OpenStack Cloud Storage

Sam Fineberg
HP Storage Division
fineberg@hp.com

What is OpenStack®

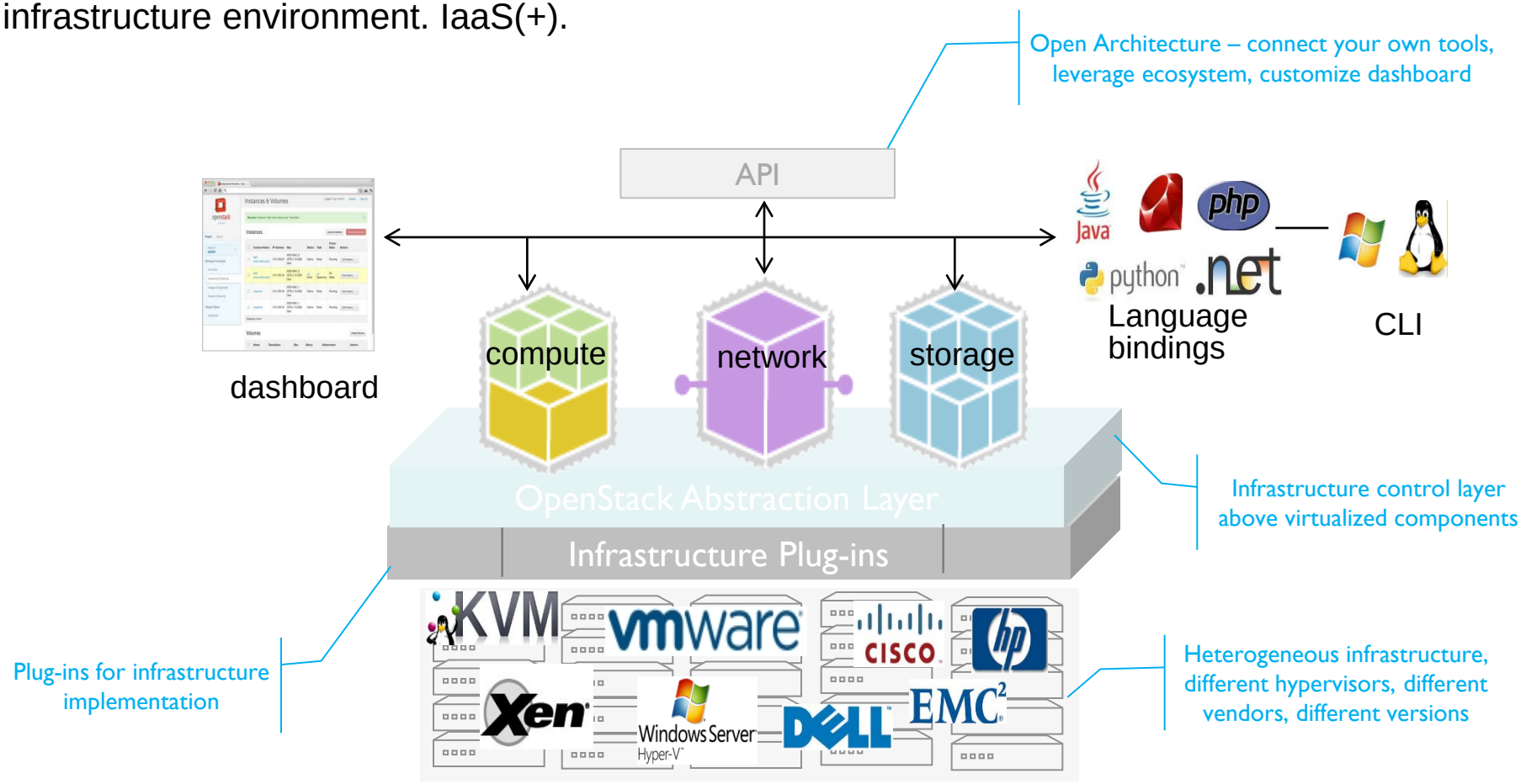
Free open source (Apache license) software governed by a non-profit foundation (corporation) with a mission **to produce the ubiquitous Open Source Cloud Computing platform that will meet the needs of public and private clouds regardless of size, by being simple to implement and massively scalable.**

- **Massively scalable** cloud operating system that controls large pools of **compute, storage**, and **networking** resources
- **Community open source** with contributions from **1000+ developers** and **180+** participating **organizations**
- **Open** web-based API ***Programmatic*** **Infrastructure** as a Service
- **Plug-in architecture**; allows different hypervisors, block storage systems, network implementations, hardware agnostic, etc.

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

What is OpenStack®

A series of interrelated projects that control pools of compute, storage, and networking infrastructure exposed as a consistent and open layer (API) for a heterogeneous infrastructure environment. IaaS(+).



2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

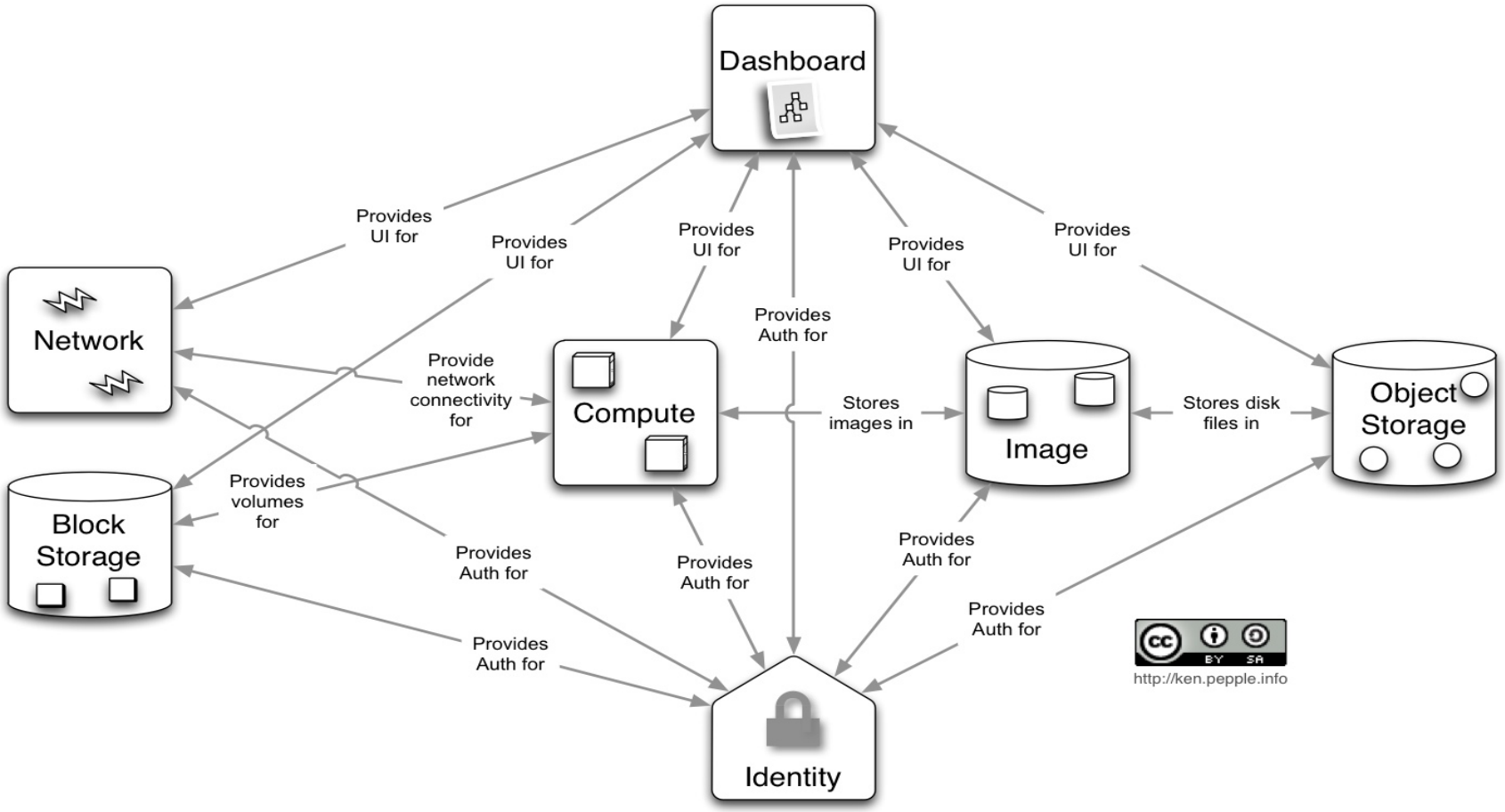
What is OpenStack®

7 core components with several others in incubation

- **Object Store** (codenamed "Swift") provides object (file) storage that is API accessible and URL referenceable. It allows you to store or retrieve files (but not mount directories like a fileserver).
- **Image** (codenamed "Glance") provides a catalog and repository for virtual disk images. Images can be private, shared, or public
- **Compute** (codenamed "Nova") provides virtual servers upon demand.
- **Dashboard** (codenamed "Horizon") provides a web-based user interface for most of the OpenStack services where users can perform most cloud operations like launching an instance, assigning IP addresses and setting access controls.
- **Identity** (codenamed "Keystone") provides authentication and authorization for all the OpenStack services. It also provides a service catalog of services within a particular OpenStack cloud.
- **Network** (codenamed “Neutron") provides network connectivity services between interface devices managed . The service works by allowing users to create their own networks and then attach interfaces to them.
- **Block Storage** (codenamed "Cinder") provides persistent block storage volumes to guest VMs.

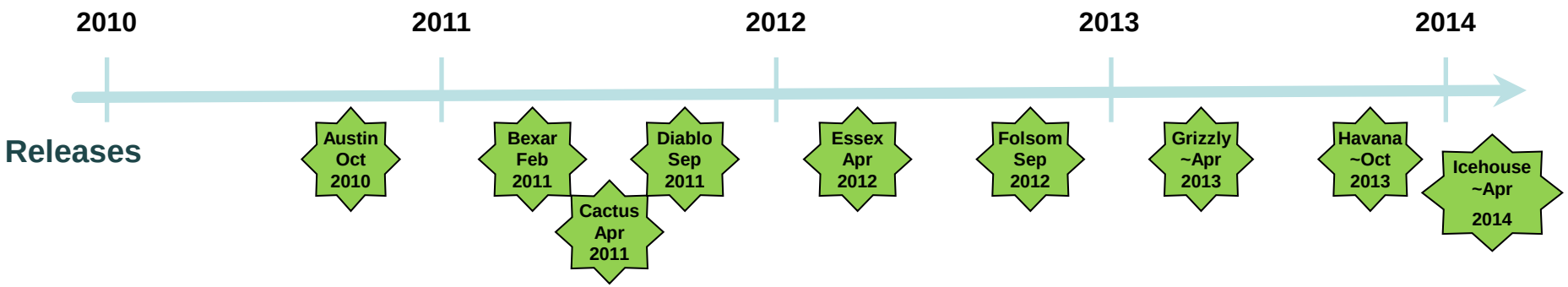
2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

OpenStack Components



2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

OpenStack Releases



6-Month Release Cycles

- Typically spring and fall releases for predictable availability
- Loosely coupled to Ubuntu releases and Ubuntu Enterprise Cloud
- Release names use alphabetic naming

Planning

- Developers plan the next release at the Design Summit
- Sessions are selected by projects leads and are generally driven by blueprint topics
- Project Technical Leads (PTLs) accept a number of blueprints into the release plan

Development &Testing

- Milestone iterations (commonly 5 weeks) are defined and followed
- Development/documentation occurs per the blueprints and plan
- End of cycle testing with a defined Release Criteria process

Release

- Release Candidate is made available
- Next release is open for development
- After hardening, release occurs

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cloud Storage 101

Block Storage

- Generally SCSI protocol based, organized by LUNs
- Boot volumes for VMs
- Ephemeral vs. Persistent
- Not directly consumed by applications, usually used to hold a filesystem
- Low level storage abstraction upon which file and object storage is built

File Storage

- Files organized in directory hierarchy and accessed by pathname
- File-based NAS protocols like NFS and CIFS
- Rich and complex application support: random access, in-place file updates, locking, etc.

Object Storage

- Efficient flat namespace: objects organized by accounts, containers, object keys, and metadata
- HTTP / REST / URL based – easily scriptable, many language choices
- Relatively simple interface compared to file storage
- Scalable to very high object counts
- More easily scaled across multiple geographies
- Ideal for relatively static data

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cloud Storage 101

Block Storage

- Generally SCSI protocol based, organized by LUNs
- Boot volumes for VMs
- Ephemeral vs. Persistent
- Not directly consumed by applications, usually used to hold a filesystem
- Low level storage abstraction upon which file and object storage is built

File Storage

- Files organized in directory hierarchy and accessed by pathname
- File-based NAS protocols like NFS and CIFS
- Rich and complex application support: random access, in-place file updates, locking, etc.

Object Storage

- Efficient flat namespace: objects organized by accounts, containers, object keys, and metadata
- HTTP / REST / URL based – easily scriptable, many language choices
- Relatively simple interface compared to file storage
- Scalable to very high object counts
- More easily scaled across multiple geographies
- Ideal for relatively static data

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

OpenStack Block Storage

- Nova Volume
  - Originally OpenStack Compute (Nova) included support for ephemeral volumes
    - Used for boot/runtime storage of VMs
    - Ephemeral – volumes only existed as long as VMs
    - Volumes were typically backed by VM server files
  - Nova Volume had limited support persistent volumes on iSCSI
- Beginning with the Folsom release, a separate persistent block stoage service, Cinder, was created
  - Cinder is a core part OpenStack project
  - Consists of a plug-in interface for supporting various block storage devices

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cinder Core Functionality

- Volumes
  - Create, Show, Update, Delete Volume
  - List Volume Summaries/Details
- Snapshots
  - Create, Show, Update, Delete Snapshot
  - List Snapshot Summaries/Details
- Volume types
  - List/Show volume types
- Extensions
  - Backups

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cinder Supported Devices

- Drivers are available for
  - Ceph
  - EMC (SMI-S)
  - GlusterFS
  - HP 3Par
  - HP Lefthand
  - NetApp
  - Nexenta
  - Solidfire
  - NFS
- Interfaces
  - iSCSI
  - Fiber Channel
  - NFS/shared file system

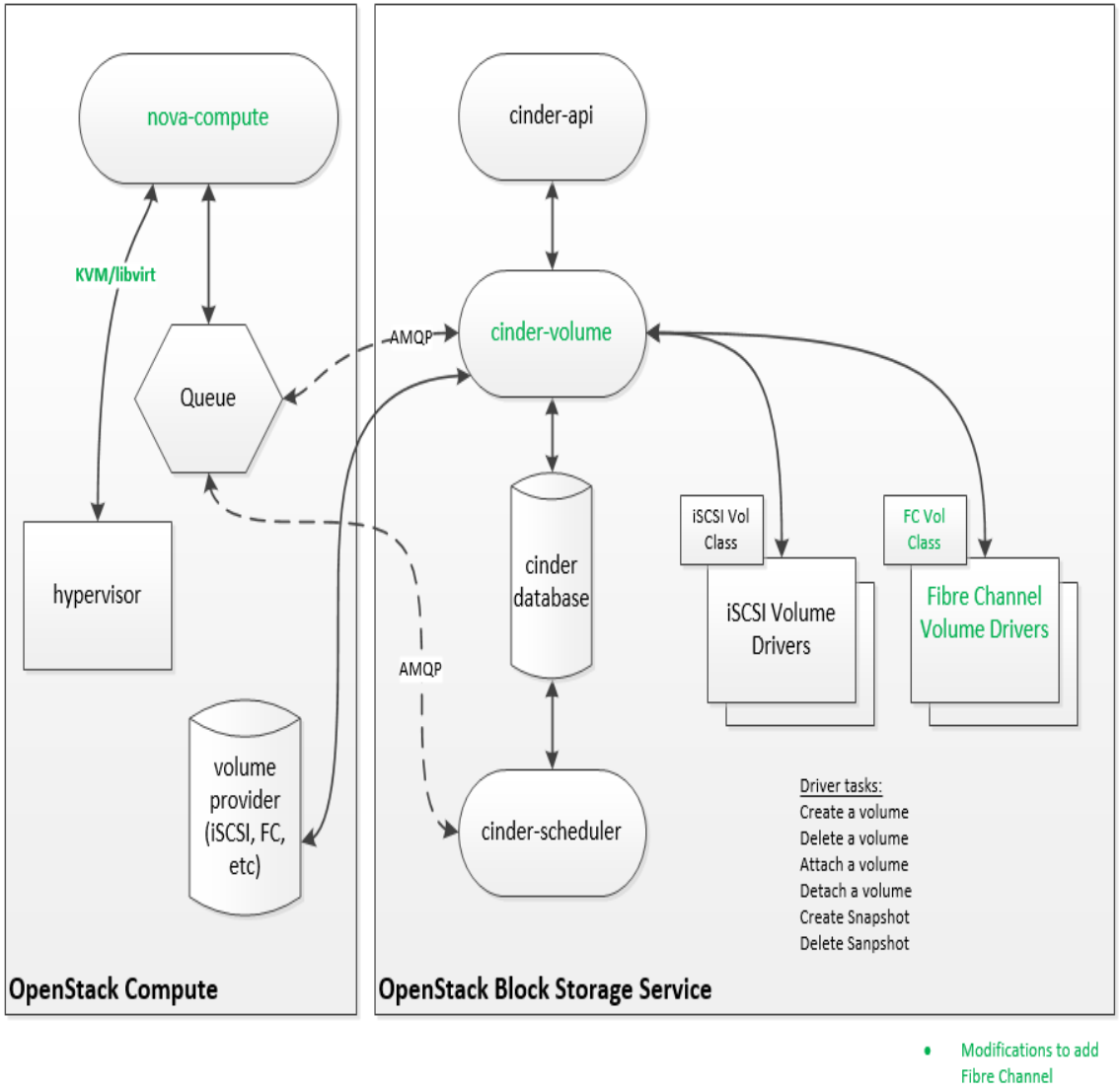
2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cinder volume attachment

- NFS/shared file system
  - Similar to Nova ephemeral volumes
  - Volume file created on file share, attached through libvirt or other VM specific mechanism
  - All VM hosts must already be attached to file share
- iSCSI
  - Attach to iSCSI volume over TCP/IP network
  - Cinder provisions volumes on target, coordinates initiator and target connection

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cinder Fibre Channel Support in Grizzly



- Developed by a group of vendors, led by HP and Brocade
- Cinder changes to add support for FC volume class driver
- Cinder changes to add support for FC volume attach
- Nova changes to support FC volume attach
  - KVM hypervisor only
- Reference Cinder FC volume driver
  - HP 3Par array driver
- Future releases
  - Zoning
  - Security
  - Additional hypervisors
  - Additional arrays

Cinder Futures

- Additional drivers
- Additional arrays
- Better FC support
- QoS
- Shared volumes

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cloud Storage 101

Block Storage

- Generally SCSI protocol based, organized by LUNs
- Boot volumes for VMs
- Ephemeral vs. Persistent
- Not directly consumed by applications, usually used to hold a filesystem
- Low level storage abstraction upon which file and object storage is built

File Storage

- Files organized in directory hierarchy and accessed by pathname
- File-based NAS protocols like NFS and CIFS
- Rich and complex application support: random access, in-place file updates, locking, etc.

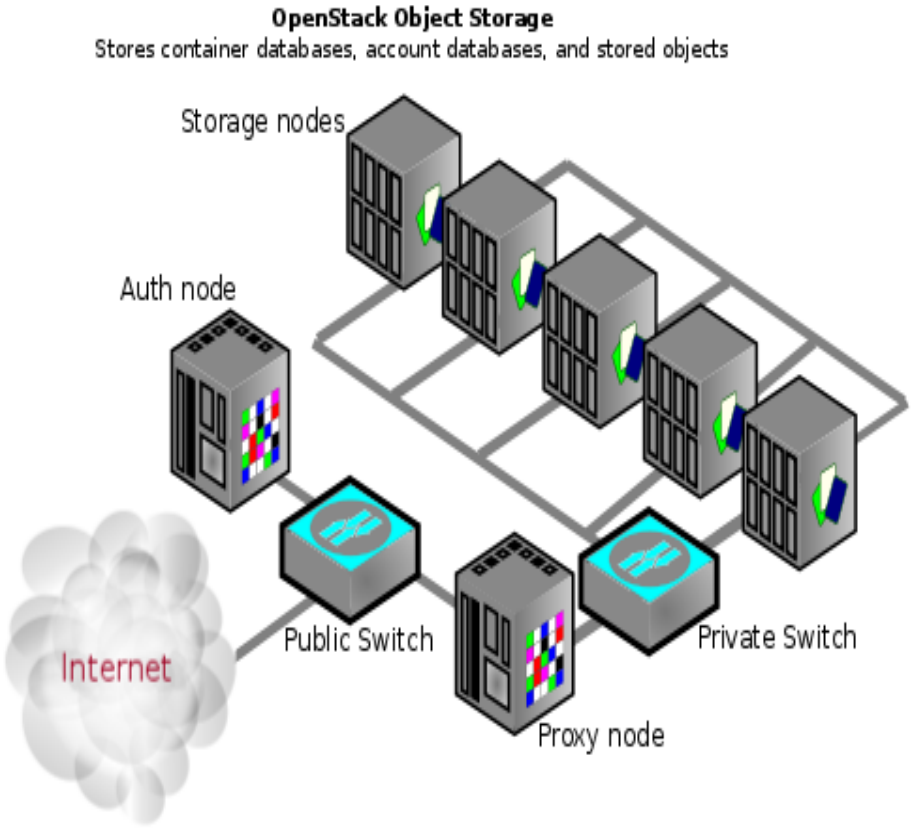
Object Storage

- Efficient flat namespace: objects organized by accounts, containers, object keys, and metadata
- HTTP / REST / URL based – easily scriptable, many language choices
- Relatively simple interface compared to file storage
- Scalable to very high object counts
- More easily scaled across multiple geographies
- Ideal for relatively static data

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

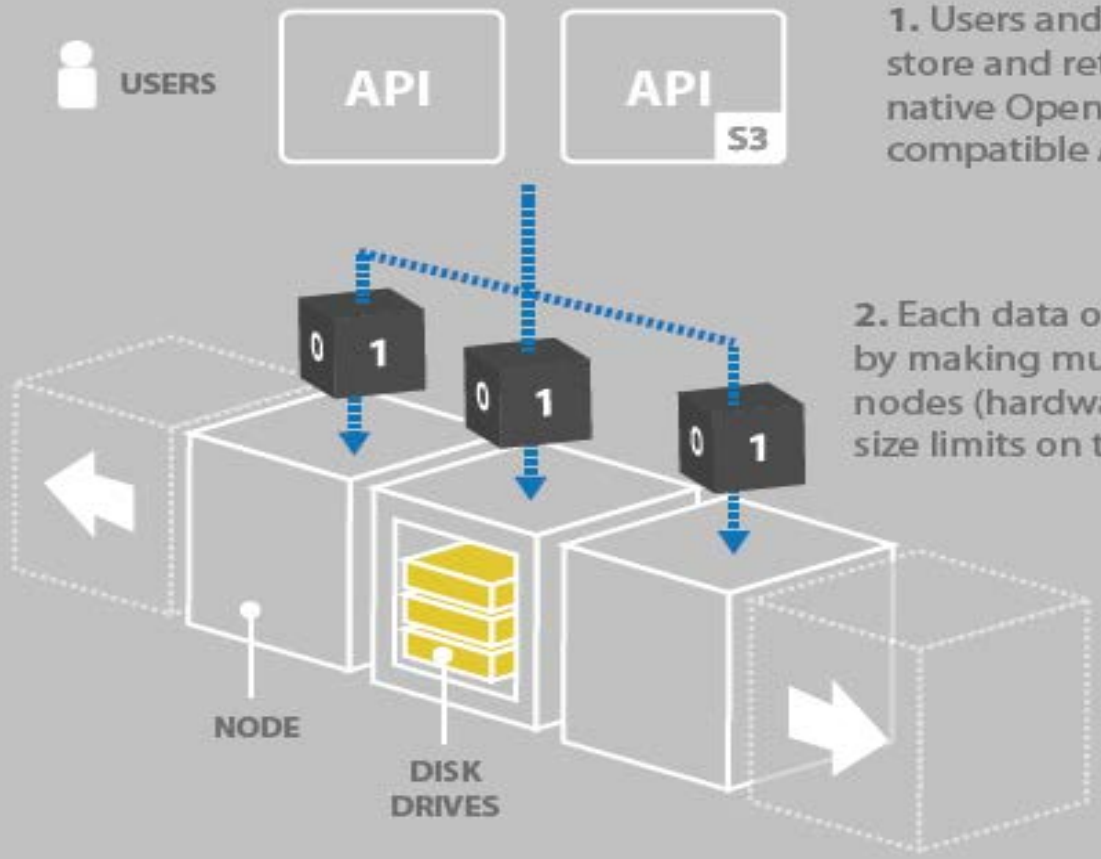
OpenStack Object Storage - Swift

- Swift was one of the original components of OpenStack
  - Originally developed by RackSpace, in production in their data center (Cloud Files)
  - Distributed scale-out shared-nothing object storage
  - Web APIs for data access
    - Swift API
    - "S3" API
- Swift is open source code, not an API standard
  - While some vendors may implement a Swift compatible API, there is no compatibility suite and the API is subject to change
- Storage model
  - Users authenticate with Keystone
  - Each account owns a set of containers
  - Containers hold a set of objects
    - Containers also have metadata, both system and user defined
    - Access permissions are set at the container level
  - Objects contain data
    - Size is limited, but composite objects are supported for larger objects
    - Objects also have system and user defined metadata
    - User specifies object name inside container
    - Name can contain pseudo-paths



2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

OPENSTACK OBJECT STORAGE



1. Users and applications request to store and retrieve data through the native OpenStack API or the Amazon S3 compatible API

2. Each data object is stored redundantly by making multiple copies to different nodes (hardware devices). There are no size limits on the objects stored.

3. Just by adding more nodes, clusters are massively scalable to multi- petabyte size and billions of objects

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Swift Details

- The Ring
  - A ring is a hash map between entity names and the physical location of servers
  - There are separate rings for accounts, containers, and objects
  - The Ring maintains this mapping using zones, devices, partitions, and replicas.
  - Each partition in the ring is replicated, by default, 3 times across the cluster
- Proxy Server
  - The Proxy Server is responsible for tying together the rest of the Object Storage architecture.
  - For each request, it will look up the location and route the request accordingly.
  - The public developer API is also exposed through the Proxy Server.
- Container Server
  - Primary job is to handle listings of objects
  - It doesn't know where those object's are, just what objects are in a specific container
  - The listings are stored as sqlite database files, and replicated across the cluster Account Server
- Account Server
  - Very similar to the Container Server, excepting that it is responsible for listings of containers
  - The listings are stored as replicated sqlite database files
- Object Server
  - A very simple server that can store, retrieve and delete objects stored on local devices
  - Objects are stored as binary files on the filesystem with metadata stored in extended attributes

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Swift Details (cont.)

- Replication
  - Designed to keep the system in a consistent state in the face of temporary error conditions
  - Replication processes compare local data with remote copies to ensure they contain the latest version.
  - Replication updates are push based, updating is just a matter of rsyncing files to the peer.
  - Account and container replication push missing records over HTTP or rsync whole database files.
- Updaters
  - There are times when container or account data can not be immediately updated.
  - If an update fails, it is queued locally on the file system, and the updater will process it.
  - This is where an eventual consistency comes into to play.
- Auditors
  - Auditors crawl the local server checking the integrity of the objects, containers, and accounts.
  - If corruption is found (in the case of bit rot, for example), the file is quarantined, and replication will replace the bad file from another replica.
  - If other errors are found they are logged

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Swift New Features/Futures

- Grizzly
  - Global clusters
    - Support for geo-dispersed zones – region tier
    - Differing replica counts per region
    - Read from closest replica – based on timing
  - Bulk requests
    - Auto extract
    - Bulk delete
  - Quotas
- Havana+
  - Erasure coding
  - Container sharding

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Cloud Storage 101

Block Storage

- Generally SCSI protocol based, organized by LUNs
- Boot volumes for VMs
- Ephemeral vs. Persistent
- Not directly consumed by applications, usually used to hold a filesystem
- Low level storage abstraction upon which file and object storage is built

File Storage

- Files organized in directory hierarchy and accessed by pathname
- File-based NAS protocols like NFS and CIFS
- Rich and complex application support: random access, in-place file updates, locking, etc.

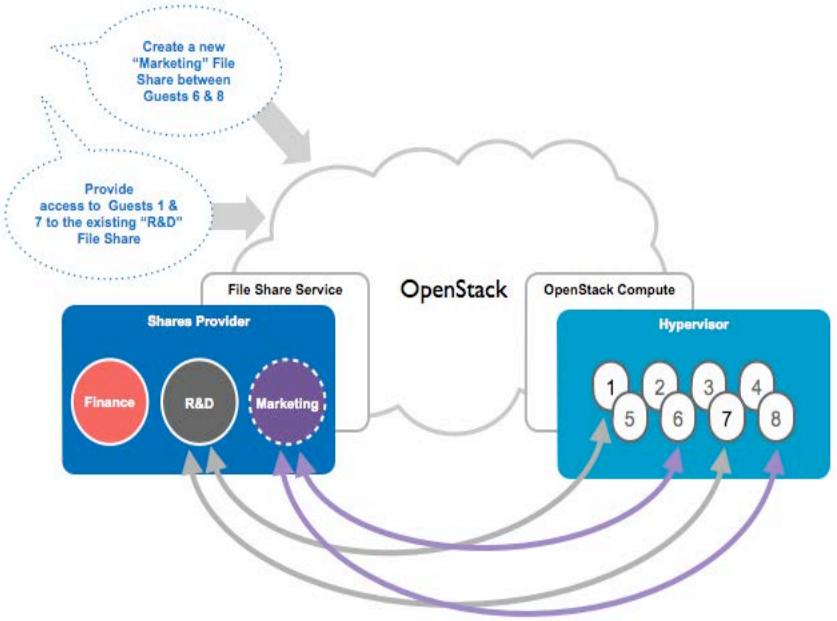
Object Storage

- Efficient flat namespace: objects organized by accounts, containers, object keys, and metadata
- HTTP / REST / URL based – easily scriptable, many language choices
- Relatively simple interface compared to file storage
- Scalable to very high object counts
- More easily scaled across multiple geographies
- Ideal for relatively static data

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

OpenStack Shared File Storage

- Not originally planned for OpenStack,
  - Clouds like AWS don't have shared file storage
  - Security and networking issues are difficult
- Demand has continued for this type of a service
  - For legacy applications that are not object storage enabled
  - Other options like running NFS on a VM or Gluster across nodes are complex to get right
- Originally proposed as an extension to Cinder
  - Proposal from NetApp
  - Implemented in Grizzly Cinder



2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Manila

- Manila emerged as a separate project over Summer 2013
  - Plan is to make an incubator project in Havana
  - Full project in Icehouse
  - Still very early in development, and looking for more participation
- Goal is to provide a file system that is shared between VMs
  - Multiple potential implementations
  - NFS/CIFS direct, VM mapped, …
  - Strong tie-in with Neutron
- Manila core functionality
  - Create or delete a share
  - Show or list shares
  - Allow/deny access to a share
  - Create/delete share snapshots, list snapshots

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Summary

- OpenStack is free open source software for implementing public and private clouds
  - Many companies and individuals are involved in development
  - Products and services based on OpenStack exist today, and more are appearing as it matures
- OpenStack has support for block, object, and file storage
  - Both open source and commercial storage products are supported
  - These provide basic functionality and are maturing quickly
- As a developer, you can get involved with OpenStack
  - There is plenty more work to do
  - The development process is designed to be open to anyone who wants to help
    - Get involved!

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.

Questions?

- More on OpenStack
  - http://openstack.org
  - http://docs.openstack.org
- Contact me:
  - fineberg@hp.com

2013 Storage Developer Conference. © Hewlett-Packard Company. All Rights Reserved.