Tests of Cloud Computing and Storage System features requested by H1 Collaboration Data Preservation model

Bogdan Lobodzinski
Session: Grid & Cloud Middleware
21 October 2010
Why do we need a data analysis for old HEP experiments?

- to repeat existing measurements with a newly developed analysis techniques,
- to look for new phenomena found today,
- to check new or improved theoretical MC models,
- more details about Data Preservation project:
  - “Data Preservation in High Energy Physics” (PL-12),
  - “HERA Data Preservation Plans & Activities (DESY Data Preservation Group)” (PS41-2-357)
Cloud Computing concept for data analysis in data preservation model

- support of heterogeneous storage, OSs and VMs,
- idea of a private cloud computing - at some point the H1 software will be forced to use OS without valid security patches,
- easy base to maintain a possible migration of the H1 software to new platforms,
- common storage area for all virtual instances,
- scalability and centralized IT support,
- cost reduction - nodes can be shared with other projects,
Requirements of the project:

- Cloud manager:
  - open source,
  - compatibility with Amazon EC2,
  - support of multiple clusters,
  - possibility to start of SL5.X images,
  - easy way to build the interface between batch system and Cloud middleware based on the dynamical execution virtual instances recognition,
  - access to the common storage area,
Cloud Computing configuration – the H1 requirements

Requirements of the project:
- Storage system:
  - open source,
  - data safety based on the file replication,
  - usual operations on files/directories (editing, updating, saving, reading & writing) should be possible directly on the storage area,
  - non disturbing way to add additional storage units,
Cloud Computing test configuration

• The status & choices:
  • Open Source Cloud Middleware applications Eucalyptus 2.0
    (http://www.eucalyptus.com),
  • Distributed File System CEPH
    (http://ceph.newdream.net),
    - Is a part of of the kernel 2.6.34
      (http://kernelnewbies.org/Linux_2_6_34)
    - Offers easy scalability (+ default FS features: high performance & reliability)
**Cloud Computing test configuration**

- **Test-bed setup:**
  - **Eucalyptus:**
    - 4 nodes with Ubuntu 10.04.1 LTS,
    - on a single node: Cloud Controller (CC), Cluster Controller (CLC) & Storage Controller (SC)
    - 3 other nodes as Node Controllers (NC),
    - each node with 4 core Intel Xeon 3GHz CPU, 8 GB of RAM, 60 GB SATA disk, 1 GE NIC,
    - configured as a separate private subnet,
• Test-bed setup:
  - CEPH FS (Won’t go into the CEPH architecture details see http://ceph.newdream.net):
    • 3 nodes with Ubuntu 10.04.1 LTS OS,
    • 2 monitoring (MON) & Metadata servers (MDS),
    • 2 Object Storage Devices (OSD) (1 OSD shared with MDS & MON),
    • On each node 2 SATA disks: 2 x 500 GB,
    • Maintained via NFS to the Eucalyptus test-bed
Cloud Computing test configuration

- Test-bed setup:

Diagram showing public network, private switch, CC, CS, CLC, NC Nodes, CEPH storage area, MON, MDS, MON, MDS, ODS, ODS, NFS mount.
Eucalyptus configuration

- Eucalyptus installation & post-installation is rather well described. For Ubuntu 10.04.1 LTS: [http://open.eucalyptus.com/wiki/EucalyptusInstallationDebian_v2.0](http://open.eucalyptus.com/wiki/EucalyptusInstallationDebian_v2.0)
  - We didn’t get running our test installation with libvirt user as “eucalyptus”, as it is strongly suggested by Eucalyptus team. We were not able to start any instance. On the NC node we had (nc.log):

```
[...][005698][EUCAERROR ] libvirt: monitor socket did not show up.: No such file or directory (code=38)
[...][005698][EUCAFATAL ] hypervisor failed to start domain
[...][005698][EUCAERROR ] libvirt: Domain not found: no domain with matching name 'i-548A0913' (code=42)
[...][005698][EUCAINFO  ] stopping the network (vlan=10)
```

**checking** /var/log/libvirt/qemu/i-548A0913.log you can find more clear information:

```
libvir: QEMU error: cannot set ownership on /var/lib/eucalyptus/instances//admin/i-548A0913/kernel: Permission denied
```

**libvirt had to be configured to run as user “root” & group “root”**
Eucalyptus Network configuration

- **MANAGED mode:**
  - In theory, it allows to use all features supported by Eucalyptus 2.0:
    - Security groups & VM isolation (between VM started by different security groups) - useful if you need to share the cloud infrastructure with different project,
    - Metadata Service - should be available via http://169.254.169.254, we didn’t find it running
  - Networking based on virtual ethernet devices assigned by DHCP service (Eucalyptus part) with IP address from a given range of IP addresses
    - initial configuration must be created by hand, before Eucalyptus start-up on each Front-end and NC nodes with a given IP addresses (example):

```
% vconfig add eth0 10
% ifconfig eth0.10 10.10.1.2 up
% ip a sh eth0.10
```

<- check
Eucalyptus Network configuration

- `/etc/eucalyptus/eucalyptus.conf`:

  **NC node:**

  ```
  # Affects: NC
  NC_PORT=“8775"
  HYPERVERSOR=“kvm"
  MANUAL_INSTANCES_CLEANUP=0
  VNET_INTERFACE=“eth0”
  VNET_BRIDGE=“br0”
  INSTANCE_PATH=“/home/eucalyptus/instances/"
  
  # Affects: CLC, Walrus, SC
  DISABLE_DNS=“Y”
  DISABLE_ISCSI=“N”
  JVM_MEM=“512m”

  # Affects: CC, NC
  ENABLE_WS_SECURITY=“Y”
  LOGLEVEL=“DEBUG”
  VNET_PUBINTERFACE=“eth0”
  VNET_PRIVINTERFACE=“eth0”
  VNET_MODE=“MANAGED”

  # Affects: CC
  CC_PORT=“8774”
  SCHEDPOLICY=“ROUNDROBIN”
  POWER_IDLETHRESH=“300”
  POWER_WAKEHRESH=“300”
  NC_SERVICE=“axis2/services/EucalyptusNC”
  VNET_DHCPDAEMON=“/usr/sbin/dhcpd3”
  VNET_DHCPUSER=“dhcpd”
  VNET_ADDRSPPERNET=“32”
  VNET_SUBNET=“10.10.0.0”
  VNET_NETMASK=“255.255.0.0”
  VNET_DNS=“131.169.XX.XXX”
  VNET_PUBLICIPS=“192.168.XX.21-192.168.XX.23”
  
  # Affects: All
  EUCALYPTUS=“/”
  EUCA_USER=“eucalyptus"

  # Affects: Front-End node:

  **Front-End node:**

  ```
OS Image creation

- **Image creation (kvm):**

  ```
  kvm -m 256 -cdrom <location_of_the_OS_iso_files>/SL.54.110309.DVD.i386.disc1.iso -drive if=scsi,file=SL54.img,boot=off
  ```

  - *Introduce changes, for example starting:*
    ```
    kvm -drive file=SL54.img,if=scsi,boot=on
    ```

  - **Add working user, modify sshd configuration appropriately, disable SELinux, remove HWADDR from the network configuration (/etc/sysconfig/network-scripts/ifcfg-eth0), disable iptables & ip6tables, add missing libraries for future applications,**

  - **Create a new ramdisk:**

    ```
    /sbin/mkinitrd --preload=virtio_pci --preload=virtio_blk --preload=virtio_net --preload=mptspi --preload=acpi php /tmp/initrd-2.6.18-164.2.1.el5.private.img 2.6.18-164.2.1.el
    ```
OS Image creation

- Download file system, ramdisk & kernel to the Walrus (a storage controller compatible with Amazon S3 and responsible for storage, backup) & register with Eucalyptus: http://open.eucalyptus.com/wiki/EucalyptusImageManagement_v2.0

- Result:

<table>
<thead>
<tr>
<th>% euca-describe-images</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMAGE eri-F2051520</td>
</tr>
<tr>
<td>admin available public</td>
</tr>
<tr>
<td>IMAGE eki-EA5D1044</td>
</tr>
<tr>
<td>available public</td>
</tr>
<tr>
<td>IMAGE emi-21240C85</td>
</tr>
<tr>
<td>available public</td>
</tr>
<tr>
<td>eri-F2051520</td>
</tr>
</tbody>
</table>

Bogdan Lobodzinski, CHEP 2010, Taipei, Taiwan
Run Virtual Machines

- **Start the SL5 instance:**
  
  ```
  euca-run-instances -t c1.xlarge -k mykey -n 1 emi-21240C85 --addressing private
  
  Problem (not yet solved ...): executing euca-run-instances without "--addressing private" we cannot start any instance:
  ```
  
  ```
  % euca-run-instances -t c1.xlarge -k mykey -n 1 emi-21240C85
  FinishedVerify: Not enough resources available: addresses (try --addressing private)
  ```

- **Result:**
  
  ```
  % euca-describe-instances
  RESERVATION     r-4A3907BA      admin   default
  INSTANCE        i-3C31079E      emi-21240C85    10.10.1.2       10.10.1.2       running
  mykey          0               c1.xlarge
  2010-10-12T15:49:52.04Z     cluster1        eki-EA5D1044    eri-F2051520
  ```
Volumes on running virtual instances

- **Attach volume to the SL5 instance:**
  
  ```
  % euca-attach-volume -i i-49EB0867 -d /dev/sdb vol-58540612
  VOLU ME vol-5854061
  ```

- **On the VM:**
  
  ```
  % /sbin/fdisk -l
  Disk /dev/sda: 1610 MB, 1610612736 bytes
  255 heads, 63 sectors/track, 195 cylinders
  Units = cylinders of 16065 * 512 = 8225280 bytes

  Device Boot Start End Blocks Id System
  /dev/sda1   * 1 195 1566306 83 Linux

  Disk /dev/sdb: 9663 MB, 9663676416 bytes
  64 heads, 32 sectors/track, 9216 cylinders
  Units = cylinders of 2048 * 512 = 1048576 bytes

  Disk /dev/sdb doesn't contain a valid partition table
  ```
Volumes on running virtual instances

- On the VM:

% /sbin/mkfs.ext3 /dev/sdb

% mkdir /opt/vol

% mount -t ext3 /dev/sdb /opt/vol

% mount h1farm183.desy.de:/ceph /ceph
Volumes on running virtual instances

- **Snapshot handling:**
  - Idea: use already prepared volumes created from a snapshot as a common input for Monte Carlo production on each started VM
  - *Cannot be realized: creation of snapshot forever in pending state:*

<table>
<thead>
<tr>
<th>% euca-describe-snapshots</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNAPSHOT</td>
</tr>
</tbody>
</table>

Creation of the input Volume for each started VM

Snapshot of the common input files

Volume with input files for a given VM

Bogdan Lobodzinski,  
CHEP 2010, Taipei, Taiwan
Volumes on running virtual instances

- **Snapshot handling (cloud-error.log):**

  16:47:29 [JDBCExceptionReporter:pool-11-thread-4] ERROR Couldn't perform the operation prepareStatement: You can't perform any operations on this connection. It has been automatically closed by Proxool for some reason (see logs).


  org.hibernate.exception.GenericJDBCException: could not insert: [edu.ucsb.eucalyptus.cloud.entities.AOEVolumeInfo]

  ...

  16:47:30 [JDBCTransaction:pool-11-thread-4] ERROR Could not toggle autocommit java.sql.SQLException: Couldn't perform the operation setAutoCommit: You can't perform any operations on this connection. It has been automatically closed by Proxool for some reason (see logs).

  ...

  16:47:30 [JDBCTransaction:pool-11-thread-4] ERROR JDBC rollback failed java.sql.SQLException: Couldn't perform the operation rollback: You can't perform any operations on this connection. It has been automatically closed by Proxool for some reason (see logs).

  ...


  16:49:10 [DefaultServiceExceptionStrategy:connector.VM.0.dispatcher.9] ERROR

---

Bogdan Lobodzinski, CHEP 2010, Taipei, Taiwan
MC application on running virtual instance

- Working MC simulation model:

  **CEPH Storage Area**

  Updates, Modifications of Input files from outside of Cloud Test-bed

  **VM**

  Attached Volume

  Input files

  Output files

  Users access to the Output files

Bogdan Lobodzinski, CHEP 2010, Taipei, Taiwan
Eucalyptus test configuration – other troubles

- **EUCALYPTUS 2.0**
  - loosing NC nodes: sometimes, after reconfiguration we cannot register NC nodes
    euca_conf -register-nodes “IP1 IP2 ..” doesn’t work,
  - solution: be sure that keys are in the right place: 
    /var/lib/eucalyptus/keys, deregister nodes and register again
  - Volumes and snapshots are not removed after starting euca-delete-[volume|snapshot]
  - Non-modular: difficult for potential modifications,
  - Log files: lot of entries & difficult to encrypt,
  - Commercial (Enterprise) release have much bigger development momentum than the open source version
• CEPH FS: installation is based on http://ceph.newdream.net/wiki/Debian
  Version: ceph-0.21
• Strongly suggested to use btrfs as a base FS for CEPH
• Replication level:
  can be adjusted independently for metadata and data pools with command:
  ceph osd pool set <metadata | data> size <nr_of_replicas>
  (for details see: http://ceph.newdream.net/Adjusting_replication_level)
CEPH File System

- **Configuration (ceph.conf):**

  **monitoring:**

  ```
  [global]
  pid file = /var/run/ceph/$name.pid
  debug ms = 1
  keyring = /etc/ceph/keyring.bin
  
  [mon]
  mon data = /x01/mon$id
  debug mon = 20
  debug paxos = 20
  mon lease wiggle room = 0.5
  
  [mon0]
  host = h1farm182
  mon addr = 131.169.XX.XXX:6789
  
  [mon1]
  host = h1farm183
  mon addr = 131.169.XX.XXX:6789
  ```

  **Metadata Server:**

  ```
  [mds]
  debug mds = 10
  mds log max segments = 2
  keyring = /etc/ceph/keyring.$name
  
  [mds0]
  host = h1farm182
  
  [mds1]
  host = h1farm183
  ```

  **Object Storage Device:**

  ```
  [osd]
  sudo = true
  keyring = /etc/ceph/keyring.$name
  osd data = /x02/osd$id
  osd journal = /x02/osd$id/journal
  osd journal size = 100
  debug osd = 20
  debug journal = 20
  debug filestore = 20
  
  [osd0]
  host = h1farm183
  btrfs devs = /dev/sdb1
  
  [osd1]
  host = h1farm184
  btrfs devs = /dev/sdb1
  ```
CEPH File System - tests

• test command:
  `svn co https://root.cern.ch/svn/root/trunk`
  after download you will get 63886 files - good & fast way for tests of any FS.

• CEPH build on ext3 FS:
  test command was stuck in the middle of download with kernel panic (`ext3 xattr bug` - exists in the kernel version: 2.6.32, not yet fixed in 2.6.34):

  [99048.568767] kernel BUG at /build/buildd/linux-2.6.32/fs/ext3/balloc.c:1384!
  [99048.568767] invalid opcode: 0000 [#1] SMP
  [99048.568767] last sysfs file:
  /sys/devices/pci0000:00/0000:00/0000:00:00.0/device

Bogdan Lobodzinski, CHEP 2010, Taipei, Taiwan
**CEPH File System - tests**

- **CEPH build on btrfs - the test command finished but we got kernel panic later:**

  ```
  Sep 2 05:44:42 h1farm183 kernel: [72426.976029] ceph: mds0 caps stale
  Sep 2 05:44:57 h1farm183 kernel: [72441.976037] ceph: mds0 caps stale
  Sep 2 05:45:27 h1farm183 kernel: [72472.066320] ceph: mds0 reconnect start
  Sep 2 05:45:27 h1farm183 kernel: [72472.069681] Modules linked in: nfs lockd nfs_acl auth_rpcgss sunrpc ceph btrfs zlib_deflate crc32c libcrc32c ppdev lp parport openafs(P) ipt_MASQUERADE iptable_nat nf_nat nf_conntrack_ipv4 nf_defrag_ipv4 xt_state nf_conntrack ipt_REJECT xt_tcpudp iptable_filter ip_tables x_tables bridge stp fbcon tileblit font bitblit softcursor vga16fb vgastate radeon ttm mptctl drm_kms_helper bnx2 drm usbhid i5000_edac hid dell_wmi shpchp edac_core agpgart i2c_algo_bit i5k_amb dcmdbas psmouse serio_raw mptsas mptscsih mptbase scsi_transport_sas [last unloaded: kvm]
  Sep 2 05:45:27 h1farm183 kernel: [72472.072332]
  Sep 2 05:45:27 h1farm183 kernel: [72472.072332] Pid: 6184, comm: ceph-msgr/1
  Tainted: P (2.6.32-24-generic-pae #42-Ubuntu) PowerEdge 1950
  Sep 2 05:45:27 h1farm183 kernel: [72472.072332] EIP: 0060:[<c01ea907>] EFLAGS: 00010246 CPU: 1
  ```

- **Patch for this bug is done (ceph-0.21.3) - not tested yet**
CEPH File System – tests

- Unstable connections between CEP parts mounted by `cfuse` mount tool (stuck connections). Better stability with standard `mount` system command
- Stable work if not too much files are involved into the process
Cloud Computing test configuration - summary

- General: Eucalyptus Cloud manager & CEPH look promising but not ready for production mode - too poor stability & reliability,
- CEPH Petabyte FS is truly experimental - all tests with bigger number of files failed,
- In both cases management is difficult,
- It is too early to present any benchmarks and performance,
- Results are encouraging, allow for optimistic anticipation of the future

Bogdan Lobodzinski, CHEP 2010, Taipei, Taiwan