Apache CloudStack

-- Architecture Overview

@CloudStack中国

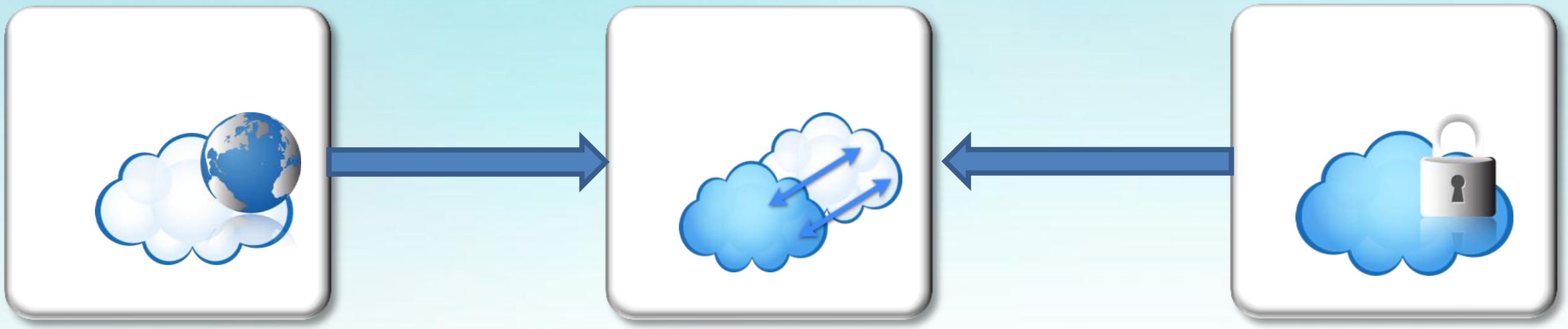
Outline

- CloudStack Memo
- CloudStack Briefly Go
- CloudStack Network Deep Dive
- CloudStack API
- Q & A

Memo

Date	Event
2008	VMOps Founded who developed CloudStack
May, 2010	VMOps Changed name to Cloud.com CloudStack2.0 released under several licenses
Jul, 2011	Citrix acquired Cloud.com Developed CloudStack 3.0
Apr, 2012	CloudStack was donated to ASF, align with ASL2.0
Nov, 2012	Apache CloudStack 4.0 released The first released version by community
Dec, 2012	CCC12 in Las Vegas

Clouds



Public Cloud

- Multi-tenant
- Shared/Mixed Resource
- Elastic Scaling
- Pay as you go
- Public network

Hybrid Cloud

- Hosted Enterprise
- Dedicated Resource
- Secure
- SLA
- 3rd party Operation

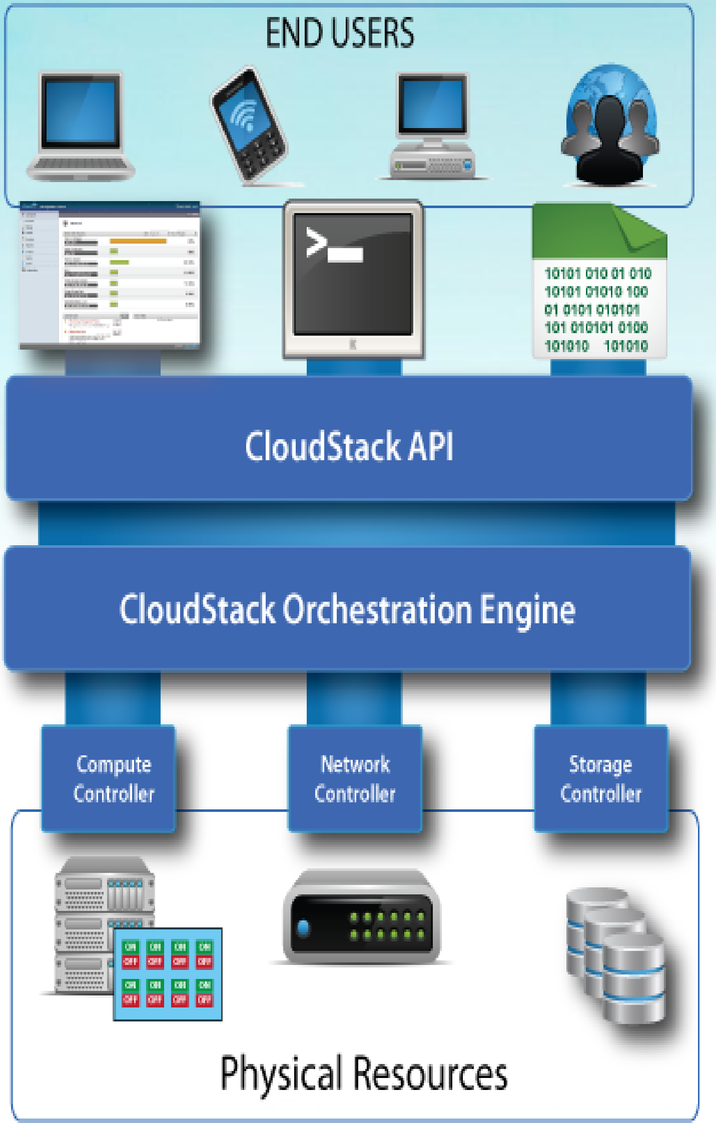
Private Cloud

- Dedicated Resource
- Secure
- Total Control
- Internal Network
- Managed by IT dept. internally

What is CloudStack?

cloudstack

- IaaS Orchestration platform
- Multi-tenant
- Scalable
- Open Source
- Resource Control
  - Cloud (IaaS)
    - Public (Multi-tenant)
    - Private (On-premise internally)
    - Hybrid (Host Enterprise)
  - Resource
    - Virtual & Physical
    - Compute
    - Storage
    - Network



Picture from Geralyn Miller

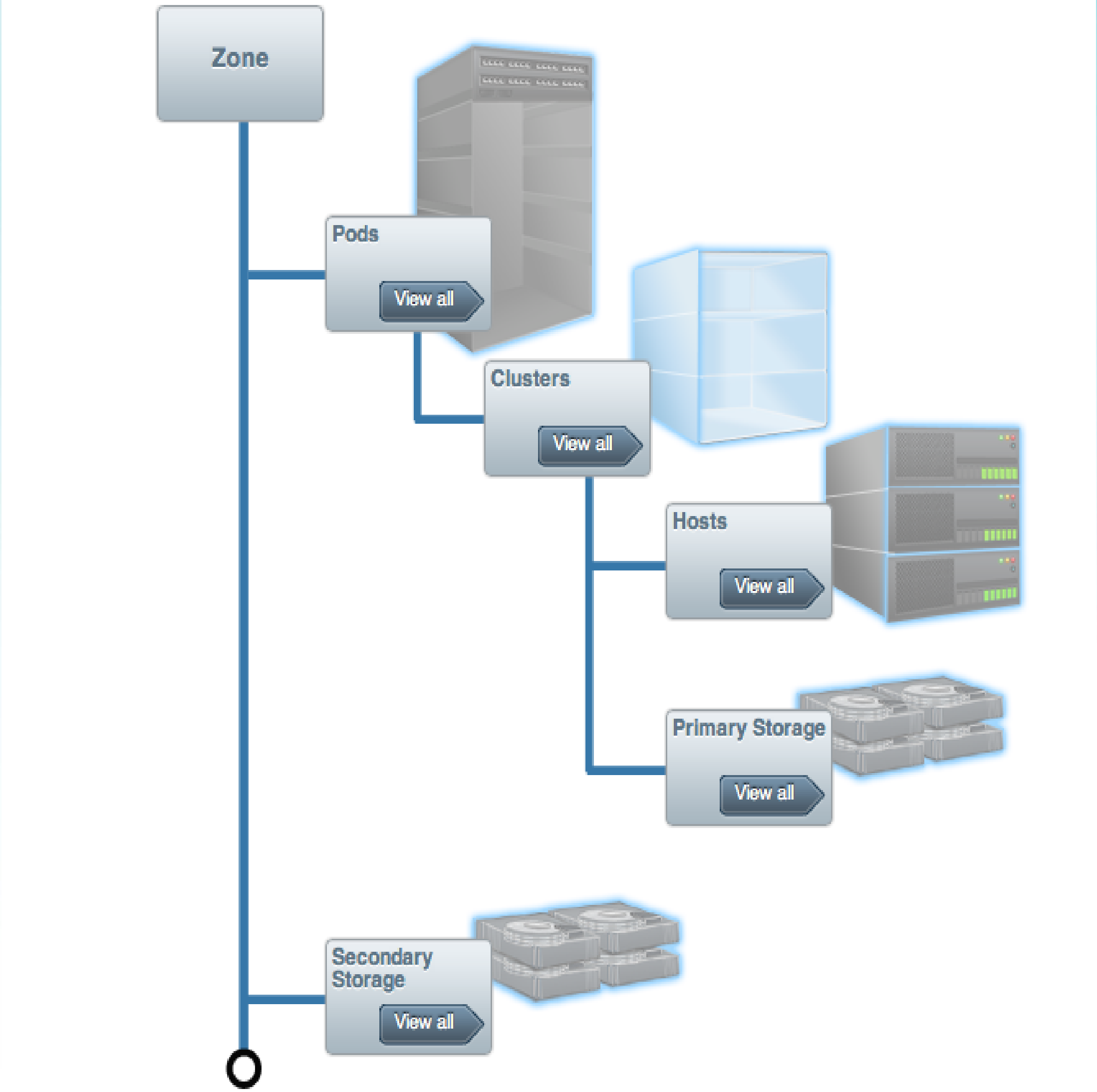
What Can CloudStack Really Do?

- Multi-tenants separation
- Allocate compute resources as pre-configured
- Services auto provisioning to end user in a controlled manner (VLAN allocation, firewall rules, load balancer deployment, VM creation, etc)
- VM HA
- Compute resource scale out
- Resource limitation modification (*dynamically*)
- Usage data measurable

CloudStack Briefly Go

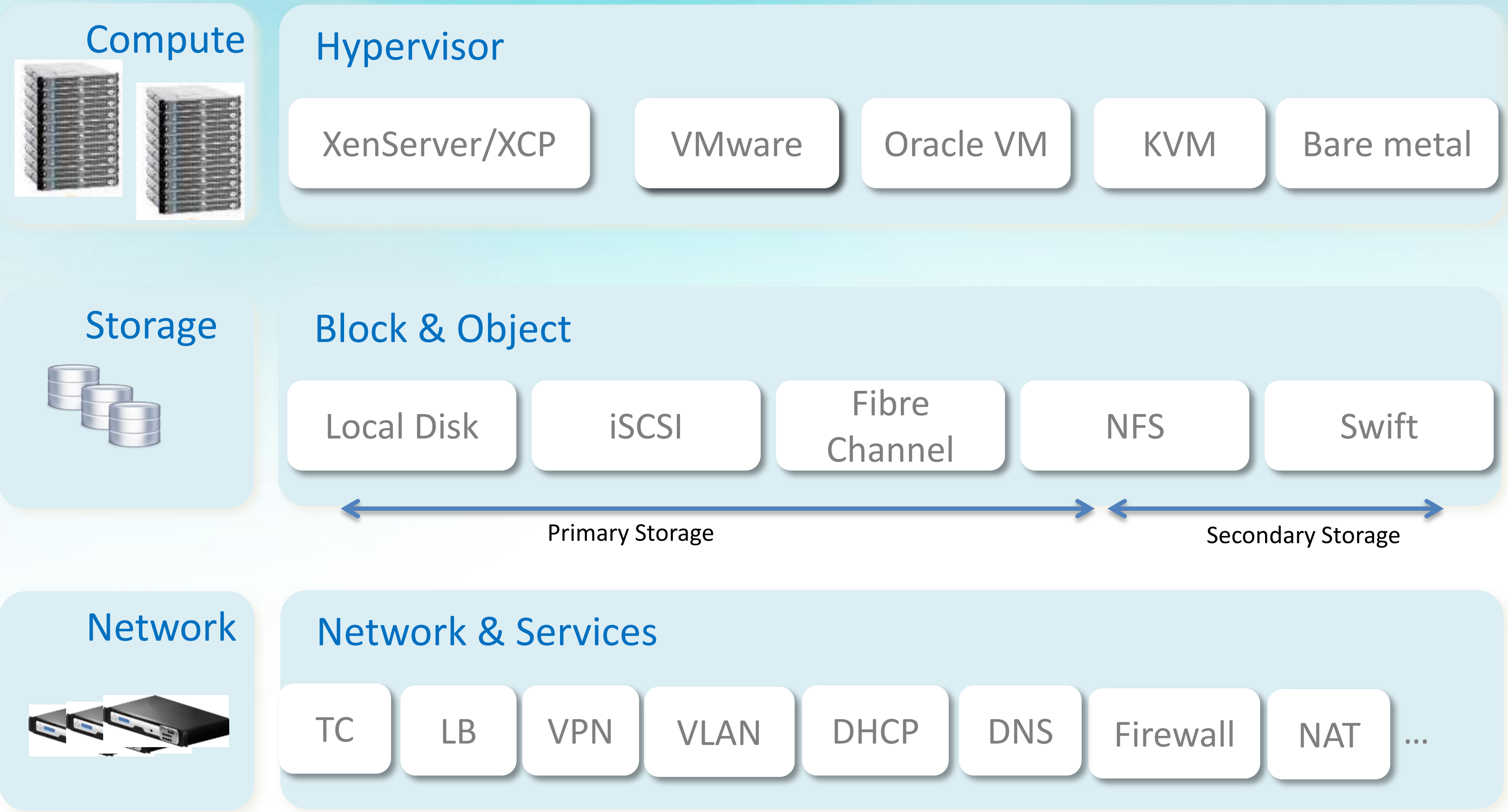
- Components – High Level
- Flexibility
- Scalability
- Reliability
- Hypervisor
- Storage
- System VM
- Networks

Components – High Level



Zone: Availability zone, aka Regions. Could be worldwide.
Pod: Rack in a data center
Cluster: Group of machines with a common type of Hypervisor
Host: A Single server
Primary Storage: Shared storage across a cluster
Secondary Storage: Shared storage in a single Zone

Flexibility



Compute

Hypervisor

XenServer/XCP

VMware

Oracle VM

KVM

Bare metal

Storage

Block & Object

Local Disk

iSCSI

Fibre Channel

NFS

Swift

Primary Storage

Secondary Storage

Network

Network & Services

TC

LB

VPN

VLAN

DHCP

DNS

Firewall

NAT

...

Scalability

- One management server can handle 10k resources
- Scales out horizontally without StatusCollector
- Real production deployment of tens of thousands of resources
- Software simulators up to 30k physical resources with 30k VMs managed by 4 management servers
- Improvement in progress

Reliability

Anything at any time in any places is unreliable

Active methods:

- Live Migration
- Maintenance

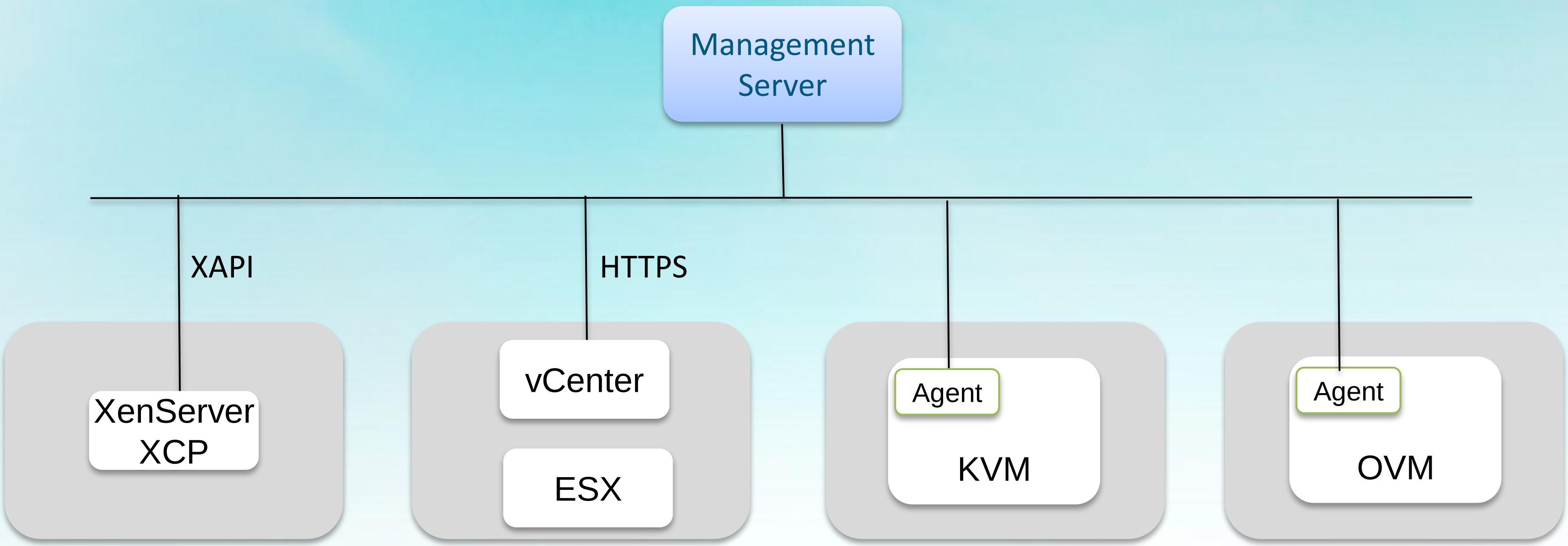
Passive Solution

- Service Offerings for VM HA
- Dedicated Host for HA enabled VM

HA in CloudStack

- HA is good for virtualization industry.
- CloudStack HA is workable and useful but not fantastic
  - Investigating needs time
  - Fencing needs time
  - May failed at last
- CloudStack will watch for HA-enabled VMs to ensure that they are up, and that the hypervisor it’s on is up – and will restart on another hypervisor if it goes down.
- More robust solution is redundant router

Hypervisor



- XS 5.6, 5.6FP1, 5.6 SP2, 6.0, XCP1.0, XCP1,1, XCP1.5Beta
- Incremental Snapshots
- VHD
- NFS, iSCSI, FC & Local disk
- Storage over-provisioning: NFS

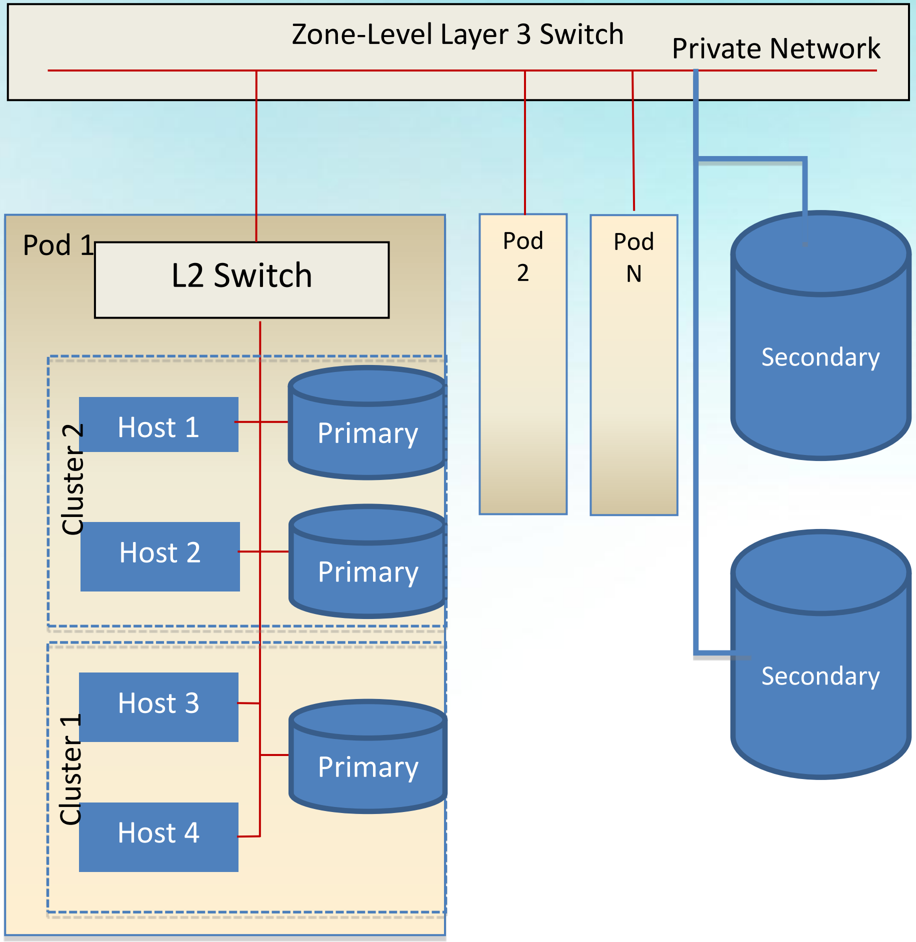
- ESX 4.1, 5.0
- Full Snapshots
- VMDK
- NFS, iSCSI, FC & Local disk
- Storage over-provisioning: NFS, iSCSI

- RHEL 6.0, 6.1, 6.2 , Ubuntu 12.04
- Full Snapshots (not live)
- QCOW2
- NFS, iSCSI & FC
- Storage over-provisioning: NFS

- OVM 2.2
- No Snapshots
- RAW
- NFS & iSCSi
- No storage over-provisioning

cloudstack

Storage



- Primary Storage
  - Block device to the VM
  - IOPs intensive
  - Accessible from host or cluster wide
  - Supports storage tier
- Secondary Storage
  - Write Once Read Many Times Pattern
  - For templates, ISO, and snapshot archiving
  - High capacity
- CloudStack manages the storage between the two to achieve maximum benefit and resiliency

Networks -- Terminology

- **Public**: Internet or public access. If CloudStack is completely in private environment (inside a company network), the address assign to vrouter and all traffic pass through via NAT, this only appeared in advenced network
- **Management**: Where the hypervisors and management server lives in and communicate with each other
- **Guest**: The network and VLAN created for guest VM within a domain/project/account.
- **Storage**: Optional network dedicated to secondary storage. Will use management network by default if not specified.
- **Link-local**: The special virtual interface exists between the host and the inside VMs. All system VM has this interface for secure interaction. Refer to RFC3927 for more.

Networking

- Network modules broken down by:
  - Method of isolation (VLAN, Security Groups)
  - Physical hardware or virtual
- CloudStack manages network services:
  - DHCP
  - VLAN allocation
  - Firewall
  - NAT/Port forwarding
  - Routing
  - VPN
  - LB
- CloudStack manages physical devices:
  - F5-Big IP
  - NetScaler
  - Juniper SRX

Security Groups

- Traditional layer 2 isolation via VLAN
- VLAN scaling problems
  - Standard has a hard limit of 4096 VLANs
  - High cost if keep up to 4096 VLANs
  - People are not will to be limited what they can do
- Use Layer 3 isolation like Amazon (Security Groups)
  - Trust layer 2 networks, which only hypervisor attached
  - Filtering/isolation occurs at bridge device
    - iptables/ebtables
  - Deny by default

System VM

- Common Features
  - Stateless, can be destroyed/recreated
  - HA
  - Interact with mgmt server via mgmt network
  - Usually 3 nics (link-local, mgmt and public)
- CPVM (Console Proxy VM)
  - Access VM via Web Console uses Ajax https
  - Scale out
  - Zone level
- SSVM (Secondary Storage VM)
  - For template/snapshot/iso upload and download
  - For VM deployment
  - Scale out
  - Zone level
- VRouter/DomR (Virtual Router/Domain Router)
  - NaaS module provide rich network function
  - Redundancy via VRRP
  - Account level

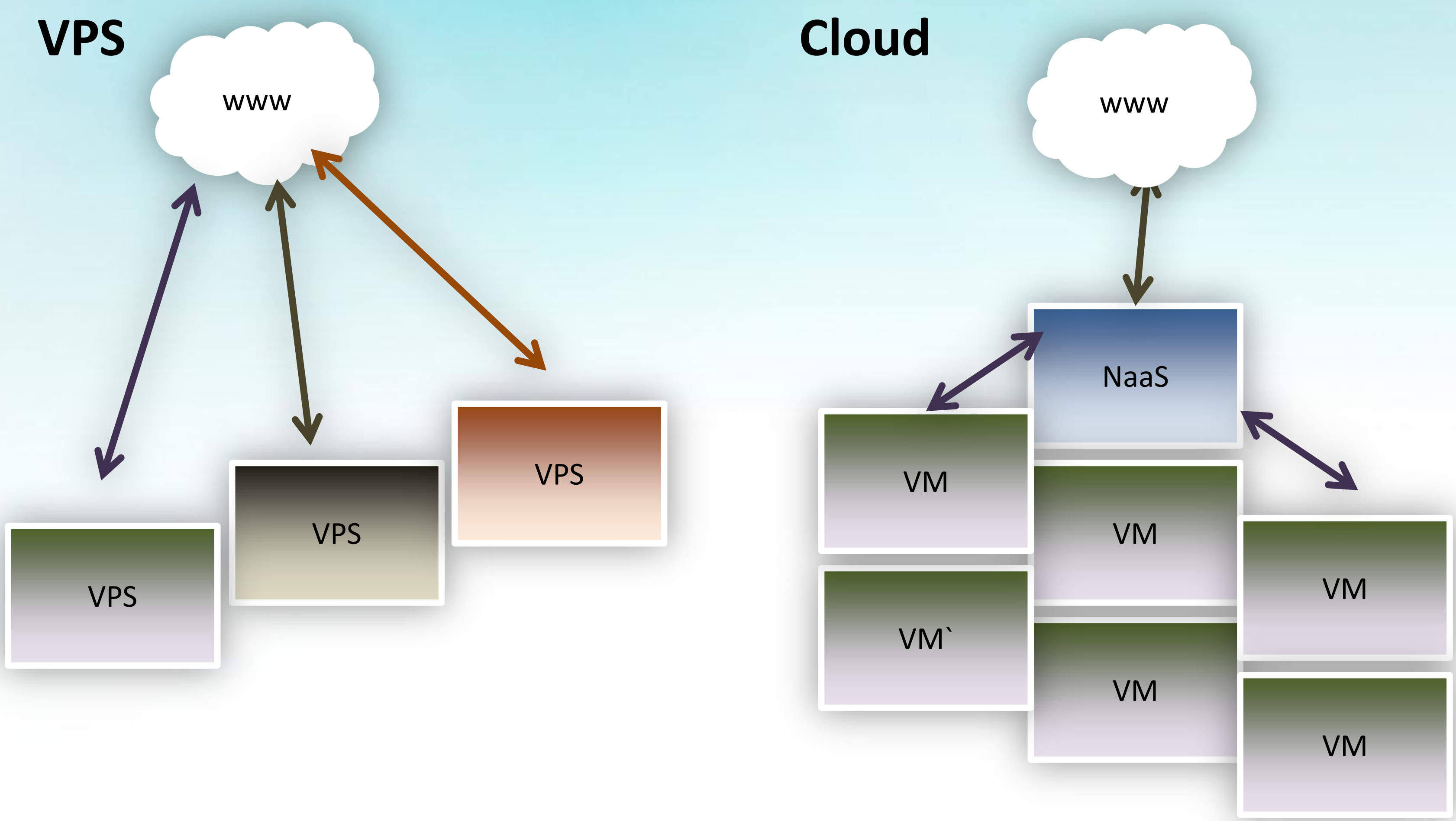
System VM Spec.

- Debian 6.0 ("Squeeze"), 2.6.32 kernel 32bit
- Essential software only:
  - haproxy, iptables, ipsec, jre ,etc.
  - printing, ftp, telnet, X, kudzu, dns, sendmail are not installed.
- SSHd service to access via hypervisor
  - SSHd only listens on the private/link-local interface.
  - SSH port changed to 3922.
  - SSH logins only using unique keys which generated at install time
- pvops kernel for performance optimization:
  - with Xen paravirt drivers
  - KVM virtio drivers
  - VMware tools for optimum performance on all hypervisors.
- Same vm works on XS, KVM, VMWare

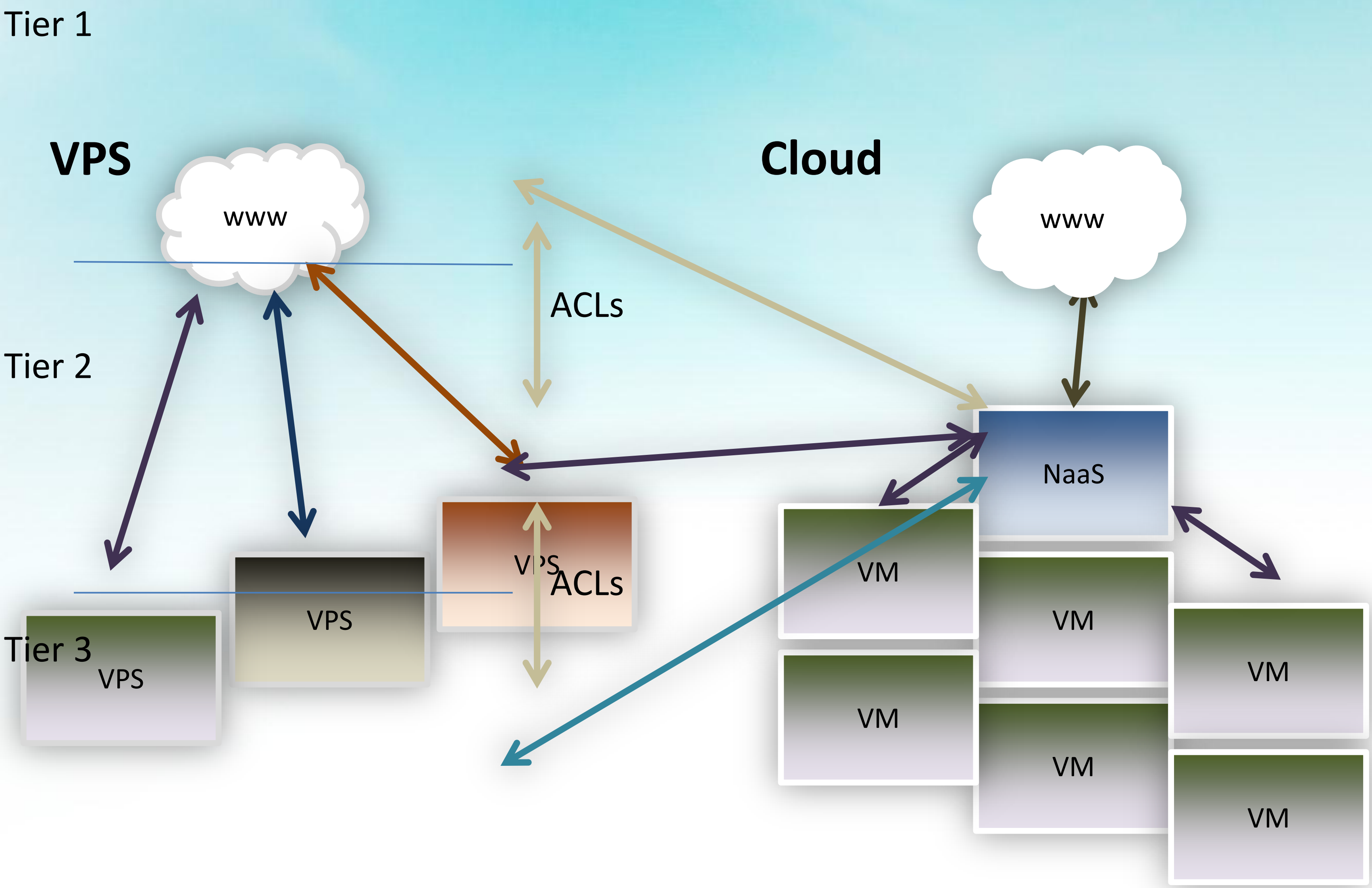
CloudStack Network Deep Dive

- Use Case
- Basic Networking
- Advanced Networking
- System VM Networking

Use Case



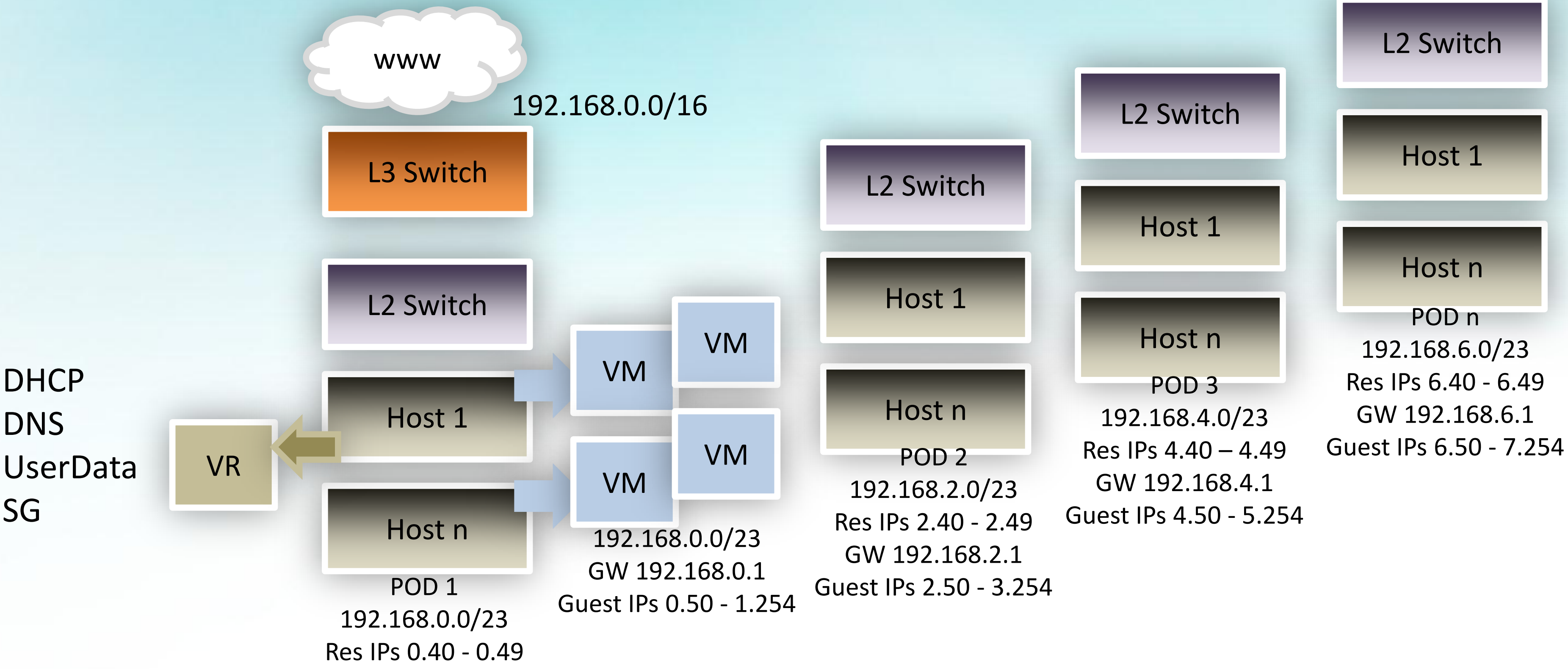
Use Case


Tier 1
VPS
Cloud
www
www
ACLs
Tier 2
NaaS
VM
VPS
ACLs
VM
VM
Tier 3
VPS
VPS
VM
VM
VM
VM

Basic Networking

- AWS Style L3 isolation – Massive Scale
- Simple Flat Network
- Only 1 Physical NIC per Host
- Guest Instances and Hosts Share IP Schema*
- Each POD has a unique CIDR
- Guest Isolation via Security Groups
- NetScaler Integration - Elastic IPs and Elastic Load Balancing

Basic Networking Models



www

192.168.0.0/16

L3 Switch

DHCP
DNS
UserData
SG

VR

L2 Switch

Host 1

Host n

VM

VM

VM

VM

POD 1
192.168.0.0/23
Res IPs 0.40 - 0.49

192.168.0.0/23
GW 192.168.0.1
Guest IPs 0.50 - 1.254

L2 Switch

Host 1

Host n

POD 2
192.168.2.0/23
Res IPs 2.40 - 2.49
GW 192.168.2.1
Guest IPs 2.50 - 3.254

L2 Switch

Host 1

Host n

POD 3
192.168.4.0/23
Res IPs 4.40 – 4.49
GW 192.168.4.1
Guest IPs 4.50 - 5.254

L2 Switch

Host 1

Host n

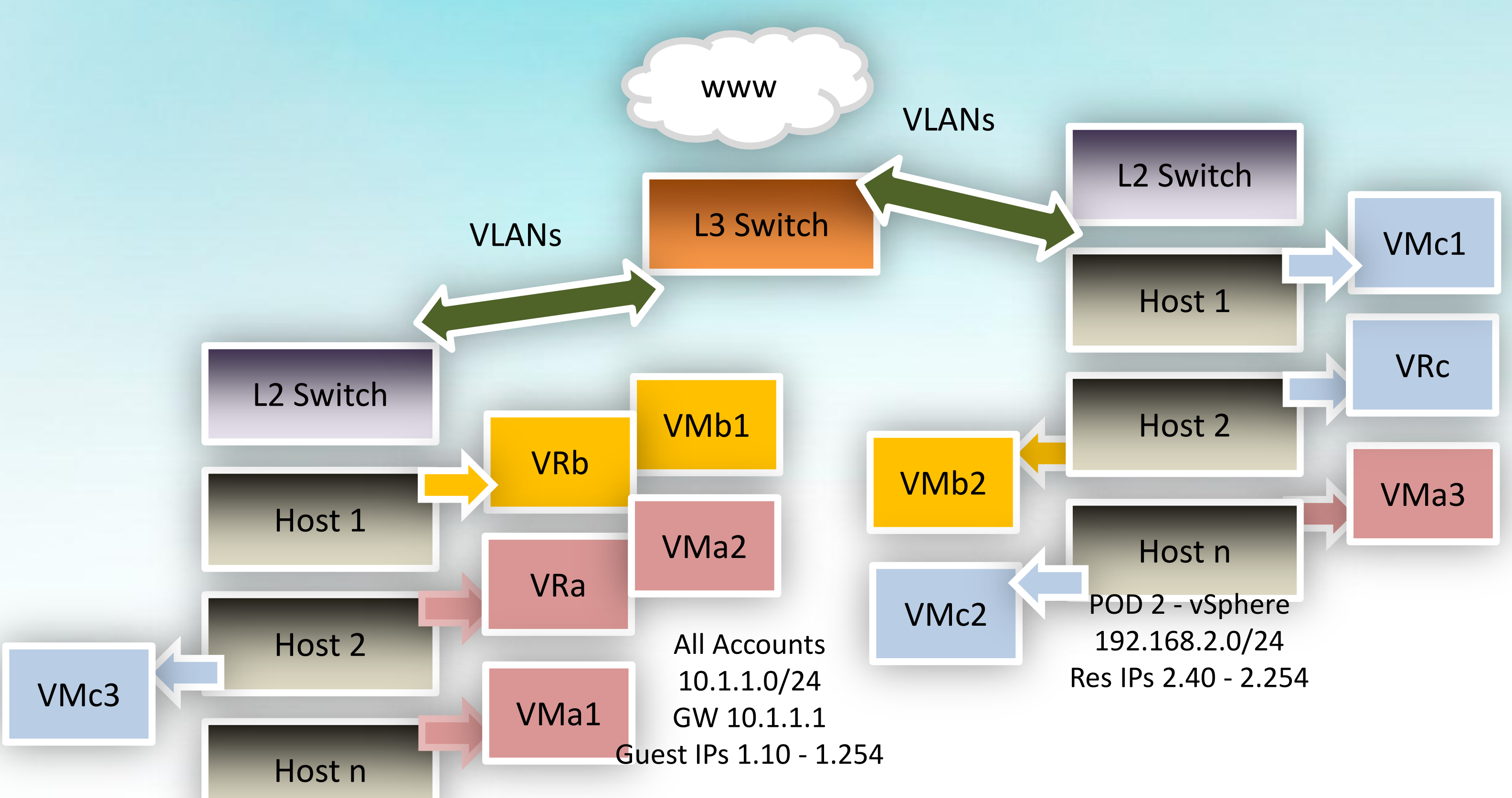
POD n
192.168.6.0/23
Res IPs 6.40 - 6.49
GW 192.168.6.1
Guest IPs 6.50 - 7.254

Advanced Networking

- Guest Networks isolated by VLANs
- Shared and Isolated Guest Networks
- Traffic spread across multiple Physical NICs which can also be Bonded
- Virtual Router for each Account / Network providing:
  - DNS & DHCP
  - Firewall
  - Client VPN
  - Load Balancing
  - Source / Static NAT
  - Port Forwarding

cloudstack

Advanced Networking Models

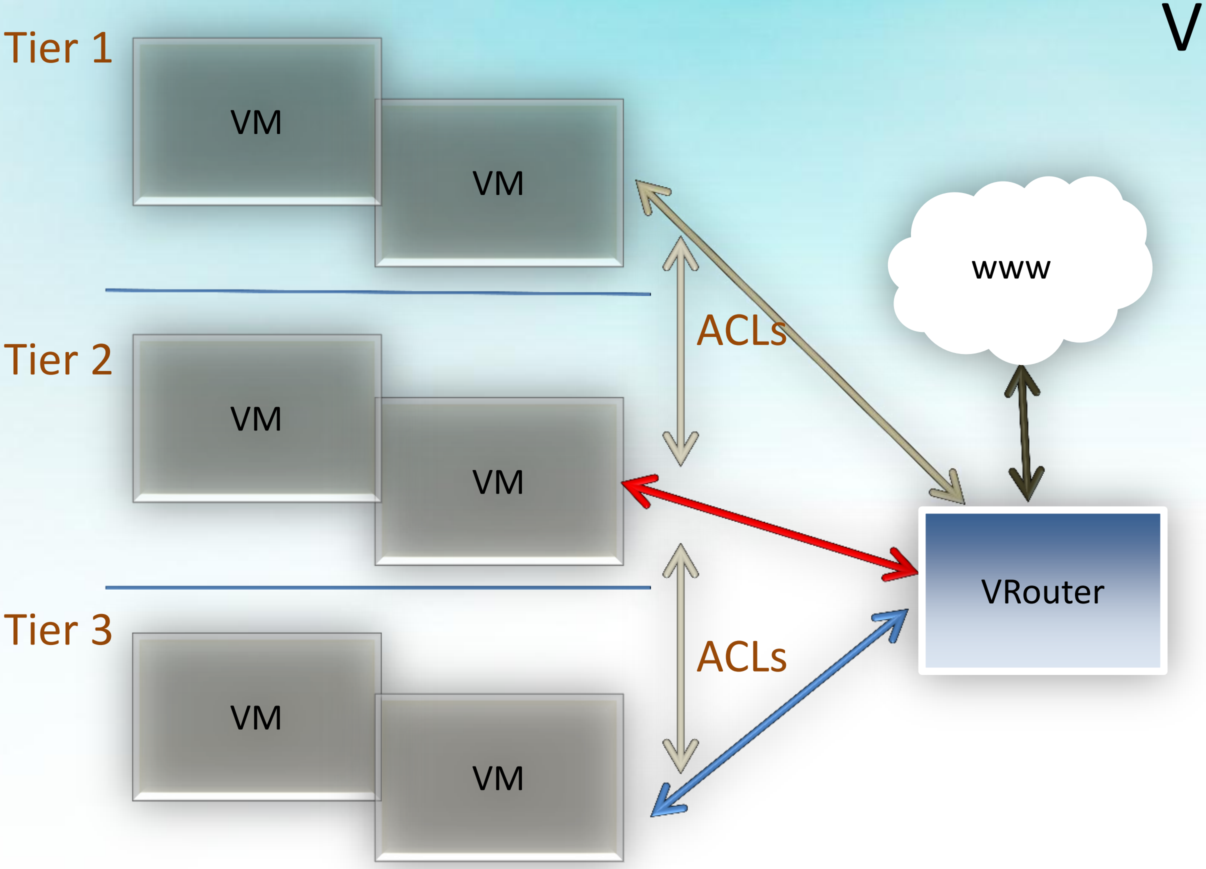


POD 1 - XenServer
192.168.0.0/24
Res IPs 0.40 - 0.49

All Accounts
10.1.1.0/24
GW 10.1.1.1
Guest IPs 1.10 - 1.254

POD 2 - vSphere
192.168.2.0/24
Res IPs 2.40 - 2.254

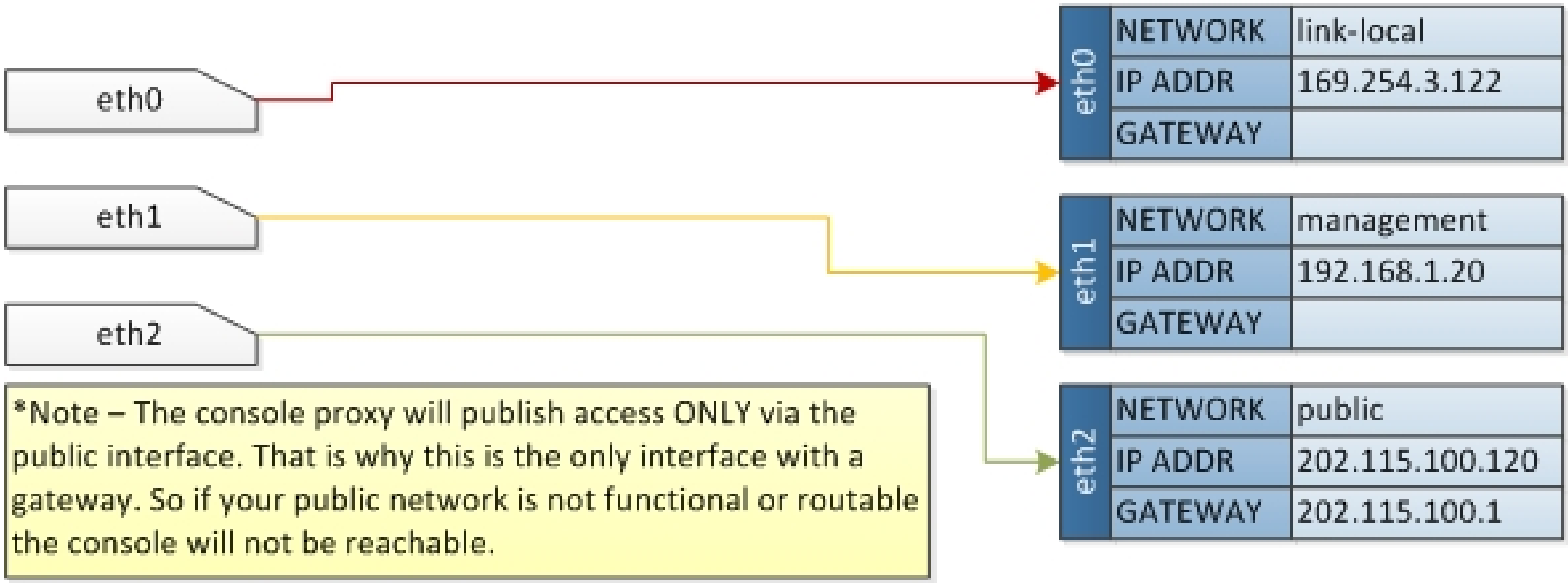
Advanced Networking VPC



Virtual Private Clouds

- Private multi-tiered Virtual Network
- Inter VLAN Routing
- Site-2-Site VPN

CPVM Networking



eth0		
	NETWORK	link-local
	IP ADDR	169.254.3.122
	GATEWAY	

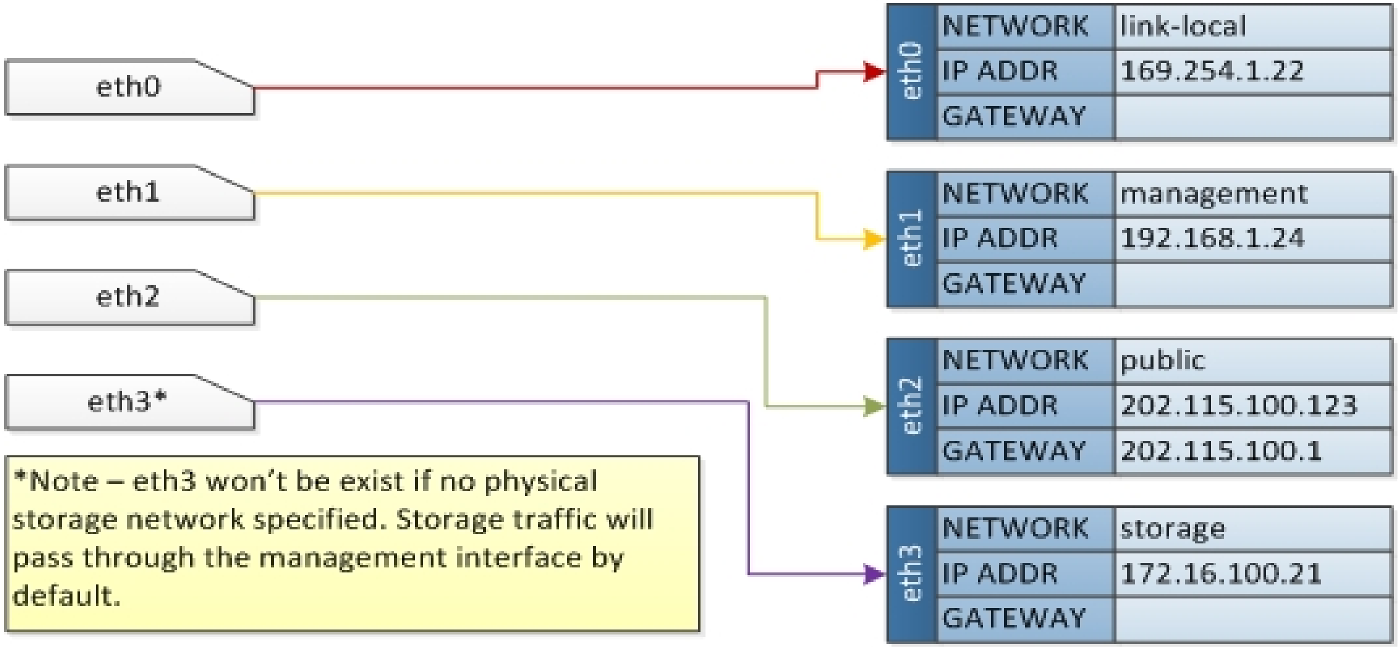
eth1		
	NETWORK	management
	IP ADDR	192.168.1.20
	GATEWAY	

eth2		
	NETWORK	public
	IP ADDR	202.115.100.120
	GATEWAY	202.115.100.1

*Note – The console proxy will publish access ONLY via the public interface. That is why this is the only interface with a gateway. So if your public network is not functional or routable the console will not be reachable.

Public Subnet: 202.115.100.0/24
Management Subnet: 192.168.1.0/24
Storage Subnet: 172.16.100.0/24
Default Guest Subnet: 10.10.100.0/24

SSVM Networking



eth0	NETWORK	link-local
	IP ADDR	169.254.1.22
	GATEWAY	

eth1	NETWORK	management
	IP ADDR	192.168.1.24
	GATEWAY	

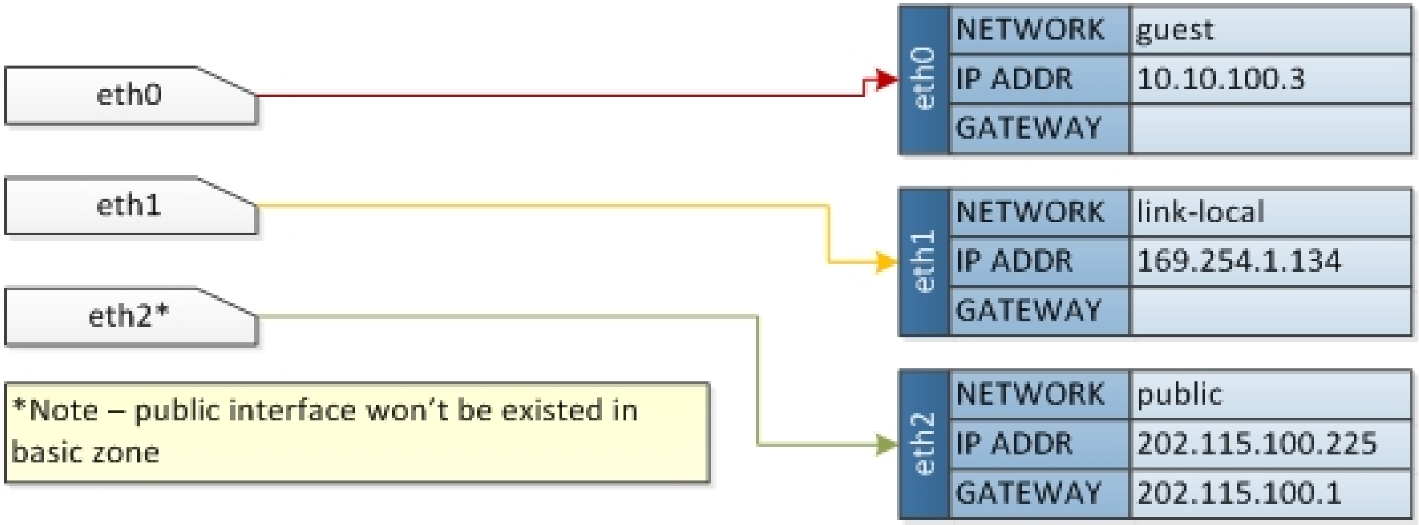
eth2	NETWORK	public
	IP ADDR	202.115.100.123
	GATEWAY	202.115.100.1

eth3	NETWORK	storage
	IP ADDR	172.16.100.21
	GATEWAY	

*Note – eth3 won't be exist if no physical storage network specified. Storage traffic will pass through the management interface by default.

Public Subnet: 202.115.100.0/24
Management Subnet: 192.168.1.0/24
Storage Subnet: 172.16.100.0/24
Default Guest Subnet: 10.10.100.0/24

Vrouter/DOMR Networking



eth0	NETWORK	guest
	IP ADDR	10.10.100.3
	GATEWAY	

eth1	NETWORK	link-local
	IP ADDR	169.254.1.134
	GATEWAY	

eth2	NETWORK	public
	IP ADDR	202.115.100.225
	GATEWAY	202.115.100.1

*Note – public interface won't be existed in basic zone

Public Subnet: 202.115.100.0/24
Management Subnet: 192.168.1.0/24
Storage Subnet: 172.16.100.0/24
Default Guest Subnet: 10.10.100.0/24

API

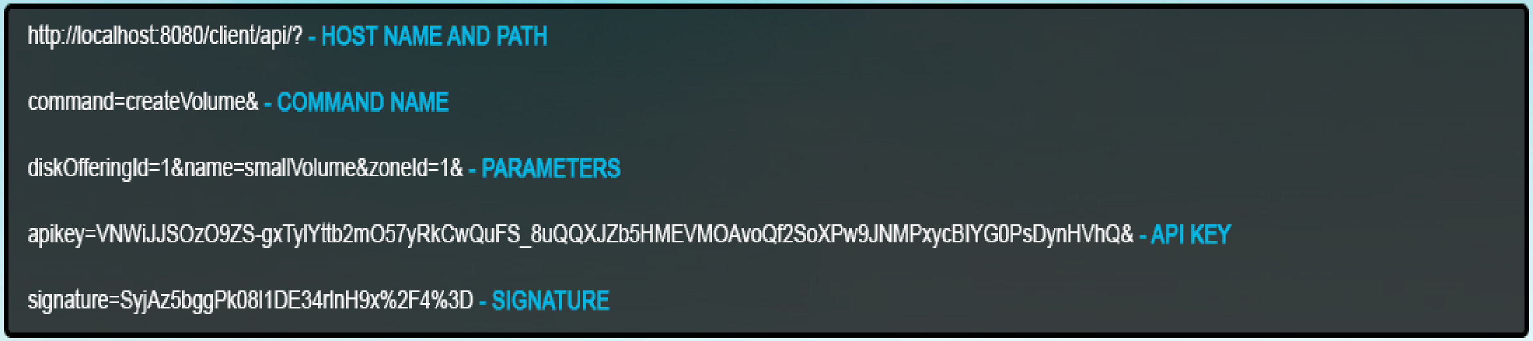
API Overview

- RESTful API interface
- UI/API pieces are stateless
- State is stored in MySQL database.
- All UI functionality is an API call
- Support xml/json as response type

Session-based Auth vs API Key Auth

- CloudStack supports two ways of authenticating via the API.
- Session-based Auth
  - Uses default Java Servlet cookie based sessions
  - Use the “login” API to get a JSESSIONID cookie and a SESSIONKEY token
  - All API commands require both cookie and token to authenticate
  - Has a timeout as configured within Tomcat
- API Key Auth
  - Works similarly to AWS API
  - Requires a bit more coding to generate the signature
  - All API commands require a signature hash

SIGNING REQUEST WITH API KEY / SECRET KEY



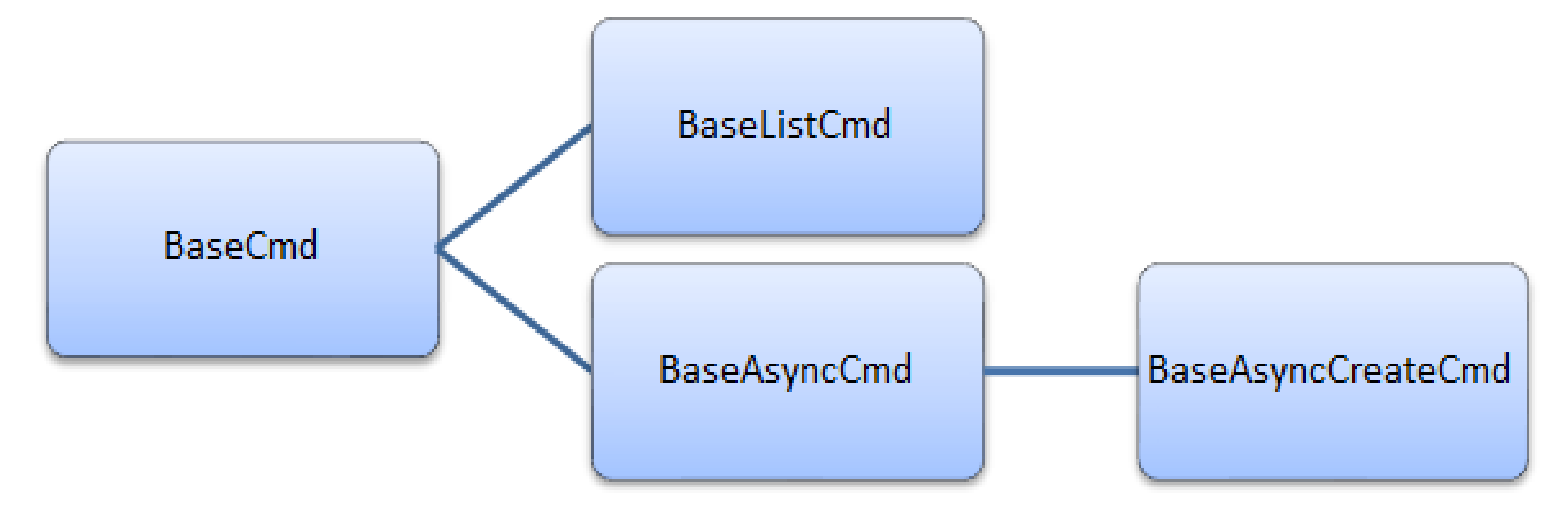
```
http://localhost:8080/client/api/? - HOST NAME AND PATH

command=createVolume& - COMMAND NAME

diskOfferingId=1&name=smallVolume&zoneId=1& - PARAMETERS

apikey=VNWiJJSOzO9ZS-gxTylYttb2mO57yRkCwQuFS_8uQQXJZb5HMEVMOAvoQf2SoXPw9JNMPxycBlYG0PsDynHVhQ& - API KEY

signature=SyjAz5bggPk08I1DE34rlnH9x%2F4%3D - SIGNATURE
```

Step 1:
commandString = command name + parameters + api key
URL encode each field-value pair within the commandstring

Step 2:
Lower case the entire commandString and sort it alphabetically via the field for each field-value pair.
sortedCommandString :
apiKey=vmwijj…&command=createvolume&diskofferingid=1&name=smallvolume=zoneid=1

Step 3:
Take the sortedCommandString and run it through the HMAC SHA-1 hashing algorithm (most programming languages offer a utility method to do this) with the user's Secret Key. Base64 encode the resulting byte array in UTF-8 so that it can be safely transmitted via HTTP. The final string produced after Base64 encoding should be SyjAz5bggPk08I1DE34lnH9x%2f4%3D

Commands



Asynchronous Commands

- CRUD (Create, Read, Update, Delete) of any first class objects in CloudStack, CUD are automatically asynchronous. R is synchronous.
- Rather than returning a response object, it will return a job ID.
- If it is a “Create” command, it will also return the object ID.
- With the job ID, you can query the async job status via the *queryAsyncJobResult* command.
- The *queryAsyncJobResult* response will return the following possible job status code:
  - 0 - Job is still in progress. Continue to periodically poll for any status changes.
  - 1 - Job has successfully completed. The job will return any successful response values associated with command that was originally executed.
  - 2 - Job has failed to complete. Please check the <jobresultcode> tag for failure reason code and <jobresult> for the failure reason.

RESPONSE FORMAT

CloudStack supports two formats as the response to an API call.

The default response is XML. If you would like the response to be in JSON, add **&response=json** to the Command String., Sample XML Response:

```
<listipaddressesresponse>
   <allocatedipaddress>
   <ipaddress>192.168.10.141</ipaddress>
   <allocated>2012-12-18T13:16:10-0700</allocated>
   <zoneid>4</zoneid>
   <zonename>Work</zonename>
   <issourcenat>true</issourcenat>
</allocatedipaddress> </listipaddressesresponse>
```

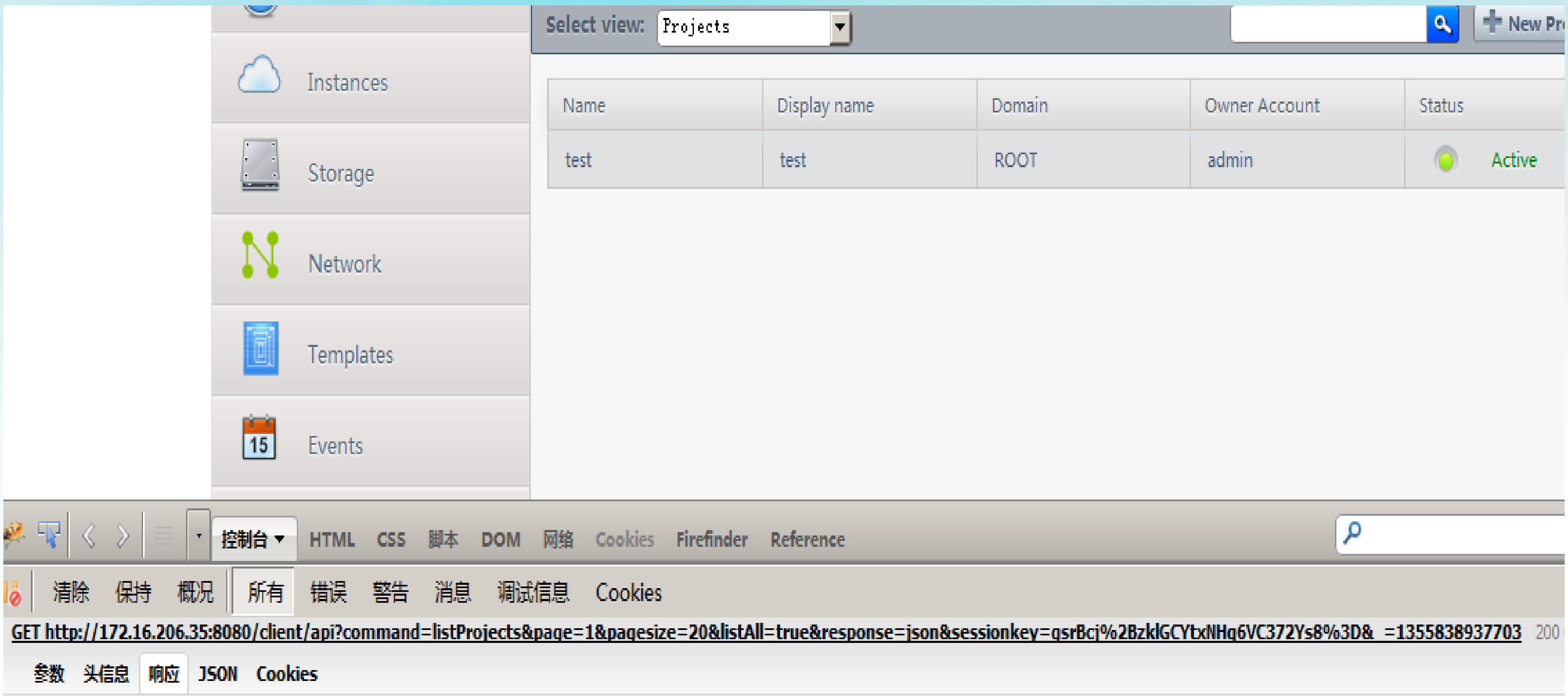
Sample JSON Response:

```
{ "listipaddressesresponse" : { "allocatedipaddress" : [ { "ipaddress" : "192.168.10.141",
"allocated" : "2012-12-18T13:16:10-0700", "zoneid" : "4", "zonename" : "Work", "issourcenat" :
"true" } ]
```

Pagination

- Using the page and pagesize parameter
  - page defines the current cursor to the list
  - pagesize defines the number of items per request
  - Pagesize is limited by the administrator
  - Sample:
    - listVirtualMachines&page=1&pagesize=500
    - listVirtualMachines&page=2&pagesize=500

Testing – From Web (Firebug, etc)

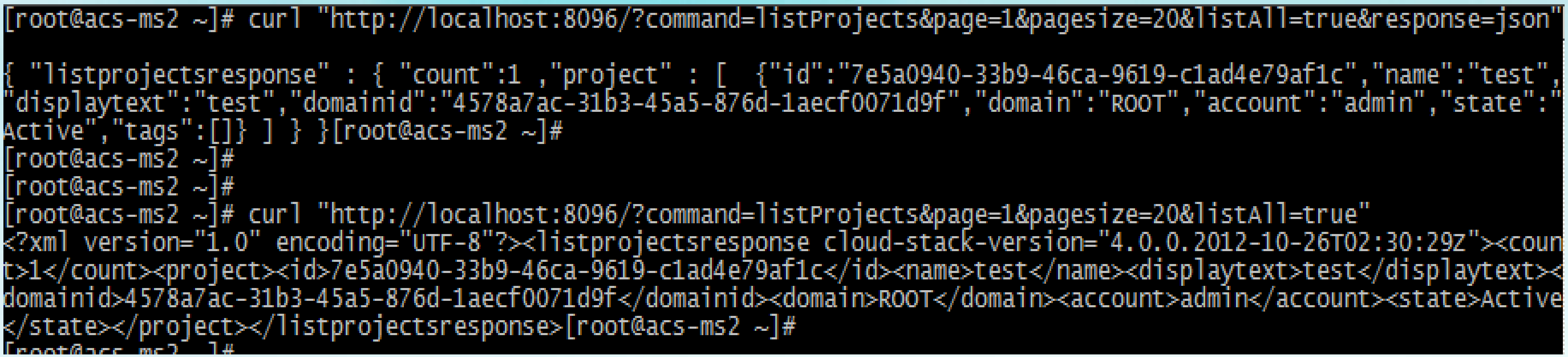

Instances
Storage
Network
Templates
Events
Select view: Projects
New Pr
控制台 HTML CSS 脚本 DOM 网络 Cookies Firefinder Reference
清除 保持 概况 所有 错误 警告 消息 调试信息 Cookies
参数 头信息 响应 JSON Cookies

Name	Display name	Domain	Owner Account	Status
test	test	ROOT	admin	● Active

GET http://172.16.206.35:8080/client/api?command=listProjects&page=1&pagesize=20&listAll=true&response=json&sessionkey=qsrBcj%2BzklGCYtxNHq6VC372Ys8%3D&_=1355838937703 200

{ "listprojectsresponse" : { "count":1 ,"project" : [{"id":"7e5a0940-33b9-46ca-9619-c1ad4e79af1c","name"
:"test","displaytext":"test","domainid":"4578a7ac-31b3-45a5-876d-1aecf0071d9f","domain":"ROOT","account"
:"admin","state":"Active","tags":[]}] } }

Testing – From API Server

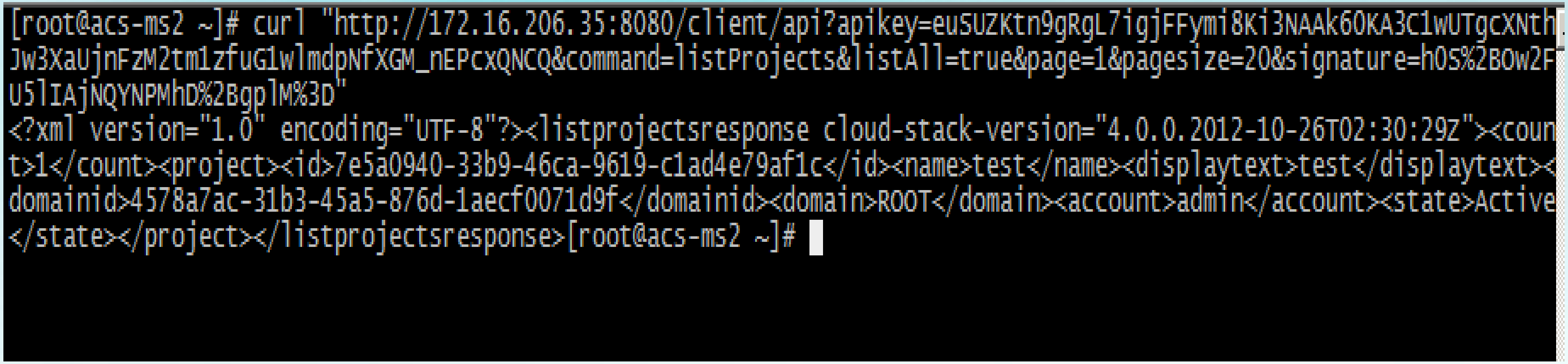


```
[root@acs-ms2 ~]# curl "http://localhost:8096/?command=listProjects&page=1&pagesize=20&listAll=true&response=json"

{ "listprojectsresponse" : { "count":1 ,"project" : [  {"id":"7e5a0940-33b9-46ca-9619-c1ad4e79af1c","name":"test",
"displaytext":"test","domainid":"4578a7ac-31b3-45a5-876d-1aecf0071d9f","domain":"ROOT","account":"admin","state":"
Active","tags":[]} ] } }[root@acs-ms2 ~]#
[root@acs-ms2 ~]#
[root@acs-ms2 ~]#
[root@acs-ms2 ~]# curl "http://localhost:8096/?command=listProjects&page=1&pagesize=20&listAll=true"
<?xml version="1.0" encoding="UTF-8"?><listprojectsresponse cloud-stack-version="4.0.0.2012-10-26T02:30:29Z"><coun
t>1</count><project><id>7e5a0940-33b9-46ca-9619-c1ad4e79af1c</id><name>test</name><displaytext>test</displaytext><
domainid>4578a7ac-31b3-45a5-876d-1aecf0071d9f</domainid><domain>ROOT</domain><account>admin</account><state>Active
</state></project></listprojectsresponse>[root@acs-ms2 ~]#
[root@acs-ms2 ~]#
```

- Port number, default 0 means disabled
- Suggest using inside Management Server
- Unsecure, must take extra caution

cloudstack

Testing – From Signature



```
[root@acs-ms2 ~]# curl "http://172.16.206.35:8080/client/api?apikey=euSUZKtn9gRgL7igjFFymi8Ki3NAAk60KA3C1wUTgcXNth
Jw3XaUjnFzM2tm1zfuG1wlmdpNfXGM_nEPcxQNCQ&command=listProjects&listAll=true&page=1&pagesize=20&signature=h0S%2BOW2F
U5lIAjNQYNPMhD%2Bgp1M%3D"
<?xml version="1.0" encoding="UTF-8"?><listprojectsresponse cloud-stack-version="4.0.0.2012-10-26T02:30:29Z"><coun
t>1</count><project><id>7e5a0940-33b9-46ca-9619-c1ad4e79af1c</id><name>test</name><displaytext>test</displaytext><
domainid>4578a7ac-31b3-45a5-876d-1aecf0071d9f</domainid><domain>ROOT</domain><account>admin</account><state>Active
</state></project></listprojectsresponse>[root@acs-ms2 ~]#
```

- Good for automation testing
- Signature
  - – generated via cmd, para and secretkey
  - – Encoded by HmacSHA1
- Secure but inconvenient