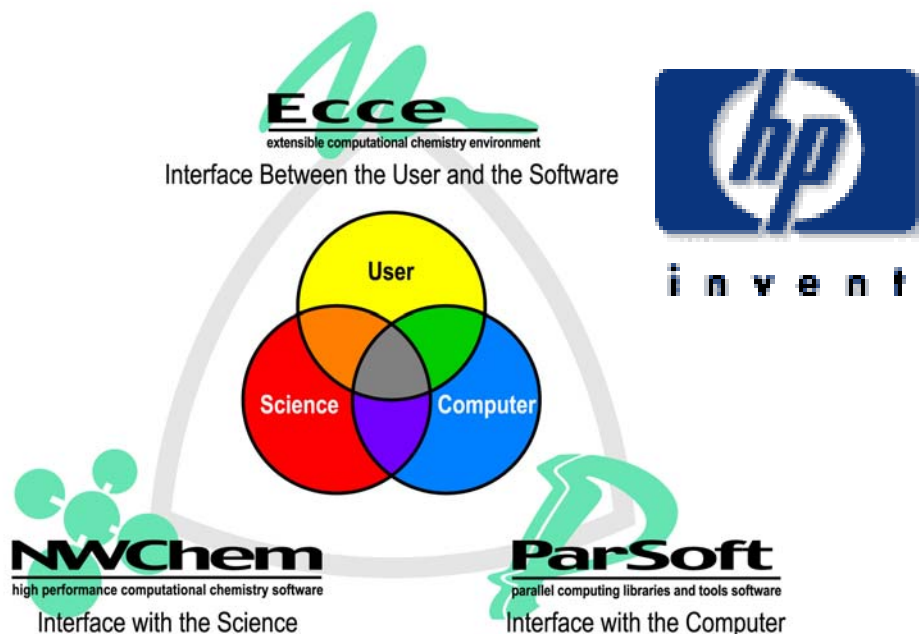


- The chemistry community is extensive - a wide range of experimental, computational, and theoretical approaches are used.
- Chemistry is one of the base sciences on which many applications are built.
- Extensive use of basic chemical measurement techniques in a wide range of areas including:
  - atmospheric measurements
  - geochemical measurements
  - combustion and chemical process measurements
  - cellular observation
- Computational chemistry covers a wide range of areas
  - accurate calculations on small molecules/processes such as heats of formation of radicals, chemical reactions and electron scattering
  - intermediate accuracy calculations for the study of large molecules, separation systems and catalysts,
  - molecular dynamics simulations of complex systems such as biomolecules and materials.

# MSCF: HP & OBER's Flagship High Performance Computing Facility

- Prototypical Topical Computing Facility with integrated hardware, software (applications, tools, & systems), support, training, graphics and visualization, data storage and management, & software distribution
- Worlds Fastest & Largest Linux Cluster: 9.1TF(0.25TF); 173TB Disk – 3500 drives (6TB); 1400 processors in 700 nodes with 3.8 TB memory (150% of NWmpp1) \$24.5M for a 3 year lease
- Impact to EMSL users
  - More than 36 times more powerful. We can go from 20 to 25GFlops sustained per ~10 Grand Challenge projects to 300 to 500 Gflops sustained for 16 to 20 GC projects



# **Computational Chemistry**

## **Electronic structure**

- This area is a huge consumer of computational resources in terms of cpu cycles, memory, and I/O but will not require large networking resources for moving data around.
- For computing across the Grid, the networking needs change significantly in terms of latency and bandwidth.

## **Molecular dynamics**

Based on the new EMSL/MSCF HP supercomputer, a 9+ TFlop system, we estimate that we will be generating ~20 Gbytes/hr leading to 400 to 500 Gbytes/day/~10 Tflop machine.

## Experimental Science

- **Mass spectroscopy for high throughput proteomics** – current estimate is ~100 Gbytes/day/mass spec based on a 50% improvement of the current prototype data collection system. For 20 to 50 mass specs for the entire complex, this leads to 2 to 5 TBytes/day.
- **Imaging** – When running at full speed, a FRET analysis of a cell will be generating a Megapixel/msec of data. A Megapixel is  $\sim 24 \times 10^6$  bits leading to  $3 \times 10^9$  bytes for a 1000 step for a 1 sec trajectory. For 50 runs a day, this will generate  $\sim 150$  Gbyte/spectrometer. We are building 6 spectrometers currently leading to  $\sim 1$  TB at 1 laboratory. If this is spread across multiple labs, this number could easily reach 10 to 20 TB/day.

## Examples of imaging techniques

- **Real-time Dual-Wavelength Confocal Microscopy.**

- allows ratiometric imaging of live cells at 30 fps. Applications, such as FRET in live cells and dynamic colocalization of multiple fluorescent probes can be conducted on timescales of minutes to hours.

- **CARS (coherent anti-stokes Raman spectroscopy)/Two-Photon Confocal Combined Microscope.**

- visualize molecules based on their vibrational properties and simultaneously by their fluorescent properties. CARS is useful for visualizing selected molecular species, such as lipids and deuterated compounds. Spatial resolution is 0.3 microns. The present temporal resolution is about 1 min. per image, but will improve with new laser technology.

- 0.5 and 1.5 Gbytes of image data/day

- **MRM/OM**

- combined magnetic resonance imaging/spectrograph system with simultaneous confocal microscopy capability.

- The magnetic resonance provides unique physical and chemical information on a distance scale as small as 10 microns, with no limitations on sample opacity, and is both non-destructive and non-invasive.

- The rather slow MRM imaging time (10s of minutes, depending on resolution) is supplemented with optical fluorescence images. Confocal images are recorded in seconds with micron resolution, simultaneously with MRM imaging.

- 0.5 and 1.5 Gbytes of image data/day

- **AFM-Enhanced Fluorescence Imaging Microscopy**
  - nanoscale characterization combined microscope that uses the Optical/AFM approach and a high-sensitivity far-field microscope to provide unperturbed measurements of reaction rates at sites that have been characterized by atomic force microscopy (AFM) imaging.
- **Single-Molecule Spectroscopy and Imaging Microscopy**
  - simultaneous structural and spectroscopic analyses of single biomolecular complexes and their reaction (interaction and association) rates, providing insight into the relationship between structure and function of cell signaling proteins and enzymes.
  - The information collected from these approaches is typically lost or hidden in the measurements using current conventional ensemble-averaged methodologies.
- **Single-Molecule Patch-Clamp/Optical Confocal Imaging Microscopy**
  - combining a confocal scanning linear/non-linear optical microscope with state-of-the-art patch clamp technologies.
  - This instrumentation significantly enhanced the diagnostic and investigative capabilities of both methods in the characterization of ion channel/receptor dynamics and mechanism in a living cell.
  - Understanding the correlated conformational dynamics of a single ion channel/receptor will be a significant step in deciphering the molecular mechanisms of ion channels functions and the process of ligand-receptor and receptor-lipid interactions for individual molecules in membranes from living cells.

- **Nuclear magnetic resonance**

- A protein structure experiment generates 1 to 1.5 Gbyte/run and one can do 2 to 3 runs/day leading to 3 to 5 Gbyte/spectrometer. Many spectrometers could easily lead to 100's of Gbytes/day.

- **Cryo electron microscopy(EM)/EM crystallography of single molecules**

- digitized electron micrograph (photographic film), consists of approximately 6000X6000 pixels, usually digitized in 12 bits and go to 1000 images/day/spectrometer. This is 54 Gbytes/spectrometer.

- **Synchrotron data**

- Produces large data sets.

- Experimentalists have been hampered by the lack of local computing facilities at the synchrotron source which prevents analysis of data and can hold up experiments.

- **Real time, remote control of experiments**

- This requires significant networking resources that are always available to the remote user.

- The experiment transmits data back to the user which enables the user to make decisions about how to control the experiment.

- The amount of data transmitted *by* the user is usually small but a high integrity network is needed with low latency.

- The amount of data transmitted *to* the user may be much larger so bandwidth and latency are important for real time control.

## *2 General Scenarios*

1. Move data around to everyone from a small number of very high performance computers and store data from everyone at a small number of very large data storage centers
2. Place many modest size computers working locally on the data with large amounts of local storage. Computers and storage are co-resident with the data Only summary data is passed around on the network.

Scenario 1 requires high network speeds with reduced manpower needs.

Scenario 2 requires significant manpower needs to manage all of the computer systems.

# Issues

Networking and other information technology research needed to enable distribution of data, analysis, and collaboration including:

- Multicast service delivered to multiple remote centers with diverse firewall filters
- Network error rate and robustness control *without* impacting the experiment's data acquisition system
- Massive applications software – e.g., 1 million lines of code in NWChem
- Interfaces of databases with the network and storage
- Technology improvements including:
  - Computing technologies
  - Computer science
  - Applied mathematics
  - Software development
  - Networking
  - Computing system-to-network interfaces
  - Fiber technologies
  - Data storage
  - Data management

*Questions:*

*How will the different sets of data be processed across the networks?*

*How often does one really want to move the data (i.e. will go to a repository once or will it be moved around more often)?*

*Will only a subset of the data be moved (needing data mining technology)?*

*Will the data only be visualized remotely?*

*What fraction of those results are moved to the user's home institution for analysis?*

- The number of locations with big chemistry data sources will always be larger than the number of places with big computers, especially as physical chemistry is used in biological analysis (e.g., NMR, mass spec).
- We are always going to need high performance networks to bring big data, computers and people together.
- We need to reduce data size before we transmit it, which will require new algorithms that are broadly supported.