

An analytical method of assessment of RemoteFX as a Cloud Gaming platform

Etienne Depasquale, Audrey Zammit, Michael Camilleri, Saviour Zammit, Adrian Muscat
Digital Gaming Platform Research Group
Dept. of Communications and Computer Engineering
Faculty of ICT, University of Malta
Msida MSD2080, Malta

Pierre Mallia, Stefan Scerri
iMovo Limited,
Tower Business Centre
Swatar, Birkirkara BKR 4013, Malta

Abstract— This paper describes a cloud computing platform that executes graphics intensive programs, such as computer games, with the support of a GPU to render the graphics, and stream the ensuing video to a mobile device over bandwidth constrained channels. Controls on the mobile device allow the user to interact with the game remotely. Initial results based on a Microsoft Windows platform using the Hypervisor in Windows Server 2012, with RemoteFX are reported.

Keywords—Cloud computing; GPU; Computer Games; Hypervisor; RemoteFX

I. INTRODUCTION

The visual quality of current PC and console based games has advanced at a very rapid pace and is yielding ultra realistic graphics to challenge even the new highest definition displays. Although the mobile industry is providing devices with increasingly better displays, these still cannot support such games.

However, the growing popularity of multimedia-oriented mobile devices, coupled with high data rate 802.11g/n WiFi networks and 3G and 4G, point towards a solution where CPU intensive applications are processed on powerful remote servers and streamed to the mobile end-users via a high speed wireless connection.

Remote processing of interactive games is still in its early stages, yet it is generally accepted that there are two techniques worth considering (1) the streaming of computer graphic commands, and (2) real-time video streaming.

In this paper we shall describe a system using a real-time video streaming method we have developed based on the concepts originally presented in [1]. Amongst the major differences to existing solutions, the system developed can service a number of clients through the use of virtualization.

The rest of this paper is organized as follows. Section II examines previous literature published on the subject. Section III introduces the system architecture adopted in the Digimocloud project. Initial results are reported in section IV. Conclusions and future work are described in Section V.

II. BACKGROUND

Recent literature includes many proposals for gaming clouds.

Sponsors: This work is sponsored by the MCST through grant [R&I-2011-010: Digital Gaming Clouds for Mobile Users](#) (DigiMoCloud).

These vary in their scope from systems that are implemented on existing public clouds, to private clouds which serve a more localized set of clients.

Cloud gaming user experience depends on a variety of subjective and objective factors. These can ultimately be grouped into two parameters – the quality of the video received by the client, and the round-trip response time, which is the time necessary for a frame resulting from a user's command to be perceived on the client's screen. [2]

The relative impact of these two parameters depends on the nature of the game being played. In the case of fast games, the Quality of Experience (QoE) decreases more rapidly with increased latency than with increased packet loss. On the other hand, slower games are more sensitive to packet loss than latency. Furthermore, packet loss and latency have a greater effect on the gaming experience when they occur from the server to the client, than in the opposite direction. [3]

Parameterisation of the user's experience of a game

1) Conventional gaming parameters

Frame rate is one primary differentiator between gaming systems in a user's experience of game play. Resolution is another primary differentiator. Frame rate and resolution both compete for resources used in processing graphics and therefore their combination introduces a design constraint given a finite set of such resources. Some benchmarks, such as 3DMark from FutureMark, focus on frame rate. Others, like Performance Test from PassMark add jitter rate to the performance measure. Good frame rate performance is highly valued in gaming communities.

Other conventional gaming parameters include antialiasing, refresh rate, response time, triple buffering and vertical synchronisation. Of these, refresh rate, response time and vertical synchronisation are of concern to the client display and are not expected to have ramifications on the feasibility of game processing on the server.

Antialiasing and triple buffering require attention, since they relate to processing of graphics. It was decided to turn off antialiasing in the first characterisation of the cloud computing infrastructure, in anticipation of carrying development forward until the time when such rough edges in game play could be

afforded attention. The act of turning off antialiasing at the server's GPU has a net effect of reduction in resource consumption during game play. Control over the number of frame buffers is not a part of RemoteFX vGPU architecture. Frames are rendered to a single frame buffer and are then processed in the capture and encoding processes before dispatch over Hyper-V's VMBus to user-mode RDP in the VM where the source of the graphics is being played. Since multiple buffers are an aspect of design in circumstances where frame rendering by the application is expected to be higher than the refresh rate of the display, consideration of this CvGP can be deferred until such high frame delivery rate to the client is actually encountered.

Finally, it may be worth referring that within the particular context of RemoteFX, Microsoft has specified 15 fps at a resolution of 1280 by 1024 pixels as the minimum performance required of a hardware RemoteFX decoder to secure the RemoteFX Enabled certification [8].

During these works, frame rate and frame resolution were the CvGPs selected as the object of study. Future work will expand the scope to include other CvGPs; in particular, jitter rate.

2) Measurement of conventional gaming parameters in this Cloud Computing infrastructure

Microsoft has added instrumentation for RemoteFX to its Performance Monitoring framework on the RDVH. The instrumentation is organised within the following objects: RemoteFX Root GPU Management, RemoteFX Software, RemoteFX Graphics and RemoteFX Network. Apart from the RDVH, the virtual desktop's Performance Monitoring Framework is also augmented when RemoteFX is enabled, by the addition of RemoteFX VM vGPU Performance.

III. SYSTEM ARCHITECTURE

RemoteFX is a term used to refer to a number of technologies that are used to extend the functionality of the Remote Desktop Protocol. A logical representation of the network infrastructure used during the proceedings of work outlined in this section is shown in Fig.1.

A. Server-side RemoteFX technologies

The technologies that are relevant to performance may be divided into two groups: one group affects the **computing** of games in the cloud servers and the other group that affects their **delivery** to the mobile client:

- Group 1: Technologies that enhance or facilitate computing
 - RemoteFX vGPU
- Group 2: Technologies that enhance or facilitate delivery
 - RemoteFX Adaptive Graphics
 - Codec selection at runtime, with selection according to classification of frame regions
 - Facility to automatically adjust encoding loss according to detected network bandwidth
 - RemoteFX for WAN

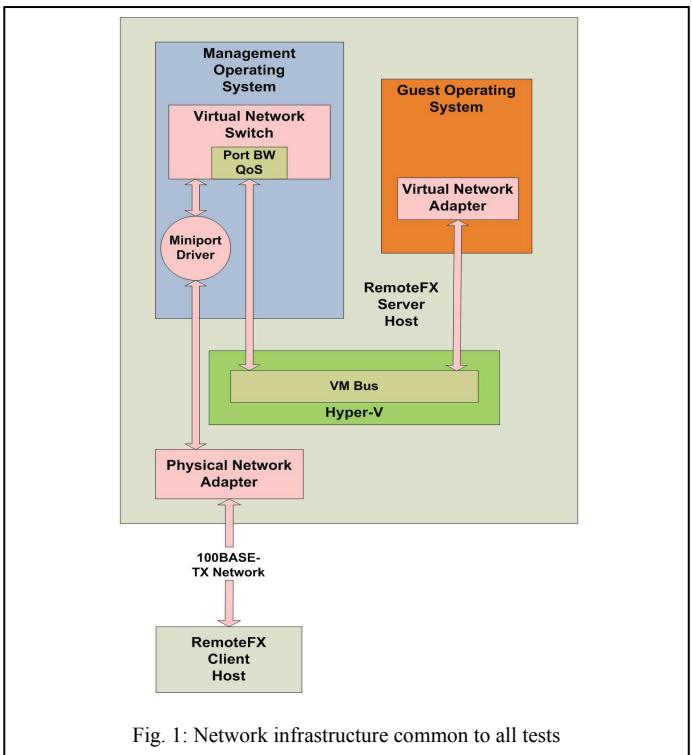


Fig. 1: Network infrastructure common to all tests

The RemoteFX vGPU is inseparable from Hyper-V; it uses Hyper-V's VMBus for inter-process communication between the components of the vGPU that reside on the guest partition and the components of the vGPU that reside on the root partition. A simplification of the RemoteFX vGPU architecture is illustrated in Fig. 2, showing the essential flow in the process of frame delivery. [4]

B. Client-side RemoteFX technologies

A complete consideration of system architecture requires assessment of those components of RemoteFX that relate to the client's interaction with the platform; these components will be dealt with in future work.

IV. SCOPE OF WORK TO DATE

The work presented here is concerned with server-side RemoteFX technologies. The work may be classified as:

1) Preliminary investigations of the user experience under constrained and relaxed bandwidth between gaming server and gaming client.

Work within this scope consisted of a basic assessment of the validity of RemoteFX Adaptive Graphics for the delivery of HD video over conditions of constrained bandwidth. This work made use of QoS bandwidth management introduced in Windows Server 2012 to set an upper bound on the bandwidth available to HD video source applications for delivery of content to clients.

2) Parameterisation of the user's experience of a game

Work was divided into four phases.

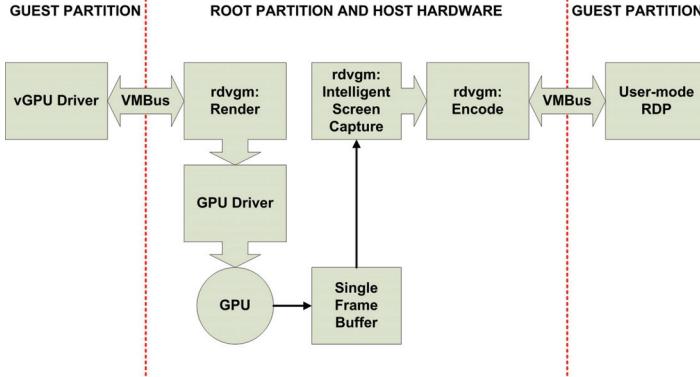


Fig. 2: RemoteFX vGPU architecture (simplified)

a) *Phase 1: Identification of a set of parameters that quantify the user experience in conventional gaming infrastructure*

Analysis and design of gaming infrastructure requires the identification of parameters that vary in a known manner with respect to user experience (UX). The task of identifying the UX parameters consisted of research in the background of conventional gaming. Henceforth, these will be referred to as **conventional gaming parameters**(CvGPs).

b) *Phase 2: A first association between the CvGPs and the context of gaming on a computer system infrastructure for cloud computing.*

Henceforth, this will be abbreviated to **cloud computing infrastructure** but it is conceded that in general, cloud computing infrastructure includes the network between the service's point of presence (server) and the service's consumer (client).

This phase consisted of exercise of effort in four directions:

- Reduction of the infrastructure between gaming client and server to a set of elements that eliminate network bandwidth as a determinant. This effort reduced the infrastructure of concern solely to the server.
- Microsoft's Performance Monitor provides a variety of Performance Monitor Objects and counters that pertain to RemoteFX technologies. These counters were used to identify variables in server-side RemoteFX that affect the CvGPs. Henceforth, these will be referred to as RFXVs.
- Mapping of the CvGPs to the RFXVs.
- Identification of variables internal to the cloud computing infrastructure that affect the RFXVs. Henceforth, these variables will be referred to as **infrastructure variables** (IVs).

Products of this phase were:

- A first association between CvGPs and RFXVs.
- Establishment of a bundle of IVs that constitute a suitable set of resources to assign to a single

virtual machine (VM). The criterion selected for suitability was that no substantial improvement in the RFXVs would be observed by further increases in the IVs allocated to the VM.

c) *Phase 3: Observation of the RFXVs.*

Documentation regarding RFXVs and RemoteFX Performance Monitor counters is limited and attempts at eliciting further documentation were not successful. Work in this phase consisted of execution of various graphics benchmarks in order to exercise the RemoteFX vGPU and thereby observe:

- Significant RFXVs i.e. those that manifested substantial variations during graphics benchmark execution
- The instrumentation points in RemoteFX vGPU processing where the RFXVs are located
- Relationships between the significant RFXVs in the inner workings of the RemoteFX vGPU.

d) *Phase 4: Modelling the performance of the cloud computing infrastructure*

The output of the infrastructure was considered in terms of the RFXVs that had been associated to the CvGPs. The model was described in terms of the RFXVs that reflect the inner workings of the RemoteFX vGPU (henceforth referred to as iRFXVs), resulting in a relationship between CvGPs and iRFXVs.

V. TESTS AND RESULTS

A. *Basic assessment of the validity of RemoteFX Adaptive Graphics*

1) *Varying bandwidth*

a) *Test conditions*

Two HD video sources ([5], [6]) were played on Hyper-V VMs in RDP sessions with the following clients:

- RDP 6.1 client
- RDP 8.0 client

Each client was connected under the following conditions of bandwidth:

- 100Mb/s

- 8Mb/s

Image Quality policy on the Remote Desktop Virtualization Host was set to “Medium”.

b) Results

Results are shown in Fig. 3. All graphs show a consecutive series of 30-second averages of the outbound bandwidth consumption at the VM’s virtual network adapter. Playback time was varied arbitrarily.

2) Varying Image Quality

a) Test conditions

- A well-known graphics benchmark [7] was processed on a Hyper-V Windows 8 Enterprise Edition (EE) VM in an RDP session with an RDP 8.0 client.
- The client was connected by a 100Mb/s bandwidth connection.
- RemoteFX Adaptive Graphics Image Quality policy on the Remote Desktop Virtualization Host was successively set to:
 - Medium
 - Lossless

b) Results

Fig. 4 shows two runs of 3DMark06 in an RDP 8.0 session with a Windows 8 EE VM. An average bandwidth saving of over 50% is evident.

B. Identification of significant parameters

The counters within the various objects were inspected and data collected therefrom to identify those counters relevant to the limited context intended for the experiments, namely that of constrained server resources within a relaxed bandwidth (100Mb/s end-to-end). The counters that affect the CvGPs under this experiment design were found to be those shown in the list below. Effect was determined through observation of significant variation with perceived frame rate.

- RemoteFX Software
 - Capture Rate for monitor [1-4]
 - Delayed Frames/sec
 - GPU response time from Capture
- RemoteFX VM vGPU Performance
 - Data: Invoked presents/sec
 - Data: Outgoing presents/sec
- RemoteFX Graphics
 - Average Encoding Time
 - Frames Skipped/second – Insufficient Client Resources
 - Frames Skipped/second – Insufficient Network Resources
 - Frames Skipped/second – Insufficient Server Resources
 - Input Frames/second
 - Output Frames/second

This observation permitted reduction in the set of relevant parameters, as follows. It is remarkable that the counter RemoteFX Software: GPU response time from Render consistently returned zero values. Furthermore, since the test

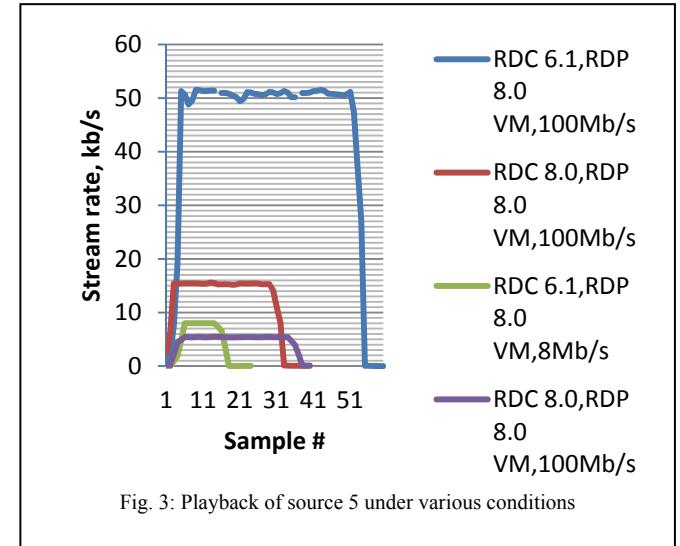


Fig. 3: Playback of source 5 under various conditions

client was a Windows 8 Professional Edition desktop computer, frames skipped/s due to insufficient client resources were insignificant during the initial assessment of the RFXVs. Apart from elimination of parameters by observation of independence from perceived frame rate, delayed frames/s were not considered. These frames are sent to the client but are not sent within a time interval that is proportional to the inter-frame delay. Due to the subjectivity of jitter rate, frame jitter was ignored so as to confine the work expected to ensue within bounds that would permit a first detection of patterns and construction of models.

a) Elimination of network bandwidth as a determinant

3DMark06, 3DMark11 and HD Video Sources([10], [11]) were used as source applications. Inspection of RemoteFX Graphics: Frames Skipped/second – Insufficient Network Resources under an end-to-end bandwidth of 8Mb/s revealed a significant frame rate loss. Expansion to 100Mb/s end-to-end consistently (over a number of runs) eliminated this loss.

b) Mapping of the CvGPs to the RFXVs through observation of the RFXVs

Frame rate and resolution have been selected as the object of study. Minimum frame resolution for HD video (720p) was chosen. This restricted the task of mapping the CvGPs to the task of mapping the delivered frame rate (frame rate seen by

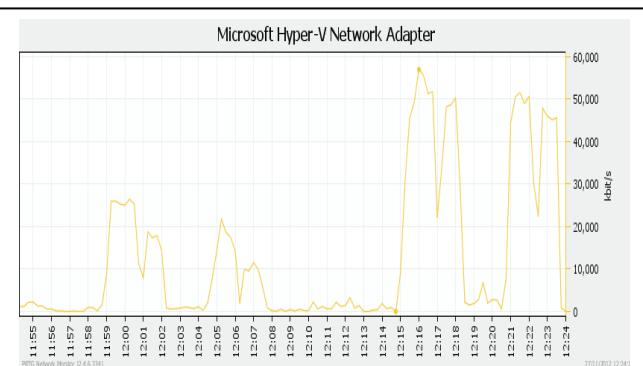


Fig. 4: Windows 8 EE VM running 3DMark06 under medium (left) and lossless (right) image quality

the client) to the RFXVs.

3DMark06 was processed on a Windows 8 EE VM and the RFXVs were collected using Performance Monitor. The results are shown in Table 1.

The exact time bounds within which the various 3DMark06 scores were calculated was not known but a very good approximation was achieved through observation of the real-time plotting of the counters in Performance Monitor and identification of the beginning and end of the four graphics tests. FutureMark's 3DMark11 graphics tests and PassMark's Performance Test graphics software tests confirmed the observation that the frame rates measured by these tests correspond to the "RemoteFX Graphics:Input frames/s" (IFR) counter.

"RemoteFX VM vGPU Performance>Data: Outgoing presents/sec" (OGPR) follows IFR closely. The exception occurs when the frame contents stagnate for a certain period of time. This period of time was not measured but it is well under one hour. In this case, the outgoing presents are buffered and presented in a spike to the capture threads – which drop them all. The result is that the client's display freezes, as witnessed by the clock time at which the OGPR stopped following the IFR. This behaviour is shown in Fig. 5.

Table 1 also shows two other RFXVs: output frame rate (OFR) and monitor capture rate (MCR). The real-time values

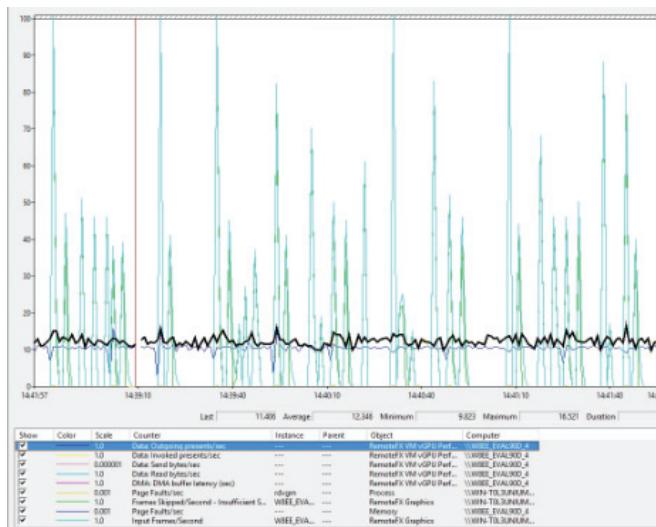


Fig. 5: OGPR (black trace) deviates from IFR (spikes) when frames do not change for a long period of time

TABLE I	
Measure	Windows 8 Medium IQ
gt1: 3DMark06	13.61
gt1: Perfmon IFR	13.78
gt1: Perfmon OFR	9.03
gt1: Perfmon MCR	9.03
<hr/>	
gt2: 3DMark06	7.38
gt2: Perfmon IFR	7.49
gt2: Perfmon OFR	5.54
gt2: Perfmon MCR	5.54
<hr/>	
hdr1: 3DMark06	8.79
hdr1: Perfmon IFR	8.67
hdr1: Perfmon OFR	6.83
hdr1: Perfmon MCR	6.80
<hr/>	
hdr2: 3DMark06	10.36
hdr2: Perfmon IFR	10.29
hdr2: Perfmon OFR	8.93
hdr2: Perfmon MCR	8.80

of these variables follow each other closely; these correspond to the frames that the user sees in this set of experiments, where both losses due to network insufficiencies and client insufficiencies have been eliminated. The exact instrumentation point of the two is unclear: basing on description of OFR and naming convention, it is assumed that the MCR is measured at the output of the capture stage and the

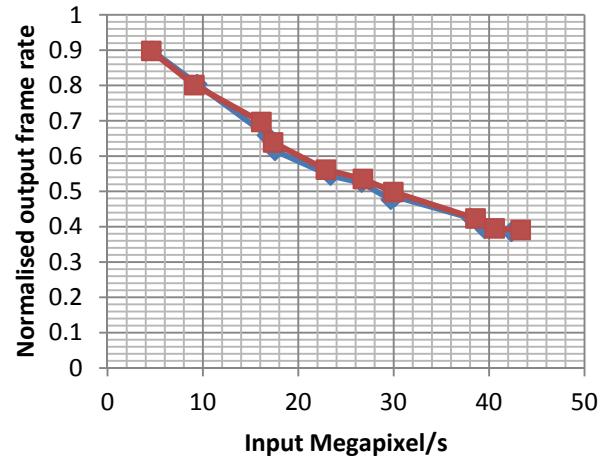


Fig. 7: Effect of variation of memory clock frequency on normalised output frame rate. Brown (darker) line: 1333MHz; Blue (lighter) line just visible below brown line: 800 MHz

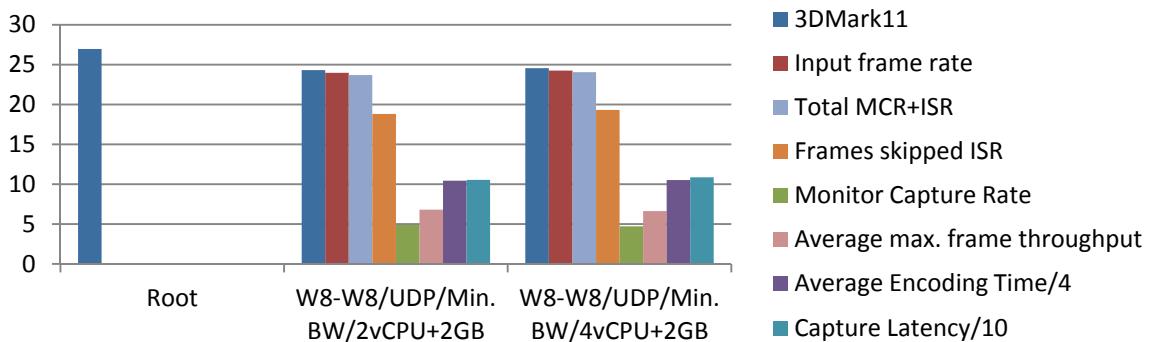


Fig. 6: 3DMark11 GT3 on the root partition, a guest running on 2 vCPUs and a guest running on 4 vCPUs

OFR is measured at the output of the encoder stage.

The difference between the IFR and the MCR is accounted for by “RemoteFX Graphics:Frames Skipped/second – Insufficient Server Resources” (ISR). This is apparent in measurements carried out while using 3DMark11. Fig. 6 shows that the sum of MCR and ISR is equal, within limits of measurement error, to the IFR.

These insights permit the location of the instrumentation points on the RemoteFX vGPU architecture diagram presented in Section 3, in the positions shown in Fig. 8.

C. Optimal resource set for a single VM for processing graphics

Some work was carried out in an attempt to establish an optimum set of resources, i.e. the point of diminishing performance returns. Apart from the use for which it has been introduced, Fig. 6 also serves the purpose of identifying the futility of providing more than 2 vCPUs for a graphics-intensive workload. This balance changes if the VM is loaded with media that require decoding on the CPU; in this case, the optimum balance will require more vCPUs. An example of this genre would be playback of encoded video.

The choice of workload is critical in interpretation of results. Most of the work presented here was intended to identify the demands placed by the RemoteFX vGPU on a cloud unit (the server); therefore workload must avoid stressing the CPU, in the understanding that it is the scalability of the graphics subsystem of the cloud unit that is being evaluated.

A first analysis was carried out of the performance in response to changes in memory clock frequency. A third HD video source [9] was played at various resolutions. The OFR and IFR were measured and plotted against the product of the average IFR and resolution (average input megapixel/s). The result is shown in Fig. 7. More work is required before conclusions can be drawn with high levels of confidence about the effect of memory bandwidth on performance.

The result of this limited work was the selection of a 2 vCPU, 2GB VM as the subject for further test. The CPU cores underlying the vCPUs are those in the Intel Xeon E5645 at

2.4GHz.

D. Modelling the performance of the cloud computing infrastructure

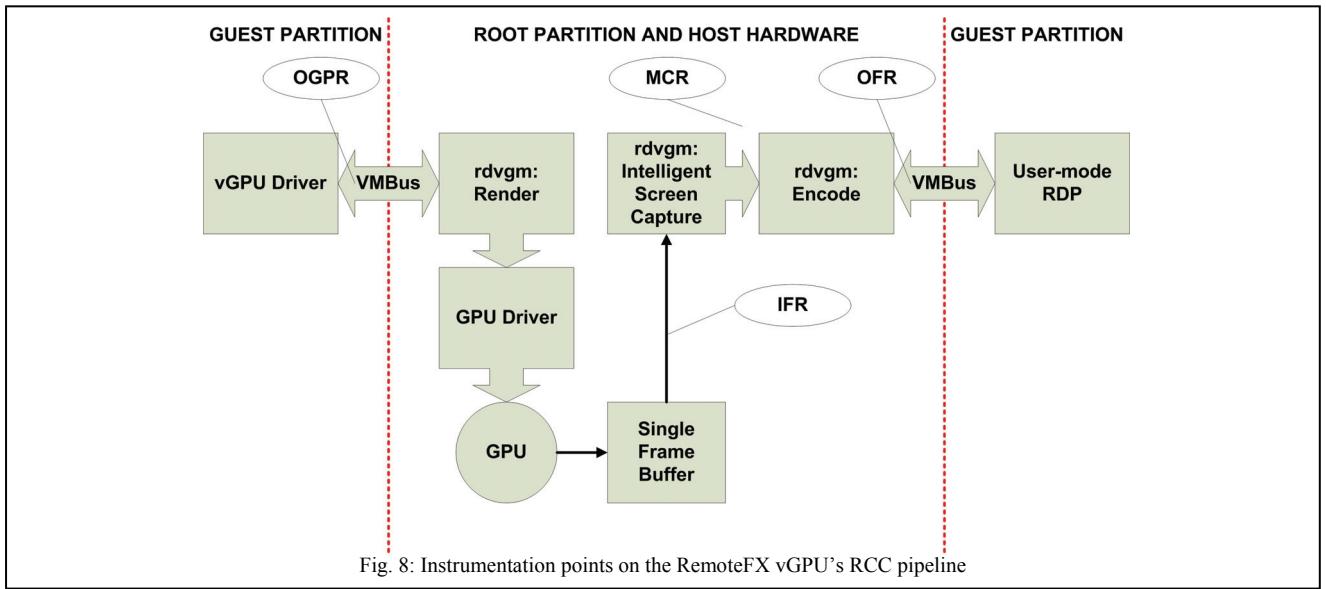
The service provider’s concern lies in the game frame rate that is being delivered to the client. This work has focused on the portion of the infrastructure that processes the games: the cloud unit, or server, with particular emphasis on its graphics subsystem as the engine pushing the frames out onto the Remote Desktop Connection. This is the CvGP on which the work presented here has focused, specifically within the technological context of RemoteFX. Once a usable grasp of the mechanics of the RemoteFX vGPU had been acquired, along with an understanding of its instrumentation (the iRFXVs) and their relation to the CvGP, it then remained to model the relationship between the CvGP and the iRFXVs. The process started with a recapitulation of the salient variables:

- GPU response time from capture, t_C
- Average encoding time per frame, t_E
- rdvgm process page fault rate, \dot{e}
- Input frame rate, \dot{x}
- Frames skipped per second due to insufficient server resources, \dot{y}_S

These variables were then classified into one of two categories:

- Whether the variable was likely to be dependent on others within the set
- Whether the variable was likely to be independent of others within the set

The classification process was based on understanding of the mechanics of the RemoteFX vGPU and previous attempts, not presented here, at identifying dependencies between these variables. The result of the classification was that only \dot{y}_S was identified as dependent; all the others were identified as independent of any other within the set.



Eleven data points were gathered, using 3DMark11 and Performance Test 8.0 DX9, DX10 and DX11 tests. Polynomial fitting of order 1 resulted in the relationship:

$$\frac{\dot{y}_S}{\dot{x}} = 0.0075\dot{x} + 2.0469t_C + 6.9542t_E + 2.1298 \cdot 10^{-6}\dot{e} - 0.0287$$

The weight of the encoding time is substantial. This result seems to strengthen the case for accelerated encoding, delegated to dedicated ASICs.

VI. CONCLUSIONS

A. Basic assessment of the validity of RemoteFX Adaptive Graphics

This component of the RemoteFX set exhibited good adaptation to network condition and administrative policy. When network conditions were reduced to 8Mb/s, the RDP 6.1 client took all available bandwidth and gave unusable frame rate delivery but the RDP 8.0 client adapted, took less than the full bandwidth and played the two 720p HD videos while delivering a user experience that showed some jitter but none of the tearing or freezing evident in the RDP 6.1 client. Furthermore, the RDP 8.0 service adapted well to administrative policy, dropping average bandwidth consumption by over 50% after being configured to reduce image quality from lossless to “medium”.

Visual inspection of the output from delivering the graphics-intensive benchmark revealed severe frame loss. This gave cause to model the processing chain with the objective of identifying the source of this loss. Modelling efforts have revealed the severity of the impact of encoding on the frame loss. Queries addressed to Microsoft’s Senior Program Manager from the “Remote Desktop Virtualization” team have revealed that the facility to delegate encoding to dedicated accelerators has been dropped from Windows Server 2012. The possibility of its re-introduction was conceded.

B. Parameterisation of the user’s experience of a game

Performance in terms of frame rate and frame resolution was selected as the scope of this work. Conventional benchmarks do not reflect the user’s experience of frame rate, since they measure the frame present rate (OGPR, or IFR, which under non-stagnant conditions, are equal), which lies at the input to the RemoteFX vGPU’s RCC pipeline but the user experiences the output of the pipeline, which is measured by the output frame rate (OFR). This does not eliminate the use of benchmarks, since they provide a workload specifically for GPUs but use of the benchmark is limited to corroboration of the assertion that they match the OGPR and the IFR.

The RCC pipeline is a good general-purpose high fidelity machine for processing graphics. It has two major weaknesses, one of which – encoding time - has structural provision for reduction in the form of dedicated ASICs. There is no known provision for the other major weakness, namely the Intelligent Screen Capture. Circumstances where this took over 200ms on average per frame were encountered. This is generally considered to be a strength, not a weakness, but for the

particular case where media type is known beforehand to be video, the use of a control to turn off its classification function may prove to be useful. There is no known data, nor are there any known instrumentation methods, regarding the division of the time which the capture function spends on difference capture and content classification. Since Microsoft claim a mix of algorithmic techniques, heuristics and application-provided hints, it seems reasonable that the content classification function may be a sizable contributor to the capture function’s latency. If, on the other hand, the “application-provided hints” are good enough, then this de-activation of classification may be automatically fired through detection of game content. Admittedly, this is purely speculative thinking.

The model provided conceals the basic fact that the latency in the capture and encoding functions is rooted in hardware capacity and this capacity is not represented directly. Therefore, the essential question of any service provider, which is how little is required to give good performance, is not answered here. What the model does answer, is what loads the graphics system; the rdvgm process (one per VM) is both the engine and the bottleneck.

The model also shows the importance of accelerated encoding and may also answer the question of why the normalised output frame rate, shown in Section IV, does not improve more with faster memory. The low coefficient of page fault rates, reduces the impact of transfers to and from system memory during operation of the rdvgm process. Further investigation may be warranted here.

The identified relationship may serve as a good base to extend modelling of the cloud computing infrastructure by reaching out into the added dimensions of VM- and virtualisation host resource set.

REFERENCES

- [1] Saviour Zammit, Adrian Muscat, and George Gauci, “Mobile gaming on a virtualized infrastructure,” 16th IEEE Mediterranean Electrotechnical Conference, Hammamet, Tunisia, 25-28 March 2012.
- [2] S. Wang, & S. Dey, (2009, November). “Modeling and characterizing user experience in a cloud server based mobile gaming approach” in Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE (pp 1-7). December 2009.
- [3] M. Jarschel, D. Schlosser, S. Scheuring, & T. Hoßfeld, (2011, June) “An evaluation of QoE in cloud gaming based on subjective tests” in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on (pp 330-335). IEEE
- [4] “Microsoft RemoteFX for Virtual Desktop Infrastructure: Architectural Overview”, (pp 11-12), Microsoft Corporation, January 2011.
- [5] “Clip 380250: Medium shot of column obelisk pantheon”, Clipcanvas, <http://www.clipcanvas.com/video-clip-380250-column-obelisk-pantheon-rome-tourist>
- [6] “Clip 328230: Animation of Bethlehem Star Loop”, Clipcanvas, <http://www.clipcanvas.com/video-clip-328230-bethlehem-star-loop-christmas-nativity>
- [7] 3DMark06, Futuremark, <http://www.futuremark.com/benchmarks/3dmark06/>
- [8] R. Williams (Microsoft), “RDP in Server 2012 – Advancing the Protocol”, Microsoft Plugfest 2012, Taipei
- [9] Clip 74552, Clipcanvas, <http://www.clipcanvas.com/free-footage/wswmedia-download.htm>