2022 APAC HPC-AI

Team NTHU-1 Presentation

Hao-Lung, Hsiao
Hsin-Ping, Peng
Pin-Syuan, Lee
Chun-Mu, Weng
Hsin-Cheng, Tu
Jing-Yu, Yang

Dept. of Computer Science, Nat'l Tsing Hua U.

October 14, 2022

Section 1

High Performance Computing with QUANTUM ESPRESSO

Single Node Performance of Gadi module

Average of 5 Times

# CPUs (`np`)	# pools (`npool`)	# linear algebra groups (`ndiag`)	CPU time [s]
48	24	4	1m53.138s
48	24	1	1m53.54s
40	20	4	1m52.794s
40	20	1	1m52.941s

Single Node Performance of Intel Compiler + Intel MPI

Average of 5 Times

# CPUs (`np`)	# pools (`npool`)	# linear algebra groups (`ndiag`)	CPU time [s]
48	24	4	1min58.916s
48	24	1	1min59.004s
40	20	4	1min56.944s
40	20	1	1min57.084s

Summary

Script

```
#!/bin/bash
#PBS -l walltime=00:10:00
#PBS -l ncpus=40
#PBS -l mem=190GB
#PBS -l software=qe
#PBS -l wd
#PBS -P jx00
#PBS -N QE-single

module load qe
export OMP_NUM_THREADS=1
mpirun -np 40 pw.x -npool 20 -ndiag 4 -inp CeO2.in
```

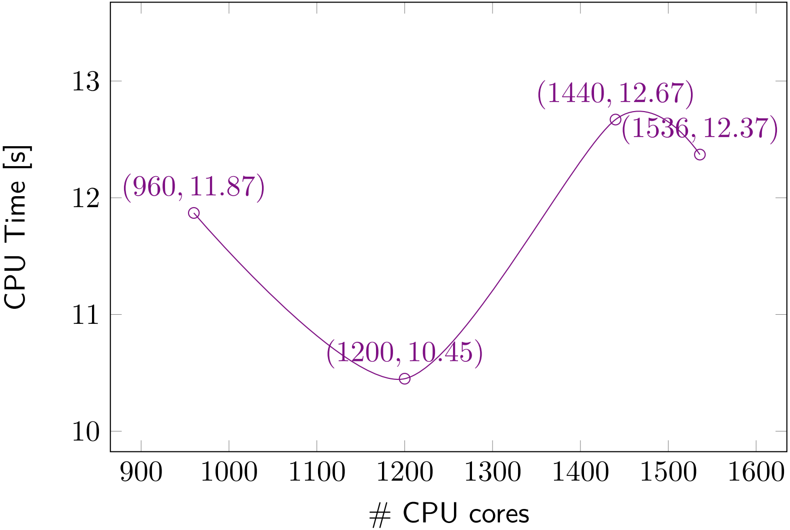
Summary (cont.)

Result

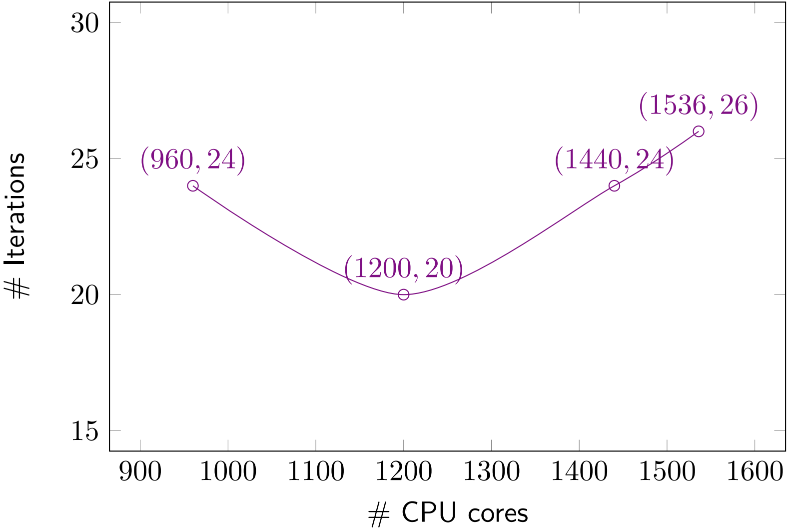
```
nevikw39 — cw2590@gadi-login-09:~/qe — ssh gadi — 80×24
     g_psi        :      0.14s CPU      0.15s WALL (     175 calls)

     Called by h_psi:
     h_psi:calbec :      3.77s CPU      3.78s WALL (     202 calls)
     vloc_psi     :     38.07s CPU     38.32s WALL (     202 calls)
     add_vuspsi   :      4.26s CPU      4.27s WALL (     202 calls)

     General routines
     calbec       :      4.97s CPU      4.98s WALL (     228 calls)
     fft          :      5.15s CPU      5.25s WALL (     349 calls)
     ffts         :      0.26s CPU      0.28s WALL (      53 calls)
     fftw         :     32.90s CPU     33.13s WALL (   19368 calls)
     interpolate  :      0.70s CPU      0.73s WALL (      27 calls)

     Parallel routines

     PWSCF        :   1m52.55s CPU   1m58.50s WALL


   This run was terminated on:   0:21:13  13Oct2022

=------------------------------------------------------------------------------=
   JOB DONE.
=------------------------------------------------------------------------------=
```

CPU Time vs. # CPU cores

`npools` were 20, 20, 20, 24 resp.; `ndiags` were left as default

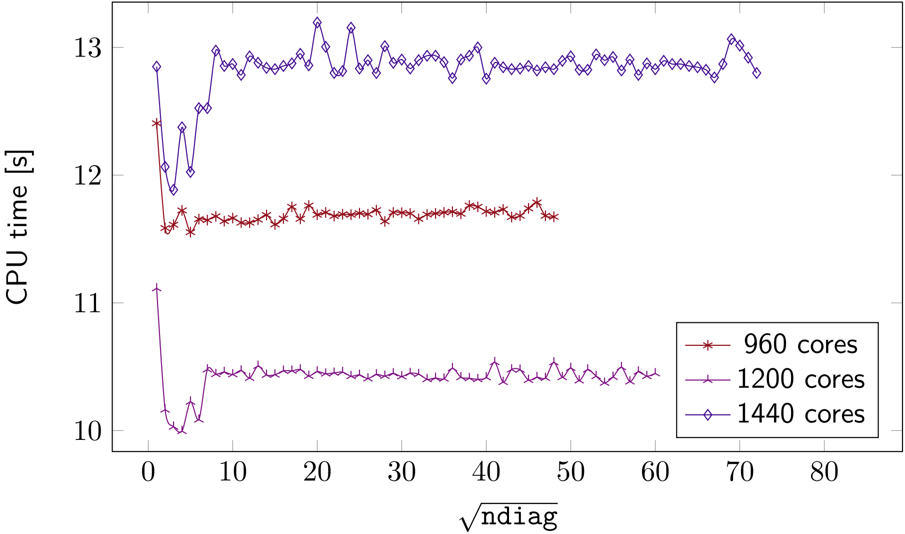


Iterations vs. # CPU cores



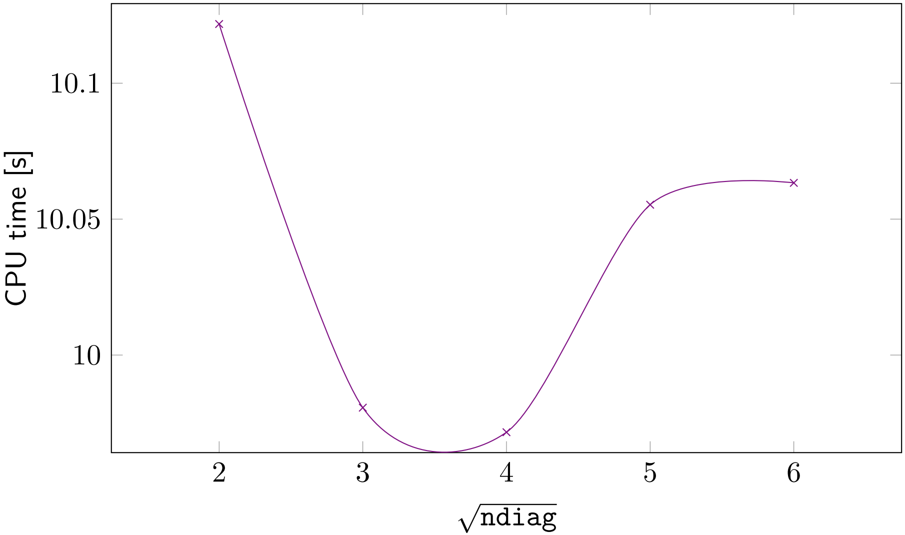
CPU time vs. `ndiag` of different CPU cores

Average of 5 Times



CPU time vs. `ndiag` of 1200 CPU cores

A closer, deeper insight



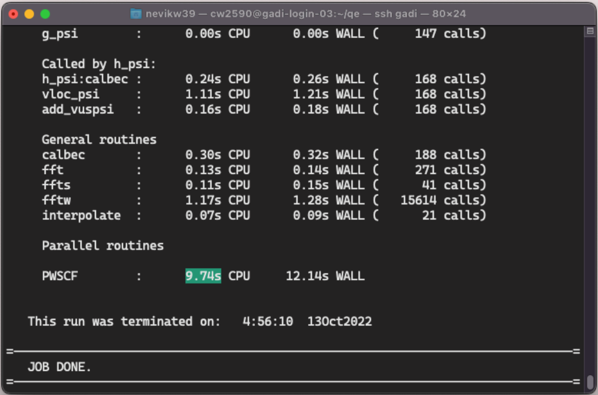
Conclusion

Script

```
#!/bin/bash
#PBS -l walltime=00:10:00
#PBS -l ncpus=1200
#PBS -l mem=760GB
#PBS -l software=qe
#PBS -l wd
#PBS -P jx00
#PBS -N QE-multi

module load qe
export OMP_NUM_THREADS=1
mpirun -np 1200 pw.x -npool 20 -ndiag 16 -inp CeO2.in
```

Conclusion (cont.)

Result



```
nevikw39 — cw2590@gadi-login-03:~/qe — ssh gadi — 80×24
     g_psi        :      0.00s CPU      0.00s WALL (     147 calls)

     Called by h_psi:
     h_psi:calbec :      0.24s CPU      0.26s WALL (     168 calls)
     vloc_psi     :      1.11s CPU      1.21s WALL (     168 calls)
     add_vuspsi   :      0.16s CPU      0.18s WALL (     168 calls)

     General routines
     calbec       :      0.30s CPU      0.32s WALL (     188 calls)
     fft          :      0.13s CPU      0.14s WALL (     271 calls)
     ffts         :      0.11s CPU      0.15s WALL (      41 calls)
     fftw         :      1.17s CPU      1.28s WALL (   15614 calls)
     interpolate  :      0.07s CPU      0.09s WALL (      21 calls)

     Parallel routines

     PWSCF        :      9.74s CPU     12.14s WALL


   This run was terminated on:   4:56:10  13Oct2022

=------------------------------------------------------------------------------=
   JOB DONE.
=------------------------------------------------------------------------------=
```

Section 2

Communications Performance with UCX

Baseline

Average of 10 iterations, small data set (i.e., each chunk with 10^6 rows), running on 16 GPUs over 4 Gadi Volta nodes.

Throughput

- 4.27 GiB/s

Enable Hardware Tag Matching

Avg. of 10 iterations, small data set, Gadi Volta nodes

Enable hardware tag matching for both *Reliable Connected (RC)* and *Dynamically Connected (DC)* so that these works are offload to NICs.

Config

```
export UCX_RC_MLX5_TM_ENABLE=y
export UCX_DC_MLX5_TM_ENABLE=y
```

Throughput

- 4.36 GiB/s
- 102.1% speedup

Enable various optimizations intended for homogeneous environment

Avg. of 10 iterations, small data set, Gadi Volta nodes

Enabling this mode implies that the local transport resources/devices of all entities which connect to each other are the same.
Nevertheless, this option would be conflict to the *rendezvous* scheme we would choose.

Config

```
export UCX_UNIFIED_MODE=y
```

Throughput

- 4.39 GiB/s
- 102.8% speedup

Increase the amount of buffers added every time the receive / send memory pool grows

Avg. of 10 iterations, small data set, Gadi Volta nodes

The default values were 8.
Nonetheless, we found that this option would hardly give rise to ideal promotion in combination with others.

Config

```
export UCX_TCP_RX_BUFS_GROW=16
export UCX_TCP_TX_BUFS_GROW=16
```

Throughput

- 4.68 GiB/s
- 109.6% speedup

Use **mutex** instead of **spinlock** for multithreading support in UCP

Avg. of 10 iterations, small data set, Gadi Volta nodes

Config

```
export UCX_USE_MT_MUTEX=y
```

Throughput

- 4.71 GiB/s
- 110.3% speedup

Enable UCX-Py non-blocking mode

Avg. of 10 iterations, small data set, Gadi Volta nodes

Config

```
export UCXPY_NON_BLOCKING_MODE=1
```

Throughput

- 4.96 GiB/s
- 116.1% speedup

Set *Rendezvous* protocol to use *Active Messages* scheme

Avg. of 10 iterations, small data set, Gadi Volta nodes

This option is not documented in detail, but we found that it brought significant improvement in performance.

Config

```
export UCX_RNDV_SCHEME=am
```

Throughput

- 5.54 GiB/s
- 129.7% speedup

Miscellanies

- UCX_IB_GPU_DIRECT_RDMA
- UCX_RNDV_THRESH
- UCX_TCP_TX_SEG_SIZE, UCX_TCP_RX_SEG_SIZE

Optimal Combination of Configurations

Config

```
export UCX_RC_TM_ENABLE=y
export UCX_DC_TM_ENABLE=y

export UCX_USE_MT_MUTEX=y

export UCXPY_NON_BLOCKING_MODE=1

export UCX_RNDV_SCHEME=am

export UCX_IB_GPU_DIRECT_RDMA=y
export UCX_RNDV_THRESH=1024
export UCX_TCP_TX_SEG_SIZE=64k
export UCX_TCP_RX_SEG_SIZE=512k
```

Overall Throughput Result

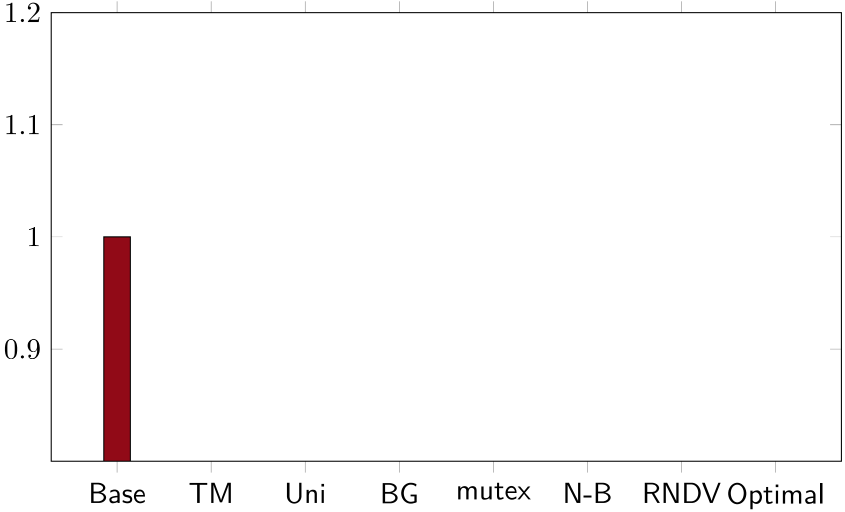
Avg. of 100 iterations on 16 GPUs over 4 Gadi Volta nodes

Small Data Set 9.28 GiB/s,
217.3% speedup in comparison to baseline

Large Data Set 12.37 GiB/s,
281.1% speedup in comparison to baseline (large one, 4.40 GiB/s)

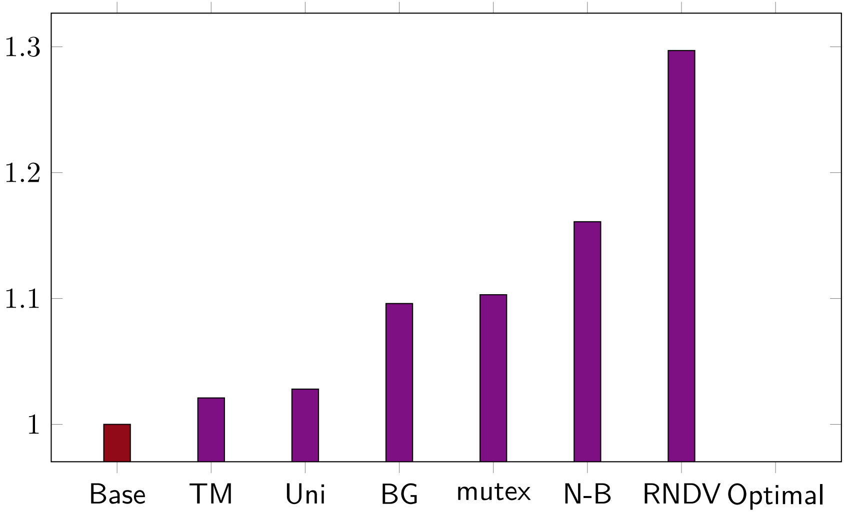
Bar graph of speedup

Avg. of 10 iterations, small data set, Gadi Volta nodes



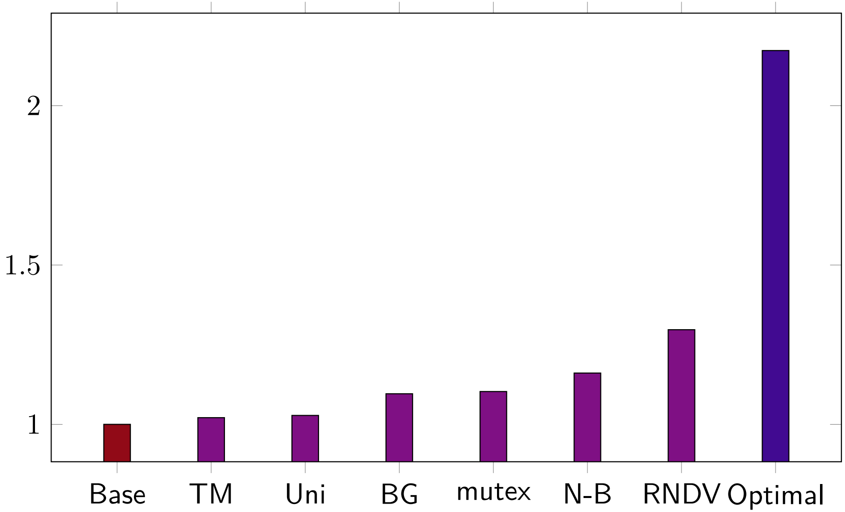
Bar graph of speedup

Avg. of 10 iterations, small data set, Gadi Volta nodes



Bar graph of speedup

Avg. of 10 iterations, small data set, Gadi Volta nodes



Overall Throughput Result on DGX-A100 nodes

Avg. of 100 iterations on 16 GPUs over 2 Gadi DGX-A100 nodes

For the small data set, the throughput with default config was 16.66 Gib/s while the throughput was increase slightly to 16.69 GiB/s when all optimized options enabled.

Overall Throughput Result on DGX-A100 nodes

Avg. of 100 iterations on 16 GPUs over 2 Gadi DGX-A100 nodes

For the small data set, the throughput with default config was 16.66 Gib/s while the throughput was increase slightly to 16.69 GiB/s when all optimized options enabled.
When it comes to the large data set, the throughput with default config was 88.29 Gib/s. Nevertheless, if all options we previously found effective were enabled, the throughput dropped to 34.89 GiB/s drastically. Then if we switch *Rendezvous* protocol back to default scheme, the throughput became 89.00 GiB/s.

Section 3

Deep-Learning-based DNA Sequence fast decoding

This is a title

This is a subtitle

- item
- item

1. 1
2. 2

Figure: Caption

Block Name

This is a block.

Alert Block Name

This is an alert block.

Example Block Name

This is an example block.