NVIDIA CUDA-aware MPI

PAX-HPC CUDA/OpenACC Workshop

Michael Bareford
m.bareford@epcc.ed.ac.uk

OpenMPI

UCX

THE UNIVERSITY of EDINBURGH

cirrus

Reusing this material

NVIDIA HPC SDK

- Several versions of the NV HPC SDK are installed on Cirrus and accessed via TCL (Tool Command Language) module files.

```
module load nvidia/nvhpc-nompi/24.5
module load openmpi/4.1.6-cuda-12.4
```

- The SDK will contain specific CUDA versions (e.g., 11.8, 12.4) that should be compatible with the underlying GPU driver (currently v550.144.03).

- OpenMPI and UCX libraries must be built against the correct CUDA API version. Otherwise, you may see errors like the following when you run your application.

> **CUDA error**: the provided PTX was compiled with an unsupported toolchain.

nvidia-smi

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 550.144.03    Driver Version: 550.144.03    CUDA Version: 12.4   |
|                                                                             |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  Off  | 00000000:1A:00.0 Off |                  Off |
| N/A   39C    P0    56W / 300W |      0MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  Off  | 00000000:1C:00.0 Off |                  Off |
| N/A   38C    P0    60W / 300W |      0MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  Off  | 00000000:88:00.0 Off |                  Off |
| N/A   38C    P0    60W / 300W |      0MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  Off  | 00000000:8A:00.0 Off |                  Off |
| N/A   39C    P0    58W / 300W |      0MiB / 16384MiB |      2%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
```

What is CUDA-aware MPI?

- Pointers to GPU device memory can be handled directly by MPI calls.

```
module load nvidia/nvhpc-nompi/24.5
module load openmpi/4.1.6-cuda-12.4
```

- Otherwise, memory operations have to go through host memory.

What is CUDA-aware MPI?

- Pointers to GPU device memory can be handled directly by MPI calls.

```
module load nvidia/nvhpc-nompi/24.5
module load openmpi/4.1.6-cuda-12.4
```

- Otherwise, memory operations have to go through host memory.

```
if (rank=0) MPI_Send(d_buf, bufsize, MPI_INT, 1, 0, MPI_COMM_WORLD);

if (rank=1) MPI_Recv(d_buf, bufsize, MPI_INT, 1, 0, MPI_COMM_WORLD, &status);
```

What is CUDA-aware MPI?

- Pointers to GPU device memory can be handled directly by MPI calls.

```
module load nvidia/nvhpc-nompi/24.5
module load openmpi/4.1.6-cuda-12.4
```

- Otherwise, memory operations have to go through host memory.

```
if (rank=0) MPI_Send(d_buf, bufsize, MPI_INT, 1, 0, MPI_COMM_WORLD);

if (rank=1) MPI_Recv(d_buf, bufsize, MPI_INT, 1, 0, MPI_COMM_WORLD, &status);
```

```
MPI_Reduce(d_sendbuf, d_recvbuf, count, MPI_INT, MPI_SUM, root, comm);
```

What is CUDA-aware MPI?

- Pointers to GPU device memory can be handled directly by MPI calls.

- Otherwise, memory operations have to go through host memory.

- The four NVIDIA V100 GPUs on each Cirrus GPU node are connected via NVLink2.
  - each GPU is connected to the other three
  - each NVLink2 connection has total bi-directional bandwidth of 100 GB/s

- Off-node GPU-to-GPU comms is handled via Infiniband interconnect.

Cirrus GPU node topology

```
[mrb@cirrus-login2]$ srun --exclusive --nodes=1 --time=00:20:00 \
                         --partition=gpu --qos=gpu --gres=gpu:4 \
                         --account=[budget code] --pty /usr/bin/bash --login
```

```
[mrb@r2i5n0]$ nvidia-smi topo -m

        GPU0    GPU1    GPU2    GPU3    mlx5_0  mlx5_1  mlx5_2  mlx5_3  CPU Affinity    NUMA
Affinity
GPU0     X      NV2     NV2     NV2     PIX     NODE    SYS     SYS     0-19,40-59      0
GPU1    NV2      X      NV2     NV2     PIX     NODE    SYS     SYS     0-19,40-59      0
GPU2    NV2     NV2      X      NV2     SYS     SYS     PIX     NODE    20-39,60-79     1
GPU3    NV2     NV2     NV2      X      SYS     SYS     PIX     NODE    20-39,60-79     1
mlx5_0  PIX     PIX     SYS     SYS      X      NODE    SYS     SYS
mlx5_1  NODE    NODE    SYS     SYS     NODE     X      SYS     SYS
mlx5_2  SYS     SYS     PIX     PIX     SYS     SYS      X      NODE
mlx5_3  SYS     SYS     NODE    NODE    SYS     SYS     NODE     X
```

Cirrus GPU node topology

NV#: connection traversing a bonded set of # NVLinks
PIX: connection traversing at most a single PCIe bridge
NODE: connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node
SYS: connection traversing PCIe as well as the SMP interconnect between NUMA nodes

```
[mrb@r2i5n0]$ nvidia-smi topo -m

        GPU0    GPU1    GPU2    GPU3    mlx5_0  mlx5_1  mlx5_2  mlx5_3  CPU Affinity    NUMA
Affinity
GPU0     X      NV2     NV2     NV2     PIX     NODE    SYS     SYS     0-19,40-59      0
GPU1    NV2      X      NV2     NV2     PIX     NODE    SYS     SYS     0-19,40-59      0
GPU2    NV2     NV2      X      NV2     SYS     SYS     PIX     NODE    20-39,60-79     1
GPU3    NV2     NV2     NV2      X      SYS     SYS     PIX     NODE    20-39,60-79     1
mlx5_0  PIX     PIX     SYS     SYS      X      NODE    SYS     SYS
mlx5_1  NODE    NODE    SYS     SYS     NODE     X      SYS     SYS
mlx5_2  SYS     SYS     PIX     PIX     SYS     SYS      X      NODE
mlx5_3  SYS     SYS     NODE    NODE    SYS     SYS     NODE     X
```

Compiling UCX 1.16.0

UCX is an open-source optimized comms library that supports multiple networks, including InfiniBand.
Is the Point-to-point Management Layer (PML) within OpenMPI.

```
...

module load gcc/10.2.0
module load nvidia/nvhpc-nompi/24.5

./configure CC=gcc CXX=g++ FC=gfortran \
    --with-knem=/opt/knem-1.1.4.90mlnx2 \
    --with-cuda=${NVHPC_ROOT}/cuda/12.4 \
    --with-mlx5-dv --enable-mt \
    --prefix=${PRFX}/ucx/1.16.0-cuda-12.4

make -j 8
make -j 8 install
```

https://github.com/hpc-uk/build-instructions/blob/main/libs/ucx/build_ucx_1.16.0_cirrus_gcc10.md

Compiling OpenMPI 4.1.6

```
...

module load gcc/10.2.0
module load nvidia/nvhpc-nompi/24.5

./configure CC=gcc CXX=g++ FC=gfortran \
  CFLAGS="-I${PMI2_ROOT}/include" LDFLAGS="-L${PMI2_ROOT/lib" \
  --enable-mpi1-compatibility --enable-mpi-fortran \
  --enable-mpi-interface-warning --enable-mpirun-prefix-by-default \
  --with-slurm --with-knem=/opt/knem-1.1.4.90mlnx2 \
  --with-ucx=${UCX_ROOT} --with-pmi=${PMI2_ROOT} --with-pmi-libdir=${PMI2_ROOT}/lib \
  --with-cuda=${NVHPC_ROOT}/cuda/12.4 \
  --with-libevent=${PRFX}/libevent/2.1.12 \
  --prefix=${PRFX}/openmpi/4.1.6-cuda-12.4

make -j 8
make -j 8 install
```

https://github.com/hpc-uk/build-instructions/blob/main/libs/openmpi/build_openmpi_4.1.6_cirrus_gcc10.md

OSU Micro-Benchmarks 2.7

A small suite of tools for benchmarking specific MPI operations (e.g., point-to-point comms, collectives).

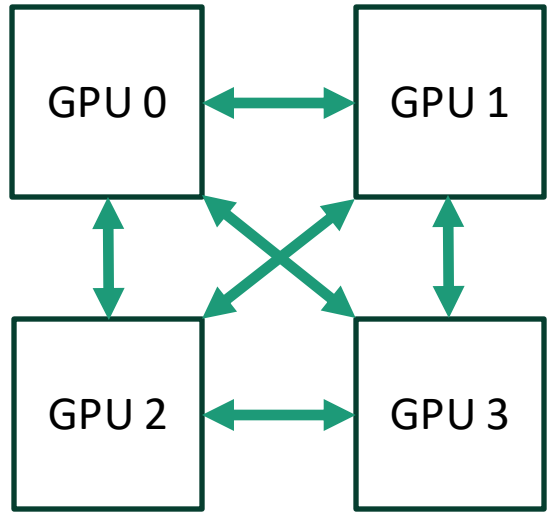
- handles CPU and GPU platforms
- https://mvapich.cse.ohio-state.edu/benchmarks/

Cirrus compilation

```
...

module load nvidia/nvhpc-nompi/24.5
module load openmpi/4.1.6-cuda-12.4

./configure --prefix=${PRFX}/osu/mb/7.2 \
  CC=mpicc CXX=mpicxx \
  --enable-cuda \
  --with-cuda=${NVHPC_ROOT}/cuda/12.4 \
  --with-cuda-include=${NVHPC_ROOT}/cuda/12.4/include/ \
  --with-cuda-libpath=${NVHPC_ROOT}/cuda/12.4/lib64/stubs \
  --with-nccl=${NVHPC_ROOT}/comm_libs/12.4/nccl

make -j 8
make -j 8 install
```

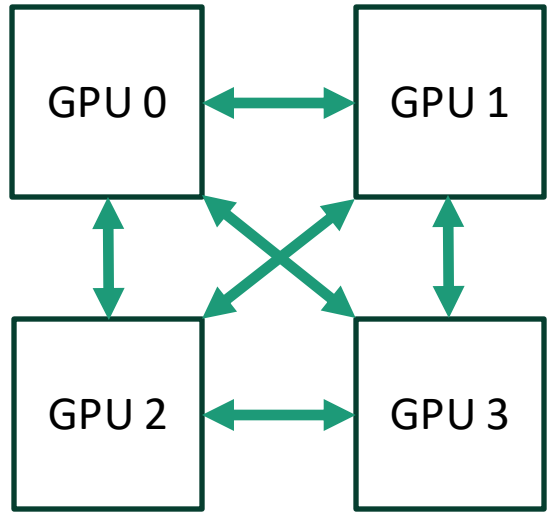
Intra GPU-node P2P comms

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PATH}
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=1 --ntasks=2 --hint=nomultithread"
OSU_BW_PARAMS="-m $((16*1024*1024)):$((16*1024*1024)) D D"

CUDA_VISIBLE_DEVICES=0,1
srun ${SRUN_PARAMS} osu_bw ${OSU_BW_PARAMS}

CUDA_VISIBLE_DEVICES=1,0
srun ${SRUN_PARAMS} osu_bw ${OSU_BW_PARAMS}

...
```



Intra GPU-node P2P comms

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PATH}
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=1 --ntasks=2 --hint=nomultithread"
OSU_BW_PARAMS="-m $((16*1024*1024)):$((16*1024*1024)) D D"

CUDA_VISIBLE_DEVICES=0,1
srun ${SRUN_PARAMS} osu_bw ${OSU_BW_PARAMS}

CUDA_VISIBLE_DEVICES=1,0
srun ${SRUN_PARAMS} osu_bw ${OSU_BW_PARAMS}

...
```



16 MiB message size

48.121 - 48.130 GB/s

Inter GPU-node P2P comms

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PATH}
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=2 --ntasks=2 --tasks-per-node=1 --hint=nomultithread"
OSU_BW_PARAMS="D D"

srun ${SRUN_PARAMS} osu_bw ${OSU_BW_PARAMS}
```

Inter GPU-node P2P comms

```
...

module -s load nvidia/nvhpc-nompi/24
module -s load openmpi/4.1.6-cuda-12

export LD_LIBRARY_PATH=/lib64:${LD_LI
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=2 --ntasks=2 --t
OSU_BW_PARAMS="D D"

srun ${SRUN_PARAMS} osu_bw ${OSU_BW_P
```

Size (Bytes)	Bandwidth (MB/s)
1	0.14
2	0.27
4	0.55
8	1.09
16	2.18
32	4.23
64	8.46
128	16.86
256	33.78
512	67.28
1,024	132.03
2,048	253.41
4,096	479.73
8,192	843.92
16,384	854.45
32,768	880.56
65,536	5,355.35
131,072	7,400.88
262,144	8,942.71
524,288	10,051.51
1,048,576	10,157.78
2,097,152	10,208.57
4,194,304	10,182.65

Intra GPU-node allreduce

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PATH}
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=1 --ntasks=4 --cpus-per-task=10 --hint=nomultithread"
OSU_ALLREDUCE_PARAMS="-m $((1*1024*1024)) -d cuda"

srun ${SRUN_PARAMS} osu_allreduce ${OSU_ALLREDUCE_PARAMS}
```

Intra GPU-node allreduce

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PA
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=1 --ntasks=4 --cpus-per-
OSU_ALLREDUCE_PARAMS="-m $((1*1024*1024)) -d

srun ${SRUN_PARAMS} osu_allreduce ${OSU_ALLRE
```

Size (bytes)	Avg. Latency (us)
1	21.33
2	21.35
4	25.55
8	25.52
16	21.93
32	22.11
64	22.29
128	22.74
256	22.83
512	23.41
1,024	24.51
2,048	26.63
4,096	30.19
8,192	38.82
16,384	53.69
32,768	83.75
65,536	140.32
131,072	247.15
262,144	467.78
524,288	828.20
1,048,576	1,590.98

Inter GPU-node allreduce

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PATH}
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=2 --ntasks=4"
SRUN_PARAMS="${SRUN_PARAMS} --tasks-per-node=2 --cpus-per-task=20"
SRUN_PARAMS="${SRUN_PARAMS} --hint=nomultithread"

OSU_ALLREDUCE_PARAMS="-m $((1*1024*1024)) -d cuda"

srun ${SRUN_PARAMS} osu_allreduce ${OSU_ALLREDUCE_PARAMS}
```

Inter GPU-node allreduce

```
...

module -s load nvidia/nvhpc-nompi/24.5
module -s load openmpi/4.1.6-cuda-12.4

export LD_LIBRARY_PATH=/lib64:${LD_LIBRARY_PA
export OMP_NUM_THREADS=1

SRUN_PARAMS="--nodes=2 --ntasks=4"
SRUN_PARAMS="${SRUN_PARAMS} --tasks-per-node=
SRUN_PARAMS="${SRUN_PARAMS} --hint=nomultithr

OSU_ALLREDUCE_PARAMS="-m $((1*1024*1024)) -d

srun ${SRUN_PARAMS} osu_allreduce ${OSU_ALLRE
```

Size (bytes)	Avg. Latency (us)
1	23.56
2	23.51
4	28.54
8	28.70
16	23.77
32	23.75
64	23.84
128	25.02
256	25.15
512	25.77
1,024	27.10
2,048	29.43
4,096	33.00
8,192	43.59
16,384	58.79
32,768	87.09
65,536	143.79
131,072	253.22
262,144	473.06
524,288	830.02
1,048,576	1,594.32

Intra GPU-node allreduce

Size (bytes)	Avg. Latency (us)
1	21.33
2	21.35
4	25.55
8	25.52
16	21.93
32	22.11
64	22.29
128	22.74
256	22.83
512	23.41
1,024	24.51
2,048	26.63
4,096	30.19
8,192	38.82
16,384	53.69
32,768	83.75
65,536	140.32
131,072	247.15
262,144	467.78
524,288	828.20
1,048,576	1,590.98

Inter GPU-node allreduce

Size (bytes)	Avg. Latency (us)
1	23.56
2	23.51
4	28.54
8	28.70
16	23.77
32	23.75
64	23.84
128	25.02
256	25.15
512	25.77
1,024	27.10
2,048	29.43
4,096	33.00
8,192	43.59
16,384	58.79
32,768	87.09
65,536	143.79
131,072	253.22
262,144	473.06
524,288	830.02
1,048,576	1,594.32