UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Mechanical Engineering

OPTIMIZATION OF FAN LOAD VIA CORE CLOCKING OF A TRIPLE FAN GPU SYSTEM

Haziq Haiqal bin Hanif

Master of Mechanical Engineering

2021

OPTIMIZATION OF FAN LOAD VIA CORE CLOCKING OF A TRIPLE FAN GPU SYSTEM

HAZIQ HAIQAL BIN HANIF

A thesis submitted
in fulfillment of the requirements for the degree of
Master of Mechanical Engineering

Faculty of Mechanical Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2021

DECLARATION

I declare that this thesis entitled " Optimization of Fan Load Via Core Clocking of a Triple Fan GPU System" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : 

Name : HAZIQ HAIQAL BIN HANIF

Date : 7/3/2022

APPROVAL

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Mechanical Engineering.

Signature : 

Supervisor Name : DR. MUHAMMAD ZULFATTAH BIN ZAKARIA

Date : 7 MAC 2022

DEDICATION

This project I dedicated to my beloved family and supervisor upon their nonstop support and motivation in every part of my life

ABSTRACT

Throughout the decade, the graphics processing unit (GPU) has seen numerous innovative advancements. The processing capacity of a GPU may compete with that of an existing Central Processing Unit (CPU) when it comes to running high-profile software. Any electronics that pass current through it, on the other hand, will generate heat. The biggest issue with GPUs is the thermal issue. The best GPU performance requires the lowest fan speed and core temperature while maintaining the most efficiency (hashing power). The goal of this study was to use the Design-Expert software's optimization function to get the best response at a specific core clock and memory clock. The reaction was recorded when the GPU was receiving a consistent quantity of power. The GPU used is the ASUS TUF RTX3060 OC, which is a triple fan GPU. The link between GPU responses (fan speed, core temperature, hash rate) and clocking (core and memory) was discovered. Central Composite Design (CCD) came up with a single equation for each fan speed, core temperature, and hash rate as a result of this. The optimization method then proposes a number of new clock settings that outperform the existing setting. The best core and memory clock were chosen for the validation and confirmation phase. The validation results show that the expected response was accurate, with a margin of error of less than 2%.

PENGOPTIMUMAN BEBAN KIPAS MELALUI PENGOLAHAN TERAS SISTEM GPU TIGA KIPAS

ABSTRAK

Sepanjang dekad ini, unit pemprosesan grafik (GPU) telah melihat banyak kemajuan yang inovatif. Kapasiti pemprosesan GPU mungkin bersaing dengan Unit Pemprosesan Pusat (CPU) sedia ada semasa menjalankan perisian berprofil tinggi. Mana-mana elektronik yang melaluinya, sebaliknya, akan menghasilkan haba. Isu terbesar dengan GPU ialah isu haba. Prestasi GPU terbaik memerlukan kelajuan kipas dan suhu teras yang paling rendah sambil mengekalkan kecekapan yang paling (kuasa pencincang). Matlamat kajian ini adalah untuk menggunakan fungsi pengoptimuman perisian Design-Expert untuk mendapatkan respons terbaik pada jam teras dan jam memori tertentu. Reaksi telah direkodkan apabila GPU menerima kuantiti kuasa yang konsisten. GPU yang digunakan ialah ASUS TUF RTX3060 OC, iaitu GPU tiga kipas. Pautan antara respons GPU (kelajuan kipas, suhu teras, kadar cincang) dan masa (teras dan memori) ditemui. Reka Bentuk Komposit Pusat (CCD) menghasilkan persamaan tunggal untuk setiap kelajuan kipas, suhu teras dan kadar cincang hasil daripada ini. Kaedah pengoptimuman kemudiannya mencadangkan beberapa tetapan jam baharu yang mengatasi tetapan sedia ada. Jam teras dan memori terbaik telah dipilih untuk fasa pengesahan dan pengesahan. Keputusan pengesahan menunjukkan bahawa respons yang dijangkakan adalah tepat, dengan margin ralat kurang daripada 2%.

ACKNOWLEDDGEMENTS

In the Name of Allah, the Most Merciful

To begin, I'd like to express my gratitude and praise to Allah the Almighty, my Creator and Sustainer, for everything I've received throughout my life. Additionally, I'd like to express my gratitude to Universiti Teknikal Malaysia Melaka (UTeM) for providing the platform software for taught courses.

My sincere appreciation to Dr. Muhammad Zulfattah Bin Zakaria of Universiti Teknikal Malaysia Melaka (UTeM) for his encouragement, counsel, and inspiration. His unwavering patience in guiding and imparting priceless insights will be cherished for eternity. Additionally, I would like to express my gratitude to my project colleague at Universiti Teknikal Malaysia Melaka (UTeM), Mr. Muhammad Hazim and Mr. Muhammad Sufi Ridhwan, for their continuous support throughout my journey. Finally, I want to express my appreciation to Dr. Muhd Ridzuan Bin Mansor for their aid and support during this process.

Finally, and perhaps most significantly, I want to convey my deepest gratitude to my devoted parents, Mr. Hanif bin Hanafi and Mrs. Rajuna binti Hasan, for their encouragement and for serving as a pillar of strength throughout my endeavours; their unwavering support, love, and prayers are invaluable. Additionally, I'd want to express my lifelong appreciation to Muhd Razfy, Muhd Hamizan, Uvaraaj, Hakim Naqiuddin, Khairil, Ru vern, and Magdelson for their patience and understanding. Additionally, I'd want to express my gratitude to my coworkers for their encouragement and aid in shouldering the load of my job throughout my study time. Finally, I want to convey my appreciation to everyone who aided, encouraged, and motivated me to complete my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

GPU	Graphic Processing Unit
CPU	Central Processing Unit
IBM	International Business Machines
HPC	High Performance Computing
R&D	Research & Development
RSM	Response Surface Methodology
VRAM	Video Random Access Memory
VRM	Voltage Regulator Module
RAM	Random Access Memory
CCD	Central Composite Design
BBD	Box-Behnken Design
ANOVA	Analysis of Variance
DOE	Design of Experiment
OC	Overclocking
C.V	Coefficient of Variation
LOF	Lack of Fit
PC	Personal Computer

LIST OF SYMBOLS

°C	Degree Celsius
°F	Degree Fahrenheit
β_0	y-intercept
β_1	Slope population model
y	Regression model
X	Random Variables
E	Error
N	Normal Distributed
Gbps	Gigabyte per second
MHz	Mega Hertz
R	Regression
A	Core Clocking
B	Memory Clocking

CHAPTER 1

INTRODUCTION

1.1 Background

Cooling uses around 40% of the total energy consumed by a typical information technology system. In this article, we will discuss the GPU and its computational capabilities. The GPU (Graphics Processing Unit) architecture is exceptionally multi-threaded and is widely utilised for both graphical and now non-graphical tasks. The primary benefit of GPUs over CPUs is that they can execute substantially more floating-point operations (FLOPs) per unit time. GPU computing expands the capabilities of the hardware and boosts its programmability. By offering a competitive pricing or performance advantage, core-GPU can be utilised as a viable alternative and complement to multi-core servers. Indeed, multi-core CPUs and multi-core GPUs may be utilised concurrently to do network coding. Additionally, it is utilised in media streaming servers, which serve hundreds of peers continuously. GPU computing is the method of accelerating general-purpose scientific and engineering applications by combining a GPU (graphics processing unit) and a CPU. NVIDIA was the first company to produce a GPU. CPUs have a limited number of cores that are dedicated to serial computing, but GPUs contain hundreds of tiny cores that are more efficient and optimised for parallel processing. As a result, CPU Plus GPU is a formidable combo. When the code is executed on the system, the CPU executes the serial section and the GPU executes the parallel piece. The GPU is utilised for general-purpose tasks like as mathematics and for gaming.

Most consumers nowadays are dissatisfied with existing technology since it does not fulfil their expectations. Furthermore, in order to develop sophisticated applications, a lot of computing power required. For example, games and high-performance applications such as Solid-Works, CATIA, and other programmed of a similar kind, which are all reliant on strong technologies such as cryptocurrency mining, need the use of more processing power. CPU-bound programmed, on the other hand, would not be able to handle real-time multimedia applications. For the overwhelming majority of personal computers, video

processing is not fast enough to be done by a CPU due to the present architecture of the processor.

In essence, overclocking a GPU improves performance by increasing the speed at which the graphics processor operates. While all GPUs are configured to operate at a particular speed, referred to as the base clock, individual cards often have the capability to exceed the manufacturer's specification. By increasing the clock speed, GPUs temperature will rise and it will use more power. It's critical to strike a balance between increase performance and maintaining a steady temperature for GPU.

1.2 Problem Statement

The overclocking capability of the GPU ensure that a PC can operate at a higher speed or increase its performance without overheating. Essentially, when a PC is accelerated, it generates heat, necessitating the usage of a powerful cooling system. Furthermore, excessive heat can shorten the life of GPUs and affect the performance as well. The GPU's high overclocking capability may provide an increase in speed up to 20% allowing a PC to execute larger applications more quickly. However, if GPUs become too hot when playing games, it will have a significant impact on the Frames Per Second (FPS), as the power will be diverted away from improving its game and toward cooling the GPU. Worst-case scenario, it would just shut down due to overheating without notifying. CPU temperatures should not exceed 80°C (176°F) for gaming these days, and they should be maintained between 75°C and 80°C (167°F and 176°F) on average.

To address this critical issue, a study was conducted to determine how to keep the GPU cool utilising a dual-fan cooling system. Some parameters were given priority in order to ensure that the GPU's thermal efficiency is maintained for maximum performance at an ideal temperature. Although a wide variety of fans were used, just one type of fan was studied. The design of the heat sink and heat pipe are two more aspects of the GPU cooling component that must be mentioned. The core clock and memory clock of a Triple-Fan Graphic Processing Unit (GPU) on a server rack were investigated using an optimization tool to identify the ideal fan speed, core temperature, and highest potential efficiency (hashing power).

1.3 Research Objectives

The objectives are:

a) To analyze validate the relationship between clocking (core and memory) and GPU responses (fan load, core temperature, power consumption, and hash rate).

b) To locate determine the optimal fan speed, core temperature and the highest possible efficiency (hashing power) at a certain clock and memory clock via optimization tool.

1.4 Research Scopes

The scope of the study focuses on fan speed load during overclocking the GPU. The process will run in the room temperature with an open-air case setup RIG. The power consumption for the setup to run is fixed which is 110.4 Watt. Additionally, these studies only optimize the triple fan GPU data without comparing with other types of cooling systems. Furthermore, using software Design-Expert to reduce the number of experiments to get the new equation for optimizing.

CHAPTER 2

LITERATURE REVIEW

2.1 History of GPUs

By definition, a GPU is a Graphics Processing Unit. It is a specially designed circuit, or single-chip processor, used to accelerate the output of pictures from a frame buffer to a display. These devices were referred to alternatively as video processing devices in the late 1990s. They began by depicting individual pixels within a two-dimensional display using triangles. Each time a video scene (each video frame) was redone, lighting effects and objects were formed. These are mathematically intensive operations that would impose a large strain on the Central Processing Unit (CPU) or another system-level processor otherwise. By removing this demand from the CPU, clock cycles may be allocated to other tasks. A CPU's processor takes a set of instructions and executes them. Numerous modules inside the CPU perform arithmetic operations such as integer multiplication, memory or register operations, and activities such as finding a number's square root. These modules are referred to as "arithmetic logic units" (ALU). The ALU offloads function independently of the major logic cores of the CPU. A GPU core is similar to an ALU. They sought to speed the fulfilment of certain duties assigned to them. GPUs are extremely efficient in manipulating computer images and are often more efficient than general-purpose CPUs at matrix manipulation. GPU technology enable parallel processing of massive amounts of data.

Nvidia Inc. invented the GPU in 1999. The GeForce 256 GPU was capable of billions of calculations per second, could process at least ten million polygons per second, and included about 22 million transistors, compared to the Pentium III's 9 million transistors at the time (Wikipedia, 2021). Over the last decade, GPU applications have expanded to encompass physics, three-dimensional texturing, shading/shadowing, Bitcoin mining, hash-code manipulation, and encryption, all of which require repetitive or recursive mathematical processes.

2.2 Purpose for GPUs

GPGPU (O. O. Sudakov et. al.,2017) stands for General-Purpose Computation on Graphics Processing Units, more commonly abbreviated as GPU Computing. GPUs are multi-core, high-performance computers capable of extremely fast processing and data transfer rates. Originally designed for computer graphics and notoriously difficult to programmed, GPUs have evolved into general-purpose parallel processors with straightforward programming interfaces and support for industry-standard languages such as C. Developers that migrate their code to GPUs frequently see speed benefits of orders of magnitude above optimized CPU versions.

While microcontrollers and microprocessors are analogous to graphics processing units (GPUs), they are not true parallel processors. The majority of mainstream CPUs, such as Intel's Xeon product line, contain up to 24 cores and are capable of executing a huge number of instructions. Several, including Intel's Knight's Landing (nextplatform,2016) , act as coprocessors, offloading certain instructions from the parent CPU. Smaller cores, such as those present in GPUs, combined with a high number of pipelining threads, are advantageous for recurring and similar calculations that do not need considerable memory access or a large number of instructions. A GPU may include hundreds of cores, including cores inside cores, each of which is capable of doing a small number of parallel calculations. Nvidia's Compute Unified Device Architecture (CUDA), a programming language equivalent to C, was developed to facilitate the migration of traditional code from CPU-based operating systems to GPU hardware. A unified programming interface may make it a more attractive choice than FBGA, which synthesizes the GPU's produced code for a certain code architecture (i.e., Open CL 1.1) rather than for a specific hardware manufacturer. Standard languages enable code to be compiled in a similar fashion to the x86 architecture, including programmed compilation on Intel, AMD, and other x86 device suppliers.

Computer graphics have become significantly more realistic during the last decade. Triangles have been replaced by meshing polygons. The graphics engine may calculate shaders, shadows, textures, and even fluid flow. Improved pipeline and core design has been created. With these enhanced ALUs, these machines were capable of running general native languages (e.g., C). Nvidia's CUDA programming language is an example of this. It was

intended to provide a standardized application programming interface (API) for offloading code to the graphics processing unit (GPU).

2.3 Overclocking GPUs

Historically, many manufacturers of motherboards, graphics cards, and CPUs were not pleased of the term "overclocking" or the activity it referred to, and with cause. Numerous components met an untimely end as a consequence of their owners' overlocking attempts. There have always been budget-conscious customers who purchase low-end gear and use it in high-end environments. To the consumer, this is an undeniably appealing idea. On the other hand, manufacturers gain nothing from this scenario; rather, consumers hold manufacturers accountable when their overclocked computers become unstable or when a component is destroyed as a consequence of reckless overlocking (Maekinen, S. ,2006).

Today, more people than ever are overclocking their systems or individual components, and ATI was one of the first firms to actively promote this community. Often, manufacturers whose graphics cards, or CPUs support the highest levels of overclocking will find their goods at the top od enthusiast product bestseller lists. As a consequence, there are now a plethora of device that a pre-configured for overclocking, making overclocking experiments simpler and, most importantly, safer.

Overclocking is the process of raising the clock rate of a component, thereby operating it faster than it was intended to operate. Increased clock rate enables a component to execute more operations per second, but it also generates more heat. While overclocking may help extract more performance from GPU components, it often necessitates additional cooling and maintenance. If GPUs do not offer extra cooling, it may get physically damaged or become unstable, causing the computer to blue-screen or restart (Leng, et. al., 2015)

2.4 GPUs Heat Dissipate Performance

When it comes to GPU heat dissipation performance, there are many options, including air and water coolers. Essentially, heat dissipates performance through the GPU cooling system, which aids in the smooth operation for the PC. Air coolers usually provide excellent entry-level heat dissipation to address increased heating. However, although water

coolers GPU are effective at transferring heat, they are not the greatest choice (Siricharoenpanich, et, al.,2021).

Aside from these two kinds, the number of fans used to remove heat from the GPU may affect its performance. Basically, to restore quickly PC build with no heat snags, the GPU cooling system is a must-have feature to consider. In addition, GPUs with air coolers often feature one, two, or three fans to guarantee adequate cooling performance.

2.5 Components Contribute Heat in GPUs

Indeed, GPUs are designed to operate in the worst-case scenario for process, temperature, and voltage fluctuation (Zamani, et. al.,2020). A GPU is inherently heated, and it is perfectly normal for it to get that warm. This is because some components may be contributing to the system's heat production. The Control Processing Unit is the first component (CPU). The CPU is a critical component of the GPU since it serves as the computer's control unit and core. The term "control unit" refers to the part of the computer that specifies all commands and processing data. CPUs are built utilizing billions or even trillions of transistors; each transistor has a finite, non-zero resistance and is electrically linked to the rest of the circuitry through current flow that is dependent on the operation and state of the circuit to which it is attached. Because the CPU is fueled by electricity, it may switch on the flow of electric impulses by enabling or disabling the small transistors. As a consequence, electrical current either flows through the CPU or is held up, generating heat. By passing energy via thermoelectric materials, temperature fluctuations are produced (Annapragada, el. Al.,2012) When current travels through the processor material, a tiny quantity of heat is produced, impairing the thermoelectric module's performance (Krishnan, Garimella, Chrysler, & Mahajan, 2007). Overclocking GPUs has also increased the amount of heat generated by the CPU. While raising the CPU's frequency may assist decrease execution time, it can also be detrimental to the CPU. This will result in the CPU being damaged as a result of the increased heat-generating capability.

The power supply unit is the second component that contributes to GPU heat production (PSU). When it comes to powerful personal computers, the graphics card is one of the most energy-intensive components. The power needs will be lower as long as you use a graphics card with a budget or entry-level GPU. However, since mid-range and high-end

GPUs have greater GPU needs, they will use somewhat more power. While mid-range and high-end graphics cards are powered via a PCI-E x16 slot on the motherboard, entry-level or budget graphics cards are powered directly from the slot and are restricted to 75W. As a result, power-hungry cards need electricity from the power supply or power supply unit, even if the system consumes very little power (6-pin and 8-pin PCIe connectors). A system failure and/or component damage may occur as a result of a power supply overheating as a result of overloading, fan failure, or insufficient airflow outside and within the system. Historically, expansion slots have been the primary cause of power supply overload issues. Overloading the system power supply is a common occurrence when several hard drives, CD-ROM drives, and floppy drives are attached to the same power source.

Another component that contributes to heat production is video random-access memory (VRAM). VRAM is critical for GPUs since the operating system, game textures, and lighting effects are all stored in VRAM, allowing the device's CPU to access them without difficulty. Additionally, random-access storage, such as a hard disc drive (HDD), a solid-state drive (SSD), or an optical drive, is often simpler to read from and write to than other types of storage, such as a hard disc drive (HDD), a solid-state drive (SSD), or an optical drive. Additionally, VRAM generates heat because it contains a chip comprised of transistors and capacitors. When electricity travels through these components, the chip generates heat.

Voltage Regulator Module is the last component (VRM). A VRM's primary function is to convert the 12-volt electricity supplied by the computer's power supply to a useable voltage. Typically, CPUs operate between 1.1V and 1.3V. A voltage that is too high may cause irreversible harm to sensitive devices. Often, powering a CPU needs accurate voltage delivery, and the provided voltage must match the voltage requirements exactly. VRM always refers to these components as undervalued; they are not insignificant. Because without a properly configured VRM, the CPU or GPU will not get power at a stable voltage. A CPU's total output will be reduced by an underperforming VRM, and the processor will be unable to handle increasing demands. Overclocking the GPU has the potential to overheat the VRM, resulting in unintentional system shutdowns.

By and large, it can be determined that the CPU, VRAM, VRM, and power supply all contribute to the GPU's heat dissipation. To maintain an optimal temperature, the CPU, VRAM, and VRM are all placed in the core of the GPU, allowing for the use of a wide

variety of cooling agents that are also GPU components. These components are usually located in the core of the GPU to ensure that the GPU operates at the optimal temperature to avoid damage.

2.6 ASUS TUF RTX 3060 OC Edition

The ASUS TUF RTX 3060 OC Edition Graphics Card is a top choice for high-end cooling solutions. They have excellent thermal dissipation performance with good airflow in the chassis, starting with the single fan GPU. They just have one fan, which is fast yet inefficient in removing heat from the case (Asus, 2021).

The single fan GPU really helps to transfer heat in one direction, which misses numerous places in order to keep components cool. Not only do they have a lower cooling capacity, but they also do not improve performance. On the contrary, a triple fan GPU often has three fans that readily distribute air around the chassis to keep all of the components cool. They also feature bigger heatsinks and copper heat pipes to keep the GPU cool while running the PC.

Figure 2.1: ASUS TUF RTX 3060 OC Edition

2.7 Overclocking Software

MSI understands the value of increased performance and has developed its own overclocking software feature within Afterburner to streamline the process. The application enables near-complete control over the performance of a GPU, including clock speeds, core voltage, power limit, temperature limit, memory clock speed, and fan speed. Afterburner has all of the tools necessary for fine-tuning GPUs to meet specific criteria, while other

applications does not. Afterburner is compatible with any available graphics card using OC Scanner, a programmed that uses an Nvidia-developed algorithm to determine the GPU model (whether it's a 960 or a 30 series). After identifying the video card being used, the OC Scanner will compute the device's maximum and most stable overclocking settings. The card is then promptly modified, resulting in instantaneous performance gains (wepc,2021) Additionally, Afterburner enables users to fine-tune their fan profile by selecting a custom fan speed curve that changes in reaction to the GPU's temperature. The fan will spin faster when the temperature is higher; the fan will spin slower when the temperature is lower.

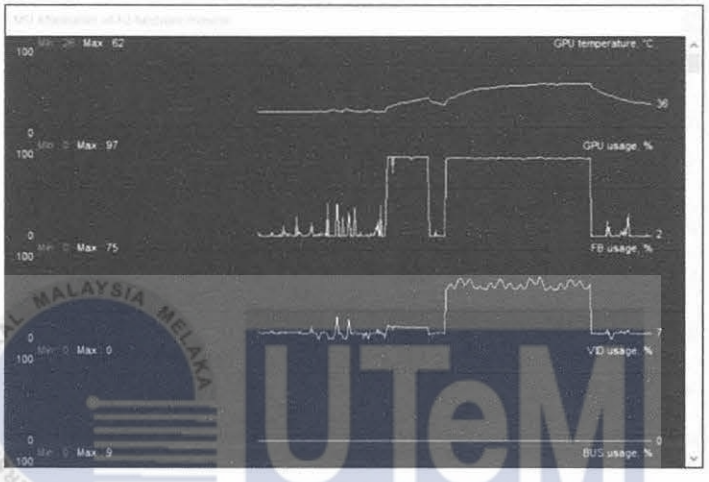

Figure 2.2: On-Screen Display

The on-screen display makes use of the Rivatuner statistics server, which may be configured for a variety of different purposes. Monitoring GPUs activity while gaming is a significant component of today's benchmarks, and the monitoring function enables to do so with live, on-screen updates.

2.8 Design-Expert

Design-Expert provides sophisticated tools for creating the optimum experiment by blending, combining, or combining parts and components (DOE). When in doubt, build a strong structure by calculating in-line power and adding blocks and center points. This is made easy by Design-design Expert's wizards and logical layouts such as the two-level structure of stoplights. It is straightforward to examine what appears to be statistically significant using analysis of variance (ANOVA) and how the findings are modelled most