VBS-Lustre: A Distributed Block Storage System for Cloud Infrastructure

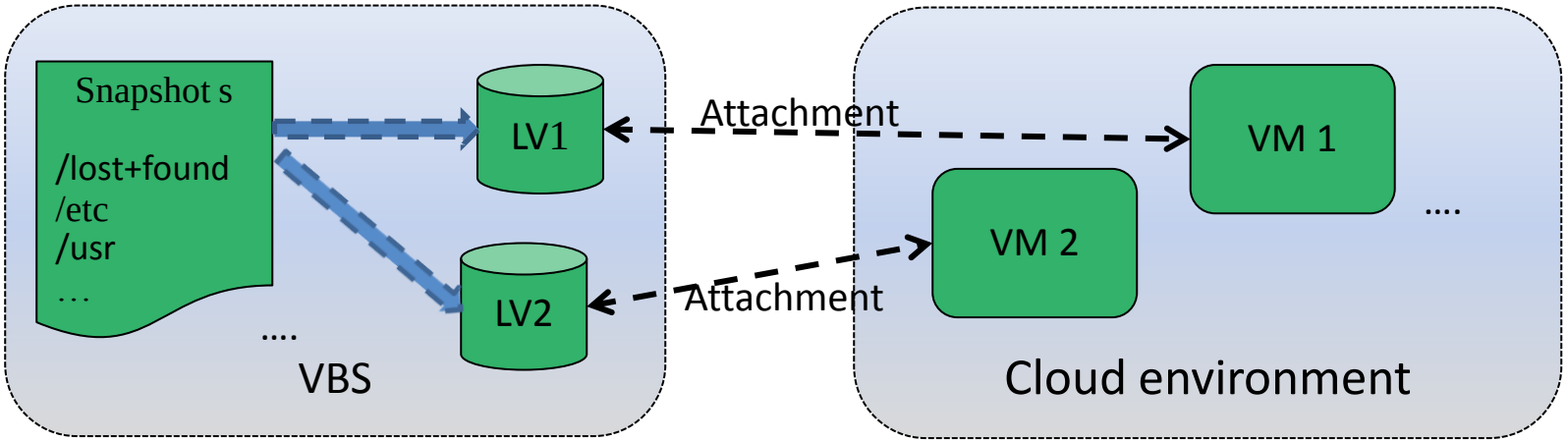
Xiaoming Gao, gao4@indiana.edu
Yu Ma, yuma@indiana.edu
Marlon Pierce, mpierce@cs.indiana.edu
Mike Lowe, jomlowe@iupui.edu
Geoffrey Fox, gcf@indiana.edu

Outline

- Introduction to VBS and VBS-Lustre
- The Lustre file system
- VBS-Lustre architecture
- Workflows
- Security and access control
- Read-only volume sharing
- Preliminary performance test
- Future work

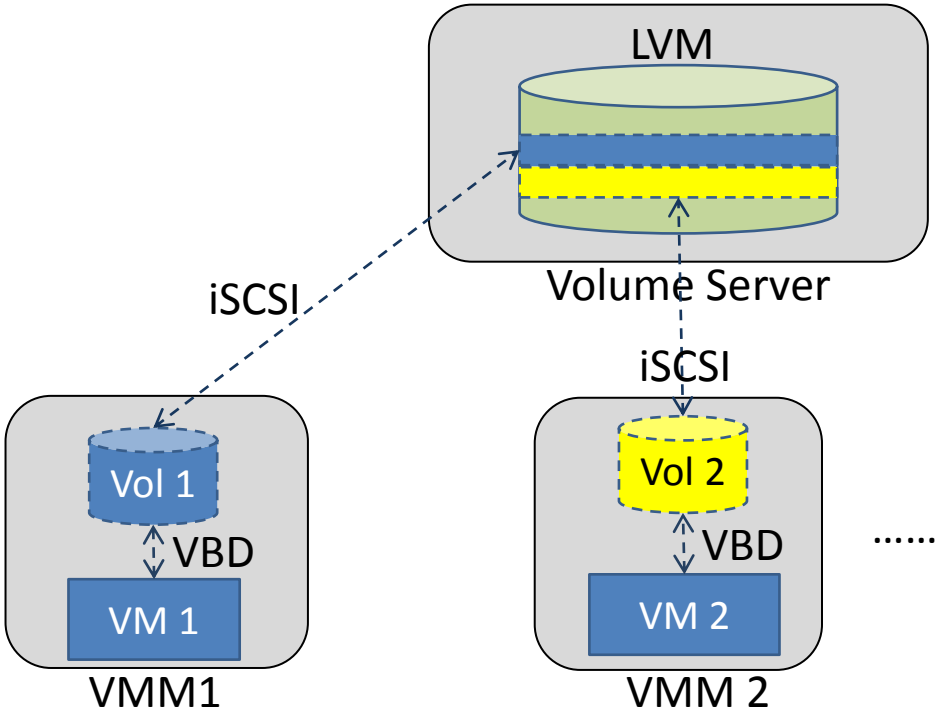
Introduction - VBS

- The Virtual Block Store (VBS) system is a block storage system that provide persistent virtual volumes to virtual machines in clouds.
- Similar functionality to Amazon Elastic Block Store (EBS): volume/snapshot creation and deletion, volume attachment and detachment



LV: logical volume
VM: virtual machine
Snapshot: a static “copy” of a logical volume at a specific time point

Introduction – VBS architecture

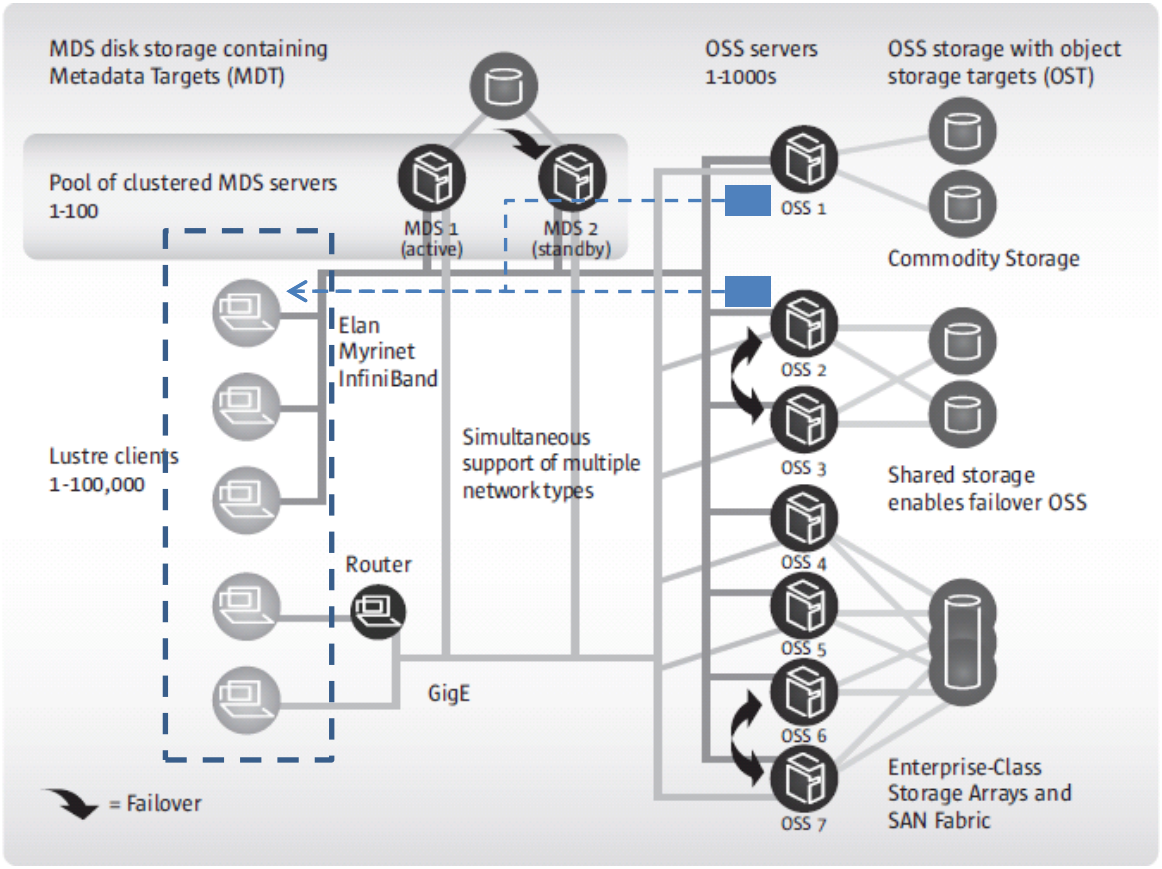


LVM: Logical Volume Manager
iSCSI: internet SCSI protocol
VBD: Virtual Block Device
VM: Virtual Machine
VMM: Virtual Machine Manager

- Single point of failure on volume server
- Not scalable
- Solution: VBS-Lustre

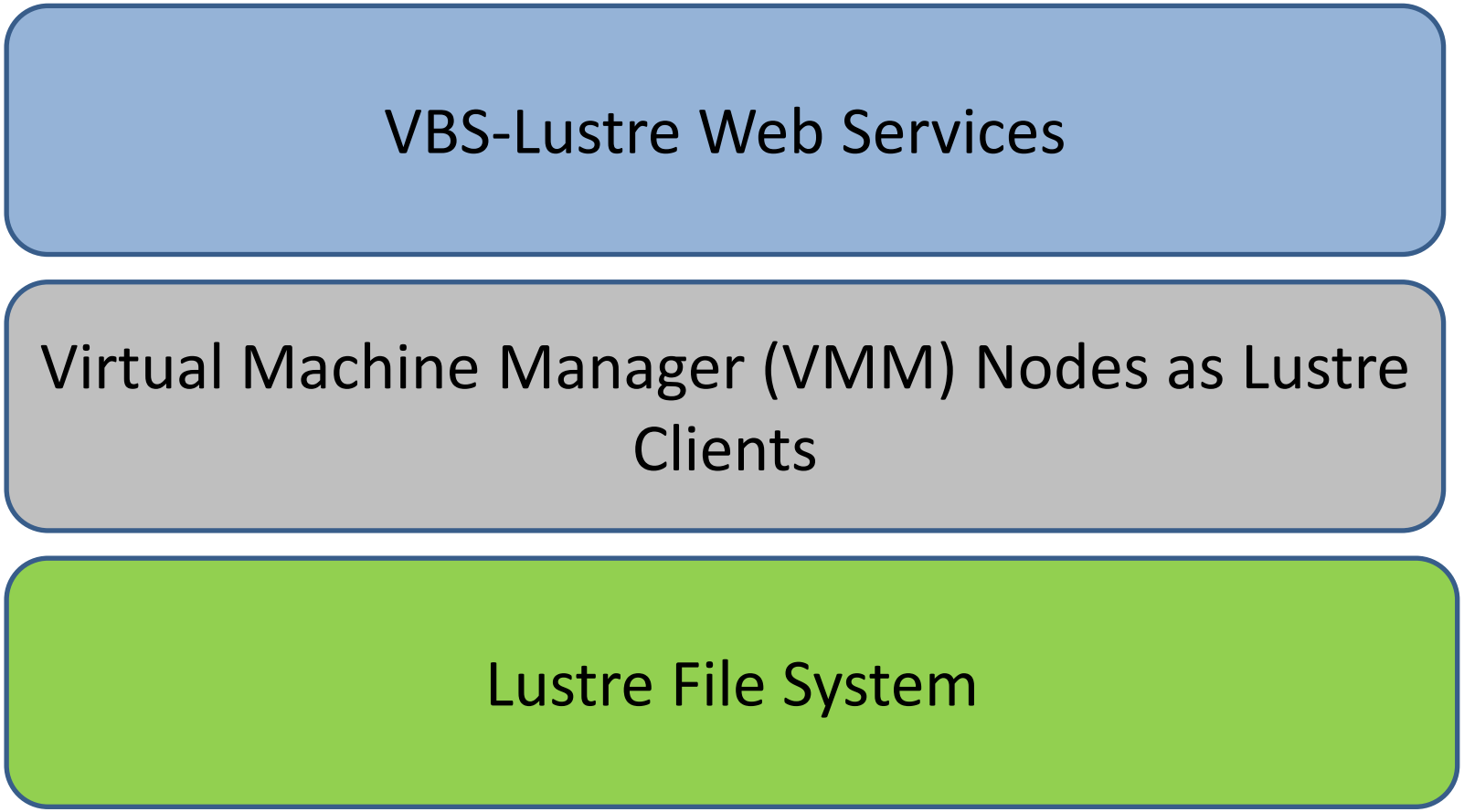
Lustre file system

- Developed by Oracle and Sun
- Scale to petabytes of storage and hundreds of gigabytes of I/O throughput

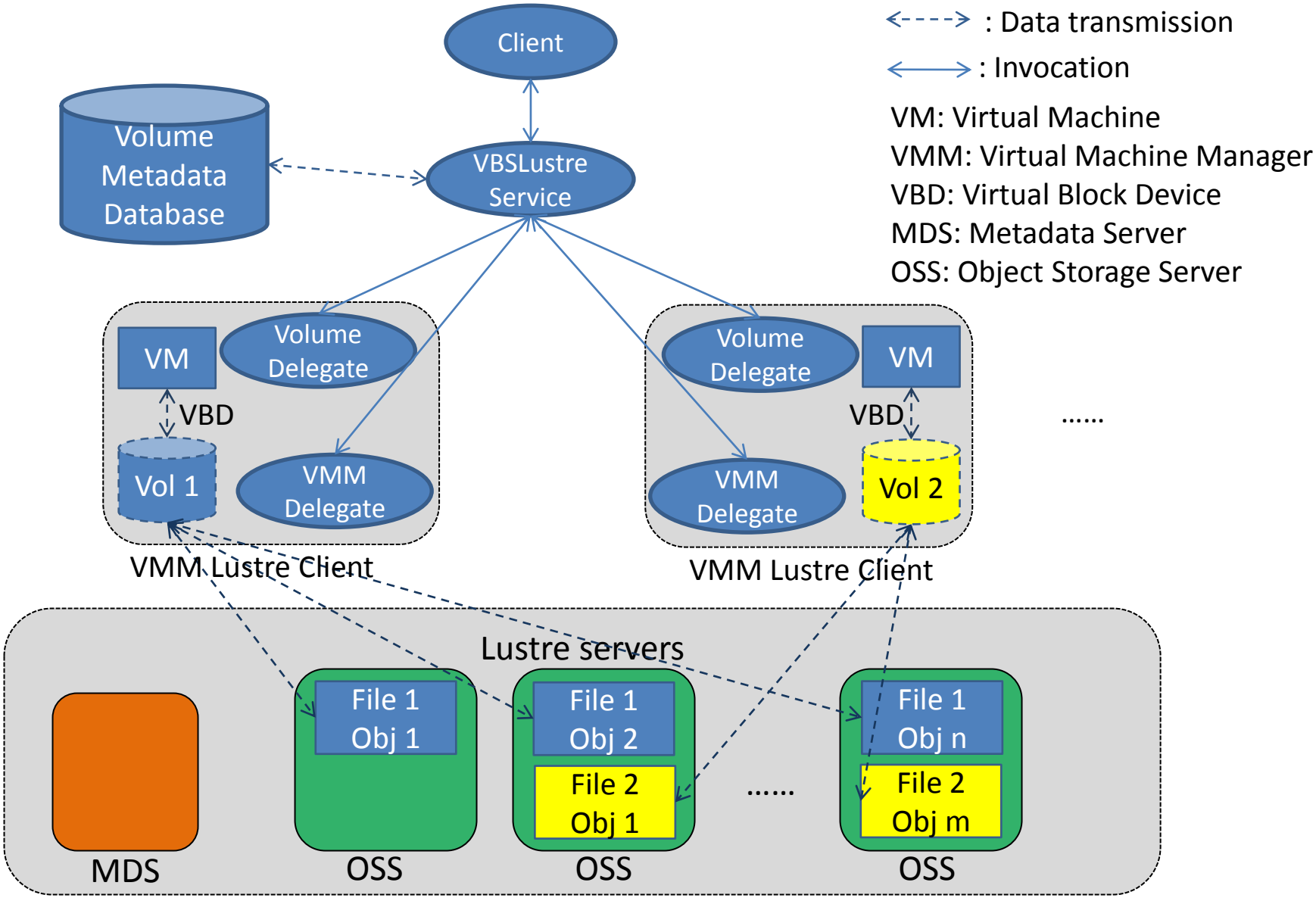


(Picture from the Lustre white paper 2008)

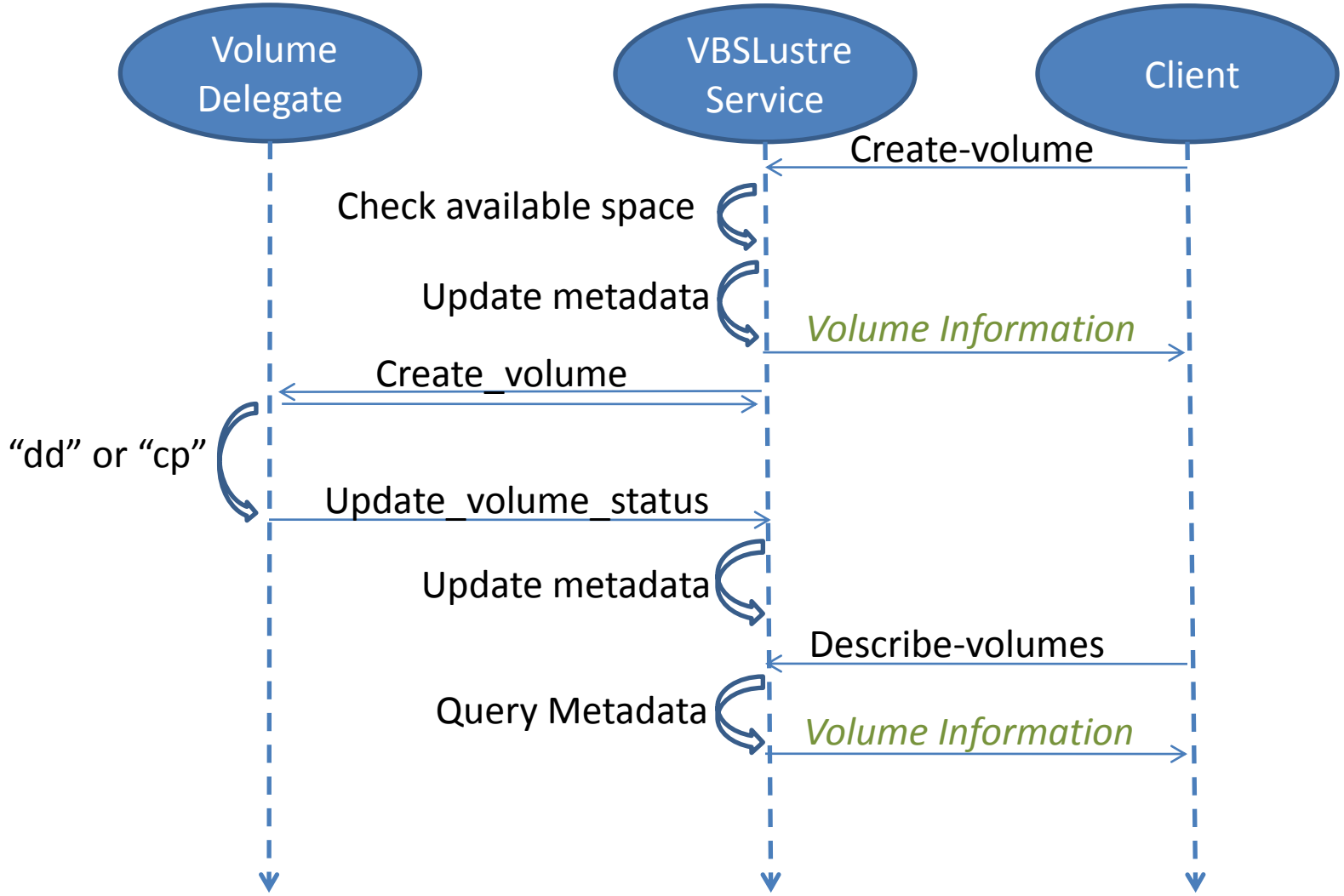
VBS-Lustre architecture



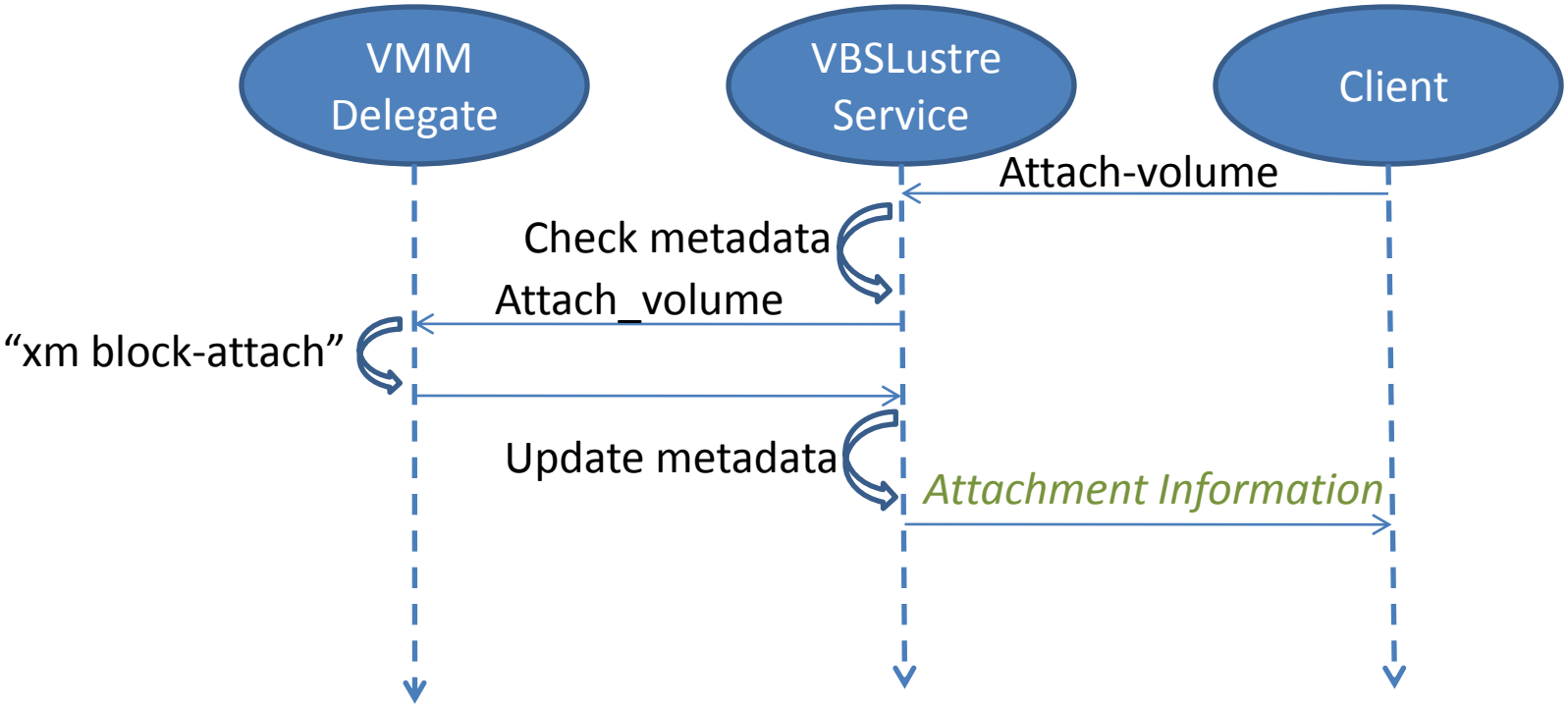
VBS-Lustre architecture



Workflows – create and describe volume



Workflows – attach volume

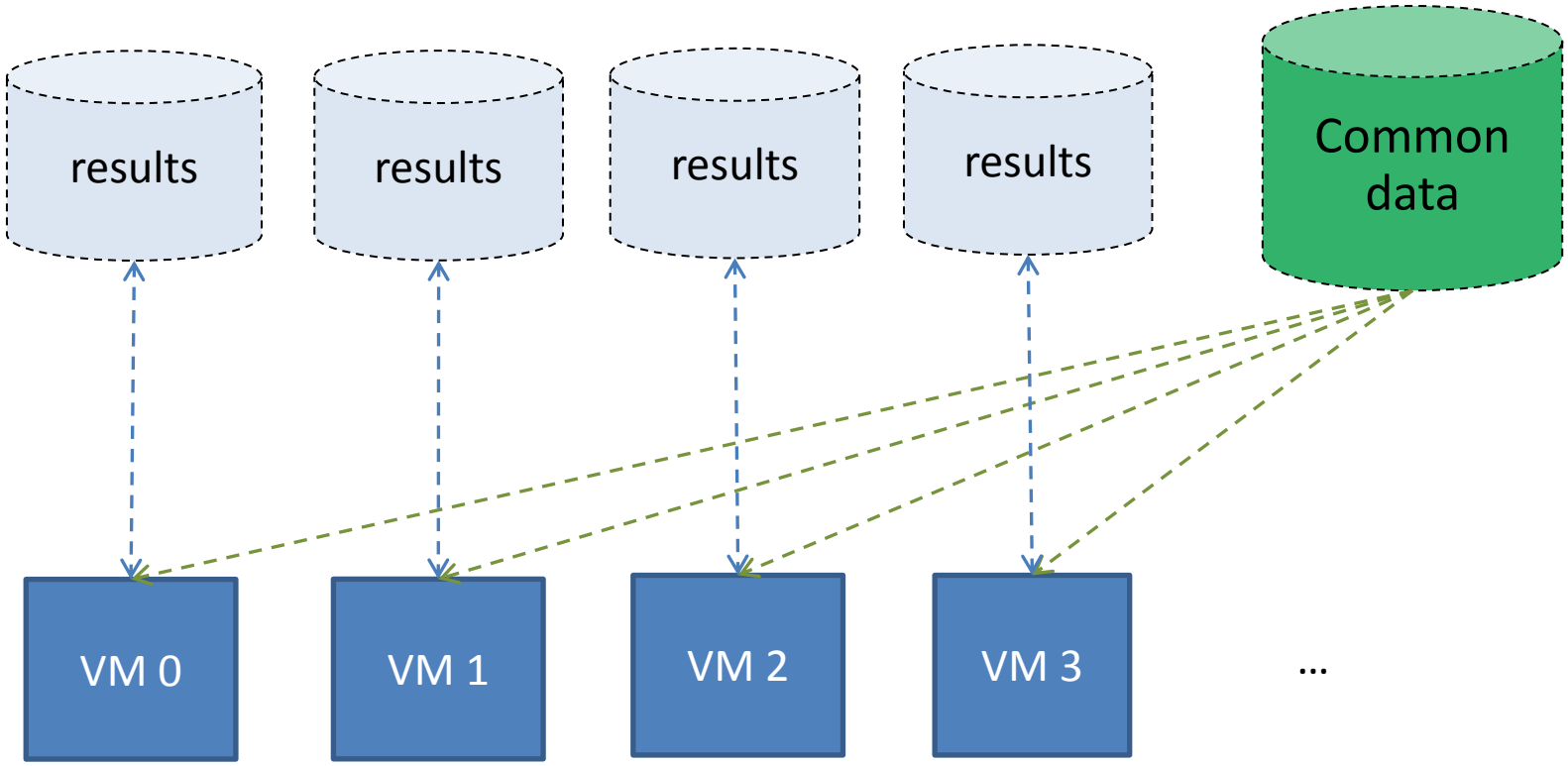


Security and access control

- Web service accesses protected with HTTPS channels
- Public key user authentication: users only allowed to access their own volumes
- New accounts created by adding new users' certificates to services' trusted certificate store

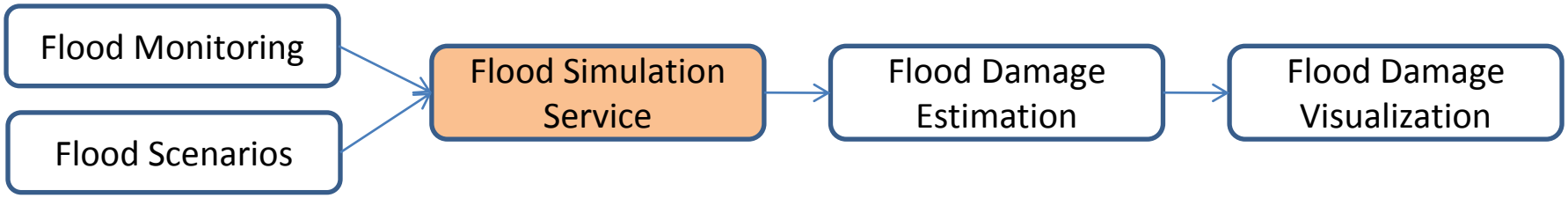
Read-only volume sharing

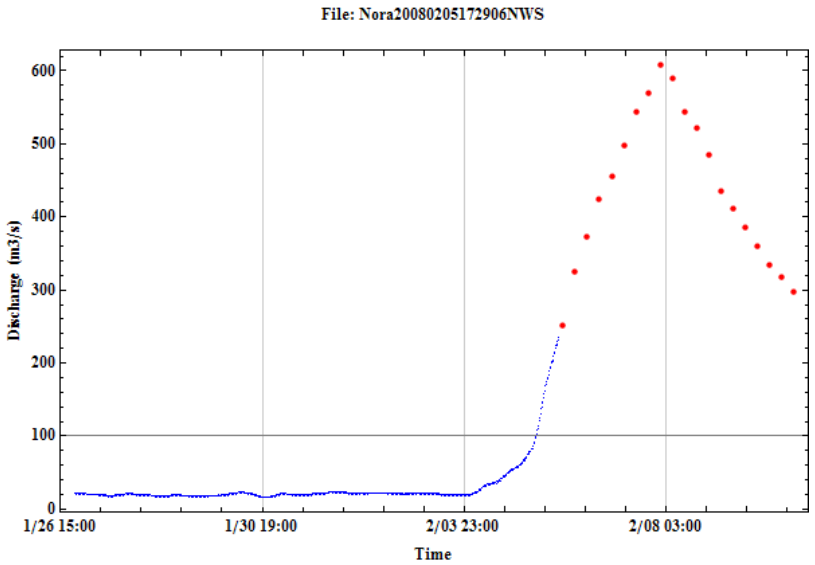
Definition: attaching one volume to multiple VM instances in read-only mode at the same time.

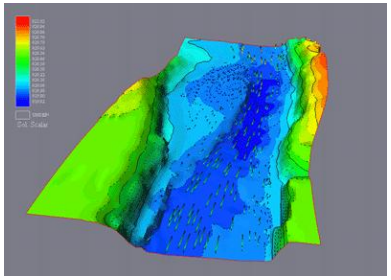


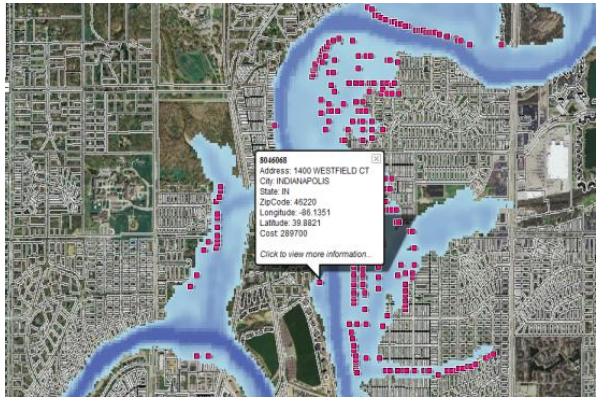
Experience with FloodGrid

- FloodGrid: an integrated platform for inundation modeling, property loss estimation, and visual presentation.

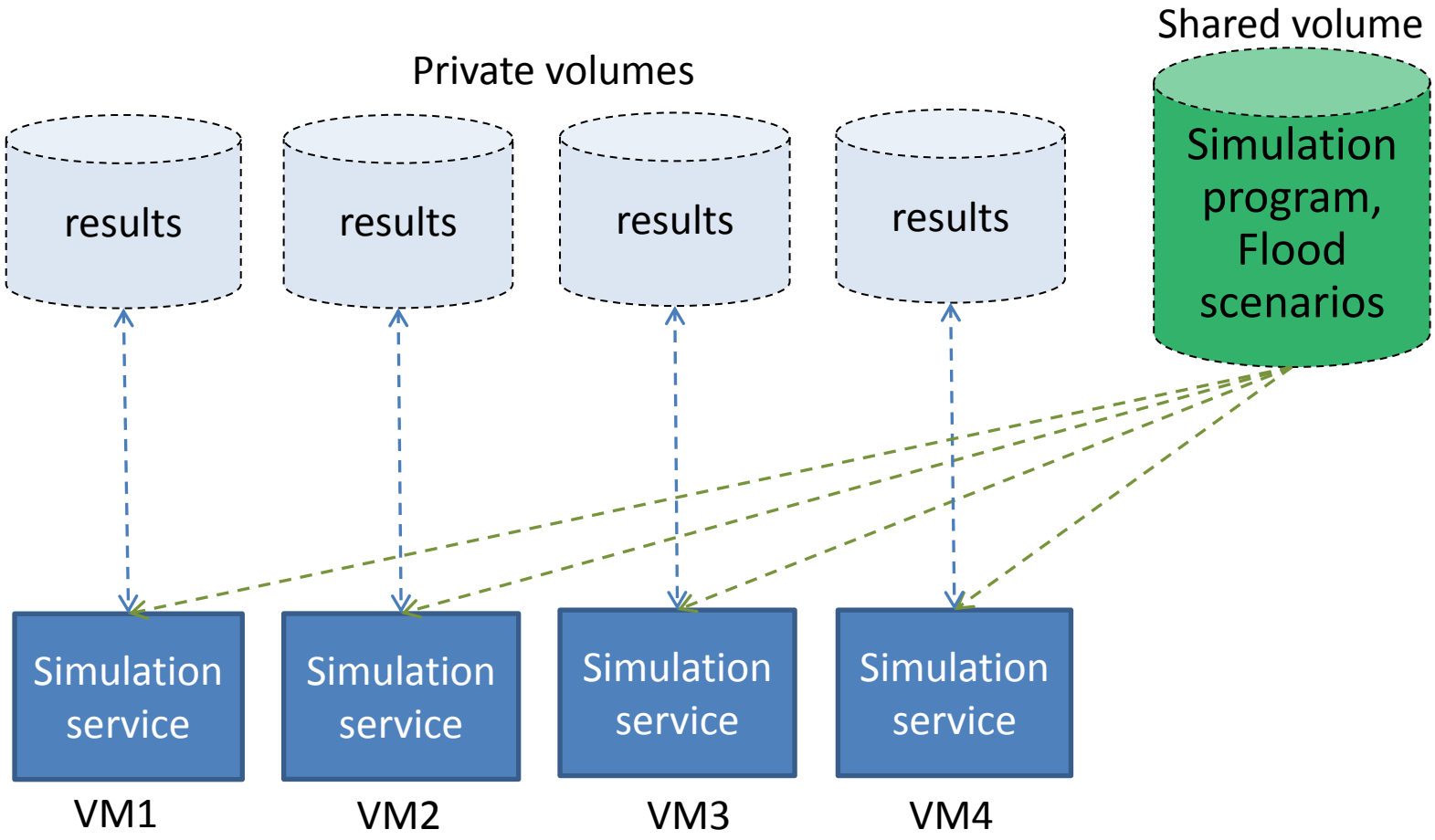






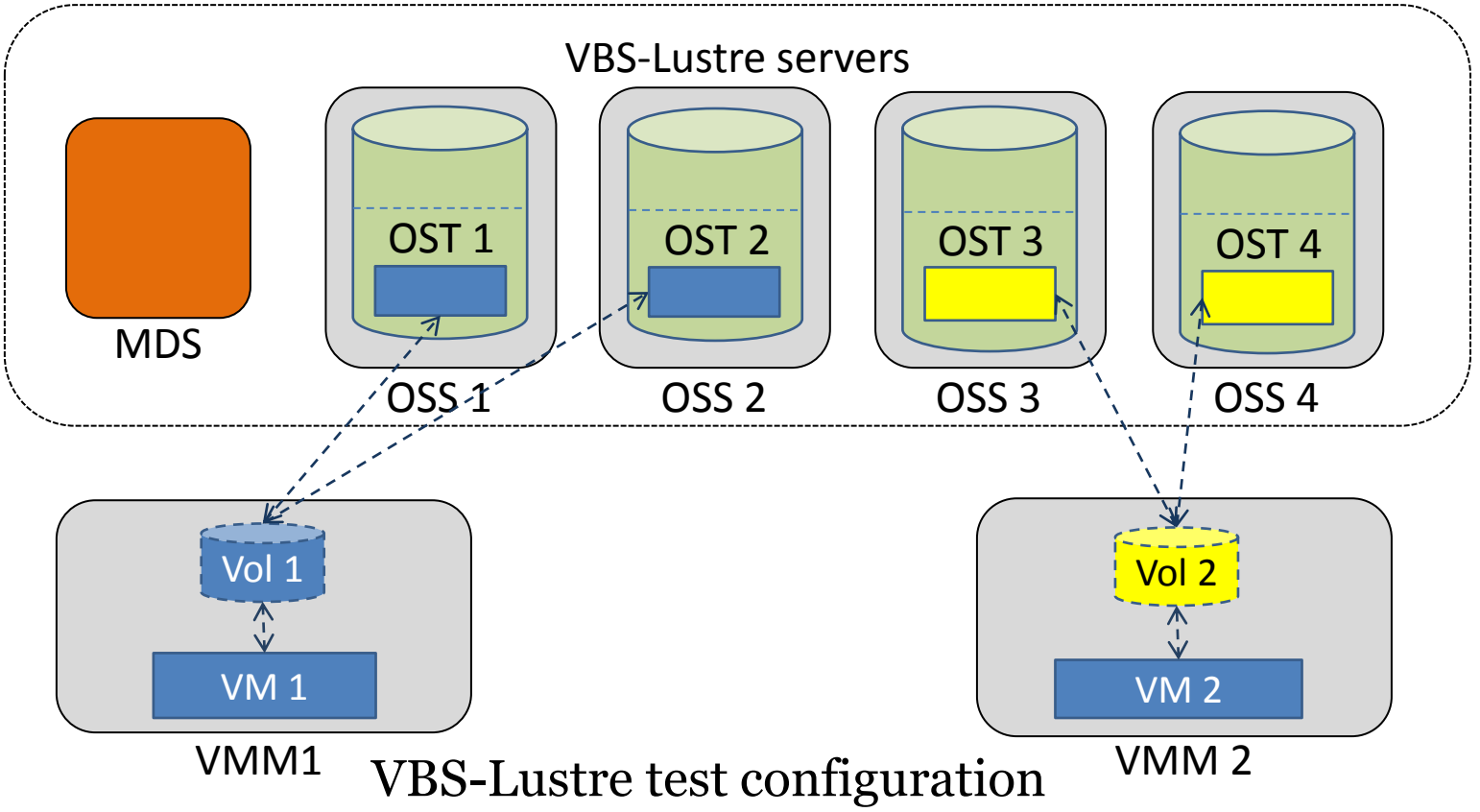


Experience with FloodGrid



- Analysis for 10 flood scenarios takes 205 minutes; in comparison, it takes 739 minutes if only 1 VM is used.

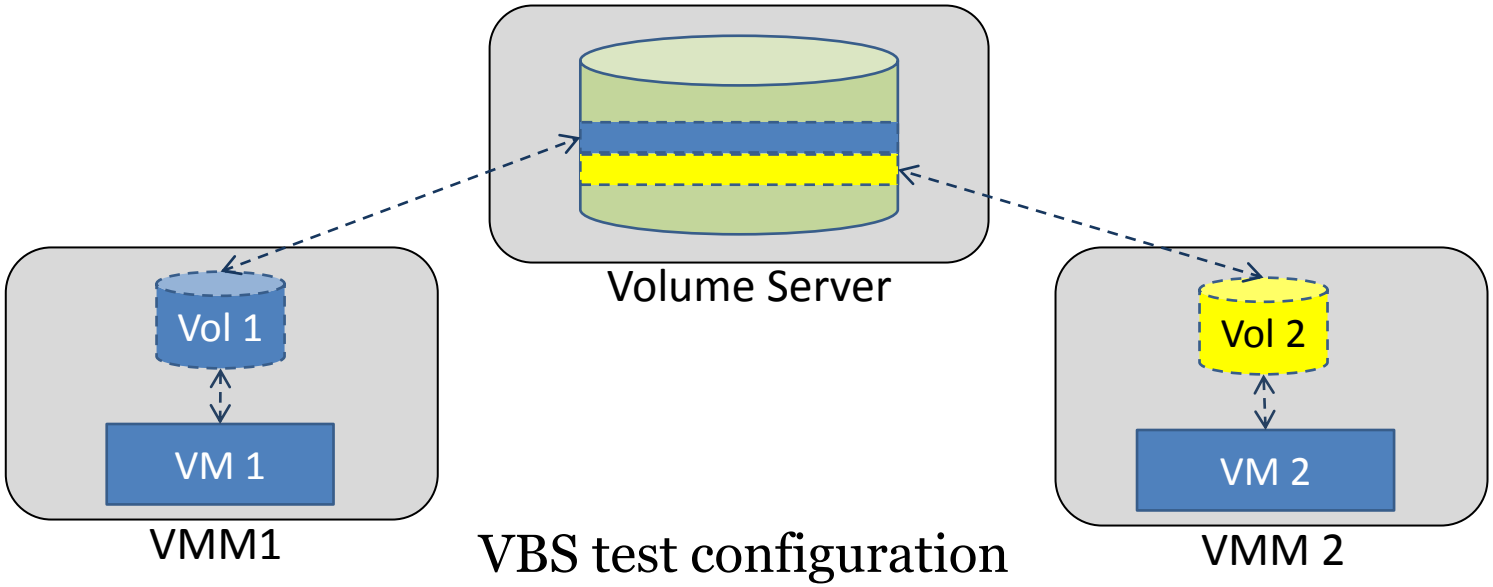
Preliminary performance tests



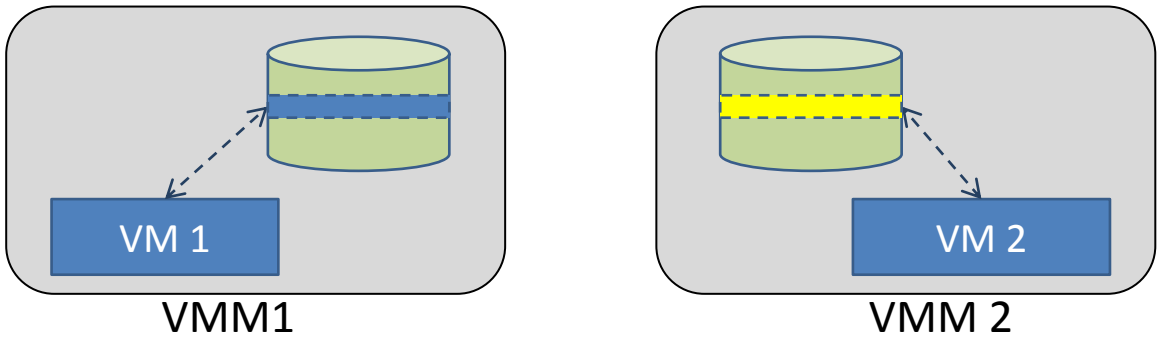
VBS-Lustre test configuration

MDS: 4 * Intel Xeon 2.8G CPU, 512MB, and 2 * 147GB 10K RPM.
OSS and VMM: 2 * AMD Opteron 2.52G CPU, 2GB, and 1 * 73GB 10K RPM.
VM: 1 * AMD Opteron 2.52G CPU, 256MB, and a 4GB disk image.
Volume size: 5GB.
All nodes connected to a 1Gb Ethernet LAN.

Preliminary performance tests

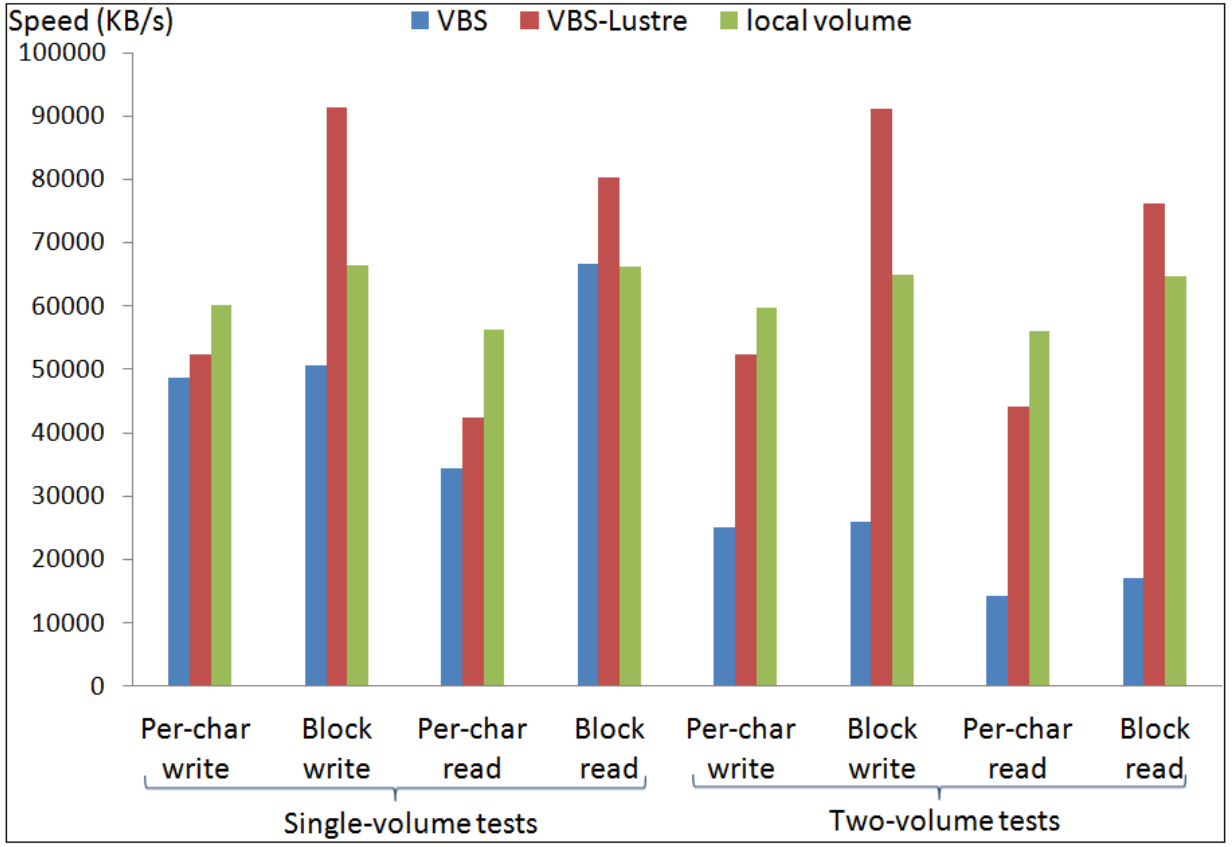


VBS test configuration



Local volume test configuration

Preliminary performance test



I/O throughput tests done with Bonnie++

Preliminary performance test

- VBS-Lustre metadata performance (files/s)

Test type	Sequential create	Random create	Random delete
single-volume	6629	6654	23211
two-volume VM1	6510	6724	23312
two-volume VM2	6565	6771	23274
two-volume Aggregate	13075	13495	46586

Future work

- Larger scale tests using data capacitor
- More efficient volume and snapshot creation
- Accommodate commodity hardware: using Distributed Replicated Block Device (DRBD) and Hadoop Distributed File System (HDFS)?
- Address issues with Lustre, such as metadata maintenance and small file access.

References

[1] X. Gao, M. Lowe, Y. Ma, M. Pierce, "Supporting Cloud Computing with the Virtual Block Store System", *Proceedings of e-Science 2009*, Oxford, UK, Dec. 2009.

[2] Amazon EBS, http://aws.amazon.com/ebs/

[3] Lustre file system white paper, Oct. 2008.

[4] Yang, R., "Flood Grid" *The 2009 International Symposium on Collaborative Technologies and Systems (CTS 2009)* , Baltimore, MD, 05/2009.

[5] bonnie++ http://www.coker.com.au/bonnie++/.

[6] LVM, http://tldp.org/HOWTO/LVM-HOWTO/.

[7] The iSCSI protocol, http://tools.ietf.org/html/rfc3720.

[8] The VBD technology of Xen, http://www.xen.org/.

[9] Eucalyptus, http://open.eucalyptus.com/.

[10] DRBD, http://www.drbd.org/.

[11] The Hadoop Distributed File System, http://hadoop.apache.org/hdfs/

Questions?