THERMAL LOAD OF A DUAL-FAN GRAPHIC PROCESSING UNIT (GPU) SYSTEM ON A SERVER RACK

MUHAMMAD HAZIM BIN KAMARUDIN

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

THERMAL LOAD OF A DUAL-FAN GRAPHIC PROCESSING UNIT (GPU) SYSTEM ON A SERVER RACK

MUHAMMAD HAZIM BIN KAMARUDIN

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

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Faculty of Mechanical Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

JANUARY 2022

DECLARATION

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

Signature : .....

Name : .....

Date : .....

UTeM
اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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 : .....

UTeM
اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

DEDICATION

This project I dedicated to my dearly loved family and supervisor upon their non-stop support and motivation in every part of my life.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ABSTRACT

Graphics Processing Unit (GPU) has gone through many revolutionary changes throughout the decade. GPU's processing power can 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 certain core clock and memory clock via the optimization tool of Design-Expert software. The response was recorded while applying constant amount of power supplied to the GPU. The GPU used is in the category of dual fan which is Gigabyte GeForce GTX 1070 WINDFORCE OC 8G. The relationship between the clocking (core and memory) and GPU responses (fan speed, core temperature, hash rate) were 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. For the validation and confirmation process, the best one core and memory clock was selected. The results from the validation process prove that the predicted response was precise with less than 2% deviation.

اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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*. Tidak balas* GPU *telah direkodkan sepanjang kuasa yang malar dibekalkan kepada* GPU. GPU *yang digunakan dalam penyelidikan ini adalah* GPU *dua kipas yang bernama* GeForce GTX 1070 WINDFORCE OC 8G. *Hubungan 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%.*

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ACKNOWLEDGEMENT

All praise and thanks to Allah and His blessing for providing me with the opportunity to finish my thesis. First and foremost, I want to thank my supervisor, Dr. Muhammad Zulfattah bin Zakaria, for his eternal encouragement, patience and, most importantly, for guiding me through completing this thesis. He offers many suggestions, advice, valuable knowledge, and help for me to complete my project completely and successfully, especially in writing this report. It has been a great delight and honor to have him as my supervisor.

A massive thanks to all the parties involved either directly or indirectly. I have been able to complete this project by receiving many helps and supports from all parties around me. My family Kamarudin, Atika, Mahirah and Irfan also supported me in terms of financially and emotionally. I would like to express my gratitude to my project colleague at Universiti Teknikal Malaysia Melaka (UTeM), Mr. Sufi and Mr. Aizuddin, for their continuous support throughout my journey. Finally, I would like to express my appreciation toward the panels. Without their comments, suggestions, improvising ideas, and tips, I cannot spot and improve my weaknesses and mistakes, especially project presentations. Hopefully, this research will benefit and guide other students and researchers. I can't mention everyone I appreciate since it would take a lifetime but knowing that I am not alone in this journey makes me feel wonderful. Thank you very much. Trust me, all of you will always be in my prayers

اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

TABLE OF CONTENTS

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UTeM
اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF TABLES

UTeM
اونيۏرسيتي تيکنيکل مليسيا ملاک
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF FIGURES

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF ABBREVIATIONS

GPU	Graphic Processing Unit
CPU	Central Processing Unit
RSM	Response Surface Methodology
BBD	Box-Behnken Design
CCD	Central Composite Design
LAN	Local Area Network
WAN	Wide Area Network
VPU	Visual Processing Unit
AI	Artificial Intelligence
CUDA	Compute Unified Device Architecture
VRAM	Video Card Random-Access Memory
VRM	Voltage Regulator Module
PSU	Power Supply Unit
PCI	Peripheral Component Interconnect
ANOVA	Analysis of Variance
DOE	Design of Experiment
OC	Overclocking
RAM	Random Access Memory
PC	Personal Computer
LOF	Lack of fit

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

C.V	Coefficient of variation

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF SYMBOLS

%	= Percent
V	= Voltage
W	= Watts
Qmax	= Maximum heat transfer capacity
$Q_{conduction}$	= Rate of heat conduction, W
k	= Thermal conductivity, W/m.K
ΔT	= Temperature difference, K
Δx	= Thickness, m
Q	= Heat capacity, J
m	= Mass of substance, kg
cp	= Specific heat, J/kg.K
ΔT	= Temperature difference, K
$Q_{convection}$	= Rate of heat convection, W
h	= Convection heat transfer coefficient, W/m^2.K
A	= Surface area for heat transfer, m^2
Q_{rad}	= Rate of heat radiation, W
ε	= emissivity
σ	= Stefan-Boltzmann constant, 5.67 x 10^{-8} W/m^2.K^4

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

T_s	= Absolute Temperature of Surface, K
T_{surr}	= Absolute Temperature of Surrounding, K
CFM	= Cubic Feet Per Minute
rpm	= Revolutions per minute
°C	= Degree Celsius
R^2	= Coefficient of determination
MH/s	= Mega hash per second

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

CHAPTER 1

INTRODUCTION

1.1 Background

A server is a computer program that provides data or services to another computer, such as computer software and hardware. Servers are created to perform multiple essential data processing for client servers with access through the internet or local networks. A server does not have a screen or keyboard, and it is a computer that transfers data from one computer to another. The technology used in delivering the data can be employed to operate on a local area network (LAN) or over a wide area network (WAN). Site, mail, and other clients are all different forms of servers, and there are several options for each.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Server has a sensitive part and would lead to technical issues to the entire system and have many wire connections. A server rack is a storage and organizing system for electronics equipment such as GPU. Server racks help them organize the wire and protect the server and their parts from external damage. Stacking servers and other electronics equipment in a rack makes it easier to keep things organised and monitor airflow. The design of the server racks also plays an important role in increasing the server's efficiency (Gao et al., 2016).

The server's major component is the motherboard which connects all other components. Next is a processor and it was divided into two types: the Central Processing Unit (CPU) and the Graphic Processing Unit (GPU). GPU have become more programmable in recent years, allowing them to be used in many different fields. Since they're faster and more energy-efficient than a conventional CPU, GPUs can perform even larger tasks with greater effectiveness (Erik et al., 2014). To maximise the speed of the computer, parallel processing is used. In parallel computing, GPU are being used frequently. GPUs have much more parallelization capacities than CPUs as they have many cores compared to CPU (Ebubekir et al., 2018). Servers usually use several GPUs to have enhanced processing power. GPU was developed to simulate 3D graphics such as for AutoCAD, CATIA and SolidWorks software (Sadrieh et al., 2011). When time went by, they were more malleable and more competent, and they improved their skills. With their enhanced capabilities, graphic programmers could use advanced lighting and shadowing methods to improve the look of visuals. The field of deep learning has used GPUs to speed up other workloads. Also, a GPU can do several calculations simultaneously, which helps it advance the overall efficiency of the server and process large numbers of numbers quickly.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

A GPU can be found on servers connected to the circuit where the CPU is located or integrated into the motherboard. There are multiple big brands in the GPU industry, including the NVIDIA, AMD, INTEL, and ARM franchises. At present, NVIDIA and AMD are known as the two most prominent GPU vendors. In the case of NVIDIA and AMD, they design and supply GPUs, which are passed on to third-party vendors such as MSI, ASUS, and Sapphire, who can make their customised changes without altering the GPU's chip in some way (Stewart., 2021). A GPU is a so-called multi-processor that performs multiple tasks simultaneously, while a conventional

processor does not enable this. This is because the GPU processor comprises hundreds or thousands of small cores or units that handle multiple tasks simultaneously for complicated graphics processing.

The speed of the GPU's VRAM is determined by the memory clock, whereas the core clock determines the speed of the GPU's chip. The GPU's core clock may be compared to a gaming PC's CPU and RAM clocks. The core clock often has a greater impact on gaming performance than the memory clock. Games' visual effects are temporarily stored in virtual memory (VRAM) on the graphics processing unit (GPU). More VRAM means the graphics card can process images faster, and faster VRAM means the users can store more assets. As a result, the games will render more quickly if the memory clock speed is greater. The clock speed of the graphics processor's GPU core determines how quickly it can process graphics.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

1.2 Problem Statement

Recently, keeping electronic chips cold has been one of the biggest concerns due to the infinite number of systems requiring more efficient and less power-consuming technology (Jose-Carlos et al., 2018). The research is primarily looking for the best way to cool a GPU in the server rack since it releases a lot of heat. A significant heat source in electronic devices is the GPU (Siricharoenpanich et al., 2021). An issue related to the GPU heat of the system is semiconductor chips. If the GPU is assigned a workload including gaming and highly intensive apps, it can produce a great deal of heat. There has been a rise in interest in GPU as researchers realise, they need to find the cooling solution of the devices to satisfy the market's needs (Al-Rashed et al.,

2016). Several GPUs server designs have been upgraded to ensure good air circulation and prevent overheating. There are several methods of cooling systems such as passive and active to ensure the graphics card temperatures remain low. In active cooling, the fan is used, but for passive cooling, the heat sink is used (Svasta et al., 2017). However, server chassis has its limitations such as construction costs and it needed excessive room within for placement of the system.

Thus, to deal with this major difficulty, an analysis was carried out to keep the GPU cool using a dual-fan cooling system. Priority was given to some factors to ensure that the GPU maintains its thermal efficiency for maximum performance at an ideal temperature. A wide variety of fans was employed, but only a specific type of fans was being analysed. Any other points must be pointed out regarding the cooling component in GPU are the design of heat sink and heat pipe. While in this research, core clock and memory clock of a Dual-Fan Graphic Processing Unit (GPU) on a server rack were examined to determine the optimal fan speed, core temperature and the highest possible efficiency (hashing power) via optimization tool.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

1.3 Objectives

The core objectives of this project are as follow:

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

1.4 Scope of Project

The scopes of this project are:

1. Gathering literature review of GPUs component.
2. Getting the Dual-Fan Graphic Processing Unit GTX 1070 response (fan speed, core temperature and hash rate) at a certain core and memory clock using Afterburner and PhoenixMiner software.
3. Finding solutions using numerical optimization using Design-Expert software.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter aimed to explain in detail and get a better understanding related to the project title. In this Literature Review, the Graphic Processing Unit (GPU) and the thermal management are the main keywords. The evidence and findings on this project will be investigated to their fullest to ensure complete understanding. All of the fundamentals and equations that will be applied to this project can be found in this chapter. This research aims to locate the optimal fan speed, core temperature, and the highest possible efficiency (hashing power) at certain core clock and memory clock via optimization tool.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2.2 Introduction of GPU

In today's world, maximum processing speed is needed. While the progress made by the Central Processing Unit (CPU) over the last two decades has been enormous, it has now reached a halt. To address this, NVIDIA in 1999 introduced the Graphics Processing Unit (GPU) or the Visual Processing Unit (VPU), a modern