Carnegie Mellon University

Parallel Advances in Modern Computing Hardware and Video Game Development

A glimpse at how video games have given us powerful computational abilities

Advaith Sethuraman

76-101: Section RR

Mr. Craig Stamm

12/14/2017

Abstract

The public opinion on the effect and place of video games in society are that they are primarily a tool for recreation. However, further analysis of how certain computer hardware developers have responded to the increasingly complicated nature of video games has led to discoveries that video game development influences computer hardware directly. Innovations in computer hardware has direct implications in the fields of Machine Learning, Cryptography, and Parallel Computing. By finding profound applications of these technologies in modern day society, we can argue that the development of video games has led to tangible difference in society, albeit in a different sector than originally intended by video game developers.

Introduction

Since the beginning of video games, the main driving factor for video game companies was to create a game that would be enticing enough to warrant players spending physical money on a digital consumable. The primary motivation for companies to create a video game was to generate profit. Furthermore, this puts the purpose of video games into question. Although some people consider video games to be the distribution of digital art and a vessel in which developers can communicate their message to the players, there exists the counterargument that video games are not art and serve no useful purpose in modern society (Ebert). Although video games can be written off as only for recreational purposes and as having no immediate impact to the important functions of society, the effects and acceleration of society through the development of various video game hardware is profound. Namely, advances in the field of graphics processing units in desktop and mobile computers have far reaching social and economic effects in our immediate lives.

Graphics Cards and their Development

Within the gaming desktop archetype, the performance of a computer is highly dependent on the hardware present within it. In fact, the requirements of modern day video games require highly specialized hardware made specifically for the purposes of rendering and outputting high quality graphics. Often, the hardware required to perform the intense computations of rendering and processing various video game elements such as shaders, textures, and animations is in the form of a graphics card.

When the term "graphics processing unit" or GPU was first coined by the computer hardware company Nvidia in 1999, the state of computer hardware was very different from what it is at now. However, it is important to look backwards to get a better idea of how the advances in graphics card technology has affected very important parts of our society. The first graphics cards produced by Nvidia (Geforce 3) were unique because of the fact that they were the only ones on the market to include programmable shading, a technique that involves running a specific script for each pixel before outputting to the display which results in a variety of textures and color outputs that were not possible before. Features like this in Nvidia's line up foreshadowed their emergence as the market share leader in graphics card technologies. Between the years of 1999, major development in computer hardware occurred. According to Moore's Law, the number of millions of transistors on an Integrated Circuit doubles every 2 years which in turn means that processing power experiences significant increases as well (Aizcorbe, Kortum). The same principle can be applied to the development of graphics cards. Modern day GPUs are far more powerful than those found 10 years ago. Indeed, the specialized hardware architecture of graphics cards lends them easily to niche tasks that are performed with incredible ease when using GPUs but take far more computational power when performed with traditional processing units.

Before we discuss the specific tasks that graphics cards excel at over traditional computing hardware, we must fully understand the hardware architecture of a graphics card. The fundamental difference between a CPU (Central Processing Unit) found in all computers and the specialized hardware found in a GPU is the fact that there are hundreds of cores operating simultaneously to render and compute each pixel. In a Central Processing Unit, there are a limited number of cores (usually 4 or 2), which significantly limits the throughput from the unit. Furthermore, the standard of measuring computing power is called the FLOP (Floating Point Operations per second). The numerical computational power of a CPU is in the range from 25-40 Giga FLOPs, whereas the computational power of a modern GPU is pushing the boundaries at 11 tera FLOPs. Clearly, the difference in computational power in significant. Since graphics cards were originally developed for use in graphics and video game consumption, the incredible performance aspects of the hardware are a positive side effect of the demand created by video game developers. Thankfully, companies such as Nvidia and AMD have filled the demand with their developments in computing hardware (Zhou).

It is important to note the cause of the development of graphics card technology. The research behind creating new graphics card hardware is very closely related to the changes and innovations in newly released video games. We can track the morphing of video games from 2D to 3D and consolidate them with new graphics card releases to create a clearer correlation between the two. The first widely recognized 3D video game was *Wolfenstein 3D* which was released in 1992. This set the precedent for further 3D game development. Subsequent video games include *Quake* which included boss battles with complex 3D characters and graphics.

Furthermore, it is important to note the specific change in complexity in 3D video game development. The complexity of the calculations the computer needed to perform did not change. Rather, as graphics quality increased, the number of computations the computer had to perform increased. As video games began including new graphics properties such as shaders, which were individual mathematical functions that changed individual pixels on the final rendered image, it became clear that a more powerful processing unit was not effective. Rather, a processing unit that could handle multiple redundant calculations was in demand. As Clark expands on his paper on how capitalist organizations handle demand for a new product in a market, "capitalist production is not marked by the subordination of social production to social need, even as that is expressed in the restricted form of 'effective demand' in the market, for the purpose of the capitalist is not to meet social need, but to expand his capital" (Clark). Clearly, the motivation for computer hardware developers making graphics cards was not to provide for a social good such as furthering scientific research or providing hardware for financial and pharmaceutical companies to create their statistical models upon. Instead, graphics card companies saw the demand in the market, and chose to "expand their capital" by fulfilling that demand (Clark). As Adam Smith expands upon in his book *The Wealth of Nations*, development of technology is beneficial to the public because it is "led by an invisible hand to promote an end which was no part of his intention" (Smith). Essentially, when video game developers saw the potential for gaining capital in the form of the video game market, they chose to fulfill it and as a result had positive side effects in sectors of society that they had no intention of affecting.

The demand for better hardware for video game developers to run their games on created a new market for graphics hardware that is motivated by capitalist gains. Although development in computer hardware is mainly motivated by capitalist values, the uses for the resulting hardware are far reaching and more profound than what one can call recreational.

Unintended Uses for Graphics Cards

The number of people who use computers continues to increase each year as it becomes easier and easier to purchase and maintain machines. Furthermore, as the world moves into the digital age, leaving paper and books behind, the applications for computer technology only continues to grow. Given the architectural limitations of graphics cards, there exists a very niche sector in which their technology can be applied. However, although their applications are very narrow, they excel in each of those specific roles. Graphics card usage in the fields of Machine Learning, Parallel Computing and Cryptography are directly applicable to our everyday lives.

First, we must consider the generic use case of Parallel Computing. Simply put, parallel computing is any application when a computer executes tasks in tandem, or in parallel instead of sequentially. The benefits of this are the ability to handle various streams of information at the same time or to output various calculations on different variables without having to wait for the previous operation to terminate before starting a new one. As such, parallel computing has many uses when it comes to cryptography of statistical modelling. Specifically, we will be focusing on the use of GPU acceleration to handle intense computations for a method called the Monte Carlo Method. The Monte Carlo Simulation is a statistical package which has widespread uses in finance, biology, and game theory. At its core, it is a method which uses repeated randomized sampling to produce a data set from which algorithms can draw conclusions upon. Research that takes place using Monte Carlo methods are very important for financial and medicinal simulations for applications like stock investment and pharmaceutical trials.

Given the applications of the Monte Carlo method, time is of the essence. Any hardware that can reduce the time required to run the simulation saves money and sometimes lives. The reason GPUs are so effective in running the Monte Carlo simulation is highly technical. Since GPU's have "up to 30 multiprocessors per card in response to commercial demand for real-time graphics rendering", they have evolved to become very efficient for running redundant algorithms (Lee). Furthermore, there are various reasons why using a GPU for statistical research is convenient. Since graphics cards were made for the consumer market, the are "readily obtainable from consumer-level computer stores" and their low energy consumption means that researchers do not need specialized hardware or have massive computer clusters put into use (Lee). Researchers then use CUDA (Compute Unified Device Architecture) to accelerate the computations by splitting them between all the multiprocessing units. The usage of GPU acceleration increases the speed of calculations by "500 times on the 8800 GT and 800 times on the GTX 280", where 8800 GT and GTX 280 are both Nvidia graphics cards at the time of publication (Lee).

Specific applications of parallel Monte Carlo method simulations include direct applications to financial modeling. Many investment firms use a subset of mathematics called stochastic volatility modeling. In short, this model is used to predict a market when there are changing variances, or constantly changing variables that the model is dependent upon. The speedup using a GTX 280 CPU when compared to a flagship CPU is 8x when using 8192 as the number of inputs for the randomization (Lee). Although the time cut down is impressive at 8192, the true improvement occurs when increasing the number of inputs for randomization. At 131,072, the GPU offers a 10x speed up factor (Lee). When dealing with large data sets, this means that the time required to produce a statistical model for a market will be cut down from 10 days to 1 day. Clearly, if investment firms create faster and more accurate models for the market, it will allow them to generate more profit and stimulate growth in the economy. Again, as Adam Smith says, the side effects of investment firms' practices include generating profit for shareholders and end users alike, although that will not be the immediate intention when investing.

Another important pillar of society is being able to transmit information securely. The field of cryptography involves converting information to an unrecognizable form and reverting it to a usable format. Cryptography has profound implications in everyone's daily lives. Every time an individual uses a credit card, accesses their bank account, or accesses a government website, they are making use of cryptography.

The basis of encryption used in modern day devices involves the factorization of a number into smaller primes, or numbers that have factors of 1 and themselves. Taking RSA encryption as an archetype, we can see that the sender has a public and private key. The private key is sent to the receiver so that they can decrypt the message. The encryption algorithm itself works by using a function that gives the remainder when divided by a number and by using two large prime numbers, which is called modular arithmetic. The point of using RSA encryption is that any brute force method of guessing the prime numbers that will allow a hacker to decrypt the message without the private key takes more time than is practical (on the order of a few lifetimes).

Another example of a popular encryption algorithm is called Elliptical Curve Method for factorization. As Bernstein details in his research papers, "modular arithmetic is the main bottleneck in computing scalar multiplication in ECM" (Bernstein). Clearly there is a shortcoming when using CPUs for modular arithmetic. GPUs can optimize this by using a "8-

way modular arithmetic unit that is capable of carrying out 8 modular arithmetic operations simultaneously" (Bernstein). Again, we see the parallel computational abilities of the GPU reduce the time required to perform certain calculations. The factoring of primes occurs at 22.66 * 10^6 operations per second using a GTX 280 GPU versus 7.85*10^6 operations per second using an Intel Core 2 Duo CPU (Bernstein). The consequence of this is a 3x speed up factor.

From here there are two possibilities. The Elliptic Curve Method is used for finding prime with which you can decrypt encrypted messages. It is not necessary if you are provided with the private key. There are two cases when decryption time becomes essential. Case 1: suppose the government is intercepting messages from a foreign entity that has immediate consequence to national security. It would be extremely important to be able to decrypt the message in the shortest amount of time possible since national security is at risk. Clearly, GPUs provide the speed factor necessary to make this possible. Case 2: suppose a hacker is using GPU acceleration to decrypt the credit card numbers of United States citizens. Again, a GPU would provide the speed factor necessary to perform this operation, but is it ethical? The development of technology often occurs without foresight into what its applications are. This is true of the atom bomb (technology developed by Einstein but not intended for destruction) and we must leave it to the end users to decide on the ethicality of discrete uses for technology.

Finally, graphics cards have found a new home in the field of Machine Learning. Machine Learning is using previously recorded information to train and predict future occurrences of the same event. The algorithm involves loading the "training data onto the GPU" then "running the GPU shaders to make up the learning algorithm" (Steinkraus). Machine Learning takes advantage of a technology that was originally developed to help create varied textures for individual pixels for video games. The far reaching effects of graphics card technology are now evident as we observe the usage of shaders for Machine Learning algorithms. The results of the experiments carried out by Bernstein and his colleagues at Microsoft Research are clear: they yield a 3x speedup for a 2-level neural network. The difference given by GPUs can be seen in various applications of Machine Learning. One example is the usage of Machine Learning in medical diagnoses. As Kononenko elaborates in his research on how Machine Learning techniques can be effectively applied to the medical field, Neural Networks can be used to diagnose diseases in "oncology, liver pathology, thyroid diseases, rheumatology, craniostenosis syndrome, cardiology, neuropsychology, gynecology, and perinatology" (Kononenko). Specifically, the usage of techniques similar to the Monte Carlo Method such as Bayesian Classifiers yields considerable insights into medical data. The usage of graphics cards in the field of Machine Learning and its applications leads to performance increases that can result in the potential diagnosis of complicated diseases in a much shorter time period than traditional methods.

Clearly, the usage of specialized video game hardware is extremely useful in applications to statistical modeling, and is completely the side effect of consumer demand for better graphics rendering. The unintended uses of graphics cards in the financial sector, medical diagnoses, and cryptography all have immediate and profound consequences in our daily lives.

The Future of Graphics Processing Units and their Limitations

The primary reason GPUs have an upper hand over comparable CPUs is because of the number of cores and the fact that computations that need massive parallelization run much faster on the leaner, simpler cores of a GPU. However, following Moore's Law, which states that the number of transistors on a given die of processing unit will double every 2 years, we can expect to reach a limit (Aizcorbe, Kortum). There is a physical limitation to the microscopic size to which transistors can be made. Once this limit is reached, chip designers will either need to increase the area of the processing card, or look towards alternate computer architecture to achieve the same results. Currently, a GPU may look like a perfect alternative to a CPU, but there are several drawbacks. First, a GPU excels in a certain subset of computations that a computer needs to perform to function properly. Given a task that requires multiple threads (or different calculations that converge), a GPU will perform better than a CPU. However, the limitations exist in the amount of cache (or free memory) that a GPU has to operate with compared with a CPU. Since a CPU has more cache, it can perform more labor intensive calculations that a GPU does not have the resources to perform.

Since GPUs have already made an impression on academics and researchers globally, we can predict that they are here to stay in the field of serious scientific research. Another example of how Nvidia responds to market conditions can be seen in the release of the Titan V GPU. Nvidia has developed their own Nervana chips made specifically for Artificial Intelligence and Machine Learning algorithm implementation. Thus we can see Nvidia coming full circle and underlining the theory that the development of video games led to an emerging market of various uses for GPUs. Nvidia, motivated by capitalist gains, responds well to the demands set forth by the market, as illustrated by their pivoting or horizontal expansion into hardware made specifically for the unintended uses for their original paradigm.

Although the innovation presented by Nvidia seems to be benefiting multiple people in society, we must be wary of monopolization. Intel currently has the lion's share of the market at 67% while Nvidia has 20%. When researches begin turning towards Nvidia for more specialized applications for processing units, it creates more demand for Nvidia over Intel, since they are mutually exclusive (you cannot run an Intel and Nvidia GPU at the same time). As Nvidia's market share increases and Intel cannot infringe upon their patented GPU architecture, it will naturally create a monopoly. Although Sherman Anti-Trust laws will prevent a complete monopoly, Nvidia achieving a lion's share of the market will lead to unfair regulations and throttling of computer hardware, which may not be a future the free consumer wishes to live in.

The future seems bright when there is a strong connection between the people who create profound change in the world (scientific researchers) and the industry. Their interactions can be exemplified by a symbiotic relationship in nature. As the industry continues to improve their computer hardware, the academics will continue to find new innovative uses that the industry never considered. As emerging markets develop, the industry will continue to make more specialized hardware to fill the demand to further their capitalist gains. This positive feedback loop is a win-win situation for both society and industry. Clearly, the relationship between scientific researchers and the industry in the context of graphics cards is a de facto example of how Adam Smith's 'invisible hand' provides a means for the market to benefit and reciprocate the innovation found in industry.

Works Cited

1. Aizcorbe, Ana, and Samuel Kortum. “Moore's Law and the Semiconductor Industry: A Vintage Model.” *The Scandinavian Journal of Economics*, vol. 107, no. 4, 2005, pp. 603–630. *JSTOR*, JSTOR, www.jstor.org/stable/3441017.

3. Bernstein, Daniel J, et al. “ECM on Graphics Cards" *Department of Computer Science University of Illinois at Chicago* pp.484-501 https://link.springer.com/content/pdf/10.1007/978-3-642-01001-9_28.pdf

4. Clarke, Simon “Marx and the Market” *University of Warwick, Coventry* pp.1-31 https://homepages.warwick.ac.uk/~syrbe/pubs/LAMARKW.pdf

9. Ebert, Roger. "Video Games Can Never Be Art." *Roger Ebert*. N.p., 16 Apr. 2010. Web. Apr. 2015.

5. Hutchison, Terence. “Adam Smith and The Wealth of Nations.” *The Journal of Law & Economics*, vol. 19, no. 3, 1976, pp. 507–528. *JSTOR*, JSTOR, www.jstor.org/stable/725079.

7. Igor Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective”, In Artificial Intelligence in Medicine, Volume 23, Issue 1, 2001, Pages 89-109, ISSN 0933-3657, https://doi.org/10.1016/S0933-3657(01)00077-X. http://www.sciencedirect.com/science/article/pii/S093336570100077X

2. Lee, Anthony, et al. “On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods.” *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, 2010, pp. 769–789. *JSTOR*, JSTOR, www.jstor.org/stable/25765373.

8. Steinkraus, “Using GPUs for Machine Learning Algorithsm/” *IEEE* 2005 http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1575717&tag=1

6. Zhou, Hua, et al. “Graphics Processing Units and High-Dimensional Optimization.” *Statistical Science*, vol. 25, no. 3, 2010, pp. 311–324. *JSTOR*, JSTOR, www.jstor.org/stable/41058950.

Revision Report

Based on Instructor Feedback:

- Removed ‘in conclusion’ from essay
- Expanded about how the market created graphics cards
- Added 2 more examples
- Elaborated upon why GPUs are not replacements for CPUs
- Counter argument for uses of GPUs

Parallel computing is the Computer Science discipline that deals with the system architecture and software issues related to the concurrent execution of applications. It has been an area of active research interest and application for decades, mainly the focus of high performance computing, but is now emerging as the prevalent computing paradigm due to the semiconductor industryâ€™s shift to multi-core processors. A Brief History of Parallel Computing.Â nToday, parallel computing is becoming mainstream based on multi-core processors. Most desktop and laptop systems now ship with dual-core microprocessors, with quad-core processors readily available. Parallel Computing is an international journal presenting the practical use of parallel computer systems, including high performance architecture, system software, programming systems and tools, and applications. Within this context the journal covers all aspects of high-end parallel computing that use multiple nodes and/or multiple accelerators (e.g., GPUs). Parallel Computing features original research work, tutorial and review articles as well as novel or illustrative accounts of application experience with (and techniques for) the use of parallel computers. We also welcome studies reproduc Parallel computing allows us to solve large problems by splitting them into smaller ones and solving them concurrently. Parallel computing was considered for many years the â€œholy grailâ€ for solving data-intensive problems encountered in many areas of science, engineering, and enterprise computing; it required major advances in several areas, including algorithms, programming languages and environments, performance monitoring, computer architecture, interconnection networks, and last but not least, solid-state technologies. Parallel hardware and software systems allow us to solve problems deman