Efficient Processing of Videos in a Multi-Auditory Environment Using Device Lending of GPUs

Konstantin Pogorelov[1], Michael Riegler[1], Jonas Markussen[1], Håkon Kvale Stensland[1]
Pål Halvorsen[1], Carsten Griwodz[1], Sigrun Losada Eskeland[3], Thomas de Lange[23]

[1]Simula Research Laboratory and University of Oslo
[2]Cancer Registry of Norwayy [3]Vestre Viken Hospital Trust

konstantin@simula.no

ABSTRACT

In this paper, we present a demo that utilizes Device Lending via PCI Express (PCIe) in the context of a multi-auditory environment. Device Lending is a transparent, low-latency cross-machine PCIe device sharing mechanism without *any* the need for implementing application-specific distribution mechanisms. As workload, we use a computer-aided diagnosis system that is used to automatically find polyps and mark them for medical doctors during a colonoscopy. We choose this scenario because one of the main requirements is to perform the analysis in real-time. The demonstration consists of a setup of two computers that demonstrates how Device Lending can be used to improve performance, as well as its effect of providing the performance needed for real-time feedback. We also present a performance evaluation that shows its real-time capabilities of it.

CCS Concepts

•**Information systems** → **Information retrieval; Multimedia and multimodal retrieval;**

Keywords

Medical Multimedia; Information Systems; Classification

1. INTRODUCTION

Colonoscopy is a medical procedure, during which specialists in bowel diseases (gastroenterologists), investigate and operate on the colon through minimally invasive surgery by using flexible endoscopes. These examinations are usually done in a special examination room as depicted in figure 1(a). A standard hospital normally has several of these rooms in their gastroenterology department. These rooms contain screens for the doctors that show the video stream from the camera, a bed for the patient, the endoscopic processor, a desktop computer for reporting and some medical treatment supplies. The endoscopes can vary in their attributes like the thickness of the endoscope or its length, but also in the resolution of the videos. Figure 1(b) shows a collection of different endoscopes. Endoscopes are frequently moved between examination rooms to fit the requirements of a specific examination. From the tip of the endoscope (figure 1(c)), a video is transmitted, and the gastroenterologist relies on the video stream to diagnose disease and apply treatments. To control the endoscope, the control unit that is part of every endoscope is used. As one can see in figure 1(d), this is a complex mechanism that requires a lot of concentration from the doctor during the whole procedure, lasting up to 2 hours depending on the findings. The camera can be seen as the virtual the eye of the gatroentologists, and the video stream is all they perceive. Usually, doctors get "third eye" support from their nurses to support them during the examinations and increase the number of findings.

Recently, computer-aided diagnostic systems are more and more used in gastroenterology. The most recent and best

(a) The examination room where the endoscopies are performed. A usuall hospital has several of these rooms.

(b) Different endoscopes for different examinations and patients. For example the very small one is for children.

(c) The tip of the endoscope. It is very flexible and can be moved by the gastroentologist in every possible direction.

(d) The controll unit of the endoscope the gastroentologiest uses to controll the endoscope in terms of zoom, rotation, etc.

Figure 1: These images show an auditorium and endoscopic equipment in the Bærum Hospital in Norway where our system will be used.

MMSys'16 May 10-13, 2016, Klagenfurt, Austria

ACM ISBN 978-1-4503-4297-1/16/05.

DOI: http://dx.doi.org/10.1145/2910017.2910636

working system is Polyp-Alert [10]. This computer-aided diagnostic system helps to determine the quality of the colonoscopy during the procedure. It reaches very high accuracy and sensitivity, but it only reaches near real-time and not full real-time feedback. This is not optimal for live examinations where the medical expert controls the camera manually and cannot rely on a system that introduces delays. Even though real-time performance can be reached by using multiple GPUs in one sufficiently powerful desktop machine, placing such noisy and costly machines in the examination rooms of a hospital is impractical. A more realistic scenario is therefore to have or to use already installed smaller machines in each room and to use Device Lending whenever more resources are needed. Here, Device Lending is a concept where computers interconnected in a PCI Express network can share devices using a transparent cross-machine device sharing system without any special effors to use remote resources locally. It is a low-latency, high-throughput solution for distributed computing, utilizing common hardware already present in all modern computers and requiring little additional interconnection hardware.

In this paper, we will present a demo that utilizes Device Lending of GPUs in combination with our own computer-aided diagnosis system. With this demo, we address two main challenges. First, we will show that real-time support is possible using this technology. Second, we demonstrate the possibility of having one mainframe that can lend the devices to different computers based on the computational demands. This can be an important advantage and even required for scenarios where no room for large machines exists. Further, it can be important for setups where the requirements change fast and often on the fly (e.g., an examination room in a hospital changes the used endoscopes several times during the day; endoscopes with a very high resolution need more processing power than those with lower resolution).

2. REAL-TIME COMPUTER AIDED DIAGNOSIS SUPPORT

Automatic detection of polyps in colonoscopies has been in focus of research for a long time [9]. However, few complete systems exist that are able to do real-time detection, or that can support endoscopists by computer-aided diagnosis for colonoscopies in real-time and at the same time maintain a high detection accuracy. The most recent and best working approach is Polyp-Alert [10] that is able to give near real-time feedback during colonoscopies. Visual features and a rule based classifier are used to detect the edges of polyps, and a performance of 97.7% correctly detected polyps is reported. However, real-time support is limited as they reach only 10 frames per second.

To target the real-time performance, we have proposed EIR [8, 7, 6] medical experts supporting system for the task of detecting diseases and anatomical landmarks in the gastrointestinal (GI) tract, which used in this demo as a use case. It has several key attributes, i.e., EIR (i) is easy to use, (ii) is easy to extend to different diseases, (iii) can do real time handling of multimedia content, (iv) is able to be used as a live system and (v) has high classification performance with minimal false negative classification results. Compared to Polyp-Alert, our detection accuracy is slightly below. The classification performance of the polyp detection in our EIR system lies around a precision of 0.903 and a recall of 0.919, but it is tested on a different dataset, meaning that the numbers are not directly comparable.

Currently, the system consists of two parts, the detection subsystem that detects irregularities in video frames and images and the localisation subsystem that localises the exact position of the disease. The detection can not determine the location of the found irregularity. The location determination is done by the localisation subsystem. The localisation subsystem uses the output of the detection system as input. After the automatic detection and analysis of the content, the output has to be presented in a meaningful way to the gastroentologists. Therefore, the system has a visualisation subsystem that is reliable, robust and easy to understand also under stressful situations that can occur during a live examination. Moreover, it supports easy search and browsing through a large amount of data after the examination. In this demo, we do not focus on EIR but rather using Device Lending and how it can improve performance. EIR itself is just a relevant use case.

2.1 GPU Implementation

Parts of EIR had to be improved and changed to run on multiple GPUs and allow the system to perform in real-time. Therefore, the most compute-intensive parts have been ported to CUDA, a computation support framework for nVidia graphic cards. To achieve this, parts of the system had to be built as a heterogeneous processing subsystem. The GPU framework supports at the moment a number of features, namely Joint Composite Descriptor (JCD), which includes Fuzzy Color and Texture Histogram (FCTH) and Color and Edge Directivity Descriptor (CEDD), and Tamura, but we are working on increasing the supported features.

A main processing application interacts with a modular image processing subsystem. Both of these are implemented in Java. A multi-threading architecture is used by the image processing unit to handle multiple processing and feature extraction requests at the same time. A shared library that is responsible for maintaining connection with and stream data to the stand-alone CUDA-enabled processing server is implemented in C++. To ensure high data transfer performance and reduce excessive data copy operations, shared memory has been used, while sending requests and receiving status responses uses local UNIX sockets. A CUDA server implemented in C++ runs in the background and performs computations on GPU. The whole system can easily be extended with multiple CUDA servers running locally or on a number of remote servers. This is also valid for the processing server, which can be extended with new feature extractors and advanced image processing algorithms, and utilize multi-core CPU and GPU resources concurrently.

2.2 Device Lending

Device Lending is a concept where computers interconnected in a PCI Express [5] network can share devices. It provides transparent, low-latency cross-machine PCIe device sharing without *any* need to implement application-specific distribution mechanisms or modify native device drivers. As the workload increases or decreases, the system can allocate and de-allocate additional resources.

Today, PCIe is the most common interconnection network inside a computer, and with PCIe non-transparent bridges (NTB) [1], it can be turned into an interconnection network

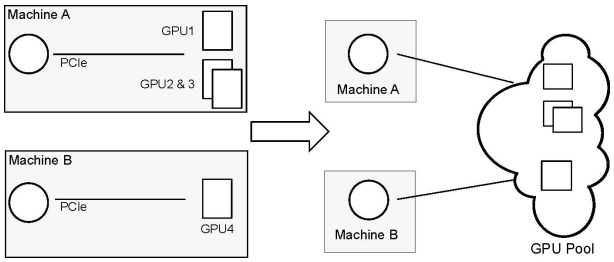


Figure 2: Pooling of devices attached in the PCIe network in the experimental setup.

for multiple machines. In PCIe, all devices connected to the computer are considered part of one common resource pool (figure 2). All devices resources in PCIe are represented by addresses that can be mapped into a remote memory space by an NTB. Device Lending is implemented [3] using Dolphin Interconnect Solutions NTB software [1].

For the EIR system, Device Lending enables the combination of multiple GPUs through CUDA's own peer-to-peer communication model, instead of either writing a distributed system, using rCUDA [2] or MPI [4].

2.3 Performance Evaluation

To evaluate the performance of our system and also to show that Device Lending in our scenario works as intended, we performed 4 different experiment sets. An overview of the hardware used and the performed experiments can be found in table 1. For all configurations, we used the same CPU (Intel Core i7-4820K 3.7GHz) and RAM (16GB Quad Channel DDR3). The test setup consists of 2 computers (Machine A and B, see figure 2), where the host code of the tests runs on one of them. The second one lends a GPU to it. Experiment E1 uses one local GPU, E2 uses two local GPUs and E3 uses three local GPUs. In E4, we borrowed one GPU from the second computer in addition to three local GPUs. With the current machine setup it is not possible to lend more that one GPU because of software limitations in the motherboard's BIOS.

In the experiments, we performed polyp classification and real-time feedback on the video for up to 16 parallel video streams. All video streams are full HD (1920x1080) videos from colonoscopies. We measured the performance from capturing the video up to showing the output on the screen. The complete evaluation is shown in figure 3.

Figure 3(a) shows the performance in terms of processing time per frame for all streams simultaneous. The results

Device	Type	E1	E2	E3	E4
GPU1	Nvidia Tesla K40c	*	*	*	*
GPU2	Nvidia Quadro K2200		*	*	*
GPU3	Nvidia GeForce GTX 750			*	*
GPU4	Nvidia Tesla K40c				*

Table 1: This table shows the used hardware combinations of the different experiments. GPU 1 to 3 are local GPUs. GPU4 is lend via Device Lending.

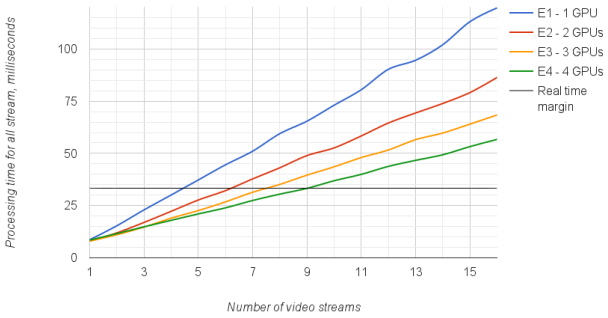
reveal that for up to 7 parallel full HD streams, the 3 local GPUs are fast enough. For more than 7 streams, GPU lending is required. The graph shows that the more parallel streams are processed, the better is the performance gain from the borrowed GPU. We assume that this is due to the excessive overhead for transferring small amount of data, which hinders Device Lending to reach its full potential. This becomes less important when we have more parallel streams, and that Device Lending can indeed improve performance.

The plot in figure 3(b) shows the overall system performance. The evaluation shows that Device Lending can indeed improve the system performance. The maximum overall frames per second we reach when using 4 GPUs at the same time is 30 fps for 9 parallel full HD streams, which is equivalent to 270 fps for a single video stream. Further, this graph shows that the borrowed GPU does not increase the performance for a smaller number of videos, but for 5 and more videos the increase is higher. This is another indicator that Device Lending can increase performance a lot for large scale processing.

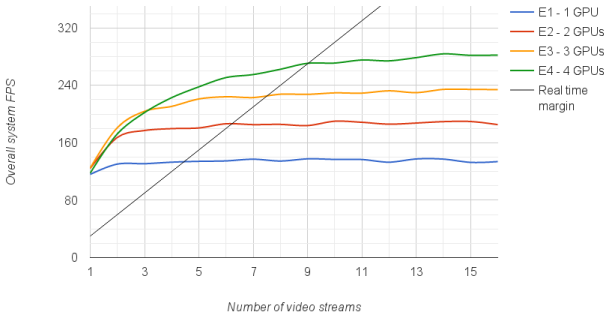
All in all, the experiments showed two important things: (i) Device Lending does not make sense for small amounts of data, but if the data to process is large it can give a large performance boost, and (ii) Device Lending makes sense in a multi-auditory scenario like we present with our demo.

3. DEMONSTRATION SETUP

The above experiments show the performance of EIR on powerful machines and that Device Lending works efficiently, i.e., high performance and low latencies at a very low overhead. However, placing such a setup in the many examination rooms in a hospital is impractical for a number of reasons like high costs and noisy machines. A more realistic scenario is therefore to have smaller machines in each room and use Device Lending whenever more resources are needed.



(a) Frame processing time for several full HD streams in parallel.



(b) Overall system performance for multiple full HD steams in parallel.

Figure 3: System performance evaluation in terms of processing time per frame and maximum performance using 4 different configurations described in table 1. Each video stream is a full HD video.

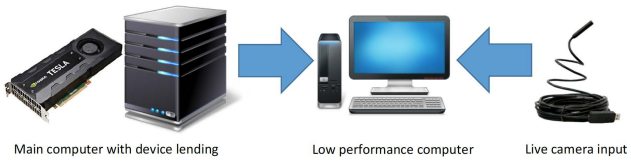


Figure 4: A compete overview of the demo setup. The demo consists of 2 computers, 1 Dolphin interconnect device, 1 screen, an artificial colon and a flexible camera. The users can use the camera in the flexible colon and will get real-time feedback about possible findings. Furthermore, the demo can be switched between Device Lending on and off to demonstrate the effect of it more clear.

To demonstrate the usefulness of Device Lending, we therefore use the above scenario. In the demo, users can use a flexible camera to perform a colonoscopy in an artificial colon, and the system will support them in real-time with analysis and feedback. The complete demo setup is depicted in figure 4. During the demo, the camera can be used to examine the artificial colon and the output of the system will be shown in real-time on the screen. The demo will show the performance increase when a GPU can is borrowed from another machine. Therefore, the demo application can be switched between lending and not lending a GPU. An example of the output for detected polyps can be seen in figure 5. This setup is similar to our real world setup of the system for live colonoscopy with videos as shown to the doctors. Thus, the processing will be done on a very weak computer that is not able to perform the complicated analysis in real-time. Therefore, it is connected to another PC via a Dolphin interconnect device and uses Device Lending to allocate the required processing power. The demo will clearly show the visible differences when Device Lending is used and when not. We also would like to point out, that the presented demonstration is based on the findings in [3] which describes the Device Lending in more detail for further reading.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a demo for Device Lending for computer-aided diagnosis that can assist medical doctors to analyse colonoscopy videos in a multi-auditory scenario. We proved that we can reach high performance in terms of processing time for several full HD video streams in parallel which make it possible to use the proposed system during several and parallel live colonoscopies. We showed that running multiple classifiers in parallel by offloading the processing to multiple machines connected through a PCI Express network and using GPU lending works in our scenario. This optimized version of the application will be able to dynamically allocate, distribute and release compute resources on demand from a pool of available GPUs. For future work, we would like to improve the scheduling of tasks within our lending network. This would include decisions for what and how much to lend to which part of the system using different input information like the required support level of doctors and the endoscope used. We also think that this idea is applicable to other scenarios like for example in cinemas where a less powerful PC in each saloon allocates GPUs based on the quality of the movie to show, e.g., one room shows 4k, one 3D and another one full HD.

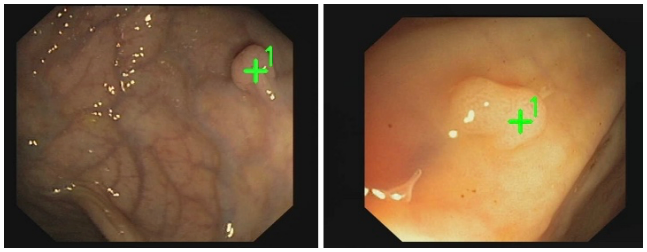


Figure 5: This figure shows 2 examples of what the doctor will see on the screen and what we will show during the demo. In both pictures, the system detected polyps and marked them with a cross. If nothing is detected, the corners of the screen are marked green for feedback.

5. ACKNOWLEDGMENT

This work has been performed in context of the FRINATEK project *EONS* (#231687) and the BIA project *PCIe* (#235530) funded by the Research Council of Norway (RCN). The authors also acknowledge Lars Bjørlykke Kristiansen and Dolphin Interconnect Solutions for assistance with Device Lending and PCIe interconnect equipment. We also would like to thank Mathias Lux from the University of Klagenfurt for "lending" us hardware at the conference venue.

6. REFERENCES

[1] Dolphin Interconnect Solution PXH810 NTB Adapter, 2015.

[2] J. Duato, A. Pena, F. Silla, R. Mayo, and E. Quintana-Ortí. rCUDA: Reducing the number of GPU-based accelerators in high performance clusters. In *Proc. of HPCS*, pages 224–231, 2010.

[3] L. B. Kristiansen, J. Markussen, H. K. Stensland, M. Riegler, H. Kohmann, F. Seifert, R. Nordstrøm, C. Griwodz, and P. Halvorsen. Device lending in PCI Express Networks. In *Proc. of NOSSDAV*, 2016.

[4] NVIDIA Corporation. *Developing a Linux Kernel Module using GPUDirect RDMA*, 2015.

[5] PCI-SIG. *PCI Express 3.1 Base Specification*, 2010.

[6] K. Pogorelov, M. Riegler, P. Halvorsen, P. T. Schmidt, C. Griwodz, D. Johansen, S. L. Eskeland, and T. de Lange. GPU-accelerated real-time gastrointestinal diseases detection. In *Proc. of CBMS*, 2016.

[7] M. Riegler, K. Pogorelov, P. Halvorsen, T. de Lange, C. Griwodz, P. T. Schmidt, S. L. Eskeland, and D. Johansen. EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies. In *Proc. of CBMI*, 2016.

[8] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, and S. L. Eskeland. Computer aided disease detection system for gastrointestinal examinations. In *Proc. of MMSys*, 2016.

[9] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen. Near real-time retroflexion detection in colonoscopy. *IEEE BMHI*, 17(1):143–152, 2013.

[10] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. de Groen. Polyp-alert: Near real-time feedback during colonoscopy. *CMPBM*, 120(3):164–179, 2015.