The Graphics Processing Unit (GPU) revolution

Ramu Anandakrishnan
Outline

- The need for parallel processing
- Basic parallel processing concepts
- The GPU – a massively parallel processor
- Overview of GPU hardware architecture
- Introduction to GPU programming
- Performance considerations
- Homework assignment

Lecture 1

Lecture 2
A typical scientific computing problem

- The DNA in the human genome is made up of ~300 billion atoms.
- How is this 3 m long DNA strand packed into the 3 µm nucleus of a cell?
- In this tightly packed environment how is a specific gene transcribed, as and when needed?
- Experimental tools are not yet sophisticated enough to answer such questions.
Molecular dynamics simulations can be used to study these systems.

- \( N = 3 \times 10^6 \) atoms
- \( T = 10^{-15} \) sec per step (simulation time)
- \( F = 10^{12} \) FLOPS (teraflop)
- Time simulated / year = \( \left( \frac{T}{N^2} \right) \times F \times 3600 \times 8 \times 365 \)
- \( \sim 1 \) nanosecond
- Not long enough to “observe” meaningful activity

\[
F_i = F_{i,\text{elec}} + F_{i,\text{bond}} + F_{i,\text{vdw}} = m_i a_i
\]

\[
r_i(t+\Delta t) = r_i(t) + v_i(t+\Delta t) + \frac{1}{2} a_i t^2
\]

\[
F_{i,\text{elec}} = \frac{q_i}{4 \pi \epsilon_0} \sum_{j=1}^{N} \frac{q_j}{r_{ij}^2}
\]
The “power wall” limits how much faster individual processors can run

- In 1965 Intel founder Gordon Moore predicted that the number of components in a processor would double every year.
- To get a corresponding increase in computing power (FLOPS), you need to increase clock speed.
- But the resulting increase in heat dissipation is limiting how fast you can run the processors – the power wall.
- The answer: parallel processing
Instruction level parallelism (ILP)
Limitations of instruction level parallelism

- The Pentium 4 has a 20 stage pipeline
- Typically every 5th or 6th instruction is a branch instruction – longer pipelines have a larger branch mis-prediction penalty
- Different instructions require different number of clock cycle and occur at different frequencies
Application level parallelism
Application level parallelism - implementation models

Shared address space (OpenMP)
HokieOne: 492 cores

Message Passing (MPI)
HokieSpeed: 2448 cores (204 nodes)
Limitations of supercomputer architecture
Gaming machines were way ahead of supercomputers when it came to parallel processing

- General purpose central processing units (Intel/AMD) were just not up to the task of rapidly rendering realistic images
- Rendering graphics involves a large number of computations
- But each pixel can be computed more or less independently
- So the computer graphics were rendered using dedicated graphical processing units (GPUs)
GPUs – Massively parallel processors

GeForce GTX 690 Specifications

- CUDA Cores: 3072
- Base Clock: 915 MHz
- Boost Clock: 1019 MHz
- Memory Config: 4GB / 512-bit GDDR5
- Memory Speed: 6.0 Gbps
- Power Connectors: 8-pin + 8-pin
- TDP: 300W
- Outputs: 3x DL-DVI, Mini-DisplayPort 1.2
- Bus Interface: PCI Express 3.0

(www.nvidia.com)
3 of the 5 fastest supercomputers in the world use GPUs (June 2012)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RIKEN Advanced Institute for Computational Science (AICS) Japan</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu</td>
</tr>
<tr>
<td>2</td>
<td>National Supercomputing Center in Tianjin China</td>
<td>NUDT YH MPP, Xeon X5670 6C 2.93 GHz, <strong>NVIDIA 2050</strong> NUDT</td>
</tr>
<tr>
<td>3</td>
<td>DOE/SC/Oak Ridge National Laboratory United States</td>
<td>Cray XT5-HE Opteron 6-core 2.6 GHz Cray Inc.</td>
</tr>
<tr>
<td>4</td>
<td>National Supercomputing Centre in Shenzhen (NSCS) China</td>
<td>Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, <strong>NVIDIA 2050</strong> Dawning</td>
</tr>
<tr>
<td>5</td>
<td>GSIC Center, Tokyo Institute of Technology Japan</td>
<td>HP ProLiant SL390s G7 Xeon 6C X5670, <strong>Nvidia GPU</strong> NEC/HP</td>
</tr>
</tbody>
</table>
5 of the top 10 “green” supercomputers use GPUs (June 2012)

<table>
<thead>
<tr>
<th>Green500 Rank</th>
<th>MFLOPS/W</th>
<th>Site</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2026.48</td>
<td>IBM - Rochester</td>
<td>BlueGene/Q, Power BQC 16C 1.60 GHz, Custom</td>
</tr>
<tr>
<td>2</td>
<td>2026.48</td>
<td>IBM Thomas J. Watson Research Center</td>
<td>BlueGene/Q, Power BQC 16C 1.60 GHz, Custom</td>
</tr>
<tr>
<td>3</td>
<td>1996.09</td>
<td>IBM - Rochester</td>
<td>BlueGene/Q, Power BQC 16C 1.60 GHz, Custom</td>
</tr>
<tr>
<td>4</td>
<td>1988.56</td>
<td>DOE/NNSA/LLNL</td>
<td>BlueGene/Q, Power BQC 16C 1.60 GHz, Custom</td>
</tr>
<tr>
<td>5</td>
<td>1689.86</td>
<td>IBM Thomas J. Watson Research Center</td>
<td>NNSA/SC Blue Gene/Q Prototype 1</td>
</tr>
<tr>
<td>6</td>
<td>1378.32</td>
<td>Nagasaki University</td>
<td>DEGIMA Cluster, Intel i5, <strong>ATI Radeon GPU</strong>, Infiniband QDR</td>
</tr>
<tr>
<td>7</td>
<td>1266.26</td>
<td>Barcelona Supercomputing Center</td>
<td>Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, <strong>NVIDIA 2090</strong></td>
</tr>
<tr>
<td>8</td>
<td>1010.11</td>
<td>TGCC / GENCI</td>
<td>Curie Hybrid Nodes - Bullx B505, <strong>Nvidia M2090</strong>, Xeon E5640 2.67 GHz, Infiniband QDR</td>
</tr>
<tr>
<td>9</td>
<td>963.70</td>
<td>Institute of Process Engineering, Chinese Academy of Sciences</td>
<td>Mole-8.5 Cluster, Xeon X5520 4C 2.27 GHz, Infiniband QDR, <strong>NVIDIA 2050</strong></td>
</tr>
<tr>
<td>10</td>
<td>958.35</td>
<td>GSIC Center, Tokyo Institute of Technology</td>
<td>HP ProLiant SL390s G7 Xeon 6C X5670, <strong>Nvidia GPU</strong>, Linux/Windows</td>
</tr>
<tr>
<td>11</td>
<td>928.96</td>
<td><strong>Virginia Tech</strong></td>
<td>HokieSpeed, SuperServer 2026GT-TRF, Xeon E5645 6C 2.40GHz, Infiniband QDR, <strong>NVIDIA 2050</strong></td>
</tr>
</tbody>
</table>