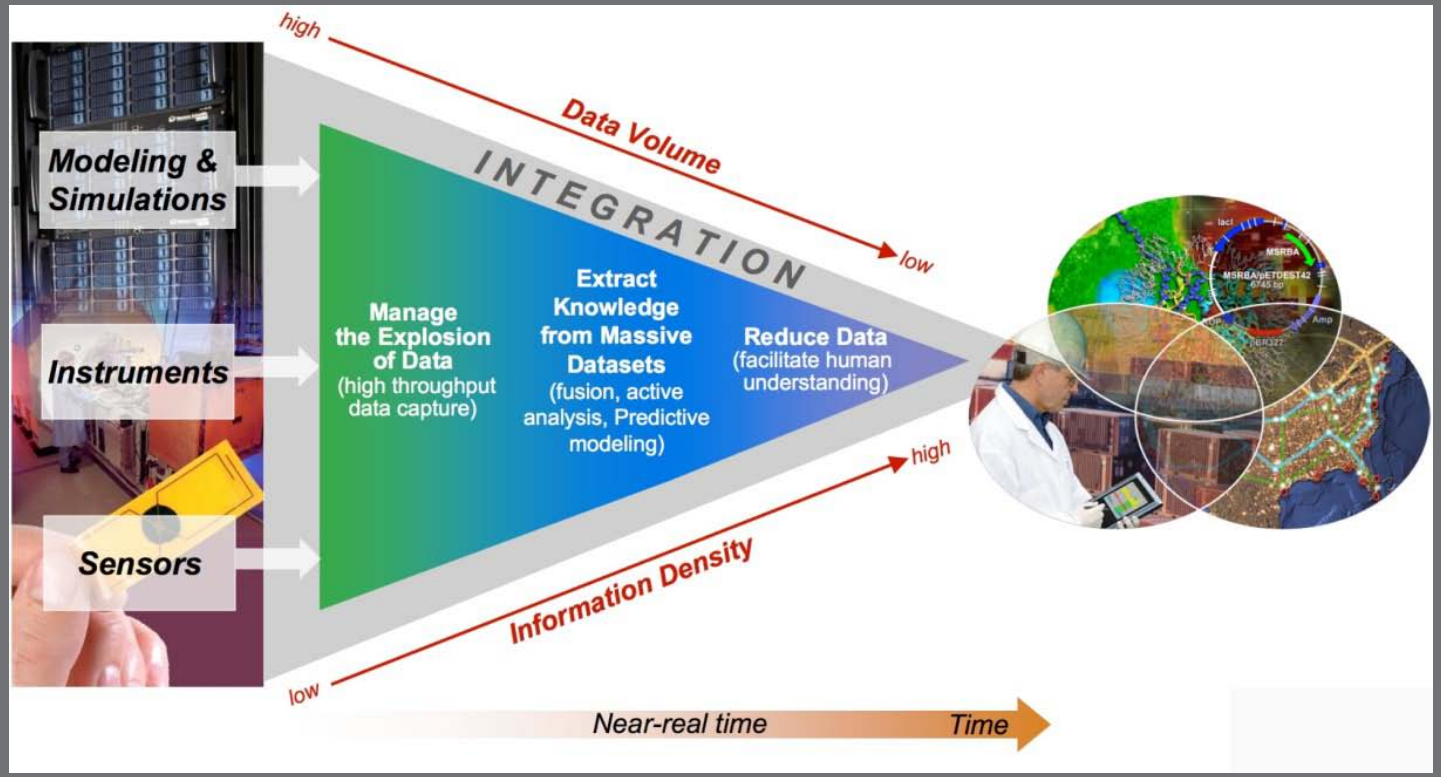
Data Intensive Computing

- PNNL is a leader in research for Data Intensive Computing methods
  - Signature area developing algorithms, applications and visualization capabilities for science and defense domains where data volume cripples conventional applications.
  - Allows analysts and scientists to operate on large-scale data for real-time awareness and hypothesis discovery.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Operating on large-scale data for real-time awareness and hypothesis discovery



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Architectures for Data Intensive Computing

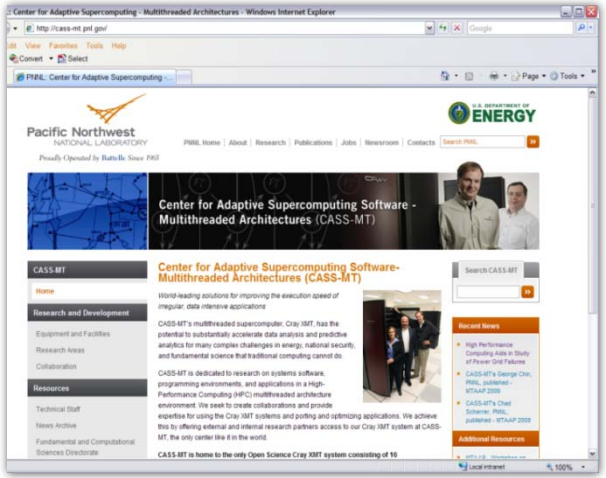
- Knowledge Discovery
- Information Analysis
- Data Intensive Computer Systems

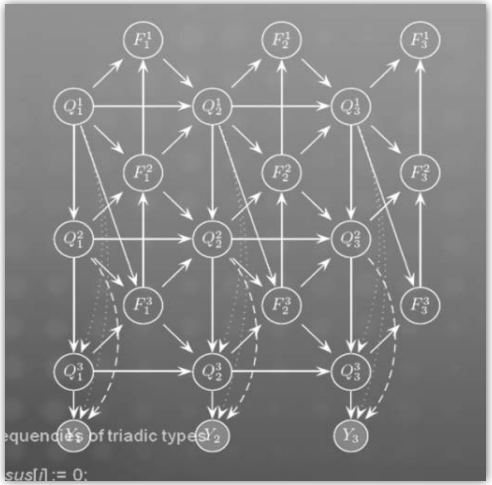
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Center for Adaptive Supercomputer Software

http://cass-mt.pnl.gov/

- Working with government, academic, and industry partners to develop applications, libraries, and next generation systems to solve the most challenging irregular problems in knowledge discovery





Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Cougar XMT

- The largest, open CRAY XMT system supporting government, academic, and commercial research and application development
  - 128 compute processors
  - 1 TB shared memory
  - 7.7 TB disk space

ASK US ABOUT GAINING ACCESS TO THE XMT

CASS Partners

Georgia Tech

Ohio State

Washington State University

Cray
THE SUPERCOMPUTER COMPANY

Least Squares Software, Inc.

Berkeley Lab

Sandia National Laboratories

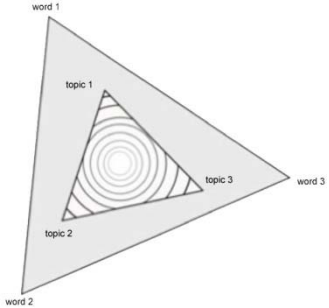
Pacific Northwest National Laboratory

Proudly Operated by Battelle Since 1965

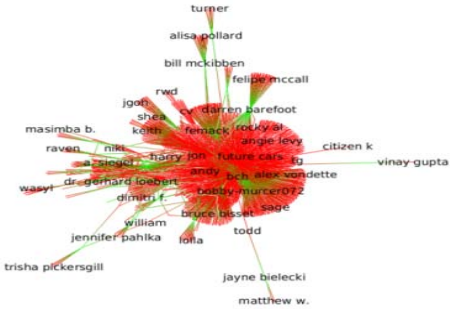
CASS Application Development Projects

complex multi-scale video analysis

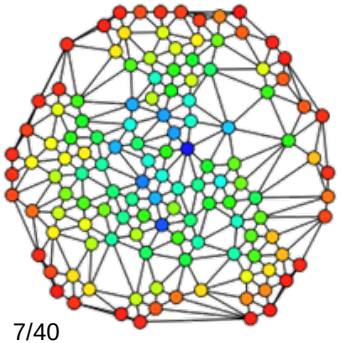
understanding text documents



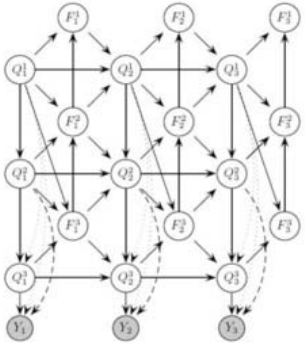
navigating blog communities



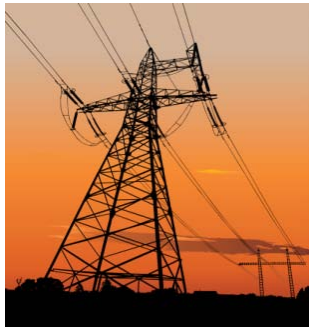
social networks



Bayesian networks



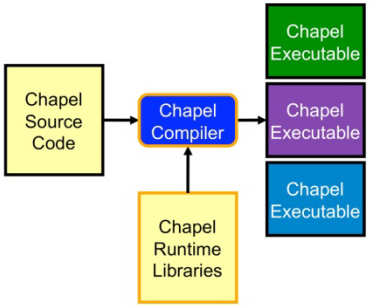
contingency analysis of electrical power grids

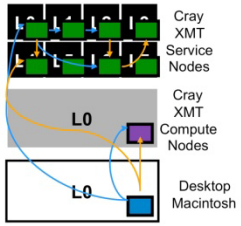


Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

CASS Software Projects

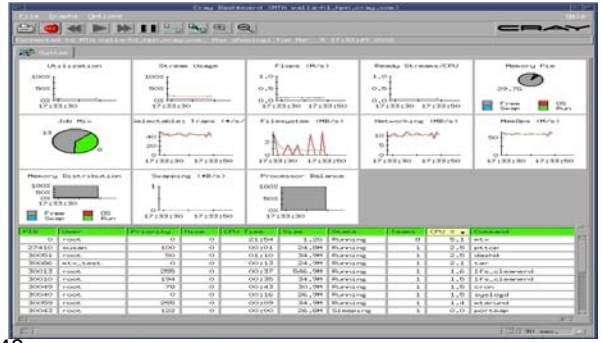
communication layer for hybrid systems



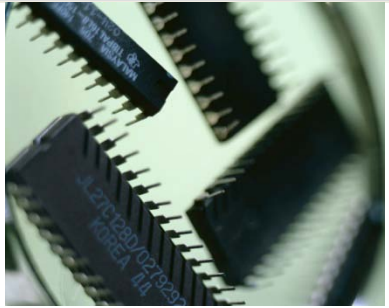


MT compilers and runtimes

MT performance tools



next generation MT processors



Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Computing Resources

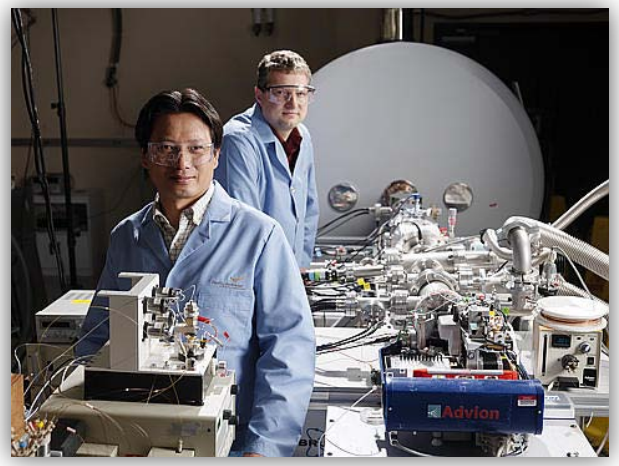
- CHINOOK
  2310 node HP cluster

Feature	Detail
Interconnect	DDR InfiniBand (Voltaire, Mellanox)
Node	Dual Quad-core AMD Opteron 16 GB memory
Local Scratch	400 MB/s, 924GB/s aggregate 440 GB per node. 1 PB aggregate
Global Scratch	30 GB/s 250 TB total
User /home	1 GB/s 20 TB total

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Environmental Molecular Sciences Laboratory (EMSL) Capabilities

- EMSL houses an unparalleled collection of state-of-the-art capabilities that are used to address scientific challenges
- Chinook **cluster** solves computing challenges in these areas:
  - Deposition and microfabrication
  - Kinetics and reactions
  - Mass spectrometry
  - Microscopy
  - NMR and EPR
  - Spectroscopy and diffraction
  - Subsurface flow and transport



Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Themes, Modeling, and Experimentation

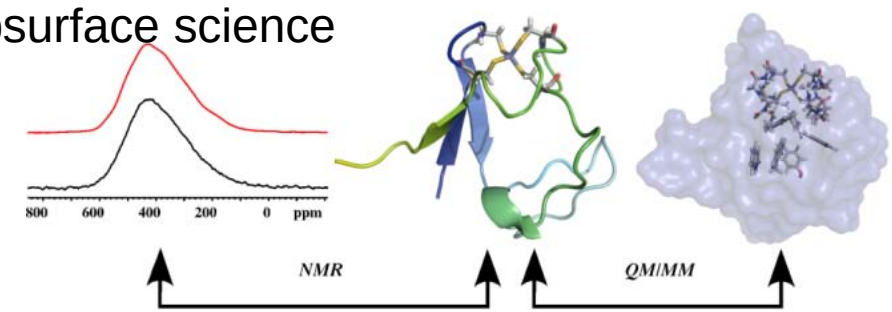
- Capabilities are available to users, generally at **no cost** to the user, for research in four science theme areas

Biological interactions and dynamics

Geochemistry/biogeochemistry and subsurface science

Science of interfacial phenomena

Atmospheric aerosol chemistry
(Emerging Theme)



Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Molecular Science Computing Software

EMSL provides software to enable scientific research. The **Molecular Science Software Suite** is comprised of three packages:

MS³
MOLECULAR SCIENCE SOFTWARE SUITE

NWCHEM
HIGH-PERFORMANCE COMPUTATIONAL CHEMISTRY SOFTWARE

Computational chemistry software suite designed to perform Quantum and Classical Mechanics calculations on parallel supercomputers

GA TOOLS
PARALLEL COMPUTING LIBRARIES AND SOFTWARE TOOLS

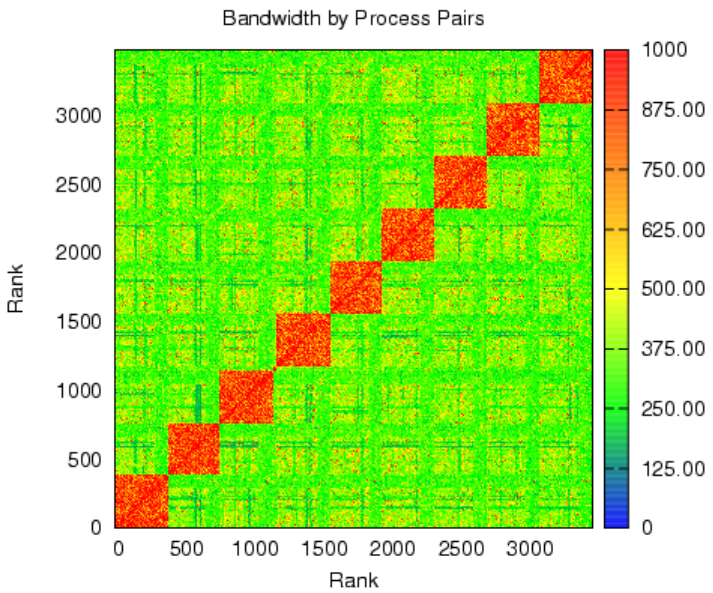
Efficient and portable shared-memory programming interface for distributed-memory computers

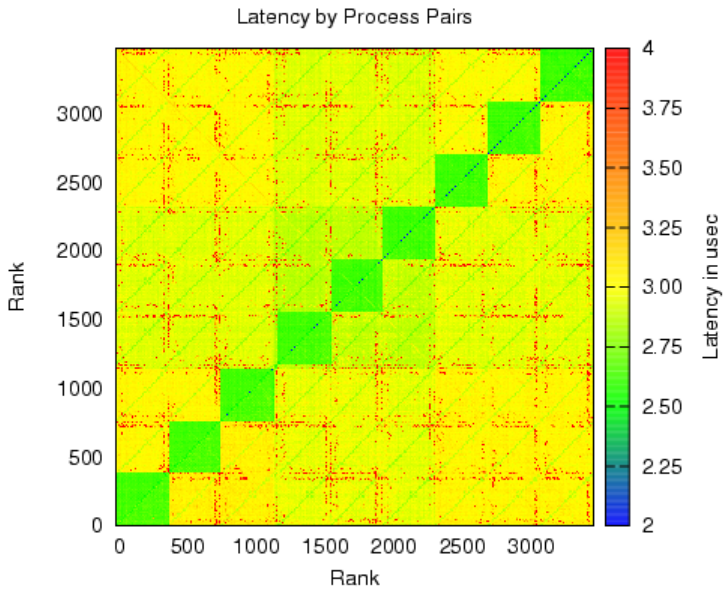
ECCE
EXTENSIBLE COMPUTATIONAL CHEMISTRY ENVIRONMENT

Graphical user interface, scientific visualization tools, and underlying data management framework

Chinook Interconnection Latency

- MPI Latency (in µsec) and Bandwidth (in MB/s) with two tasks/node on 1738 nodes.





Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Algorithms for Data Intensive Computing

- Scientific Discovery
  - Data analysis, supervised/unsupervised learning, visualizations for hypothesis discovery
- Real-time processing
  - Rapid streaming processing for instrument control
- Data management

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Data Intensive Algorithms

- Data-intensive computing is collecting, managing, analyzing, and understanding data at volumes and rates that push the frontier of current technologies.

Computing strategies and implementations to help deal with the data tsunami

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Leadership in Data Intensive Algorithms

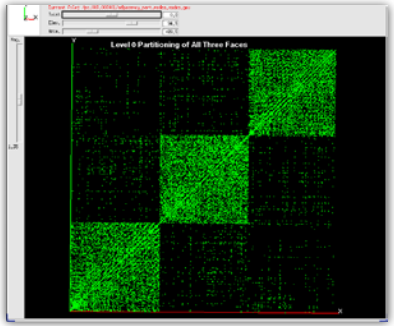
- Community involvement
  - Supercomputing 2007 SciNET Chair
  - DARPA's Next-Generation Hardware Program Review Panel
  - IPDPS: Workshop on Performance Optimization for High-Level Languages and Libraries
  - SC2008 HPC Analytics Challenge
- Publications
  - S IEEE Computer Special Issue on Data-Intensive Computing – April 2008
  - Over 80 journal papers, conference papers, and 'magazine' articles
  - Invited paper in SciDAC Review on Cray XMT Applications
  - Best paper runner up award at IEEE VizSec Symp., Oct 2007 for CLIQUE
  - Monthly article for Scientific Computing

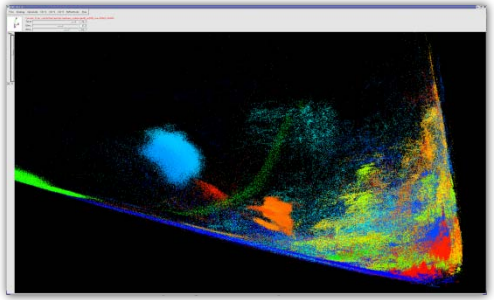
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Data Intensive Algorithms – Demonstrating impact on the community

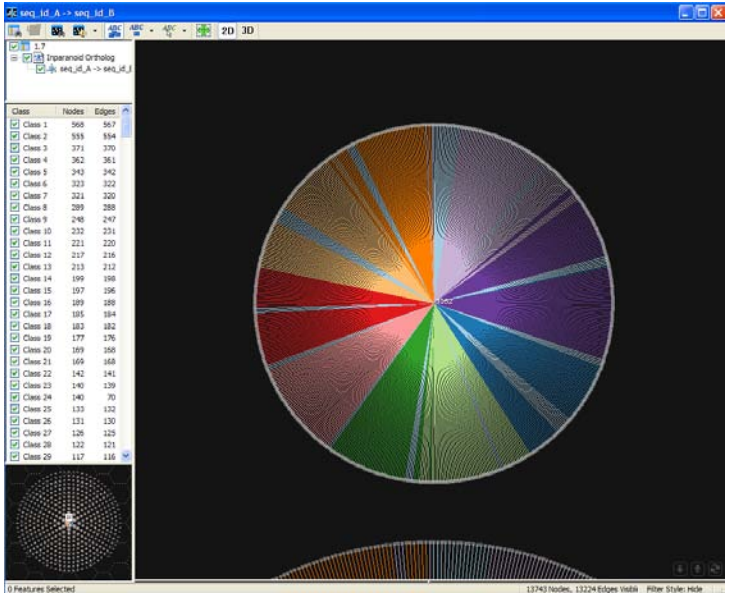
- Demonstration projects in a variety of science and defense application domains
  - Groundwater
  - Biology
  - Biodefense
  - Image/video processing
  - Cyber security
  - Instrument control





Data Intensive Computing in Bioinformatics (SHOT, SVM Hustle)

- SHOT is a support vector machine (SVM, machine learning) implementation of protein homology detection
  - Improved sensitivity over industry standard BLAST, PSI-BLAST
  - Requires training on hundreds of thousands to millions of protein pair vectors
- SVM-Hustle is a SVM-based implementation of protein family classification

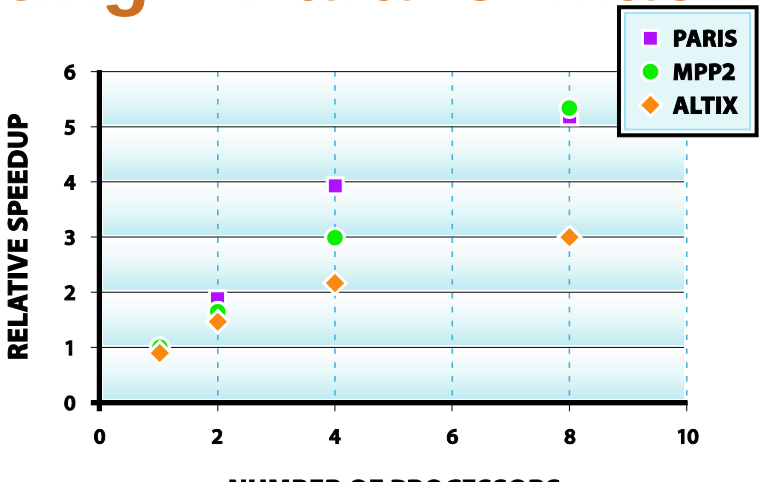


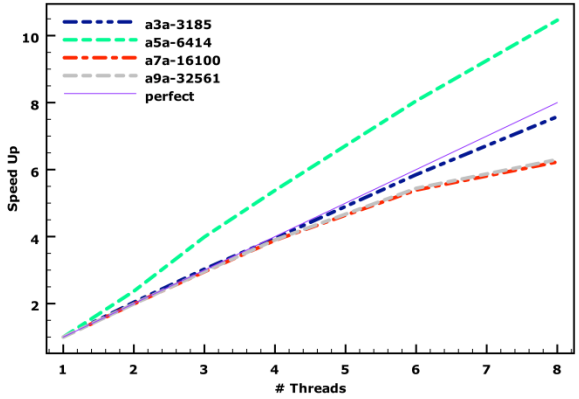
Proteins from 10 Shewanella species (indicated by different colors) are found to be related to a single evolutionary parent (central protein) using SHOT.

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Data Intensive Computing – Parallel Machine Learning

- Evaluation of existing parallel methods for Support Vector Machine (SVM) training
- Design and implementation of threaded and parallel SVM training, classification methods



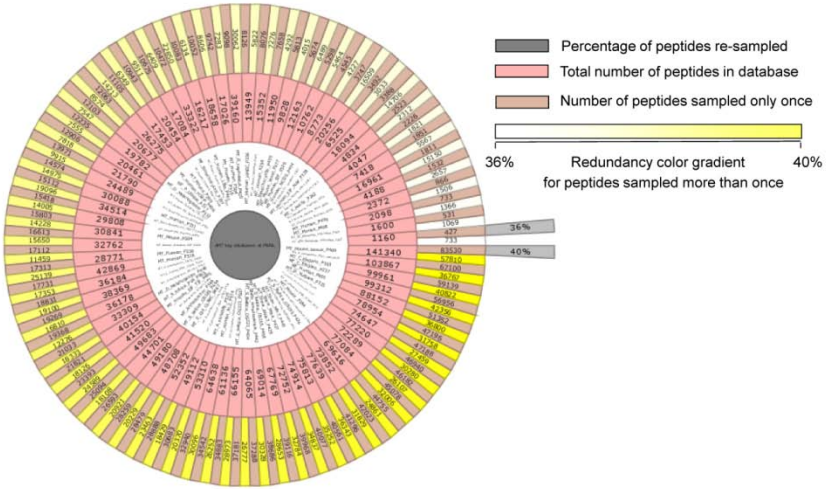


Pacific Northwest NATIONAL LABORATORY

Operated by Battelle Since 1965

Data Intensive Computing- Data Analysis for Proteomics

- Data intensive application for dealing with redundancy in Accurate Mass and Time (AMT) tag database
  - Removing redundant peptides from the analytical pipeline increases throughput by focusing compute cycles on peptides not previously analyzed.

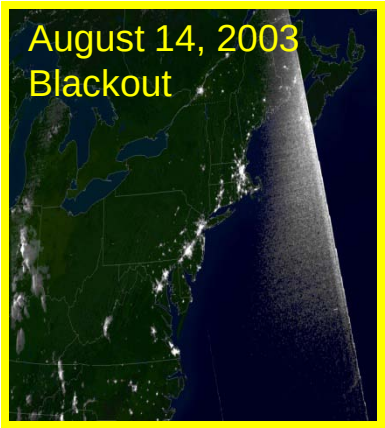


A diagram illustrating the number of uniquely identified peptides, the number of repeatedly observed peptides, and the total number of peptides in existing AMT tag databases generated using tandem LC-MS experiments. 36% to 40% of peptides in each of the databases is redundant using current analysis methods.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

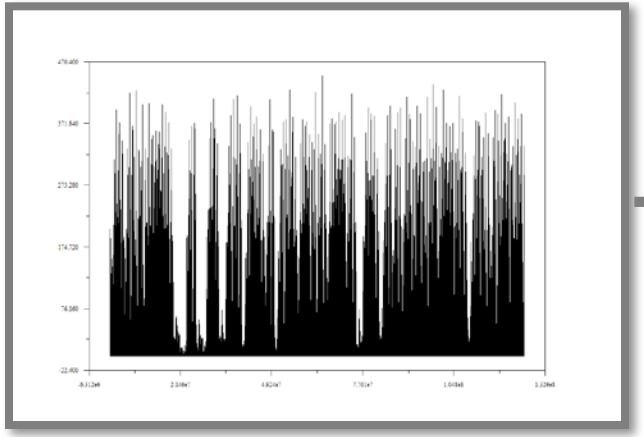
Data Intensive Computing: Power state estimation

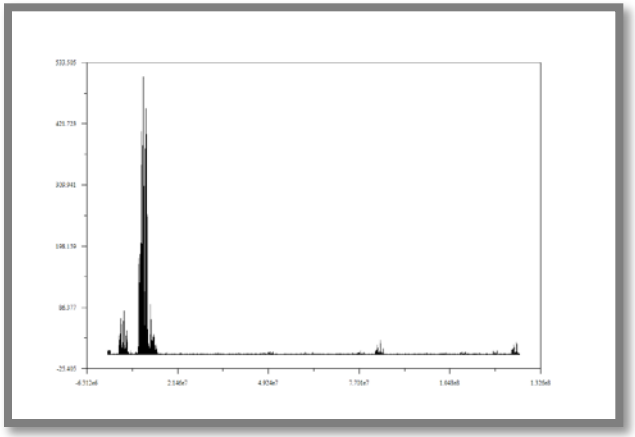
- Power system state estimation (PSE) is a critical element of the software used by power grid control centers
  - Under real-time constraints (< 10 seconds)
  - Commercial PSE solvers are not commonly parallel
- Power system state estimation
  - **Input**: power grid topology, telemetry on line flows, bus injections or bus voltages
  - **Output**: systems state estimation
  - minimize WLS: via Newton-Raphson
  - Western Electricity Power Grid = 14K bus system with up to 1M leafs



FPGA-Based Signal Processing for Mass Spectrometry

- Advanced mass spectrometers (such as ion mobility) are costly to build and operate
- Current usage of mass spectrometers limits capabilities
- Algorithm overlaps signal sampling from several instances of the same biological sample
- Designed and implemented an FPGA version of deconvolution algorithm for high-throughput ion mobility data
- Computation is 6x faster than software executing on a 2.4GHz Opteron

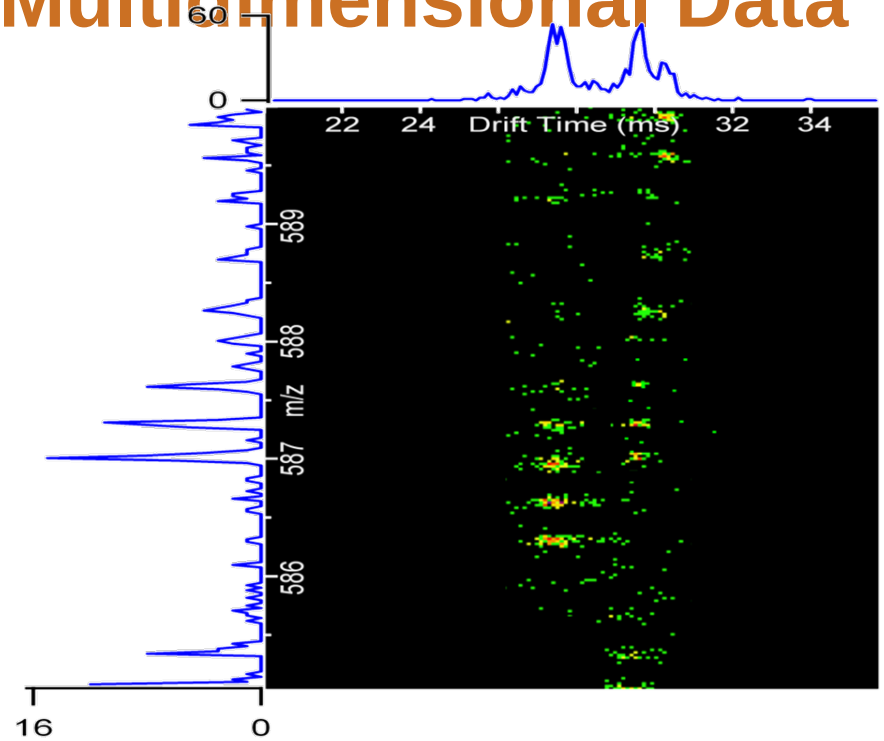




Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Compression and Organization for Multidimensional Data



Multidimensional view allows for separation of biomarker signatures from mass-spectrometry data

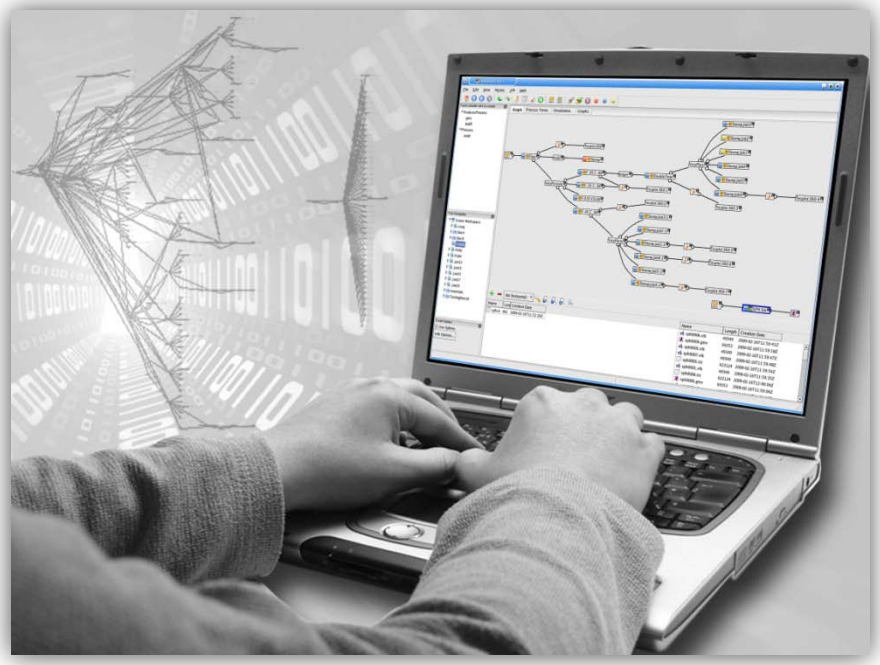
- **Resolving peptides in multiple dimensions is prohibitive without compression**
  - 9 TB of data!
  - Data compression reduces extraction time from >24 hours to 20 minutes

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Data Virtualization Architecture

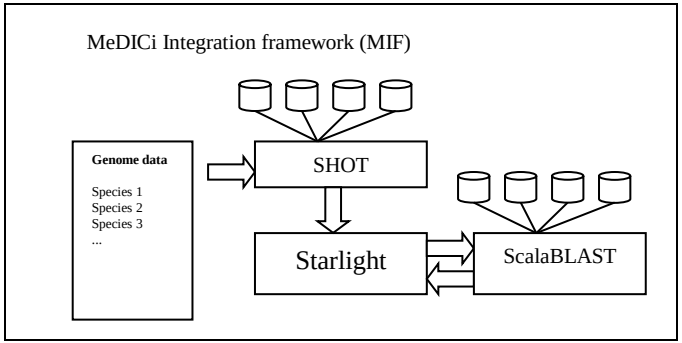
- Data virtualization allows users to track complex provenance dependencies with a common framework
  - Tracks dependencies between analytical processes, data collection and value-added products

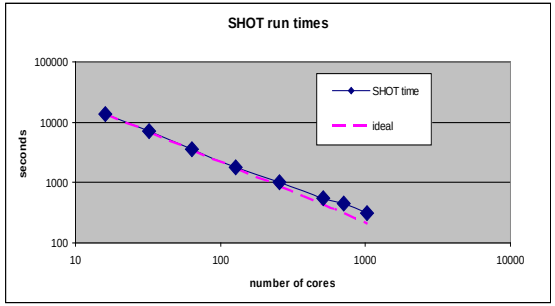


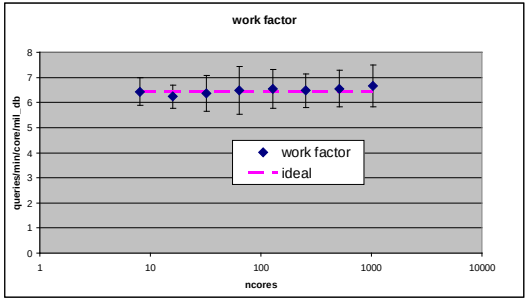
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

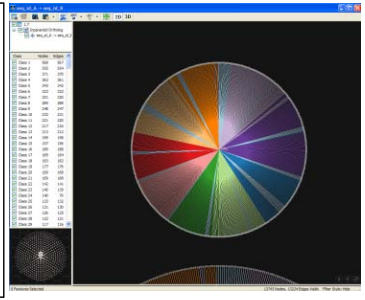
Data Intensive Computing: Bringing together applications and architectures for real-time hypothesis discovery

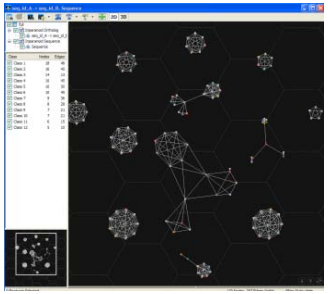
- MeDICi framework used to integrate HPC algorithms SHOT and ScalaBLAST with browsable Starlight visualization
  - Winner of the SC08 HPC Visual Analytics Challenge











Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

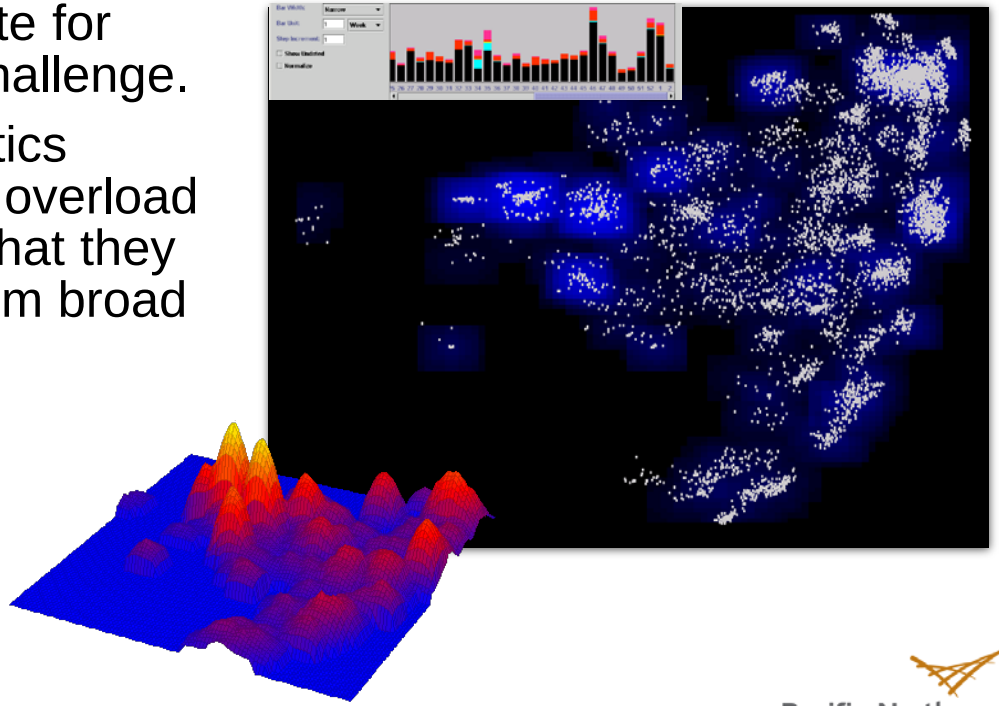
Visual Analytics

- Interactive visual analysis help humans make sense of vast amounts of heterogeneous information.
- A leader in visual analytics, PNNL is developing new techniques for exploration and discovery in real-time, high-volume data.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Visual Analytics: From data to decisions

- Creating interfaces appropriate for human analysis is a critical challenge.
- PNNL's scalable visual analytics approaches helps avoid data overload while allowing users to find what they need in their information – from broad overviews to specific details.

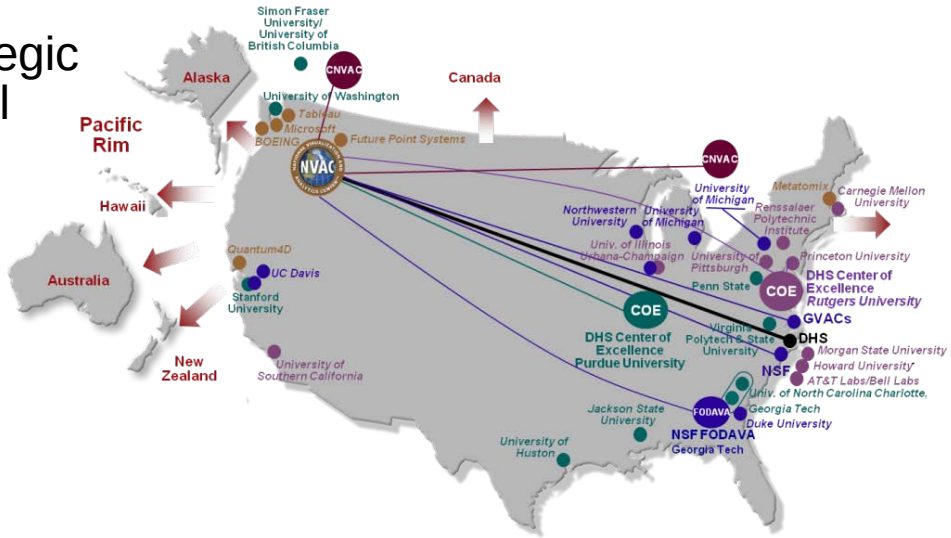


Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

National and International Leadership in Visual Analytics

- The National Visualization and Analytics Center (NVAC) is a national and international resource providing strategic leadership and coordination for visual analytics technology and tools.
- NVAC deploys data-intensive visual analytics capabilities with operational customers



Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Collaborations in Visual Analytics

- PNNL operates the Visualization and Analytics Centers Consortium, which brings together researchers and practitioners from across government, industry, and academia to develop and deliver novel visual analytics technology.
- The Consortium coordinates activities among collaborators to define and execute on a community-wide research agenda.
- http://nvac.pnl.gov/

Visual Analytics: The next generation of scalable applications

- Visual analytics is fundamentally interdisciplinary; basic advances inform technology across domains, and solutions developed for one domain are often transitioned to others.

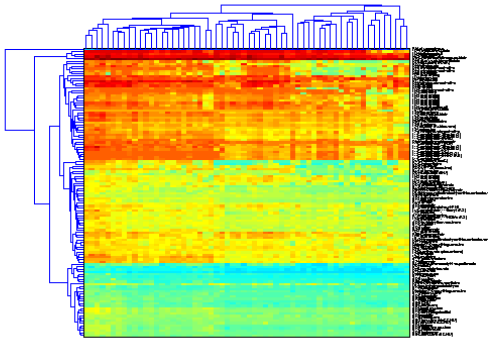
- Performing analysis in real time requires the ability to visualize and detect features in data as they are collected. These techniques lead to a new generation of scalable visual analysis applications.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Dynamic Aggregation of Massive Data Sets

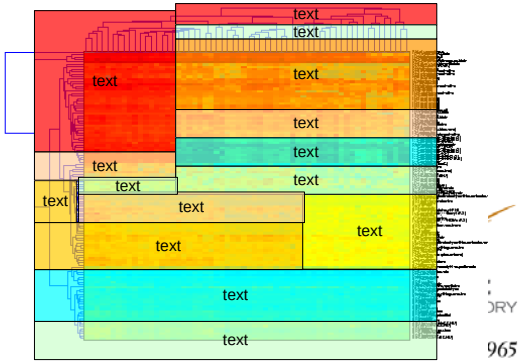
- High-throughput instrumentation for genomics and cyber security is producing more data than can be visualized readily in its raw form. Heat maps are a common way to visualize this data, but they typically do not scale well.
- Decomposing massive heat maps into multiple levels of abstraction helps represent the underlying macrostructures while supporting near real-time navigation of the space.

Massive Data



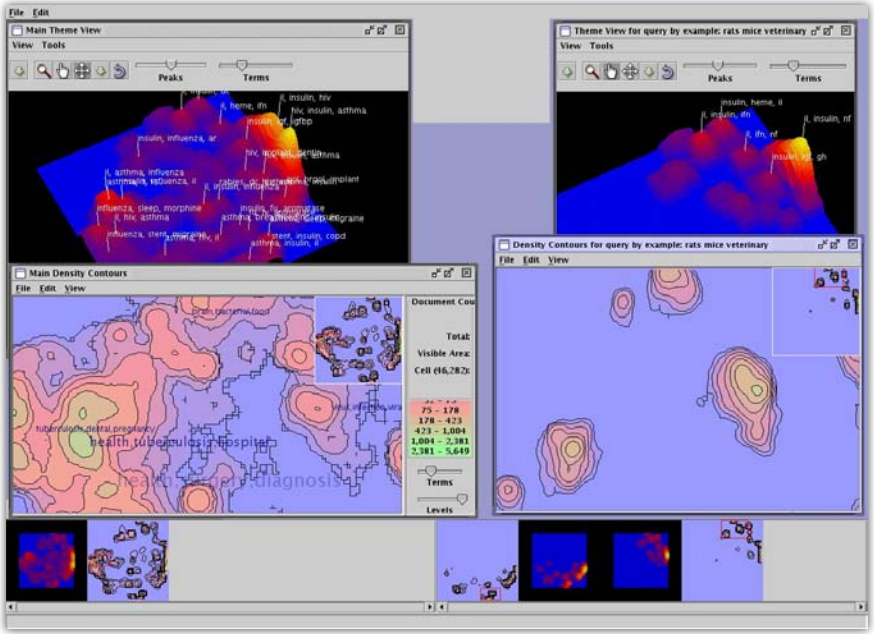
Information Segmentation

Discovery & Exploration



High-Performance Text Analysis

- Analysts must often understand the potential relevance of hundreds of thousands or millions of documents to their tasks.
- Creating text analysis algorithms that best benefit from parallelization can enable visual analysis of more documents than is currently possible.
- A parallel text processing architecture resulted in a **ten-fold improvement in document ingest speed and the ability to visualize relationships in 6 million document records.**

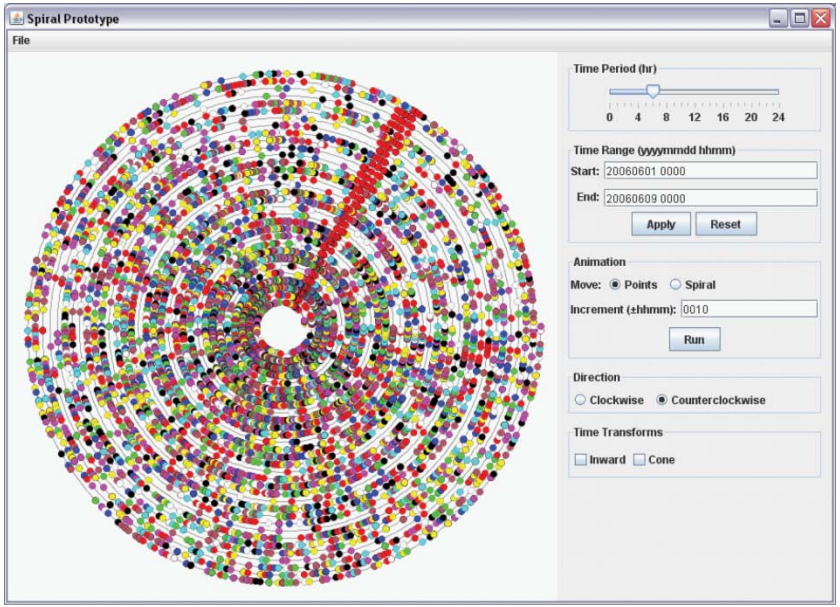


Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Visual Content Analysis of Dynamic Data Streams

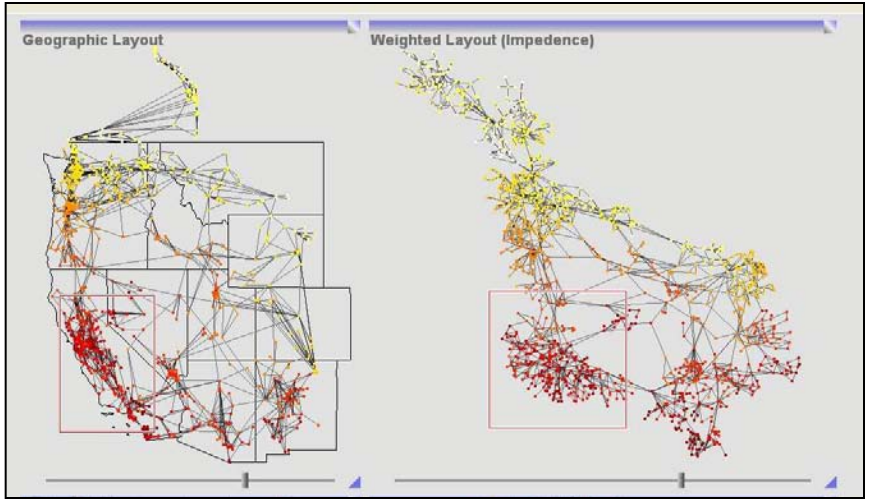
- Visualizations for real-time data need to be able to represent dynamic behavior.
- The Dynamic Visualization Tool Suite (DVTS) includes
  - Easy-to-understand visualizations organize data points based on scientific parameters and attributes
  - Multiple perspectives of data that support more comprehensive and integrative analysis, revealing patterns and structure in data.

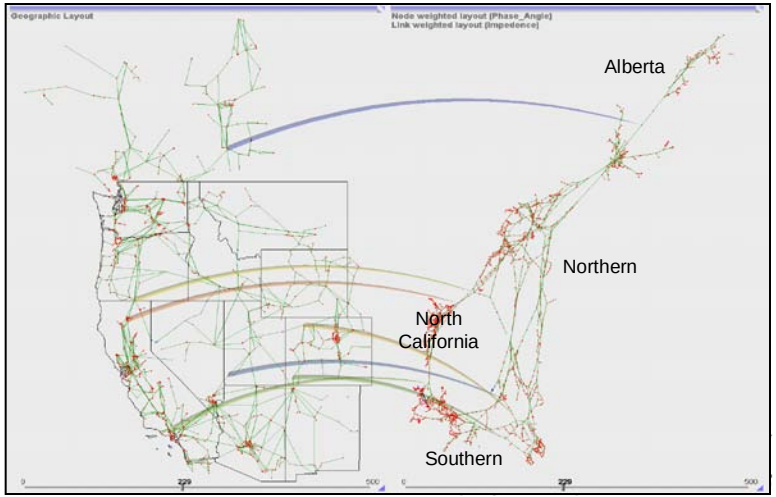


Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Visualizing Large Graphs

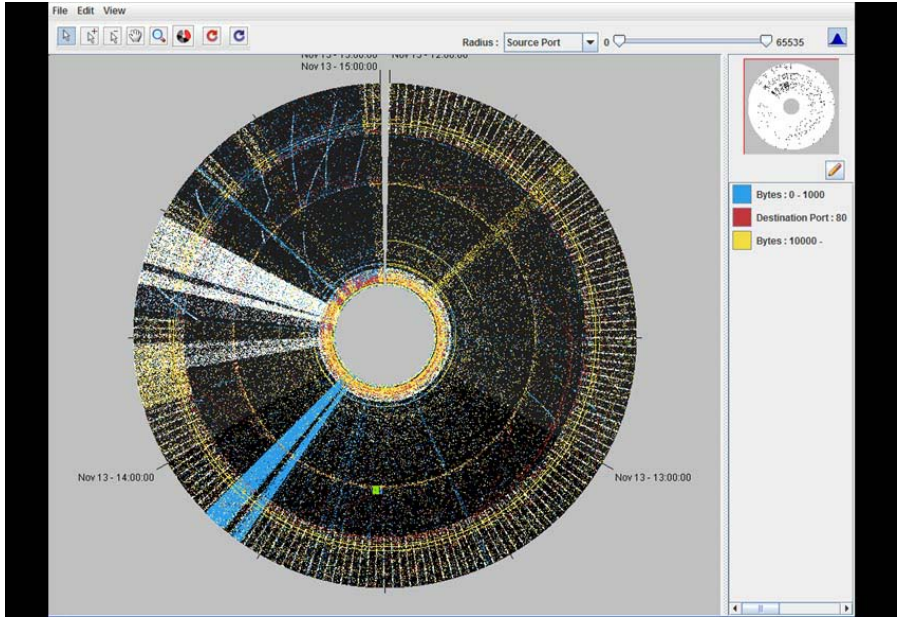
- Exploring and understanding node-link data sets with millions of nodes and tens of millions of edges requires new scalable graph visualization techniques and new methods of interaction.
- PNNL developed a groundbreaking ability to interactively navigate large graphs and automatically extract features to depict multiple perspectives on a network.
- A visual environment for critical infrastructure protection and risk assessment, built on this capability, is used to detect anomalous events in the power grid.





Visualizing Streaming Computer Network Data

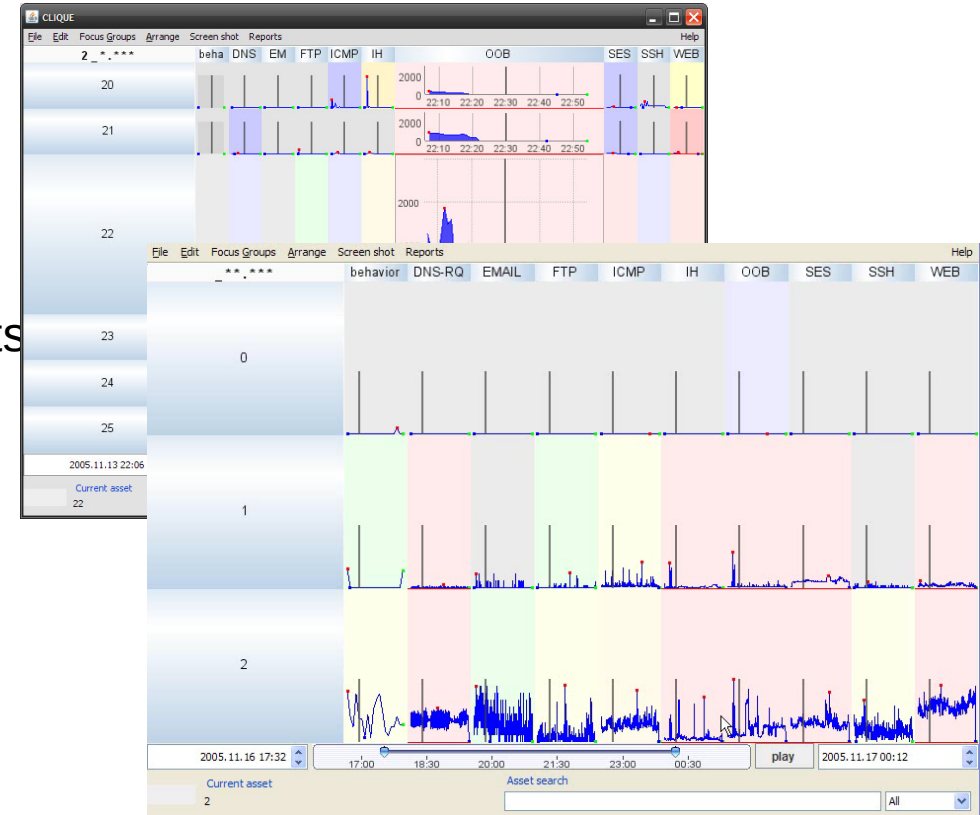
- Cyber security analysts need situational awareness from massive amounts of transactional data
- Typical raw log visualization tools do not scale to operational data volumes (which might include billions of transactions per day)
- PNNL, in partnership with Stanford University and US-CERT, is developing and deploying streaming network flow visualization capabilities that help analysts discover cyber threats as they occur.



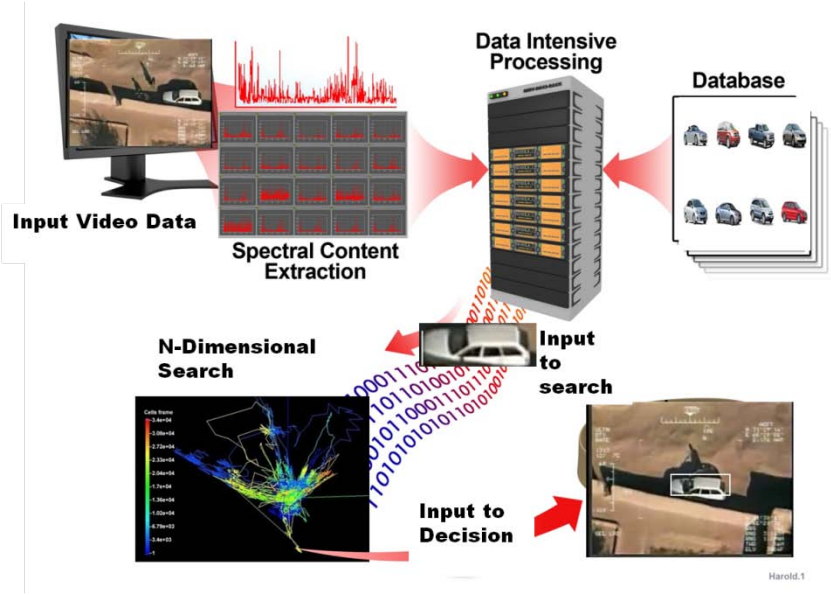
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

CLIQUE: Real-time cyber data aggregation and event detection

- **PNNL's CLIQUE** provides a scalable interaction technique for behavioral modeling and event detection in real-time cyber data.
- CLIQUE uses PNNL's MeDICi to distribute data to analysis components and visual clients.



Hybrid Computing for Video Analytics

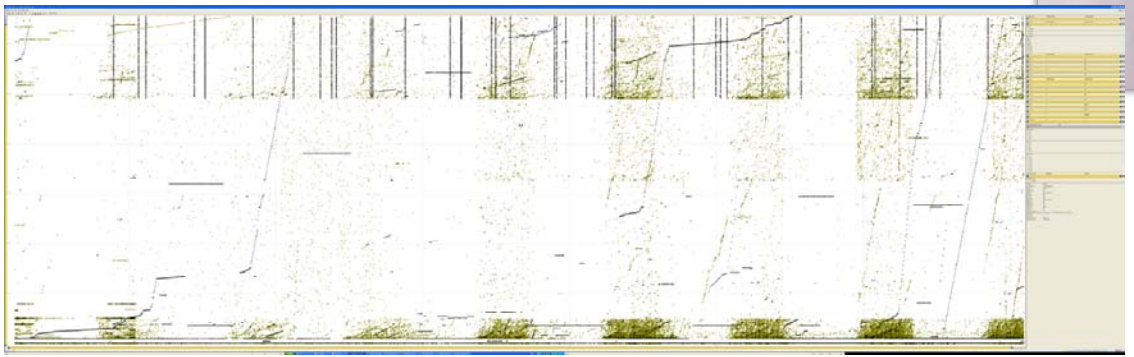


- Summarizing, indexing, and searching large video archives requires new algorithms and a hybrid parallel computing infrastructure.
- PNNL's solution incorporates quantitative representations of the information content of video or image data.
- GPUs, cloud computing, and the Cray XMT large shared-memory multithreaded system are part of an integrated video analysis workflow.

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Large, High-Resolution Information Spaces

- Large, high-resolution displays can give users the “space to think” about complex problems
- **PNNL**, in collaboration with Virginia Tech, is working with high-megapixel displays to explore new visualization and interaction techniques suited for data-intensive analysis problems in cyber security and other fields.



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Data Intensive Computing

- PNNL continues to push the envelope of Data Intensive Computing capabilities with research in
  - Emerging hardware architectures
  - Algorithms for driving analysis at extreme scales
  - Visualization and data management solutions
- MeDICi middleware for constructing data intensive workflow from components is freely available
  - http://medici.pnl.gov/

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965