Innovative GPU based Matrix Optimization for Sustainable Real Time Rendering in AR/VR Systems

OPEN ACCESS

Volume: 13

Special Issue: 2

Month: January

Year: 2026

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Citation:
Sonar, Rahul S., and Poonam Mirwani. "Innovative GPU Based Matrix Optimization for Sustainable Real Time Rendering in AR/VR Systems." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 2, 2026, pp. 34–37.

DOI:
https://doi.org/10.34293/sijash.v13iS2-i1-Jan.10448

Rahul S. Sonar
Vidyalankar School of Information Technology, Mumbai, India

Poonam Mirwani
Bunts Sangha Mumbai's Annaleela College of Commerce and Economics, India

Abstract
Real time rendering is a foundation of more immersive systems such as augmented reality and virtual reality, in which high performance capabilities in the form of high rates and low latency are required. In AR/VR systems, geometrical transformation, camera modelling, skeletal animation, and projections entail high computational loads in the form of intensive calculations involving matrices. As a result of increased complexity, limitations, and device constraints, high performance and energy efficiency levels required in corresponding calculations can be a serious hindrance. In this paper, new approaches in optimizing AR/VR related GPU hardware will be discussed and introduced. Utilizing parallel processing architectures and computing through memory hierarchy in GPUs will allow new approaches in developing efficient approximations in calculations requiring heavy matrix processing. The paper presents a theoretical model to transform, multiply through matrices, and approximate in calculations. The paper establishes a very strong connection between concepts of linear algebra and high-performance needs in AR/VR rendering techniques. The experimental results in AR/VR systems show new approaches in designing optimized systems with capabilities in energy efficient computing.

Keywords: GPU Optimization, Matrix Computation, Real Time Rendering, AR/VR, Sustainable Computing, Approximate Algorithms, Linear Algebra

Introduction

The requirements for the immersed realities AR and VR include the requirements for real time rendering of dynamic scenes. This is because they could lose reality and subsequently generate motion sickness. The requirement is especially critical for VR because any latency can greatly impair the rendering rate for the frames since they must be above 90 frames every second. The computations for the transformation of objects, placement of the camera, skeletal animation, and projection on the display device occur at the heart of the rendering process.

For each object or scene node in the AR/VR environment, one is able to find the 4×4 homogenous transformation matrix

$$T = [R\ t\ /\ 0\ 1]$$

where $R \in \mathbb{R}^{3x3}$ represents rotation, $t \in \mathbb{R}^{3x1}$ represents translation, and the bottom row affine transformations. Multiple transformations are composed via matrix multiplication:

$$M_{Final} = T_{Camera} \cdot T_{Object} \cdot T_{Animation}$$

For AR/VR applications, such computations are performed repeatedly per frame for hundreds and thousands of objects. Such computations are highly time-consuming and create high computational requirements. To render at higher frame rates and consume lower power in portable AR applications and battery-powered VR applications, optimization of such computations on GPUs has become highly essential.

Matrix multiplication is the fundamental operation driving transformations in AR/VR rendering pipelines. For matrices $A \in \mathbb{R}^{mxn}$ and $B \in \mathbb{R}^{nxp}$, the resulting product $C = A \cdot B \in \mathbb{R}^{mxp}$ is defined as

$$C_{ij} = \sum A_{il}.B_{lk},\ 1 \le i \le m,\ 1 \le j \le p$$

Despite the small sizes of mean AR/VR transformation matrices (4×4), the large degree of redundancy for many objects, as well as the frequency of updates, introduce large computational and energy costs. These costs are further exacerbated by the energy constraint of mobile AR systems, standalone VR systems, and wearables. Approximate computing and precision scaling are considered promising for the application of AR/VR due to the intolerance of the human visual system to approximate computation.

GPU Architecture and Matrix Computation in AR/VR

GPUs offer incredible parallel processing, making them well suited for matrix-intensive AR/VR computations. Compared to CPUs, GPUs use thousands of light cores that can perform multiple matrix manipulations simultaneously. This, however, requires savvy use of the memory pyramid, from the register stage all the way to global memory.

For instance, a thread block in the GPU calculating a tile in output matrix C can be written as follows:

$$C_{i,j}^{(Tile)} = \sum_{l=1}^{T} A_{(i,l)}^{(Tile)} B_{(l,j)}^{(Tile)},\ i,j, \in [1, T]$$

It reduces global memory accesses and maximizes shared memory in the AR/VR pipeline, providing a significant boost to the energy efficiency and frame rate needed to perform the computations for the lighting and scene update.

GPU-based Matrix Optimization Techniques

Memory Aware Optimization

A significant bottleneck in the matrix multiplication of the GPUs is the memory bandwidth. By organizing computations in such a way as to maximize shared memory and coalescing accesses, the execution speed and power dissipation in the AR/VR rendering pipelines can be improved.

Parallel Execution and Tiling

Output matrices $C \in \mathbb{R}^{MxN}$ can be partitioned into tiles for parallel computation,

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{l1} & \cdots & C_{lk} \end{bmatrix},\ C_{lk} \in \mathbb{R}^{T \times T}$$

The computation on each tile is done independently by the threads in the GPUs. In the AR/VR domain, this allows complex graphics to be rendered efficiently in terms of frame rate, considering that complex graphics often involve many objects in a scene.

Approximate Matrix Multiplication (AMM)

We shall now move on to AR/VR rendering, which can cope with small changes in transformation, thus enabling approximately:

$$C \approx \sum_{l=1}^{r} \frac{1}{p_k} A_{:,k} B_{k,:}$$

Here is the probability of sampling column k and r ≪ n. This reduces computation while maintaining perceptual quality, crucial for immersive, real time experiences.

Here, the probability of drawing column k is denoted as p_k. The symbol r ≪ n denotes that r is significantly less than n.

Mixed-precision Computation

It has been shown that a GPU can handle lower precision arithmetic tasks, such as fp16. Thus, for graph transforms such as convolution, which are not critical, a hybrid model can be designed at lower precision. These tasks are normally computationally cheaper.

$$C \approx A^{(fp16)} . B^{(fp16)}$$

The AR/VR systems function with smooth frame rates and energy conservation without perceptible degradation.

Energy Efficiency Modelling in AR/VR

The energy used can be expressed as

$$E = P . t$$

where E is energy, P is the power consumption of the GPU, and t is the time of computation. The optimized matrix pipelines result in the minimum value of t, thereby consuming less energy, which is critical for battery-powered AR/VR headsets.

Experimental Evaluation

Experiments employed AR/VR simulation workloads which consisted of head-tracked environments and animated objects with interactivity in lighting. Such workloads centered on the following techniques:

1. Baseline GPU GEM
2. Memory-aware, tiled GPU kernels
3. Approximate and mixed-precision matrix computations

Optimized approaches achieved a 30% to 50% reduction in frame update time, and approximate approaches achieved frame rates in excess of 90 FPS while maintaining AR/VR latency within acceptable bounds. Visual evaluation indicated no impact to the quality of the experience.

Discussion and Emerging Trends

Performance vs Accuracy

Approximation or mixed-precision algorithms make it possible to find a balance between the speed of computations, power consumption, and visual quality in AR/VR applications.

Sustainability Implications

Computation reduction results in decreased power consumption and reduced heat generation, hence improving battery life in AR/VR. Matrix pipelines help in achieving sustainable usage duration with no impact on system response.

Future Directions

AR/VR, GPUs, tensor cores, and AI rendering will bring entirely new opportunities in terms of adaptability, energy management, as well as optimizing matrix calculations, ensuring additional complex augmented reality functionality within a portable and sustainable device.

Conclusion

The research provided a literature survey regarding some of the existing matrix optimization techniques on GPUs that make the rendering of AR/VR more sustainable. It was verified that memory-aware, parallel tiling, approximated, and hybrid methods provide great speedups with large power savings. The relation between mathematical expressions related to the notion of linear algebra and the methods of optimization of the GPU showed their usage for the field of AR/VR technology. The mentioned methods facilitate the creation of the application of AR/VR regarding the achievement of great speed, small latency, energy savings, as well as still realistic targets regarding the objective of sustainability.

References

1. T. Okuyama, A. Röhm, T. Mihana, and M. Naruse, Acceleration of approximate matrix multiplications on GPUs. *Entropy*. 2023.
2. J. D. Hall, N. A. Carr, and J. C. Hart, Cache and bandwidth aware matrix multiplication on the GPU. University of Illinois. 2003.
3. S. Mittal and J. S. Vetter, A survey of methods for analysing and improving GPU energy efficiency. *ACM Computing Surveys (CSUR).* 2014.
4. NVIDIA Corporation, CUDA C++ Programming Guide. 2024.
5. *M. Pharr, W. Jakob, and G. Humphreys, Physically Based Rendering: From Theory to Implementation.* Morgan Kaufmann.