Content Adaptation In Game On Demand Environment

Ritthichai Jitpukdeebodintra

A Thesis Submitted in Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Computer Engineering
Prince of Songkla University
2016

Thesis Title	Content Adaptation In Game On Demand Environment
Author	Mr.Ritthichai Jitpukdeebodintra
Major Program	Computer Engineering

Major Advisor

....
(Asst.Prof.Dr.Suntorn Witosurapot)

Examining Committee :

....Chairperson
(Dr.Somchai Limsiroratana)

....Commitee
(Asst.Prof.Dr.Supachate Innet)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfillment of the requirements for the Doctor of Philosophy in Computer Engineering

....
(Assoc.Prof.Dr.Teerapol Srichana)
Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

.... Signature
(Dr. Suntorn Witosurapot)
Major Advisor

.... Signature
(Mr. Ritthichai Jitpukdeebodintra)
Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

……………………………………………… Signature
(Mr. Ritthichai Jitpukdeebodintra)
Candidate

ชื่อวิทยานิพนธ์ กลไกปรับแต่งข้อมูลในสภาวะแวดล้อมของบริการเกมออนดีมานด์
ผู้เขียน นาย ฤทธิชัย จิตภักดีบดินทร์
สาขาวิชา วิศวกรรมคอมพิวเตอร์
ปีการศึกษา 2558

บทคัดย่อ

บริการเกมออนดีมานด์ได้รับการออกแบบเพื่อรองรับการเล่นเกมที่ต้องการความละเอียดของภาพสูง แต่สามารถใช้งานบนอุปกรณ์คอมพิวเตอร์ที่มีทรัพยากรในการประมวลผลต่ำได้ (เช่น สมารท์โฟน หรือแท็บเล็ต เป็นต้น) โดยใช้เครื่องเซิร์ฟเวอร์เสมือนบนคลาวด์คอมพิวติงมาช่วยประมวลผลและรับภาระงานเกมทั้งหมด ทำให้เหลือภาระงานเฉพาะด้านการแสดงผลกราฟิกในลักษณะของวิดีโอแบบสตรีมมิง ซึ่งส่งผ่านเครือข่ายอินเตอร์เน็ตความเร็วสูงแทน ดังนั้น ในกรณีที่ทรัพยากรของเครื่องเซิร์ฟเวอร์เสมือนไม่เพียงพอที่จะรองรับภาระงานที่ถูกร้องขอจากเครื่องคอมพิวเตอร์ของผู้เล่นเกมได้ จะทำให้ประสิทธิภาพในการให้บริการลดลงอย่างมีนัยสำคัญ กลไกการจัดการภาระงานของบริการเกมออนดีมานด์จึงมีความสำคัญมาก แต่ปัจจุบันยังขาดกลไกดังกล่าวที่มีประสิทธิภาพ เนื่องจากไม่ได้พิจารณาภาพรวมของการจัดสรรทรัพยากรที่เกี่ยวข้องทั้งหมด จึงอาจทำให้เกิดความขาดแคลนของทรัพยากรที่เกี่ยวข้องอื่นๆ ที่ไม่ได้นำมาพิจารณาร่วมกัน ในงานวิจัยเพื่อวิทยานิพนธ์นี้ นำเสนอผลของการศึกษาวิเคราะห์เพื่อเสนอแนะแนวทางในการจัดสรรทรัพยากรของบริการเกมออนดีมานด์อย่างรอบด้าน ตามพื้นฐานของสถาปัตยกรรมแบบไคลเอนต์/เซิร์ฟเวอร์ในประเด็นต่างๆ ดังต่อไปนี้

1) การศึกษาและวิเคราะห์ถึงผลกระทบของปัจจัยต่างๆ ทางด้านเครื่องผู้ใช้ ที่มีต่อปริมาณของภาระงานที่เกิดขึ้นยังเครื่องเซิร์ฟเวอร์ ซึ่งพบว่าสามารถนำไปใช้ในการประมาณความสัมพันธ์ของปัจจัยดังกล่าวในลักษณะของสมการเชิงเส้นแบบพื้นฐานได้ จึงสะดวกในนำไปใช้รองรับการคาดการณ์ปริมาณทรัพยากรต่างๆ อย่างแม่นยำ จนนำไปสู่การจัดสรรทรัพยากรโดยรวมอย่างมีประสิทธิภาพได้

2) การศึกษาเพื่อปรับปรุงกลไกการจัดเตรียมภาระงานให้ทรัพยากรต่างๆ บนเครื่องเซิร์ฟเวอร์ให้บริการเกมออนดีมานด์ โดยให้นำทรัพยากรทั้งหมดเข้ามาพิจารณาร่วมกัน โดยนำแนวทางการแก้ปัญหาแบบการจัดของใส่ถังแบบหลายตัวแปร มาใช้ในกำหนดสัดส่วนที่เหมาะสมของการใช้ทรัพยากร ภายในหน่วยประมวลผลแบบ CPU และ GPU รวมถึงทรัพยากรเครือข่าย ซึ่งพบว่าได้ผลดีกว่าวิธีการที่นำเสนอในงานวิจัยอื่นๆ ทั่วไปในแง่ของการรักษาระดับของทรัพยากรต่างๆ ได้อย่างพอเพียงกับการให้บริการ โดยมีหลักฐานสนับสนุนจากผลการทดลองจากเกมหลากหลายประเภท

3) การศึกษาเพื่อนำเสนอกลไกการปรับแต่งภาระงาน ให้ทำงานอยู่บนพื้นฐานของประโยชน์ทั้งด้านของผู้ใช้และเซิร์ฟเวอร์ให้บริการ โดยได้นำเสนออัลกอริทึมเพื่อหาความเหมาะสมในการประนีประนอมผลประโยชน์ที่ทั้งสองด้าน ซึ่งมีประสิทธิภาพสูงกว่าวิธี

พื้นฐานแบบอื่นๆ ซึ่งมักจะคำนึงถึงประโยชน์แต่เพียงด้านเดียวเท่านั้น โดยมีหลักฐานสนับสนุนจากผลการทดลองเกมหลากหลายประเภทเช่นกัน

ผลของการศึกษาหัวข้อทั้งสามนั้นสามารถจะสนับสนุนแนวความคิดใหม่ในการพิจารณาอย่างรอบด้านหรือครบถ้วนทั้งในด้านของทรัพยากรที่เกี่ยวข้องกับการประมวลผลเกมออนดีมานด์โดยรวม หรือประโยชน์ของผู้เกี่ยวข้องทั้งหมด ได้เป็นอย่างดี และคาดว่าจะเป็นประโยชน์ต่อวงการผู้พัฒนาเกมประเภทดังกล่าวได้ต่อไปในอนาคต

Thesis Title	Content Adaptation In Game On Demand Environment
Author	Mr. Ritthichai Jitpukdeebodintra
Major Program	Computer Engineering
Academic Year	2015

ABSTRACT

Game on demand is designed for supporting high resolution gaming on low computing resource devices (such as smartphone or tablets). In game on demand, the virtual server will take care all of game workload computation. Then, the result will be shown on client machine display in form of video streaming over high speed internet. Hence, working in this manner, if the demanded resources cannot be met at the server, the significant degradation of service performance will be resulted. The mechanism of resource management in game on demand is very important. However, the current mechanisms are inefficiency, because they do not consider all related resources in comprehensive manner. That leads to insufficiency of the other related resources which are not considered altogether. The research in this thesis presents the analytical study to suggest the workload management method for game on demand in all views of client/server architecture basis.

1) This thesis study and analysis on the effect of client machine variable to the volume of workload on the server. By an experimental result, it can be found that these relations can be modeled with simple linear function. Then, the accuracy performance to approximate the game on demand workload which lead to efficient resource management is possible.

2) This thesis study to improve the provisioning workload mechanism in game on demand servers, by apply the complete view of all related resources. In this study, the multi-variable bin-packing is used to find the suitable ratio of CPU, GPU and network resource utilization. According to an experimental result from many example games, these ideas lead to more efficient workload provisioning in term of maintaining the acceptable level for game on demand services.

3) This thesis study to improve the adaptation mechanism for game on demand workload based on mutual benefits of client and server basis. Depending on the experiment result from many example games, the compromise of benefits for both parties will lead to greater adaptation

performance than the current method which usually aim for single prime objective.

The result of above three studies conclude that the new idea of combining all game on demand related resources can bring more efficient workload management and expect to give benefits to game on demand developer in near future.

ACKNOWLEDGEMENT

The successful completion of this thesis would not have been possible without the help of many people in many ways. First, I would like to express my deepest gratitude to Asst.Prof.Dr.Suntorn Witosurapot, my thesis supervisor, not only for his extensively idea, strong feedback and invaluable direction on this thesis, but also his guidance for me to be more complete mankind.

I would like to thanks to all colleague at Wireless Information Group (WIG) Lab especially Atchara Rueangprathum whom I spend good time with during my study.

I am extremely grateful to Nida Simapatanapong, MD for all her kindness, unlimited love and great support.

I also would like to express my special thanks to my family include, Dr.Surachai Jitpukdeebodintra, Asst.Prof.Dr.Somrutai Jitpukdeebodintra and Chaiyarit Jitpukdeebodintra, MD for their unconditioned love, devoting, understanding and constant encouragement through my entire study. Without these, this work would not have been possible.

Ritthichai Jitpukdeebodintra

CONTENT

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATIONS

Mbps	Megabit Per Seconds
FPS	Frame Per Seconds
ms	Millisecond
MPixel	Million Pixel
MB	Megabyte
GB	Gigabyte
GHz	Gigahertz
CPU	Central Processing Unit
GPU	Graphic Processing Unit
FLOPS	Floating-point Operations Per Second

LIST OF PAPERS

Paper 1 R. Jipukdeebodintra and S. Witosurapot, "**A study on the impact of client display resolutions in cloud gaming workloads**", Proceeding of 6th Annual International Conference on ICT-BDCS 2015, 2015.

Paper 2 R. Jipukdeebodintra and S. Witosurapot, "**Efficient Cloud Gaming Resource Provision Via Multi-dimensional Bin-Packing**", "GSTF International Journal on Computing (JoC)", Vol.4 No.4, March 2016.

Paper 3 R. Jipukdeebodintra and S. Witosurapot, "**Hybrid method for adaptive cloud gaming contents**", "GSTF International Journal on Computing (JoC)", Vol.4 No.2, March 2015.

PERMISSION FROM PUBLISHERS

Permission for paper 2: “Efficient Cloud Gaming Resource Provision Via Multi-dimensional Bin-Packing” which publish on “GSTF International Journal on Computing (JoC)”, Vol.4 No.4, March 2016.

GSTF is committed to keeping articles published in its journals in full compliance with Open Access principles and practices through GSTF Digital Library (http://dl4.globalstf.org) as well as Springer’s Open Access Platform Global Science Journal (www.springer.com/globalsciencejournals), for everyone. By default, GSTF publishes these articles under a Creative Commons Attribution Non Commercial (CC BY-NC 3.0) licence that allows reuse subject only to the use being non-commercial and to the article being fully attributed (http://creativecommons.org/licenses/by-nc/3.0) to GSTF. Articles funded by certain organisations that mandate publication with a Creative Commons Attribution (CC BY 3.0) licence, which may require reuse for commercial purposes are allowed, subject to the article being fully attributed to GSTF.

GSTF requires the author(s) to grant exclusive world-wide license in perpetuity, in all forms, format and media to a) publish b) distribute c) display d) store e) reprint f) translate into other languages and reproduce the paper g) the inclusion of electronic links of the paper to third party material where-ever it may be hosted h) and/or license to any third party to do any or all of the above.

The author(s) may use their own papers for non-commercial purposes by acknowledging first publication in GSTF Journal and giving full reference and/or web link as appropriate.

Permission for paper 3: “Hybrid method for adaptive cloud gaming contents” which publish on “GSTF International Journal on Computing (JoC)”, Vol.4 No.2, March 2015.

GSTF is committed to keeping articles published in its journals in full compliance with Open Access principles and practices through GSTF Digital Library (http://dl4.globalstf.org) as well as Springer’s Open Access Platform Global Science Journal (www.springer.com/globalsciencejournals), for everyone. By default, GSTF publishes these articles under a Creative Commons Attribution Non Commercial (CC BY-NC 3.0) licence that allows reuse subject only to the use being non-commercial and to the article being fully attributed (http://creativecommons.org/licenses/by-nc/3.0) to GSTF. Articles funded by certain organisations that mandate publication with a Creative Commons Attribution (CC BY 3.0) licence, which may require reuse for commercial purposes are allowed, subject to the article being fully attributed to GSTF.

GSTF requires the author(s) to grant exclusive world-wide license in perpetuity, in all forms, format and media to a) publish b) distribute c) display d) store e) reprint f) translate into other languages and reproduce the paper g) the inclusion of electronic links of the paper to third party material where-ever it may be hosted h) and/or license to any third party to do any or all of the above.

The author(s) may use their own papers for non-commercial purposes by acknowledging first publication in GSTF Journal and giving full reference and/or web link as appropriate.

Chapter 1
Introduction

Nowadays, the direct streaming of video and music contents from cloud services to heterogeneous devices, such as TV, PC and Tablets, has drastically gained its popularity. For example, the video streaming service like YouTube[1], Netflix[2] or the music streaming service like Spotify[3], Apple Music[4] have widely recognized and become essential services in the cloud service era. As a consequence, the software industry has shown the attention to apply the sort of principle for service provisioning to the other applications as well, such as Cloud gaming [1].

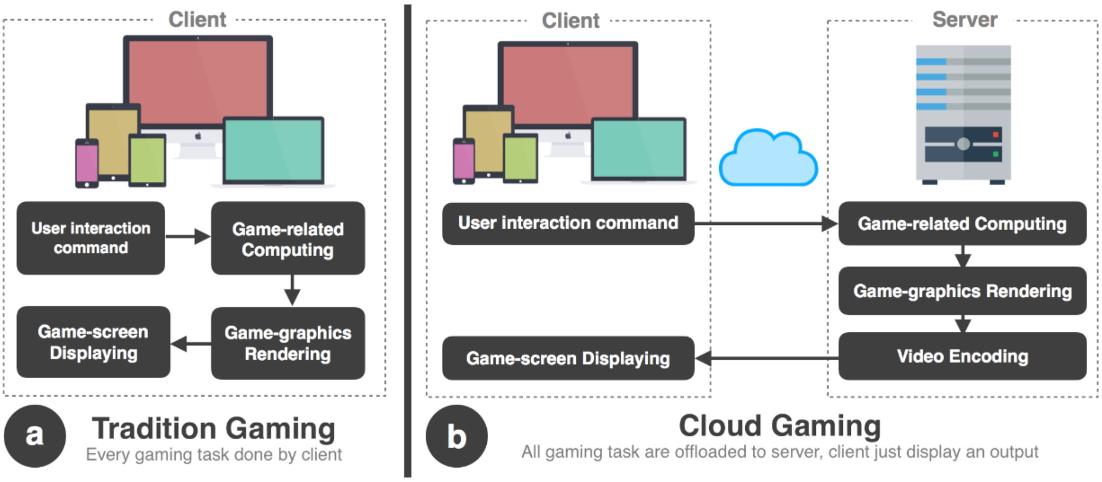


Figure 1-1 Architectural Comparison of Traditional network gaming application (a) and Cloud gaming service (b)

Based on the Fig. 1-1, we can see that the architecture of cloud gaming service (also called Gaming On Demand (GoD) or Gaming- As-A-Service (GaaS) [2]) in Fig. 1-1 (b) are significantly different from that of traditional gaming application in Fig. 1-1 (a). It

[1] http://www.youtube.com
[2] http://www.netflix.com
[3] http://www.spotify.com
[4] http://www.apple.com/music

can be noticed that the computation of actual game-codes and the rendering of game graphics in cloud gaming will be completely offloaded to the powerful machine of cloud gaming server (as shown in the right side of Fig. 1-1(b)). Afterwards, the resulted video streams are encoded and sent straightforwardly to render on the gaming client over the high-speed internet connection. Therefore, the cloud gaming clients perform merely the functions related to the user inputs and the rendering of video display, without involving any gaming process at all. By working in this manner, the cloud gaming service exhibits key advantages as follows:

- It enables devices with low-computing capability, e.g. smartphones or tablets, to access for high-computing demanded game application without problems,
- It requires no pre-installation on the cloud gaming device, and hence making no assumption of underlying hardware or operating system.

Unfortunately, these dominant advantages cause the performance of cloud gaming to rely on the following issues:

- A high-speed network connection between client and server machine becomes essential, in order to stream high-resolution video efficiently.
- A high computing power of cloud gaming server is inevitable, in order to support all client requests adequately.

It becomes clear that the service quality of cloud gaming can deteriorate to an unacceptable level, if the network or computing resources are not well managed. This motivates a number of efficient mechanisms in the past years. Nevertheless, we can classify them into two broad approaches of which details are in the following section.

1.1 Two existing approaches for handing disadvantages of cloud gaming

1.1.1 Approach I: Workload provision

This approach works on the assumption that the quality of cloud gaming service is always acceptable as long as the sufficient resources can make available. In this regard, the issue of “sufficient provisioning of resource facilities for gaming workload” will become a prime consideration (as shown in Fig. 1-2). By following this direction, many mechanisms are proposed in such a way that the available resources, such as Network in [1-2], Graphics Processing Unit (GPU) in [3-4] and Central Processing Unit (CPU) in [5-14], should be utilized or managed in a somewhat efficient manner.

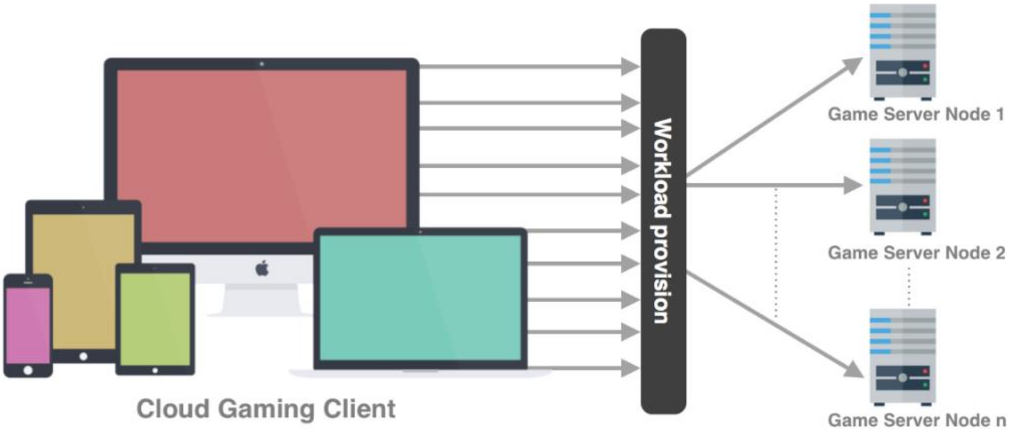


Figure 1-2 Workload provision approach for cloud gaming

1.1.2 Approach II: Workload adaptation

In contrast to the first approach, this approach assumes the resources are always limited and hence the workload adaptation will be essential to enable the best of service quality from the limited resource availability, such the screen size and resolution of cloud gaming devices as shown in Fig. 1-3. In the past years, many techniques belonging to the workload adaptation have been studied, but they can be classified according to the obtained benefits whether for the server or the client. While the target of server-centric benefit aims to decrease the computing workload at server [15,16] that of client-centric benefit aims instead to reduce the network requirement at the client [17-20].

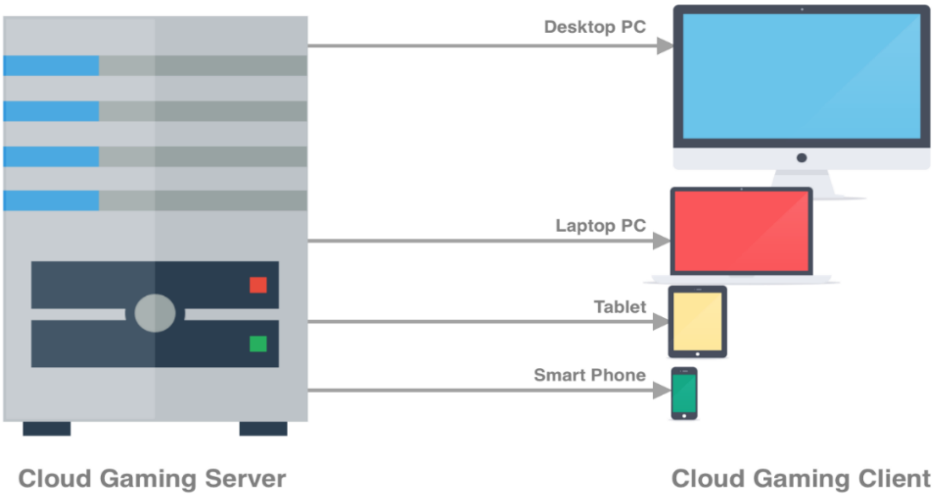


Figure 1-3 Workload adaptation approach for cloud gaming

1.2 Weaknesses of existing approaches

Although the workload provision and workload adaptation approaches seems to be capable of handling the inherit characteristics of typical cloud gaming system, neither of them has their own weaknesses in somewhat level. Table 1-1 summarizes some side effects during the implementation of these approaches.

Table 1-1 Weaknesses of existing approaches

Approach	Resource in focus or Viewpoint	Side effects
Workload provision	CPU	Possible shortage of GPU and Network resources
	GPU	Possible shortage of CPU & Network resources
	Network	Possible shortage of CPU & GPU resources
Workload adaptation	Server-benefit	Degradation of graphic quality at the gaming client
	Client-benefit	Increase of server workload

In the case of **workload provision**, managing resource on a single issue of CPU, GPU or Network resources is indeed not practical for the cloud gaming service environment, since they are all required for cloud gaming executions. Placing a burden on some cloud gaming resource on the server machine will surely affect to the other resources in somewhat level. Therefore, poor resource management should be aware since it may lead to the collapse of cloud gaming services.

In the case of **workload adaptation**, attempting to reduce on the server machine by merely changing the graphic resolutions in trans-rendering engine can be worsening the user's perception. In contrast, improving the graphic quality at the client machine in the transient periods will unnecessarily increase the server's workload as well.

1.3 Research Questions

In order to alleviate the disadvantages of current approaches that attempt to handle some intrinsic limitations of cloud gaming architecture as described above, the following research questions have arisen:

Research Question1:

How can the workload provision approach on the cloud gaming server be augmented to include the effects from client display resolution?

In the past, the cloud gaming resource management relying on the sole consideration of current server's workload can be applicable for the client machine with a few set of standard display resolutions. However, as we have already experienced a diverse set of device capability in smartphones at present, those legacy approaches are then unlikely adequate. In this thesis, we believe that an effective resource management in cloud gaming environment should take into account of both the current server's workload and the actual capability of mobile devices. In this regard, it demands for the omni-viewpoint of workload provision approach to be devised.

Research Question2:

How can the workload provision approach be enhanced by realizing the workload utilization on both CPU and GPU resources?

While an increased number of research works on cloud gaming workload provision can be found in the literature, they often overlook the coupling of CPU and GPU computing resources, due to the limited capability of GPU technology in the past (see Fig. 1.4). Moreover, some of them had a misassumption that cloud gaming service can be simply classified as CPU-based or GPU-based games in a similar way to the traditional game playing [36-38]. Indeed, the cooperation of CPU and GPU operations are important for maintaining the service quality in cloud gaming and experimental results in [38] can be well-served as evidence. As a result, those previous works rarely take an omni-viewpoint of CPU and GPU workloads as well as network resource into consideration and, hence, an effective resource management in the cloud gaming system cannot be always guaranteed. Unlike the work in this thesis, we will explore a more effective resource management algorithm that will allow us to have a realistic control of all coupled resources.

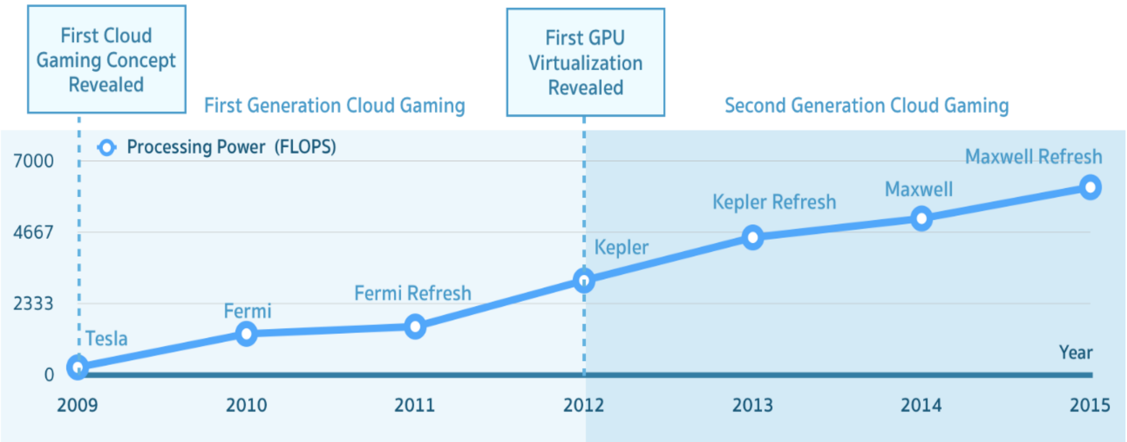


Figure 1-4 Timeline of cloud gaming and the GPU evolution

Research Question3:

How should the workload adaptation approach be improved by working on the mutual benefits of client and server?

Although, a number of work related to the workload adaptation technique in the field of cloud gaming can be found in the literature, they often take a certain focus on obtained benefits for either the client or server machines. As stated earlier in the section 1.2, either of them will eventually have its own weakness. Then, it would be interesting to investigate how (and in what way) the knowledge of user requirement (in terms of preferred display resolution) can devise the better cloud gaming resource management (in terms of better utilization of computing and network resources) on the basis of user and server cooperation.

1.4 Thesis summary

The organization of this thesis can be illustrated in a form of table (as shown in Table 1-2) and structured as follows:

Table 1-2 Summary of this thesis

	Introduction	Literature Review	Theory	Experimental	Conclusion
Chapter 1	✓				
Chapter 2	✓	✓			
Chapter 3	✓		✓	✓	✓
Chapter 4	✓		✓	✓	✓
Chapter 5	✓		✓	✓	✓
Chapter 6					✓

In chapter 1, we introduce and define the basic terminology used in this thesis. The contents divide into four sections. Section 1 describes the characteristic and the problems of cloud gaming service, following with the explanation of two existing approaches, the workload provision and adaptation, that attempt to solve the cloud gaming problem. Section 2 identifies the weaknesses of existing approaches that motivate our research works. Section 3 declares our research questions and thesis summarization is given in Section 4.

In chapter 2, we give background and review of literatures related to this thesis. The contents divide into two sections. Section 1 illustrates the activities of cloud gaming service, and explain it technical problems. Section 2 deals with the reviews of literature relates to workload provision and adaptation. In addition, we explain why those works in the literature cannot apply in realistic situation.

In chapter 3, we concentrate on the relation of client variables and characteristics of cloud gaming workload. A simple linear equation is then proposed to model this relation. The contents divide into seven sections. Section 1 is introductory to the issue. Section 2 gives some background of cloud gaming workload. Section 3 deals with study methodology. In Section 4-6, our studies are detailed and the evaluation results are described respectively. In Section 7, we conclude this chapter.

In chapter 4, we propose the omni-viewpoint based cloud gaming workload provision. The contents divide into five sections. Section one the introduction to this section will be given. Section 2 details the decision making process for service permission in cloud gaming. Section 3 models our approach to provision cloud gaming workload based on the bin-packing problem. Section 4 examines the performance of our propose approach. In Section 5 concludes the chapter.

In chapter 5, we propose the cloud gaming workload adaptation based on omni-viewpoint. This contents divide into six sections. Section 1 introduces this issue. Section 2 declares our newly proposed mechanism for mutual resource utilization, following with the examination and result of our mechanism performance in Section 3 -5 respectively. Section 6 concludes this chapter.

Conclusions are drawn in chapter 6. The main goals of the thesis that have been reached are concluded. Our suggestion is that performances of both approaches of workload provision and adaptation applied for resource management in cloud gaming service environment can be potentially improved if the global knowledge of the current workload on all coupling resources on the server machine and the clients' device capability are known and brought into consideration. At last, some directions to extend our research works are given.

Chapter 2
Background & Literature Review

2.1 Workflows of Cloud gaming

As mentioned in the chapter 1, basic architecture of cloud gaming service is relatively different from that of traditional network gaming application. Therefore, the tasks on the client device will be minimal, but relies instead on gaming operations on the server machine. In this regard, overall activities of cloud gaming operations after client asks for cloud gaming service can be illustrated as an activity diagram in Fig. 2-1. Noticed that there are two consecutive phases of the initialize phase and rendering phase in the diagram.

In the initialize phase, when a gaming client requests for a service, it is the duty of service controller to find the most suitable gaming server node through the admission control mechanism, where the sufficient resource availability of each gaming server will be examined and ranked. After finishing the determination of a qualified server, a session will be established between the client and the selected server node.

In the rendering phase, a control loop of operations related to cloud gaming computation and video-encoding tasks will be took place. In each loop, when receiving the user’s commands or inputs from the client device, the server will perform associated issues accordingly, such as intelligence of non-player characters, score calculation, or collision detection. The results will be drawn on graphic frames by GPU in rendering sub-system, which will be later encoded into high definition stream and sent to the cloud gaming client over the network in video encoding sub-system. By working in this manner, it is not surprised why the performance of cloud gaming service will rely greatly on the adequacy of both computing and network resources.

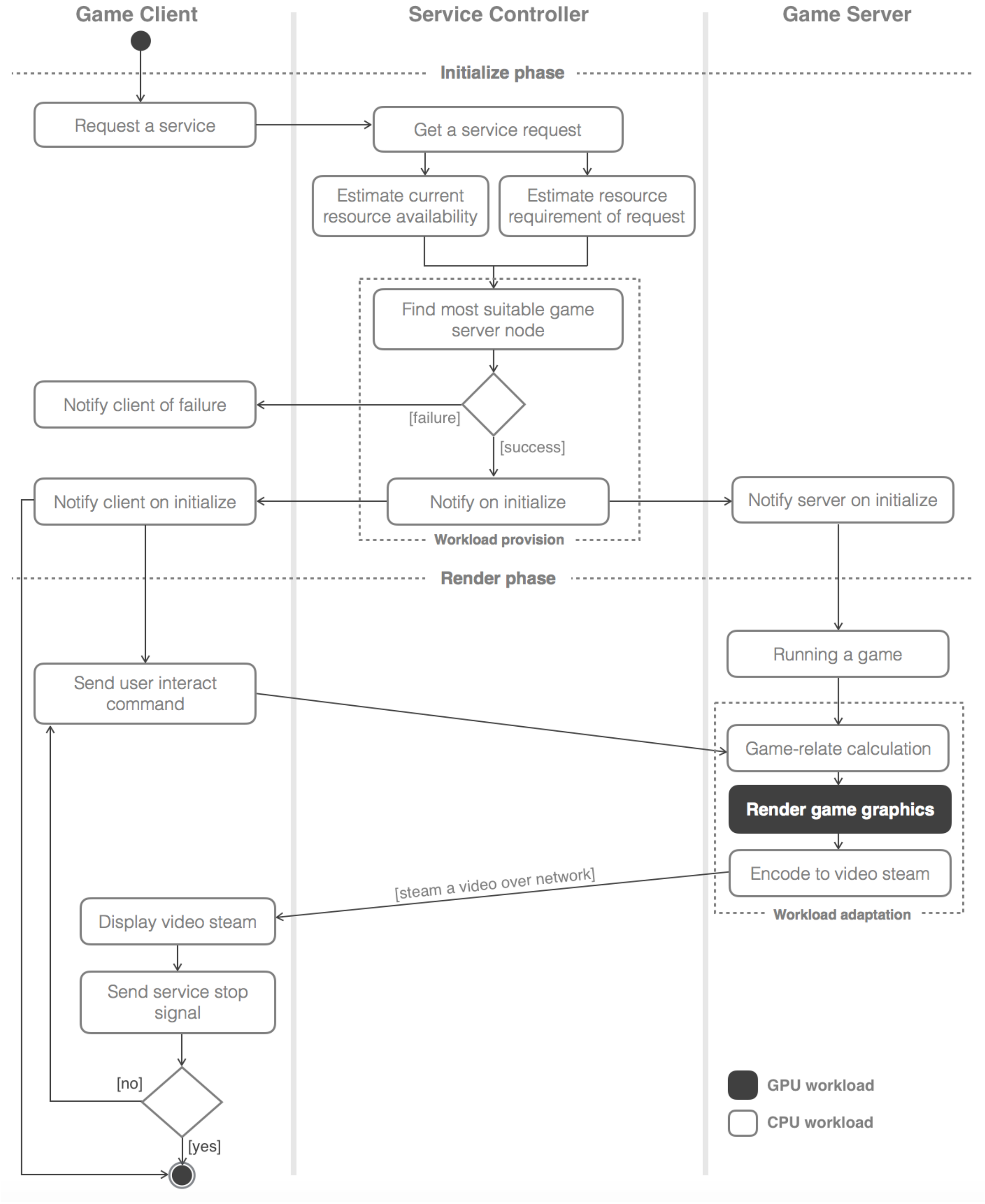


Figure 2-1 An activity diagram of cloud gaming workflow

In order to provide an evidence of this matter, an “Assassin's Creed (AC) : Syndicate” game had been experimented under various conditions. The obtained results are shown as graphs in Fig. 2-2.

- In Fig. 2-2(a), when both computing and network resources are ample, the game will be played back smoothly on the client machine at the acceptable rate at 30 frames-per-second (FPS).
- Fig. 2-2(b), (c) and (d) represent three different cases when network, CPU or GPU resources are in scarcity accordingly. As expected, the resulted games will be rather fluctuated at the rates lower than the acceptable one in all cases. Particularly, the unpleasant effect of frame skip can occur if network is so scarce until the required frames are delay to arrive within the dateline during the video playback.

Based on the evidence above, it becomes clear why both network and computing resources must always maintain to stay above the level of minimum requirement, otherwise the performance degradation of cloud gaming service will be realized at the client machine.

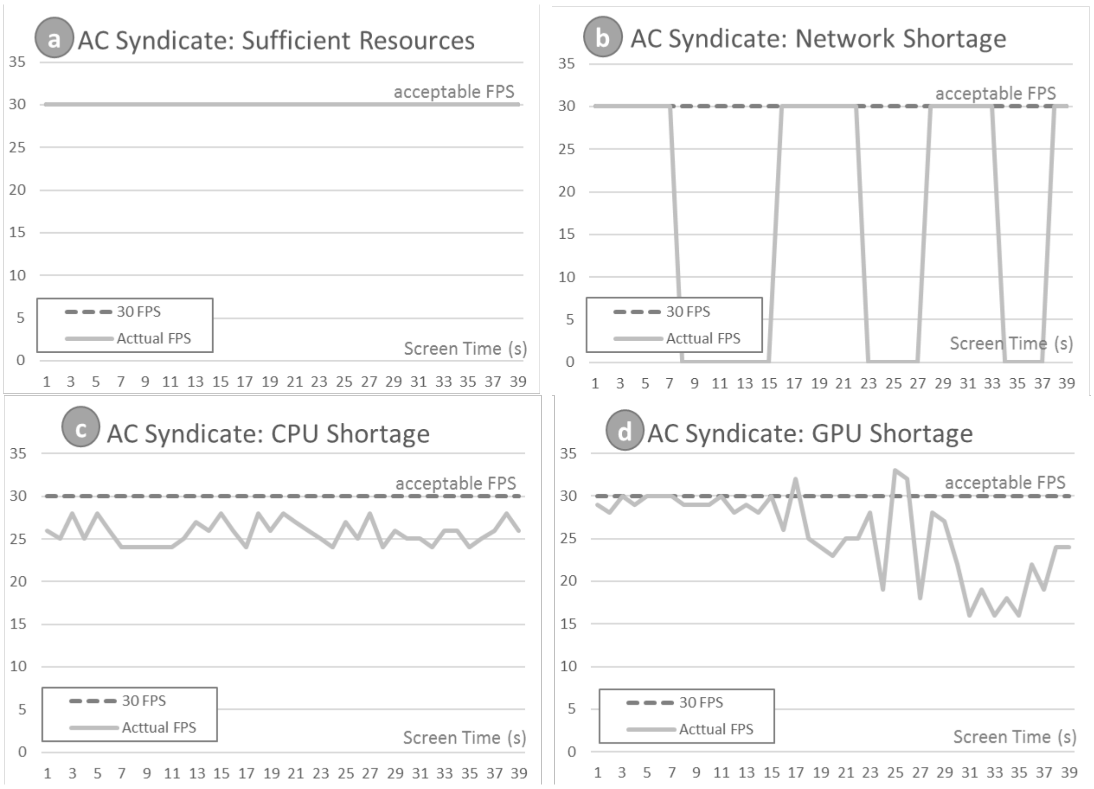


Figure 2-2 Frame rate (FPS) of cloud gaming under the shortages of some resource.

2.2 Review of solutions for handling resource inadequacy in cloud gaming

In order to maintain the available resources to stay in the efficient level, a number of solutions can be possible, depending on the working phases of cloud gaming (referred to the Section 2.1) that is especially interested. However, the classified approaches of "resource provision" and "resource adaptation" that are mentioned earlier in section 1.1, can be applicable here.

2.2.1 Literature reviews on the workload provision approach

In general, the workload provision approach aims at providing an efficient management of currently available resources against the incoming workload; hence, many algorithms in the field of optimization can be directly applicable for allocating the requested workload into the most optimal resource slot.

As seen in Fig. 2-3, since the role of GPU during the years of 2009 – 2012 was still limited, most of the research works applying the workload provision approach rather aimed specifically at optimizing the CPU resource for video-encoding task and some other game-related issues. A literature survey of V.Vinothina et al. [5] in 2012 suggested many possible strategies that can achieve the aim, such as using Auction-based method, Service-level agreement (SLA) method, or mathematical heuristic methods. In 2010, G. Wei et al. [6] explained the success of game-theory method for managing CPU resource in cloud computing environment.

In the 2012, the GPU virtualization technology and a quantum leap of GPU performance were introduced and provided a great potential for handling many concurrent sessions in a cloud gaming server. Since then, the consideration of cloud gaming workload provision becomes an interesting subject of research. For instance, the work of Zhang et al. [3] in 2014 revealed how the scheduling approach can be also applied for optimizing GPU resource utilization in cloud gaming environment.

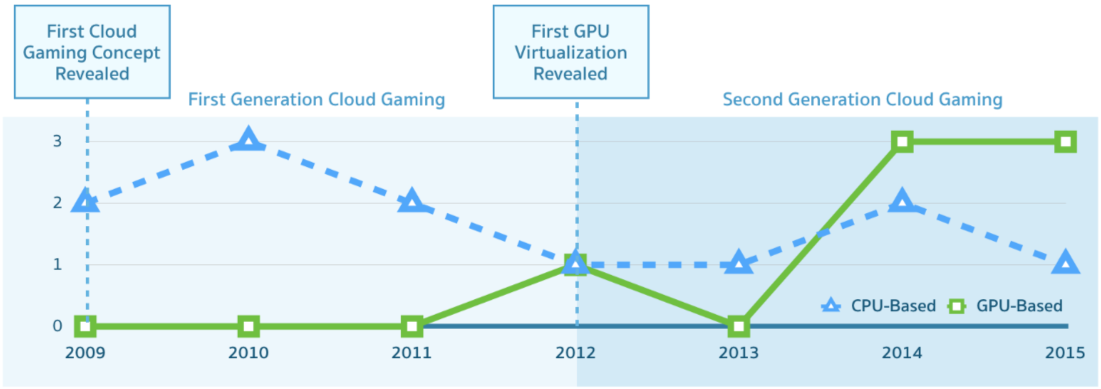


Figure 2-3 Number of reviewed literatures in the thesis by the CPU or GPU resources

We summarize some previous works, which are related to the application of workload provision approach for different kind of resources (i.e. CPU, GPU and network resources), in Table 2-2. However, these works considered the CPU and GPU resources in orthogonal, which is not realistic in the cloud gaming environment nowadays. Therefore, they are different than our work in this thesis, which argue on the coupling of CPU and GPU in many tasks (e.g. that of video encoding) in the modern cloud gaming service. Details of this issue will be described later in the chapter 4.

Table 2-1 Summary of literatures related to workload provision in cloud gaming

Resource in Focus	Ref. work	Year	Technique	Application Domain
GPU	[3]	2014	Adaptive scheduling algorithm	Cloud Gaming
	[4]	2014	SLA-Aware scheduling	Cloud Gaming
CPU	[5]	2012	Time-series approach	Cloud Computing
	[6]	2010	Game theoretic method	Cloud Computing
	[7]	2009	Most Fit Processor Policy	Cloud Computing
	[8]	2010	Matchmaking strategy	Cloud Computing
	[9]	2011	Utility function	Cloud Computing
	[10]	2010	Based on demand prediction	Cloud Computing
	[11]	2009	Machine learning technique	Cloud Computing
	[12]	2011	Annealing algorithm	Cloud Computing
	[13]	2010	Stochastic approach	Cloud Computing
	[14]	2012	Prediction based	Cloud Computing
	[14]	2014	Bin-Packing problem solver approach	Cloud Gaming
	[30]	2013	Machine Learning Based Prediction	Cloud Computing
Network	[29]	2013	Bin-Packing problem solver approach	Cloud Computing

2.2.2 Literature reviews on the workload adaptation approach

In general, the workload adaptation approach will take place in the rendering phase of cloud gaming session (referred to the Section 2.1) on the server machine, although there will exist some work [38] that proposed this approach at the client device. In essence, the aim of this approach is to provide a matching of the output target resolution of cloud gaming server and the display resolution of the client device. Regarding the adaptive feature of heterogeneous devices, this approach is considered importance, since it can be used as a tool for tuning the efficient level of resource utilization in the network, while the acceptable resolution at the client machine is still maintained. For instance, it can be noticed in Fig. 2-4 that running games (i.e. Batman AK or AC Syndicate) at different resolutions (e.g. 1080p and 4K) will consume different GPU workloads, In similar, for a given GPU workload, running "Batman Arkham Knight" game at 1080p resolution can produce the game-rate of 85 FPS and 4k resolution at the lower rate of 27 FPS (as seen in Fig.2-5).

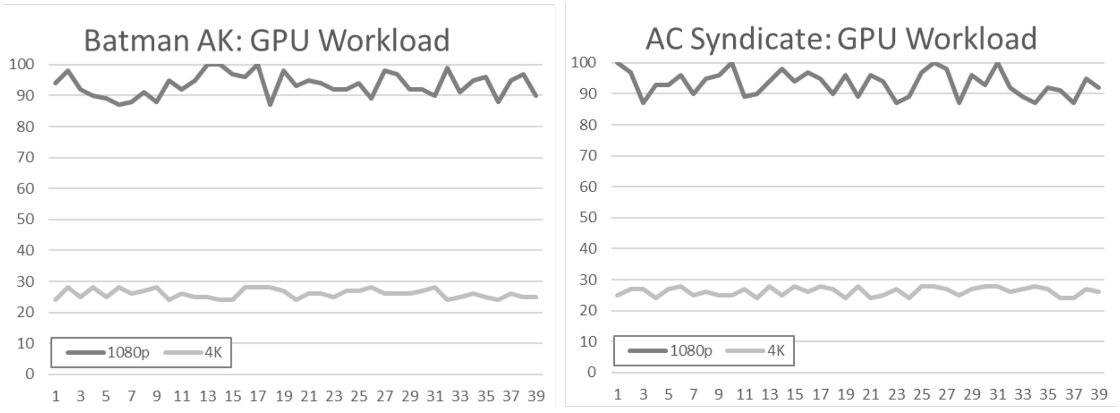
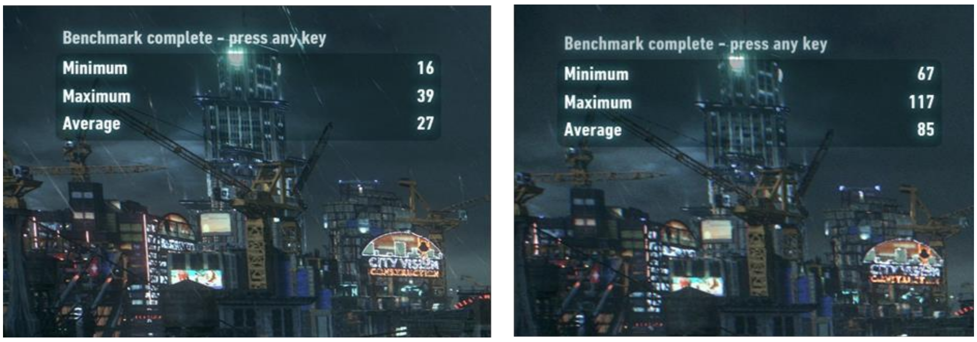


Figure 2-4 GPU utilization when playing games at 1080p and 4K resolution



Benchmark complete – press any key
Minimum 16
Maximum 39
Average 27

Benchmark complete – press any key
Minimum 67
Maximum 117
Average 85

Figure 2-5 The FPS of Batman Arkham Knight at 4k (left) 1080p (right)

In what follows, we review the literature related to the workload adaptation in cloud gaming services. They can be classified into two groups using the returned benefits of the server or the client.

For the server-centric benefits, the objective of adaptive mechanism taking place at the server is to decrease the workload that is involved into the game-rendering process on the server machine. A widely known technique is that of "Trans-rendering" technique, where the less volume of contents can be obtained by lowering graphic resolutions or removing some texture details in video games, as shown in Fig. 2-6. Example works can be seen in [15] and [16]. In the latter work, they suggested on how the computing workloads can be reduced by adjusting 5 gaming render factors consisting of Realistic effect, View distance, Texture detail, Environment detail and Rendering Frame Rate.

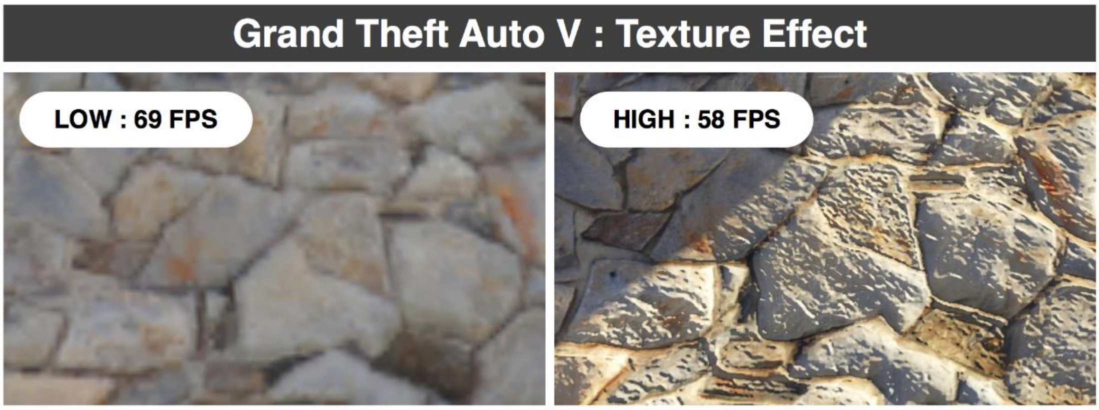


Figure 2-6 Example of graphics using trans-rendering technique.

From the client-centric benefits, the objective of adaptive mechanism at the server machine is to decrease the network usage via lowering video resolutions of the output stream. A widely known technique is that of "Trans-scaling", where the video frames are scaled by somewhat manner. For instance, Winter et al. [18] suggested using different video codec for the entire game scene. Aparicio-Pardo et al. [19] suggested using the contexts of each game scene to select a right configuration of video encoding. The similar work of M. Hemmati et al. [20] also recommended changing the bitrate configuration of video encoding, in order to reduce the cloud gaming latency. More works can be found in the research field of video-on-demand (such as [32-35]).

In Table 2-3, we summarize some of previous works, which are related to the application of workload adaptation approach in the cloud gaming environment. Nevertheless, the work in this thesis is significantly different from these works, where either the server-centric benefits or client-centric benefits are separately concerned. Therefore, it is unavoidable to face the undesirable side-effects as mentioned earlier in the section 1.2 of the last chapter.

Table 2-2 Literature review related to workload adaptation in cloud gaming

Method	in	Year	Field	Key statement
Trans-scaling	[17]	2009	Cloud gaming	Showed the effects of different codec algorithms on the resource in use
	[18]	2008	Cloud gaming	Showed the effects of different codec algorithms on the cloud gaming latency
	[19]	2011	Cloud gaming	Used rendering context to select frames needed for encoding
	[20]	2013	Cloud gaming	Advice how to adapt the output bitrate, rather than the output resolution
	[32]	2007	Video on Demand	Suggested using network bandwidth to calculate the output resolution
	[33]	2008	Video on Demand	Quality of Server (QoS) based adaptation
	[34]	2007	Video on Demand	Showed how to adapt an output resolution over the running-time
	[35]	2005	Video on Demand	Showed how standard technology (MPEG-21, SOAP and HTTP) can be applied in adaptive video streaming
Trans-rendering	[16]	2015	Video on Demand	Applied an integer linear program (ILP) to calculate the output resolution
	[15]	2010	Cloud gaming	Used a level selection algorithm to calculate the output rendering

While the workload adaptation for **server-centric benefits** can solve an insufficiency of server's GPU resources by decreasing the computing workload, it can worsen the quality of output video stream as well, such as blurring the screen, and pixelate the texture detail on the client machine. As results, these frames are not bearable for showing on the high-resolution display (see Fig. 2-7 for example illustration).

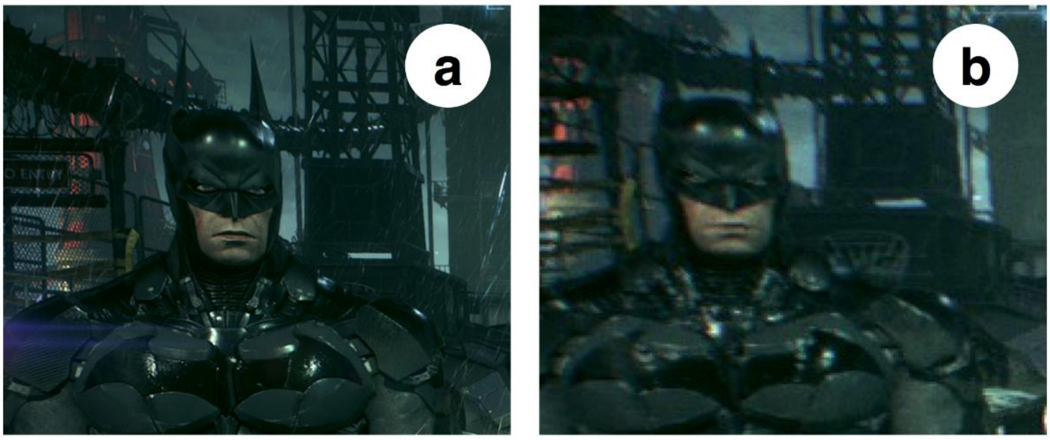


Figure 2-7 Gaming at high resolution (a) compare to low resolution (b)

While the workload adaptation for **client-centric benefits** can reduce the network consumption requiring for sending video stream to the client machine, it demands a great burden of computing tasks at server as well. This can cause the degradation of service quality, especially during the prime time when the server becomes in a serious congestion condition with a large number of game clients. Evidences can be seen from our simple experiment, where the “Sleeping Dog AVG” is simultaneously accessed by the different number of users. In Fig. 2-8, it can be seen clearly that, as the number of users are increased, the capability of trans-rendering engine will be lower and lower until the unacceptable frame rate (i.e. less than 30 frames per second) may be in results. Based on the evidence, we argue that an effective technique for workload adaptation in cloud gaming service should span across the benefits of both client and server. This will be explained in details later in the chapter 5.

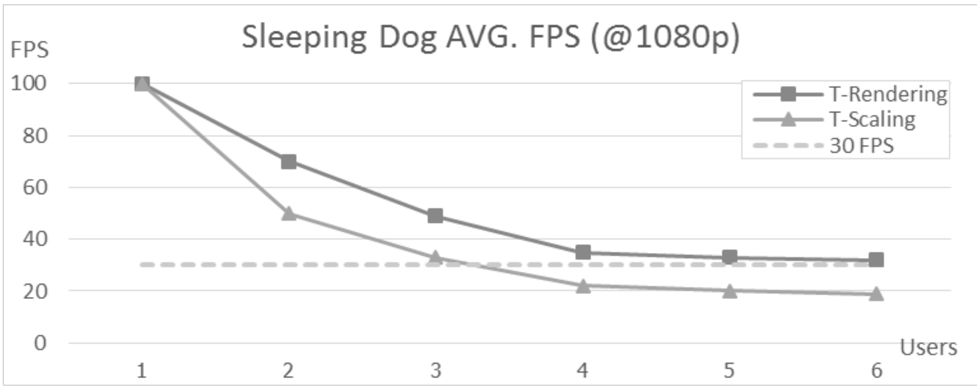


Figure 2-8 Average Frame per second for 1080p cloud gaming

Chapter 3
Effects of client variables to cloud gaming workload

3.1 Introduction

Cloud gaming workload characteristics have been widely studied in the past, but they do not take into account the client variables, such as the factor of display resolution at the client machines. As a result, the management of computer and network resources at the cloud gaming server cannot be efficient in realistic situations, where devices with heterogeneous display capabilities are found, and contents can be varied according to targeted resolution. In this section, the study on an effect of client variable to cloud gaming workload will be exposed. The main contribution of this section can be provided into two fold;

- First, we provide evidence (via empirical study) to confirm that the factor of display resolution must be concerned and taken into account the Cloud gaming workload characteristics.
- Second, we suggest an empirical equation that can be used to coarsely approximate the relation of client-resolution and cloud gaming workload.

This chapter is organized as follows. In section 3.2, we will explain what is cloud gaming workload. In section 3.3, we will declare what is the possible effects of display resolution to cloud gaming workload, following with our study methodology in section 3.4. In section 3.5 and 3.6, our performance studies are detailed and the evaluation results are described respectively. In section 3.7, we conclude this chapter.

3.2 The cloud gaming workload

Unlike traditional network game playing that all game workloads are handled at the client machine, Cloud gaming service will take care of the game workload computation at the cloud gaming server, and then output the result back to the client as a form of “High-definition video streaming” for each gaming session. By working in this manner, serious tasks for game graphic rendering can be avoided at the client machine. That is why low-end computers can be smoothly run graphics-intensive computer games without problems.

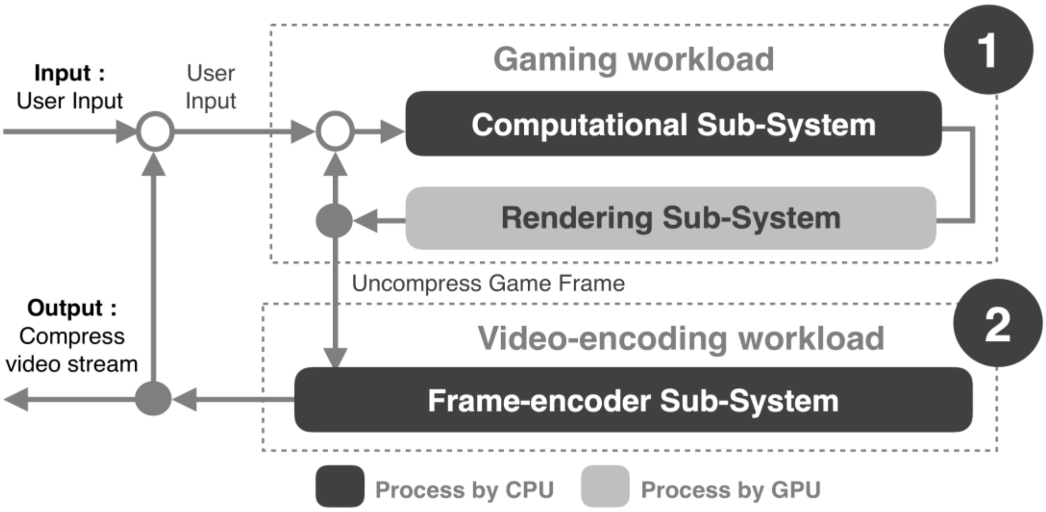


Figure 3-1 Simplified architecture of cloud gaming server

As shown in Fig. 3-1, the process of workload computation can be seen as an infinite loop on two different computing tasks for “Gaming workload” (Number 1) and “Video-encoding workload” (Number 2). In each loop of execution occurring when game application receives a user interaction command, all game-related issues, such as game intelligence, score calculation, collision detection, will be executed by the basic Central Processing Unit (CPU) in the “Computational sub-system”, however, the result of these issues will be processed to draw on many graphic frames by the Graphic Processing Unit (GPU) in the “Rendering sub-system”. Later, these frames will be managed to send to the client over the network. While many choices for video compression or even no compression can be applied, the decision is upon implementation. However, they will surely affect to the data size of resulted video frames, as seen in Table 3-1 as an example.

Table 3-1 Comparison of frame bandwidth in a certain game

Setup	Bandwidth
Uncompressed: (59.94 fps - 8-bit)	372.9 MB/sec
Compress with H.264	65.3 MB/sec

3.3 Possible effects of display resolution to Cloud gaming workloads

To the best of our knowledge, we found that client display resolution has been overlooked in many studies of cloud gaming workload characteristics. However, the following papers can be seen as evidences for support our idea that attempts to take client display resolution into account cloud gaming workload characteristics.

- For realizing the actual gaming workload, the work in [23] showed that game workload and frame complexity have indeed somewhat relationship, and can be formulated as a model for roughly prediction such as a case of Linear-In-Parameter (LIP) function in [24].
- For determining the actual video-ending workload, the work in [25] showed the relationship between encode video resolution and CPU usage, which can be also predicted by a proper mathematical approach such as that in [26].

3.4 Experimental Methodology

We setup our testing equipment as shown on Fig. 3-2, which we used Cisco 2960G switch and Cisco 3845 router to establish the gigabit-connection speed between the end-points.

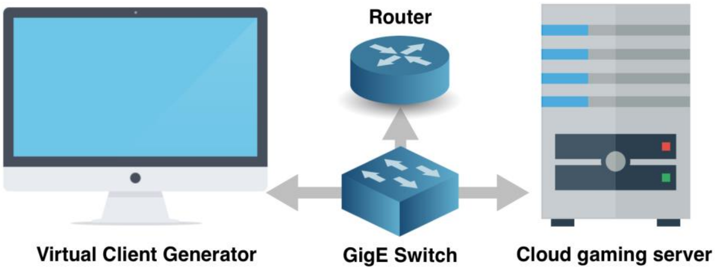


Figure 3-2 Equipment setup

At cloud gaming server, we developed a sample cloud gaming application whose workflow is shown in Fig. 3-3. This application ran on server, which contained of 3.5 GHz 6-Core Xeon E5-1650v2 processor, 16GB DDR3 main memory and Dual NVIDIA GeForce GTX Titan GPU. It would run 15 sampling games: Race Driver GRiD2, Need for speed: Rivals, Crysis 3, Metro 2033, Battlefield 4, Watch_dog, Grand Theft Auto IV, Resident Evil 6, Resident Evil: Revelation, Tomb Raider, The Elder scroll: Skyrim, Batman Arklam Origin, Titanfall, Bioshock: Infinite and Dragonball: Xenoverse.

At the client side, we used a PC which contains of 4.0 GHz Quad-core Intel Core i7, 16GB DDR3 main memory, to generate a sample game request with each resolution to cloud gaming server. Then, we pulled out working log and used it to explain the characteristic of cloud gaming workload hereafter.

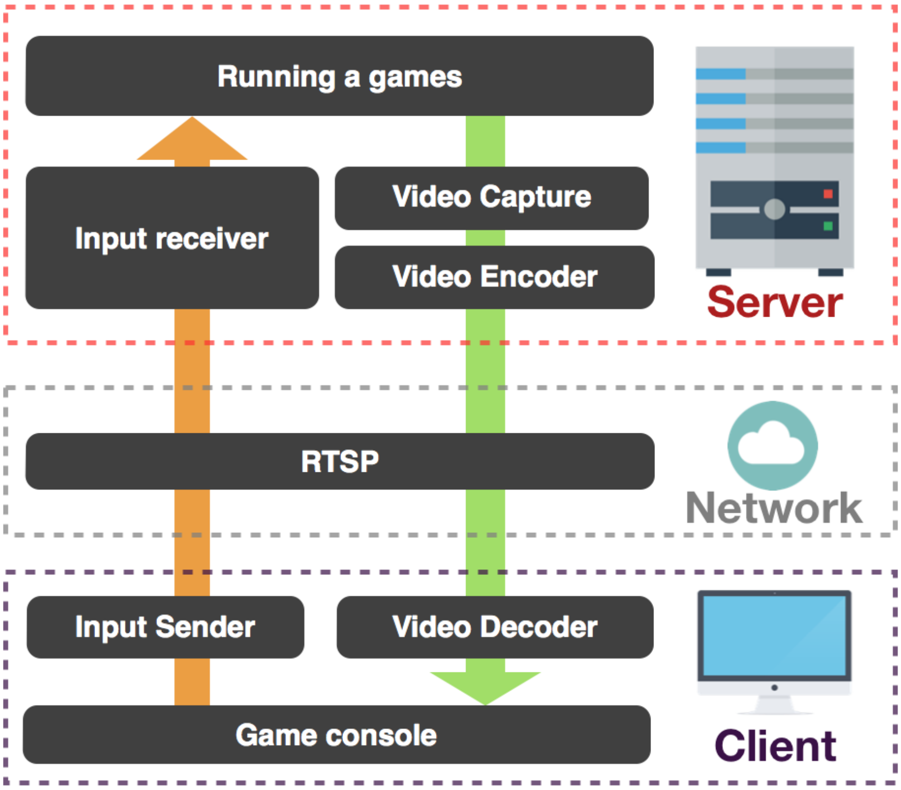


Figure 3-3 Our sample cloud gaming application workflow

3.5 Experimental Result

To clearly explain an effect from client display resolution to cloud gaming characteristic, we will show an experimental result by 4 aspects in this section, i.e. GPU Workloads, CPU Workload, Memory consumption and Network usage.

3.5.1 GPU Workload

GPU takes responsible on game rendering sub-system. If, GPU resources demand cannot be met. A drawing of graphic frames will be unable to catch a user interaction input, which affected to responsiveness and smoothness of the game session. In Fig. 3-4, we show the GPU workload when all adapted game contents are adjusted by display resolutions.

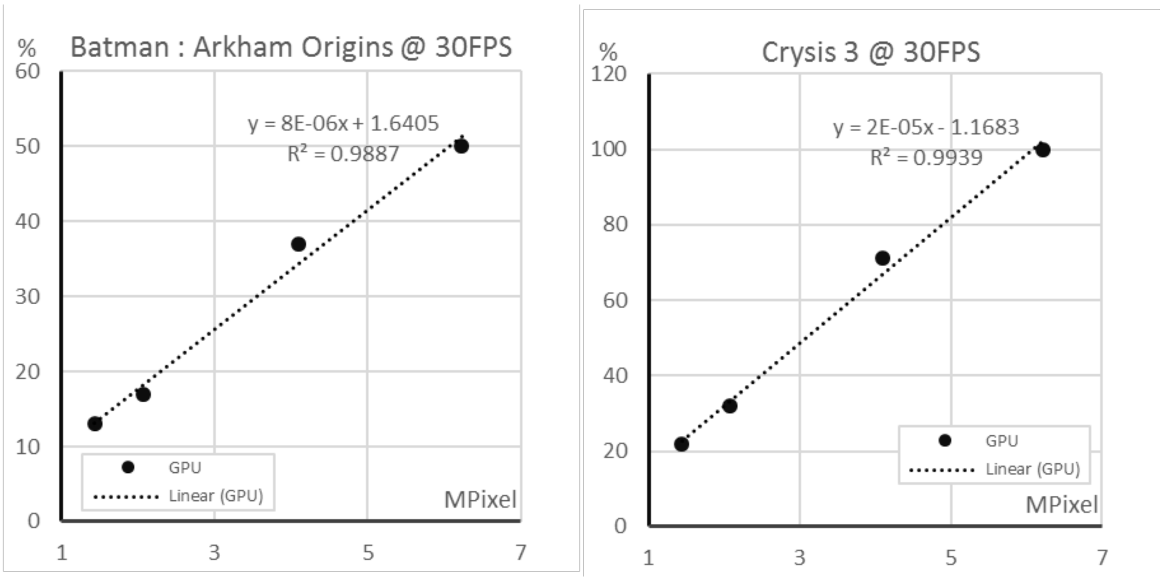


Figure 3-4 GPU workload

3.5.2 CPU Workload

In cloud gaming services, CPU mostly take responsibility on computational & frame-encoding sub-system. Hence, the inanition of CPU workload will cause game delay. In Fig. 3-5, we shown the CPU workload when display resolutions changed.

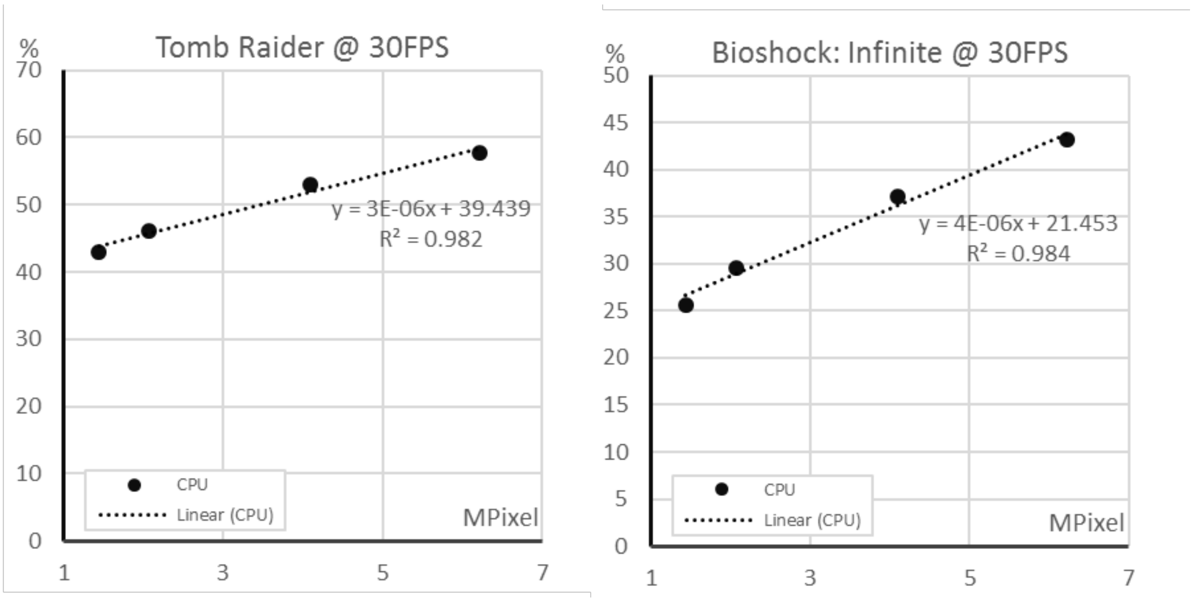


Figure 3-5 CPU workload

3.5.3 Memory Usage

Since most high-definition games will not only cache binary-related data temporally on the system memory, but also the GPU memory, therefore, both kinds of memory need to be studied. Here, the consumption of system memory and GPU memory are shown in Fig. 3-6 and Fig. 3-7 respectively.

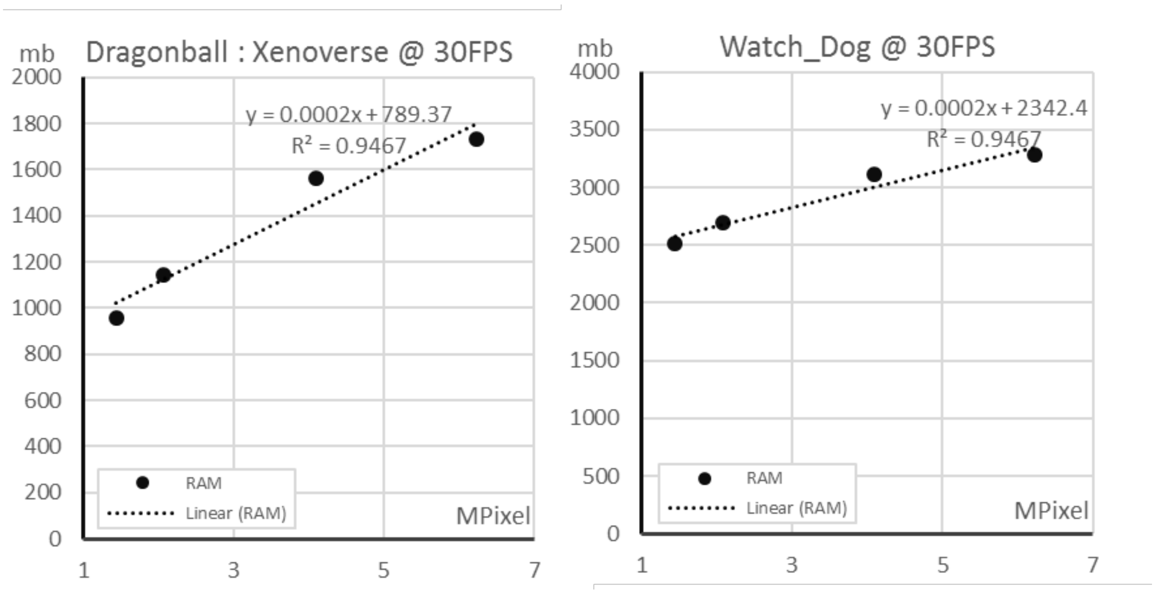


Figure 3-6 System Memory

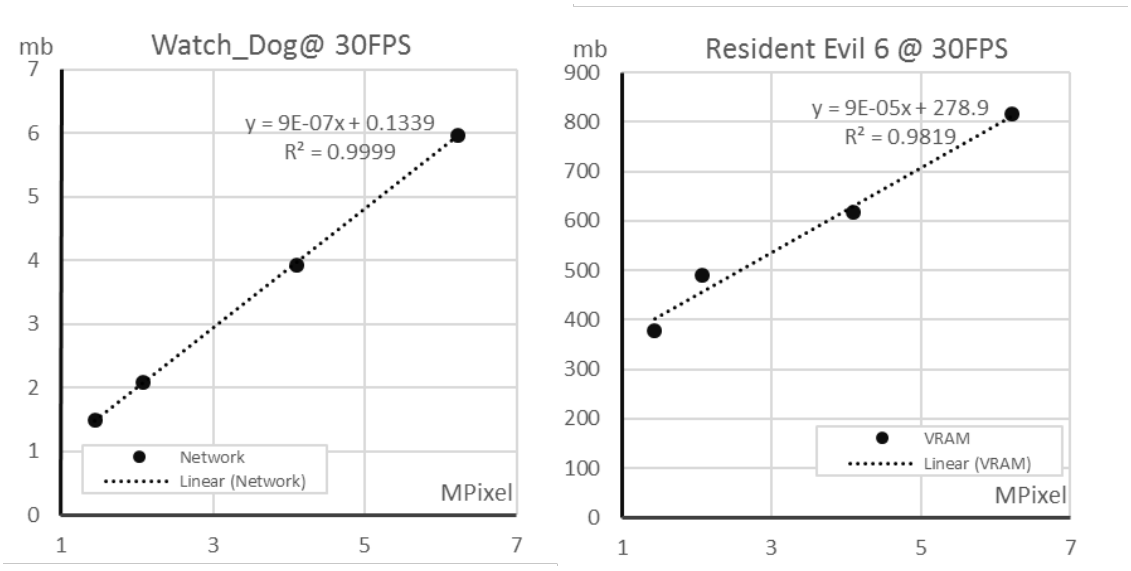


Figure 3-7 GPU Memory

3.5.4 Network Consumption

An insufficient of network bandwidth cause the delay of the game graphics transfer. In Fig. 3.8, we give a result of network consumption when client display resolution changed.

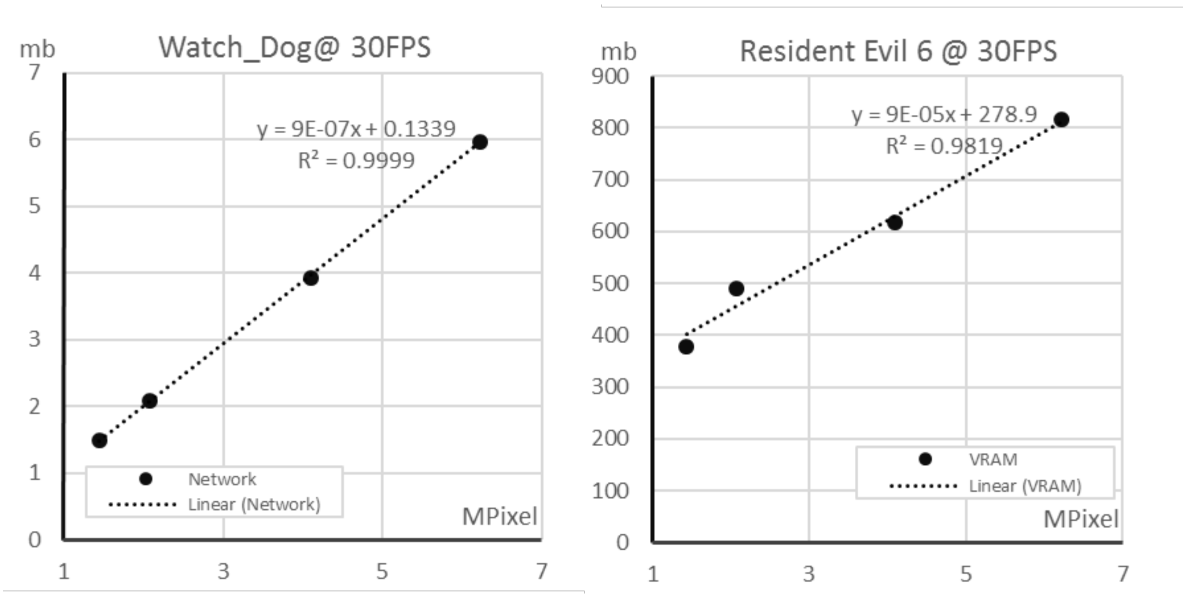


Figure 3-8 Network Consumption

3.6 Experimental Conclusion

From the empirical experiment in section 3.5, we can conclude the results into twofold.

- First, it can be seen clearly that client resolution is an important factor that crucially affects to cloud gaming workloads as well.
- Second, the relationship of cloud gaming workload and client display resolution can be coarsely approximated with a linear function. This can be confirmed by our experiments with less than 13% of error on average.

3.7 Conclusion

In this chapter, we have described how various display resolutions of cloud gaming clients should be taken into consideration for realizing the realistic resource utilization of heterogeneous devices. Based on our empirical study, we provide evidence that client display resolution should be a prime concern on efficient workload approximation. We also suggest that the approximate resource utilization of each gaming session can be coarsely made through a linear equation, which is derived through our experiment results. Our empirical study looks promising and give benefits straightforwardly to the cloud gaming workload provision and adaptation approach state in next section.

Chapter 4
Effects of omni-viewpoint to cloud gaming workload provision

4.1 Introduction

In order to enable the acceptable level of service quality for cloud gaming services, sufficient resources should be always maintained and workload provisioning is then necessary. However, taking only a limited set of server-related parameters, but ignoring the client-related parameters, in the typical formulation of optimization problem cannot yield for optimal workload provision in the cloud gaming environment nowadays, where the game server's computing processors can be driven by both the CPU and GPU, and the client devices' display resolutions are largely heterogeneous. In this section, the study on how the omni-viewpoint of all service related factors can be providing more efficient cloud gaming workload provision will be expose. The main contribution of this section can be provided into two fold;

- Technically, the problem of resource provision on traditional cloud gaming server can be formulated as an optimization problem so that many techniques for finding optimal solutions can be applicable. However, the bin-packing optimization which shows its attractive capability appeared in many works (such as [3, 5]). Thus, in this section, we will show how to integrate the omni-viewpoint of all CPU and GPU resources and the client devices' display resolution in the formulation of bin-packing problem.
- In addition, we will show an evidence that significantly improved benefits can be obtained by our formulation.

This chapter is structured as follows. In section 4.2, we describe some background of optimization-based service provisioning in the cloud gaming domain. Then, we give a detail of our proposed Multi-dimensional bin-packing optimization formulation in section 4.3, following with the performance evaluation of the improved system via experiments and the discussion on results in section 4.4 and 4.5 respectively. Finally, we conclude the chapter in section 4.6.

4.2 The decision making process of cloud gaming

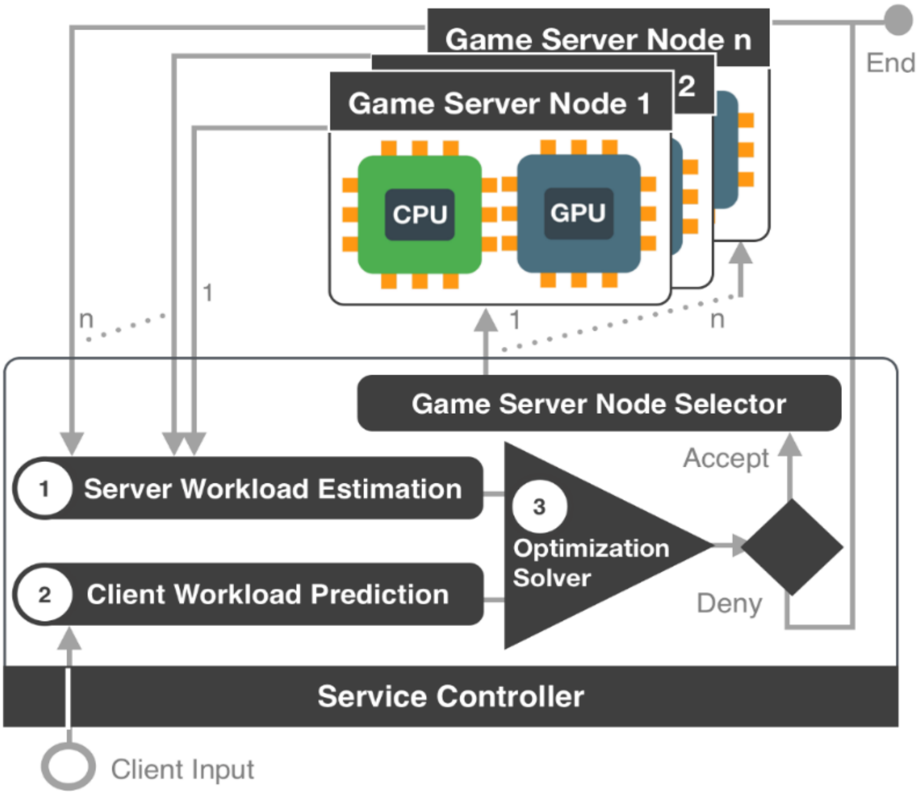


Figure 4-1 Conceptual diagram of the decision making process for admitting a client request

In typical cloud gaming, the decision making process for call admission or rejection when a request of game client asks for a connection can be illustrated in Fig. 4-1. In this regard, the server will decide whether or not a client should be admitted by the result of comparison between the resource availability of current server workload (Number 1) and the predicted resource utilization of client connection (Number 2), which is performed by the optimization solver (Number 3). Here, the optimization problem will well-served for resolving the optimal selection of game server node, and informing the game server node selector accordingly. Hence, it is obvious that the resource optimization problem should be properly formulated and effectively solved.

While a number of studies have attempted to find an optimal cloud gaming server resource provision by using different methods in the reviewed literature, they all share a common perspective of single resource optimization. More crucially, they are not efficient for implementing in the present cloud gaming technology, due to the following reasons:

- Firstly, they do not take into account a key factor of the display resolution of heterogeneous client device, which is proved to be a key influence on different levels of cloud gaming resource consumptions in [9].
- Secondly, they have misassumption that the cloud gaming service can be simply classified as "CPU-based” or "GPU-based" games in the similar way as the traditional game playing [36-38]. Unfortunately, this is utterly difference for cloud gaming because there are many tasks of cloud gaming that utilize both GPU and CPU such as the video encoding task. As evidence, an experiment in [38] can be used to confirm the importance of co-existed CPU and GPU operations for maintaining the service quality in cloud gaming.

Hence, in order to obtain the efficient implementation, it is required that the more suitable optimization problem should bring many types of cloud gaming resources into consideration.

4.3 Bin-Packing workload provision problem

In this section, we describe two forms of bin-packing optimization problems that have a potential for solving the cloud gaming workload provision; single-dimensional and multi-dimensional bin-packing problem.

4.3.1 Single-dimensional Bin-packing Optimization Problem

The first form is called Single-dimensional Bin-Packing problem (SBP), which has a primary objective to minimize the resource waste of single-constraint bin, while the total resource consumption from object doesn’t exceed bin capability.

In order to solve the provision problem of cloud gaming service, SBP may be declare as the formal statement in the following:

$$\text{Minimize:} \quad C - \sum_{i=1}^{n} c_i \qquad (1)$$

$$\text{Subject to:} \quad \sum_{i=1}^{n} c_i \leq C \qquad (2)$$

$$\forall_i \in \{1, \ldots, n\} \qquad (3)$$

where:

□ C is either GPU or CPU resources capacity of each server.

□ $\square_{\square}$ is GPU or CPU usage of each game workload.

Noticed that the primary objective of SBP as showed in (1) concerns only one resource. Hence, it will be calculated twice for two concerned resources, e.g. the first consideration is for CPU resource and then the other is for GPU resource.

4.3.2 Multi-dimensional Bin-packing Optimization Problem

The second form is called Multi-dimensional Bin-Packing problem (MBP), which takes a similar objective as the SBP above, but here many constrained bins can be involved. For a case of CPU, GPU and network resource consideration, the general form of MBP can be given below:

$$\text{Minimize:}\quad C \cdot G \cdot W - \sum_{i=1}^{n} c_i \cdot g_i \cdot w_i \qquad (4)$$

$$\text{Subject to:}\quad \sum_{i=1}^{n} c_i \leq C\,, \sum_{i=1}^{n} g_i \leq G, \sum_{i=1}^{n} w_i \leq W \qquad (5)$$

$$\forall_i \in \{1, \ldots, n\} \qquad (6)$$

where:

- *C, G, W* is the total resource of CPU, GPU and Network respectively.
- $\square_{\square}$, $\square_{\square}$ and $\square_{\square}$ is the requested workload of CPU, GPU and Network, which can be estimated by means of a linear function that expresses the relationship between the client resolution and the cloud gaming workload [9].

Noticed that the primary objective of MBP as showed in in (4) aims to minimize the waste of 3 server resources including CPU, GPU and Network for the allocation of new resource requirement, under the constraint of total workload in (5).

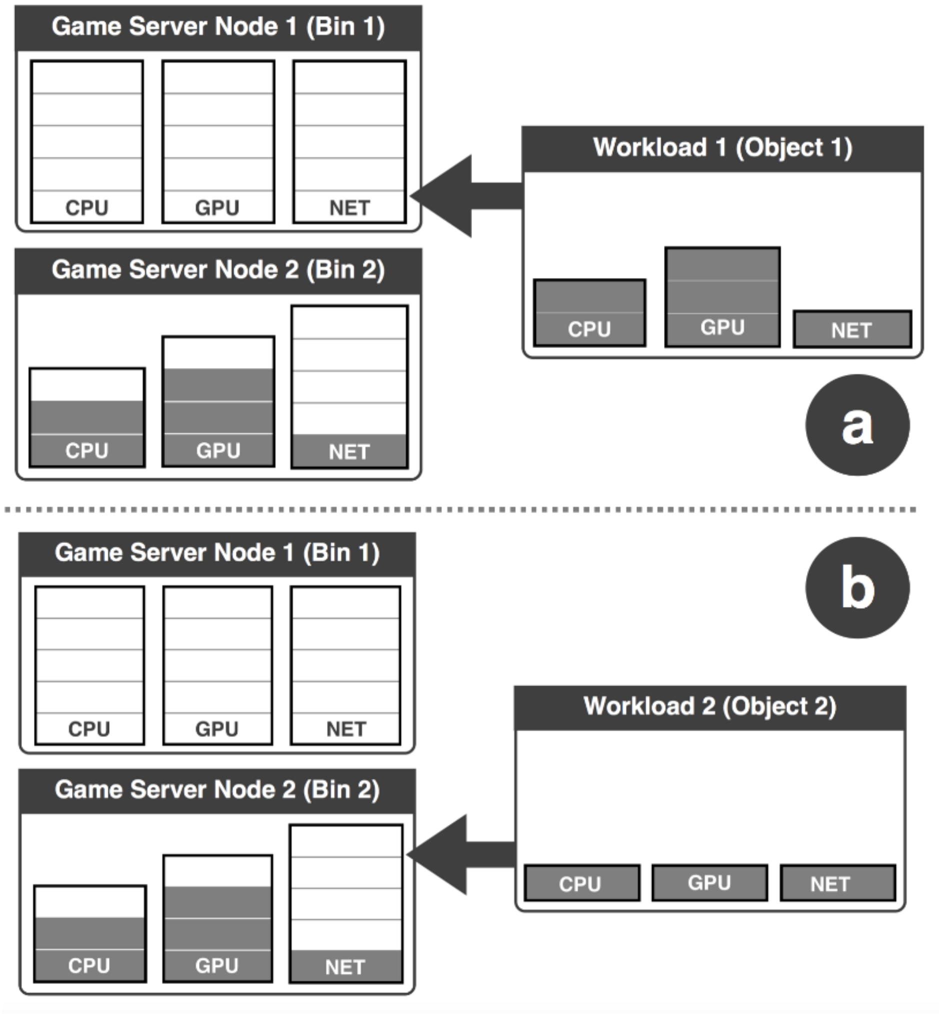

Game Server Node 1 (Bin 1)
CPU
GPU
NET
Game Server Node 2 (Bin 2)
CPU
GPU
NET
Workload 1 (Object 1)
CPU
GPU
NET
a
b
Game Server Node 1 (Bin 1)
CPU
GPU
NET
Game Server Node 2 (Bin 2)
CPU
GPU
NET
Workload 2 (Object 2)
CPU
GPU
NET

Figure 4-2 A simple case of Multi-dimensional Bin-Packing Problem

In Fig. 4-2, the illustration aimed to explain the MBP process in the simple manner. In Fig. 4-2(a), the Workload 1 (Object 1) will be assigned to the Bin 1, due to the insufficient size of Bin 2. In contrast, the Workload 2 (Object 2) in Fig. 4-2(b) will be assigned to the Bin 2, since the lower waste of resource will be obtained.

In order to solve MBP, a number of possible algorithms (e.g. first-fit, best-fit or first-fit decreasing algorithm) can be possibly used. However, in this paper, the first-fit decreasing algorithm will be interested particularly, due to the dominant feature of fast computation and effectiveness in solving this sort of problem [39-41]. In essence, this algorithm will firstly sort objects by the decreasing order, then attempt to place each object into the first possible accommodate bin.

4.4 Performance Evaluation

4.4.1 Experimental Setup

The cloud gaming experimental infrastructure as shown in Fig. 4-3 will be supported throughout the experiments. In this infrastructure, a gigabit connection will be served as a backbone by the Cisco ISR 3845 as a router and Cisco catalyst 4500-E with Supervisor engine VI as a core switch so that all machines, including all 5 servers, a service controller, and a virtual client generator.

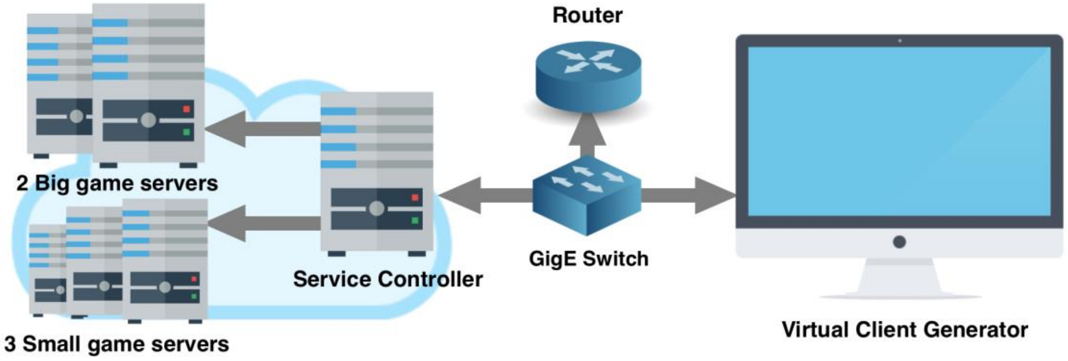


Figure 4-3 Experimental infrastructure

The group of game servers for performing game executions can be classified into two groups; Small server and Big server, for simulating the heterogeneous game server capabilities. The specification of the small servers consists of 3.4 GHz 4-core Core i5, 16GB DDR3 main memory and NVIDIA GeForce GTX 960 GPU, while that of big servers consists of 3.4 GHz 6-core Core i7, 16GB DDR3 main memory and NVIDIA GeForce GTX 970 GPU.

The service controller plays a crucial role of cloud gaming resource provision (referred to the functionality in Fig. 4-1). The specification of service controller consists of 2.7 GHz 4-core Core i5 and 4GB DDR3 main memory.

For a virtual client generator, it will be used to emulate many client machines. The specification of service controller consists of 4.0 GHz Quad-core Intel Core i7, 16GB DDR3 main memory. Here, the 5 sets of client machines will be generated according to the scale-down ratio of Internet connected devices collected statistically from 4 different sites (Stream online game store, Statista.com, Apprepim.com and Statcounter.com). The goal is to maintain the similar ratio of heterogeneous machines found in those cloud gaming sites. A different set of client machines with various display resolutions can be found in Table 7.

Table 4-1 Emulated Client Setup

Display Resolution	The number of client machines				
	Set 1	Set 2	Set 3	Set 4	Set 5
1080p	3	2	1	4	2
720p	2	3	1	2	4
640p	2	3	4	1	1
480p	1	0	2	1	1

4.4.2 Testing Method

We choose 9 best-selling games in 2014 (according to Forbes [42]), i.e. Call of Duty: Advanced Warfare, Madden NFL 15, Destiny, Grand Theft Auto 5, Minecraft, NBA 2K15, Watch Dogs, FIFA 15 and Call of Duty Ghosts, to provide cloud gaming service on our 5 game servers. In each set of client machines, each game will be requested in sequence and all service parameters (like Server usage, GPU waste, CPU waste and Network usage) on each game machine running different optimization problems, i.e. SBP for GPU, SBP for CPU and MBP, will be recorded accordingly.

4.4.3 Experimental Results

The 9 cloud gaming services have been tested, however only 3 different game characteristics will be selectively showed below, due to the page limit. Here are our choices; the Minecraft is represented for heavy-consumed CPU resource game, Destiny for heavy-consumed GPU resource game and Grand Theft Auto V for heavy-consumed CPU and GPU resource game. The resource consumptions of these games can be seen in Fig. 4-4

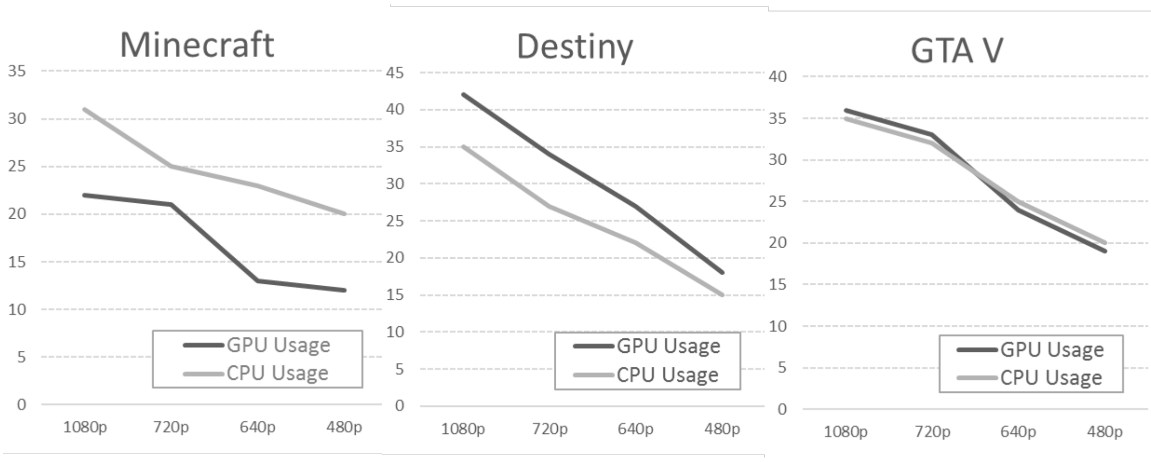


Figure 4-4 The resource consumption of Minecraft, Destiny and GTA V

Table 8 shows the number of game server machines that are required for running each game. As expected, the number of CPU and GPU resources will greatly depend on the game characteristics. Since the Minecraft is CPU-oriented game, it takes the total of 3 servers in the case of SBP-CPU method, which is higher than that of SBP-GPU method (i.e. 2 servers). However, it is not the same in the case of SBP-CPU and MBP, since it shares the same number of required resources. This is because both of CPU and GPU resources will be concerned together in the resource allocation problem solving by the MBP method.

Table 4-2 Optimization results: required servers for each game

Game	Formulation method	Number of game server require		
		Small game server	Big game server	Total
Minecraft	SBP-CPU	1	2	3
	SBP-GPU	1	1	2
	MBP	1	2	3
Destiny	SBP-CPU	2	0	2
	SBP-GPU	1	3	4
	MBP	2	0	2
Grand Theft Auto V	SBP-CPU	1	3	4
	SBP-GPU	2	1	3
	MBP	2	2	4

Fig. 4-5 shows the comparison results of resource utilization calculated by using the MBP, SBP-CPU and SBP-GPU methods. It is obvious that the given resources calculated by the MBP method will be sufficient in all cases. This is in contrast to the other methods, which can be the SBP-CPU or SBP-GPU method depending on the game type whether it heavily consumes on CPU or GPU resources.

For instance, the Minecraft demands the more resource of CPU than the GPU, the SBP-CPU method will be used to find the optimal CPU resource provision (depicted as the bar with diagonal lines), which will be later determined the number server machines and the volume of other resources (i.e. GPU and network) by looking up the values in Table 9. However, the result of SBP-GPU method is also given in this case for the clear performance comparison of these 3 methods

Table 4-3 Resources volumes for each kind of game server

Server Type	GPU	CPU	Network
Big game server	100	100	50
Small game server	40	60	50

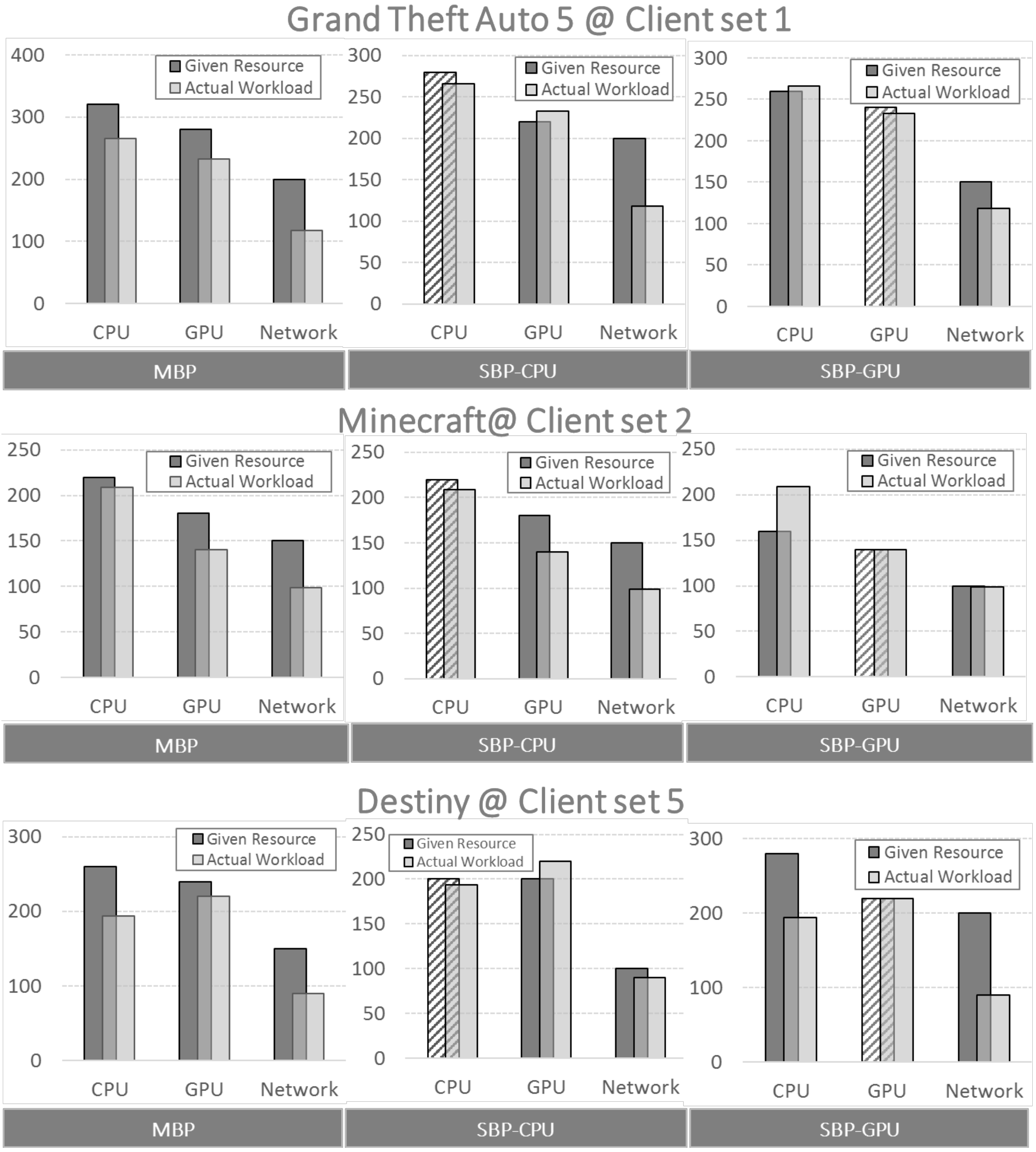


Figure 4-5 Experimental Result

In essence, by taking into consideration of all available resources in the bin-packing optimization problem like the MBP method will yield a better and sufficient resources for all game types taken into the experiments. This can effectively avoid the quality degradation, such as low-frame rate or frame-skip, due to the insufficiency of provided resources in the cloud gaming server as showed in Fig. 4-6.

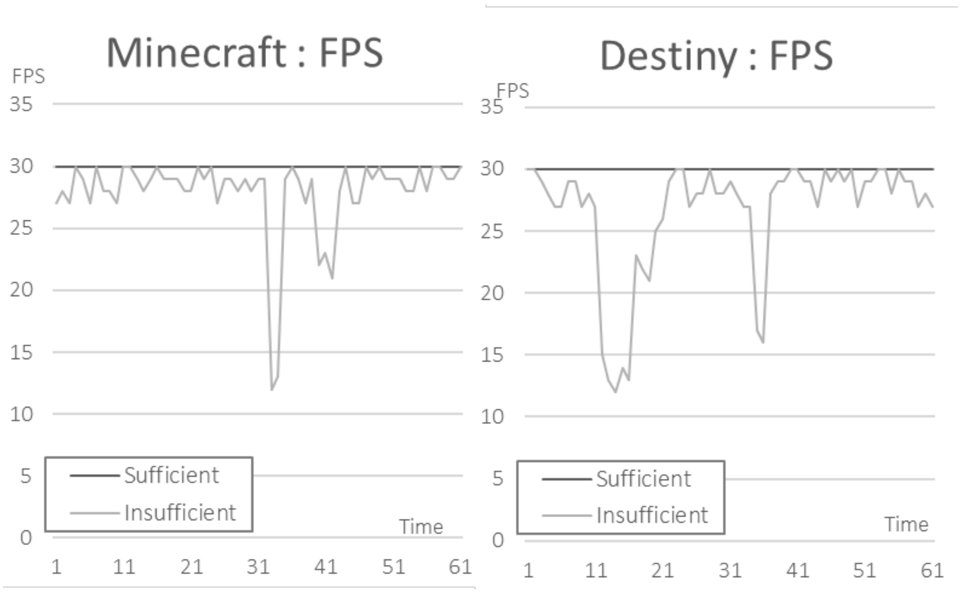


Figure 4-6 Effect of insufficient resource to game framerates

4.5 Conclusion

In this chapter we advocate on the use of Multi-dimensional Bin-Packing problem for determining the optimal workload provisioning in Cloud gaming servers, since a complete view of all resource availability will be taken into consideration for several advantages. This will yield a far more efficient resource utilization than the single-dimensional Bin-Packing problem as showed in our experiment results. As a result, the game service quality can be expected. Based on the evidence given in this paper, the MBP formulation method is extremely interesting and hence should be extensively used by cloud gaming service providers, or investigated further on improved performances by researchers in the cloud gaming community.

Chapter 5
Effects of omni-viewpoint to cloud gaming workload adaptation

5.1 Introduction

Methods for adapting cloud gaming workload can be divided according to client-centric or server-centric benefits. In this section, we argue that the better method should be performed in such a way that both client and server should share common benefits. By working on the omni-viewpoint, the basic of trans-rendering and trans-scaling mechanism are proposed to work together for obtaining better utilizations of network bandwidth and graphic processor altogether. The main contribution of this section can be provided into twofold;

- We will show how directional viewpoint of current client and server benefits method can be working altogether.
- We will show many evidences that significantly improved benefits can be obtained by our proposed method.

This chapter will be structured as follows. In section 5.2, some background of workload adaptation mechanisms in cloud gaming services will be provided, following with our newly proposed mechanism for mutual resource utilization in section 5.3. In section 5.4 and 5.5, details of our tests for performance evaluation and results will be described respectively. In section 5.6, we conclude the chapter.

5.2 Hybrid content-adaptation method

The basic workflow of Hybrid method as shown in Fig. 5-1. It begins with the server monitoring process, where the current status of server's GPU is observed and a decision for next sequential process according to the GPU status. If the GPU is in the non-overloaded status, the adaptation process will be directly executed. In contrast, if the GPU becomes overloaded, the newly proposed Mutual-benefits algorithm will be performed so that the balance point of the benefits between the client and the server can be computed.

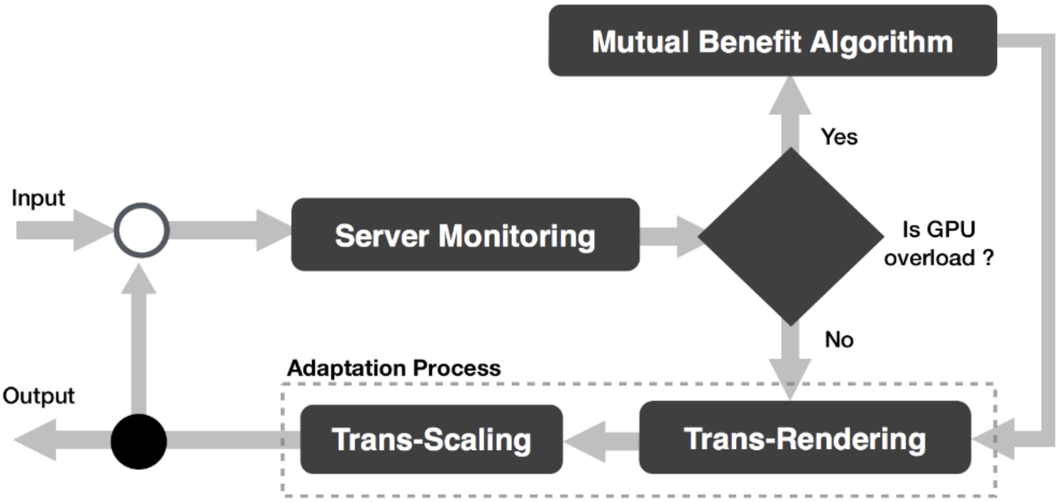


Figure 5-1 The workflow of Hybrid method

5.2.1 Adaptation process

This process is basically the combined process of trans-rendering and trans-scaling methods, where the ultimate goal is aimed at compromising the resulted benefits for both client and server. Therefore, the Trans-rendering and Trans-scaling will be executed accordingly. However, to understand the process in an easily manner, the illustration of adapted game scene from 640p to 1080p will be used as an example. In Fig. 5-2(a), the output obtained from the pure trans-rendering method will be lower the picture quality in order to maintain the picture size. This is in contrast to the pure trans-scaling in Fig. 5-2(b), where the picture size is not important, but the resolution will be the prime concern and hence the high GPU load is also demanding accordingly. Then, it becomes clear that our combined method of trans-rendering and trans-scaling can produce the higher quality of picture than the pure trans-rendering one, but lower demand of GPU load as seen in Fig. 5-2(c).

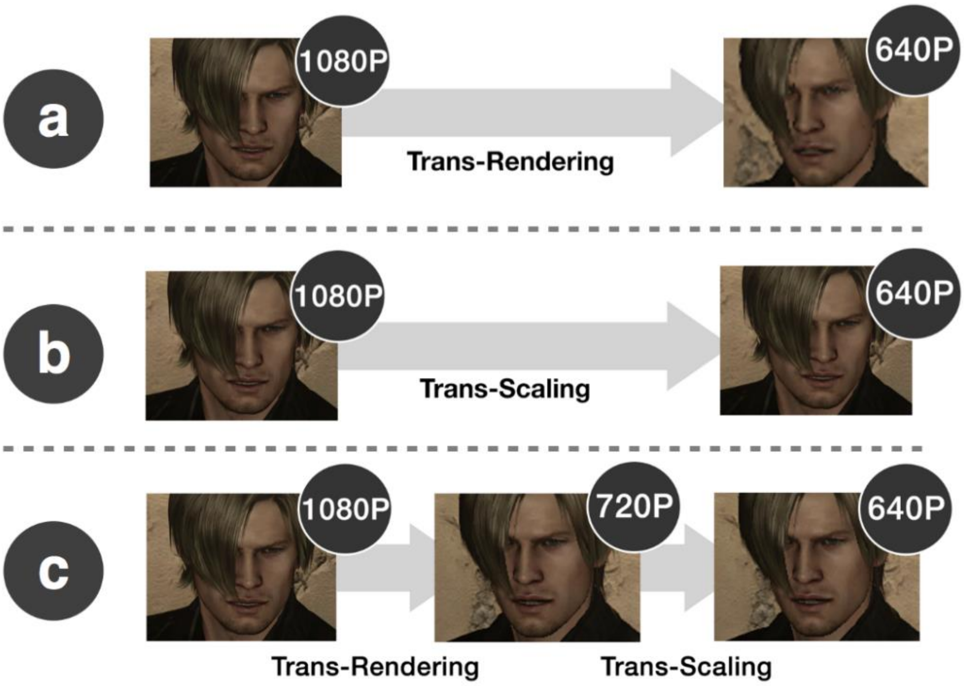


Figure 5-2 Comparison of a) pure trans-rendering, b) pure trans-scaling, and c) Combined trans-rendering and trans-scaling

5.2.2 Mutual Benefits Algorithm

The objective of Mutual-benefits algorithm is to find the balance of relation between server-centric benefits (GPU usage) and client-centric benefits (Render resolution), which are the linear relation due to our research conclusion in chapter three. In this regard, we can plot a linear relationship between game workload and resolution as a linear graph shown in Fig. 5-3, which is expressed as a function in (1). Therefore, we can always find the suitable resolution of cloud gaming, regarding the current GPU workload and the client screen size.

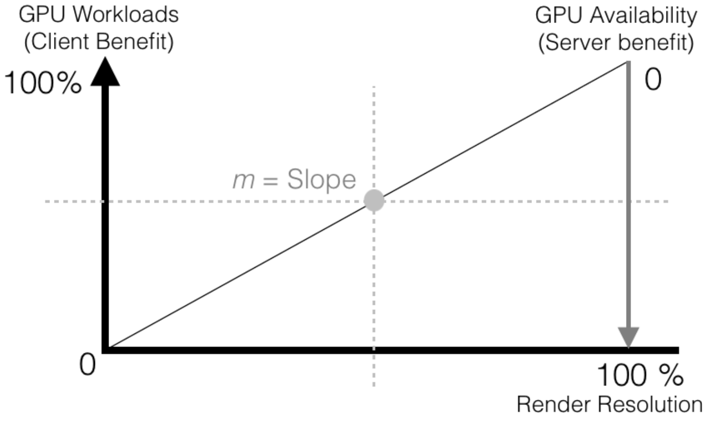


Figure 5-3 Relational between GPU Workload and Rendered resolution

$$x = \left(\frac{y}{m}\right) *ptarget \qquad (1)$$

Where y is percentage of server’s available GPU resource (e.g. 80%), m is slope of the relation between GPU workload and render resolution which depends on each game, and ptarget is pixel number of client display (e.g. 921,600 pixel for 1280x720 display). We can find the x from (1), that will be the compartments point for the adaptation process we state in section III a.

5.3 Experimental Methodology

To illustrate performance of hybrid method, we setup testing equipment as show in Fig. 5-4. On network infrastructure, we use Cisco 3845 router and Cisco 2960G gigabit switch to create gigabit connection between end-points.

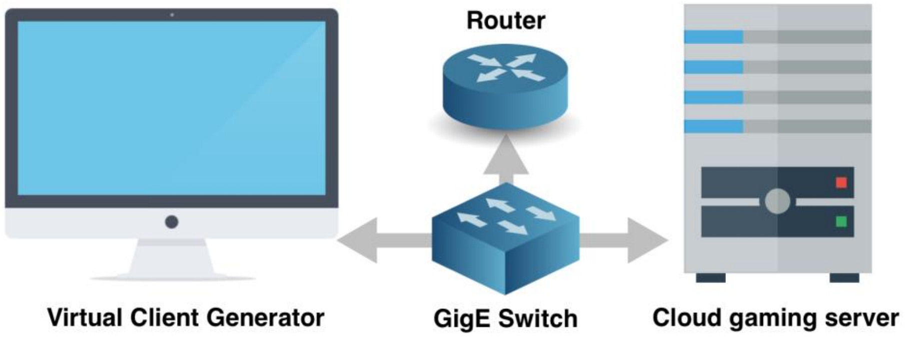


Figure 5-4 Network Scenario

On cloud gaming server, we use a server contains of 3.5 GHz 6-Core Xeon E5-1650v2 processor, 16GB DDR3 main memory and Dual NVIDIA GeForce GTX Titan GPU.

This server will run 5 chosen sampling games which represent the most popular game genres in nowadays, including "Race driver Grid 2" for racing genre, "Metro 2033" for the first person shooter genre, "Resident Evil 6" for the third person shooter genre, "The elder scrolls V: Skyrim" for role-playing and "Grand Theft Auto IV" for open-world genre. We will give detail about game setting in table 10.

Table 5-1 Experimental Sample Game

Games	Setup
Resident Evil 6	VSYNC On, FXAA3HQ, Motion Blur ON, Shadow Quality High, Texture Quality High, Screen Quality High
The Elder Scrolls V: Skyrim	Ultra Quality, AA OFF
Grand Theft Auto IV	Texture Quality: High, Render Quality: High, View Distance 33, Detail Distance 33, Vehicle Density 26, Shadow Density 3
Race Driver Grid 2	Ultra Quality, AA OFF, Advance Lighting
Metro 2033	DX11 Very High Quality + AA OFF

Table 5-2 Client Sampling

Tier	Sample	Steam	Web PC	Web PC+M	MobileSale	Avg
1	1920x1080	4	1	1	1	2
2	1600x900	2	1	1	0	1
3	1280x720	4	7	5	1	4
4	800x600	0	1	2	4	2
5	320x240	0	0	1	4	1
Total		10	10	10	10	10

Note: "**Steam**" group come from number of steam user, "**Web PC**" and "**Web PC+M**" come from client who surf the web on PC and on PC combine with mobile device respectively, "**Mobile Sale**" come from mobile devices sale on 2014 and "**Avg**" is an average.

On cloud gaming client, for convenience to collect results from sampling, we use a high-end desktop PC to emulate a sampling device rather than a real hardware. We generate 5 sampling groups. Each group consists of 10 emulated clients, which estimates from 4 global statistic on internet connected devices (Stream online game store, Statista.com, Apprepim.com and Statcounter.com). As we show in table 5-2.

We use these above emulated clients to send command to sampling server. Then we pull out working log from both sides and use it to explain the efficiency of hybrid method hereafter.

5.4 Experimental Result

To explain the efficiency of hybrid method, we will show an experimental result by two aspects, client and server benefit aspects.

5.4.1 Client-centric benefits

For obtaining client-centric benefits, cloud gaming needs to render game graphics at the compatible resolution for displaying on the client machine, and at the same time, responsiveness and smoothness of the game session must be maintained. It also needs to limit network usage.

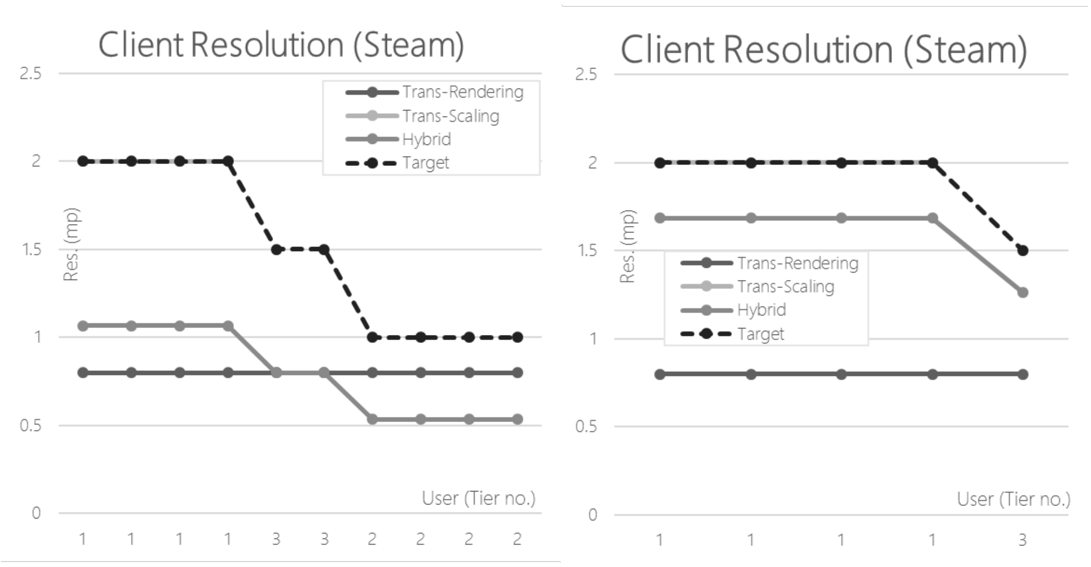


Figure 5-5 Render resolution for each client. Sample by "Steam" client scenario; congested (Left) and uncongested (Right)

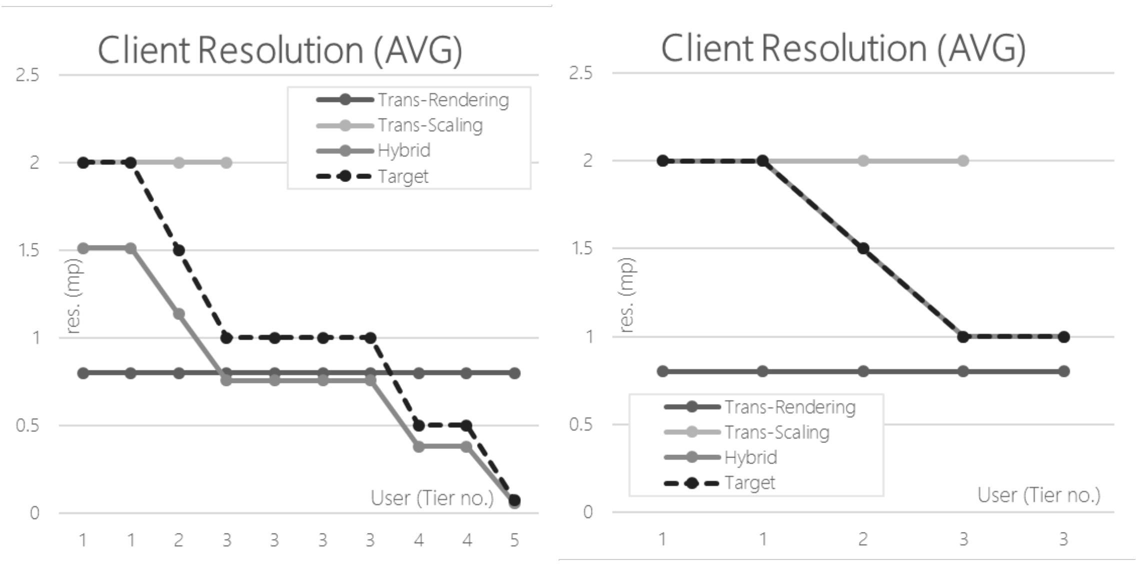


Figure 5-6 Render resolution for each client. Sample by "Average" client scenario; congested (Left) and uncongested (Right)

Fig. 5-5 and Fig. 5-6 shows a rendered resolution for each user on both scenarios of uncongested and congested loads. During uncongested periods, the hybrid method will render every pixels requested by the client. During congestion, this method will instead render at near client resolution, rather than those resolutions obtained from trans-rendering and trans-scaling methods. For enabling the precise analyst, we use "Root Mean Square Error" (RMSE) to determine the dependence between content adaptation method and client display resolution. The results are shown in table 5-3, which can be noticed that the hybrid method has the lowest RMSE. This means the output of hybrid method have more congruence to device resolution than the other methods.

Table 5-3 Relative between content adaptation method and target resolution By root mean square error.

Method	RMSE (M. Pixel)
Trans-Rendering	0.597
Trans-Scaling	0.794
Hybrid	0.405

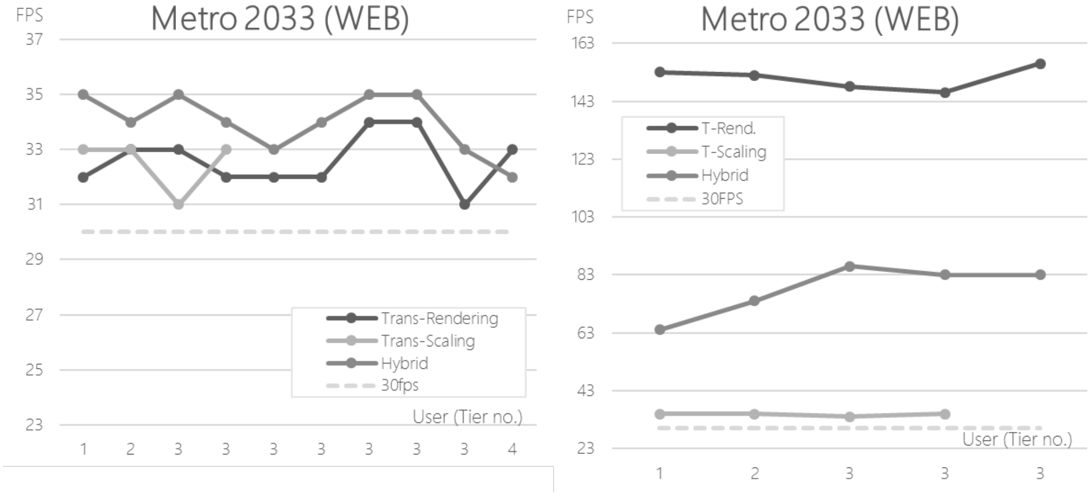


Figure 5-7 Average FPS for each client. Sample by "Web" client scenario. congest (Left) and uncongested (Right)

Beside the render resolution, the clients yet require the other service quality, e.g. smoothness and responsiveness of the game session, which have a close relationship with the values of frame-per-second (FPS) and the latency. Fig. 5-7 reveals number of an average FPS, which shows the hybrid method's ability to comfortably deliver more than 30 FPS in congest and more than 60 FPS in uncongested situation. That means the clients will always get reasonable smooth game experience on both situations.

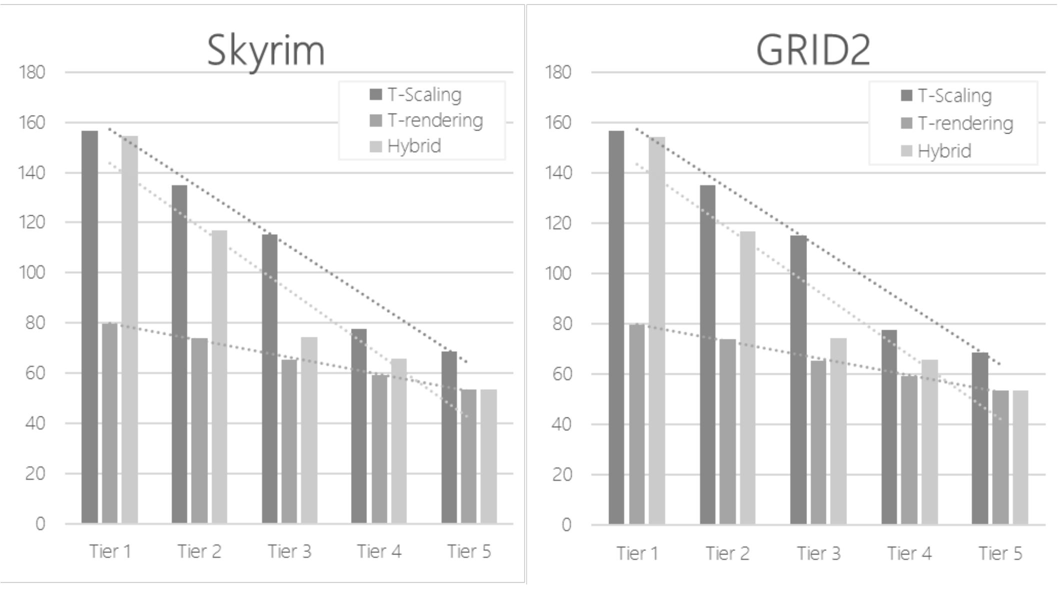


Figure 5-8 Game latency for each type of user (in millisecond)

Fig. 5-8 shows a number of game latency, which explains that an implementation of hybrid method does not cause significant latency compared to the other method. Most importantly, it can keep the game latency below the acceptable region of 200ms, which surely has no effect on human visual perception [19]. It can be noticed in Fig. 5-9 that this feature can be obtained on the network usage that is competent to the other methods

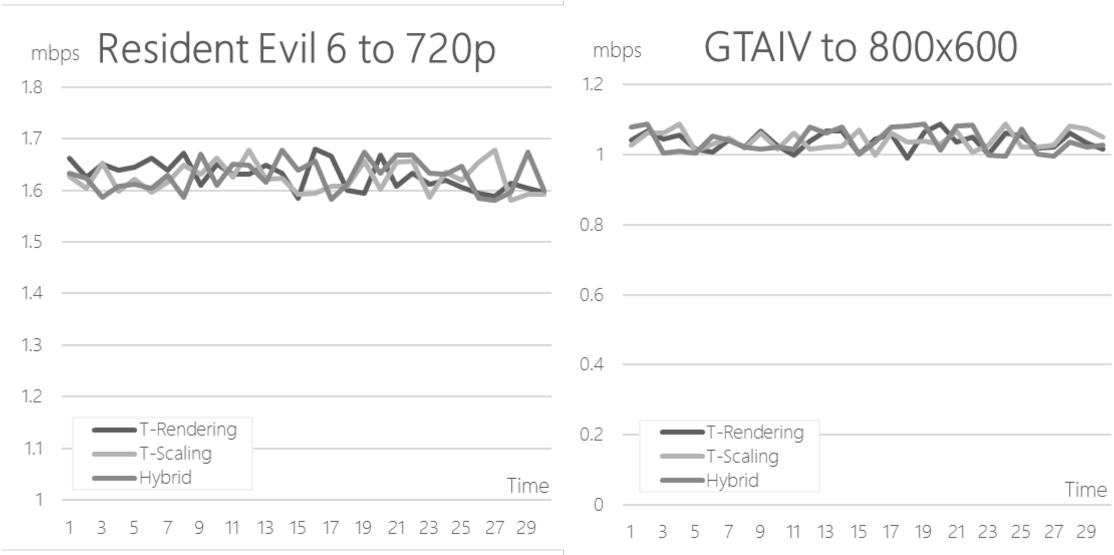


Figure 5-9 Network Usage

5.4.2 Server-centric benefits

For obtaining server-centric benefits, the game server will attempt to increase their user capability, lower the GPU usage, and maintain the network bandwidth, when the number of users are increased.

Because of the server capability depends on the number of client FPS and latency, it is showed in Fig. 5-10 and Fig. 5-11 that hybrid method can deliver the comparative number of clients as the trans-rendering method, but significantly better than those of trans-scaling method. In addition, when looking at the GPU utilization in Fig. 5-12, the hybrid method is yet slowly increasing, compared with the trans-scaling method.

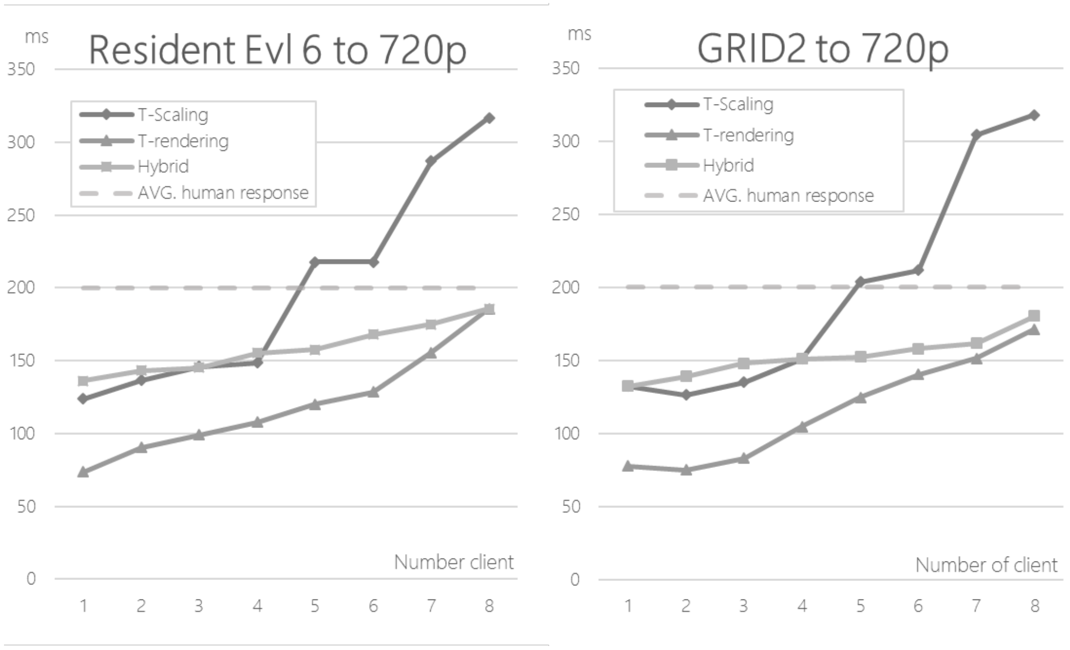


Figure 5-10 Game latency by the number of users.

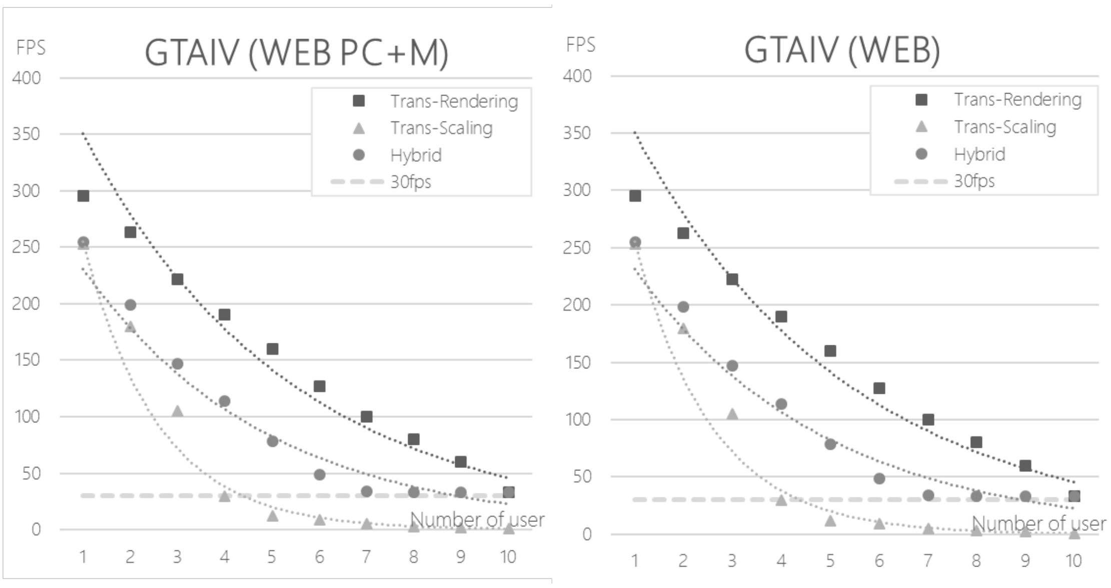


Figure 5-11 Average FPS by the number of users.

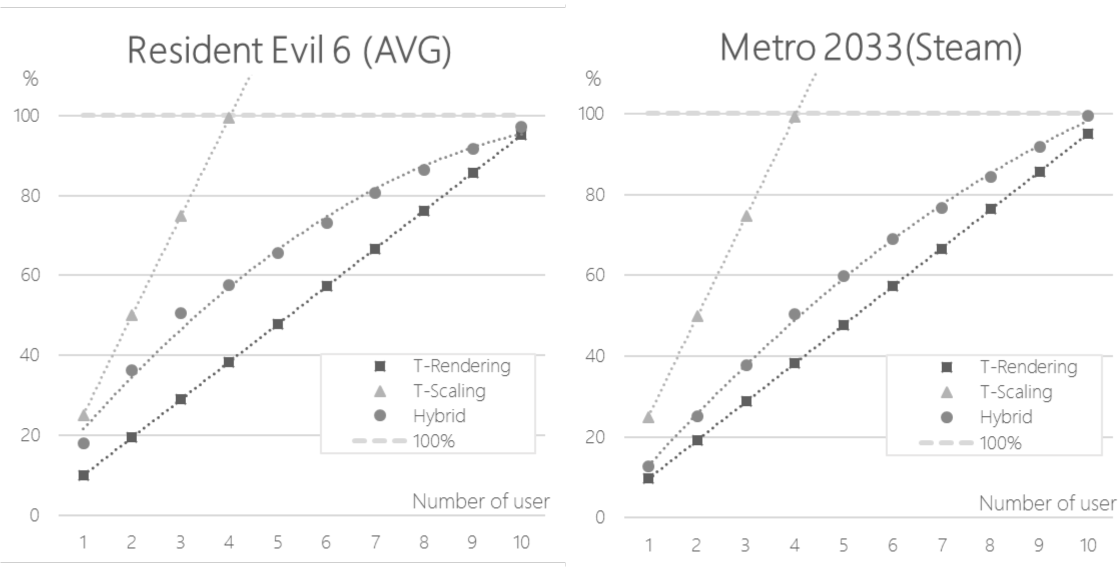


Figure 5-12 GPU Workload by the number of users.

5.5 Experimental Conclusion

Based on all experiments explained above, we can give a complete picture of performance issues in Fig. 5-13. It can be clearly seen that the hybrid method are outstanding on many issues better than other two counterparts, except for the average game FPS and the game latency.

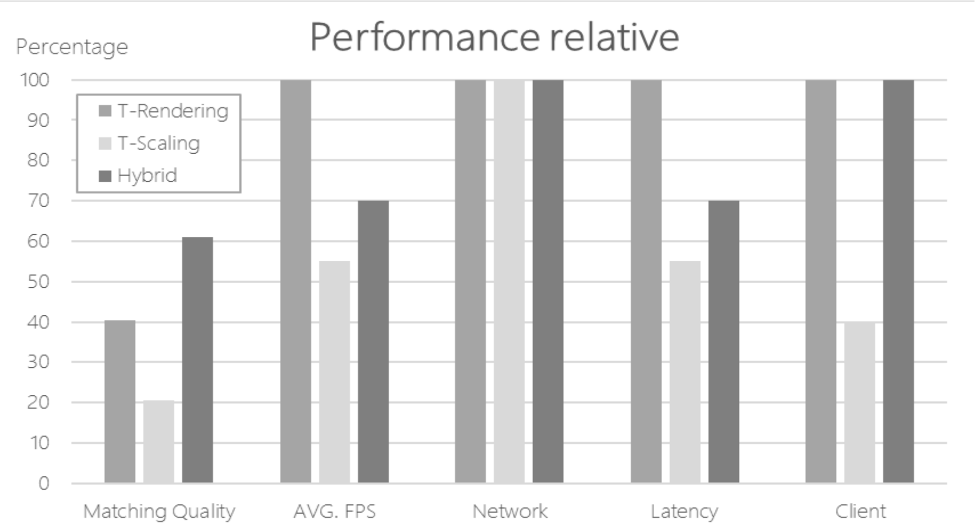


Figure 5-13 Experiment summary.

5.6 Conclusion

In this chapter, we introduce the combination of trans-rendering and trans-scaling methods aiming at adaptation for cloud gaming content and taking both server-centric and client-centric benefits into consideration. By means of experiments, we showed evidences that our proposed method is outstanding in terms of picture quality and associated GPU workloads, in comparison with trans-rendering and trans-scaling methods performing alone. We believe the content of this section will give substantial benefits for researchers, who also work in the field of cloud gaming.

Chapter 6
Conclusion & Future Works

6.1 Conclusion

In this thesis, we argue that the facing problems of current workload provision and workload adaptation approaches in cloud gaming services are caused by the partial view of resource management (i.e. computing and network resources). Our assumption is that an efficient solution for resolving these problems must be taken into consideration the global viewpoint of computing resources under management. In order to approve this assumption, the research studies have been prudently performed, and the conclusion can be given in the followings:

- In chapter 3, we perform the relationship of server-side and client-side parameters so that the full understanding of their relationships affects to the characteristics of cloud gaming workload can be defined. After analyzing the experimental results, we can propose a linear function that can be well-used for predicting the future workload, which is necessary for determining the demanded resource of a newly request connection.
- In chapter 4, the key issue is on the workload provision approach in cloud gaming environment. We take the proposed model (obtained from the chapter 3) to relax the burden of demanded resource computation for a new gaming connection. It is in the sense that the increase of CPU, GPU and network loads can be easily approximated on the server machine. Based on these calculation, an optimal resource allocation on each resource can be made possible. Then, we give a demonstration of effective resource utilization results via multi-variable bin-packing optimization technique.
- In chapter 5, the key issue is on the workload adaptation approach in cloud gaming environment. On the favor of global viewpoint of resource management, we combine the use of client-centric and server-centric benefits for fulfilling the workload adaptation. Our hybrid method is then proposed and show the outstanding in terms of better graphics quality and more efficient server resource utilization than the legacy one.

6.2 Future Works

The work in this thesis can extend in some possible directions as follows.

6.2.1 Better modeling the characteristics of cloud gaming workload

Since our proposed model in chapter 3 works is approximate in nature via a basis of linear function, it can just provide an approximation of resource in a short period of time on the demand that needs to be reserved for a new gaming request. It would be better if the finer-granular approximation techniques are replaced since the more precision result of resource allocation can be achieved than the current one. However, this may achieve on the expense of calculation time.

6.2.2 Exploring the other optimization techniques for the workload provision

The first-fit decreasing algorithm for finding the optimal solution of bin-packing problem are simply selected for giving a demonstration, due to the sake of simplicity. In this regard, it would be interesting to investigate on the application of other bin-packing algorithms so that more efficient results can be expected.

References

[1] K.-T. Chen, C.-Y. Huang, and C.-H. Hsu, “Cloud gaming onward: research opportunities and outlook,” in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–4.

[2] S. Choy, B. Wong, G. Simon, and C. Rosenberg, “The brewing storm in cloud gaming: A measurement study on cloud to end-user latency,” in *2012 11th Annual Workshop on Network and Systems Support for Games (NetGames)*, 2012, pp. 1–6.

[3] J. Y. Chao Zhang, “vGASA: Adaptive Scheduling Algorithm of Virtualized GPU Resource in Cloud Gaming,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 11, pp. 3036–3045, 2014.

[4] M. Yu, C. Zhang, Z. Qi, J. Yao, Y. Wang, and H. Guan, “VGRIS: Virtualized GPU Resource Isolation and Scheduling in Cloud Gaming,” in *Proceedings of the 22Nd International Symposium on High-performance Parallel and Distributed Computing*, New York, NY, USA, 2013, pp. 203–214.

[5] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, “A game-theoretic method of fair resource allocation for cloud computing services,” *J Supercomput*, vol. 54, no. 2, pp. 252–269, Jul. 2009.

[6] V. P. Anuradha and D. Sumathi, “A survey on resource allocation strategies in cloud computing,” in *2014 International Conference on Information Communication and Embedded Systems (ICICES)*, 2014, pp. 1–7.

[7] D. Gmach, J. Rolia, and L. Cherkasova, “Satisfying Service Level Objectices in a Self-Managing Resource Pool,” in *Proceedings of the 2009 Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems*, Washington, DC, USA, 2009, pp. 243–253.

[8] A. Kochut, K. Beaty, H. Shaikh, and D. G. Shea, “Desktop workload study with implications for desktop cloud resource optimization,” in *2010 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW)*, 2010, pp. 1–8.

[9] H. Goudarzi and M. Pedram, “Multi-dimensional SLA-Based Resource Allocation for Multi-tier Cloud Computing Systems,” in *2011 IEEE International Conference on Cloud Computing (CLOUD)*, 2011, pp. 324–331.

[10] J. Li, M. Qiu, J.-W. Niu, Y. Chen, and Z. Ming, “Adaptive resource allocation for preemptable jobs in cloud systems,” in *2010 10th International*

Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 31–36.

[11] K. H. Kim, A. Beloglazov, and R. Buyya, “Power-aware Provisioning of Cloud Resources for Real-time Services,” in *Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science*, New York, NY, USA, 2009, pp. 1:1–1:6.

[12] K. Kumar, J. Feng, Y. Nimmagadda, and Y.-H. Lu, “Resource Allocation for Real-Time Tasks Using Cloud Computing,” in *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, 2011, pp. 1–7.

[13] K.-C. Huang and K.-P. Lai, “Processor allocation policies for reducing resource fragmentation in multi-cluster grid and cloud environments,” in *Computer Symposium (ICS), 2010 International*, 2010, pp. 971–976.

[14] S. Islam, J. Keung, K. Lee, and A. Liu, “Empirical Prediction Models for Adaptive Resource Provisioning in the Cloud,” *Future Gener. Comput. Syst.*, vol. 28, no. 1, pp. 155–162, 2012.

[15] S. Wang and S. Dey, “Rendering Adaptation to Address Communication and Computation Constraints in Cloud Mobile Gaming,” in *2010 IEEE Global Telecommunications Conference (GLOBECOM 2010)*, 2010, pp. 1–6.

[16] A. Jurgelionis, P. Fechteler, P. Eisert, F. Bellotti, H. David, J. P. Laulajainen, R. Carmichael, V. Poulopoulos, A. Laikari, P. Perälä, A. De Gloria, C. Bouras, A. Jurgelionis, P. Fechteler, P. Eisert, F. Bellotti, H. David, J. P. Laulajainen, R. Carmichael, V. Poulopoulos, A. Laikari, P. Perälä, A. De Gloria, and C. Bouras, “Platform for Distributed 3D Gaming, Platform for Distributed 3D Gaming,” *International Journal of Computer Games Technology, International Journal of Computer Games Technology*, vol. 2009, 2009, p. e231863, Jun. 2009.

[17] P. Eisert, “Remote rendering of computer games,” in *in Proc. Internation Conference on Signal Processing and Multimedia Applications (SIGMAP*, 2007.

[18] D. De Winter, P. Simoens, L. Deboosere, F. De Turck, J. Moreau, B. Dhoedt, and P. Demeester, “A Hybrid Thin-client Protocol for Multimedia Streaming and Interactive Gaming Applications,” in *Proceedings of the 2006 International Workshop on Network and Operating Systems Support for Digital Audio and Video*, New York, NY, USA, 2006, pp. 15:1–15:6.

[19] R. Aparicio-Pardo, K. Pires, A. Blanc, and G. Simon, “Transcoding Live Adaptive Video Streams at a Massive Scale in the Cloud,” in *Proceedings of*

the 6th ACM Multimedia Systems Conference, New York, NY, USA, 2015, pp. 49–60.

[20] M. Hemmati, A. Javadtalab, A. A. Nazari Shirehjini, S. Shirmohammadi, and T. Arici, “Game As Video: Bit Rate Reduction Through Adaptive Object Encoding,” in *Proceeding of the 23rd ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, New York, NY, USA, 2013, pp. 7–12.

[21] V. Delwadia, S. Marshall, and I. Welch, “The Effect of User Interface Delay in Thin Client Mobile Games,” in *Proceedings of the Eleventh Australasian Conference on User Interface - Volume 106*, Darlinghurst, Australia, Australia, 2010, pp. 5–13.

[22] A. Pathania, Q. Jiao, A. Prakash, and T. Mitra, “Integrated CPU-GPU power management for 3D mobile games,” in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2014, pp. 1–6.

[23] B. Dietrich, S. Nunna, D. Goswami, S. Chakraborty, and M. Gries, “LMS-based low-complexity game workload prediction for DVFS,” in *2010 IEEE International Conference on Computer Design (ICCD)*, 2010, pp. 417–424.

[24] M. Shafique, M. U. K. Khan, and J. Henkel, “Power efficient and workload balanced tiling for parallelized high efficiency video coding,” in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 1253–1257.

[25] Y. Huang, A. Tran, and Y. Wang, “A Workload Prediction Model for Decoding Mpeg Video and Its Application to Workload-scalable Transcoding,” in *Proceedings of the 15th ACM International Conference on Multimedia*, New York, NY, USA, 2007, pp. 952–961.

[26] M. van der Schaar and Y. Andreopoulos, “Rate-distortion-complexity modeling for network and receiver aware adaptation,” *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 471–479, Jun. 2005.

[27] T. Lan, Y. Chen, and Z. Zhong, “MPEG2 decoding complexity regulation for a media processor,” in *2001 IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp. 193–198.

[28] Z. Song, “Adaptive Resource Provisioning for the Cloud Using Online Bin Packing,” *Computers, IEEE Transactions on*, vol. 63, no. 11, pp. 2647–2660, 2014.

[29] K. Sembiring and A. Beyer, “Dynamic Resource Allocation for Cloud-based Media Processing,” in *Proceeding of the 23rd ACM Workshop on Network*

and Operating Systems Support for Digital Audio and Video, New York, NY, USA, 2013, pp. 49–54.

[30] A. Khan, X. Yan, S. Tao, and N. Anerousis, “Workload characterization and prediction in the cloud: A multiple time series approach,” in *2012 IEEE Network Operations and Management Symposium (NOMS)*, 2012, pp. 1287–1294.

[31] J. Brandt and L. Wolf, “Adaptive Video Streaming for Mobile Clients,” in *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, New York, NY, USA, 2008, pp. 113–114.

[32] M. Eberhard, L. Celetto, C. Timmerer, E. Quacchio, H. Hellwagner, and F. S. Rovati, “An interoperable multimedia delivery framework for scalable video coding based on MPEG-21 Digital Item Adaptation,” in *2008 IEEE International Conference on Multimedia and Expo*, 2008, pp. 1607–1608.

[33] N. Cranley, P. Perry, and L. Murphy, “Dynamic content-based adaptation of streamed multimedia,” *Journal of Network and Computer Applications*, vol. 30, no. 3, pp. 983–1006, 2007.

[34] A. Hutter, P. Amon, G. Panis, E. Delfosse, M. Ransburg, and H. Hellwagner, “Automatic adaptation of streaming multimedia content in a dynamic and distributed environment,” in *IEEE International Conference on Image Processing, 2005. ICIP 2005*, 2005, vol. 3, pp. III–716–9.

[35] “Can you, list which games are ‘heavily CPU-based’, ‘heavily GPU-based’, or both? - Forums - PCPartPicker.” [Online]. Available: https://pcpartpicker.com/forums/topic/87771-can-you-list-which-games-are-heavily-cpu-based-heavily-gpu-based-or-both. [Accessed: 08-Feb-2016].

[36] “The game is CPU-bound, not GPU-bound.□:: Dying Light,” *Steam Community*. [Online]. Available: https://steamcommunity.com/app/239140/discussions/0/604941528466981109/. [Accessed: 08-Feb-2016].

[37] S. Peak, “Quad-Core Gaming Roundup: How Much CPU Do You Really Need?,” *PC Perspective*, 06-Jul-2015. [Online]. Available: http://www.pcper.com/reviews/Systems/Quad-Core-Gaming-Roundup-How-Much-CPU-Do-You-Really-Need. [Accessed: 08-Feb-2016].

[38] R. Lewis, “A general-purpose hill-climbing method for order independent minimum grouping problems: A case study in graph colouring and bin

packing," *Computers & Operations Research*, vol. 36, no. 7, pp. 2295–2310, 2009.

[39] M. David and S. Johnson, "A 71/60 theorem for bin packing," *J. Complexity*, vol. 1, no. 1, pp. 65–106, 1985.

[40] G. Dósa, "The Tight Bound of First Fit Decreasing Bin-Packing Algorithm Is FFD(I) ≤ 11/9OPT(I) + 6/9," in *Combinatorics, Algorithms, Probabilistic and Experimental Methodologies*, B. Chen, M. Paterson, and G. Zhang, Eds. Springer Berlin Heidelberg, 2007, pp. 1–11.

[41] Erik Kain, "The Top Ten Best-Selling Video Games Of 2014," *Forbes*. [Online]. Available: http://www.forbes.com/sites/erikkain/2015/01/19/the-top-ten-best-selling-video-games-of-2014/#5e73c7aa6b9d. [Accessed: 08-Feb-2016].

[42] C. Huang, C. Hsu, Y. Chang, and K. Chen, "GamingAnywhere: An Open Cloud Gaming System," in *Proceedings of the 4th ACM Multimedia Systems Conference*, New York, NY, USA, 2013, pp. 36–47.

[43] N. Singhal, I. Park, and S. Cho, "Implementation and optimization of image processing algorithms on handheld GPU," in *2010 17th IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 4481–4484.

[44] B. Dietrich, D. Goswami, S. Chakraborty, A. Guha, and M. Gries, "Time Series Characterization of Gaming Workload for Runtime Power Management," *IEEE Transactions on Computers*, vol. 64, no. 1, pp. 260–273, 2015.

[45] Y. Lee, K. Chen, H. Su, and C. Lei, "Are all games equally cloud-gaming-friendly? An electromyographic approach," in *2012 11th Annual Workshop on Network and Systems Support for Games (NetGames)*, 2012, pp. 1–6.

[46] S. Shi, C. Hsu, K. Nahrstedt, and R. Campbell, "Using Graphics Rendering Contexts to Enhance the Real-time Video Coding for Mobile Cloud Gaming," in *Proceedings of the 19th ACM International Conference on Multimedia*, New York, NY, USA, 2011, pp. 103–112.

[47] R. Shea, J. Liu, E. C. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *IEEE Network*, vol. 27, no. 4, pp. 16–21, Jul. 2013.

[48] A. Ojala and P. Tyrvainen, "Developing Cloud Business Models: A Case Study on Cloud Gaming," *IEEE Software*, vol. 28, no. 4, pp. 42–47, Jul. 2011.

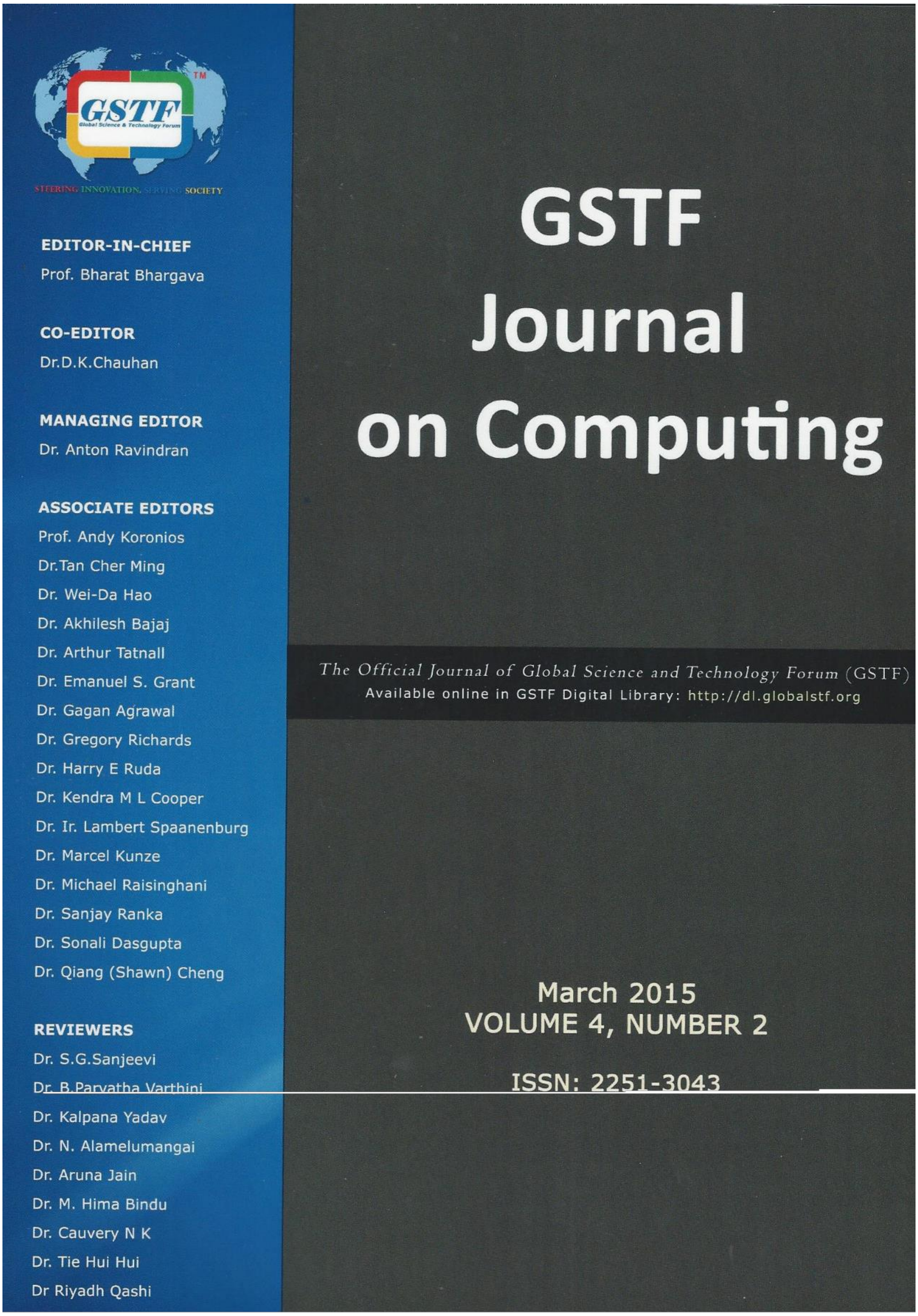
[49] R. Suselbeck, G. Schiele, and C. Becker, "Peer-to-peer support for low-latency Massively Multiplayer Online Games in the cloud," in *2009 8th Annual Workshop on Network and Systems Support for Games (NetGames)*, 2009, pp. 1–2.

[50] M. Manzano, J. Hernandez, M. Uruena, and E. Calle, “An empirical study of Cloud Gaming,” in *2012 11th Annual Workshop on Network and Systems Support for Games (NetGames)*, 2012, pp. 1–2.

[51] G. Olds, “Will Nvidia ‘n’ pals pwn future gaming?,” *The Register*. [Online]. Available: http://www.theregister.co.uk/2012/05/22/will_nvidia_tech_change_gaming/. [Accessed: 08-Feb-2016].

Appendix

Research Paper 1 :
Hybrid method for adaptive cloud gaming contents



GSTF
Global Science & Technology Forum
STEERING INNOVATION, SERVING SOCIETY

GSTF Journal on Computing

The Official Journal of Global Science and Technology Forum (GSTF)
Available online in GSTF Digital Library: http://dl.globalstf.org

March 2015
VOLUME 4, NUMBER 2

ISSN: 2251-3043

DOI 10.7603/s40601-014-0005-9
GSTF Journal on Computing (JoC) Vol.4 No.2, March 2015

Hybrid Method for Adaptive Cloud Gaming Contents

R. Jitpukdeebodintra and S. Witosurapot

Received 30 Dec 2014 Accepted 27 Jan 2015

***Abstract*—Methods for adapting cloud gaming content can be divided according to client- or server-returned benefits. In this paper, we argue that the better method should be performed in such a way that both client and server should share common benefits. By working on a cooperative scheme, the basic of trans- rendering and trans-scaling mechanism are proposed to work together for obtaining better utilizations of network bandwidth and graphic processor altogether. Based on our experiments of running cloud gaming services in simulated environment, we show many evidences that significantly improved benefits can be obtained as planned. Our proposed hybrid adaptation method is therefore clearly helpful for the game research community.**

Keywords-component; cloud gaming; content adaptation; trans-rendering; trans-scaling

I. INTRODUCTION

Cloud gaming services are particularly designed to support for playing highly sophisticated games on low-end computers and hand-held devices. This can be made possible through the exploitation of cloud computing model, where the cloud server play a key role of storing, executing the game contents, and delivering the game outputs in a form of video streaming to the clients so that substantial workloads on graphics execution and GPU (Graphics Processing Unit) power can be lessened at the end devices. By working in this manner, the content adaptation technique [1] for matching the video streaming to the target resolution of the end device is extremely important. Several solutions for content adaptation in cloud gaming services have been largely exercised in the literature on either aspects of server or client benefits. While the prime objective of server benefits is on decreasing the computing workload at server, that of client benefits is alternatively on reducing the network requirement. As a result, these solutions may not be necessarily efficient as claimed.

In this paper, we argue that an effective solution for content adaptation technique in cloud gaming service should be worked on mutual benefits of server and client machines in a cooperative manner. In the server-beneficial approach, adaptively varying the graphic resolutions in trans-rendering engine from the sole consideration of server workloads may deteriorate the graphic quality of the client devices. Alternatively, in the client-beneficial approach, updating the video codecs of graphic display under the limited view of client machine on the transient network delay can increase the server's workload unnecessarily. Therefore, by realizing a complete view of mutual benefits on both client and server, we believe that a better utilization of computing and network resources can be achieved.

This paper will be structured as follows. In section 2, some background of content adaptation mechanisms in cloud gaming services will be provided, following with our newly proposed mechanism for mutual resource utilization in section 3. In section 4 and 5, details of our tests for performance evaluation and results will be described respectively. In section 6, we conclude the paper.

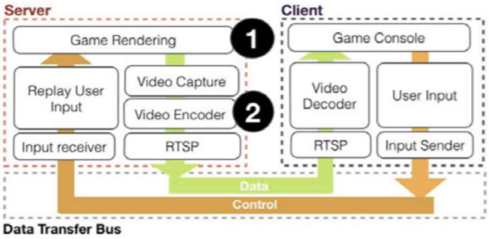


Figure 1. Cloud gaming service general architecture

II. BACKGROUND AND RELATED WORK

In cloud gaming service as show in Fig. 1, it is the function of server to compute all current workloads of game running at the client with the highest graphic resolution, and therefore the high resolution video streaming will be usually sent to the client. Unfortunately, this may cause two key problems, i.e. the insufficient GPU resources at the server [1] and the high bandwidth consumption at the client [2]. To response to these problems, many adaptation mechanisms have been proposed in a number of works, but can be classified into two groups based on the server or client viewpoint of returned benefits.

From the viewpoint of sever-benefits, the objective of adaptive mechanism at the server will be aimed at decreasing workload, in the "game rendering" (Number 1) on graphic processing at the server machine. This technique is widely known as "Trans-rendering" process, which can be done by lowering graphic resolutions or removing some texture details in video games. An example of the research work for cloud gaming services can be seen in [4], where computing workloads can be managed to reduce by adjusting 5 gaming render factors, which are Realistic effect, View distance, Texture detail, Environment detail and Rendering Frame Rate.

From the viewpoint of client-benefits, the focus of adaptive mechanism at the server is on decreasing the network usage through various resolutions of output video stream (Number 2). This technique is usually known as "Trans-scaling". An example of the research work for cloud gaming services can be seen in [5], where the H.264 video encoding parameter was adjusted to match for the client device capability. More examples [6-8] can be found in the field of video-on-demand.

DOI: 10.5176/2251-3043_4.2.318

Published online: 21 August 2015

However, no matter what viewpoint of the returned benefit is, either of them has its own weakness as in follow:

A. *Weakness of Content Adaptation for Server-benefits*

While the content adaptation for server-benefits can solve an insufficiency of server's GPU resources by decreasing the computing workload, it can worsen to the quality of output video stream, such as blurring the screen, and pixelate the texture detail on the client machine, in such a way it cannot be bearable for showing on the high-resolution display [4].

B. *Weakness of Content Adaptation for Client-benefits*

While the content adaptation for client-benefits can reduce the network consumption requiring for sending video stream to the client machine, it demands a great burden of computing tasks at server. This can cause the degradation of service quality especially during the prime time when the server becomes congested with a number of clients [5]. Some of evidence can be seen from our recent experiment. In Fig. 2, it can be seen clearly that the more users are, the less capability of trans-rendering engine can be produced the acceptable frame rate (i.e. less than 30 frames per second).

Therefore, based on the above-mentioned weaknesses, we argue that an effective method for content adaptation in cloud gaming service should span across the benefits of both client and server. We will refer to this method as "Hybrid method" hereafter.

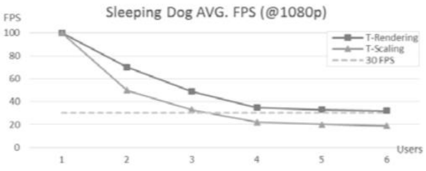


Figure 2. Average Frame per second for 1080p cloud gaming

III. Proposed Mechanism

The basic workflow of Hybrid method as shown in Fig. 3. It begins with the server monitoring process, where the current status of server's GPU is observed and a decision for next sequential process according to the GPU status. If the GPU is in the non-overloaded status, the adaptation process will be directly executed. In contrast, if the GPU becomes overloaded, the newly proposed Mutual-benefits algorithm will be performed so that the balance point of the benefits between the client and the server can be computed.

A. *Adaptation Process*

This process is basically the combined process of trans-rendering and trans-scaling methods, where the ultimate goal is aimed at compromising the resulted benefits for both client and server. Therefore, the Trans-rendering and Trans-scaling will be executed accordingly. However, to understand the process in an easily manner, the illustration of adapted game scene from 1080p to 640p will be used as an example. In Fig. 4(a), the output obtained from the pure trans-rendering method will be lower the picture quality in order to maintain the picture size. This is in contrast to the pure trans-scaling in Fig. 4(b), where the picture size is not important, but the resolution will be the prime concern and hence the high GPU load is also demanding accordingly. Then, it becomes clear that our combined method of trans-rendering and trans-scaling can produce the higher quality of picture than the pure trans-rendering one, but lower demand of GPU load as seen in Fig.4(c).

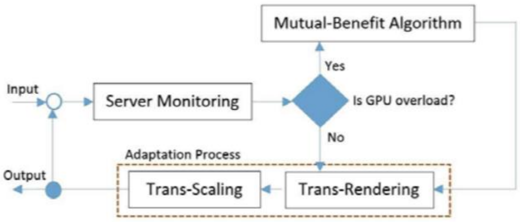


Figure 3. The workflow of Hybrid method

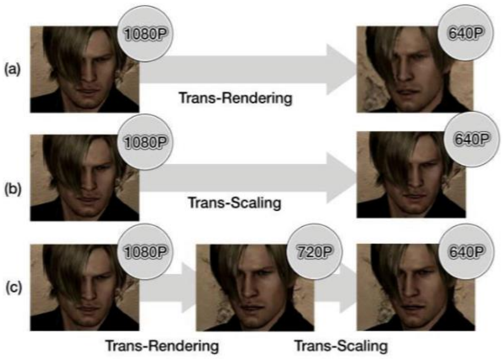


Figure 4. Comparison of a) pure trans-rendering, b) pure trans-scaling, and c) Combined trans-rendering and trans-scaling

B. *Mutual Benefits Algorithm*

The objective of Mutual-benefits algorithm is to find the balance of relation between server-benefits (GPU usage) and client-benefits (Render resolution), which are the linear relation due to many works state on [9-14]. In this regard, we can plot a linear relationship between game workload and resolution as a linear graph shown in Fig. 5, which is expressed as a function in (1). Therefore, we can always find the suitable resolution of cloud gaming, regarding the current GPU workload and the client screen size.

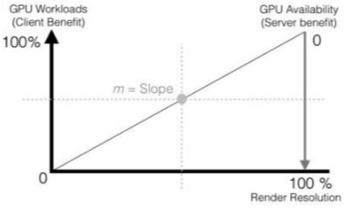


Figure 5. Relational between GPU Workload and Rendered resolution

$$x = \left(\frac{y}{m}\right) * ptarget \qquad (1)$$

Where y is percentage of server's available GPU resource (e.g. 80%), m is slope of the relation between GPU workload and render resolution which depends on each game, and *ptarget* is pixel number of client display (e.g. 921,600 pixel for 1280x720 display). We can find the x from (1), that will be the compartments point for the adaptation process we state in section III a.

IV. Experimental Method

To illustrate performance of hybrid method, we setup testing equipment as show in Fig. 6. On network infrastructure, we use Cisco 3845 router and Cisco 2960G gigabit switch to create gigabit connection between end-points.

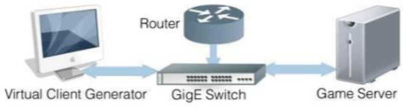


Figure 6. Network Scenario

On cloud gaming server, we use a server contains of 3.5 GHz 6-Core Xeon E5-1650v2 processor, 16GB DDR3 main memory and Dual NVIDIA GeForce GTX Titan GPU. This server will run 5 chosen sampling games which represent the most popular game genres in nowadays, including "Race driver Grid 2" for racing genre, "Metro 2033" for the first person shooter genre, "Resident Evil 6" for the third person shooter genre, "The elder scrolls V: Skyrim" for role-playing and "Grand Theft Auto IV" for open-world genre. We will give detail about game setting in table I.

TABLE I. EXPERIMENTAL SAMPLE GAME

Games	Setup
Resident Evil 6	VSYNC On, FXAA3HQ, Motion Blur ON, Shadow Quality High, Texture Quality High, Screen Quality High
The Elder Scrolls V: Skyrim	Ultra Quality, AA OFF
Grand Theft Auto IV	Texture Quality: High, Render Quality: High, View Distance 33, Detail Distance 33, Vehicle Density 26, Shadow Density 3
Race Driver Grid 2	Ultra Quality, AA OFF, Advance Lighting
Metro 2033	DX11 Very High Quality + AA OFF

On cloud gaming client, for convenience to collect results from sampling, we use a high-end desktop PC to emulate a sampling device rather than a real hardware. We generate 5 sampling groups. Each group consists of 10 emulated clients, which estimates from 4 global statistic on internet connected devices (Stream online game store, Statista.com, Apprepim.com and Statcounter.com). As we show in table II.

TABLE II. CLIENT SAMPLING

Tier	Sample	Steam	Web PC	Web PC+M	Mobile Sale	Avg
1	1920x1080	4	1	1	1	2
2	1600x900	2	1	1	0	1
3	1280x720	4	7	5	1	4
4	800x600	0	1	2	4	2
5	320x240	0	0	1	4	1
Total		**10**	**10**	**10**	**10**	**10**

Note: "Steam" group come from number of steam user, "Web PC" and "Web PC+M" come from client who surf the web on PC and on PC combine with mobile device respectively, "Mobile Sale" come from mobile devices sale on 2014 and "Avg" is an average.

We use these above emulated clients to send command to sampling server. Then we pull out working log from both sides and use it to explain the efficiency of hybrid method hereafter.

V. EXPERIMENTAL RESULTS

To explain the efficiency of hybrid method, we will show an experimental result by two aspects, client and server benefit aspects.

A. Client Benefits

For obtaining client-benefits, cloud gaming needs to render game graphics at the compatible resolution for displaying on the client machine, and at the same time, responsiveness and smoothness of the game session must be maintained. It also needs to limit network usage.

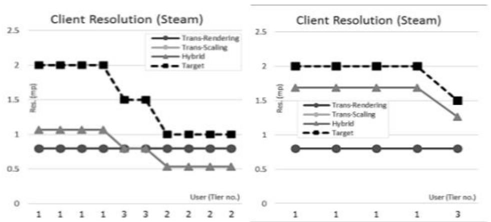


Figure 7. Render resolution for each client. Sample by "Steam" client scenario; congested (Left) and uncongested (Right)

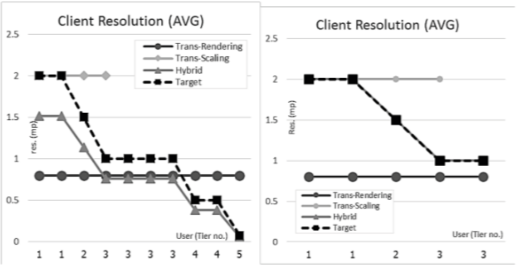


Figure 8. Render resolution for each client. Sample by "Average" client scenario; congested (Left) and uncongested (Right)

Fig. 7 and Fig. 8 shows a rendered resolution for each user on both scenarios of uncongested and congested loads. During uncongested periods, the hybrid method will render every pixels requested by the client. During congestion, this method will instead render at near client resolution, rather than those resolution obtained from trans-rendering and trans-scaling methods. For enabling the precise analyst, we use "Root Mean Square Error" (RMSE) to determine the dependence between content adaptation method and client display resolution. The results are shown in table III, which can be noticed that the hybrid method has the lowest RMSE. This means the output of hybrid method have more congruence to device resolution than the other methods.

TABLE III. RELATIVE BETWEEN CONTENT ADAPTATION METHOD AND TARGET RESOLUTION BY ROOT MEAN SQUARE ERROR.

Method	RMSE (M. Pixel)
Trans-Rendering	0.597
Trans-Scaling	0.794
Hybrid	0.405

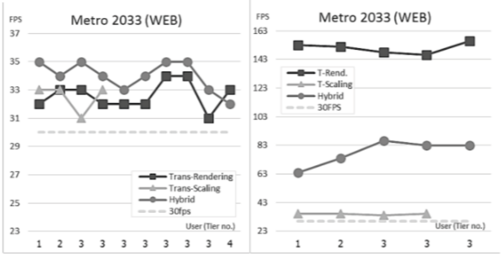


Figure 9. Average FPS for each client. Sample by "Web" client scenario. congest (Left) and uncongested (Right)

Beside the render resolution, the clients yet require the other service quality, e.g. smoothness and responsiveness of the game session, which have a close relationship with the values of frame-per-second (FPS) and the latency. Fig. 9 reveals number of an average FPS, which shows the hybrid method's ability to comfortably deliver more than 30 FPS in congest and more than 60 FPS in uncongested situation. That means the clients will always get reasonable smooth game experience on both situations.

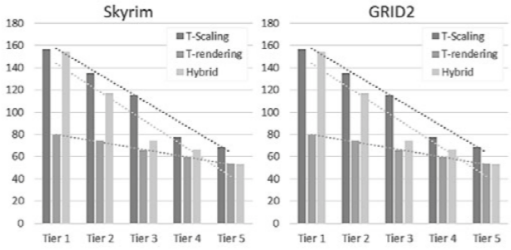


Figure 10. Game latency for each type of user. (in millisecond)

Fig. 10 shows a number of game latency, which explains that an implementation of hybrid method does not cause significant latency compared to the other method. Most importantly, it can keep the game latency below the acceptable region of 200ms, which surely has no effect on human visual perception [19]. It can be noticed in Fig. 11 that this feature can be obtained on the network usage that is competent to the other methods.

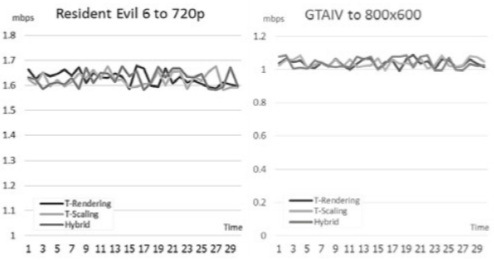


Figure 11. Network Usage

B. *Server Benefits*

For obtaining server-benefits, the game server will attempt to increase their user capability, lower the GPU usage, and maintain the network bandwidth, when the number of users are increased.

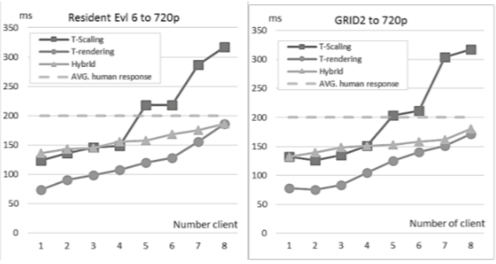


Figure 12. Game latency by the number of users.

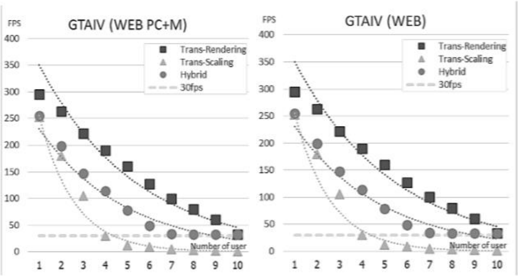


Figure 13. Average FPS by the number of users

Because of the server capability depends on the number of client FPS and latency, it is showed in Fig. 12 and Fig. 13 that hybrid method can deliver the comparative number of clients as the trans-rendering method, but significantly better than those of trans-scaling method. In addition, when looking at the GPU utilization in Fig. 14, the hybrid method are yet slowly increasing, compared with the trans-scaling method.

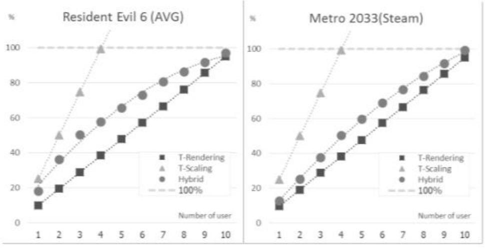


Figure 14. GPU Workload by the number of users.

C. *Discussion*

Based on all experiments explained above, we can give a complete picture of performance issues in Fig. 15. It can be clearly seen that the hybrid method are outstanding on many issues better than other two counterparts, except for the average game FPS and the game latency.

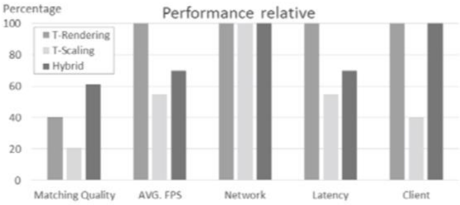


Figure 15. Experiment summary.

VI. CONCLUSION

In this paper, we introduce the combination of trans-rendering and trans-scaling methods aiming at adaptation for cloud gaming content and taking both server-and client-benefits into consideration. By means of experiments, we showed evidences that our proposed method are outstanding in terms of picture quality and associated GPU workloads, in comparison with trans-rendering and trans-scaling methods performing alone. We believe the content of this paper will give substantial benefits for researchers, who also work in the field of cloud gaming.

REFERENCES

[1] P. Eisert and P. Fechteler, "Remote Rendering of Computer Games"

[2] Kuan-Ta Chen Chun-Ying Huang and Cheng-Hsin Hsu, "Cloud Gaming Onwards: Research Opportunities and Outlook"

[3] D.De Winter P.Simoens L.Deboosere F.De Turck J.Moreau, B.Dhoedt and P.Demeester, "A Hybrid Thin-Client protocol for Multimedia Streaming and Interactive Gaming Applications" , NOSSDAV '06, 2006

[4] Shaoxuan Wang and Sujit Dey, "Rendering Adaptation to Address Communication and Computation Constraints in Cloud Mobile Gaming", IEEE Globecom 2010, 2010

[5] A.Jurgelionis F.Bellotti A.De Gloria P.Eisert J.-P. Laulajainen and A.Shani, "Distributed video game streaming system for pervasive gaming", 2008

[6] J.Brandt and L.Wolf, "Adaptive video streaming for mobile clients", NOSSDAV '08, 2008

[7] N.Cranley P.Perry and L.Murphy,"Dynamic content-based adaptation of streamed multimedia", Journal of Network and Computer Applications 30 (2007), 2007

[8] M.Eberhard L.Celetto C.Timmerer E.Quacchio H.Hellwagner and F.Rovati, "An interoperable multimedia delivery framework for scalable video coding based on MPEG-21 Digital Item Adaptation", IEEE International Conference on Multimedia and Expo 2008, 2008

[9] M Jarschel D.Schlosser S.Scheuring and T.Hossfeld, "An evaluation of QoE in cloud gaming based on subjective tests", Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on, 2011

[10] K.Chen Y.Chang P.Tseng C.Huang3 and C.Lei "Measuring The Latency of Cloud Gaming Systems", MM '11 Proceedings of the 19th ACM international conference on Multimedia, 2011 ,pp. 1269-1272

[11] S.Choy B.Wong G.Simon and C.Rosenberg, "The brewing storm in cloud gaming: a measurement study on cloud to end-user latency", Proceedings of the 11th Annual Workshop on Network and Systems Support for Games (NetGames '12), 2012

[12] C.Huang C.Hsu Y.Chang, and K.Chen, "GamingAnywhere: an open cloud gaming system", Proceedings of the 4th ACM Multimedia Systems Conference (MMSys '13), 2013

[13] Nitin Singhal, In Kyu Park and Sungdae Cho, "Implementation and optimisation of image algorithm processing algorithm on handheld GPU", Proceedings of 2010 IEEE 17th International Conference on Image Processing, 2010

[14] Dietrich B. Goswami, D. Chakraborty, S. Guha, A. and Gries, M, "Time Series Characterization of Gaming Workload for Runtime Power Management", IEEE Transactions on Computer,(Volume:PP Issue: 99), 2013

[15] Y.Lee K.Chen H.Su C.Lei, "Are all games equally cloud-gaming-friendly? An electromyographic approach," Network and Systems Support for Games (NetGames), 2012

[16] H.Shi C.Hsu K.Nahrstedt and R.Campbell, "Using graphics rendering contexts to enhance the real-time video coding for mobile cloud gaming", Proceedings of the 19th ACM international conference on Multimedia (MM '11), 2011

[17] R.Shea J.Liu H.Ngai and Y.Cui, "Cloud gaming: architecture and performance," Network, IEEE vol.27, no.4, pp.16,21,2013

[18] A.Ojala and P.Tyrväinen, "Developing cloud business models: A case study on cloud gaming", IEEE Software, 2011

[19] R.Suselbeck G.Schiele and C.Becker, "Peer-to-peer support for low-latency Massively Multiplayer Online Games in the cloud," Network and Systems Support for Games (NetGames), 2009 8th Annual Workshop on, 2009

[20] M.Manzano J.Hernández M.Uruenña, and E.Calle, "An empirical study of cloud gaming", Proceedings of the 11th Annual Workshop on Network and Systems Support for Games (NetGames '12), 2012

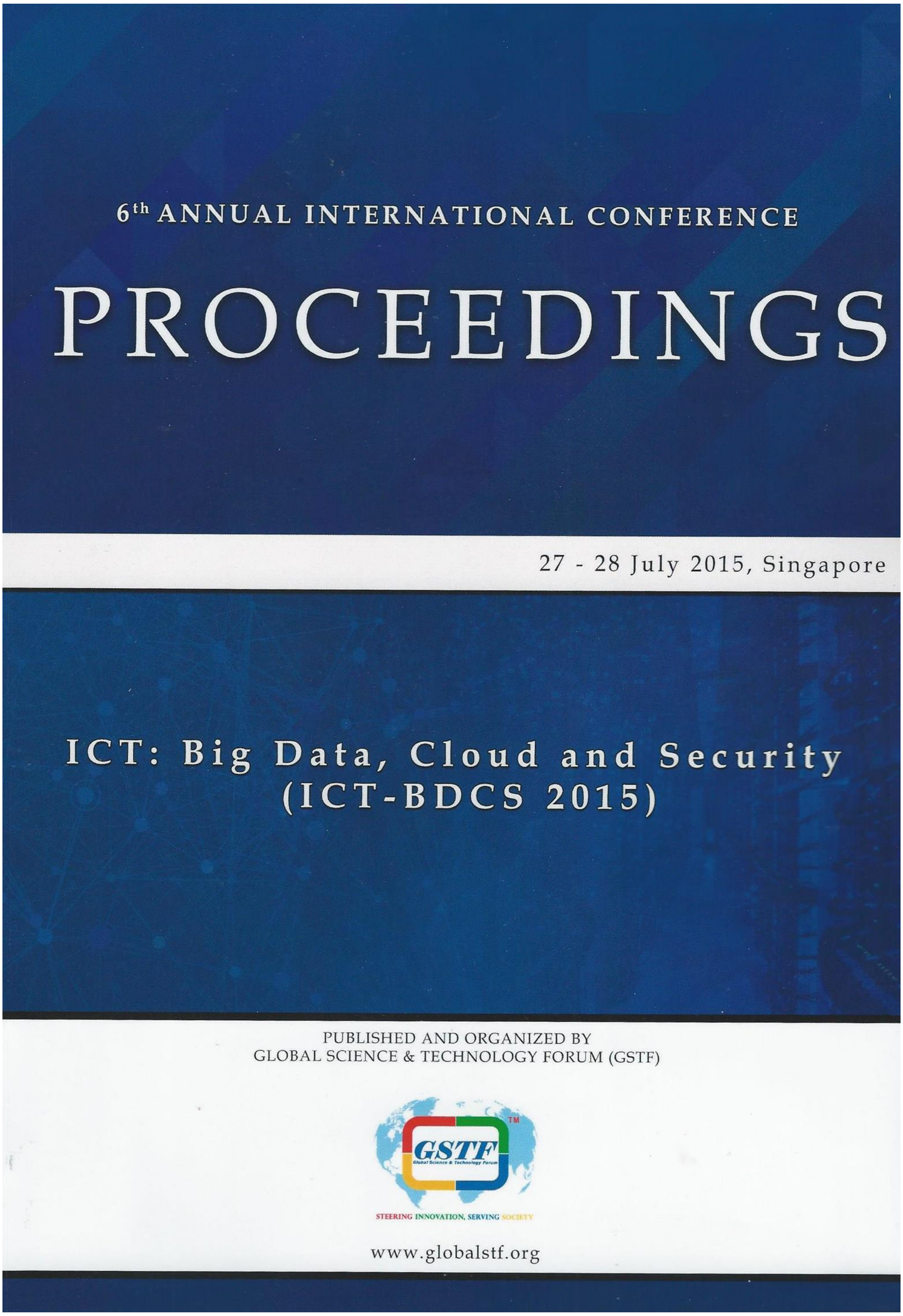
[21] D. Olds, "Will Nvidia 'n' pals pwn future gaming?", The Register, [Online] 2012, http://www.theregister.co.uk/2012/05/22/will_nvidia tech_change_gaming/ , (Accessed: 9 September 2014)

AUTHORS' PROFILE

Ritthichai Jitpukdeebodintra is currently a doctorate candidate in the field of computer engineering at faculty of engineering, Prince of Songkhla University. His research interests include technology in computer games, computer graphics, and cloud computing.

Dr. Suntorn Witosurapot is an Assistant Professor in department of Computer Engineering, Faculty of Engineering, in Prince of Songkla University (PSU), Thailand. He received the bachelor and Master degrees in Electrical Engineering from PSU, Thailand and Ph.D. degree from Swinburne University of Technology, Australia, with the thesis topics related to resolving network resource competition in the Internet. His research interests include Web engineering and applications, semantic Web, and management of information technology. Currently, most of his research work revolves around information engineering in smart home network, smart grid infrastructure, and active games for people with visual disabilities.

Research Paper 2 :
A study on the impact of client display resolutions in cloud gaming workloads



6th ANNUAL INTERNATIONAL CONFERENCE

PROCEEDINGS

27 - 28 July 2015, Singapore

ICT: Big Data, Cloud and Security
(ICT-BDCS 2015)

PUBLISHED AND ORGANIZED BY
GLOBAL SCIENCE & TECHNOLOGY FORUM (GSTF)

GSTF™
Global Science & Technology Forum

STEERING INNOVATION, SERVING SOCIETY

www.globalstf.org

A study on the impact of client display resolutions in cloud gaming workloads

Ritthichai Jitpukdeebodintra
Department of Computer Engineering
Faculty of Engineering, Prince of Songkla University
Hatyai, Songkhla, Thailand
rittichai@capsuledna.com

Suntorn Witosurapot
Department of Computer Engineering
Faculty of Engineering, Prince of Songkla University
Hatyai, Songkhla, Thailand
wsuntorn@coe.psu.ac.th

***Abstract*—Cloud gaming workload characteristics have been widely studied in the past, but they do not taken the factor of display resolution at the client machines into account. As a result, the management of computer and network resources at the cloud gaming server cannot be efficient in realistic situations, where devices with heterogeneous display capabilities are found, and contents can be varied according to targeted resolution. In this paper, we present our empirical study on the effect of different client resolutions and the server workloads. In addition, we give a coarse approximation of their relationship with a linear fashion so that a prediction of resources in the cloud gaming environment can be done in the easily manner. This will give utterly benefits to service providers for suitable and effective cloud gaming resource management.**

Keywords-component; cloud gaming workload; client display resolution; content adaptation; linear approximation

I. Introduction

Cloud gaming [1] can be considered as on-demand gaming as a Service (GaaS) that empowers low-end computers to run graphics-intensive computer games without problems. This is due to the great burden of computing workload and resource consumption is actually at the cloud server. Then, it becomes crucial that cloud server must provide sufficient service quality to all connected clients. In fact, if the number of connections becomes excessive or demanded resources cannot be met at the Cloud gaming server, the performance degradation will be definitely resulted at the clients. Therefore, a better plan of resource utilization management at the cloud server is required. This, in turn, demands the good knowledge of cloud gaming workload characteristics so that resource requirement can be approximated and sufficient resource reservation can be done beforehand.

While the characteristics of cloud gaming workload has been widely researched in past years (such as [1-5]), it is rather limited to the server-specific or server-related parameters, such as the latency of cloud gaming server that caused from different locations [1, 3-5] and the impact of Graphics Processor Unit (GPU)'s scheduling algorithms under various workloads on resource utilization at the server machine [2]. In this paper, we argue that more suitable characteristics of cloud gaming workloads should be included the factor of display resolutions at the client machines into consideration. Since varying display resolutions of adapted contents ranging from high-definition to low-fidelity display will have a substantially effect to the volume of data and the burden of associated computing [6]. The contributions of this paper are as follows: First, we provide evidence (via our empirical study) to confirm that the factor of display resolution must be concerned and taken into account the Cloud gaming workload characteristics. Second, we suggest that a linear equation can be used to coarsely approximate the relation of client-resolution and cloud gaming workload.

The remainder of the paper is organized as follows. In section 2, some background of cloud gaming workload will be given, following with our study methodology in section 3. In section 4 and 5, our performance studies are detailed and the evaluation results are described respectively. In section 6, we conclude the paper.

II. Background And Related Work

In this section, we give the background of cloud gaming workload characteristics and the reasons why it should be extended to include various display resolution into account.

A. *Cloud gaming architecure*

Unlike traditional network game playing that all game workloads are handled at the client machine, Cloud gaming service will take care of the game workload computation at the cloud gaming server, and then output the result back to the client as a form of "High-definition video streaming" for each gaming session. By working in this manner, serious tasks for game graphic rendering can be avoided at the client machine. That is why low-end computers can be smoothly run graphics-intensive computer games without problems.

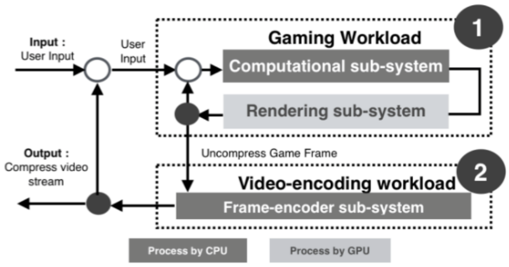


Figure 1. Simplified architecture of cloud gaming server.

As shown in Fig. 1, the process of workload computation can be seen as an infinite loop on two different computing tasks for “Gaming workload” (Number 1) and “Video-encoding workload” (Number 2). In each loop of execution occurring when game application receives a user interaction command, all game-related issues, such as game intelligence, score calculation, collision detection, will be executed by the basic Central Processing Unit (CPU) in the “Computational sub-system”, however, the result of these issues will be processed to draw on many graphic frames by the Graphic Processing Unit (GPU) in the “Rendering sub-system”. Later, these frames will be managed to send to the client over the network. While many choices for video compression or even no compression can be applied, the decision is upon implementation. However, they will surely affect to the data size of resulted video frames, as seen in Table 1 as an example.

TABLE I. COMPARISION OF FRAME BANDWIDTH IN A CERTAIN GAME

Setup	**Bandwidth**
Uncompressed: (59.94 fps - 8-bit)	372.9 MB/sec
Compress with H.264	65.3 MB/sec

B. Effects of display resolution to Cloud gaming workloads

To the best of our knowledge, we found that client display resolution has been overlooked in many studies of cloud gaming workload characteristics. However, the following papers can be seen as evidences for support our idea that attempts to take client display resolution into account cloud gaming workload characteristics.

- For realizing the actual gaming workload, the work in [7] showed that game workload and frame complexity have indeed somewhat relationship, and can be formulated as a model for roughly prediction such as a case of Linear-In-Parameter (LIP) function in [8].

- For determining the actual video-ending workload, the work in [9] showed the relationship between encode video resolution and CPU usage, which can be also predicted by a proper mathematical approach such as that in [10].

III. EXPERIMENT METHODOLOGY

We setup our testing equipment as shown on Fig. 2, which we used Cisco 2960G switch and Cisco 3845 router to establish the gigabit-connection speed between the end-points.

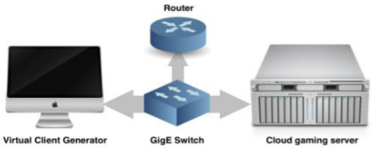


Figure 2. Equipment setup.

At cloud gaming server, we developed a sample cloud gaming application whose workflow is shown in Fig. 3. This application ran on server, which contained of 3.5 GHz 6-Core Xeon E5-1650v2 processor, 16GB DDR3 main memory and Dual NVIDIA GeForce GTX Titan GPU. It would run 15 sampling games: Race Driver GRiD2, Need for speed: Rivals, Crysis 3, Metro 2033, Battlefield 4, Watch_dog, Grand Theft Auto IV, Resident Evil 6, Resident Evil: Revelation, Tomb Raider, The Elder scroll: Skyrim, Batman Arklam Origin, Titanfall, Bioshock: Infinite and Dragonball: Xenoverse.

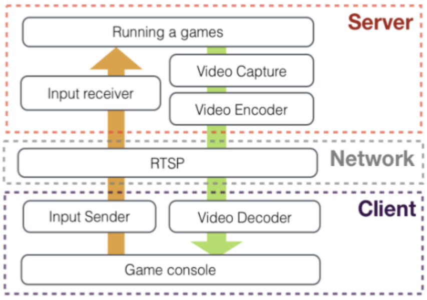


Figure 3. Our sample cloud gaming application workflow.

At the client side, we used a PC which contains of 4.0 GHz Quad-core Intel Core i7, 16GB DDR3 main memory, to generate a sample game request with each resolution to cloud gaming server. Then, we pulled out working log and used it to explain the characteristic of cloud gaming workload hereafter.

IV. EXPERIMENT RESULT

To clearly explain an effect from client display resolution to cloud gaming characteristic, we will show an experimental result by 4 aspects in this section, i.e. GPU Workloads, CPU Workload, Memory consumption and Network usage.

A. GPU Workloads

GPU takes responsible on game rendering sub-system. If, GPU resources demand cannot be met. A drawing of graphic frames will be unable to catch a user interaction input, which affected to responsiveness and smoothness of the game session. In Fig. 4, we show the GPU workload when all adapted game contents are adjusted by display resolutions.

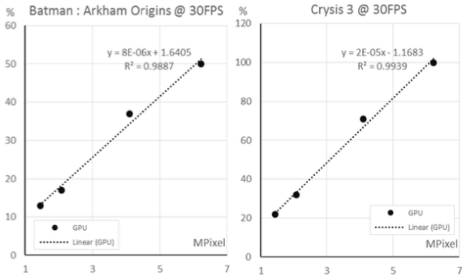


Figure 4. GPU workload

B. *CPU Workloads*

In cloud gaming services, CPU mostly take responsibility on computational & frame-encoding sub-system. Hence, the inanition of CPU workload will causes game delay. In Fig. 5, we shown the CPU workload when display resolutions changed.

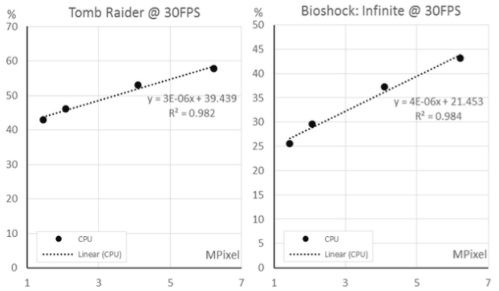


Figure 5. CPU workload

C. *Memory usage*

Since most high-definition games will not only cache binary-related data temporally on the system memory, but also the GPU memory, therefore, both kinds of memory need to be studied. Here, the consumption of system memory and GPU memory are shown in Fig. 6 and Fig. 7 respectively.

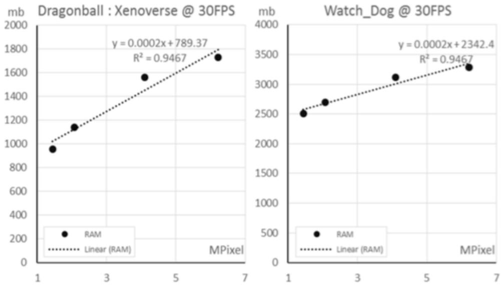


Figure 6. System Memory

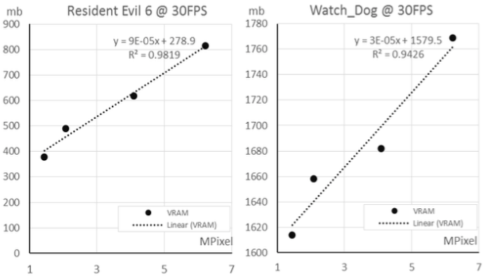


Figure 7. GPU Memory

D. *Network Consumption*

An insufficient of network bandwidth cause the delay of the game graphics transfer. In Fig. 8, we give a result of network consumption when client display resolution changed.

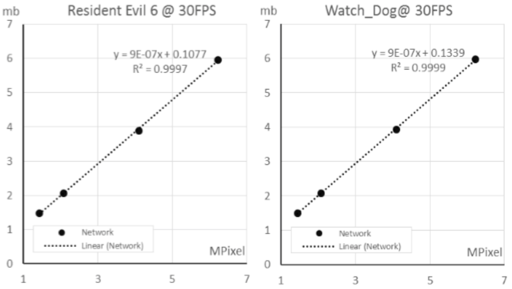


Figure 8. Network Consumption

V. Experimental Conclusion

From the empirical experiment in the previous section, we can conclude the results into twofold.

□ First, it can be seen clearly that client resolution is an important factor that crucially affects to cloud gaming workloads as well.

□ Second, the relationship of cloud gaming workload and client display resolution can be coarsely approximated with a linear function. This can be confirmed by our experiments with less than 13% of error on average.

VI. Conclusion

In this paper, we have described how various display resolutions of cloud gaming clients should be taken into consideration for realizing the realistic resource utilization of heterogeneous devices. Based on our empirical study, we provide evidence that client display resolution should be a prime concern on efficient workload approximation. We also suggest that the approximate resource utilization of each gaming session can be coarsely made through a linear equation, which is derived through our experiment results. Our empirical study looks promising and give benefits straightforwardly to the cloud gaming service industry.

References

[1] Choy, S.; Wong, B.; Simon, G.; Rosenberg, C., "The brewing storm in cloud gaming: A measure-ment study on cloud to end-user latency," Network and Systems Support for Games (NetGames), 2012 11th Annual Workshop on , vol., no., pp.1,6, 22-23 Nov. 2012

[2] Zhengwei Qi, Jianguo Yao, Chao Zhang, Miao Yu, Zhizhou Yang, and Haibing Guan, "VGRIS: Virtu-alized GPU Resource Isolation and Scheduling in Cloud Gaming", ACM Trans. Archit. Code Optim. 11, 2, Article 17 (July 2014)

[3] Wei Cai; Min Chen; Conghui Zhou; Leung, V.C.M.; Chan, H.C.B., "Resource management for cognitive cloud gaming," Communications (ICC), 2014 IEEE International Conference on , vol., no., pp.3456,3461, 10-14 June 2014

[4] V. Delwadia, S. Marshall, and I. Welch, "The effect of user interface delay in thin client mobile games. In Proceedings of the Eleventh Australasian Conference on User Interface - Volume 106 (AUIC '10), Christof Lutteroth and Paul Calder (Eds.)", Vol. 106. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 5-13.

[5] V. Clincy and B. Wilgor, "GSTF International Journal on Computing (JoC)", Vol.3 No.1, March 2013

[6] R. Jipukdeebodintra and S. Witosurapot, "Hybrid method for adaptive cloud gaming contents" "GSTF International Journal on Computing (JoC)", Vol.4 No.2, April 2015

[7] Pathania A., Jiao Q., A. Prakash, and Mitra T., "Integrated CPU-GPU Power Management for 3D Mobile Games. In Proceedings of the 51st Annual Design Automation Conference (DAC '14)". ACM, New York, NY, USA, , Article 40

[8] Dietrich, B.; Nunna, S.; Goswami, D.; Chakraborty, S.; Gries, M., "LMS-based low-complexity game workload prediction for DVFS," Computer Design (ICCD), 2010 IEEE International Conference on , vol., no., pp.417,424, 3-6 Oct. 2010

[9] Shafique, M.; Khan, M.U.K.; Henkel, J., "Power efficient and workload balanced tiling for parallelized high efficiency video coding," Image Processing (ICIP), 2014 IEEE International Conference on , vol., no., pp.1253,1257, 27-30 Oct. 2014

[10] Y. Huang, A. Tran, and Y. Wang, "A workload prediction model for decoding mpeg video and its ap-plication to workload-scalable transcoding. In Proceedings of the 15th international conference on Multimedia (MULTIMEDIA '07)". ACM, New York, NY, USA, 952-961.

[11] S. Wang and S. Dey, "Rendering Adaptation to Address Communication and Computation Constraints in Cloud Mobile Gaming", IEEE Globecom 2010, 2010

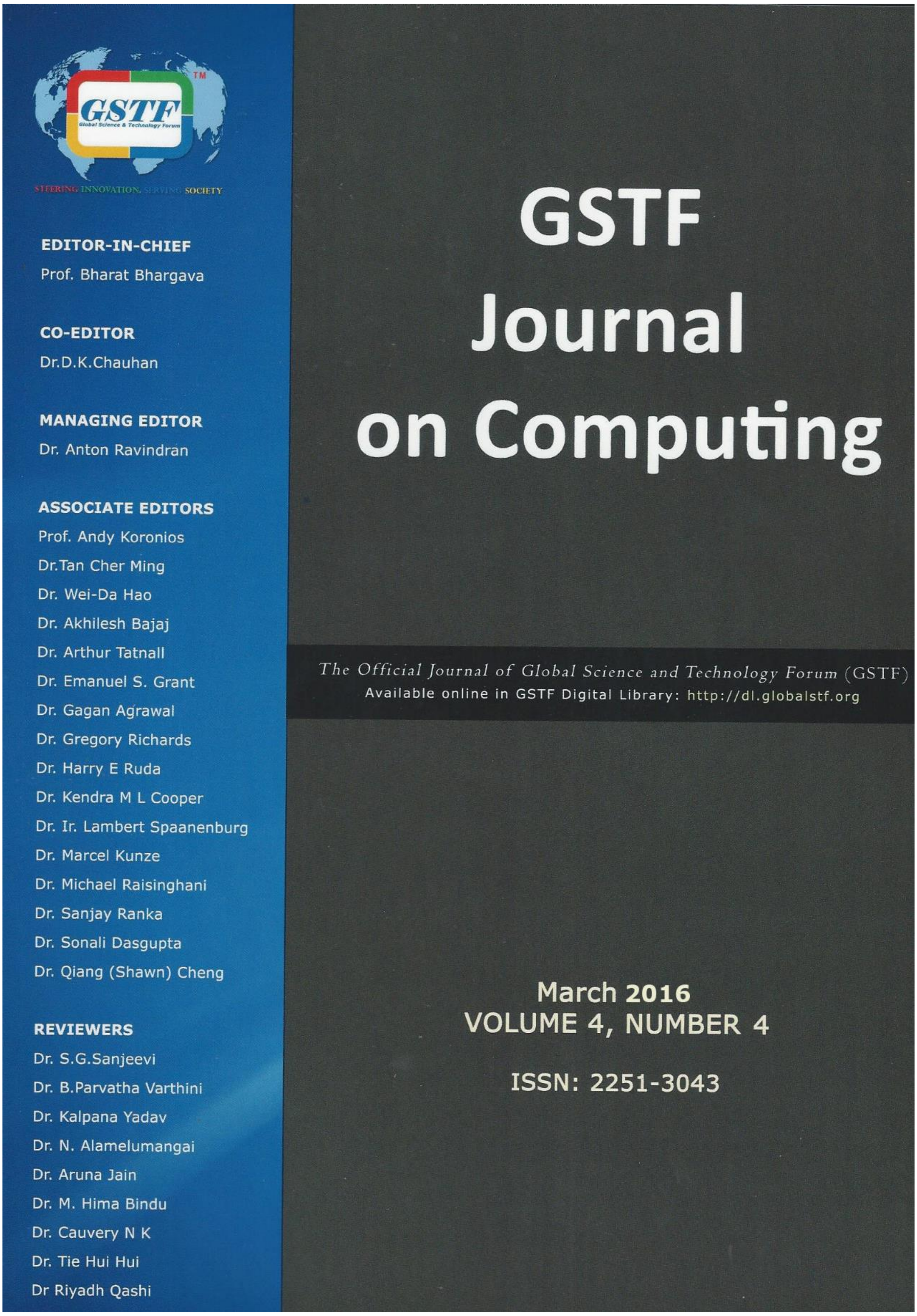
[12] A.Jurgelionis F.Bellotti A.De Gloria P.Eisert J.-P. Laulajainen and A.Shani, "Distributed video game streaming system for pervasive gaming", 2008

[13] J.Brandt and L.Wolf, "Adaptive video streaming for mobile clients", NOSSDAV '08, 2008

[14] N.Cranley P.Perry and L.Murphy,"Dynamic content-based adaptation of streamed multimedia", Journal of Network and Computer Applications 30 (2007), 2007

[15] M.Eberhard L.Celetto C.Timmerer E.Quacchio H.Hellwagner and F.Rovati, "An interoperable multi-media delivery framework for scalable video coding based on MPEG-21 Digital Item Adaptation", IEEE International Conference on Multimedia and Expo 2008, 2008

Research Paper 3 : Efficient Cloud Gaming Resource Provision Via Multi-dimensional Bin-Packing



STEERING INNOVATION, SERVING SOCIETY

GSTF Journal on Computing

The Official Journal of Global Science and Technology Forum (GSTF)
Available online in GSTF Digital Library: http://dl.globalstf.org

March 2016
VOLUME 4, NUMBER 4

ISSN: 2251-3043

Efficient Cloud Gaming Resource Provision Via Multi-dimensional Bin-Packing

R. Jitpukdeebodintra and S. Witosurapot

Abstract— **In order to enable the acceptable level of service quality for cloud gaming services, sufficient resources should be always maintained and optimal resource management is then necessary. However, taking only a limited set of server-related parameters, but ignoring the client-related parameters, in the typical formulation of optimization problem cannot yield for optimal resource allocation in the cloud gaming environment nowadays, where the game server's computing processors can be driven by both the CPU and GPU, and the client devices' display resolutions are largely heterogeneous. In this paper, we describe how better cloud gaming resource utilization can be achieved through a formulation of Multi-dimensional bin-packing optimization problem. Based on the experiment results, our proposed mechanism looks promising for realistic cloud gaming services, where the adaptive feature must be taken as a prime consideration for efficient cloud gaming resource management.**

Keywords-component; cloud gaming; resource allocation; optimization; bin-packing method

I. Introduction

Cloud gaming service [1] allows the graphic-intensive computer games to be displayed on any low computation capability devices, since it off-loads all computing burden from the clients to the cloud server and then sends the output back in a form of high-resolution video streaming. By working in this manner, the less demand on resource consumption for game computation can be expected at the client machine, and the more computing workload can be occurred on the server machine. Therefore, in order to avoid the overloaded condition of network and computing resources due to the excessive demand, the problem of resource utilization must be carefully investigated and efficiently solved on the server machine; otherwise the degraded service quality of all admitted connections will be resulted.

Technically, the problem of resource utilization on traditional cloud gaming server can be formulated as an optimization problem (such as [1-6]) so that many techniques for finding optimal solutions can be applicable. However, the bin-packing optimization of cloud gaming resource provision will be especially concerned, due to its attractive capability appeared in many works (such as [7-10]). Nevertheless, we argue that these works rather take a limited view of single resource utilization in their formulations. Indeed, there exist available resources, e.g. those of Central Processing Unit (CPU) and Graphics Processor Unit (GPU), which must be taken into account altogether. Since placing a burden on a cloud gaming resource on the server will surely affect the other resources in somewhat level [9, 10], poor resource management can potentially lead to the collapse of cloud gaming services. Therefore, we argue that dealing only with partial server resource utilization in an optimization problem will not yield for practical solutions in the realistic cloud gaming environments. In addition, we advocate on the inclusion of client-based parameter related to the device's display resolution into our Multi-dimensional bin-packing optimization formulation so that better resource utilization can be obtained, since this new parameter has been firmly proved as a factor having a significant influence on the server resource consumption in our recent study [10].

This paper is structured as follows. In section 2, we describe some background of optimization-based service provisioning in the cloud gaming domain. Then, we give a detail of our proposed Multi-dimensional bin-packing optimization formulation in section 3, following with the performance evaluation of the improved system via experiments and the discussion on results in section 4. Finally, we conclude the paper in section 5.

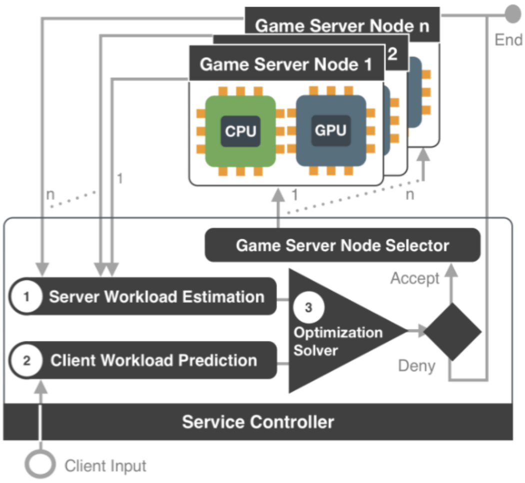


Figure 1. Conceptual diagram of the decision making process for admitting a client request

II. Problem Statement

In typical cloud gaming, the decision making process for call admission or rejection when a request of game client asks for a connection can be illustrated in Fig. 1. In this regard, the server will decide whether or not a client should be admitted by the result of comparison between the resource availability of current server workload)Number 1(and the predicted resource utilization of client connection)Number 2(, which is performed by the optimization solver)Number 3(. Here, the optimization problem will well-served for resolving the optimal selection of game server node, and informing the game server node selector accordingly. Hence, it is obvious that the resource optimization problem should be properly formulated and effectively solved.

While a number of studies have attempted to find an optimal cloud gaming server resource utilization by using different methods in the literature, they all share a common perspective of single resource optimization, such as using the scheduling approach to find an optimal GPU resource in [1-2], exercising mathematical heuristics to find an optimal CPU resource in [4-5], or exploiting bin-packing problems to find an optimal solution of GPU resource in]7[or CPU resource in [8] or Memory resource in [14]. More crucially, they are not efficient for implementing in the present cloud gaming technology, due to the following reasons:

□ Firstly, they do not take into account a key factor of the display resolution of heterogeneous client device, which is proved to be a key influence on different levels of cloud gaming resource consumptions in our recent study in [9].

□ Secondly, they have misassumption that the cloud gaming service can be simply classified as "CPU-based" or "GPU-based" games in the similar way as the traditional game playing [11-13]. Unfortunately, this is utterly difference for cloud gaming because there are many tasks of cloud gaming that utilize both GPU and CPU such as the video encoding task. As evidence, our experiment in]10[can be used to confirm the importance of co-existed CPU and GPU operations for maintaining the service quality in cloud gaming.

Hence, in order to obtain the efficient implementation, it is required that the more suitable optimization problem should bring many types of cloud gaming resources into consideration.

III. Bin-Packing Resource Optimization Problems

In this section, we describe two forms of bin-packing optimization problems that have a potential for solving the cloud gaming resources provision; single-dimensional and multi-dimensional bin-packing problem.

A. Single-dimensional Bin-packing Optimization Problem

The first form is called Single-dimensional Bin-Packing problem (SBP), which has a primary objective to minimize the resource waste of single-constraint bin, while the total resource consumption from object doesn't exceed bin capability.

In order to solve the provision problem of cloud gaming service, SBP may be declare as the formal statement in the following:

$$\text{Minimize:} \quad C - \sum_{i=1}^{n} c_i \qquad (1)$$

$$\text{Subject to:} \quad \sum_{i=1}^{n} c_i \leq C \qquad (2)$$

$$\forall_i \in \{1, \ldots, n\} \qquad (3)$$

where:

□ C is either GPU or CPU resources capacity of each server.

□ c_i is GPU or CPU usage of each game workload.

Noticed that the primary objective of SBP as showed in (1) concerns only one resource. Hence, it will be calculated twice for two concerned resources, e.g. the first consideration is for CPU resource and then the other is for GPU resource.

B. Multi-dimensional Bin-packing Optimization Problem

The second form is called Multi-dimensional Bin-Packing problem (MBP), which takes a similar objective as the SBP above, but here many constrained bins can be involved. For a case of CPU, GPU and network resource consideration, the general form of MBP can be given below:

$$\text{Minimize:} \quad C \cdot G \cdot W - \sum_{i=1}^{n} c_i \cdot g_i \cdot w_i \qquad (4)$$

$$\text{Subject to:} \quad \sum_{i=1}^{n} c_i \leq C, \sum_{i=1}^{n} g_i \leq G, \sum_{i=1}^{n} w_i \leq W \qquad (5)$$

$$\forall_i \in \{1, \ldots, n\} \qquad (6)$$

where:

□ *C, G, W* is the total resource of CPU, GPU and Network respectively.

□ c_i, g_i and w_i is the requested workload of CPU, GPU and Network, which can be estimated by means of a linear function that expresses the relationship between the client resolution and the cloud gaming workload [9].

Noticed that the primary objective of MBP as showed in in (4) aims to minimize the waste of 3 server resources including CPU, GPU and Network for the allocation of new resource requirement, under the constraint of total workload in (5).

In Fig. 2, the illustration aimed to explain the MBP process in the simple manner. In Fig. 2(a), the Workload 1 (Object 1) will be assigned to the Bin 1, due to the insufficient size of Bin 2. In contrast, the Workload 2 (Object 2) in Fig. 2(b) will be assigned to the Bin 2, since the lower waste of resource will be obtained.

In order to solve MBP, a number of possible algorithms (e.g. first-fit, best-fit or first-fit decreasing algorithm) can be possibly used. However, in this paper, the first-fit decreasing algorithm will be interested particularly, due to the dominant feature of fast computation and effectiveness in solving this sort of problem [14-16]. In essence, this algorithm will firstly sort objects by the decreasing order, then attempt to place each object into the first possible accommodate bin.

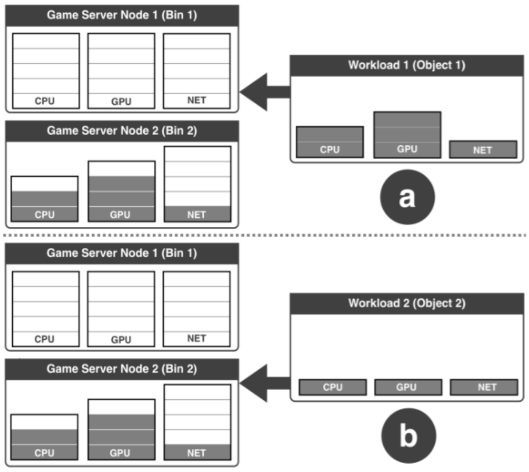

Game Server Node 1 (Bin 1)
CPU
GPU
NET
Game Server Node 2 (Bin 2)
CPU
GPU
NET
Workload 1 (Object 1)
CPU
GPU
NET
a
Game Server Node 1 (Bin 1)
CPU
GPU
NET
Game Server Node 2 (Bin 2)
CPU
GPU
NET
Workload 2 (Object 2)
CPU
GPU
NET
b

Figure 2. Illustration for explaining to equation (4)

IV. PERFORMANCE EVALUATION

A. *Experimental Setup*

The cloud gaming experimental infrastructure as shown in Fig. 3 will be supported throughout the experiments. In this infrastructure, a gigabit connection will be served as a backbone by the Cisco ISR 3845 as a router and Cisco catalyst 4500-E with Supervisor engine VI as a core switch so that all machines, including all 5 servers, a service controller, and a virtual client generator.

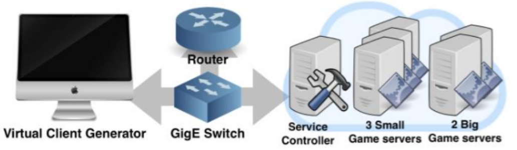


Figure 3. Experimental infrastructure

The group of game servers for performing game executions can be classified into two groups; Small server and Big server, for simulating the heterogeneous game server capabilities. The specification of the small servers consists of 3.4 GHz 4-core Core i5, 16GB DDR3 main memory and NVIDIA GeForce GTX 960 GPU, while that of big servers consists of 3.4 GHz 6-core Core i7, 16GB DDR3 main memory and NVIDIA GeForce GTX 970 GPU.

The service controller plays a crucial role of cloud gaming resource provision (referred to the functionality in Figure 1). The specification of service controller consists of 2.7 GHz 4-core Core i5 and 4GB DDR3 main memory.

For a virtual client generator, it will be used to emulate many client machines. The specification of service controller consists of 4.0 GHz Quad-core Intel Core i7, 16GB DDR3 main memory. Here, the 5 sets of client machines will be generated according to the scale-down ratio of Internet connected devices collected statistically from 4 different sites (Stream online game store, Statista.com, Apprepim.com and Statcounter.com). The goal is to maintain the similar ratio of heterogeneous machines found in those cloud gaming sites. A different set of client machines with various display resolutions can be found in Table 1.

TABLE I. EMULTAED CLIENT SETUP

Display Resolution	The number of client machines				
	Set 1	Set 2	Set 3	Set 4	Set 5
1080p	3	2	1	4	2
720p	2	3	1	2	4
640p	2	3	4	1	1
480p	1	0	2	1	1

B. *Testing Method*

We choose 9 best-selling games in 2014 (according to Forbes [17]), i.e. Call of Duty: Advanced Warfare, Madden NFL 15, Destiny, Grand Theft Auto 5, Minecraft, NBA 2K15, Watch Dogs, FIFA 15 and Call of Duty Ghosts, to provide cloud gaming service on our 5 game servers. In each set of client machines, each game will be requested in sequence and all service parameters (like Server usage, GPU waste, CPU waste and Network usage) on each game machine running different optimization problems, i.e. SBP for GPU, SBP for CPU and MBP, will be recorded accordingly.

C. *Experimental Results*

The 9 cloud gaming services have been tested, however only 3 different game characteristics will be selectively showed below, due to the page limit. Here are our choices; the Minecraft is represented for heavy-consumed CPU resource game, Destiny for heavy-consumed GPU resource game and Grand Theft Auto V for heavy-consumed CPU and GPU resource game. The resource consumptions of these games can be seen in Fig. 4

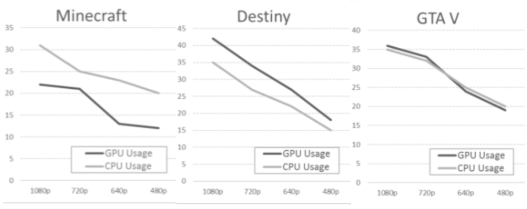


Figure 4. The resource consumption of Minecraft, Destiny and GTA V

Table II shows the number of game server machines that are required for running each game. As expected, the number of CPU and GPU resources will greatly depend on the game characteristics. Since the Minecraft is CPU-oriented game, it takes the total of 3 servers in the case of SBP-CPU method, which is higher than that of SBP-GPU method (i.e. 2 servers). However, it is not the same in the case of SBP-CPU and MBP, since it shares the same number of required resources. This is because both of CPU and GPU resources will be concerned together in the resource allocation problem solving by the MBP method.

TABLE II. OPTIMIZATION RESULTS: REQUIRED SERVERS FOR EACH GAME

Game	Formulation method	Number of game server require		
		Small game server	*Big game server*	*Total*
Minecraft	SBP-CPU	1	2	3
	SBP-GPU	1	1	2
	MBP	1	2	3
Destiny	SBP-CPU	2	0	2
	SBP-GPU	1	3	4
	MBP	2	0	2
Grand Theft Auto V	SBP-CPU	1	3	4
	SBP-GPU	2	1	3
	MBP	2	2	4

Fig. 5 shows the comparison results of resource utilization calculated by using the MBP, SBP-CPU and SBP-GPU methods. It is obvious that the given resources calculated by the MBP method will be sufficient in all cases. This is in contrast to the other methods, which can be the SBP-CPU or SBP-GPU method depending on the game type whether it heavily consumes on CPU or GPU resources.

For instance, the Minecraft demands the more resource of CPU than the GPU, the SBP-CPU method will be used to find the optimal CPU resource provision (depicted as the bar with diagonal lines), which will be later determined the number server machines and the volume of other resources (i.e. GPU and network) by looking up the values in Table III. However, the result of SBP-GPU method is also given in this case for the clear performance comparison of these 3 methods.

TABLE III. RESOURCES VOLUMES FOR EACH KIND OF GAME SERVER

Server Type	GPU	CPU	Network
Big game server	100	100	50
Small game server	40	60	50

In essence, by taking into consideration of all available resources in the bin-packing optimization problem like the

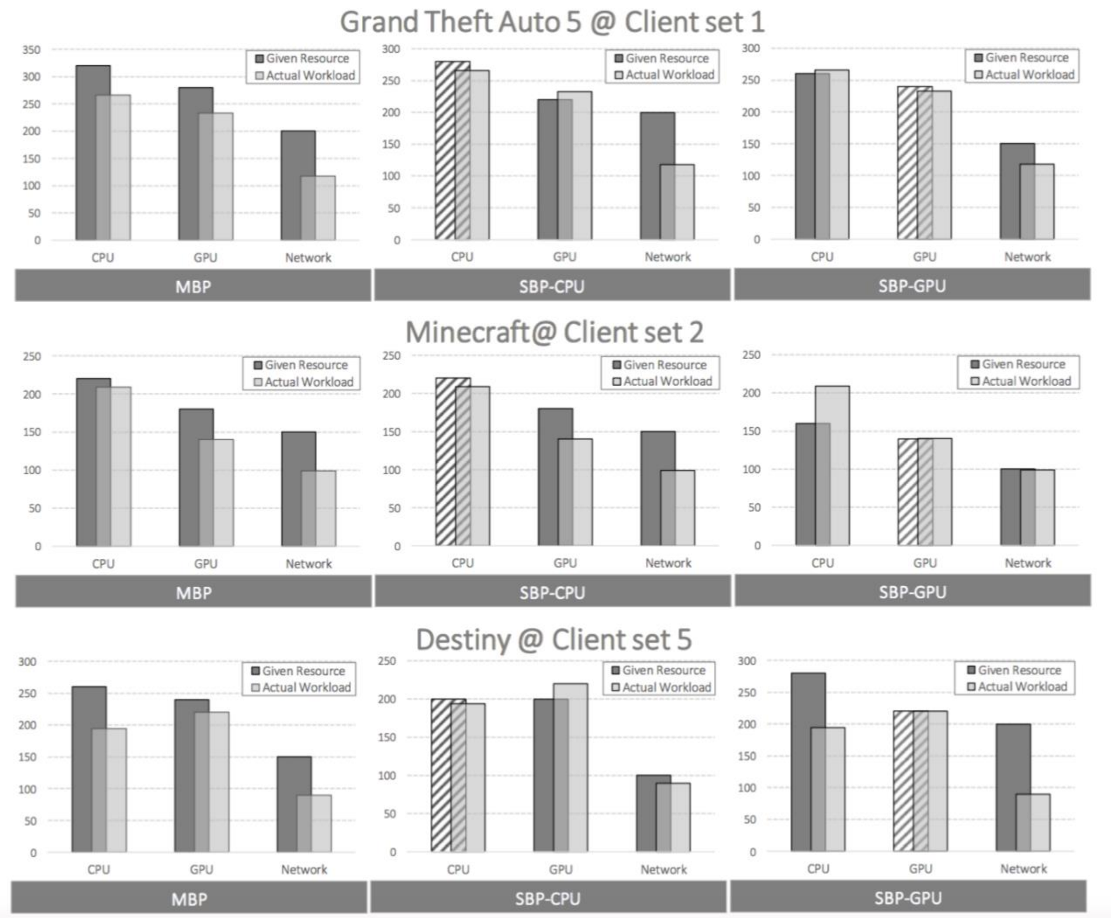


Figure 5. Experimental results

MBP method will yield a better and sufficient resources for all game types taken into the experiments. This can effectively avoid the quality degradation, such as low-frame rate or frame-skip, due to the insufficiency of provided resources in the cloud gaming server as showed in Fig. 6.

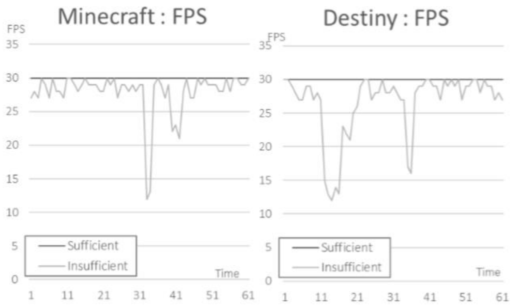


Figure 6. Effect of insufficient resource to game framerates

V. CONCLUSION

In this paper, we advocate on the use of Multi-dimensional Bin-Packing problem for determining the optimal resource provisioning in Cloud gaming servers, since a complete view of all resource availability will be taken into consideration for several advantages. This will yield a far more efficient resource utilization than the single-dimensional Bin-Packing problem as showed in our experiment results. As a result, the game service quality can be expected. Based on the evidence given in this paper, the MBP formulation method is extremely interesting and hence should be extensively used by cloud gaming service providers, or investigated further on improved performances by researchers in the cloud gaming community.

REFERENCE

[1] C. Zhang et al., "vGASA: Adaptive Scheduling Algorithm of Virtualized GPU Resource in Cloud Gaming", 2014.

[2] M. Yu et al., "VGRIS: Virtualized GPU Resource Isolation and Scheduling in Cloud Gaming", 2014.

[3] A. Khan et al., "Workload Characterization and Prediction in the Cloud : A Multiple Time Series Approach".

[4] G. Wei et al., "A game-theoretic method of fair resource allocation for cloud computing services", J Supercomput (2010) 54: 252–269, 2010.

[5] V.Vinothina et al., "A Survey on Resource Allocation Strategies in Cloud Computing", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No.6, 2012, 2012.

[6] S. Islam et al., "Empirical prediction models for adaptive resource provisioning in the cloud", Future Generation Computer System, 2012.

[7] Y. Li et all, "On Dynamic Bin Packing for Resource Allocation in the Cloud", SPAA'14, June 23–25, 2014.

[8] W. Song et all, "Adaptive Resource Provisioning for the Cloud Using Online Bin Packin". IEEE TRANSACTIONS ON COMPUTERS, 2013.

[9] R. Jipukdeebodintra and S. Witosurapot, "Hybrid method for adaptive cloud gaming contents", "GSTF International Journal on Computing (JoC)", Vol.4 No.2, April 2015.

[10] R. Jipukdeebodintra and S. Witosurapot, "A study on the impact of client display resolutions in cloud gaming workloads", Proceeding of 6th Annual International Conference on ICT-BDCS 2015, 2015.

[11] "Can you, list which games are "heavily CPU-based", "heavily GPU-based", or both?". [Online]. From: https://pcpartpicker.com/forums/topic/87771-can-you-list-which-games-are-heavily-cpu-based-heavily-gpu-based-or-both. [Acessed on 26 September 2015].

[12] "The game is CPU-bound, not GPU-bound" [Online]. From: https://steamcommunity.com/app/239140/discussions/0/604941528466981109/. [Acessed on 26 September 2015].

[13] S. Peak, "Quad-Core Gaming Roundup: How Much CPU Do You Really Need?". [Online]. From: http://www.pcper.com/reviews/Systems/Quad-Core-Gaming-Roundup-How-Much-CPU-Do-You-Really-Need. [Acessed on 26 September 2015].

[14] R. Lewis, "A General-Purpose Hill-Climbing Method for Order Independent Minimum Grouping Problems: A Case Study in Graph Colouring and Bin Packing", Computers and Operations Research 36 (7): 2295–2310. doi:10.1016/j.cor.2008.09.004, 2013.

[15] R. Michael et al., "A 71/60 theorem for bin packing", Journal of Complexity 1: 65–106, doi:10.1016/0885-064X(85)90022-6, 1985.

[16] G. Dósa, "The Tight Bound of First Fit Decreasing Bin-Packing Algorithm Is FFD(I)≤(11/9)OPT(I)+6/9", in Combinatorics, Algorithms, Probabilistic and Experimental Methodologies, Springer Berlin / Heidelberg, pp. 1–11, doi:10.1007/978-3-540-74450-4, ISBN 978-3-540-74449-8, ISSN 0302-974, 2007

[17] E Kain, "The Top Ten Best-Selling Video Games Of 2014", [Online], From: http://www.forbes.com/sites/erikkain/2015/01/19/the-top-ten-best-selling-video-games-of-2014. [Acessed on 26 September 2015].

AUTHORS' PROFILE

Ritthichai Jitpukdeebodintra is currently the doctorate Candidate in field of computer engineering at faculty of engineering, Prince of Songkhla University. His research interests include technology in computer games, computer graphics, graphics process and cloud computing.
Email: amethystxx@gmail.com

Suntorn Witosurapot is an Assistant Professor in department of Computer Engineering, Faculty of Engineering, in Prince of Songkla University (PSU), HatYai, Thailand. He received the bachelor and Master degrees in Electrical Engineering from PSU, Thailand and Ph.D. degree from Swinburne University of Technology, Melbourne, Victoria, Australia, with the thesis topics related to resolving network resource competition in the Internet.

His research interests include Web engineering and applications, semantic Web, and management of information technology. Currently, most of his research work revolves around Information engineering in smart home network, smart grid infrastructure, and active games for people with visual disabilities.
Email: wsuntorn@psu.ac.th

VITAE

Name Ritthichai Jitpukdeebodintra

Student ID 5310130027

Educational Attainment

Degree	Name of Institution	Year of Graduation
Master of Engineering	Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University	2010
Bachelor of Engineering	Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University	2008

List of Publication and Proceedings

R. Jipukdeebodintra and S. Witosurapot, “Hybrid method for adaptive cloud gaming contents”, “GSTF International Journal on Computing (JoC)”, Vol.4 No.2, March 2016.

R. Jipukdeebodintra and S. Witosurapot, “A study on the impact of client display resolutions in cloud gaming workloads”, Proceeding of 6th Annual International Conference on ICT-BDCS 2015, 2015.

R. Jipukdeebodintra and S. Witosurapot, “Efficient Cloud Gaming Resource Provision Via Multi-dimensional Bin-Packing”, “GSTF International Journal on Computing (JoC)”, Vol.4 No.4, March 2016.