Overview of a virtual cluster using OpenNebula and SLURM

Ismael Farfán Estrada
ifarfane0900@ipn.mx

October 27, 2011
Definitions

Objective
Use virtual machines as means for dynamic fractional resource management and load balancing in a batch cluster environment.

Dynamic Fractional Resource Management
Allow one or more physical nodes to execute different number of jobs depending on its load and the one of the cluster.

Load Balancing
Move jobs from busy nodes to idle nodes.
Software

**SLURM**
A resource manager, it’s job is to receive batch scripts (job requests) and allocate them in the available (virtual) nodes for execution.

**OpenNebula**
A virtual machine management service, makes it easy to work with VMs providing an easy way to create, destroy, migrate, contextualize, etc. virtual machines.

**Thiao**
An ad-hoc software developed to make SLURM and OpenNebula work together.
Physical connection

- The main/login server executes the SLURM resource manager and OpenNebula.
- The job of the worker nodes will be to host running VMs.
- The VMs execute the batch jobs.
- The VMs connect to the physical network using bridged connections.

**Figure: Client/server topology for the virtual cluster**
Why power saving mode?

Using SLURM’s power saving mode we can:

- Schedule jobs
- Launch new VMs when there are jobs waiting for execution
- Automatically shutdown idle VMs
- Share nodes/resources by launching many VMs
- Register the VMs to use

SLURM knows which VM is working and which isn’t
Enabling power saving mode

Slurm calls thiao’s “Resume” program when it needs to launch a VM and calls “Suspend” when it doesn’t need it any more

**slurm.conf**

```
SuspendProgram=/opt/thiao/bin/Suspend
ResumeProgram=/opt/thiao/bin/Resume
ResumeTimeout=180
SuspendTime=60
# Register 11 VMs
NodeName=fg[0-10] Procs=1 State=UNKNOWN
PartitionName=virtual Nodes=fg[0-10] State=UP
```
Resume program

1. Thiao’s Resume program receives the name of the VMs to launch from SLURM
2. Search in /opt/thiao/examples/ for a file named [vm-name].one
3. Asks OpenNebula to launch the VM
4. The VM says “hi” to SLURM
5. The VM starts to execute batch jobs

SLURM calls Suspend program to shuts down VMs when they are not needed any more
Resume diagram

Figure: The power saving mode tells which VM to resume to execute a job
Template file

The [hostname].one file is a regular OpenNebula VM template file

fg0.one
NAME = "fg0"
CPU = 1
MEMORY = 3000
DISK = [
    source = "/img/rhel6.img",
    target = "hda",
    readonly = "no",
    clone = "yes"
]
...

Job file

No modification needed in the job files

```
o5m.sh
#!/bin/bash
#SBATCH --time=30
#SBATCH --nodes=4
#SBATCH --ntasks-per-core=1
#SBATCH --job-name=o5m

date
cd /pub/user/
time mpirun ./o5m 180
date
```
Load balancing program

- OpenNebula monitors the hosts load
- Thiao’s “balancer” program retrieves the load
- balancer migrates running VMs from busy nodes to idle ones through OpenNebula

Figure: Load balancing
Load balancing diagram

Retrieve host/VM information using Remote Procedure Calls and balance the load by migrating VMs
Test environment

- A Debian SLURM and OpenNebula server
- 2 RHEL 6.0 hosts with 8-cores and 32Gb RAM (Thanks FutureGrid folks :)
- A shared NFS directory
- Debian/testing VM/KVM image
- A local SQLite database
Virtual cluster

Flexible schedule: 1 available VM

Overview of a virtual cluster

Ismael Farfán Estrada ifarfane0900@ipn.mx

October 27, 2011 14 / 16

```
308 virtual-f o5m ismael PD 0:00  4 (Priority)
309 virtual-f o5m ismael PD 0:00  4 (Priority)
310 virtual-f o5m ismael PD 0:00  4 (Priority)
311 virtual-f o5m ismael PD 0:00  4 (Priority)
312 virtual-f o5m ismael PD 0:00  4 (Priority)
300 virtual-f o5m ismael R 0:35  4 fg[0-3]
```

```
ismael@kriemhild:~$ sinfo
PARTITION AVAIL TIMELIMIT NODES STATE NODELIST
virtual-* up 3:00:00 17 down* fg[4-6,8-10],vm[2-12]
virtual-* up 3:00:00 1 idle~ fg7   There's no use on starting this one
virtual-* up 3:00:00 4 alloc fg[0-3] Only 4 node jobs in the queue
```

```
ismael@kriemhild:~$ onevm list
ID USER NAME STAT CPU MEM HOSTNAME TIME
230 oneadmin fg-0 runn 183 2.9G s80r.idp.sdsc.f 00 00:13:02
231 oneadmin fg-1 runn 210 2.9G s81r.idp.sdsc.f 00 00:13:03
232 oneadmin fg-2 runn 209 2.9G s80r.idp.sdsc.f 00 00:13:02
233 oneadmin fg-3 runn 208 2.9G s81r.idp.sdsc.f 00 00:13:01
```

```
ismael@kriemhild:~$
```

```
ismael: bash
localhost (QEMU)
```

```
Name CPU usage Host CPU usage Disk I/O Network I/O
```

```
localhost (QEMU)
```

```
s80r.idp.sdsc.futuregrid.org (QEMU)
```

```
one-230 Running
```

```
one-231 Running
```

```
one-232 Running
```

```
one-233 Running
```

```
s81r.idp.sdsc.futuregrid.org (QEMU)
```

Overview of a virtual cluster
Submit some 1 node jobs

The 1 node jobs can be backfilled by starting the dormant VM

We have to wait a little while it boots

Backfilled!
More information

Thiao’s project page
https://github.com/scarmiglione/thiao

Screenshots
https://github.com/scarmiglione/thiao/tree/master/doc/screenshots

SLURM
http://www.schedmd.com/

OpenNebula
http://opennebula.org/