Cloud Computing & Big Data

PARALLEL & SCALABLE MACHINE LEARNING & DEEP LEARNING

Prof. Dr. – Ing. Morris Riedel

Adjunct Associated Professor

School of Engineering and Natural Sciences, University of Iceland

Research Group Leader, Juelich Supercomputing Centre, Germany

SHORT LECTURE 13

OpenStack Cloud Operating System

November 22th, 2018

Room Endurmenntun – Naustið

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES

FACULTY OF INDUSTRIAL ENGINEERING,
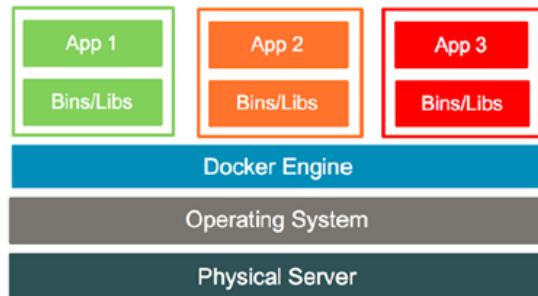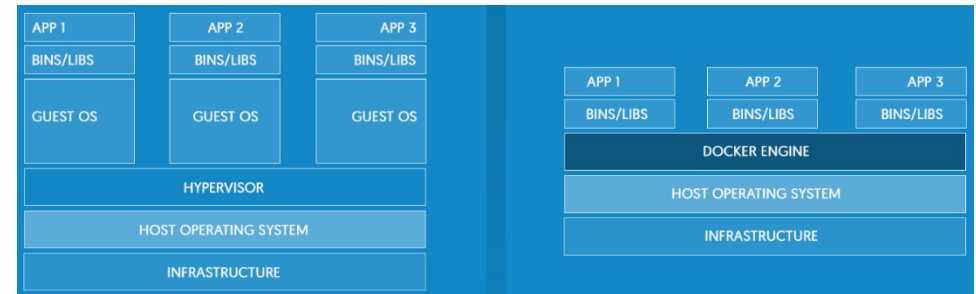MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH
Forschungszentrum

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

DEEP
Projects

Review of Lecture 12 – Docker & Container Management

- **Containerization of 'Software'**

[1] Docker Web page

(difference virtualization vs. containers?)
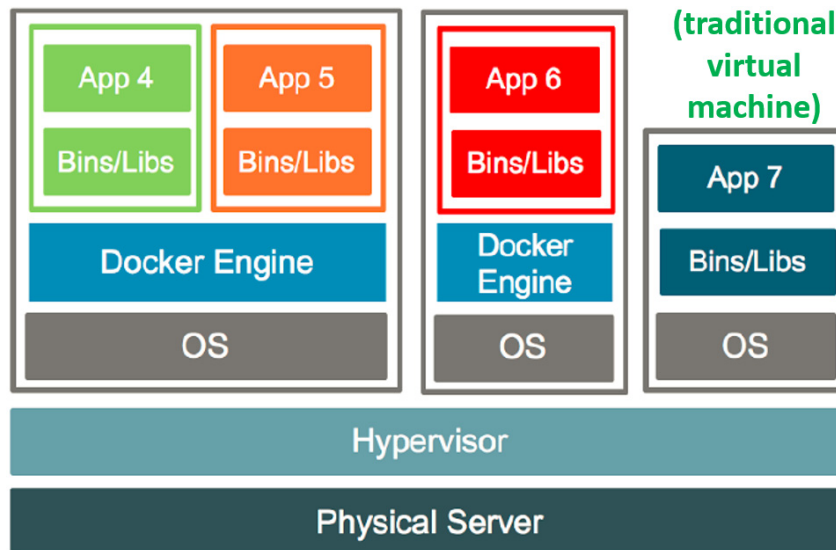
(traditional virtual machine)

- **Many Benefits (e.g. storage)**

- **Important tools & Clouds**
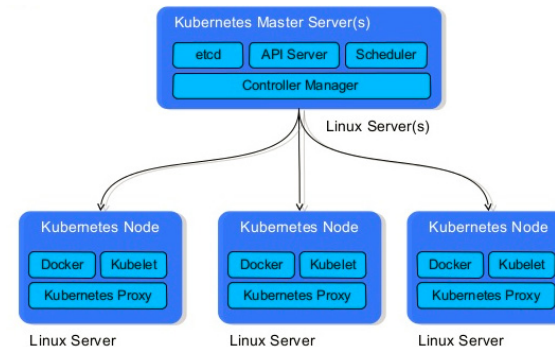
[2] Kubernetes Web page

[3] Apache Mesos Web page

(overall scheduling & resource management)

Outline of the Course

1. Cloud Computing & Big Data

2. Machine Learning Models in Clouds

3. Apache Spark for Cloud Applications

4. Virtualization & Data Center Design

5. Map-Reduce Computing Paradigm

6. Deep Learning driven by Big Data

7. Deep Learning Applications in Clouds

8. Infrastructure-As-A-Service (IAAS)

9. Platform-As-A-Service (PAAS)

10. Software-As-A-Service (SAAS)

11. Data Analytics & Cloud Data Mining

12. Docker & Container Management

13. OpenStack Cloud Operating System

14. Online Social Networking & Graphs

15. Data Streaming Tools & Applications

16. Epilogue

+ additional practical lectures for our

hands-on exercises in context

- Practical Topics

- Theoretical / Conceptual Topics

Outline

- This is only a Short Lecture
  - Goal is to provide a few pointers to other advanced related university courses/topics
  - 'OpenStack as Cloud Operating System' needs a full course & substantial tutorial
  - Links previous Lectures & Practical Lectures with further material to study & research

- OpenStack Cloud Operating System
  - Challenges for Virtualized IT Cloud Resources
  - Core Services Landscape
  - Service Interaction & Virtualization
  - OpenStack API & Usage Models
  - Business Cases & Adoption in Clouds

- Promises from previous lecture(s):
- *Lecture 1:* Lecture 13 provides in-depth insights into other features of the OpenStack cloud operating system
- *Lecture 4:* Lecture 13 provides more examples & applied scalability for different services in OpenStack system
- *Lecture 5:* Lecture 13 will provide more details on using the OpenStack Cloud Operating System for Clouds

OpenStack Cloud Operating System
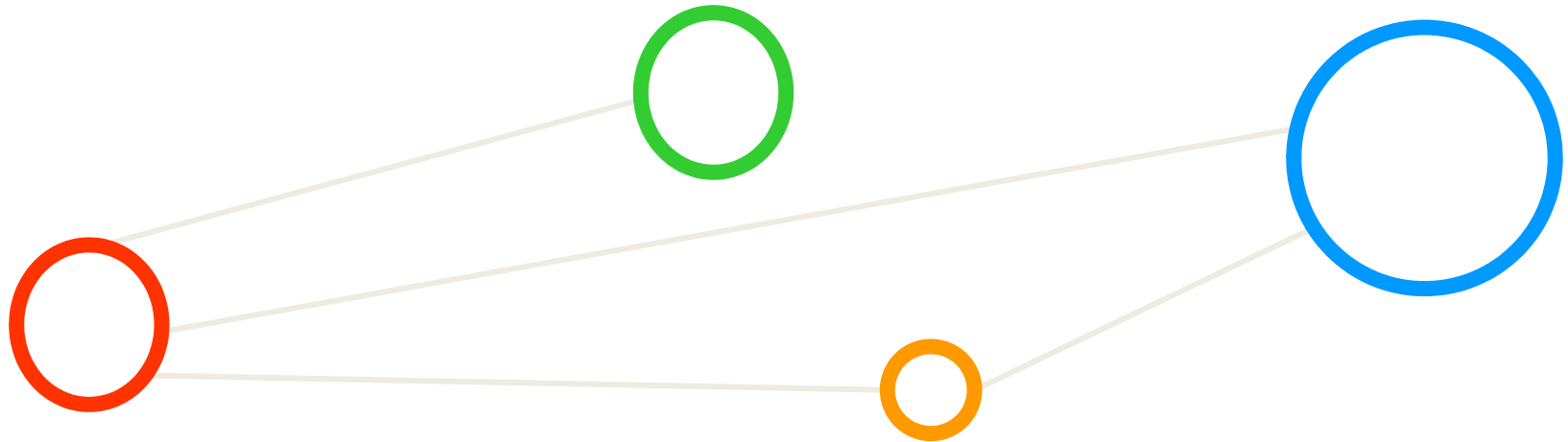
Virtualized Resources in Data Centers – Revisited

- Virtualized Infrastructure
  - Enabled by hardware virtualization concept
  - Virtualized servers, storage, and network combined in virtualized platforms
- Cloud computing
  - Enable elastic cloud systems via server virtualization
  - VMs installed on a virtualized platform are often used for hosting third-party programs (e.g. content streaming server)

(basis is still physical hardware)
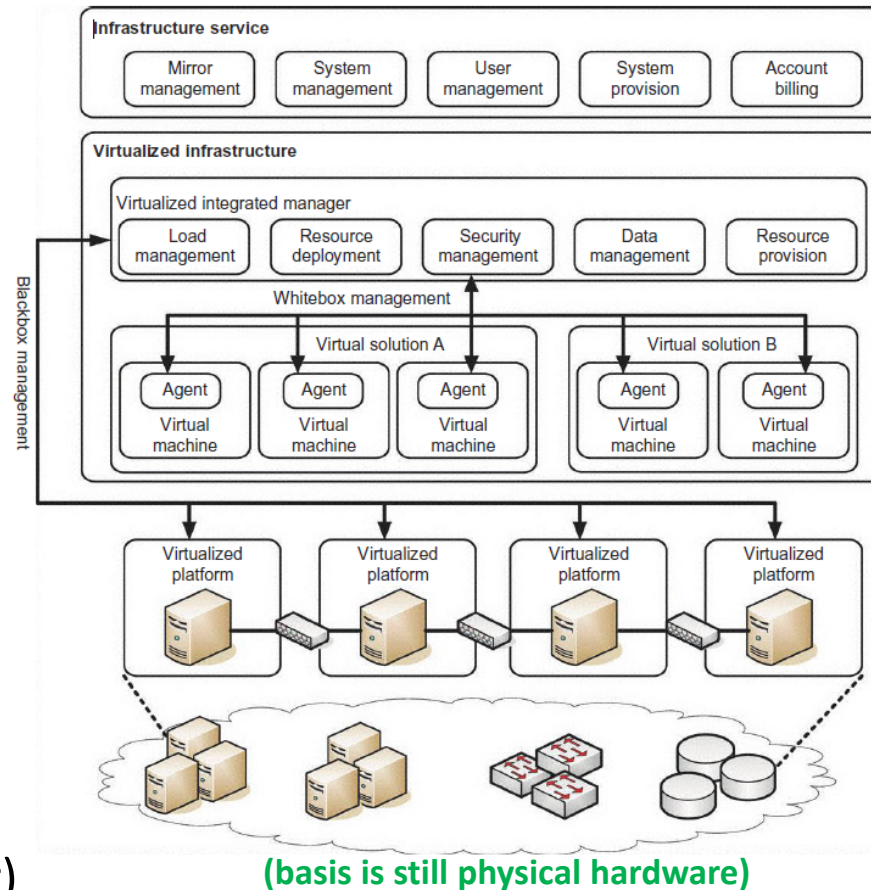
[3] Distributed & Cloud Computing Book

- **OpenStack Cloud Operating System manages & controls (geographically distributed) cloud resources**

Data Center Metrics – Application & Technology Scalability

- **Application scalability: Instead of increasing machine size, application enlarges problem size**
- **Technology scalability: Systems that can adapt to changes in building technologies (e.g. new CPUs)**

- ## Application scalability
  - Application has a problem size that affects the size of the data set or the workload increase for a data center server (problem size vs. machine size)
  - E.g. server for counting words gets one document & is 98% idle: add docs
  - E.g. storage is only filled 4%, application can further scale until storage full
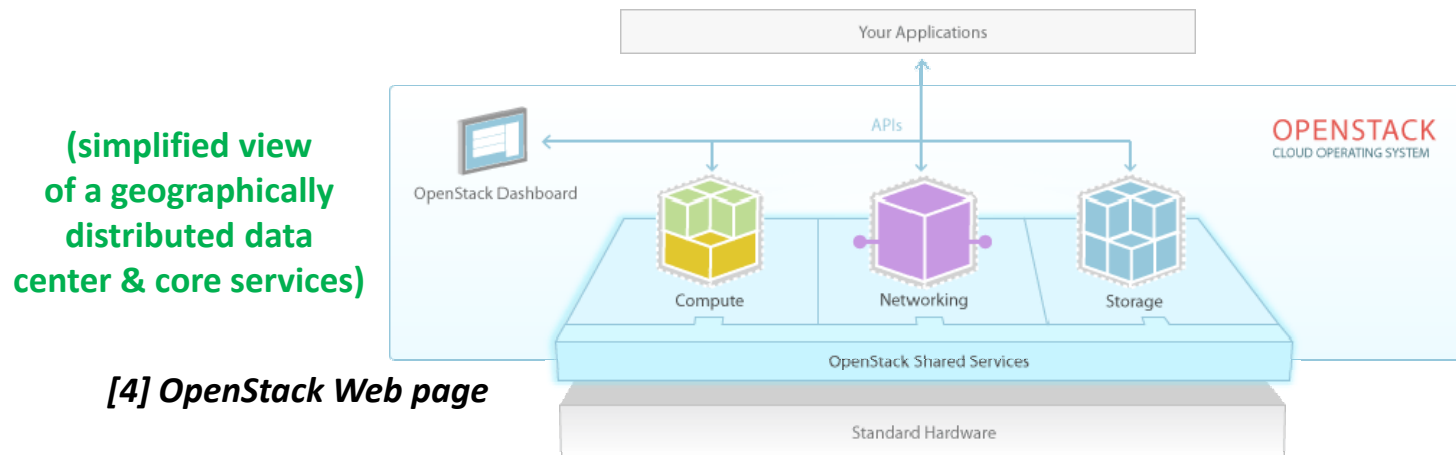 - E.g. application uses just one core from a multi-core chip, but can use more

- ## Technology scalability
  - Technologies: processor components, networking technologies, storages
  - E.g. new-generation CPU means impact to motherboard and power supply
  - E.g. packaging & energy concerns when porting among different suppliers
  - E.g. use of heterogenous hardware/software components from n vendors

OpenStack Cloud Operating System

- Open Source Software with API
  - Manages and controls large resource pools of compute, storage, and networking resources
  - Provides services with the key functionality of many cloud applications
  - Attractive dashboard that provides administrators with a tool to manage all resources via one graphical user interface (easy overview)
  - Marketplace that offers an easy way to find related products/services: e.g. existing distributions, appliances, consultants, trainings, etc.
  - Well curated sample configurations based on real-world applications

(simplified view of a geographically distributed data center & core services)

[4] OpenStack Web page

[5] Big Data Tips 'OpenStack'

Motivation for using OpenStack

67832 people

182 countries

651 supporting companies

20M+ lines of code

One of the fastest growing open source communities in the world.

The OpenStack project is a global collaboration of developers and cloud computing technologists producing the open standard cloud computing platform for both public and private clouds. Backed by a vibrant community of developers and some of the biggest names in the industry.

Get Involved ▶

openstack™ CLOUD SOFTWARE

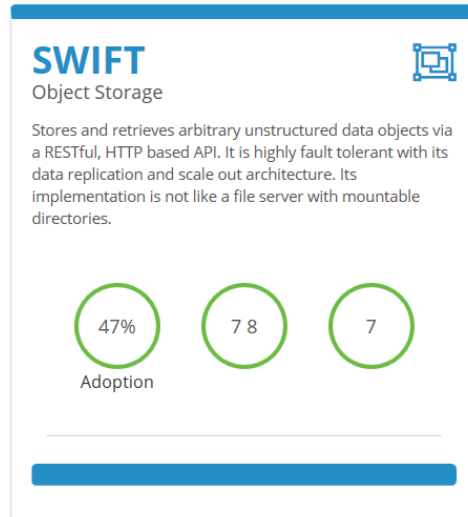Join us at the OpenStack Summit Boston, May 8-11

Whether you are pursuing a private, public or multi-cloud approach, the OpenStack Summit is the place to network, skill up and plan your cloud strategy.

[4] OpenStack Web page

(OpenNebula, Nimbus, or Eucalyptus are similar systems, but OpenStack is mostly used)

- **OpenStack provides six core shared services to manage cloud resources: Nova (compute), Neutron (network), Swift (object storage), Cinder (block storage), Keystone (identity), Glance (image service)**
- **OpenStack core services represent cloud functionality used over and over again in many applications**

Object Storage vs. Block Storage

SWIFT
Object Storage

Stores and retrieves arbitrary unstructured data objects via a RESTful, HTTP based API. It is highly fault tolerant with its data replication and scale out architecture. Its implementation is not like a file server with mountable directories.

47% 7 8 7

Adoption

CINDER
Block Storage

Provides persistent block storage to running instances. Its pluggable driver architecture facilitates the creation and management of block storage devices.

88% 8 8 5

Adoption

- **Object storage manages data as objects (not like file/folder hierarchy) whereby object = data, metadata & unique identifier**

- Object Storage architecture
  - Easy management like data replication/distribution
  - E.g. storing photos on Facebook
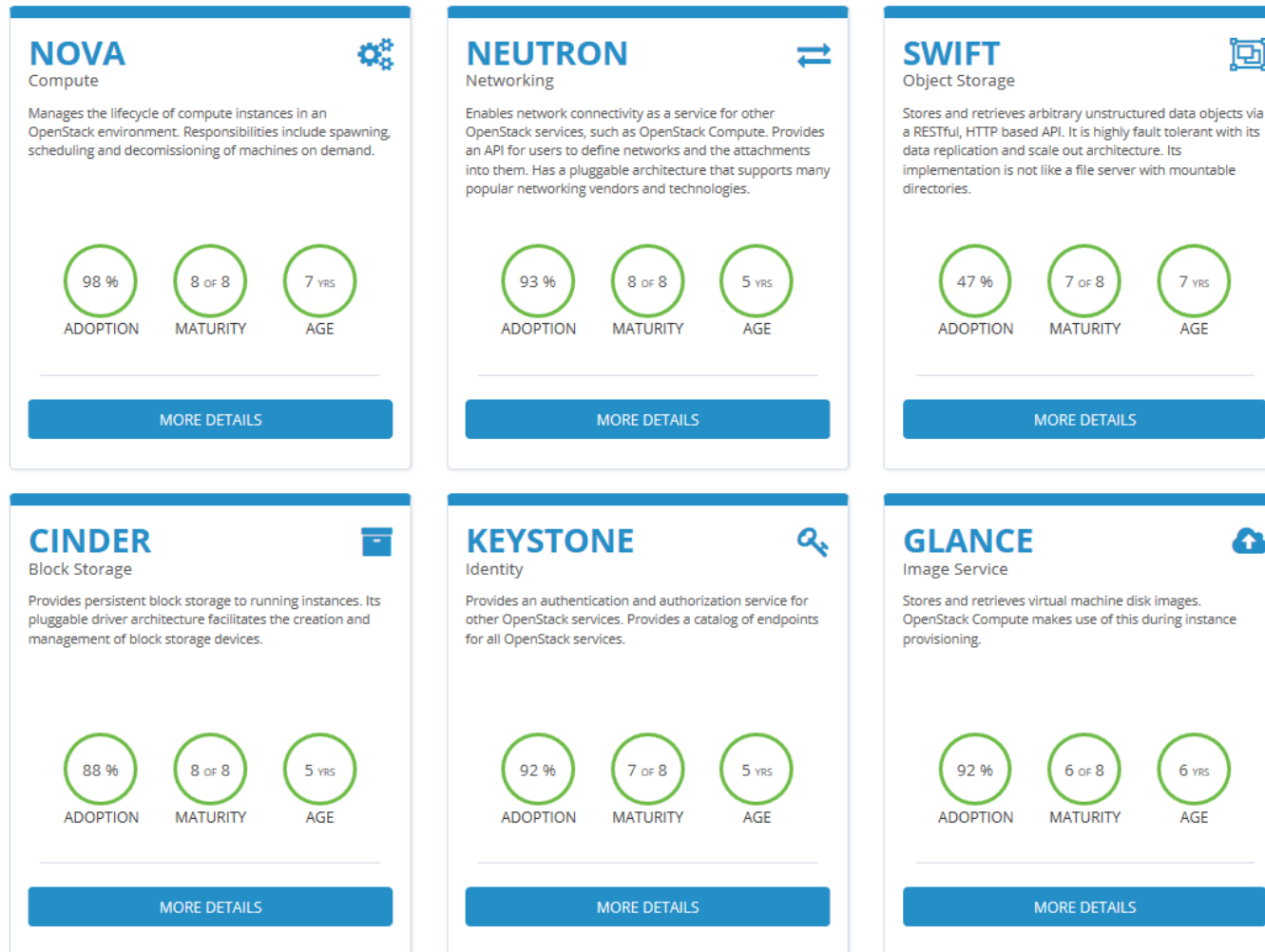 - E.g. retrieving songs on Spotify
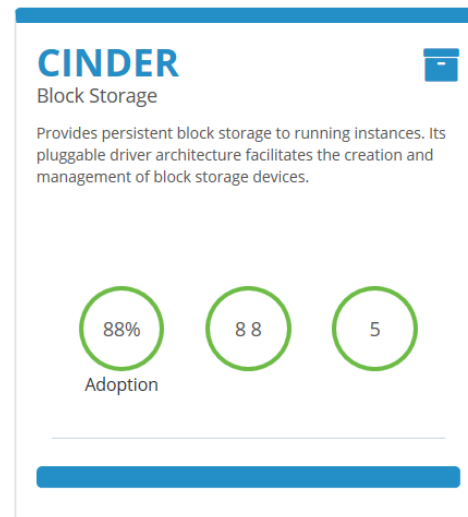
- **Block storage manages data as blocks and split files up into raw blocks of data whereby each block is controlled like a hard drive**

- Block Storage architecture
  - Provides API to request and consume resources w/o requiring knowledge of where storage is
  - E.g. directly accessed by the operating systems

[6] Druva Blog

OpenStack Core Services Landscape

NOVA ⚙️
Compute

Manages the lifecycle of compute instances in an OpenStack environment. Responsibilities include spawning, scheduling and decomissioning of machines on demand.

98 %	8 OF 8	7 YRS
ADOPTION	MATURITY	AGE

MORE DETAILS

NEUTRON ⇄
Networking

Enables network connectivity as a service for other OpenStack services, such as OpenStack Compute. Provides an API for users to define networks and the attachments into them. Has a pluggable architecture that supports many popular networking vendors and technologies.

93 %	8 OF 8	5 YRS
ADOPTION	MATURITY	AGE

MORE DETAILS

SWIFT
Object Storage

Stores and retrieves arbitrary unstructured data objects via a RESTful, HTTP based API. It is highly fault tolerant with its data replication and scale out architecture. Its implementation is not like a file server with mountable directories.

47 %	7 OF 8	7 YRS
ADOPTION	MATURITY	AGE

MORE DETAILS

CINDER
Block Storage

Provides persistent block storage to running instances. Its pluggable driver architecture facilitates the creation and management of block storage devices.

88 %	8 OF 8	5 YRS
ADOPTION	MATURITY	AGE

MORE DETAILS

KEYSTONE 🔑
Identity

Provides an authentication and authorization service for other OpenStack services. Provides a catalog of endpoints for all OpenStack services.

92 %	7 OF 8	5 YRS
ADOPTION	MATURITY	AGE

MORE DETAILS

GLANCE ☁️
Image Service

Stores and retrieves virtual machine disk images. OpenStack Compute makes use of this during instance provisioning.

92 %	6 OF 8	6 YRS
ADOPTION	MATURITY	AGE

MORE DETAILS

[4] OpenStack Web page

OpenStack Optional Services

Horizon	Dashboard	6 of 8	5 Yrs	92 %
Ceilometer	Telemetry	1 of 8	4 Yrs	62 %
Heat	Orchestration	6 of 8	4 Yrs	66 %
Trove	Database	3 of 8	3 Yrs	17 %
Sahara	Elastic Map Reduce	3 of 8	3 Yrs	11 %
Ironic	Bare-Metal Provisioning	5 of 8	3 Yrs	25 %
Zaqar	Messaging Service	4 of 8	3 Yrs	2 %
Manila	Shared Filesystems	5 of 8	3 Yrs	11 %
Designate	DNS Service	3 of 8	3 Yrs	19 %
Barbican	Key Management	2 of 8	3 Yrs	5 %
Magnum	Containers	2 of 8	2 Yrs	13 %
Murano	Application Catalog	1 of 8	2 Yrs	12 %
Congress	Governance	1 of 8	2 Yrs	2 %

(cf. Lecture 5)

[4] OpenStack Web page

OpenStack Service Interaction & Virtualization
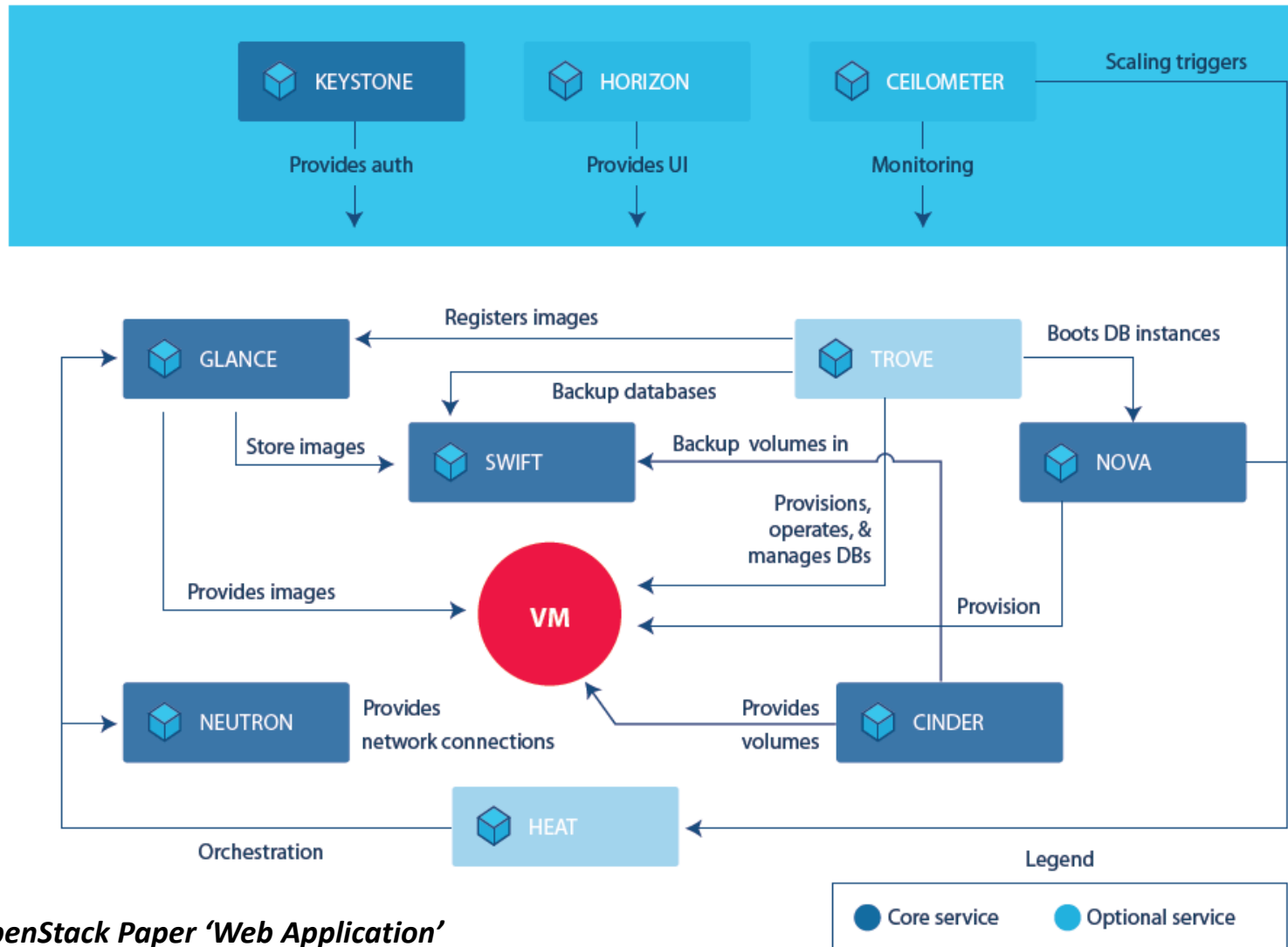
[7] OpenStack Paper 'Web Application'

Case Studies & Sample Configurations

- **OpenStack Sample Configurations are based on real-world reference architectures across selected industries with different types of workloads**
- **OpenStack case studies with configurations provide a way to understand which core and optional cloud services are used for different commercial and academic environments**

	Identity (keystone)	Networking (neutron)	Compute (nova)	Object Storage (swift)	Image Service (glance)	Block Storage (cinder)
Video Processing & Content Delivery	yes	yes	yes	yes		
Web applications	yes	yes	yes	yes	yes	yes
Big Data	yes	yes	yes		yes	yes
eCommerce	yes	yes	yes		yes	yes
High Throughput Computing	yes		yes		yes	yes

Video Processing & Content Delivery – Business Case

- **Real-world** OpenStack Example in Entertainment Industry
    - Leading post-production company
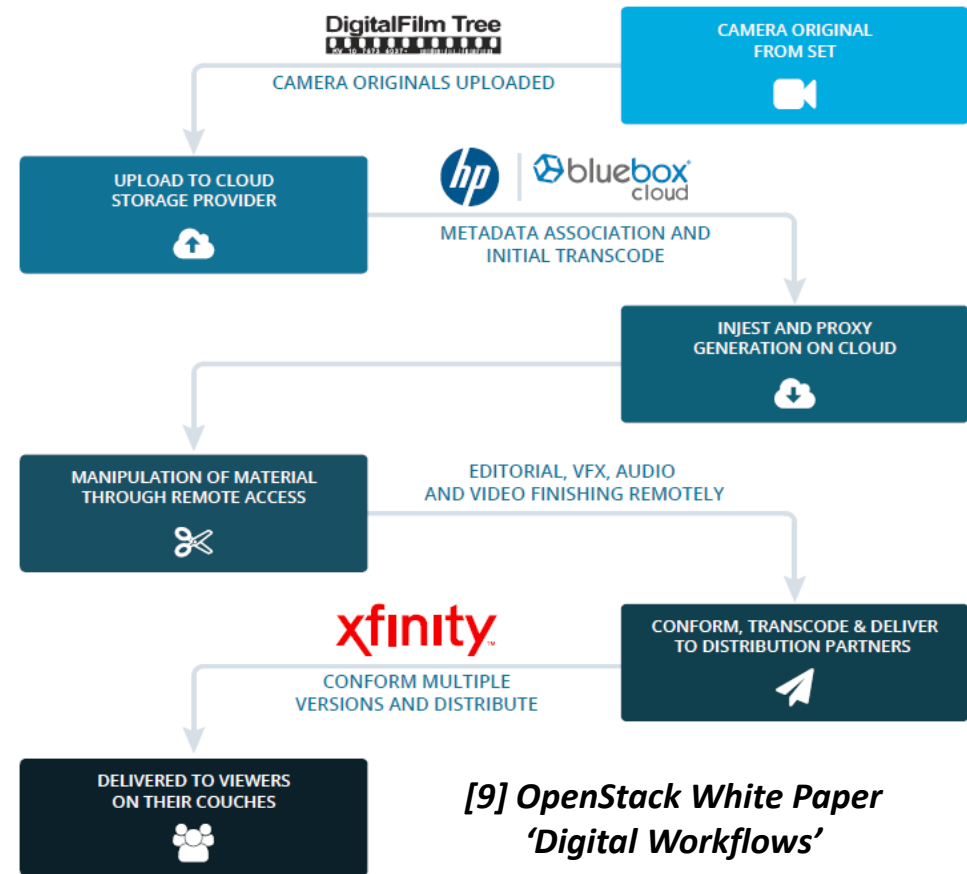   for television and film   *[8] Web page*

- **Business Goals**
    - Moves digital footage through a workstream that includes initial filming, editing, special effects creation, post-production, and review
    - Workstream is a 'digital workflow' across many collaborators
    - Interoperability among OpenStack clouds enables DigitalFilm Tree workloads to flow seamlessly using the resources needed at each stage
    - Leverage public and hybrid clouds to adapt to change at will in workflows
    - Security from end to end with a single point of identity control

- **OpenStack provides services used for video processing and content delivery that enable the DigitalFilm Tree company to scale on demand (e.g. scale for a 2-3 month period for the pilot season, then scale back down, some series are larger than others with larger formats like 4k/8k video, etc.)**
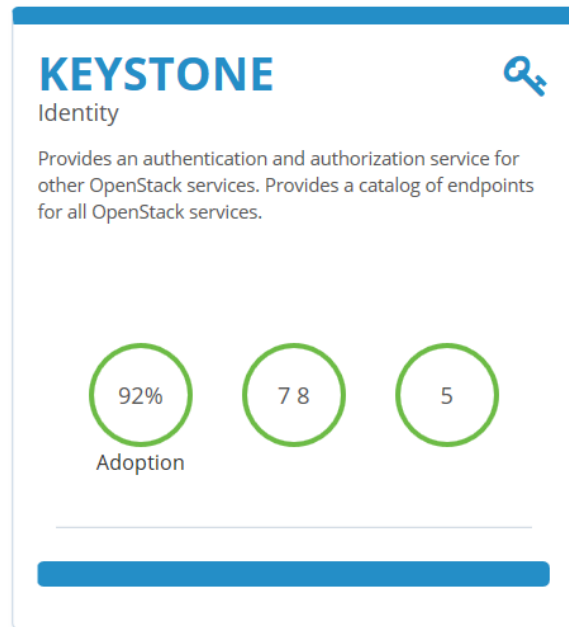
Video Processing & Content Delivery – Workflow

- Workflow: 'From Camera to Couch'

  - Video processing and content delivery across regions

  - Seamlessly move video content across private and public clouds

  - Use the same dashboard and security authentication infrastructure all the time

  - Delivery of the application service for viewers depends on many contributors from multiple companies

  - Forms 'end-to-end workflow' with storage, computing, networking and security

DigitalFilm Tree

CAMERA ORIGINAL FROM SET

CAMERA ORIGINALS UPLOADED

UPLOAD TO CLOUD STORAGE PROVIDER

hp | bluebox cloud

METADATA ASSOCIATION AND INITIAL TRANSCODE

INJEST AND PROXY GENERATION ON CLOUD

MANIPULATION OF MATERIAL THROUGH REMOTE ACCESS

EDITORIAL, VFX, AUDIO AND VIDEO FINISHING REMOTELY

xfinity

CONFORM, TRANSCODE & DELIVER TO DISTRIBUTION PARTNERS

CONFORM MULTIPLE VERSIONS AND DISTRIBUTE

DELIVERED TO VIEWERS ON THEIR COUCHES

[9] OpenStack White Paper 'Digital Workflows'

Federated Identity with OpenStack Keystone

KEYSTONE
Identity

Provides an authentication and authorization service for other OpenStack services. Provides a catalog of endpoints for all OpenStack services.

92%   7 8   5

Adoption

(enables 'single sign-on')

- Identity provider
  - Central source of identity and identities stay the same across multiple clouds
  - E.g. DigitalFilm Tree maintains the identity for each employee, vendor, or contractor

- Identity consumer
  - E.g. cloud service provider HP Cloud
  - Keystone confirms incoming requests by validating a set of user credentials (userid and password) with the identity provider

- Give access to applications, resources and/or content once credentials are approved

[9] OpenStack White Paper 'Digital Workflows'

> - **OpenStack Keystone provides IAM: Identity (authentication) & Access Management (authorization)**
> - **Keystone includes at least two parties: one identity provider & one-more identity consumers**
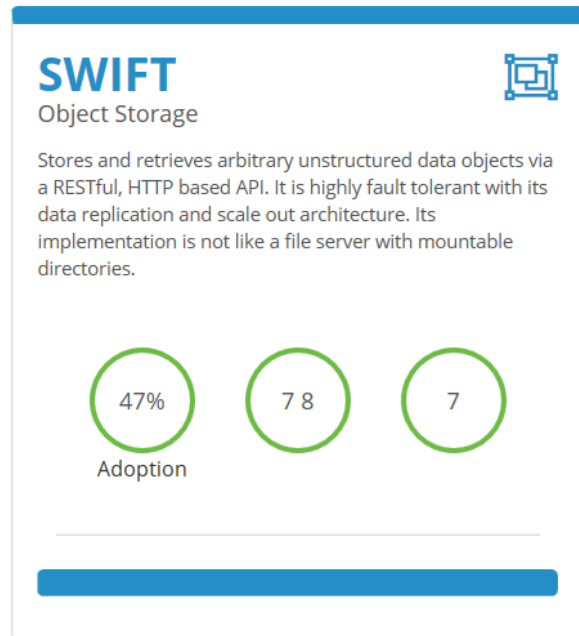
Video Processing & Content Delivery – Involved Services

- Using OpenStack Services
  - Federated identity across many OpenStack clouds with Keystone
    - Footage is uploaded to a Swift container in a data center in Vancouver
    - Data in Swift is replicate geographic regions whe can be accessed locally b production staff in Vancouver & Los Angeles
  - Nova computes the transcode & rendering processes in HP & Blue Box clouds for scale & performance

  [9] OpenStack White Paper 'Digital Workflows'
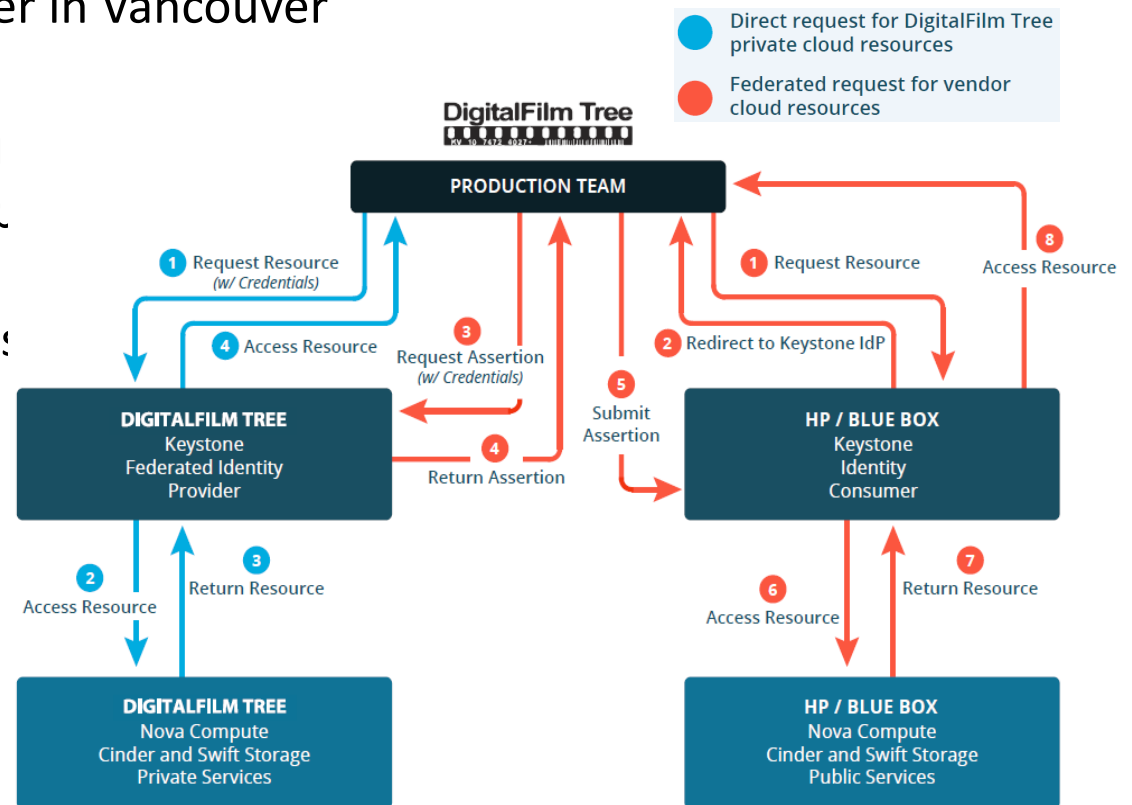
Object Storage with OpenStack Swift

SWIFT
Object Storage

Stores and retrieves arbitrary unstructured data objects via a RESTful, HTTP based API. It is highly fault tolerant with its data replication and scale out architecture. Its implementation is not like a file server with mountable directories.

47%
Adoption

7 8

7

(enables object storage in form of Object = data, metadata, unique ID)

- Case study
  - Data in Swift is replicated across geographic regions where it can be accessed locally by production staff in Vancouver & Los Angeles
- Powerful & robust storage service
  - High durability and high availability
  - Concurrency accross the entire data sets
- Examples
  - E.g. storing unstructured big data that significantly grows on a daily basis

[10] Big Data Tips 'OpenStack Swift'

[9] OpenStack White Paper 'Digital Workflows'

- **OpenStack Swift is an object storage that is a cost effective storage with properties to scale-out**
- **Swift includes features such as high fault tolerance since its architecture includes data replication**

Web Applications – Business Cases

- **Real-world** OpenStack Examples   *[7] OpenStack Paper 'Web Application'*

    - Workday **(finance and human resource system)**

    - Betfair **(online betting)**

    - Ancestry.com **(family trees & history records)**

    - JFE steel **(large steel company in Japan)**

    - Her Majestry's Revenue and Customs **(UK Government)**
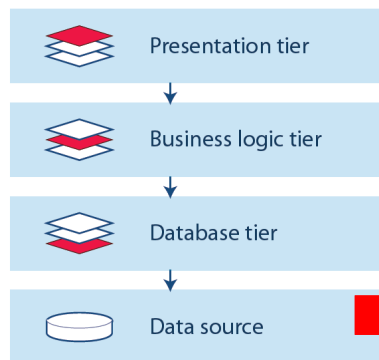
    - LivePerson **(Life chat software)**

- **Business Goals**

    - **Interact** with employees, customers & partners online, using applications such as online banking, human resources, tax filing, pet adoption, etc.

    - Build a **three-tier web application** as an OpenStack-based cloud

> - **OpenStack provides services used for interactive Web applications at scale providing IT resources needs for Web applications that often fluctuate with user demand (predictably or unpredictably)**
> - **OpenStack can dynamically add/remove resources that impacts customer satisfaction and sales**

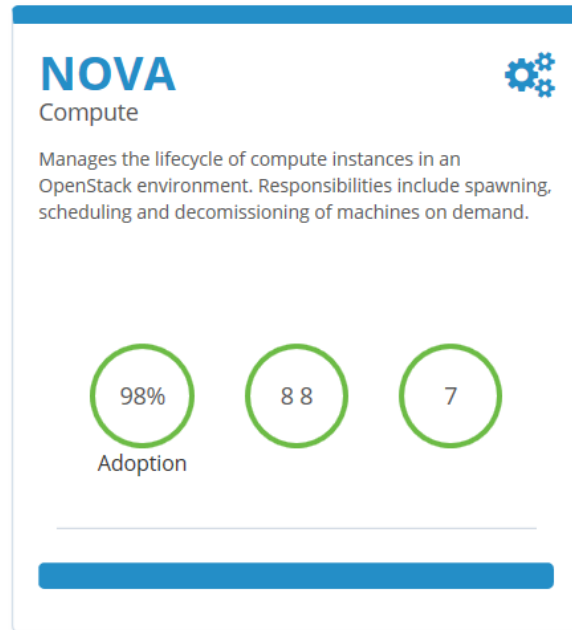# Web Applications – Three-Tier Reference Architecture

- Web presentation tier (Nova, Neutron)
  - Cluster of web server used to render content for the web browser

- Application tier (Nova, Neutron)
  - Cluster of application servers
    used to process content & business logic

- Database tier
  (Nova, Neutron, Cinder, Swift)
  - Cluster of database servers used
    to store data persistently

modified from [7]
OpenStack Paper
'Web Application'
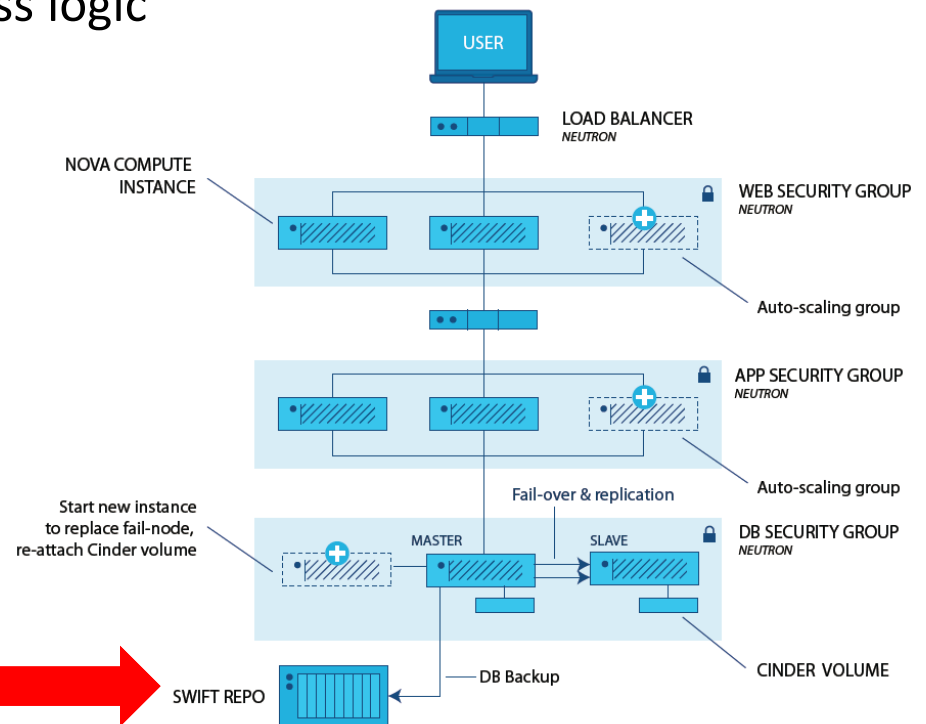
Compute & Processing with OpenStack Nova

NOVA
Compute

Manages the lifecycle of compute instances in an OpenStack environment. Responsibilities include spawning, scheduling and decomissioning of machines on demand.

98%
Adoption

8 8

7

(performs VM instance provisioning in combination with OpenStack Glance)

- Case study
  - E.g. host Web application server with business logic as a pay-per-use service for customers
- Powerful computing service
  - Manages and automates pools of computer resources using virtualization (cf. Lecture 3)
  - Also configurable for bare metal and High Performance Computing (HPC), cf. Lecture 2
- Examples
  - Perform big data analytics (cf. Lecture 1) processing jobs that scale with big data

[4] OpenStack Web page

- **OpenStack Nova provides massively scalable, on demand, and self service access to pools of compute resources & manages them using known virtualization software (e.g. KVM, Vmware, XEN)**

eCommerce – Business Cases

- **Real-world** OpenStack Examples

  - eBay **(consumer-to-consumer & business-to-consumer services)**

  - Walmart **(multinational retailing corporation)**

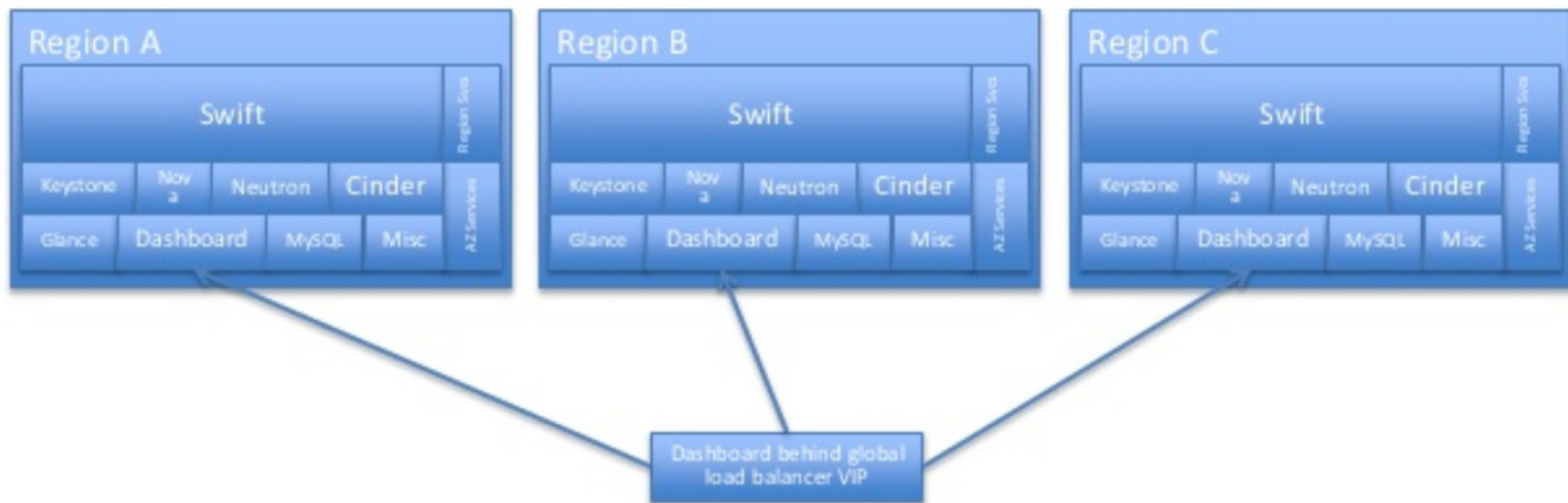 - Best Buy **(multinational consumer electronics)**

- **Business Goals**

  - Use a multi-tenant, multi-region, self-service cloud that hosts the company's customer-facing business-critical applications
    (e.g. eBay OpenStack-based private cloud hosts eBay key website)

  - Use a cloud for platforms the developers use for software development

  - Provide application software with a distributed architecture for a high degree of resiliency and reliabity including powerful databases

[12] Computerweekly eBay & OpenStack

[13] Walmartlabs on OpenStack

- **OpenStack provides services used as backends for eCommerce at scale via flexible API & dashboard**

- **OpenStack service and APIs enable to adapt to fast-moving customer demands including agile development for product managers to think outside the box and to iterate over product ideas**
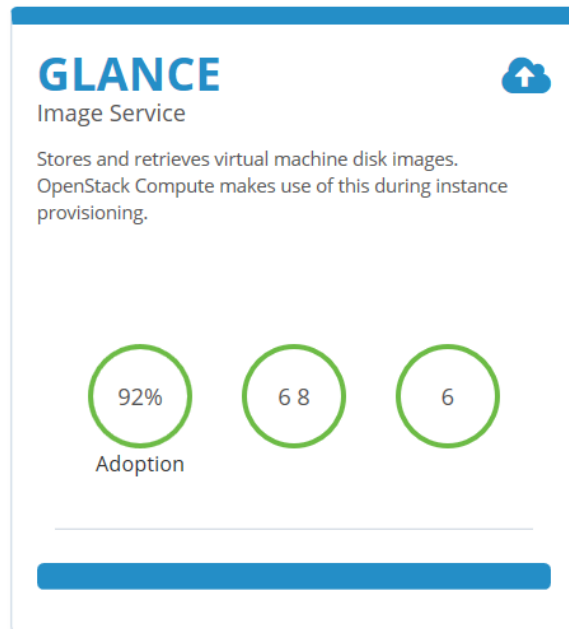
eCommerce – eBay Example

- ### Region-based deployments

  - Numerous OpenStack services that work together

  - Swift, Keystone, Nova, Cinder, Glance, Dashboard/Horizon

  - Other services like mysql

[14] Production eBay

Virtual Image Service with OpenStack Glance

GLANCE
Image Service

Stores and retrieves virtual machine disk images.
OpenStack Compute makes use of this during instance
provisioning.

92%
Adoption

6 8

6

**(performs VM instance provisioning
in combination with OpenStack Nova)**

- Case study
  - Use a cloud for platforms the developers use for software development
- Flexible virtual image service
  - Discovering, registering, and retrieving of virtual machine (VM) images (cf. Lecture 3)
  - Add & querying VM image metadata
- Examples
  - E.g. storing pre-installed software development environments and integrated development environments like e.g. Eclipse as virtual image

[4] OpenStack Web page

- **OpenStack Glance provides a service where users can upload and discover data assets in the form of virtual images that are meant to be used with other services such as the Nova compute service**

Dashboard with OpenStack Horizon

[15] YouTube, 'OpenStack Newton Release Demo'

- ## Case study
  - eBay runs business-critical applications such as its Website or PayPal

- ## Flexible dashboards
  - Created/used dynamically w/o in-depth knowledge of the services
  - E.g. dashboards help to identify if there are heavy users causing trouble

- **OpenStack Horizon provides a Web-based self-service portal to interact with OpenStack services such as launching an instance, assigning IP addresses, configuring access controls, and monitoring**

OpenStack API & SDK

- OpenStack software development resources
  - Ensure that when a developer writes an application for one OpenStack cloud, it will work across other OpenStack private/public clouds
  - This includes workload and data portability as well

- Software Development Kit (SDK)
  - Used to create OpenStack cloud applications or to interact with clouds
  - Contains code, examples, documentation in many programming languages (e.g. Java, Python, Ruby, .NET, PHP, etc.)

- Application Programming Interface (API)
  - Access (remote) service capabilities through predefined functions
  - E.g. use OpenStack APIs to launch server instances, create images, assign metadata to instances and virtual images, create storage containers and objects, and complete other actions in OpenStack clouds for applications
  - OpenStack's REST APIs is particularly easy to learn and program against
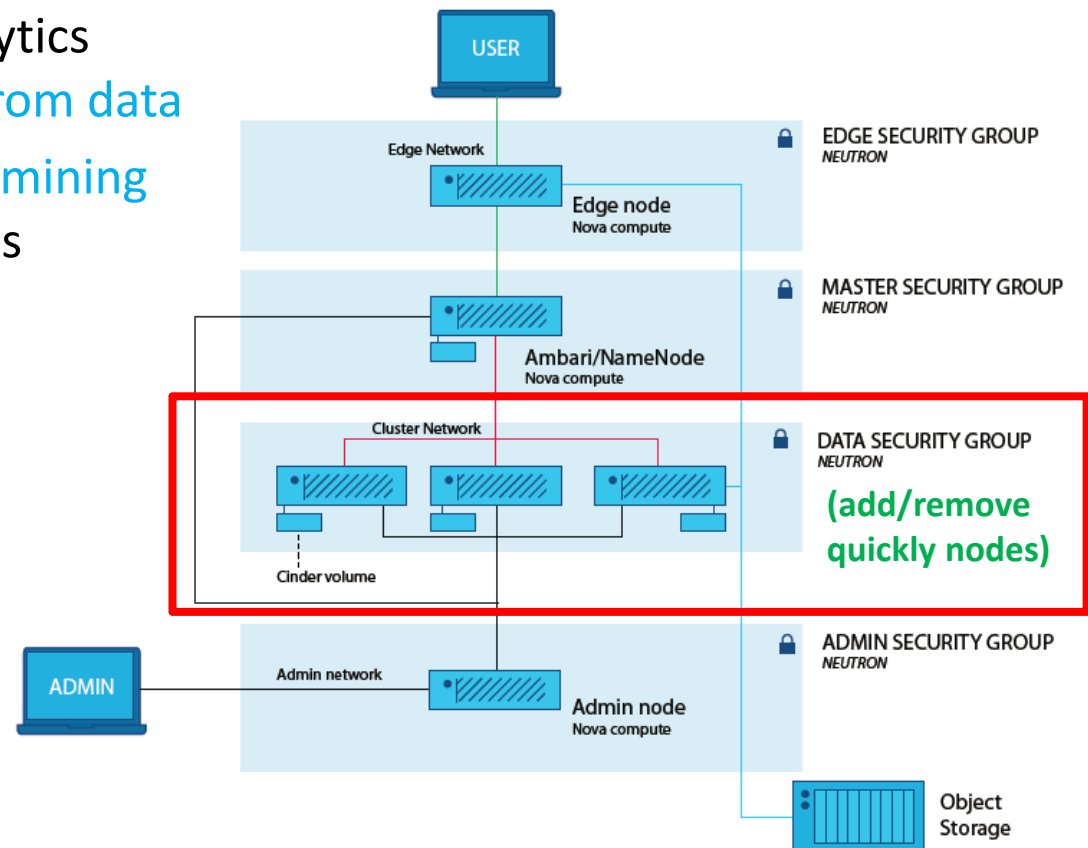
Handling Big Data with Scalability

- Scalable Big Data Analytics

  - Supports enhanced business models (e.g. understanding customers better)

  - Traditional data processing tools often inadequate to deal with 'big data'

  - E.g. use of predictive analytics in order to extract value from data

  - E.g. frequent pattern (FP) mining in customer shopping carts

- Need for scalability

  - Tackle dynamic situation caused by high volumes of analytics requests

  - Means rapid deployment time requirement in an automated manner
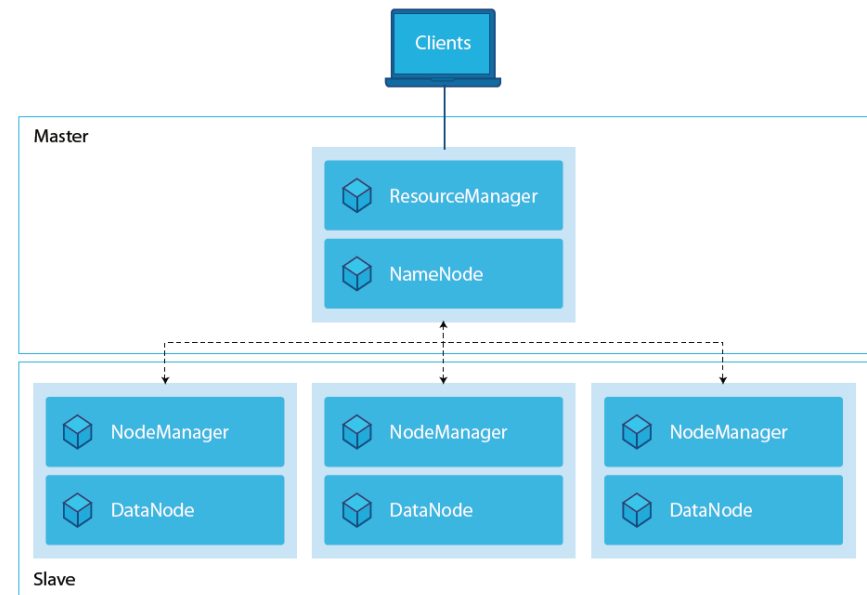
[16] OpenStack Paper 'Big Data'

USER

Edge Network

Edge node
Nova compute

🔒 EDGE SECURITY GROUP
NEUTRON

Ambari/NameNode
Nova compute

🔒 MASTER SECURITY GROUP
NEUTRON

Cluster Network

Cinder volume

🔒 DATA SECURITY GROUP
NEUTRON

(add/remove quickly nodes)

ADMIN

Admin network

Admin node
Nova compute

🔒 ADMIN SECURITY GROUP
NEUTRON

Object Storage

Big Data Analytics with Map-Reduce & Hadoop

- Apache Hadoop (based on map-reduce)   *[14] Apache Hadoop*

  - Big Data infrastructure that scales out both compute and storage resources

  - Provides secure and automated capabilities for virtual cluster deployment

  - Multiple Hadoop clusters are
    often deployed to respond to
    an enterprise's needs across
    different data centers
    (cf. Lecture 2)

- Apache Ambari   *[17] Apache Ambari*

  - Open source software to
    provision, manage and
   monitor Hadoop clusters

  **(OpenStack Sahara is an emerging service that provides
  simple means to provision and scale many Hadoop Cluster)**


[16] OpenStack Paper 'Big Data'

Big Data Analytics with Apache Hadoop on OpenStack

- **Resource Manager (uses Nova)**
  - Scheduler Yarn to allocate resources to various applications on the cluster

- **NameNode (uses Nova & Cinder)**
  - Metadata about the data blocks are stored in the NameNode
  - Provides lookup functionality and tracking for all data or files in the Hadoop cluster

- **NodeManager (uses Nova)**
  - Takes instructions from Yarn and responsible to execute and monitor applications

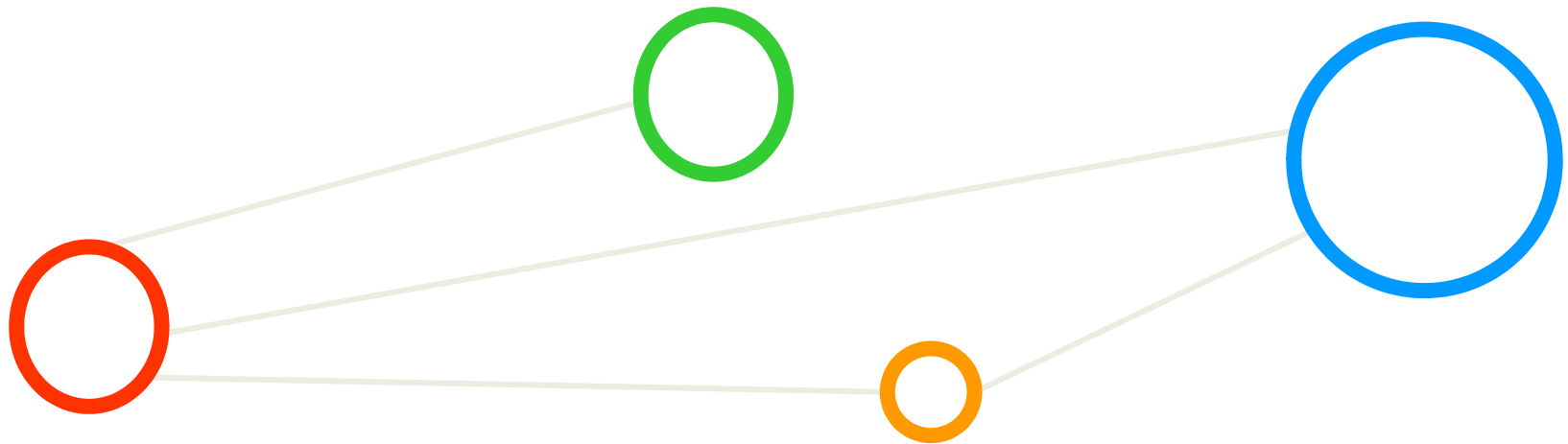- **Datanode (uses Nova & Cinder)**
  - Store and process the data
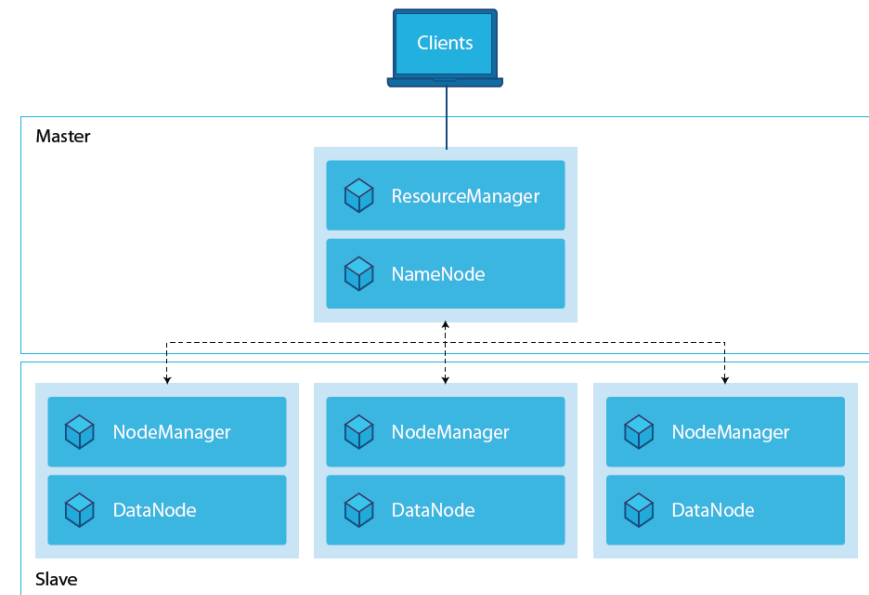
[16] OpenStack Paper 'Big Data'

[VIDEO] OpenStack Newton Release Demo
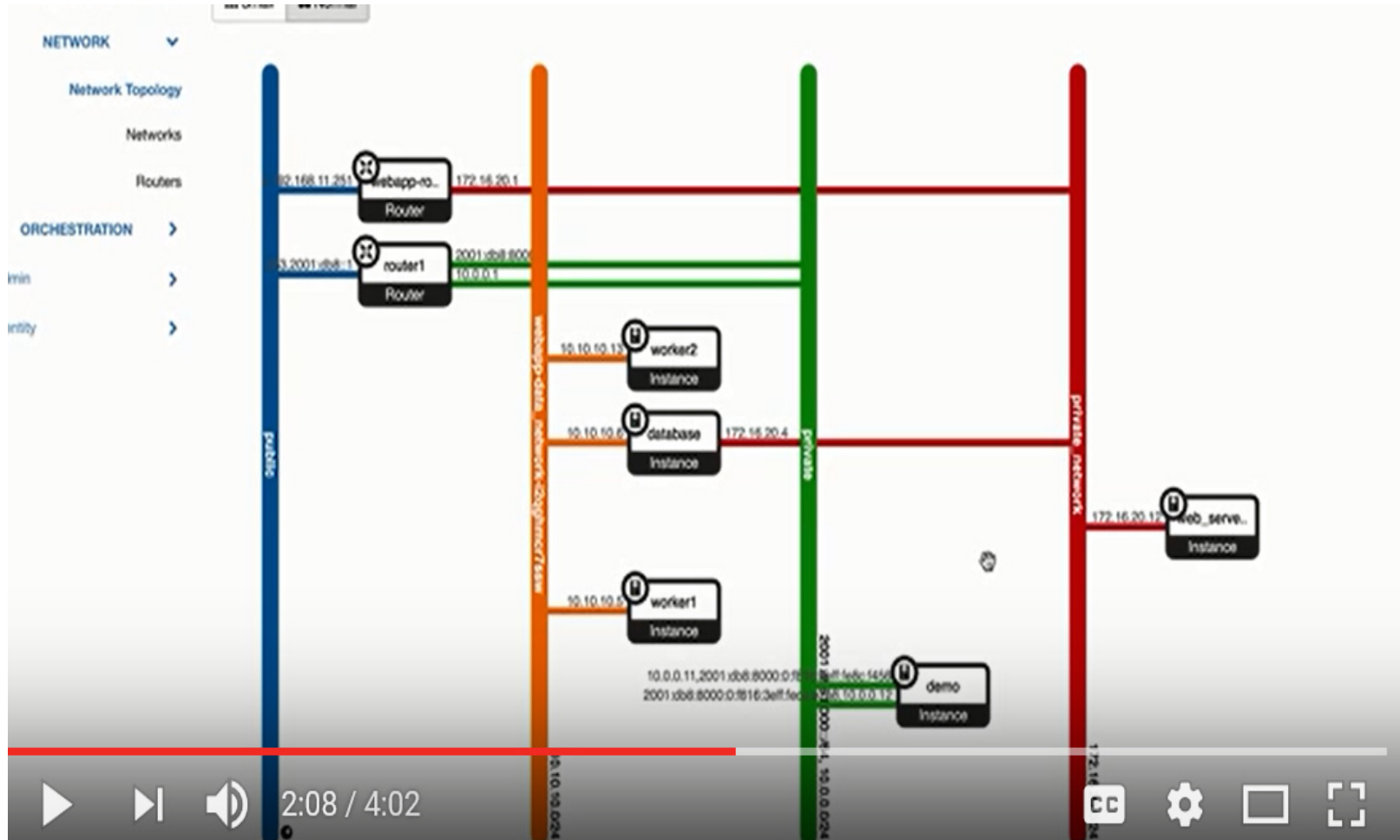
[11] YouTube, 'OpenStack Newton Release Demo'

Lecture Bibliography

Lecture Bibliography (1)

- [1] Docker Web page,
  Online: https://www.docker.com/

- [2] Kubernetes Web page,
  Online: https://kubernetes.io/

- [3] K. Hwang, G. C. Fox, J. J. Dongarra, 'Distributed and Cloud Computing', Book,
  Online: http://store.elsevier.com/product.jsp?locale=en_EU&isbn=9780128002049

- [4] OpenStack Web Page,
  Online: https://www.openstack.org/software/

- [5]Big Data Tips, 'OpenStack',
  Online: http://www.big-data.tips/openstack

- [6] Druva Blog, 'Object Storage versus Block Storage: Understanding the Technology Differences',
  Online: http://www.druva.com/blog/object-storage-versus-block-storage-understanding-technology-differences/

- [7] OpenStack Paper 'OpenStack Workload Reference Architecture: Web Applications',
  Online: https://www.openstack.org/assets/software/mitaka/OpenStack-WorkloadRefArchWebApps-v7.pdf

- [8] DigitalFilm Tree Company Web Page,
  Online: https://www.digitalfilmtree.com/

- [9] OpenStack White Paper 'Expediting Digital Workflow with OpenStack ',
  Online: https://www.openstack.org/assets/pdf-downloads/OpenStack-Workflow-White-Paper-Letter-Final.pdf

- [10] Big Data Tips, 'OpenStack Swift',
  Online: http://www.big-data.tips/openstack-swift

Lecture Bibliography (2)

- [11] YouTube Video, 'OpenStack Newton Release Demo',
  Online: https://www.youtube.com/watch?v=z6ftW7fUdp4

- [12] Computerweekly.com, 'Case study: How eBay uses ist own OpenStack private cloud ',
  Online: http://www.computerweekly.com/news/2240222899/Case-study-How-eBay-uses-its-own-OpenStack-private-cloud

- [13] Walmartlabs, 'Why we chose OpenStack for Walmart Global eCommerce',
  Online: http://www.walmartlabs.com/2015/02/why-we-chose-openstack-for-walmart-global-ecommerce/

- [14] Slideshare, 'OpenStack Trove in Production at eBay - Trove Day 2014',
  Online: http://www.slideshare.net/tesoracorp/4-open-stack-trove-day-ebay-final?qid=76eb7cb1-9d82-4ad3-a3c3-3ec6e6ab474e&v=&b=&from_search=1

- [15] YouTube Video, 'OpenStack Newton Release Demo',
  Online: https://www.youtube.com/watch?v=z6ftW7fUdp4

- [16] OpenStack Paper, 'OpenStack Workload Reference Architecture: Big Data',
  Online: https://www.openstack.org/assets/software/mitaka/OpenStack-WorkloadRefBigData-v4.pdf

- [17] Apache Ambari Web page,
  Online: https://ambari.apache.org/