Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

"Technologies in massive, complex graphs are essential for big data applications in eScience and national security. We are using the strengths of multi-threaded environments to drive innovations providing data analysts with powerful new capabilities to find meaningful patterns in reasonable times as the size and complexity of data sets inevitably grow."

– Pacific Northwest National Laboratory Task Lead Cliff Joslyn

U.S. DEPARTMENT OF ENERGY

Semantic Database Systems

HIGH PERFORMANCE COMPUTING FOR SEMANTIC DATABASE PROFESSIONALS AND ANALYSTS

At a glance

Massive data sets possess not just great size, but also great complexity, with many heterogeneous elements linked by relationships of many different types. Graph theoretical representation can be superior to optimize the ability of users and data owners to find and identify meaningful patterns and query results in such data sets. They can be especially powerful when tasks involve pattern finding in networks which are sparse in attributes, with qualitative attributes (labels, categories), rich in random-access connections, and with patterns and paths of indefinite length. The new semantic graph database (SGD) paradigm uses ontological systems for typing schema; large labeled, directed graphs for data; graph pattern matching for query; and recursive query languages for graph analysis. One popular SGD format uses the prominent OWL/RDF/SPARQL software capabilities; we position these technologies against related models and languages such as Datalog. The Semantic Database Task for CASS develops high-performance software platforms for SGDs in massively multi-threaded environments; advances scalable methods for analyzing and representing SGDs in terms of their structural and semantic properties; and performs R&D in hybrid data environments to pair multi-threaded platforms with traditional relational databases (RDBs), key-value stores, and distributed cloud environments.

Empowering analysts to master their massive semantic graph and hybrid databases

What we do

Focusing on SGDs, novel analytical techniques will assist in identifying features of multi-relational data at both the structural and semantic levels, identifying new opportunities for knowledge discovery and efficient query. Massive shared memory, multi-threaded platforms like the Cray XMT make it an ideal candidate for hosting SGDs. Our overall goal is to provide a software and algorithms base that will make it easy to use multi-threaded architectures for semantic database applications. Our three foci include:

- **Research** in combinatorial and information theoretical approaches to identifying prominent semantic patterns in data; graph compression and automorphism; and

methods in hybrid graph/relational hybrid data, queries, formalisms, platforms, data models, and languages.

- **Engineering** of prototype semantic graph database capability for large memory, multi-threaded platforms.
- **Analysis** of massive SGD data and seeking out characteristic benchmark data and test suites.

How we do it

Multi-threaded architectures are known to be good at solving graph problems with sparseness and irregularity. Semantic graphs are a generalization of this where edges are directional and nodes and edges have types and other attributes. In turn, these types have a logical structure as reflected in an ontological typing system, supported semantically constrained query and inference. We are developing efficient graph data structures capable of supporting these types of general graphs and algorithms that search these graphs for user-specified patterns. Performing optimization techniques on the search queries is critical to the viability of the system.

Within the framework of an SGD, functions beyond the building of the graph and searching for patterns (querying) are essential to realizing the full potential of the knowledge representation and information retrieval system. Some of these functions include inferencing such as RDFS closure, Owl Horst Semantics, and rule-based languages. In addition to these foundational features, extensions will be investigated such as expressing (e.g., in SPARQL or Datalog) and executing path and other subgraph queries.

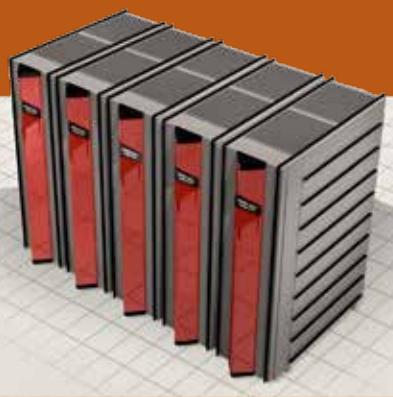
Applications

Our prototype capabilities are being applied to a range of data sets especially aiming at benchmark standards, but also including massive compendia of computational and systems biology information.

CASS is studying challenging irregular problems in search, knowledge discovery, cybersecurity, complex network, and natural language understanding. It is driving development of next generation software platforms, programming models, runtime systems, and high-performance computing systems that support global shared memory, hardware multi-threading, and fine-grain synchronization.

The Center manages a variety of computer systems with the potential to substantially accelerate data analysis and predictive analytics beyond the limitations of traditional computing. Our systems allow multiple, simultaneous processing, helping researchers find solutions to the world's most complex challenges faster. For example, our 128 processor, 2 TB Cray XMT can process at scale irregular, data-intensive applications that have random memory access patterns. 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.

We provide user accounts on all our systems, and can help you port and optimize your application. We seek collaborations in all our research areas of interest, and look forward to working with internal and external research partners.



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

Cliff Joslyn
Task Lead
High-Performance Computing
206-528-3042
cliff.joslyn@pnnl.gov

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*