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.

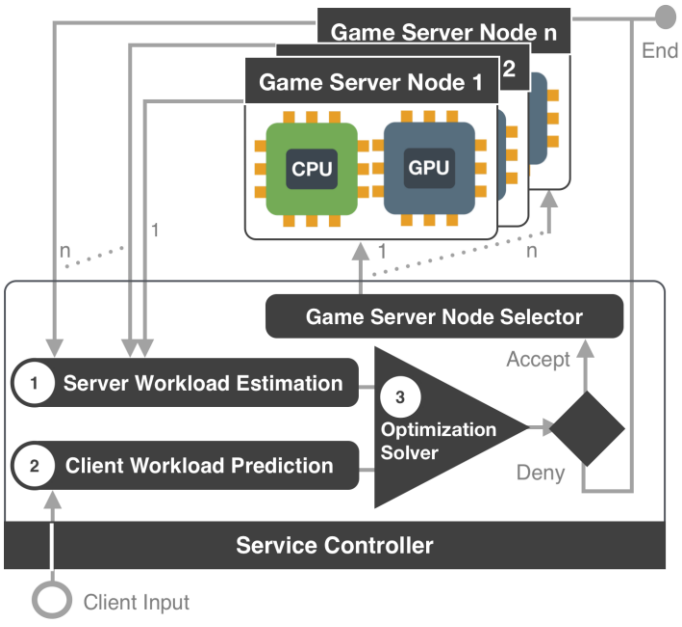

Figure 1. Conceptual diagram of the decision making process for admitting a client request

DOI: 10.5176/2251-3043_4.4.340

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:

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

$$\text{Subject to:} \quad \sum_{i=1}^{n} c_i \le C \tag{2}$$

$$\forall_i \in \{1, \dots, n\} \tag{3}$$

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:

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

$$\text{Subject to:} \quad \sum_{i=1}^{n} c_i \le C\,, \sum_{i=1}^{n} g_i \le G\,, \sum_{i=1}^{n} w_i \le W \tag{5}$$

$$\forall_i \in \{1, \dots, n\} \tag{6}$$

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.

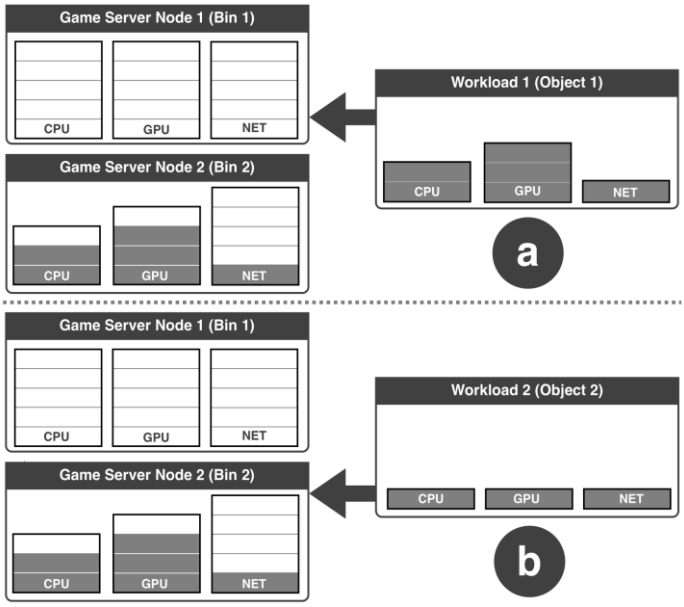

Figure 2. Illustration for explaining to equation (4)

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.

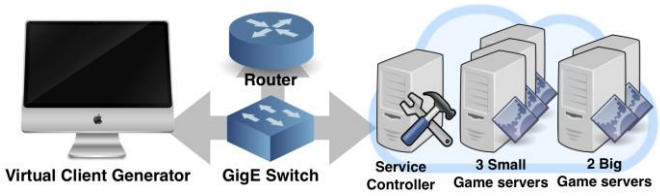

Figure 3. Experimental infrastructure

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

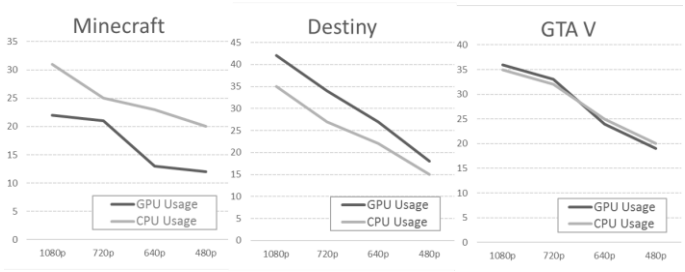

Figure 4. The resource consumption of Minecraft, Destiny and GTA V

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

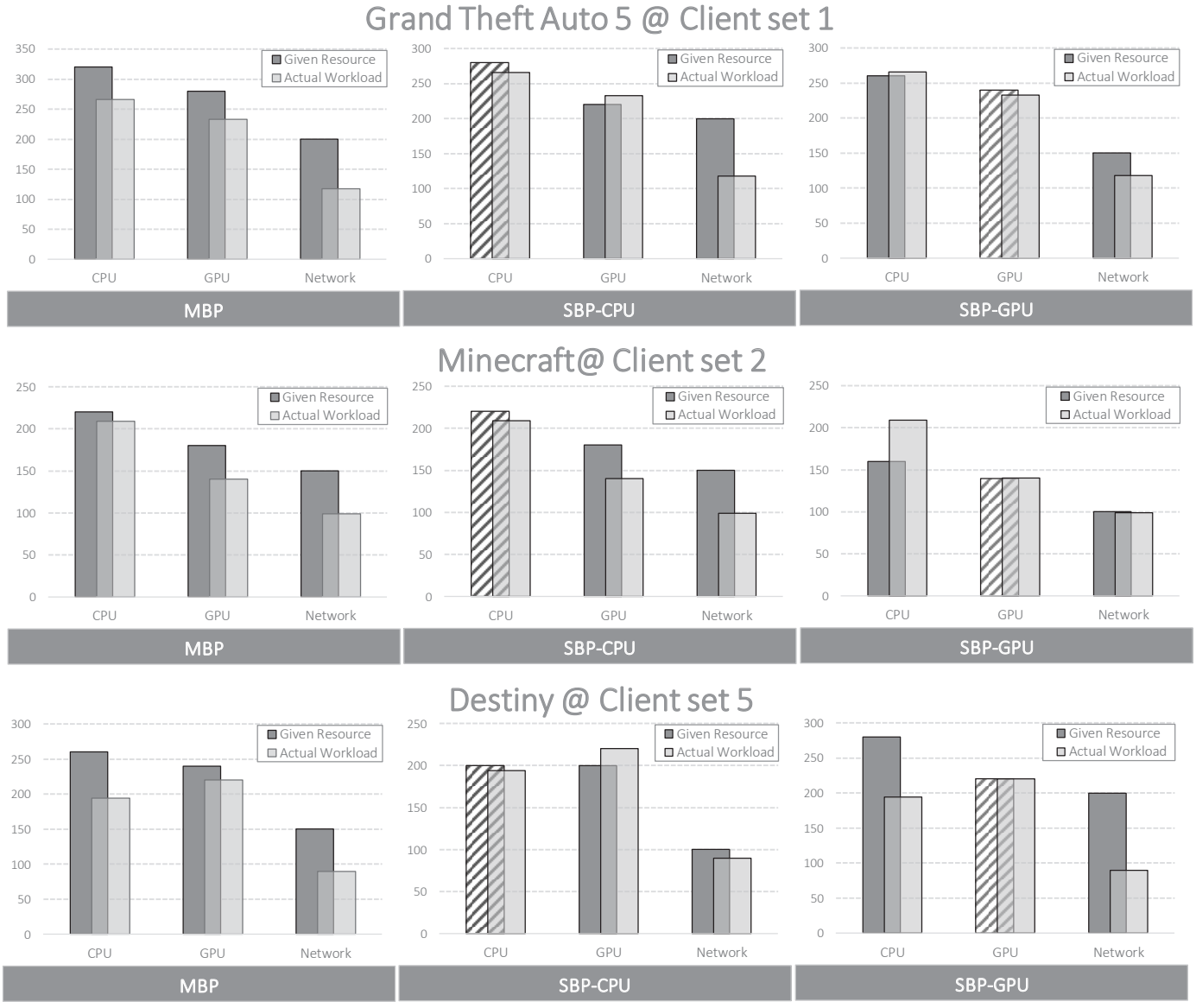

Figure 5. Experimental results

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.

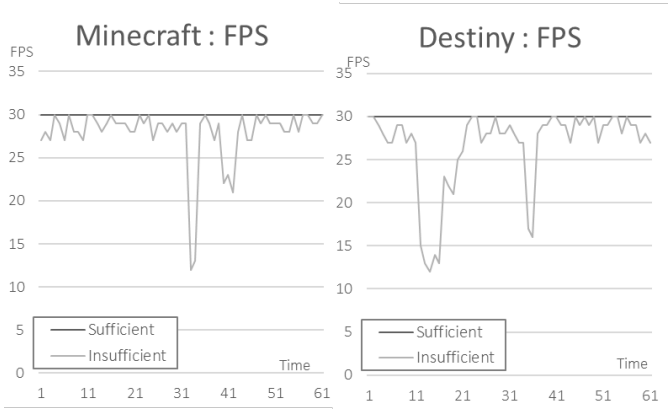

Figure 6. Effect of insufficient resource to game framerates

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