Grid Computing as a Platform for Parallel Computing

Mark P. Wachowiak, Ph.D.
Department of Computer Science and Mathematics
Nipissing University

June 5, 2007

Grid computing

- Definition:
  - “Grid computing is distributed computing performed transparently across multiple administrative domains” (P.V. Coveney).
  - Distributed high-performance computing.
  - Large geographically distributed networks of computers.
- Provides a means for marshalling distributed resources to solve large problems.
- “What the Web did for communication, grids endeavor to do for computation.”

Grid computing (2)

- Very general computing applications:
  - Database searches and queries.
  - Simulation, visualization, computer-aided manufacturing.
  - Access to computer-controlled laboratory experiments.
  - Many other applications.
- Transparency:
  - Distributing computational resources among multiple and widely separated sources and users is a difficult ***algorithmic*** problem.

Characteristics of grids

- Grids coordinate resources that are not subject to centralized control.
- Grids use standard, open, general-purpose protocols and interfaces.
- Grids deliver nontrivial qualities of service.

Grid vs. parallel computing

- Grid computing is distinguished from parallel computing on one or more multiprocessors:
  - Parallel computing: locally “clustered” machines or large supercomputers.
  - Grid computing: computation across different administrative domains.

Characteristics of grid applications

- Data acquired by scientific instruments.
- Data are stored in archives on separate, perhaps geographically-separated sites.
- Data are managed by teams belonging to different organizations.
- Large quantities of data (tera- or petabytes) are collected.
- Software used to analyze and summarize the raw data.

Recent Applications

- Earthquake forecasting (Gorder, IEEE CSE, 2007).
- Ecosystem modeling (Wang et al., IEEE CSE, 2005).
- Multiobjective optimization (Nebro et al., Soft Computing, 2007).
- Human Proteome Folding Project.
  - First phase officially completed on July 18, 2006.
  - Phase II focuses on better protein models – more computationally intensive.
- Data-driven reservoir studies (Parashar et al., 2005).
- Smallpox research.
- Molecular modeling (Sudholt et al., 2005).
- SETI@Home.

Grid Middleware

- Globus toolkit
  - Most widely-used middleware for grids.
  - Open source toolkit for building computing grids.
  - Provides a standard platform upon which other services build.
  - Provides directory services, security, and resource management.

Grid Middleware (2)

- Condor
  - Eliminates need for re-configuring software.
  - Existing code linked with specialized libraries.
  - Checkpointing for robustness in the event of failures.
  - Advanced CPU scavenging.
  - Provides transparency for security issues.
- gLite
  - Part of the European EGEE (Enabling Grids for E-Science) project.
- Many others.

Future Directions

- More advanced CPU scavenging.
- Dynamic load balancing.
- Security issues.
- Robustness in the face of component failures.
- Peer-to-peer systems (e.g. Caromel et al., 2007).
- New parallel computational applications.