Compute Unified Device Architecture (CUDA) Computing

BTech 451
End of First Semester Report

University of Auckland
New Zealand

HaoYu, Gan
UPI: hgan006
ID: 1184198

E-mail: hgan006@aucklanduni.ac.nz

Abstract

Compute Unified Device Architecture (CUDA) is a parallel computing architecture developed by Nvidia for graphics processing. CUDA is the computing engine in NVIDIA graphics processing units (GPUs) that is accessible to software developers through variants of industry standard programming languages.

The GPU, as a specialized processor, addresses the demands of real-time high-resolution 3D graphics compute-intensive tasks. As of 2012 GPUs have evolved into highly parallel multi core systems allowing very efficient manipulation of large blocks of data. This design is more effective than general-purpose CPUs for algorithms where processing of large blocks of data is done in parallel.

(Wikipedia)

The project is divided into two parts, first part is researching on CUDA and benchmarking on certain GPU cards from NVIDIA by using COTS(commercial off the shelf) applications, gathering all the results and analysis the performance in order to gain a fully knowledge of CUDA.

Second part will continue testing without COTS, but with using Semi-Global Matching (SGM), as we know that SGM employs CUDA as the computing tool, and this SGM is currently a running project in our University.

Contents Page

1. Introduction

This is 4th year final project for student who majoring in Bachelor of Technology (IT) degree. The project carries weight of two semester University courses, which it is Btech451 Part A and Part B for semester 1 and semester 2 respectively.

Student should guarantee 8-10 hours of work per week in the first semester, and 16-20 hours of work per week in the second semester. This report is not the final version, it only keep tracks of first half of the project progress. Full version of this report will be provided at the end of this year 2012.

2. Project Overview

This project is an appraisal of a computing environment based on NVIDIA ARM and Intel for COTS and research applications. We will benchmark COTS (commercial off the shelf) applications including a few that have been optimized and certified for NVIDIA CUDA(Quadro and Tesla). Testing based on Quadro 2000 graphics card and Tesla C2075 companion card, also included a new technology called Maximus which combines Quadro and Tesla products. As NVIDIA likes to reiterate to their customers it's not a new product, it's a new technology – a new way to use NVIDIA's existing Quadro and Tesla products together. There's no new hardware involved, just new features in NVIDIA drivers and new hooks exposed to application developers.

2.1 Project goal

Gain a good understanding of how CUDA Computing works, and how does Semi-Global Matching employs CUDA as the computing tool (it will begin in Semester 2).
To understand today's industry emphasis, in both commercial and academically ways, also hoping to gain more knowledge about Hardware's design and architectures, not just only Software.

2.2 Company Information

This project is sponsored by the company Compucon New Zealand.

Compucon NZ is part of an International Computer manufacturing group of companies founded in 1989 in Sydney.

The NZ operation is registered as Modern Technology NZ Ltd and has established a reputation for technical excellence based on sound engineering and other knowledge based practices. All manufacturing processes are certified by Telarc ISO 9002 quality standards at our Albany assembly plant in Auckland NZ.

The Compucon team contributes to the success of our customers through our knowledge, excellence, commitment and supply of computing platforms and solutions meeting or exceeding customer expectations.

3. Research

Before it starts, some researches on CUDA are necessary, any topics related to CUDA are encouraged for further studying. For this project, i will need to set up a computer for benchmarking, so hardware knowledge such as motherboard, CPU, GPU, Memory RAM and Hard disk drive will be part of the area need to research as well.

3.1 CPU serial vs CUDA parallel Computing

CPU serial:

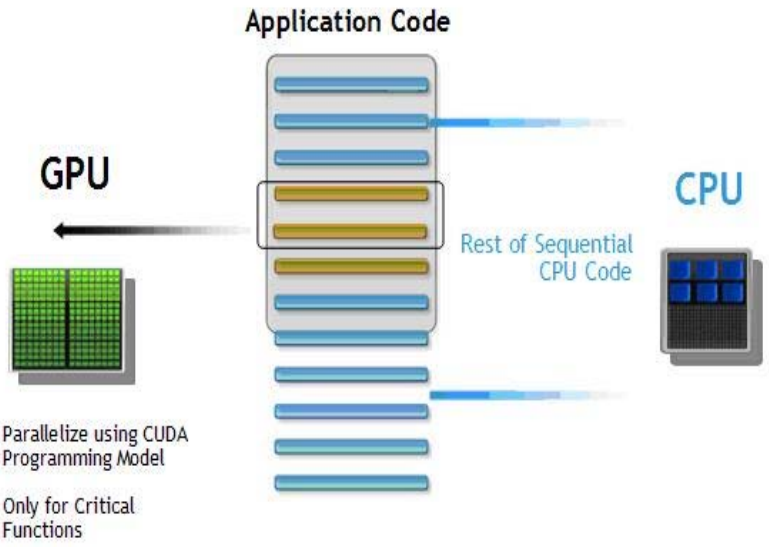
CPU serial Computing refers of a computer system that carries out the instructions of a computer program, to perform the basic arithmetical, logical, and input/output operations of the system by using central processing unit. It simply means that most of the thing is done by CPU alone.

CUDA parallel Computing:

CUDA is a parallel computing platform and programming model invented by NVIDIA. It enables dramatic increases in computing performance by harnessing the power of the graphics processing unit (GPU).
Computing is evolving from "central processing" on the CPU to "co-processing" on the CPU and GPU. To enable this new computing paradigm, NVIDIA invented the CUDA parallel computing architecture that is now shipping in GeForce, Quadro, and Tesla products, representing a significant installed base for application developers. (NVIDIA website)

With the development of the graphics card, the GPU is more powerful, somehow it has gone beyond the general-purpose CPU. If such a powerful chip is only for processing graphics, then it is too wasted, so NVIDIA launched CUDA that enabled graphics card can be used for purposes other than the image calculation.

GPU works as following:
The diagram shows how an application that normally runs in the CPU of a PC is ported over to the GPU.



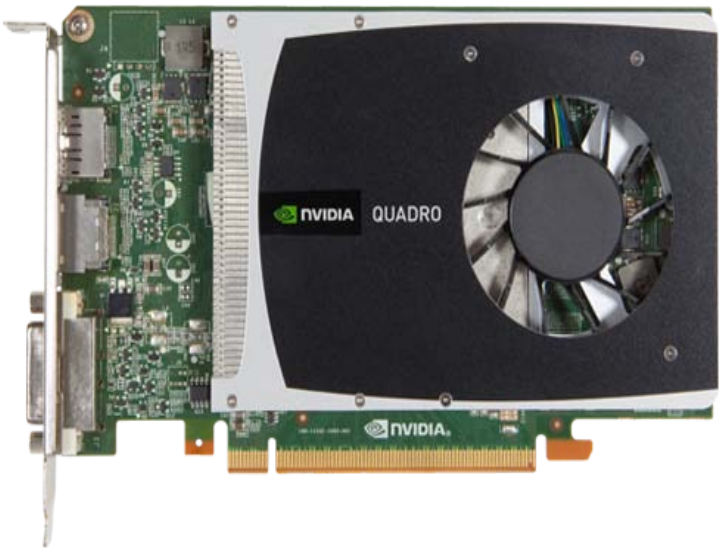
A software program is made up of application codes. The 1st step is to partition some sections of the code of a repeating nature to run on the GPU and other sections to remain on the CPU. Use C, C++ or other supported languages with special keywords. The sections of the code allocated to the GPU must be highly computing intensive and they tend to be the core algorithm and thus the critical parts of the application. A minor effort of code porting will result in significant performance gains. At the center of this parallel computing model is CUDA (Compute United Device Architecture) which is NVIDIA's parallel computing hardware and programming model.
So in general words', processing on CUDA is that GPU provides help/support for CPU.

CUDA parallel computing: To be parallel run using CPU + GPU
Example of CUDA processing flow
1. Copy data from main mem to GPU mem
2. CPU instructs the process to GPU
3. GPU execute parallel in each core
4. Copy the result from GPU mem to main memory

3.2 Quadro 2000

GPU Specifications:	
NVIDIA Quadro GPU	Quadro 2000
CUDA Cores	192
Form Factor	4.376” H x 7” L Single Slot
GPU Memory Specifications:	
Total Frame Buffer	1 GB GDDR5
Memory Interface	128-bit
Memory Bandwidth (GB/sec)	41.6 GB/s

The Quadro 2000 is based on Nvidia’s Fermi architecture [1, 2], and is equipped with 192 CUDA parallel processing cores. Accompanying these is 1GB of GDDR5 RAM running over a 128-bit memory interface, and offering 41.6GB of memory bandwidth. The Quadro is compute and graphics intensive, since the algorithm/equation handles graphical tasks.



http://www.nvidia.com/object/product-quadro-2000-us.html

3.3 Tesla C2075

GPU Specifications:	
NVIDIA Tesla GPU	Tesla C2075
CUDA Cores	448
Form Factor	9.75” PCIe x16 form factor
GPU Memory Specifications:	
Total Frame Buffer	6 GB GDDR5
Memory Interface	384-bit
Memory Bandwidth (GB/sec)	144 GB/s

The NVIDIA Tesla C2075 companion processor [3] is built for GPU computing. It features 448 application-acceleration cores per board, dramatically increasing performance compared to a traditional workstation. By adding a Tesla companion processor, engineers, designers, and content creation professionals can add over one Teraflop of computing potential to their workstation.
So Tesla is not like Quadro, it is compute processing intensive, Tesla is still a GPU and the cores are being used exclusively for general computing purposes to offload work from the CPU while the Quadro half of the equation handles graphical tasks.

http://www.nvidia.com/object/workstation-solutions-tesla.html#

4. Requirements and Set up

4.1 Configuration of the test system Model

Our test system model's information for this project is shown below:

Motherboard: Asus P9X79
http://www.asus.com/Motherboards/Intel_Socket_2011/P9X79/#overview

CPU: Intel i7-3930K LGA2011
http://ark.intel.com/products/63697/Intel-Core-i7-3930K-Processor-%2812M-Cache-up-to-3_80-GHz%29

Memory RAM: 8GB ~16GB DDR3-1333 Quad Channel

Graphics Card: Quadro 2000, Tesla C2075, or Maximus(Q+T)

HDD: Single SATA with two different models

1. WESTERN DIGITAL WD5000AAKX Caviar Blue 500GB 7200 RPM 16MB cache SATA 6.0Gb/s 3.5" internal hard drive
2. Seagate Barracuda 7200.7 ST3120827AS 120GB 7200 RPM 8MB Cache SATA 1.5Gb/s

For a comparison purpose, we will compare the results from previous results done by Joseph, Joseph used to be one of the staff that working for the Compucon Company. He have done similar benchmarks for Quadro 2000 by using various of different systems configuration.

4.2 Benchmarking applications

This first half of the project is mainly on benchmarking. Hence we will use some of the benchmarking tools that are optimized and certified for CUDA graphics cards. Most of them are COTS(commercial off the shelf) which mean they are not open sources and pay to use. The following are the applications that we will be benchmark:

1. SPECviewperf 11: http://www.spec.org/
2. Cadalyst c2012 AutoCAD: http://www.cadalyst.com/benchmark-test
3. PPBM5 test for Adobe Premiere Pro CS5.5: http://ppbm5.com/

More details will be discussed when each comes to the real testing.

5. SPECviewperf 11

The first phase of testing was done using SPECviewperf 11 from the Standard Performance Evaluation Corporation. SPECviewperf is a benchmarking application that uses viewsets from various CAD applications such as Autodesk Maya, SolidWorks and Siemans NX to simulate daily CAD usage.

Since testing began I have found that SPECviewperf is designed to isolate the graphics subsystem and is only reliable for comparing graphics cards and not other system components such as CPU and memory. I have since found other applications to test overall system performance for CAD work (see next section). SPECviewperf is still useful for comparing different graphics cards, the following configurations were tested and the results are below:

- Xeon x5506 / 6GB DDR3-1333 / Quadro 600
- Xeon x5506 / 6GB DDR3-1333 / Quadro 2000
- Xeon x5506 / 6GB DDR3-1333 / GeForce GTS 450
- Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000 (our system)

The GTS 450 card was used as a comparison as it is the closest specced desktop card to the Quadro 2000 (they have the same number of active CUDA cores and the same amount of memory). Results are also included for the Quadro 600 card: this is the cheapest Quadro Fermi card available, it has 96 CUDA cores and 1GB memory. Testing was done at **1280x960 resolution** with **8x multi-sampling** enabled in the benchmark application.

	X/6GB Quadro 600	**X**/6GB Quadro 2000	**X**/6GB Geforce
CATIA	9.26	14.13	3.19
EnSight	8.18	12.72	15.90
LightWave	23.94	23.15	5.78
Maya	20.53	26.07	2.90
Pro/ENGINEER	5.11	5.16	0.89
SolidWorks	20.35	25.41	5.53
Siemens TCVis	8.98	11.80	0.61
Siemens NX	7.52	10.03	2.35

As expected, the type of graphics card used greatly affected the benchmark score. The Quadro 2000 card received benchmarking scores up to almost 20x better than the GeForce card. There is, however, an outlier in this situation: EnSight gained a performance increase when using the GeForce card; from this we can conclude that it is not optimized for use with the Quadro model card and relies on raw performance which the GeForce has more of.

As a new configuration is added: Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000
So a new Testing was done again at 1280x1024 resolution with 8x multi-sampling by HaoYu, another testing was done at 1024x768 resolution with 8x multi-sampling by Celestino.

The following things need to be aware:
Different resolution and muli-sampling was using since the SPECviewperf 11 that we installed did not include 1280x968 resolution which Joseph has done previously. I decided to choose 1280x1024 resolution that it is the closest value.
Different configuration was used for this test, a better CPU core and 2GB memory higher than before.
From the previous test, Joseph has mentioned that he have found SPECviewperf is designed to isolate the graphics subsystem and is only reliable for comparing graphics cards and not other system components such as CPU and memory. If he is right, the new configuration of Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000 compares with Xeon x5506 / 6GB DDR3-1333 / Quadro 2000 should come out with a similar score, else we should expect a higher score.

The result is shown below:

	1280x1024 8x multi-sampling	**1024x768 8x multi-sampling**	**1280x968 8x multi-sampling**
CATIA	15.29	14.48	14.13
EnSight	19.09	20.21	12.72
LightWave	21.79	21.42	23.15
Maya	15.57	15.21	26.07
Pro/ENGINEER	4.58	4.53	5.16
SolidWorks	22.97	23.03	25.41
Siemens TCVis	8.05	7.7	11.80
Siemens NX	24.32	24.45	10.03
Configuration	Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000	Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000	Xeon x5506 / 6GB DDR3-1333 / Quadro 2000

A interesting result has came out, The color marked in green represents a higher score than before, the color marked in purple means a lower score. A significant decreasing performance score for "Maya", it dropped from 26.07 to 15.57, which it doesn't make sense.
As we all known that SPECviewperf 11 is a 3rd party benchmarking software, this will lead to us some bias in some situations. Bias is always a concern with testing.

In this case, we have concluded our first hypothesis of bias that it might causes this result, since the new configuration was using a brand new mother board Asus P9X79 with the newest CPU chipset LGA 2011, these hardware are only released about few months ago, the SPECviewperf 11 may has not updated to the newest version that support our hardware.

6. Cadalyst c2012 AutoCAD

The next phase of testing was done using the Cadalyst benchmark test for AutoCAD 2011. This is not an independent application like SPECviewperf and is instead run from inside a fully installed version of AutoCAD. This way it is mimicking actual CAD usage in a proper CAD application and should give us the most consistent and realistic benchmark we can hope for. A 30 day trial of AutoCAD 2011 was used as it is compatible with this test.

There have been some instabilities experienced when running this benchmark, this would be due to the fact that it is a 3rd party benchmarking test. The first group of tests was done using the base Quadro driver supported by AutoCAD; the second group was done with the additional AutoCAD performance driver by NVIDIA installed and there was a large increase in 3D rendering performance. The following systems were tested:

	DXA Workstation 1366	Superhawk Plus 1366	Superhawk 1155	Superhawk 1155	Our system for this project
Motherboard	Supermicro X8DAi	Asus P6X58D-E	Asus P8P67 LE	Asus P8P67 LE	Asus P9X79
CPU	Xeon X5680, 6C/12T,12MB, 3.33GHz	i7-950, 4C/8T, 8MB, 3.06GHz	i3-2100, 2C/4T, 3MB, 3.10GHz	i7-2600K, 2C/4T, 3MB, 3.10GHz	i7-3930K, 6C/12T, 12MB, 3.20GHz
Memory	6GB DDR3-1333 Triple Channel	6GB DDR3-1333 Triple Channel	4GB DDR3-1333 Dual Channel	4GB DDR3-1333 Dual Channel	8GB DDR3-1333 Quad Channel
Graphics Card	Quadro 2000	Quadro 2000	Quadro 2000	Quadro 2000	Quadro 2000
HDD	Single SATA 3Gbps	Single SATA 3Gbps	Single SATA 3Gbps	Single SATA 3Gbps	Single SATA 6Gbps

The majority of the price different comes from the CPU and motherboard used. Do not use these to calculate the price increase (e.g. 300% cost increase) as the cost of parts such as the PSU, chassis and more expensive storage solutions aren't factored in and these would change the final cost ratio.

The tests were performed on newly installed operating systems using the system configuration recommended by the Cadalyst benchmark as well as installing a tweak to remove the info centre from AutoCAD (this sub-application appears to have been adding to the instability during testing). The testing was performed at 1024x768 instead of the suggested resolution of 1280x1024 because the higher resolution was also causing instability during testing.

	i7-950 / Base	x5680 / Base	i3-2100 / Base	i7-2600K / Base	i7-950 / Perf	i3-2100 / Perf	i7-2600K / Perf
3D	762	799	842	992	4133	4140	4408
2D	305	334	298	383	300	300	377
CPU	226	244	229	289	224	232	285
HDD	147	137	145	145	144	146	144
Total Score	**360**	**378**	**379**	**452**	**1201**	**1205**	**1304**
Total Time (Mins)	18	17	17	14	14	13	11

I have made a new table for comparing previous tests done by Joseph with my currently tests:
Testing was done using the Cadalyst benchmark test for AutoCAD 2012 instead of AuctoCAD 2011.
The first group of tests were done using the base Quadro driver supported by AutoCAD where we called it a "Base" testing; the second group called "Performance" testing was done with the additional AutoCAD performance driver by Nvidia.
PS: the result is tested by two different Driver versions, for instance: (8.17.12.9573) vs (8.17.12.6570)

	i7-3930K / Base	i7-3930K / Base	i7-3930K / Perf
3D	1149	983	will be test when 10th of April
2D	409	417	
CPU	294	288	
HDD	175	159	
Total Score	**507**	**462**	
Total Time (Mins)	11	12	
Driver version	8.17.12.9573	8.17.12.6570	

After the first group of tests were done by using the base Quadro driver, then lately we found out that testing with performance drive is unable if we continue using AutoCAD 2012. Since Nvidia currently does not support performance drive for AutoCAD 2012. The link is shown here:
http://www.nvidia.com/object/AutoCAD_PD_workstation.html

We suspect that Nvidia has hidden the drivers for AutoCAD 2012 this time. When Joseph found the drivers for AutoCAD 2011, Joseph did note that the driver was hidden somewhere on the Nvidia website.

Similar scores where gained despite the varying system components, with the least expensive system performing fractionally better (due to the new architecture). The performance driver results in a 3D performance increase of over 400%. Please note, however, that the performance driver has the following limitation with AutoCAD 2011:

- The "Advanced Material Effects" option introduced with AutoCAD 2011 is not currently supported by the NVIDIA AutoCAD Performance Driver. The setting controlling this graphics mode (in the Manual Performance Tuning dialog accessed by the GraphicsConfig command) is greyed-out when the Performance Driver is active.
- Procedural Materials and Maps introduced with AutoCAD 2011 will only display with the material's diffuse colour.
- Materials and Maps used in drawings coming from earlier AutoCAD versions are supported as they would have displayed in AutoCAD 2010.

7. Premiere Pro Benchmark 5 (PPBM5)

The next phase of testing was done using the PPBM5 benchmark test for Adobe Premiere Pro CS5.5. Since our goal is to test out Maximus functionality (Quadro combines Tesla) that whether it can provide a cost effective configuration, and our two previous tests done by both SPECviewperf 11 and AutoCAD 2012 do not have the feature to support Maximus technology [5].

Complete directions are included in the ZIP file downloaded from the website. Create a directory called PPBM on your Premiere project disk and download the PPBM5 file and unzip it in that directory. The zip file also includes a timing/information gathering script which writes the Output.txt file.

There are four tests in PPBM5.

1 Render the Timeline to create Preview files (Pressing Enter).

This test may have to be done twice, once with Hardware MPE acceleration and once with Software MPE only.

2 Export the Timeline with Abode Media Encoder to a MPEG2 DVD file.
3 Export the Timeline with Adobe Media Encoder to a H.264 file.
4 Export the Timeline with Adobe Media Encoder to a Microsoft DV AVI file.

DISK I/O test:
The overriding factor is disk speed here. The test uses many small reads and a large sequential write (nearly 13 GB). Number of cores makes no real difference (it is not well multithreaded), but clock speed does.

MPEG2 DVD test:
The two overriding factors here are amount of memory and number of cores. More is better here. Additionally the location and speed of the pagefile can be important especially if you have a small amount of RAM.

H.264 test:
Here the speed of CPU/RAM communication is king. Number of cores, clock speed and the amount of CPU cache are very important. Dual processor systems are hampered by the 2 chip communication.

CPU / GPU Test Result:
This is almost solely based on the video card and whether hardware or software MPE is used.

MPE Gain:
This shows how much faster hardware MPE rendering is than software only rendering. The minimum score is of course 1, since if there is no hardware MPE available, there is no performance gain.

Total Time:
The Total Time is the sum total of the individual test scores, where each test score is calculated by seconds, so the lower the score the better.
Below are the results for benchmarking PPBM5 by using two different configurations:

1. Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000 only
2. Intel i7-3930K LGA2011 / 8GB DDR3-1333 / Quadro 2000 + Tesla C2075 (Maximus technology)

The two tables shown below are the time taken for running the test, it is basically the score for each test.

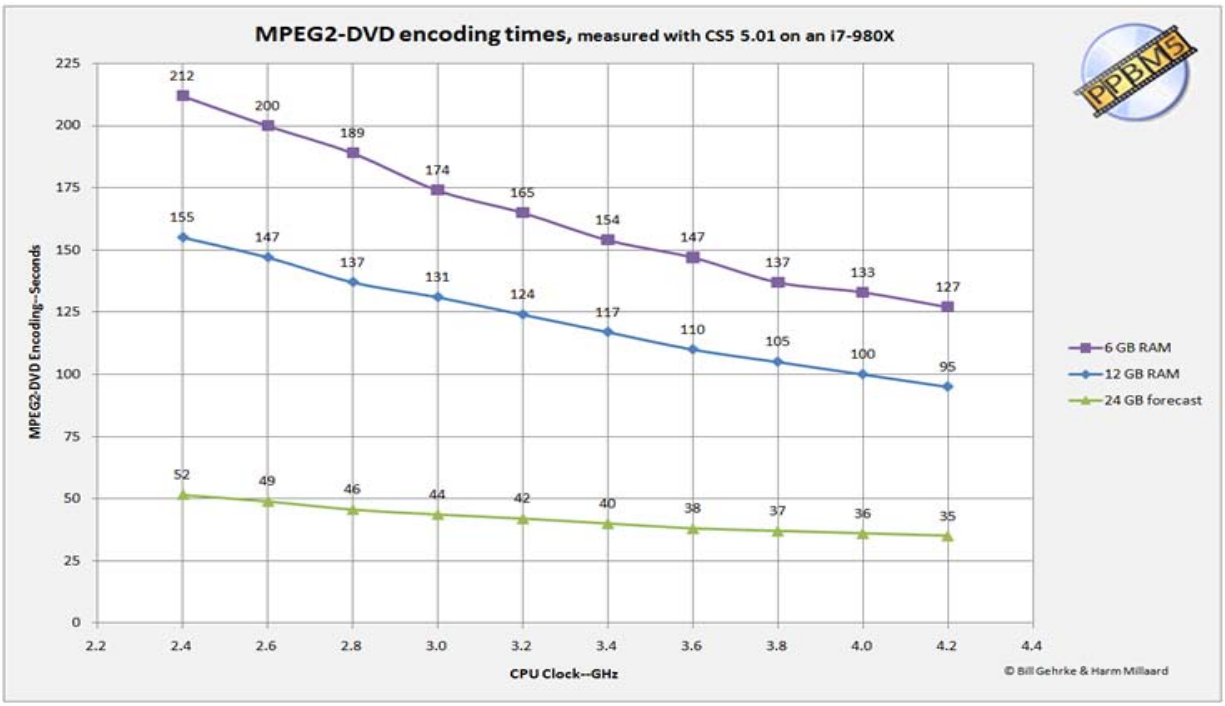
Quadro 2000 only	Format	Preset	Encoding time
H.264 test	H.264 Blu-ray	Custom	1.13 sec
Disk test	Microsoft AVI	PAL DV	5.44 sec
MPEG2-DVD test	MPEG2-DVD	Custom	3.46 sec

Quadro 2000 + Tesla C2075	Format	Preset	Encoding time
H.264 test	H.264 Blu-ray	Custom	1.03 sec
Disk test	Microsoft AVI	PAL DV	5.26 sec
MPEG2-DVD test	MPEG2-DVD	Custom	2.26 sec

Now these are the scores for each test from Output.txt file: This test has to be done twice, once with Hardware MPE acceleration (MPE-On) and once with Software MPE only (MPE-Off). Hardware acceleration is useful with rendering, previewing, and on certain parts of the export process, i.e. scaling, frame rate adjustments, blurring and blending, but not with encoding.

Mercury Playback Engine GPU Acceleration	Quadro 2000	Quadro 2000 + Tesla C2075	Mercury Playback Engine Software Only	Quadro 2000	Quadro 2000 + Tesla C2075
Disk I/O	345	359	Disk I/O	345	359
MPEG2-DVD	226	140	MPEG2-DVD	226	140
H.264	73	62	H.264	73	62
MPE-On	10	8	MPE-off	99	96

Our result for Quadro + Tesla(MPE-ON) ----> Total scores = 359+140+62+8 = 569
it is so high because of the HDD(Disk IO score) which we are using an old one from TSD. Our MPEG2-DVD (we got 140) is also quite high which I do not quite understand at the beginning. But after I did some research on the website, I found out that the score is affected by the amount of memory. See the diagram I attached below:



Rank	Motherboard	CPU Model	Cores	# RAM	Version	Video card	Total	Disk I/O	MPEG2 DVD	H.264 BR	MPE On	MPE Off	OS Disk	# HDD's
1	SR-2	Xeon X5680	12	48	5.03	GTX 580	131	67	21	37	6	64	SSD	6
2	P9-X79PRO	i7-3930K	6	32	5.04	GTX 570	133	71	22	35	5	56	SSD	2
3	GA-X58A-UD7 rev2	i7-970	6	24	5.04	GTX 480	134	59	18	53	4	67	SSD	16
4	GA-UD5-X79	i7-3930K	6	32	5.04	GTX 590	134	66	23	39	6	66	SSD	5
5	P6T WS Supercomputer	i7-980X	6	24	5.03	GTX 580	139	58	21	55	5	65	10000	9
6	X79-UD5	i7-3930K	6	32	5.03	GTX 570	142	75	24	38	5	60	SSD	5
7	Saberthooth X79	i7-3930K	6	32	5.5	GTX 570	144	46	58	36	4	49	10000	3
8	P8P67 EVO	i7-2600K	4	16	5.03	GTX 470	145	56	28	57	4	66	SSD	4
9	X58 Xtremen	i7-950	4	24	5.03	GTX 460	146	75	24	42	5	79	7200	4
10	X79 UD5	i7-3930K	6	32	5.5	GTX 580	146	48	57	37	4	55	SSD	4
11	X79	i7-3960X	6	32	5.5	GTX 580	146	47	58	36	5	49	7200	3
12	P8P67 EVO	i7-2600K	4	16	5.03	GTX 460	148	56	29	58	5	65	SSD	6
13	GA-X58A UDR3	i7-980X	6	24	5.5	GTX 680	149	51	55	39	4	61	SSD	5
14	SR-2	Xeon X5650	12	24	5.03	GTX 470	150	69	23	52	6	74	7200	6
15	X79 Extreme9	i7-3930K	6	64	5.5	GTX 480	150	46	62	38	4	52	SSD	16
16	SR-2	Xeon E5660	12	48	5.5	GTX 580	152	52	58	37	5	56	SSD	8
17	P6T WS Supercomputer	i7-980X	6	24	5.03	GTX 580	153	67	22	59	5	71	15000	7
18	Z67 Exreme4 Gen3	i7-2600K	4	16	5.03	GTX 570	153	65	27	56	5	74	SSD	5
19	P6X58D-E (bios 602)	i7-980X	6	24	5.03	GTX 480	154	68	25	56	5	71	SSD	7
20	Rampage III Extreme	i7-970	6	24	5.03	GTX 570	155	64	27	59	5	69	10000	9

Resource from: http://ppbm5.com/DB-PPBM5-1.php

As we can see from the top 20 results, they all had a very high amount of RAM, especially for the ranking No.1, it has number of 48 RAM for the configuration. For our set up, we had only 8GB RAM.

And for the GPU card GTX580, it has 512 CUDA cores comparing with our Quadro 2000 that contains only 192 CUDA cores. If we look further for Maximus, Quadro 2000 combines with Tesla C2075 with 448 CUDA cores. We should expect a high improvement result by using Maximus functionality. But in fact, our result comes out with not much improve overall even though our H.264 score and MPE-On score are actually very close to the top 20 result.

So after we have compared our results with the Top 20 scores, we decided to increase our 8GB RAM to 16GB RAM and test again by using exactly the same configuration. Hopefully, we should expect a better result related to MPEG2-DVD score.

Mercury Playback Engine GPU Acceleration	Quadro 2000	Quadro 2000 + Tesla C2075	Quadro 2000	Mercury Playback Engine Software Only	Quadro 2000	Quadro 2000 + Tesla C2075	Quadro 2000
Disk I/O	345	359	370	Disk I/O	345	359	370
MPEG2-DVD	226	140	234	MPEG2-DVD	226	140	234
H.264	73	62	75	H.264	73	62	75
MPE-On	10	8	9	MPE-off	99	96	96
Memory RAM used	8GB	8GB	16GB	Memory RAM used	8GB	8GB	16GB

**Important Note:
This test was still using the old/same Hard disk drive Model: Seagate Barracuda 7200.7 ST3120827AS 120GB 7200 RPM 8MB Cache SATA 1.5Gb/s 3.5" Hard Drive
Why these needs to be pay a special attention? I will talk about this later after we had another tests on a New Hard disk drive Model:

Now, let's check the table, it shows the fact that no matter how we changed the Memory RAM does not give any improvement to our scores. Hence, we did not continue on testing with Quadro2000 + TeslaC2075, it is meaningless for doing it since RAM does not apply better result. So we had one conclusion by now:
1. Even though Adobe PPBM5 explained/proved that higher Memory RAM will increase performance, it does not work for our configuration with Quadro2000 graphics card.

We don't think we can expect much from them given that they are third-party solutions. After we contacted with Joseph, he then gave us some suggestions:

One possible solution to the issue would be to bypass the benchmarks and test the performance manually. Possibly you could find a few common tasks with the application you wish to benchmark (rendering a model in AutoCAD for example) and them manually running the task and timing how long it takes to finish. This way you get a 'real world' performance result (i.e. instead of saying "it received a benchmark score of 398" which is abstract, you could say "it rendered the model 20 seconds faster" or "it took half as long to encode a 20 minute video". The problem with that I guess would be that you wouldn't be able to compare it with scores online and would have to run the tests on older system setups to compare performance differences.

Next step we are trying to improve the performance for Disk I/O, as i mentioned previously about OLD hard disk, now we have changed to new Model:
- Western Digital Caviar Blue WD5000AAKX 500GB 7200 RPM 16MB Cache SATA 6.0 GB/s 3.5" Internal Hard Drive

Mercury Playback Engine GPU Acceleration	Quadro 2000	Quadro 2000 + Tesla C2075	Quadro 2000	Quadro 2000	Mercury Playback Engine Software Only	Quadro 2000	Quadro 2000 + Tesla C2075	Quadro 2000	Quadro 2000
Disk I/O	345	359	133	370	Disk I/O	345	359	133	370
MPEG2-DVD	226	140	217	234	MPEG2-DVD	226	140	217	234
H.264	73	62	76	75	H.264	73	62	76	75
MPE-On	10	8	8	9	MPE-off	99	96	80	96
Memory RAM used	8GB	8GB	8GB	16GB	Memory RAM used	8GB	8GB	8GB	16GB
Hard Disk	Old	Old	New	Old	Hard Disk	Old	Old	New	Old

This time successfully improves score of Disk I/O as expected, reducing the time from 340-360 sec to 133 sec, which leads to our second conclusion:
2. Disk I/O is working perfectly fine with our configuration, better the Hard disk, better the result.

The next phase of testing was done using Tesla C2075 alone, Quadro 600 alone and Maximus(TeslaC2075 + Quadro 600) with different range of RAM.
PS: Both Tesla C2075 and Quadro 600 do not support Mercury Playback engine. However, Maximus enables MPE on.
Explanation of MPE from Adobe website:
'Mercury Playback Engine' is a name for a large number of performance improvements in any version above Premiere Pro CS5. Those improvements include the following:
- 64-bit application
- Multi-threaded application
- processing of some things using CUDA

Everyone who has Premiere Pro CS5 has the first two of these. Only the third one depends on having a specific graphics card.
Confusingly---because of one of our own early videos that was just plain unclear---a lot of people think that 'Mercury' just refers to CUDA processing. This is wrong. To see that this was not the original intent, you need look no further than the project settings UI strings 'Mercury Playback Engine GPU Acceleration' and 'Mercury Playback Engine Software Only', which would make no sense if 'Mercury' meant "hardware" (i.e., CUDA).

The official and up-to-date list of the cards that provide the CUDA processing features is here:
http://www.adobe.com/products/premiere/systemreqs/

Mercury Playback Engine GPU Acceleration	Quadro 2000	Quadro 2000 + Tesla C2075	Quadro 2000	Quadro 2000	Quadro 2000	Quadro 600 + Tesla C2075
Disk I/O	345	359	133	370	138	120
MPEG2-DVD	226	140	217	234	217	126
H.264	73	62	76	75	73	62
MPE-On	10	8	8	9	9	7
Memory RAM used	8GB	8GB	8GB	16GB	16GB	16GB
Hard Disk	Old	Old	New	Old	New	New

As long as MPE is on, the result always comes out around 10 sec, which it is a good score.
With the new Hard disk, our 1st conclusion still apply to the Quadro2000 card, RAM does not improve at all, in real, it should not be the case, because Memory RAM does affect both TeslaC2075 and Quadro600 as we can see the following table:
Note: The following table is MPE-OFF, since TeslaC2075 and Quadro600 do not respond to MPE.

Mercury Playback Engine Software Only	Quadro 2000	Tesla C2075	Tesla C2075	Quadro 600	Quadro 600 + Tesla C2075
Disk I/O	133	130	124	126	120
MPEG2-DVD	217	105	54	41	126
H.264	76	98	97	96	62
MPE-off	80	78	77	81	80
Memory RAM used	8GB	8GB	16GB	16GB	16GB
Hard Disk	New	New	New	New	New

TeslaC2075 with 16GB RAM reduced the time from 105sec to 54 sec. Same effect apply to Quadro 600 even though we did not test Quadro 600 with 8GB RAM, but the data shows a fact that 16GB RAM working perfectly fine with Quadro 600, and leads to a result better than TeslaC2075(41sec < 54 sec).
This draws us to a deep thinking of why will Quadro 2000 perform lower score regarding to Quadro 600 if we do not consider the special feature "MPE-On/Off. This could be the reasons where graphics card's driver version or 3rd part solution.
Recently, Adobe official announces Adobe Premiere Pro CS6 is released, this could be more reliable and accuracy for testing Quadro2000 and Maximus functionality.

A Quadro 6000 and Tesla C2075 are not identical but they are very similar and you can expect similar performance. There are a few reasons you might want to use a Maximus configuration for Premiere Pro rather than a single Quadro 6000:
1. Having both a Quadro and Tesla GPU in the system means when the Tesla is cranking full-out on Mercury Playback Engine the Quadro is unaffected, so you can, say, open After Effects or other application that may take advantage of the Quadro, and system performance on that app will be better than if it was competing for resources with MPE on a single GPU.
2. In the future, we expect many users will want to run an animation application (using the Quadro) and a simulation application (on the Tesla) at the same time to provide animators with a level of interactivity they don't have without Maximus technology. Example video is here. (http://youtu.be/_LagqqsVO28)
3. It costs less. A typical Maximus configuration has a mid-range Quadro (e.g. a Quadro 2000) and a Tesla C2075, which in that instance costs hundreds of dollars less than a single Quadro 6000 and offers similar performance plus the workflow advantage listed above.Of course, some users may want to run a Quadro 6000 and a Tesla C2075 and get maximum performance, but others can actually get the best MPE acceleration for less money with Maximus technology. (Resource from Adobe forums)

Maximus is a technology that essentially marries a graphics-intensive Quadro card with a Tesla card, which is all compute, inside a workstation to meet that challenge. There's also a software stack at the driver level that allocates the code within any application you're using to CUDA, routing it over to Tesla to handle the compute processing and the Quadro to handle graphics.

Sum up all the conclusions so far:
1. Even though Adobe PPBM5 explained/proved that higher Memory RAM will increase performance, it does not work for our configuration with Quadro2000 graphics card.
2. Disk I/O is working perfectly fine with our configuration, better the Hard disk, better the result.
3. Maximus feature: (Q2000 + C2075) vs (Q600 + C2075), result came out as Q600 + C2075 is better.

8. Future work and Plan

For several years already, high end graphics processors have been supporting high performance applications. However, programs on these GPUs were limited to the capabilities of the specialized hardware. And now, we know modern GPU are usable as high-speed coprocessors for general purpose computational task, as in 2007, NVIDIA introduced the Compute Unified Device Architecture (CUDA) that combines a new hardware concept(built around just one type of programmable processor) with a new and more flexible programming model.

Furthermore, Semi-Global Matching [4] employs CUDA as the computing tool which it is the main task for this project. Because currently, we have done enough research and benchmarking for CUDA by using COTS. It is time for us to bypass the benchmarks and test the performance manually. Semi-global matching (SGM) is one of the best ways for doing stereo matching in computer vision currently. Stereo vision has been an intensive research area in the last decades. The solutions proposed were originally split into two main categories, local and global methods. Later a third category was introduced to separate some of the algorithms from the global methods. This third category is Semi-global methods which reduced the computational complexity to allow real-time implementation and based on global optimizations.

Most of the COTS tools we have benchmarked were not Open sources that we cannot investigate deeper into the concept for CUDA computing. So for next semester, we would like to have our own experience of applying CUDA to speed up SGM, and the comparison could also be drawn with belief propagation SGM on CUDA.

9. Bibliography

[1] Fermi Architecture for High-Performance Computing | NVIDIA
http://www.nvidia.com/object/fermi-architecture.html

[2] NVIDIA's Next Generation CUDA Compute Architecture
http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

[3] NVIDIA® TESLA™ C2075 COMPANION PROCESSOR CALCULATE RESULTS EXPONENTIALLY FASTER
http://www.nvidia.com/docs/IO/43395/NV-DS-Tesla-C2075.pdf

[4] GPU optimization of the SGM stereo algorithm. Istvan Haller, Sergiu Nedevschihttp
http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5606438&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F5598248%2F5606391%2F05606438.pdf%3Farnumber%3D5606438

[5] Maximus Technology
http://www.nvidia.com/object/maximus.html