BarraCUDA: A Supercomputer for Student Research

Matt Redmond

December 18, 2009

1 Abstract

BarraCUDA (Beginner's Analytic Research and Rendering Architecture) will be a small-scale home-brew supercomputer designed around the NVIDIA CUDA architecture. The project is meant to serve as a platform for student-researchers to develop parallel-programming skills, and design programs that showcase concepts in modern science research through real-time visualization. After building this supercomputer, I intend to develop a comprehensive and accurate three dimensional simulation of electrodynamics suitable for demonstrating the principles of electromagnetism to high school physics students. I am seeking funds and/or donated parts to help make this dream a reality.

2 Motivation

I'm currently a Senior in high school, and through my Applied Science Research and Computer Science classes, I've gradually become aware of the growing importance of high performance computational power in the fields of physical, chemical, and biological research. I'm also quite interested in learning more about fluid dynamics and visual representations of physical modeling problems. Current research in these computationally expensive fields includes [1], [2], and [3].

Last summer, I worked in Caltech's materials physics lab on the DANSE project. The DANSE project uses distributed computational analysis to model the theoretical results of neutron scattering experiments before actually conducting them. At Caltech, I was exposed to the elegant concept and power of a Beowulf cluster. I realized that our high school could make good use of such a powerful tool (on a smaller scale) for instructional purposes, in the demonstration/visualization of scientific concepts as well as the programming process. I envision a class dedicated to high-performance computing, where students create programs for the science department to be used for various purposes. It would be an excellent way to prepare students for careers as science researchers.

I'd like to build a supercomputer based largely on NVIDIA's CUDA (C Unified Device Architecture) platform, because CUDA represents the new direction that high performance computing is moving in: instead of waiting for a single fast CPU to perform heavy calculations sequentially, these calculations are farmed out in parallel to hundreds of individual GPU cores for computation. The actual performance speed-up (across diverse applications) is reported by various sources [4] to be anything from 5x to 2600x the time it would take

a CPU-based system to do the same tasks. Furthermore, the CUDA architecture is alive and well-supported by the community, with various development tools available, and many new applications released frequently. CUDA is supported on a few different chipsets, but the two most suitable architectures for this project are the NVIDIA GeForce GTX295 and the NVIDIA Tesla C1060. Ideally, I would be able to use the Tesla boards, as they are optimized for supercomputing, but unless I am able to get them donated by NVIDIA, they are out of my price range.

Completing this project would provide me with a fair background in supercomputer design and construction, as well as an opportunity to learn the parallel programming skills that I'll need to be a successful researcher in the future. The programming API OpenMP is well-known and widely supported in the scientific computing community, and I'd like to gain further experience working with it on a platform that can unlock all of its power. When I finish the project, I'll donate the supercomputer to my school, giving other interested students an opportunity to learn the basics of high-power computing.

3   Design

Although I initially toyed with the idea of building a CPU-based 4-node cluster computer, after speaking with Dr. Kayvon Fatahalian (a post-doctoral researcher working at Stanford and NVIDIA on the CUDA project), I realized that I could actually get a lot more computational power for less money with a single-node, multi-GPU system. The NVIDIA cards like the GTX295 and the Tesla C1060 offer a compelling performance per dollar ratio, allowing this design to reach speeds greater than one TFLOP of processing power for substantially less money than an equivalent CPU-based system would cost.

The motherboard was chosen because it utilizes the latest X58 chipset from Intel, with a LGA1366 socket that will support a quad core Bloomfield CPU. Furthermore, this particular motherboard was chosen because it supports double-wide PCI Express cards (the GTX295 and the TeslaC1060 are both twice as wide as a standard PCI-E card, so any motherboard used for this project needs to meet this requirement). The motherboard also supports the fastest (current as of December 2009) DDR3 memory, so the computer will not bottleneck during intensive memory read-write operations.

The Intel Core i7-920 CPU was chosen because it provides the largest performance / price ratio of all quad-core CPUs on the market. Other options for this component include the Core i7-950, i7-960, and i7-975, but these processors offer only marginal performance gains in exchange for substantially higher costs.

The CUDA processors can be found on the NVIDIA GeForce GTX285 and 295, as well as the NVIDIA Tesla C1060. As mentioned above, the Tesla boards are are outside of the budget for this project unless they are donated. The GTX295 offers twice the performance of the GTX285, but it costs less than twice as much. Therefore, it was selected for this project.

Supercomputing can put a high demand on the memory of any system, and I wanted to ensure that the BarraCUDA system wouldn't bottleneck here. 12 GB of memory ensures that the CUDA boards have enough working space (in addition to each of their 1.7GB on-board memory) to conduct any calculations easily. In addition to having high memory capacity, the system uses the fastest possible memory (DDR3 1600). This doesn't come cheaply, but it is worth the extra money to ensure the best possible performance.

The hard drive is selected for its rapid data transfer rate. The Velociraptor drive does not provide a lot of storage space, but this is not a relevant concern given the applications that BarraCUDA will be used for.

The biggest bottleneck in any supercomputing is generally I/O speed, and this drive helps to alleviate some of that pressure by providing rapid seek/write times for all sectors.

An excellent introduction to the CUDA architecture (and indeed, a computer designed similar to the one described above) can be found at [5].

4 Component List

Component	Description	Cost
Motherboard	ASRock X58 Extreme LGA1366	169.99
CPU	Intel Core i7-920 Quad Core Bloomfield 2.66 Ghz	288.99
CUDA Boards	2x Gigabyte GV-N295-181-B GTX295	2x 494.99
Memory	OCZ Gold 12GB (6 x 2gb) DDR3 1600	310.99
Hard Drive	Western Digital Velociraptor 300GB SATA @ 10,000 RPM	199.99
Power Supply	Tuniq Ripper 1000W	149.99
Total Cost		$2,109.93

5 Funding

- Dr. James Dann $500

- Menlo School .... $???

- NVIDIA Corporation ... 2x GTX-295 or 2x Tesla C1060?

6 Citations

[1] CUDA-Based Incompressible Navier-Stokes Solver. Julien Thibault and Inanc Senocak.

<http://coen.boisestate.edu/senocak/files/BSU_CUDA_Res_v5.pdf>

[2] TeraFlop Computing on a Desktop PC with GPUs for 3D CDF. J. Tolke and M. Krafczyk.

<http://www.irmb.bau.tu-bs.de/UPLOADS/toelke/Publication/toelkeD3Q13.pdf>

[3] CUDA Acceleration of Molecular Dynamics. David Kirk and Wen-mei W. Hwu.

<http://www.ks.uiuc.edu/Research/gpu/files/lecture8casestudies.pdf>

[4] NVIDIA CUDA Project Repository. <http://www.nvidia.com/object/cuda_home.html#>

[5] Asrock Supercomputer Video. <http://www.youtube.com/watch?v=_87a6P0-Xjw&feature=related>