Cloud Computing for ADMI

ADMI Board Meeting and faculty workshop
Elizabeth City State University
December 16 2010
Geoffrey Fox
gcf@indiana.edu

Director, Digital Science Center, Pervasive Technology Institute
Associate Dean for Research and Graduate Studies, School of Informatics and Computing
Indiana University Bloomington
Talk Components

• Important Trends
• Clouds and Cloud Technologies
• Applications in Bioinformatics
• FutureGrid
Important Trends

• **Data Deluge** in all fields of science
• **Multicore** implies parallel computing important again
  – Performance from extra cores – not extra clock speed
  – GPU enhanced systems can give big power boost
• **Clouds** – new commercially supported data center model replacing compute **grids** (and your general purpose computer center)
• **Light weight clients**: Sensors, Smartphones and tablets accessing and supported by backend services in cloud
• **Commercial efforts** moving much faster than academia in both **innovation** and **deployment**
<table>
<thead>
<tr>
<th>Benefit</th>
<th>Years to Mainstream Adoption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformational</td>
<td>Cloud Computing, Cloud Web Platforms, Media Tablet</td>
</tr>
<tr>
<td>High</td>
<td>Mobile Application Stores, Predictive Analytics</td>
</tr>
<tr>
<td>Moderate</td>
<td>Consumer-Generated Media, Pen-Centric Tablet PCs</td>
</tr>
<tr>
<td>Low</td>
<td>3D Flat-Panel TVs and Displays</td>
</tr>
<tr>
<td></td>
<td>4G Standard, Public Virtual Worlds</td>
</tr>
<tr>
<td></td>
<td>Tangible User Interfaces</td>
</tr>
<tr>
<td></td>
<td>As of August 2010</td>
</tr>
</tbody>
</table>
Data Centers Clouds & Economies of Scale I

Range in size from “edge” facilities to megascale.

Economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K server center.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cost in small Data Center</th>
<th>Cost in Large Data Center</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>$95 per Mbps/month</td>
<td>$13 per Mbps/month</td>
<td>7.1</td>
</tr>
<tr>
<td>Storage</td>
<td>$2.20 per GB/month</td>
<td>$0.40 per GB/month</td>
<td>5.7</td>
</tr>
<tr>
<td>Administration</td>
<td>~140 servers/Administrator</td>
<td>&gt;1000 Servers/Administrator</td>
<td>7.1</td>
</tr>
</tbody>
</table>

2 Google warehouses of computers on the banks of the Columbia River, in The Dalles, Oregon

Such centers use 20MW-200MW (Future) each with 150 watts per CPU
Save money from large size, positioning with cheap power and access with Internet
Data Centers, Clouds & Economies of Scale II

- Builds giant data centers with 100,000’s of computers; ~200-1000 to a shipping container with Internet access
- “Microsoft will cram between 150 and 220 shipping containers filled with data center gear into a new 500,000 square foot Chicago facility. This move marks the most significant, public use of the shipping container systems popularized by the likes of Sun Microsystems and Rackable Systems to date.”
Amazon offers a lot! The Cluster Compute Instances use hardware-assisted (HVM) virtualization instead of the paravirtualization used by the other instance types and requires booting from EBS, so you will need to create a new AMI in order to use them. We suggest that you use our Centos-based AMI as a base for your own AMIs for optimal performance. See the EC2 User Guide or the EC2 Developer Guide for more information.

The only way to know if this is a genuine HPC setup is to benchmark it, and we've just finished doing so. We ran the gold-standard High Performance Linpack benchmark on 880 Cluster Compute instances (7040 cores) and measured the overall performance at 41.82 TeraFLOPS using Intel's MPI (Message Passing Interface) and MKL (Math Kernel Library) libraries, along with their compiler suite. This result places us at position 146 on the Top500 list of supercomputers.

The input file for the benchmark is here and the output file is here.
• **SaaS:** Software as a Service imply software capabilities (programs) have a service (messaging) interface
  - Applying systematically reduces system complexity to being linear in number of components
  - Access via messaging rather than by installing in `/usr/bin`
• **IaaS:** Infrastructure as a Service or **HaaS:** Hardware as a Service – get your computer time with a credit card and with a Web interface
• **PaaS:** Platform as a Service is IaaS plus core software capabilities on which you build SaaS
• **Cyberinfrastructure** is “Research as a Service”
Philosophy of Clouds and Grids

• **Clouds** are (by definition) commercially supported approach to large scale computing
  – So we should expect *Clouds to replace Compute Grids*
  – Current Grid technology involves “non-commercial” software solutions which are hard to evolve/sustain
  – Maybe Clouds ~4% IT expenditure 2008 growing to 14% in 2012 (IDC Estimate)

• **Public Clouds** are broadly accessible resources like Amazon and Microsoft Azure – powerful but not easy to customize and perhaps data trust/privacy issues

• **Private Clouds** run similar software and mechanisms but on “your own computers” (not clear if still elastic)
  – Platform features such as Queues, Tables, Databases currently limited

• **Services** still are correct architecture with either REST (Web 2.0) or Web Services

• **Clusters** are still critical concept for MPI or Cloud software
Grids MPI and Clouds

- **Grids** are useful for managing distributed systems
  - Pioneered service model for Science
  - Developed importance of Workflow
  - Performance issues – communication latency – intrinsic to distributed systems
  - Can never run large differential equation based simulations or datamining

- **Clouds** can execute any job class that was good for Grids plus
  - More attractive due to platform plus elastic on-demand model
  - MapReduce easier to use than MPI for appropriate parallel jobs
  - Currently have performance limitations due to poor affinity (locality) for compute-compute (MPI) and Compute-data
  - These limitations are not “inevitable” and should gradually improve as in July 13 Amazon Cluster announcement
  - Will probably never be best for most sophisticated parallel differential equation based simulations

- **Classic Supercomputers** (MPI Engines) run communication demanding differential equation based simulations
  - MapReduce and Clouds replaces MPI for other problems
  - Much more data processed today by MapReduce than MPI (Industry Informational Retrieval ~50 Petabytes per day)
Cloud Computing: Infrastructure and Runtimes

- **Cloud infrastructure**: outsourcing of servers, computing, data, file space, utility computing, etc.
  - Handled through Web services that control virtual machine lifecycles.
- **Cloud runtimes or Platform**: tools (for using clouds) to do data-parallel (and other) computations.
  - Apache Hadoop, Google MapReduce, Microsoft Dryad, Bigtable, Chubby and others
  - MapReduce designed for information retrieval but is excellent for a wide range of science data analysis applications
  - Can also do much traditional parallel computing for data-mining if extended to support iterative operations
  - MapReduce not usually on Virtual Machines
Higher Education 2020

Motivating Issues
- job / education mismatch
- Higher Ed rigidity
- Interdisciplinary work
- Engineering v Science, Little v. Big science

Computational Thinking
- Modeling & Simulation
- C(DE)SE

Internet & Cyberinfrastructure

C4 Stewards
- C4 Curricula, programs
- C4 Experiences (delivery mechanism)
- C4 REUs, Internships, Fellowships

NSF
- Educate “Net Generation”
- Re-educate pre “Net Generation” in Science and Engineering
- Exploiting and developing C4
C4 EMERGING VISION

While the internet has changed the way we communicate and get entertainment, we need to empower the next generation of engineers and scientists with technology that enables interdisciplinary collaboration for lifelong learning.

Today, the cloud is a set of services that people intently have to access (from laptops, desktops, etc). In 2020 the C4 will be part of our lives, as a larger, pervasive, continuous experience. The measure of success will be how “invisible” it becomes.

C4 Education Vision

*C4 Education will exploit* advanced means of communication, for example, “Tabatars” conference tables, with real-time language translation, contextual awareness of speakers, in terms of the area of knowledge and level of expertise of participants to ensure correct semantic translation, and to ensure that people with disabilities can participate.

C4 Society Vision

While we are no prophets and we can’t anticipate what exactly will work, we expect to have high bandwidth and ubiquitous connectivity for everyone everywhere, even in rural areas (using power-efficient micro data centers the size of shoe boxes).
MapReduce

- Implementations (Hadoop – Java; Dryad – Windows) support:
  - Splitting of data
  - Passing the output of map functions to reduce functions
  - Sorting the inputs to the reduce function based on the intermediate keys
  - Quality of service

A hash function maps the results of the map tasks to reduce tasks
MapReduce “File/Data Repository” Parallelism

Map = (data parallel) computation reading and writing data
Reduce = Collective/Consolidation phase e.g. forming multiple global sums as in histogram

Iterative MapReduce

Map Map Map Map
Reduce Reduce Reduce Reduce
All-Pairs Using DryadLINQ

- Calculate pairwise distances for a collection of genes (used for clustering, MDS)
- Fine grained tasks in MPI
- Coarse grained tasks in DryadLINQ
- Performed on 768 cores (Tempest Cluster)

Hadoop VM Performance Degradation

Perf. Degradation = \( \frac{T_{vm} - T_{baremetal}}{T_{baremetal}} \)

15.3% Degradation at largest data set size

Perf. Degradation On VM (Hadoop)
Sequence Assembly in the Clouds

**Cap3 Parallel Efficiency**

**Cap3 – Time** Per core per file (458 reads in each file) to process sequences
Cap3 Performance with Different EC2 Instance Types

- **Amortized Compute Cost**
- **Compute Cost (per hour units)**
- **Compute Time**

The graph compares the performance of different EC2 instance types based on their compute time and cost. The instance types include Large-8x2, Xlarge-4x4, HCXL-2x8, HCXL-2x16, HM4XL-2x8, and HM4XL-2x16.
Cap3 Cost

Cost ($) vs. Num. Cores * Num. Files

- Azure MapReduce
- Amazon EMR
- Hadoop on EC2

<table>
<thead>
<tr>
<th>Num. Cores * Num. Files</th>
<th>64 * 1024</th>
<th>96 * 1536</th>
<th>128 * 2048</th>
<th>160 * 2560</th>
<th>192 * 3072</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost ($)</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>60</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>70</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
</tr>
</tbody>
</table>
SWG Cost

Cost ($)

64 * 1024 96 * 1536 128 * 2048 160 * 2560 192 * 3072

Num. Cores * Num. Blocks

AzureMR
Amazon EMR
Hadoop on EC2
Smith Waterman: Daily Effect

Time (s)

EMR

Azure MR Adj.
US Cyberinfrastructure Context

• There are a rich set of facilities
  – Production TeraGrid facilities with distributed and shared memory
  – Experimental “Track 2D” Awards
    • FutureGrid: Distributed Systems experiments cf. Grid5000
    • Keeneland: Powerful GPU Cluster
    • Gordon: Large (distributed) Shared memory system with SSD aimed at data analysis/visualization
  – Open Science Grid aimed at High Throughput computing and strong campus bridging
TeraGrid • ~2 Petaflops; over 20 PetaBytes of storage (disk and tape), over 100 scientific data collections
FutureGrid and clouds for ADMI?

- Clouds could be used by ADMI in
  - Research
  - Education
  - Institutionally

- FutureGrid can be vehicle for
  - Supporting CS Research
  - Experimenting with cloud approaches for any of 3 modes

- We could set up a customized ongoing support activity on FutureGrid for ADMI

- We could offer a hands-on tutorial or summer school
  - See Jerome Mitchell proposal

- FutureGrid valuable to ADMI for HPC Grids and Clouds

http://futuregrid.org
FutureGrid key Concepts I

• FutureGrid is an **international testbed** modeled on Grid5000

• Supporting international **Computer Science** and **Computational Science** research in cloud, grid and parallel computing
  – **Industry** and Academia
  – **Prototype software development** and Education/Training
  – Mainly **computer science, bioinformatics, education**

• The FutureGrid testbed provides to its users:
  – A flexible development and testing platform for middleware and application users looking at **interoperability, functionality and performance**, exploring new computing paradigms
  – Each use of FutureGrid is an **experiment** that is **reproducible**
  – A rich **education and teaching** platform for advanced cyberinfrastructure classes
  – Support for users experimentation
FutureGrid key Concepts II

• Rather than loading images onto VM’s, FutureGrid supports Cloud, Grid and Parallel computing environments by dynamically provisioning software as needed onto “bare-metal” using Moab/xCAT
  – Image library for all the different environments you might like to explore ….

• Growth comes from users depositing novel images in library
• FutureGrid has ~4000 (will grow to ~5000) distributed cores with a dedicated network and a Spirent XGEM network fault and delay generator
• Apply now to use FutureGrid on website www.futuregrid.org
FutureGrid Partners

- **Indiana University** (Architecture, core software, Support)
  - Collaboration between research and infrastructure groups
- **Purdue University** (HTC Hardware)
- **San Diego Supercomputer Center** at University of California San Diego (INCA, Monitoring)
- **University of Chicago**/Argonne National Labs (Nimbus)
- **University of Florida** (ViNE, Education and Outreach)
- **University of Southern California Information Sciences** (Pegasus to manage experiments)
- **University of Tennessee Knoxville** (Benchmarking)
- **University of Texas at Austin**/Texas Advanced Computing Center (Portal)
- **University of Virginia** (OGF, Advisory Board and allocation)
- **Center for Information Services and GWT-TUD from Technische Universität Dresden.** (VAMPIR)
- **Red institutions** have FutureGrid hardware
# Compute Hardware

<table>
<thead>
<tr>
<th>System type</th>
<th># CPUs</th>
<th># Cores</th>
<th>TFLOPS</th>
<th>Total RAM (GB)</th>
<th>Secondary Storage (TB)</th>
<th>Site</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM iDataPlex</td>
<td>256</td>
<td>1024</td>
<td>11</td>
<td>3072</td>
<td>339*</td>
<td>IU</td>
<td>Operational</td>
</tr>
<tr>
<td>Dell PowerEdge</td>
<td>192</td>
<td>768</td>
<td>8</td>
<td>1152</td>
<td>30</td>
<td>TACC</td>
<td>Operational</td>
</tr>
<tr>
<td>IBM iDataPlex</td>
<td>168</td>
<td>672</td>
<td>7</td>
<td>2016</td>
<td>120</td>
<td>UC</td>
<td>Operational</td>
</tr>
<tr>
<td>IBM iDataPlex</td>
<td>168</td>
<td>672</td>
<td>7</td>
<td>2688</td>
<td>96</td>
<td>SDSC</td>
<td>Operational</td>
</tr>
<tr>
<td>Cray XT5m</td>
<td>168</td>
<td>672</td>
<td>6</td>
<td>1344</td>
<td>339*</td>
<td>IU</td>
<td>Operational</td>
</tr>
<tr>
<td>IBM iDataPlex On Order</td>
<td>64</td>
<td>256</td>
<td>2</td>
<td>768</td>
<td>On Order</td>
<td>UF</td>
<td>Operational</td>
</tr>
<tr>
<td>Large disk/memory system TBD</td>
<td>128</td>
<td>512</td>
<td>5</td>
<td>7680</td>
<td>768 on nodes</td>
<td>IU</td>
<td>New System TBD</td>
</tr>
<tr>
<td>High Throughput Cluster</td>
<td>192</td>
<td>384</td>
<td>4</td>
<td>192</td>
<td></td>
<td>PU</td>
<td>Not yet integrated</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1336</strong></td>
<td><strong>4960</strong></td>
<td><strong>50</strong></td>
<td><strong>18912</strong></td>
<td><strong>1353</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FutureGrid: a Grid/Cloud/HPC Testbed
Typical Performance Study

Linux, Linux on VM, Windows, Azure, Amazon Bioinformatics

![Graph showing performance comparison between different systems]

- Hadoop VM
- EC2
- Azure
- Hadoop Bare Metal
- DryadLINQ

Time to process 458 reads (1 files) per core (sec) vs. Number of Files
<table>
<thead>
<tr>
<th>Owner Name</th>
<th>Title</th>
<th>Institution</th>
<th>Date Submitted</th>
<th>Date Started</th>
<th>Results</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sumin Mohanan</td>
<td>Policy based distributed computing</td>
<td>University of Minnesota, Department of Computer Science and Engineering.</td>
<td>11.17.2010</td>
<td>11.17.2010</td>
<td>SAGA Grids</td>
<td>Clouds Policy</td>
</tr>
<tr>
<td>Yonggang Liu</td>
<td>Parallel File System</td>
<td>University of Florida ACIS, Advanced Computing and Information Systems Laboratory.</td>
<td>11.11.2010</td>
<td>5.1.2010</td>
<td>Parallel I/O</td>
<td>Parallel File Systems Nimbus QuS</td>
</tr>
<tr>
<td>Takion Wu</td>
<td>Survey of Open-Source Cloud Infrastructure using FutureGrid Testbed</td>
<td>Indiana University, Pervasive Technology Institute.</td>
<td>11.09.2010</td>
<td>5.1.2010</td>
<td>Clouds</td>
<td>OpenNebula Nimbus Eucalyptus</td>
</tr>
<tr>
<td>Jenett Tillotson</td>
<td>Comparing Moab metascheduling to Condor and MCP (Modified Critical Path)</td>
<td>Indiana University.</td>
<td>11.09.2010</td>
<td>11.09.2010</td>
<td>Moab Nimbus</td>
<td>Condor Metascheduling</td>
</tr>
<tr>
<td>Mariusz Mamonksi</td>
<td>Interoperability tests between OGF HPC-BasicProfile endpoints</td>
<td>Poznan Supercomputing and Networking Center.</td>
<td>11.09.2010</td>
<td>11.09.2010</td>
<td>OGS-SE</td>
<td>Interoperability Genesis II SMOA Unicore</td>
</tr>
<tr>
<td>Gideon Jue</td>
<td>Running workflows in the cloud with Pegasus</td>
<td>University of Southern California, Information Sciences Institute</td>
<td>11.05.2010</td>
<td>11.05.2010</td>
<td>Cloud Workflow</td>
<td>Pegasus</td>
</tr>
<tr>
<td>Anthony Chronopoulos</td>
<td>Integrating High Performance Computing in Research and Education for Simulation, Visualization and RealTime Prediction</td>
<td>University of Texas at San Antonio, Department of Computer Science</td>
<td>11.05.2010</td>
<td>11.05.2010</td>
<td>Education Research Clouds</td>
<td>MapReduce</td>
</tr>
<tr>
<td>Warren Smith</td>
<td>Publish/Subscribe Messaging as a Basis for TeraGrid Information Services</td>
<td>Texas Advanced Computing Center.</td>
<td>11.05.2010</td>
<td>11.05.2010</td>
<td>Nimbus TeraGrid Information Services</td>
<td>Publish/Subscribe</td>
</tr>
</tbody>
</table>
ViNe provided the necessary inter-cloud connectivity to deploy CloudBLAST across 5 Nimbus sites, with a mix of public and private subnets.
300+ Students learning about Twister & Hadoop MapReduce technologies, supported by FutureGrid.

July 26-30, 2010  NCSA Summer School Workshop
http://salsahpc.indiana.edu/tutorial
User Support

- Being upgraded now as we get into major use

- **Regular support:** there is a group forming FET or “FutureGrid Expert Team” – initially 13 PhD students and researchers from Indiana University
  - User requests project at [http://www.futuregrid.org/early-adopter-account-project-registration](http://www.futuregrid.org/early-adopter-account-project-registration)
  - Each user assigned a member of FET when project approved
  - Users given accounts when project approved
  - FET member and user interact to get going on FutureGrid
  - Could have identified ADMI support people

- **Advanced User Support:** limited special support available on request
  - Cummins engine simulation supported in this way