Facility-based Clouds using OpenStack

John Hover, Xin Zhao

OSG All-Hands Meeting 2013

Indianapolis, Indiana

Outline

Rationale/Benefits

Limitations

Openstack Overview

- Components
- Networking
- BNL Openstack Instance
- General prospects
- New Openstack Features (v5 Folsom)

Discussion

Rationale

Expose Site Resources via Standard EC2 API

- Allows uniform access to Cloud-oriented workload systems.
- Gives users capability of sophisticated usage (not just worker nodes).
- Dynamic partitioning of facility resources (standard grid cluster, user purposes, testbeds, virtual Tier 3s).
  - Facility becomes customer of its own resources.
- Flexible facility management
  - Reboots, migration
  - Testing

Limitations

Using Cloud in OSG facility contexts will require:

- Some X509 authentication mechanism or gateway: Current platform implementations all require username/passwords.
  - x509 auth, a la Fermilab and OpenNebula
  - HTCondor-CE
- Accounting mechanism.
- Automated, supported install and configuration.

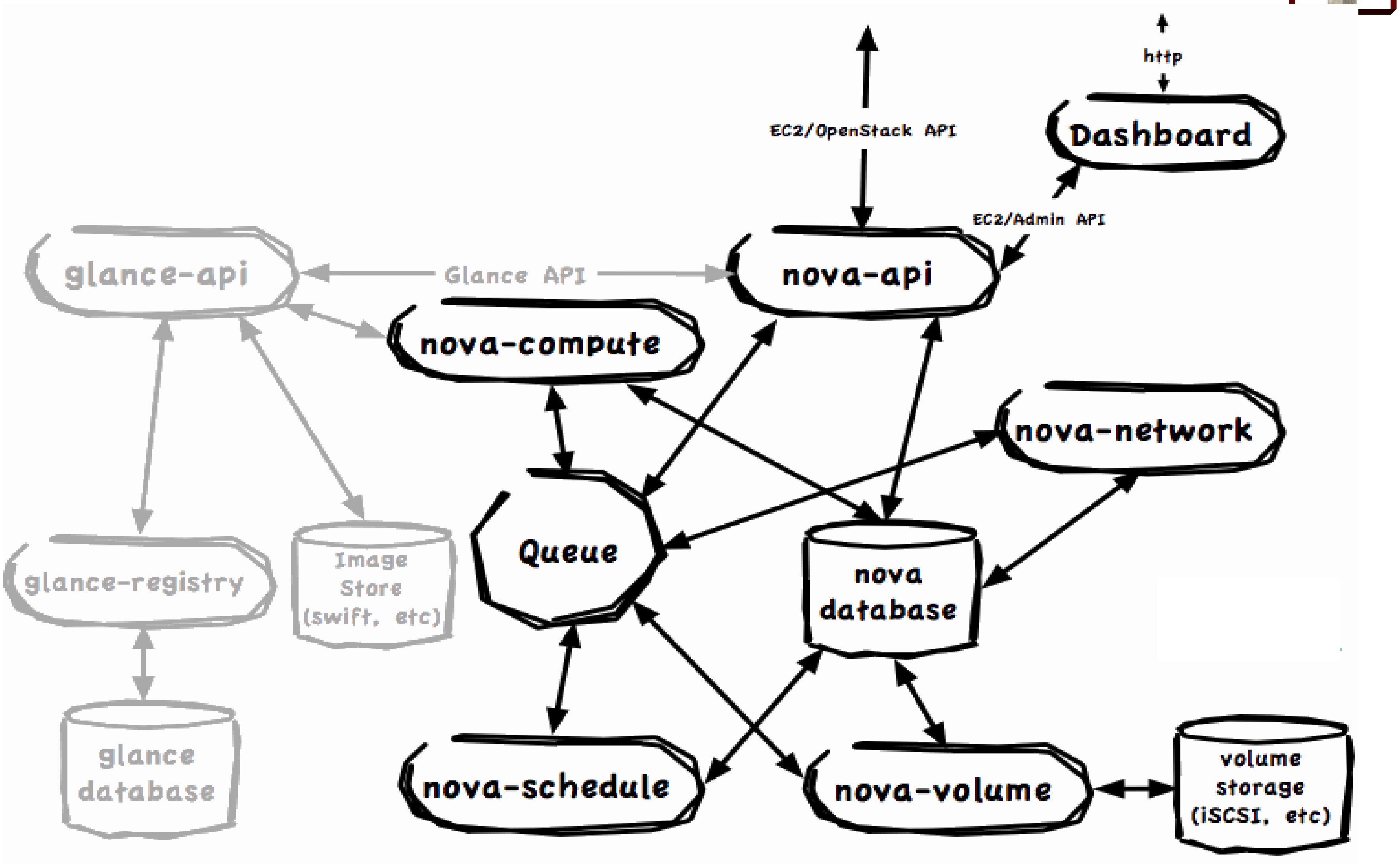
Intrusive: Fundamental change

- Does represent a new lowest-level resource management layer.
- But, once adopted all current management can still be used.

Networking and Security

- Public IPs require some DNS delegation, may also require additional addresses. (Limited public IPs at BNL).
- Some sites may have security issues with the Cloud model. Public IPs the issue at BNL.

Openstack v4 Components



Components

nova-api = EC2

  – External EC2 interface

nova-compute

  – Runs VMs

nova-schedule

  – Scheduler component

nova-volume

  – Internal/ephemeral storage management

swift = S3

  – Persistent storage management

glance

  – VM image management

Networking

nova-network: Network Manager Tasks

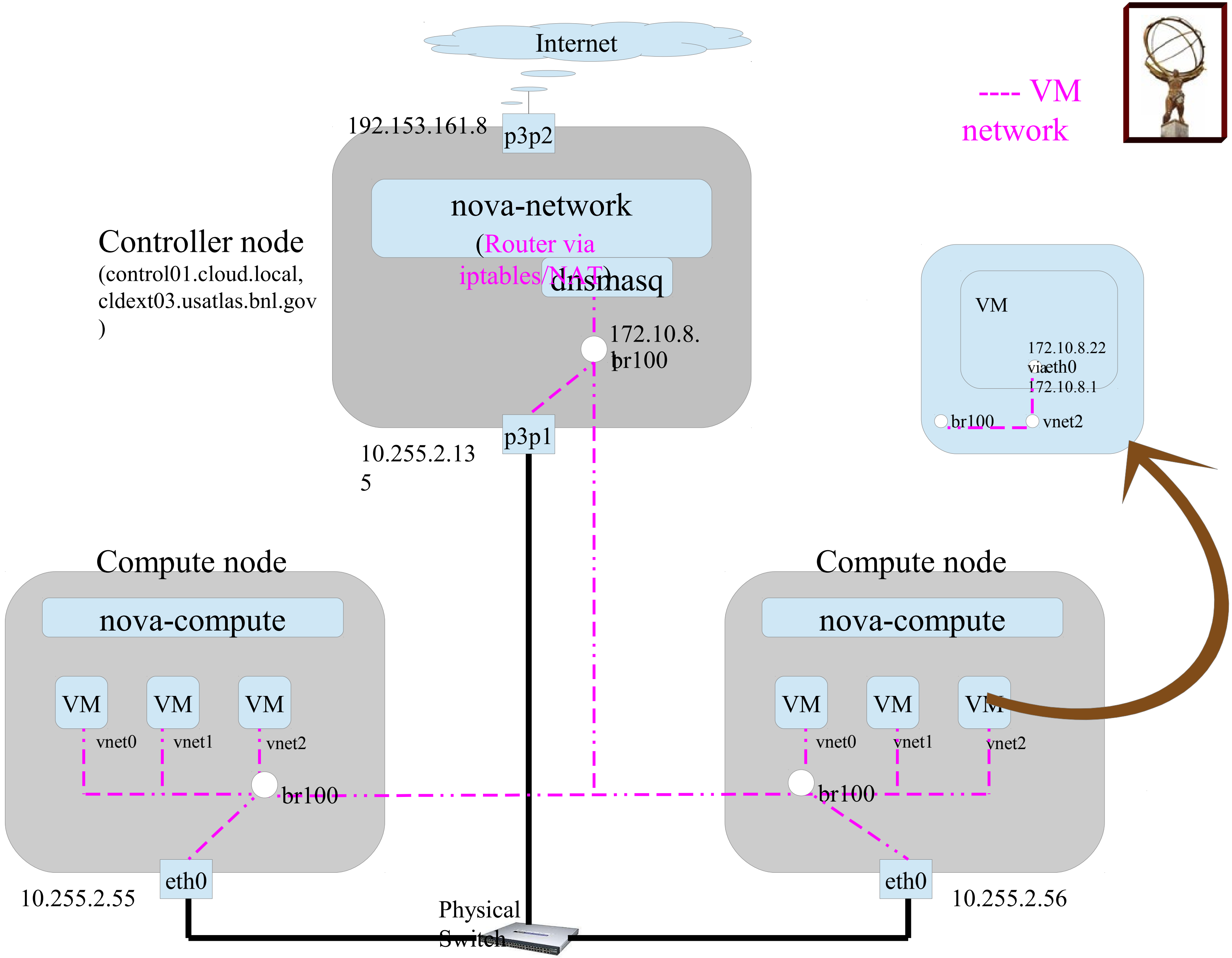
- IP allocation to instances
- Creating linux bridges (bridge-utils)
- Plugging instances into linux bridges
- Providing DHCP services for instances
- Configuring VLANs
- Providing external connectivity to instances

Handles by manipulating host iptables

Networking Types

Network manager determines layout:

- Flat Network Manager
  - One large IP pool. Shared by tenants.
  - Plugs instances into predefined bridge.
- Flat DHCP Network Manager
  - Adds DHCP server for VMs
- VLAN Network Manager
  - Manages multiple IP subnets, with tenant isolation.
  - Runs a dedicated bridge for each network
  - Switch requires support for 802.1Q tagged VLANs


Internet
192.153.161.8
p3p2
Controller node
(control01.cloud.local,
cldext03.usatlas.bnl.gov
)
nova-network
(Router via
iptables/NAT)
dnsmasq
172.10.8.
1
br100
p3p1
10.255.2.13
5
---- VM
network
VM
172.10.8.22
via eth0
172.10.8.1
br100
vnet2
Compute node
nova-compute
VM
VM
VM
vnet0
vnet1
vnet2
br100
eth0
10.255.2.55
Compute node
nova-compute
VM
VM
VM
vnet0
vnet1
vnet2
br100
eth0
10.255.2.56
Physical
Switch

Concerns about Testbed Network

Networking single point of failure

- – nova-network is down, no internet connectivity
- – Fix?: multi-host networking mode
    - Run nova-network on every worker node host
    - Each worker node has its own gateway, dnsmasq, NAT for its own VMs
    - Requires outbound connectivity on all worker nodes

Single big IP pool

- – No isolation between tenants (security concern,...)
- – Fix?: VLAN Manager

System puppet iptables vs. Openstack iptables

Administration

Nova CLI admin commands, e.g.

- nova add-fixed-ip
- nova add-floating-ip
- nova delete <server>
- nova flavor-create
- nova image-list
- nova boot
- nova x509-create-cert

Glance service-specific CLI

- glance index
- glance add < image.raw
- glance delete

BNL Openstack Instance

Openstack 4.0 (Essex)

- 1 Controller, 100 execute hosts (~300 2GB VMs), fairly recent hardware (3 years), KVM virtualization w/ hardware support.
- Shared cluster nova-network on controller (10Gb throughput shared)
- Provides EC2 (nova), S3 (swift), and image service (Glance).
- Essex adds keystone identity/auth service, Dashboard.
- Programmatically deployed, with configs publically available.
- Fully automated compute-node installation/setup (Puppet)
  - http://svn.usatlas.bnl.gov/svn/atlas-puppet/
- Enables 'tenants'; partitions VMs into separate authentication groups, such that users cannot terminate (or see) each other's VMs. Three projects currently.

BNL Openstack 2

Use FlatDHCPManager

- Nova-network runs on controller (control01)

Physical network

- Controller has dual NICs, one internal, one out- facing the internet
- All worker nodes have single NIC, which is on the internal network (10.255.2.0/24)

VM network

- VM network IP pool (172.10.8.0/21)
- Outbound internet connection from instances goes through controller node, where the VM network gateway is located.
- Inbound connectivity to instances can be achieved by using “floating IPs” (6 from 192.153.X.X subnet)

Prospects/ Future

Ubuntu Adoption

- Began packaging and distributing Openstack in 2011

CERN switching to OpenStack

- Tim Bell, Infrastructure Manager at CERN IT, on Openstack council
- ATLAS using Openstack at P1. CMS?

BNL sent 2 people to Openstack Summit 2012,

- CERN attended.
- Conference attended by 1200, up from 200 a couple years ago.

Rapid adoption, ambitious roadmap, and aggressive release cycle bode well for progress.

- Open source rivals?

Release Schedule

OpenStack adopts a 6 months release cycle, starting from the Cactus release

Release name	Release date
Grizzly	?
Folsom	October 2012
Essex	April 2012
Diablo	October 2011
Cactus	April 2011
Bexar	March 2011
Austin	October 2010

Openstack v5 (Folsom) Quantum

A New Networking Platform

- Network API
  - Flexible API for service providers *or their tenants* to manage OpenStack network topologies
  - Evolves independently of the Nova compute API
- Plugin Architecture
  - Separates the description of network connectivity from its implementation
  - Linux bridges, VLAN, iptables, OpenFlow, ...
- A Platform for integrating Advanced solutions
  - If interested in customized network technology (eg Infiniband), one can extend the API and provide their own plugin.

Quantum Architecture

- Quantum-server
  - API: for tenants to define their network
  - On controller or standalone host
- Agents: responsible for directly managing the network
  - Plugin agent
    - On every worker nodes and network devices to perform local network configuration
  - DHCP agent
    - Provide DHCP service to tenant networks
  - L3 agent
    - L3/NAT forwarding for external network access for VMs on tenant networks

Currently Available Plugins in Quantum

Open vSwitch

Linux Bridge

Cisco (UCS Blade + Necus)

Nicira NVP

Ryu OpenFlow controller

NEC ProgrammableFlow Controller

Questions/Discussion

How many sites running Openstack

- BNL, Nebraska, Chicago?

Largest deployment?

- BNL=300 VMs. Larger?
- ATLAS P1 still 1 compute node prototype.

Interest in OSG-mediated deployment?