Available online at www.prace-ri.eu

Partnership for Advanced Computing in Europe

HPC Programming to Generate Electron-molecule Resonance Data for DNA Radiation Damage Studies.

Andrew Sunderland*, Martin Plummer

STFC Daresbury Laboratory, Warrington, United Kingdom

Abstract

DNA oxidative damage has long been associated with the development of a variety of cancers including colon, breast and prostate, whilst RNA damage has been implicated in a variety of neurological diseases, such as Alzheimer's disease and Parkinson's disease. Radiation damage arises when energy is deposited in cells by ionizing radiation, which in turn leads to strand breaks in DNA. The strand breaks are associated with electrons trapped in quasi-bound 'resonances' on the basic components of the DNA. HPC usage will enable the study of this resonance formation in much more detail than in current initial calculations. The associated application is UKRmol [1], a widely used, general-purpose electron-molecule collision package and the enabling aim is to replace a serial propagator (coupled PDE solver) with a parallel equivalent module.

1. Introduction and goals of the project

Energy deposited in cells by ionizing radiation is then channeled into production of free secondary electrons. Reactions of such electrons induce substantial yields of strand breaks in DNA and RNA. The strand breaks are associated with electrons trapped in quasi-bound 'resonances' on the basic components of the DNA. These resonances may be found and studied using the general-purpose low-energy electron-molecule collisions package UKRmol [1]. However while the main collision (scattering matrices and cross sections) code has been parallelised, the particular modules for resonance finding and fitting remain serial.

This project has optimized and parallelised the resonance modules, which will provide enhanced computational resources for collaborating research groups to study biological molecular resonance formation in much more detail than current initial (though impressive) calculations [2,3,4,5] on DNA and RNA base molecules. In addition to the main thrust of this work on applications directly related to cell radiation damage and strand-breaking, the general application of the enhanced package will have impact in other socio-economic fields (for example, dissociative recombination studies relevant to atmospheric and aeronautical physics and chemistry).

2. Structure and technical aspects

The specific part of UKRmol adapted is the module TIMEDEL, which explores and illuminates resonance features of electron molecule interactions. This module is called for various molecular geometries and sets of incident electron energies. The full R-matrix simulation takes in several separate stages. Configuration space is divided into

Corresponding author, e-mail address: andrew.sunderland@stfc.ac.uk

two regions by a sphere which contains all the 'target' electrons. A full CI calculation involving continuum states and orbitals as well as bound states and orbitals takes place within the sphere. This is independent of scattering energy and is performed once for each geometry: it principally involves the setting-up and diagonalization of a large Hamiltonian matrix.

The outer region calculations take place separately, and quite possibly on another machine. The theory is of one electron moving in a multichannel potential arising from all of the CI states and channels included in the inner region calculation. Various options may be chosen to calculate collision parameters and other quantities. The TIMEDEL module can be run as a self-contained outer region package, particularly for the intense resonance fitting tasks of interest to our collaborators, and this is how we have been testing it. The full outer region package consists of a wrapper main program that calls various alternative or combined modules from a set of modules which make up the main outer region code. A more limited version of TIMEDEL can be combined with, for example, an initial solution of the multichannel collision problem using a predefined energy grid, but our collaborators prefer the automatic energy grid generation as resonances are searched for (each energy point requires solution of the multichannel collision problem).

Stage 1 of the optimization work introduced an overall parallelism for these calculations with a parallel harness to run distinct geometries and ranges, invoking the serial code simultaneously for each dataset in separate folders. Several sets of scripts and codes were developed for the control programs, as our two alternative target hardware platforms (IBM BlueGene/Q and IBM iDataplex) necessitated different approaches:

1. MPI-based C control program incorporating the *exec()* family of Linux functions which enable the replacement of current process images with new process images.
2. MPI-based Fortran 2003 control program with *iso_c_binding* calls to C functions for command line operations.

The first approach is suitable for most HPC systems with Linux kernels operating on the application nodes. However IBM utilize a 'Lightweight Compute Node Kernel' on the Blue Gene application nodes that do not support the required functions, hence the development of the second approach for this and other platforms. We note that the second approach fits reasonably clearly with the structure of the code, as the overall MPI control and *iso_c_binding* calls to C code only required editing of the 'wrapper' main program from which the TIMEDEL modules are called, with the possibility of introducing further sub-communicators for standard MPI operations within the modules.

Stage 2 of the parallelisation work attempted to identify inner loops in the TIMEDEL module suitable for MPI or OpenMP parallelisation. An analysis of the code identified the automated resonance search as having a suitable loop structure for parallelisation. Here the energy grid is sub-divided into clusters of energies in which resonances are expected, in particular around the thresholds for inelastic collisions into excited states of the molecule. It is also possible to input targeted grids of energies at which an earlier run has located evidence of resonances. The resonance location and subsequent fitting in each of these subdivisions is independent and to first order, equivalent, and may be parallelised. As a first stage of parallelisation of this nature we have split the energy grid appropriately with an MPI communicator. Further work will improve load-balancing within the sub-divisions by allocating the MPI tasks according to a more precise calculation of the number of energies requiring K-matrices around each resonance.

To facilitate progress in the time available, the PRACE WP7 developments have been tested on a relatively simple and known case (e-N_2^+ dissociative recombination resonances [6,7]) in conjunction with scientists from the AMOPP research group at University College London (UCL) [8]. Studies have been made to devise a 'production' job submission algorithm that can optimize the ratio of cores used for the 'external' harnessing and the 'internal' programmed parallelisation, with final verification tests on a DNA/RNA base (adenine, guanine, uracil) to be set up following this project. The parallelised code will then be used to enable new more realistic and detailed DNA/RNA base resonance studies, which are planned but are not currently feasible.

The PRACE Tier-0 machine Juqueen BlueGene/Q [9] along with the Hartree Centre [10] systems Blue Joule (IBM BlueGene/Q) and Blue Wonder (IBM iDataplex) in the UK have been targeted for the development work, optimization and benchmarking.

3. Performance Results

All performance tests have been undertaken on a large set of geometries from the e-N_2^+ dissociative resonance test case provided by the AMOPP Group at UCL.

3.1. Stage 1 Parallelisation

Benchmarking the performance of the parallel harness over multiple geometries associated is complicated by the varying characteristics of the computation across geometries. The different geometry calculations contain energy thresholds at different levels and this determines both the resolution of the scattering energy grid, the number of scattering energy grid points and the computational load in the 'fitting' function. This is an intrinsic load-balance problem that will need to be addressed in future versions of the harness. For now, the performance results for runs on less than 1024 cores are therefore sampled from across the full geometry set range in an attempt to be as representative as possible.

Using the provided datasets, parallel performance tests on the Hartree Centre IBM BlueGene/Q have been undertaken and analysed. The Blue Joule system consists of 6 racks, each rack containing 1,024 16-core, 64 bit, 1.60 GHz A2 PowerPC processors, with 16GB of system memory per node. The high-level parallelisation involves looping over data representing 1024 different geometries for dissociative e-N_2^+ scattering calculations. The code is compiled with thread-safe mpi-enabled versions of the IBM compilers *mpixlc_r, mpixlf95_r,* with optimization level set at *–O3*. The IBM ESSL library is linked with the code to provide highly optimized mathematical functions and linear algebra operations. In each case the amount of geometries increases with core count and therefore the weak scaling properties of this approach are reported in Table 1.

Number of Blue Gene/Q Cores Used	Number of BlueGene/Q Nodes Used	Geometry Calculations (full energy range)	Time to Completion (secs)	Time taken relative to 16 core case
16	1	16	4781	1
64	4	64	6637	1.39
256	16	256	7106	1.49
1024	64	1024	7839	1.64

Table 1. Speed-up obtained from Stage 1 parallelisation of the 1024 geometry calculations

For this large-scale dataset, the approach scales well up to 1024 cores on the IBM BlueGene/Q, with the 1024 geometry calculation time around 64% slower than the 16-geometry calculation. Around 3 GBytes of output is required per geometry dataset, and other parallel tests have shown that this does not have a significant impact on overall parallel performance. Rather, it is the varying computational load across geometries which impacts upon parallel efficiency at higher core counts.

3.2. Stage 2 Parallelisation

In order to scale to higher numbers of cores, the second stage parallelisation scheme across energy clusters within each geometry calculation (described in Section 2) has to be invoked. Evidently this would ideally be used in conjunction with the geometry parallelisations described in Stage 1, leading to another order of parallel scalability. The dataset is once again the dissociative e-N_2^+ scattering calculation that provides four clusters of scattering energies per geometry when run over the required energy range for this problem. The testing was undertaken on an IBM iDataplex machine at the Hartree Centre, UK. The IBM iDataPlex, known as Blue Wonder, comprises 512 nodes each with two 8 core 2.6 GHz Intel SandyBridge processors making 8,192 cores in total. The nodes each have 32 GB of memory.

The Stage 2 MPI parallelisation results demonstrate reasonable scaling across the four clusters of energies for our dataset on a single node of the IBM iDataplex. The number of clusters remains the same across all geometries and they can be pre-determined with a short pre-processing step before embarking on a large-scale parallel run. However the number of energies per cluster does vary and this evidently has a major impact on load-balancing the computations efficiently between the MPI tasks.

Number of MPI Tasks	Elapsed Time for Calculation (1 Geometry with 4 Energy Clusters) (secs)	Speed-up
1	3690.8	1
2	2171.4	1.7
4	1117.0	3.3

Table 2. Speed-up obtained from Stage 2 parallelisation of the scattering energy clusters.

4. Summary

Two distinct levels of MPI-based parallelisation have been introduced into the serial TIMEDEL resonance finding and fitting code which have enabled large-scale electron-molecule calculations to exploit the computational power of hundreds, or even thousands of cores on modern HPC NUMA architectures. This code development facilitates a more detailed representation of electrons trapped in quasi-bound ‘resonances’ on the base components of DNA, which are implicated in cell damage. We have developed two alternative mechanisms for harnessing the calculations for different geometries across parallel nodes. On most Linux-based platforms either approach can be used, but on heavily stripped down Linux compute kernels the Fortran 2003/C interoperability may be required. The varying computational load across geometries impacts overall scaling at higher core counts with the parallel harness. A second stage MPI parallelisation has been applied to the energy loops at the computational core of the calculations

and this will enable runs on larger core counts to be undertaken efficiently. It has also been observed that the large amount of results data generated impacts overall scaling at high core counts and discussions with the scientific community have been initiated to reduce these overheads, for example by reducing the amount of repeated formatted output.

5. Future Work

Future versions are planned to introduce more sophisticated approaches to task farming, where either geometries could be bunched according to their pre-determined physical properties or shared counters could be exploited to load-balance the scheduling of the irregular parallel tasks. As the resonances are associated with the target state thresholds, we may in principle subdivide the set of geometries in advance in order that each group involves calculations with a similar number of resonances, as we know in advance the target state energy-geometry map (the included target states are known input to the collision/resonance calculation). Thus each group will perform a load-balanced calculation. This is more straightforward for the test case diatomic molecule but will be extended to more general molecules in collaboration with the scientific partners.

6. Bibliography

In addition to the references below, a recent overview of radiation damage to biomolecular systems may be found in the text '*Radiation damage to bio-molecular systems'*, editors G G Gómez-Tejedor and M C Fuss, Springer 2012, ISBN 978-94-007-2563-8 e-ISBN 978-94-007-2564-5. DOI 10.1007/978-94-007-2564-5

The general application of molecular R-matrix theory is described in:
'*Electron-molecule collision calculations using the R-matrix method'*, J Tennyson, Physics Reports, Volume 491, Issues 2–3, Pages 29-76 (June 2010).

Acknowledgements

This work was financially supported by the PRACE project funded in part by the EUs 7th Framework Programme (FP7/2007-2013) under grant agreements. The work has been achieved using the PRACE Research Infrastructure resources FZJ [9]. The authors also wish to thank Lucian Anton from STFC for his help in developing the parallel control mechanisms and the AMOPP group at UCL, particularly Duncan Little, for providing representative datasets.

References

1. UKRmol: JM. Carr, PG Galiatsatos, JD Gorfinkiel, AG Harvey, MA. Lysaght, D Madden, Z Mašín, M Plummer, J Tennyson, and HN Varambhia, Eur. Phys. J. D 66, 58 (2012)
2. A Dora, L Bryjko, T van Mourik and J Tennyson, J. Chem. Phys. 146 (2012) 024324
3. A Dora, L Bryjko, T van Mourik, and J Tennyson, J. Phys. B 45, 175203 (2012)
4. Z Masin and JD Gorfinkiel, J. Chem. Phys. 137, 204312 (2012).
5. Z Masin and JD Gorfinkiel, Eur. Phys. J. D 68, 112 (2014)
6. DA Little and J Tennyson, J. Phys. B 46, 145102 (2013)
7. DA Little and J Tennyson, J. Phys. B 47, 105204 (2014)
8. Atomic, Molecular, Optical and Positron Physics, University College London, http://www.ucl.ac.uk/phys/amopp.
9. http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html
10. http://www.stfc.ac.uk/hartree