SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Boosting performance of virtualized desktop infrastructure with physical GPU and SPICE

Shupan Li[1,2], Limin Xiao[1,2*], Chungang Shi[3], Liequan Che[3], Changyou Zhang[4] & Yuanzhang Li[5]

[1]*State Key Laboratory of Software Development Environment, Beihang University, Beijing* 100191*, China;*
[2]*School of Computer Science and Engineering, Beihang University, Beijing* 100191*, China;*
[3]*Beijing Jinghang Computation and Communication Research Institute, Beijing* 100074*, China;*
[4]*Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China;*
[5]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing* 100081*, China*

Appendix A Experiment Environment

The experiment environment includes five parts, the physical machine (PM), the virtual machine (VM), the terminal, the network, and the encoding type of the the display information (DI). The configurations of the experiment environment are listed below:

(1) The configuration of the PM: The CPU is Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz * 2, 256KB L2 cache, 35840KB L3 cache. The memory is DDR4 2400 MHz with 16GB * 8. The type of the hard disk is SATA, and the size of it is 3TB. The operating system (OS) is Ubuntu 18.04. The virtual machine monitor (VMM) is KVM and Qemu, and the versions of them both are 2.11.1. The physical GPU (pGPU) is AMD Radeon WX4100.

(2) The configuration of the VM: The CPU of the VM has 2 cores. The memory size is 4GB. The image format is Qcow2 with 50GB. The OS is windows 7. The pGPU is assigned to VM. The virtual GPU (vGPU) is QXL.

(3) The configuration of the terminal: The CPU is Intel(R) Core(TM) i5-4210H CPU @ 2.90GHz, The memory is DDR3 with 8GB. The OS is Ubuntu 18.04. The type of the hard disk is SATA, and the size of it is 1TB.

(4) The configuration of the network: The network is 1000Mb/s.

(5) The encoding type of the DI: When the VM uses pGPU, the encoding type of the DI is H.264. When the VM uses vGPU, the encoding type of the DI is MPEG.

Appendix B Function Evaluation

Function evaluation is used to verify the correctness of the optimized SPICE. We evaluate the function of the optimized SPICE from two aspects: One is to evaluate whether the GPUs support 3D and the performances of the GPUs. The other is whether the original SPICE and the optimized SPICE can display the DI from the pGPU to the user.

Appendix B.1 3D supporting and the performance of the GPUs

We use 3Dmark to evaluate whether the GPUs support 3D and we also evaluate the performances of the GPUs. 3DMark is a performance testing tool for GPU, and the performance is indicated by a score. The higher the score the GPU gets, the better. When the VM uses QXL, the tests in 3DMark could not be run. When the VM uses pGPU and the PM uses pGPU, we were able to measure the performance using 3DMark. Therefore, QXL does not support 3D; however, the pGPU can support 3D. The scores from the GPU performance tests are shown in Fig. B1. The scores of the VM using the pGPU are less than that of the PM using the pGPU, for the same test cases. This is because that the vt-d and Qemu cause a loss of performance.

* Corresponding author (email: xiaolm@buaa.edu.cn)

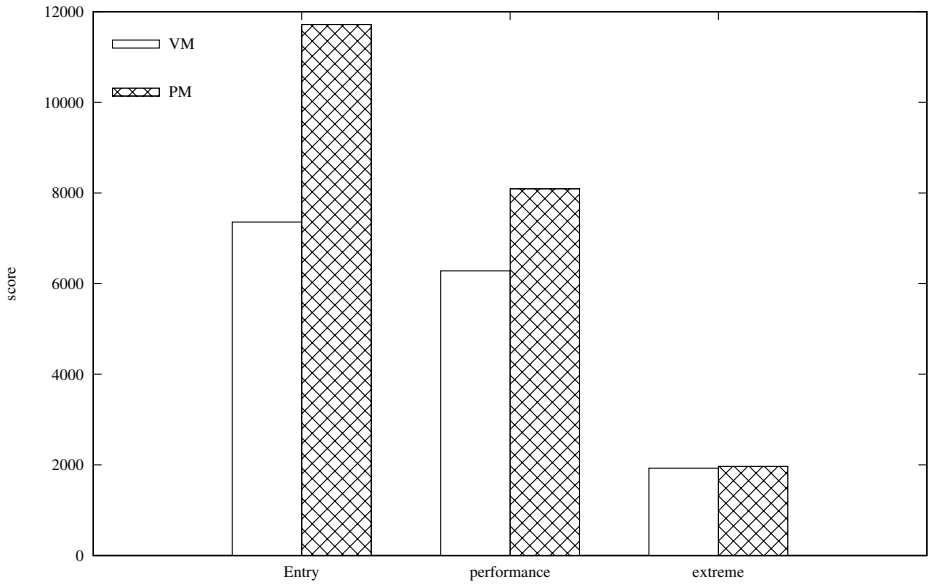


Figure B1 The performance of the GPUs.

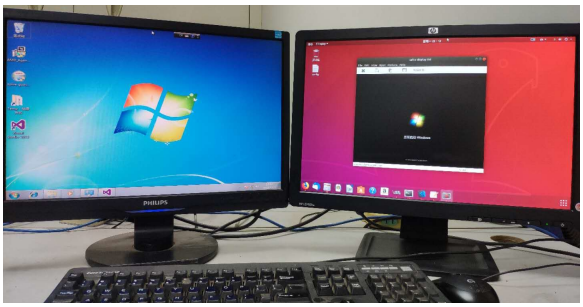


Figure B2 The DI in the SPICE client when the VDI uses the original SPICE.

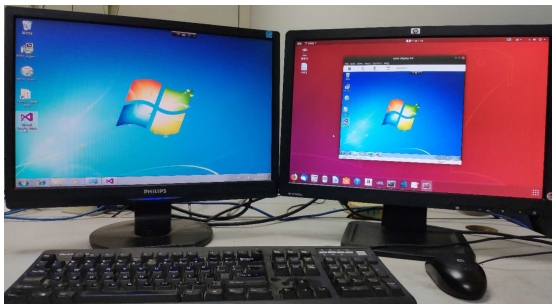


Figure B3 The DI in the SPICE client when the VDI uses the optimized SPICE.

Appendix B.2 Transferring the DI from the pGPU to the user

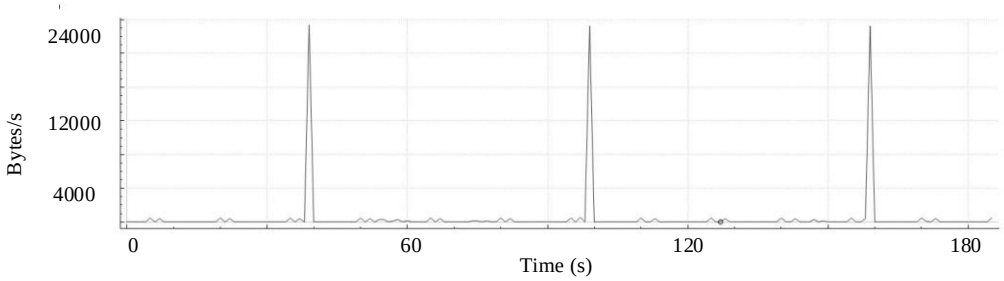
We evaluate whether the original SPICE and the optimized SPICE can transfer the DI from the driver of pGPU to SPICE client when the VM uses a pGPU. Figure. B2 shows the result when the VM uses the original SPICE and Fig. B3 shows the result when the VM uses the optimized SPICE. The monitors on the left in Fig. B2 and Fig. B3 receive the DI directly from the mini DP port of the pGPU, and the SPICE clients on the right in Fig. B2 and Fig. B3 receive the DI from the SPICE server. We can see that the original SPICE cannot transfer the DI from pGPU to the SPICE client; however, the optimized SPICE can.

Appendix C Performance Evaluation

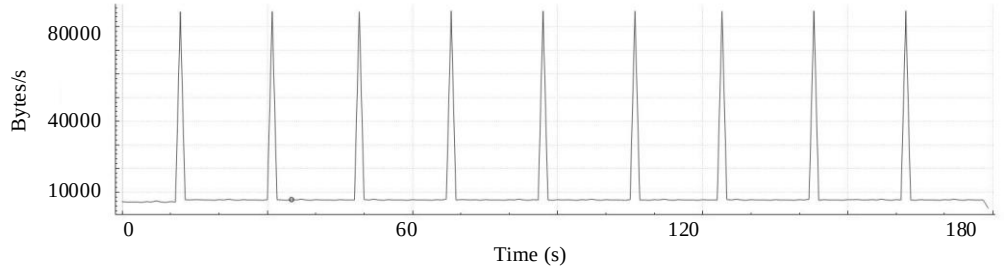
We evaluate the performance of the optimized SPICE from three aspects, network traffic, CPU utilization, and latency of the DI.

Appendix C.1 The network traffic

We use WireShark to evaluate the network traffic, as the DI is sent to the SPICE client over the network. In general, the smaller the network traffic, the better. We compare the network traffic under two conditions. One in which the DI does not change, and another in which the DI changes rapidly (the VM plays a 1080P video). The results are shown in Fig. C1 and Fig. C2, that the results review the following: (1) In the former condition: (a)the network traffic when the VM uses the vGPU is always less than that when the VM uses the pGPU. For a VM using avGPU, there are two DI buffers in the SPICE server and client, and SPICE server does not send the DI, if the DI do not change. (b)The network traffic is



（a）The network traffic when the VM uses vGPU



（b）The network traffic when the VM uses pGPU

Figure C1 The network traffic when the DI does not change.

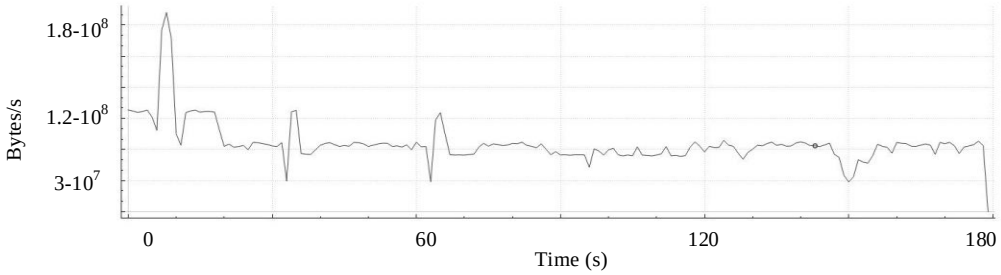
periodic. The period of the network traffic when the VM uses the vGPU is approximately 60s, and that when the VM uses the pGPU is approximately 35s. (c)Sometimes, the network traffic can sometimes spike. The reason is that it is used to send the whole frame of the DI to avoid the errors after losing some frames of the DI. (2) For the latter condition, we have the following: According to the type of encoding, the network traffic when the VM uses a vGPU is much more than that when the VM uses a pGPU, and the variation of the network traffic when the VM uses a vGPU is larger than that when the VM uses a pGPU. The MPEG encoding is lossless and uncompressed. The sizes of the frames in MPEG encoding are almost same. The H.264 encoding is lossy and just recodes the relationship between the adjacent frames of DI. The greater the changes between two frames in H.264 encoding, the more data needs to be recorded.

Appendix C.2 The CPU utilization

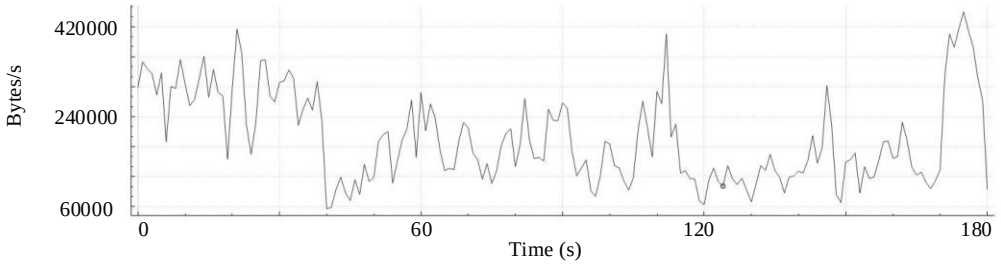
The processing power of the CPU is an important resource. We use CPU utilization to measure the amount of the CPU resource that are used. In general, the lower the CPU utilization is occupied, the better. Therefore, we compare the CPU utilizations in the VM and the PM for two conditions. One is when the VM does not run any program, and the other is when the VM plays a 1080P video. In Linux, we use the command "top" to get the CPU utilization, and the accuracy of the CPU utilization is 0.1%. In Windows, we get the CPU utilization from the "Windows Task Manager", and the accuracy of the CPU utilization is 1%. If the CPU utilization is less than the accuracy, the value is "0". This result is shown in Fig. C3, we can see that the following: (1) When the VM does not run any program, the CPU utilizations of the PM and the VM when the VM uses the vGPU are the same as that when the VM uses the vGPU. (2) When the VM plays 1080P video, the CPU utilizations of the PM and the VM when the VM uses the pGPU are both less than that when the VM uses the vGPU. This is because that the vGPU is simulated by the VMM and it uses the CPU resource; however, the pGPU is hardware and it uses its own resource.

Appendix C.3 The latency of the DI

When the users use a VDI, the DI has a latency, as the DI is transferred over the network, compared to when DI is obtained directly from the DP port of a pGPU. We evaluated the latency of the DI, when the VM plays a 1080P video. The latency is defined as the time interval between when we begin capturing the packet of DI until when the packet arrives at the terminal. The latency is approximately 600ms and varies with the sizes of the DI packages. This latency is acceptable.



（a）The network traffic when the VM uses vGPU



（b）The network traffic when the VM uses pGPU

Figure C2 The network traffic when the DI changes rapidly (the VM plays 1080P video).

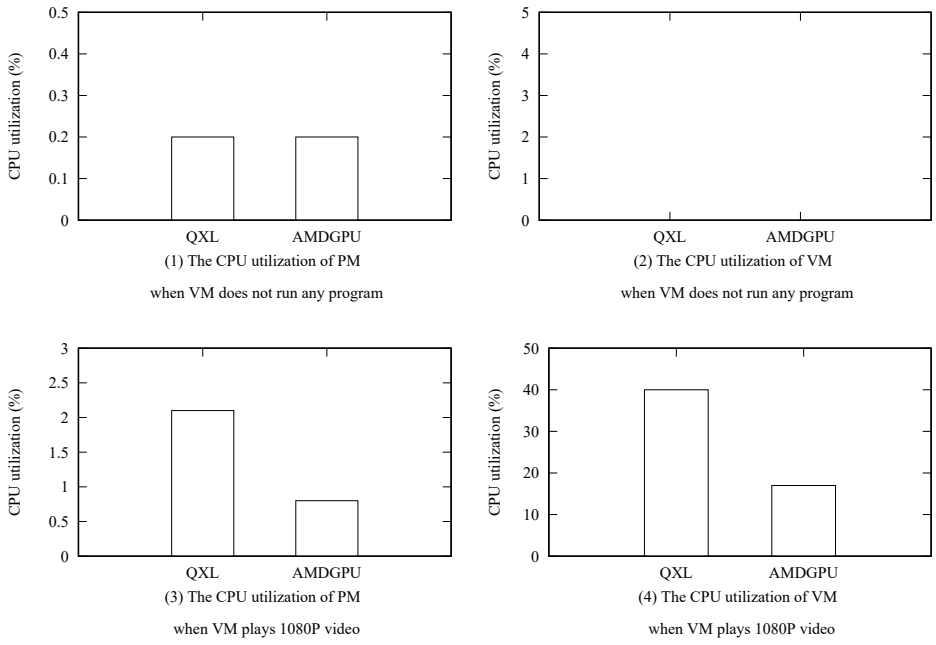


Figure C3 The CPU utilization.