Performance of Large-Scale Scientific Applications on the IBM ASCI Blue-Pacific System

Arthur A. Mirin
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

Highlights

- Accelerated Strategic Computing Initiative (ASCI)
- IBM Machine Configuration
- Programming and Performance Issues
- Programming Model Comparison
- Three-Dimensional Turbulence using SPPM
- Other IBM Applications
- Conclusions

Accelerated Strategic Computing Initiative (ASCI)

- Helps to maintain safety and reliability of nuclear stockpile in absence of nuclear testing, through leading-edge computing and simulation
- **One program, three laboratories**
  - Lawrence Livermore National Laboratory
  - Los Alamos National Laboratory
  - Sandia National Laboratories
- **ASCI Alliances**
  - California Institute of Technology
  - University of Chicago
  - University of Illinois
  - Stanford University
  - University of Utah

ASCI Computing Platforms

- **Blue–Pacific**: IBM SP
  Lawrence Livermore National Laboratory
  - distributed/shared, many nodes, few processors per node

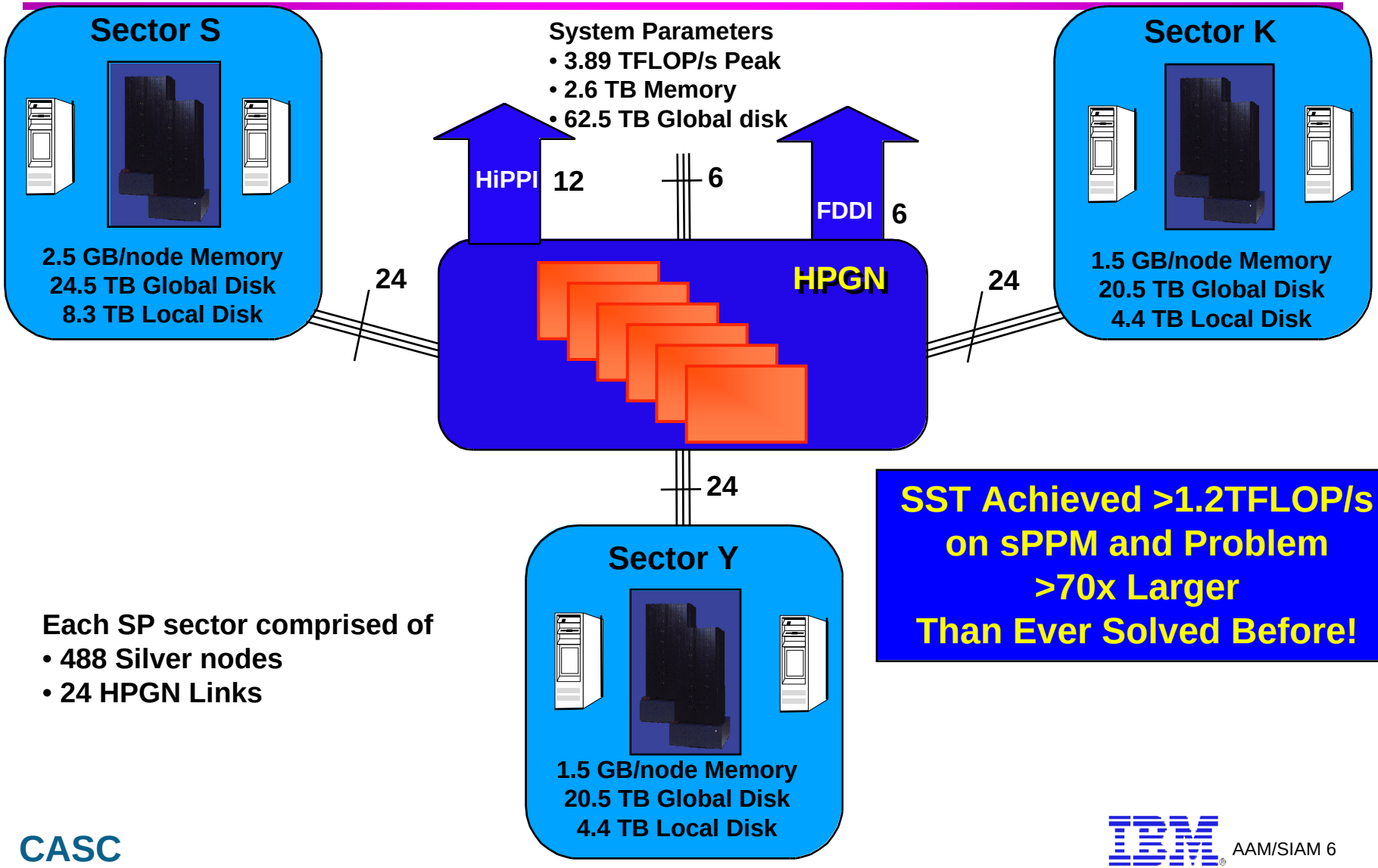
- **Blue Mountain**: SGI Origin
  Los Alamos National Laboratory
  - distributed/shared, few nodes, many processors per node

- **Red**: Intel Tflops
  Sandia National Laboratories
  - traditional massively parallel processor

IBM Sustained Stewardship TeraOPS (SST) System

- Three sectors, each comprising 488 "silver" nodes
- Each node is 4-way SMP based on 332 MHz PowerPC 604e, with 1.5–2.5 GB local memory, and a 9.1–18.2 GB local disk
- Peak throughput = 3.9 Tflops (aggregate)
- Processor-to-memory bandwidth = 2.1 Tbyte/s (aggregate)
- 62.5-Tbyte global disk
- IO to local disk bandwidth = 10.5 Gbyte/s (aggregate)
- Multi-level switching system (HPGNs connect sectors)
- Node-to-node bandwidth = 150 Mbyte/s (bi-directional)

SST Builds on TR Success

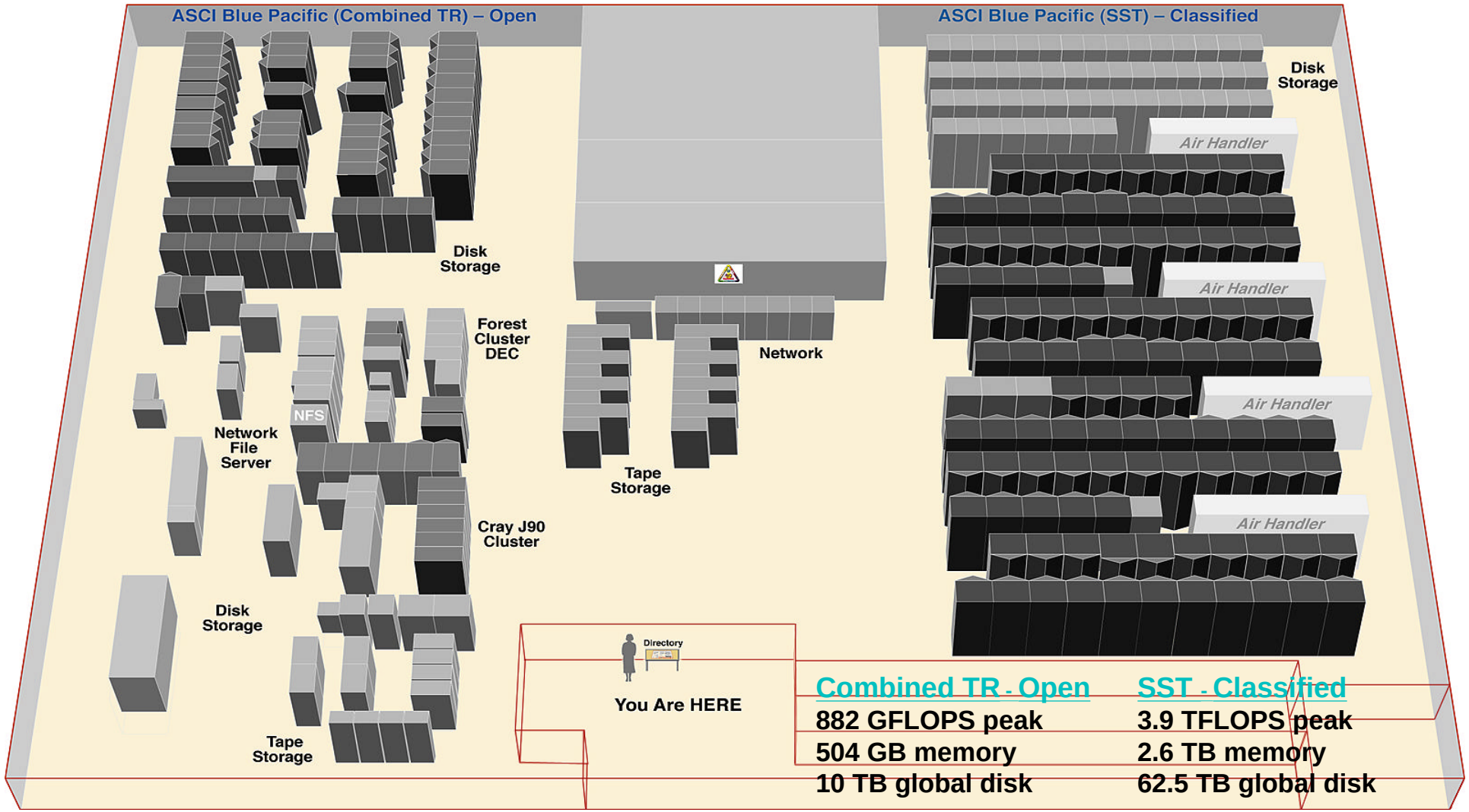


Each SP sector comprised of
- **488 Silver nodes**
- **24 HPGN Links**

SST Achieved >1.2TFLOP/s on sPPM and Problem >70x Larger Than Ever Solved Before!

Two-Stage SST Delivery in 1998: S&K on October 15, and Y on December 27

SST Sited at LLNL Six Months Early!



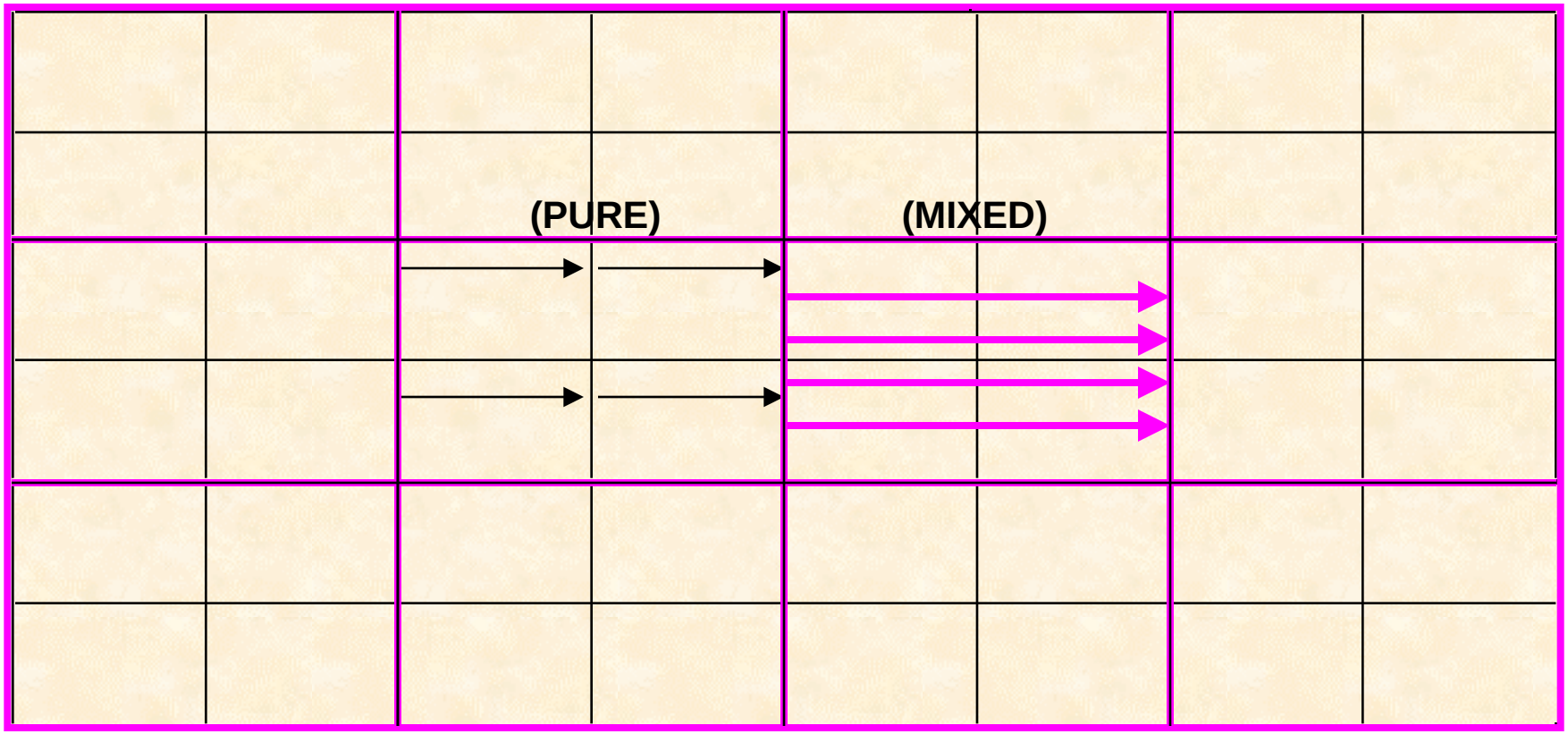
Programming Models

- **Mixed model**
  - distributed memory parallelism across nodes, shared memory parallelism on-node (MPI with OpenMP)
- **Pure message-passing**
  - one task per processor, pretend each processor has its own local memory (MPI with on-node emulation)

Programming Model Considerations

- Pure message passing is more simple.
- Mixed model offers greater parallelization and memory utilization freedom.
- Mixed model results in less communication overhead.
- Idle threads (mixed model) versus redundant computations (pure model).

Pure versus Mixed Model Domain Decomposition



IBM Communication Models

- **Internet Protocol (IP)**
  - involves operating system
- **User Space (US)**
  - direct communication with switch, bypasses operating system kernel
  - on-node communication goes through node adapter and does not take advantage of shared memory

Piecewise Parabolic Method (PPM) Code — Pure versus Mixed Programming

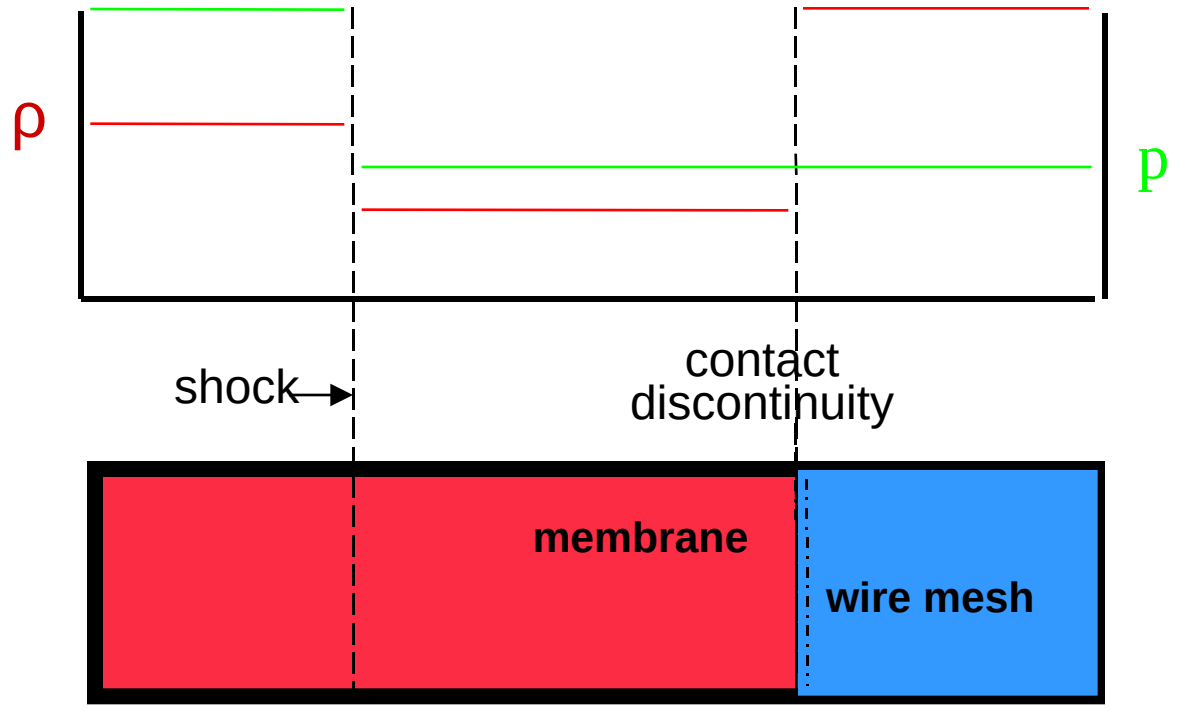
Baseline case: 256-cubed mesh, 16 nodes	980 seconds
OpenMP, single thread	1100 seconds (12% overhead)
OpenMP, 4 threads (guided)	340 seconds (2.9 speedup)
Pure MPI	290 seconds (3.4 speedup)
MPI across 64 nodes at 1 processor/node	260 seconds (12% overhead)

- Pure MPI runs 17% faster than OpenMP.
- Each has 12% overhead.

Simulations of Three-Dimensional Turbulence using SPPM Code

R.H. Cohen, B.C. Curtis, W.P. Dannevik, A. M. Dimits, D.E. Eliason, A.A. Mirin, S.E. Anderson, D.H. Porter, P.R. Woodward

RM Mixing Can Be Explored via Shock Tube Experiments



SPPM Code

- Simplified Piecewise Parabolic Method (Colella and Woodward)
  - Godunov method
  - Lagrangian plus remap (effectively Eulerian)
- Three-dimensional domain decomposition
- Posix threads plus MPI
- Fortran 77
- 32-bit arithmetic

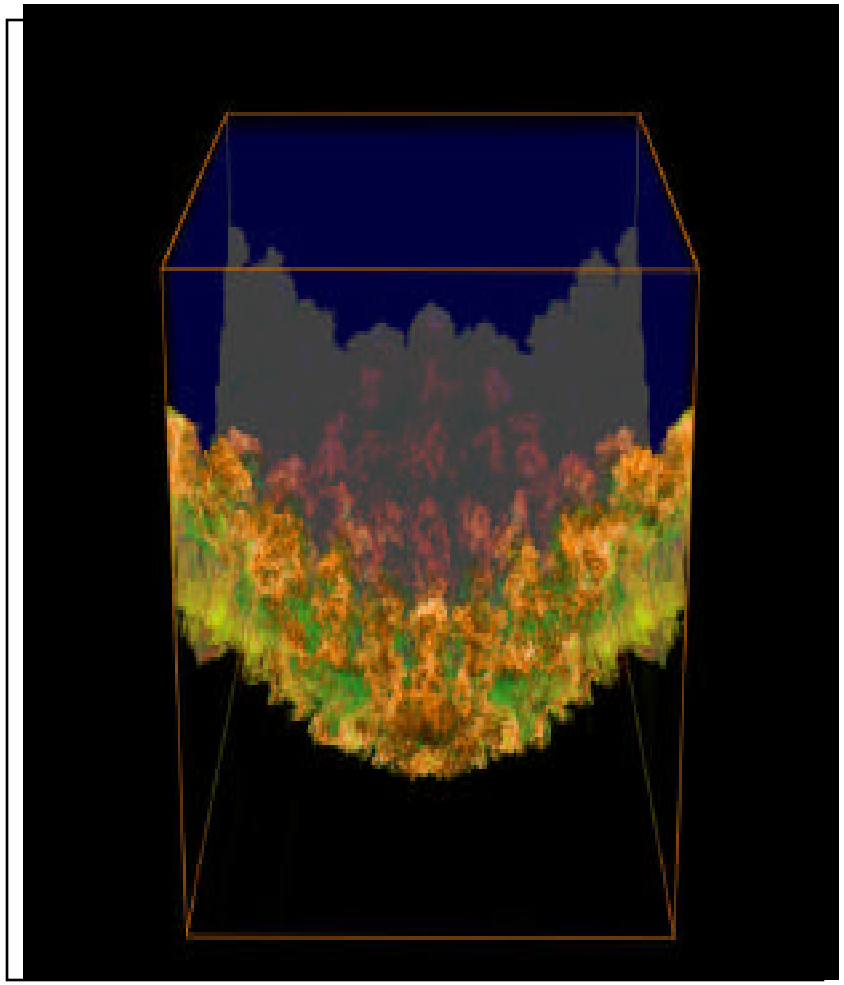
The SPPM Simulation on the IBM SST System

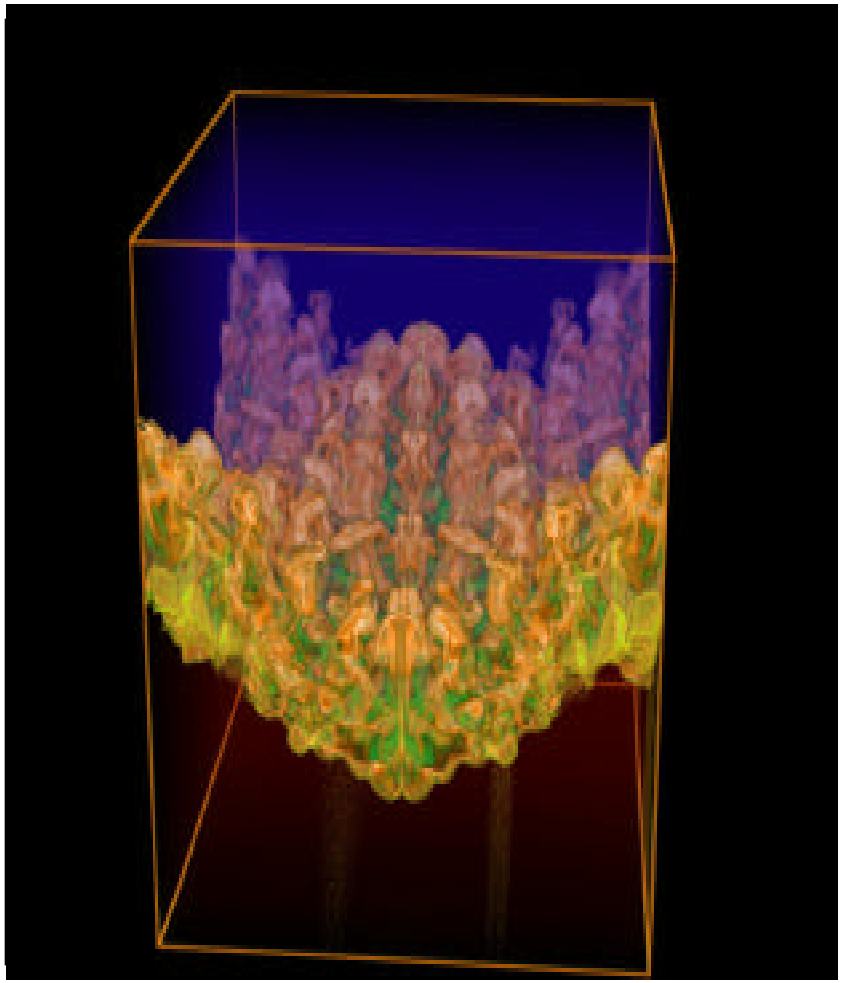
- 960 nodes of IBM SST
- 2048 X 2048 X 1920 mesh
- 8 X 8 X 15 domain decomposition
- 256 X 256 X 128 local mesh
- 27,000 timesteps
- 173 hours of full machine time, spread over 226 wall clock hours
- 129 MFlops (sustained) per processor
- 494 GFlops sustained throughput

Output Procedures and Statistics

- Restart dumps
  - 960 nodes X 196 MB = 188 GB
  - backup copy on neighboring node
- Bob dumps (movie frames)
  - 274 frames X 960 nodes X 8.4 MB = 2,210 GB
  - 10:1 compression results in 221 GB
- Compressed data dumps (16-bit integer)
  - 10 dumps X 960 nodes X 84 MB = 806 GB
- 275,000 files to store
- Data flow: local disk, to GPFS, to Riptide, to FAST storage

High-Resolution and Low-Resolution Volume Renderings





Expected Impact

- Explores nonlinear interactions between short and long wavelength energy transfer and resulting effects on mixing.
- Largest calculation of its type.
- High resolution allows capture of fine-scale physics, e.g., possible multiple transitions from coherent to turbulent states with increasing Reynolds number.
- Elucidates vital differences between 3-D and 2-D turbulence.
- Simulation diagnostics will provide tests of sub-grid scale parameterization model performance.

Ab Initio Simulations of Hydrogen Fluoride / Water Mixture

F. Gygi

G. Galli

F. Ree

Hydrogen Fluoride / Water Mixture

- Highly corrosive (limited experimental data)
- Present in detonation products of insensitive explosives

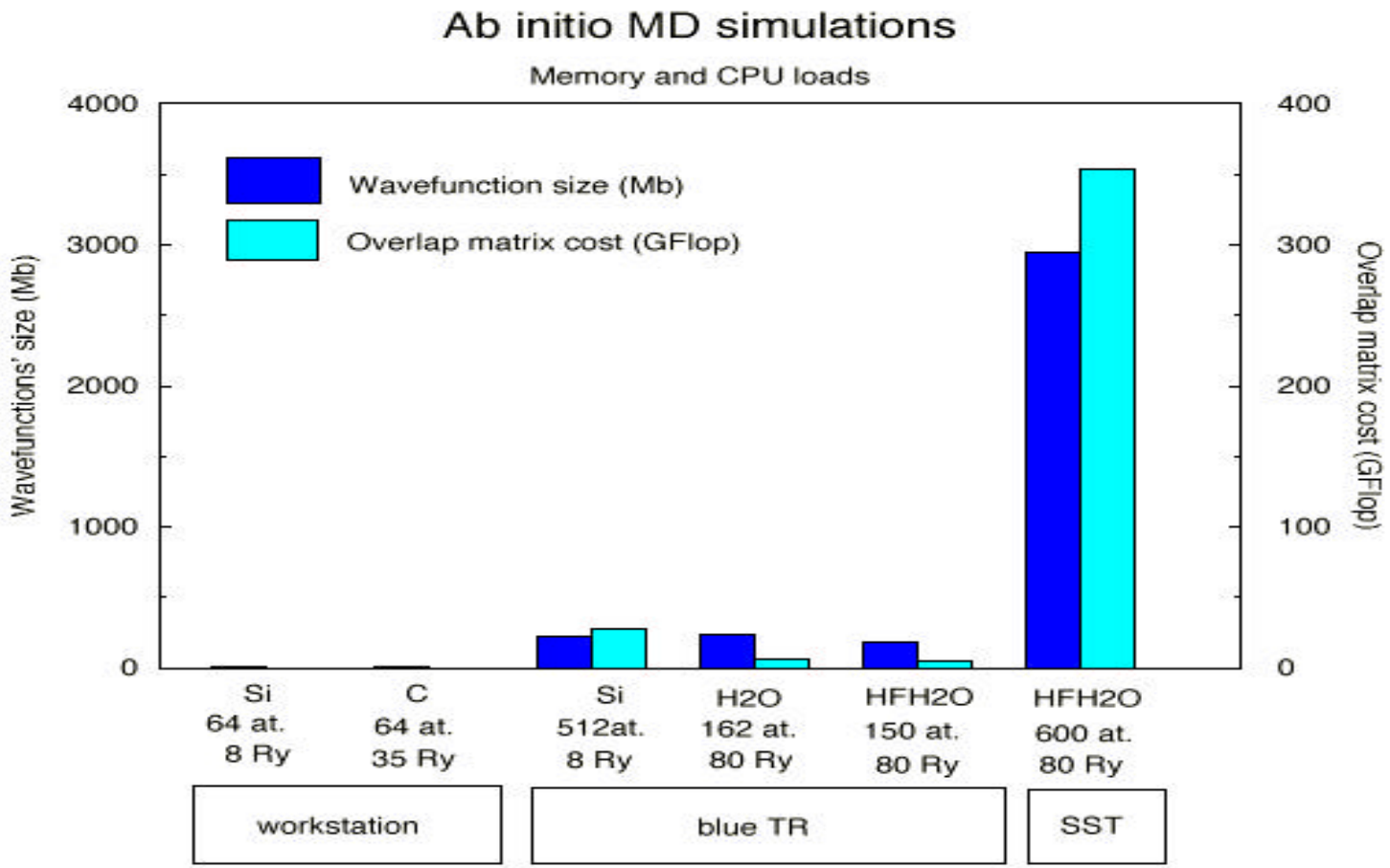
JEEP Molecular Dynamics Code

- Three-Dimensional
- Density Functional Theory to solve Schröedinger equation
- Applicable to high-temperature and high-pressure physics and to biochemistry
- Linear algebra and FFTs
- MPI plus multithreading

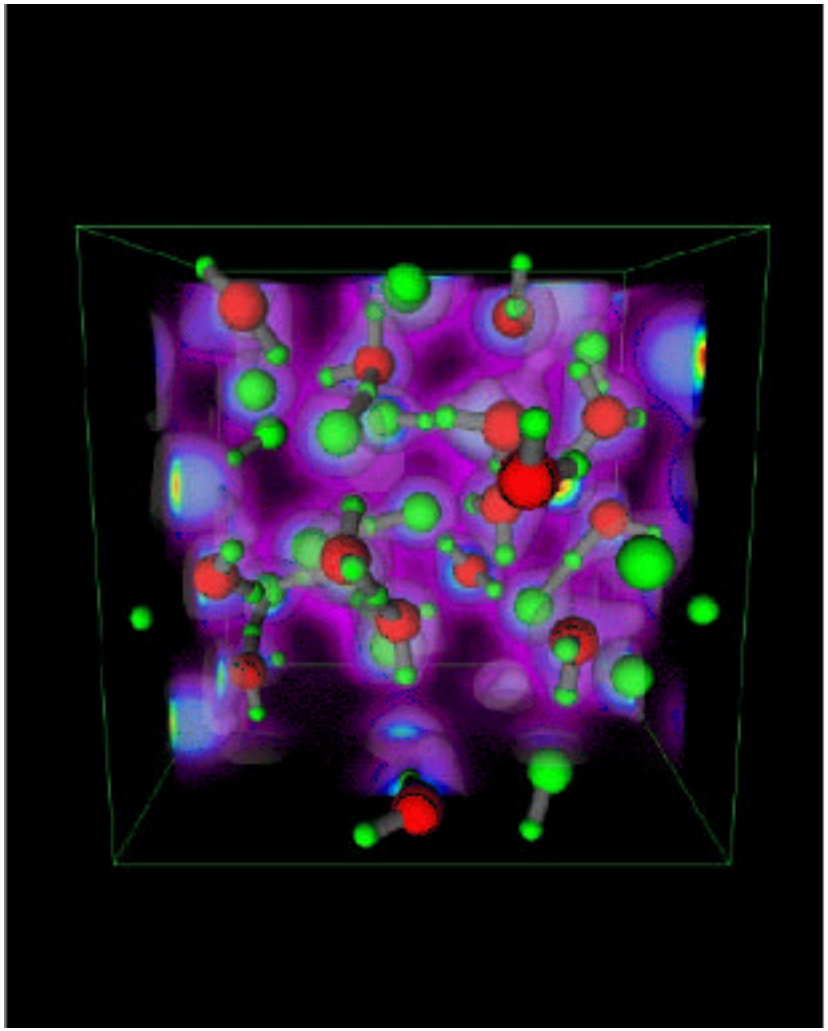
JEEP Simulation

- 240 molecules (600 atoms)
- 480 nodes
- Good scalability up to 480 nodes
  - poor scalability to more than one sector due to collective communication calls

Ab Initio MD Simulations



Visualizing the Simulation



Neutron Transport in the NOVA Laser

P.N. Brown, B. Chang, U. Hanebutte,

S. Smith, M. Dorr, R. Buck, J. Hall,

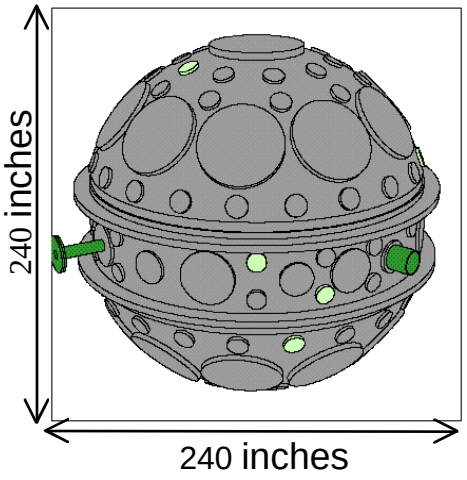
S. Post, J. Ferguson, J. Rogers, P. Nowak,

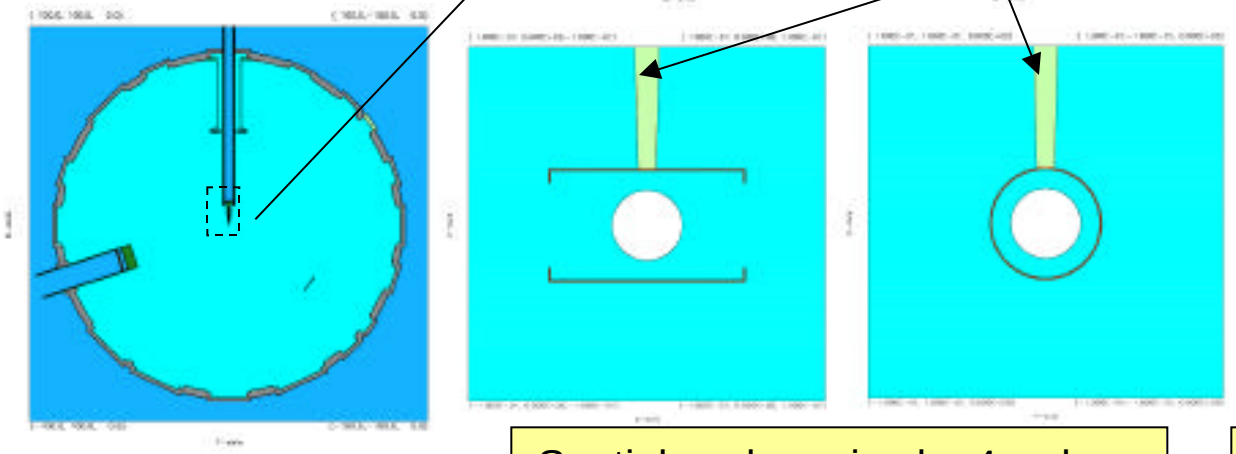
M. Zika and S. Hadjimarkos

Computation of Neutron Flux in NOVA

- Important to know neutron flux distribution in order to provide adequate shielding for experimentalists
- Target chamber has complicated 3-D geometry
- Construct numerical prototype of test chamber

Full System Run on Blue SST NOVA Test Chamber





Aluminum
Air at 760 Torr
Air at 10^{-5} Torr
Glass
Stainless Steel
Brass
Plastic
Copper
Helium
Gold

Spatial scale varies by 4 orders.

Material properties vary by 14 orders

ARDRA
Neutron and Radiation Transport Code

- Solves Boltzmann equation using either S_N or P_N approach
- Parallel in space, direction, and energy
- Handles complex geometries via interface to geometry engine of Monte Carlo code COG
- Optionally uses multigrid preconditioning based on Diffusion Synthetic Acceleration for optically thick problems
  - Scalable for S_N, but not P_N
- MPI with POSIX threads

High-Resolution ARDRA Simulation

- 160 million zones
- 4 moments
- 23 energy groups
- 15 billion unknowns
- 960 nodes
- Unprecedented detail and accuracy

Arbitrary Lagrangian Eulerian (ALE) Calculations

- Finite element method for treating fluid and elastic-plastic response on unstructured, hexahedral grid
- Domain decomposition with MPI across nodes
- MPI or OpenMP within node
- See the paper at this meeting entitled "Coupled Mechanical/Heat Transfer Simulation on MPP Platforms using a Finite Element/Linear Solver Interface," by C.J. Aro, E.I. Dube, W.S. Futral and J.D. Maltby.

Quantum Chemistry Computations

C.L. Janssen

I. Nielsen

E. Seidl

M. Colvin

Massively Parallel Quantum Chemistry Program (MPQC) Developed by SNL and LLNL Researchers

MPQC production level QC program:

- Closed and open shell Hartree-Fock energies and structures and frequencies
- Closed and open shell second order Møller-Plesset perturbation theory energies and structures and frequencies
- Coupled cluster, DFT and solvent models being implemented

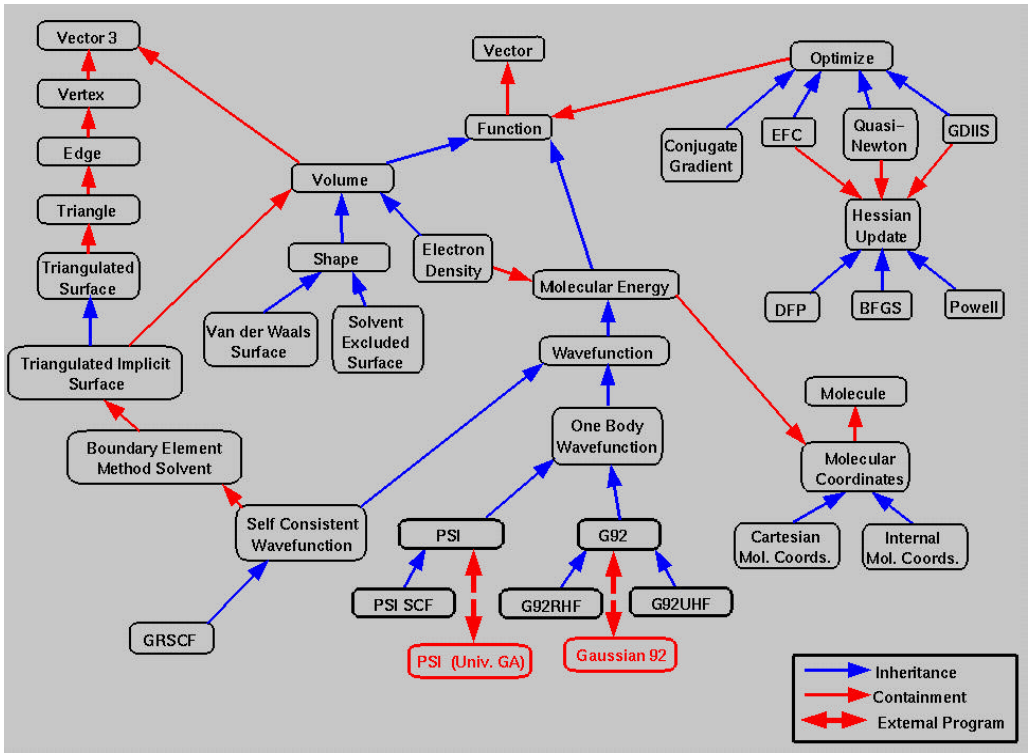
Object-oriented Implementation:

- 94,000 lines C++
- 34,000 lines C
- 115,000 lines machine-generated C++

Efficient on many computers:

- ASCI Blue and IBM SP2
- ASCI Red and Intel Paragon
- Shared memory multiprocessors

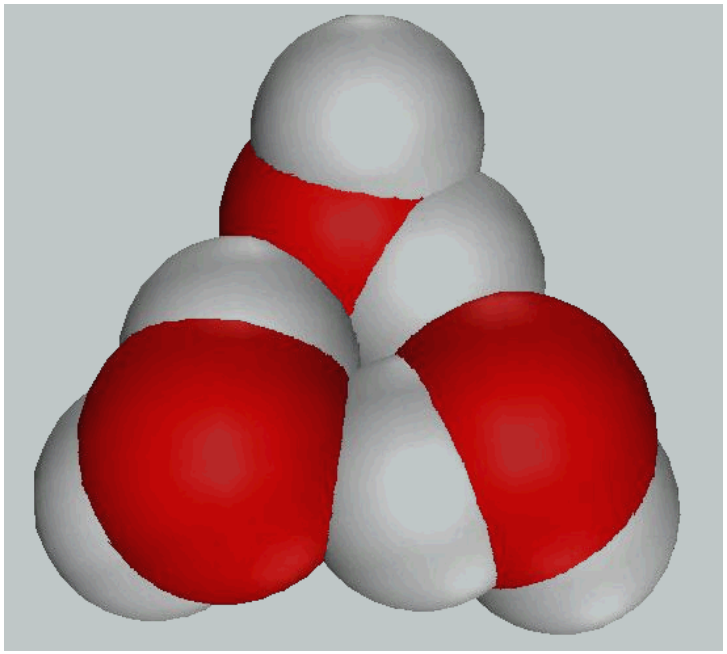
C++ Class Hierarchy for MPQC



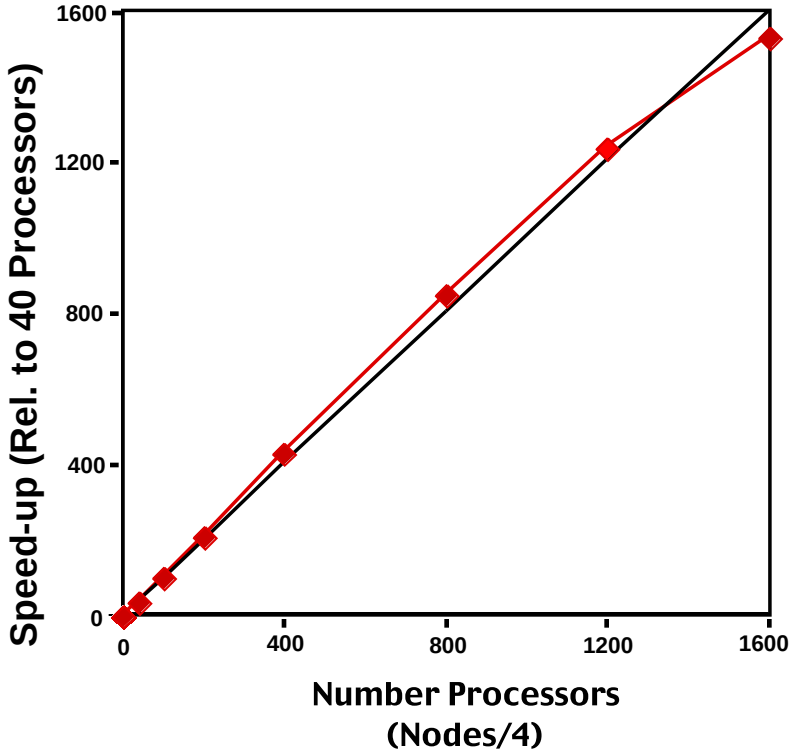
MPQC Used for Largest Quantum Chemical Perturbation Theory Calculation Ever Performed

Calculation performed by C.L. Janssen and I.M.B. Nielsen, SNL

Water Trimer: Accurate data on clusters needed for bulk liquid model potential



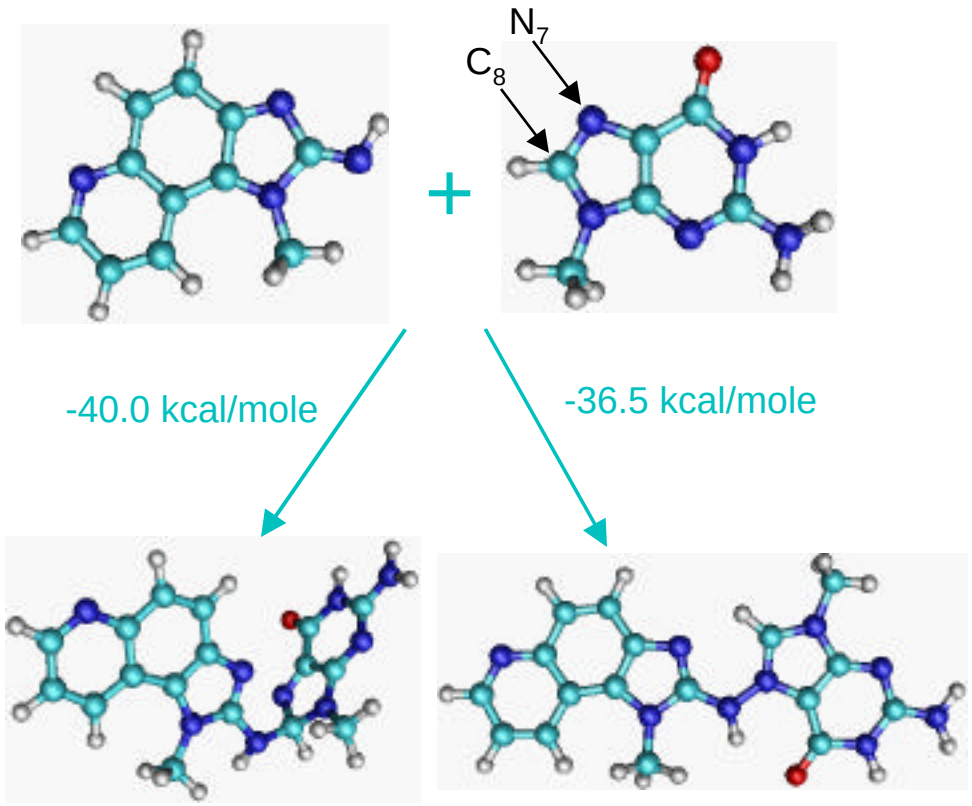
Excellent speed-up for MPQC up to 1600 processors



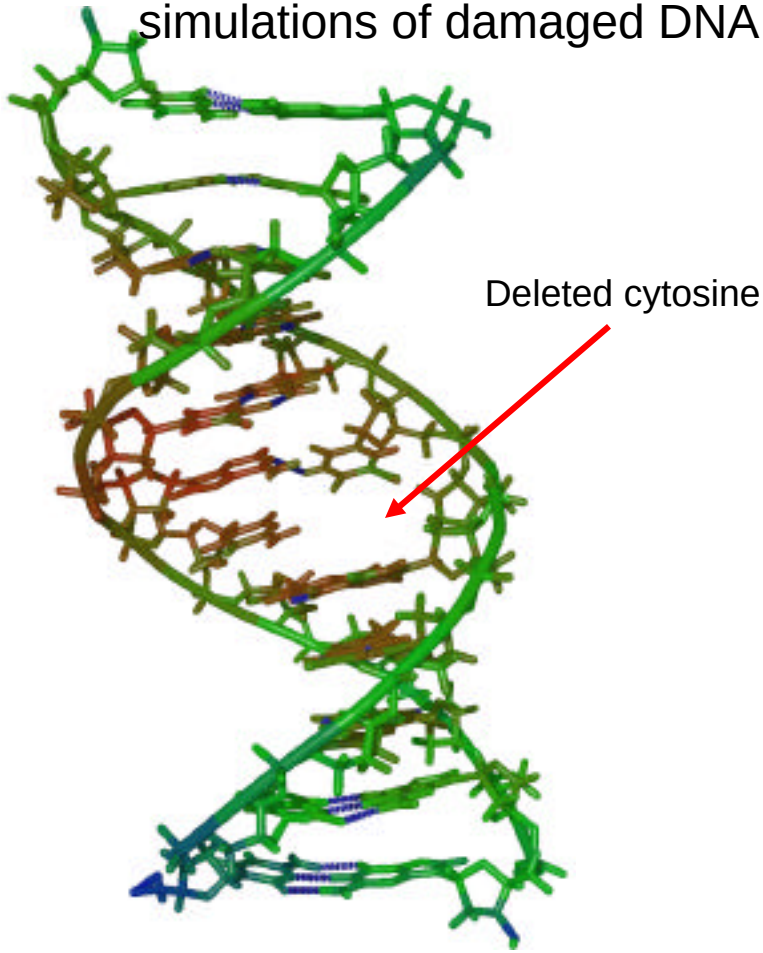
1329 basis functions--largest MP2 calculation ever done (1.4 hours on 1600 processors)

MPQC and CHARMM
Biochemical Simulations Being Run on ASCI Blue

Quantum chemical simulations
of mutagen-DNA binding



Classical dynamics
simulations of damaged DNA



Computational Gene Discovery using Massively Parallel Genomic Similarity Search (MPGSS)

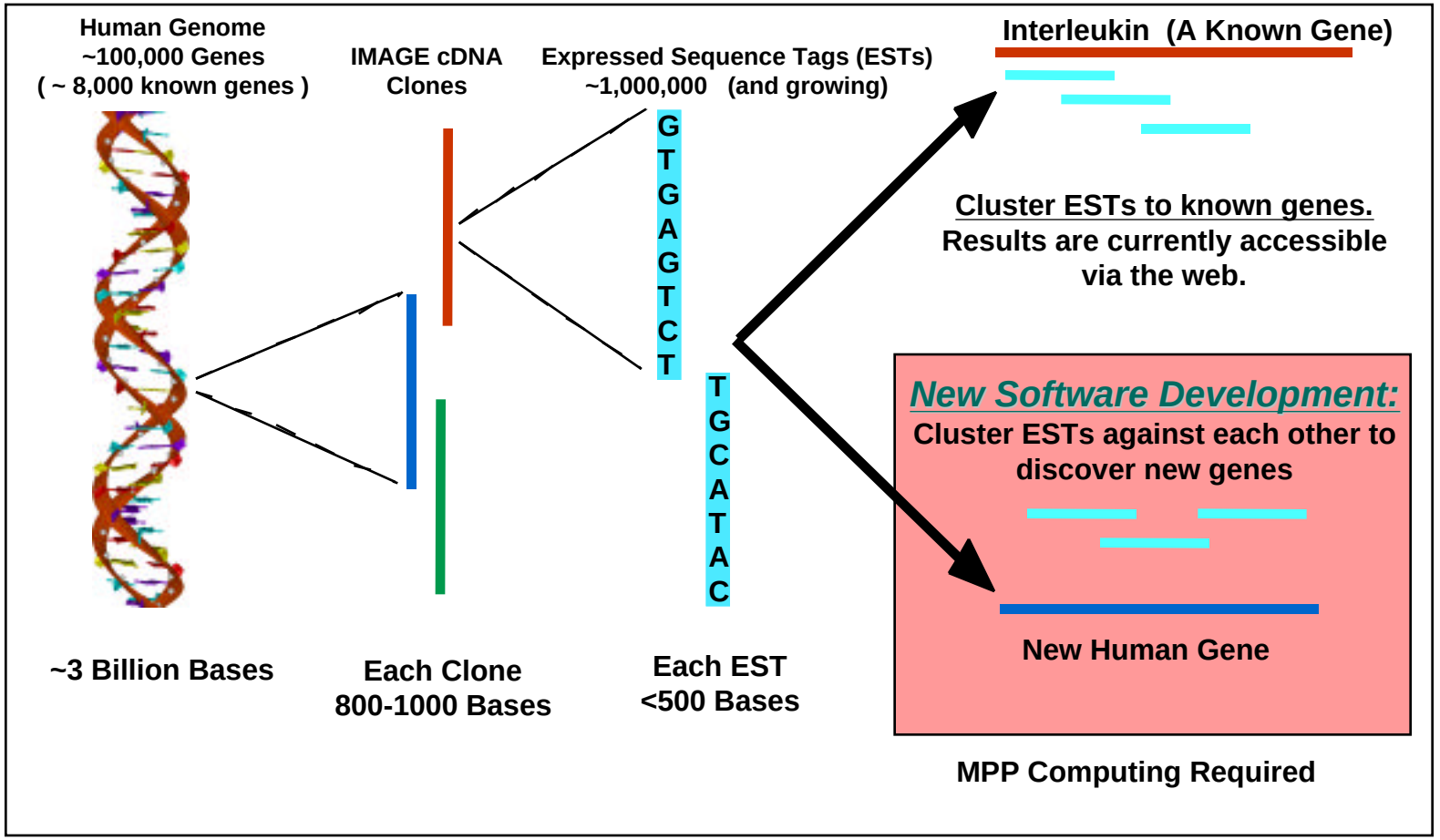
T.A. Kuczmarski

MPP Calculations in Human Gene Research

- ~ 3 billion bases (chemical subunits, ACGT, that describe the genetic code) in the 46 human chromosomes.
- Only about 1-6% of the bases are grouped into the biologically active and interesting entities called genes.
- Estimated to be ~100,000 human genes, less than 8000 discovered.
- We home in on areas of real genetic interest.
- Ignore the “junk” DNA sequences.
- Use MPP calculations to discover new human genes.

Computational Gene Discovery

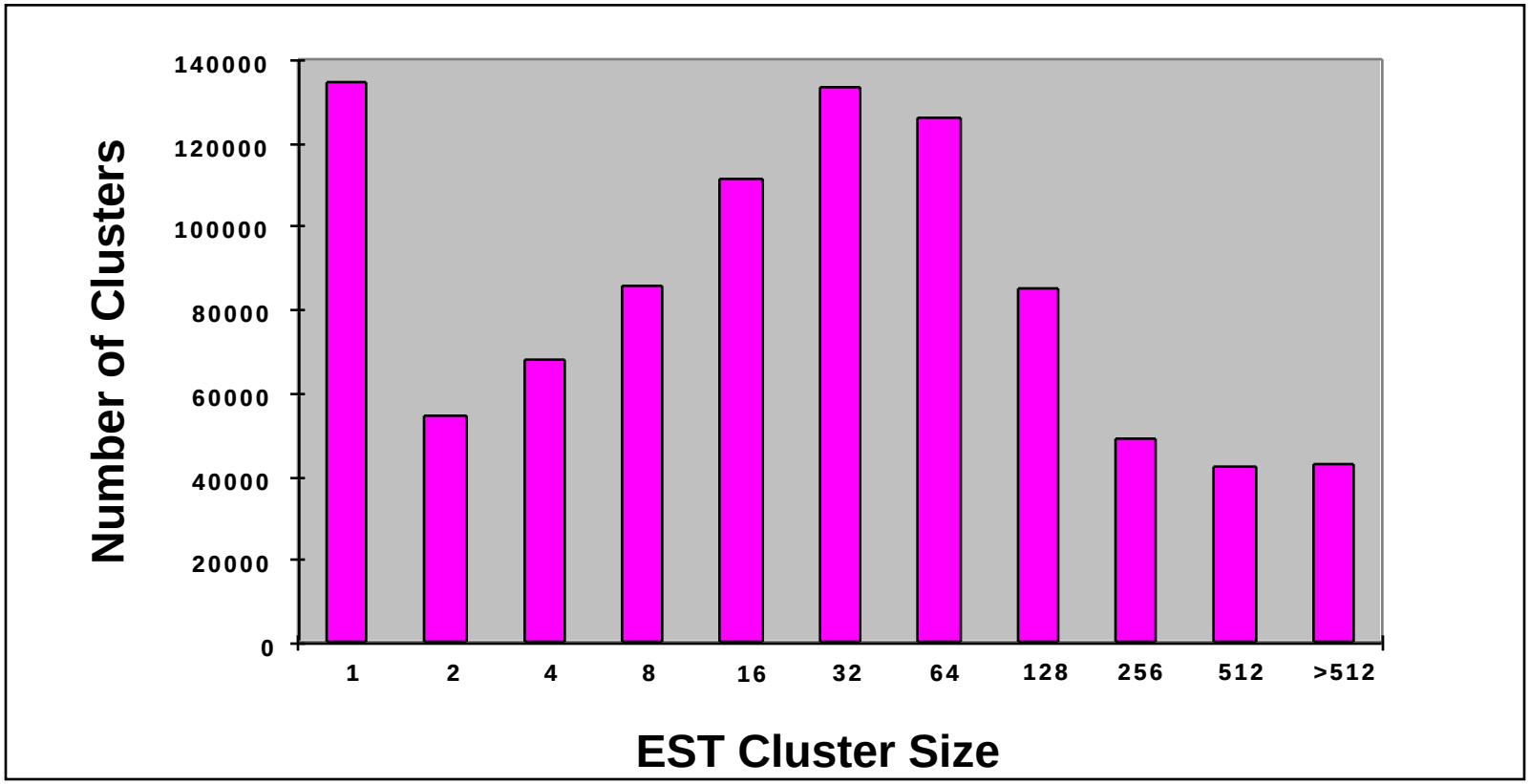
Mathematically Discover new Genes from
Information in National Biotechnology Databases



Computational Procedure

- Framework controls parallel executions of genetic similarity search program BLAST (National Center for Biotechnology Information).
- Framework uses Message Passing Interface.
- BLAST executes using threads.
- Results in very high parallel efficiency.

10^6 x 10^6 EST Cluster Run — 1920 CPUs on LLNL ASCI SST (6 Hours)



Requires ~ 6 months on 2 CPUs of a Sun Enterprise 4000

Conclusions

- Breakthrough calculations are being performed on the IBM ASCI machine in a variety of scientific areas.
- Many of these computations have been shown to scale to thousands of processors.
- It is too early to judge the superiority of either the mixed programming model or the pure message passing model.