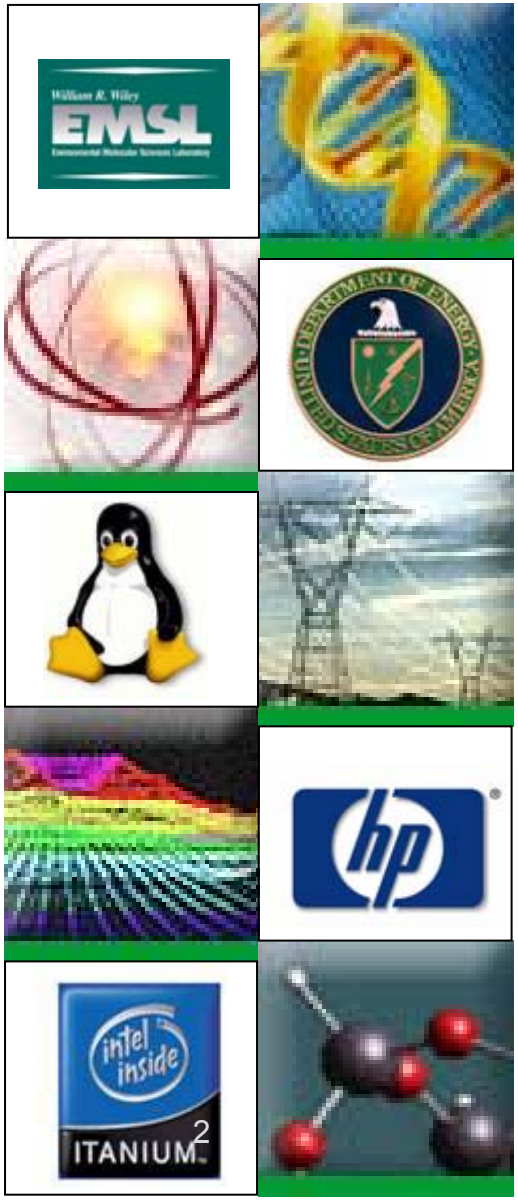


Planned Machines Panel

HPCS2 (AKA the HP cluster at PNNL)

R. Scott Studham
Molecular Science Computing Facility

studham@pnl.gov

Major Applications

Largest use by far is computational chemistry (NWChem, ADF, etc.)

Note we use atomic and molecular basis sets (not grids) so many important kernels are unfamiliar to non-chemists

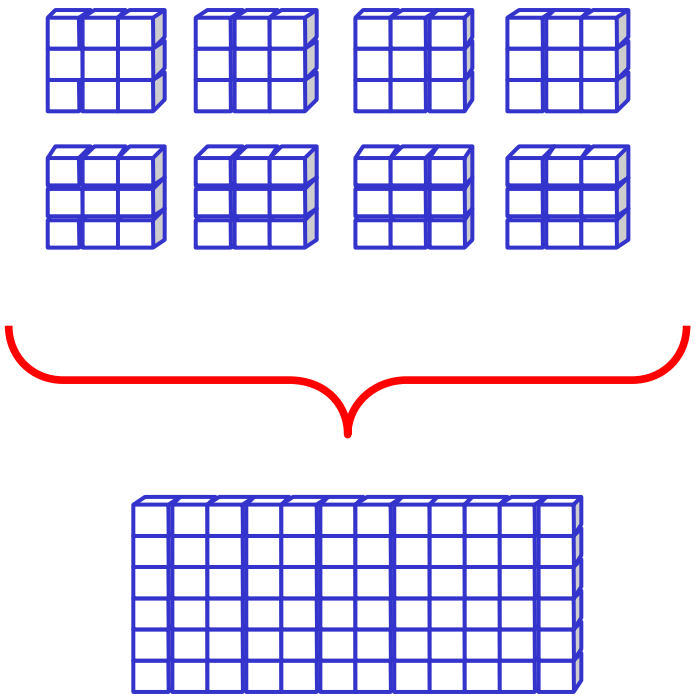
- High-accuracy electronic structure
- Low- and medium-accuracy electronic structure
- Molecular dynamics for computational biology

Other codes

- Atmospheric chemistry, regional climate modeling
- Subsurface transport
- Fluid dynamics, structural mechanics

Global Arrays

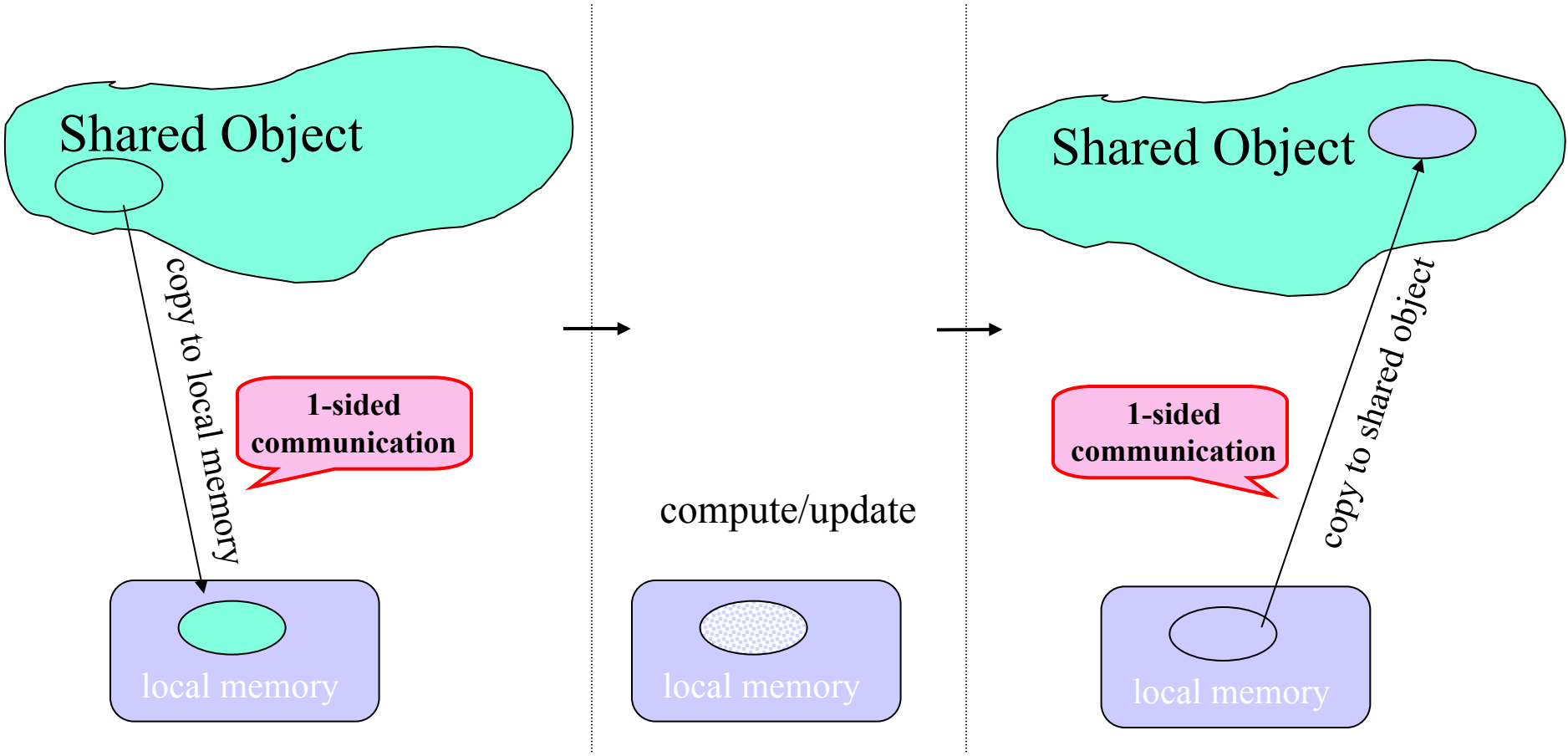
Physically distributed data



Single, shared data structure

- **Shared-memory-like model**
  - **Fast local access**
  - **NUMA aware and easy to use**
  - **MIMD and data-parallel modes**
  - **Inter-operates with MPI, …**
- **BLAS and linear algebra interface**
- **Ported to major parallel machines**
  - **IBM, Cray, SGI, clusters,...**
- **Used by most major chemistry codes, financial futures forecasting, astrophysics, CFD, computer graphics**
- **Supported by DoE MICS base program on Programming Models for Scalable Parallel Computing, and DOE 2000 ACTS in the past (J. Nieplocha)**

Shared memory model of computation in GA



- One-sided access to shared data increases ease of composition, reduces synchronization, and leads to more scalable programs

Gaussian Integral Evaluation

$$(ij\,|\,kl) = \int dr_1 dr_2 \phi_i(r_1)\phi_j(r_1) r_{12}^{-1} \phi_k(r_2)\phi_l(r_l)$$

N=O(10^3 – 10^4) – have O(N^4) integrals

Large code(s) with many critical loops

- Up to 80% of time for medium precision electronic structure calculations (SCF, DFT, MP2)
- Vectorizable recursions; short DAXPY (y(j)=a*x(j)+b*y(j)) and similar operations;
  gather and scatter; compiler generated code
- Some small level-2 and level-3 BLAS
- Tunable blocking to improve cache locality
- DAXPY-like operations must run well out of cache

SCF+MP2 analytic gradient

Semidirect algorithm

- Integrals+DGEMM+I/O
- Reduce re-computation by saving expensive AO/part transformed integrals on disk
- SCF requires high bandwidth and high capacity I/O
  - Write integrals once read many; sequential+large records
- MP2 requires lower bandwidth and less capacity
  - Out-of-core transformation/transposition
  - Strided write in blocks, sequential read
  - Blocks to use available local/shared memory and disk
  - Similar steps in coupled cluster computation

Summary of key computational kernels

CCSD(T)	DGEMM
Integrals	DAXPY-like in cache
Fock-matrix	Index eval, gather/scatter
Spectral	FFT 2-D
Mesh	Short sparse DDOT/DAXPY
MD	FFT, recip, sqrt, exp
SCF	Fast sequential IO
MP2	Fast sequential and random IO

Level-3 BLAS

DGEMM and similar operations

- Matrix multiplication
- Majority of wall time for high-precision electronic structure calculations (CCSD-T) designed to be in DGEMM
- A component of Gaussian and plane-wave DFT calculations
- Related kernels in DFT quadrature, FMM and other fast algorithms (e.g., MRA)

Achieving near theoretical peak processor speed on Itanium2 (FLAME efficiency on Itanium2)

SGEMM 99.0% of peak	CGEMM 99.0% of peak
DGEMM 98.4% of peak	ZGEMM 98.5% of peak

Let's answer Question2 then Question1.

EMSL/BER codes → Many competing algorithms, superset chosen → our balance

Q2: What characterizes your applications?

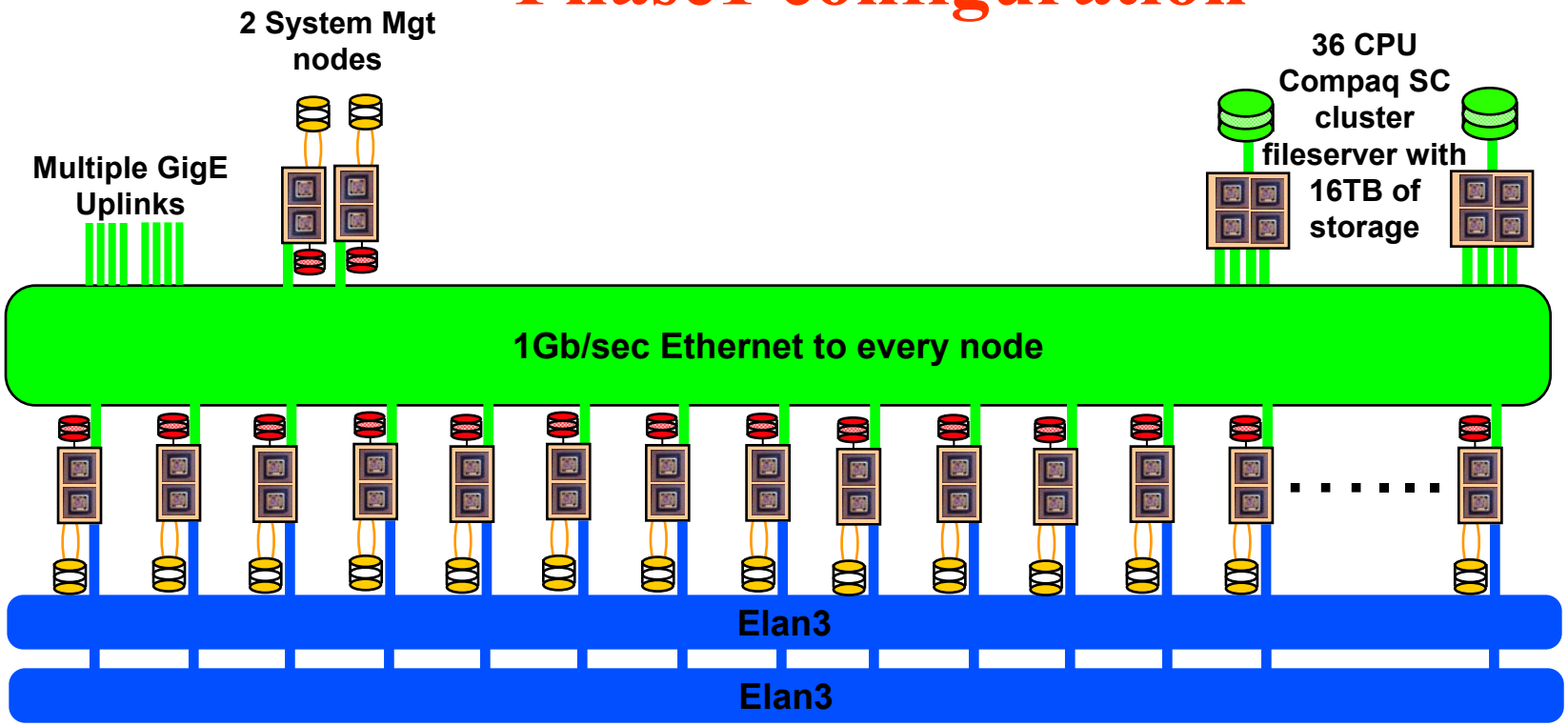
1. Global Arrays allows us to view all 7TB of RAM as a single shared object.
2. We can use local IO to cache integrals instead of recomputing
3. Applications spend large fraction of wall clock in BLAS kernels.
4. Programming models makes use of RDMA

Q1: What is unique in structure and function of your machine?

1. Require >>232 address space → 64bit everything
2. 292GB filesystem sustaining 190MB/s <u>on each node</u>
3. High efficiency in BLAS3 → Itanium2 & good math libraries
4. Quadrics Elan4 Interconnect for low latency, low overhead, RDMA.

256 Intel® Itanium® 2 processors

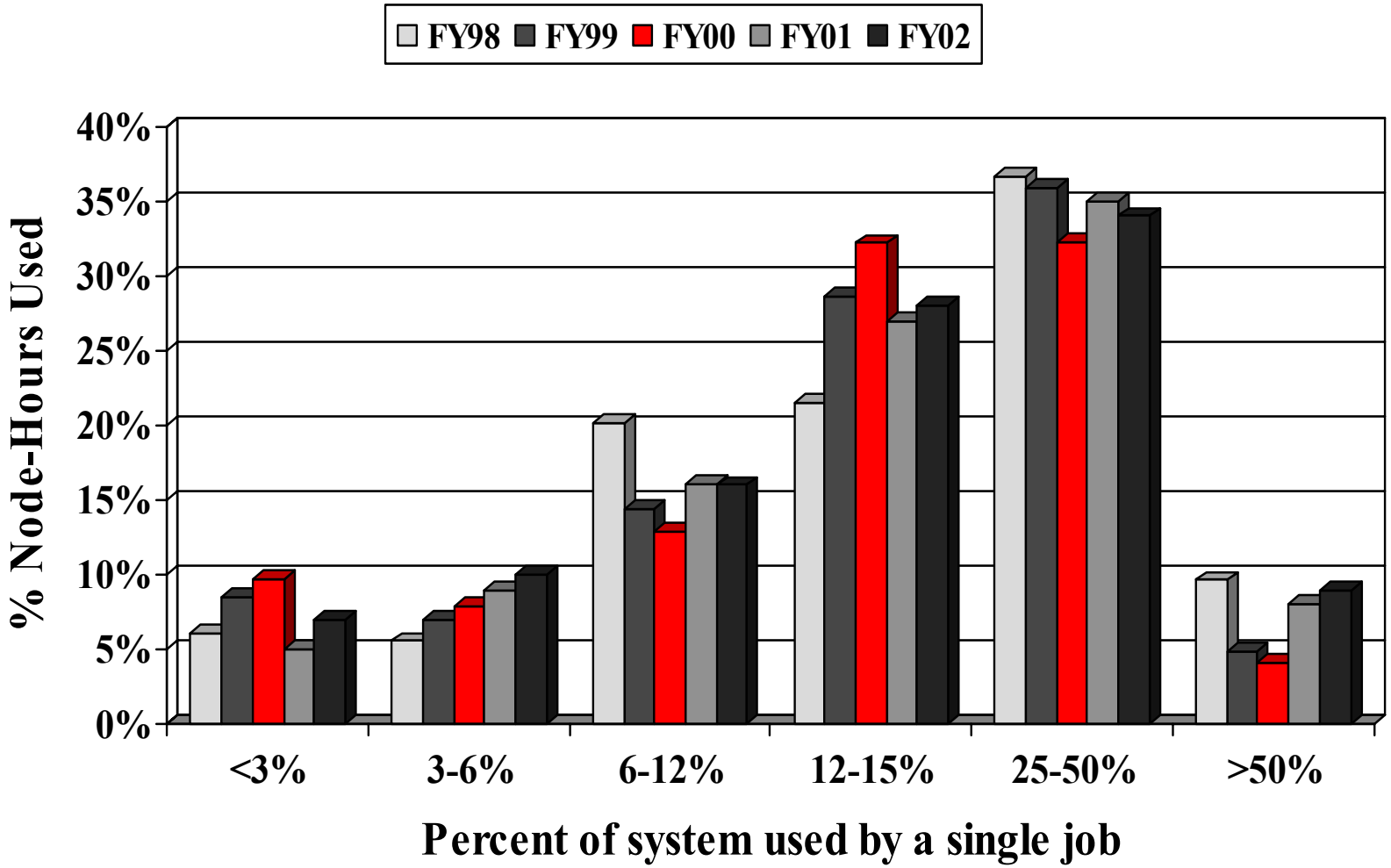
Phase1 configuration



1TF
1.5TB Memory
16TB Global FS
25GB/s Aggregate IO

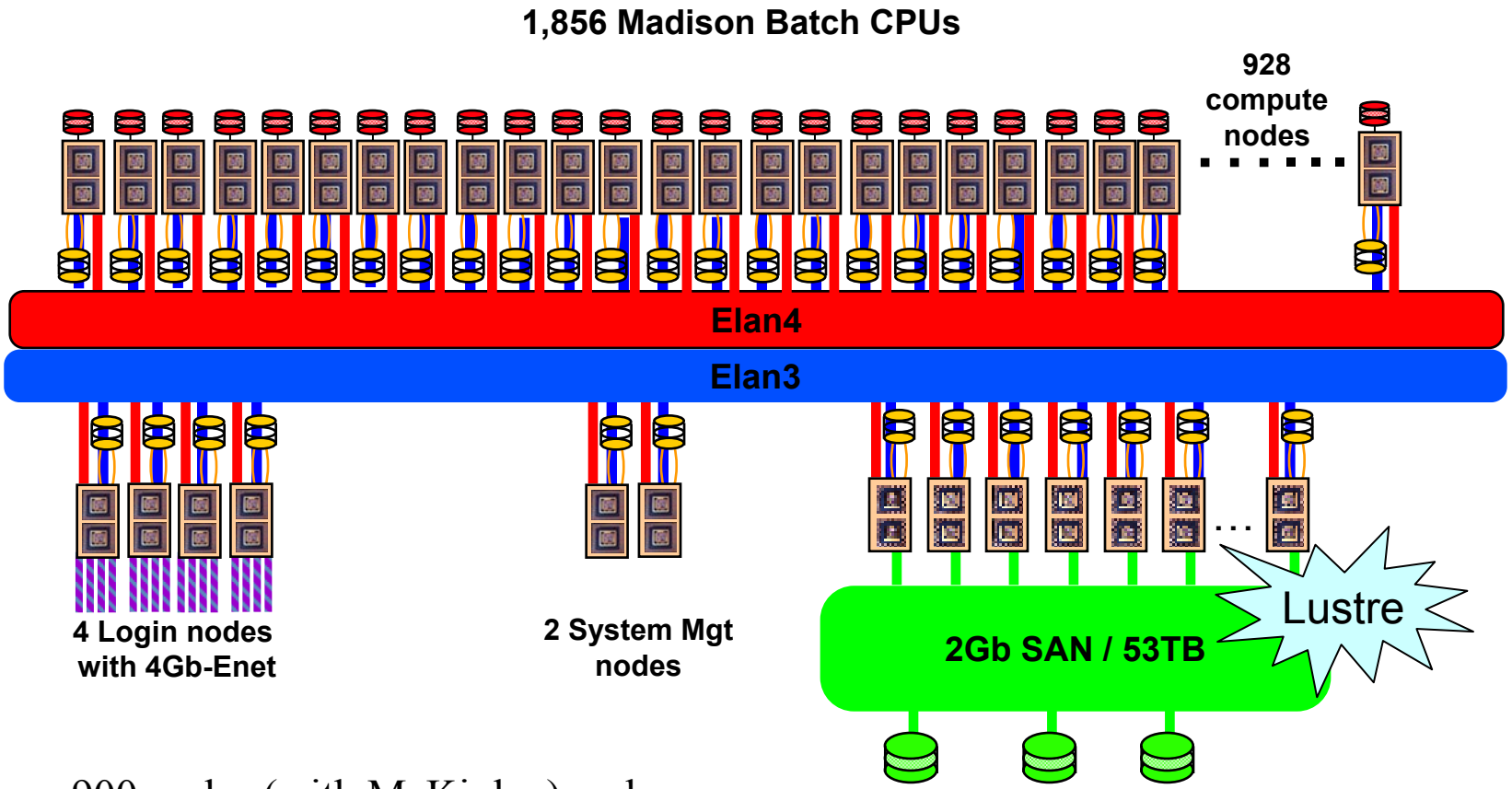
MSCF Usage

Most of the usage is in large parallel jobs



1,900+ next generation Itanium® processors

Phase2 Configuration



900 nodes (with McKinley) and Elan3 being delivered this week. Upgrade this summer.

11.4TF
6.8TB Memory

Q3: What prior experience guided you to this choice?

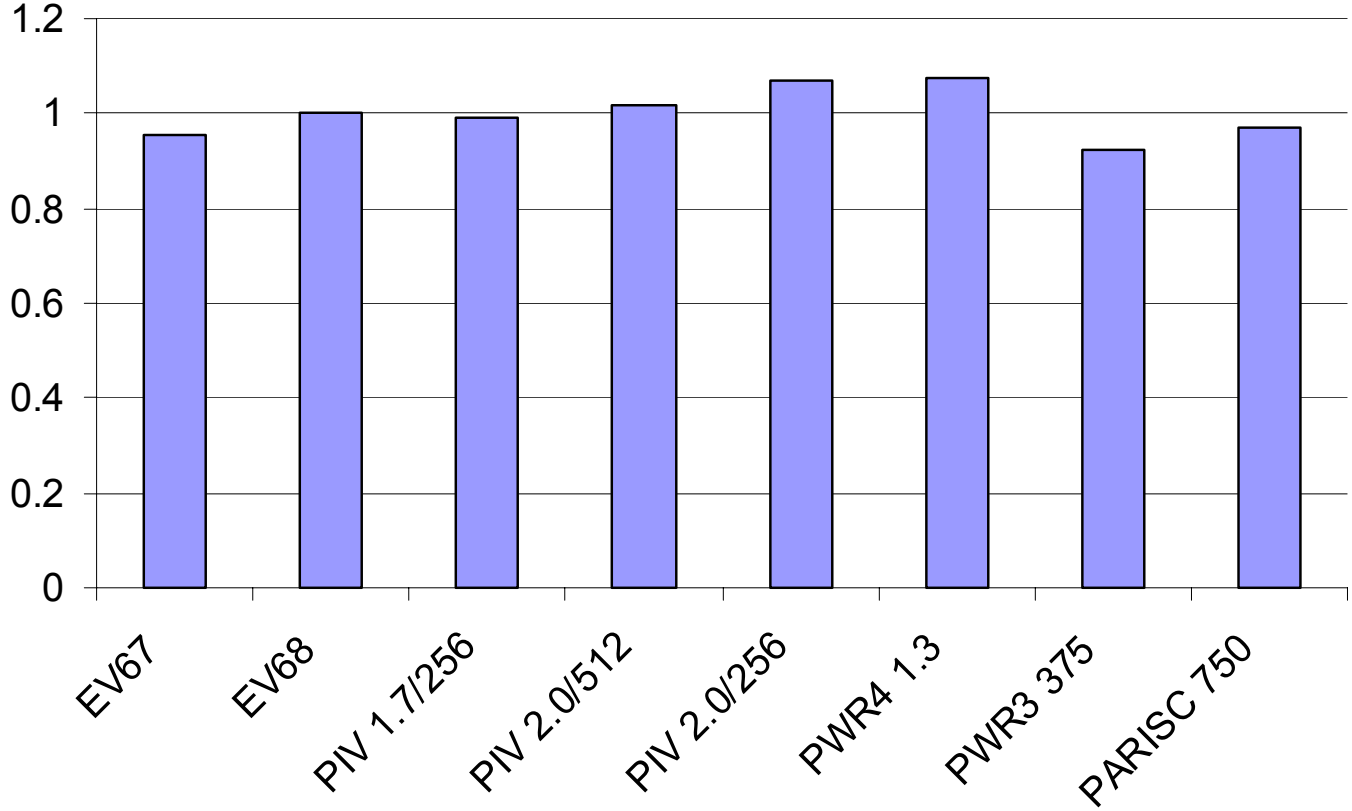
1. 512 node IBM SP2 with similar balance
   - Local IO
   - Good balance of compute to interconnect
2. Application developers were cornerstone to system choice.
3. Correlation between industry standard benchmarks and our benchmarks
4. Benchmarks on different systems with different balances.

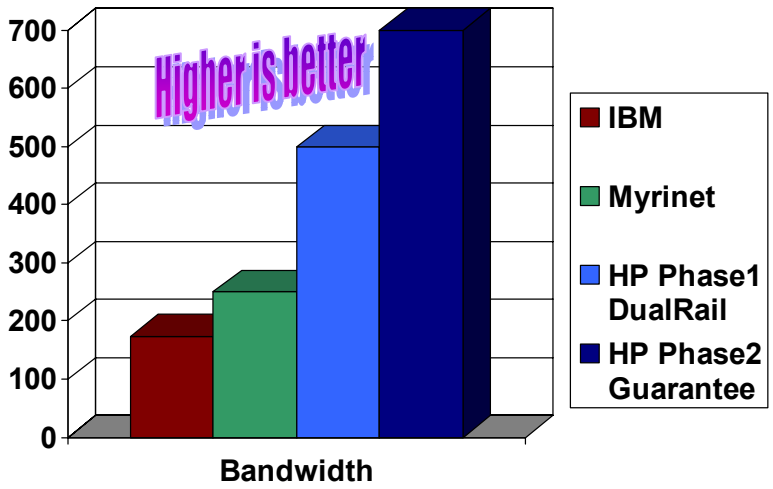
Role of Standard Benchmarks

Linpack 100/1000, SpecFP, Stream, …

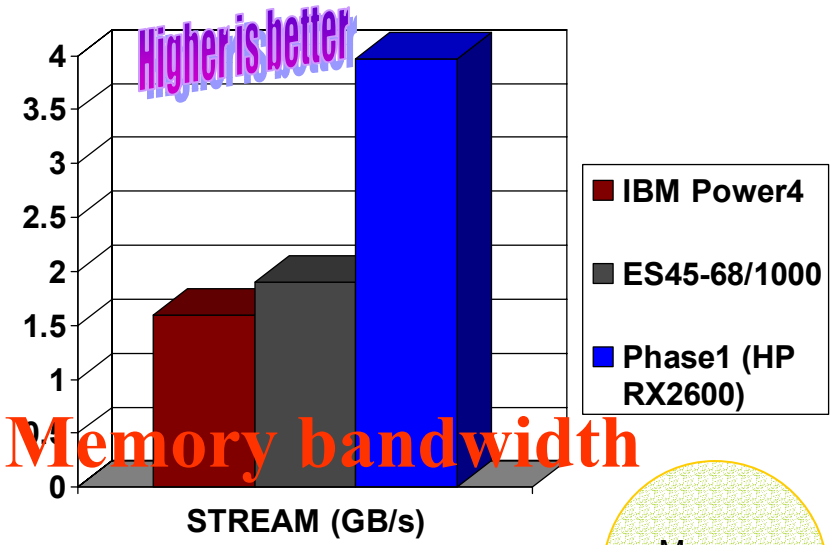
- Enable comparison between vendors with some limited interpretation
- Correlates well with selected applications
- Partially eliminates impact of immature compilers, etc.

Correlation between SpecFP and DFT SiOSi3 benchmark



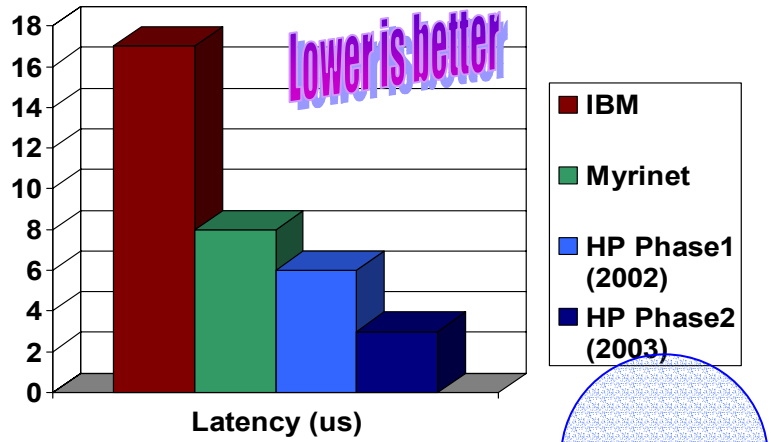


Interconnect

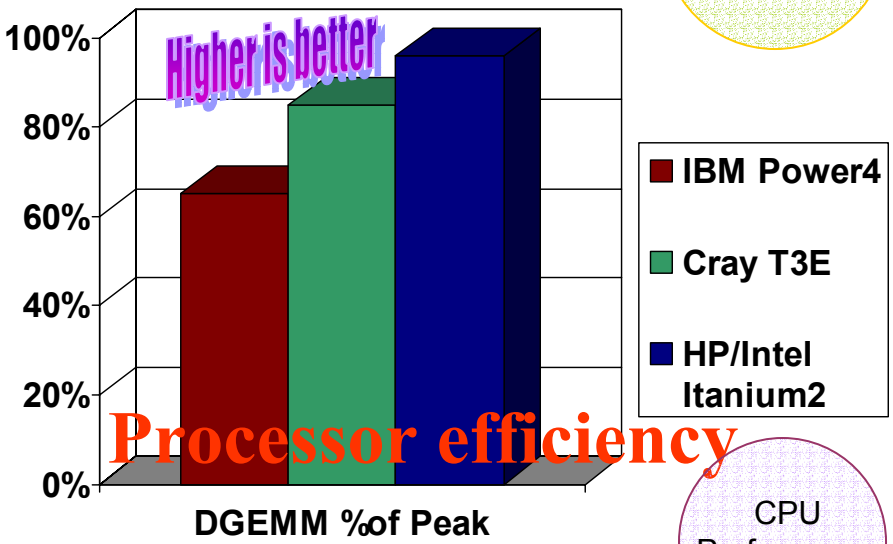


Memory bandwidth

Memory Bandwidth



Interconnect

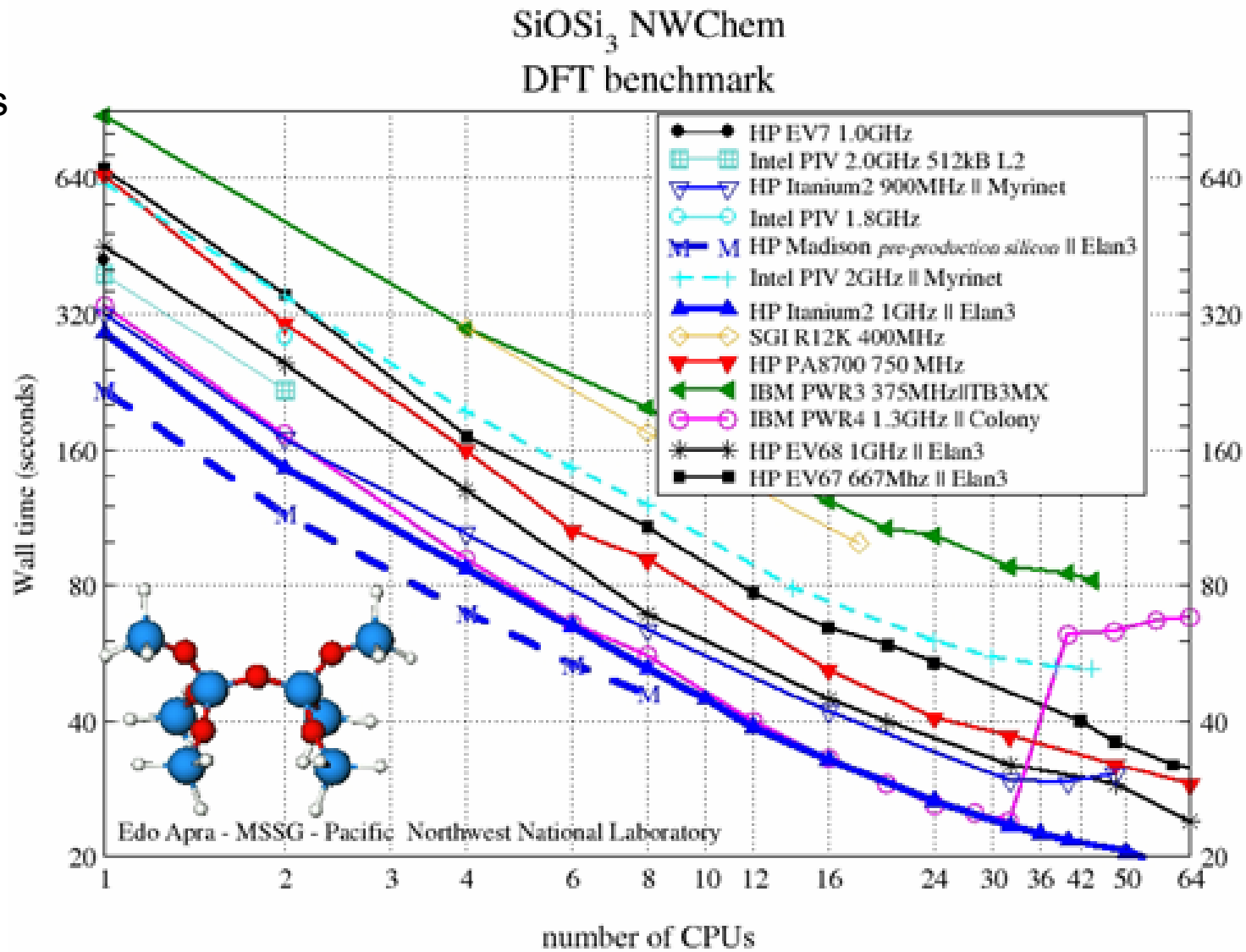


Processor efficiency

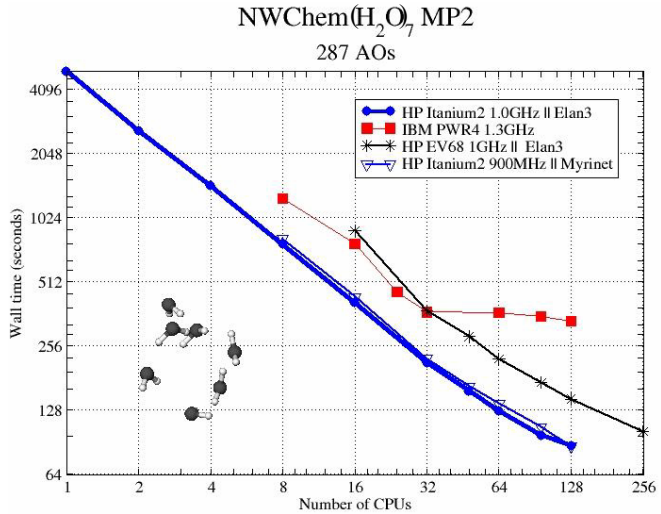
CPU Performance

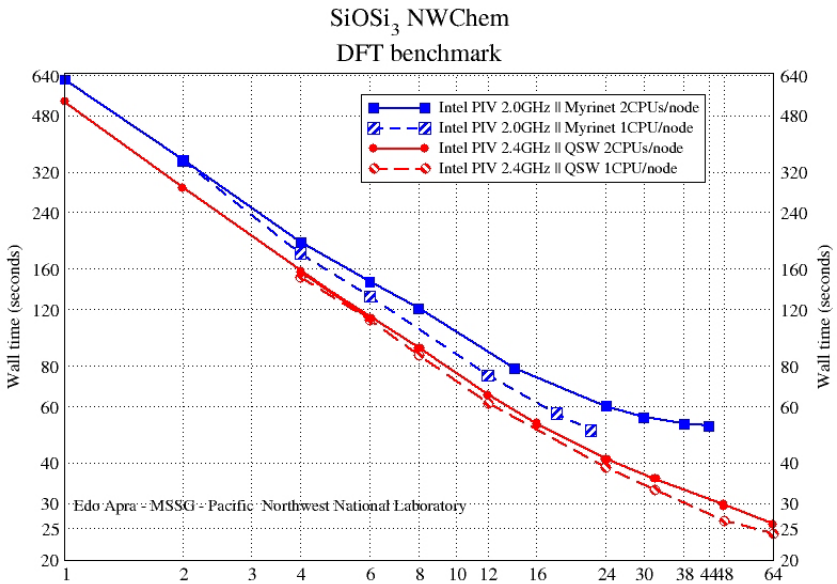
Parallel scaling of NWChem

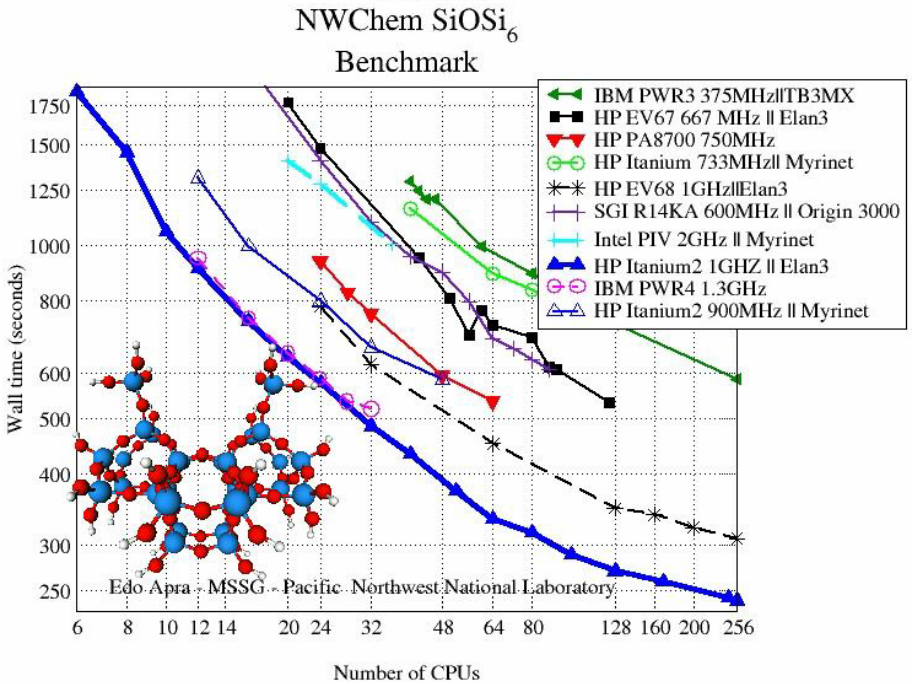
Prototype Madison's have demonstrated a greater than clock rate improvement in performance.

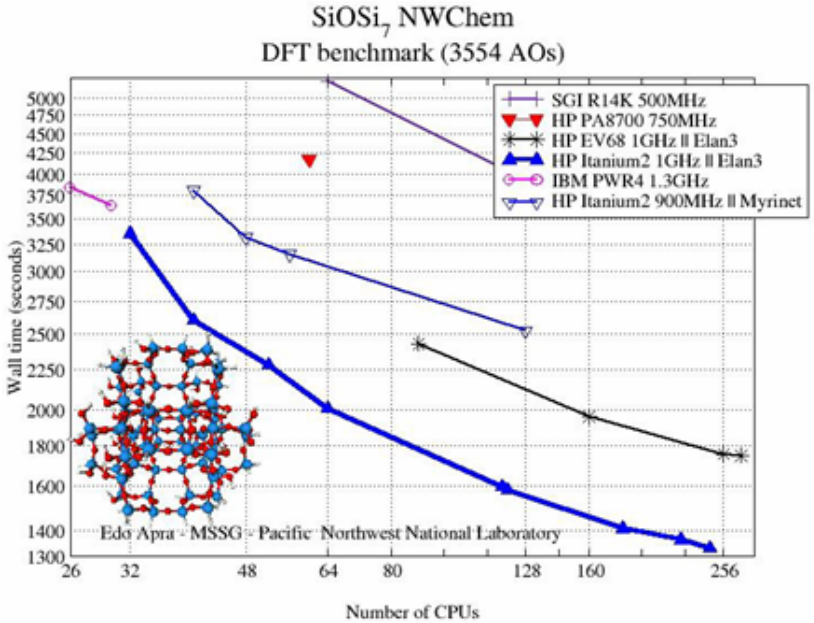


Parallel scaling of NWChem









Q4: Other than your own machine, for your needs what are the best and worst machines? And, why?

Best

EV79, Elan4 with local IO

(until somebody buys one the PSC cluster will do)

Worst

IA32, no good RDMA, without local IO

Acknowledgments

U.S. Department of Energy's Office of Biological and Environmental Research
and many others

Alan Geller
Angela Grossman
Ari Patrinos
Balaji Veeraghavan
Barry Crume
Bert DeJong
Bill Pitre
Bill Rogers
Bob Gobielle
Bruce Cambell
Caroline Nold
Chip Ritchie
Chris Heiter
Chris Maestas
Chris Schafer
Dale Flowers
Dan Fogel
Dave Brawn
Dave Field
Dave Greenaway
Dave Sweetser
David Cowley
David Dixon
David Field
David Geary
David Shurtlef
Desna Weber
Doug Ray
Edo Apra
Evan Felix
Everett Darco
Frank Baetke
Gary Skouson
Gary Thunquest
Gene Saxman
Gene Stott
George Fann
Glen Rowe
Greg Astfalk
Gregg Syrovatka
Harold Trease
Henry Schwindt
Hsin-Ying Lin
Jack Burris
James Marsh
Jarek Nieplocha
Jeane Bosse
Jeff Day
Jeff Nichols
Jim Baumgartner
Jim Gianotti
Jim Zafarana
Joanne Blanding
John Laroco
John Lenthall
John Porter
Kathy Wheeler
Keith Bazarnik
Keith Poirier
Kevin Carson
Kevine Wilson
Kwang Lim
Larry Mahoney
Leel Peesapati
Lisa Burke
Lisa Hobson
Logan Sankaran
Mark Hillstead
Mark Jensen
Mark Woodbury
Matt Hageman
Melanie Soulas
Merritt Smith
Mike Moran
Nathan Tenney
Operations
Pam Dreizen
Pam Tresslar
Patti Howe
Paul Bayer
Paul Martin
Paul Martin
Ralph Wescott
Rich Christman
Richard Pinos
Richard Searles
Rick Soctt
Rob Lucke
Robert Harrison
Ron Fisher
Rudy Anderson
Ryan Wright
Sam Rector
Scott Jackson
Scott Studham
Shelly Beck
Sherri Rossnack
Stephen Squires
Steve Joachims
Steve Langdon
Steve Leventer
Stuart Yoshida
Terry Caracuzzo
Terry Kjemperud
Theadora Carson
Theresa Fryberger
Theresa Windus
Tim Witteveen
Tjerk Straatsma
Tom McGarry
Toni Quackenbush
Vicki Niccum
And many others