NVBLAS LIBRARY

DU-06702-001_v11.0 | August 2020

User Guide

Chapter 1. INTRODUCTION

The NVBLAS Library is a GPU-accelerated Libary that implements BLAS (Basic Linear Algebra Subprograms). It can accelerate most BLAS Level-3 routines by dynamically routing BLAS calls to one or more NVIDIA GPUs present in the system, when the charateristics of the call make it to speedup on a GPU.

Chapter 2. OVERVIEW

The NVBLAS Library is built on top of the cuBLAS Library using only the CUBLASXT API (See the CUBLASXT API section of the cuBLAS Documentation for more details). NVBLAS also requires the presence of a CPU BLAS lirbary on the system. Currently NVBLAS intercepts only compute intensive BLAS Level-3 calls (see table below). Depending on the charateristics of those BLAS calls, NVBLAS will redirect the calls to the GPUs present in the system or to CPU. That decision is based on a simple heuristic that estimates if the BLAS call will execute for long enough to amortize the PCI transfers of the input and output data to the GPU. **Because NVBLAS does not support all standard BLAS routines, it might be necessary to associate it with an existing full BLAS Library. Please refer to the Usage section for more details.**

Chapter 3.
GPU ACCELERATED ROUTINES

NVBLAS offloads only the compute-intensive BLAS3 routines which have the best potential for acceleration on GPUs.

The current supported routines are in the table below :

Routine	Types	Operation
gemm	S,D,C,Z	multiplication of 2 matrices.
syrk	S,D,C,Z	symmetric rank-k update
herk	C,Z	hermitian rank-k update
syr2k	S,D,C,Z	symmetric rank-2k update
her2k	C,Z	hermitian rank-2k update
trsm	S,D,C,Z	triangular solve with multiple right-hand sides
trmm	S,D,C,Z	triangular matrix-matrix multiplication
symm	S,D,C,Z	symmetric matrix-matrix multiplication
hemm	C,Z	hermitian matrix-matrix multiplication

Chapter 4.
BLAS SYMBOLS INTERCEPTION

Standard BLAS Library implementations usually expose multiple symbols for the same routines. Let say **`func`** is a BLAS routine name, **`func_`** or/and **`func`** are usually defined as extern symbols. Some BLAS Libraries might also expose some symbols with a proprietary appended prefix. NVBLAS intercepts only the symbols **`func_`** and **`func`**. The user needs to make sure that the application intended to be GPU-accelerated by NVBLAS actually calls those defined symbols. Any other symbols will not be intercepted and the original BLAS routine will be executed for those cases.

Chapter 5.
DEVICE MEMORY SUPPORT

Starting with Release 8.0, data can be located on any GPU device, even on GPU devices that are not configured to be part of the computation. When any of the data is located on a GPU, the computation will be exclusively done on GPU whatever the size of the problem. Also, this feature has to be used with caution : the user has to be sure that the BLAS call will be indeed intercepted by NVBLAS, otherwise it will result on a crash when the CPU Blas tries to execute it.

Chapter 6. SECURITY PRECAUTION

Because the NVBLAS Library relies on a symbols interception mechanism, it is essential to make sure it has not been compromised. In that regard, NVBLAS should never be used from a process running at elevated privileges, such as Administrator on Windows or root on Linux.

Chapter 7. CONFIGURATION

Because NVBLAS is a drop-in replacement of BLAS, it must be configured through an ASCII text file that describes how many and which GPUs can participate in the intercepted BLAS calls. The configuration file is parsed at the time of the loading of the library. The format of the configuration file is based on keywords optionally followed by one or more user-defined parameters. At most one keyword per line is allowed. Blank lines or lines started by the character **#** are ignored.

7.1. NVBLAS_CONFIG_FILE environment variable

The location and name of the configuration file must be defined by the environment variable **`NVBLAS_CONFIG_FILE`**. By default, if **`NVBLAS_CONFIG_FILE`** is not defined, NVBLAS will try to open the file **`nvblas.conf`** in the current directory. For a safe use of NVBLAS, the configuration file should have have restricted write permissions.

7.2. Configuration keywords

The configuration keywords syntax is described in the following sub-sections.

7.2.1. NVBLAS_LOGFILE

This keyword defines the file where NVBLAS should print status and error messages. By default, if not defined, the standard error output file (e.g stderr) will be used. It is advised to define this keyword early in the configuration to capture errors in parsing that file itself.

7.2.2. NVBLAS_TRACE_LOG_ENABLED

When this keyword is defined, every intercepted BLAS calls will be logged into the NVBLAS_LOGFILE. This feature, even though intrusive, can be useful for debugging purpose.

7.2.3. NVBLAS_CPU_BLAS_LIB

This keyword defines the CPU BLAS dynamic library file (e.g .so file on Linux or .dll on Windows) that NVBLAS should open to find the CPU BLAS symbols definitions. This keyword must be defined for NVBLAS to work. Because CPU Blas libraries are often composed of multiple files, even though this keyword is set to the full path to the main file of the CPU library, it might still be necessary to define the right path to find the rest of the library files in the environment of your system. On Linux, this can be done by setting the environment variable LD_LIBRARY_PATH whereas on Windows, this can be done by setting the environment variable PATH.

For a safe use of NVBLAS, the following precautions are strongly advised:

- the CPU BLAS Library should be located where ordinary users do not have write permissions.
- the path specified should be absolute, not relative.

7.2.4. NVBLAS_GPU_LIST

This keyword defines the list of GPUs that should participate in the computation of the intercepted BLAS calls. If not defined, only GPU device 0 is used, since that is normally the most compute-capable GPU installed in the system. This keyword can be set to a list of device numbers separated by blank characters. Also the following wildcard keywords are also accepted for simplicity :

Keyword	Meaning
ALL	All compute-capable GPUs detected on the system will be used by NVBLAS
ALL0	GPU device 0, AND all others GPUs detected that have the same compute-capabilities as device 0 will be used by NVBLAS

Note : In the current release of CUBLAS, the CUBLASXT API supports two GPUs if they are on the same board such as Tesla K10 or GeForce GTX690 and one GPU otherwise. Because NVBLAS is built on top of the CUBLASXT API, NVBLAS has the same restriction. If access to more GPUs devices is needed, details of the licensing are described at cublasXt.

7.2.5. NVBLAS_TILE_DIM

This keyword defines the tile dimension that should be used to divide the matrices involved in the computation. This definition maps directly to a call of the cublasXt API routine **`cublasXtSetBlockDim`**. Refer to cuBLAS documentation to understand the tradeoffs associated with setting this to a larger or a smaller value.

7.2.6. NVBLAS_GPU_DISABLED_<BLAS_FUNC_NAME>

This keyword, appended with the name of a BLAS routine disables NVBLAS from running a specified routine on the GPU. This feature is intended mainly for debugging purposes. by default, all supported BLAS routines are enabled.

7.2.7. NVBLAS_CPU_RATIO_<BLAS_FUNC_NAME>

This keyword, appended with the name of ta BLAS routine defines the ratio of the workload that should remain on the CPU in the event that the NVBLAS decides to offload work for that routine on the GPU. This functionality is directly mapped to the cublasXt API routine **`cublasXtSetCpuRatio`**. By default, the ratio is defined to zero for all routines. Please refer to the cuBLAS Documentation for details and for the list of routines which support this feature.

7.2.8. NVBLAS_AUTOPIN_MEM_ENABLED

This keyword enables the Pinning Memory mode. This functionality is directly mapped to the cublasXt API routine **`cublasXtSetPinningMemMode`**. If this keyowrd is not present in the configuration file, the Pinning Memory mode will be set to CUBLASXT_PINNING_DISABLED.

There are some restrictions to use this feature as specified in the cuBLAS documentation of the underlying routine **`cublasXtSetPinningMemMode`**. Specifically when NVBLAS is used in a multi-threaded applications, this option should not be used if there is a chance that matrices used by different threads overlaps while calling NVBLAS. Please refer to the cuBLAS Documentation of the routine **`cublasXtSetPinningMemMode`** for details.

7.2.9. Config file Example

The example below shows a typical NVBLAS configuration file :

```
# This is the configuration file to use NVBLAS Library
# Setup the environment variable NVBLAS_CONFIG_FILE to specify your own config
 file.
# By default, if NVBLAS_CONFIG_FILE is not defined,
# NVBLAS Library will try to open the file "nvblas.conf" in its current
 directory
# Example : NVBLAS_CONFIG_FILE  /home/cuda_user/my_nvblas.conf
# The config file should have restricted write permissions accesses

# Specify which output log file (default is stderr)
NVBLAS_LOGFILE  nvblas.log

# Enable trace log of every intercepted BLAS calls
NVBLAS_TRACE_LOG_ENABLED

#Put here the CPU BLAS fallback Library of your choice
#It is strongly advised to use full path to describe the location of the CPU
 Library
NVBLAS_CPU_BLAS_LIB  /usr/lib/libopenblas.so
#NVBLAS_CPU_BLAS_LIB  <mkl_path_installtion>/libmkl_rt.so

# List of GPU devices Id to participate to the computation
# Use ALL if you want all your GPUs to contribute
# Use ALL0, if you want all your GPUs of the same type as device 0 to contribute
# However, NVBLAS consider that all GPU have the same performance and PCI
 bandwidth
# By default if no GPU are listed, only device 0 will be used

#NVBLAS_GPU_LIST 0 2 4
#NVBLAS_GPU_LIST ALL
NVBLAS_GPU_LIST ALL0

# Tile Dimension
NVBLAS_TILE_DIM 2048

# Autopin Memory
NVBLAS_AUTOPIN_MEM_ENABLED

#List of BLAS routines that are prevented from running on GPU (use for debugging
 purpose
# The current list of BLAS routines supported by NVBLAS are
# GEMM, SYRK, HERK, TRSM, TRMM, SYMM, HEMM, SYR2K, HER2K

#NVBLAS_GPU_DISABLED_SGEMM
#NVBLAS_GPU_DISABLED_DGEMM
#NVBLAS_GPU_DISABLED_CGEMM
#NVBLAS_GPU_DISABLED_ZGEMM

# Computation can be optionally hybridized between CPU and GPU
# By default, GPU-supported BLAS routines are ran fully on GPU
# The option NVBLAS_CPU_RATIO_<BLAS_ROUTINE> give the ratio [0,1]
# of the amount of computation that should be done on CPU
# CAUTION : this option should be used wisely because it can actually
# significantly reduced the overall performance if too much work is given to CPU

#NVBLAS_CPU_RATIO_CGEMM 0.07
```

Chapter 8. INSTALLATION

The NVBLAS Library is part of the CUDA Toolkit, and will be installed along all the other CUDA libraries. It is available on 64-bit operating systems. NVBLAS Library is built on top of cuBLAS, so the cuBLAS library needs to be accessible by NVBLAS.

Chapter 9. USAGE

To use the NVBLAS Library, the user application must be relinked against NVBLAS in addition to the original CPU Blas (technically only NVBLAS is needed unless some BLAS routines not supported by NVBLAS are used by the application). To be sure that the linker links against the exposed symbols of NVBLAS and not the ones from the CPU Blas, the NVBLAS Library needs to be put before the CPU Blas on the linkage command line.

On Linux, an alternative way to use NVBLAS Library is to use the LD_PRELOAD environment variable; this technique has the advantage of avoiding the relinkage step. However, the user should avoid defining that environment variable globally because it will cause the NVBLAS library to be loaded by every shell command executed on the system, thus leading to a lack of responsiveness of the system.

Finally mathematical tools and libraries often offer the opportunity to specify the BLAS Library to be used through an environment variable or a configuration file. Because NVBLAS does not support all the standard BLAS routines, it might be necessary to pair NVBLAS with a full BLAS library, even though your application only calls supported NVBLAS routines. Fortunately, those tools and libraries usually offer a way to specify multiple BLAS Libraries. Please refer to the documentation of the appropriate tools and libraries for details.

Notice

Trademarks

Copyright