Molecular simulations: Electronic structure calculations

Electronic structure simulations are broadly used throughout the HPC community to address problems in molecular science, materials and biology. In all cases, the Schrödinger equation for the motion of electrons is solved using a wide range of approximations from semi-empirical molecular orbital (SEMO) and density functional theory (TB-DFT) approaches to weakly parameterized DFT to correlated molecular orbital theory to quantum Monte Carlo to exact solutions such as full configuration interaction (CI) with a finite basis set. The codes used for such simulations are often coupled with molecular dynamics (MD) codes to study the motion of the nuclei as a function of time by either solving Newton's equation of motion or by solving the Schrödinger equation of nuclear motion. The MD calculations can use derivatives from electronic structure calculations (ab initio = aiMD), fitted potential energy surfaces from such calculations, classical force fields (molecular mechanics = MM), or some combination (QM/MM). The electronic structure calculations can treat additional effects such as those due to relativity either explicitly or implicitly, for example, the latter with effective core potentials. The computational scaling ranges from pseudo linear in some cases for electronic structure calculations (although these all involve a matrix diagonalization step of $O(N^3)$) to $O(N^7)$ for CCSD(T) to factorial for full CI where N is the number of basis functions. An issue is that the high accuracy calculations require large basis sets to converge to the complete basis set (CBS) limit although new efforts to explicitly include electronic coordinates in the wavefunction are reducing the size of the basis set needed to reach the CBS limit, albeit at increased cost.

There are many electronic structure codes which can perform electronic structure calculations including Gaussian09, NWChem, GAMESS, MOLPRO, MOPAC, DIRAC, VASP, Elk, Siesta, CP2K, and QMPot. The most popular code world-wide is Gaussian but it does not perform well on the latest HPC architectures as it was not designed for such systems and originates from about 1970 from the Pople group. Our focus is NWChem, which was initiated under the EMSL project in the early 1990's. At that time, it was decided to design an electronic structure and MD code for massively parallel architectures. When NWChem development was initiated, the predominant architecture for HPC was Cray computers with fast vector but limited number of processors and global memory access. Chemists were just beginning to use DFT to solve chemical problems due to the advent of new exchange-correlation functionals so the plan of a massively parallel code which also had DFT capabilities was a paradigm shift. The development of NWChem was led by staff in the EMSL at PNNL and included chemistry software developers, applied mathematicians, and computer scientists, and a team of internal and external users to provide input about the functionality. The main parts of the code were developed in Fortran and C++. The team developed their own tools in the ParSoft toolkit including Global Arrays and the Parallel Eigensolver (PeIGS) to maintain portability to different architectures and to optimize performance for each new architecture separate from the main solvers in NWChem. The development of the parallel performance toolkit enabled NWChem to be developed by a community of developers not all of whom were located in the EMSL. Tools were developed to integrate external contributions including nightly builds to check for compatibility to ensure that all parts of the code worked together and that issues could be found quickly. The initial distribution model was a no cost license to any user with the source and sole general purpose copy maintained in the EMSL. This was done to maintain an always 'correct' and working version; this model has since morphed into an open-source distribution, but the EMSL still

maintains a core version. The project has involved $10M's with about $2M/year currently being spent on the project. The project involved 100's of people including graduate students, postdoctoral fellows, university faculty, and core scientific and technical staff at national laboratories. There have been as many as 8 to 10 FTEs in the NWCHem group and there are currently 5 to 6 FTEs at PNNL.

NWChem has a wide range of capabilities that each rely on a different mathematical formulations. The high accuracy coupled cluster relies on DGEMM and large tensor algebra, coupled with the calculation of complex Gaussian basis function derived integrals. The plane wave aiMD relies on parallel FFT and matrix-vector and matmul operations. Both methods require iterative solvers to obtain self-consistency in the calculations. The parallel programming model used within NWChem is different for the coupled cluster methods and the plane wave methods. The coupled cluster methods utilize the Global Arrays Toolkit with libraries that been run successfully on DOE leadership computing platforms. The toolkit uses a global address space programming model, utilizing one-sided communication where possible. The plane wave method uses MPI for its communication framework. The latest NWChem 6.5 release has a coupled cluster capability that can be run on NVIDIA GPUs using IIRC and CUDA. NWChem can be compiled with standard Fortran and C compilers. NWChem uses standard BLAS, LAPACK, SCALAPACK which are available on current leadership computing platforms. The CCSD(T) single point calculations can scale to 100K cores. For example, hard scaling of the high accuracy single-reference coupled cluster with perturbative triples (CCSD(T)) method for a large organic system to 210,000 processors on the ORNL Jaguar machine is close to linear. Benchmarks on ORNL's Cray-XK6 have demonstrated scaling to at least 80,000 processor cores. Graphs showing the strong scaling can be found on the NWChem benchmark site(http://www.nwchem-sw.org/index.php/Benchmarks) as well as the results of benchmarking done for the CUDA enabled coupled cluster method. The plane wave aiMD module for a 576 atom system using exact exchange shows scalability beyond 90,000 processor cores on the NERSC Hopper supercomputer.

We are currently developing (ANL and UA) with ANL LDRD funds a fast parallel eigensolver for SEMO and TB-DFT methods for very large numbers of molecules to aid in materials synthesis design. The Shift-and-Invert Parallel Spectral Transformations (SIPs) is a computational approach to solve sparse eigenvalue problems. The largest example studied is a 128,000 atom nanotube for which ~330,000 eigenvalues and eigenfunctions are obtained in ~190 seconds when parallelized over 266,144 Blue Gene/Q cores. The basic software stack is shown in Figure 1. This effort has involved a postdoctoral fellow with the help of 3 staff at ANL and one faculty member from UA.

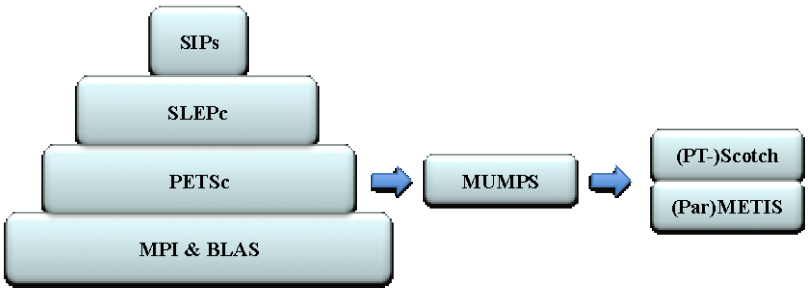

Figure 1. Software structure of SIPs. The layered structure indicates the requirements of the individual software packages, and arrows indicate the interfaces to external libraries.