CUDA Introduction

Piotr Luszczek

October 30, 2017

Per-Core CPU Performance

Individual core speed no longer increases
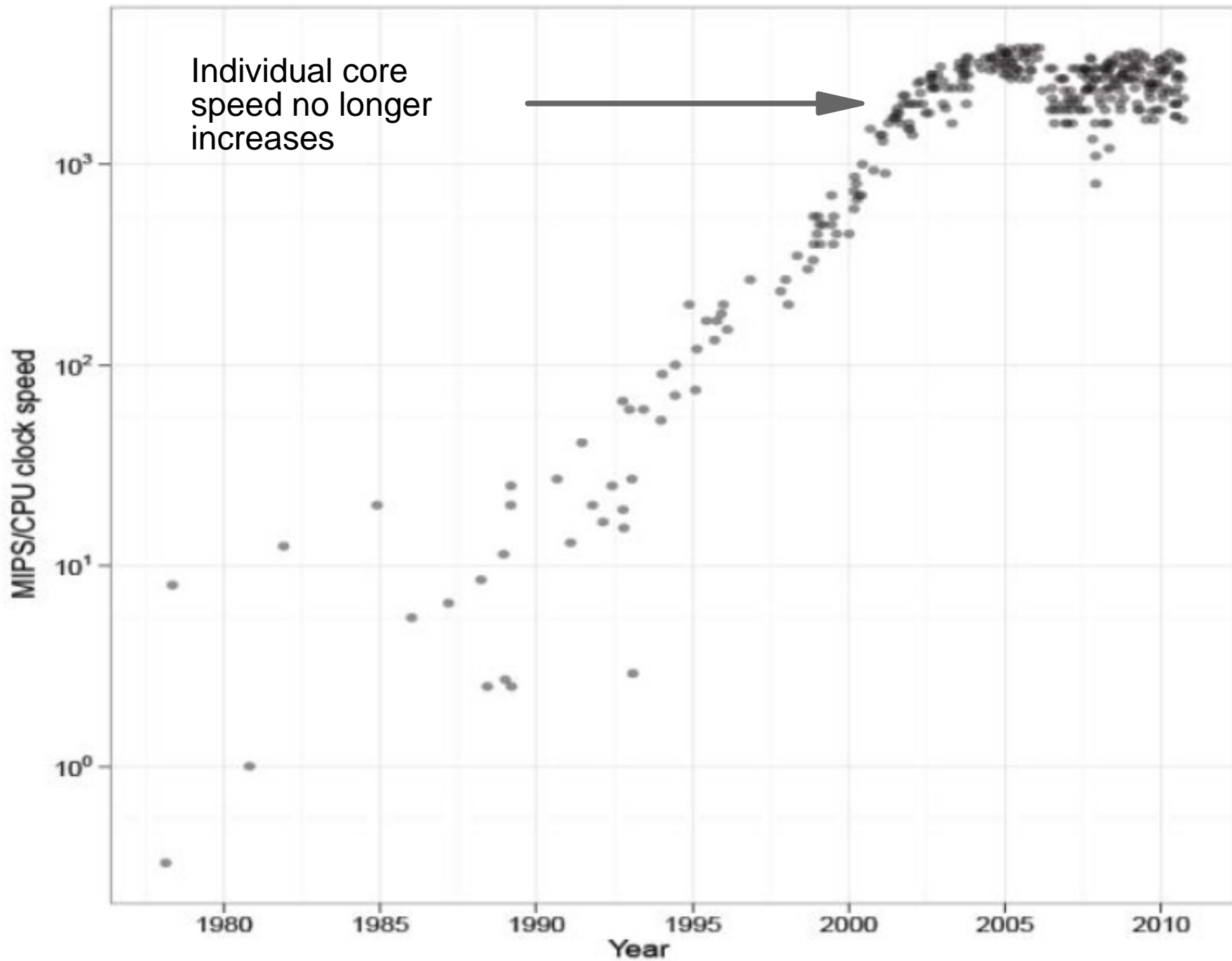
GPU vs. CPU Performance over Years
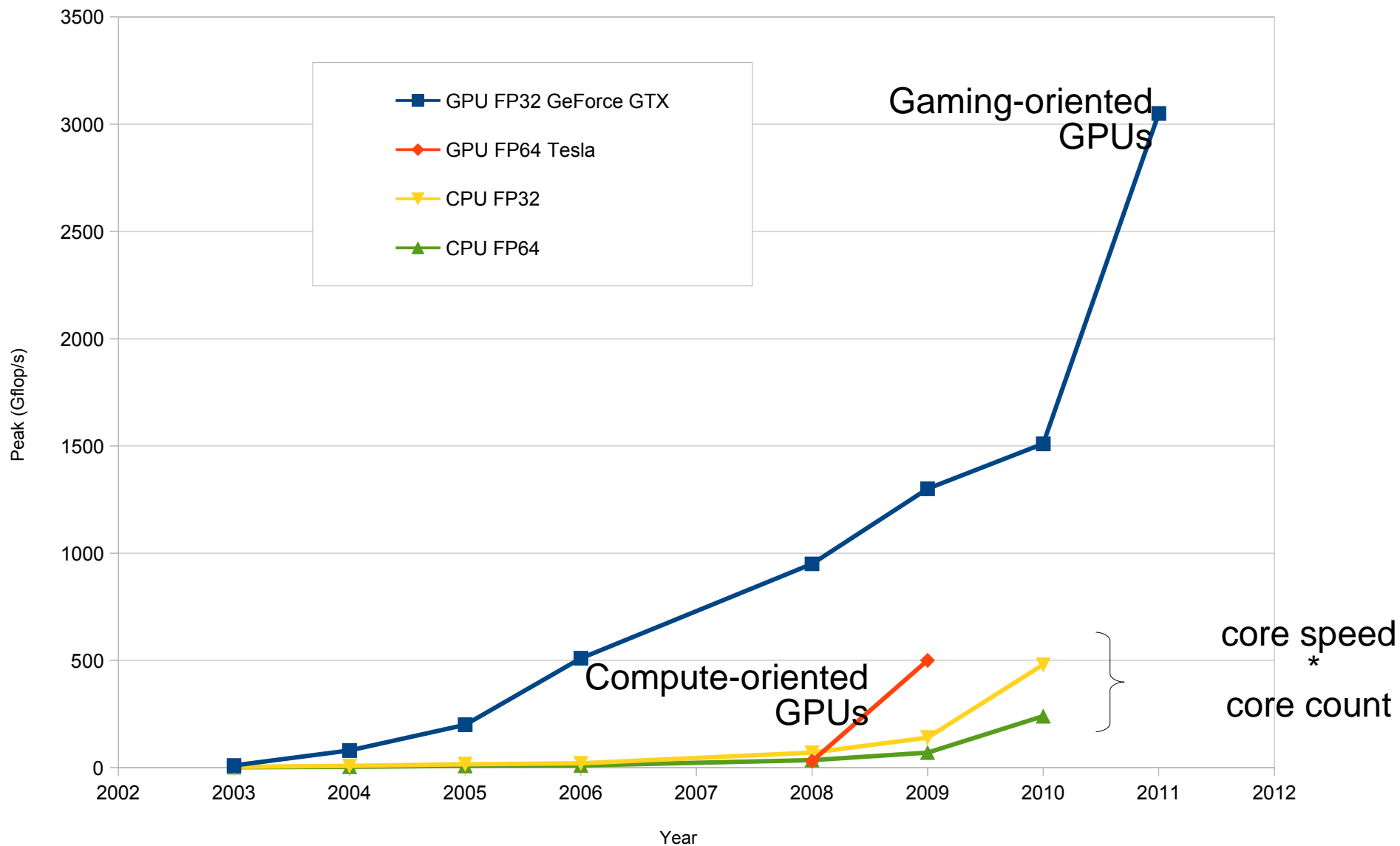
GPU and GPGPU: The Origin Story

- Programmable graphics pipeline
  - GLSL (shader language)

- Interpolation vs. dynamic range
  - Colors in graphics look better in floating-point

- Early attempts at programming
  - Cg, Brook, …

- Modern standards or de facto standards
  - CUDA (currently 8, 9alpha)
    - Compute Unified Device Architecture
  - OpenCL (currently 2)

- High-level languages
  - OpenMP 4.5 and offload directives
  - OpenACC (like OpenMP but GPU oriented)

Hardware: CPU vs. GPU

Socket 0   Socket 1

QPI

Socket 2   Socket 3

GPU Chip 0   GPU Chip 1

SLI

Long latency
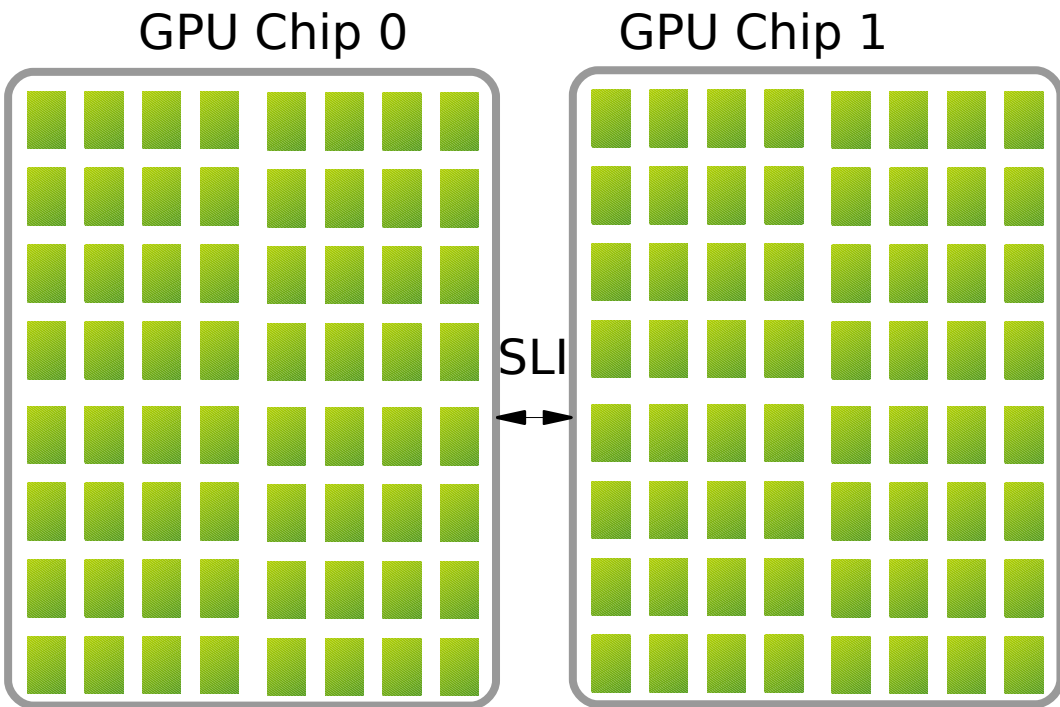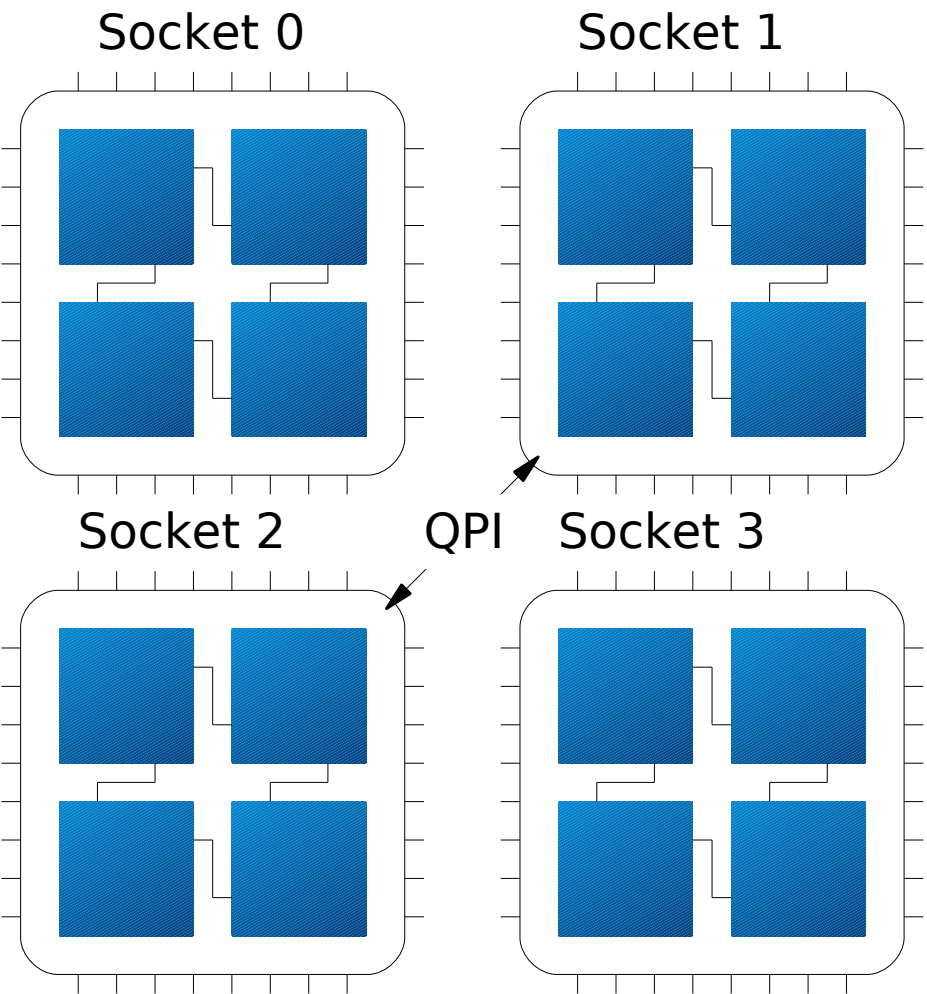
Main Memory RAM: DDR3 or DDR4
Size: ~500 GiB
Bus width: 64 bits
Speed: ~100 GB/s

PCIexpress

15 GB/s

GPU Memory RAM: GDDR5
Size: ~10 GiB
Bus width: 384 bits
Speed: ~200 GB/s

Software: CPU + GPU

CPU code

GPU code

Ideal: CPU and GPU codes overlap