Infrastructure: Data Dissemination, Submission, On-Premise Private Cloud

Martial Michel
Jim Golden
Oleg Aulov
Open Science Big Data Analytic Technology R&D Model
(“Bring the experiment to the data”)
openstack

Project 1
- Write Only 1
- Compute Nodes 1
- Virtual Private Network
- User Management
- Front End
- Firewall

Project 2
- Shared Read Only Data Sets
- Compute Nodes 2
- Virtual Private Network

Your Applications
- APIs
- Compute
- Networking
- Storage

OpenStack Shared Services

Virtual Machines
- Containers
- Bare Metal
- Compute Nodes
- Private Network
- Data Access
- User separation
- Human Interface
- Network Separation

Standard Hardware
Discussion

- VM vs Containers vs Bare Metal?
- Benchmark model?
- Data access methodology?
- Compute paradigm?
- Data reusability if given ...

<table>
<thead>
<tr>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes: 16</td>
</tr>
<tr>
<td>CPU (cores): 32 (496)</td>
</tr>
<tr>
<td>HDD: 467.8 TB</td>
</tr>
<tr>
<td>RAM: 2.1 TB</td>
</tr>
</tbody>
</table>
How to push my code?

- We provide a template VM that runs with Virtual Box and has sample data (Ubuntu, CentOS)
  - Instructions on where to pull the data from (local to VM, will be “mounted” on the fly in production)
- Other option: Docker containers (easier use of accelerators)
- No Bare Metal option because of security concerns
- No interactivity/network access because of NIST regulations and data exfiltration concerns
  - Facilitate protection of Intellectual Property
- No Windows or Mac OS, (hypervisor compatibility and licensing are a limitation)
How to test and upload my VM?

- Upload using secure file transfer or upload it via sharing a link
- Testing can be done on AWS, Virtual Box, etc.
How to use specialized cloud software?

- Hadoop – user can create a template using OpenStack Sahara configurations to deploy it on EMS
- OpenStack App catalog
- Docker has GPU and CPU images for Tensorflow
Issues with Dataset Access

- A small dataset can be made available for download
- Big dataset for training is problematic (can be shared on S3), needs further discussion
Is there a development mode allowing to get back experimental data?