HPC with Multicore and GPUs

Stan Tomov
Electrical Engineering and Computer Science Department
University of Tennessee, Knoxville

CS 594 Lecture Notes
January 22, 2014
Outline

- Introduction
  - Hardware trends

- Challenges of using multicore+GPUs

- How to code for GPUs and multicore
  - An approach that we will study

- Introduction to CUDA and the cs954 project/library

- Conclusions
Speeding up Computer Simulations

Better numerical methods

- Exploit advances in hardware

E.g. a posteriori error analysis: solving for much less DOF but achieving the same accuracy

http://www.cs.utk.edu/~tomov/cflow/

Performance Development in Top500

- Manage to use hardware efficiently for real-world HPC applications
- Match LU benchmark in performance!
Why multicore and GPUs?

Hardware trends

- Multicore

- GPU Accelerators

Power is the root cause of all this!

A hardware issue just became a software problem.

(Source: slide from Kathy Yelick)

<table>
<thead>
<tr>
<th></th>
<th>GeForce GTX 280</th>
<th>GeForce GTX 260</th>
<th>Tesla C1060</th>
<th>Tesla S1070</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form Factor</td>
<td>Dual slot card</td>
<td>Dual slot card</td>
<td>Dual slot card</td>
<td>Rackmount</td>
</tr>
<tr>
<td>TPCs</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>4x10</td>
</tr>
<tr>
<td>SMs</td>
<td>30</td>
<td>24</td>
<td>30</td>
<td>4x30</td>
</tr>
<tr>
<td>SPs</td>
<td>240</td>
<td>192</td>
<td>240</td>
<td>4x240</td>
</tr>
<tr>
<td>Graphics Freq.</td>
<td>602 MHz</td>
<td>575 MHz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processor Freq.</td>
<td>1296 MHz</td>
<td>1242 MHz</td>
<td>1300 MHz</td>
<td>1500 MHz</td>
</tr>
<tr>
<td>Memory Freq.</td>
<td>1107 MHz</td>
<td>999 MHz</td>
<td>800 MHz</td>
<td>800 MHz</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>141.7 GB/s</td>
<td>127.9 GB/s</td>
<td>102.4 GB/s</td>
<td>4x102.4 GB/s</td>
</tr>
<tr>
<td>Memory Capacity</td>
<td>1 GB</td>
<td>866MB</td>
<td>4 GB</td>
<td>4x4 GB</td>
</tr>
<tr>
<td>Power</td>
<td>236W TDP</td>
<td>183W TDP</td>
<td>160W &quot;Typical&quot;</td>
<td>700W &quot;Typical&quot;</td>
</tr>
<tr>
<td>SP GFLOP/s</td>
<td>622.1</td>
<td>476.9</td>
<td>624.0</td>
<td>4x720.0</td>
</tr>
<tr>
<td>SP GFLOP/s (w/MUL)</td>
<td>933.1</td>
<td>715.4</td>
<td>966.0</td>
<td>4x1080.0</td>
</tr>
<tr>
<td>DP GFLOP/s</td>
<td>77.8</td>
<td>59.6</td>
<td>73.0</td>
<td>4x72.0</td>
</tr>
</tbody>
</table>

(Source: "NVIDIA's GT200: Inside a Parallel Processor")
Main Issues

- **Increase in parallelism** *1
  How to code (programming model, language, productivity, etc.)?

- **Increase in commun. cost (vs computation)** *2
  How to redesign algorithms?

- **Hybrid Computing** *3
  How to split and schedule the computation between hybrid hardware components?

Despite issues, **high speedups** on HPC applications are reported using GPUs
(from NVIDIA CUDA Zone homepage)

- **CUDA architecture & programming:** *1
  - A data-parallel approach that scales
  - Similar amount of efforts on using CPUs vs GPUs by domain scientists demonstrate the GPUs' potential

- **Processor speed** improves 59% / year but memory bandwidth by 23%
  latency by 5.5% *2

- **e.g., schedule small** non-parallelizable tasks on the CPU, and large and parallelizable on the GPU *3
Evolution of GPUs

**GPUs**: excelling in graphics rendering

- Scene model → **Streams of data** → Graphics pipelined computation → Final image

Repeated fast over and over: e.g., TV refresh rate is 30 fps; limit is 60 fps

This type of computation:
- Requires **enormous computational power**
- Allows for **high parallelism**
- Needs **high bandwidth vs low latency**
  (as low latencies can be compensated with deep graphics pipeline)

Obviously, this pattern of computation is common with many other applications
Challenges of using multicore+ GPUs

- **Massive parallelism**
  Many GPU cores, serial kernel execution
  [e.g. 240 in the GTX280; up to 512 in *Fermi* – to have concurrent kernel execution]

- **Hybrid/heterogeneous architectures**
  Match algorithmic requirements to architectural strengths
  [e.g. small, non-parallelizable tasks to run on CPU, large and parallelizable on GPU]

- **Compute vs communication gap**
  Exponentially growing gap; persistent challenge
  [Processor speed improves 59%, memory bandwidth 23%, latency 5.5%]
  [on all levels, e.g. a GPU Tesla C1070 (4 x C1060) has compute power of $O(1,000)$ Gflop/s but GPUs communicate through the CPU using $O(1)$ GB/s connection]
How to Code for GPUs?

- Complex question
  - Language, programming model, user productivity, etc

- Recommendations
  - Use CUDA / OpenCL
    [already demonstrated benefits in many areas; data-based parallelism; move to support task-based]
  - Use GPU BLAS
    [high level; available after introduction of shared memory – can do data reuse; leverage existing developments ]
  - Use Hybrid Algorithms
    [currently GPUs – massive parallelism but serial kernel execution; hybrid approach – small non-parallelizable tasks on the CPU, large parallelizable tasks on the GPU ]

Typical order of acceleration:
- dense matrix-matrix $O(1)$ $X$
- dense matrix-vector $O(10)$ $X$
- sparse matrix-vector $O(100)$ $X$
An approach for multicore+GPUs

- Split algorithms into **tasks** and **dependencies** between them, e.g., represented as DAGs
- Schedule the execution in parallel without violating data dependencies

### Algorithms as DAGs
(small tasks/tiles for homogeneous **multicore**) e.g., in the **PLASMA** library for Dense Linear Algebra http://icl.cs.utk.edu/plasma/

### Hybrid CPU+GPU algorithms
(small tasks for multicores and large tasks for GPUs) e.g., in the **MAGMA** library for Dense Linear Algebra http://icl.cs.utk.edu/magma/
An approach for multicore+GPUs

- Split algorithms into tasks and dependencies between them, e.g., represented as DAGs
- Schedule the execution in parallel without violating data dependencies

Want to develop libcs594:
A framework for parallel programming of linear algebra algorithms for systems of multicores accelerated with GPUs

e.g., in the MAGMA library for Dense Linear Algebra
http://icl.cs.utk.edu/magma/
A Parallel Programming Framework Inspired by CUDA / OpenCL

- To experiment with techniques
- To try to develop hybrid algorithms
- To be built as part of homeworks (maybe final projects?)

- Data-parallel tasks (on GPUs, multicores or both)
- To be started asynchronously from a “master” going over the critical path
- The critical path to be overlapped with the data-parallel tasks?
How to program in parallel?

- There are many parallel programming paradigms (to be covered w/ Prof. George Bosilca), e.g.,

- In reality applications usually combine different paradigms

- CUDA and OpenCL have roots in the data-parallel approach (now adding support for task parallelism)

Compute Unified Device Architecture (CUDA) Software Stack

(CPU) Application

(CUDA) CUDA Libraries

(CUDA) CUDA Runtime

(CUDA) CUDA Driver

(GPU)

CUBLAS, CUFFT, MAGMA, ...

C like API

(Source: NVIDIA CUDA Programming Guide)
CUDA Memory Model

(Source: NVIDIA CUDA Programming Guide)
CUDA Hardware Model

(Source: NVIDIA CUDA Programming Guide)
CUDA Programming Model

- **Grid of thread blocks**
  (blocks of the same dimension, grouped together to execute the same kernel)

- **Thread block**
  (a batch of threads with fast shared memory executes a kernel)

- **Sequential code launches asynchronously GPU kernels**

---

```c
// set the grid and thread configuration
Dim3 dimBlock(3,5);
Dim3 dimGrid(2,3);

// Launch the device computation
MatVec<<<dimGrid, dimBlock>>>( . . . );

__global__ void MatVec( . . . ) {
  // Block index
  int bx = blockIdx.x;
  int by = blockIdx.y;

  // Thread index
  int tx = threadIdx.x;
  int ty = threadIdx.y;

  . . .
}
```

(Source: NVIDIA CUDA Programming Guide)
Introduction to the cs594 library

- **Context initialization**
  Initialize a pool of hardware resources for a data-parallel computation (could be many contexts/nested, etc.)
  ```c
  cs594_context *context = cs594_init(8, &argc, argv);
  ...
  cs594_finish(context);
  ```

- **The library will be developed w/ Prof. George Bosilca**
  Starting point is a sequential implementation
  To add parallel algorithm implementations

- **Data-parallel kernels development**
  Launch a kernel:
  ```c
  cs594_grid( 512 );
  context->sgemv_kernel<<< grid >>>( . . . );
  ```
  A matrix-vector product kernel example:
  ```c
  void sgemv_kernel(cs594_context_t *context, ...) {
    // 1. Get from context where the data is for this kernel invocation
    // 2. Start the sequential code processing the particular data
    ...
  }
  ```
  (see the sources for further information)

- **Hybrid algorithms for multicore and GPU computing**
  (to be developed using the library)
Conclusions

Hybrid Multicore+GPU computing:

- Architecture trends: towards heterogeneous/hybrid designs
- Can significantly accelerate linear algebra [vs just multicores];
- Can significantly accelerate algorithms that are slow on homogeneous architectures