Open Cirrus Overview

Presentation at the 6th Open Cirrus Summit
Atlanta, GA, October 12, 2011

Dejan Milojicic
Senior Research Manager and Researcher
Managing Director, Open Cirrus
Hewlett-Packard Laboratories
dejan.milojicic@hp.com
Outline

• Old slide set on Open Cirrus with some analysis
• More recent research examples
• Lessons learned
• If we were to do it all over again today…?
• Summary
• What Next
Open Cirrus™ Cloud Computing Testbed

Shared: research, applications, infrastructure (11K cores), data sets

Global services: sign on, monitoring, store. Open source stack (prs, tashi, hadoop)

Sponsored by HP, Intel, and Yahoo! (with additional support from NSF)

- 15 sites currently, target of around 20 in the next two years.
Open Cirrus

• Objectives
  • Create an ecosystem for Cloud services modeling
  • Foster systems research around cloud computing
  • Expose research community to enterprise level requirements
  • Provide realistic traces of cloud workloads
  • Vendor-neutral open-source stacks and APIs for the cloud

• How are we unique
  • Support for systems research and applications research
  • Federation of heterogeneous datacenters
  • Interesting data sets
Process

• Central Management Office, oversees Open Cirrus

• Governance model
  – Research team
  – Technical Team
  – New site additions
  – Support (legal (export, privacy), IT, etc.)

• Each site
  – Runs its own research and technical teams,
  – Contributes individual technologies
  – Operates some of the global services

• E.g. HP Site supports: Portal and PRS
Requesting Access to OpenCirrus

The resources provided through the OpenCirrus Cloud Computing Testbed are a finite resource and are intended to be used for research purposes only.

Consequently, OpenCirrus computing resources are allocated to research projects that must be approved by one or more of the OpenCirrus Centers of Excellence. Project proposals are submitted by a Principal Investigator who is typically a university faculty member, senior staff member, or industrial researcher/technologist. Once a project is approved the Principal Investigator is able to identify additional team members who should be granted access as part of the project. This organization is similar to the arrangement for PlanetLab and nearly identical to the one used for Emulab.

Project Proposal Process

The process for proposing a project is relatively straightforward.

1. First, the Principal Investigator (PI) should select one of the OpenCirrus Centers of Excellence to serve as the Home Site for a project.

2. The PI should email a brief description of the project to the Project Coordinator at the Home Site. This description should include at least (1) the research goals of the projects, (2) a high-level description of the OpenCirrus resources that would be involved, and (3) the expected project start/end dates. See a sample here. The research coordinators for each site are listed below:

- HP Labs Site - Martha Lyons, martha.lyons@hp.com
- Intel Pittsburgh Research Site - Michael Kozach, email@intel.com
- Yahoo! Research - Thomas Kwan, email@yahoo.com
- UIUC -
- KIT -
- Singapore IDA -
Intel Research BigData Cluster

45 Mb/s T3 to Internet

Switch 48 Gb/s

Switch 48 Gb/s

Switch 48 Gb/s

Switch 48 Gb/s

1 Gb/s (x4)

1 Gb/s (x4)

1 Gb/s (x4)

1 Gb/s (x4)

1 Gb/s (x5 p2p)

Rack of 40 blade compute/storage nodes

Node: 8 Core2 cores (4x2), 8GB RAM, 0.3TB disk (2x150GB)

Rack of 40 blade compute/storage nodes

Node: 8 Core2 cores (4x2), 8GB RAM, 0.3TB disk (2x150GB)

Rack of 15 1u compute/storage nodes

Node: 8 Core2 cores (2x4), 8GB RAM, 2TB disk (2x1TB) x2

Rack of 15 2u compute/storage nodes

Node: 8 Core2 cores (2x4), 8GB RAM, 6TB disk (6x1TB) x3

Rack of 5 3u storage nodes

Node: 12TB disk (12x1TB)

---

10 nodes: 8 Core2 cores (4x2), 8GB RAM, 0.3TB disk (2x150GB)
20 nodes: 1 Xeon core, 6GB RAM, 366GB disk (36+300GB)
10 nodes: 4 Xeon cores (2x2), 4GB RAM, 150 GB disk (2x75GB)

<table>
<thead>
<tr>
<th>Nodes/cores</th>
<th>RAM (GB)</th>
<th>Storage (TB)</th>
<th>Spindles</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/140</td>
<td>240</td>
<td>11</td>
<td>80</td>
</tr>
<tr>
<td>40/320</td>
<td>320</td>
<td>12</td>
<td>80</td>
</tr>
<tr>
<td>30/240</td>
<td>240</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>45/360</td>
<td>360</td>
<td>270</td>
<td>270</td>
</tr>
<tr>
<td>[totals]</td>
<td></td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

[155/1050]  [1150]  [413]  [550]
UIUC cluster network topology

- Console switches: connect to 1 port/head node
- Links to external UIUC network connect either to both head nodes, or to both core switches
- Each storage node has 1x1Gb/s to each core switch
- Each compute node has 1x1Gb/s link to each core switch
- Core switches have 2x10Gb/s inter-switch links between them

Diagram:
- 1x1Gb/s uplink per compute node
- 1x1Gb/s uplink per rack for console LO100i
- 1x10Gb/s uplink per storage node
- 2x10Gb/s inter-switch link
# Open Cirrus Sites

<table>
<thead>
<tr>
<th>Site</th>
<th>#Cores</th>
<th>#Servers</th>
<th>Public partition</th>
<th>Memory Size</th>
<th>Storage Size</th>
<th>Spindles</th>
<th>Network</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP</td>
<td>1,024</td>
<td>256</td>
<td>178</td>
<td>3.3TB</td>
<td>632TB</td>
<td>1152</td>
<td>10G internal 1Gb/s x-rack</td>
<td>Hadoop, Cells, PRS, scheduling</td>
</tr>
<tr>
<td>IDA</td>
<td>2,400</td>
<td>300</td>
<td>100</td>
<td>4.8TB</td>
<td>43TB+16TB SAN</td>
<td>600</td>
<td>1Gb/s</td>
<td>Apps based on Hadoop, Pig</td>
</tr>
<tr>
<td>Intel</td>
<td>1060</td>
<td>155</td>
<td>145</td>
<td>1.16TB</td>
<td>353TB local 60TB attach</td>
<td>550</td>
<td>1Gb/s</td>
<td>Tashi, PRS, MPI, Hadoop</td>
</tr>
<tr>
<td>KIT</td>
<td>2048</td>
<td>256</td>
<td>128</td>
<td>10TB</td>
<td>1PB</td>
<td>192</td>
<td>1Gb/s</td>
<td>Apps with high throughput</td>
</tr>
<tr>
<td>UIUC</td>
<td>1024</td>
<td>128</td>
<td>64</td>
<td>2TB</td>
<td>~500TB</td>
<td>288</td>
<td>1Gb/s</td>
<td>Datasets, cloud infrastructure</td>
</tr>
<tr>
<td>Yahoo</td>
<td>3200</td>
<td>480</td>
<td>400</td>
<td>2.4TB</td>
<td>1.2PB</td>
<td>1600</td>
<td>1Gb/s</td>
<td>Hadoop on demand</td>
</tr>
</tbody>
</table>
Access Model

- At a minimum, sites must expose a ssh gateway
- Sites may also provide additional external connections
  - Some provision for web services is highly recommended
- Sites may also be divided into resource pools by service
  - Some services may require a front-end machine (e.g. hadoop)
Open Cirrus Software Stack

End-user services

Platform services

Applications

Hadoop

environment save/restore

data

Virtual Resource Sets

Cloud infrastructure services

IT infrastructure layer
(Zoni, Physical Resource Sets)

Cloud application services

Eucalyptus

Tashi

Tycoon

EC2

NFS storage service

S3 storage

cells-as-a-service

PRS service
How do users get access to Open Cirrus sites?

• Project PIs apply to each site separately

• Contact email addresses on the Open Cirrus portal
  – http://opencirrus.org

• Each Open Cirrus site decides which users and projects get access to its site

• A *global sign on* for all sites
  – Users are able to login to each OpenCirrus site for which they are authorized using the same login and password.
What kinds of research projects are Open Cirrus sites looking for?

• Open Cirrus™ is seeking research in the following areas (different centers will weight these differently)
  • Datacenter federation
  • Datacenter management
  • Web services
  • Data-intensive applications and systems
  • Hadoop map-reduce applications

• The following kinds of projects are of less interest
  • Traditional HPC application development
  • Production applications that just need lots of cycles
  • Closed source system development
Metrics of Success

- Community
  - Technology used
  - # Sites, Projects, (Vibrant) Users
  - Research Productivity (Shared Cost of Research), # papers published
  - Cross-collaboration (Portal traffic)
  - # New open source components
  - Global presence

- Technical
  - Utilization of Open Cirrus, TCO
  - Ease of use (e.g. provision 50% of OC nodes in < 30sec)
  - Federation transparency/adoptions
  - Reliability
# Open Cirrus v. Other Testbeds

<table>
<thead>
<tr>
<th>Type of research</th>
<th>Open Cirrus</th>
<th>IBM/Google</th>
<th>TeraGrid</th>
<th>PlanetLab</th>
<th>EmuLab</th>
<th>Open Cloud Consortium</th>
<th>Amazon EC2</th>
<th>LANL/NSF cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach</strong></td>
<td>Systems &amp; applications</td>
<td>Data-intensive applications research</td>
<td>Scientific applications</td>
<td>Systems and services</td>
<td>Systems</td>
<td>interoperab. across clouds using open APIs</td>
<td>Commer. use</td>
<td>Systems</td>
</tr>
<tr>
<td><strong>Participants</strong></td>
<td>HP, Intel, IDA, KIT, UIUC, Yahoo!</td>
<td>IBM, Google, Stanford, U.Washington, MIT</td>
<td>Many univ. &amp; orgs</td>
<td>Many univ &amp; organizations</td>
<td>University of Utah</td>
<td>4 centers –</td>
<td>Amazon</td>
<td>CMU, LANL, NSF</td>
</tr>
<tr>
<td><strong>Distribution</strong></td>
<td>15 sites</td>
<td>1 site</td>
<td>11 partners in US</td>
<td>&gt;700 nodes world-wide</td>
<td>&gt;300 nodes univ@Utah</td>
<td>480 cores, distributed in four locations</td>
<td>1000s of older, still useful nodes at 1 site</td>
<td></td>
</tr>
</tbody>
</table>
Open Cirrus Research Summary

**HP**
- Mercado
- Policy Aware Data Mgmt
- Wikipedia Mining & tagging
- SPARQL Query over Hadoop (UTD)
- N-tier App Benchmark (GaTech)

**Intel**
- Everyday Sensing and Perception
- SLIPstream/Sprout
- Parallel Machine Learning
- NeuroSys
- Computational Health
- FastBeat (w/France Telecom)

**Cloud application frameworks and services**

**Cloud infrastructure services**

**IT infrastructure layer**

**HP**
- Economic Cloud Stack
- Parallel Data Series
- OpenNet
- Exascale Data Center

**Intel**
- Tashi (with CMU, Yahoo)
- PRS (with HP)
OpenNet on OpenCirrus

• OpenNet
  – Programmable, open layer-2 network
  – Features for
    • Robust, adaptive routing over redundant layer-2 networks
    • VM machine migration *without* dropping connections
    • In-situ network monitoring
    • Quality-of-Service guarantees
  – Installed on OpenCirrus cluster at HP Fall 2009

• OpenNet on OpenCirrus
  – Full bisection bandwidth
  – Virtual machine migration
  – Platform for high energy efficiency in the Data Center
  – Based on SPAIN (HP Labs), PortLand (UC San Diego)
  – Joint project between HP Labs, UC San Diego (funded by HP Open Innovation Program)
OpenCirrus on GENI

• GENI: Global Environment for Network Innovations
  − Major National Science Foundation program to provide a national-scale experimental facility for computer science researchers
  − Currently entering Spiral Two prototyping phase

• OpenCirrus on GENI
  − Give access to GENI researchers to the OpenCirrus platform (PlanetLab Control Framework for OpenCirrus)
  − Give OpenCirrus users access to GENI resources

• Key technological challenges
  − Mutual authentication between PlanetLab Control and OpenCirrus
  − Exchange of authorization and access functions
  − Resource allocation

• Status
  − Joint proposal to GENI Project Office by HP Labs (Kevin Lai, Rick McGeer) and UC San Diego (Alex Snoeren, Amin Vahdat)
  − Accepted by GENI Project Office (GPO) for Spiral Two Funding
  − Part of GPO proposal to NSF for Spiral Two (decision early Sept)
SPARQL Query over Hadoop for Very Large RDF Datasets

- Provide a semantic web framework using Hadoop which scales for large RDF data sets.
  - Use the Lehigh University Benchmark (LUBM) data (provides 14 queries) to measure SPARQL queries implemented over Map/Reduce framework provided by Hadoop.
  - Goal: to find the best possible way to query the data (SPARQL) by Map/Reduce programming.
N-tier Application Benchmark & Evaluation over Open Cirrus

- Generate, deploy, and run N-tier application benchmarks (including non-stationary workloads)
  - Collect data on standard and custom N-tier application benchmarks such as RUBiS (e-commerce) and RUBBoS (bulletin board) over a wide range of settings and configurations (both hardware and software)
  - Collect, analyze, and evaluate performance data using statistical software tools.
  - Apply the experimental evaluation results to cloud management applications such as configuration planning and adaptive reconfiguration
## Cloud Sustainability Dashboard (CSD)

<table>
<thead>
<tr>
<th>Open Cirrus Site</th>
<th>Economical ($)</th>
<th>Ecological</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IT</td>
<td>cooling</td>
<td>ntwk</td>
</tr>
<tr>
<td>Site 1</td>
<td>$0.72</td>
<td>$0.35</td>
<td>$0.16</td>
</tr>
<tr>
<td>Site 2</td>
<td>$1.27</td>
<td>$0.59</td>
<td>$0.21</td>
</tr>
<tr>
<td>Site 3</td>
<td>$1.05</td>
<td>$0.47</td>
<td>$0.12</td>
</tr>
<tr>
<td>Site 4</td>
<td>$0.75</td>
<td>$0.35</td>
<td>$0.12</td>
</tr>
<tr>
<td>Site 5</td>
<td>$0.27</td>
<td>$0.13</td>
<td>$0.05</td>
</tr>
<tr>
<td>Site 6</td>
<td>$1.82</td>
<td>$0.77</td>
<td>$0.11</td>
</tr>
<tr>
<td>Site 7</td>
<td>$1.23</td>
<td>$0.54</td>
<td>$0.11</td>
</tr>
<tr>
<td>Site 8</td>
<td>$0.55</td>
<td>$0.26</td>
<td>$0.10</td>
</tr>
<tr>
<td>Site 9</td>
<td>$1.01</td>
<td>$0.44</td>
<td>$0.10</td>
</tr>
<tr>
<td>Bricks-and-Mortar (US)</td>
<td>$0.58</td>
<td>$0.70</td>
<td>$0.12</td>
</tr>
</tbody>
</table>
CSD Summary

• A systematic approach for representing and assessing sustainability of Clouds
  • Derived from a comprehensive model (economical, ecological, social)

• Automated, real-time Cloud Sustainability Dashboard
  • Express, assess and display run-time sustainability of Cloud & Cloud services
  • Preference-based customization

• Opportunities for integration with different enterprise tools
**Motivation**

- Towards Exascale – Impressive growth in cores #
- Increase in output data with increase of cores #
- GTC Fusion Modeling:
  - 50k processes, 200 MB/process, 3TB written every 30 minutes, 5 minutes for each step
- I/O performance – Huge perform. limiting factor
- Power limits local/distrib. memory DRAM buffers

**Research Questions**

- How to reduce compute to storage data move?
- How to reduce checkpoint cost, increase checkpoint frequency (faster restarts!)
- How to reduce I/O interference of Data written to storage then read for analytics/diagnosis/etc?
- Different output data -- different storage lifetime
- Should we move all data across IO layers? e.g. checkpoint data–max retention few hours

**Hypothesis**

**Benefits**

- Reduced I/O Block time using Active NVRAM
- Reducing checkpoint latencies by asynchronous mini checkpoints to local NVRAM instead of disk
- Reduction in memory usage and power reduction
- NVRAM – More complex in situ operations possible compared to SSD
A Case for Running HPC Applications on Cloud
Abhishek Gupta abhishek.gupta5@hp.com, gupta59@illinois.edu

**Motivation** (Why Clouds for HPC ?)
- No: startup/maintenance cost, create cluster time
- Time-to-solution - no job queue
- The true purpose of Cloud for eScience: “Ease, Availability, and Uniformity-,” Satoshi Matsuoka
- Network is a bottleneck for HPC apps on Cloud

**Research Questions**
- HPC apps Cloud-Supercomp perf.-cost tradeoffs
- HPC apps sensitivity (e.g. communication)
  - Latency vs. Bandwidth bound
  - Point to Point vs. Collective communication
  - Improving app perf. (strong/weak scaling) ?
- Is an SC optimized app good for Cloud
- Make apps cloud-friendly, eg infrequent load balance. In state-space search
- A combination of Cloud and supercomputer?

**Hypothesis**

$ vs Performance and Latency Sensitivity

**Benefits**
- Clouds offer best cost/performance for some HPC apps (latency sensitivity, perf. guarantees)
- Power savings should be possible
- Benefits from economy of scale
- Clouds can save job queue time for HPC apps (demand and supply)
- Clouds are cost – effective (substitute/addition) for some academic and commercial HPC users
Lessons Learned

- Entered 4th year, growing, 15 members, 6 summits, IEEE Computer paper
- 16 HPL projects across different Labs, similar in other institutions
- Of them, 4 directly OpenCirrus related, most remaining Cloud related

- **Sujata Banerjee**, “Our NSDI paper would not have happened without Open Cirrus”
- **Rick McGeer**, “We couldn’t have done demo at GEC-8 without Open Cirrus”
- **Prakash Reddy**, “BookPrep prepared 1M books on Open Cirrus on demand”
- **Jerry Liu** “Open Cirrus enabled us to build performance models for Article Clipper“

++ A lot of external excitement, people “get it,” want to join, use it, associate with it
+ Exhaustive use by individual sites, excellent setting for external collaboration
– Feels like a research project but not resourced like one
– – Getting hardware without development resources stifles progress (HP, UIUC, IDA)
If I were to start it all over again…

• Apply for external funding for common development resources (Or make sure internal funding is secured at each site)

• Align with the existing open source efforts (Open Stack, others)

• Start planning and aligning across multiple organizations
  • Business impact combined with IP leverage
  • Simplify IP management
  • Strong program management required

• Research-wise, do the same again
Summary

- Successful as a community
- Ahead of the time as a Cloud Stack
- Contributions internally, less so across organizations
- Created a lot of IP and enabled a lot of research

- Invitation to Open Cirrus Summits
  - 5th Summit in Moscow (June), 6th Summit in Atlanta (October)
    IEEE Co-sponsored, papers will appear in IEEE Digital Library
  - 7th Summit in Beijing, China, June 21-22, 2012
  - 8th Summit in Palo Alto, CA, collocated with ICAC, Fall 2012
What Next

- Question for this audience… … where do you want to take Open Cirrus next
- HP and Intel are committed to Open Cirrus
- New license agreement in final stages
- Should we adopt Open Stack?
- Replace portal with Facebook, twitter, Wordpress …
- How to encourage members to cross-collaborate?
Notes

• Karsten:
  • Open Stack natural
    − Works good with KVM, not with Xen
  • Funding:
    − Bigger joint funding, NSF, mid-size efforts, 2-3 partner team
    − Cloud Intern money
    − International efforts
    − Open Cirrus internship award + list the projects and what they have done!

• Calton:
  • Microsoft offered Azure hours, Open Cirrus did not (visibility)

• Roy:
  • Internet 2 advisory, need services to be used; connections to Europe and China; Scaled up OC activity; GENI; Commercial
Notes, cont.

• Mike:
  • Things that need to be added to infrastructure, e.g. TPM enabled systems, add a new rack every new month;
  • 6 months cadence. Raise an issue if you want to have something
  • Power instrumentation, TPMs, etc.

• Ling Liu:
  • Offer hours, but also offer to users to study

• Eduardo and Karsten
  • REUs have 70k-100k per year for 5 years
  • REUs Research Experience for Undergraduates
  • Open Cirrus REU program