Social Implications of Graphics Processing Units

MMX-0601

An Interactive Qualifying Project Report
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science
by

Muhammad Kashif Azeem
Rohit Jagini
Mandela Kiran
Kaushal Shrestha

Date: May 1, 2007

Advisors:

Professor Murali Mani, Advisor
Professor Emmanuel Agu, Co-Advisor

Table of Contents

Cover Page i

Table of Contents ii

List of Figures iv

Abstract vi

Executive Summary vii

Acronyms ix

1. Introduction 1

2. Background Research 3

3. History of GPUs 5

4. Market Research 7

4.1. The Rise of Nvidia 9

4.2. Intel's Entrance in GPU Market 10

4.3. Market Power: Nvidia and ATI 11

4.4. Market Analysis of Nvidia 13

4.5. Market Analysis of ATI (AMD) 14

4.6. Nvidia in Today's Market 16

5. GPU Architecture and Model 18

5.1. Architecture of a CPU 18

5.2. Architecture of a GPU 19

5.3. GPUs and CPUs: Similarities and Differences 21

5.4. GPU Stream Model 23

6. Primary Application of GPUs 27

6.1. Video Games 27

6.2. Virtual Reality Therapy 31

7. General Purpose Applications of GPUs 32

7.1. Scientific Computation in GPUs 32

7.2. Medical Application 33

7.3. Database Operations on GPUs 33

7.4. Bioinformatics 34

7.5. Computer Vision 34
7.6. Applications of GPUs in Protein Folding Simulation 35

8. GPUs in Database Applications 37

8.1. Fast Computation of Database Operations 37
8.1.1. Predicate Evaluation 37
8.1.2. Range Query 38
8.1.3. Multi-Attribute Query and Semi Linear Query 39
8.1.4. Accumulation 41
8.2. Sorting Algorithm using GPUs: GPU Terasort 42
8.3. Hardware Acceleration for Spatial Database Operations 44

9. Medical Applications 51

9.1. Medical Imaging 51
9.2. Visualization in Medicine 54

10. Future of GPU Applications 56

11. GPU: A Disruptive Technology 57

12. Survey on GPU Awareness 59

13. Conclusion 70

Bibliography 72

Appendix A: Glossary 79

Appendix B: Survey Questions 81

Appendix C: GPU Awareness Flyer 82

List of Figures

FIGURE 1: GPU CORES FROM NVIDIA AND ATI 2
FIGURE 2: GPU MARKET SHARES [DAMIEN] 9
FIGURE 3: ARCHITECTURE OF AN INTEL 80386 MICROPROCESSOR [BRAUGH] 19
FIGURE 4: ARCHITECTURE OF ATI X800 [VANBUR] 20
FIGURE 5: GENERAL ARCHITECTURE OF A COMPUTER—THE CPU AND THE GPU [KILGAR]. 21
FIGURE 6: PERFORMANCE COMPARISON IN GIGAFLOPS OF 3.0 GHZ DUAL-CORE P4, ATI R420 GPU, AND NVIDIA'S G70 AGAINST MOORE'S LAW. [GEER] 23
FIGURE 7: BLOCK DIAGRAM OF THE ARCHITECTURE OF A GPU [FIALKA]. 24
FIGURE 8: GRAPHICS PROCESSOR COMPUTATION PIPELINE [LEFOHN]. 25
FIGURE 9: JUSTICE BUREAU STATISTICS RELATING GROWTH OF VIDEO GAMES AND JUVENILE CRIME [THEECO] 29
FIGURE 10: AMERICAN PLAYERS OF VIDEO GAMES BY AGE GROUP [THEECO]. 30
FIGURE 11: MRI SCAN OF A SKULL [LEFOHN]. 35
FIGURE 12: EXECUTION TIME OF A PREDICATE EVALUATION WITH 60% SELECTIVITY BY A CPU-BASED AND A GPU-BASED ALGORITHM [GOVIND] 38
FIGURE 13: EXECUTION TIME OF A RANGE QUERY WITH 60% SELECTIVITY USING A GPU-BASED AND A CPU-BASED ALGORITHM [GOVIND] 39
FIGURE 14: EXECUTION TIME OF A MULTI-ATTRIBUTE QUERY WITH 60% SELECTIVITY FOR EACH ATTRIBUTE AND A COMBINATION OF AND OPERATOR. TIME *I* IS THE TIME TO PERFORM A QUERY WITH *I* ATTRIBUTES.[GOVIND] 40
FIGURE 15: EXECUTION TIME OF A SEMI-LINEAR QUERY USING FOUR ATTRIBUTES OF THE TCP/IP DATABASE [GOVIND]. 41
FIGURE 16: TIME REQUIRED TO SUM THE VALUES OF AN ATTRIBUTE ON THE CPU AND BY THE GPU-BASED ACCUMULATOR ALGORITHM [GOVIND]. 42
FIGURE 17: PERFORMANCE OF GPUTERASORT ON GPUS AND HIGH END CPUS 43
FIGURE 18: ORACLE SPATIAL QUERY MODEL [BANDI] 46
FIGURE 19: QUADTREE INDEX STORAGE FOR *PRISM* AND *HYDRO* (IN LOG SCALE) 48
FIGURE 20: QUADTREE INDEX CREATION TIME FOR *PRISM* AND *HYDRO* (IN LOG SCALE) 49
FIGURE 21: TIMING RESULTS FOR SELECTION OVER *PRISM* 49
FIGURE 22: TIMING RESULTS FOR SELECTION OVER *HYDRO* 50
FIGURE 23: TIMING RESULTS FOR JOIN OF *COUNTY X HYDRO* 50
FIGURE 24: 3D VISUALIZATION OF A HUMAN SPINE [SGI]. 52
FIGURE 25: DETAILED VIEW OF A SKELETON [SGI]. 53
FIGURE 26: DR. GRAZIA MANCINI USES 3D VISUALIZATION ST. ERASMUS MC [REID] 55
FIGURE 27: SURVEY DISTRIBUTION BY GENDER 59
FIGURE 28: SURVEY DISTRIBUTION BY OWNERSHIP OF A PC/LAPTOP 60

FIGURE 29: SURVEY DISTRIBUTION BASED ON OWNERSHIP OF CONSOLES 60
FIGURE 30: SURVEY DISTRIBUTION BASED ON VIDEO GAME CONSOLE OWNERSHIP 61
FIGURE 31: SURVEY DISTRIBUTION BASED ON GAMERS VS. NON-GAMERS 61
FIGURE 32: SURVEY DISTRIBUTION BASED ON GRAPHICS CARD USED IN PCS 62
FIGURE 33: SURVEY DISTRIBUTION BASED ON HAVING PROGRAMMING SKILLS 62
FIGURE 34: SURVEY DISTRIBUTION BASED ON PROGRAMMING LANGUAGES PEOPLE KNEW 63
FIGURE 35: SURVEY DISTRIBUTION BASED ON KNOWLEDGE OF GPU 63
FIGURE 36: SURVEY DISTRIBUTION BASED ON FAVORITE GAMES OF THE PEOPLE SURVEYED 64
FIGURE 37: SURVEY DISTRIBUTION BASED ON INCREASED TIME IN PLAYING GAMES DUE TO IMPROVEMENT IN GRAPHICS 64
FIGURE 38: SURVEY DISTRIBUTION BASED ON PEOPLE INTERESTED IN LEARNING MORE ABOUT GPUS 65
FIGURE 39: SURVEY DISTRIBUTION BASED ON STUDENTS' INTEREST IN THE FUTURE APPLICATIONS OF GPUS 65
FIGURE 40: SURVEY DISTRIBUTION BASED ON KNOWLEDGE ABOUT RENDERING OF GRAPHICS 66
FIGURE 41: SURVEY DISTRIBUTION BASED ON WAYS STUDENTS WANT TO LEARN ABOUT GPUS 66
FIGURE 42: SURVEY DISTRIBUTION BASED ON STUDENTS' MAJORS 67
FIGURE 43: PERCENTAGE OF CS MAJORS WHO KNOW WHAT A GPU IS 67
FIGURE 44: PERCENTAGE OF ECE MAJORS WHO KNOW WHAT A GPU IS 68
FIGURE 45: PERCENTAGE OF STUDENTS FROM OTHER MAJORS WHO KNOW WHAT A GPU IS 68
FIGURE 46: TYPES OF GPUS BEING USED IN CONSOLES 69

Abstract

A Graphics Processing Unit (GPU) is a processor deemed as a complete replacement of CPUs for intensive computational purposes. Due to the constant improvements in speed and number of flops, GPUs have led to many technological breakthroughs in medical research, database enhancement, military computations, etc. With GPUs being used more and more in our day-to-day life for general purpose computations, the role they play on our society have become significant. This project is pertaining to the societal impacts of GPUs.

Executive Summary

One of the major components of today's modern computers are their graphics cards, or to be more specific, the graphic processing units (also commonly known as GPUs). Owing to the growing demand for 3D graphics, GPUs are now being extensively used in many fields. The history of GPUs dates back to the 1970s, and since then the demand for better graphics and the development of graphics cards have together risen exponentially. Out of the several manufacturers of GPUs that exist in the market, Nvidia, ATI and Intel stand out with impressive fractions of the market share in comparison to the other companies. Nvidia and ATI produce GPUs of the best quality due to their reputation and reliability followed by Intel's integrated GPU. While Nvidia and ATI compete neck to neck with each other, Intel has the highest market share solely due to the embedded (integrated) GPUs in their motherboards (Intel chipsets). This phenomenal growth of the GPU market is putting pressure on the CPU market which results in business moves such as AMD's acquisition of ATI.

GPUs are essentially processors which perform complex mathematical computations at a very high speed. GPUs are based on stream architecture, which has led many researchers to consider processing massive volumes of data through high speed and high performance capabilities. GPUs are now being used in various fields to speed up *database applications*, *bioinformatics*, *protein sequence matching*, and *medical imaging* along with their more traditional use in video games and the new Virtual Reality Therapy.

The goal of this Interactive Qualifying Project is to perform preliminary research on the societal impact of GPUs, whether it is regarding the primary applications of the device in the gaming industry or for general purpose computations. We were able to perform in-depth market analysis of the two major players—Nvidia and ATI. The idea of general purpose computations

on GPUs became more prominent to us as we continued to work on this project. After realizing the capability of the ever improving GPUs and the number of applications where it can be used to make the world better than what it is today, we were interested and curious to know if people were aware of this fact. In order to see how the students of our university (Worcester Polytechnic Institute) responded to the idea of GPUs being used for video game consoles and for various general purpose applications, we conducted a survey which provided us with more concrete results.

The survey was conducted keeping in mind our target population which comprised of students and staff of the university. The most interesting insight was the fact that gamers increase the number of hours they play depending on how good the game is and not on how good the graphics are. Also of the people surveyed, only a small group of people who have programming background knew what a GPU was and how graphics are rendered to a PC.

As a supplement to this project, and realizing that the WPI community is not fully aware of the capabilities of GPUs, the group decided to spread awareness and share the knowledge gained during the course of the project with fellow community members by preparing and handing out flyers in the final week of the project.

Acronyms

2D	Two-Dimensional
3D	Three-Dimensional
ALU	Arithmetic and Logic Unit
BLAS	Basic Linear Algebra Subprograms
CAM	Computer Aided Manufacturing
CPU	Central Processing Unit
CT	Computer Tomography
DVD	Digital Versatile Disk
GB	Giga Bits
GPGPU	General Purpose GPU
GPU	Graphics Processing Unit
HPC	High Performance Computing
I/O	Input / Output
LCD	Liquid Crystal Display
MRI	Magnetic Resonance Imaging
PC	Personal Computer
PET	Positron Emission Tomography
PS	Play Station
RAM	Random Access Memory
SIMD	Single Instruction Multiple Data
VCD	Video Compact Disk

1. Introduction

Graphics Processing Units (GPUs), also known as Video Processing Units, are devices mainly used for manipulating and displaying graphics for computers and video game consoles. They have been around in the market for a number of years and are mostly used in modern PCs for graphics rendering. The main goal of this *Interactive Qualifying Project* was to research the social impacts of the Graphics Processing Units in our day to day lives. In order to do so we targeted companies which specialize in manufacturing graphics rendering devices like Nvidia, ATI and Intel. Also we looked into the various applications that have improved since the introduction of this technology. GPUs no longer only play a major role in graphics rendering but are nowadays also being used as co-processors for highly intensive computations and massive calculations. Owing to the high demand for higher computational power, the growth of GPUs has been remarkable. With the GPU market growing rapidly, there is intense competition between the big players in the market like Nvidia, ATI and Intel. We discuss in this report the societal impact of GPUs through their applications and a detailed analysis of its growing market.

The technical definition of a GPU is "a single chip processor with integrated transform, lighting, triangle setup/clipping, and rendering engines that is capable of processing a minimum of 10 million polygons per second" [NVIGPU]. According to the sources, August 31, 1999 marked the introduction of the Graphics Processing Unit for the PC industry [NVIGPU]. It was on this date that six of the top independent graphic accelerator manufacturers, namely Creative Technology, Ltd., ELSA Inc., Guillemot Corporation, Canopus Co. Ltd., ASUSTeK Computer Inc. and Leadtek Research Inc., selected GeForce 256 (TM) graphics processing unit by Nvidia for their add-in-card product lines [BUSWIR].

A GPU is a single-chip processor which is dedicated to processing images and runs in parallel with a CPU. This helps in lifting off burden from the CPU and enabling them to use their processing power for other intensive tasks. GPUs are based on the Single Instruction Multiple Data (SIMD) principle which enables them to perform many more tasks in comparison to that of an older computer [WIKI03]. SIMD refers to the principle in which same operations are performed on multiple records or data elements.

A GPU is not a compulsory part of a computer, but when used, improves graphical processing significantly. Today, GPUs have become very common in PCs and in addition to that they have now started to appear in most of the electronic devices with display, such as game consoles (Xbox, PS2, etc), hand-held Personal Digital Assistants (PDAs), mobile phones, Game Boy, calculators, mp3 players, TV, DVD & VCD players, etc. Two of the world's most popular graphics cores are shown in Figure 1.

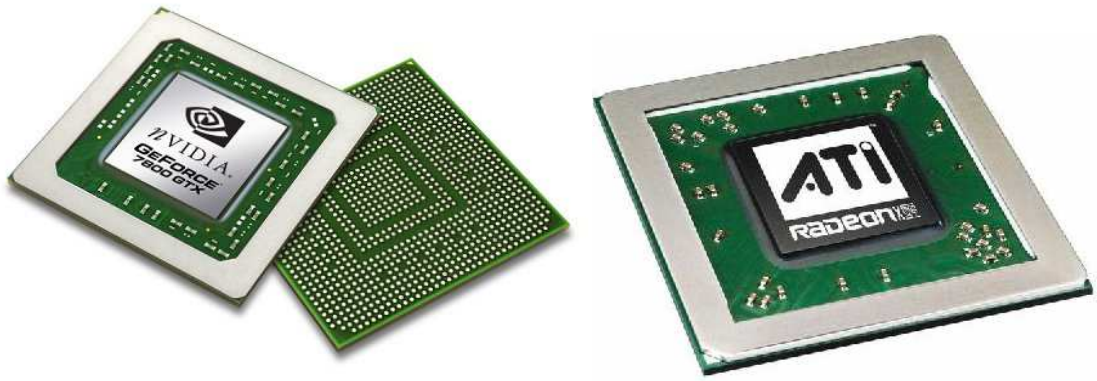

Figure 1: GPU Cores from Nvidia and ATI

2. Background Research

The crisp and clear images that can be seen on the monitors of PCs essentially consist of tiny dots termed as pixels. Nowadays, monitors display over a million pixels on the screen depending on the resolution settings. A CPU after processing data needs to decide on how to create an image on the screen and hence needs a translator which can convert binary data from the CPU into a picture viewable on the screen. Unless there is graphics capability built-in on the motherboard, this task is served by the graphics card or specifically a GPU. Computer graphics such as photos, videos and games require a lot of processing power. To rephrase it in technical terms, it uses a lot of cycles (processing time) of the CPU therefore resulting in slow processing speed. In order to speed up the graphic functions of computers, two-dimensional (2D) accelerators were introduced in the market in the very early 90's. Soon with the introduction of DVDs and VCDs the demand for high resolution and hardware accelerated three-dimensional (3D) graphics grew. Thus, including an entire dedicated processing unit was needed. These units added more execution units making deeper pipelines, wider superscalar architectures and cache.

If a CPU is processing graphics too along with the rest of the operations, a lot of processing speed will be used up in rendering graphics resulting in lower processing speed. With a GPU running in parallel, CPU processing speed for other operations increases by a large factor. GPU's highly developed parallel structure enables it to carry out mathematically intensive tasks using a range of complex algorithms. This capability makes the graphic run faster than drawing the functions from the CPU itself. Its dedicated functionality to compute 3D functions such as lighting effects and 3D motion also helps the CPU run more efficiently. It frees up more cycles for the CPU which can be utilized to run other and more important and intensive applications.

GPUs mainly deal with computations and calculations that are required for rendering graphics and videos only, which makes them more efficient than CPUs in graphic processing. In other words the GPUs carry out only arithmetic calculations compared to the CPUs which carry out arithmetic, control and logistics processing. Also*, dedicated graphic cards* have their own RAM to utilize and hence are more effective than *integrated graphic cards* which utilize the system's RAM (Random Access Memory). The demand for GPUs has since increased great concern about its development in recent years.

In summary the need for higher quality output, high pixel precision and bit depth is driving the requirement for more parallelism and more number of pixel pipelines which the GPU is able to provide. GPUs are used in many fields where certain applications need high speed processing such as medical imaging, military computations, databases and videogames. GPUs have impacted many such fields with their computational power and it is important that their social implications along with their growing market be analyzed.

3. History of GPUs

Early computer graphics dealt with only two-dimensional (2D) graphics, a task that a general purpose CPU (microprocessor) was capable of accomplishing. With the advancement of the microprocessors, in its speed and performance, the demand for the visual output also increased to 3D graphics. Anything that we vision in our everyday life involves 3D graphics, and hence there was a need for creating computer graphics look as real as possible. All these requirements indirectly translate into more geometry and more of crunching numbers. This demand for realism in the graphics added more load to the CPU and to overcome this situation resulted in the concept and design of GPUs for PCs.

GPUs, however, are not a recent device; and have been used extensively since the late 1970's. At that point of time they were used in the Atari consoles. Also, a lot of graphics boards like the TMS340 were used for PC's and workstations to implement drawing functions mainly for CAD purposes. They have progressed decade after decade with IBM launching the Commodore Amiga in 1980's as the first video card to implement 2D graphics as accelerators. This was the first device to be recognized as a full video accelerator which took pressure of the CPU hardware.

In the early 1990's because of the rise in Microsoft Windows, a better graphic processor was required which could process high speed, high resolution 2D bitmapped graphics. Therefore in 1991, S3 Graphics came up with the single-chip 2D accelerator commonly known as S3 86C911. As the internet grew in popularity and several technological advancements such as the VCD and DVD came into being the help of GPUs continued to be sought to improve video acceleration. Also, in the 1990's, 3D graphics became more common in computer and console

games and hence more companies started taking interest in this field of Graphics processing as they saw huge potential in this technology. The Playstation and the Nintendo 64 consoles served as the earliest examples of the use of this technology that were sold in the market.

In the 2000's, GPUs improved further by adding the programmable shading functionality to their capabilities. In this regard Nvidia was the first company to market a chip called GeForce 3, with the program shading functionality. The Nvidia GeForce 3 then received their first competition with the introduction of the ATI Radeon 9700 in October 2002. With the advent of Microsoft Windows Vista, DirectX10 was released and this technological breakthrough should take GPUs to the next higher level [WIKI02].

4. Market Research

Today, many segments of the graphics industry are investigating 3D graphics processing technologies as they do not find the CPU fast enough for their applications. Due to growing demand of 3D graphics, GPUs have now become extremely important. They are extensively used in gaming (both video game consoles and PC), CAM softwares, and flight simulators, which are used by the army for training purposes, and in many other fields. “Many companies have taken interest in the development of GPUs out of which Nvidia, ATI, S3, Intel and Microsoft are the most noticeable” [WIKI01].

GPU market dates back to 1970’s when the first video game consoles with graphic chips such as Atari were introduced to the market which gained immense popularity. Many laser printers by Apple at that time were also shipped with image processors. As the chip technology improved, graphics chips became cheaper and easier to make. In 1980’s, the first mass-production of video cards with ‘blitter’ took place in the form of *Commodore Amigo* computer. This was soon followed by the world’s first video card for PC with 2D primitives by IBM. The demand for GPUs suddenly climbed in 1990 with the release of Windows 3.1, which had a Graphical User Interface (GUI). This version of Windows was the first by Microsoft to achieve a huge commercial success by selling over 2 million copies in the first 6 months [WIKI02].

In 1991, S3 came up with the world’s first single chip GUI accelerator. The company showed a remarkable profit and was the GPU market leader until 1996, when the trend shifted towards 3D graphics with the release of Windows '95 and later Windows '98. S3 released a series of 3D graphics card but they failed in the market due to their poor performance. After struggling for a couple of years the company sold off its core graphics portion to VIA Technologies for

$323 million in 2001. Since then S3 have become a part of VIA motherboard integrated graphics which has helped it to gain 10% share of the overall PC graphics market [SALVAT].

Intel entered GPU market in 1998 with its first graphics chip known as *Intel i740*. It aimed to become one of the major players in the discrete graphics card market but rivalry from companies like Nvidia, ATI and Matrox made the competition fierce. Later, Intel embedded its graphics chip onto its motherboard which became well-known as Intel Extreme Graphics. This way the graphic cards were sold with the motherboards giving Intel firms hold on the GPU market share. In 2002, Extreme Graphics chipsets were replaced with Intel Integrated Graphics with improved functionality, first of which is the GMA 900 and GMA 950 series embedded on 910G and 915G chipset motherboards. Today, Intel has the largest share in the GPU market of around 35% solely because of its high-selling motherboards. Although, the integrated graphics are not good enough to play demanding 3D games as they share memory with the system (motherboard), they are sufficient for business applications like word processing and spreadsheet manipulation. In addition, they are dramatically cheap in comparison to other discrete graphic cards like ATI's Radeon and Nvidia's GeForce which makes it popular for normal home and office use [EXTEDT].

Certain interesting facts are noteworthy from the ever-growing GPU market. Many experts believe that the companies that manufacture GPUs are going to play a key role in computer manufacturing industry. Hence there is a sense of threat and need for the big companies such as AMD and Intel to buy the now growing GPU companies or launch themselves into the GPU market in an impressive fashion. By buying ATI, AMD has definitely

cemented its position in the future race for which company can build the best combination of CPU and GPU,

Intel is a much bigger company than AMD but it is not feasible for them to buy a growing GPU company like AMD did. As of today, Intel has the largest share of 35% in the GPU market, due to the on-board integrated graphics. It is followed by Nvidia with a 24% market share and then ATI with 23%. Intel has been trying to launch itself into the high end GPU market by launching improved graphic cards but has failed so far owing to tough competition. The pie-chart shown in Figure 2 summarizes the market share of various GPU manufacturing companies.

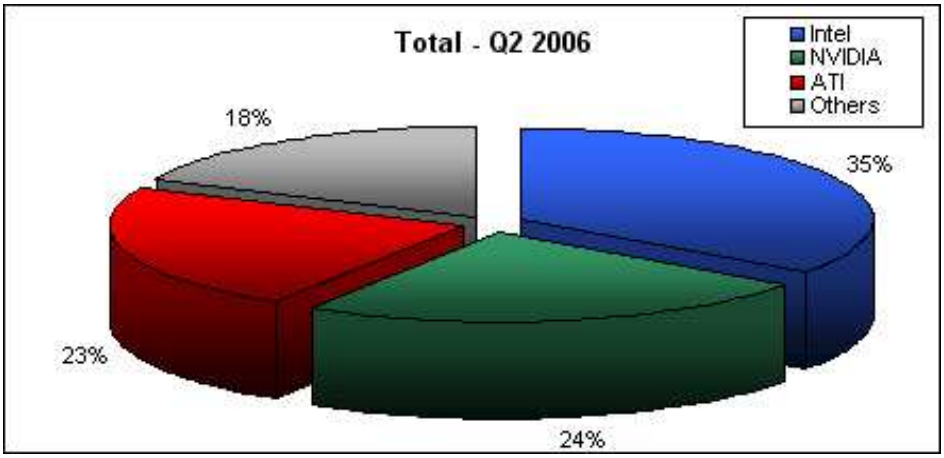

Figure 2: GPU Market Shares [DAMIEN]

4.1. The Rise of Nvidia

Nvidia, initially a tiny company staked a substantial claim in the huge market for videogame players. This was achieved through Nvidia's clever deals with several partners; the main idea is to let the other companies do the lifting. It costs up to about a billion dollars to set

up a semiconductor chip factory and Nvidia has become successful in becoming one of the main players in the GPU manufacturing industry without setting up one.

Huang, the cofounder of Nvidia produced a multimedia accelerator with a mere $7.5 million in venture capital and two dozen young engineers from companies like LSI Logic and Sun Microsystems. They came up with specialized chip designs at its office in Sunnyvale, California and then Nvidia contracted with other firms to have them made from silicon. Nvidia after this big step did not make the game-playing circuit either; Huang arranged for Diamond Multimedia Systems to buy the chips and put them in multimedia accelerator boards. Finally, Nvidia's third deal and the most important of all was the endorsement from Sega, a popular videogame company. Sega translated its traditional console and arcade games to the personal computer format for the first time using Nvidia chips for its highest-quality games. However this had been tried before, the people at Sega were not satisfied with the nonstandard world of PCs. With Nvidia's chip they finally had a multimedia platform that was as good as their own game machines [YOUNG].

In addition to all this, Nvidia has even outsourced its marketing. Its contract with Sega enabled it to launch itself on Sega's name. Nvidia is primarily focused on consumer products market. Because of the seasonality in this market (back-to-school and holiday seasons), stronger revenue performance is expected in the second half of the calendar year [COMTEX].

4.2. Intel’s Entrance in GPU Market

Intel has entered a very competitive and fast growing market. In spite of being a manufacturing giant, it consistently has been fighting for its place in the GPU industry. Intel has

not eclipsed the current market yet; it is competitive in the market because of its integrated graphics chipset and therefore does have an impact on the industry, especially pricing.

Richard Brune, an industry expert predicts that in order to remain competitive in the market, Intel needs a wide range of graphics chip parts. He also states that its strategy is too predictable as it does with its processors. According to him, Intel will probably keep its i740 in the market till it comes out with a product that is at a higher performance level and it will bring down the price of the i740. So a customer like Compaq Computer Corporation will be selling computers with different graphic chip parts at different prices. Intel's competitors will be introducing next-generation parts which will eclipse the performance of Intel's chip [BRUNER].

Intel has certain strengths which play to its advantage. One of these is its exceptionally strong relationship with customers. This factor alone might have been the motivating factor for them to enter the market. There are many who depend on Intel's products and product development roadmap for their own products. This makes them different from other graphics chip vendors.

4.3. Market Power: NVIDIA and ATI

In recent years, there have been clearly two major players competing in the Graphics processors market, Nvidia and ATI. Over the years, Nvidia had been the performance leader with ATI running very close. The company took in $687.5 million for the three months ending July 30, 2006, a massive jump from the $574.8 million it racked up for the same period last year. In 2003, ATI introduced its R300 Graphics processor also known as Radeon 9700 pro which turned

the tables and gave ATI an edge over Nvidia. Since then ATI and Nvidia have been competing head-to-head, introducing their new and improved chips to the market after every few months.

On an average, GPU processing power is doubled every six months. An important reason for this market to expand rapidly is the increase in number of the game enthusiasts. Microsoft's new Xbox 360 has already sold over 6 million units in just about half a year's time since the product was launched, powered by ATI graphics processor. This is an outstanding sales figure and is projected to reach 10 million at the end of 2006 by Microsoft. But this prediction has been proved incorrect by the launch of Sony's PS3 in November 2006, which uses the graphics chip provided by Nvidia.

Graphics chip for Xbox was provided by Nvidia till August 2003, when Microsoft broke the contract with Nvidia. Microsoft wanted to cut down prices for Xbox after they experienced lower sales than expected. This issue resulted in a disagreement and Microsoft dissolved the contract with Nvidia in favor of its rival, ATI Technologies, for its next video game console—the Xbox 360. This was a significant blow to Nvidia Corporation as Xbox business accounted for 21 percent of its $1.9 billion revenue in 2002 [YI]. But the very next year, December of 2004, this blow was subsidized as Nvidia signed a deal with Sony Corporation to be the graphic chips provider for their next video game console, Play Station 3. ATI has been providing graphic chips for Nintendo which has a significant contribution to the gaming market.

Recently, ATI was bought by AMD for approximately $5.4 billion. This gave ATI a better opportunity to fight its competitors in the open market. Also, the leading notebook sellers like Dell, IBM, Sony and Toshiba, have agreed to include ATI graphics chips in their notebooks, which further secures market for ATI. On the other hand, Nvidia acquires Portal Player for $357

million, which was a primary supplier of system-on-chips for Apple's famous iPod music player [DANG]. Also it has recently become Apple's default high performance desktop graphics provider after the introduction of Apple's iMac with GeForce2 MX. Within a short period of time it has become the top to bottom supplier of GPUs for the Macintosh desktops. One of its promising markets is Microsoft Windows Vista, where it provides a secondary LCD interface with its Preface technology. Nvidia has also aimed towards hand-held devices with graphic chips [DANG].

4.4. Market Analysis of Nvidia

Nvidia Corporation is a worldwide leader in programmable graphics processor technologies and is headquartered in Santa Clara, California. It has a total number of 2,737 employees with 1,654 full-time employees only in the Research and Development department. It is primarily based in Asia and the Americas. In recent years the company has shown a remarkable increase in market share. With the successful launch of its GPU series GeForce 6800 and 6600 in 2005, its market share increased from 18% to 67% of the Performance DirectX 9.0-compatible graphics controller segment. Subsequently, with the launch of GeForce 7800 in June 2005 its market share further increased to 79% in its performance. The discrete graphics space accounts for only about 55% of the overall GPU market, with the rest taken over by the Intel integrated chipset graphics [DATAMO].

The company recorded revenue of $2,375.7 million at the end of 2006 fiscal year, which is an increase of 18.5% over that of 2005. The operating profit of the company increased to $340.1 million during 2006 from $113.6 million in 2005, with the net profit being $302.6 million in fiscal year 2006 compared to $100.4 million in 2005. Overall, the company has shown an

outstanding progress in terms of market leadership and is currently a world leader in discrete graphics GPUs [DATAMO].

4.5. Market Analysis of ATI (AMD)

Advanced Micro Devices is the world's second largest manufacturers of processors after Intel, and is based in Sunnyvale, California. In October 2006, it acquired ATI Technologies for $5.4 billion to get an edge over Intel in processors market. But the predictions turned out to be wrong. For the financial quarter ended December 31st, AMD announced a net loss of $574 million which is $1.08 per share. Currently, the graphics sector of AMD, ATI Technologies, alone has 3,469 total numbers of employees [AMDFR].

This loss was in sharp contrast to a net income of $96 million, or 21 cents per share, which the company experienced last year at the same time. Sales for the last financial quarter were $1.77 billion compared to $1.84 billion a year ago. Wall Street had predicted revenues of $1.73 billion which turned out into a loss of $574 million.

The purchase of ATI resulted in a charge of $550 million amounting to $1.04 per share. This was further increased by a $27 million expense over employee stock-based compensation. Even after subtracting the acquisition expenses the operating profit of the company for the quarter was $63 million which was 77% less than what it was a year ago and 56% less compared to the previous quarter. AMD also experienced gross margins of 40% compared to 57% last year at the same time of the year [AMDFR].

AMD said that its server business had been the same as compared to the third quarter of 2006. Since November, AMD is facing a severe price competition from its rival, Intel, which

introduced several new Xeon quad-core processors, has led to some loss in the server business in the last few months of 2006 [AMDFR]. The summary of the finances of AMD have been summarized in Table 1 below.

Table 1: AMD – Finances summary at the end of 2006

Employees	16,500
Market Cap	$7.1 billion
Total Debt	$3.7 billion
Cash Balance	$1.3 billion
Revenue	$5.6 billion
Revenue Growth (1yr)	-3.4%
Revenue Growth (5yr)	7.7%
Operating Margin	-0.8%
Operating Margin Growth (1yr)	4.0%
Operating Margin Growth (5yr)	-4.6%
Net Margin	-2.4%

AMD is expected to launch its first line of processors that combines the x86 processor with graphics chip technology from ATI Technologies. The AMD 690 series of processors includes features from GPU or ATI Radeon x1250, known as 690G, and a lower-priced 690V. These chipsets are seen as first step towards AMD's Fusion processors, which merges x86 and ATI graphics onto one chip. The 690 series is scheduled to come out in 2009 and will initially be available for notebooks, and later for desktops. The new 690 series integrates HDMI (high definition multimedia interface) and DVI output (digital visual interface) further enhancing visual output and giving more display choices to the computers it powers. Also, these chipsets are built to support advanced graphics of Microsoft Windows Vista, which was released in Jan 2006 for home customers. By acquiring ATI, AMD hopes that integrating ATI graphics with its

CPUs would improve sales to the commercial desktops and laptops markets long dominated by Intel [AMDFA].

4.6. Nvidia in Today's Market

Nvidia uses a fabless manufacturing strategy where all phases of the manufacturing process are delegated to different vendors. This system allows the company to avoid any major risks associated with owning and operating manufacturing operations. This allows Nvidia to focus more on its product design, quality assurance, marketing and customer support. This strategy is one of the critical reasons for the company's outstanding success.

The company's focus on R&D has increased in the last few years. In the fiscal year 2006, the number of full time employees engaged in R&D has increased to 1654 from 1231. The company incurred R&D expenditures of 270 million in 2004, 335.1 million in 2005 and 352.1 million in 2006. This is another competitive advantage because of which it enables the company to develop innovative products, which confer a competitive advantage.

At the end of 2006, it had increased its market share in various sub segments of programmable graphics processor technologies in comparison to the years before. The successful production of its GeForce 6800 and GeForce 7800 has increased its market share in the performance segment considerably along with investor's confidence.

The revenue of Nvidia is derived from a very limited number of customers and this aspect of the company's sales strategy reduces the bargaining power of the company. The two largest customers of Nvidia are *Edom Technology* and *Asustek Computer* which account for 14% and 12% of the total revenue respectively. Sales to these two largest customers accounted for

approximately 36%, 31% and 26% of the company's revenue during fiscal years 2004, 2005 and 2006, respectively.

A decrease of 34.5% was observed in the company's consumer electronics business generated revenue from 260 million in 2005 to 170.2 million in 2006. This decrease is because of lower and discontinued sales of Xbox related products. Decrease in revenue from this segment has adversely affected the company's total revenue.

Nvidia faces fierce competition from its prime rival ATI. The discrete graphics space (55% of the GPU market) has intense competition between ATI and Nvidia. Both have very similar market share and size profiles. This makes it difficult for either company to gain a long term edge. Also several competitors have increasingly put pressure on prices, including XGI, S3, and Matrox. In addition, Intel also offers severe competition with its integrated chipset offering basic integrated 3D graphics. Any delay on Nvidia's part to bring products to market on time could impact the company's market share and profitability significantly.

5. GPU Architecture and Model

A GPU in terms of its working is similar to a CPU, and at the same time there exists significant differences between them. The fundamental function of both GPUs and CPUs is data processing; however the only major difference lies in the way data processing is implemented. CPUs are more of general processors that need to handle different types of data and process them in different ways as instructed by the programs. GPUs on the other hand are more specific in terms of data that is to be processed and the instructions written for data processing. By exploring more on CPUs and GPUs by researching into their individual architecture their similarities and differences can be discussed.

5.1. Architecture of a CPU

The CPU generally aggregates a number of hardware components, and is generally referred to the core microprocessor of a computer system. A microprocessor can be divided into different functional blocks. Figure 3 depicts a typical 80386 microprocessor from Intel divided into its different functional blocks.

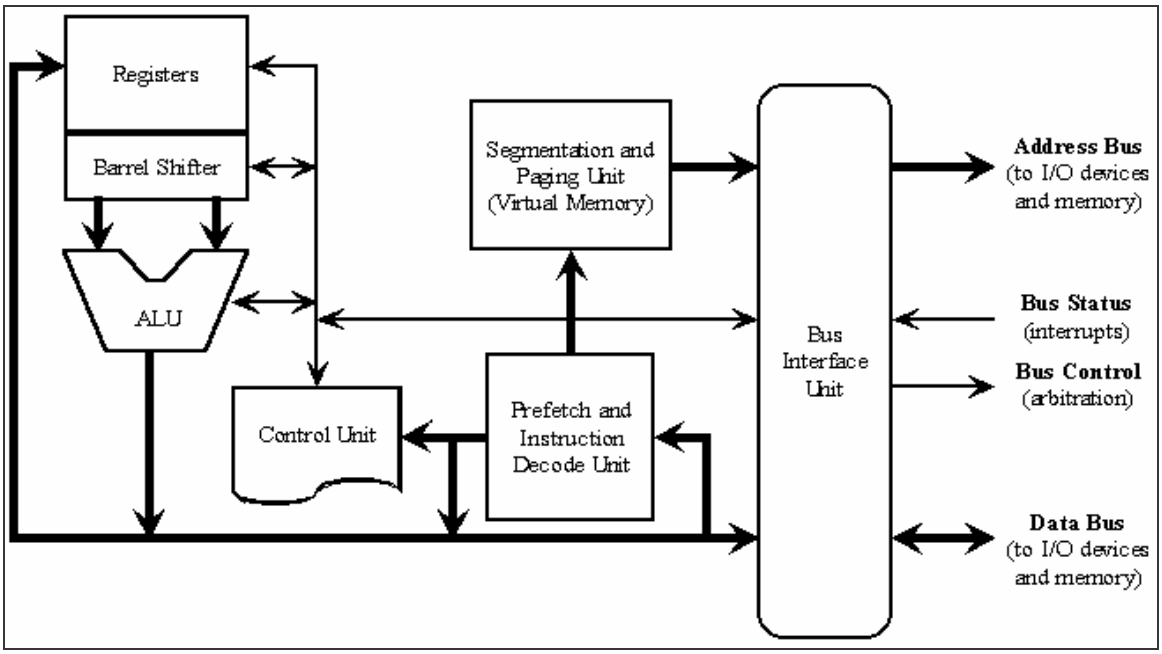

Figure 3: Architecture of an Intel 80386 microprocessor [BRAUGH]

The most important blocks in reference to our discussion are the *Registers*, *Control Unit*, *ALU*, and *Address and Data Buses*. Registers are temporary memory locations used by the processor to store the data before or after processing. The control unit controls the flow of data between the ALU, registers, I/O devices and memory while the ALU behaves as the heart of the computer where all the arithmetic computations and logical decision-making operations are handled. The address lines or address bus controls the I/O devices that are attached to the CPU while the data lines or data bus are responsible for data transfer between the CPU and the I/O devices.

5.2. Architecture of a GPU

GPUs, as mentioned earlier, are microprocessors designed for a more specific task of running number-crunching algorithms on huge volume of data. In short, GPUs are dedicated

processors. The general architecture of a typical GPU can be realized by observing the architecture of the ATI RADEON X800 GPU shown below in Figure 4.

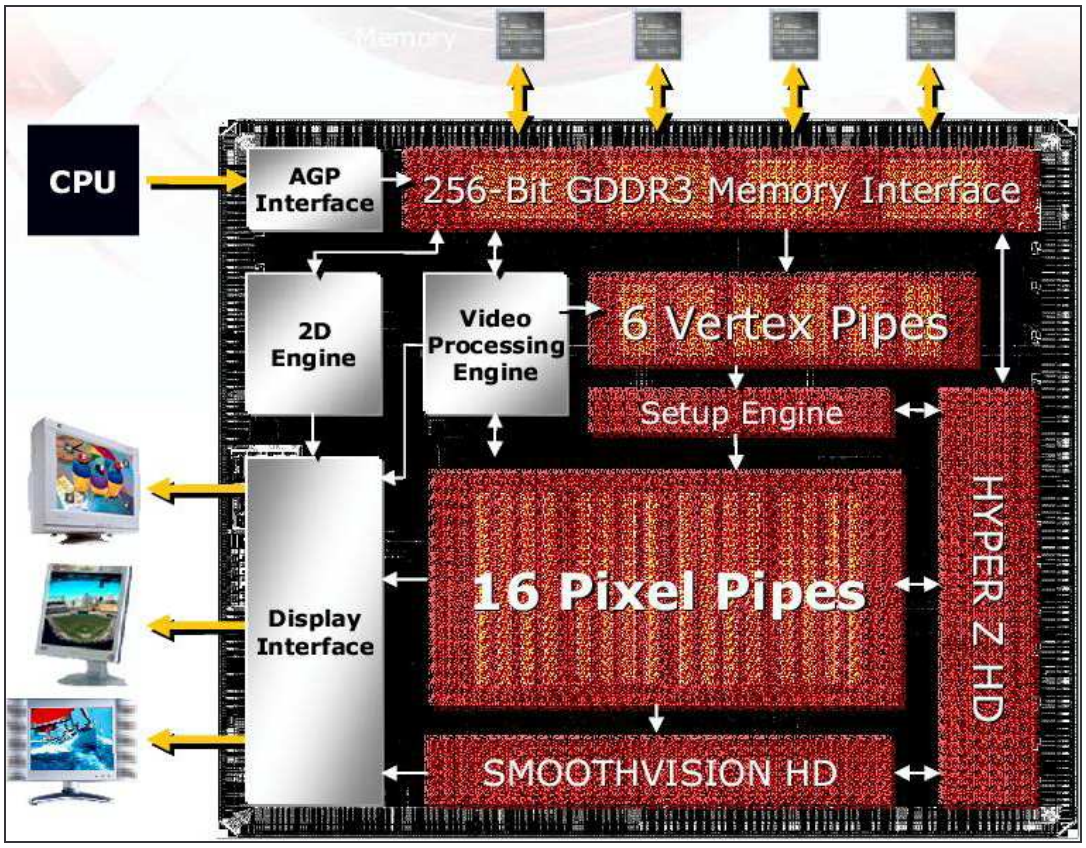

Figure 4: Architecture of ATI X800 [VANBUR]

Like the CPU, even GPUs have I/O interfaces in terms of *Display Interface*, the *Video Processing Engine* is the core processor and is analogous to the *ALU* of a CPU, and the *Memory Interface* of a GPU is not much different from the *Bus Interface Unit* of a CPU.

Figure 5 gives us an idea of how the CPU and GPU of a computer system are linked together with other components like system RAM, graphics RAM, bridges etc.

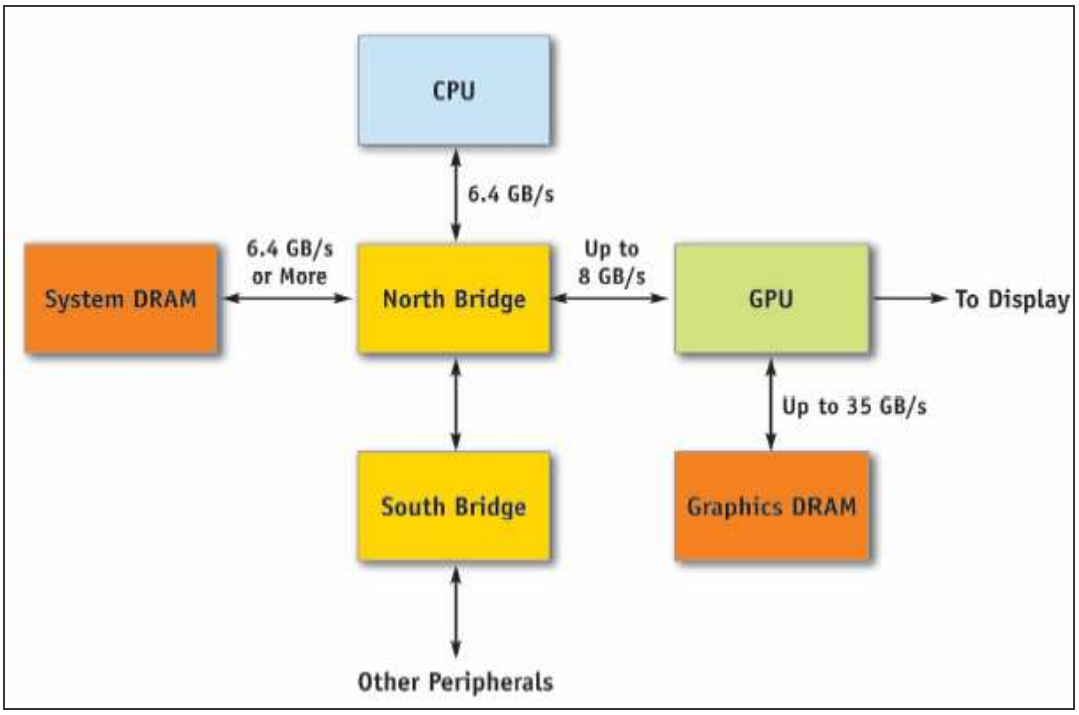

Figure 5: General Architecture of a computer—the CPU and the GPU [KILGAR].

5.3. GPUs and CPUs: Similarities and Differences

Although they are very similar to CPUs in some way, GPUs differ from one another in terms of their architecture and capabilities. Both CPUs and GPUs need memory access for data processing. Thus, as the pattern suggests new versions of GPUs have improved performance by decreasing their memory latency. Both GPU and CPU have limited data transfer speed and bandwidth, and the building blocks of both devices are flops or transistors. GPUs, however, are single-chip processors, that are fully dedicated for handling number crunching algorithms to render realistic computer graphics involving very complex mathematical computations for lighting effects and object transformations. This helps in removing the burden from the CPU and enabling it to use its processing power for other more important tasks. The instruction bus and the data bus of GPUs are independent of each other unlike that of the CPU. GPUs are therefore

considered a parallel computing device in regards to the CPU, which turns out to be a sequential computing device. GPUs implement floating point arithmetic and hence are made up of a number of pixel pipelines.

Today, the new GPUs come in the flavor of very high speed and bandwidth. The GPU Memory Interface has a bandwidth of 35 GB/sec compared to the CPU Memory Interface which has only 6.4 GB/sec bandwidth [KILGAR].

As the CPUs are developing over the last three decades in terms of clock frequency and the transistor manufacturing process (basically the ever decreasing gate width, now in terms of nanometers with the nano-technology), their GPU counterparts have been developing far more rapidly. GPUs have been doubling their performance almost every six months over a period of a decade or so in terms of execution speed. Nvidia, for example, over the period of 5 years have come up with the GeForce family namely GeForce2 (2000), GeForce3 (2001), GeForce4 (2002), GeForce FX (2003) and GeForce6 (2004). This family of graphics processors maintained its clock frequency at 400-500 MHz; however the memory speed increased from 2x230 MHz (GeForce2) to 2x550 MHz (GeForce6) [UJALDO].

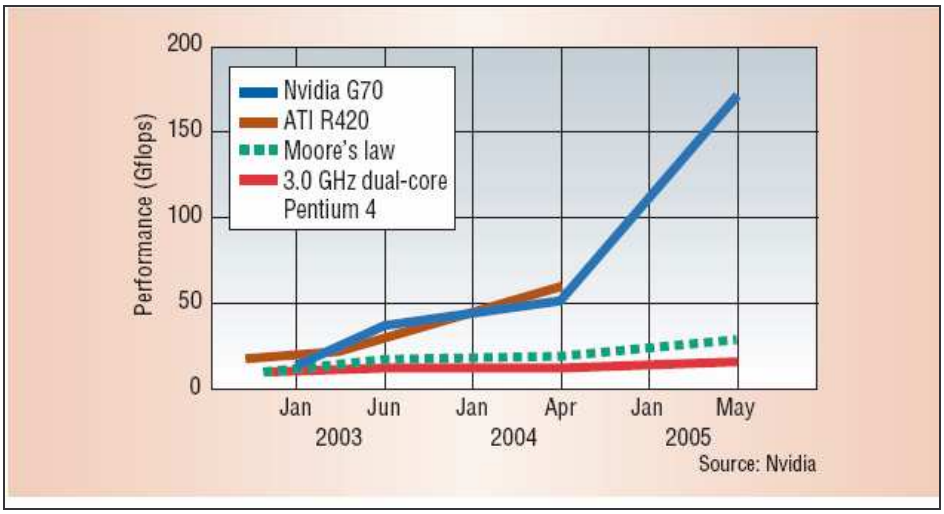

Figure 6: Performance comparison in Gigaflops of 3.0 GHz dual-core P4, ATI R420 GPU, and Nvidia's G70 against Moore's law. [GEER]

5.4. GPU Stream Model

CPUs come under the general hardware category whereas GPUs come under the specialized category as it performs specific functions mainly pertaining to graphics processing and computing operations. Even though these computing operations are very simple, and require very little memory, they require the ability to perform many computations extremely fast and in parallel. Parallel computing is an important concept because it refers to simultaneous execution of one particular task by further sub-dividing it in order to achieve faster results. This whole idea uses the fact that any problem, big or small, can be further divided into smaller tasks and solved easily and faster, with proper coordination. Below is a simple block representation for the graphics-specific stream model that is implemented with some variations in different GPUs by different manufacturers. The GPU has a stream of input data (typically known as vertices), and all of the vertices go through the same process of vertex shading, triangle assembly, culling, rasterization, pixel shading before dumped into the frame buffer and applying the textures to give an image as an output.

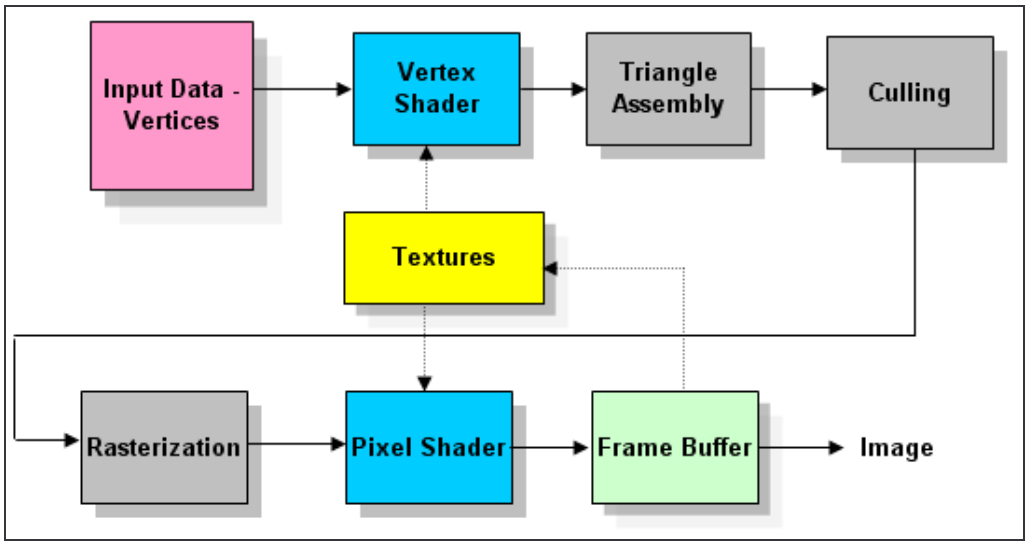

Figure 7: Block diagram of the architecture of a GPU [FIALKA].

Stream modeling was hence used in the GPUs with parallel processing being its base logic as it helped them achieve the efficiency demanded with minimal effort. In comparison to the present architectures in the market, the stream processors are able to provide twenty times the performance at similar conditions [WIKI02].

As soon as researchers realized that performing a computation on a graphics card would be far faster than performing it on a CPU (a GPU can process over 50 million triangles and 4 billion pixels in one second), the demand for GPUs increased in the market. Graphics pipeline is similar to a manufacturing assembly line with each stage adding something to the previous one. As mentioned before we know a GPU works mainly on the basis of parallel computing because of this pipeline architecture. The graphics pipeline hence accepts some form of a three-dimensional image as an input and transforms it into a two dimension screen. The programmable parts of a GPU are the *vertex* and the *fragment processors* which execute the vertices and fragment shaders respectively. A major advantage of functions like the *vertex and the fragment shaders* is that they allow the GPUs to do the operations in parallel as the result in each vector is independent from the other.

Most of the programmable computational power of a GPU is within the fragment processor. For example, there are 16 pipelines in the fragment processor of the Nvidia GeForce 6 Series GPUs. According to the information provided by Nvidia, each one of these 16 pipelines can handle a maximum of 4 floating point operations in parallel with one another at 450 MHz. On further testing they found out that the total computational power of the GeForce 6 Series GPUs was around 50 *Gigaflops*. They are expecting this performance to grow with a suitable increase in the number of pipelines.

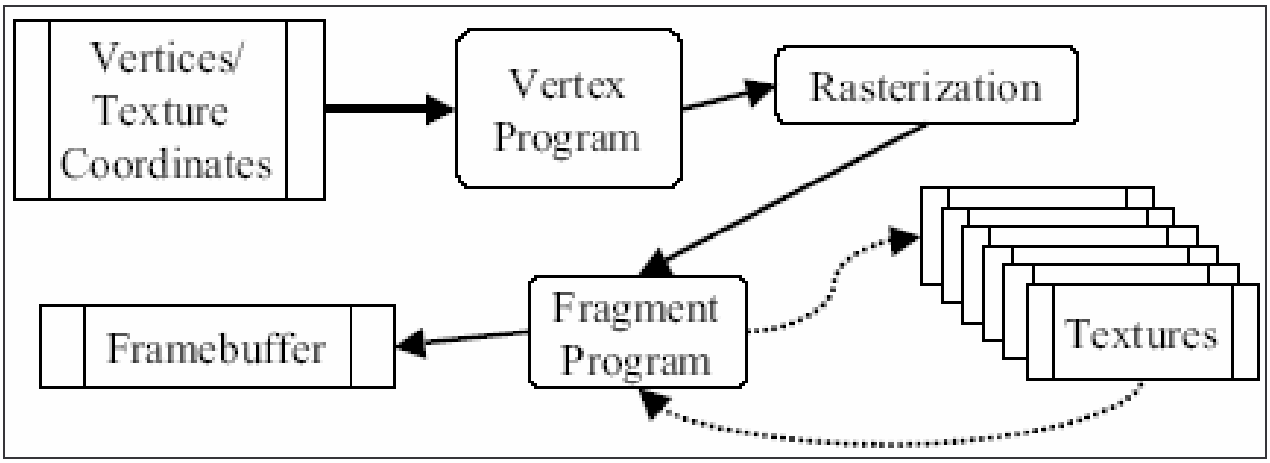

Figure 8: Graphics Processor Computation Pipeline [LEFOHN].

According to Sony Computer Entertainment Inc., the recently released PS3 has a floating point performance of somewhere around 2 teraflops whereas the Xbox 360 has a floating point performance of 1 teraflops. This essentially means that these video game consoles can behave as supercomputers because currently the most general purpose PCs has a floating point performance of a maximum of 10 gigaflops. 3D graphics operations are good examples of a problem which can easily be dealt with, by dividing the task between different execution units and pipelines, allowing a high speed.

Another important feature of a GPU is that they do not perform *branch* or decision making operations like ‘if/else’ statements which determine what will be executed based on the value of a piece of data unlike a CPU. Also we know that CPUs access data more unpredictably than the GPUs which require them to include an on-chip memory called a *cache* for quick random access.

6. Primary Application of GPUs

GPUs came into the market as a graphics rendering device. GPUs are primarily used in PCs and video games for rendering high resolution graphics. In this section we discuss briefly about two of the applications of GPUs that involve rendering of high resolution and quality graphical images.

6.1. Video Games

With the use of modern technology and a competitive sellers market, video games are not the same as they were 25 years ago. The gamer generation is much bigger than the baby boomers now with more than 90 million [SHERMA]. It has won 30% of the US toy market earning more than 8.8 billion, a share which is bigger than Hollywood box-office gross [MEDIAL].

Video games are now more complex and challenging. George Lewis, an MSNBC correspondent, believes that the development of strategic and critical thinking skills, balanced with the need for change and participation in other activities, make it acceptable that video games can have a positive impact upon society for some people. Video games played in moderation can help young people develop mental skills such as problem solving and careful risk taking [LEWIS]. They have already been integrated into the daily routines of 65% of US households [MEDIAL]. But there is also a negative side to it. About one out of eight gamers develop all of the patterns similar to an addiction, according to Dr. David Walsh with the National Institute on Media and the Family. He also believes that the time spent on gaming replaces time spent on outdoor sports and social interaction. The positive and negative effects of excessive gaming are often debated by experts.

Modern GPUs allow for larger, more realistic and impressive worlds to be presented to a gamer. These games with the help of GPUs simulate real time high quality graphics which result in emotional involvement of the gamer [DELVES]. Nvidia and Havok, the game industry's leading supplier of cross platform middleware recently demonstrated a physics effects solution that completely runs on a GPU. This collaboration is resulting in a new software called Havok FX which can simulate dramatically detailed physical phenomenon. Presently Havok FX, GPUs can simulate the interactions of thousands of colliding rigid bodies using a fundamental technique in physics computation seen in many PC games today. But with the new software it is now possible to compute the components of friction, collisions, gravity, mass and velocity to form the basis of rigid body physics. This allows game developers to implement sophisticated facts such as debris, smoke and liquids which add immense detail to game environments. The software is designed for Nvidia GeForce 6 and 7 Series GPUs. The effects of this reality of video games in society is often debated [SBERT].

It is a popular opinion that video games drive gamers to be more aggressive and violent. Contrarily Steven Johnson, a cultural critic, points out in a recent book, "Everything Bad Is Good for You", that gaming is so widespread now that if it did make people violent, its effects should be obvious. He also points out that in America violent crime actually fell in the 1990s, just when the use of video and computer games was growing rapidly. This is backed by the federal crime statistics, according to which the rate of juvenile violent crime in the United States is at a 30-year low as indicated by the plot below [THEECO],

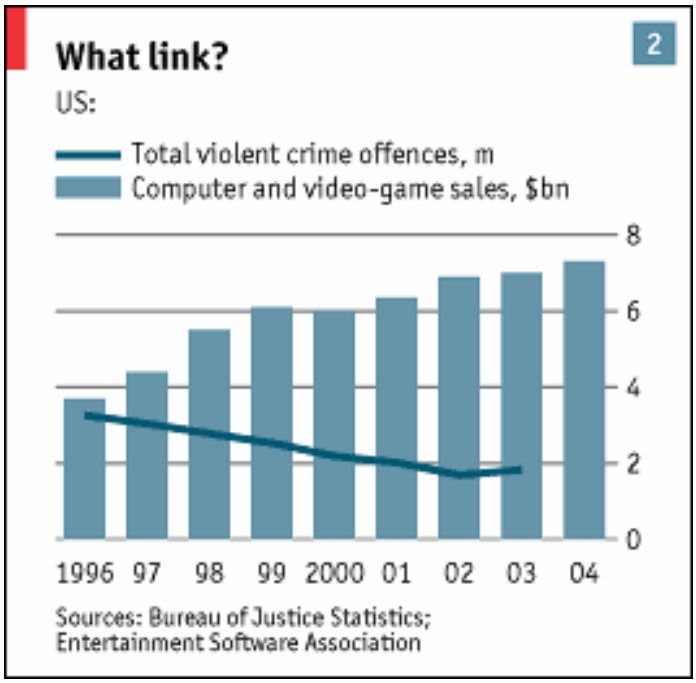

Figure 9: Justice Bureau Statistics relating growth of video games and juvenile crime [THEECO]

Another interesting fact observed by researchers is that people serving time for violent crimes typically play less videogames before committing their crimes than the average person. It's true that young offenders include gamers but young people in general are more likely to be gamers [THEECO, JENKIN].

The biggest concern is probably the amount of time children spend playing games. Spending time playing video games can eliminate social activities [JOHNSO]. Growing global marketing efforts helps in development of millions of loyal followers of the gaming culture who prefer interacting in cyber play rather than spending time with friends or play street sports according to a study by Media Analysis Laboratory, Simon Fraser University, Burnaby B.C. [MEDIAL]. Contrary to this, Henry Jenkins, a MIT professor considers video game play as a social activity. He points out that 60 percentage of frequent gamers play with friends. Thirty-three percentage play with siblings and 25 percent play with spouses or parents [JENKIN].

The economist reports that age plays an important role in attitudes towards gaming. Most of these games are played by young adults and only a third of the gaming population is under 18 according to Marc Prensky of ***Games2Train***, a firm that promotes the educational use of games. The average age of an American gamer is 30. Though half of America plays video games or computer games of which 76% are under 40, most of the critics are over 40 years old according to Nielson a market research firm. This clearly implies an entire gaming generation that began playing as children continued to play. Figure 10 below gives us a better idea of the distribution of age groups of the gaming generation [THEECO].

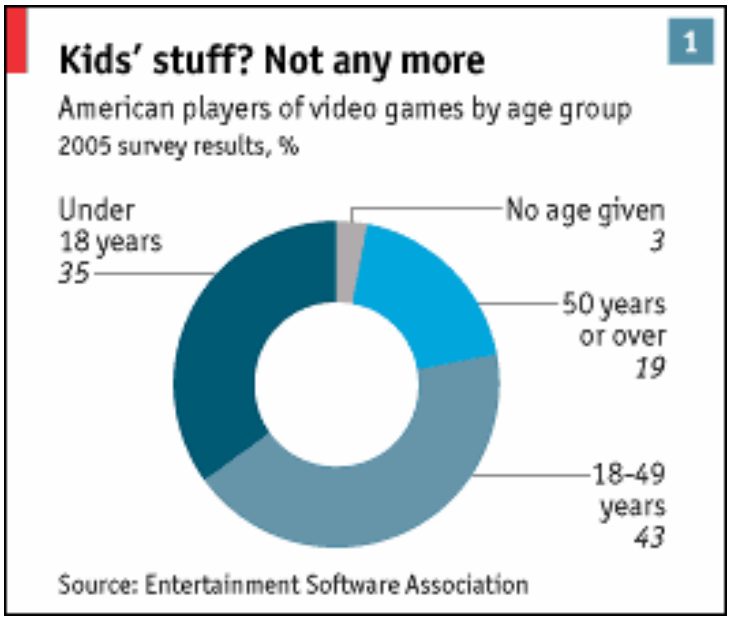

Figure 10: American players of video games by age group [THEECO].

This discussion also provides interesting insights to a question frequently asked about gaming: when will it become a truly mainstream form of entertainment? Above given figures indicate that it already is a mainstream entertainment for people under 40 [THEECO].

6.2. Virtual Reality Therapy

Burn injuries are some of the most painful injuries imaginable. Patients are required to undergo physical therapy to stretch the newly healed skin which has to be cleaned regularly. Many patients despite being on medications report excruciating pain during wound care. This trait of wound care has always been known and not much could be done about it.

Imprint Interactive Technology, based in Seattle works on virtual reality simulations and games for training, therapy and medical research. Dr. Hunter Hoffman and Dr. David Patterson at the University of Washington's Harborview Burn Center collaborated with Imprint and recently came up with an immersive simulation called Snow World using Nvidia graphics and Virtools™ 3D. This design aims at distracting the patients to offset excessive pain during therapy. The patients are made to wear a helmet connected to the virtual world which makes them feel as if they are floating through an icy 3D canyon. Also controlled studies show a considerable 40-50% reduction in pain ratings when Snow World was used. VR pain distraction results are encouraging according to Dr. Hoffman. A growing number of burn centers are now pioneering the use of Snow World using Nvidia graphics and Virtools™ 3D [NVIIMP]. This is one of the success stories of the use of GPUs in the medical field.

7. General Purpose Applications of GPUs

Today, there is no uncertainty in mentioning that the GPUs have come a long way in terms of what they are capable of doing, and in terms of competing with their CPU counterparts. There are many different projects nowadays that try to comply with the *stream model* of GPUs so that some specific applications can be run at a faster speed in comparison to a regular CPU. This concept of using the power of fast computation of GPUs is termed as General Purpose GPUs [GPGPU]. GPUs nowadays are being used in many cases, for many different applications that do not involve graphics but involve a lot of computations. GPUs have been used to accelerate many highly parallel applications. Some of them are listed below.

- Scientific Computing
- Medical Applications
- Protein Sequence Matching
- Database Applications
- Bioinformatics
- Computer Vision
- Application of GPUs in protein simulation

7.1. Scientific Computation in GPUs

With the power of GPUs, which is essentially fast computation of floating point numbers, it can be looked at as a powerful vector co-processor to the CPU. The intermediate values during a computation (using float buffers) are no longer clamped. Additionally, another good reason to use them as a co-processor is its parallel nature at the rasterization stage (pixel-level)

[HOUSTO]. The GPU stream model is still valid for this application as the texture-images (data) become matrices of values to do computations. In addition to normal computations, nowadays, GPUs are being used for linear algebra computations as well. There are, however, some limitations in GPUs which hinder scientific computations. The limited instructions and register space plays a big hand in making the program as simple as possible with the least number of lines of code. Absolutely no branching or conditionals or multipasses are allowed in GPU programs.

Basic Linear Algebra Subprograms (BLAS) are programs directly dependent on the GPUs. They perform many complex computations like vector-vector, matrix-vector or matrix-matrix operations through softwares like MATLAB [HOUSTO].

7.2. Medical Application

Medical field have been making use of the extreme computational power of GPUs for a while. Since medical field is a vast topic and area where GPUs are currently being used—esp. Medical Imaging, Visualization in Machine, and are discussed later separately in Section 9.

7.3. Database Operations on GPUs

As mentioned earlier in Section 6.1, Applications of GPUs, the main idea behind the use of GPUs in various other fields for general purpose applications is to make use of its fast computational power, thus being used as co-processors to CPUs. Capable of processing tens of millions of geometric primitives per second, GPUs can be used in performing fast computation of database operations. These operations like *predicate evaluation, range query, and accumulation* have been further discussed in greater detail in Section 8.

7.4. Bioinformatics

Bioinformatics applications are one of the most relevant and computationally demanding applications in today's day and age. In order to accelerate a bioinformatics application, graphics accelerators such as GPUs are used. GPUs are used for such applications as they are inexpensive and based on high performance SIMD architecture in comparison to the other substitutes in the market. Initially GPUs were only used for graphics applications whereas the ones nowadays can be programmed and used for many different purposes. In terms of bioinformatics, one of the common examples where GPUs come into use is in porting a bioinformatics application called RAxML which is a program for inference of phylogenetic trees from a DNA sequence data based on the Maximum Likelihood [CHARAL].

7.5. Computer Vision

Computer vision tasks are extremely intensive and exceed the capability of a CPU. As a result of this computer vision tasks are mapped differently on GPGPUs by mapping mathematical operations of the computer vision onto the modern computer graphics architecture. A large number of computer vision algorithms are written and implemented because of the programmability feature of the GPUs which was earlier not possible using a CPU.

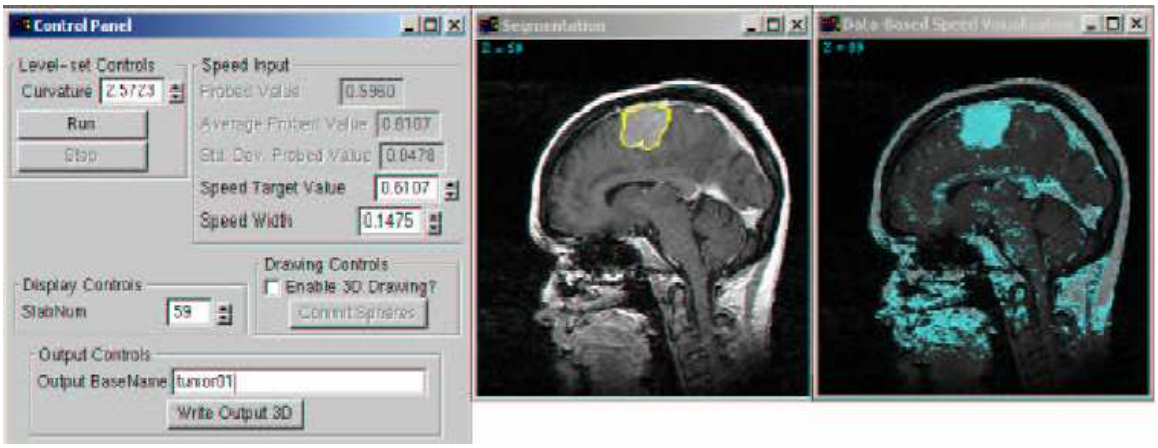

Figure 11: MRI Scan of a Skull [LEFOHN].

The Computer vision algorithm analyzes images to create numerical representations of a scene whereas on the other hand computer graphics analyzes the numerical description of a scene to create an image as illustrated in Figure 11. The SIMD property of the GPU comes into use here as many similar computations are implemented at the same time. By not using the CPU for these applications many more important tasks can be implemented on the CPU [FUNG].

7.6. Applications of GPUs in Protein Folding Simulation

The use of PS3 in protein folding simulation is a very interesting general purpose application of GPUs. Proteins, before performing any function fold together. This folding process is very critical and fundamental to all of the biological functions. This process is not clearly understood and thus remains a mystery. This is because proteins are difficult to observe as they are extremely small and the process speed is very fast, nearly about one millionth of a second. Any discrepancy in the folding can cause serious diseases such as Alzheimer's, Huntington's and Parkinson's disease.

Scientists hence use computer simulations but it has its own limitations. A normal computer would take a day to create a nanosecond of simulation and so it would take about 30 years to complete one complete simulation. Sony is working with Stanford University's *Folding@home* project to harness the PS3's technology to help study how proteins are formed or sometimes incorrectly formed in the human body. This is done by harnessing power from many PS3s connected to the internet when not being used to play games [WILLIA]. The cell microprocessor of the PS3 will help perform calculations to simulate protein folding and its graphic chip will be used to show the folding process in real time from different angles with good looking interface using new image technologies. With the use of new GPUs they are likely to attain high performance to the 10 teraflops scale.

Presently this *Folding@home* project uses a network of about 200,000 personal computers to simulate protein folding. The division of complicated calculations into smaller manageable packets and sending them to participating machines enables this project to do calculations which would even be a strain to combined resources of all the supercomputers in a country. These calculations being very challenging, a network of PS3s would increase the speed of the simulation by a considerable factor. According to Folding@home, a network of 10000 PS3s would increase the speed by a factor of 50 of what is possible today. This could reduce a couple of years of work to nearly a month [WILLIA]. The general consensus is that a number of Play Station 3consoles along with their GPU's will help them achieve calculations at the 10 petaflop scale which easily outperforms the fastest supercomputer in the market today.

8. GPUs in Database Applications

GPUs have been used to increase the speed of data access and sorting data from a database for a while. This section discusses the various areas within databases where GPUs have been used for general purpose computation hence speeding up the various processes.

8.1. Fast Computation of Database Operations

Database operations include predicates, Boolean combinations, and aggregations [GOVIND]. The paper published by researchers at University of North Carolina, Chapel Hill, Fast Computation of Database Operations using Graphics Processors [GOVIND] discuss the implementation of various algorithms on GPUs for selection queries on one or more attributes and generic aggregation queries including selectivity analysis on large databases. The paper also talks more in detail about the algorithms that the authors have implemented for database operations. Below we discuss in more detail the experiments performed and the results obtained for each of the database operations. All experiments were performed on a high end Dell Precision Workstation with dual 2.8GHz Intel Xeon Processors and an Nvidia GeForceFX 5900 Ultra graphics processor. The graphics processor had 256MB of video memory with a memory data rate of 950MHz which can process up to 8 pixels at processor clock rate of 450 MHz. This GPU can perform IEEE single precision floating point operations in fragment programs.

8.1.1. Predicate Evaluation

Each of the experiments was conducted on the first attribute of each record in the database. Figure 12 shows the plot of the time taken to compute a single predicate for an attribute

such that the selectivity was 60%. The plot is a simple comparison of a compiler generated SIMD optimized CPU code against a simple GPU implementation [GOVIND]. The computational time for evaluations of the predicate and the time taken to copy the attribute into the depth buffer are both considered under GPU timings. It is quite obvious from the result of the experiment that GPU timings are nearly 3 times faster than the CPU timings (20 times faster if we just consider the computational time).

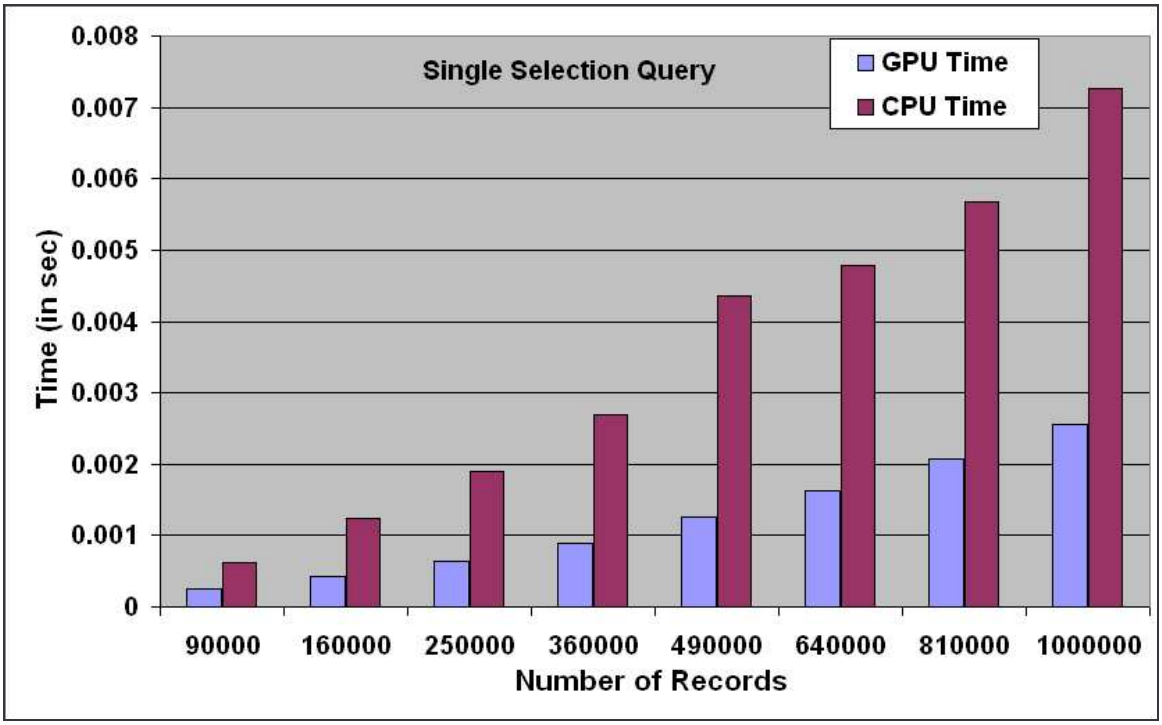

Figure 12: Execution time of a predicate evaluation with 60% selectivity by a CPU-based and a GPU-based algorithm [GOVIND]

8.1.2. Range Query

As predicate evaluations, the performances of range queries were also tested with 60% selectivity. The 60% selectivity was ensured by setting valid range of values between the 20th percentile and 80th percentile of data values. Figure 13 summarizes the time taken for a simple

GPU implementation and a compiler-optimized SIMD implementation on CPU. Again, the overall performance of GPU is seen to be nearly 4 times faster than the CPU implementation (20 times faster considering the computational time alone) [GOVIND].

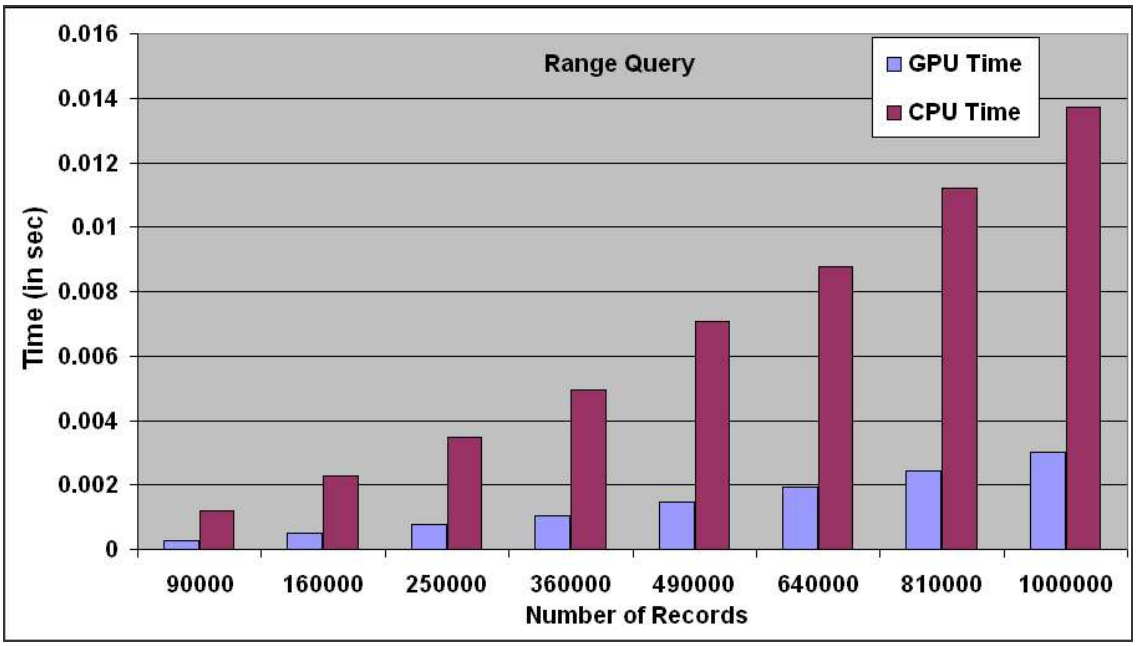

Figure 13: Execution time of a range query with 60% selectivity using a GPU-based and a CPU-based algorithm [GOVIND]

8.1.3. Multi-Attribute Query and Semi Linear Query

The performance of the hardware based multi-attribute queries were tested by varying the number of attributes and the number of records in the database. Again queries with the 60% selectivity for each attribute were used in conjunction with the AND operator on the result of each attribute [GOVIND]. The tests used up to four attributes per query and for each attribute per query, the GPU implementation copied the data values from the attribute's texture to the frame-buffer. Similar to multi-attribute query, semi-linear queries were also performed on the four attributes by using a linear combination of four random floating point values and comparing it against an arbitrary value [GOVIND].

Figure 14 gives us a summary of the results for multi-attribute query, which again illustrates GPU timings to be nearly 2 times faster than the CPU implementation. Figure 15 follows up with the results of timings for semi-linear queries where we observe that the GPU timings are 9 times faster than that of an optimized CPU implementation.

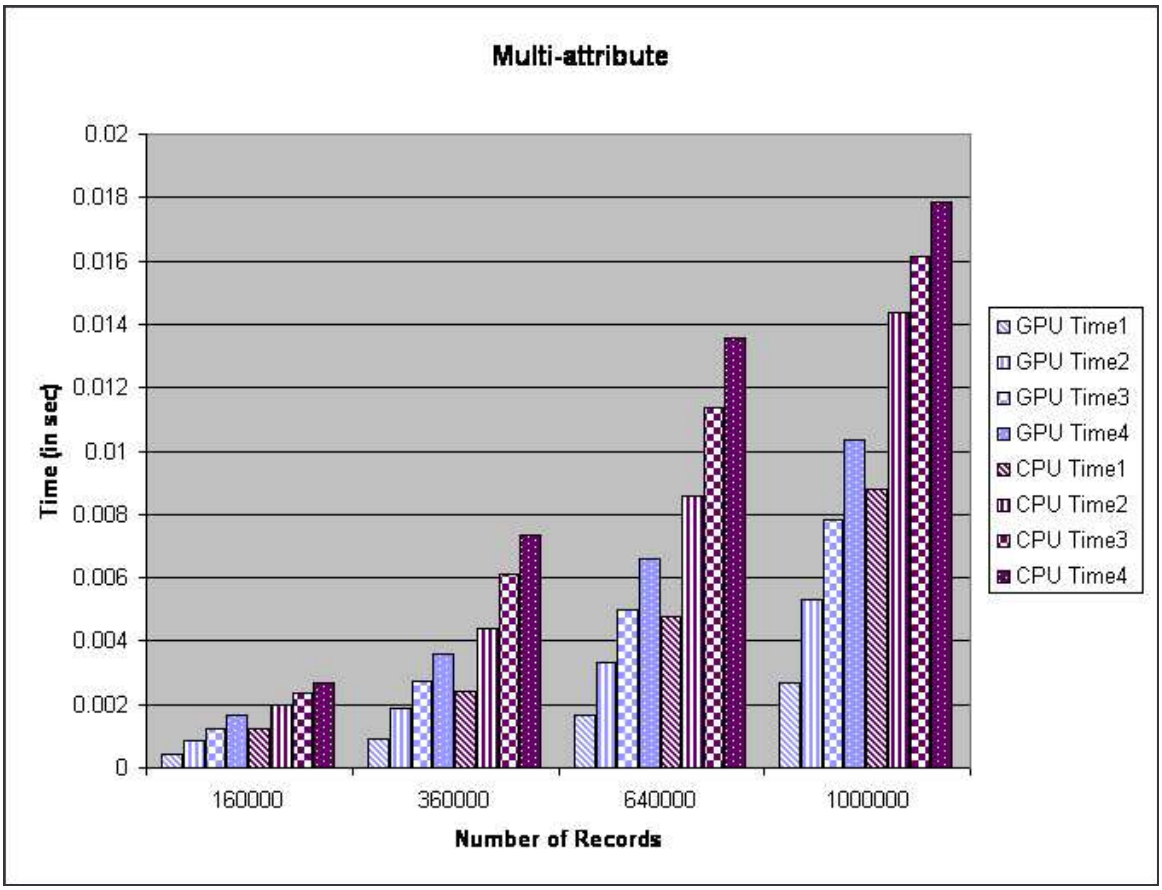

Figure 14: Execution time of a multi-attribute query with 60% selectivity for each attribute and a combination of AND operator. Time *i* is the time to perform a query with *i* attributes.[GOVIND]

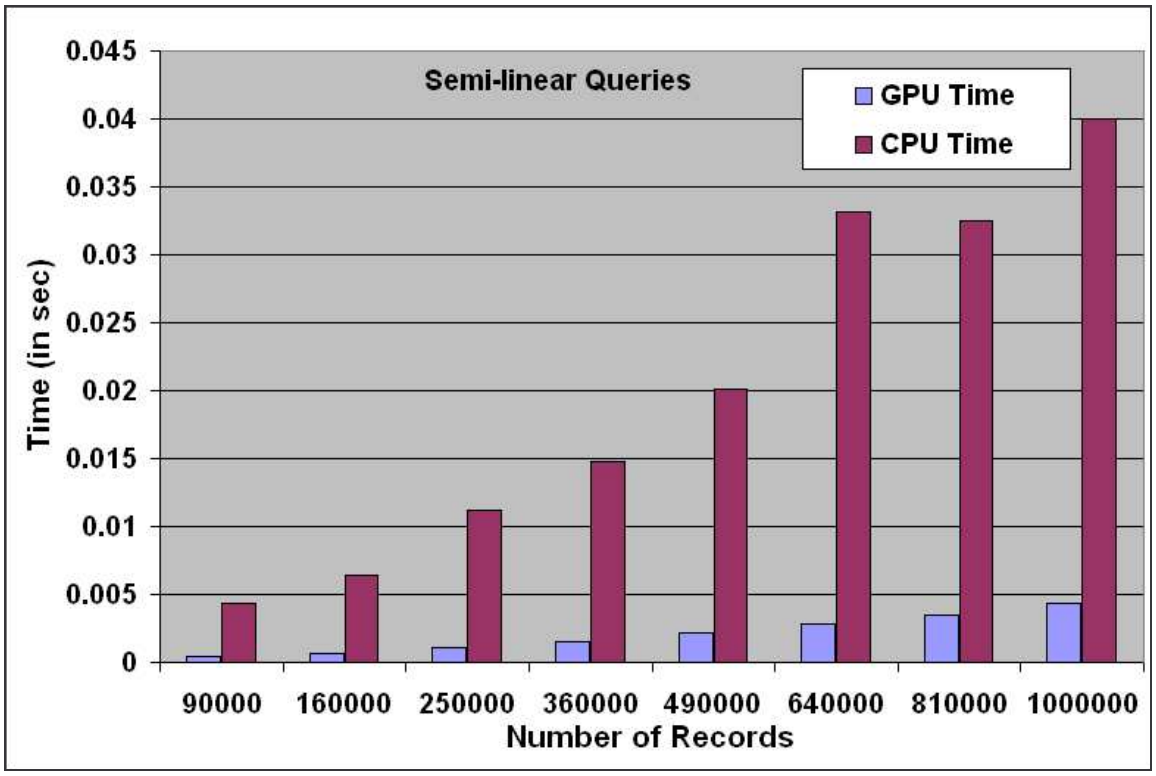

Figure 15: Execution time of a semi-linear query using four attributes of the TCP/IP database [GOVIND].

8.1.4. Accumulation

Accumulation is one area where the GPU seems to be slower than a CPU. Figure 16 demonstrates the performance of an accumulator on the GPU and a compiler-optimized SIMD implementation of accumulator on the CPU. The experiments [GOVIND] shown in the figure demonstrates that the GPU algorithms were `nearly 20 times slower than the CPU implementation when including the copy time'. This is because the CPUs have a much higher clock rate as compared to the GPU.

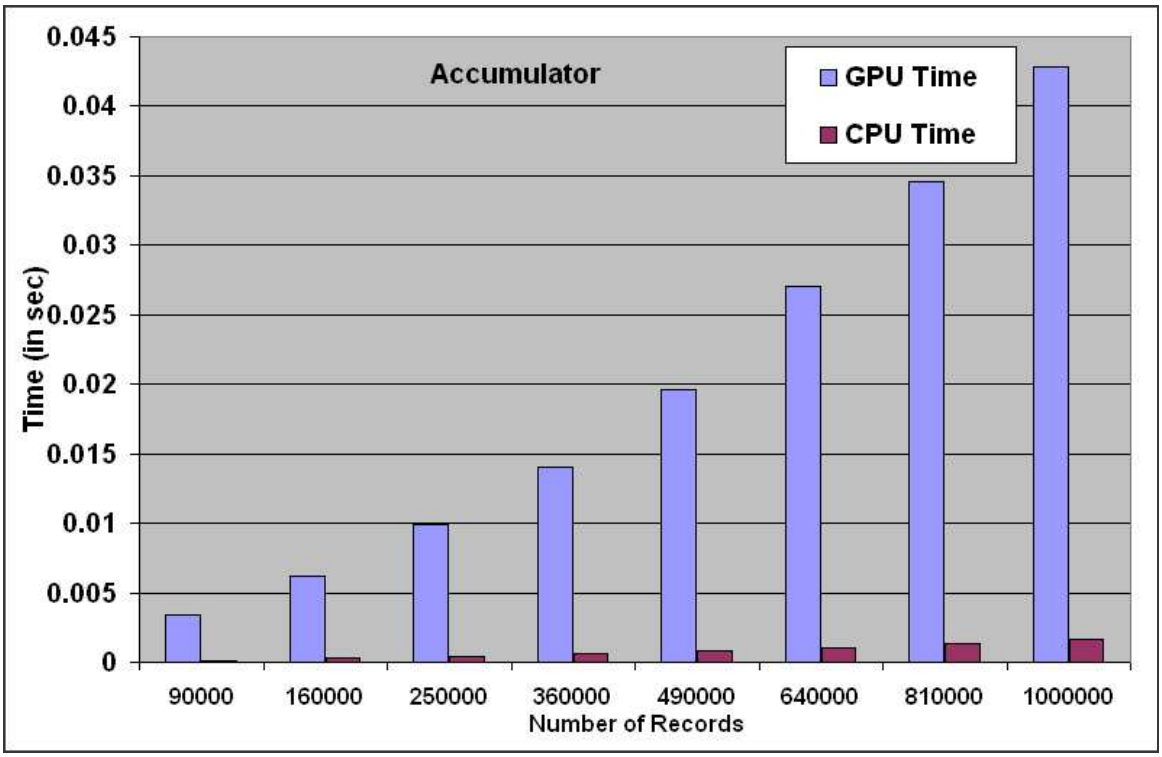

Figure 16: Time required to sum the values of an attribute on the CPU and by the GPU-based Accumulator algorithm [GOVIND].

8.2. Sorting Algorithm using GPUs: GPU Terasort

The term *GPUTeraSort* is self-explanatory and relates to sorting of hundreds of gigabytes using GPU cores. It is an external sorting algorithm developed in University of North Carolina, Chapel Hill in collaboration with Microsoft Research that uses GPUs on large databases that consists of billions of records. The designers of this algorithm have used the data and task parallelism power of GPUs to perform memory-intensive and computation-intensive tasks with the CPU performing the regular task of I/O and resource management. This sorting architecture controls multiple memory interfaces on the same computer using the GPU's high bandwidth memory interface along with the general CPU main memory interface. This high bandwidth of GPU memory interface and low bandwidth of the CPU main memory interface, together achieve higher memory bandwidth than purely CPU-based algorithms.

The GPUTeraSort involves two distinct phases. The first phase consists of a task pipeline that comprises read disk, build keys, sort using the GPU, generate runs, and write disk. The second phase comprises read, merge, write. The design takes into account the limited communication bandwidth between the two processors—the CPU and the GPU, and reduces data transfer between the two. It also pipelines data transfer from disks into the memory, resulting in a very high I/O performance. The algorithm has been tested on the standard Sort benchmark at hundred Gigabyte scale.

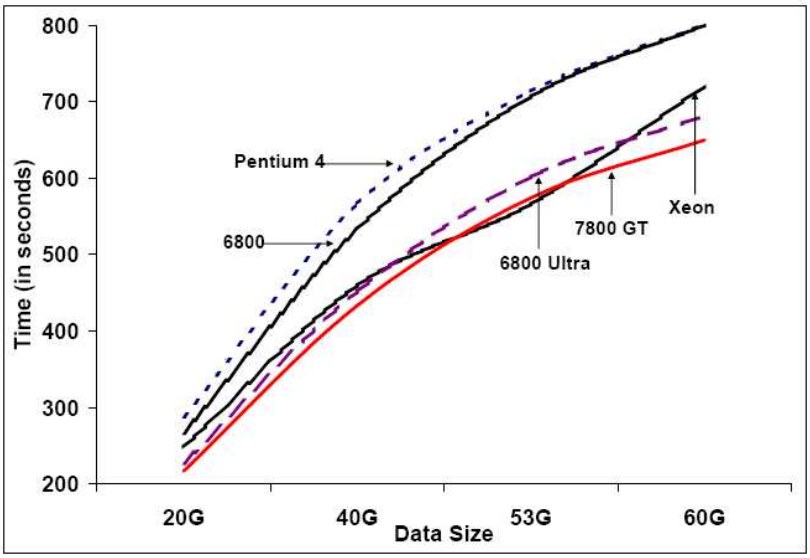

Figure 17: Performance of GPUTeraSort on GPUs and high end CPUs

The results of the experiments and test performed on GPUs and high-end CPUs are shown in Figure 17. It is interesting to see how the trend of GPU performance increases from Nvidia 6800 to 6800 Ultra and from Nvidia 6800 Ultra to 7800 GT. The Nvidia 7800 GT has a much better performance than CPU-based algorithm running on a 3 GHz Pentium IV processor and a significant performance improvement over optimized CPU-based algorithm is observed on the high-end PC with 3.6 GHz Dual Xeon processors once the data size goes beyond 55 G.

Overall, the results indicate that using a GPU as a co-processor can significantly improve the performance of sorting algorithms on large databases.

8.3. Hardware Acceleration for Spatial Database Operations

It has been a tradition in database systems to focus on reducing the cost of I/O operations as far as possible. I/O is the bottleneck for many operations in a computer. With databases being increasingly accepted in areas such as GIS (Geographic Information Systems) and Bio-informatics, there has been an increasing demand of commercial databases having to support data types for complex data such as spatial geometrics and protein structures. New challenges are presented by these non-conventional data types and their associated operations, in particular the computational cost of some spatial operations, can be orders of magnitude higher than the I/O cost of the PC itself. Innovative solutions have started to emerge in order to improve the performance of spatial query processing. One of the new ideas include the use of a graphics processing unit to increase the speed of spatial queries in commercial databases [BANDI].

There are essentially two distinct factors that a DBMS query comprises—I/O cost and computational cost. The *I/O cost* refers to the time that the CPU spends in loading the actual data from secondary storage devices like hard drives into the primary memory (RAM) whereas the *computational cost* refers to the time spent by a database management system in processing the data in RAM and returning the result. I/O being a major bottleneck in typical computer architecture, traditional databases have always focused on reducing the I/O cost until very recently. With the increased uses of databases in various fields—specifically GIS and Bio-informatics, all of a sudden there is a need for DBMS to support complex data types that are used in these applications which are contradictory to classical databases. These new data types bring

along intense challenges and one of them is support for efficient data storage and retrieval of spatial data. Spatial data typically consists of large datasets that represent real world GIS information (along with CAD information).

Spatial databases are typically evaluated in two steps, namely *filtering* step and *refinement* step. In filtering step, the Minimum Bounding Rectangles (MBRs) of the objects and spatial indexes are used to quickly determine a set of candidate results. The refinement step is when the final results are determined by retrieving the actual geometrics of the previous step. The retrieved candidates from filtering step are compared to either query geometry or to each other. For some complex geometries like polygons, the cost of this step (refinement step) takes up most of the query cost. At ACM SIGMOD international conference in 2003, C. Sun, D. Agrawal, and A. El Abbadi, presented the idea of using graphics hardware to speed up the refinement step in spatial query operations. With the current computer architecture, algorithmic advancements are unlikely to significantly reduce the cost of geometry-geometry comparison. However, on the other hand, with the up-coming advancements in graphics technologies, graphics cards are capable of handling thousands of polygons in real time. Both graphics hardware and spatial databases work on geometrics such as points, lines, and polygons, and deal with geometric relations such as intersection and containment in a 2-dimensional or 3-dimensional space; the obvious reason to exploit the computational power of GPU to speed up spatial database operations [BANDI].

Oracle is a commercial database management system that has integrated set of functions and procedures that enables spatial data to be stored, accessed and analyzed promptly and in a very efficient manner. The name given to this collection is *Oracle Spatial*. The Oracle Spatial's

data model is a simple hierarchical structure consisting of elements, geometries, and layers that correspond to spatial data representation. The supported spatial elements (basic building block of a geometry) are points, line strings, and polygons. A geometry is modeled as an ordered set of primitive elements and is the representation of a spatial feature. Collection of geometries having the same attribute set is known as layer and each layer's geometries and the associated spatial indices are stored in standard tables of database [BANDI].

Oracle Spatial uses a multi-stage query model as shown in Figure 18. The first stage refers to the *primary filter* or the filtering step as discussed previously. In this step, the spatial index is used for query filtering and candidate geometries are identified on the basis of a given query criterion. In *the immediate filter* stage, candidate geometries from the first stage are compared with the query geometry. Immediately after this step follows the final stage, referred to as the *secondary filter*, in which the crucial *refinement* takes place as discussed previously. Optimization of this stage is what can make spatial queries faster.

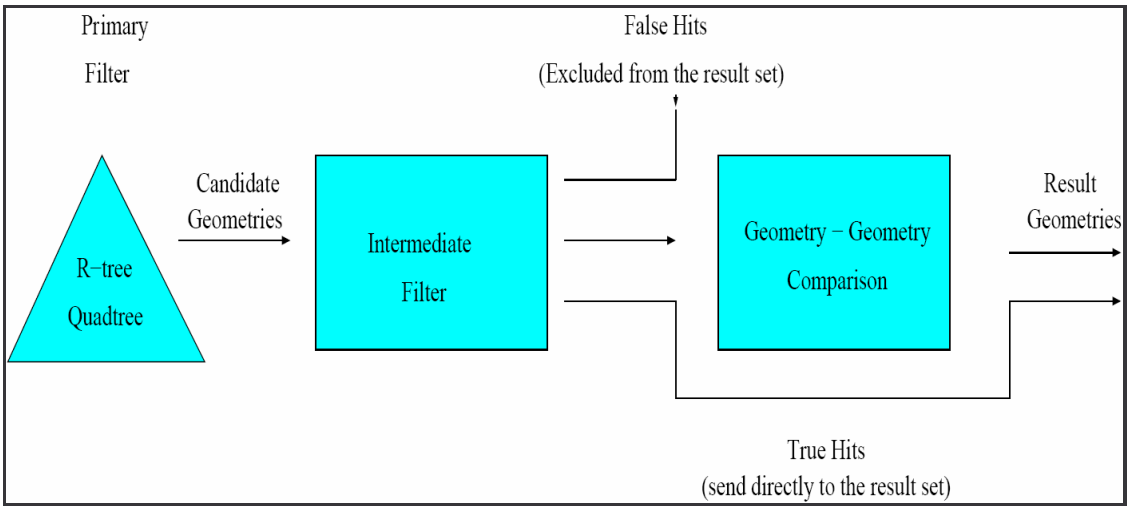

Figure 18: Oracle Spatial Query Model [BANDI]

The performance of the hardware filter (implemented with the use of GPU) integrated with an Oracle database was evaluated in the experiments performed by C. Sun, D. Agrawal, and

A. El Abbadi. The performances of intersection queries with and without the hardware filter were compared. The experimental setup comprised a desktop PC with an AMD AthlonXP 1800+ CPU and 1GB DDR (Double Data Rate) RAM. The GPU used was GeForce Ti4600 processor with 128MB on-board memory from Nvidia. The Oracle database was a version 9.2.0.1 instance running on Linux operating system. The hardware filter was coded in C++, compiled to a shared library using *g++* and was integrated with Oracle using the *Dual Thread* architecture. The experiments were conducted with real world datasets. The datasets that were involved in the experiments are listed below.

- **PRISM**: Average annual precipitation in the contiguous United States at 1:2,000,000 scale for the period 1961-1990.
- **HYDRO**: Hydrological unit boundaries for the United States, Puerto Rico and the US Virgin Islands at 1:2,000,000 scale.
- **COUNTY**: The boundaries of the US counties at 1:2,000,000 scale.
- **STATES50**: The boundaries of the main land boundaries of the 50 US states at 1:2,000,000 scale
- **LSOVER**: The boundaries of Landslide Incidence and Susceptibility distribution in the United States at 1:2,000,000.

Table 2: Statistics of experimental Datasets [BANDI]

Dataset	N	Number of Vertices Per Polygon		
		Min	Max	Average
STATES50	50	91	70238	4416
PRISM	6243	4	45854	94
HYDRO	5348	4	12450	218
COUNTY	4933	4	10838	139
LSOVER	2814	4	91752	92

Some of the results of the experiments are presented below in figures (Figure 19 – Figure 23) which signify that the use of GPUs helped in increasing the performance of the Spatial database [BANDI].

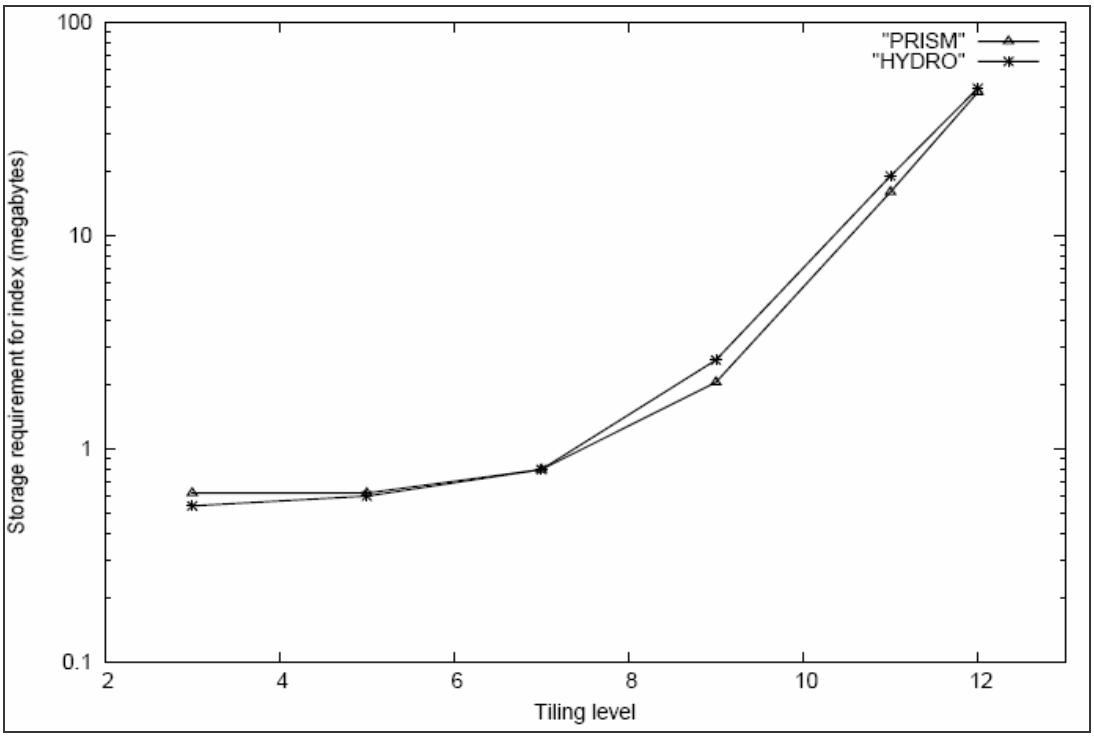

Figure 19: Quadtree index storage for *PRISM* and *HYDRO* (in log scale)

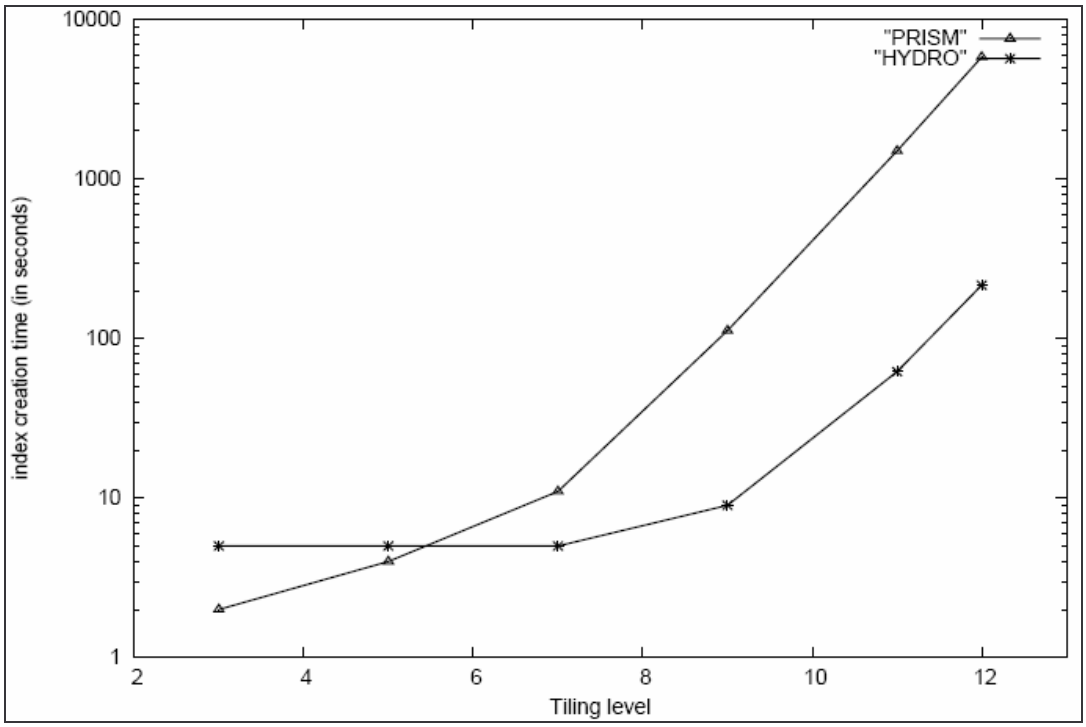

Figure 20: Quadtree index creation time for *PRISM* and *HYDRO* (in log scale)

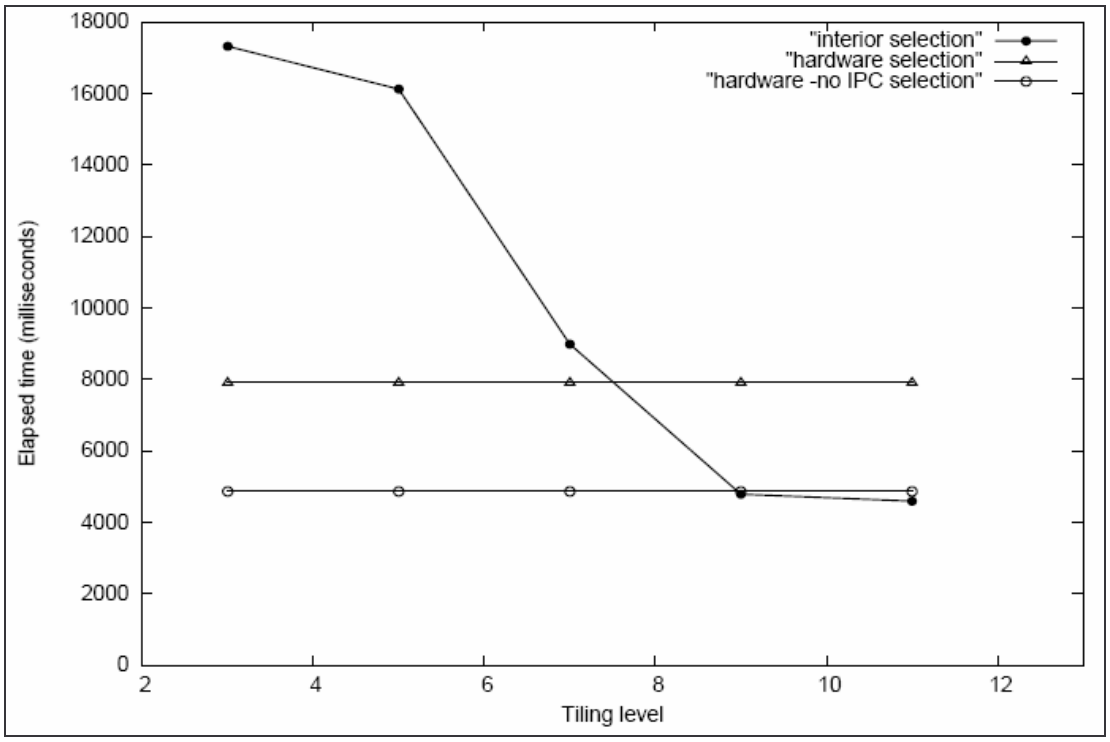

Figure 21: Timing results for selection over *PRISM*

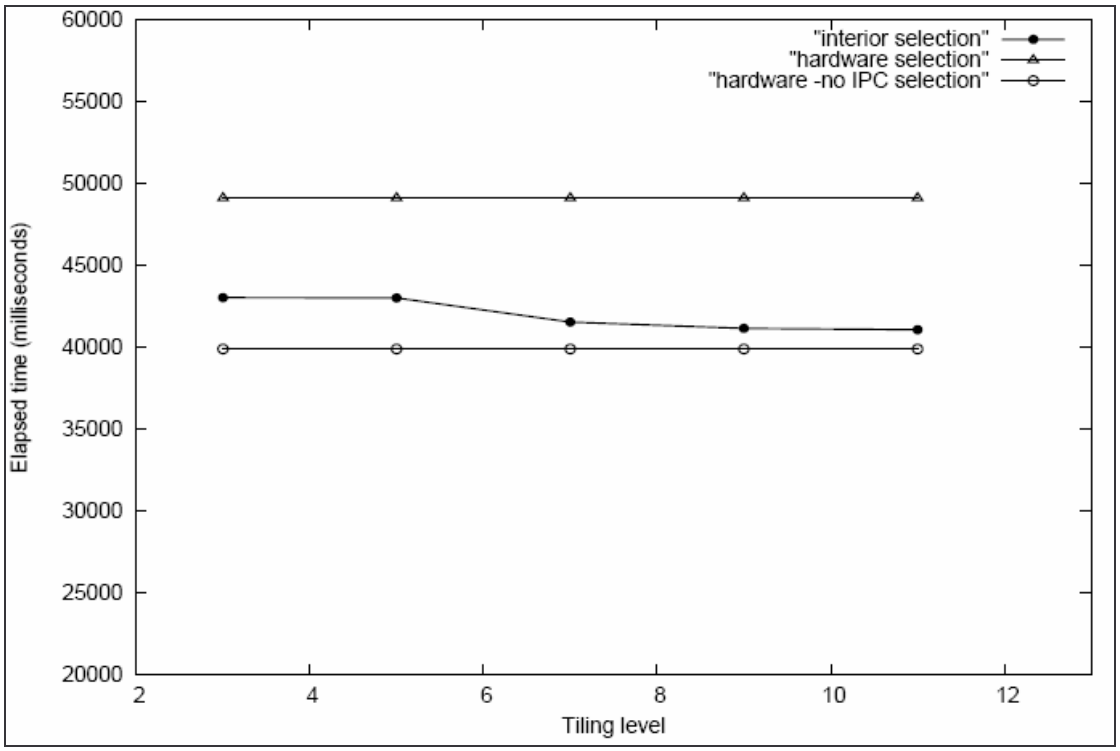

Figure 22: Timing results for selection over *HYDRO*

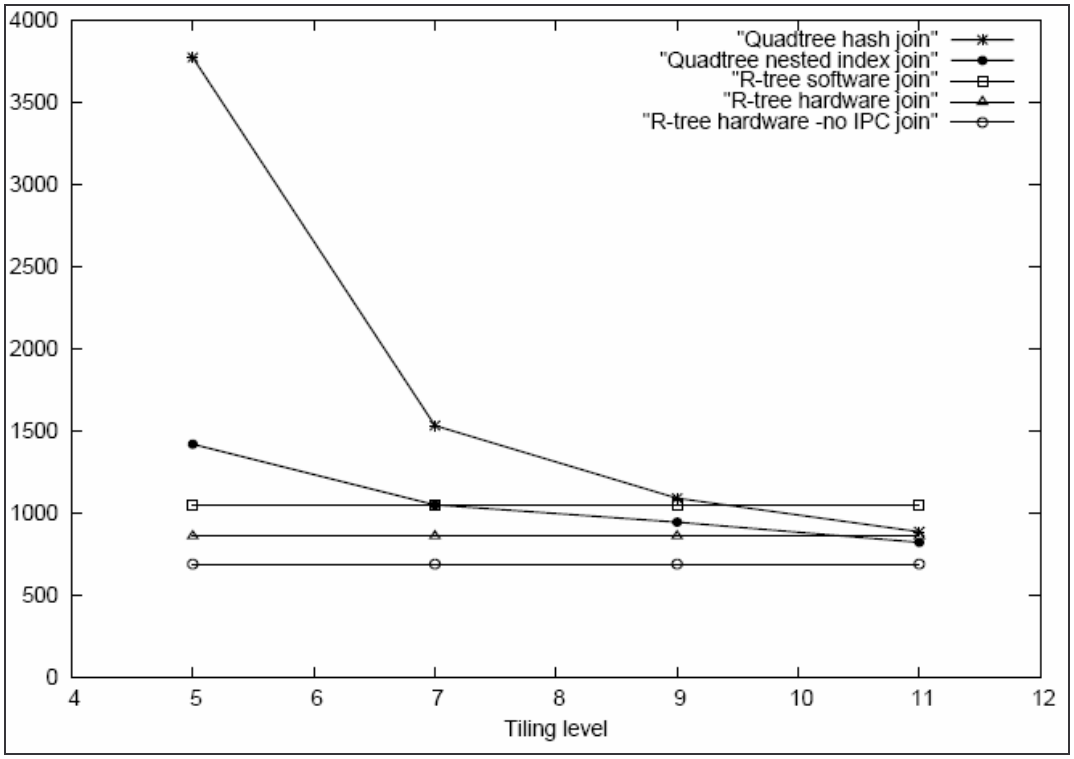

Figure 23: Timing results for join of *COUNTY X HYDRO*

9. Medical Applications

As previously discussed in Section 7, medical field has been making use of the fast computational power of GPUs for a while. This section briefly discusses two of the special areas of medical industry where GPUs have been used significantly.

9.1. Medical Imaging

In today's HPC (High Performance Computing) markets such as health care, users are not able to visualize large volumes of data. Nowadays the scanning and measurement devices are so advanced that they capture more information than that can be processed by a generic computer. They record information on multiple 2D planes and capture information about specific areas of particular structure within a large volume of data. After processing, GPUs allows the researchers to view this data in 3D by using various transformations to learn about the specific material in greater detail. As large volumes of data can be combined from various sources, visualization software aligns data that can be represented in 3D. This task of *volume visualization* requires powerful graphics and computing resources which can only be provided by GPUs [SGI].

In order to process this data in 3D a lot of pressure is put upon the CPUs. As this turns out to be an extremely time consuming procedure which yields slow results, GPUs have been introduced as a counterpart. As the medical field does not have any room for error, it is extremely important that it receives all the facilities in order to make medical breakthroughs in minimal time. For example, in order to detect cancerous cells in a timely manner with immense

accuracy we need a very powerful processor that can handle enormous computations in a short period of time.

Figure 24: 3D visualization of a Human Spine [SGI].

Today's GPUs are severely memory constrained, with each graphics processor offering no more than 512MB of memory. So even if we use four GPUs in a single system, we can see that only 2GB of memory would be available towards visualizing a data set. Although this may cater to the needs of some individuals, it would be considered to be a failure in the High performance computing industry. Recent market trends have also shown that in the immediate future visualizing data sets of 2GB will be rendered useless. Currently, this trend is clearly evident in the field of medical imaging as the industries are facing an explosion in the volume of data captured by Computer Tomography (CT) scans, Positron Emission Tomography (PET), fluoroscopy devices, Magnetic Resonance Imaging (MRI) scans, and two-photon microscopes which require a large amount of memory [SGI].

Figure 25: Detailed View of a Skeleton [SGI].

In the 1960's scientists came up with a rendering technique called ray casting which was later used to simulate nuclear penetration effects and create a computer animated movie called ***Tron*** [SGI]. This eventually came to be known as ***Ray Tracing***. The way Ray Tracing works is that it projects a ray through every pixel on the screen and traces it to the 3D scene image in high quality using a GPU through stream processing. This algorithm then calculates how the ray and the light source intersect with an object in the scene, and get reflected, refracted or absorbed. Ray tracing was used to achieve shadows and illumination at a higher level of precision with the introduction of the GPUs. These ray traced scenes and images can provide realistic detail even if it is obtained from a scanner or some other animation. According to the results a GPU based Ray

Tracer has the potential to easily outperform CPU based algorithms without needing new hardware to perform similar operations [SGI].

9.2. Visualization in Medicine

In today's day and age people argue that medical imaging suffers from a wealth of capability. Digital medical imaging could become more efficient if the captured data could be analyzed in a more effective manner. This additional use of 3D visualization in trauma and orthopedics will become more common in the near future [REID].

3D imaging in real time is referred to as 4D imaging which is clearly relevant in cardio applications. One area that will benefit most from digital imaging in general, and 3D visualization is mammography which is the process of using low-dose X-rays to examine the human breast to look for different types of tumors and cysts. In order to obtain all of this information through MRIs, CT scans and ultrasound, a generic CPU is aided by the latest generation of programmable Graphics Processing Units. In order to get faster results for the process of mammography, Mercury designed a solution which was based on Nvidia graphics technology using a single Nvidia Quadro graphics board that was able to generate imagery. It could accomplish this in under five minutes in comparison to the former 32 processor cluster which yields a reduction of 50% for the image processing time. By using this technology, life of doctors, radiologists and patients could become a lot easier.

In particular, radiologists can take maximum advantage from this technology as they would require minimal training to analyze the results as cancers can be more easily seen and differentiated from benign lesions. Even the patient's benefit from this as the exam requires only

one breast compression, as opposed to two with traditional mammography. Also due to the use of GPUs, results obtained are faster and more accurate which allows complete breast exams to be done in a single visit.

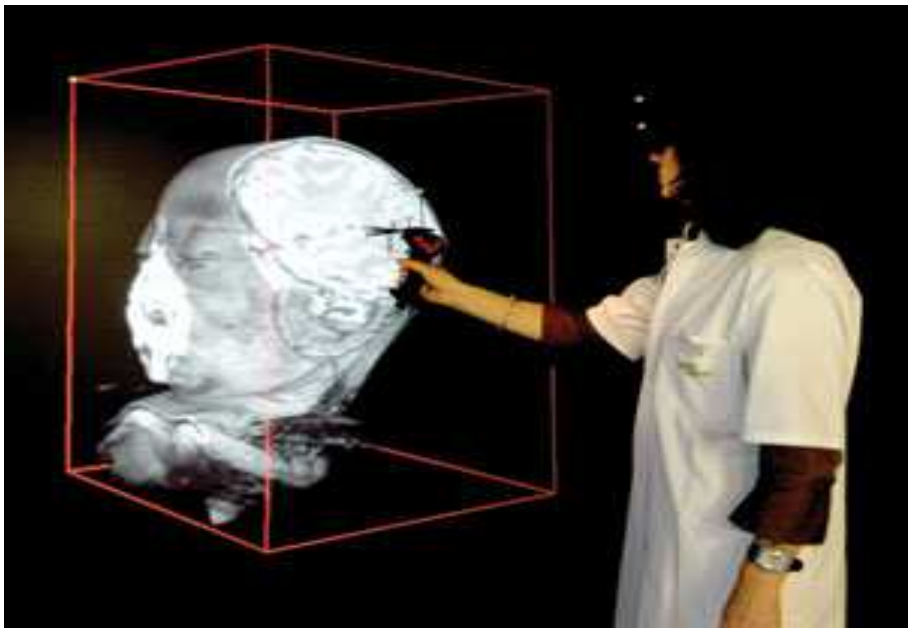

Figure 26: Dr. Grazia Mancini uses 3D visualization St. Erasmus MC [REID]

10. Future of GPU Applications

BionicFX has announced the release of a revolutionary technology for music production that turns Nvidia video cards into audio effects processors. Audio Video Exchange (AVEX) converts digital audio into graphics data, and then performs effect calculations using the 3D architecture of the GPU. The latest video cards from Nvidia are capable of more than 40 gigaflops of processing power compared to less than 6 gigaflops on the Intel and AMD CPUs. AVEX represents a major technological achievement that will allow music hobbyists and professional artists to run studio quality audio effects at high sample rates on their desktop computer [GPGPUAS]. This is just an example of what awaits this technology in the future.

11. GPU: A Disruptive Technology

The reason why some people are deeming GPUs as a disruptive technology is because of the recent similarities of the GPU architecture with that of a generic CPU. According to major hardware providers the new upcoming processors will not be user friendly to programmers who write processor specific code. Hence, in order to utilize these new hardware designs the programmers will be needed to change the way they write software which would be highly inconvenient as they would have to program in a complete new style. The processors from now on will feature non uniform and complex memory hierarchies which rapidly increase the core counts with integration of special acceleration units [PAPAKI].

The way software programmers will be affected by these new designs is that before they could change hardware designs through compilers and libraries with standardized interfaces but now, in order to take advantage of these newly designed processors the programmers will have to extract and explicitly express parallelism in their program. The best way to program these processors is through a method called stream programming which would enable superior productivity, performance, and efficiency, increasing the similarities to the GPU architecture. Hence, unless people start stream programming such innovative technologies could pose a serious problem to software engineers in the way they approach programming the processors. As it changes the norm of customary computer programming and deem the older technologies useless, GPUs could turn out to be a disruptive technology [PAPAKI].

AMD was the first company to break into the multi-core processor market when it released its dual-core processor by incorporating both the memory interface and the processor

interconnect interface on the same processor. Intel has since followed suit and is releasing the very first quad-core processor. According to some experts the numbers of cores on a single chip will double every eighteen months in accordance with Moore's Law. This will cause programming issues while making multithreaded applications due to the programmer's lack of knowledge. As both GPUs and the CPUs are going to become multi core with complex memory hierarchies, we can say that they will be similar in their architectures and design. The only difference in them would be in the features of the CPU like the larger cache, branch prediction, and speculative execution and in features of the GPU like signal processing, voice recognition, and medical image analysis [PAPAKI].

12. Survey on GPU Awareness

This survey aims at finding the general awareness of the GPUs with WPI students and staff as its primary audience. It will be interesting to know what percentage of the gaming population at WPI know what a GPU is and how graphics are rendered. By knowing what kind of games they play and what console they own through the survey a number of conclusions can be made. The type of games they are more attracted to, which graphics card do they use the most and the corresponding demographic information are useful in analyzing the obtained results.

This survey was conducted for 132 students at WPI. It should be noted that the target audience of this survey are only WPI students. The distribution of the survey population by gender is shown below in Figure 27.

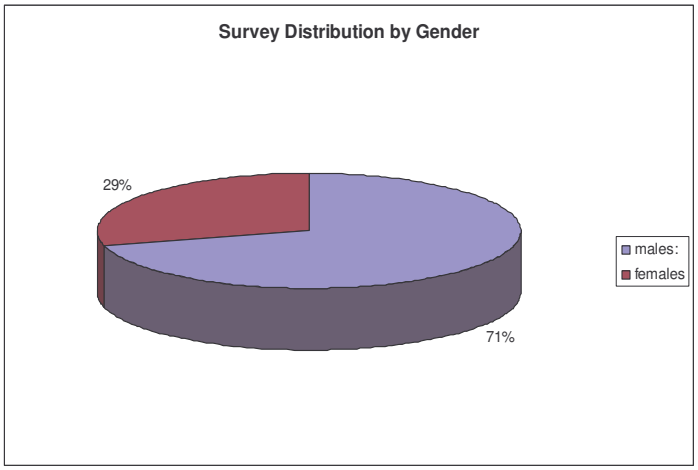

Figure 27: Survey Distribution by Gender

Out of the 130 students surveyed, 29% were females and 71% were males. This information is part of the demographic data obtained. This ratio is very typical to that of WPI as a whole.

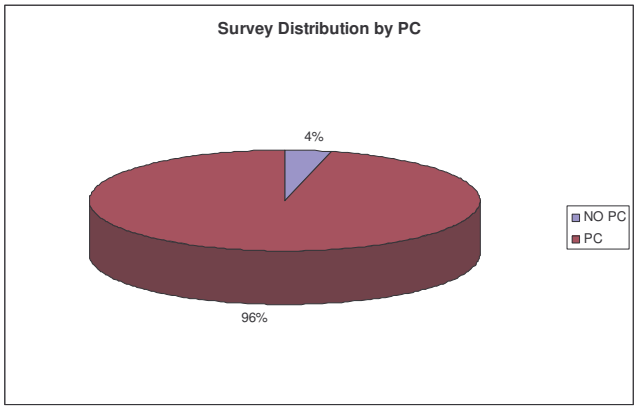

Figure 28: Survey distribution by ownership of a PC/laptop

Of all survey participations, only 4% did not own a personal computer. The distribution of the population depending on whether or not the surveyed participants own a PC is shown above in Figure 28. Majority being engineering students, it is understandable that only 4% do not own personal computers.

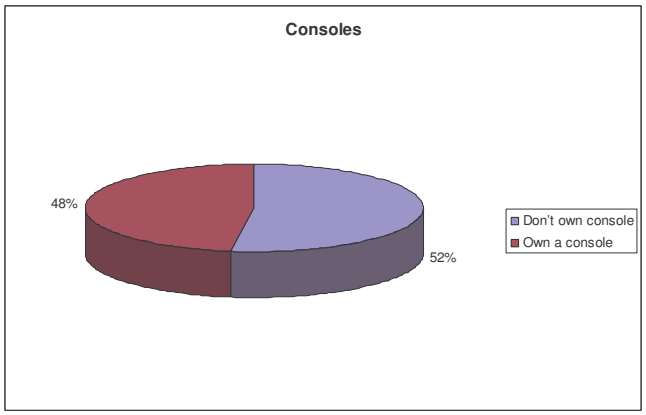

Figure 29: Survey distribution based on ownership of consoles

Figure 29 shows the popularity of video games played on consoles at WPI. Of the people surveyed 52% do not own any video game console. These include people who play games on their PC's and those who do not own a console but still play video games either on their friends PC's or at a gaming parlor.

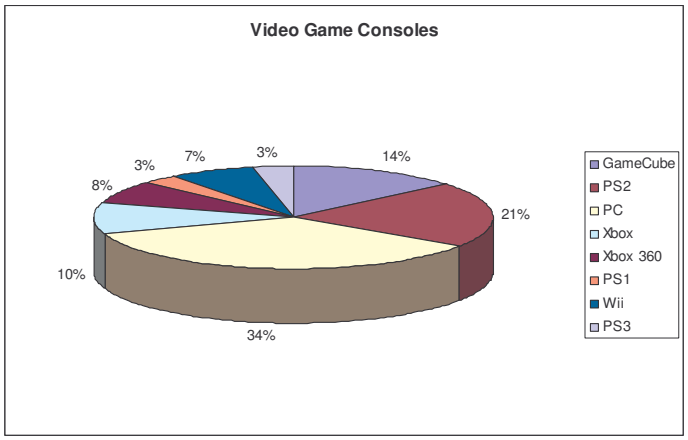

Figure 30: Survey distribution based on video game console ownership

Displayed in Figure 30 is an interesting distribution of popularity of each kind of video game console. Clearly most of the students' still play games on PCs followed by PlayStation 2 and GameCube. Note that PS 2 and GameCube were released a few years ago and were the most popular in the previous age of video game consoles.

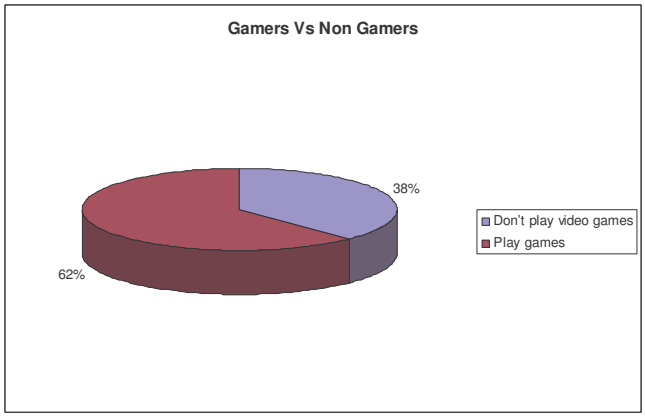

Figure 31: Survey distribution based on gamers vs. non-gamers

As the chart in Figure 31 suggests, a large percentage of students do not play games at all which is about 38%. The 62% includes people who own a console and those who do not own a console but still play video games.

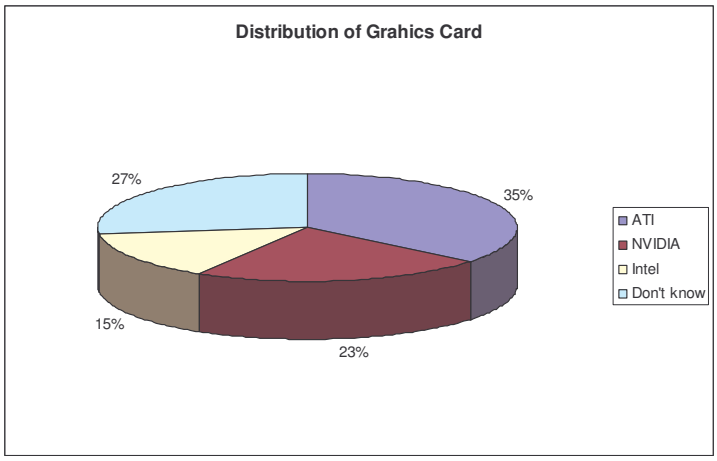

Figure 32: Survey distribution based on graphics card used in PCs

Figure 32 above demonstrates that most of the population knew about the graphics card they have in their PC/Laptops. 35% of the total number had ATI, followed by Nvidia with 23%. These percentages are quite close to the global graphics market percentages in 2005, when ATI was leading in discrete graphics card market. Another interesting thing to observe is that 27% of the people surveyed do not know what kind of graphics card is used in their PCs. Majority of the people in this group major in non-engineering fields such as Biology/Biotechnology. Students who major in BME and Chemical Engineering also rarely knew about their Graphics card.

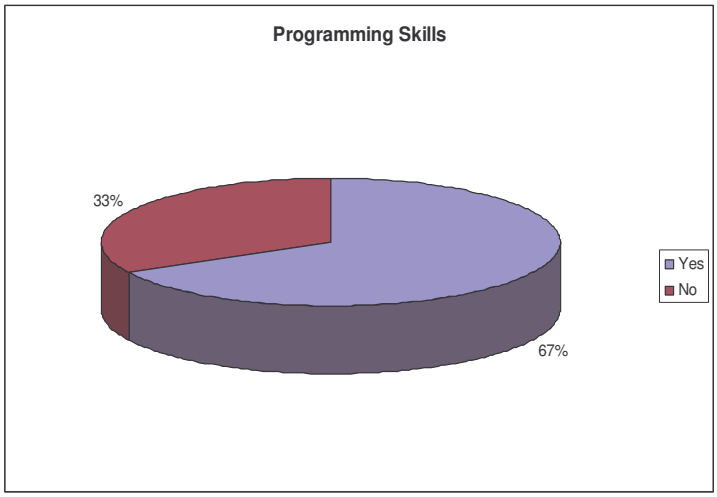

Figure 33: Survey distribution based on having programming skills

Figure 33 illustrates that out of all the people surveyed 67% of them knew programming. Of these people it will be interesting to find out how many know about GPUs. As expected majority of these 67% people have a CS or an ECE background.

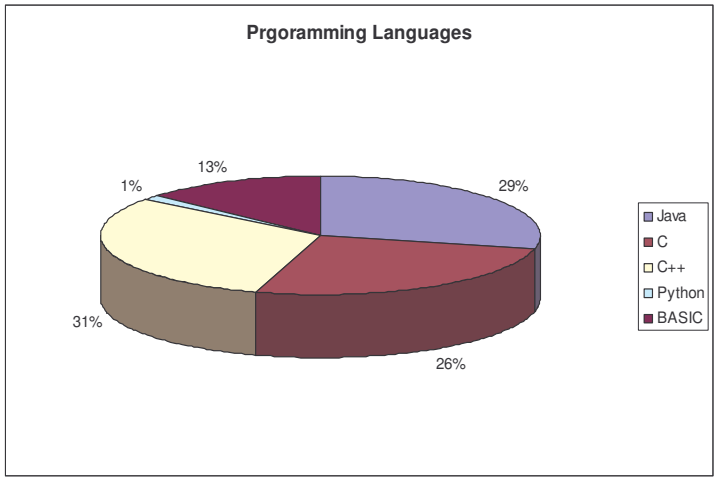

Figure 34: Survey distribution based on programming languages people knew

Out of the people who knew programming, most of them could program in C, C++ and Java (26%, 31% and 29% respectively) while only a very few knew BASIC (13%) as illustrated by Figure 34.

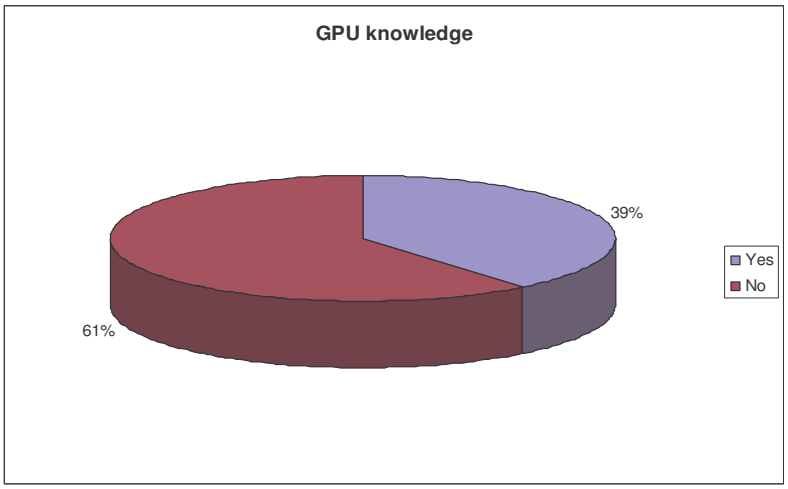

Figure 35: Survey distribution based on knowledge of GPU

Only 39% of the students knew about GPUs compared to the figure that 61% of them were programmers and were mostly from Computer Science department as shown in Figure 35.

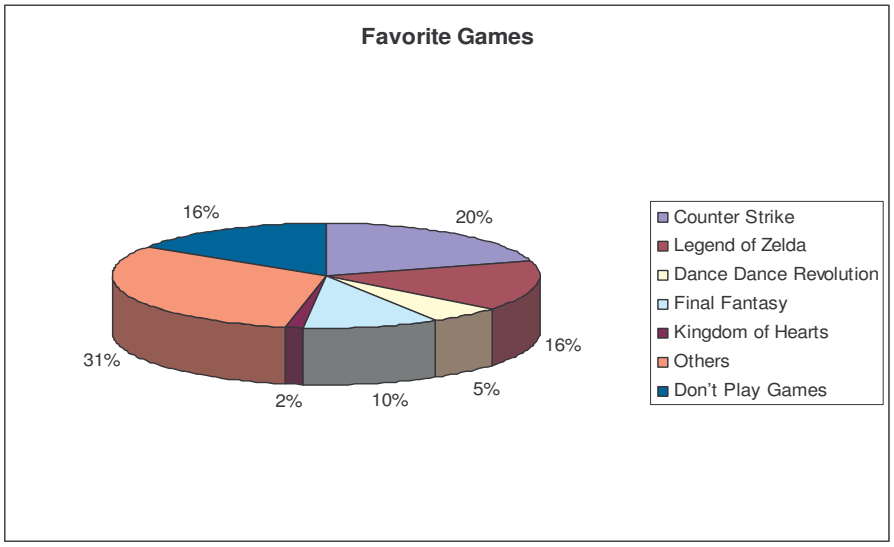

Figure 36: Survey distribution based on favorite games of the people surveyed

Figure 36 describes popularity of different games. It was surprising to see a large variation in the types of games played which tells us that gaming graphics market is not dependent on popularity of a few games. More importantly it was established that students tend to play games which are appealing rather than games which have better graphics.

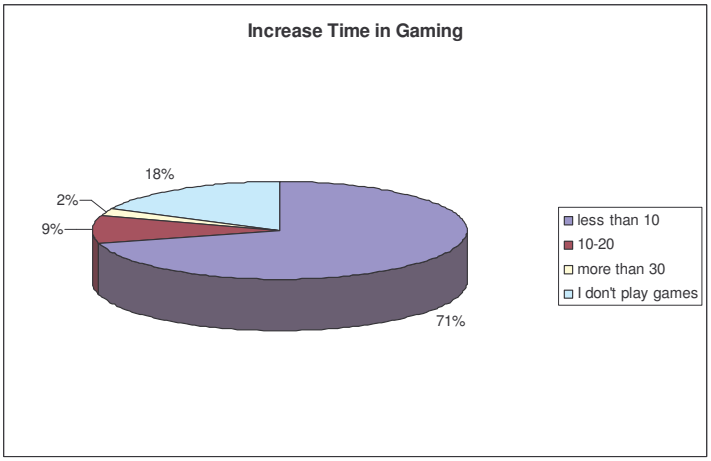

Figure 37: Survey distribution based on increased time in playing games due to improvement in graphics

The graph in Figure 37 tells us how much time they will spend on gaming if the graphics were improved. Surprisingly, a large population was not interested in graphics quality as most of them would still play less than 10 hours a week (71%). Many have commented that the number of hours played by them only increase when good games are released and not when the graphics capability increases.

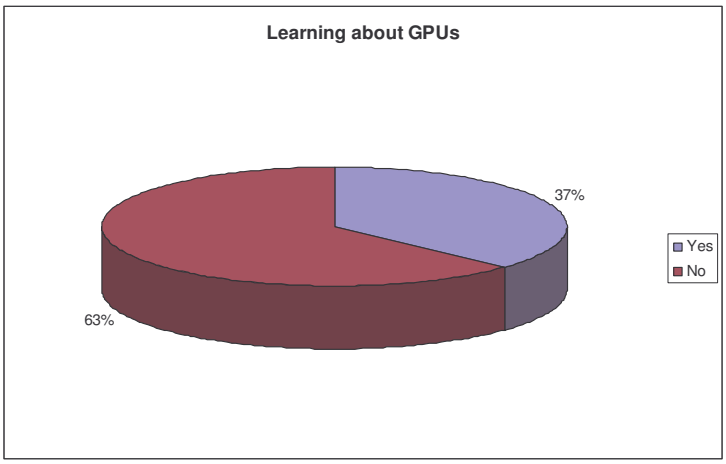

Figure 38: Survey distribution based on people interested in learning more about GPUs

Given an option, it was very interesting to notice that only 37% of the people surveyed were interested in learning more about GPUs as illustrated in Figure 38.

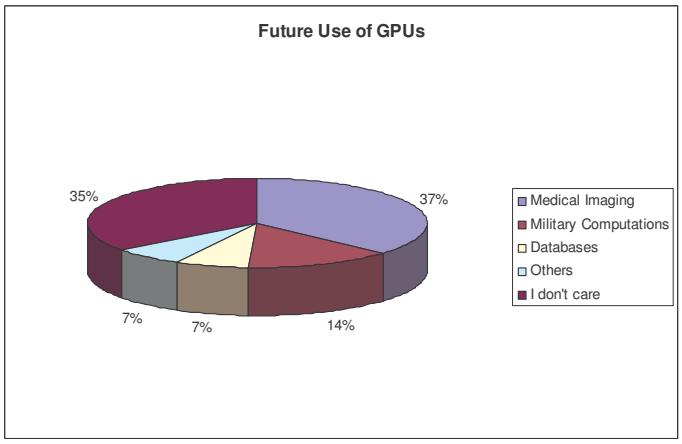

Figure 39: Survey distribution based on students' interest in the future applications of GPUs

Figure 39 shows that a large population of students thought that GPUs should be used for medical imaging and military computations (72% in total) in the near future than anything else.

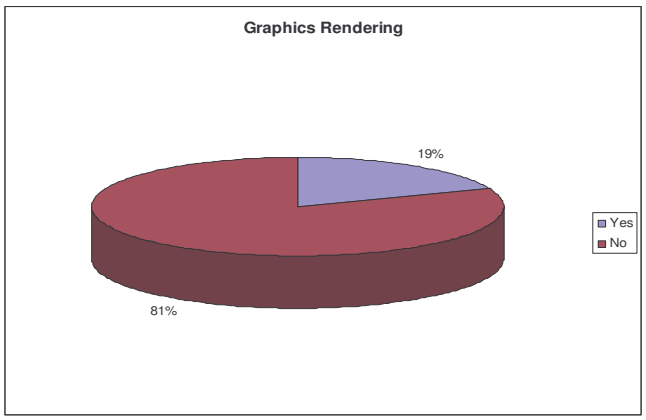

Figure 40: Survey distribution based on knowledge about rendering of graphics

As Figure 40 shows, only 19% of the population knew about general graphics rendering. This implies that whatever knowledge students have about Graphic cards is only about the market and not about the functioning and core of a graphics card.

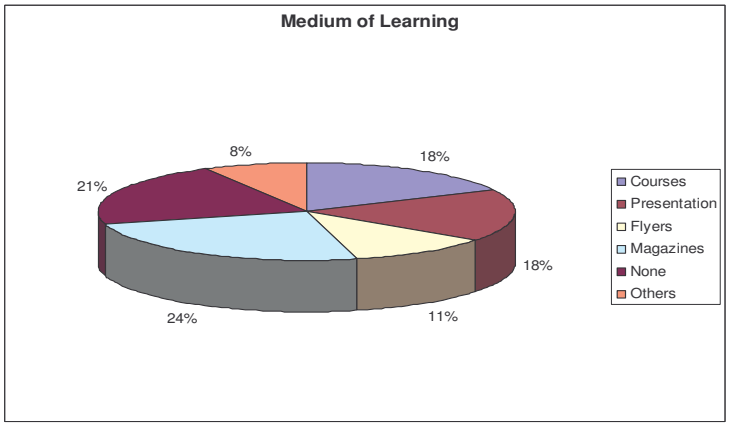

Figure 41: Survey distribution based on ways students want to learn about GPUs

Figure 41 demonstrates that out of the people who wanted to learn about GPUs most of them preferred magazines as the medium which shows that most people rely on magazine articles for information. Interestingly 18% of the people wanted to learn about GPUs through

courses. This suggests that a certain group of people are interested in knowing a lot more than what a flyer or a magazine can tell them about GPUs.

Of the 130 people surveyed, the pie chart in Figure 42 illustrates the distribution of the population by their major field of study. It can be inferred that one third of the surveyed participants were majoring in Electrical and Computer Engineering (ECE) while 19% were majoring in Biomedical Engineering (BME). This result was inconsistent with what we were expecting.

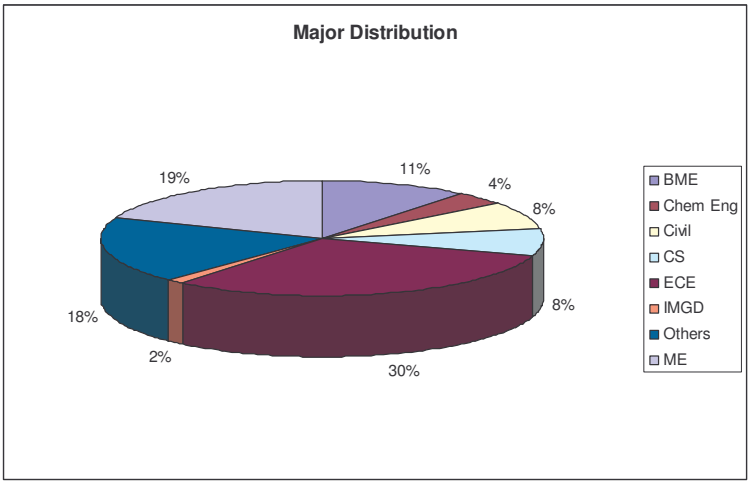

Figure 42: Survey distribution based on students' majors

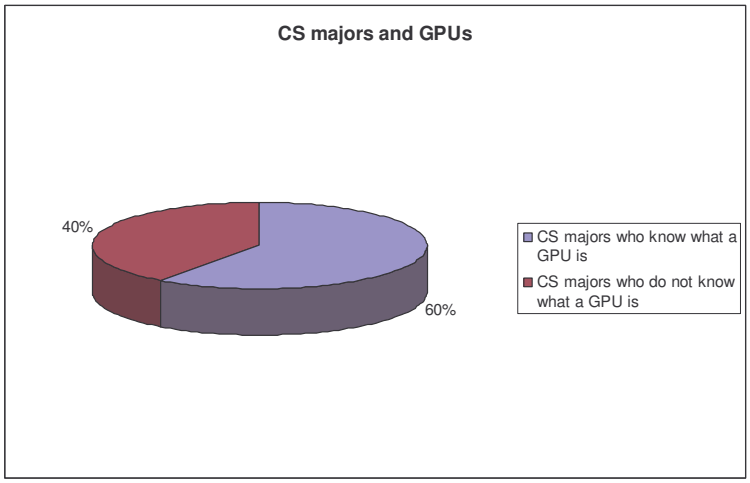

Figure 43: Percentage of CS majors who know what a GPU is

The chart in Figure 43 shows that 60% of the people surveyed who major in CS do know about GPUs. Of the people who know what a GPU is are the ones who had some idea on how graphics are rendered and this has been shown in Figure 44.

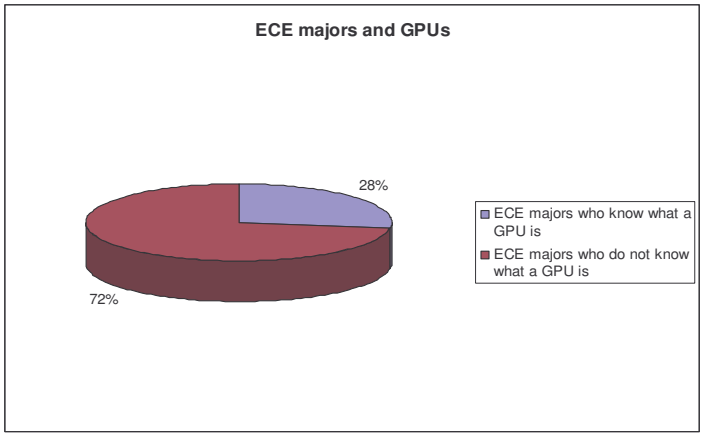

Figure 44: Percentage of ECE majors who know what a GPU is

Similarly the chart in Figure 44 suggests that almost one third of the people who major in ECE know what a GPU is. The above two charts clearly suggests that students who have good background in computer science will have a better chance of understanding the working of a GPU. The statements made above are supported by the pie chart in Figure 45 which shows us that only 10% of the people from other majors know what a GPU is.

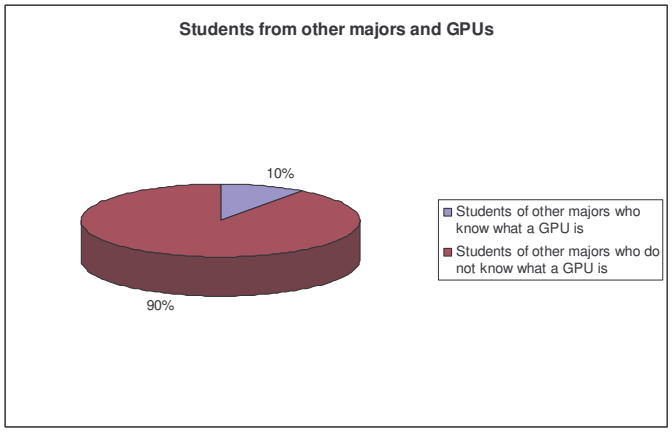

Figure 45: Percentage of students from other majors who know what a GPU is

The chart in Figure 46 shows what kind of graphics cards are being used the most these days. The findings are very interesting. About 52% of the people who own video game consoles have an ATI GPU in them. This is interesting as Nvidia is a bigger company which ships more units when compared to ATI.

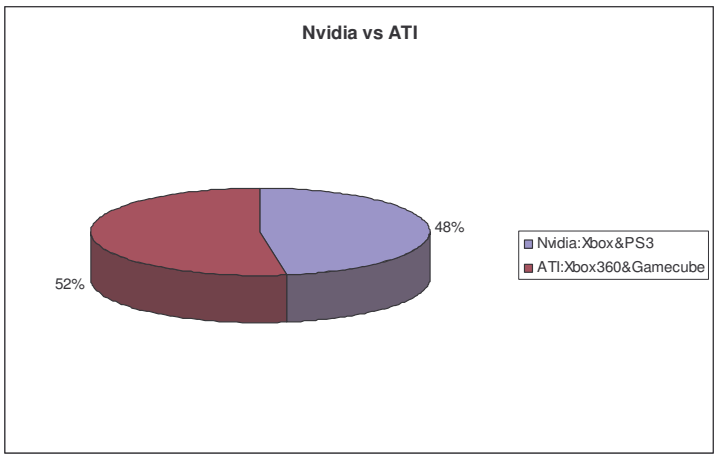

Figure 46: Types of GPUs being used in consoles

This survey has helped us understand that GPUs are not as popular as their applications are. Also the types of GPUs used in video game consoles are discussed owing to the tough competition in the GPU market. The market share of Nvidia and ATI are compared in WPI. Interestingly, ATI has a slightly higher share than Nvidia in WPI in spite of Nvidia shipping more units than ATI. It was also observed that many students from CS and ECE departments knew what a GPU is. This is probably because their majors are closely related to GPUs. There were more students than expected who play on PC rather than on video game consoles. Though very few people knew what a GPU is, many people are interested in knowing more about a GPU from magazines, flyers and presentations. A good percentage of the students also wanted to know more about GPUs through classes.

13. Conclusion

Initially, GPUs started off as an aid to CPUs in relieving their load from graphics processing with their main usage then being in the gaming industry. Later on, its importance in PC industry increased when the gaming trend transformed from 2D to 3D. Introduction of video game consoles further increased the number of gamers and catalyzed the growth of PC games as well. Observing this rapid growth in gamers and a growing market for GPUs, many companies entered this field. This initiated a GPU war between companies in producing better and more powerful graphics cards in order to capture the market. Due to this intense competition many small companies went bankrupt and vanished over time. Only a few companies were able to survive and have since stood firm in this GPU market. The key players are ATI, Nvidia and Intel. Over the past five years Intel has been the leader in graphics market due to its integrated graphics card on their motherboard. Since, there has been fierce competition between ATI and Nvidia for the second pole position. These graphic giants have been so efficient and rapid in their research and development that at an average of every six months they introduce graphics chips with twice the improvement in processing power. As a result, the processing power of a GPU today has surpassed the processing power of the current fastest CPU by a large factor.

Recent developments have proved that the processing power of GPU can be harnessed for tasks other than graphics processing, also known as General Purpose GPU. A very important breakthrough was in finding a cure for the Alzheimer's, Huntington's and Parkinson's disease, where protein folding is simulated with calculations being carried out at a rate of 10 gigaflops per second. Similarly, GPUs are being used by military in flight simulators, databases and many other applications. Recently, algorithms have been suggested which are claimed to simulate

weather of a region. It means that man will be aware of any natural disaster, such as storms and tornadoes, long before they come. This shows that there is a very high potential in this technology and can possibly revolutionize many fields.

A technology which can be placed next to invention of computers in order of significance should not be underestimated by the current generation of people. It is important that people have some knowledge about this amazing chip technology and its immense potential. This way more people can help in its development and its uses as GPGPU and help in the betterment of mankind and the world we live in.

Our survey at WPI showed that only a few people knew about this technology and rarely anyone knew about its potential as a powerful computational device. Most of the students who knew about GPUs were only ECE or CS majors. The most interesting result that we obtained from the survey was the curiosity shown by most of the students in wanting to learn more about GPUs. After the completion of this project we feel that there is a great need for an awareness program for GPUs not only in technological universities but also in the industrial world.

Bibliography

[AMDFR] AMD. "AMD Financial Report." AMD Announces Q4 Losses Amid Pricing War. PC Magazine Online, Found 24 January 2007.

[AMDFA] AMD. "AMD Launches First Combo Chipset From ATI Acquisition[." Advanced Micro Devices.'s 690 series, ATI Technologies Inc. Information Week, Found February 2007.

[BRAUGH] Associate Professor Braught, Grant William. Computer Science 354, Operating Systems. Fall Semester 1997. <http://dickinson.edu/~braught/courses/cs354f97/Classes/Class02/Class02LN.html>.

[BANDI] Bandi, Nagender, et al. "Hardware Acceleration in Commercial Databases: A Case Study of Spatial Operations." 30th VLDB Conference. Toronto, Canada, 2004. 1-12.

[BRUNER] Bruner, Richard W. In graphics chips, even Intel has to prove itself (i740 chip) (Product Information). Vol. 24 n5. Electronic Business. Cahners Publishing Company, May 1998.

[BUSWIR] Business Wire. Nvidia Announces Design Wins For Industry-Leading Geforce 256 Graphics Processing Unit, GPU. Palm Springs, California, August 1999.

[CHARAL] Charalambous, Maria, Pedro Trancoso and Alexandros Stamatakis. Initial Experience Porting a Bioinformatics Application to a GPU. 2006. Dept. of Computer Science. Found October 2006. <http://www.cs.ucy.ac.cy/~pedro/publications/pci05-raxml.pdf>.

[COMTEX] Comtex. EDGAR Online-Glimpse. Found 16 March 2007. <http://infotrac.galegroup.com/itw/infomark/469/355/95408311w1/purl=rc1_BCPM_0_A160648239&dyn=53!xrn_27_0_A160648239?sw_aep=mlin_c_worpoly>.

[DAMIEN] Damien. BE Hardware. 27 July 2005. <http://www.behardware.com/news/7718/gpu-market-shares.html>.

[DANG] Dang, Alan. History of NVIDIA. 9 February 2001. <http://www.firingsquad.com/features/nvidiahistory/>.

[DATAMO] Data Monitor. "Nvidia Corporation - Company Profile." Nvidia Corporation - Company Profile. Data Monitor, August 2006. 4-6.

[DELVES] Delves, James. Nvidia and Havok Demonstrate World's First GPU-Powered Game Physics Solution at Game Developer's Conference. Found 20 March 2006. <http://www.nvidia.co.uk/object/IO_30476.html>.

[EXTEDT] ExtremeTech-Editor. Intel-Integrated Graphics. Found November 2006. <http://www.extremetech.com/article2/0,1697,1741173,00.asp>.

[FIALKA] Fialka, Ondřej. Tone Mapping Operators on GPU. Found October 2006. <http://netra.felk.cvut.cz/Zope/cgg/publications/diplom/FialkaOndrej/abstract_html>.

[FUNG] Fung, James. "Computer Vision on the GPU." GPU Gems 2. Ed. Matt Pharr. Addison Wesley, 2005. 649-665.

[GEER] Geer, David. "Taking the Graphics Processor beyond Graphics." September 2005. IEEE Computer Society. Found October 2006. <http://ieeexplore.ieee.org/iel5/2/32339/01510560.pdf?arnumber=1510560>.

[GOVIND] Govindaraju, Naga K., et al. Fast Computation of Database Operations using Graphics Processors. ACM SIGMOD. June 2004. <http://gamma.cs.unc.edu/DB/main.pdf>.

[GPGPUAS] GPGPU / Audio and Signal Processing. BionicFX uses GPU as Powerful Audio Effect Processor. Found March 2007. <http://www.gpgpu.org/cgi-bin/blosxom.cgi/Audio%20and%20Signal%20Processing/index.html>.

[GPGPU] GPGPU. General Purpose GPUs. Found October 2006. <http://www.gpgpu.org>.

[HOUSTO] Houston, Mike. Genaral Purpose Computation on Graphics Processors (GPGPU). 2005. Found October 2006. <http://graphics.stanford.edu/~mhouston/public_talks/R520-mhouston.pdf>.

[JENKIN] Jenkins, Henry. "Eight Myths About Video Games Debunked." The Video Game Revolution. Found November 2006. <http://www.pbs.org/kcts/videogamerevolution/impact/myths.html>.

[JOHNSO] Johnson, Steven Berlin. "Serious Games." 10 October 2005. IT Conversations. Found October 2006. <http://www.itconversations.com/shows/detail774.html>.

[JONES] Jones, K.C. "Game consoles to power cancer, Alzheimer's research." 29 August 2006. iTnews.com.au. Found April 2007. <http://www.itnews.com.au/newsstory.aspx?CIaNID=36364&s=GPU>.

[KILGAR] Kilgariff, Emmett and Randima Fernando. "The GeForce 6 Series GPU Architecture." 2005. GPU Gems 2, Nvidia Corporation. Found October 2006. <http://download.nvidia.com/developer/GPU_Gems_2/GPU_Gems2_ch30.pdf>.

[LEFOHN] Lefohn, Aaron, Joshua Cates and Ross Whitaker. "Interactive, GPU-Based Level Sets for 3D Brain Tumor Segmentation." 2003. School of Computing, University of Utah, Salt Lake City. Found October 2006. <http://www.cs.utah.edu/research/techreports/2003/pdf/UUCS-03-004.pdf>.

[LEWIS] Lewis, George. "Researchers tout positive effects of video games." 19 May 2005. MSNBC. <http://www.msnbc.msn.com/id/7912743/>.

[MEDIAL] Media Analysis Laboratory, Simon Fraser University, Burnaby B.C. Video Game Culture: Leisure and Play Preferences of B.C. Teens. October 1998. Found November 2006. <http://www.media-wareness.ca/english/resources/research_documents/studies/video_games/video_game_culture.cfm>.

[NVIGPU] Nvidia (GPU). Graphics Processing Unit (GPU). Found October 2006. <http://www.nvidia.com/object/gpu.html>.

[NVIIMP] Nvidia (Imprint). University of Washington and Imprint Interactive Technology. Found December 2006. <http://www.nvidia.com/object/uw_imprint_success.html>.

[PAPAKI] Papakipos, Matthew. "Converging Design Features in CPUs and GPUs." HPC Wire. Found April 2007 <http://www.hpcwire.com/hpc/1209133.html>.

[REID] Reid, Keith. "Visualization in Medicine." 29 September 2006. Advanced Imaging <http://www.advancedimagingpro.com/publication/article.jsp?pubId=3&id=3301>. Supplement. Found April 2007.

[ROGERS] Rogers, Phil. In The Hot Seat: ATI's Phil Rogers ExtremeTech.com. Ziff Davis Media Inc, 14 March 2007.

[SALVAT] Salvator, Dave. The Return of S3. Found November 2006. <http://www.extremetech.com/article2/0,3973,1417276,00.asp>.

[SBERT] Sbert, Mateu and Jordi Palau. GameTools Advanced Tools for Developing Highly Realistic Computer Games. Universitat de Girona, Girona, Spain. Found November 2006. <http://www.gametools.org/archives/publications/GameTools_MateuSbert_ITRA.pdf>.

[SGI] SGI. Tackling Large Volume Visualization Challenges /w Real Time Ray-Tracing. January 2005. October 2006. <http://www.sgi.com/pdfs/3883.pdf>.

[SHERMA] Sherman, Dr. Richard C. "Video Games." 17 April 2002. Psybersite at Miami University. Found October 2006. <http://www.units.muohio.edu/psybersite/cyberspace/onlinegames/video.shtml>.

[THEECO] The Economist. "Chasing the dream." 4 August 2004. Found on October 2006. <http://www.economist.com/displaystory.cfm?story_id=4246109>.

[UJALDO] Ujaldon, Manuel, Simon Gregor Ebner and Joel Saltz. On the capabilities of the GPU for general purpose computing. 2004. Found October 2006. <http://bmi.osu.edu/resources/techreports/osubmi_tr_2004_n18.pdf>.

[VANBUR] VanBuren, Brian G. Graphics Processing Units. 10 November 2004. <http://www.rit.edu/~bgv5143/gpu.pdf>.

[WIKI01] Wikipedia [GPU Manufacturers]. GPU Manufacturers. Found October 2006. <http://en.wikipedia.org/wiki/Graphics_processing_unit>.

[WIKI02] Wikipedia [GPU]. GPU. Found October 2006. <http://en.wikipedia.org/wiki/GPU>.

[WIKI03] Wikipedia [SIMD]. Single Instruction Multiple Data. Found October 2006. <http://en.wikipedia.org/wiki/SIMD>.

[WILLIA] Williams, David E. PlayStation's Serious Side: Fighting Disease. Found October 2006. <http://www.cnn.com/2006/TECH/fun.games/09/18/playstation.folding/index.html>.

[YI] Yi, Matthew. San Francisco Chronicle. 15 August 2003. <http://sfgate.com/cgi-bin/article.cgi?f=/c/a/2003/08/15/BU190365.DTL>.

[YOUNG] "The Outsourcer." Young, Jeffrey. Nvidia Corp's outsourcing success / Company Business and Marketing. Forbes, 1995. 344.

Appendix A: Glossary

Algorithm A step-by-step problem-solving procedure used for solving a problem in a finite number of steps.

Bioinformatics The use of computer science, mathematics, and information theory to model and analyze biological systems, especially systems involving genetic material.

Blitter Blitter (from BLIT or Block Image Transfer) is a co-processor chip dedicated to memory data transfers, usually independently of the CPU using bit blit methods.

Branch These are programming techniques like the IF-THEN-ELSE statements which jump form one part of the program to the other depending on the conditions.

Cache This is the memory of the CPU that can be accessed in very little time.

Cycles Cycles (or instruction cycles) refers to the time period during which one instruction is fetched from memory and executed when a computer receives a machine language instruction; or the sequence of actions that a CPU performs to execute each machine code instruction in a program.

Databases A collection of data arranged for ease and speed of search and retrieval by a computer.

Data Registers These are used as accumulators and to store integers in temporary memory.

GFlops This is an acronym for Giga Floating Point Operations per second which is a measure of a computers computational performance when tedious calculations are made.

Pipeline	A chain of processing elements (processes, threads, co-routines, etc.), arranged so that the output of each element is the input of the next. Usually some amount of buffering is provided between consecutive elements.
Pixel	The basic unit of the composition of an image on a television screen, computer monitor, or similar display.
RAxML	Randomized Axelerated Maximum Likelihood.
Vertex	It is the representation of a point in the GPU.

Appendix B: Survey Questions

Survey

Gender: ☐ Female ☐ Male

Major: ☐ Electrical & Computer ☐ Computer Science ☐ IMGD ☐ Others _____

Class Year: ☐ Freshmen ☐ Sophomore ☐ Junior ☐ Senior ☐ Others _____

1. Do you own a computer?
   ☐ Yes ☐ No
2. If your answer to Question 1 is yes, what is the name of the manufacturer of the graphics card in your PC?
   ☐ Nvidia ☐ ATI ☐ Intel ☐ Others ☐ I don't know
3. Do you play video games?
   ☐ Yes ☐ No
4. Do you own a video game console?
   ☐ Yes ☐ No
5. If your answer to Question 3 is yes, on what platform do you play your game on?
   ☐ GameCube ☐ PlayStation1 ☐ Xbox ☐ Wii ☐ PC
   ☐ PlayStation2 ☐ PlayStation3 ☐ Xbox 360
6. What is your favorite game?
   ☐ Counter Strike ☐ World of Warcraft ☐ NFL Madden ☐ Legend of Zelda
   ☐ Dance Dance Revolution ☐ Final Fantasy Others _____
7. Do you know any kind of computer programming?
   ☐ Yes ☐ No
8. What programming language are you proficient in?
   ☐ Java ☐ C ☐ C++ ☐ Python ☐ BASIC
9. Do you know the process by which graphics are drawn to your PC?
   ☐ Yes ☐ No
10. Do you know what a hardware Graphics Processing Unit is?
    ☐ Yes ☐ No
11. By how many hours per week have you increased playing video games because of improvement in graphics?
    ☐ less than 10 ☐ 10-20 ☐ 20-30 ☐ more than 30
12. In which field do you think that GPU's should be used in the near future?
    ☐ Medical Imaging ☐ Military Computations ☐ Databases
    ☐ I don't care ☐ Others_____
13. Do you want to learn more about Graphics Processing Units?
    ☐ Yes ☐ No
14. In what way would you like to learn more about Graphics Processing Units?
    ☐ Courses ☐ Presentations ☐ Flyers ☐ Magazines ☐ Others_____

Appendix C: GPU Awareness Flyer