HPL Assignment

ShanghaiTechU, GeekPie_ HPC

Create cluster

1. make a reservation for 1 head node and 2 compute nodes
2. bind a floating ip to the head node
3. set up a nfs server on the head node
4. install `spack` on the head node in the nfs directory
5. install hpl package
6. tuning

Problem 1

4(calculate 4 float64 in one instruction) * 2(FMA instructions fuse multiply and add) * 2(IPC) * 2.6(GHz) * 10(Cores) * 2(CPUs) = 832 GFLOPs

Problem 2

BLAS Library

According to Intel's performance benchmark and CUHK's benchmark, Intel oneAPI MKL is faster.

Our experiment on indyscc_comopute_node:

openblas + intel-oneapi-mpi 465 GFLOPs

intel-oneapi-mkl + intel-oneapi-mpi 492GFLOPs

So, We chose Intel MKI since it is significantly faster.

MPI Library

Our experiment on indyscc_comopute_node:

intel-oneapi-mkl + openmpi 478 GFLOPs

intel-oneapi-mkl + intel-oneapi-mpi 491 GFLOPs

```
spack install -j 20 hpl ^intel-oneapi-mkl ^intel-oneapi-mpi
spack install -j 20 hpl ^openblas ^intel-oneapi-mpi
spack install -j 20 hpl ^intel-oneapi-mkl ^openmpi
```

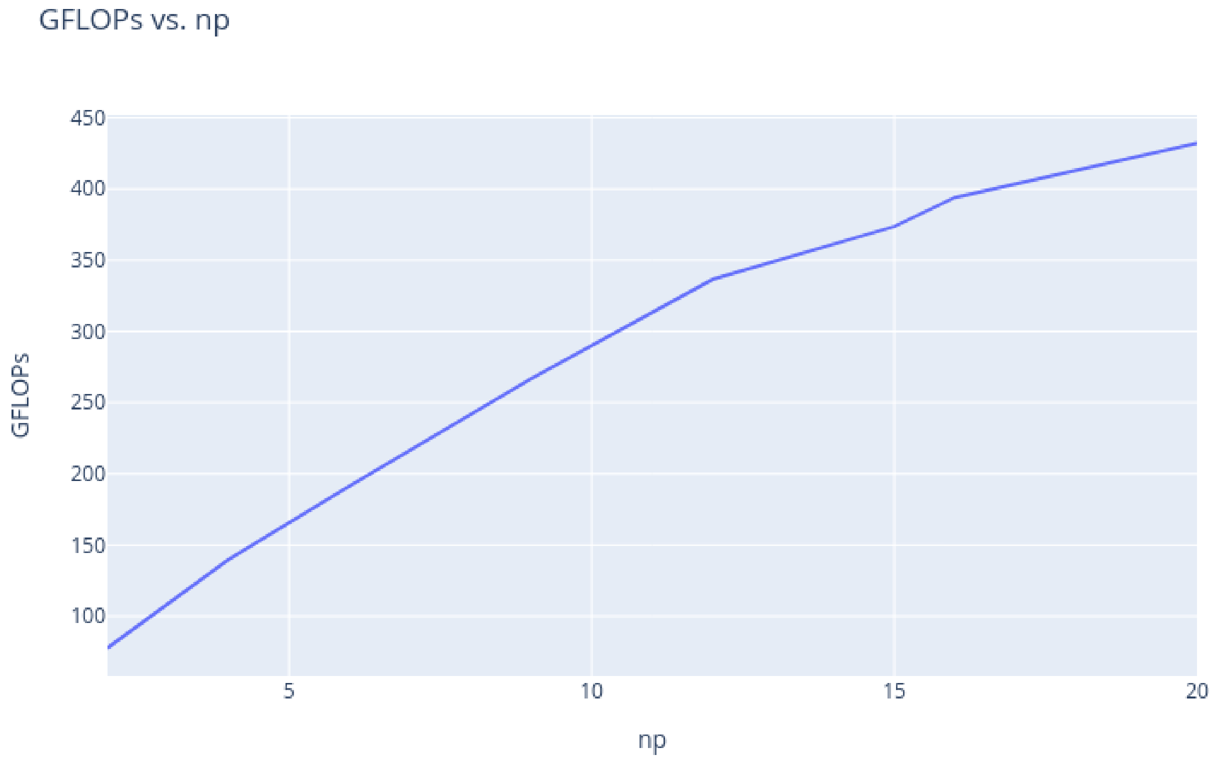
HPL.dat:

```
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out      output file name (if any)
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
20000        Ns
1            # of NBs
100          # of problems sizes (N)
0            MAP process mapping (0=Row-,1=Column-major)
1            # of process grids (P x Q)
4            Ps
5            Qs
16.0         threshold
1            # of panel fact<
2            PFACTs (0=left, 1=Crout, 2=Right)
1            # of recursive stopping criterium
4            NBMINs (>= 1)
1            # of panels in recursion
2            NDIVs
1            # of recursive panel fact.
1            RFACTs (0=left, 1=Crout, 2=Right)
1            # of broadcast
1            BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1            # of lookahead depth
1            DEPTHs (>=0)
2            SWAP (0=bin-exch,1=long,2=mix)
64           swapping threshold
0            L1 in (0=transposed,1=no-transposed) form
0            U in (0=transposed,1=no-transposed) form
1            Equilibration (0=no,1=yes)
8            memory alignment in double (> 0)
```

Problem 3

Run environment: indyscc_compute_node, CentOS

```
hpl@2.3%gcc@8.5.0~openmp arch=linux-centos8-haswell
    ^intel-oneapi-mkl@2022.1.0%gcc@8.5.0~cluster~ilp64+shared arch=linux-centos8-
haswell
        ^intel-oneapi-tbb@2021.6.0%gcc@8.5.0 arch=linux-centos8-haswell
    ^intel-oneapi-mpi@2021.6.0%gcc@8.5.0~external-libfabric~generic-names~ilp64
arch=linux-centos8-haswell
```



P	Q	np	GFLOPs
1	2	2	7.7712e+01
2	2	4	1.3986e+02
2	3	6	1.9141e+02
2	4	8	2.4142e+02
3	3	9	2.6672e+02
3	4	12	3.3658e+02
3	5	15	3.7365e+02
4	4	16	3.9404e+02
4	5	20	4.3216e+02

```
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out     output file name (if any)
6           device out (6=stdout,7=stderr,file)
1           # of problems sizes (N)
10000       Ns
1           # of NBs
100         # of problems sizes (N)
0           MAP process mapping (0=Row-,1=Column-major)
1           # of process grids (P x Q)
4           Ps
5           Qs
16.0        threshold
1           # of panel fact<
2           PFACTs (0=left, 1=Crout, 2=Right)
1           # of recursive stopping criterium
4           NBMINs (>= 1)
1           # of panels in recursion
2           NDIVs
1           # of recursive panel fact.
1           RFACTs (0=left, 1=Crout, 2=Right)
1           # of broadcast
1           BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1           # of lookahead depth
1           DEPTHs (>=0)
2           SWAP (0=bin-exch,1=long,2=mix)
64          swapping threshold
0           L1 in (0=transposed,1=no-transposed) form
0           U in (0=transposed,1=no-transposed) form
1           Equilibration (0=no,1=yes)
8           memory alignment in double (> 0)
```

Problem 4

590.29 GFLOPs

We tried to use the intel-oneapi-compilers but hpl it compiles cannot pass the test.

`mpich` and `mvapich2` cannot run.

```
[mpiexec@hpl-head.novalocal] control_cb (pm/pmiserv/pmiserv_cb.c:206): assert
(!closed) failed
[mpiexec@hpl-head.novalocal] HYDT_dmxu_poll_wait_for_event
(tools/demux/demux_poll.c:76): callback returned error status
[mpiexec@hpl-head.novalocal] HYD_pmci_wait_for_completion
(pm/pmiserv/pmiserv_pmci.c:160): error waiting for event
[mpiexec@hpl-head.novalocal] main (ui/mpich/mpiexec.c:325): process manager error
waiting for completion
```

According to the memory space, we calculate the best Ns.

We tried to adjust the value of NB.

```
================================================================================
HPLinpack 2.3  --  High-Performance Linpack benchmark  --   December 2, 2018
Written by A. Petitet and R. Clint Whaley,  Innovative Computing Laboratory, UTK
Modified by Piotr Luszczek, Innovative Computing Laboratory, UTK
Modified by Julien Langou, University of Colorado Denver
================================================================================

An explanation of the input/output parameters follows:
T/V    : Wall time / encoded variant.
N      : The order of the coefficient matrix A.
NB     : The partitioning blocking factor.
P      : The number of process rows.
Q      : The number of process columns.
Time   : Time in seconds to solve the linear system.
Gflops : Rate of execution for solving the linear system.

The following parameter values will be used:

N      :  102144
NB     :     192
PMAP   : Row-major process mapping
P      :       4
Q      :       5
PFACT  :   Right
NBMIN  :       4
NDIV   :       2
RFACT  :   Crout
BCAST  :  1ringM
DEPTH  :       1
SWAP   : Mix (threshold = 64)
L1     : transposed form
U      : transposed form
EQUIL  : yes
ALIGN  : 8 double precision words

--------------------------------------------------------------------------------

- The matrix A is randomly generated for each test.
- The following scaled residual check will be computed:
      ||Ax-b||_oo / ( eps * ( || x ||_oo * || A ||_oo + || b ||_oo ) * N )
- The relative machine precision (eps) is taken to be               1.110223e-16
- Computational tests pass if scaled residuals are less than                16.0

================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR11C2R4      102144   192     4     5            1203.62             5.9029e+02
HPL_pdgesv() start time Tue Sep 20 14:30:42 2022
HPL_pdgesv() end time   Tue Sep 20 14:50:45 2022

--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=   2.87742271e-03 ...... PASSED
================================================================================

Finished      1 tests with the following results:
              1 tests completed and passed residual checks,
              0 tests completed and failed residual checks,
              0 tests skipped because of illegal input values.
--------------------------------------------------------------------------------

End of Tests.
================================================================================
```

```
hpl@2.3%gcc@8.5.0~openmp arch=linux-centos8-haswell
    ^intel-oneapi-mkl@2022.1.0%oneapi@2022.1.0~cluster~ilp64+shared arch=linux-centos8-haswell
        ^intel-oneapi-tbb@2021.6.0%oneapi@2022.1.0 arch=linux-centos8-haswell
    ^intel-oneapi-mpi@2021.6.0%gcc@8.5.0~external-libfabric~generic-names~ilp64 arch=linux-centos8-haswell
```

HPL.dat

```
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out      output file name (if any)
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
102144         Ns
1            # of NBs
192           NBs
0            PMAP process mapping (0=Row-,1=Column-major)
1            # of process grids (P x Q)
4            Ps
5            Qs
16.0         threshold
1            # of panel fact
2            PFACTs (0=left, 1=Crout, 2=Right)
1            # of recursive stopping criterium
4            NBMINs (>= 1)
1            # of panels in recursion
2            NDIVs
1            # of recursive panel fact.
1            RFACTs (0=left, 1=Crout, 2=Right)
1            # of broadcast
1            BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1            # of lookahead depth
1            DEPTHs (>=0)
2            SWAP (0=bin-exch,1=long,2=mix)
64           swapping threshold
0            L1 in (0=transposed,1=no-transposed) form
0            U  in (0=transposed,1=no-transposed) form
1            Equilibration (0=no,1=yes)
8            memory alignment in double (> 0)
##### This line (no. 32) is ignored (it serves as a separator). ######
0                               Number of additional problem sizes for PTRANS
1200 10000 30000                values of N
0                               number of additional blocking sizes for PTRANS
40 9 8 13 13 20 16 32 64        values of NB
```

run_hpl.sh

```
spack load hpl -openmp ^intel-oneapi-mkl ^intel-oneapi-mpi % gcc
# spack load hpl -openmp ^intel-oneapi-mkl ^openmpi % gcc

echo 3 > /proc/sys/vm/drop_caches
echo 1 > /proc/sys/vm/compact_memory
echo 0 > /proc/sys/kernel/numa_balancing
echo 'always' > /sys/kernel/mm/transparent_hugepage/enabled
echo 'always' > /sys/kernel/mm/transparent_hugepage/defrag
sleep 10
sudo cpupower frequency-set -g performance

mpi_options="$mpi_options --bind-to core --map-by core:PE=1"
OMPI_ALLOW_RUN_AS_ROOT=1 OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1 mpirun $mpi_options -np 20 xhpl
```

Problem 5

780 GFLOPs

No, the GFLOPs number is not exactly twice that of your single-node performance.

1. transfering data between nodes is costly
2. scheduling overhead

run command

```
mpirun -hostfile hostfile --bind-to core --map-by core:PE=1 --report-bindings -np 40 xhpl
```

HPL.dat

```
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out      output file name (if any)
6            device out (6=stdout,7=stderr,file)
1            # of problems sizes (N)
30000         Ns
1            # of NBs
192           NBs
0            PMAP process mapping (0=Row-,1=Column-major)
1            # of process grids (P x Q)
8            Ps
5            Qs
16.0         threshold
1            # of panel fact
2            PFACTs (0=left, 1=Crout, 2=Right)
1            # of recursive stopping criterium
4            NBMINs (>= 1)
1            # of panels in recursion
2            NDIVs
1            # of recursive panel fact.
1            RFACTs (0=left, 1=Crout, 2=Right)
1            # of broadcast
1            BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1            # of lookahead depth
1            DEPTHs (>=0)
2            SWAP (0=bin-exch,1=long,2=mix)
64           swapping threshold
0            L1 in (0=transposed,1=no-transposed) form
0            U  in (0=transposed,1=no-transposed) form
1            Equilibration (0=no,1=yes)
8            memory alignment in double (> 0)
##### This line (no. 32) is ignored (it serves as a separator). ######
0                               Number of additional problem sizes for PTRANS
1200 10000 30000                values of N
0                               number of additional blocking sizes for PTRANS
40 9 8 13 13 20 16 32 64        values of NB
```

hostfile

```
10.20.30.96
10.20.31.251
```

HPL and Dependencies

```
Input spec
--------------------------------
hpl~openmp
    ^intel-oneapi-mkl
    ^openmpi%gcc

Concretized
--------------------------------
hpl@2.3%gcc@8.5.0~openmp arch=linux-centos8-haswell
    ^intel-oneapi-mkl@2022.1.0%oneapi@2022.1.0~cluster~ilp64+shared arch=linux-centos8-haswell
        ^intel-oneapi-tbb@2021.6.0%oneapi@2022.1.0 arch=linux-centos8-haswell
    ^openmpi@4.1.4%gcc@8.5.0~atomics~cuda~cxx~cxx_exceptions~gpfs~internal-hwloc~java~legacylaunchers~lustre~memchecker+romio+rsh~singularity+static+vt+wrapper-rpath fabrics=none schedulers=none arch=linux-centos8-haswell
        ^hwloc@2.8.0%gcc@8.5.0~cairo~cuda~gl~libudev+libxml2~netloc~nvml~oneapi-level-zero~opencl+pci~rocm+shared arch=linux-centos8-haswell
            ^libpciaccess@0.16%gcc@8.5.0 arch=linux-centos8-haswell
                ^libtool@2.4.7%gcc@8.5.0 arch=linux-centos8-haswell
                    ^m4@1.4.19%gcc@8.5.0+sigsegv patches=9dc5fbd,bfdffa7 arch=linux-centos8-haswell
                        ^diffutils@3.8%gcc@8.5.0 arch=linux-centos8-haswell
                            ^libiconv@1.16%gcc@8.5.0 libs=shared,static arch=linux-centos8-haswell
                        ^libsigsegv@2.13%gcc@8.5.0 arch=linux-centos8-haswell
                ^pkgconf@1.8.0%gcc@8.5.0 arch=linux-centos8-haswell
                ^util-macros@1.19.3%gcc@8.5.0 arch=linux-centos8-haswell
            ^libxml2@2.10.1%gcc@8.5.0~python arch=linux-centos8-haswell
                ^xz@5.2.5%gcc@8.5.0~pic libs=shared,static arch=linux-centos8-haswell
                ^zlib@1.2.12%gcc@8.5.0+optimize+pic+shared patches=0d38234 arch=linux-centos8-haswell
            ^ncurses@6.3%gcc@8.5.0~symlinks+termlib abi=none arch=linux-centos8-haswell
        ^numactl@2.0.14%gcc@8.5.0 patches=4e1d78c,62fc8a8,ff37630 arch=linux-centos8-haswell
            ^autoconf@2.69%gcc@8.5.0 patches=35c4492,7793209,a49dd5b arch=linux-centos8-haswell
                ^perl@5.34.1%gcc@8.5.0+cpanm+shared+threads arch=linux-centos8-haswell
                    ^berkeley-db@18.1.40%gcc@8.5.0+cxx~docs+stl patches=26090f4,b231fcc arch=linux-centos8-haswell
                    ^bzip2@1.0.8%gcc@8.5.0~debug~pic+shared arch=linux-centos8-haswell
                    ^gdbm@1.19%gcc@8.5.0 arch=linux-centos8-haswell
                        ^readline@8.1.2%gcc@8.5.0 arch=linux-centos8-haswell
            ^automake@1.16.5%gcc@8.5.0 arch=linux-centos8-haswell
        ^openssh@9.0p1%gcc@8.5.0+gssapi arch=linux-centos8-haswell
            ^krb5@1.19.3%gcc@8.5.0+shared arch=linux-centos8-haswell
                ^bison@3.8.2%gcc@8.5.0 arch=linux-centos8-haswell
                ^gettext@0.21%gcc@8.5.0+bzip2+curses+git~libunistring+libxml2+tar+xz
```

```
arch=linux-centos8-haswell
                    ^tar@1.34%gcc@8.5.0 zip=pigz arch=linux-centos8-haswell
                        ^pigz@2.7%gcc@8.5.0 arch=linux-centos8-haswell
                        ^zstd@1.5.2%gcc@8.5.0+programs compression=none
libs=shared,static arch=linux-centos8-haswell
                ^openssl@1.1.1q%gcc@8.5.0~docs~shared certs=mozilla patches=3fdcf2d
arch=linux-centos8-haswell
                    ^ca-certificates-mozilla@2022-07-19%gcc@8.5.0 arch=linux-centos8-
haswell
            ^libedit@3.1-20210216%gcc@8.5.0 arch=linux-centos8-haswell
        ^pmix@4.1.2%gcc@8.5.0~docs+pmi_backwards_compatibility~restful arch=linux-
centos8-haswell
            ^libevent@2.1.12%gcc@8.5.0+openssl arch=linux-centos8-haswell
```

Reference:

HPL Benchmark

Tune HPL dat file