THERMAL LOAD OF THREE-FAN COOLING SYSTEM OF A GRAPHIC PROCESSING UNIT(GPU)

AIZUDDIN AZRI BIN CHE AZIZ

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

THERMAL LOAD OF THREE-FAN COOLING SYSTEM OF A GRAPHIC PROCESSING UNIT(GPU)

AIZUDDIN AZRI BIN CHE AZIZ

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 Three-Fan Cooling System of a Graphic Processing Unit” 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

To my dear beloved father, mother, friends and my supportive supervisor that endlessly encouraging me to keep going forward.

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 three-fan which is ASUS ROG Strix GTX1070. 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.

ABSTRAK

Unit Pemprosesan Grafik (GPU) telah melalui banyak perubahan yang besar 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 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 GPU tiga kipas yang bernama ASUS ROG Strix GTX1070. Perkaitan 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 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%.

ACKNOWLEDGEMENT

First and foremost, I'd like to express my gratitude and praise to Allah the Almighty, my Creator and Sustainer, for the opportunity to finish this project despite of all the hardship and circumstances throughout the journey. Additionally, I'd like to express my gratitude to Universiti Teknikal Malaysia Melaka (UTeM) for providing the best supervisor for me. My dear supervisor, Dr Muhammad Zulfattah bin Zakaria has given me continuous support, encouragement, counsel and inspiration that has been a great help for me. His unwavering patience in guiding will always be cherished through the eternity.

Additionally, sincere thanks to my friends, Mr Hazim and Mr Sufi that always helps and cooperate with me during discussion and studies. They have provided a huge amount of helps since the starting of this project. Lastly, to my father, Mr. Che Aziz and my mother, Mrs. Faridah for their prayers, love and support that are priceless and heavily appreciated. They have been accompanying me through the ups and downs of this semester.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

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 Varience
DOE	Design of Experiment
OC	Overclocking
C.V	Coefficient of Variation
LOF	Lack of Fit
PC	Personal Computer

LIST OF SYMBOLS

W =	Watts
Q_{max} =	Maximum heat transfer capacity
Q_{cond} =	Rate of heat conduction
k =	Thermal conductivity
A =	Cross sectional area
$T_1 - T_2$ =	Temperature difference
Δx =	Thickness
Q_{conv} =	Rate of heat convection
h =	Convection heat transfer coefficient
A_s =	Surface area where heat convection takes place
T_s =	Surface temperature
T_∞ =	Temperature of the fluid sufficiently far from the surface
Q_{rad} =	Rate of heat radiation
ε =	Emissivity
σ =	Stefan-Boltzmann constant
T_s =	Surface temperature
T_{surr} =	Absolute temperature of surrounding
CFM=	Cubic feet per minute
rpm=	revolutions per minute
ºC=	Degree Celcius
MH/s=	Mega hash per second

CHAPTER 1

INTRODUCTION

1.1 Background

Graphic Processing Unit (GPU) was originally more like a video display card that was made by International Business Machines (IBM) in 1981. Monochrome Display Adapter (MDA) was originally in single monochrome to allow high-resolution text and symbol display at 80 x 25 characters that was useful for drawing forms. Years later, IBM came out with first graphic card with full-color display (PC, 2019). After that, the technology developed and evolved into having 8nm chip and having memory of 24GB that able to support hardware-raytracing, variable-rate shading and more. Mainly GPU is used to improve the performance of 3D rendering. Time after time, affected by modern technology, it able to produce stunning visual effects and natural scenes with advanced lighting and shadowing methods. Developers has also started to greatly increase the capability of GPU to boost its performance so that it able to do deep learning, high performance computing (HPC) and more in a shorter period of time (Intel, 2020).

Graphic cards are at the mercy of names like Nvidia and AMD, which exert great control over the market. EVGA, ASUS, and MSI are a couple of the many third-party graphic card manufacturers that purchase their chips from these two companies (Crider, 2018). Following that, these third-party companies will come out with their

own package with hardware such as cooler, lighting, plus additional video ports or maybe a bit more attractive lighting option. At this particular point is where single, dual even three-fan GPU were invented. In the absence of anything that could be done to help reduce the heating issues, these various GPU producers realized the value of R&D in figuring out the best method to deal with it.

GPUs significantly outperformed CPUs in terms of speed, making them a valuable medium for computing (Breitbart, 1999). GPUs consist of chips and RAM (random access memory) which chips are used to process the data received from the input port and RAM are used to store the data of the image processed (S. Gillis, 2020). Now with the latest GPU that can generate almost 1695MHz at boost clock, it must be a challenging task to keep maintaining the GPU's temperature at its best to keep on having the best performance.

1.2 Problem Statement

Electronic devices such as central processing unit (CPU) or graphic processing unit (GPU) have a high heat generating capacity which equivalent to nuclear reactors, which makes it difficult to control their heat output on an individual component basis (Hirasawa et al.,2005). When the GPU becomes too hot, its performance degrades rapidly. However, most GPUs are equipped with thermal insulation which will lessen the amount of heat generated to maintain full efficiency of it. Normally, there are a few factors contributing to the overheating issue such as heatsink mount, fan speed or maybe the fans are clogged with dust. Moreover, it is a difficult task to get rid of the

heat and also to monitor the temperature around the chip since GPU is close-packed (Feng & Li, 2013).

It is obvious that thermal management is the main issue if it is about GPU. However, with the nonstop research that has been done to resolve the issue, little by little discoveries of ways to maintain the GPU at its best temperature have been figured out. There is a strong relationship between the clocking (core and memory) to the GPU response. Hence, there is a need to examine how clocking will affect the GPU responses such as fan speed, core temperature and hashing power.

1.3 Objectives

The core objectives of the project are as follow:

1. To analyze 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 scope of the project is:

1. Collecting literature review for GPU cooling system.
2. Get the Three-Fan Graphic Processing Unit GTX 1070 response (fan speed, core temperature and hash rate) at certain core and memory clock using Afterburner and Phoenixminer software.
3. Finding solutions using numerical optimization using Design-Expert software

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter will explain in depth of how the GPU really works. It will briefly explain which components that contribute to the heat released and which components that helps to counter the problem of heat throttling that affecting the performance of the GPU. Also included in this section is the fundamental knowledge and equations related to this project. In this chapter, detailed information to be applied onto this project will be provided.

Figure 2.1: AORUS GeForce RTX™ 3090 XTREME 24G

2.2 Component Contributing to Heat

Pushing better performance meaning higher heat will be dissipated from the GPU itself. Then when its temperature rises to a certain degree, the GPU's performance will be dropped to prevent the GPU from overheating and damaging the GPU permanently. The normal allowable temperature of a normal operating processor is below 70°C but the increment of 2°C will make it 10% less reliable than its normal working condition (Choi et al., 2012). Moreover, temperature is another critical consideration because high processor temperatures can damage performance, shorten the life of the processor and lead to reliability issues (Kang et al., 2011).

The main contributor of heat on a GPU is the processor itself. To differentiate the processor from the GPU, it usually located in the center of the GPU having a square shape. All processors are Integrated Circuits (IC) which means that it will generate heat as soon as it is switched on. There are two company in control of GPU market, Nvidia and AMD with their GPU architecture of Pascal and Polaris respectively (Verma, 2017). The processor can be considered as the brain of the GPU that will handle all the calculation, simple arithmetic, mathematical, input/output and control operations based on the instruction given to it. As electricity is running into a GPU processor, it can get hot in the same way as a CPU does. The addition of Watts would raise the temperature much further, in addition to the potential to overclock the GPU.

Second component that contribute to the heat is the Video Random Access Memory (VRAM) of the GPU. This component can be considered as the second most important component on a GPU. In this component is the location where both graphics data and game textures are placed for GPU processing. RAM is made up out of

hundreds of billions of transistors that are packed onto the chipset to transform with an order of command. These transistors act like a switch that turns on and off according to commands given to them. As we know these devices have certain current resistance that will heats up, so the amount of heat produced is determined by the command it is executing such as running graphics-intensive software with additional background software running at the same time.

The third one would be Voltage Regulator Module (VRM). Living up to its name, it is used to transform higher voltage from the power supply into lower voltage ranges for GPU usage. It usually transforms 12V to about 1V to 1.5V approximately, which is the voltage range at which GPUs typically run. Each graphics card has different numbers of VRM (Evanson, 2020). GPU with higher power will requires more VRM to control the voltage input so it able to turn down the voltage when idling to keep the heat down and to save power (Verma, 2017). VRMs may get extremely hot, much hotter than the GPU, so they need adequate cooling to prevent the graphics card from shutting down.

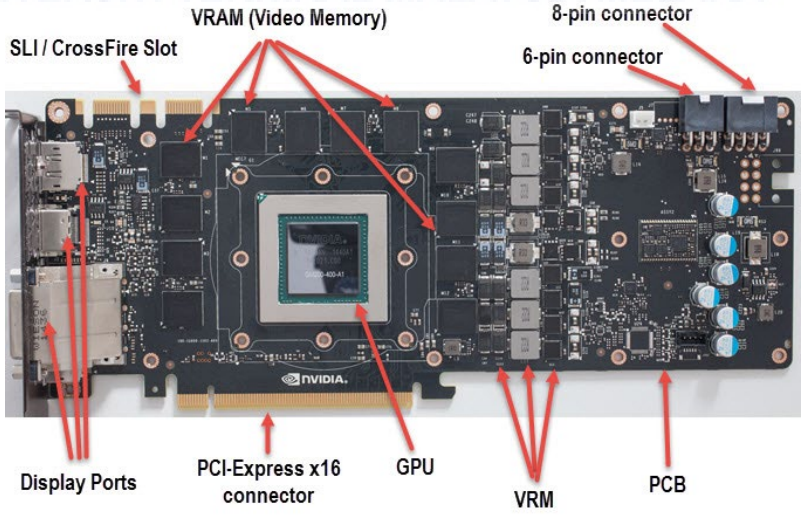

Figure 2.2: Components on a GPU

2.3 Component to Counter Heat Problem

When power is delivered to any semiconductor, heat will be generated through the process. Since the 90s, cooling systems for graphic card have gone through evolutions because in the early days, GPU needs no cooling system since the chips were bare. Then in 2001, heatsink was introduced to GPUs attached by using thermal paste. Graphics cards have become more power hungry as their computational power has increased. Heatsinks and the fans that cooled them had grown in size since then, but they were still small and light. Larger heatsinks and fans that also cool RAM were introduced in 2003. In the same year, Nvidia released the FX 5800 Ultra, a GPU with a blow-out cooling system that turned out to be the loudest in history. Then, there was also new model introduced with heat pipes later that year. Hence, there are main components which are heatsink, fans, thermal paste and heat pipes.

Figure 2.3: Rage Pro with no cooling, 1997

Figure 2.4: Radeon 7000 with heatsink, 2001

Figure 2.5: Radeon 9800 XT with larger heatsink plus fans, 2003

Figure 2.6: FX 5800 Ultra with blow-out style fan, 2003

2.3.1 Heatsink

Heatsink can be considered as the most important component for cooling the GPU processor. It usually located on top of any component that generate heat to reject most of heat generated to maintain the performance of GPU. The performance of active GPU processor cooling heatsinks primarily depends on the forced air convection created by computer fans. These components are used in electronics as well as high-power electrical components and regarded to be the best and most cost-effective cooling option. It comes in a variety of forms and base materials, each having its own flowing flow (Khattak & Ali, 2019). In most cases, heat created by processors is transmitted to a heat sink by conduction and subsequently to the environment by natural, mixed, or forced convection. Moreover, heatsink plays an important role when it comes to overclocking the GPU since heatsink is the closest component to the GPU processor. Overclocking is a process where the input Watts is increased to get more performance from the GPU. As the temperature increases, the heat sink's low heat removal efficiency may cause damage to the electronic component. As a result, if the excessive heat created by the electronic element is not evacuated in a timely manner, chip reliability and service life are jeopardized. (Hamdi et al., 2018). Best way to enhance the heat dissipation is by adding heat pipes or by increasing the fin area of the heatsink. However, both options would result in a directly increase the cost and system size (Choi et al., 2012).

2.3.2 Heat Pipes

Heat pipes are metal pipes which filled with a liquid such as ammonia, acetone, or water. It began to be used on a larger scale around year of 2003. The liquid evaporates as it heats up and travels through the pipe. As it travels, it loses heat and cools down, returning to a liquid state. The heat pipe is an excellent way to transfer heat from one point to another because this loop is infinite. The pipes are usually connected to a cooling plate above the GPU and then transfer the heat to a heatsink that is located away from the GPU. A heat pipe can be thought of as a passive heat pump that removes heat due to physical laws. (Badalan & Svasta, 2017). A heat pipe is a device with excellent thermal conductivity that allows heat to be transported while keeping a nearly constant temperature across its heated and cooled parts (Jouhara & Hussam, 2018). The cooling system's performance can be improved by adding heat pipes (Choi et al., 2012).

According to Jouhara & Hussam (2018), a heat pipe is a passive thermal transfer device that uses phase-change processes and vapour diffusion to transmit huge quantities of heat over relatively long distances with no moving elements. Heat pipes are made out of an evacuated tube that is partly filled with a working fluid that can transform in both liquid and vapour phases. The evaporator is on the left side of the heat pipe, while the condenser is on the right. When the fluid enters the evaporator with high temperature, it will evaporate and turn into vapour that can flow to the condenser with high velocity. The vapour condenses and releases its latent heat as soon as it reaches the condenser part. After that, it flows back to the evaporator either by