A Scientific Paper
Reproducible Environment
with Overlay Cloud Architecture

Shigetoshi Yokoyama*, Yoshinobu Masatani*, Tazro Ohta†, Osamu Ogasawara‡, Nobukazu Yoshioka*, Kento Aida*

*National Institute of Informatics,
†Database Center for Life Science,
‡National Institute of Genetics
Background
Reproducibility of Scientific Activities

Data and Infrastructure are ready to Reproduce

Data Centric Science ⇒ Reproducibility of IT environments in Scientific Activities

- Open Database for Genome Analysis
- Sharing Experimental Data
- Genome Analysis Pipelines Reproducibility
- Bare metal Cloud for Reproducibility

CloudLab

Genomics Virtual Lab

NII National Institute of Informatics
Issues in Reproducibility of Scientific Activities

Data don’t move and Processes don’t, either.

1) Huge Scientific Data (Distributed)
2) Huge Scientific Computation
3) Complex Scientific Computation Software

Huge Data
Huge Computation
Complex Software
Huge Data
Huge Data
Huge Data
Cloud Federation Approaches

Cloud Standardization is not the only way to go.

**Cloud Standard Approach**

Cloud Standards make cloud federations possible.

**On-demand Cloud Approach**

Separate hardware providers from cloud providers.

Cloud providers deploy clouds on-demand.

Cloud-A Provider

Cloud-B Provider

Cloud-C Provider

Hardware-1 Provider

Hardware-2 Provider

SINET

Separation

Cloud-A

Cloud-B

Cloud-C

Application Cluster-α

Application Cluster-β

Application Cluster-γ

Network with Glue software

Hardware-1

Hardware-1 Provider

Hardware-2

Hardware-2 Provider
On-demand Cloud: Academic Inter-Cloud
(academic community cloud hub)

- Univ. cloud
  - Cluster -A
  - Cluster -B
  - Cluster -C
  - Cluster -D

- Academic Inter-Cloud (compute)
  - Academic Inter-Cloud cluster -A
  - Academic Inter-Cloud cluster -B
  - Academic Inter-Cloud cluster -C
  - Academic Inter-Cloud cluster -D

- Academic Inter-Cloud (storage)
  - Inter-cloud Object Store Service

Mapping university’s cluster VLAN_ID and cluster ID in the cloud

Cooperate

As if a cluster is added to university cluster
As if multiple universities use one data center
As if an object store is same as local

Cooperate
On-demand cloud approach works but …

Need to Federate beyond Community Clouds.

• On-demand cloud approach works fine in our academic community cloud environment with a limitation.

Public clouds do not always support bare metals.
Overlay Cloud
## Container Revolutions in Two Industries

<table>
<thead>
<tr>
<th>Transport Industries</th>
<th>Contents</th>
<th>Containers</th>
<th>Container Management System</th>
<th>International Container Management System</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT Industries</td>
<td>Applications</td>
<td>Virtual Macines</td>
<td>Cloud</td>
<td>Inter-cloud</td>
</tr>
</tbody>
</table>

1950s -

2000s -
Separation of Concerns in Transport Industries

Separation of Concerns
Separation of Concerns in Transport Industries

New System

Separation of Concerns
Container Revolution in Transportation Industries

Containers wrap cargo

Separation of Concerns

Inside of container

Outside of container

Container Management

Infrastructure User

Infrastructure Provider

Separation of Concerns
Container Revolution in IT Industries

Containers wrap software

Inside Container
Outside Container
Overlay Cloud
Cloud User
Cloud Provider

Applications

Separation of Concern

L2Tunnels
Cluster
BM
VM

Container Engine, BM: Bare-metal Machine, VM: Virtual Machine
“Container” = Container

Overlay Cloud Approach is the way to go.

Overlay Cloud Approach is the way to go.

Applications

New System

Separation of Concerns

IT Industries
Virtual Cloud Provider
Virtual Cloud Provider (VCP)

VCP is the Middleware for Overlay Cloud.

- Virtual Cloud Provider
  - Functions (Deploy/Manage/Monitor)
  - Registry (Container images/Cluster Templates)

Virtual Cloud α
Virtual Cloud β
Virtual Cloud γ
VCP manages Two level Containers.
- Base container contains virtual cloud platform
- Application container contains application software
VCP Prototype

Quick Prototype with Open Source Products.

1. Real Cloud Resource Mgt.
3. Container Network Mgt.
4. Registry

Docker hub
https://hub.docker.com

Terraform Providers

Resources

- `bm_network`
  - name
  - vlan_id
  - subnet
  - network_addr

- `bm_instance`
  - name
  - image
  - flavor
  - networks

- `base_container`
  - host
  - image
  - run_commands
  - ipv4_address
  - ipv4_gw_address
  - volumes
  - network_tag

- `host`
  - name
  - type
  - interface

- `base_net`
  - name
  - subnet
  - gateway

- `overlay_net`
  - name
  - subnet
  - gateway
Case Study of Bioinformatics Application
Genome Sequencing

• application of the inter-cloud system to genomic sequence annotation and analysis workflows
Genome Sequencing On Overlay Cloud

Workflow Repositories
- Github Workflow Repo.
- DockerHub Container Repo.

Mesos/Aurora Cluster
- Mesos Master
- Mesos Slave #1
- Mesos Slave #2
- Mesos Slave #3

Containerized Galaxy Tool
- Aurora Scheduler
- Mesos Master
- zookeeper

File Server
- Raw Data copy from DDBJ
- Experiments Result copy from Riken
- Galaxy Job Working Dir.

Galaxy
- Workflow(a)
- Workflow(b)
- Workflow(c)

Base Container
- Mesos Slave #1
- Mesos Slave #2
- Mesos Slave #3

Docker Daemon Executor
- Mesos Master
- Aurora Scheduler
- Mesos Master
- zookeeper

Run Workflow
- Workflow(a)
- Workflow(b)
- Workflow(c)

Import
- Github Workflow Repo.
- DockerHub Container Repo.

Copy
- Raw Data from DDBJ
- Experiments Result from Riken
Workflow Example (RNA-Seq)

![Workflow diagram with exec. Time graph]
Conclusion
Conclusion

Overlay Cloud Approach Works.

• Proposed overlay cloud architecture.
• Developed middleware, VCP (Virtual Cloud Provider).
• Case study shows VCP enables the user to deploy computing environment that produces reproducible results.
History: The Road to Cloud Computing

Inter-Net (Networks)
Internet
TCP/IP, DNS

Inter-Web (Contents)
Web
hyperlink(http), Search Engine

Inter-Cloud (Services)
Cloud Computing
RESTful Web Service(http), Service Image Registry, Service Discovery, Service Monitoring

Loosely Connected Independently Governed Domains Work Together
Recent: New Wave is Coming

• The 3\textsuperscript{rd} Wave is coming …
  (1\textsuperscript{st} Wave) Public Cloud (AWS, …)
  (2\textsuperscript{nd} Wave) Open Cloud (OpenStack, …)
  (3\textsuperscript{rd} Wave) Containerization (Docker, …)

• They work together and make inter-cloud real.
Inter-Cloud Everywhere …

- Third Wave of Cloud Revolution is coming…
  
  (1\textsuperscript{st} Wave) Public Clouds (AWS, …)
  
  (2\textsuperscript{nd} Wave) Open Clouds (OpenStack, …)
  
  (3\textsuperscript{rd} Wave) Containers on Clouds (Docker, …)

- They co-exist and realize inter-cloud everywhere
Thank you