INTRODUCTION TO PARALLEL AND GPU COMPUTING

Carlo Nardone, Sr. Solution Architect, PSG
AGENDA

1 Intro
2 Processors Trends
3 Multiprocessors
4 GPU Computing
PARALLEL COMPUTING, ILLUSTRATED
FROM MULTICORE TO MANYCORE
PARALLEL COMPUTING IS OLD

Source: Tim Mattson
Processors Trends
Gordon Moore (Intel co-founder) predicted in 1965 that transistor density of semiconductor chips would double every 18 months.

Moore’s Law

Source: U.Delaware CISC879 / J.Dongarra
THE POWER WALL

Source: S. Borkar (Intel)
MICROPROCESSORS PERFORMANCE TREND

(SPECint)
Uniprocessor
Performance

Pentium 4, 3.0 GHz,
20 stage, 3 CISC
issue (6 uop issue)

???%/year

Pentium 4, 3.6 GHz,
31 stage, 6 uop
issue, 3 CISC issue

52%/year

Sparc V7-RISC
5-stage
Sun 4/260
16.7 MHz

Vax "Star", CISC
Vax-11/780

25%/year

Vax "Nautilus",
CISC, Vax 8700

PowerPC 604, 100
MHz
7 stage, 4 issue

From Hennessy and Patterson, Computer Architecture: A
MICROPROCESSORS PERFORMANCE TREND

source: CA5 Hennessy/Patterson
MULTICORE SOLUTION TO POWER WALL

Input → Processor

Output

Processor

Input

f/2

Processor

Output

f

f/2

Capacitance = 2.2C
Voltage = 0.6V
Frequency = 0.5f
Power = 0.396CV^2f

Capacitance = C
Voltage = V
Frequency = f
Power = CV^2f

Source: Tim Mattson
THE PROCESSOR-MEMORY GAP

source: CA5
Hennessy & Patterson
Arithmetic intensity, specified as the number of floating-point operations to run the program divided by the number of bytes accessed in main memory [Williams et al. 2009]. Some kernels have an arithmetic intensity that scales with problem size, such as dense matrix, but there are many kernels with arithmetic intensities independent of problem size. Source: CA5 Hennessy & Patterson
Multiprocessors
FLYNN’S TAXONOMY (1)

source: CA5 Hennessy & Patterson
FLYNN’S TAXONOMY (2)

source: CA5 Hennessy & Patterson
FLYNN’S TAXONOMY (3)

source: CA5 Hennessy & Patterson
FLYNN’S TAXONOMY (4)

source: CA5 Hennessy & Patterson
FLYNN’S TAXONOMY (5)

source: CA5 Hennessy & Patterson
FLYNN’S TAXONOMY (6)

source: CA5 Hennessy & Patterson
MULTIPROCESSORS

source: CA5
Hennessy & Patterson
MULTIPROCESSOR SMP CLASS

source: CA4
Hennessy & Patterson
MULTIPROCESSOR DISTRIBUTED MEM CLASS

source: CA4
Hennessy & Patterson
X86 ARCHITECTURE

Intel Clovertown

Core 2
4MB Shared L2

Front Side Bus

10.6 GB/s

Chipset (4x64b controllers)

21.3 GB/s (read)

667MHz FBDIMMs

10.6 GB/s (write)

10.5 GB/s

AMD Opteron

Opteron
1MB victim

Opteron
1MB victim

SRI / crossbar

120b memory controller

4GB/s (each direction)

Opteron

SRI / crossbar

120b memory controller

667MHz DDR2 DIMMs

10.66 GB/s

Uniform Memory Access

Non-uniform Memory Access

Adapted from Sam Williams, John Shalf, LBL/NERSC et al.
GPU Computing
Why GPU Computing?

![Graphs showing performance trends over time for different GPU models and CPU models.](image-url)

- **GFlops/sec**
  - Single Precision: NVIDIA GPU
  - Single Precision: x86 CPU
  - Double Precision: NVIDIA GPU
  - Double Precision: x86 CPU

- **GBytes/sec**
  - NVIDIA GPU
  - X86 CPU
  - ECC off
The Graphics Pipeline

- Vertex
- Tessellation
- Geometry
- Rasterize
- Pixel
- Test & Blend
- Framebuffer

Hardware **used to look like this**

- Vertex, pixel processing became programmable
- New stages added

**GPU architecture increasingly centers around shader execution**
Vertex shaders, pixel shaders, etc. become *threads* running different programs on a flexible core.
ACCELERATED COMPUTING

10x Performance & 5x Energy Efficiency for HPC

CPU
Optimized for Serial Tasks

GPU Accelerator
Optimized for Parallel Tasks
GEFORCE 8: 1\textsuperscript{ST} MODERN GPU ARCH
GT200 GPU ARCH
Kepler GK110 Block Diagram

Architecture
- 7.1B Transistors
- 15 SMX units
- > 1 TFLOP FP64
- 1.5 MB L2 Cache
- 384-bit GDDR5
KEPLER SMX DIAGRAM
GPU Accelerate Science

- Medical Imaging
  - U of Utah
- Molecular Dynamics
  - U of Illinois, Urbana
- Video Transcoding
  - Elemental Tech
- Matlab Computing
  - AccelerEyes
- Astrophysics
  - RIKEN

- Financial Simulation
  - Oxford
- Linear Algebra
  - Universidad Jaime
- 3D Ultrasound
  - Techniscan
- Quantum Chemistry
  - U of Illinois, Urbana
- Gene Sequencing
  - U of Maryland
U.S. DoE to Build Two Flagship Supercomputers for National Labs

Major Step Forward on the Path to Exascale

- Summit
  - 150-300 PFLOPS Peak Performance
  - IBM POWER CPU + NVIDIA Volta GPU
  - NVLink High Speed Interconnect
  - >40 TFLOPS per Node, >3,400 Nodes
  - 2017

Sierra
Top500: Performance from Accelerators

Chart showing the total performance (PFLOPS) from 2007 to 2012, with a significant increase in performance after 2010.
THANKS!
Carlo Nardone
+39 335 5828197
cnardone@nvidia.com