

CentraleSupélec  

Mineure HPC-SBD

Bibliothèque CUBLAS

Stéphane Vialle

 université Sciences et technologies de l'information et de la communication (STIC)  

Stephane.Vialle@centralesupelec.fr
<http://www.metz.supelec.fr/~vialle>

CentraleSupélec

Bibliothèques BLAS et CUBLAS

- 1 – Les bibliothèques BLAS
- 2 – CUBLAS vs BLAS
- 3 – La fonction *cublasDgemv*
- 4 – Format des données de *cublasDgemv*
- 5 – Espace de stockage pour *cublasDgemv*

CentraleSupélec

Bibliothèques BLAS et CUBLAS

Les bibliothèques « BLAS »

BLAS : Basic Linear Algebra Subprograms

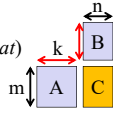
- Ensemble de fonctions d'algèbre linéaire
- Prototypes publiés en 1979
https://fr.wikipedia.org/wiki/Basic_Linear_Algebra_Subprograms
- 3 niveaux de BLAS :
 - niveau 1 : opérations sur les vecteurs
 - niveau 2 : opérations de type matrice – vecteur
 - niveau 3 : opérations de type matrice – matrice
- **UNE API standardisée et DES implantations**
 - implantations open-sources
 ex : OpenBLAS, ATLAS (très efficaces)
 - implantations propriétaires ciblées sur un type de matériel
 ex : bibliothèque MKL d'Intel
 - **installer en recompilant sur la machine cible**

Bibliothèques BLAS et CUBLAS

Les bibliothèques « BLAS »

La Fonction `cblas_dgemm` :

- Fonction de produit de matrices denses
- Travaille sur des *double* (`cblas_sgemm` pour les *float*)

$$C = \alpha.op(A) \times op(B) + \beta.C$$


```
void cblas_dgemm (
  const CBLAS_LAYOUT Layout,
  const CBLAS_TRANSPOSE transa,
  const CBLAS_TRANSPOSE transb,
  const int m, const int n, const int k,
  const double alpha,
  const double *a, const int lda,
  const double *b, const int ldb,
  const double beta,
  double *c, const int ldc)

```

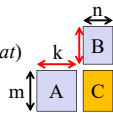
Stockage en row major ou column major

Bibliothèques BLAS et CUBLAS

Les bibliothèques « BLAS »

La Fonction `cblas_dgemm` :

- Fonction de produit de matrices denses
- Travaille sur des *double* (`cblas_sgemm` pour les *float*)

$$C = \alpha.op(A) \times op(B) + \beta.C$$


```
void cblas_dgemm (
  const CBLAS_LAYOUT Layout,
  const CBLAS_TRANSPOSE transa,
  const CBLAS_TRANSPOSE transb,
  const int m, const int n, const int k,
  const double alpha,
  const double *a, const int lda,
  const double *b, const int ldb,
  const double beta,
  double *c, const int ldc)

```

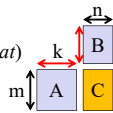
Transposition « au vol » (ou pas) de A et B

Bibliothèques BLAS et CUBLAS

Les bibliothèques « BLAS »

La Fonction `cblas_dgemm` :

- Fonction de produit de matrices denses
- Travaille sur des *double* (`cblas_sgemm` pour les *float*)

$$C = \alpha.op(A) \times op(B) + \beta.C$$


```
void cblas_dgemm (
  const CBLAS_LAYOUT Layout,
  const CBLAS_TRANSPOSE transa,
  const CBLAS_TRANSPOSE transb,
  const int m, const int n, const int k,
  const double alpha,
  const double *a, const int lda,
  const double *b, const int ldb,
  const double beta,
  double *c, const int ldc)

```

Tailles des matrices

Bibliothèques BLAS et CUBLAS

Les bibliothèques « BLAS »

La Fonction `cblas_dgemm` :

- Fonction de produit de matrices denses
- Travaille sur des *double* (`cblas_sgemv` pour les *float*)

$$C = \alpha.op(A) \times op(B) + \beta.C$$

```

void cblas_dgemm (
  const CBLAS_LAYOUT Layout,
  const CBLAS_TRANSPOSE transa,
  const CBLAS_TRANSPOSE transb,
  const int m, const int n, const int k,
  const double alpha,
  const double *a, const int lda,
  const double *b, const int ldb,
  const double beta,
  double *c, const int ldc)

```

lda, ldb, ldc : tailles des zones de stockage (voir plus loin)

Bibliothèques BLAS et CUBLAS

2 – CUBLAS vs BLAS

- Mêmes fonctionnalités d'algèbre linéaire que les BLAS
- Mais quelques différences :
 - Besoin d'initialiser l'usage de la bibliothèque**
 - Récupération d'un « handle » sur la bibliothèque
 - Passage de ce *handle* en paramètre de toutes les fonctions CUBLAS
 - Le format « column major » (style FORTRAN) est imposé**
 - Possibilité de convertir les données d'entrée au vol
 - Besoin de convertir les données de sortie
 - La bibliothèque ne se charge pas des transferts CPU/GPU**
 - les transferts restent à la charge du développeur
 - Quelques nouvelles fonctions disponibles**
 - des « extensions des BLAS » : non standard mais pratiques !

Bibliothèques BLAS et CUBLAS

2 – CUBLAS vs BLAS

Utilisation des CUBLAS

- Ajout de `#include <cublas_v2.h>` au début du fichier src Cuda
- Ajout de `-lcublas` à l'édition de lien
- Initialisation des CUBLAS dans le code Cuda


```

cublasHandle_t handle;
cublasCreate(&handle); // Return a cudaStatus_t
                        // CUBLAS_STATUS_SUCCESS ?
cublasXxxx(handle,...); // Cublas usage
cublasDestroy(handle); // End of cublas usage

```

Rmq : L'initialisation (`cublasCreate`) peut prendre un peu de temps!
Ne la faire qu'une seule fois au début du pgm.

Bibliothèques BLAS et CUBLAS

3 – La fonction *cublasDgemm*

API :

$$C = \alpha \cdot op(A) \times op(B) + \beta \cdot C$$

```

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

```

Comparé aux BLAS :

- On passe le « *handle* » des cublas (1^{er} paramètre)
- Les valeurs « *alpha* » et « *beta* » ne sont pas des constantes mais des adresse de variables (constantes)
- On ne précise pas le format des données :
→ le stockage « *column major* » est imposé

Bibliothèques BLAS et CUBLAS

3 – La fonction *cublasSgemm*

API :

$$C = \alpha \cdot op(A) \times op(B) + \beta \cdot C$$

```

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)

```

Comparé aux BLAS :

- On passe le « *handle* » des cublas (1^{er} paramètre)
- Les valeurs « *alpha* » et « *beta* » ne sont pas des constantes mais des adresse de variables (constantes)
- On ne précise pas le format des données :
→ le stockage « *column major* » est imposé

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Format des données (matrices) :

$$C = \alpha \cdot op(A) \times op(B) + \beta \cdot C$$

```

cublasDgemm(
cublasHandle_t handle,
cublasOperation_t transa,
cublasOperation_t transb,
int m, int n, int k,.....)

```

A est une matrice :

- mathématiquement de *m* lignes × *k* colonnes,
- considérée stockée au format *column major* (imposé),
- que l'on peut transposer « au vol » :

$$op(A) = \begin{cases} A & \text{if transa} == \text{CUBLAS_OP_N} \\ A^T & \text{if transa} == \text{CUBLAS_OP_T} \\ A^H & \text{if transa} == \text{CUBLAS_OP_C} \end{cases} \rightarrow \text{matrice conjuguée}$$

Idem pour B : matrice de *k* lignes × *n* colonnes

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemv*

Format des données (matrices) :

$C = \alpha.op(A) \times op(B) + \beta.C$

```
cublasDgemv(
  cublasHandle_t handle,
  cublasOperation_t transa,
  cublasOperation_t transb,
  int m, int n, int k,.....)
```

A est une matrice :

- mathématiquement de *m lignes* × *k colonnes*,
- **considérée stockée au format *column major*** (imposé),
- que l'on peut transposer « au vol » :

En langage C/C++ (comme en CUDA) un tableau 2D est stocké en *row major*...

...en demandant la transposition au vol d'une matrice (avec `transa = CUBLAS_OP_T`) on l'obtient en *column major* !

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemv*

Format des données (matrices) :

$C = \alpha.op(A) \times op(B) + \beta.C$

```
cublasDgemv(
  cublasHandle_t handle,
  cublasOperation_t transa,
  cublasOperation_t transb,
  int m, int n, int k,.....)
```

Mais la matrice résultat (C) :

- est générée au format *column major* (imposé)
- donc on obtient C^T au format naturel du langage C/C++/CUDA

Pour obtenir C au format du C/C++/CUDA on peut :

- Transposer C avec un kernel écrit pour l'occasion
- Transposer C avec la fonction CUBLAS `cudaDgemv` qui est une extension des BLAS
- Organiser les calculs matriciels pour calculer mathématiquement C^T et donc obtenir C dans le format du C/C++/CUDA

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemv*

Kernel de transposition simpliste :

```

__global__ void TransposeKernel_v0(float *MT, float *M,
                                  int mLig, int nCol)
{
  int lig = threadIdx.y + blockIdx.y*B_SIZE_XY_KT0;
  int col = threadIdx.x + blockIdx.x*B_SIZE_XY_KT0;
  if (lig < mLig && col < nCol)
    MT[col*mLig + lig] = M[lig*nCol + col];
}
```

Accès NON coalescent

Accès coalescent

Solution : passer par la shared memory pour être coalescent à la lecture ET à l'écriture...

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Kernel de transposition coalescent :

Diagram illustrating a coalescent transposition kernel. It shows a memory layout M with a 1 warp access pattern. A shaded region of size shm is shown in a matrix of size m by n . An arrow points to the resulting matrix of size n by m .

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Kernel de transposition coalescent :

Diagram illustrating a coalescent transposition kernel. It shows a memory layout M and M^T with a 1 warp access pattern. A shaded region of size shm is shown in a matrix of size m by n . An arrow points to the resulting matrix of size n by m .

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Kernel de transposition coalescent :

Diagram illustrating a coalescent transposition kernel. It shows a memory layout M and M^T with a 1 warp access pattern. A shaded region of size shm is shown in a matrix of size m by n . A red arrow points to a region where "Ce thread ne travaille pas" (this thread does not work) and a green arrow points to a region where "Ce thread travaille" (this thread works).

→ Deux conditions différentes par thread
 → Deux « boundaries » par threads

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Kernel de transposition coalescent :

```

__global__ void TransposeKernel_v1(float *MT, float *M,
                                   int mLig, int nCol)
{
    int firstLibBlock = blockIdx.y*BSIZE_XY_KT1;
    int firstColBlock = blockIdx.x*BSIZE_XY_KT1;
    int lig = firstLibBlock + threadIdx.y;
    int col = firstColBlock + threadIdx.x;
    int ligT = firstColBlock + threadIdx.y;
    int colT = firstLibBlock + threadIdx.x;
    __shared__ float shM[BSIZE_XY_KT1][BSIZE_XY_KT1];
    if (lig < mLig && col < nCol) // Condition 1
        shM[threadIdx.y][threadIdx.x] = M[lig*nCol + col];
    __syncthreads();
    if (ligT < nCol && colT < mLig) // Condition 2
        MT[ligT*mLig + colT] = shM[threadIdx.x][threadIdx.y];
}

```

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Transposition par *cublasDgeam* :

Réalise le calcul de : $C = \alpha.op(A) + \beta.op(B)$

```

cublasStatus_t cublasDgeam(
    cublasHandle_t handle,
    cublasOperation_t transa,
    cublasOperation_t transb,
    int m, int n,
    const double *alpha,
    const double *A, int lda,
    const double *beta,
    const double *B, int ldb,
    double *C, int ldc)

```

Voir aussi `cublasSgeam(...)`

→ Permet le calcul (très rapide) d'une transposée de matrice

Rmq : on peut passer un ptr NULL si on ne sert pas d'une matrice et qu'on ne fait pas d'opération dessus TP 4

Bibliothèques BLAS et CUBLAS

4 – Format des données de *cublasDgemm*

Calcul de M^T en stockage *column major* :

Calculer M^T au lieu de M avec les CUBLAS

- Donc obtenir M^T stockée en *column major*
- Donc obtenir M dans un tableau C/C++/CUDA en *row major*

Rappel : $(A \times B)^T = B^T \times A^T$

.....

TP 4

Bibliothèques BLAS et CUBLAS

5 – Espace de stockage pour *cublasDgemv*

Autres paramètres (doc NVIDIA)

Param.	Memory	In/out	Meaning
handle		input	handle to the CUBLAS library context.
transa		input	operation op(A) that is non- or (conj.) transpose.
transb		input	operation op(B) that is non- or (conj.) transpose.
m		input	number of rows of matrix op(A) and c.
n		input	number of columns of matrix op(B) and c.
k		input	number of columns of op(A) and rows of op(B).
alpha	host or device	input	<type>- scalar used for multiplication.
A	device	input	<type>- array of dimensions lda \times k with lda \geq max(1,m) if transa == CUBLAS_OP_N and lda \times n with lda \geq max(1,k) otherwise.
lda		input	leading dimension of two-dimensional array used to store the matrix A.
B	device	input	<type>- array of dimension ldb \times n with ldb \geq max(1,k) if transb == CUBLAS_OP_N and ldb \times n with ldb \geq max(1,n) otherwise.
ldb		input	leading dimension of two-dimensional array used to store matrix B.
beta	host or device	input	<type>- scalar used for multiplication. If beta=0, C does not have to be a valid input.
C	device	in/out	<type>- array of dimensions ldc \times n with ldc \geq max(1,m).
ldc		input	leading dimension of a two-dimensional array used to store the matrix C.

Les tableaux de stockages de A, B et C, peuvent être plus grands que les matrices

On doit préciser leur dimension principale...

Bibliothèques BLAS et CUBLAS

5 – Espace de stockage pour *cublasDgemv*

Les espaces de stockage peuvent être plus grand que les matrices :

- Réutilisation d'un espace alloué précédemment
- Espace de stockage dimensionné au maximum
- Calcul sur des sous-matrices
- ...

La fonction `cblas_dgemv` doit connaître :

- le nombre de colonnes de la matrice
- le nombre de colonnes de l'espace de stockage

→ Pour savoir combien d'espace sauter entre deux lignes

Ex : en *row major* sans transposer A et B :

il faut sauter $lda - k$, avec : $lda \geq k$, $ldb \geq n$, $ldc \geq n$

Bibliothèques BLAS et CUBLAS

5 – Espace de stockage pour *cublasDgemv*

Les espaces de stockage peuvent être plus grand que les matrices :

- Réutilisation d'un espace alloué précédemment
- Espace de stockage dimensionné au maximum
- Calcul sur des sous-matrices
- ...

La fonction `cublasDgemv` impose un stockage en *column major* :

- Sans transposer A et B il faut : $lda \geq m$, $ldb \geq k$, $ldc \geq m$
- En transposant A il faut : $lda \geq k$
- En transposant B il faut : $ldb \geq n$

Bibliothèques BLAS et CUBLAS

5 – Espace de stockage pour *cublasDgemv*

Spécification de lda et ldb

A	device	input	<type> array of dimensions $lda \times k$ with $lda \geq \max(1, m)$ if $transa == CUBLAS_OP_N$ and $lda \times m$ with $lda \geq \max(1, k)$ otherwise.
lda		input	leading dimension of two-dimensional array used to store the matrix A.
B	device	input	<type> array of dimension $ldb \times n$ with $ldb \geq \max(1, k)$ if $transa == CUBLAS_OP_N$ and $ldb \times k$ with $ldb \geq \max(1, n)$ otherwise.
ldb		input	leading dimension of two-dimensional array used to store matrix B.

- Un tableau A d'un espace total de $m \times k$ éléments suffira
- Si on n'a pas transposé A on indiquera $lda = m$ (ou plus), Sinon on indiquera $lda = k$ (ou plus)
- Un tableau B d'un espace total de $k \times n$ éléments suffira
- Si on n'a pas transposé B on indiquera $ldb = k$ (ou plus), Sinon on indiquera $ldb = n$ (ou plus)

Bibliothèques BLAS et CUBLAS

5 – Espace de stockage pour *cublasDgemv*

Spécification de ldc

C	device	in/out	<type> array of dimensions $ldc \times n$ with $ldc \geq \max(1, m)$.
ldc		input	leading dimension of a two-dimensional array used to store the matrix c.

- Un tableau C d'un espace total de $m \times n$ double suffira
- C sera (forcément) en *column major*
- On indiquera $ldc = m$ (ou plus)

Bibliothèques BLAS et CUBLAS
