VmBatch, a dynamic virtualization tool for running grid jobs

Bjarte Kileng

Bergen University College

March 20, 2015

What is VmBatch

- System for running batch jobs using virtual machines.
- Less complex than a full private cloud infrastructure.
- Has been tested in our private AliEn testbed to run AliEn jobs.
- Currently, only support for Torque with XEN3 and XEN4 as hypervisors.
- Has been tested with SLC5, CentOS6, CernVM2 and CernVM3.
  - Still no XEN tools in CentOS7.

How it works

- Uses Torque **prologue** and **epilogue** scripts:
  - **prologue** is run when a job starts and **epilogue** when a job finishes.
- Script **remoteshell** returns a connection to the virtual machine.
  - **remoteshell** uses as default a SSH connection into the running guest.
  - Experimental support for using the console of the guest.
- Guest parameters (e.g.: notification that the guest is alive and ready, and IP of guest) are transferred to VmBatch using a NFS share.
- Configuration by configuration files and VmBatch tags added to the libvirt domain XML file.

Contextualization

VmBatch will prepare the guest before running the job:

- The submitting user must be created.
- The job must be copied to the guest.
- The NFS share must be set up inside the guest.
- If not DHCP, the guest network must be configured.
- VmBatch allows tar archives to be unpacked inside the guest prior to running the job.
- If **remoteshell** is configured to connect to the console of the guest, the guest will be configured to auto-login as the submitting user.
- VmBatch supports three different approaches for guest contextualization, *CDROM*, *DISK* and *ROOTSSH*.

The *CDROM* contextualization approach

- VmBatch will create a custom CDROM ISO image which is attached to the guest.
- The ISO image will contain the job, VmBatch parameters, tar files and init-scripts.
- VmBatch can unpack a tar archive onto the CDROM file system.
  - E.g. with a CernVM contextualization file **context.sh**.
- The guest must be set up to mount the CDROM and run a **prolog.sh** and **epilog.sh** script located on the CDROM.
  - A stock CernVM image can be used as VmBatch guest.

Extract of domain XML with *CDROM* contextualization

```
<!-- Libvirt stuff -->
<vbconfig>
  <vbpreparemethod>CDROM</vbpreparemethod>
  <vbhypervisor>XEN4</vbhypervisor>
  <vbtarfiles>
    <vbtarfile tarfile='alice.tar' path='.' strip='1' option='cdrom'/>
  </vbtarfiles>
</vbconfig>
<devices>
  <disk type='file' device='disk'>
    <!-- The boot disk -->
    <driver name='tap2' type='aio'/>
    <vbdisksource source='http://.../ucernvm-prod.1.18-13.cernvm.x86_64.fat' copy_method='download'/>
    <vbcheckhash hashfunction='/usr/bin/md5sum' hashvalue='826164d1748929445e805b57929c8f7e'/>
    <source file='/vmbatch/images/cernvm3.fat'/>
    <target dev='xvdb' bus='xen'/>
  </disk>
  <disk type='file' device='disk'>
    <!-- The 20GByte disk with the persistent CernVM-FS cache -->
    <driver name='tap2' type='vhd'/>
    <vbdisksource source='cernvm-hd.vhd' size='20480' create='always' format='vhd-util-vhd'/>
    <source file='/vmbatch/images/cernvm3-hd.vhd'/>
    <target dev='xvdaa' bus='xen'/>
  </disk>
  <disk type='file' device='disk'>
    <!-- CDROM with contextualization data -->
    <driver name='tap2' type='aio'/>
    <vbpreparecdrom source='context.iso'/>
    <source file='/vmbatch/images/context.iso'/>
    <target dev='xvdc' bus='xen'/>
    <readonly/>
  </disk>
  <!-- More libvirt devices stuff -->
</devices>
```

The *DISK* contextualization approach

- VmBatch will prepare the disk image(s) prior to starting the guest.
- The job, VmBatch parameters, tar files and init-scripts are copied to the image(s).
- VmBatch init-script are configured to run when the guest boots.
- The disk layout must be specified in the libvirt domain XML file:

```
<domain type='xen'>
  <!-- Libvirt stuff -->
  <vbconfig>
    <vbpreparemethod>DISK</vbpreparemethod>
    <vbjobpath>/torque/mom_priv/jobs</vbjobpath>
  </vbconfig>
  <devices>
    <vbdisks>
      <!-- If DISK contextualization, VmBatch must know the
           disk layout -->
      <vbdisk name='rootdisk' dev='xvda' partition='xvda2'/>
      <vbdisk name='jobdisk' dev='xvda' partition='xvda3'/>
    </vbdisks>
    <!-- Libvirt and VmBatch devices stuff -->
  </devices>
</domain>
```

The *ROOTSSH* contextualization approach

- Prior to using the guest with VmBatch, the guest must be set up to run VmBatch init scripts at boot.
- The guest is prepared for the job by using a root ssh-connection into the running guest.
- Username, group, uid and gid of submitting user, and some other VmBatch parameters can be transferred to the guest through the kernel boot line.

VmBatch image formats

- Unless configured read-only, a disk image can not be shared between simultaneously running virtual machines.
- For larger images, copy-on-write images must be used.
- Copy-on-write in XEN4 must use VHD.
  - The XEN4 documentation mentions also QCOW, but the format is not supported by the blktap2 driver.
- Copy-on-write in XEN3 must use QCOW created with XEN tools.
  - VMware images might be supported by blktap, but not by VmBatch.

Downloading images

- VmBatch can download images from the net:
  - Once for each job, or
  - once for each image.
- If downloading image once, the first job will lock the image while downloading.
  - All jobs will work with a copy or a copy-on-write version of the downloaded image, or the image must be readonly:

```
<disk type='file' device='disk'>
  <driver name='tap2' type='aio'/>
  <vbdisksource
    source='http://cernvm.cern.ch/releases/.../ucernvm-prod.1.18-13.cernvm.x86_64.fat'
    copy_method='shared'/>
  <source file='/vmbatch/images/ucernvm-prod.1.18-13.cernvm.x86_64.fat'/>
  <vbcheckhash hashfunction='/usr/bin/md5sum'
    hashvalue='826164d1748929445e805b57929c8f7e'/>
  <target dev='xvdb' bus='xen'/>
  <readonly/>
</disk>
```

Creating empty disk for the persistent CernVM-FS cache

- VmBatch can create empty disks and attach them to a guest:
  - Once for each job, or
  - once for each image.
- If creating disk only once, concurrent jobs will have to wait for the first job to finish and release the image.
  - All but the first job will work with a copy, or a copy-on-write version of the image.

```
<disk type='file' device='disk'>
  <driver name='tap2' type='vhd'/>
  <vbdisksource source='cernvm-hd.vhd'
    size='20480'
    create='once'
    format='vhd-util-vhd'/>
  <source file='/vmbatch/images/cernvm3-hd.vhd'/>
  <target dev='xvdaa' bus='xen'/>
</disk>
```

Allowing submitter to specify guest images

- VmBatch can allow the submitter to specify image(s) when submitting a job:

```
<disk type='file' device='disk'>
  <driver name='tap2' type='aio'/>
  <vbdisksource
    source='http://cernvm.cern.ch/.../ucernvm-prod.1.17-12.cernvm.x86_64.fat'
    copy_method='download' id='cernvm'/>
  <vbuserimage/>
  <vbcheckhash hashfunction='/usr/bin/md5sum'
    hashvalue='ec1b2ecde8c3cd9b897f18331833012b'/>
  <source file='/vmbatch/images/ucernvm-prod.1.17-12.cernvm.x86_64.fat'/>
  <target dev='xvdb' bus='xen'/>
</disk>
```

- Tag *vbuserimage* allow submitter to override the image path, as Torque *other* resources:

```
qsub -q vmbatch_short -S /share/vmbatch/bin/remoteshell \
  -l "other=vbuserimage=http://...1.18-13...fat#cernvm#826164d1748929445e805b57929c8f7e" \
  jobscript.sh
```

- AliEn LDAP for a CE can include guest image URLs as part of the *submitcmd* attribute.

CernVM3 with XEN – Selecting the boot target

- Libvirt ignores the boot tag for PV guests.
  - Not possible to specify the boot device or boot order.
- Both libvirt and CernVM will sort the targets:
  - Libvirt uses the first target as the boot disk, whereas
  - CernVM3 uses the first target as the persistent CernVM-FS cache.
- Must pick device names which CernVM3 and libvirt sort differently:

```
<disk type='file' device='disk'>
  <driver name='tap2' type='raw'/>
  <source file='/vmbatch/images/cernvm3.fat'/>
  <!-- The boot disk -->
  <target dev='xvdb' bus='xen'/>
</disk>
<disk type='file' device='disk'>
  <driver name='tap2' type='vhd'/>
  <source file='/vmbatch/images/cernvm3-hd.vhd'/>
  <!-- The persistent CernVM-FS cache -->
  <target dev='xvdaa' bus='xen'/>
</disk>
```

The XEN bootloader of SLC5 and CernVM3

- PyGrub wrongly identifies the CernVM3 image as a disk with MBR.
  - The CernVM3 XEN image is a FAT partition.
  - FAT partitions have a boot signature equal to that of MBR at address 0x1fe:0x200 $\rightarrow$ PyGrub identifies the FAT as an MBR.
- VmBatch is provided with a corrected XEN3 bootloader:

```
<domain type='xen'>
  <bootloader>/share/vmbatch/bin/pygrub-vmbatch</bootloader>
  <!-- More libvirt stuff -->
</domain>
```

- The VmBatch bootloader also corrects an error with the boot line parameters.
  - The XEN3 bootloader does not add the boot line parameters to the kernel line.

Some network comments

- VmBatch guests will usually use DHCP, but VmBatch can also configure guests with a static network.
- If network setup is static, a list of IPs must be provided, or VmBatch will use a sequence of successive IPs.
- Parameters for a static network will be copied to the disk image, the CDROM or transferred to the guest through the kernel boot line.
- A VmBatch init script will configure the network early in the boot process.

Conclusions

Conclusions

A realistic use of a lightweight tool for dynamic virtualization in a batch processing setup has been proven successful.

Further work

- Add support for SLURM.
- Add support for KVM and VirtualBox.
- Add option to allow a running guest to be reused for a new job.
- Add support for CentOS7 and SLC7.
- Improve the support for running jobs through the guest console.
  - Should also work with KVM.
  - Have only succeeded to read from a VirtualBox console.
- Add support for file staging.
  - Files can be staged on a disk image which is then attached to guest.
  - When job completes, the image is detached and output files fetched.
- Allow submitter to override more VmBatch options, e.g.:

```
qsub -l "other=domainxml=http://.../domain.xml#9d912309d453d9dd28e6c61c718351b1" \
     -q vmbatch_short -S /share/vmbatch/bin/remoteshell jobscript.sh
```

- More descriptive error messages.
- If DISK contextualization, add support for guests using LVM.
- More testing, e.g. sending jobs from AliEn to CernVM3 guests.
- Create a project web page.

VmBatch download

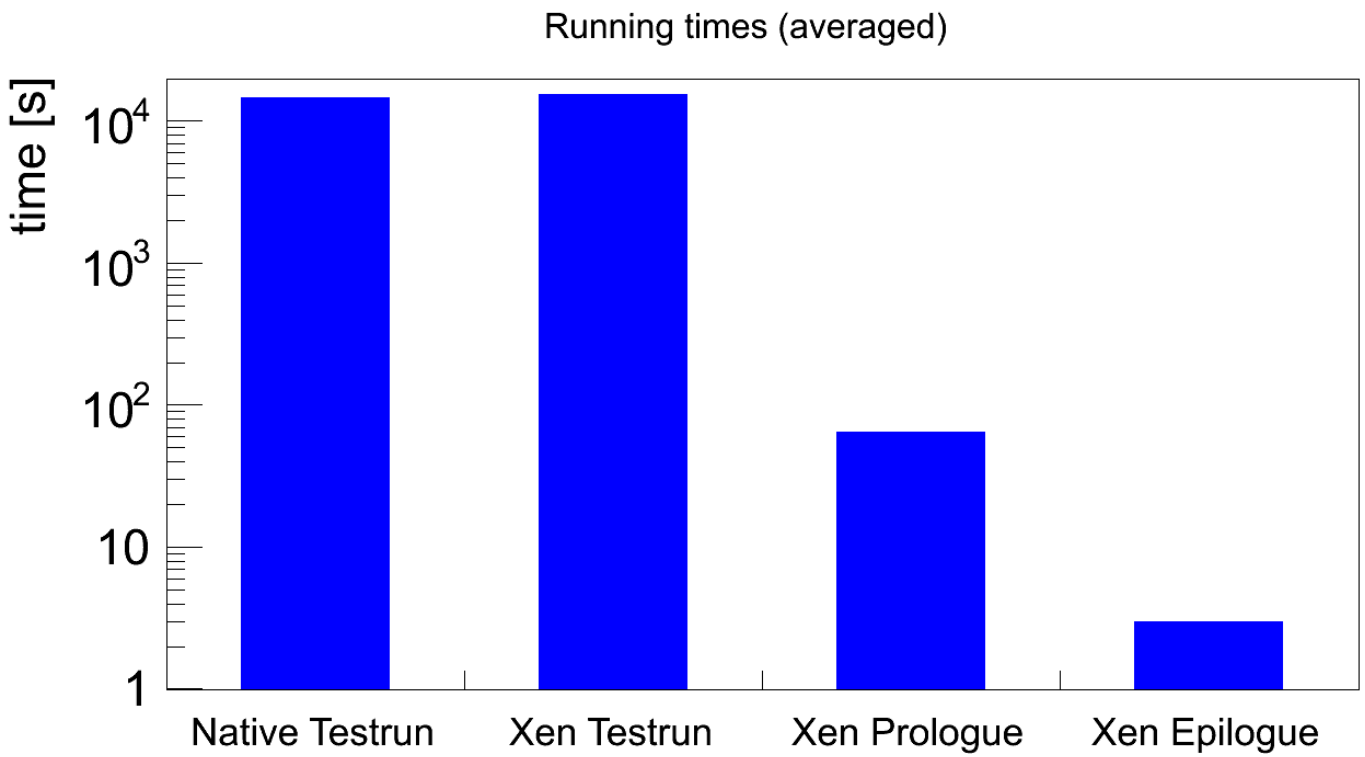
Subversion repository

http://eple.hib.no/svn/vmbatch/tags/latest

Backup – Measuring VmBatch performance

- VmBatch performance has been measured using (parts of) PbPbbench.
- Both host and guest OS are Scientific Linux 5.
- NFS caching was turned off.
  - With the default values, file attributes on a client may lag the NFS server for up to 60 seconds.

Backup – Vmbatch performance



- *Xen Prologue* is the time to setup the virtual environment.
- *Xen Testrun* is the time used to run the job.
- *Xen Epilogue* is the time to clean up after job completion.