iRODS in the Cloud: SciDAS and NIH Helium Commons

Claris Castillo
RENCI, UNC Chapel Hill
Not Scaling up Data Analysis is Not an Option

DatAPocaLypse Prediction (Genomics):

In 20 years, every CVS, subway, hospital, research lab, public health facility, police station, etc will have a DNA sequencer generating Exabytes of data in aggregate each week.

- Limited computational skills (What is a C library?)
- Poor use of advanced networks (We need more HDs to mail!)
- Limited access to computational resources (awareness, $$$)
- Unpredictable time to compute result (queue times, queue times, queue times, broken nodes, segfaults, OOM, data geography)
- Missing skillsets (I only know Perl)
- Data must be organized and good stuff deleted (Data policies)

*How many bioinformaticists are on the CVS payroll?*
*How many faculty recruitments failed because campus X research computing resources are stuck in 2015?*
*How many adverse drug reactions were not predicted because of limited/broken cyberinfrastructure?*

*Alex Feltus*
Heterogeneous and Complex CI Ecosystems

Community data sharing platforms

Compute infrastructure

Advance networks

Storage infrastructure

...
Commoditization of Cloud computing and the convergence of compute, storage, data and network technologies enables the ‘illusion’ of a single large computer consisting of widely distributed systems.
Breakdown: One Layer at A Time -- Data

iRODS team connected iRODS to a MariaDB Galera Cluster to provide a multi-master, distributed iRODS catalog over the WAN.

Breakdown: One Layer at A Time -- Compute

Apache Mesos: A layer of abstraction, to utilize an entire data center as a single large server.
Scientific applications will be available in the form of SciApps “virtual appliances” (NSF CC-ADAMANT, [works15])
SciDAS: Bringing it All Together Into One System

- Network aware placement
  - Optimize for data locality
- Capability aware resource aware placement
  - GPU able nodes
- Authentication and authorization infrastructure
  - CiLogon

Improving scientific productivity by the numbers

Comparison of KINC Workflow Runtime

<table>
<thead>
<tr>
<th>Platform</th>
<th>Runtime (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SciDAS</td>
<td>2500</td>
</tr>
<tr>
<td>Chameleon</td>
<td>4000</td>
</tr>
<tr>
<td>AWS</td>
<td>4000</td>
</tr>
<tr>
<td>Azure</td>
<td>4000</td>
</tr>
</tbody>
</table>

Platforms:
- SciDAS: (7 workers, 144 cores)
- Chameleon: (1 worker, 48 cores)
- AWS: (3 workers, 48 cores)
- Azure: (3 workers, 48 cores)

Provisioning Time Comparison among Computing Platforms

- SciDAS
- AWS
- Chameleon

<table>
<thead>
<tr>
<th>Platform</th>
<th>Lease creation</th>
<th>Resource acquisition</th>
<th>Configuration</th>
<th>Clean up</th>
</tr>
</thead>
<tbody>
<tr>
<td>SciDAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AWS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chameleon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time (seconds)
Virtualization system

Metadata to encode rich information

Rule engine programmed with rules to enact policies

Data Federation

High-level descriptor of applications

Intelligent decision (cloud aware)

Provision & deploy

Access/write data anywhere

Make results discoverable

Virtualization system

Metadata to encode rich information

Rule engine programmed with rules to enact policies

Data Federation

High-level descriptor of applications

Intelligent decision (cloud aware)

Provision & deploy

Access/write data anywhere

Make results discoverable
iRODS enables powerful data sharing models in the Commons

BYOD: Cloud storage can be added as storage resources

Data Federation (default): continuous virtual system while retaining control of each endpoint

Extended data collaboration (BYODS): Seamless integration with data hosted on external data services

BYOD: Cloud storage can be added as storage resources

Data Federation (default): continuous virtual system while retaining control of each endpoint

Extended data collaboration (BYODS): Seamless integration with data hosted on external data services
Thank you!

claris@renci.org