INNOVATIVE 3D MMO SERVERS ARCHITECTURES BASED ON GPGPU

ASAVEI, V[ictor]; MOLDOVEANU, A[lin] D[ragos] B[ogdan]; MOLDOVEANU, F[lorica]; BOIANGIU, C[ostin] A[nton]; MORAR, A[nca] - A[ndreea] & EGNER, A[lexandru] I[onut]

Abstract: Virtual spaces are becoming more and more present as a medium of social interaction on the internet. The biggest challenge for a massively multiuser online (MMO) application is to be able to accommodate a large number of users while maintaining a high degree of immersion for the users of the virtual space. Current 3D MMO Servers architectures try to solve this challenge by using traditional models such client-server or peer-to-peer but suffer from limitations regarding performance, fault tolerance and cost. In this paper we propose an innovative 3D MMO Server Architecture that uses GPGPU (General Purpose programming on Graphical Processing Units) in order to solve some of these limitations.
Key words: MMO, Virtual Reality, Parallel Processing

1. INTRODUCTION

Massively Multiuser Online (MMO) applications are becoming increasingly popular and the number of worldwide users for such applications is growing permanently. A user of a MMO application assumes the role of a fictive character and then is able to control its actions. The fictive character represents the user "avatar". MMO applications offer their users the possibility to accomplish tasks with well defined objectives. In contrast with single-user applications, in MMOs the users can do the tasks either collaborating or in a competitive manner against other users.

As examples of types of virtual 3D MMO spaces we can mention: real world simulations (e.g.: *Second Life*), e-learning applications, edutainment (education + entertainment), military simulators for training, etc.

Also another aspect to take into consideration is the fact that virtual spaces have become very attractive also from a financial viewpoint, for example the MMO *World of Warcraft* has a user base that exceeds 10 million of players and annual revenue of more than 1 billion dollars (Carless, 2009).

This growing popularity of 3D MMO virtual spaces has lead inevitably to an increase of the number of users that current 3D MMO servers architectures need to accommodate at the same time. The techniques that are used to solve this challenge by the traditional 3D MMO server architectures have drawbacks regarding scalability, reliability, redundancy, cost and the realism of the virtual world.

In this paper we first present the current architectures, identify the limitations introduced by them and finally propose a new architecture that tries to solve or at least reduce these limitations.

2. CURRENT SOLUTIONS

An important aspect of a MMO is the fact that the virtual worlds usually span over a large geographical area that cannot be maintained by a single server. This has lead to the development of methods that split the virtual world into regions and sub-regions that can be managed more easily by a single server.

Currently there are two distinct methods to obtain such a separation:
- Zoning
- Instancing

Zoning splits the virtual world into several distinct sub-spaces that can be assigned to a server or a group of servers to manage.

Instancing represents a special area that creates a copy of itself, hence the name, for a group of users that wish to access it. The respective group of users will access the copy of the zone practically being isolated from the other users of the virtual space because the content and state of the instance is visible only in the context of the group of users that accessed it.

Usually, the majority of MMO applications use client-server architecture. The clients access the virtual space through a *connection server* that after the process of authentication redirects them to a *shard server*.

The shards are independent versions of the same virtual world that run in parallel. This method of splitting the user base is used to achieve some degree of scalability. Usually the shards are not synchronized and users that are on different shards cannot interact. This fact contributes to a decrease in the realism of the simulated virtual world.

Some MMO applications such as for example *Eve Online* (Eve Online, 2003), try to implement "shard-less" solutions that don't split the user base and to allow all the users to be inside of a single huge virtual world. For the moment such approaches are very strongly dependent of the characteristics of the virtual world they try to simulate.

Although Client-Server solutions have inherently advantages such as security and centralized control they also introduce the following important disadvantages:
- Cost : because a large number of servers is used for the simulation of the virtual world, the maintenance costs after the launch can reach up to 80% of the revenues (Fan et al., 2007)
- Reliability : servers can be points of failure in the whole system
- Scalability : in order to achieve horizontal-scalability a large number of servers is used to manage the virtual world but this approach after a certain point can also represent a performance bottleneck

MMO applications on a reduced scale usually use P2P approaches. The Peer-To-Peer model offer several inherent advantages thus attracting interest for research in how to use this approach for MMO applications (Chen & Muntz, 2006)

Reduced scale applications usually share between the participants the load for the simulation of the virtual medium. In this way, using a P2P approach, it makes possible for the system to use resources from the new users and to accommodate the supplementary load that these new participants put on the virtual space.

Another advantage of this model is given by the fact that if one of the resources fails the other participants can take over the tasks allocated to that resource and thus making the system

more robust. Also using a P2P model reduces the costs of maintenance for the owner of the virtual space.

Unfortunately, even if P2P architectures solve some of the limitations of the client-server model they also have drawbacks that hinder their adoption for large-scale MMO applications:

- P2P by itself has no mechanism to ensure the state persistence of the virtual world; when a user disconnects from the virtual medium all the resources that he provided, including storage data, become unavailable
- Te fact that there is no central mechanism of authority makes the update of the virtual space very difficult and also can lead to security problems
- The resources that are shared by the participants are heterogeneous each client machine having its own computational limit; without a simple and efficient load balancing system this could lead to an overload of the resources of the peers

3. OUR SOLUTION

In recent years, the computing hardware, CPU and GPU, has evolved dramatically regarding the architecture used and computing power (Breitbart, 2008).

The current GPU is a very versatile hardware that uses a parallel multi-core architecture. This architecture that includes GPUs, multi-core CPUs from Intel and AMD, CELL processors, differentiates from the classical CPU architecture in the following way: they are designed to prioritize operations that can be executed in parallel over a large quantity of data compared to single task operations that have low latency.

These techniques are not unique for the graphical processing units, but when compared with CPUs, the GPUs designs take these architectures to the extreme.

As an example, the NVIDIA GeForce GTX280 has a total of 240 Streaming Processors that operate at a frequency of 1.3 GHz, and has a peak rate of 933 GFLOPS. For comparison, a high-end Core2Quad processor from Intel, which operates at 3.0 GHz and has 4 cores, achieves a peak rate of 96 GFLOPS (Fathalian & Houston, 2008).

Our solution tries to achieve scalability, both vertical and horizontal, and to reduce the limitations of the traditional 3D MMO server architectures proposing a new architecture with the following innovative characteristics:

- Using a hybrid architecture by introducing P2P techniques where they are adequate in order to offload the workload of the core servers; in this way we try to keep the advantages of both models whilst reducing their inherent limitations imposed on the system
- Offload heavy computational operations from the CPUs of the servers and implement them as GPGPU programs
- Using an architecture that is event and task driven and that is centred around the location and actions of the user entities in the virtual world because they are the ones that consume computational power and not the data of the virtual world by itself
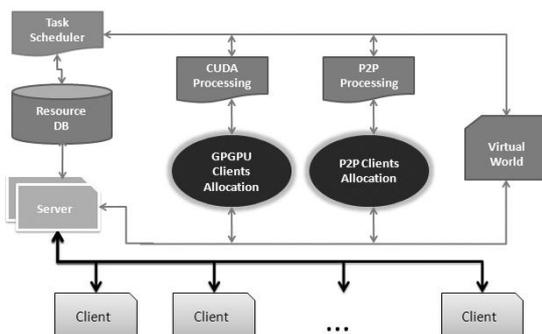

Fig. 1. 3D MMO Server Architecture

For our prototype we have used NVidia GPU hardware and CUDA (Compute Unified Device Architecture) to implement GPGPU operations. The following operations have been implemented as GPGPU programs:

- World Physics : collision detection and computations regarding forces (gravity and ground collision)
- Zoning computations using a spatial (based on the locations of the users in the virtual world) heuristic allocation of resources

Also, the direct communication between user entities has been successfully implemented using a P2P method.

The following important modules have been implemented in order to execute and schedule the GPGPU computations and P2P operations in the virtual world (see Fig. 1):

- Task scheduler module : responsible with the creation, scheduling and propagation of the CUDA and P2P tasks
- CUDA processing module : responsible with the execution of the GPGPU programs on the graphics hardware
- P2P processing module : responsible with the management of the P2P resources shared by the clients and the execution of the P2P on these resources

4. CONCLUSION AND FUTURE WORK

Preliminary tests based on a prototype that uses the architecture proposed have provided encouraging results proving that it is feasible to offload heavy computational operations from the virtual world servers as GPGPU programs. In terms of future scalability, taking into account the current evolution of GPUs, using GPGPU programs for 3D MMO servers can achieve an important degree of vertical scalability.

Also using P2P techniques is possible in order to move certain operations from the virtual world servers directly to the client involved in these operations.

As future work, we need to further investigate how P2P techniques can be extended to be used in a large scale 3D MMO application not only for direct communication messages between the clients but also for other classes of operations. We also need to develop mechanisms and protocols to ensure the security of operations that are managed using P2P.

Also as future work regarding operations that can be offloaded as GPGPU programs and given the fact that usually in a MMO application there is a large number of computer controlled entities we can take into consideration the possibility to implement basic AI operations as GPGPU programs.

5. REFERENCES

Breitbart J. (2008). Case studies on GPU usage and data structure design. *Dept. of Computer Science and Electrical Engineering*, Universitat Kassel, 2008

Carless, S. (2009). Interview: Screen Digest On Subscription MMO Growth, Blizzard's Next, *Available from:* http://www.gamasutra.com/view/news/23003/Interview_Screen_Digest_On_Subscription_MMO_Growth_Blizzards_Next.php *Accessed:* 2010-03-15

Chen, A. & Muntz R. (2007). Peer clustering: a hybrid approach to distributed virtual environments, *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*, Article No. 11, ISBN: 1-59593-589-4, Singapore, ACM, New York

Fan, L.; Taylor H. & Trinder P. (2007). Mediator: a design framework for P2P MMOGs, *Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games*,pp.43-48,ISBN:978-0-9804460-0-5,ACM,New York

Fathalian K. & Houston M. (2008). A closer look at GPUs. *Communications of the ACM, ACM Press*, 2008

*** (2003) http://www.eveonline.com – Eve Online, *Accessed on: 2010-03-1*