CUBLAS

Linear algebra using CUDA

presentation : https://developer.nvidia.com/cuBLAS
toolkit documentation : http://docs.nvidia.com/cuda/cublas/#axzz3YKIBNyuA

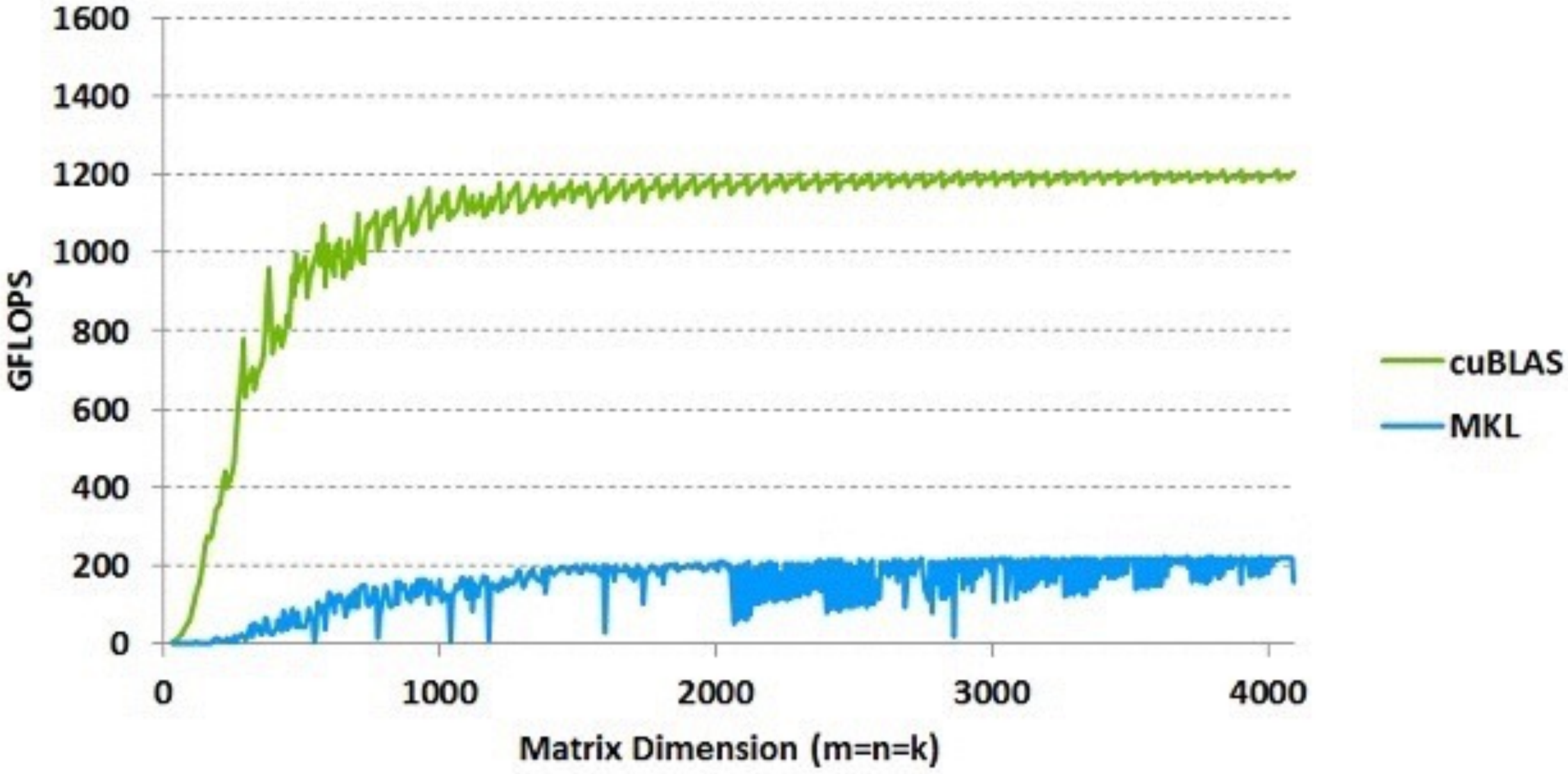
BLAS

- Level-1 operations : scalars and vectors.
    - Ex : sums, find min and max, copies, scales, swaps, axpy, dot product, norm, applies Givens rotation matrix
- Level-2 operations : vectors and matrices
    - Products Ax+y (gemv), rank updates,
    - Works with triangular matrices, packed, banded, hermitians…
- Level-3 operations : matrices and matrices
    - Matrices products, linear triangular system solving (left or right), linear equations systems (batch) solving, transpositions, LU decomp., conversion between packed/unpacked
    - Declinations with triangular matrices, symmetrics…
- Operations generally exist for complex, doubles…

CUBLAS

- ~152 functions

~Benchmark against MKL



ZGEMM, cuBLAS 6.5 on K40m, ECC ON, input and output data on device. MKL 11.0.4 on Intel IvyBridge single socket 12 -core E5-2697 v2 @ 2.70GHz

Basic use

1.Create :
  cublasHandle_t handle; (one per CPU thread AND device !)

2.cublasCreate(&handle);

3.Copy on device or use cublasSetVector or cublasSetMatrix

4.Use a function on allocated memory

5.Retrieve result with cublasGet{Vector,Matrix}

6.Clean up :
  cublasDestroy(handle)

- Apart from math functions :
  - cublasSetVector(int n, int elemSize, const void *x, int incx, void *y, int incy) : copies from CPU to GPU
    - also : cublasGetVector, setmatrix, getmatrix…
- Errors are returned as cublasStatus_t
- WARNING : expecting column-major format ! (c and c++ use row-major)
  Number of **rows** is the leading dimension

TP

Matrix-vector multiplication

- Compute the matrix vector product A x = y, where y is a vector of size N with all elements equal to 1, and A a MxN matrix with all elements of the i-th row equal to i, for i between 1 and M.