MSc Scientific Computing Dissertation
ARM Cluster Linpack Benchmark

John Duffy

August 2020

1 Introduction

Objectives.

1. Best achievable High-Performance Linpack performance.

2. An entry in 'The Green500' list!?

3.

etc

This is also intended to be a complete guide for reproducing the build (hardware and software) of the ARM cluster and also for reproducing the results.

Top Gflops.

Describe setup, referencing appendices for detail.

Diagram of setup.

node1 is *master* node.

Code at:

```
https://github.com/johnduffymsc/phas0077.git
```

2 Theoretical Maximum Performance (Gflop/s)

The Raspberry Pi Model 4B uses the Broadcom BCM2711 System on a Chip (Soc).

4 cores

1.5 GHz

NEON 128 bit SIMD

4 GB memory (our chosen model)

Caches...

Pipeline...

Simplistically, ...

This ignores instructions pipelining benefits...

3 High-Performance Linpack (HPL)

https://www.netlib.org/benchmark/hpl/

3.1 Installation

Reference Appendix... with specifics.

3.2 Tuning

Describe methodology for configuration file parameters

Tool: http://hpl-calculator.sourceforge.net

3.3 Baseline Benchmark

As per software installation from Ubuntu Server 20.04 LTS.

OpenBLAS

OpenMP

OpenMPI

HPL-2.3 compiled with Make.rpi4.baseline

NB = 128 (mid-range guess; to be tuned based on L1 cache size)

N for 80% efficiency (from tool)

Recommended: P×Q as square as possible, with Q > P

4 cores per node 1.5 GHz clock speed 4 GB memory per node 2 instructions per cycle (estimated as NEON is 128 bits)

From tool:

Run:

```
mpirun -host node1:4 -np 4 xhpl
mpirun -host node1:4,node2:4 -np 8 xhpl
mpirun -host node1:4,node2:4,node3:4 -np 12 xhpl
etc
```

TABLE RESULTS - Gflops vs node count, time vs node count GRAPH RESULTS - Gflops vs node count, time vs node count

Discussion...

NB size...

P
Q Ratio...

Node number scaling...

3.4 OpenMPI without OpenMP

Describe processor grid layout.

hosts-with-slots file.

3.5 OpenMPI with OpenMP

Describe processor grid layout.

hosts-no-slots file.

4 Performance Optimisation

4.1 Methodology

1. Measure
2. Study results and propose theory
3. Change something based on 2.
4. Measure
5. Repeat steps 1 - 4

5 Build Kernel with Jumbo Frames Support

Standard MTU is 1500 bytes...

Maximum payload size is 1472 bytes...

NB of 184 (x 8 bytes for Double Precision) = 1472 bytes...

NB > 184 => packet fragmentation => reduced network efficiency...

This causes drop of in performance???...

Max MTU on Raspberry Pi 4 Model B is set at build time to 1500...

Not configurable above 1500...

TODO: EXAMPLE OF ERROR MSG...

Need to build the kernel with higher MTU...

Make source packages available...

```
sudo touch /etc/apt/sources.list.d/picluster.list
sudo vim /etc/apt/sources.list.d/picluster.list...
    deb-src http://archive.ubuntu.com/ubuntu focal main
    deb-src http://archive.ubuntu.com/ubuntu focal-updates main
sudo apt update
```

Create a kernel build directory with the correct access permissions to prevent source download warnings.

```
mkdir kernel
sudo chown _apt:root kernel
cd kernel
```

Install the kernel build dependencies...

```
sudo apt-get build-dep linux linux-image-$(uname -r)
```

Download the kernel source...

```
sudo apt-get source linux-image-$(uname -r)
```

Make the required changes to the source... as per REFERENCE

```
cd linux-raspi-5.4.0

sudo vim include/linux/if_vlan.h...
    #define VLAN_ETH_DATA_LEN   9000
    #define VLAN_ETH_FRAME_LEN  9018

sudo vim include/uapi/linux/if_ether.h...
    #define ETH_DATA_LEN        9000
    #define ETH_FRAME_LEN       9014

sudo vim drivers/net/ethernet/broadcom/genet/bcmgenet.c...
    #define RX_BUF_LENGTH       10240
```

Add a Jumbo Frames identifier, "+jf", to the new kernel name...

```
sudo vim debian.raspi/changelog...
    linux (5.4.0-1013.13+jf) focal; urgency=medium
```

Build the kernel...

```
sudo LANG=C fakeroot debian/rules clean
sudo LANG=C fakeroot debian/rules binary
```

Install the new kernel...

```
sudo sudo dpkg -i linux*5.4????????.deb
```

6 Single Core Optimisation

6.1 Block Size Optimisation

The block size, NB tuning parameter, is used for matrix calculations and also for network transport.

The most efficient block size is related to the L1 cache size. Describe...

7 Single Node Optimisation

8 Cluster Optimisation

8.1 Recompile HPL for Cortex-A72

Block size!

RESULTS

8.2 Recompile OpenBLAS with OpenMP Support

As advised by the DebianScience/LinearAlgebraLibraries, OpenBLAS should be recompiled from source for best performance.??? See comflicting statement below.

The Ubuntu/Debian OpenBLAS package is an ARM64 multi-architecture build of OpenBLAS which includes the ARM Cortex-A72. As stated in the README.Debian, performance improvements will be minimal by compiling from source.

However, Ubuntu/Debian build does not include support for OpenMP. So, to test the combination of OpenMPI/OpenMP it is necessary to recompile OpenBLAS from source.

To download the same version of OpenBLAS as Ubuntu 20.04 Server LTS (v0.3.8):

```
cd ~/phas0077/downloads
wget https://github.com/xianyi/OpenBLAS/archive/v0.3.8.zip -O OpenBLAS-0.3.8.zip
cp OpenBLAS-0.3.8.zip ~/phas0077/projects
cd ~/phas0077/projects
unzip OpenBLAS-0.3.8.zip
rm OpenBLAS-0.3.8.zip
cd OpenBLAS-0.3.8
cp Makefile.rule Makefile.rule.original
vim Makefile.rule
make
make install
make clean
```

PROCESSOR AFFINITY - include later.

Need to edit 'Makefile.rule'.

The file 'Makefile.arm64' already has the following so there is no need to add specific architecture flags to 'Makefile.rule'.

```
ifeq ($(CORE), CORTEXA72)
CCOMMON_OPT += -march=armv8-a -mtune=cortex-a72
FCOMMON_OPT += -march=armv8-a -mtune=cortex-a72
endif
```

Block size!

TODO: Read DebianScience howto recompile/package

TODO: Check ARM gcc options; -mune, -march for ARM Cortex-A72 USE LD_PRELOAD trick to use recompiled libopenblas.so for testing before packaging as a Debian package. WHAT ABOUT THE OTHER NODES - How do they know about the re-compiled package on node1?

LD_PRELOAD for testing on node1.

ONCE installed as Debian package, prevent it being 'updated/upgraded'.

RESULTS

8.3 Recompile OpenMPI for Cortex-A72

Block size!

TODO: Check ARM gcc options; -mune, -march for ARM Cortex-A72

RESULTS

8.4 Network Tuning

Is is possible to improve performance looking at network parameter? MTU?

TODO: Read OpenMPI docs

RESULTS

8.5 Perf/Dtrace Linpack

See if there are any performance bottlenecks

New package OpenBLASRPI4? for RPI4 specific optimisations?

RESULTS

9 Appendix A - Raspberry Pi Cluster Build

9.1 Introduction

This appendix is intended to be a complete and self contained guide for building a Raspberry Pi Cluster. With the caveat that the cluster has the bare minimum software/functionality necessary to compile and run the High Performance Linpack (HPL) benchmark, namely the build-essential package, two BLAS libraries (OpenBLAS and BLIS), and Open-MPI. A number of performance measurement tools are also installed, such as perf and iperf. The latest version of HPL is downloaded and built from source.

It would be a relatively simple task to add... SLIRM or...

The cluster consists of the following components...

8 x Raspberry Pi 4 Model B 4GB compute nodes, node1 to node8 1 x software development and build node, node9 9 x Official Raspberry Pi 4 Model B power supplies 9 x 32GB Class 10 MicroSD cards 1 x *workstation*, in my case my MacBook Pro, macbook 1 x 8 port Gigabit Router/Firewall 1 x 16 port Gigabit switch with Jumbo Frame support

Items

Photo

9.2 Preliminary Tasks

1. Update the EE-PROM
2. Get MAC address
3. Generate keys
4. Amend macbook /etc/hosts file...

9.2.1 Update Raspberry Pi EE-PROMs

9.2.2 Get Raspberry Pi MAC Addresses

9.2.3 Generate User Key Pair

On macbook (no passphrase):

```
ssh-genkey -t rsa -C ucapjbj
```

This will create two files... in ...

9.2.4 Amend macbook /etc/hosts

On macbook, using your favourite editor, add the following to /etc/hosts:

```
192.168.0.1 node1
192.168.0.2 node2
192.168.0.3 node3
192.168.0.4 node4
192.168.0.5 node5
192.168.0.6 node6
192.168.0.7 node7
192.168.0.8 node8
192.168.0.9 node9
```

This enables...

```
ssh john@node1
```

or, the abbreviated...

```
ssh node1
```

provided the user name on the macbook is the same as the Linux user created by cloud-init.

9.2.5 Router/Firewall Configuration

Local network behind firewall/switch: 192.168.0.254

WAN address LAN address

Firewall/Switch (Netgear FVS318G)

Describe DHCP reservations mapping IP to MAC addresses.

Describe ssh access

Add relevant PDFs.

9.2.6 Create the Raspberry Pi Ubuntu Server Image

On macbook...

Download Ubuntu 20.04 LTS 64-bit pre-installed server image for the Raspberry Pi 4...

Double click to uncompress the .xz file which leaves the .img file.

Double click to mount the .img in the filesystem...

Amend /Volumes/system-boot/user-data...

```
#cloud-config

# This is the user-data configuration file for cloud-init. By default this sets
# up an initial user called "ubuntu" with password "ubuntu", which must be
# changed at first login. However, many additional actions can be initiated on
# first boot from this file. The cloud-init documentation has more details:
#
# https://cloudinit.readthedocs.io/
#
# Some additional examples are provided in comments below the default
# configuration.

# On first boot, set the (default) ubuntu user's password to "ubuntu" and
# expire user passwords
chpasswd:
  expire: false
  list:
  - ubuntu:ubuntu
  - john:john

# Enable password authentication with the SSH daemon
ssh_pwauth: true
```

```
## On first boot, use ssh-import-id to give the specific users SSH access to
## the default user
#ssh_import_id:
#- lp:my_launchpad_username
#- gh:my_github_username

## Add users and groups to the system, and import keys with the ssh-import-id
## utility
#groups:
#- robot: [robot]
#- robotics: [robot]
#- pi
#
groups:
- john: [john]

#users:
#- default
#- name: robot
#  gecos: Mr. Robot
#  primary_group: robot
#  groups: users
#  ssh_import_id: foobar
#  lock_passwd: false
#  passwd: $5$hkui88$nvZgIle31cNpryjRfO9uArF7DYiBcWEnjqq7L1AQNN3
users:
- default
- name: john
  gecos: John Duffy
  primary_group: john
  sudo: ALL=(ALL) NOPASSWD:ALL
  shell: /bin/bash
  ssh_authorized_keys:
  - ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQDGsnzP+1Q6NgeeKFTd/+Mom+UCYJTL/wzIiS9nL

## Update apt database and upgrade packages on first boot
#package_update: true
#package_upgrade: true
package_update: true
package_upgrade: true

## Install additional packages on first boot
#packages:
#- pwgen
#- pastebinit
#- [libpython2.7, 2.7.3-0ubuntu3.1]
```

```
packages:
- git
- tree
- unzip
- iperf
- net-tools
- linux-tools-common
- linux-tools-raspi
- build-essential
- gdb
- openmpi-common
- openmpi-bin
- libblis3-serial
- libblis3-openmp
- libopenblas0-serial
- libopenblas0-openmp

## Write arbitrary files to the file-system (including binaries!)
#write_files:
#- path: /etc/default/keyboard
#  content: |
#    # KEYBOARD configuration file
#    # Consult the keyboard(5) manual page.
#    XKBMODEL="pc105"
#    XKBLAYOUT="gb"
#    XKBVARIANT=""
#    XKBOPTIONS="ctrl: nocaps"
#  permissions: '0644'
#  owner: root:root
#- encoding: gzip
#  path: /usr/bin/hello
#  content: !!binary |
#    H4sIAIDb/U8C/1NW1E/KzNMvzuBKTc7IV8hIzcnJVyjPL8pJ4QIA6N+MVxsAAAA=
#  owner: root:root
#  permissions: '0755'
write_files:
- path: /etc/hosts
  content: |
    127.0.0.1 localhost
    192.168.0.1 node1
    192.168.0.2 node2
    192.168.0.3 node3
    192.168.0.4 node4
    192.168.0.5 node5
    192.168.0.6 node6
    192.168.0.7 node7
```

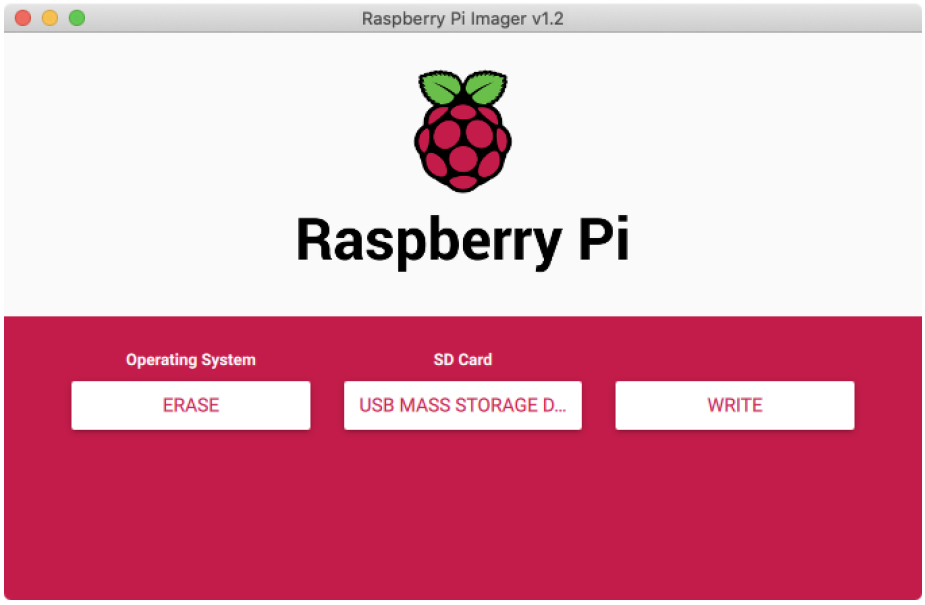

Figure 1: Using Raspberry Pi Imager to erase and format a MicroSD card.

```
    192.168.0.8 node8
    192.168.0.9 node9
  permissions: '0644'
  owner: root:root

## Run arbitrary commands at rc.local like time
#runcmd:
#- [ ls, -l, / ]
#- [ sh, -xc, "echo $(date) ': hello world!'" ]
#- [ wget, "http://ubuntu.com", -O, /run/mydir/index.html ]
runcmd:
- hostnamectl set-hostname --static node$(hostname -i | cut -d ' ' -f 1 | cut -d
- reboot
```

Eject/unmount .img file

Use Raspberry Pi Imager to erase...

Then use the Raspberry Pi Imager to write preinstalled server image to the MicroSD card...

When complete, remove the MicroSD card from the card reader, place it the Raspberry Pi and plug in the power cable.

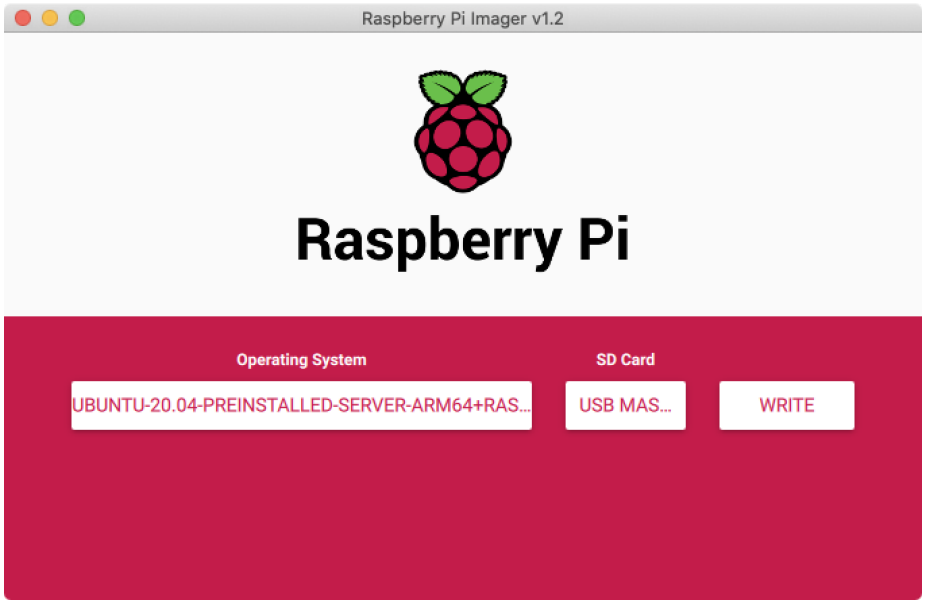

Figure 2: Using Raspberry Pi Imager to write the server image to a MicroSD card.

The cloud-init configuration process will now start. The Raspberry Pi will acquire its IP address from the router, setup users, update apt, upgrade the system, download software packages, set the hostname (based on the IP address), and finally the system will reboot.

9.3 Post-Installation Tasks

9.3.1 Enable No Password Access

This is required for Open-MPI...

Our public key was installed on each node by cloud-init. So, we can ssh into each node without a password, and use the abbreviated ssh node1, instead of ssh john@node1 (assuming john is the user name on the workstation).

We need to copy our private key to node1 (only node1)...

```
scp ~/.ssh/id_rsa node1:~/.ssh
```

Then to enable access to nodes node2 to node9 without a password from node1, we need to import the ... keys into the node1 knownhosts file...

This is easily done...

From macbook, ssh into node1...

ssh node1

and then from node1, for each of the nodes node2 to node9:

ssh node2

This will generate...

```
The authenticity of host 'node2 (192.168.0.2)' can't be established.
ECDSA key fingerprint is SHA256:5VgsnN2nPvpfbJmALh3aJdOeT/NvDXqN8TCreQyNaFA.
Are you sure you want to continue connecting (yes/no/[fingerprint])?
```

responding yes, imports the key into the node1 knownhosts file...

```
exit
```

Next node...

This is only required to be done on intial contact with nodes node2 to node9 (unless the keys on these nodes change)

9.3.2 Uninstall unattended-upgrades

The package unattended-upgrades is installed automatically...

Can potentially interferer with long running benchmarks...

Remove...

From macbookpro:

```
ssh node1 sudo apt remove unattended-upgrades --yes
ssh node2 sudo apt remove unattended-upgrades --yes
ssh node3 sudo apt remove unattended-upgrades --yes
ssh node4 sudo apt remove unattended-upgrades --yes
ssh node5 sudo apt remove unattended-upgrades --yes
ssh node6 sudo apt remove unattended-upgrades --yes
ssh node7 sudo apt remove unattended-upgrades --yes
ssh node8 sudo apt remove unattended-upgrades --yes
ssh node9 sudo apt remove unattended-upgrades --yes
```

Don't forget to update your cluster regularly at convenient times...

See update/upgrade script below...

9.3.3 Create a Project Repository

Xpand upon...

```
ssh node1
mkdir picluster
cd picluster
git init
```

Ensure you do

git add git commit git push

at regular intervals...

9.3.4 Create an System Update/Upgrade Script

Automate...

```
#!/usr/bin/bash

NODES=9

for (( i=$NODES; i>0; i-- ))
do
        echo ""
        echo "UPGRADING node$i..."
        ssh node$i sudo apt update
        ssh node$i sudo apt full-upgrade --yes
        ssh node$i sudo apt autoremove --yes
        ssh node$i sudo shutdown -r now
done
```

9.3.5 Select BLAS Library

We have installed four BLAS libraries...

Confirm all nodes are using the same one initially...

ssh node1 sudo update-alternatives –config libblas.so.3-aarch64-linux-gnu

TODO screen output...

Confirm option 0, OpenBLAS, is selected. Press return to keep this option and then exit.

10 Appendix B - High-Performance Linpack (HPL) Installation

Download and install the latest version of HPL on node1...

```
ssh node1
cd picluster
mkdir hpl
cd hpl
wget https://www.netlib.org/benchmark/hpl/hpl-2.3.tar.gz
gunzip hpl-2.3.tar.gz
tar xvf hpl-2.3.tar
rm hpl-2.3.tar
cd hpl-2.3
```

Create Make.serial file...

```
cd setup
bash make_generic
cd ..
cp setup/Make.UNKNOWN Make.serial
```

Amend Make.serial as per...

Build...

```
make arch=serial
```

This creates xhpl and HPL.dat in bin/serial

Copy xhpl to all nodes (only xhpl, and not HPL.dat)...

```
ssh node2 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node3 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node4 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node5 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node6 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node7 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node8 mkdir -p picluster/hpl/hpl-2.3/bin/serial
ssh node9 mkdir -p picluster/hpl/hpl-2.3/bin/serial

scp bin/serial/xhpl node2:~picluster/hpl/hpl-2.3/bin/serial
scp bin/serial/xhpl node3:~picluster/hpl/hpl-2.3/bin/serial
scp bin/serial/xhpl node4:~picluster/hpl/hpl-2.3/bin/serial
scp bin/serial/xhpl node5:~picluster/hpl/hpl-2.3/bin/serial
```

```
scp bin/serial/xhpl node6:~picluster/hpl/hpl-2.3/bin/serial
scp bin/serial/xhpl node7:~picluster/hpl/hpl-2.3/bin/serial
scp bin/serial/xhpl node8:~picluster/hpl/hpl-2.3/bin/serial
scp bin/serial/xhpl node9:~picluster/hpl/hpl-2.3/bin/serial
```

11 Appendix D - Hints and Tips

Hints from experience... and time savers... for building a development cluster on a local network.

11.1 IP/MAC Addresses

If IP/MAC address assignments get confused, which is easily done during initial build, view IP address assignments on the local network with:

```
arp -a
```

Then delete *incomplete* IP addresses with:

```
sudo arp -d incomplete-ip-address
```

11.2 SSH known_hosts

If *ssh* reports differing keys in 'known-hosts', and warns of a potential 'man-in-the-middle-attack', then just delete 'known-hosts':

```
sudo rm ~/.ssh/known_hosts
```

'known_hosts' will be re-populated as you log into each node.

11.3 Package 'unattended-upgrades'

Unattended upgrades can interfere/stop long running jobs. So, remove the package 'unattended-upgrades' from all nodes:

```
sudo apt remove unattended-upgrades
```

But don't forget to do regular manual upgrades!

11.4 tmux

`tmux` is your friend!

Monitoring long running jobs from a workstation, which goes to sleep after a period of no activity, for example, may interfere with the running of the jobs if a SSH connection is broken.

Use a `tmux` session to start long running jobs, and then detach from the `tmux` session. The job will quite happily run in the background on the cluster. Turn the workstation off and go to bed. In the morning, turn the workstation on and 'attach' to the `tmux` session. All will be well.

11.5 git

`git` is your best friend!

During your cluster build you will accidentally delete files, results etc. After every significant...