Pacific Northwest
NATIONAL
LABORATORY

*Proudly Operated
by* **Battelle** *Since 1965*

Center for Adaptive Supercomputing Software-Multithreaded Architectures (CASS-MT)

Hierarchical Bayesian Modeling for Text Analysis

Enabling the statistical analysis of large irregular data on the Cray XMT

"Hierarchical Bayesian modeling is a key technology that has many practical applications in national security and fundamental science. Faster, scalable statistical analysis enabled by the Cray XMT can help detect security threats, counter cyber attacks and help accelerate scientific discovery in areas like bioinformatics."

- Pacific Northwest National Laboratory Task Lead Chad Scherrer

At a glance

Hierarchical Bayesian modeling is used in many scientific settings to draw conclusions based on uncertain, often partially-observed data. Problem domains are wide-ranging, including text analysis, biological applications, and human behavior modeling.

The process of developing a hierarchical Bayesian model typically requires several iterations of a run/analyze/modify cycle; modifications to the model often require significant changes to the code, especially if performance is a consideration. Specialized code generation allows a high-level model representation to map to high-performance C code for execution on XMT, reducing development time and allowing accelerating model development.

What we do

At Pacific Northwest National Laboratory's (PNNL) CASS-MT, the primary objective of this task is to enable parallel hierarchical Bayesian modeling, with applications including text analysis and fundamental science.

How we do it

Hierarchical Bayesian modeling assigns a distributional node to every relevant value that is not directly observed. If a parameter for the distribution at one node

U.S. DEPARTMENT OF
ENERGY

is a function of another node, there is a dependence between the nodes preventing them from being sampled in parallel.

The problem is therefore to determine the best way to schedule sampling over the nodes. This can be done statically, as in a graph coloring, or dynamically, as in producer/consumer parallelism and dataflow methods. We are investigating both of these approaches.

The nature of code generation demands that we compare against hand-optimized code, which must be constructed for a particular model. Our initial model of interest is one we have developed for ASCR relating to liquid chromatography-mass spectrometry (LC-MS) proteomics.

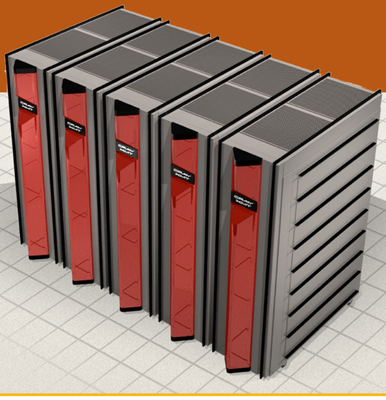
CASS-MT is dedicated to research on systems software, programming environments, and applications in a High-Performance Computing (HPC) multithreaded architecture environment.

We offer the only Open Science Cray XMT system, a one-of-a-kind supercomputer consisting of 128 multithreaded processors, 1 TB RAM, and a 7.7 TB Lustre parallel filesystem.

The Cray XMT supercomputer has the potential to substantially accelerate data analysis and predictive analytics beyond the limitations of traditional computing. Multithreaded processors allow multiple, simultaneous processing, helping researchers find solutions to the world's most complex challenges faster. The XMT can process irregular, data-intensive applications that have random memory access patterns. Unlike many applications where data delivery is dependent on memory speed, the Cray XMT's multi-threaded architecture tolerates memory access latencies by switching context between multiple threads that work continuously, overlapping the memory latency and preventing the processor from being held up while it waits for data to arrive.

The multithreaded technology powering our Cray XMT is ideally suited to perform pattern matching, scenario development, behavioral prediction, anomaly identification, and graph analysis.

Try it for yourself. We seek to create collaborations and provide expertise for porting and optimizing applications. The opportunity to use our Cray XMT system is available to internal and external research partners.



John Feo,
CASS-MT Director
(509) 375-3768
John.feo@pnl.gov
cass-mt.pnl.gov/

Chad Scherrer
Task Lead
Computational Mathematics
509-375-6308
chad.scherrer@pnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*