Response of the Canadian Association of Theoretical Chemists (CATC) to Compute Canada's Call for Whitepapers for Sustainable Planning for Advanced Research Computing (SPARC)

The Canadian Association of Theoretical Chemists (CATC) discussed Compute Canada's Call for Whitepapers for Sustainable Planning for Advanced Research Computing (SPARC) at its Annual General Meeting in Montreal on July 10th, with 24 members in attendance (out of the total membership of 94); a request was also circulated to the membership for input to the SPARC call and responses were obtained from 15 members of the association. The CATC mandated Peter Kusalik, who served as Chair of the Chemistry/Biochemistry Resource Allocation Committee for Compute Canada, to prepare the following summary representative of the needs captured in those discussions and responses. Additionally, Peter Kusalik discussed technical aspects of many of the needs described below with Rob Simmons and Paul Wellings (University of Calgary). A draft of this white paper was circulated to the whole membership electronically and 12 members provided additional comments which were included in the final version of the document, which we believe genuinely reflects the input from the entire Canadian computational chemistry and biochemistry community.

The research being undertaken by the various groups within the CATC spans a diverse range of areas and explores a wide range of important systems and phenomena, including and not restricted to:

- biomolecular chemistry involving, for example, model studies of protein folding and macromolecular assembly, membranes and membrane transporters, and enzyme activity;
- medicinal chemistry - for example, studies of ligand binding leading to rational design of new drugs;
- materials chemistry with studies exploring, for example, new materials for solar cells and for separation and/or storage of gases such as CO_2 and CH_4;
- development and testing of new models, theories and computational approaches - for example to investigate interactions of light/photons with atoms and molecules, or to probe the processes underlying organization in membranes.

The physical processes of interest in these studies easily span 12 orders of magnitude in time (from attosec. to millisec.) and 6 orders of magnitude in length (from nm to mm).

These studies utilize a wide range of approaches, including:

- classical molecular dynamics (MD) simulations utilizing empirical molecular mechanics (MM) force fields;
- quantum mechanics (QM) calculations that explore electronic and molecular structure;

- *ab initio* molecular dynamics simulations that derive molecular interactions from first principles;
- QM/MM techniques that are a hybrid of the QM and MM approaches.

In order to achieve further advances or to provide new insights, most research groups identified needs to drive research to more detailed descriptions and better models, larger systems, and/or longer timescales. Given that these approaches are in essentially all cases computationally intensive, this translates into significant need for greater computational power, with implications such as increased memory, increased storage, and increased parallelism (need for fast interconnects).

Theoretical and computational chemistry and biochemistry has traditionally been a very strong discipline in Canada, and analysis of current data shows that chemistry and biochemistry users account for roughly 30% of the total cpu time usage within Compute Canada, which is more than any other discipline (the next group, Physics, accounts for 22%). Data also indicate that chemistry and biochemistry users tend to utilize somewhat higher levels of parallelism in their jobs than average Compute Canada users (see Table 1). This is indicative of the fact that much of the software typically used by chemists (e.g. AMBER, CHARMM, GROMACS, GAUSSIAN) is rather mature code that has been highly optimized and tuned. It should also be noted that many of these codes continue to undergo methodological and software improvements (e.g. to take advantage of GPUs).

Table 1: % of total cpu time used by jobs with various degrees of parallelism.

	Serial	Moderately parallel (2-32 way)	Highly parallel (64-512 way)	Massively parallel (1024 or more way)
All users	16%	26%	48%	9%
Chem/Biochem users	2%	32%	53%	12%

Taking into account the feedback provided to the questions:
- what kinds of problems are you trying to solve?
- what kind of infrastructure is best suited to solve these problems today?
- how much of that infrastructure would be needed to meet your future needs?

the following points attempt to capture generally the anticipated future hardware needs of chemists and biochemists:

- Researchers generally noted the vital importance of Compute Canada facilities to their research programs. Projections from individuals for future needs for computational resources over the next 5 years varied, ranging from relatively constant to a thousand-fold increase (to achieve the next major scientific advances). These data aggregate to an anticipated increase in demand of one to two orders of magnitude to reasonably support the needs of the discipline over the next 5 years.

- The availability of GPUs (or similar commodity-based technologies) as a means of providing cost-effective computing was frequently expressed. Considerably larger clusters of the kind (relative to those currently available) were envisioned.
- The need for high levels of parallelism and fast interconnects was often anticipated. At the same time, almost 40% of the jobs run by chemists are serial (i.e. there is a lot of them), and another 40% of jobs are 8-way parallel (i.e. typically run on a single node), because of the very nature of the calculations. It is anticipated that needs for these jobs will continue and hence it is important that these categories continue to be well supported.
- While a few users indicated needs for large memory, up to 50-100 GB/core for some of their work (access to one large-memory ultrafast parallel machine for particularly challenging multi-dimensional QM calculations would be desirable), a majority of jobs could be accommodated with 2-4 GB/core.
- Particularly for many QM calculations, the availability of fast local scratch storage on nodes is an important factor; this need might be met with nodes having solid state disks available. Accessibility to serial (modestly parallel) machines with an intermediate amount of RAM (8-12 GB/core) is also desired for many standard QM computations.
- Storage requirements were anticipated to scale roughly with computational performance, growing to the 1 PB range for larger users. A proposal for a hierarchical (or layered) storage system arose from technical discussions. In such a system, the physical location where a file or sub-directory might be stored (e.g. local spinning disk, or remotely archived) would depend on whether it had been recently accessed, and would otherwise be transparent to the user.
- "Cloud" solutions were not seen generally as reasonable or cost-effective for most chemistry work (because of the need for dedicated resources).