Scalable First Principles Electronic Structure Methods

William A. Shelton, George Fann, and Robert J. Harrison

Computer Science and Mathematics Division

Oak Ridge National Laboratory

Acknowledgement of Sponsors

Department of Energy/Office of Science

Office of Advanced Scientific Computing Research

Mathematics, Information and Computer Science

Applied Mathematical Sciences Program

Office of Basic Energy Science

Chemical Sciences, Geosciences, and Biosciences

Oak Ridge National Laboratory

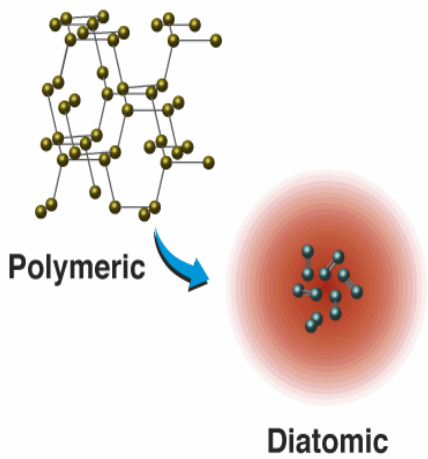
National Leadership Computing Facility
located at Oak Ridge National Laboratory

Outline

- **Motivation**
- **Linear Scaling Algorithms**
- **Conclusion**

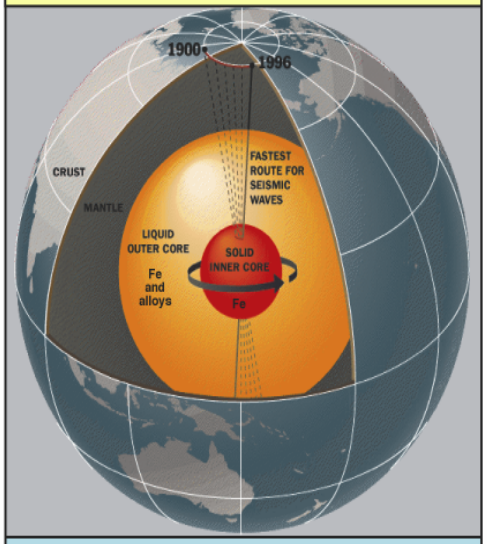
Quantum Simulation Goals: Accuracy and Predictive Capabilities

PREDICT Physical Properties



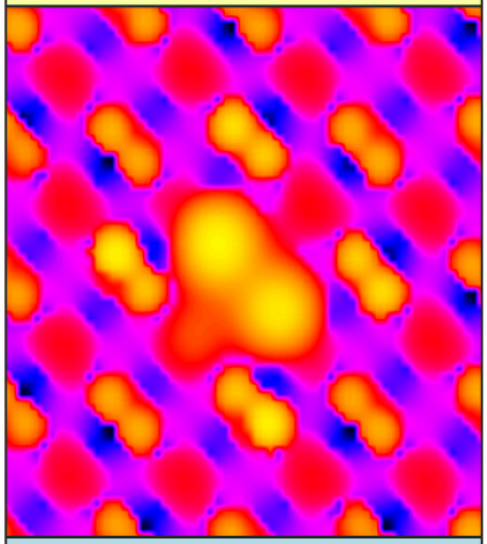
Design of novel materials

INVESTIGATE Properties Not Directly Accessible to Experiments



Matter under extreme conditions

INTERPRET and COMPLEMENT Experiments

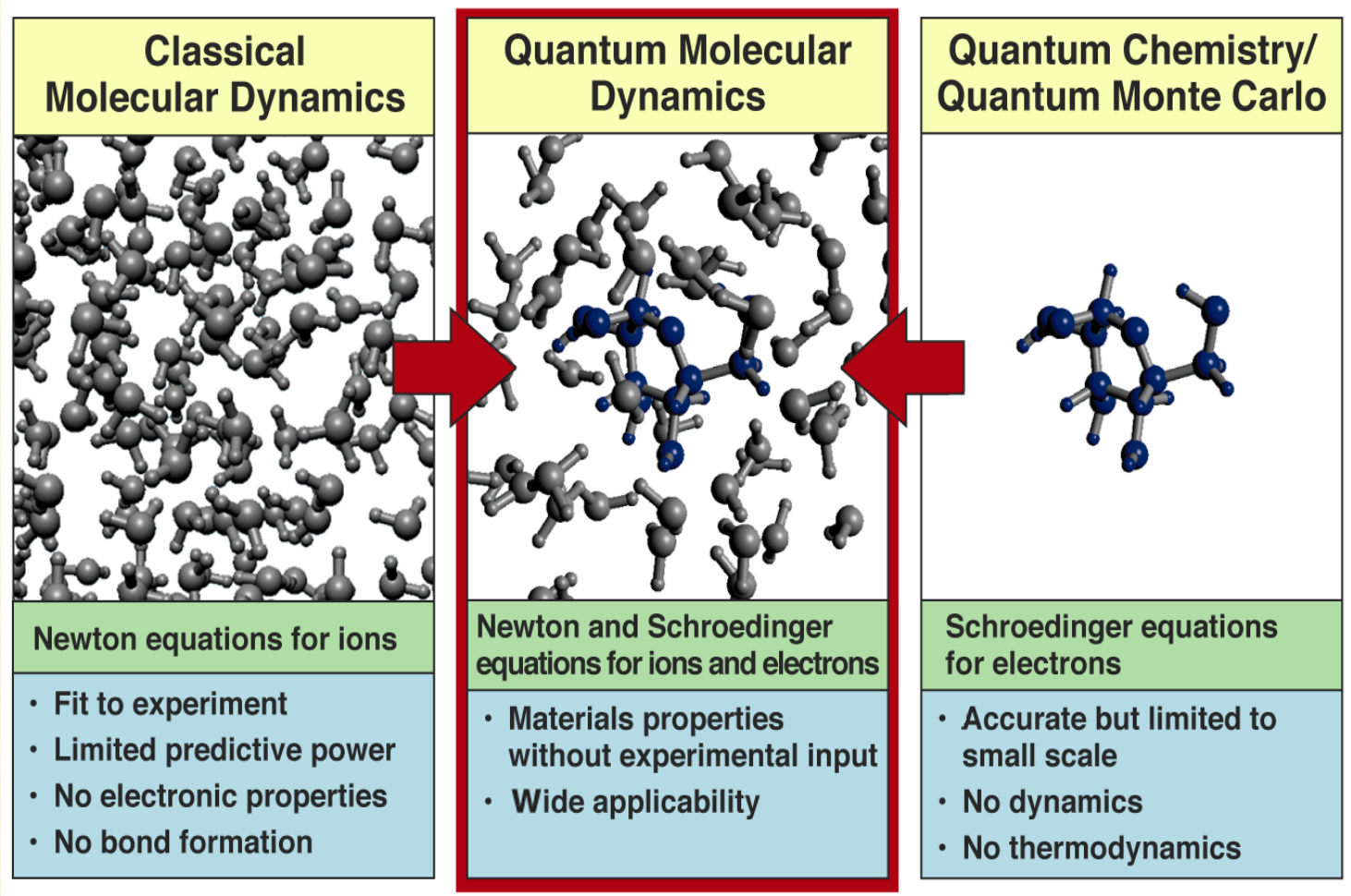


Determining surface structure

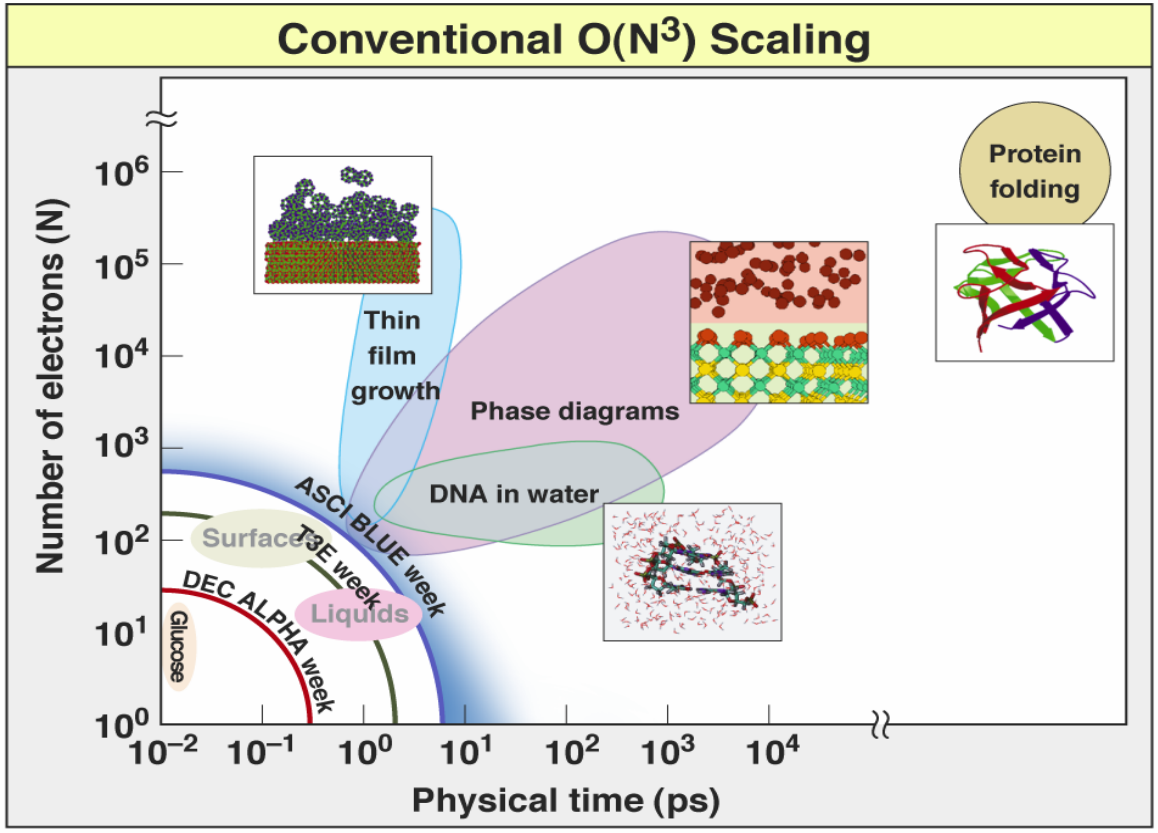
Predict properties of matter on the basis of laws of quantum mechanics

Scalable and Accurate First Principles Method

Atomistic Methods

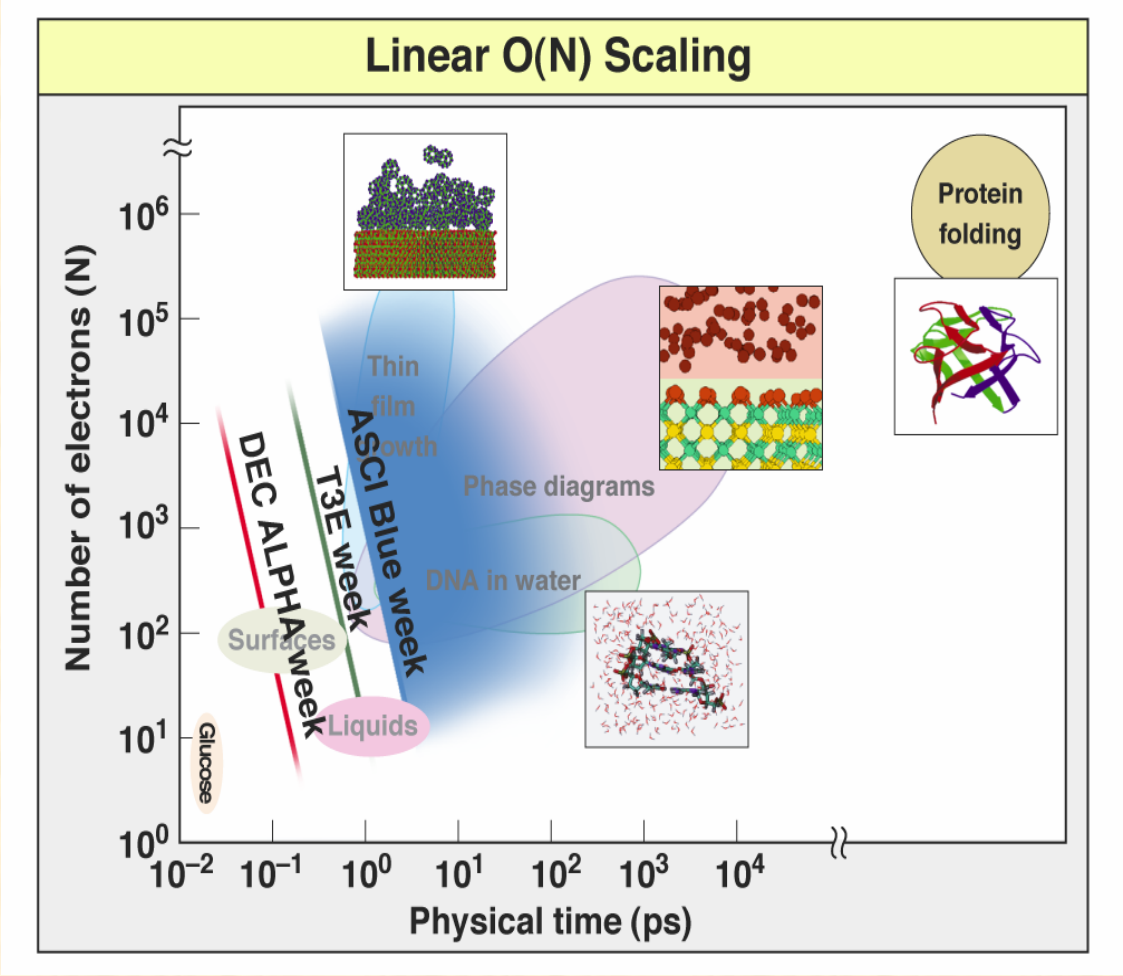


Advances In Hardware Alone Are Not Sufficient



With conventional algorithms, even large-scale machines will not provide "enabling" capability for interesting problems

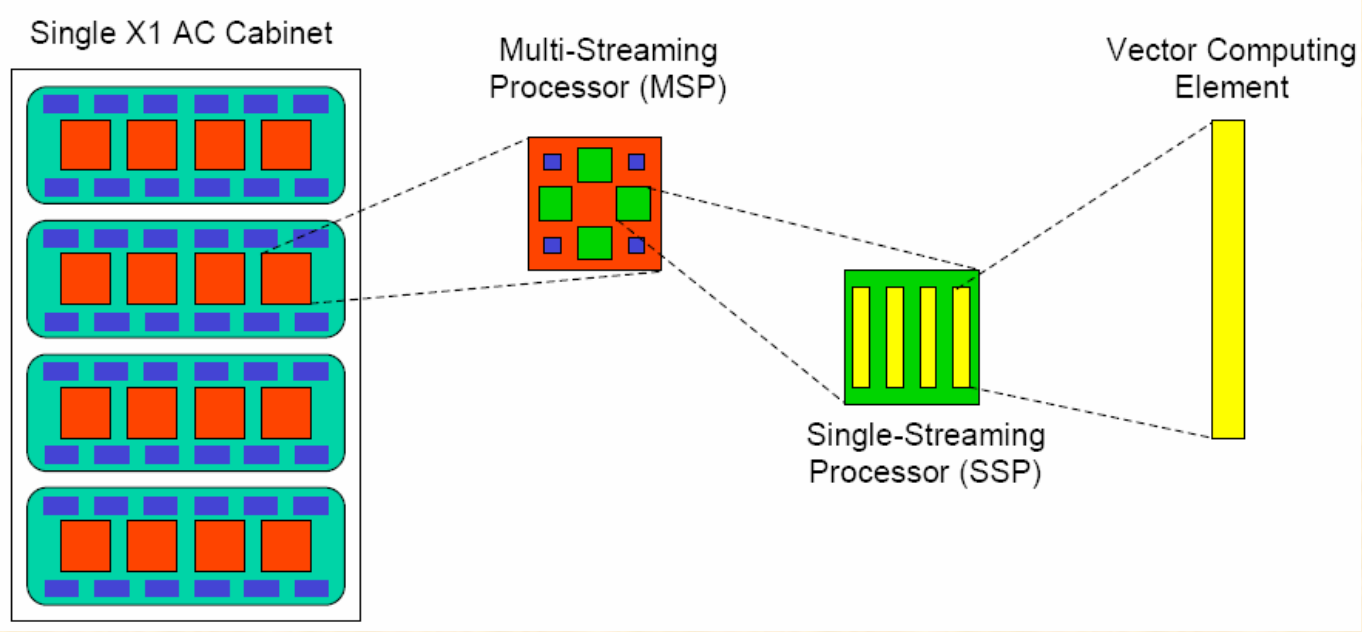
Linear Scaling Algorithms Will Enable Solutions to New Problems



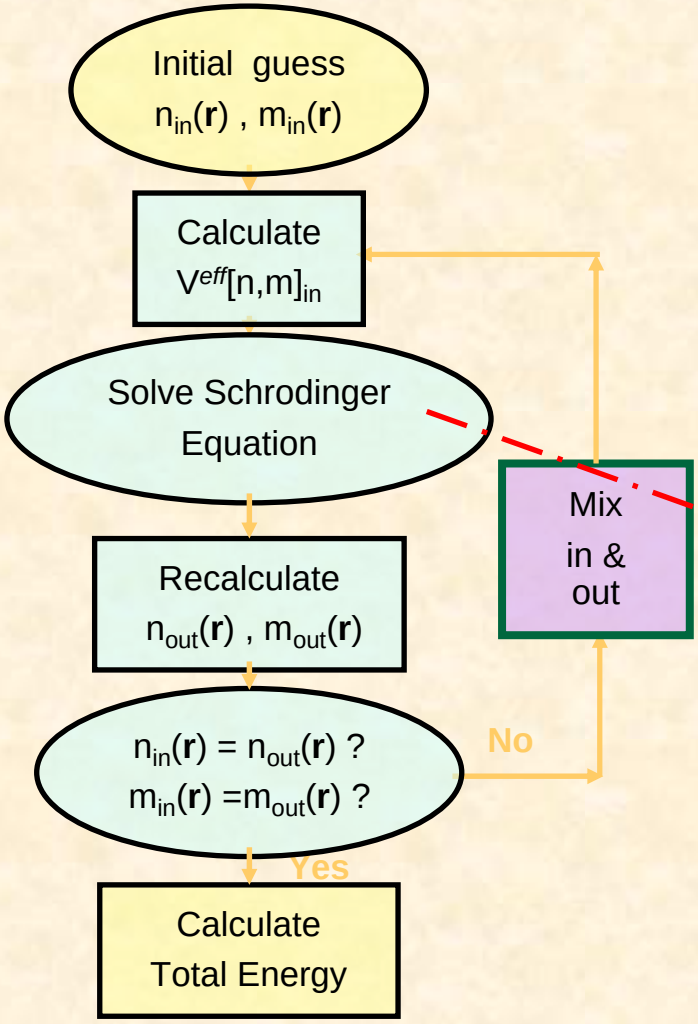
The combination of new advanced computing platforms and new scaling algorithms will open new areas in quantum-level materials simulations

Cray-X1E

- **1024 Multi-streaming vector processor (MSP)**
  - **Each MSP has 2 MB of cache and a peak computation rate of 12.8 GF**
  - **4 single-streaming processors (SSPs) form a node with 16 Gbytes of shared memory**
  - **Memory is physically distributed on individual modules**
  - **all memory is directly addressable to and accessible by any MSP in the system through the use of load and store instructions**



LSDA &Multiple Scattering Theory (MST)



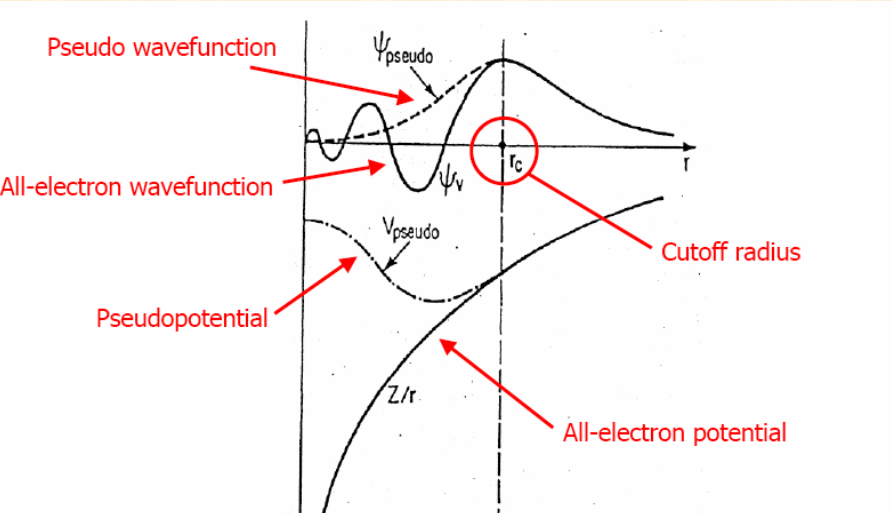
- Multiple Scattering Theory (MST)
  J. Korringa, Physica 13, 392, (1947)
  W. Kohn, N. Rostoker, PR, 94, 1111,(1954)
- MST Green function methods
  B. Gyorffy, and M. J. Stott, "Band Structure Spectroscopy of Metals and Alloys", Ed. D.J. Fabian and L. M. Watson (Academic 1972)
  S.J. Faulkner and G.M. Stocks, PR B 21, 3222, (1980)

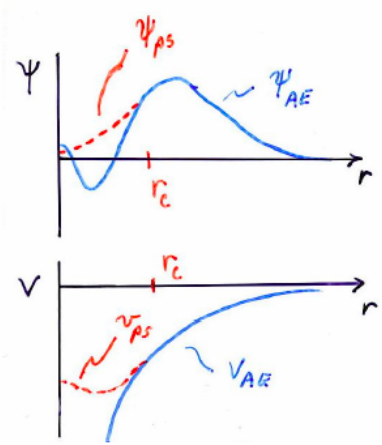
$$\{(\varepsilon + \frac{\hbar}{2m}\nabla^2)\hat{1} - \hat{V}^{eff}\}\hat{G}(\mathbf{r},\mathbf{r}';\varepsilon) = \hat{1}\delta(\mathbf{r}-\mathbf{r}')$$

$$n(\mathbf{r}) = \mathrm{Im}\frac{1}{\pi}\int d\varepsilon\, f(\varepsilon-\mu)\ \mathrm{Tr}\hat{G}(\mathbf{r},\mathbf{r}';\varepsilon)$$

$$\mathbf{m}(\mathbf{r}) = \mathrm{Im}\frac{1}{\pi}\int d\varepsilon\, f(\varepsilon-\mu)\ \sigma\mathrm{Tr}\hat{G}(\mathbf{r},\mathbf{r}';\varepsilon)$$

Pseudopotentials and Planewaves





Beyond r_c :

$\psi_{AE}(r) \rightarrow \psi_{ps}(r)$

$V_{AE}(r) \rightarrow V_{ps}(r)$

Also

$\varepsilon_{AE} = \varepsilon_{PS}$

- By construction V_{ps} has correct ε_{nl}
- Also want:
  - Norm conservation
  - Scattering properties remain pretty good for nearby ε_{nl}

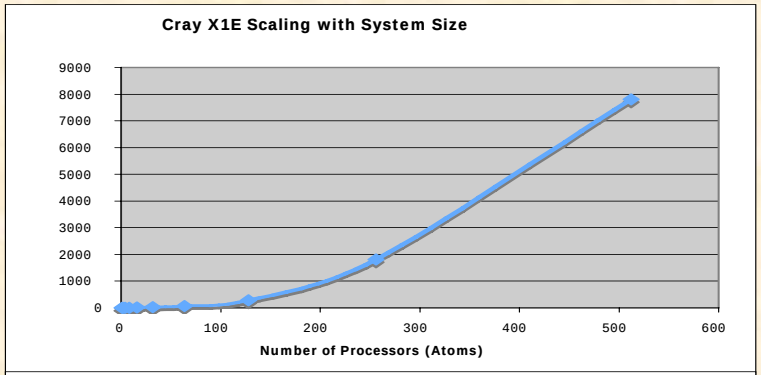
$$H\,\psi(G) = KE\,\psi(G) + (V\psi)(G)$$

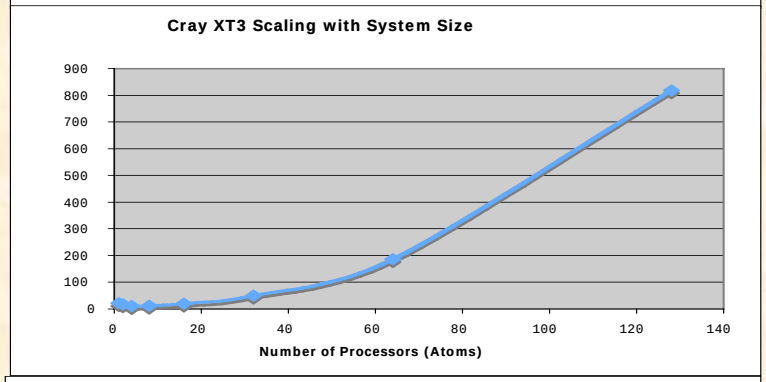
FFT ↓ FFT ↑

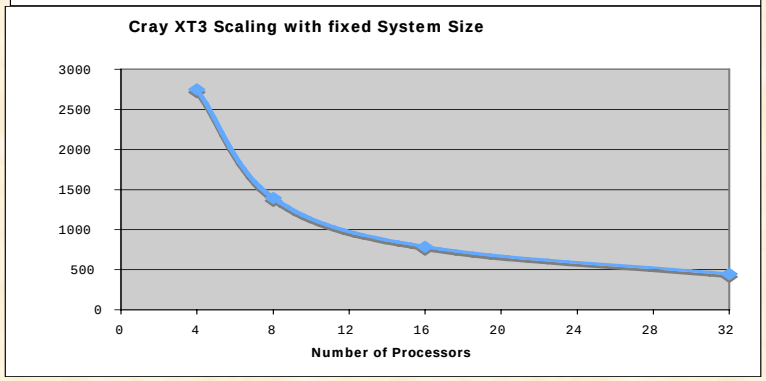
$$\psi(r) \longrightarrow (V\psi)(r)$$

Multiply on Real space mesh

Scaling on Cray X1E and XT3



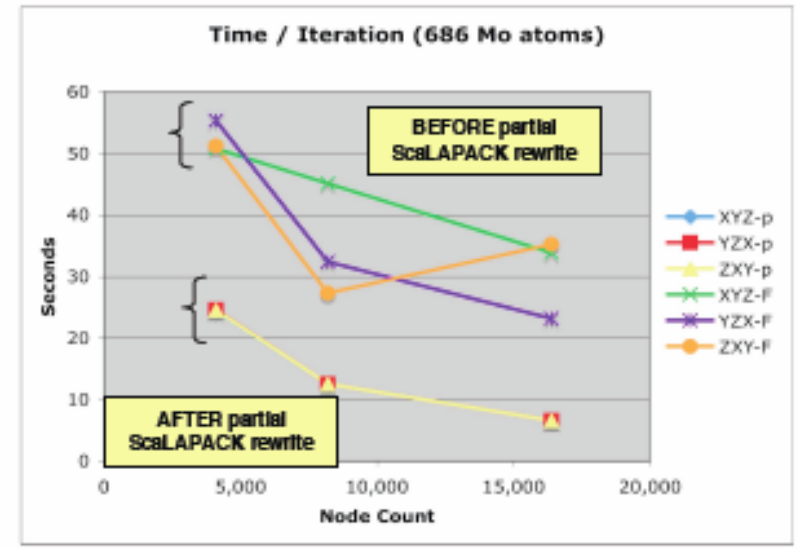


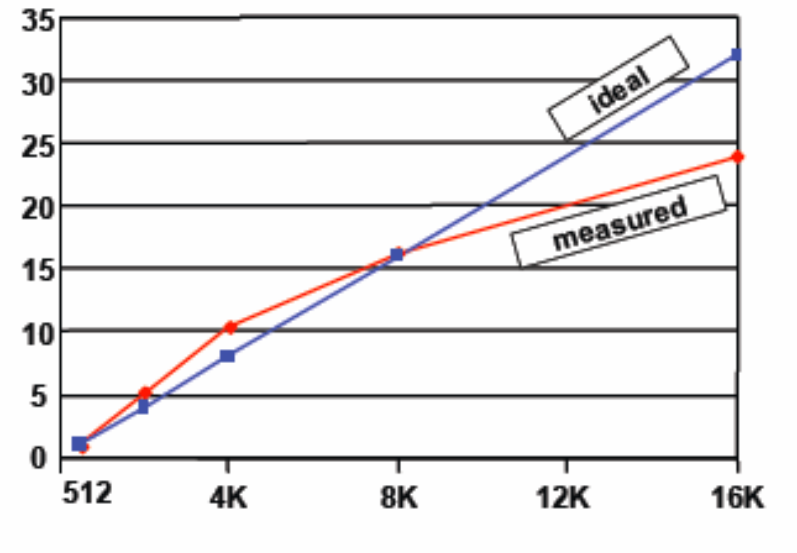


- **Vienna Ab-initio Simulation Package (VASP)**
  - **BCC Cu, 400 eV Plane wave cutoff**
- **Small system sizes**
  - **Forward and backward FFT**
- **Large system sizes**
  - **Davidson diagonalization**
  - **Forward and backward FFT**
- **Fixed number of plane waves**
  - **Changing the number of plane waves per processor**
- **Optimal density of atoms/node**
  - **For the Buckyball ~ 1.9 atoms/node**
  - **Thus, to run a 1000 atom system optimally would require 500 processors**

Qbox

- Qbox is a C++/MPI implementation of the planewave,
pseudopotential, ab initio molecular dynamics
It is developed at LLNL.
- Massively parallel C++ / MPI implementation with
  specialized 3D FFTs
  specialized ScaLAPACK
- 686-atom Mo solid and other heavy metal simulations are under way
- Scalability tests on BG/L show that Qbox can achieve a 3x speedup when solving a given problem
on 16384 nodes instead of 4096 nodes. This represents a 75% parallel efficiency. Further optimizations will provide even greater efficiency.

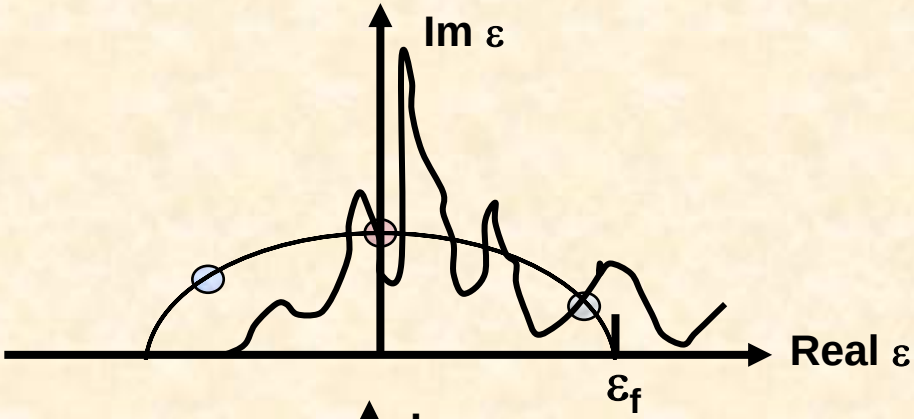




Algorithm Design for future generation architectures

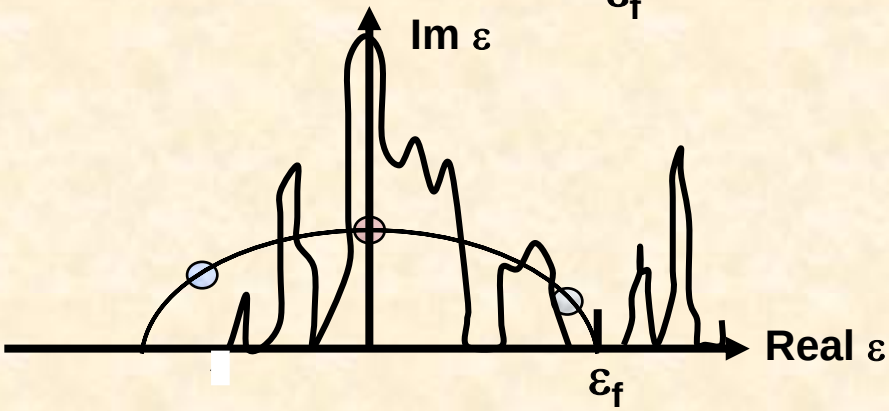
- More accurate
  - Spectral or pseudo-spectral accuracy
- Wider range of applicability
- Sparse representation
  - Memory requirements grow linearly
  - Each processor can treat thousands of atoms
- Make use of large number of processors
- Message-Passing
  - Each atom/node local message-passing is independent of the size of the system
- Time consuming step of model
  - Sparse linear solver
  - Direct or preconditioned iterative approach

Complex Energy Plane



ε_f is the highest occupied electronic state in energy

- Scattering is local since there are no states near the bottom of the energy contour
- Scattering is local since a large Im ε is equivalent to rising temperature which smears out the states
- Near ε_f scattering is non-local (metal)



Semi-conductors and insulators could work well since they have no states at ε_f

The scattering properties at complex energy can be used to develop highly efficient real-space and k-space methods

Multiple Scattering Theory

- Multiple scattering theory
  - Green function

$$G^{nm}(\vec{r},\vec{r}\,';\varepsilon)=\sum_{LL'}Z_L^n(\vec{r}_n;\varepsilon)\,\tau_{LL'}^{nm}(\varepsilon)Z_{L'}^m(\vec{r}_m';\varepsilon)-Z_L^n(\vec{r}_<;\varepsilon)J_{L'}^m(\vec{r}_>';\varepsilon)\delta_{LL'}$$

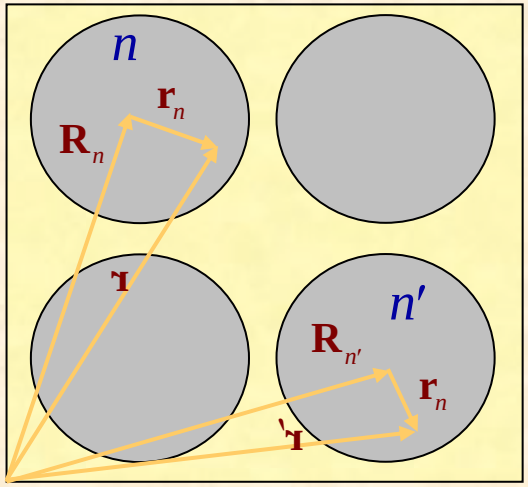
  - Scattering path matrix

$$\tau^{nm}(\varepsilon)=t_n(\varepsilon)\delta_{nm}+\sum_{n\neq m}t_n(\varepsilon)G(\vec{R}_n-\vec{R}_m)\,\tau^{nm}(\varepsilon)$$

$$\tau^{nm}(\varepsilon)=(M^{nm}(\varepsilon))^{-1}$$

$$M_{LL'}^{nm}(\varepsilon)=t_{LL'}^{-1}\delta_{LL'}\delta_{nm}-G_{LL'}(\vec{R}_n-\vec{R}_m)$$

Generalization of t-matrix. Converts incoming wave at site n into outgoing wave at site m in the presence of all the other sites



$$M(\varepsilon)=\begin{bmatrix} t_0^{-1} & -G^{01}(\varepsilon) & -G^{02}(\varepsilon) & -G^{03}(\varepsilon) \\ -G^{10}(\varepsilon) & t_1^{-1} & -G^{12}(\varepsilon) & -G^{1m}(\varepsilon) \\ -G^{20}(\varepsilon) & -G^{21}(\varepsilon) & t_2^{-1} & -G^{2m}(\varepsilon) \\ -G^{30}(\varepsilon) & -G^{31}(\varepsilon) & -G^{32}(\varepsilon) & t_m^{-1} \end{bmatrix}$$

$$G^{nm}(\varepsilon)=-\frac{1}{4\pi}\frac{e^{i\varepsilon^{1/2}\left|\vec{R}_n-\vec{R}_m\right|}}{\left|\vec{R}_n-\vec{R}_m\right|}$$

decay slowly with increasing distance
contain free-electron singularities

Screened MST Representation

Tight Binding Multiple Scattering Theory

- Notice that G_0 has no eigensolutions and decays rapidly for negative energies
- Need a reference system that supports no eigensolutions in energy range important to solid-state physics and chemistry(~+1 Ryd.)
- Embed a constant repulsive potential
- Shifts the energy zero to negative energies
- Rapidly decaying interactions
- Sparse representation

$$G^{nm,r}(\varepsilon) \propto e^{-(V^r-\varepsilon)^{1/2}|\vec{R}_n-\vec{R}_m|}$$

$$G^{nm}(\varepsilon) \propto \frac{e^{-\varepsilon^{1/2}\left|\vec{R}_n-\vec{R}_m\right|}}{\left|\vec{R}_n-\vec{R}_m\right|}$$

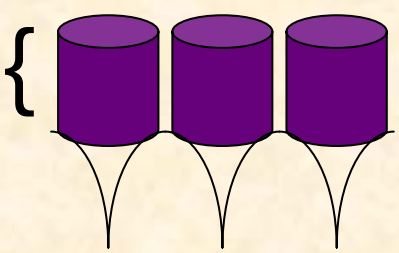
$$G = G_0 + G_0 t G_0 + G_0 t G_0 t G_0 = G_0 (I - t G_0)^{-1}$$

$$G^{-1} = G_0^{-1} - t$$

$$G^r = G_0 (I - t^r G_0)^{-1}$$

$$(G^r)^{-1} = G_0^{-1} - t^r \rightarrow G_0^{-1} = (G^r)^{-1} + t^r$$

$$G^{-1} = (G^r)^{-1} + t^r - t \rightarrow (G^r)^{-1} - \Delta t$$

2 Ryd. 

$$V^r \begin{cases} \text{Constant inside a sphere} & \acute{\mathbf{R}} \le \acute{\mathbf{R}}_{\mathbf{mt}} \\ 0 & \acute{\mathbf{R}} > \acute{\mathbf{R}}_{\mathbf{mt}} \end{cases}$$

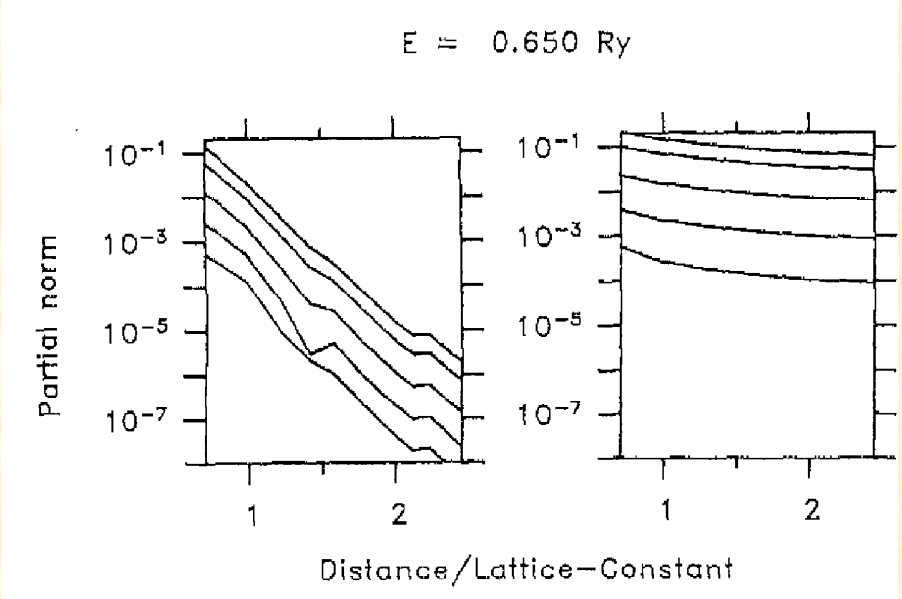
Screened Structure Constants

- Linear solve using m atom cluster that is less than the n atom system
- Easy to perform Fourier transform
  - K-space method

$$G^s(\varepsilon) = [I - t^s G^{free}(\varepsilon)]^{-1}$$

$$G(\vec{k},\varepsilon) = \sum_m G^{m,s}(\varepsilon) e^{i\vec{k}\bullet\vec{R}_m}$$

- Screened Structure Constants G_s on the left unscreened on the right
  - Screened structure constants rapidly go to zero, whereas the free space structure constants have hardly changed



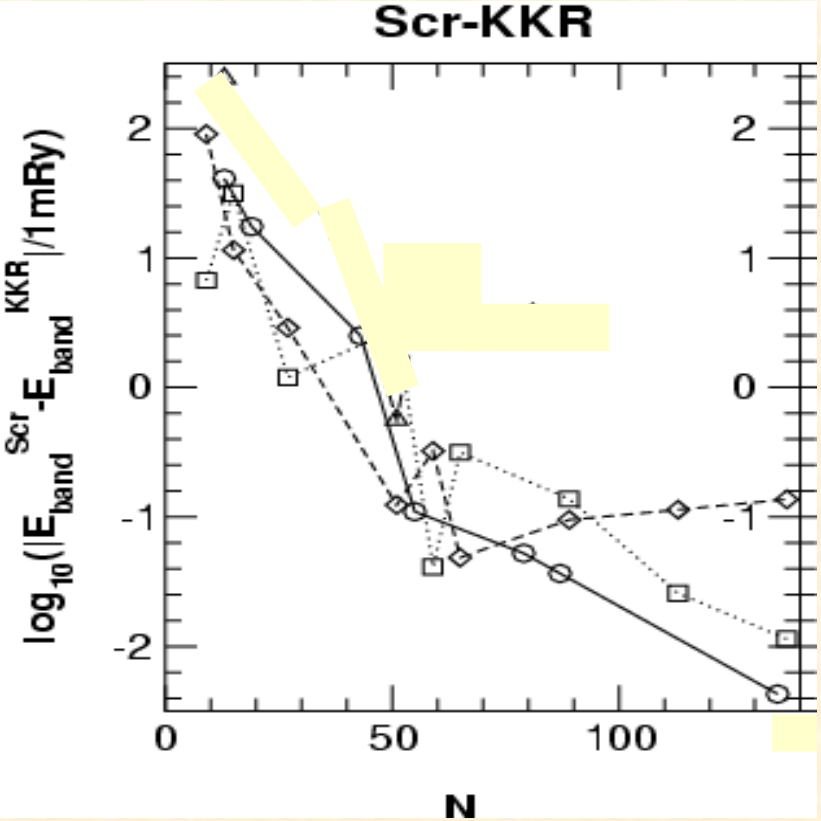
Screened MST Algorithm Design

- Linear scaling
  - Each node performs a fixed size local calculation
    - Thus each node performs the same number of flops
- Message-Passing
  - Each atom/node local message-passing is independent of the size of the system
- Time consuming step of model
  - Sparse non-symmetric iterative step
    - Highly parallel since main computation is a sparse matrix-vector or matrix-matrix operation
    - sparse BLAS level 2 or 3

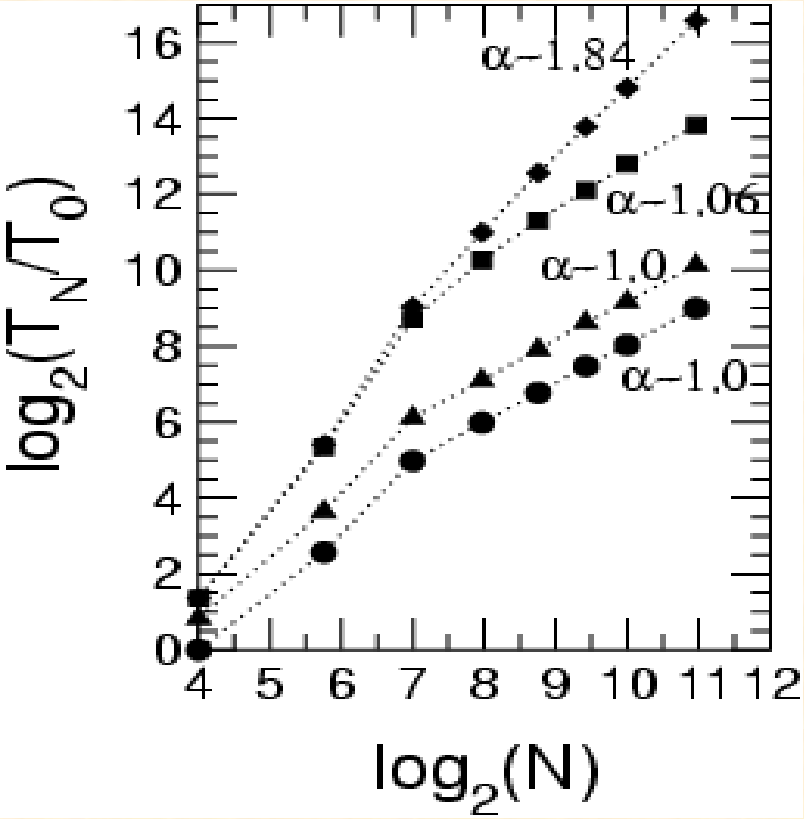
Screened MST Methods

- Formulation produces a sparse matrix representation
  - 2-D case has tridiagonal structure with a few distant elements due to periodicity
  - 3-D case has scattered elements
    - Mainly due to mapping 3-D structure to a matrix (2-D)
    - A few elements due to periodic boundary conditions
- Require block diagonals of the inverse of $\tau(\varepsilon)$ matrix
  - Block diagonals represent the site $\tau(\varepsilon)$ matrix and are needed to calculate the Green's function for each atomic site
- Sparse direct and preconditioned iterative methods are used to calculate $\tau^{ii}(\varepsilon)$
  - SuperLU
  - Transpose free Quasi-Minimal Residual Method (TFQMR)

Screened KKR Accuracy and Timing



□ fcc Cu
○ bcc Cu
◇ bcc Mo
△ hcp Co



Conclusion

- Initial benchmarking of the Screened MST method
  - SuperLU $N^{1.8}$ for finding the inverse of the upper left block of τ
  - TFQMR with block Jacobi preconditioner $N^{1.06}$ for finding the inverse of the upper left block of τ
- Extremely high sparsity (97%-99% zeros increases with increasing system size)
- Large number of atoms on a single processor
- Real-space/K-Space hybrid may provide the most efficient parallel approach for new generation architectres
- Single code contains both screened and unscreened methods
- Ideal for including DFT with Exact Exchange

Multiresolution chemistry objectives

- **Complete elimination of the basis error**
  - **One-electron models (e.g., HF, DFT)**
  - **Pair models (e.g., MP2, CCSD, …)**
- **Correct scaling of cost with system size**
- **General approach**
  - **Readily accessible by students and researchers**
  - **Higher level of composition**
  - **No two-electron integrals – replaced by fast application of integral operators**
- **New computational approaches**
- ***Fast algorithms with guaranteed precision***

How to "think" multiresolution

- **Consider a ladder of function spaces**

$$V_0 \subset V_1 \subset V_2 \subset \mathrm{L} \ \subset V_n$$

  - **E.g., increasing quality atomic basis sets, or finer resolution grids, …**

- **Telescoping series**

$$V_n = V_0 + (V_1 - V_0) + (V_2 - V_1) + \mathrm{L} \ + (V_n - V_{n-1})$$

  - **Instead of using the most accurate representation, use the difference between successive approximations**
  - **Representation on V_0 small/dense; differences sparse**
  - **Computationally efficient; possible insights**

Adaptive Refinement

- **To satisfy the global error condition**

$$\left\| f - f^{n} \right\|_2 \le \varepsilon \left\| f \right\|_2$$

- **Truncate according to**

$$\left\| d_l^n \right\|_2 \le 2^{-n/2} \varepsilon \left\| f \right\|_2$$

- **This is rather conservative – usually use**

$$\left\| d_l^n \right\|_2 \le \varepsilon$$

Separated form for integral operators

$$T * f = \int ds K(r-s) f(s)$$

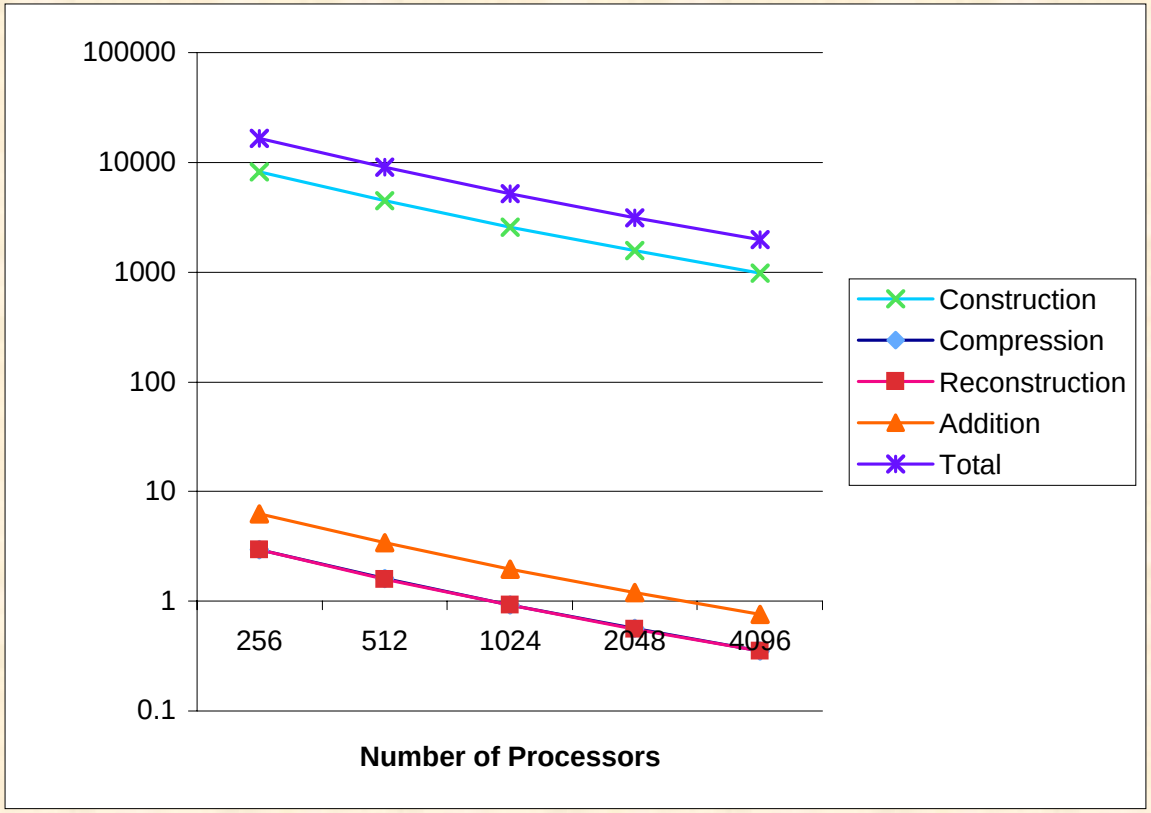
- **Approach in current prototype code**
  - **Represent the kernel over a finite range as a sum of Gaussians**

$$K(r) = \sum_i \omega_i e^{-t_i r^2} + O(\varepsilon)$$

  - **Only need compute 1D transition matrices (X,Y,Z)**
  - **SVD the 1-D operators (low rank away from singularity)**
  - **Apply most efficient choice of low/full rank 1-D operator**
  - **Even better algorithms not yet implemented**

Timing and Scaling on Cray XT3

- **4096 Cu atoms**
- **Displays near linear scaling with increasing system size**
- **Working with localized orbitals**
  - **O(1) application of operators to one orbital**
  - **O(N) computation of Coulomb potential**
  - **O(N) computation of Fock-like matrices**
  - **More robust convergence**



Summary

- **Multiresolution analysis provides a general framework for computational chemistry**
  - **Accurate and efficient with high-level composition**
  - **Multiwavelets provide high-order convergence and readily accommodate singularities/boundary conditions**
  - **General framework readily accessible to researchers**
  - **Real impact will be application to many-body models**
- **Separated form for operators and functions**
  - **Critical for efficient computation in higher dimension**
- **Precision is guaranteed**
  - **Excited states, non-linear response, …**
- **Near total rewrite in C++**
  - **Two-levels of parallelism targeting massively parallel computer using multi-processor nodes**
  - **In anticipation of highly-threaded processors**