FutureGrid Education: Using Case Studies to Develop A Curriculum for Communicating Parallel and Distributed Computing Concepts

Jerome E. Mitchell, Judy Qiu, Massimo Canonio, Shantenu Jha, Linda Hayden, Barbara Ann O’Leary, Renato Figueiredo, Geoffrey Fox
Outline

• Motivation and Introduction
• FutureGrid Education
• The Need for Community Modules
• Modules and Platform Packages
• Modules for Teaching Parallelism
• Case Studies
  – Examples of Associated Modules
• Conclusion and Future Involvement
Introduction and Motivation

- Cloud computing => 15 Million Jobs
- Data exploding faster than Moore’s Law
- The Problem:
  - Some educators lack understanding
  - Limited/nonexistence resources
- The Solution:
  - Use online/local resources coupled with teaching materials to supplement disparities
FutureGrid: a Grid/Cloud/HPC Testbed

Internet 2/ TeraGrid

Core Router

NID: Network Impairment Device

UC: 7TF IBM 672 cores
PU: 4TF Dell 384 cores
IU: 11TF IBM 1024 cores
    7TF Cray 684 cores
    5TF Disk Rich 512 cores
UCSD: 7TF IBM 672 cores
TACC: 8TF Dell 768 cores
UF: 3TF IBM 256 cores

Private
Public
FG Network
FutureGrid Education

• Executable modules – virtual appliances
  – Deployable on FutureGrid resources
  – Deployable on other cloud platforms, as well as virtualized desktops

• Community sharing – Web 2.0 portal, appliance image repositories
  – An aggregation hub for executable modules and documentation
The Need for Community Modules

• Community benefit from services and features for anyone willing to learn parallel and distributed computing

• We are developing, distributing, and supporting modules and platform packages to serve diverse needs of the community and to advance contribution in computing
A Guide for Teaching Parallelism


Developing a set of essential learning objectives for each knowledge area, to serve as a guide for incorporating parallelism topics into modules.
Modules for Teaching Parallelism

• Develop modular and flexible, teaching materials, so it can be supported in a variety of environments
  – Compatibility with many organizations approaches to learning parallel and distributed computing
  – Simple and quick to deploy in a course, workshop, summer school
  – Minimal investment required of instructors using materials, including those who are NOT specialists in parallel and distributed computing
Educational ‘Platform Packages’ Appliances

• A flexible, extensible platform for hands-on, lab-oriented education on FutureGrid
• Need to support clustering of resources
• Virtual machines + social/virtual networking to create sandboxed modules
  – Virtual ‘grid’ appliances: self-contained, pre-packaged execution environments
  – Group VPNs: simple management of virtual clusters by students and educators
Appliance Infrastructure

• Deployability: Students and users should be able to deploy modules in a simple manner, and in a variety of resources
  – Reduce barriers to entry; avoid dependences upon a particular infrastructure

• Community-oriented: Modules should be simple to share, discover, reuse, and expand
  – Create conditions for ‘viral’ growth
Case Study: Cloudy View on Computing workshop
Modules Associated with Workshop

• Modules:
  – Introduction to parallel and distributed processing
  – From functional programming to MapReduce and the Google File System (GFS)
  – Graph Algorithms with MapReduce

• Assignments
  – “Hello World” MapReduce Lab
  – Twister PageRank Lab
  – Hadoop BLAST Lab
Case Study: Distributed Scientific Computing at Louisiana State University

• FutureGrid supported activities in a semester-long course offered in Fall 2012

• A practical and comprehensive graduate course preparing students for research involving distributed scientific computing
  – Taught by Shantenu Jha
  – Topics:
    • Introduction to the practice of distributed computing
    • Cloud computing and master-worker pattern
    • Distributed application use cases
Modules Associated with Scientific Computing Course

• Modules:
  – Introduction to Numerical Methods
  – Vector Algebra, Basic Visualization Programming
  – Best Coding Practices

• Assignments
Case Study: Cloud Computing at Indiana University
Modules Associated with Cloud Computing course

• Modules
  – Introduction to Data Intensive Sciences
  – Parallel Programming/MPI vs. MapReduce/Hadoop
  – MapReduce on Multicore/GPU

• Assignments
  – Twister K-means Lab
  – Hadoop/Twister Pairwise distance Calculation using SWG Lab
Case Study: Distributed Systems at Indiana University

• FutureGrid supported activities in a semester-long course offered in Spring 2011
• A practical and comprehensive graduate course preparing students for research involving distributed systems
  – Taught by Judy Qiu
  – Topics
    • Design principles, systems architecture, and innovative applications of parallel, distributed, and cloud computing systems
Modules Associated with Distributed Course

- **Modules**
  - Computer clusters for Scalable Parallel Computing
  - Introduction to Distributed Systems, Architectures, and Communication
  - Processes, Performance Issues, and Synchronization

- **Assignments**
  - Page Rank MPI
  - Build a dynamic virtual cluster
Case Study: Cloud Computing at the University of Piemonte-Orientale
Modules Associated with the Cloud Computing course

• Modules
  – Introduction to Cloud Computing
  – Introduction to Eucalyptus, Nimbus, and OpenNebula
  – Eucalyptus: Image Management; Monitoring and Cloning
  – HybridFox

• Assignments
Conclusion

• Concurrency revolution has sparked the need to teach parallelism to a diverse community

• We are developing community-supported modules and platform packages

• Demonstrated successful teaching materials through case studies from a variety of environments
Future Plans

• Develop a ‘cloud computing handbook’ based upon
  – Distributed and Cloud Computing by Kai Hwang, Jack Dongarra, and Geoffrey Fox
  – FutureGrid

• Innovate new ways for reaching a broader audience through web 2.0 technologies
  – presentation tools
  – community tools
Science Cloud Summer School

July 30 – August 3, 2012

The Science Cloud Summer School targets education and training of graduate students and the fostering of a community around a topic that has increasing interest and relevance: the use of cloud computing technologies in science - including infrastructure-as-a-service and platform-as-a-service. Because cloud computing systems and technologies provide a considerable departure from traditional models and evolve at a rapid pace, this event would provide a basis for students to immerse in a focused, intensive curriculum to learn fundamentals and experiment with these technologies in practice. We will cover topics of interest to students with both application and computer science focus.

Organizer:
- Geoffrey Fox, Indiana University

Prerequisites:
Java and HPC experience will be beneficial

Topics:
The Science Cloud Summer School curriculum will cover both technology (computer science) and use of clouds (informatics, computational science).

Coursework will include:
- Introductory Session and Panels
- Infrastructure as a Service
- MapReduce and other cloud platforms (NOSQL) and data-intensive Applications
- Commercial Environments
- Clouds and Cyberinfrastructure
- Education and Clouds

Sites
- Indiana University, Bloomington, IN
- Louisiana State University, Center for Computation & Technology, Baton Rouge, LA
- Michigan State University, Institute for Cyber Enabled Research, East Lansing, MI
- Pennsylvania State University, State College, PA
- Princeton University, Princeton, NJ
- Rutgers University, Piscataway, NJ
- University of California Los Angeles, Los Angeles, CA
- University of Michigan, Ann Arbor, MI
- University of South Carolina, Columbia, SC
- University of Wisconsin - Milwaukee, Milwaukee, WI
Birds of a Feather (BOF)

Hosting Cloud, HPC and Grid Educational Activities on FutureGrid

Renato Figueiredo, University of Florida
Barbara Ann O'Leary, Indiana University

Today at 4:45-5:45 in THIS ROOM
Questions, Comments

Thank You