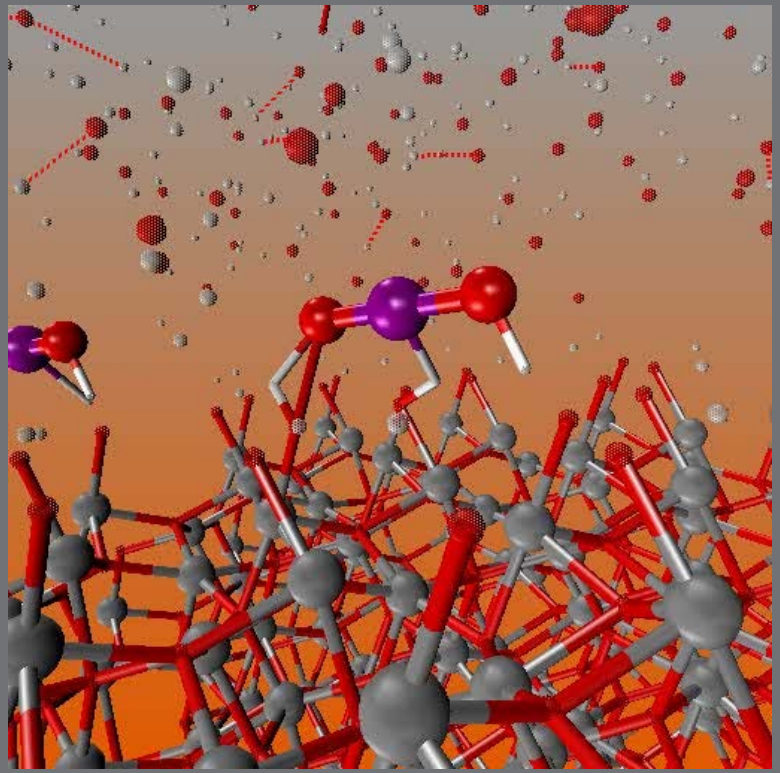
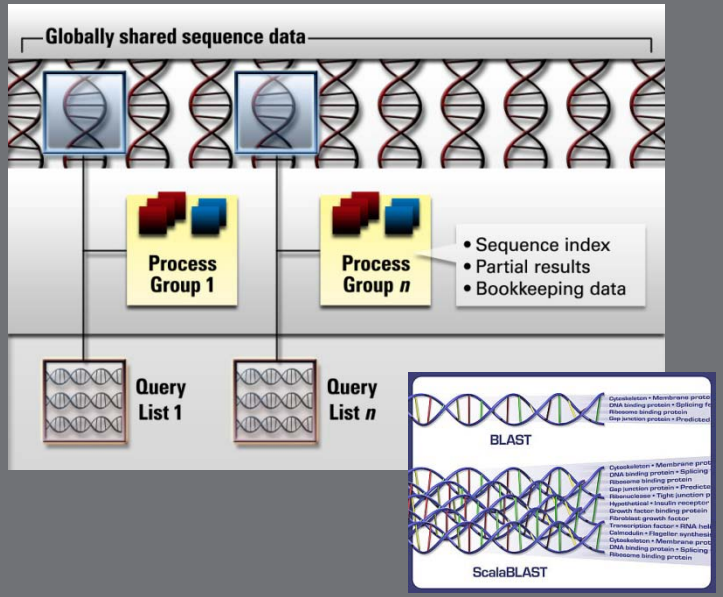
Scalable Applications

- PNNL is home to high-impact scalable applications for scientific simulations and discovery in
  - Computational Chemistry
  - Atmospheric sciences
  - Computational Biology

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Scalable applications at PNNL: Driving science toward the petascale and beyond





Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

NWCHEM
HIGH-PERFORMANCE COMPUTATIONAL CHEMISTRY SOFTWARE

NWChem

- DOE's Premier Computational Chemistry Software Package
- Provides one-of-a-kind computational chemistry modeling solutions for molecular science that are scalable with respect to scientific challenge and hardware

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Pushing the scientific envelope on large computing platforms

- Designed and developed to be a highly efficient and portable **massively parallel** computational chemistry package
- Provides computational chemistry solutions that are scalable with respect to chemical system size as well as MPP hardware size
- Designed for parallel architectures
- World-wide distribution (downloaded by 2350+ groups) 70% is academia, rest government labs and industry

MS³ MOLECULAR SCIENCE SOFTWARE SUITE

NWCHEM HIGH-PERFORMANCE COMPUTATIONAL CHEMISTRY SOFTWARE

GA TOOLS PARALLEL COMPUTING LIBRARIES AND SOFTWARE TOOLS

ECCE EXTENSIBLE COMPUTATIONAL CHEMISTRY ENVIRONMENT

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Extreme scale development

- NWChem: Scalability will be a major effort
- Scalability demonstrated to tens of thousands of processors
- Crucial for extreme scale development will be a team of
  - Computational chemists
  - Mathematicians
  - Computer scientists
  - Collaborative effort of diverse disciplines
- Efforts underway in NWChem
  - PNNL Exascale Initiative LDRD led by Karol Kowalski
  - PNNL Extreme Scale Initiative LDRD Abhivav, de Jong

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

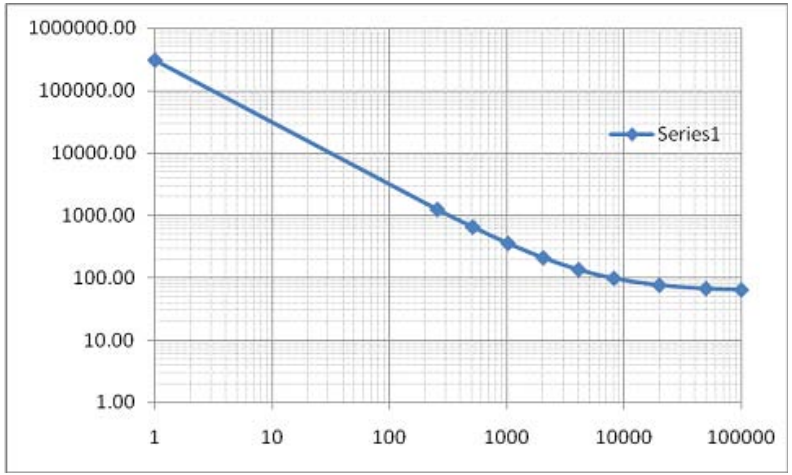
NWChem: Gordon Bell Finalist

- Show MPP2 63% peak on 1800 procs on small water system
- Show Jaguar 55% of peak on 96K procs on large water system

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Scaling high accuracy

- Improving the parallel efficiency for coupled cluster
  - Trading communication for calculations
  - Better task scheduling based on the a-priori estimates of the numerical overhead associated with each task
  - Using processor groups
  - Improving data-localization in the CC calculations:
    - Localization of the low-rank tensors in local memory,
    - Localization of the higher-rank tensors on local disks,
    - Developing CC models with selected subset of the most important cluster amplitudes

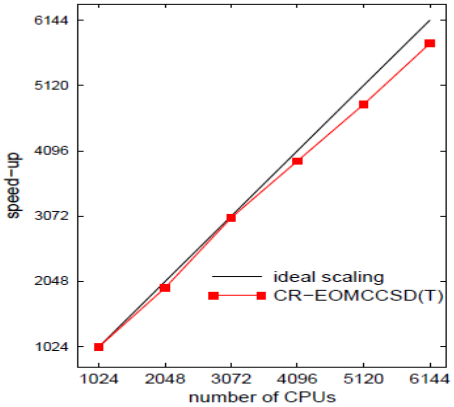


Extrapolated timings for the Zn-porphyrine calculations

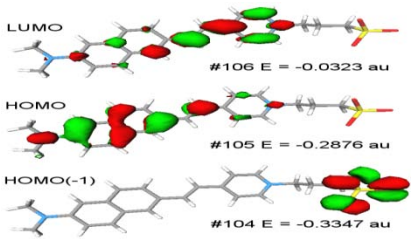
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Science with scaling high accuracy

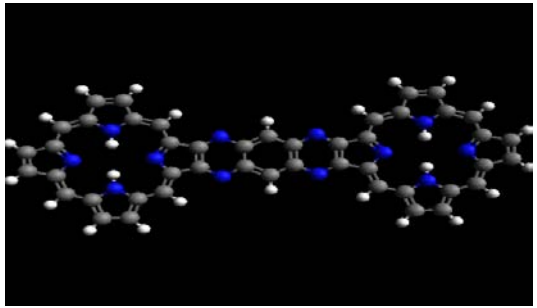
System	Time per EOMCCSD iteration (1024 CPUs)
di-8-ANEPPS	
6-31+G* (570 bsf)	1680 sec. (no symm.)
cc-pVDZ (516 bsf)	1500 sec. (no symm.)
P_2TA	
cc-pVDZ (942 bsf)	1120 sec. (D_{2h} symm.)



P2TA: scalability of the triples part ofthe CR-EOMCCSD(T) implementation.
Only orbitals with corresponding orbital energies below 1.5 Hartree were correlated (total 578 bsf)



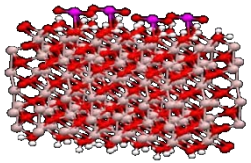
Excited-state calculations for fluorescent probe (di-8-ANEPPS) used in monitoring physiological processes in biological cell.



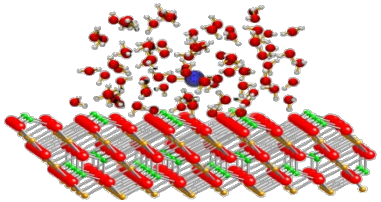
Porphyrin dimer linked by a tetraazaanthracene bridge (P_2TA)

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

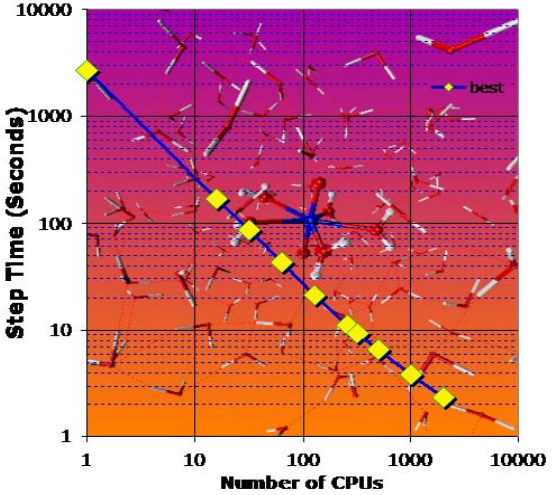
NWChem: Plane wave at the petascale



Uranyl on a hydroxylated Al_2O_3 surface



Uranyl in solution interacting with iron oxide



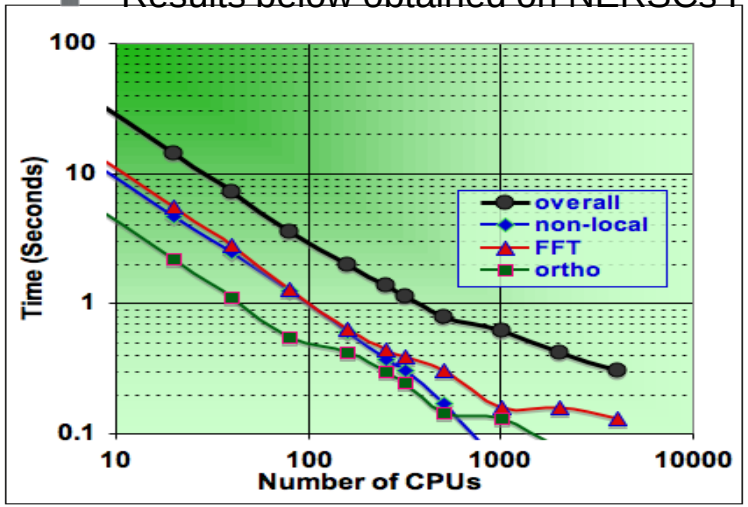
Car-Parrinello plane wave performance, PBE96 GGA Functional, -300 K thermostat, 0.121 fs time step, 122 water molecules-15.6 Å box

Pacific Northwest NATIONAL LABORATORY

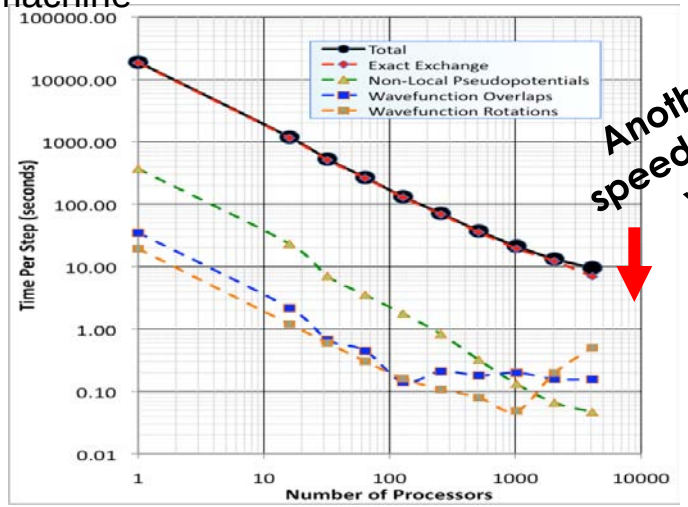
Proudly Operated by Battelle Since 1965

NWChem: Hard scaling of plane waves to 40K procs

- Extensive work done to develop parallel plane wave algorithm for hybrid-DFT solvers
  - Results below obtained on NERSCs Franklin machine



DFT calculation on $Nb_{10}O_{28}^{6-}$– O(Ne)



Hybrid DFT calculation on 80 atom cell of hematite– O(Ne*Ne)

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Moving towards exascale

- Efficient NWChem on exascale platform enables our users to
  - Tackle larger molecular complexes with greater accuracy
  - Integrate of methods and dynamics
  - Obtain their results with a faster time-to-solution
- Goal is to enable NWChem to run effectively on exascale platforms and beyond
  - Understanding scalability issues in computational chemistry
  - Software infrastructure and architecture (re)design
    - 10,000s to 100,000s processors
    - Changing computing paradigm
    - Efficient load balancing (actor based models)
    - Fault resilience and/or tolerance

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Preparing for new hardware architectures

- Multicore nodes
  - Heterogeneous nodes or heterogeneous computing in a node
    - Not all CPUs doing the same thing
    - Utilizing cache
  - Memory bandwidth issues
- GPGPU
  - Rapidly developing field
  - Mostly single precision, but in some cases it works
- Dealing with hardware failures (Assuming the OS can handle it)
  - How to recover the missing piece?
    - Recompute or RAID configuration of data or …
- Heterogeneous networks
  - Align data movement with network topology/bandwidth

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: Scalability will be a major effort

- Crucial for exascale development will be a team of
  - Computational chemists
  - Mathematicians
  - Computer scientists
  - Collaborative effort of diverse disciplines

- Efforts underway in NWChem
  - PNNL Exascale Initiative LDRD led by Karol Kowalski
  - PNNL Extreme Scale Initiative LDRD Abhivav, de Jong

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

NWChem: An open-source development platform

- Greatly expanding user base
  - Ease of access for everybody
- Expanding developers base by going open-source
  - Leverage development and resources of universities and national labs
  - Resources needed to maintain NWChem state-of-the-art exceed people resources in any group
  - Already have significant external developers base
- NWChem as the community code for computational chemistry
  - Provide infrastructure to build upon
  - Consortium model
  - Interested in being part of this effort, send me an email
- Establish more collaborative development environment

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

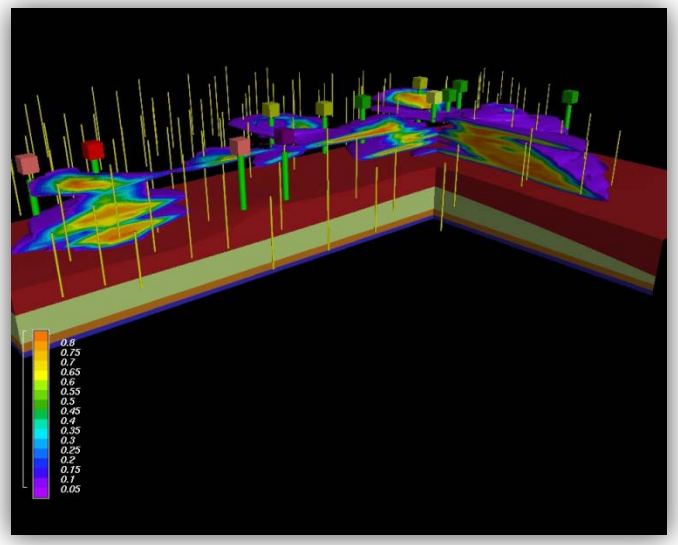
STOMP

- **S**ubsurface **T**ransport **O**ver **M**ultiple **P**hases
- High-performance subsurface flow and reactive transport simulator
- Used for a broad range of applications including contaminant transport, remediation, multiphase flow, and geologic carbon sequestration.
- http://stomp.pnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

STOMP: Scalable and efficient

- Scalable STOMP application allows efficient simulation of problems with millions to billions of variables
- Overcomes major roadblocks in subsurface simulation:
  - Geologic systems with multi-scale heterogeneity
  - Multi-phase fluids, transient flow, multi-component reactions
  - Uncertainty quantification through many alternative forward simulations or inverse modeling



Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

STOMP: Leadership

- The subsurface Science Focus Area research program at PNNL is the **largest among seven participating laboratories** at $6.5M per year;
  - STOMP is a key computational capability within the program.
- Leading two of the three multi-million-dollar Integrated Field Research Challenge (IFRC) research efforts funded by DOE.
- STOMP is licensed for application
  - at Department of Energy sites under DOE Order 414.1C (System Safety Software)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

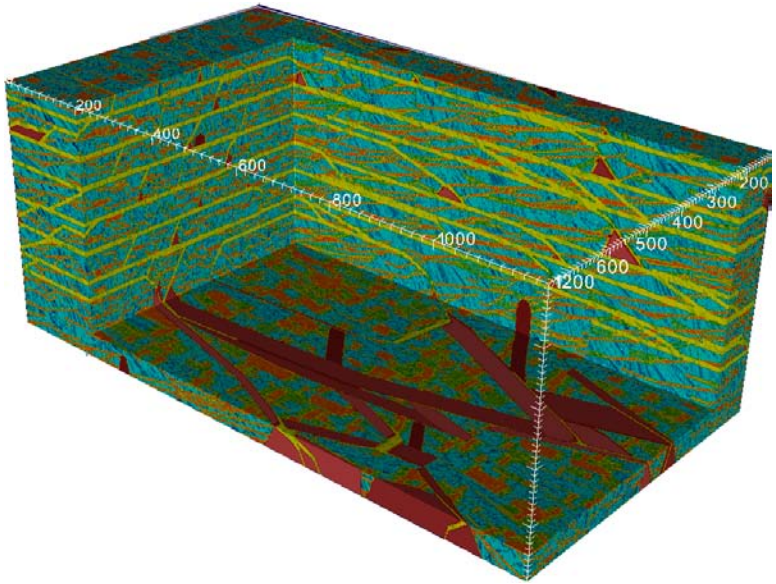
STOMP– Contributions and Collaborations

- A component-based version of STOMP has been developed under the DOE SciDAC program. PNNL leads this effort in collaboration with several SciDAC Centers and Institutes. http://subsurface.pnl.gov
- SciDAC developments are being leveraged through internal investments in STOMP under PNNL's eXtreme-Scale Computation Initiative (XSCI).

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Exteme-Scale Simulation of Multi-Scale Geologic Systems

- Analysis of upscaling of reactive transport in multi-scale heterogeneous systems will provide major contributions to one of the grand challenges in subsurface science
- Kilometer-scale domain resolved at sub-meter scale
- Geologically realistic; based on sedimentological studies of braided gravelly river systems
- **STOMP** simulation of flow and transport has been performed in 156-million element grid
- Collaboration between PNNL and Wright State University

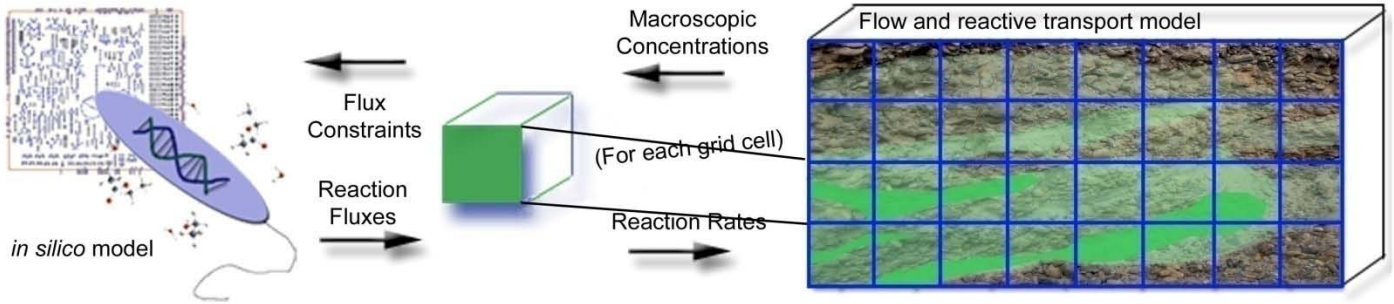


Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

STOMP: Links to computational biology

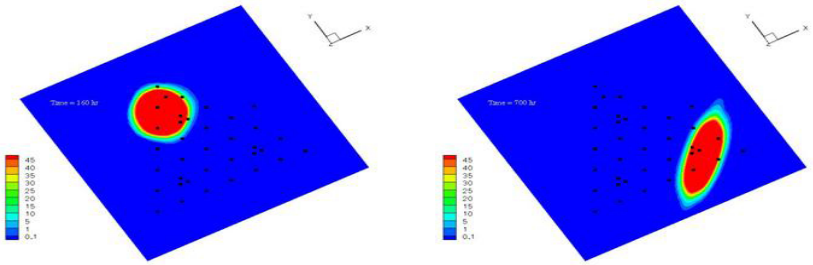
- Research is underway to develop methods for integrating the STOMP simulator with genome-scale models of microbial metabolism
- Applications exist in bioremediation of metals and radionuclides in mine tailings plumes and contaminant release sites.
- PNNL collaboration with University of Massachusetts and University of Toronto



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

STOMP and CCA: Collaboration with SciDac

- Under a PNNL-led SciDAC project, a scalable component-based version of STOMP is being developed.
- Using the Common Component Architecture (CCA) framework provides a high level of modularity and extensibility
  - This new software architecture will provide interchangeable components for structured and unstructured grids, alternative physics and chemistry modules, and multiple solvers
  - Test runs have been made based on the Hanford Site 300 Area model (see figure above)

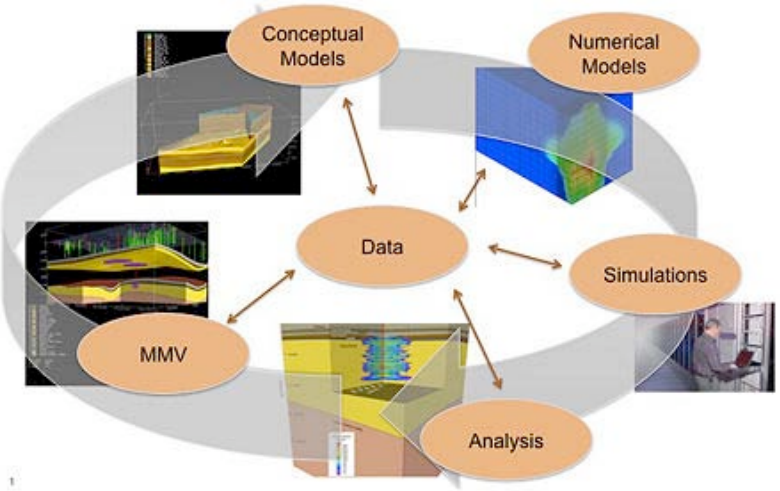


Simulation of transport of contaminant through an array of monitoring wells at the Hanford IFRC site using the component version of STOMP

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

STOMP: Geological carbon sequestration

- Deep underground storage of supercritical carbon dioxide (CO_2) captured from power plant emissions is one alternative being explored as a mitigation measure for global warming.
- A version of STOMP incorporates physics and chemistry of supercritical CO_2
- We are investing in creation of a computational platform for CO_2 sequestration modeling and a CO_2 version of the scalable STOMP code



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

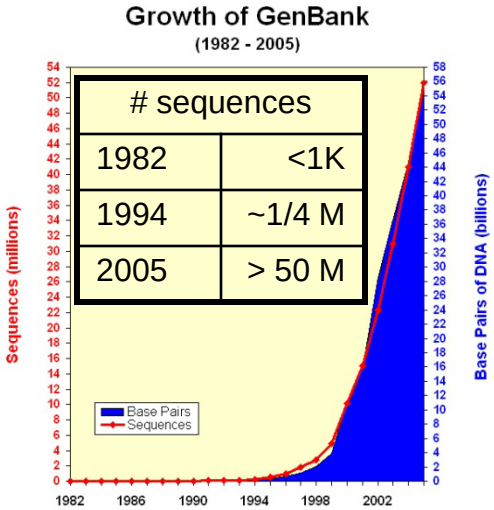
ScalaBLAST

- High Performance Driver for Biological Sequence Alignment
  - Dynamic scheduling of independent search tasks
  - Maintain copy of reference library in core, not on disk
  - Independent output for each compute core
  - Scales to *thousands of cores* on shared or distributed memory systems

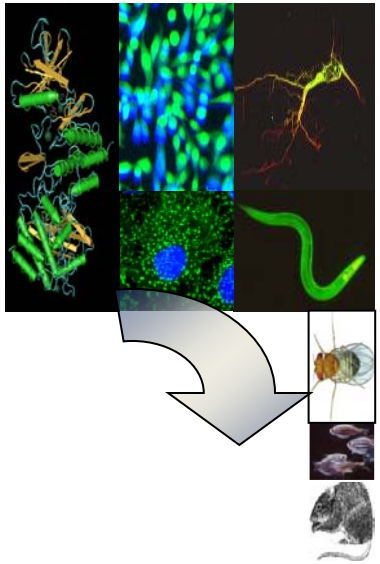
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST: Sequence and down-stream analysis

- Drives sequence analysis on the scale of multiple genomes
- Essential for analyzing metagenomes, communities, phylogeny, etc…
- Workhorse application to drive downstream analysis for
  - Genome, next-gen sequencing, gene prediction, calling
  - Protein ortholog mapping, functional prediction
  - Genome-wide studies



Increasing Data Volumes



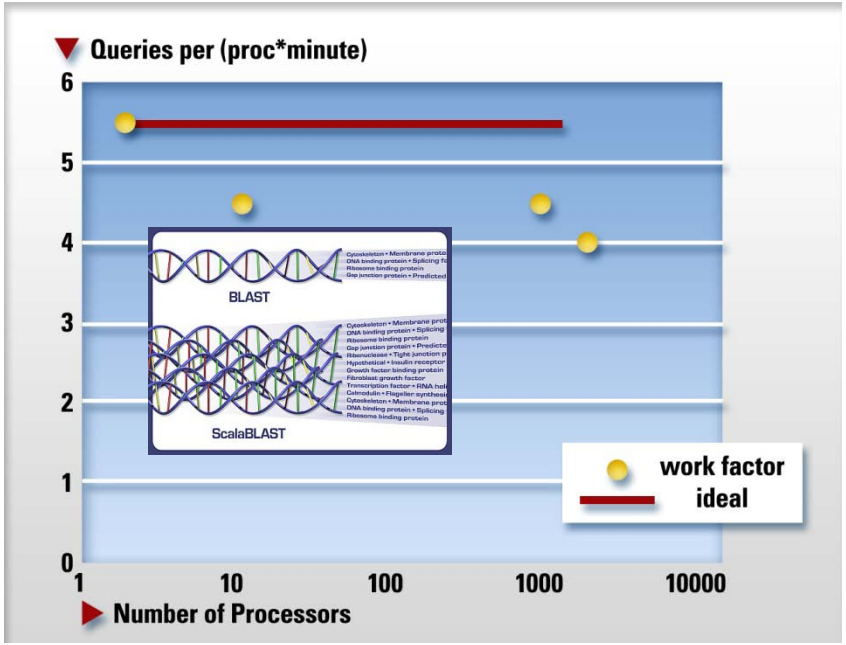
Increasing Data Complexity

http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST: Leadership in computational biology

- Leading the way in high-performance sequence analysis with
  - Ideal scaling for large and small-scale tasks on a variety of system sizes and configurations
  - Web interface for wide accessibility
  - Fault-resilient programming model to enable scaling to petascale and beyond
- ScalaBLAST was an essential component in the Joint Genome Institute's Integrated Microbial Genome (IMG) release.



Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

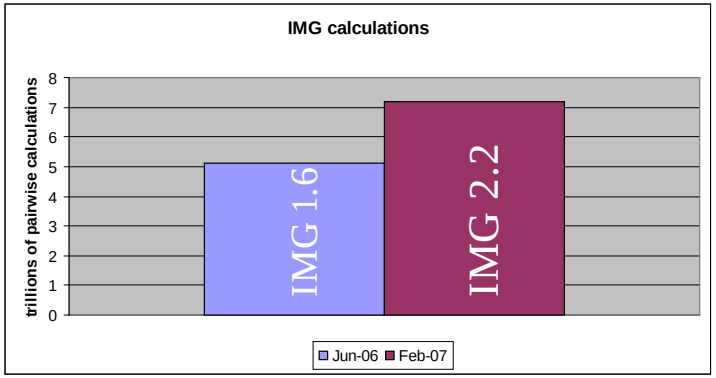
ScalaBLAST: Interdisciplinary emphasis and collaborations

- Current and emerging collaborations with major genome sequencing centers and research centers
  - Joint Genome Institute
  - University of Texas Health Science Center, San Antonio
  - Great Lakes Bioenergy Research Center
- Interdisciplinary emphasis on algorithms, efficient implementation, and science applications
- ScalaBLAST was a critical component in the **SC08 HPC Visual Analytics Challenge winning entry from PNNL**

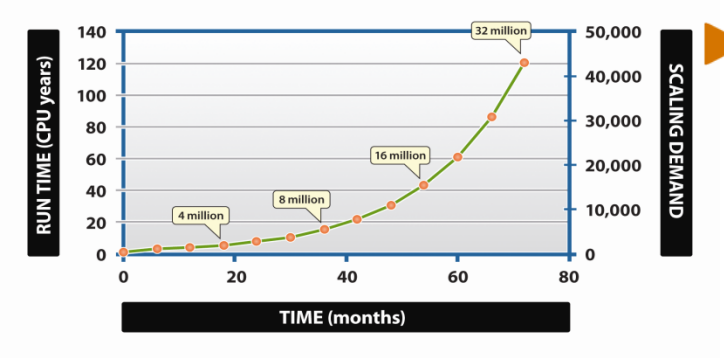
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST for Joint Genome Institute

- Collection of 3000+ microbial, eukaryal and viral genomes
  - ~thousands of proteins per genome
  - n^2 sequence pairs to align



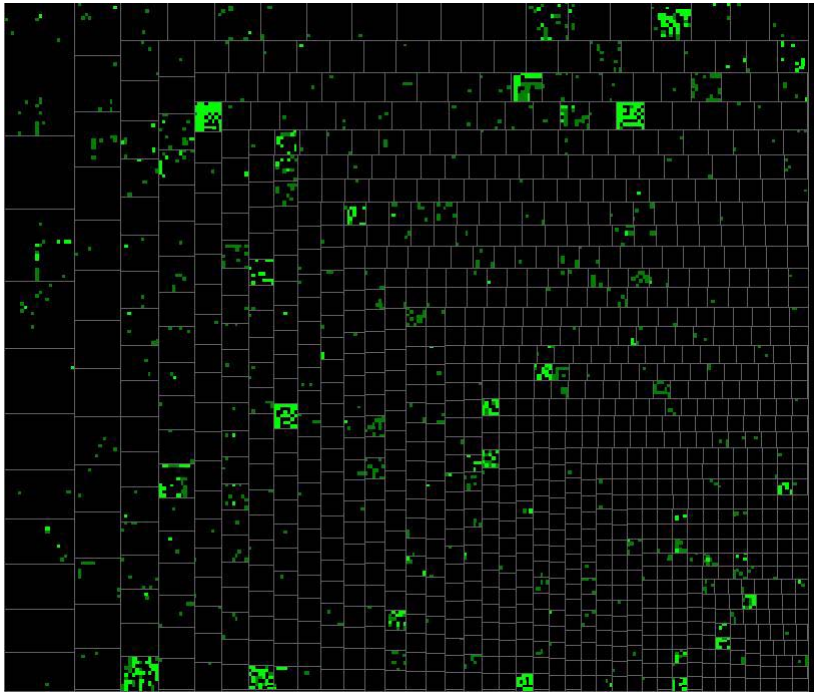
Growth in compute demand over 8 month period



- Dataset doubles every 18 months!
  - Want to continue to deliver comparison output in days
  - Will require scaling to much larger systems

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST for Proteomics

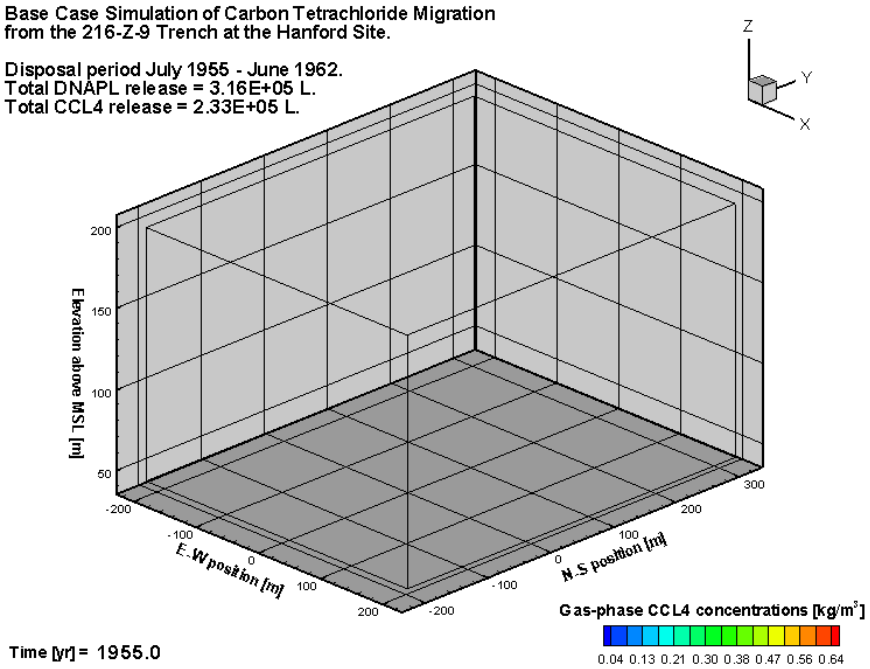


TreeMap view of protein fragments from over 100K clusters grouped using ScalaBLAST output

- Proteomics for metagenomes and communities
  - ScalaBLAST helps identify representative sequences for highly homologous proteins across species
  - ScalaBLAST output feeds into large-scale clustering calculations to enable grouping of detected protein types from community samples

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST for Real-Time Hypothesis Discovery

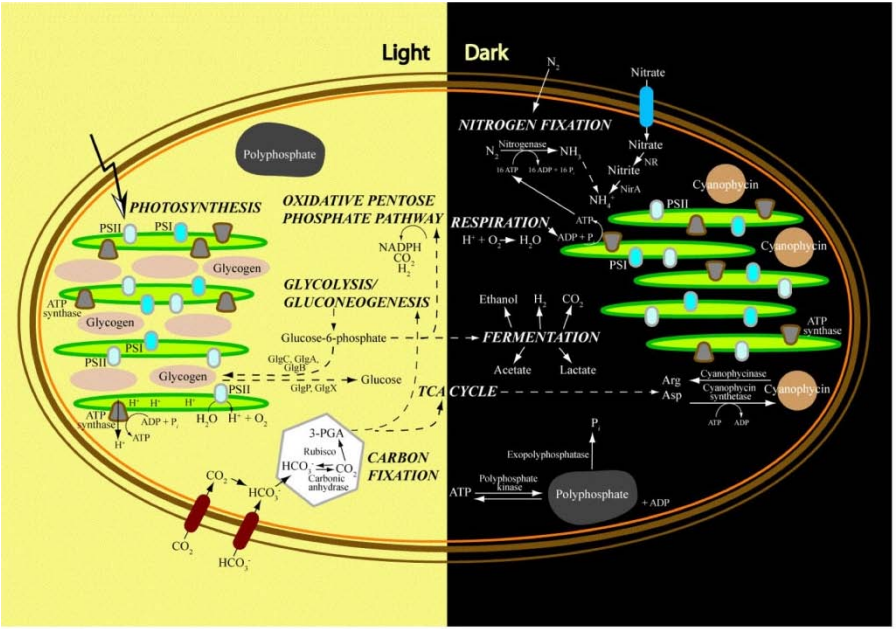


- ScalaBLAST used for iterative hypothesis discovery and refinement
  - Combined with Starlight visual analytics tool
  - Launched on HPC cluster from laptop interface using MeDICi – a PNNL developed middleware framework

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST for Genome Sequencing

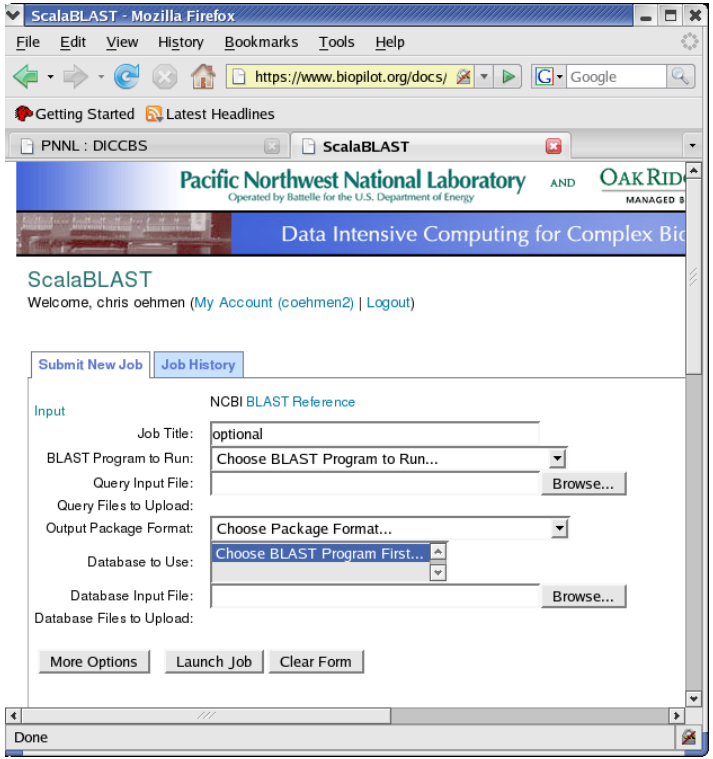
- BLAST is an algorithm used to compare biosequences
  - Gene calling relies on finding matching sequence segments in DNA
  - Predicting function of gene products involves BLAST calculations
- ScalaBLAST was used for finishing the Cyanothece genome
  - Eliminates BLAST calculations as rate-limiting step in gene calling and annotation



Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ScalaBLAST Web Interface

- Bringing high performance computing to bench biologists
  - Web front-end that allows users to upload many queries in a single file
  - Automatically schedules and notifies upon completion
  - Enables clusters to be directly usable by biologists



Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Future of Scalable Applications at PNNL

- PNNL is leading the way enabling sciences at the petascale and beyond with algorithms for
  - Graph searching and irregular memory access
  - Fault resilience and fault tolerance
  - Actor-based models for efficient load balancing

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965