SETTING UP AN EXPERIMENTAL LINUX BASED CLUSTER SYSTEM UNDER SLURM WORKLOAD MANAGER FOR NUMERICAL WEATHER PREDICTION

Prepared by: Farah Ikram

Endorsed by:

Dr. Jehangir Ashraf Awan

Dr. Muhammad Tahir Khan

Research and Development Division, Pakistan Meteorological Department,
Pitras Bukhari Road, Sector H-8/2, Islamabad, 44000

Technical Report No.: PMD-11/2019 November, 2019

Preface

This report presents setting up of a 3 node Linux based high performance cluster (HPC) system for running numerical weather prediction (NWP) models on Pakistan Meteorological Department (PMD) HPC. PMD is equipped with a 32 node HPC cluster system. In this experiment 3 nodes were utilized for setting up a separate cluster for running a NWP model. The purpose for this experiment is to understand the working and troubleshooting of the new workload manager deployed in the newly installed HPC under Specialized Medium Range Forecast Centre. This experimental setup of cluster system consists of a node as master node and two other nodes as compute nodes. Linux operating system is installed on all three nodes. The scientific libraries required by the NWP model to run on the cluster such as DWD-libgrib1 DWD library, ECMWF GRIB API, INT2LM source code, COSMO-Model source code, C-Compiler and Fortran Compiler, HDF5 library, Jasper, grib-api library, NetCDF, NetCDF Fortran and Open MPI are shared through network file system from the master node. The security realm used in this system is Munge Uid 'N' Gid Emporium which enables user id and group id authentication across a host cluster hence a service for creating and validating credentials in the HPC environment. (Simple Linux Utility for Resource Management (SLURM workload manager) for batch processing and scheduling software is installed to manage the parallel processing. It is an open source scalable cluster management and job scheduling software for Linux based cluster. C*onsortium for Small-Scale Modeling* (COSMO) is also compiled and simulated on this setup. The *COSMO-Model* is a nonhydrostatic limited-area atmospheric prediction model. It has been designed for both operational numerical weather prediction (NWP). The author would also thank for the technical guidance of Mr. Syed Ahsan Ali Bukhari in installation of COSMO Model on this cluster.

Executive Summary

The Pakistan Meteorological Department (PMD) has made significant progress in up gradation of its High Performance Computing (HPC) facility in the recent years. PMD reached a milestone of deploying its first HPC back in 2006. The specifications of this system included HP rackmount server with peak performance 0.2 TFLOPS with nine compute nodes and two cores per node. The next generation HPC was installed in 2009. This generation of HPC are DELL PowerEdge Blade Servers with peak performance of 1.7 TFLOPS having 32 Compute Nodes and 8 cores per node. The third generation of HPC cluster system is installed in 2018 under SMRFC project by JICA. This cluster has 24 nodes, 28 cores per node with 56 Gbps infiniband connectivity. It has peak performance of 22 TFLOPS and utilizes Simple Linux Unitility for Resource Manager (SLURM) workload manager for job scheduling and job management. In order to gain skill over this new HPC cluster for numerical weather prediction (NWP) and other climate modelling tasks, an experimental setup was configured on the previous PMD HPC so that the new cluster's configuration should not be troubled. To perform this task a 3 node experimental cluster was set up on the 2nd generation PMD HPC with one master/ head node and the other two as client nodes. The configuration of this HPC cluster involves seven main steps 1) Linux installation, 2) network configuration, 3) installing the security realm (Munge Uid 'N' Gid Emporium), 4) installation of SLURM workload manager, 5) Secure shell (SSH) without password, 6) configuration of the network file system (NFS) for shared libraries between master and client nodes and lastly 7) the compilation and running of the NWP COSMO model. The resulting cluster from this setup is flexible and is also applicable to set of computers whereas SLURM is an excellent, easy to use workload manager for managing parallel jobs as well as for resource management on the HPC. This report will also serve as a technical guidance for the new third generation cluster system which has been recently deployed.

Table of Contents

List of Figures

List of Tables

Introduction

A cluster computer is used to solve large computational problems which are resource intensive. The idea behind the computer cluster is staking multiple compute nodes hence, more cores more programs it can handle simultaneously. Every computer has a processor with multiple cores. Each core has multiple threads which can handle a process [(1)]. A cluster computer consist of one head/master node which sends instructions to the rest of the cluster nodes (compute node) through an isolated secure network [(6,7,8)]. High performance computing clusters are designed to use the parallel processing of multiple nodes by splitting a computational task among the nodes [(3)]. The difference between serial and parallel computing is that serial computing breaks down the computational problem into discrete set of instructions which are executed in a row sequentially. These instructions are executed on single processor with one instruction being processed at a time (figure 1a.). In contrast to serial computing, the parallel computing is simultaneous utilization of available multiple computational resources process the instruction. Therefore, a computational problem/ instruction is fragmented into discrete parts that can be solved synchronously with isntruction further split into to series of instructions. These instructions from each discrete part is executed concurrently on different processors/CPUs (figure 1b). In parallel computing clusters such as high performance computing clusters, each compute node has multi-processors which is a parallel computer, such multiple nodes are networked together with an infiniband [(13)]. The cluster architecture discussed in this paper is an asymmetric type of architecture, a master node acts as a gateway between the nodes and user. It provides high level of security as all the traffic passes through the head node. Another advantage is that the remaining nodes has minimal software hence are exclusively utilized by the cluster [(3)].

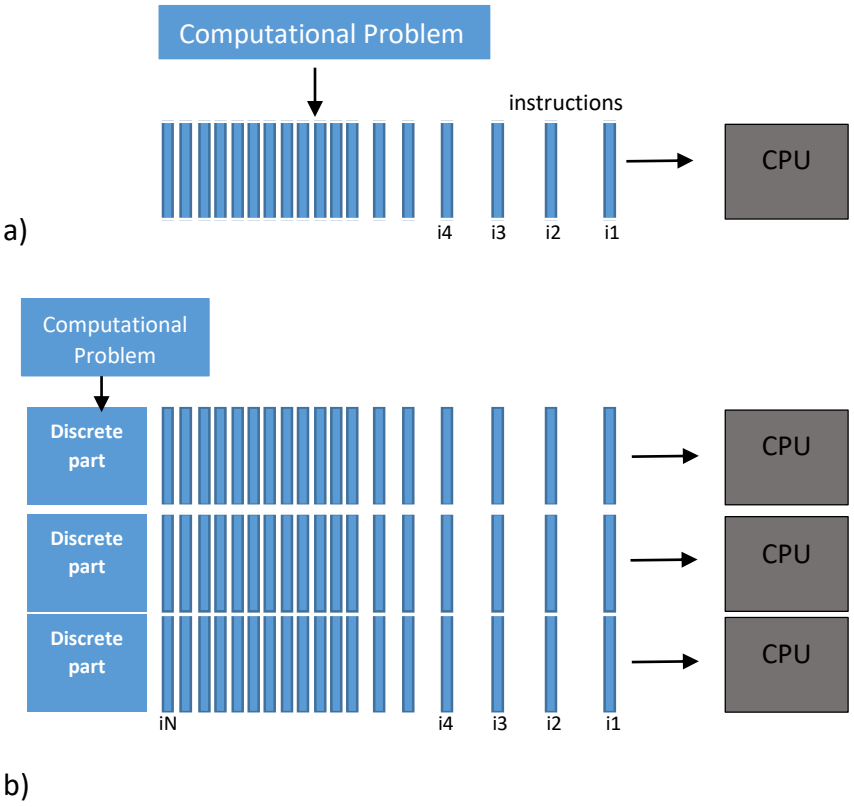

Figure 1: Serial (a) and parallel (b) computing

Parallel computing has therefore, a major application in ordinary world where complex and interrelated events which are concurrent such as weather, climate, climate change, planetary movements, plate tectonics, galaxy creation/destruction etc[13].

PMD's High Performance Computing Facility

Pakistan Meteorological Department is equipped with three generations of High Performance Computing Clusters (HPCC). The first generation of HPCC was installed in 2006. The specifications of this system include HP rackmount server with peak performance 0.2 TFLOPS with nine compute nodes and two cores per node. The storage capacity of this system is 12TB with

Gigabit Ethernet connectivity. The next generation HPC was installed in 2009. This generation of HPC are DELL PowerEdge Blade Servers with peak performance of 1.7 TFLOPS having 32 Compute Nodes and 8 cores per node. The storage capacity of this system is 50TB. The HPCC has 20Gbps Infiniband connectivity and storage has 8Gbps fiber channel connectivity. The cluster is managed through Gigabit Ethernet. The third generation of HPC cluster system is installed in 2018 under SMRFC project by JICA. This cluster has 24 nodes, 28 cores per node with 56 Gbps infiniband connectivity for HPCC and storage is connected through 10Gbps Ethernet. It has peak performance of 22 TFLOPS and utilizes SLURM workload manager for job scheduling and job management.

Numerical Weather Prediction Model

The brief history of the numerical weather prediction proposed by L.F. Richardson in 1922. His idea was forecasting weather based on time integration of basic equations of fluid mechanics that express the atmospheric circulation. In the recent years, most meteorological organizations in the world predict weather based upon numerical simulations of the equations representing global atmospheric circulations. For example, the numerical model used in Japan consists of three models: a meso-scale model which covers the region of East Asia with horizontal resolution of 10 km is nested which is a regional model of 20 km resolution. This regional model is nested with global model which has 50 km resolution. The short range forecast is based on a single run with the finest initial conditions, however the long range(1 month) forecast is based on ensemble mean of 26 different ensemble members with slightly different perturbations to initial conditions (ensemble forecast) [16] . Similarly several developed countries have established numerical weather prediction models of their own while the developing countries are using those models for weather forecasting in their region. Pakistan Meteorological Department is also among the end user of such

models and have been using the output of such models for generating the weather forecast over Pakistan. One of the NWP models is the C*onsortium for Small-Scale Modeling* (COSMO) model. The main purpose of COSMO Model is to develop, advance and sustain a non-hydrostatic limited-area atmospheric model, useful for both operational and research applications by all the members of the consortium and other licensed institutions. The emphasis made on providing accurate simulation of small-scale physical processes by using latest physical parametrizations. Presently, observations are assimilated by emplying a nudging approach. The **Consortium** was created in October 1998, at the annual meeting of DWD (Germany) and MeteoSwiss (Switzerland). Following year, a Memorandum of Understanding (MoU) for scientific collaboration in non-hydrostatic modeling was signed by the Directors of DWD (Germany), MeteoSwiss (Switzerland), ReMet (Italy) and HNMS (Greece) in spring 1999. [4,5]

Table 1: COSMO: Participating Meteorological Services

DWD	Deutscher Wetterdienst, Offenbach, Germany
MeteoSwiss	Meteo-Schweiz, Zurich, Switzerland
USAM	Ufficio Generale Spazio Aero e Meteorologia,Roma, Italy
HNMS	Hellenic National Meteorological Service,Athens, Greece
IMGW	Institute of Meteorology and Water Management, Warsaw, Poland
ARPA-SIMC	Agenzia Regionale per la Protezione Ambientale dell Emilia-Romagna Servizio Idro Meteo Clima, Bologna, Italy
ARPA-Piemonte	Agenzia Regionale per la Protezione Ambientale, Piemonte, Italy
CIRA	Centro Italiano Ricerche Aerospaziali, Italy
ZGeoBW	Zentrum f¨ur Geoinformationswesen der Bundeswehr, Euskirchen, Germany
NMA	National Meteorological Administration, Bukarest, Romania
RosHydroMet	Hydrometeorological Centre of Russia, Moscow, Russia

COSMO Model

The COSMO Model is a non-hydrostatic limited-area numerical weather prediction model of the atmosphere. It is designed for use in operational numerical weather prediction (NWP) and several other scientific/ research applications involving meso-β and meso-γ scale. The COSMO-Model is based on the primitive thermo-hydrodynamical equations representing compressible fluid flow of a moist atmosphere. The model equations are developed in rotated geographical coordinates with generalized vertical height coordinate system using terrain following vertical coordinate formulation. A wide range of physical processes are accounted for and represented by parameterization schemes. The basic version of the COSMO-Model (formerly known as *Lokal Modell (LM)*) has been developed at the *Deutscher Wetterdienst* (DWD). The COSMO-Model and the global grid point triangular mesh model GME form (in conjunction with the conforming data assimilation schemes) which are part of the Numerical Weather Prediction setup at the DWD, is operational since end of 1999. New developments/upgradations related to the model are structured within COSMO, the *Consortium for Small-Scale Modeling*. The COSMO-Model is provided free of charge for scientific research and academic purposes, particularly for COSMO members [4,5]

MUNGE Uid 'N' Gid Emporium

The security software used in this setup is MUNGE Uid 'N' Gid Emporium. This software enables uid (user id) and gid (group id) authentication across a host cluster hence a service for creating and validating credentials and is designed especially for use in HPC cluster environment. The process involves authentication of UID and GID of another remote or local process within a group of hosts having common users and groups. These hosts then form a security realm defined by a cryptographic key that is shared between these munged hosts

within the security realm. The clients therefore, within this secure realm can authenticate credentials without having the root privileges, reserving specific ports or methods implemented for particular platform. The integrity of the credentials is ensured by message authentication code (MAC). The decoding of the credential can be restricted to particular user or group ID. Internal format of these credentials are encoded in platform independent manner and are based64 encoded to allow it to be transmitted over and transport [9,10]. Along with the message passing interface software (MPI) a workload manager (SLURM) for batch processing and scheduling is installed.

SLURM Workload Manager

The acronym stands for Simple Linux Utility for Resource Management, developed by Lawrence Livermore National Laboratory in 2002 as resource manager. Its code is written in C language. It is used on many of the world's largest computers with an active global development community [10]. SLURM is an open source scalable cluster management and job scheduling software for Linux based clusters. There are three key functions of Slurm, allocation of exclusive or non-exclusive access of resources (compute nodes) to users to perform tasks over a duration of time. Secondly, it provides a framework for starting, executing and monitoring e.g parallel job on the set of allocated compute nodes. Lastly, it determines the distribution of resources by managing the pending job queue.

The architecture of this software has a centralized manager called slurmctld to monitor resources and work. The compute nodes has a slurmd daemon which takes tasks, execute them, returns status and waits for next task. Therefore slurmd provide a fault tolerant hierarchical communications. Some of the user tools of slurm include **srun** to start job, **scancel** to terminate queued or running job, **sinfo** for system status e.g nodes status and **squeue** to check task status

whereas **sbatch** is used to submit batch (parallel) jobs. An administrative **scontrol** tool is available to monitor or modify configuration and state of the cluster system [11]. Figure 2 shows the slurm components.

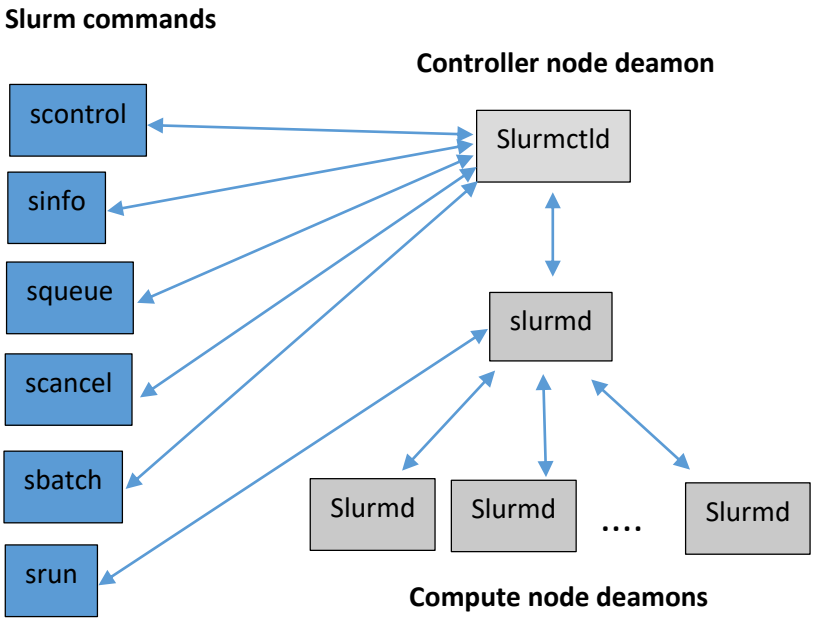

Figure 2: SLURM components

Methodology

The PMD HPC 2nd generation cluster has total 32 nodes with total memory of 256 GB. Each node has Intel(R) Xeon(R) CPU X5470 @ 3.33GHz. Total CPU cores on each node are 8 with 8 GB RAM. Each CPU has 2 quad core processors. Each node is connected through gigabit Ethernet as well as Infiniband DDR and fiber channel FC 8 Gbps. The linux distribution used for development of this cluster system is CENTOS 7, however, any linux distribution can be used for this purpose. Figure 3 illustrates the process chart of the setps involved in building the cluster and Figure 4 illustrates the layout of the experimental setup presented in this study.

- Configuration of the master node

The configuration of master node involves following steps

1) Linux installation
2) Setting up network and hostname
3) Installing Munge Uid 'N' Gid Emporium
4) Installing Slurm Workload manager
5) SSH without password
6) Configuring NFS

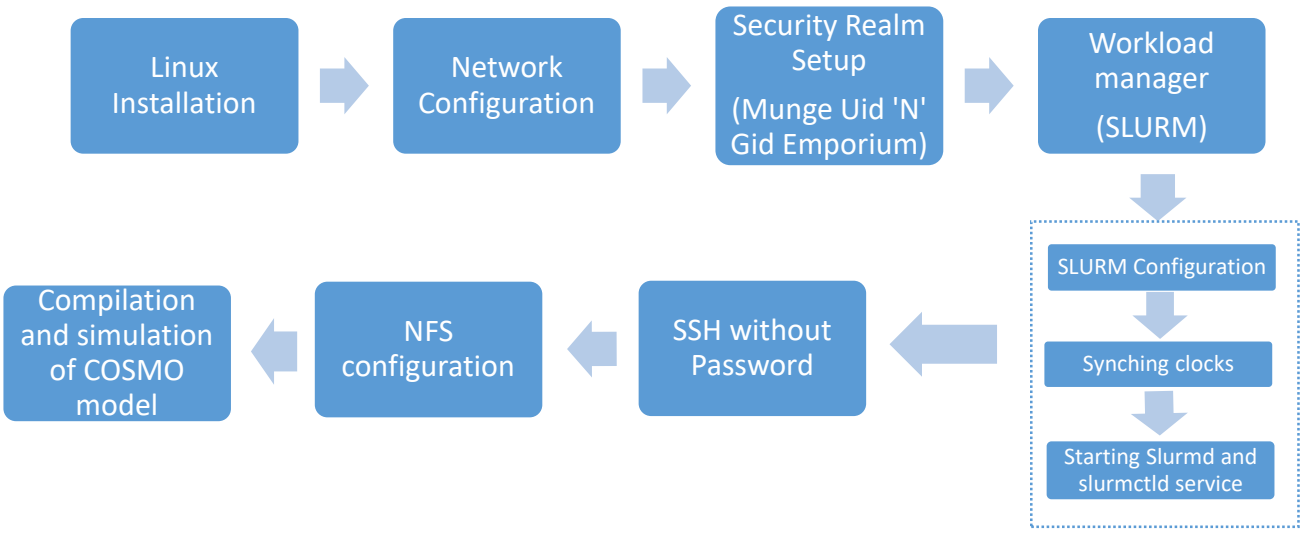

Figure 3: Process chart for HPC configuration

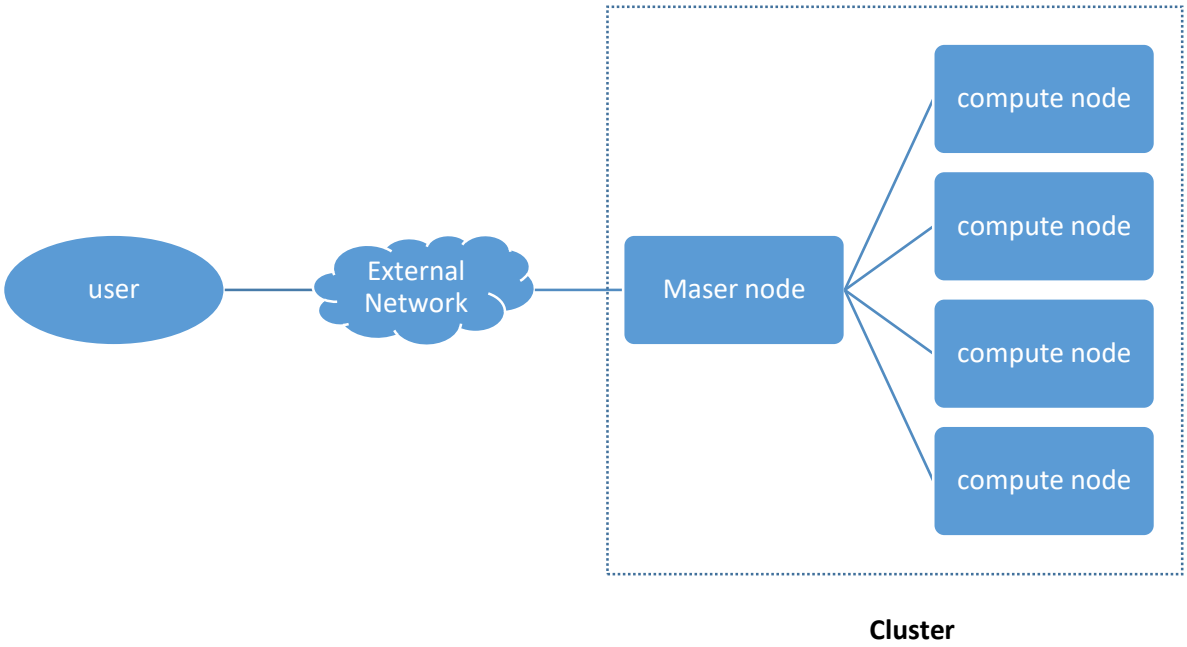

Figure 4: Layout diagram of the experimental PMD HPC cluster.

1) Linux Installation

CENTOS version 7.5 is installed on the master and compute nodes by choosing the server option. With standard partitioning. As /home, /root, /boot and /swap

2) Setting up the network

The network of the master node is configured by setting up static ip by editing the Ethernet network script as in this case /etc/sysconfig/network-scripts/ifcfg-eno1. Host name of master and compute nodes is setup by updating the /etc/hostname and /etc/hosts file. On compute node the /etc/hosts file is like

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4::1         localhost
localhost.localdomain localhost6 localhost6.localdomain6
192.168.15.105 pmdhpc1          #Masternode address
192.168.15.107 pmdhpc3          #compute node address
```

Whereas on master node /etc/hosts is

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4::1

localhost localhost.localdomain localhost6 localhost6.localdomain6

192.168.15.105 pmdhpc1 pmdhpc1.local          #Masternode address
```

The file is saved and settings are saved using following way:

 # sudo systemctl restart network

3) Installing Munge Uid 'N' Gid Emporium

3.1) creating the user and group IDs.

- Create user with unique id e.g *# export MUNGEUSER=981* (if this id is already in use the one can allot a different unique id e.g in our case it is 9991) [13]
- Add user and group munge *# groupadd -g $MUNGEUSER munge*
- Define user and group specs to the munge software

 # useradd -m -c "MUNGE Uid 'N' Gid Emporium" -d /var/lib/munge -u $MUNGEUSER -g munge -s /sbin/nologin munge

- Create slurm user *# export SlurmUSER=982*
- Add user and group *# groupadd -g $SlurmUSER slurm*
- Define user and group specs to the slurm software

 # useradd -m -c "Slurm workload manager" -d /var/lib/slurm -u $SlurmUSER -g slurm -s /bin/bash slurm

- These users can also be added to sudoers group so that later on there won't be any problems regarding permissions. Also SSH without password needs to be configured.

 # yum install munge munge-libs munge-devel

Test the configuration and installation using

 # munge –C

 # munge –M

- Generate munge key

 # dd if=/dev/urandom bs=1 count=1024 > /etc/munge/munge.key

- Changing ownership and permissions of the key

 # chown munge: /etc/munge/munge.key

 # chmod 400 /etc/munge/munge.key

- Securely propagate /etc/munge/munge.key (e.g., via Scp) to all other nodes within the same security realm

 # scp -p /etc/munge/munge.key user@xxx.xxx.xx.xxx:/etc/munge/munge.key

- Change ownership and permisions on all nodes

 # chown -R munge: /etc/munge/ /var/log/munge/

 # chmod 0700 /etc/munge/ /var/log/munge/

- enable and start the MUNGE service on all nodes

 # systemctl enable munge

 # systemctl start munge

- Testing Munge

 # munge -n

 # munge -n | unmunge # Displays information about the MUNGE key

 # munge -n | ssh user@xxx.xxx.xx.xxx unmunge (not working without password prompt)

 # remunge

4) Installing Slurm Workload Manager

- Following slurm dependencies are installed first with yum.[13]

Gcc, openssl, libssh2-devel, pam-devel, numactl, hwloc, lua, readline-devel, rrdtool-devel ncurses-devel, gtk2-devel, man2html, libibmad, libibumad perl-Switch, perl-ExtUtils-MakeMaker

- Install mysql server libraries: mariadb-server, mariadb-devel. Latest version of slurm source code (.tar.bz2 file) from https://www.schedmd.com/downloads.php e.g. slurm-17.02.11.tar.bz2 (there are two versions available slurm-17.11.7 and slurm-17.02.11), For the current PMD HPCC slurm-17.02.11 is used.

- Set the version (currently 17.02.11) and build Slurm RPMs as

 # export VER=17.02.11

 # rpmbuild -ta slurm-$VER.tar.bz2

Navigate to rpm directory (as root user):

 # /root/rpmbuild/RPMS/x86_64/

 # export VER=17.02.11

- RPMs to be installed on the head node and compute nodes

slurm-$VER*rpm, slurm-devel-$VER*rpm, slurm-munge-$VER*rpm, slurm-perlapi-$VER*rpm, slurm-plugins-$VER*rpm, slurm-torque-$VER*rpm

 # export VER=17.02.10

slurm-slurmdbd-$VER*rpm, slurm-sql-$VER*rpm, slurm-plugins-$VER*rpm

- Enable slurm service

 # systemctl enable slurmctld

users other than root then that user needs to be in the sudoers group as well

to add user in sudoers (wheel group in Centos):

 # usermod –aG wheel user

 # visudo

Add users as in the sudoers file under

```
## Allow root to run any commands anywhere
root    ALL=(ALL)       ALL
User  ALL=(ALL)       ALL
```

- Enabling slurm service as another user

 # sudo systemctl enable slurmctld

On compute nodes

slurm-pam_slurm-$VER*rpm

 # systemctl enable slurmd

Repeat this procedure on all compute nodes

4.1) Slurm configuration

File named slurm.conf is created in /etc/slurm/ using online tool available at

http://slurm.schedmd.com/configurator.easy.html.

The edited slurm.conf [12]file according to the system specification is as follows

```
ControlMachine=pmdhpc1
ControlAddr=192.168.15.105
AuthType=auth/munge
CryptoType=crypto/munge
SlurmUser=slurm
TaskPlugin=task/cgroup
ClusterName=pmdhpc
# COMPUTE NODES
NodeName=pmdhpc1 NodeAddr=192.168.15.105 CPUs=8 State=UNKNOWN
NodeName=pmdhpc2 NodeAddr=192.168.15.106 CPUs=8 State=UNKNOWN
NodeName=pmdhpc3 NodeAddr=192.168.15.107 CPUs=8 State=UNKNOWN
PartitionName=debug Nodes=pmdhpc[1-3] Default=YES MaxTime=INFINITE State=UP
```

Another file defining resources used by the cluster group is edited in following way

```
# Slurm cgroup support configuration file
#
# See man slurm.conf and man cgroup.conf for further
# information on cgroup configuration parameters
#--
CgroupMountpoint="/sys/fs/cgroup"
CgroupAutomount=yes
CgroupReleaseAgentDir="/etc/slurm/cgroup"
AllowedDevicesFile="/etc/slurm/cgroup_allowed_devices_file.conf"
ConstrainCores=no
AllowedSwapSpace=0
MaxRAMPercent=100
MaxSwapPercent=100
MinRAMSpace=30
```

This slurm.conf and cgroup.conf file is copied to each node

 # scp slurm.conf root@xxx.xxx.xx.xxx:/etc/slurm/slurm.conf

On master node required directories and log files for slurm are created as follows

 # mkdir /var/spool/slurmctld

 # chown slurm: /var/spool/slurmctld

 # chmod 755 /var/spool/slurmctld

 # touch /var/log/slurmctld.log

 # chown slurm: /var/log/slurmctld.log

 # touch /var/log/slurm_jobacct.log /var/log/slurm_jobcomp.log

 # chown slurm: /var/log/slurm_jobacct.log /var/log/slurm_jobcomp.log

Same is done on compute nodes

 # mkdir /var/spool/slurmd

 # chown slurm: /var/spool/slurmd

 # chmod 755 /var/spool/slurmd

 # touch /var/log/slurmd.log

 # chown slurm: /var/log/slurmd.log

Use the following command to make sure that slurmd is configured properly:

 # slurmd –C

Should give

```
NodeName=pmdhpc1 CPUs=8 Boards=1 SocketsPerBoard=2 CoresPerSocket=4 ThreadsPerCore=1
RealMemory=7820 TmpDisk=93740 UpTime=0-21:18:40
```

4.2) Syncing clocks and starting slurmd and slurmctld service

On each node clocks are synced through

 # yum install ntp –y

 # chkconfig ntpd on

 # ntpdate pool.ntp.org

 # systemctl start ntpd

On compute nodes slurmd.service is triggered as user “slurm”

 # sudo systemctl enable slurmd.service

 # sudo systemctl start slurmd.service

 # sudo systemctl status slurmd.service

On master node start slurmctld.service as user “slurm”

 # sudo systemctl enable slurmctld.service

 # sudo systemctl start slurmctld.service

 # sudo systemctl status slurmctld.service

5) SSH without password

RSA private key is generated in the home directory in ~/.ssh (this directory needs to be created if it doesn't exist). In order to create directory and that directory to be accessible to the slurm user, the user should have read and write permissions to that directory, in this regard, the slurm and munge users should also be in the sudoers group. For adding users to sudoers group

 # cd ~/.ssh

 # ssh-keygen -t rsa (keep pressing enter when prompted)

 # scp ~/.ssh/id_rsa.pub username@xxx.xxx.xx.xxx:~/.ssh/.

 # ssh username@xxx.xxx.xx.xxx

 # cat id_rsa.pub >> ~/.ssh/authorized_keys

 # chmod 755 /var/lib/slurm/.ssh

 # chmod 755 /var/lib/slurm/.ssh/authorized_keys

To check

 # ssh username@xxx.xxx.xx.xxx

6) Configuring the Network File System (NFS)

On master node edit file /etc/exports

/exportdir ipaddress/netmask(options)

Since on the master node required libraries to be shared with compute nodes are installed in /share/apps

/share/apps 192.168.15.105/255.255.255.0(rw,async,no_root_squash)

Restart nfs and enable it at boot time

 # systemctl restart nfs

 # systemctl enable nfs

Import these shared folders on compute node by editing /etc/fstab as follows

pmdhpc1:/share/apps /share/apps nfs rw,hard,intr 0 0

pmdhpc1 is the name of the head node, /share/apps is one of the shares defined in /etc/exports on the head node. The second /share/apps tells the OS where to mount the share on the compute node, and nfs lets the OS know that it should use nfs to mount the share. This command can also be added to /etc/rc.local to make sure the folder is mounted on startup.

 # mount /share/apps

Results

Slurm software package provides excellent tools in order to monitor the state of the cluster and to make sure every state is up *sinfo* is used. If any node has state down/up it will appear as state down/up. In order to change the state of a node e.g in the following case two node are shown in down state.

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
debug*     up   infinite     2  down* pmdhpc[2-3]
debug*     up   infinite     1   idle pmdhpc1
```

 # scontrol update NodeName=pmdhpc1 state=resume

 # sinfo

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
debug*     up   infinite     1  idle* pmdhpc3
debug*     up   infinite     1  down* pmdhpc2
debug*     up   infinite     1   idle pmdhpc1
```

 # scontrol update NodeName=pmdhpc1 state=down Reason=drained, undrained etc

For testing purpose a numerical weather prediction model, Consortium for Small Scale Modeling (COSMO) is compiled on this cluster and simulated by submitting a batch job through *sbtach* command. COSMO is a non-hydrostatic limited area atmospheric prediction model provided by Deutscher Wetterdienst (DWD), Offenbach, Germany [(4)]. Since the installation of COSMO is not the scope of this paper, hence it is not discussed here. Details about the installation of this model is available in the user guide of COSMO [(4)]. First the *INT2LM* interpolation program is run for interpolating coarse grid model data such as ICON (used in this example) to the initial and

boundary data for COSMO-Model [(5)]. Icosahedral Nonhydrostatic model (ICON) is a global numerical weather prediction model developed by joint efforts of DWD and Max Plank Institute of Meteorology in Hamburg (MPI-M) [(2)]. The ICON model data is available at 13km resolution and in this example it is downscaled at 7km resolution (see test box below). The date and time including the resolution, domain size and node specification such as number of processors in x and

y direction nprocx and nprocy etc. are specified in the namelist INPUT_ORG. The advantage of using *sbatch* is that mpi command can be executed by submitting it to *sbatch* through a shell script. While in the *sbatch* command options are available to specify comma separated nodelist as well as the extra node info **–B** option with arguments as number of sockets on each node and number of processors on each socket, which, in this cluster's case is 2:4.

```
&LMGRID
 startlat_tot  = 19.5, startlon_tot  = 55.5,
 pollat=90.0,       pollon=-180.0,
 dlat=0.0625,       dlon=0.0625,
 ie_tot=331,        je_tot=481,       ke_tot=50,
/
 &RUNCTL
hstart = 0.0,   hstop  = 6.0,    dt = 60.0,    ydate_ini='2018020800',
 nprocx = 4,  nprocy = 4,   nprocio =,
 lphys     = .TRUE.,   luse_rttov    = .FALSE.,   luseobs = .FALSE.,  leps = .FALSE.,
 lreorder   = .FALSE.,  lreproduce    = .TRUE.,   itype_timing = 4,
 ldatatypes = .FALSE.,  ltime_barrier = .FALSE.,  ncomm_type=3,
 nboundlines=3, idbg_level = 2, lartif_data=.FALSE,
 ldfi=.FALSE., ldebug_io=.FALSE., lprintdeb_all=.FALSE.,
/
 &TUNING
 c_soil   =   1.0,
 clc_diag =   0.5,
 crsmin   = 150.0,
 qc0      =   0.0,
 q_crit   =   4.0,
 qi0      =   0.0,
 rat_can  =   1.0,
 rat_lam  =   1.0,
 tur_len  = 500.0,
 v0snow   =  25.0,
 wichfakt =   0.0,
 tkhmin   =   0.4,
 tkmmin   =   0.4,
```

After creating the interpolated initial and boundary conditions through int2lm program, a shell script describing the batch job is submitted through sbatch. The shell script named cosmorun.sh is given below.

```
#!/bin/bash
mpirun lmparbin_all
```

the script is then submitted to slurm workload manager using command

 # sbatch --nodelist=pmdhpc1,pmdhpc3 –B 2:4 cosmorun.sh

Where nodelist is the name of nodes on which COSMO is running. Here –B 2:4 means extra node info where 2=number of sockets on each node and 4= number of processors on each socket (in INPUT_ORG enter total number of processors of all nodes listed). Each successful run creates a log file of output. Simulation is monitored through following command.

 # tail –f slurm_jobid.out

The result of the running job is shown in the following text box.

```
   + + + + + + + + + + + + + + + +
   + RUNNING IN DOUBLE PRECISION +
   + + + + + + + + + + + + + + + +

  SETUP OF THE LM
   INITIALIZATIONS

     Info about KIND-parameters:   iintegers / MPI_INT =          4         7
                                   int_ga    / MPI_INT =          8        11
   INPUT OF THE NAMELISTS
 *** NOTE: Old 10 digit date format is used

==== Code information used to build this binary ====
Binary name .......: lmparbin

Library name ......: lm_f90
Tag name ..........: V5_4d_4
Checkin-Date ......: 2017-05-29 12:19:25
Code is modified ..: .false.
Compile-Date ......:
Compiled by .......: uschaett
GRIB_API version ..:

Current start time : 2018-09-28 02:34
Running on nodes ..:
Data decomposition :
==== End of code information ====
   INPUT OF THE NAMELISTS FOR DYNAMICS
  DOMAIN SIZE (approx.) in m: L_x =    1472539.3829233176
                              L_y =    3301218.0398787037
   INPUT OF THE NAMELISTS FOR PHYSICS
  *** Default specifications of soil main levels are used ***
```

```
   INPUT OF THE NAMELISTS FOR DIAGNOSTICS
   INPUT OF THE NAMELISTS FOR GRIB-IO
  *** WARNING: Horizontal diffusion specified but lhordiff=F or hd_corr_trcr_xx=0***
  (Your specification: itype_diff is not respected for QV)
  *** WARNING: Horizontal diffusion specified but lhordiff=F or hd_corr_trcr_xx=0***
  (Your specification: itype_diff is not respected for QC)

   ALLOCATE SPACE
  hd_mask - SETUP: i_west  =          15
  hd_mask - SETUP: i_east  =         317
  hd_mask - SETUP: j_south =          15
  hd_mask - SETUP: j_north =         467
OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/icon_d00/laf2018020800
Note: analysis field PLCOV    with time range indicator          0  is used
Note: analysis field LAI      with time range indicator          0  is used
Note: analysis field ROOTDP   with time range indicator          0  is used
Note: analysis field VIO3     with time range indicator          0  is used
Note: analysis field HMO3     with time range indicator          0  is used
Note: analysis field T_SNOW   with time range indicator          0  is used
Note: analysis field W_I      with time range indicator          0  is used
Note: analysis field W_SNOW   with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field T_SO     with time range indicator          0  is used
Note: analysis field W_SO      with add. element number          0  is used
Note: analysis field W_SO     with time range indicator          0  is used
Note: analysis field RHO_SNOW with time range indicator          0  is used
```

```
 GRIB edition number used for atmospheric initial data:          1
 Working with reference atmosphere:          2
CLOSING grb1 FILE
OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/icon_d00/lbff00000000
CLOSING grb1 FILE
OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/icon_d00/lbff00030000
CLOSING grb1 FILE
variable T_MNW_LK is not allocated and is removed from the list
 variable T_WML_LK is not allocated and is removed from the list
 variable T_BOT_LK is not allocated and is removed from the list
variable C_T_LK is not allocated and is removed from the list
 variable H_ML_LK is not allocated and is removed from the list
 variable T_ICE is not allocated and is removed from the list
 variable H_ICE is not allocated and is removed from the list
 OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/d00/lfff00000000c
 CLOSING grb1 FILE
  INITIALIZATIONS
   INITIALIZATIONS for DYNAMICS and RELAXATION
 Minimum horizontal grid spacing (dx, dy):    4513.6202381297226       6949.9327155341125
 Minimum vertical   grid spacing (dz)    :    17.181145833333630
 Value of big time step used    :    60.000000000000000
Subr.[init_grid_metrics] ...

 Damping coefficients in Rayleigh damping layer
 level   height    damping coefficient
   1    29287.04     0.199306
   2    27886.15     0.193954
   3    26534.47     0.184027
   4    25230.43     0.170481
   5    23972.49     0.154313
   6    22759.24     0.136493
   7    21589.30     0.117914
   8    20461.36     0.099361
   9    19374.20     0.081493
  10    18326.63     0.064832
  11    17317.52     0.049774
  12    16345.79     0.036587
  13    15410.41     0.025432
  14    14510.40     0.016378
  15    13644.82     0.009411
  16    12812.78     0.004459
  17    12013.41     0.001401
  18    11245.87     0.000083

   Lowest Level with Rayleigh-damping:          18

 set  beta_s_8=beta_s_4, ... (recommended together with i_div_at_horiz_bound=1)
 divergence damping vertically implicit
```

```
   PHYSICAL PACKAGES
 ************************************************
 *   Radiative transfer calculations employ data    *
 *       provided in routine rad_aibi            *
 ************************************************
         initialize background aerosol (aerdis)
    INITIALIZE CONTROL VARIABLES AND MEAN VALUES
  TIME STEPPING
    STEP          0

CS=   348.92816041757197      DT=   60.000000000000000      DTSMAX=   9.7637103706843327
 SHORT TIME STEP of 1. RK step    5.0000000000000000     , number of steps:           4
 SHORT TIME STEP of 2. RK step    5.0000000000000000     , number of steps:           6
 SHORT TIME STEP of 3. RK step    5.0000000000000000     , number of steps:          12
 ismtstep_sum =          22  ismtstep_max =          12
  LEVELINDEX WHERE LEVELS BECOME FLAT, KFLAT =          13
  BETA_SW  =   0.40000000000000002      BETA_GW  =   0.40000000000000002
  BETA2_SW =   0.40000000000000002      BETA2_GW =   0.40000000000000002
  XKD =   0.10000000000000001
 OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/d00/lfff00000000p
 CLOSING grb1 FILE
 OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/d00/lfff00000000z
 CLOSING grb1 FILE
 OPEN: grb1-file: /home/slurm/cosmo/cosmo_7/d00/lfff00000000
 CLOSING grb1 FILE
    STEP          1
    STEP          2
    STEP          3
    STEP          4
    STEP          5
    STEP          6
    STEP          7
    STEP          8
    STEP          9
    STEP         10
    STEP         11
    STEP         12
    STEP         13
    STEP         14
```

Conclusion

A high performance computing cluster is installed on 3 nodes. The resulting cluster flexible and is also applicable for set of computers. Slurm is an excellent open source workload managing software with wide range of tools available for managing parallel jobs on HPC environment.

Following commands can be used to check the resource status

>>sinfo (for checking the up nodes state)

If any node is showing state down, in order to change the state of the node use following command

>>scontrol update NodeName=pmdhpc1 state=resume

In order to state down a node, use following command

>>scontrol update NodeName=pmdhpc1 state=down Reason=drained, undrained etc

To check network usage, use *nethogs, nload, iftop -n*

References

1. Brownell, Matthew., 2015. Building and Improving a Linux Cluster, A senior thesis submitted to the faculty of Brigham Young University Idaho in partial fulfillment of the requirements for the degree of Bachelor of Science. Brigham Young University Idaho.
2. Reinert, D., Frank, H., & Prill, F. (2015). *ICON database reference manual*. Deutscher Wetterdienst.
3. Sokunbi, Moses. (2006). Installing a 3 node Linux based cluster for scientific computation. International Journal HIT Transactions on ECCN: 0973-6875. 1. 192 – 204.
4. Schättler, U., Doms, G. and Schraff, C., 2008. A description of the nonhydrostatic regional COSMO-model part VII: user's guide. *Deutscher Wetterdienst. COSMO-Model*.
5. Schattler, U., Blahak, U., 2017. *A Description of the Nonhydrostatic Regional COSMO-Model* Part V:Preprocessing:Initial and Boundary Data for the COSMO-Model. *Deutscher Wetterdienst. COSMO-Model*
6. Vance, N. R., Poublon, M. L., & Polik, W. F. (2016). BYOC: Build Your Own Cluster, Part III-Configuration. *Linux Journal*, (279), 70.
7. Datti, A.A., Umar, H.A. and Galadanci, J., 2015. A beowulf cluster for teaching and learning. *Procedia Computer Science*, *70*, pp.62-68.
8. Al-Khazraji, S.H.A.A., Al-Sa'ati, M.A.Y. and Abdullah, N.M., 2014. Building High Performance Computing Using Beowulf Linux Cluster. *International Journal of Computer Science and Information Security*, *12*(4), p.1.
9. https://github.com/dun/munge/wiki/Man-7-munge
10. https://centos.pkgs.org/6/epel-x86_64/munge-0.5.10-1.el6.x86_64.rpm.html
11. https://slurm.schedmd.com/overview.html
12. https://slurm.schedmd.com/SLUG17/SlurmOverview.pdf
13. https://computing.llnl.gov/tutorials/parallel_comp/#Whatis
14. http://slurm.schedmd.com/configurator.easy.html
15. https://wiki.fysik.dtu.dk/niflheim/Slurm_installation
16. Kimura, R., 2002. Numerical weather prediction. *Journal of Wind Engineering and Industrial Aerodynamics*, *90*(12-15), pp.1403-1414.

Glossary

CentOS	Community Enterprise Operating System, is a distribution of the Linux operating system based on RHEL (Red Hat Enterprise Linux)
Ethernet	an array of networking technologies and systems used in local area networks (LAN), where computers are connected within a primary physical space.
Fiber Channel	a computer networking technology that is used to transfer data between one or more computers at very high speeds commonly implemented in storage networking server environments
Infiniband	an input/output (I/O) architecture and high-performance specification for data transmission between high-speed, low latency and highly-scalable CPUs, processors and storage.
TFlops	rate of computing speed that achieves one trillion floating point operations per second. "Flops" refer to "floating point operations per second."