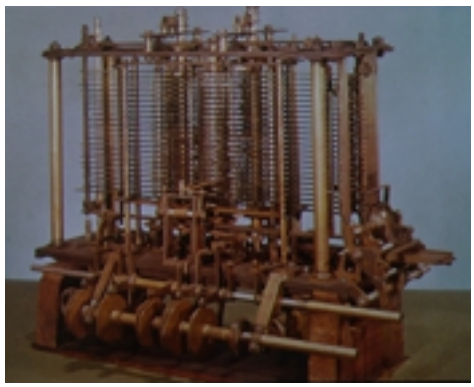
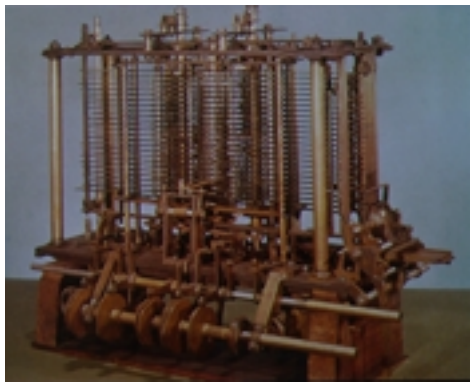
HPC Roll: Users Guide



Version 5.1 Edition



HPC Roll: Users Guide :

Version 5.1 Edition

Published Nov 2008

Table of Contents

List of Tables

Preface

The primary purpose of the HPC Roll is to provide configured software tools that can be used to run parallel applications on your cluster.

The following software packages are included in the HPC Roll:

- MPI over ethernet environments (OpenMPI, MPICH, MPICH2)
- PVM
- Benchmarks (stream, iperf, IOzone)

Chapter 1. Overview

Table 1-1. Summary

Name	hpc
Version	5.1
Maintained By	Rocks Group
Architecture	i386, x86_64
Compatible with Rocks™	5.1

Table 1-2. Roll Compatibility

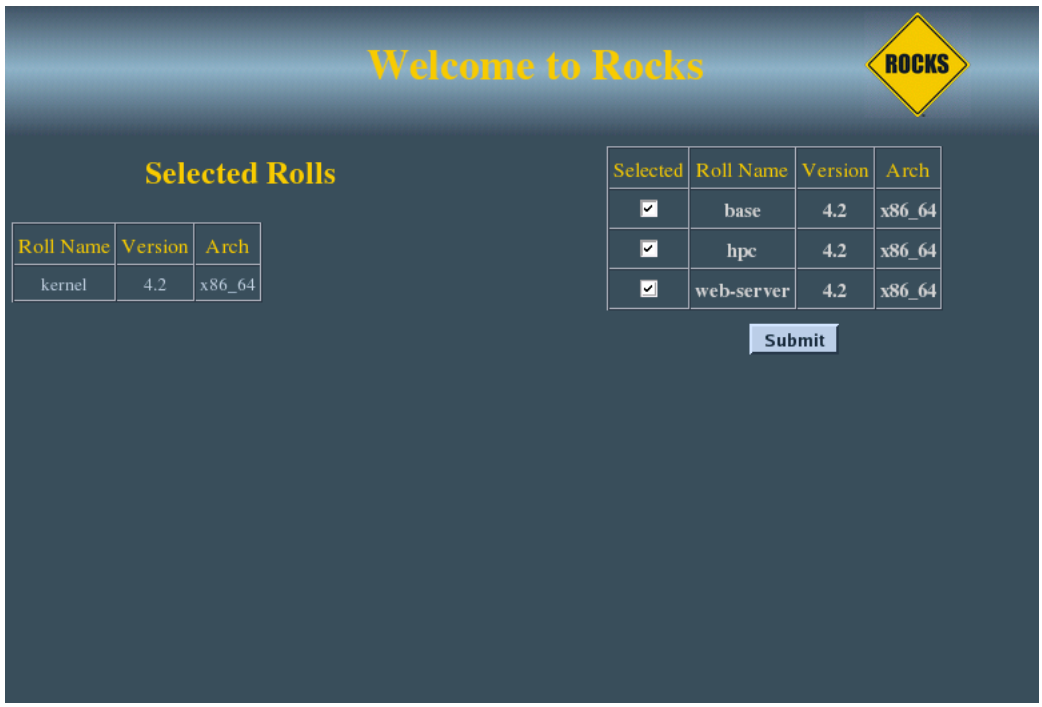
Roll	Requires [a]	Optional [b]	Conflicts
alpha		X	
area51		X	
base	X		
bio		X	
condor		X	
ganglia		X	
grid		X	
hpc		X	
java		X	
kernel	X		
os (disk 1)	X		
os (disk 2)	X		
os (disk 3)		X	
os (disk 4)		X	
os (disk 5)		X	
os (disk 6)		X	
os (disk 7)		X	
pbs		X	
service-pack		X	
sge		X	
viz		X	
web-server		X	
xen		X	

Roll	Requires [a]	Optional [b]	Conflicts
Notes: a. You may also substitute your own OS CDs for the Rocks™ OS Roll CDs. In this case you must use all the CDs from your distribution and not use any of the Rocks™ OS Roll CDs. b. Only Rolls that have been verified as compatible with this Roll are listed. Other Rolls will likely work, but have not been tested by the maintainer of this Roll.			

Chapter 2. Installing

2.1. On a New Server

The hpc Roll should be installed during the initial installation of your server (or cluster). This procedure is documented in section 1.2 of the Rocks™ usersguide. You should select the hpc Roll from the list of available rolls when you see a screen that is similar to the one below.



2.2. On an Existing Server

The hpc Roll may also be added onto an existing server (or frontend). For sake of discussion, assume that you have an iso image of the roll called `hpc.iso`. The following procedure will install the Roll, and after the server reboots the Roll should be fully installed and configured.

```
$ su - root
# rocks add roll hpc.iso
# rocks enable roll hpc
# rocks-dist dist
# kroll hpc | bash
# init 6
```

Chapter 3. Using

3.1. Using mpirun from OpenMPI

To interactively launch a test OpenMPI program on two processors:

- Create a file in your home directory named `machines`, and put two entries in it, such as:

```
compute-0-0
compute-0-1
```

- Now launch the job from the frontend:

```
$ ssh-agent $SHELL
$ ssh-add
/opt/openmpi/bin/mpirun -np 2 -machinefile machines /opt/mpi-tests/bin/mpi-ring
```

You must run MPI programs as a regular user (that is, not root).

If you don't have a user account on the cluster, create one for yourself, and propogate the information to the compute nodes with:

```
# useradd username
# rocks sync users
```

3.2. Using mpirun from MPICH

To interactively launch a test MPICH program on two processors:

- Create a file in your home directory named `machines`, and put two entries in it, such as:

```
compute-0-0
compute-0-1
```

- Compile a test program using the MPICH environment:

```
$ cd $HOME
$ mkdir mpich-test
$ cd mpich-test
$ cp /opt/mpi-tests/src/mpi-ring.c .
$ /opt/mpich/gnu/bin/mpicc -o mpi-ring mpi-ring.c -lm
```

- Now launch the job from the frontend:

```
$ ssh-agent $SHELL
$ ssh-add
$ /opt/mpich/gnu/bin/mpirun -nolocal -np 2 -machinefile $HOME/machines \
 $HOME/mpich-test/mpi-ring
```

You must run MPI programs as a regular user (that is, not root).

If you don't have a user account on the cluster, create one for yourself, and propogate the information to the compute nodes with:

```
# useradd username
# rocks sync users
```

3.3. Cluster-Fork

Cluster-Fork runs a command on compute nodes of your cluster.

Often we want to execute parallel jobs consisting of standard UNIX commands. By "parallel" we mean the same command runs on multiple nodes of the cluster. We use these simple parallel jobs to move files, to run small tests, and to perform various administrative tasks.

Rocks provides a simple tool for this purpose called `cluster-fork`. For example, to list all your processes on the compute nodes of the cluster:

```
$ cluster-fork ps -U$USER
```

By default, cluster-fork uses a simple series of ssh connections to launch the task serially on every compute node in the cluster. Cluster-fork is smart enough to ignore dead nodes. Usually the job is "blocking": cluster-fork waits for the job to start on one node before moving to the next. By using the `--bg` flag you can instruct cluster-fork to start the jobs in the background. This corresponds to the "-f" ssh flag.

```
$ cluster-fork --bg hostname
```

Often you wish to name the nodes your job is started on. This can be done by using an SQL statement or by specifying the nodes using a special shorthand.

The first method of naming nodes uses the SQL database on the frontend. We need an SQL statement that returns a column of node names. For example, to run a command on compute nodes in the first rack of your cluster exectute:

```
$ cluster-fork --query="select name from nodes where name like 'compute-1-%'" [cmd]
```

The next method of requires us to explicitly name each node. When launching a job on many nodes of a large cluster this often becomes cumbersome. We provide a special shorthand to help with this task. This shorthand, borrowed from the MPD job launcher, allows us to specify large ranges of nodes quickly and concisely.

The shorthand is based on similarly-named nodes and uses the `--nodes` option. To specify a node range *compute-0-0 compute-0-1 compute-0-2*, we write *--nodes=compute-0-%d:0-2*. This scheme works best when the names share a common prefix, and the variables between names are numeric. Rocks compute nodes are named with such a convention.

Other shorthand examples:

- Discontinuous ranges:

  compute-0-%d:0,2-3 --> `compute-0-0 compute-0-2 compute-0-3`

- Multiple elements:

  compute-0-%d:0-1 compute-1-%d:0-1 --> `compute-0-0 compute-0-1 compute-1-0 compute-1-1`

- Factoring out duplicates:

  *2*compute-0-%d:0-1 compute-0-%d:2-2* --> `compute-0-0 compute-0-0 compute-0-1 compute-0-1 compute-0-2`

```
$ cluster-fork --nodes="compute-2-%d:0-32 compute-3-%d:0-32" ps -U$USER
```

The previous example lists the processes for the current user on 64 nodes in racks two and three.

Appendix A. Rocks Copyright

Rocks(r)
www.rocksclusters.org
version 5.1 (VI)

Appendix B. Third Party Copyrights and Licenses

This section enumerates the licenses from all the third party software components of this Roll. A "best effort" attempt has been made to insure the complete and current licenses are listed. In the case of errors or ommisions please contact the maintainer of this Roll. For more information on the licenses of any components please consult with the original author(s) or see the Rocks™ CVS repository[1].

B.1. iozone

```
     Original Author: William Norcott (wnorcott@us.oracle.com)
             4 Dunlap Drive
             Nashua, NH 03060

     Enhancements: Don Capps (capps@iozone.org)
             7417 Crenshaw
             Plano, TX 75025

Copyright 1991, 1992, 1994, 1998, 1999, 2002   William D. Norcott

License to freely use and distribute this software is hereby granted
by the author, subject to the condition that this copyright notice
remains intact.  The author retains the exclusive right to publish
derivative works based on this work, including, but not limited to,
revised versions of this work.
```

B.2. iperf

```
Distributed Applications Support Team


  Iperf Copyright

--------------------------------------------------------------------
Copyright (c) 1999,2000,2001,2002,2003,2004 The Board of Trustees of the
University of Illinois
All Rights Reserved.

Iperf performance test <http://dast.nlanr.net/Projects/Iperf>
Mark Gates
Ajay Tirumala
Jim Ferguson
Jon Dugan
```

```
Feng Qin
Kevin Gibbs
National Laboratory for Applied Network Research
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
http://www.ncsa.uiuc.edu

Permission is hereby granted, free of charge, to any person obtaining a
copy of this software (Iperf) and associated documentation files (the
"Software"), to deal in the Software without restriction, including
without limitation the rights to use, copy, modify, merge, publish,
distribute, sublicense, and/or sell copies of the Software, and to
permit persons to whom the Software is furnished to do so, subject to
the following conditions:

    * Redistributions of source code must retain the above copyright
      notice, this list of conditions and the following disclaimers.

    * Redistributions in binary form must reproduce the above copyright
      notice, this list of conditions and the following disclaimers in
      the documentation and/or other materials provided with the
      distribution.

    * Neither the names of the University of Illinois, NCSA, nor the
      names of its contributors may be used to endorse or promote
      products derived from this Software without specific prior written
      permission.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS
OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF
MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT.
IN NO EVENT SHALL THE CONTIBUTORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY
CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT,
TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE
SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
```

B.3. MPICH

```
      COPYRIGHT

The following is a notice of limited availability of the code, and disclaimer
which must be included in the prologue of the code and in all source listings
of the code.

Copyright Notice
 + 1993 University of Chicago
 + 1993 Mississippi State University
```

```
Permission is hereby granted to use, reproduce, prepare derivative works, and
to redistribute to others.  This software was authored by:

Argonne National Laboratory Group
W. Gropp: (630) 252-4318; FAX: (630) 252-5986; e-mail: gropp@mcs.anl.gov
E. Lusk:  (630) 252-7852; FAX: (630) 252-5986; e-mail: lusk@mcs.anl.gov
Mathematics and Computer Science Division
Argonne National Laboratory, Argonne IL 60439

Mississippi State Group
N. Doss:  (601) 325-2565; FAX: (601) 325-7692; e-mail: doss@erc.msstate.edu
A. Skjellum:(601) 325-8435; FAX: (601) 325-8997; e-mail: tony@erc.msstate.edu
Mississippi State University, Computer Science Department &
 NSF Engineering Research Center for Computational Field Simulation
P.O. Box 6176, Mississippi State MS 39762

         GOVERNMENT LICENSE

Portions of this material resulted from work developed under a U.S.
Government Contract and are subject to the following license: the Government
is granted for itself and others acting on its behalf a paid-up, nonexclusive,
irrevocable worldwide license in this computer software to reproduce, prepare
derivative works, and perform publicly and display publicly.

      DISCLAIMER

This computer code material was prepared, in part, as an account of work
sponsored by an agency of the United States Government.  Neither the United
States, nor the University of Chicago, nor Mississippi State University, nor
any of their employees, makes any warranty express or implied, or assumes any
legal liability or responsibility for the accuracy, completeness, or
usefulness of any information, apparatus, product, or process disclosed, or
represents that its use would not infringe privately owned rights.
```

B.4. MPICH2

```
      COPYRIGHT

The following is a notice of limited availability of the code, and disclaimer
which must be included in the prologue of the code and in all source listings
of the code.

Copyright Notice
 + 2002 University of Chicago

Permission is hereby granted to use, reproduce, prepare derivative works, and
to redistribute to others.  This software was authored by:
```

Argonne National Laboratory Group
W. Gropp: (630) 252-4318; FAX: (630) 252-5986; e-mail: gropp@mcs.anl.gov
E. Lusk: (630) 252-7852; FAX: (630) 252-5986; e-mail: lusk@mcs.anl.gov
Mathematics and Computer Science Division
Argonne National Laboratory, Argonne IL 60439

B.5. OpenMPI

$HEADER$

Redistribution and use in source and binary forms, with or without
modification, are permitted provided that the following conditions are
met:

- Redistributions of source code must retain the above copyright
  notice, this list of conditions and the following disclaimer.

- Redistributions in binary form must reproduce the above copyright
  notice, this list of conditions and the following disclaimer listed
  in this license in the documentation and/or other materials
  provided with the distribution.

- Neither the name of the copyright holders nor the names of its
  contributors may be used to endorse or promote products derived from
  this software without specific prior written permission.

The copyright holders provide no reassurances that the source code
provided does not infringe any patent, copyright, or any other
intellectual property rights of third parties. The copyright holders
disclaim any liability to any recipient for claims brought against
recipient by any third party for infringement of that parties
intellectual property rights.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS
"AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT
LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR
A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT
OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT
LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE,
DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY
THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT
(INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE
OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

B.6. PVM

B.7. stream

```
        labelled whenever they are published.  Examples of
        proper labelling include:
        "tuned STREAM benchmark results"
        "based on a variant of the STREAM benchmark code"
        Other comparable, clear and reasonable labelling is
        acceptable.
    3c. Submission of results to the STREAM benchmark web site
        is encouraged, but not required.
4. Use of this program or creation of derived works based on this
   program constitutes acceptance of these licensing restrictions.
5. Absolutely no warranty is expressed or implied.
```

Notes

1. http://cvs.rocksclusters.org