

Using Remote Clouds

FIFE Workshop (June 04, 2013)

- Anthony Tiradani

Cloud Definition

- When I say Cloud, I mean Infrastructure as a Service (IaaS) – Amazon EC2
- Other “Clouds”:
 - Platform as a Service (PaaS) – OpenShift, Azure
 - Service as a Service (SaaS) – Gmail

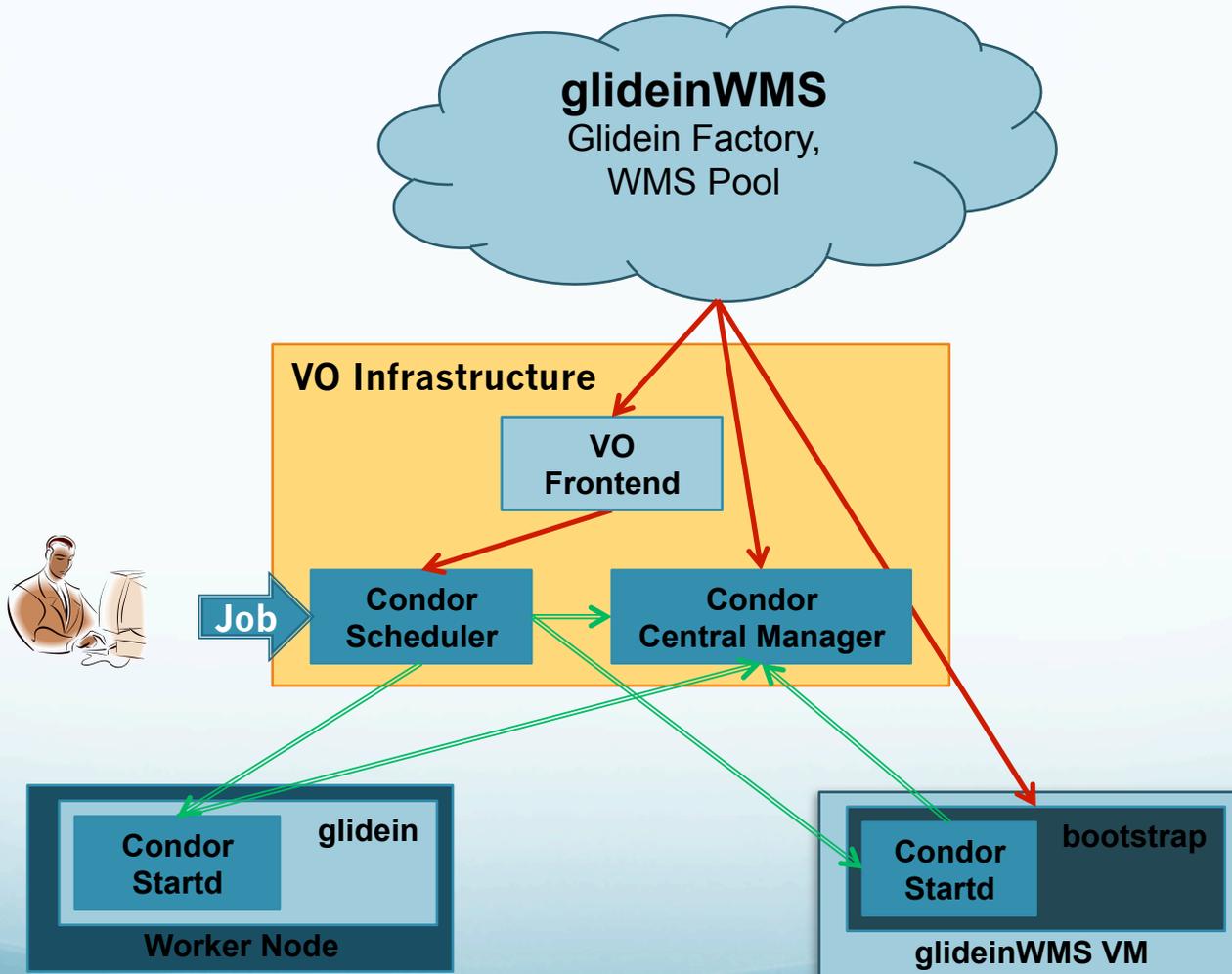
Grid vs. Cloud

- Grid
 - Site Gatekeeper
 - OSG – GT2/GT5
 - WLCG – CREAM
 - Site batch system
 - HTCondor
 - PBS
 - etc.
 - Aggregated Monitoring and Accounting infrastructure
- Cloud
 - Cloud “Controller”
 - EC2 Query API
 - Implementation specific protocols
 - No batch system
 - Must provide a VM image
 - However, can be customized to your specific requirements
 - Monitoring and Accounting specific to Cloud Site

Cloud Provisioning

- There are a few different systems in use by LHC experiments that can provision Cloud resources
- I am going to expose my bias here and discuss the glideinWMS approach

glideinWMS: Grid vs. Cloud



glideinWMS: Cloud

- glideinWMS aggregates resources to provide a virtual private HTCondor pool to the requesting VO
- glideinWMS seamlessly aggregates Grid and Cloud resources together in the same pool
- From the scientist's point of view there is no difference between a cloud slot and a grid slot
- Requires additional monitoring by VO administration to ensure run time constraints are enforced (budgetary constraints)

Grid vs. Cloud (part 2)

- Grid
 - Site owns worker node
 - Security patches
 - Installs software
 - “Guarantees” minimal software subset
 - Grid interfaces
 - Site may use different batch systems
- Cloud
 - VO or Factory operator owns VM images
 - Security patches
 - Installs software
 - Can be explicitly customized (at the expense of flexibility)
 - Cloud interfaces
 - No batch system

Cloud Cost Model

- Running on Open Science Grid is “free”
 - VO specific costs:
 - VO software maintenance
 - Submission infrastructure
 - End user support
- Running on the Cloud isn't free
 - Running on Commercial clouds (e.g. Amazon EC2) requires a credit card
 - Private Clouds have different “pay” models
 - VO specific costs
 - Same costs as OSG
 - VM image creation/maintenance
 - Cloud charges
- Commercial Cloud warnings
 - No explicit time limits like batch systems enforce
 - Cloud provider will happily let you continue to run until your credit card is maxed out

CMS Cost Measurements

- In 2010 CMS made some cost measurements to compare Amazon EC2 charges to the a Tier-2 University charges
 - Generated Monte Carlo events
 - Minimal data transfer
 - EC2 charges: \$0.872 per 1000 events
 - University charges: \$0.098 per 1000 events
- While the numbers are out of date, the conclusions remain valid today
 - Note: all of the usual caveats apply
 - Amazon EC2 is still more expensive than running a dedicated data center

What about EC2 Spot Instances?

- Amazon's description:
 - “Spot Instances allow you to name your own price for Amazon EC2 computing capacity. You simply bid on spare Amazon EC2 instances and run them whenever your bid exceeds the current Spot Price, which varies in real-time based on supply and demand.”
 - <http://aws.amazon.com/ec2/spot-instances/>
- Pricing is close to a dedicated data center
- Applications running on Spot Instances must be fault tolerant as the Instance may be terminated at any time

Bring Your Own Data

- Commercial clouds have plenty of proprietary tools for moving and storing data
 - Data movement is probably the largest line item on your bill, followed by data storage
 - No standard tools
- Must move data into and out of Cloud
 - Large data sets pose significant problems

VM Image



CMS Plans

- “Cloud Bursting”
 - Use cloud resources to handle peak demand periods
 - Costs unclear outside of Amazon EC2
- High Level Trigger farm (HLT)
 - During the long shutdown, the farm is unused
 - ~13,000 cores
 - OpenStack Essex
 - glideinWMS v3.0