

HP systems in computational chemistry: part 2

Computational case study series

Gaussian case study A: predicting infrared (IR) and Vibrational Circular Dichroism (VCD) spectra of large chiral molecules using *ab initio* Density Functional Theory (DFT)



Abstract.....	2
Introduction.....	2
VCD calculations	3
Optimum performance for large problems	4
Symmetric multiprocessing (SMP) servers.....	4
Clusters.....	4
Comparative results	5
Parallelism.....	5
Memory requirements.....	5
Computing architectures	6
Conclusion.....	7
For more information.....	8

Abstract

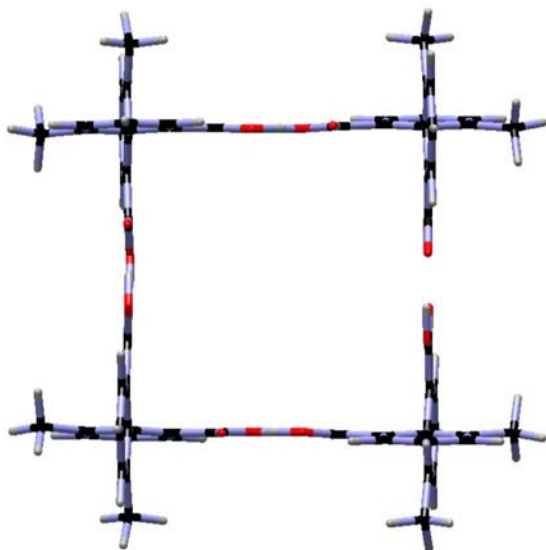
A collaborative study by USC, Gaussian Inc. and HP on VCD calculations of the supramolecular tetramer of 2,2'-dimethyl-biphenyl-6,6'-dicarboxylic acid (BT) indicate that optimizing computational variables of memory, architecture, and parallelism can significantly improve computational performance, critical for solving large problems efficiently.

Introduction

In this paper, we focus on a particular *ab initio* method, Density Functional Theory (DFT). This method, first developed by Nobel Laureate Walter Kohn et al., is based on the axiom that the total energy of a many-particle system can be explicitly and uniquely expressed as a functional of the total charge density.

One of the uses of DFT is to elucidate 3D molecular structure and properties. P.J. Stephens et al. at the University of Southern California (USC), in collaboration with Gaussian, Inc., have been developing new theoretical methods based on DFT to study the stereochemistry of chiral organic molecules, especially those of pharmacological and biological relevance, via the analysis of their Vibrational Circular Dichroism (VCD) spectra. F.J. Devlin and P.J. Stephens are currently extending the application of DFT to the field of supramolecular chemistry. In this case study, they predict the VCD spectra of the supramolecular tetramer of 2,2'-dimethyl-biphenyl-6,6'-dicarboxylic acid (BT) and compare it against experimentally determined spectra from the M. Urbanova lab (Prague) to show that this tetramer is the dominant structure at high concentrations in non-polar solvents.

The cyclic tetramer of the R enantiomer of the chiral molecule 2,2'-dimethyl-biphenyl-6,6'-dicarboxylic acid.



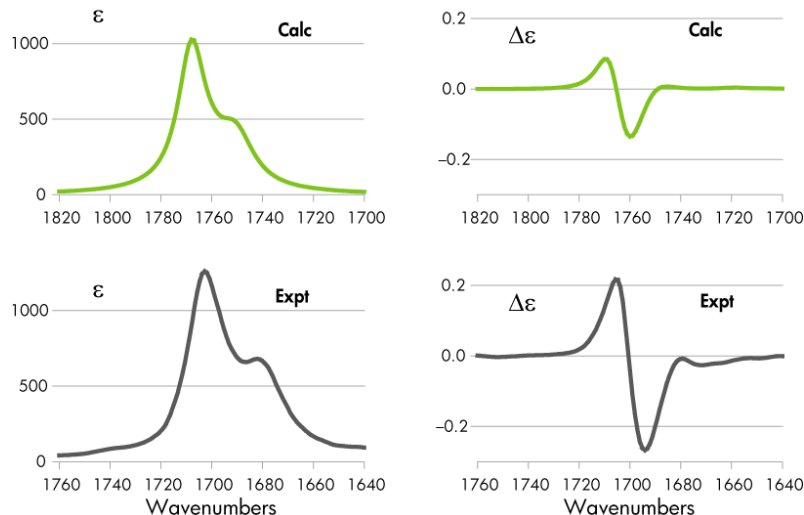
HP, working with Devlin and Stephens and Cheeseman of Gaussian, Inc., tested the computational requirements of the VCD calculations by running the BT problem on a number of different HP computational platforms, varying computing architecture, memory size, and scaling, and symmetric multiprocessing vs. clusters.

VCD calculations

Ab initio DFT calculations were carried out using the programs *Gaussian 98* and *Gaussian 03* (see sidebar below). The B3LYP functional is used throughout, together with the basis set 6-31G* (for a total of 1312 basis functions). Harmonic vibrational frequencies, dipole strengths, and rotational strengths are obtained from Hessians (force fields), Atomic Polar Tensors (APTs), and Atomic Axial Tensors (AATs). AATs are calculated using Gauge-Invariant (Including) Atomic Orbitals (GIAOs); as a result, rotational strengths are origin-independent. IR and VCD spectra are obtained from vibrational frequencies, dipole strengths and rotational strengths, assuming Lorentzian band shapes.

Gaussian 98 and the current version, *Gaussian 03*, are the products of Gaussian, Inc. They are a series of electronic structure programs derived from the basic laws of quantum mechanics. *Gaussian* predicts the energies, molecular structures, and vibrational frequencies of molecular systems, along with numerous molecular properties derived from these basic computation types. It can be used to study molecules and reactions under a wide range of conditions, including both stable species and compounds which are difficult or impossible to observe experimentally, such as short-lived intermediates and transition structures. Continually developed and supported, *Gaussian* is used by chemists, chemical engineers, biochemists, physicists, and others for research in established and emerging areas of chemical interest.

Figure 1: IR spectra (left panels) and VCD spectra (right panels) for the BT tetramer



The predicted IR and VCD spectra of BT in the C=O stretching region are shown in Figure 1. The excellent agreement confirms that the experimental spectra originate in the BT tetramer.

Optimum performance for large problems

DFT calculations can be too large to solve on a single processor system and would take too long on a small computing system. They are moderately parallel so they can take advantage of either symmetric multiprocessing or cluster computing with TCP Linda software (see sidebar below) for faster job completion.

TCP Linda is a coordination language: a standard for parallel programming in C and Fortran from Scientific Computing Associates (SCA). It provides a simple yet complete command set that enables process creation, synchronization, and communication. Every Linda software system employs powerful application optimization techniques and carefully tuned, architecture-specific run-time systems. Through many years of collaboration between SCA and Gaussian, TCP Linda technology is embedded in *Gaussian 03* to enhance its parallel performance.

In addition to parallelism, other computational variables, such as available memory or processor architecture, could have a large impact on overall performance. The characteristics of the *Gaussian 03* DFT module, in this case, as well as the specific research problems being solved, dictate how efficient and fast a particular computing solution can be.

To study the optimal balance of computational variables for running DFT calculations using the BT VCD calculation as an example of a large quantum mechanics problem, HP repeated the calculations on the following computing systems:

Symmetric multiprocessing (SMP) servers

- HP Integrity Superdome (1-64 Intel® Itanium processors, 1.5 GHz) running HP-UX, with 1920 MW (Megaword*) memory
- HP Integrity rx7620 Server (1-8 Intel Itanium processors, 1.5 GHz) running Linux® (Red Hat Enterprise Linux AS release 3), with 75, 480, or 960 MW memory
- HP Integrity rx2600 Server (1-2 Intel Itanium processors, 1.5 GHz) running Linux (Red Hat Linux Advanced Server release 2.1AS), with 240 MW memory
- HP ProLiant DL140 Server (1-2 Intel Xeon™ processors, 3.06 GHz) running Linux (Red Hat Enterprise Linux AS release 3) with 240 MW memory
- HP ProLiant DL145 Server (1-2 AMD Opteron processors, 2.2 GHz) running Linux (Red Hat Enterprise Linux AS release 3) with 480 MW memory

Clusters

- Cluster of HP Integrity rx2600 Servers (each with 2 Intel Itanium processors, 1.5 GHz) running Linux (Red Hat Linux Advanced Server release 2.1AS) with 240MW memory/server and Gigabit Ethernet interconnect

(Compilers: Intel v7.1 compilers used to build *Gaussian* on systems with Itanium processors/Linux. Intel's MKL v6.1 and PGI v5.1 compilers used to build *Gaussian* on Intel Xeon and AMD Opteron.)

* 1 Megaword = 8 Megabytes

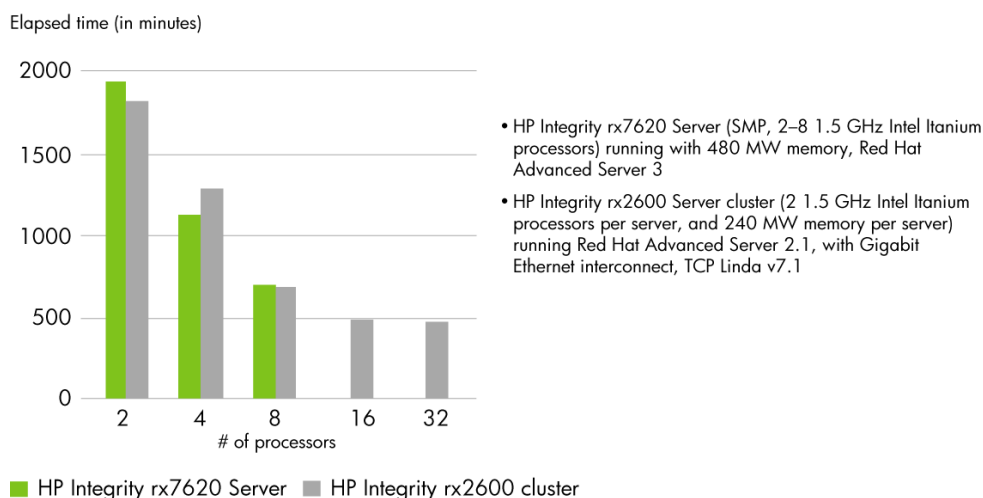
Comparative results

Parallelism

One of the key mechanisms for reducing the overall computing time of long jobs is to take advantage of any parallelism inherent in the software. Both SMP and cluster styles of computing are applicable, though they vary in their efficiency of scaling depending on the nature of the job. Clusters are usually very efficient if the processes within a job do not have to communicate significantly with each other. Scientific Computing Associates' (SCA's) TCP Linda software, used in the cluster implementation, has been optimized to run *Gaussian*. Running the BT VCD calculation in parallel does reduce job completion time: the 8-processor runs are about 2.7–2.8 times faster than the 2-processor runs. Performance appears to plateau between 16 and 32 processors. This agrees with the general rule of thumb for *Gaussian* scaling.

In this case, cluster performance and SMP performance are about the same. However, other problem sets might behave differently.

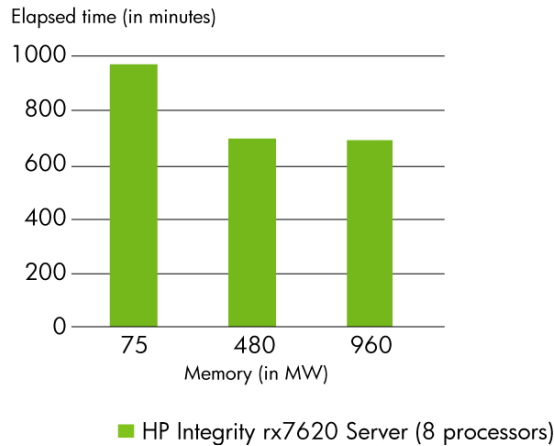
Figure 2. Results of the BT VCD calculation run in parallel



Memory requirements

Symmetric multiprocessing systems have the benefit of shared memory, accessible by all the processors in the system. This can be quite useful for certain *Gaussian* jobs writing large scratch files. If there is insufficient memory, a significant amount of time is wasted on I/O, moving data from disk to memory. In clusters, where generally smaller systems are pooled together and memory is not shared between systems, insufficient memory in individual systems can limit performance.

Figure 3: For the BT VCD calculation, a look at performance relative to available memory shows that the amount of memory clearly affects completion time.

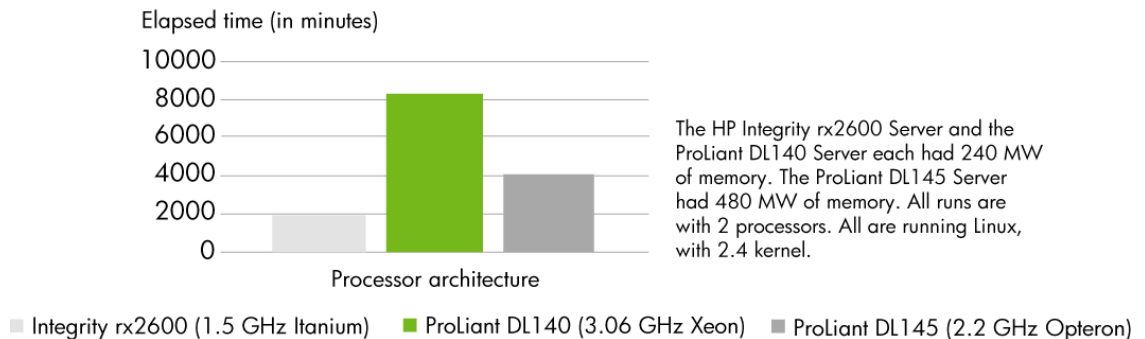


By varying the amount of memory from 75 MW (600 MB) to 480 MW (3.84 GB), we see a greater than 35% improvement in elapsed time with no additional improvement in performance at 960 MW (7.7 GB). Similar results were found on the HP AlphaServer platform.

Computing architectures

Performance differences often appear because of the native differences in processor architectures and the clock speeds of the current implementations. In Figure 4, HP servers based on 1.5 GHz Intel Itanium processors, 3.06 GHz Intel Xeon processors, and 2.2 GHz AMD Opteron processors are compared.

Figure 4. Comparison of processors



The BT calculations using 2-processor systems show the HP Integrity rx2600 Server (Intel Itanium processor) to be over 2 times faster than the ProLiant DL145 Server (AMD Opteron processor) and 4 times faster than the ProLiant DL140 Server (Intel Xeon processor). The 64-bit version of *Gaussian 03* is used for both the Itanium processor-based and the Opteron-based runs while the Xeon-based run used a 32-bit implementation of *Gaussian 03*.

Conclusion

Using the large *Gaussian 03* calculations of VCD spectra of BT as an example, we have found that memory and processor architecture have a large effect on performance, as does parallelism. Increasing the number of processors (to 16 processors) improves performance significantly. In this case, the parallel performance difference between symmetric multiprocessing and a cluster is small.

Additional comparisons are needed for different types of jobs to improve our characterization of *Gaussian 03* requirements. HP, USC, and Gaussian are continuing their joint efforts to characterize *Gaussian 03*. HP also works with SCA as a partner to ensure the successful implementation of Linda software on HP platforms.

Look for other white papers in this series in 2005.

For more information

www.hp.com/go/lifesciences

© 2004 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Intel Itanium, and Xeon are trademarks or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries. Linux is a U.S. registered trademark of Linus Torvalds.

5982-9951EN, 11/2004

