VM Images management: From Cloud to Fog

Jad Darrous
PhD candidate
Gilles Fedak, Shadi Ibrahim

March 31, 2017

ENS Lyon, INRIA, LIP, AVALON

Table of contents

Introduction

VM image management

A VM image management software is responsible for discovering, registering, and retrieving VM images.

[1] https://docs.openstack.org/developer/glance/architecture.html

VM image management

A VM image management software is responsible for discovering, registering, and retrieving VM images.

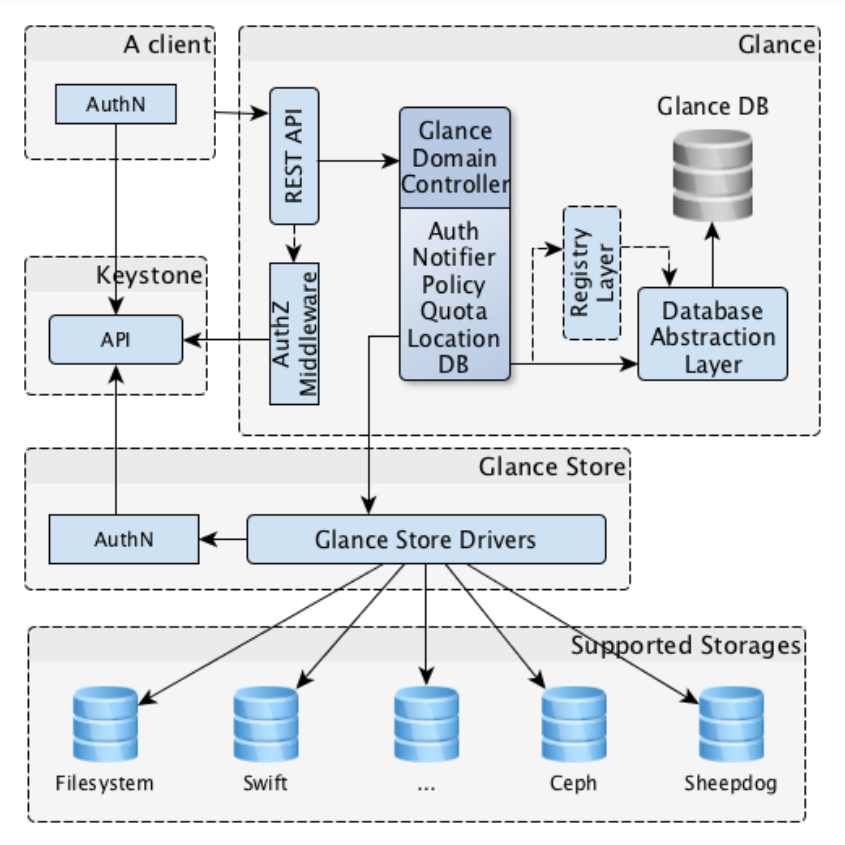

Figure 1: Glance architecture [1]

[1] https://docs.openstack.org/developer/glance/architecture.html

VM image management

The two important components are:

- **The Catalog i.e. DB:**
  - Stores images' Metadata;
  - Reads queries are the dominant;
  - Small size compared to nova DB (e.g. around 6000 public VM images at Amazon [4]).
- **The back-end storage:**
  - Stores the images;
  - Retrieve queries are the dominant;
  - Its performance is crucial for VM provisioning.

VM Image management

VM image management

- **The goal:** Efficient provisioning
- **The problem:** Huge size of VMIs (dozens of GBs)
- **The good news:** High similarities (up to 80% [5, 4])
- **Optimization:** Deduplication techniques

Intra Datacenter

Intra Datacenter

VM Images could be stored in

	Object storage	**Block storage**
OpenStack/Amazon	Swift/S3	Cinder/EBS
Protocol	HTTP	iSCSI, FC or NFS
Provisioning time	Minutes	Seconds
Durability	Ephemeral	Persistent
Could be suspended	No	Yes
Performance	Stable	Could be affected by others
Price (Amazon)	Per size (cheaper)	Per size and IO operations

Table 1: Back-end storage

Intra Datacenter

VM Images could be stored in

	Object storage	**Block storage**
OpenStack/Amazon	Swift/S3	Cinder/EBS
Protocol	HTTP	iSCSI, FC or NFS
Provisioning time	Minutes	Seconds
Durability	Ephemeral	Persistent
Could be suspended	No	Yes
Performance	Stable	Could be affected by others
Price (Amazon)	Per size (cheaper)	Per size and IO operations

Table 1: Back-end storage

Which one is better?

- It depends on the workload...

Intra Datacenter

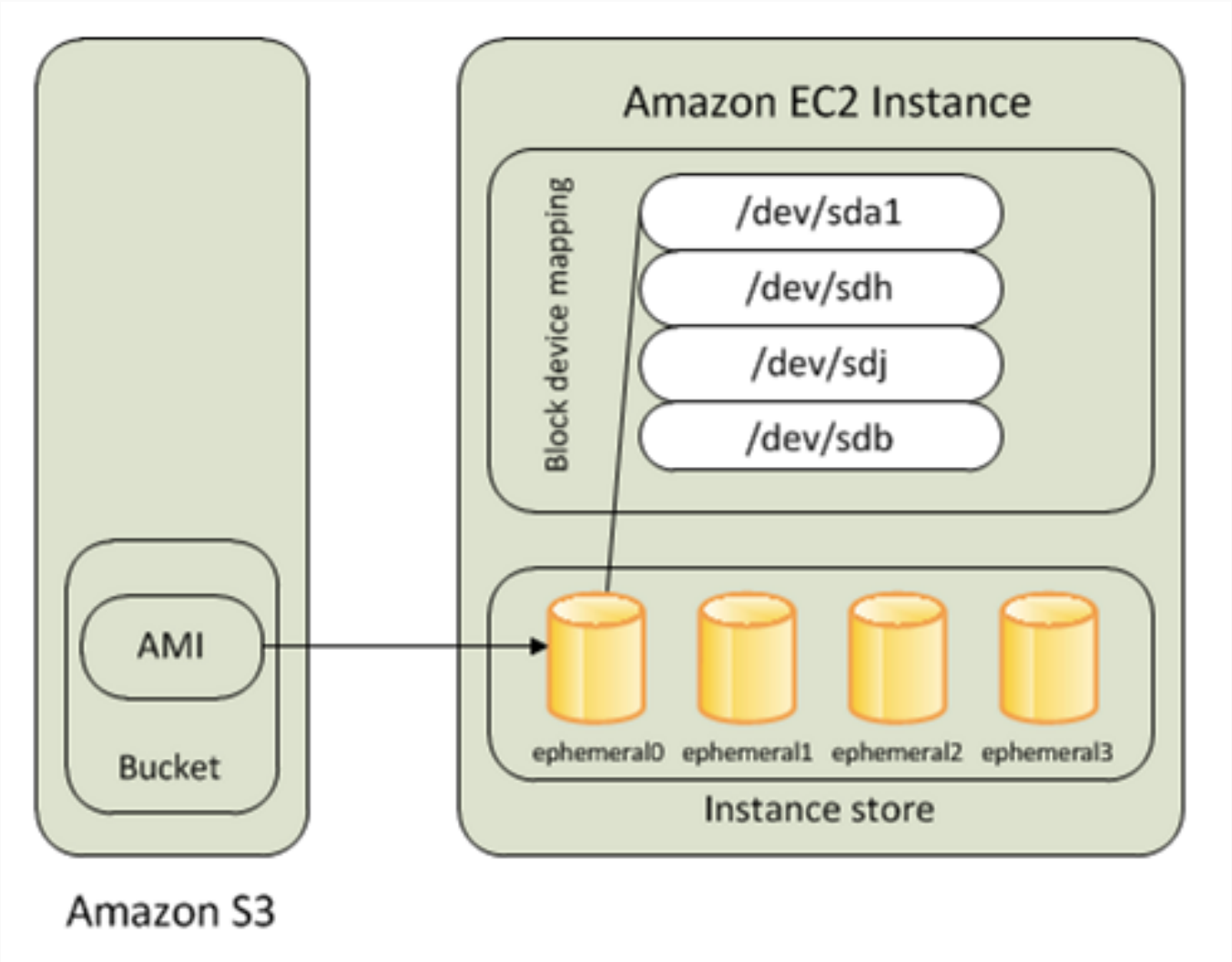

Figure 2: Object storage

Intra Datacenter

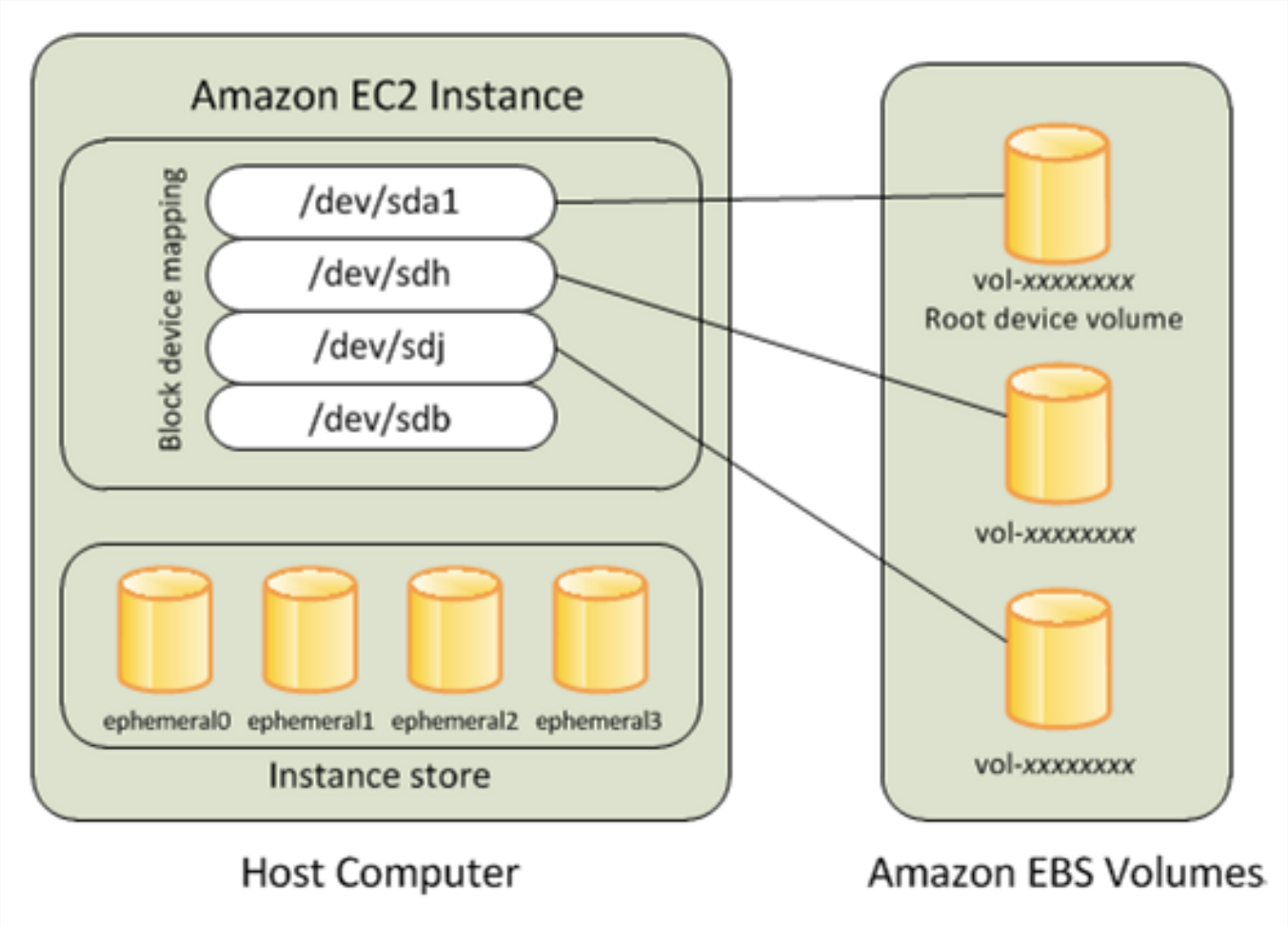

Figure 3: Block storage

Deduplication

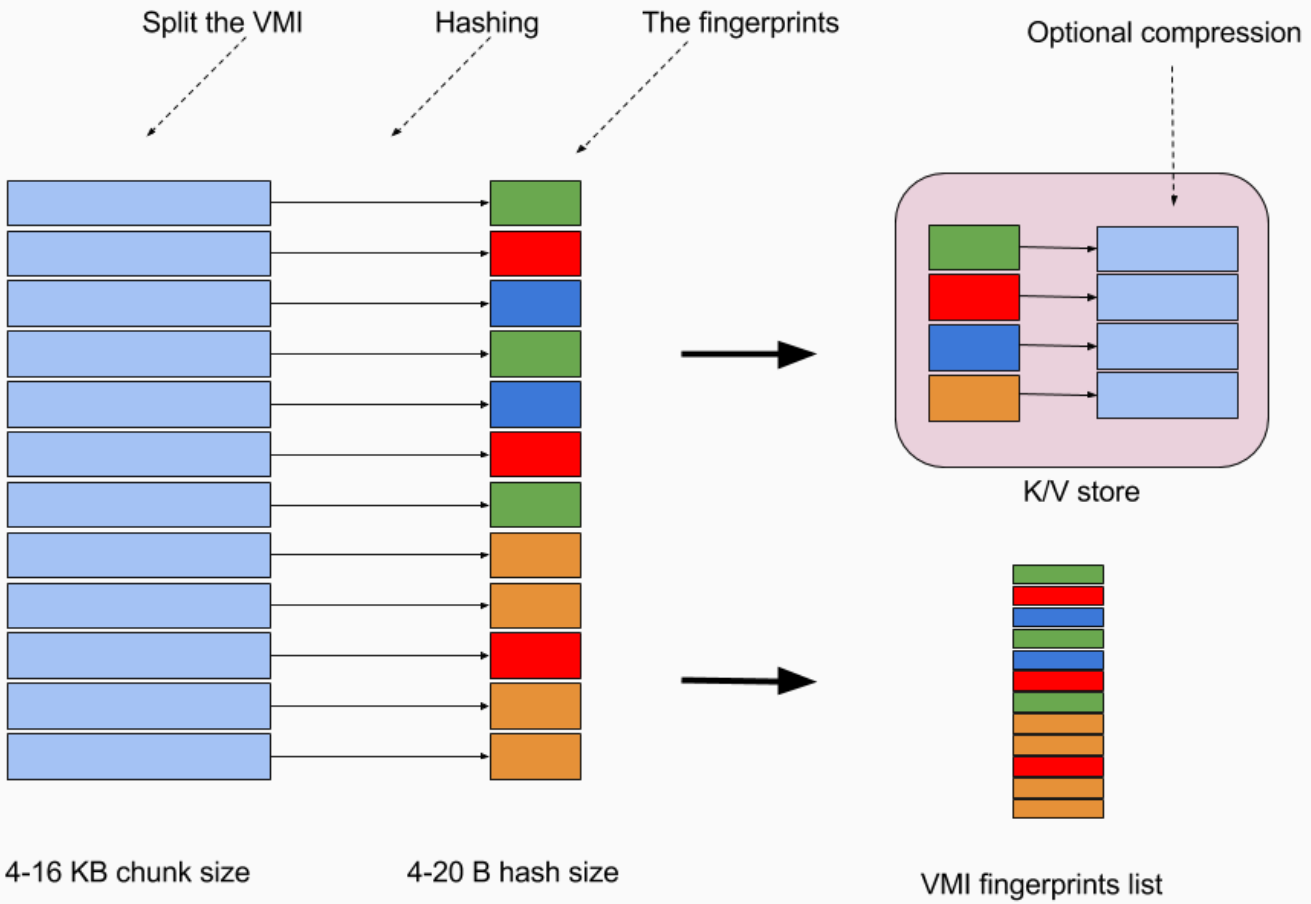

Figure 4: Deduplication mechanism

Intra Datacenter

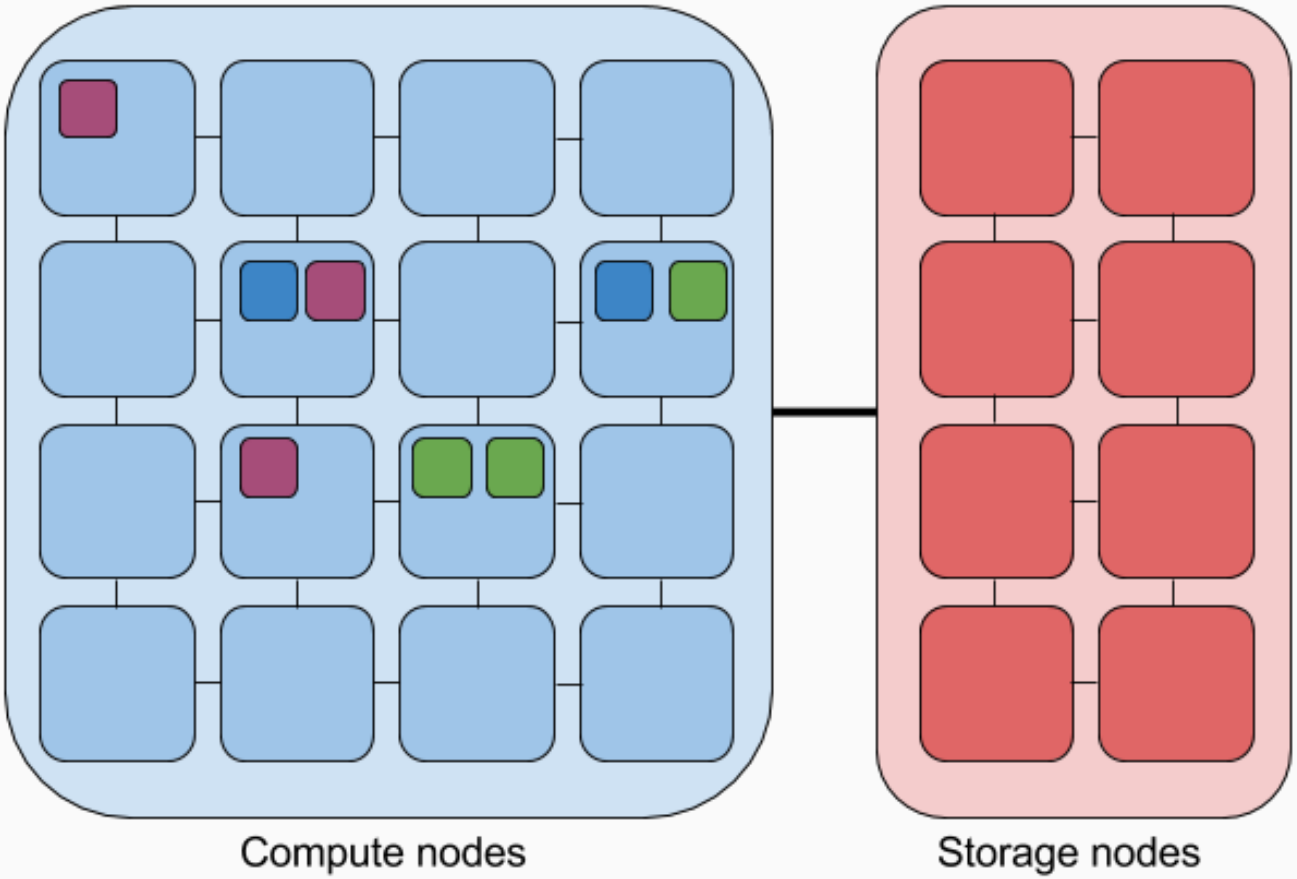

Figure 5: Datacenter schema

Intra Datacenter

Going further:

- Peer-to-peer sharing [8, 11]
  - Take network topology into account [7, 10]
- Scheduling friendly algorithms [7]
- Enhanced image formats [9, 2]

Intra Datacenter

Is it actually used?

- IBM: Probably YES.
- OpenStack: NO. (bittorrent blueprints, but not accepted)
- Amazon, Google, Microsoft: No idea!

Geo-distributed Datacenters

Geo-distributed Datacenters

Challenges:

- High WAN latency (Up to 380 ms between two Amazon sites [6])
- Heterogeneous network bandwidth
- High probability of failures and network partitioning

Geo-distributed Datacenters

Can we use the same techniques (i.e. deduplication) to distribute the images?

- Yes

Does it works?

- Yes. We will see how soon...

IPFS

InterPlanetary File System

What it IPFS?

- Peer-to-peer distributed file system
- Inspired by BitTorrent
- Provides Content Addressing
- Tolerates network partitioning

IPFS

Why choosing IPFS?

- IPFS is completely distributed system with no SPOF.
- IPFS is massively scalable.
- IPFS is designed to work in heterogeneous network latency and bandwidth.
- IPFS provides deduplication by content addressing the data.

With some drawbacks...

Evaluation

Testbed:

- Grid5000 [1] testbed using the sagittaire cluster in Lyon.
- Each host has 2 CPUs AMD Opteron 250, 1 core/CPU, 2GB RAM, 68GB HDD, and a single Ethernet interface.
- The throughput of the network links is 1 Gbps.

Dataset:

- 8 Debian cloud images [2].
- combined size of 17 GB.

Methodology:

- One node load the data into the repository
- Other node pull the data consecutively.

[2] `http://cdimage.debian.org/cdimage/openstack/archive/`

Dataset

$$Compression\ ratio = 1 - \frac{|Unique\ chunks|}{|Non\ zero\ chunks|}$$

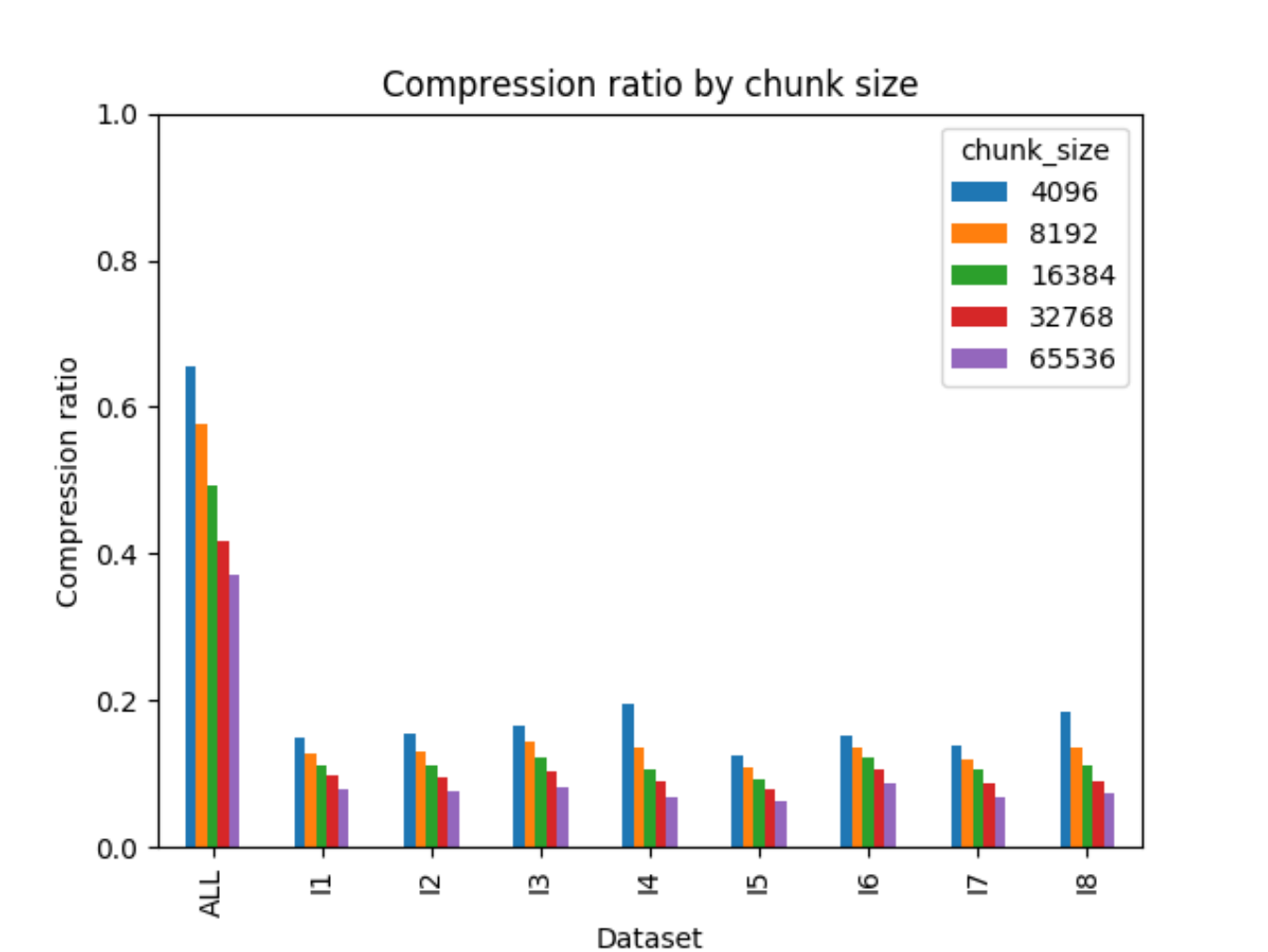

Figure 6: Compression ratio per chunk size

Experimental results

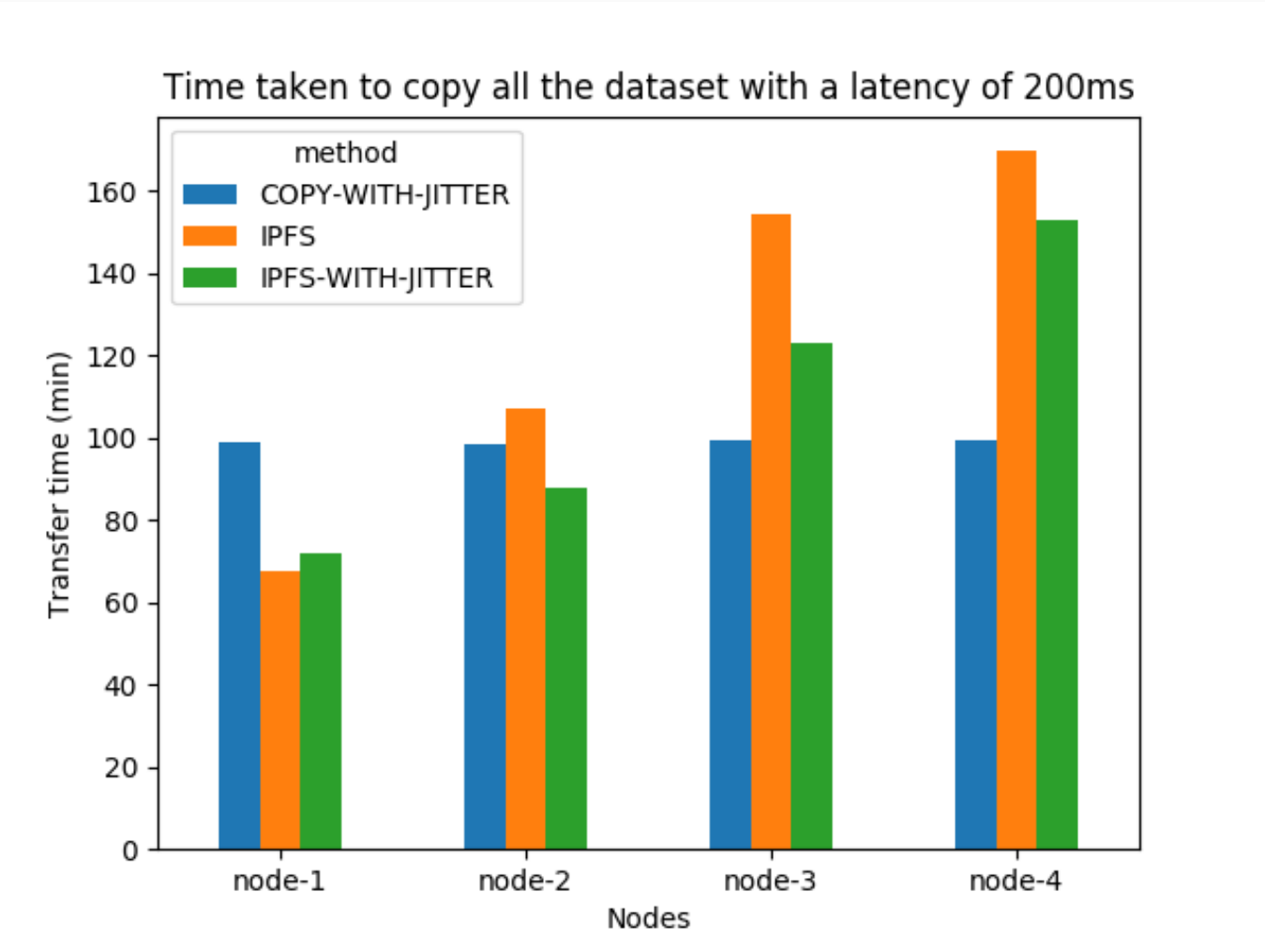

Figure 7: With 200 ms of latency

Experimental results

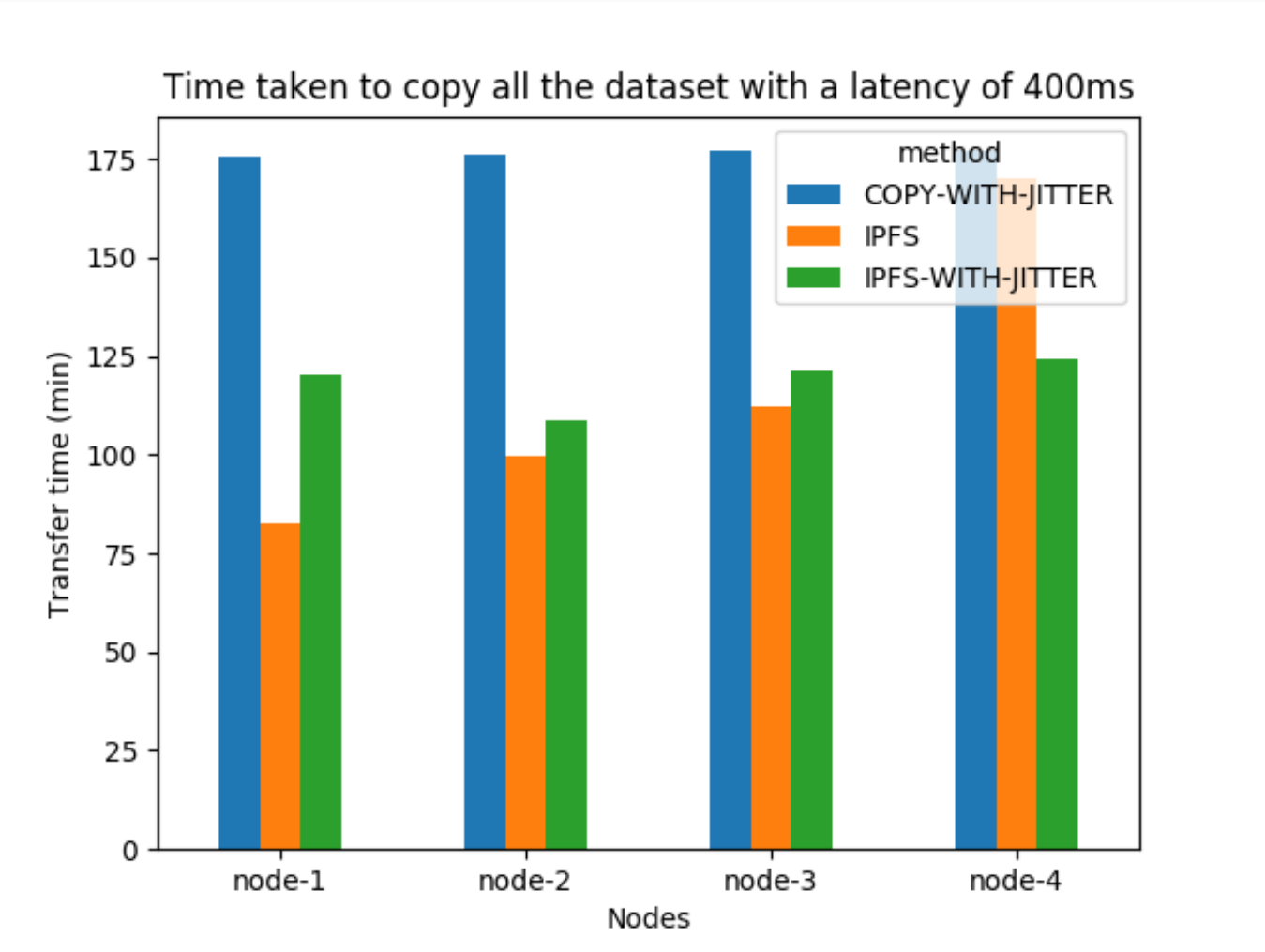

Figure 8: With 400 ms of latency

Experimental results

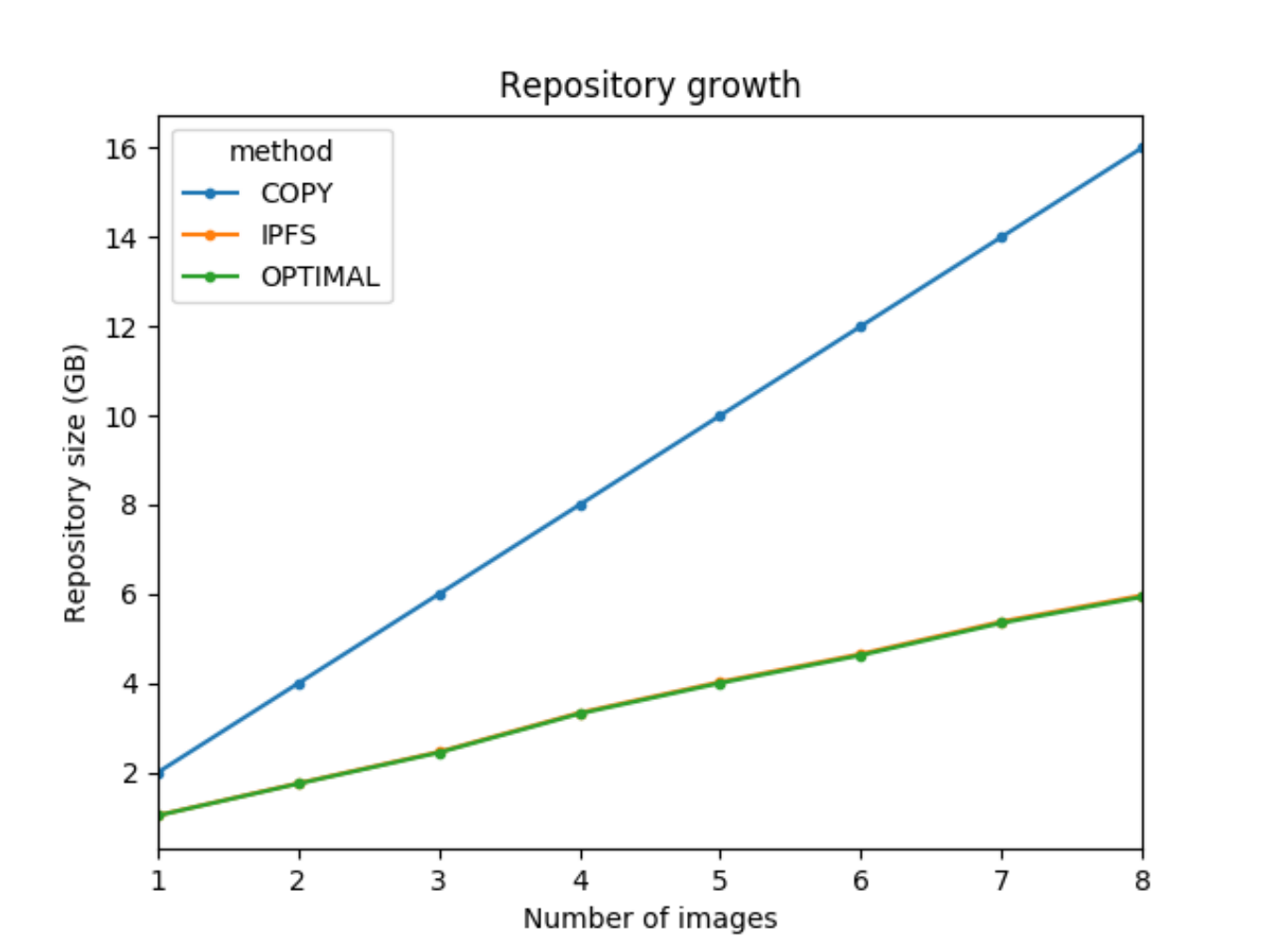

Figure 9: Repository growth

Wait

But wait!! Why we not just **rebuild** the image?!

Geo-distributed Datacenters

How to build a VM image?

1. Have the base image;
2. Reserve a compute node (VM);
3. Download required packages and softwares;
4. Install and configure the softwares;
5. Take a snapshot of the disk, and store it as an image.

Geo-distributed Datacenters

How to build a VM image?

1. Have the base image;
2. Reserve a compute node (VM);
3. Download required packages and softwares;
4. Install and configure the softwares;
5. Take a snapshot of the disk, and store it as an image.

Which is better in this case?
- It depends!

Geo-distributed Datacenters

How to build a VM image?

1. Have the base image;
2. Reserve a compute node (VM);
3. Download required packages and softwares;
4. Install and configure the softwares;
5. Take a snapshot of the disk, and store it as an image.

Which is better in this case?
- It depends!

HELP!
- Any idea about the percentage of public vs custom image usage?

Fog / Massively Distributed Clouds

Fog

Architecture

- Hierarchical by nature [3]
- Mega datacenters still there
- Extreme Fog nodes are backed by Micro DC

Fog

Can we use the same techniques (i.e. deduplication) to distribute the images?
- No.
Why?
- Deduplication is not effective with small set of images!

Fog - Network Latency

- Low latency: few ms
- "Medium" latency: 5 to 50 ms
- High latency: more than 50 ms

Fog

How to distribute the VMIs?

Fog

How to distribute the VMIs?

- Pull the image from the nearest Micro DC
  - Takes some time to transfer the image.
  - Tolerate network partitioning and variable bandwidth

Fog

How to distribute the VMIs?

- Pull the image from the nearest Micro DC
  - Takes some time to transfer the image.
  - Tolerate network partitioning and variable bandwidth
- Remotely attached block-device
  - Immediate launch
  - Strong dependency on the network
  - Low performance

Fog

Some thoughts..

- VMs and VMMs are really heavy;
- Containers are very lightweight;
- Containers are **NOT** lightweight Virtual Machines.

The Fog is not just an extension to the Cloud, it comes with new architectural design.

Conclusion

Conclusion

VMI management:

- **Intra DC:** Extensively studied.
- **Geo DC:** Questionable!
- **Fog:** ??

Thank you for your time

Questions?

References I

D. Balouek, A. Carpen Amarie, G. Charrier, F. Desprez, E. Jeannot, E. Jeanvoine, A. Lèbre, D. Margery, N. Niclausse, L. Nussbaum, O. Richard, C. Pérez, F. Quesnel, C. Rohr, and L. Sarzyniec.
Adding virtualization capabilities to the Grid'5000 testbed.
In I. Ivanov, M. Sinderen, F. Leymann, and T. Shan, editors, *Cloud Computing and Services Science*, volume 367 of *Communications in Computer and Information Science*, pages 3–20. Springer International Publishing, 2013.

G. Basu, S. Nadgowda, and A. Verma.
Lvd: Lean virtual disks.
In *Proceedings of the 15th International Middleware Conference*, Middleware '14, pages 25–36, New York, NY, USA, 2014. ACM.

References II


F. Bonomi, R. Milito, J. Zhu, and S. Addepalli.
Fog computing and its role in the internet of things.
In *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, MCC '12, pages 13–16, New York, NY, USA, 2012. ACM.


K. R. Jayaram, C. Peng, Z. Zhang, M. Kim, H. Chen, and H. Lei.
An empirical analysis of similarity in virtual machine images.
In *Proceedings of the Middleware 2011 Industry Track Workshop*, Middleware '11, New York, NY, USA, 2011. ACM.


K. Jin and E. L. Miller.
The effectiveness of deduplication on virtual machine disk images.
In *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference*, SYSTOR '09, New York, NY, USA, 2009. ACM.

References III


C. Li, D. Porto, A. Clement, J. Gehrke, N. Preguiça, and R. Rodrigues.
Making geo-replicated systems fast as possible, consistent when necessary.
In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation*, OSDI'12, Berkeley, CA, USA. USENIX Association.


C. Peng, M. Kim, Z. Zhang, and H. Lei.
Vdn: Virtual machine image distribution network for cloud data centers.
In *INFOCOM, 2012 Proceedings IEEE*, 2012.

References IV

J. Reich, O. Laadan, E. Brosh, A. Sherman, V. Misra, J. Nieh, and D. Rubenstein.

Vmtorrent: Scalable p2p virtual machine streaming.

In *Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '12, New York, NY, USA, 2012. ACM.

C. Tang.

Fvd: A high-performance virtual machine image format for cloud.

In *Proceedings of the 2011 USENIX Conference on USENIX Annual Technical Conference*, USENIXATC'11, pages 18–18, Berkeley, CA, USA, 2011. USENIX Association.

References V


X. Xu, H. Jin, S. Wu, and Y. Wang.
Rethink the storage of virtual machine images in clouds.
Future Gener. Comput. Syst.


X. Zhao, Y. Zhang, Y. Wu, K. Chen, J. Jiang, and K. Li.
Liquid: A scalable deduplication file system for virtual machine images.
IEEE Trans. Parallel Distrib. Syst., 25(5):1257–1266, May 2014.