Harness the Power of GPUs: An Introduction to GPGPU Programming
Lab 4: Matrix-Matrix Multiplication

Guido Juckeland
Visiting Scholar
Technische Universität Dresden, Germany

June 17, 2014
Task 1: Matrix-matrix multiplication

- Write a program that allocates two matrices \texttt{a\_host} and \texttt{b\_host} of type float with 1,024x1,024 elements. Calculate the matrix-matrix multiplication of both matrices on the GPU. Copy the result matrix \texttt{c\_device} back to the host and check for correctness.

- Hints:
  - The matrix multiplication for \( n = 2 \) is defined as:
    \[
    AB = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \end{pmatrix}
    \]
  - Every memory is 1D. Use a 1D representation of your 2D data structure or use the 2D CUDA support (\texttt{cudaMallocPitch, cudaMemcpy2D})
  - Keep the matrix values <10 to prevent an overflow