THERMAL LOAD OF A SINGLE-FAN GRAPHIC PROCESSING UNIT (GPU) ON A SERVER CONFIGURATION

MUHAMMAD SUFI RIDHWAN BIN MOHD KHIR

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

THERMAL LOAD OF A SINGLE-FAN GRAPHIC PROCESSING UNIT (GPU) ON A SERVER CONFIGURATION

MUHAMMAD SUFI RIDHWAN BIN MOHD KHIR

This report is submitted
in fulfillment of the requirement for the Degree of
Bachelor of Mechanical Engineering (Thermal Fluid)

Faculty of Mechanical Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

JANUARY 2022

DECLARATION

I declare that this project report entitled “Thermal Load of A Single-Fan Graphic Processing Unit (GPU) on a Server Configuration” is the result of my own work except as cited in the references

Signature : ....

Name : ....

Date : ....

APPROVAL

I hereby declare that I have read this project report and in my opinion this report is sufficient in terms of scope and quality for the award of the degree of Bachelor of Mechanical Engineering (Thermal Fluid).

Signature :. ....

Name of Supervisor : ....

Date : ....

DEDICATION

This project work is dedicated tom my beloved family and friends for always support and give me the strength throughout my life.

ABSTRACT

Graphics Processing Unit (GPU) has gone through many revolutionary changes throughout the decade. GPU's processing power is able to challenge the existing Central Processing Unit (CPU) in the task of running high-profile software. However, any electronics that flow current into it will generate heat. The thermal problem is the major problem for GPUs. The optimal operating GPU requires minimum fan speed and core temperature with the highest efficiency (hashing power). This research focused on obtaining the optimal response at a particular core clock and memory clock via the optimization tool of Design-Expert software. The responses were recorded while applying a constant amount of power supplied to the GPU. The GPU used is in the Single-Fan Graphic Processing Unit, ASUS NVIDIA® GeForce TURBO-GTX1070-8G. The relationship between the clocking (core and memory) and GPU responses (fan speed, core temperature, hash rate) was observed. By that, Central Composite Design (CCD) generated one equation for each fan speed, core temperature, and hash rate. Next, the optimization process suggests several new clocks settings that give the best performance than the current setting. Next, the optimization process offers several new clocks settings that give the best performance than the current setting. The best core and memory clock was selected for the validation and confirmation process. The results from the validation process prove that the predicted response was precise with less than a 2% deviation.

ABSTRAK

Unit Pemprosesan Grafik (GPU) *telah melalui banyak perubahan revolusioner sepanjang dekad. Kuasa pemprosesan* GPU *mampu mencabar Unit Pemprosesan Pusat* (CPU) *yang sedia ada dalam tugas menjalankan perisian berprofil tinggi. Walau bagaimanapun, apabila komponen elektronik mengalirkan arus elektrik ke dalamnya ia akan menghasilkan sejumlah haba. Masalah haba adalah masalah utama untuk* GPU. GPU *yang beroperasi secara optimum memerlukan kelajuan kipas dan suhu teras yang minimum dengan kecekapan tertinggi (kuasa* hashing*). Penyelidikan ini akan memberi tumpuan untuk mendapatkan tindak balas optimum pada jam teras dan jam memori yang tertentu menggunakan alat pengoptimuman perisian* Design-Expert. *Tindak balas* GPU *telah direkodkan sepanjang kuasa yang malar dibekalkan kepada* GPU. GPU *yang digunakan dalam penyelidikan ini adalah Kipas Tunggal* GPU *yang bernama* ASUS NVIDIA® GeForce TURBO-GTX1070-8G. *Hubung kait antara jam (teras dan memori) dan tindak balas* GPU *(kelajuan kipas, suhu teras, kadar* hash*) direkodkan. Hasil daripada ini, Reka Bentuk Komposit Pusat* (CCD) *telah menjana satu persamaan untuk setiap kelajuan kipas, suhu teras dan kadar* hash. *Seterusnya, proses pengoptimuman telah mencadangkan beberapa tetapan jam baharu yang memberikan prestasi yang lebih baik daripada tetapan semasa. Bagi proses validasi dan pengesahan pula, satu tetapan jam (teras dan memori) terbaik baharu telah dipilih. Hasil proses validasi membuktikan bahawa respons yang diramalkan adalah tepat dengan sisihan kurang daripada 2%.*

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Merciful

First, 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 giving me an opportunity to pursue the Final Year Project for 2 semesters as partial fulfillment of my requirement for my course, Bachelor of Mechanical Engineering with Hons.

I would like to thank my family, especially my mother and father, who have been my biggest supporters until my research was fully finished. My family has encouraged me attentively with their fullest and truest attention to accomplish my work with truthful self-confidence. My sincere appreciation to Dr. Muhammad Zulfattah Bin Zakaria as my supervisor 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. Hazim, Mr. Aizuddin and Mr. Haziq for their continuous support throughout my journey. Finally, I want to convey my appreciation to everyone who aided, encouraged, and motivated me to complete my education.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GPU	Graphic Processing Unit
CPU	Central Processing Unit
RAM	Random Access Memory
VRAM	Video Random Access Memory
PSU	Power Supply Unit
CFD	Computational Fluid Dynamic
ALU	Arithmetic Logic Unit
TIMs	Thermal Interface Materials
NACA	National Advisory Committee for Aeronautic
HS	Heat Sink
SN	Signal to Noise
PCIe	Peripheral Component Interconnect Express
SIMT	Single-Instruction Multiple-Thread
HPC	High Performance Computing

RAMDAC	Random Access Memory Digital-To-Analogue Converter
HDD	Hard Disc Drive
SSD	Solid-State Drive
VRM	Voltage Regulator Module
CCD	Central Composite Design
RSM	Response Surface Methodology
ANOVA	Analysis of Variance
DOE	Design of Experiment
OC	Overclocking
LOF	Lack of Fit
C.V.	Coefficient of Variation

LIST OF SYMBOLS

Q	= Heat transferred
W	= Watt
J	= Joule
K	= Kelvin
V	= Voltage
m	= Mass, kg
CFM	= Cubic Feet per Minute
c_p	= Specific heat, $\mathrm{J}/kg \cdot K$
ΔT	= Temperature difference, K
$Q_{conduction}$	= Rate of heat conduction, W
k	= Thermal conductivity, $\mathrm{W}/m \cdot K$
A	= Cross sectional area, m^2
Δx	= Thickness, m
$Q_{convection}$	= Rate of heat transfer, W
h	= Convective are for heat transfer coefficient $\mathrm{W}/m^2 \cdot K$
Pa	= Pressure
%	= Percent

$Q_{emit,max}$	= Rate of heat transfer radiation, W
ε	= Emissivity
σ	= Stefan-Boltzmann constant ($5.670\ \times 10^{-8}\mathrm{W/m^2 \cdot\ K^4}$)
rpm	= Revolution per minute
MH/s	= Mega hash per second
°C	= Degree Celsius

CHAPTER 1

INTRODUCTION

1.1 Background

A graphics processing unit (GPU) is a computer that is designed specifically to produce the image seen on the monitor. To the more significant market, it is recognized for offering consistent, more beautiful video and game graphics that users demand today. The GPU was released in 1999. In the early 1990s, all application was entirely under the control of the Central Processing Unit (CPU). A GPU combined with a CPU can assist in tasks that involve a large amount of computational work, such as rendering. Because of this, many machines, including entry-level computers, are fitted with GPUs (Jablin et al., 2012). Since the GPU can do several calculations simultaneously, this speed-up application processing. Also, advances in information technology (such as faster internet and cell phone access) have resulted in an extraordinary increase in the need for GPU for computers and computer components.

Nowadays, current technology does not meet the expectations of the majority of consumers. In addition, high processing power is also needed to build complex applications. For example, the games and higher-performance applications like Solid-Work, Catia and other programs like it are all dependent on powerful technology such as mining cryptocurrency need greater processing power. On the other hand, CPU-bound applications would not support real-time multimedia. Because of its current design, it often cannot handle multimedia applications in real-time for the vast majority of personal computers, and video processing is not quick enough to be done by a CPU.

According to Ganesh Iyer & Dipakumar Pawar (2018) in mining cryptocurrency, GPU mining is slightly profitable because CPUs cannot handle computational stresses where GPUs have the upper hand. Moreover, the use of GPUs has the following benefits instead of these technologies: (1) GPUs allow for consolidation that simplifies design (2) due to cost savings and allow the addition of a GPU to production lines to be cheap (3) because of their programmability (Muyan-Özçelik et al., 2011).

GPUs greatly enhance programmability. Purpose-designed graphics cards were designed to handle computation- and timing-demanding functions, allowing threads to be performed on a greater number of the processing unit. The efficiency of today's multicore CPUs can be surpassed when it has a huge number of threads. GPU computing has been successfully used in diverse engineering and scientific problems requiring numerical analysis (Martínez-Frutos et al., 2017). This is because GPUs blend logic and processing units, it can be built from cores, each of which has a unique feature called arithmetic logic units (ALUs). One functional unit processes each processor thread of execution in this analysis. Each thread in the application is assigned to one functional unit for processing, and those are referred to as thread processors. All thread processors in a GPU core perform the exact instructions, as they share a control unit (Smistad et al., 2015).

1.2 Problem Statement

Additionally, semiconductor devices have caused an issue with the GPU. When a GPU is intensive, it generates much heat for video and scientific simulations. A cooling system is needed to disperse the heat. However, with the other methods considered, assume the cooling system does not cooperate or is inefficient. If this is enabled, the GPU will get too hot, which will lead to a chip failure. This action will impact the graphic card's results. An item that would allow the electronics to be damaged would exist in the worst-case scenario. According to Choi et al., (2012) the typically permissible operating temperature of a CPU is below 70ºC. The reliability of the chips then decreases by 10% for every 2ºC above the allowable operating temperature. Also, Vargas-Vazquez et al., n.d. (2018) had claimed that electronic chip cooling is one of the most challenging challenges as the demand for increasingly powerful computer systems grows exponentially.

To overcome this issue, a study of heat dissipation techniques for single-fan graphics processing units (GPU) and GPU controllers is ongoing. Several external variables and factors and a few device factors must be accounted for to keep the GPU operating at its peak performance and achieve an ideal temperature. Heat sink, heat pipe and fans help the GPU processor maintain a stable temperature. While, in this particular research, core clock and memory clock were analyzed to find the optimum fan speed, low core temperature and highest efficiency (hashing power) of the Single-Fan Graphic Processing Unit (GPU) on a server configuration.

1.3 Objectives

The objectives of the project are:

1. To validate the relationship between clocking (core and memory) and GPU responses (fan load, core temperature, power consumption, and hash rate).
2. To determine the optimal fan speed, core temperature and the highest possible efficiency (hashing power) at a specific clock and memory clock via optimization tool.

1.4 Scopes of Project

The scopes of this project are:

1. Gathering literature review of GPUs component.
2. Getting the Single-Fan Graphic Processing Unit (GPU) response (fan speed, core temperature, power consumption and hash rate) at a certain core and memory clock by using Afterburner and PhoenixMiner software.
3. Finding numerical optimization solutions by using Design-Expert software.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter mainly describes getting a clearer and better understanding of the title. This literature review will be briefly divided into four main keywords: Graphic Processing Unit (GPU), Thermal Management, Design of Fan, Equipment and Software. This combination of these four keywords will be explained thoroughly to understand this project fully. All equations and fundamental knowledge that apply in this project will be demonstrated in this section. The ultimate goal for this research is to find optimal fan speed, core temperature and the highest possible efficiency (hashing power) at a certain clock and memory clock.

2.2 Introduction of Graphic Processing Unit (GPU)

Graphic Processing unit has been a key player in both business and consumer applications. The GPU is designed for parallel processing and is commonly used in graphics and video rendering applications. Though graphics cards are well known for their gaming capability, they are increasingly used in advanced AI (Artificial Intelligence). GPU consists of multiple components: graphics processor, video memory, Random Access Memory digital-to-analog converter (RAMDAC), and driver software. GPUs aim to speed up 3D rendering. As time went on, their software and hardware became more capable, they gained flexibility and generalization abilities. According to Muyan-Özçelik et al., (2011), As seen in previous studies, graphics processing units (GPUs) are multiplying in application areas outside of traditional graphics that are well suited for their capabilities. This made it possible for the more experienced graphics programmers to achieve more striking visual effects and complex lighting and shading in the modelling software such as AutoCAD, Solidworks, and Catia. Several other software developers have also discovered the High-Performance Computing (HPC) and profound learning potential of GPUs. Other significant benefits are increased performance for games and other tasks that take advantage of parallel data access and banking memory, caches, and single-instruction multiple-thread (SIMT) execution, among others (Kerr et al., 2012).

GPU is divided into Integrated Graphics Processing Units and Discrete Graphics Processing Units. An integrated GPU does not exist independently and is combined with the CPU on the same circuit board. While for discrete GPU, it has its own boards and circuit and is commonly placed on PCIe (Peripheral Component Interconnect Express) slots. Nowadays, most GPUs on the market today are currently on the integrated level of GPUs. For integrated graphics, both the graphics processing unit and the CPU work together to render the monitor's image. A motherboard CPU that includes fully integrated GPU results in thinner and lighter and more energy-efficient systems, and less powerful CPUs lowers the cost. For high resource-intensive applications, such as 3D games, it is