CUDA
Lecture 1

Manfred Liebmann
Technische Universität München
Chair of Optimal Control
Center for Mathematical Sciences, M17
manfred.liebmann@tum.de

December 14, 2015

Compute Unified Device Architecture (CUDA)

Nvidia's CUDA is a parallel computing platform and programming model that enables dramatic increases in computing performance by harnessing the power of the graphics processing unit (GPU).

Since its introduction in 2006, CUDA has been widely deployed through thousands of applications and published research papers, and supported by an installed base of hundreds of millions of CUDA-enabled GPUs in notebooks, workstations, compute clusters and supercomputers. Applications used in astronomy, biology, chemistry, physics, data mining, manufacturing, finance, and other computationally intense fields are increasing using CUDA to deliver the benefits of GPU acceleration.

Nvidia Accelerated Computing Toolkit

`https://developer.nvidia.com/accelerated-computing`

Nvidia Accelerated Computing Toolkit

Nvidia provides a vast collection of libraries for integration in traditional applications.

- **Technologies**
  - Deep learning (`https://developer.nvidia.com/deep-learning`)
  - Accelerated libraries (`https://developer.nvidia.com/gpu-accelerated-libraries`)
  - OpenACC (`https://developer.nvidia.com/openacc`)
  - CUDA Zone (`https://developer.nvidia.com/cuda-zone`)

Deep learning

Deep learning is the fastest-growing field in machine learning. It uses many-layered Deep Neural Networks (DNNs) to learn levels of representation and abstraction that make sense of data such as images, sound, and text.

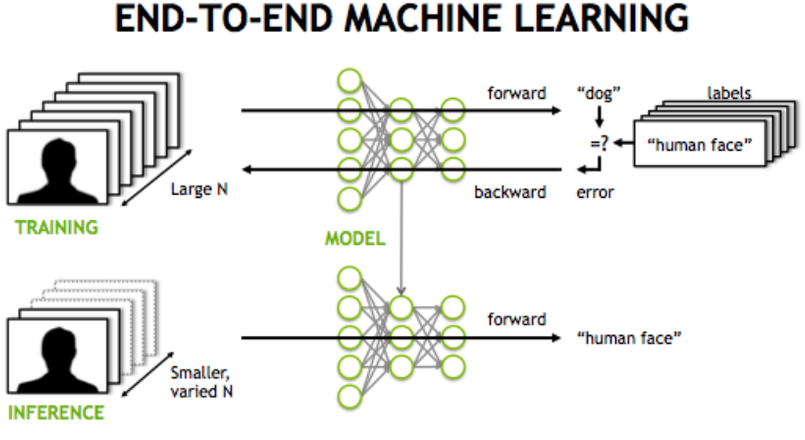

Figure 1: Deep learning pipeline

Deep learning is used in the research community and in industry to help solve many big data problems such as computer vision, speech recognition, and natural language processing.

GPU-Accelerated Libraries

GPU-Accelerated libraries provide highly-optimized algorithms and functions you can incorporate into your applications, with minimal changes to your existing code. Many support drop-in compatibility to replace industry standard CPU-only libraries such as MKL, IPP, FFTW and widely-used libraries. Some also feature automatic multi-GPU performance scaling.

AmgX, cuDNN, cuFFT, NPP, FFmpeg, CHOLMOD, CULA Tools, MAGMA, IMSL Fortran Numerical Library, cuSOLVER, cuSPARSE, cuBLAS, ArrayFire, cuRAND, CUDA Math Library, Thrust, NVBIO, Nvidia Video Codec SDK, HiPLAR, OpenCV, Geometry Performance Primitives, Paralution, Triton Ocean SDK.

Figure 2: GPU-accelerated libraries

OpenACC Toolkit

The OpenACC Toolkit from NVIDIA offers scientists and researchers a simple way to accelerated scientific computing without significant programming effort. Simply insert hints (or directives) in C or Fortran code and the OpenACC compiler runs the code on the GPU.

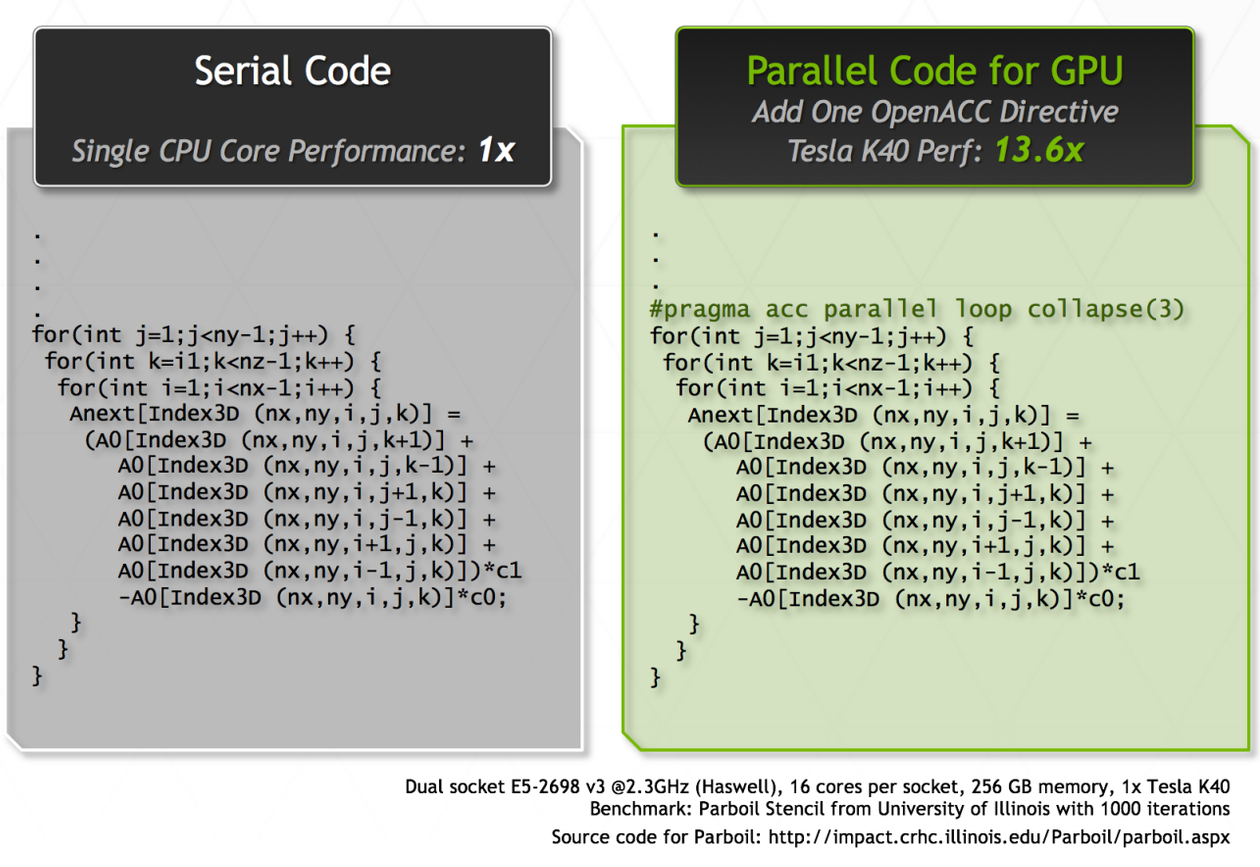

Figure 3: How OpenACC works

CUDA Toolkit

The Nvidia CUDA Toolkit provides a comprehensive development environment for C and C++ developers building GPU-accelerated applications. The CUDA Toolkit includes a compiler for Nvidia GPUs, math libraries, and tools for debugging and optimizing the performance of applications.

CUDA Zone

`https://developer.nvidia.com/cuda-zone`

CUDA 7.5 (September 2015)

`https://developer.nvidia.com/cuda-downloads`