Moving towards Desktop Grid Solutions for Large-scale Modeling Simulations in Computational Chemistry

M. Taufer, M. Crowley, J. Karanicolas, P. Cicotti, A. A. Chien, C. L. Brooks III

University of California at San Diego - The Scripps Research Institute

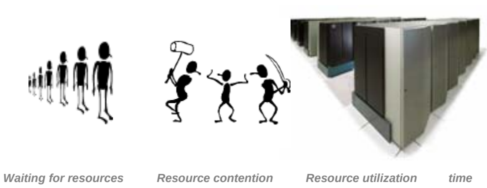
Need for more Compute Resources

The Microsecond Barrier in Molecular Dynamics

- Chemistry phenomena may occur over long timescales, e.g.:
  - protein folding needs microseconds to milliseconds
  - molecular motions of interest need microseconds or longer
  - ligand-receptor interactions converge very slowly
- Molecular dynamics calculations to study these phenomena:
  - a powerful method to investigate nature of inter-molecule and intra-molecule dynamics
  - limited to nanoseconds due to lack of compute resources

"Microsecond barrier can be overcome for molecular dynamics running on expensive supercomputers for many months" [Duan and Kollman]

Using Supercomputers for many Months



We need much more compute resources than are available at current supercomputer centers

Searching for New Compute Resources: the Desktop Grid

Characterization of Desktop Grids

- Large numbers of heterogeneous desktop PCs connected to the Internet and Intranets
- Utilization of otherwise unused compute power → 80%-90% of CPU time is idle time
- Nodes join and leave the grid frequently
- Unreliable, loosely-coupled interconnections
- Intrusive users and malicious attacks

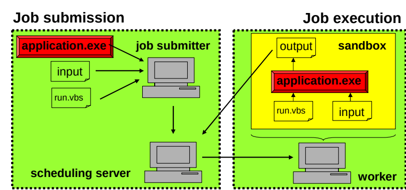
Challenges using Desktop Grids

- How can we optimally utilize a large amount of desktop PCs to stage high-quality, large-scale simulations?
- How can we deal with the heterogeneity of desktop grid platforms?
- Do the characteristics of the system (i.e. kind of resource, compute paradigms) affect the quality of the simulation results?

Our Testbed: the Entropia DCGrid at SDSC

250 heterogeneous workers (desktop PCs)

- CPU speeds: from 300MHz to 2.2GHz
- Main memory: from 256MB to 1GB
- Disk capacity: from 20GB to 100 GB
- Network interconnection:10/100 Mbps & Gigabit Ethernet
- Five major Windows OSs: Windows NT, Windows 98, Windows 2000, Windows 95, Windows XP

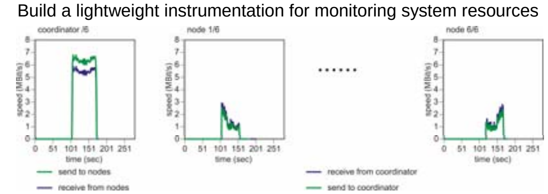


Building Effective Desktop Grids

Gain control over resource availability and resource usage for efficient matching of application requirements to desktop grid specifications

Combining Monitoring and Modeling of System Resources

Build a lightweight instrumentation for monitoring system resources



Modeling resource usage for a wide range of applications

$$time = f(CPU, memory, I/O)$$

Future Steps: Matching Applications and Resources

- Extend the range of computational chemistry problems running on desktop grid by:
  - Exploiting new programming paradigms (i.e. combining task- and data parallelism)
  - Introducing new computing models (i.e. peer-to-peer technology)
  - Adapting the resource usage at run-time based on system performance and system behavior
  - Preventing schedule and resource conflicts
- Predict viability of new applications on desktop grid:
  - Based on modeling of system resources and application requirements

The Scientific Computation Code - CHARMM (Chemistry at HARvard Macromolecular Mechanics)

- CHARMM is a computational chemistry code for studying the structure and the dynamics of relevant molecules such as proteins, DNA, RNA
- CHARMM uses classical mechanical methods for molecular dynamics simulations (MD)

$$\delta^2 s_i / \delta t^2 = F_i / m_i$$

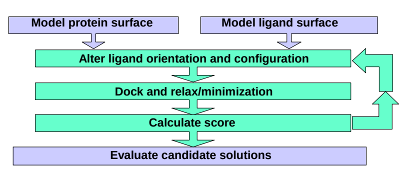
s_i displacement of atom i at time t

F_i force and m_i mass of atom i

- CHARMM MD simulations are used to investigate thermodynamic and kinetic of molecules

Beneficial Applications

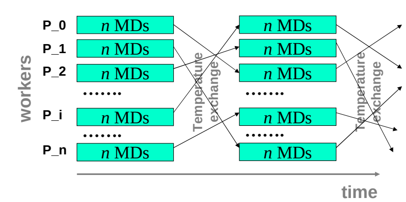
- Protein-ligand docking simulations:
  - Scoring of potential ligands to docking sites of proteins and nucleic acids



  - Efficient resource usage for computing large range of protein-ligand alternatives

- Free energy surface calculations
  - Study protein folding and functions by sampling region of conformation space with umbrella sampling method
  - Effective grouping of desktop PCs in clustered work groups to shorten turn-around time of single tasks
- Search for a structural model of amyloid fibrils
  - Study of domain-swapped aggregation between two structural elements of two monomers to build fibrils
  - Efficient resource usage for greater sampling at each possible linker region between monomers

- Replica exchange calculations:
  - More accurate sampling of energy and conformation space to study protein folding and protein functions



  - Efficient scheduling for low overhead due to required synchronization and all-to-all communication