"Scalable First Principles Methods for Electronic Transport"

W.A. Shelton*, V. Meunier, E. Aprà and M.L. Tiago
Oak Ridge National Laboratory

Summary

Developing scalable methods, able to fully utilize petascale systems, enables predictive simulations of *entire* device structures from first principles, thereby revolutionizing the design of molecular and nanoscale sensors and electronic devices. We have developed a scalable first principles approach for quantum transport where we have developed a multi-level parallel sparse iterative method for solving the non-equilibrium Green's function and implemented and optimized a first principles electronic structure method. Both methods exhibit high scalability and serial performance with increasing system size. This *virtual* design would impact many areas critical to DOE's needs, such as the development of specific biosensors with nearly single molecule detection limit, and of ultra-dense, ultra-fast molecular-sized electronic components, with very small power requirements and persistent, reprogrammable memories. Unfortunately, current first principles methods and software can only routinely handle a few hundred atoms.

A Sparse Matrix iterative Multi-level parallelization quantum transport method: In the non-equilibrium Green function calculations of transport properties, the most time-consuming part is to calculate the Green functions by inverting a matrix. By taking into consideration the fact that the Hamiltonian and overlap matrices are of sparse block-tridiagonal form a very computationally efficient sparse non-symmetric iterative algorithm has been developed to calculate the Green functions. The inversion is performed block by block and the algorithm scales linearly for both performance and memory with increasing system size. To achieve high performance on a massively parallel supercomputer requires multiple levels of parallelism and a reorganization of the algebraic formulation. In our current implementation of the transport code, we can have four levels of parallelization: (a) *Parallelization over energy integration*. The energy integration typically requires 100-500 energy points, each energy point represents a completely independent process. (b) *Parallelization over the matrix operation*. On each energy processor group, the matrices are distributed on a two-dimensional processor grid. In the block-tridiagonal representation, the matrices are decomposed into 5 blocks where each block size is ~1200. This produces an algorithm with a scalability of over 80% up to 1024 processors. Further optimization is possible by reorganizing the "all-reduce" communication. (c) *Parallelization over bias in I-V calculations*. To produce the I-V curve at least 20 values for the bias are needed and each bias

*865-576-7932, sheltonwajr@ornl.gov

calculation represents an independent process.

Scalable First principles method: The inclusion of light absorbing antenna chromophores through a covalent linkage combined with the extended π electrons of a carbon nanotube can constitute an ideal nano-assembly for generating singlet excited energy and its conversion to chemical energy. This is an example of system that is of significant importance for solar energy applications. This system represents a true computational challenge for a first principles approach since the system ranges in size from 1532 atoms to over 6000 atoms. To address the electronic structure we have implemented the Pseudopotential Algorithm for Real-Space Electronic Calculations (PARSEC) (http://www.ices.utexas.edu/parsec/index.html) on the Cray XT4 at the National Center for Computational Sciences located at Oak Ridge National Laboratory. A considerable amount of optimization had to be performed on the code to enable the simulation of these large scale systems. The optimization included a new parallel distribution of the real-space grid that allowed for better memory management and parallel performance. Without these computational enhancements the simulation becomes untenable due to poor memory management and scaling with system size. In figure-1 is the strong scaling plot as of function of number of processors for the 1532 atom system. At 4096 processors the codes runs at a sustained performance of 7.5 Tflops (36% peak performance) with 50% scalability. This plot indicates that the optimal performance is at 2048 processors where the code achieves a sustained performance of 7.5 Tflops (72% peak performance) and a scalability of over 80%. This result defines the decomposition size per processor (decomposition size on 2048 processor) for obtaining both high scalability and high serial performance as a function of increasing system size. In fact, the sustained computational performance of PARSEC at 4096 processors is 1.74 times that of QBox (http://eslab.ucdavis.edu).

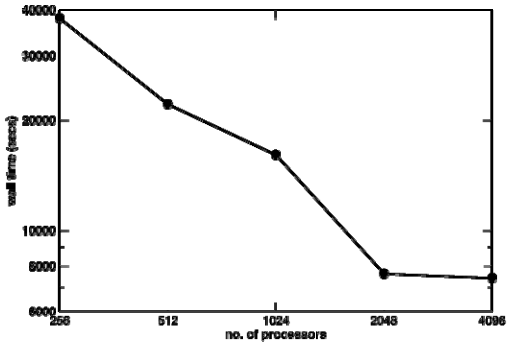

Fig. 1: Parallel scaling of PARSEC. The benchmark runs compute the DFT energy of a 3541 silicon nanocluster ($Si_{2713}PH_{828}$).

For further information on this subject contact:

Dr. W.A. Shelton
Oak Ridge National Laboratory
sheltonwajr@ornl.gov
865-576 -7932

Or

Dr. Anil Deane
Applied Mathematics Research Program
Office of Advanced Scientific Computing
Phone: 301-903-1465
deane@ascr.doe.gov