GPFS-OpenStack Integration

Dinesh Subhraveti
IBM Research

Traditional shared architecture

Shared nothing architecture

SAN

I/O bottleneck

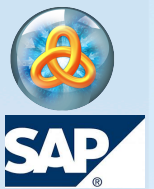Scale out performance

gpfs-san gpfs-snc

(chart values: Seq write, Seq read, Random Read, Random Write, Backward Read, Record rewrite, Stride Read)
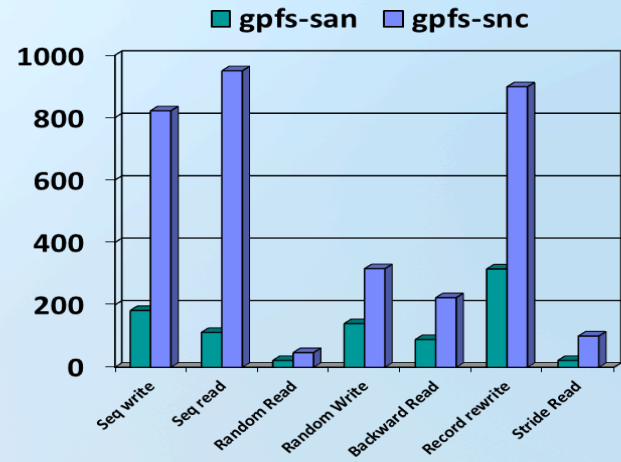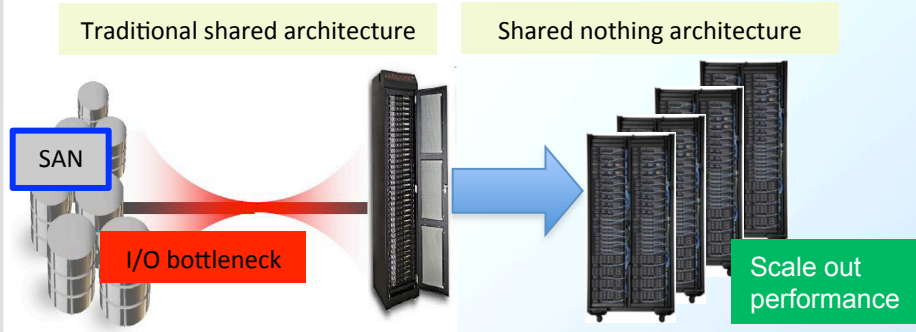
- GPFS cluster file system, enhanced with industry-leading features for shared nothing architectures.

- Winner of Supercomputing 2010 Storage Challenge

eWEEK.COM

Analytics Speeds Doubled By IBM Storage Architecture

November 22, 2010 by Darryl K. Taft

SC10 New Orleans, LA

- In production for over a year
  - GA in GPFS since Q4 2012
  - Infosphere BigInsights
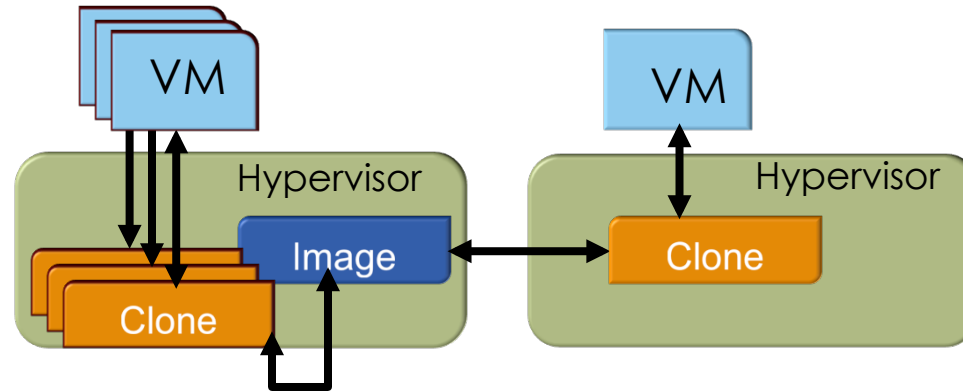 - SAP HANA appliance

SAP

GPFS User Group

Outline

- **GPFS-FPO features as they relate to cloud scenarios**

- **OpenStack overview**

- **GPFS-FPO integration with OpenStack components**

- **Demo**

*Disclaimer: Contains references to exploratory future work*

File System Level Clones

- Independently dictate the placement of master images and their clones

- Clone can be on a different node
  - Writes remain local
  - Reads are streamed

- Different replication level for the cloned instances

- Automatic dedup – both storage and in-memory

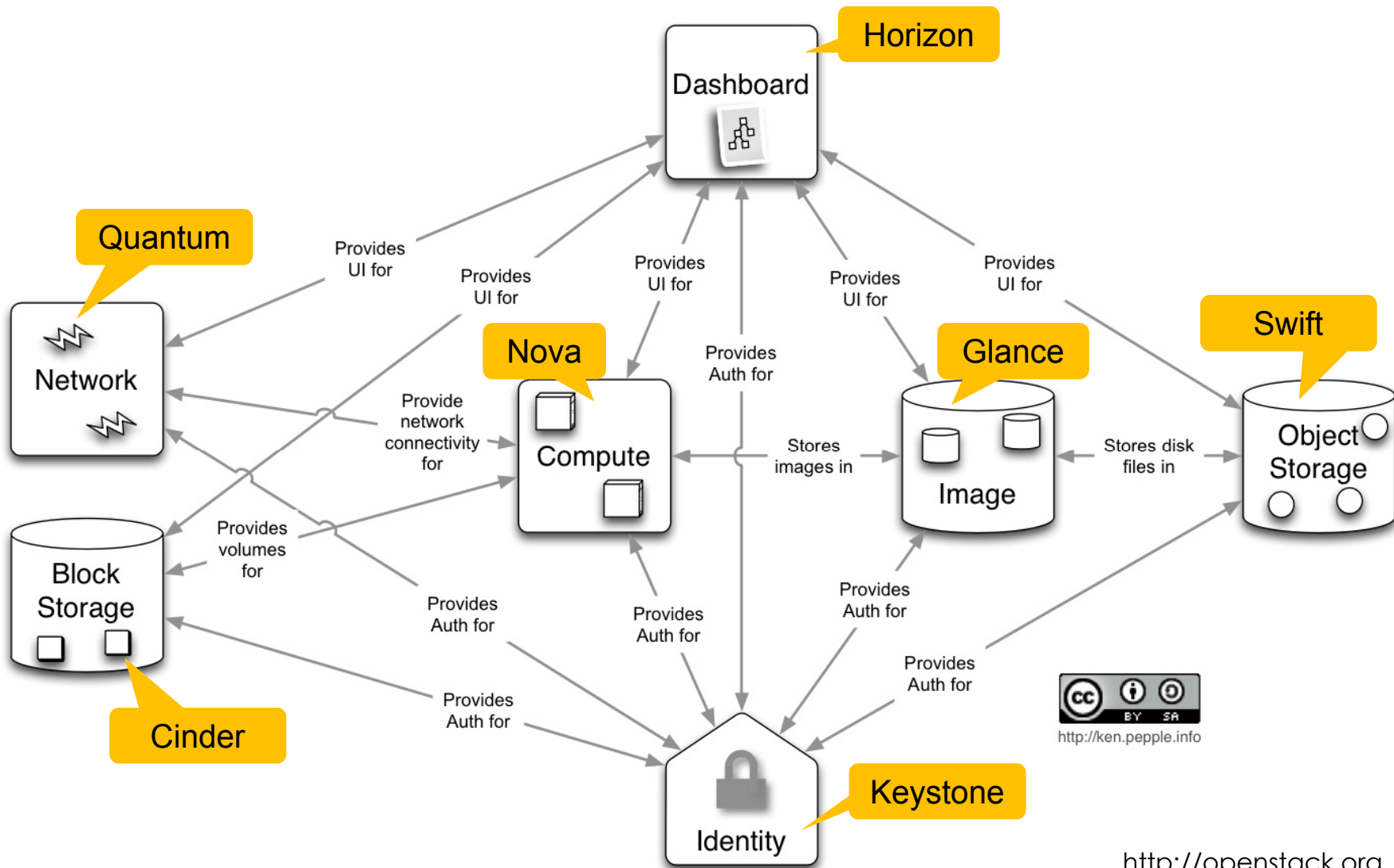- Format and hypervisor agnostic`

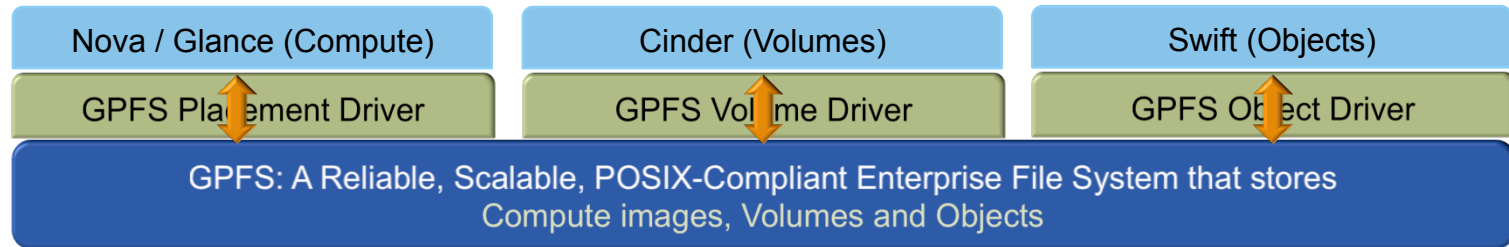# Transparent Block Level Replication and Recovery
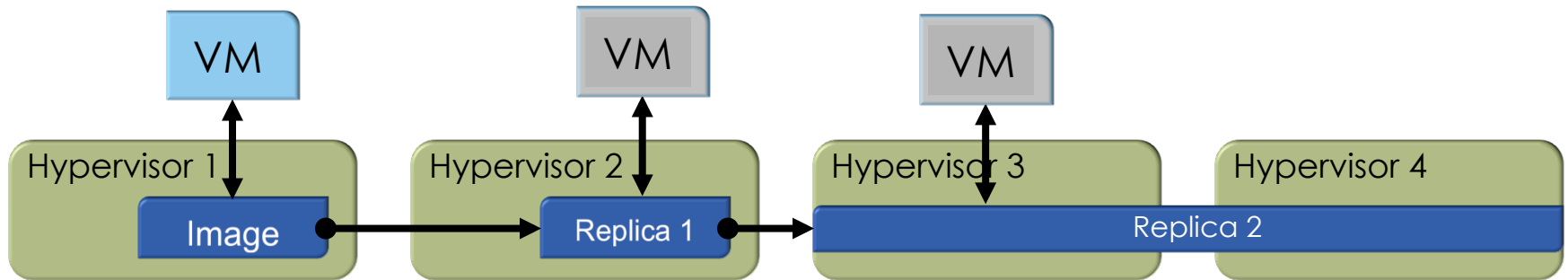
- **Pipelined replication**

- **Configurable file-level replication**
  - Store first replica on Hypervisor 2, stripe replica 2 on Hypervisors 3 and 4
  - Replicate popular master images at a higher level

- **Transparent failover**

- **Distributed restripe**

OpenStack Architecture Overview

http://openstack.org

GPFS As the Enterprise Storage Layer for OpenStack

Nova / Glance (Compute)	Cinder (Volumes)	Swift (Objects)
GPFS Placement Driver	GPFS Volume Driver	GPFS Object Driver

GPFS: A Reliable, Scalable, POSIX-Compliant Enterprise File System that stores Compute images, Volumes and Objects

- A common storage layer for images, volumes and objects
  - Avoids data copy
  - Local access of data

- Adds enterprise storage management features to OpenStack
  - Rapid volume and virtual machine provisioning using file system level copy-on-write function
  - Scale out IO performance through support for shared nothing clusters
  - Resilient volumes through transparent pipeline replication

Hypervisor Node | Storage Node

Nova	Cinder
Libvirt Volume Driver	GPFS Driver

GPFS
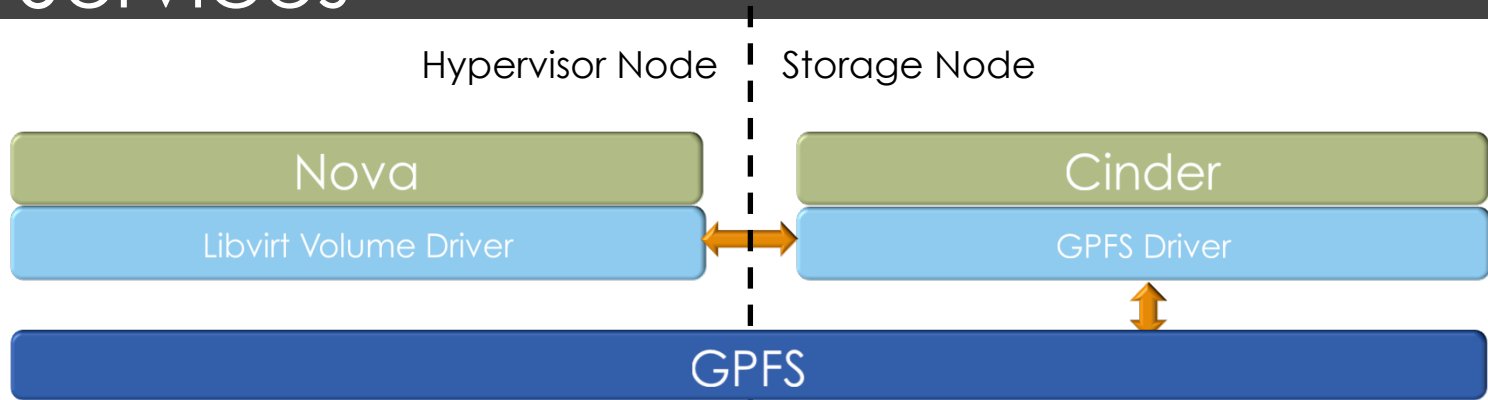
Cinder driver interface points: Create, Delete, Attach and Detach Volumes, Create Snapshot, Create volume from snapshot, Clone a volume
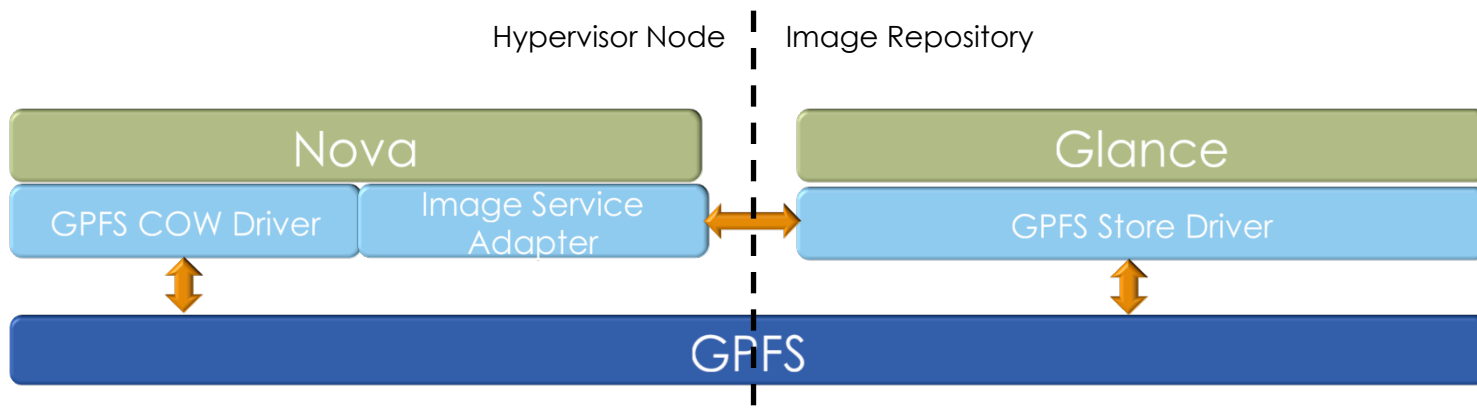
Create Volume
1. User initiates volume creation through Horizon or command line
2. The request ends up as a call to the GPFS driver
3. GPFS driver creates a (sparse) file and sets the right placement based on user parameters and policy. This operation is "instantaneous".
4. Create from snapshot is similar, except it uses GPFS COW mechanism.

Attach Volume
1. User initiates volume attach by selecting the volume and the virtual machine
2. The virtual machine and the volume objects are passed to the Nova Driver which passes them to Libvirt volume driver
3. The Libvrt volume driver invokes Libvirt interface to attach the volume file on GPFS mount point to the specified virtual machine

Hypervisor Node | Image Repository

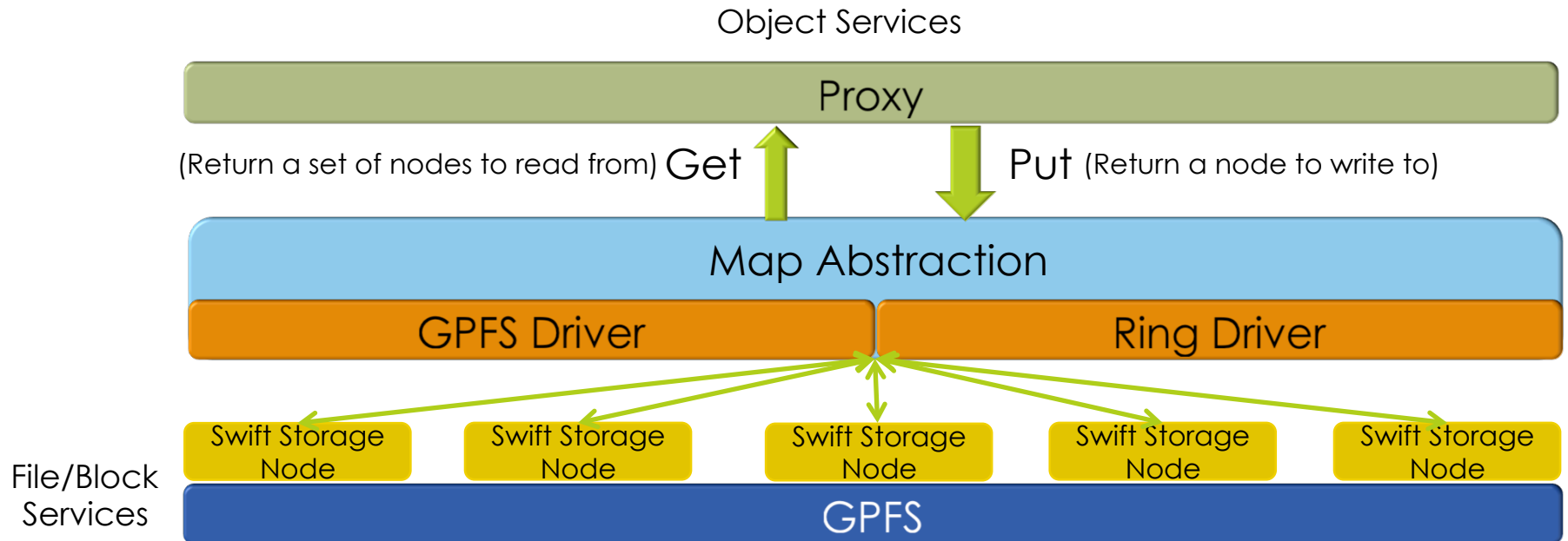| Nova	Glance
GPFS COW Driver	Image Service Adapter	GPFS Store Driver

GPFS

- **GPFS COW Driver:** Implements image creation and caching functions. Interfaces with GPFS to create instance images for provisioning.

- **Image Service Adapter:** Implements the interface with the image repository including image transport.

- **GPFS Store Driver:** Extends the Store base class to implement GPFS specific functions.

Instance Deployment

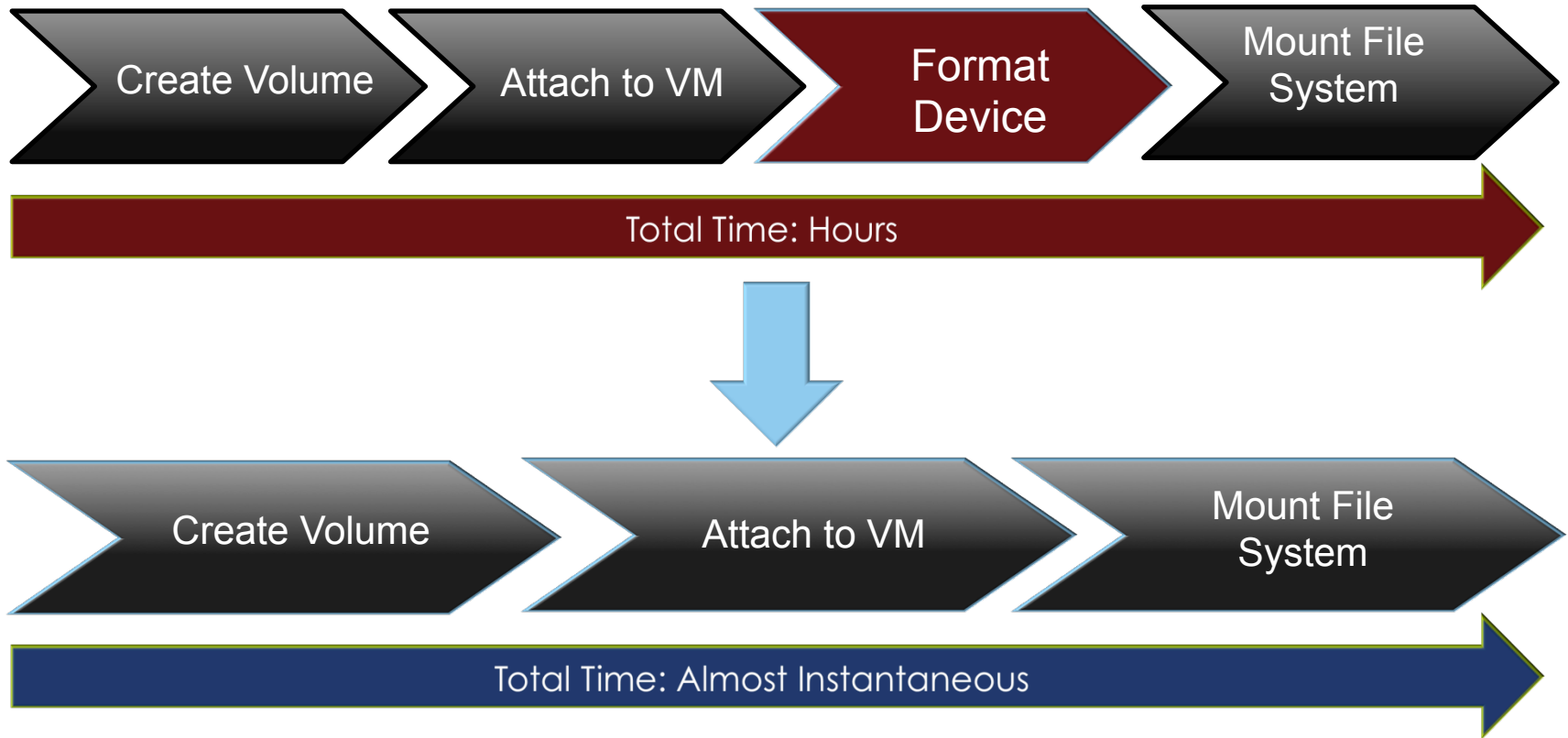1. User selects an image for instance creation – this refers to an image via the GPFS Store Driver

2. The Nova Image Service adapter creates a link from image to hypervisor cache in the same file system

3. Nova clones the image using the COW Driver for GPFS and creates a VM image instance in a per-hypervisor cache in the same file system.

4. Libvirt uses the VM image instance to deploy the VM

OpenStack Integration: Object Store

Object Services

Proxy

(Return a set of nodes to read from) Get ⬆ ⬇ Put (Return a node to write to)

Map Abstraction

| GPFS Driver | Ring Driver |

Swift Storage Node   Swift Storage Node   Swift Storage Node   Swift Storage Node   Swift Storage Node

File/Block Services

GPFS

- The robustness and features of GPFS combined with object extensions provide an enterprise-grade object store
  – High storage efficiency, Tape integration, Wide-area replication
  – Encryption, Transparent tiering, Checksums, Snapshots, ACLs
- These advanced enterprise features are available today

Demo: Quick Volume Provisioning

Create Volume → Attach to VM → **Format Device** → Mount File System

Total Time: Hours

Create Volume → Attach to VM → Mount File System

Total Time: Almost Instantaneous

Skip expensive formatting time through copy-on-write snapshot templates

http://almaden.ibm.com