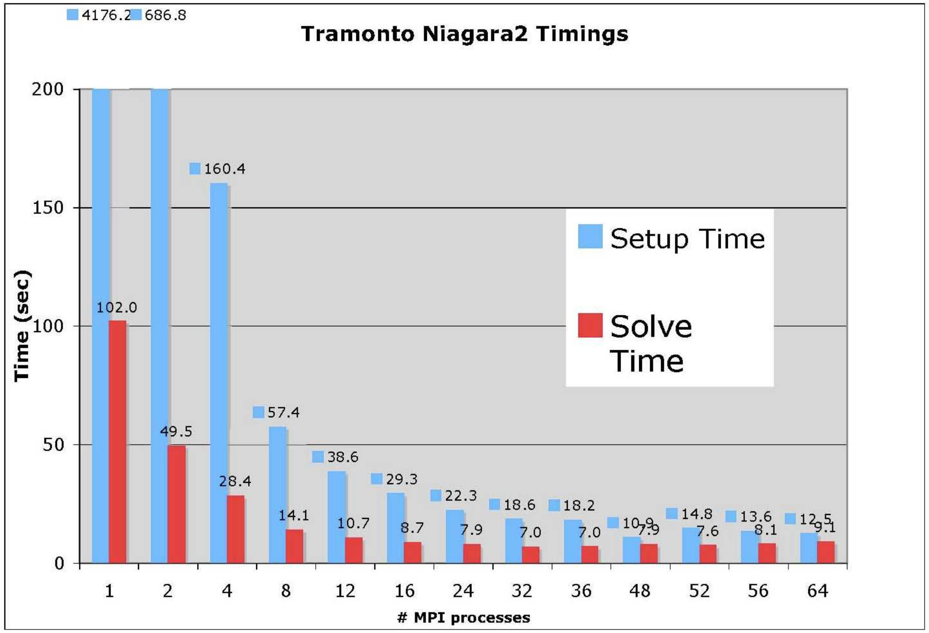


HPCCG Comparing Parallel Programming Models

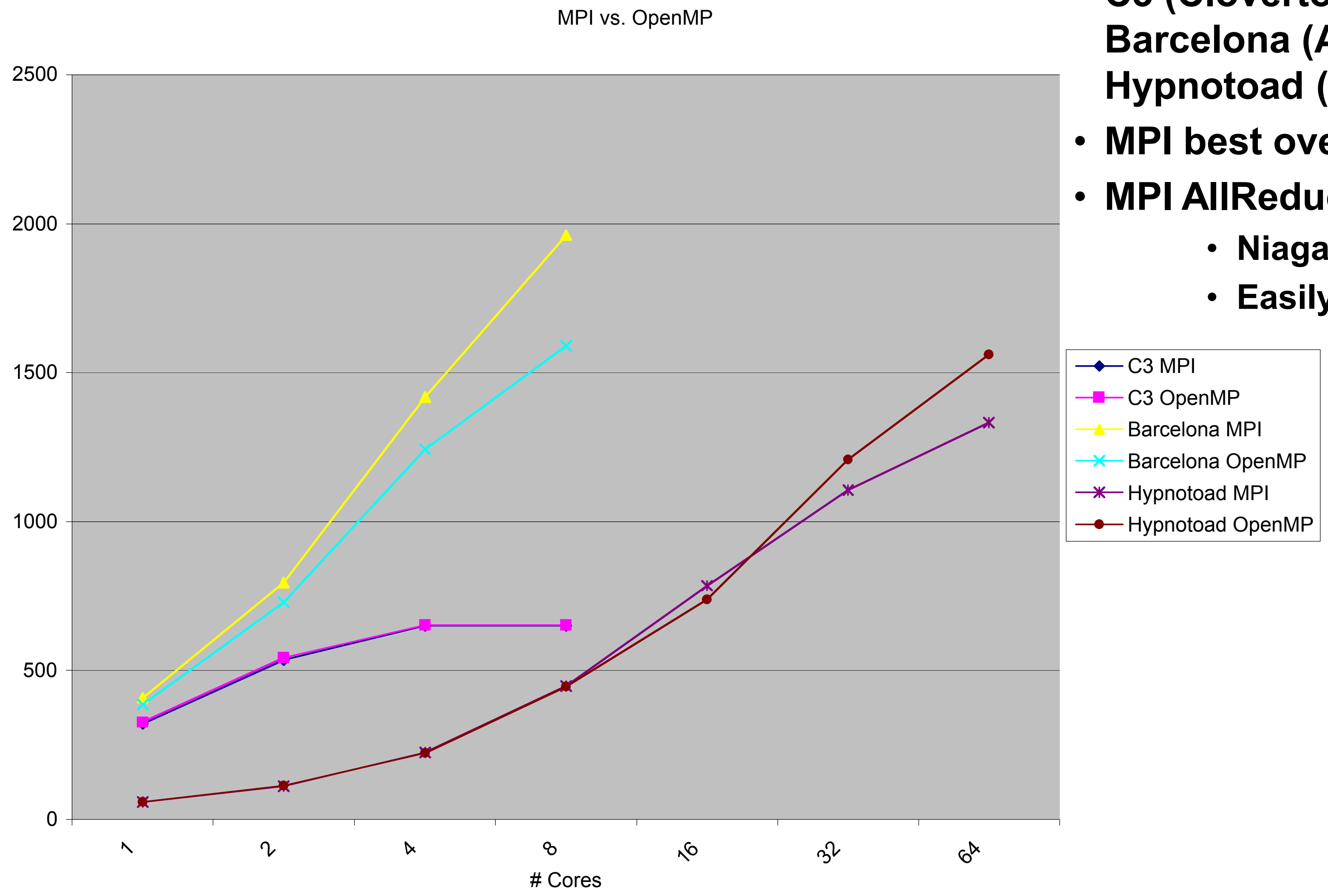


- HPCCG rewritten:
  - Qthreads: Massively threaded library.
  - BEC: Bundle-Exchange-Compute Model.
- MPI & MPI+threads.
  - App: MPI-only
  - Solver: MPI+threads





HPCCG Comparing Parallel Programming Models: MPI vs. OpenMP

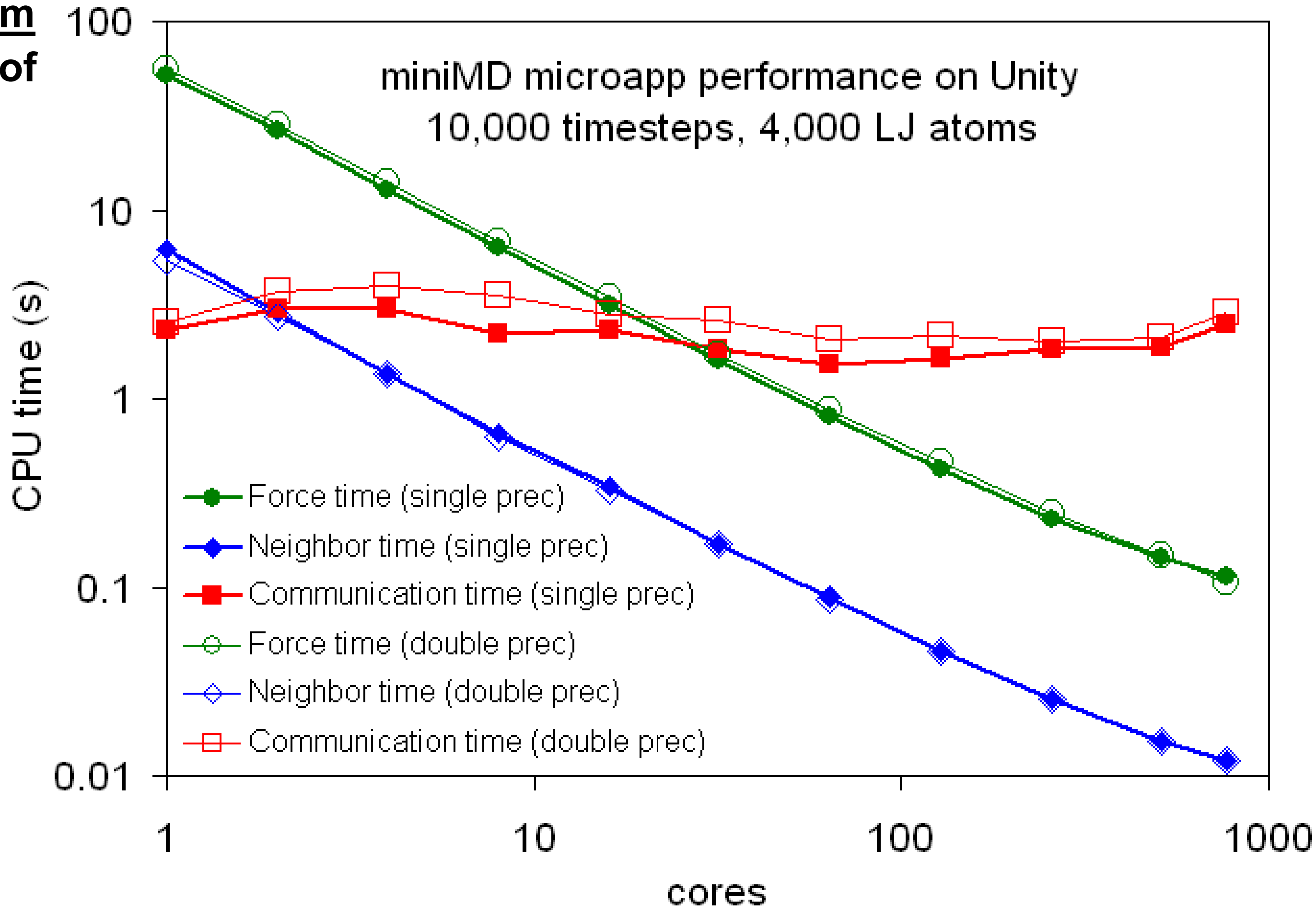


- C3 (Clovertown) Barcelona (AMD) Hypnotoad (Niagara2)
- MPI best overall
- MPI AllReduce Issues
  - Niagara (not shown).
  - Easily addressed.

miniMD Performance
(molecular dynamics microapp)

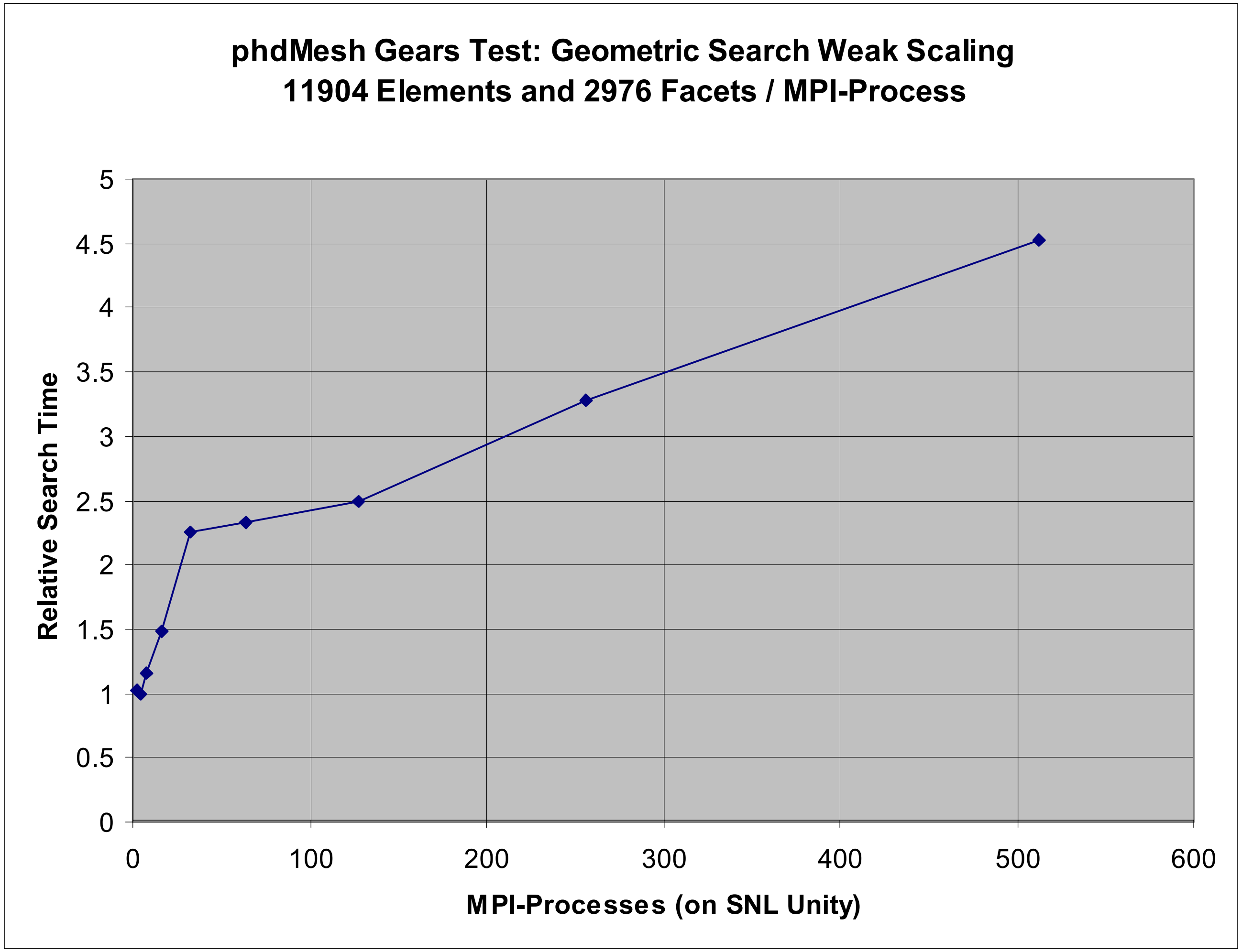
<u>**Unity: Sandia Lab System**</u>

Infiniband interconnect of 272 Nodes
x 4 sockets / node
x AMD Barcelona chips
= 4,352 cores



Pure-MPI parallel with one MPI process per core

phdMesh Gears Sample Problem
Distributed Parallel Geometric Proximity Search
Weak Scaling Study on Sandia’s Unity System



Geometric Proximity Search algorithms:

- **naively: O(N^2)**
- **practice: O(N*log(N))**
- **optimally: O(N)**

Predominant scenario: each facet is in proximity to relatively few other facets