GPU Libraries

Jonathan Halverson
PICSciE and Research Computing

Fall Break Training: Introduction to Parallel Programming
October 20, 2021

Drop-in Interfaces

NumPy

```
import numpy as np
X = np.random.randn(100, 100)
u, s, v = np.linalg.svd(X)
```

CuPy

```
import cupy as cp
X = cp.random.randn(100, 100)
u, s, v = cp.linalg.svd(X)
```

```
#SBATCH --gres=gpu:1
```

CPU and GPU Libraries

CPU

Basic Linear Algebra Subroutines (**BLAS**)

Linear Algebra Package (**LAPACK**)

OpenBLAS

Intel Math Kernel Library (**MKL**)

BLIS/libFLAME

Fastest Fourier Transform in the West (**FFTW**)

GNU Scientific Library (**GSL**)

PETSc

... and many more

GPU

cuBLAS
cuDNN
cuFFT
cuRAND
cuSOLVER
cuSPARSE
...

MAGMA

ESSL

rocBLAS
rocFFT
rocRAND
rocSOLVER
rocSPARSE
...

... and many more

Drop-in Interfaces

BLAS (CPU)

```
subroutine sgemm (
  character TRANSA,
  character TRANSB,
  integer M,
  integer N,
  integer K,
  real ALPHA,
  real, dimension(lda,*) A,
  integer LDA,
  real, dimension(ldb,*) B,
  integer LDB,
  real BETA,
  real, dimension(ldc,*) C,
  integer LDC)
```

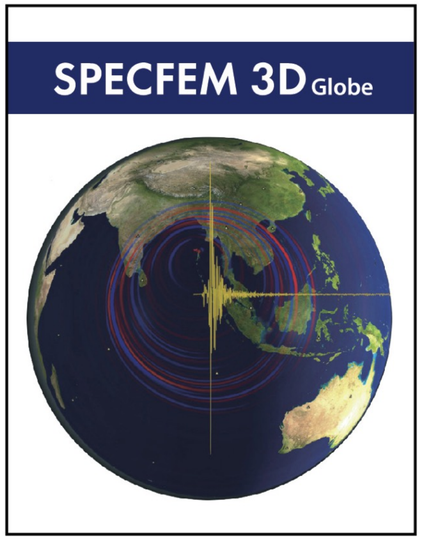
http://www.netlib.org/blas/sgemm.f

cuBLAS (GPU)

```
cublasStatus_t cublasSgemm(cublasHandle_t handle,
                           cublasOperation_t transa,
                           cublasOperation_t transb,
                           int m,
                           int n,
                           int k,
                           const float *alpha,
                           const float *A,
                           int lda,
                           const float *B,
                           int ldb,
                           const float *beta,
                           float *C,
                           int ldc)
```

https://docs.nvidia.com/cuda/cublas/index.html#cublas-lt-t-gt-gemm

SPECFEM 3D Globe




```
$ singularity pull docker://amdih/specfem3d_globe:1ee10977
$ singularity shell --rocm specfem3d_globe_1ee10977.sif
$ find /opt/rocm-4.2.0 -name "*roc*.so"
/opt/rocm-4.2.0/lib/librocalution.so
/opt/rocm-4.2.0/lib/librocalution_hip.so
/opt/rocm-4.2.0/lib/librocblas.so
/opt/rocm-4.2.0/lib/librocfft-device.so
/opt/rocm-4.2.0/lib/librocfft.so
/opt/rocm-4.2.0/lib/librocm-dbgapi.so
/opt/rocm-4.2.0/lib/librocm_smi64.so
/opt/rocm-4.2.0/lib/librocprofiler64.so
/opt/rocm-4.2.0/lib/librocrand.so
/opt/rocm-4.2.0/lib/librocsolver.so
/opt/rocm-4.2.0/lib/librocsparse.so
/opt/rocm-4.2.0/lib/libroctracer64.so
/opt/rocm-4.2.0/lib/libroctx64.so
/opt/rocm-4.2.0/rocalution/lib/librocalution.so
/opt/rocm-4.2.0/rocalution/lib/librocalution_hip.so
/opt/rocm-4.2.0/rocblas/lib/librocblas.so
/opt/rocm-4.2.0/rocfft/lib/librocfft-device.so
/opt/rocm-4.2.0/rocfft/lib/librocfft.so
/opt/rocm-4.2.0/rocm_smi/lib/librocm_smi64.so
/opt/rocm-4.2.0/rocprofiler/lib/librocprofiler64.so
/opt/rocm-4.2.0/rocrand/lib/librocrand.so
/opt/rocm-4.2.0/rocsolver/lib/librocsolver.so
/opt/rocm-4.2.0/rocsparse/lib/librocsparse.so
/opt/rocm-4.2.0/roctracer/lib/libroctracer64.so
/opt/rocm-4.2.0/roctracer/lib/libroctx64.so
```

MAGMA

Matrix Algebra for GPU and Multicore Architectures

MAGMA is a linear algebra library for multicore nodes with GPUs. It can be thought of as an improvement over BLAS/LAPACK for such nodes. MAGMA is capable of using the Tensor Cores of the V100 and A100 GPUs.

BLAS (CPU)

```
subroutine sgemm (
  character TRANSA,
  character TRANSB,
  integer M,
  integer N,
  integer K,
  real ALPHA,
  real, dimension(lda,*) A,
  integer LDA,
  real, dimension(ldb,*) B,
  integer LDB,
  real BETA,
  real, dimension(ldc,*) C,
  integer LDC)
```

cuBLAS (GPU)

```
cublasStatus_t cublasSgemm(cublasHandle_t handle,
                           cublasOperation_t transa,
                           cublasOperation_t transb,
                           int m,
                           int n,
                           int k,
                           const float *alpha,
                           const float *A,
                           int lda,
                           const float *B,
                           int ldb,
                           const float *beta,
                           float *C,
                           int ldc)
```

MAGMA (GPU)

```
void magma_sgemm (
  magma_trans_t transA,
  magma_trans_t transB,
  magma_int_t m,
  magma_int_t n,
  magma_int_t k,
  float alpha,
  magmaFloat_const_ptr dA,
  magma_int_t ldda,
  magmaFloat_const_ptr dB,
  magma_int_t lddb,
  float beta,
  magmaFloat_ptr dC,
  magma_int_t lddc,
  magma_queue_t queue)
```

ESSL

Engineering and Scientific Subroutine Library

ESSL is a numerical library by IBM for linear algebra, eigensystem analysis, Fourier transforms, convolutions and correlations, sorting and searching, interpolation, numerical quadrature and random number generation. With respect to its linear algebra routines, ESSL is not a full implementation of BLAS/LAPACK.

```
$ ssh <YourNetID>@traverse.princeton.edu
$ ls -lL /usr/include/*essl*
-rw-r--r--. 1 bin bin 171727 Feb 24  2018 /usr/include/essl.h
-rw-r--r--. 1 bin bin   4187 Jun  3  2016 /usr/include/essl_lapacke_config.h
-rw-r--r--. 1 bin bin  64882 Jan 16  2018 /usr/include/essl_lapacke.h


$ ls -lL /usr/lib64/*essl*.so
-rw-r--r--. 1 bin bin 45719787 Mar 29  2018 /usr/lib64/libessl6464.so
-rw-r--r--. 1 bin bin 53379191 Mar 29  2018 /usr/lib64/libesslsmp6464.so
-rw-r--r--. 1 bin bin 54737430 Mar 29  2018 /usr/lib64/libesslsmpcuda.so
-rw-r--r--. 1 bin bin 53925425 Mar 29  2018 /usr/lib64/libesslsmp.so
-rw-r--r--. 1 bin bin 46826939 Mar 29  2018 /usr/lib64/libessl.so
```

CUDA Toolkit and other

- **cuBLAS** - GPU-accelerated standard BLAS library
- **cuDNN** - GPU-accelerated library of primitives for deep neural networks
- **cuFFT** - GPU-accelerated library for Fast Fourier Transforms
- **cuRAND** - GPU-accelerated random number generation (RNG)
- **cuSOLVER** - Dense and sparse direct solvers for computer vision, CFD and more
- **cuSPARSE** - GPU-accelerated BLAS for sparse matrices
- **cuTENSOR** - GPU-accelerated tensor linear algebra library
- **NCCL** - Collective Communications Library for scaling apps across multiple GPUs and nodes
- **NPP** - GPU-accelerated image, video, and signal processing functions
- **nvGRAPH** - GPU-accelerated library for graph analytics

CuPy


```
$ module load anaconda3/2020.11
$ conda create --name py-gpu cupy --channel conda-forge

  _libgcc_mutex      conda-forge/linux-64::_libgcc_mutex-0.1-conda_forge
  _openmp_mutex      conda-forge/linux-64::_openmp_mutex-4.5-1_gnu
  ca-certificates    conda-forge/linux-64::ca-certificates-2021.10.8-ha878542_0
  cudatoolkit        conda-forge/linux-64::cudatoolkit-11.4.2-h00f7ccd_9
  cupy               conda-forge/linux-64::cupy-9.5.0-py39h499daff_0
  fastrlock          conda-forge/linux-64::fastrlock-0.6-py39he80948d_1
  ld_impl_linux-64   conda-forge/linux-64::ld_impl_linux-64-2.36.1-hea4e1c9_2
  libblas            conda-forge/linux-64::libblas-3.9.0-12_linux64_openblas
  libcblas           conda-forge/linux-64::libcblas-3.9.0-12_linux64_openblas
  libffi             conda-forge/linux-64::libffi-3.4.2-h9c3ff4c_4
  libgcc-ng          conda-forge/linux-64::libgcc-ng-11.2.0-h1d223b6_11
  libgfortran-ng     conda-forge/linux-64::libgfortran-ng-11.2.0-h69a702a_11
  libgfortran5       conda-forge/linux-64::libgfortran5-11.2.0-h5c6108e_11
  libgomp            conda-forge/linux-64::libgomp-11.2.0-h1d223b6_11
  liblapack          conda-forge/linux-64::liblapack-3.9.0-12_linux64_openblas
  libopenblas        conda-forge/linux-64::libopenblas-0.3.18-pthreads_h8fe5266_0
  ...
```

CuPy


```
$ cd ~/.conda/envs/py-gpu/lib
$ ls -lL libcu*.so
-rwxrwxr-x. 2 jdh4 cses 309216008 Oct  1 14:21 libcublasLt.so
-rwxrwxr-x. 2 jdh4 cses 152771648 Oct  1 14:21 libcublas.so
-rwxrwxr-x. 2 jdh4 cses    674896 Oct  1 14:21 libcudart.so
-rwxrwxr-x. 2 jdh4 cses 361308448 Oct  1 14:21 libcufft.so
-rwxrwxr-x. 2 jdh4 cses    741728 Oct  1 14:21 libcufftw.so
-rwxrwxr-x. 2 jdh4 cses   7317264 Oct  1 14:21 libcupti.so
-rwxrwxr-x. 2 jdh4 cses  83328272 Oct  1 14:21 libcurand.so
-rwxrwxr-x. 2 jdh4 cses 239124360 Oct  1 14:21 libcusolverMg.so
-rwxrwxr-x. 2 jdh4 cses 218939824 Oct  1 14:21 libcusolver.so
-rwxrwxr-x. 2 jdh4 cses 236881384 Oct  1 14:21 libcusparse.so
-rw-rw-r--. 2 jdh4 cses        31 Nov 14  2020 libncurses.so
-rw-rw-r--. 2 jdh4 cses        33 Nov 14  2020 libncursesw.so
...
```

PyTorch

```
$ module load anaconda3/2020.11
$ conda create --name torch-env pytorch torchvision cudatoolkit=11.1 -c pytorch -c nvidia

  _libgcc_mutex      pkgs/main/linux-64::_libgcc_mutex-0.1-main
  _openmp_mutex      pkgs/main/linux-64::_openmp_mutex-4.5-1_gnu
  blas               pkgs/main/linux-64::blas-1.0-mkl
  bzip2              pkgs/main/linux-64::bzip2-1.0.8-h7b6447c_0
  ca-certificates    pkgs/main/linux-64::ca-certificates-2021.9.30-h06a4308_1
  certifi            pkgs/main/linux-64::certifi-2021.10.8-py39h06a4308_0
  cudatoolkit        nvidia/linux-64::cudatoolkit-11.1.74-h6bb024c_0
  ffmpeg             pytorch/linux-64::ffmpeg-4.3-hf484d3e_0
  freetype           pkgs/main/linux-64::freetype-2.10.4-h5ab3b9f_0
  giflib             pkgs/main/linux-64::giflib-5.2.1-h7b6447c_0
  gmp                pkgs/main/linux-64::gmp-6.2.1-h2531618_2
  gnutls             pkgs/main/linux-64::gnutls-3.6.15-he1e5248_0
  intel-openmp       pkgs/main/linux-64::intel-openmp-2021.3.0-h06a4308_3350
  jpeg               pkgs/main/linux-64::jpeg-9b-h024ee3a_2
  lame               pkgs/main/linux-64::lame-3.100-h7b6447c_0
  lcms2              pkgs/main/linux-64::lcms2-2.12-h3be6417_0
  libpng             pkgs/main/linux-64::libpng-1.6.37-hbc83047_0
  ...
```

GROMACS

```
$ module load cudatoolkit/11.2
$ module load openmpi/gcc/4.0.4/64
$ cmake3 .. -DCMAKE_BUILD_TYPE=Release -DCMAKE_C_COMPILER=gcc -DCMAKE_CXX_COMPILER=g++ ...
$ make && make install
$ ldd ~/.local/bin/gmx
  linux-vdso64.so.1 (0x0000200000060000)
  libgromacs.so.6 => /home/jdh4/.local/bin/./../lib64/libgromacs.so.6 (0x0000200000080000)
  libgomp.so.1 => /lib64/libgomp.so.1 (0x00002000016b0000)
  libpthread.so.0 => /lib64/power9/libpthread.so.0 (0x0000200001720000)
  libstdc++.so.6 => /lib64/libstdc++.so.6 (0x0000200001770000)
  libm.so.6 => /lib64/power9/libm.so.6 (0x00002000019a0000)
  libgcc_s.so.1 => /lib64/libgcc_s.so.1 (0x0000200001ad0000)
  libc.so.6 => /lib64/power9/libc.so.6 (0x0000200001b10000)
  libdl.so.2 => /lib64/libdl.so.2 (0x0000200001d20000)
  librt.so.1 => /lib64/power9/librt.so.1 (0x0000200001d50000)
  libcufft.so.10 => /usr/local/cuda/targets/ppc64le-linux/lib/libcufft.so.10 (0x0000200001d80000)
  libfftw3f.so.3 => /home/jdh4/.local/lib/libfftw3f.so.3 (0x0000200017930000)
  libopenblas.so.0 => /lib64/libopenblas.so.0 (0x0000200017ab0000)
  /lib64/ld64.so.2 (0x0000200000000000)
  libgfortran.so.5 => /lib64/libgfortran.so.5 (0x00002000187d0000)
  libquadmath.so.0 => /lib64/libquadmath.so.0 (0x00002000189c0000)
  libz.so.1 => /lib64/libz.so.1 (0x0000200018a30000)
```

Hands-on Exercise 1

Run the commands below to learn about the CUDA Toolkit:

```
$ ssh <YourNetID>@adroit.princeton.edu
$ module avail cudatoolkit
$ module show cudatoolkit/11.4
$ ls -lL /usr/local/cuda-11.4/include
$ cat /usr/local/cuda-11.4/include/cusolverDn.h | less  # q to quit
$ ls -lL /usr/local/cuda-11.4/lib64/lib*.so
$ ls -lL /usr/local/cuda-11.4/bin
$ nvcc --help
-bash: nvcc: command not found
$ module load cudatoolkit/11.4
$ nvcc --help | less  # q to quit
```

Hands-on Exercise 2

```
$ ssh <YourNetID>@adroit.princeton.edu
$ git clone https://github.com/PrincetonUniversity/gpu_programming_intro.git
```

OR

```
$ cp -r /scratch/network/jdh4/gpu_programming_intro .  # live workshop only
```

THEN

```
$ cd gpu_programming_intro/06_gpu_libraries/hello_world_gpu_library
# work through the material the README file
```

Hands-on Exercise 3

```
$ cd gpu_programming_intro/06_gpu_libraries
# work through the material the README file
```