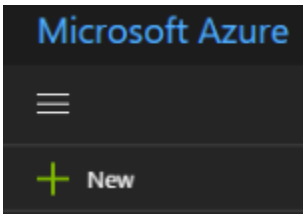
Exercise 5: Deploy Linux RDMA InfiniBand cluster with Azure CLI

Duration: 45 minutes

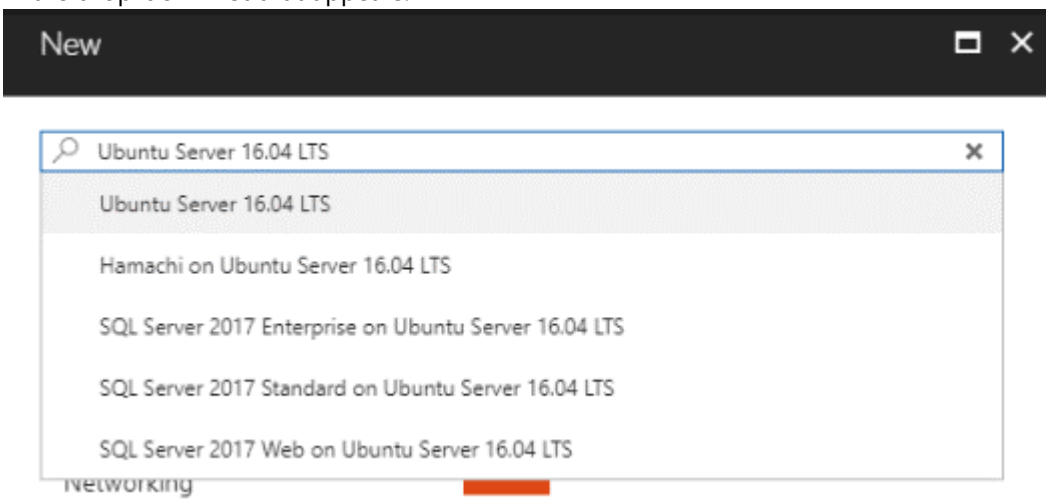
Preparation: Deploy Azure CLI Linux box

In this task you will deploy a Ubuntu 16.04 LTS VM and install Azure CLI in it, as prerequisite for executing further steps of the exercise.

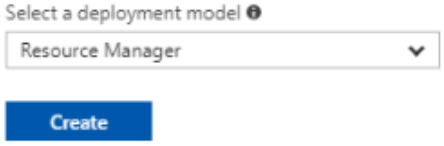
1. Using a browser, navigate to the Azure Portal.
2. Select **+ New** from the menu.



3. In the Search the Marketplace text box, enter "**Ubuntu Server 16.04 LTS**" and select the same in the drop-down list that appears.



4. On the Ubuntu Server 16.04 LTS blade, leave the selection of **Resource Manager** for Select a deployment model, and select **Create**.

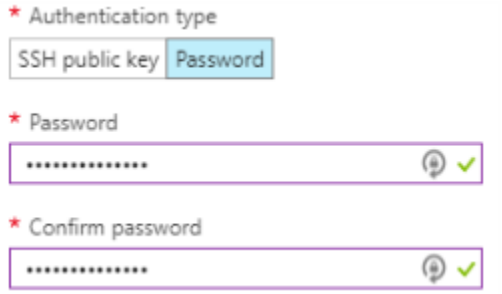


5. On the Basics blade, enter a **Name** for the new VM.

6. For VM disk type, select **HDD**.

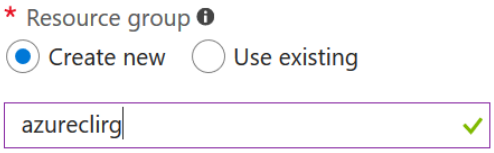
7. For User name, enter a username, e.g. **labuser**.

8. For Authentication type, select **Password** and enter a password in the fields that appear. Take note of the user name and password you supply here, as you will use it later to SSH into the VM.



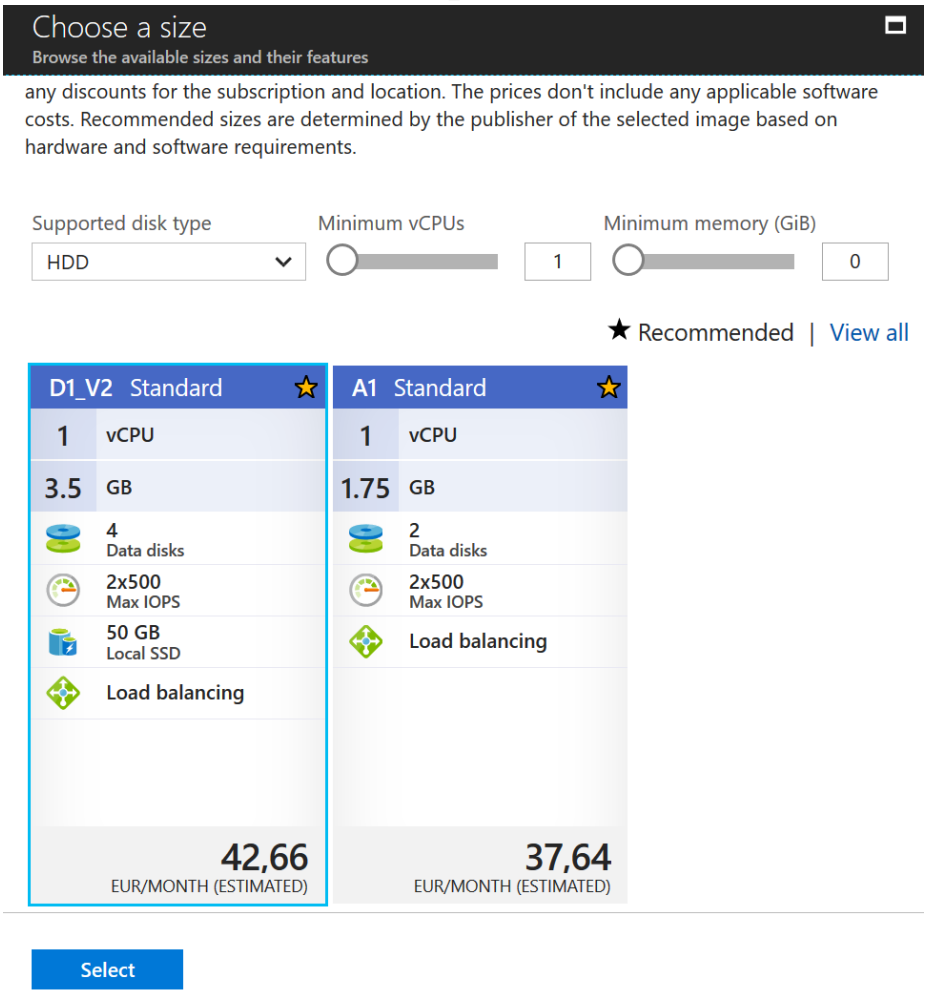
9. Choose your **Azure Subscription**.

10. For the Resource Group, select **Create new** and type the name, e.g. **azureclirg**.

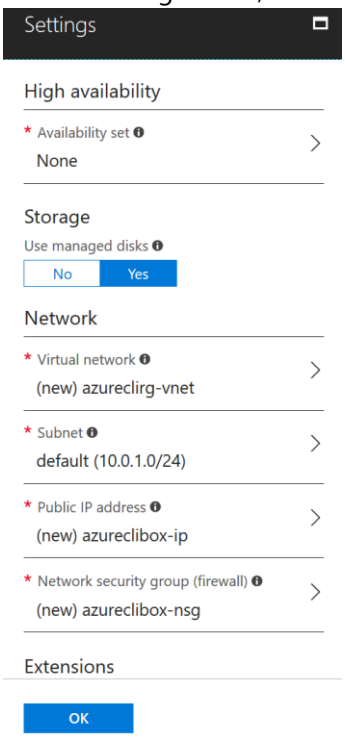


11. Choose a **location,** e.g. **West Europe**.
12. Select **OK**.

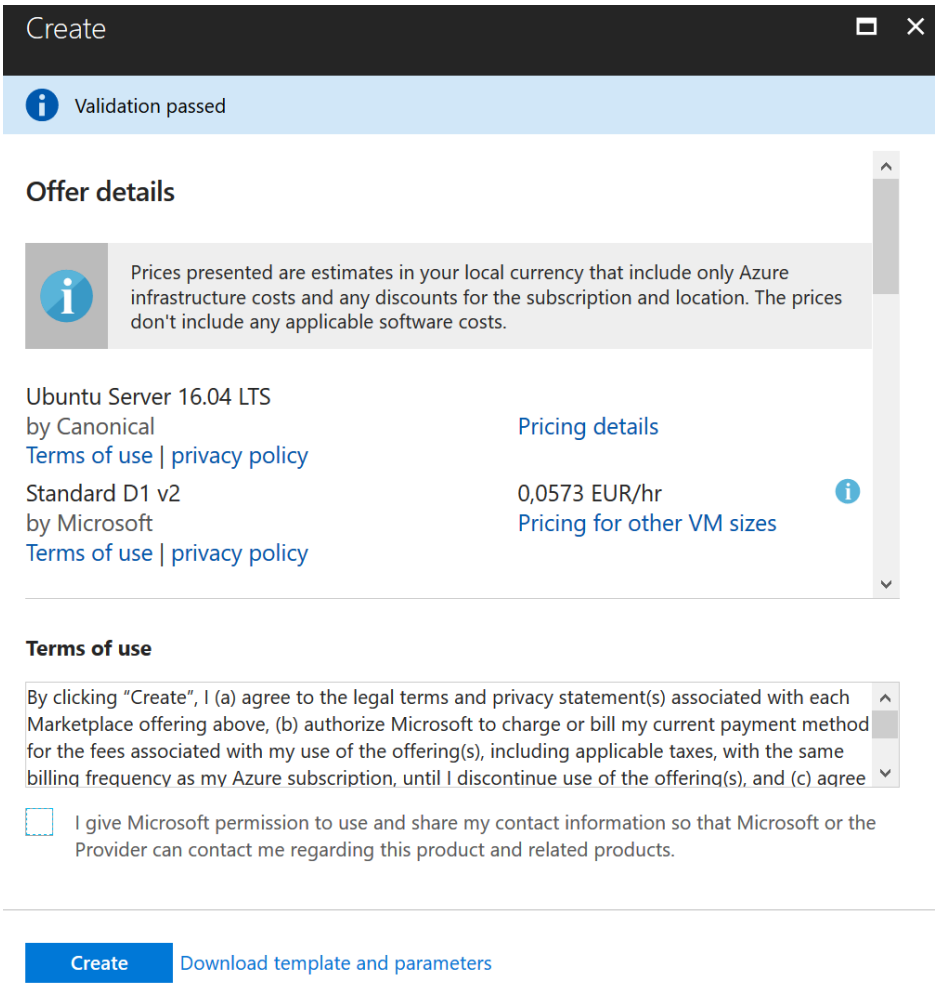
13. On the Choose a Size blade, select **D1_V2 Standard** and then click **Select**.



14. On the Settings blade, leave the defaults and select **OK**.



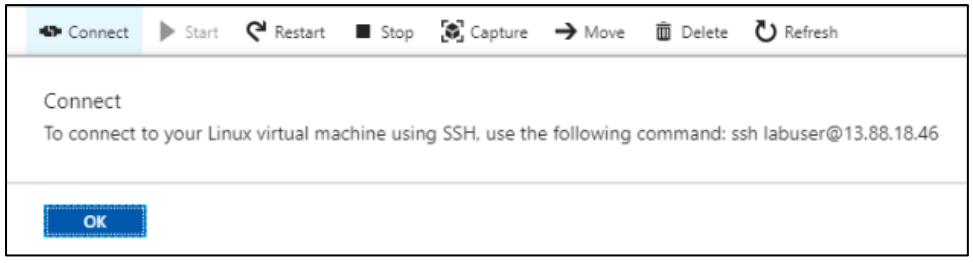
15. On the Create blade, select **Create**.



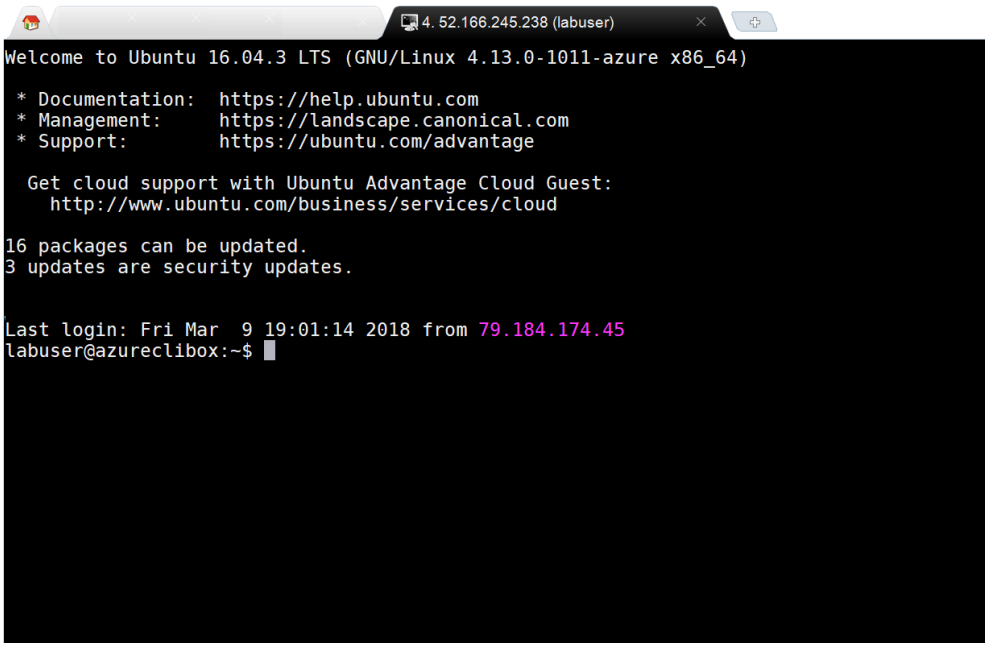
It will take about 3-5 minutes to deploy the VM.

16. Once the VM is ready, the blade for the VM will open automatically in the Azure Portal.
17. In the control bar of the VM blade, select **Connect**.

18. A dialog will appear showing the SSH command line to use to connect to the VM. Take note of the command, it includes the username (labuser in the below) and IP address (13.88.18.146 in the below) that will be used to access the VM.

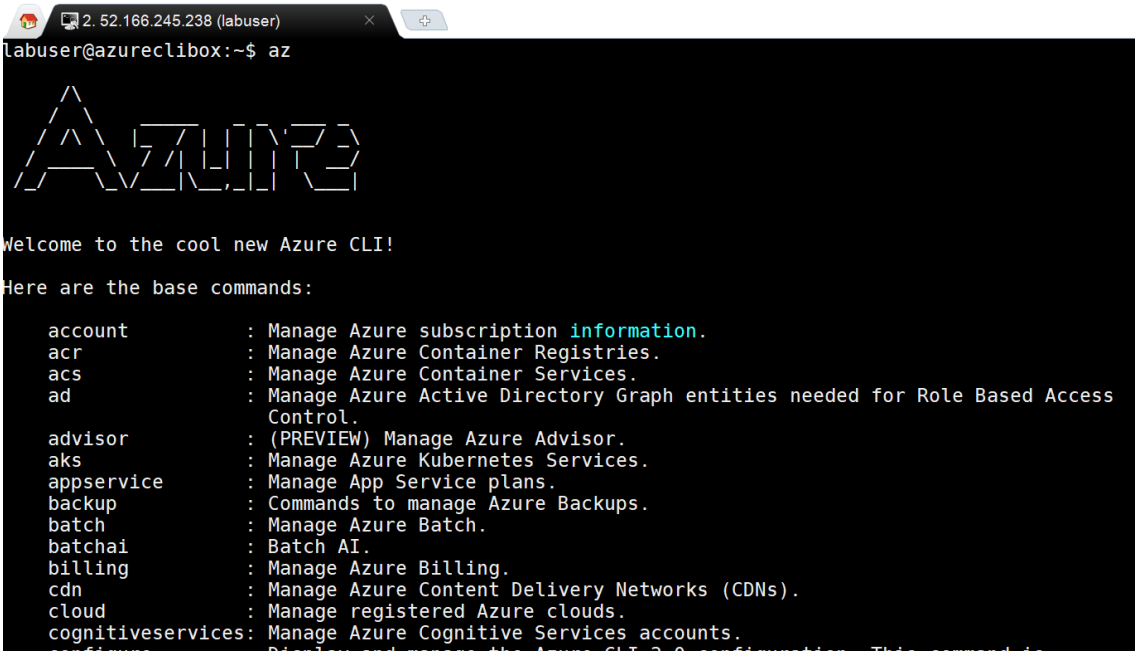


19. Using your favorite SSH client, login into the VM. Be sure to provide the username and password you specified when creating the VM.



20. Navigate to https://docs.microsoft.com/en-us/cli/azure/install-azure-cli-apt?view=azure-cli-latest and follow the instructions to install Azure CLI in your Ubuntu VM.
21. Verify that the Azure CLI is installed by running:
    `$ az`

    If you see output similar to the following the install was successful.



22. Login to your account with Azure CLI
    `$ az login`
    Follow the provided instruction – go to URL https://microsoft.com/devicelogin in a web-browser and enter the provided code string. Then login to your Microsoft account and close the web-browser page.
    After successful login you should see output similar to the following:

```
labuser@azureclibox:~$ az login
To sign in, use a web browser to open the page https://microsoft.com/d
[
  {
    "cloudName": "AzureCloud",
    "id": "1e0a1fe7-2e53-4432-8193-e2df864bb693",
    "isDefault": true,
    "name": "Azure \u2014 dost\u0119p pr\u00f3bny",
    "state": "Enabled",
    "tenantId": "91c37ad6-33d9-41f5-844c-0554c59eabba",
    "user": {
      "name": "[redacted]@outlook.com",
      "type": "user"
    }
  }
]
labuser@azureclibox:~$
```

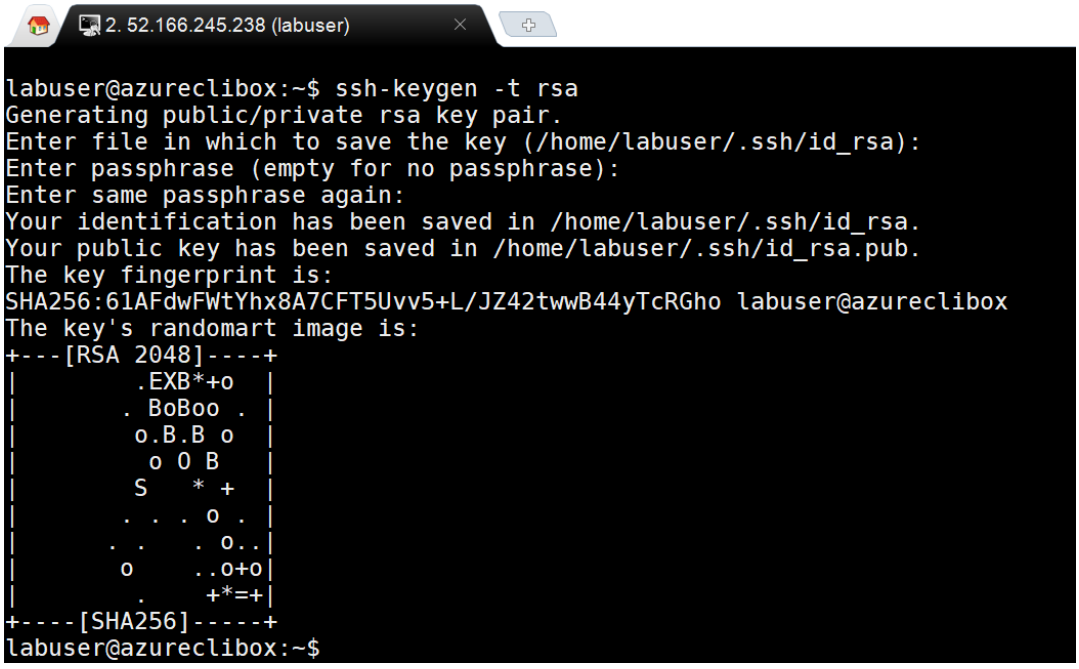
Now using Azure CLI commands you will deploy a simple CentOS7.1 cluster with RDMA InfiniBand interconnect. The cluster will be deployed using Azure VM scaleset service.

Step 1: Deploy CentOS7.1 cluster with RDMA InfiniBand

1. Open your favourite SSH client and login to Azure CLI Linux box.
2. Generate private/public SSH keys

   ```
   $ ssh-keygen -t rsa
   ```

   Press [ENTER] on all prompts.



3. Create resource group for the cluster resources:

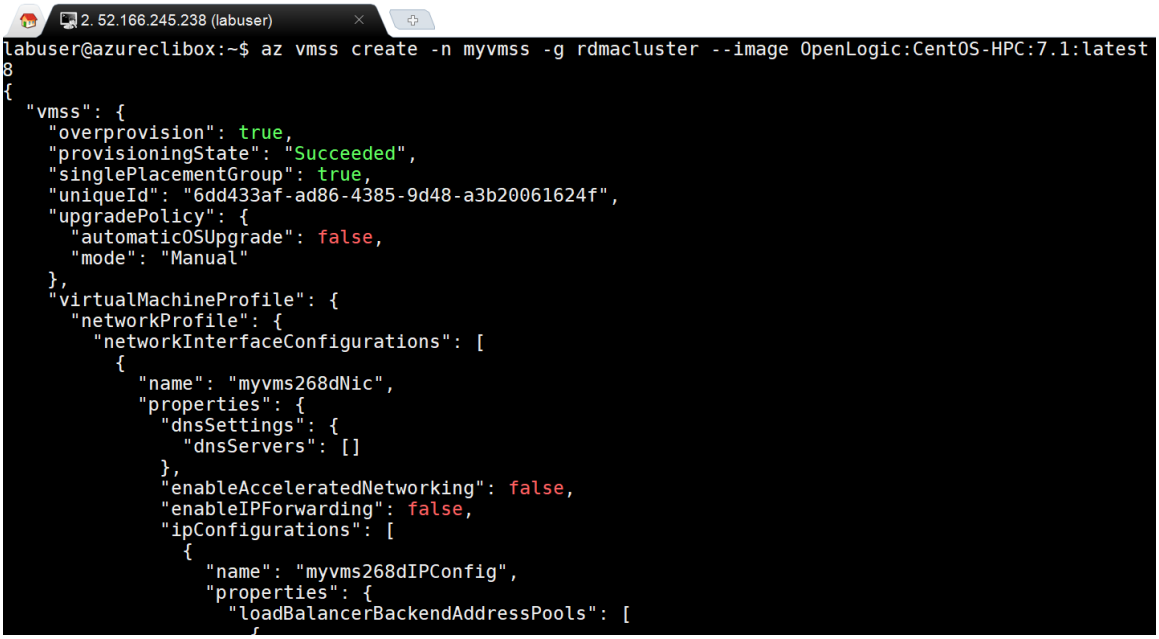
   ```
   $ az group create -n rdmacluster -l westeurope
   ```

Note: The following steps will fail with your Azure trial subscription due to low quota for RDMA enabled VMs. You will need a paid Azure subscription to be able to raise the quota for RDMA InfiniBand enabled VMs and execute the rest of this exercise. Please refer to the instructor on how to continue.

4. Create VM scaleset from Azure Marketplace CentOS7.1-HPC OS image:
   $ az vmss create -n myvmss -g rdmacluster --image OpenLogic:CentOS-HPC:7.1:latest --vm-sku Standard_H16r --instance-count 4

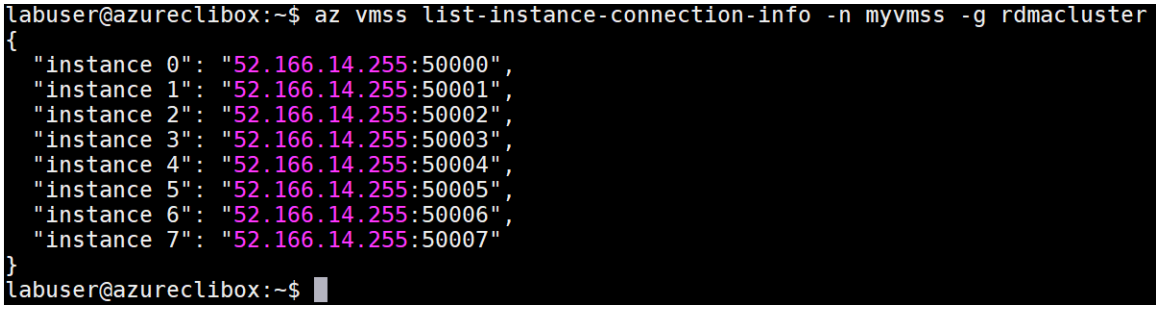
   Here is description of the used command parameters:
   -n myvmss – name of VM scaleset
   -g rdmacluster – resource group for deployment
   --image OpenLogic:CentOS-HPC:7.1:latest – the latest CentOS7.1 OS image with preinstalled InfiniBand drivers and Intel MPI library runtime
   --vm-sku Standard_H16r – InfiniBand-enabled VM instance H16r
   --instance-count 4 – number of VMs in the scaleset

   It will take about 2 minutes to deploy the VM scaleset. Your public SSH key from *~/.ssh/id_rsa.pub* on the Azure CLI host will automatically be deployed to the scaleset as an authorized key for accessing the cluster nodes.
   On successful creation of the VM scaleset you should see output similar to the following:



5. Display connection information for the VM scaleset:

```
$ az vmss list-instance-connection-info -n myvmss -g rdmacluster
```



Your scaleset cluster is deployed into a private network. Together with the scaleset Azure by default has created a load-balancer with public IP and NAT mapping for ports from 50000 upwards to subsequent nodes in the scaleset.

Note down public IP address of the load balancer which will be needed in the following steps to access the scaleset custer nodes.

6. Test access to your scaleset cluster by SSH login to any of the nodes:

```
$ ssh 52.166.14.255 -p 50000
$$ exit
$
```

Step 2: Configure SSH cluster connectivity

7. Run the below script *sshkey_copy.sh* on your Azure CLI box to distribute your private SSH key to the scaleset nodes:

sshkey_copy.sh

```
#/bin/bash
# Usage ./sshkey_copy.sh <publicip> <#nodes>

PIP=$1
COUNT=$2

if [ $# != 2 ]; then
    echo "Usage: $0 <publicip> <#nodes>"
    exit 1
fi
```

```
# compose SSH config file with strict host key checking turned off
echo 'Host *' > /tmp/config
echo 'StrictHostKeyChecking no' >> /tmp/config
chmod 400 /tmp/config

# copy private key and config to scaleset nodes
for (( i=50000; i<50000+$COUNT; i++ )); do
  scp -o StrictHostKeyChecking=no -P $i ~/.ssh/id_rsa $PIP:.ssh/
  scp -o StrictHostKeyChecking=no -P $i /tmp/config $PIP:.ssh/
done

rm -f /tmp/config
```

```
$ sshkey_copy.sh 52.166.14.255 8
```

Now your scaleset cluster nodes all share the same private key and can communicate with one another over SSH using passwordless key authentication.

Step 3: Test RDMA InfiniBand connectivity within the scaleset cluster

8. Login to any node in the scaleset, e.g. on port 50000:
   ```
   $ ssh 52.166.14.255 -p 50000
   ```

9. Install nmap package on the compute node:
   ```
   $$ sudo yum install -y nmap
   ```

10. Generate the list of cluster node IPs with the following command:
    ```
    $$ nmap -sn 10.0.0.* | grep 10.0.0. | awk '{print $5}' | tee nodeips.txt
    ```
    The list of node IPs is stored in file nodeips.txt in the home directory.

11. Use the below script *pingpong.sh* to run Intel MPI pingpong test between each pair of cluster nodes:

pingpong.sh

```
#!/bin/bash
# Example usage: ./pingpong.sh | grep -e ' 512 ' -e NODES -e usec

source /opt/intel/impi/`ls /opt/intel/impi`/bin64/mpivars.sh

for NODE in `cat ~/nodeips.txt`; do
    for NODE2 in `cat ~/nodeips.txt`; do
        echo '################################################'
&& \
        echo NODES: $NODE, $NODE2 && \
        echo '################################################'
        mpirun -hosts $NODE,$NODE2 -ppn 1 -n 2 \
            -env I_MPI_FABRICS=dapl \
            -env I_MPI_DAPL_PROVIDER=ofa-v2-ib0 \
            -env I_MPI_DYNAMIC_CONNECTION=0 \
            IMB-MPI1 pingpong
    done;
done
```

Run the following command:

```
$$ ./pingpong.sh | grep -e ' 512 ' -e NODES -e usec
```

You should expect the following output:

```
[azureuser@myvmse751000000 ~]$ ./pingpong.sh | grep -e ' 512 ' -e NODES -e usec
NODES: 10.0.0.4, 10.0.0.4
       #bytes #repetitions      t[usec]   Mbytes/sec
          512          1000         1.92       265.98
NODES: 10.0.0.4, 10.0.0.5
       #bytes #repetitions      t[usec]   Mbytes/sec
          512          1000         2.94       174.01
NODES: 10.0.0.4, 10.0.0.6
       #bytes #repetitions      t[usec]   Mbytes/sec
          512          1000         2.92       175.46
NODES: 10.0.0.4, 10.0.0.7
       #bytes #repetitions      t[usec]   Mbytes/sec
          512          1000         2.96       173.23
NODES: 10.0.0.4, 10.0.0.8
       #bytes #repetitions      t[usec]   Mbytes/sec
          512          1000         2.92       175.64
NODES: 10.0.0.4, 10.0.0.10
       #bytes #repetitions      t[usec]   Mbytes/sec
          512          1000         2.89       177.13
```

with latency measurements ~3usec proving RDMA InfiniBand connectivity between the nodes.

12. Exit the cluster node:
    `$$ exit`

Step 5: Clean up resources

13. As the last step of the exercise clean up the cluster by deleting the resource group where it belongs:
    `$ az group delete -n rdmacluster`