Understanding GPU performance e.g. using MLCommons MLBenchmarks

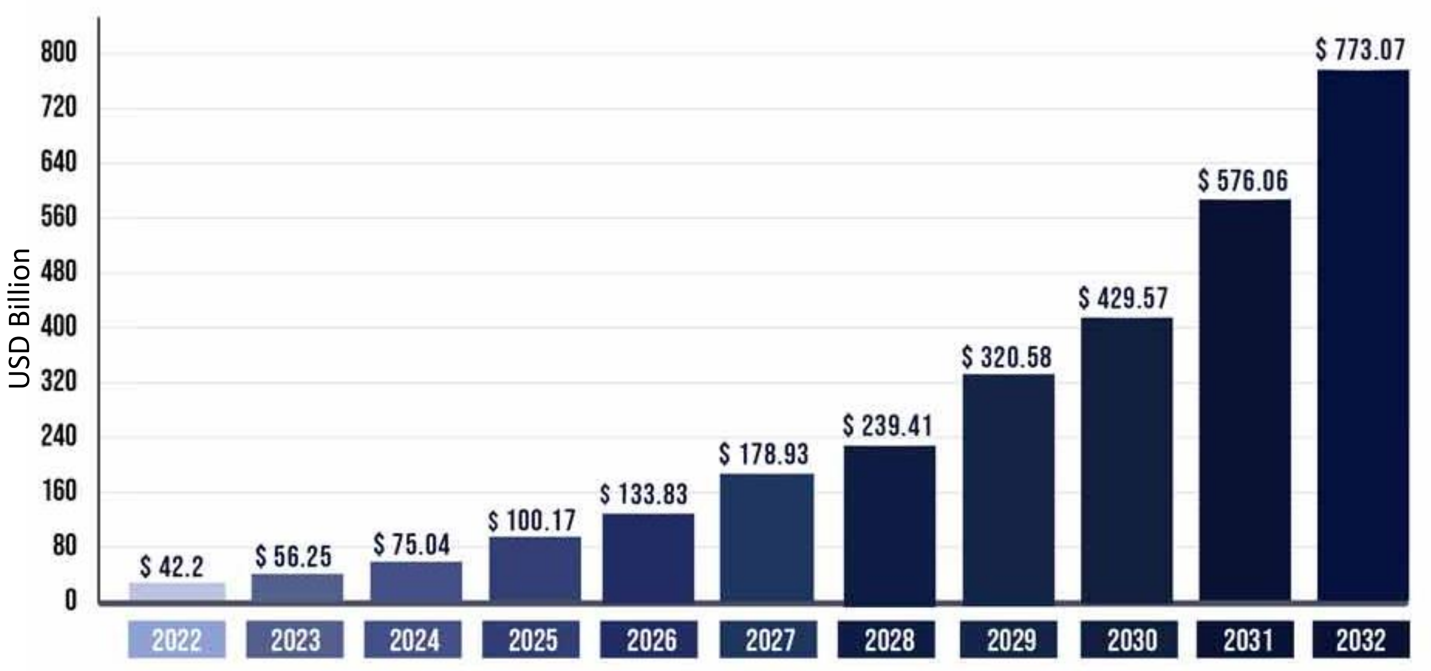
Raza, Ossama Bin

Supervisor: Chirag Mandal

Table of Contents

Introduction to GPUs [4]

- Core Specs
- Benchmarks
- Throughput
- Bandwidth
- Efficiency
- Architecture



GPU Market size, 2022 to 2032 (estimated)

Graph source: [GPU Market, Graphic Processing Unit Market Size 2023-2032 - Precedence Research]

GPU Applications [1, 4]

- Rendering Video Game Graphics
- Scientific Simulations
- Machine Learning
- Cryptocurrency Mining
- Professional Visualization

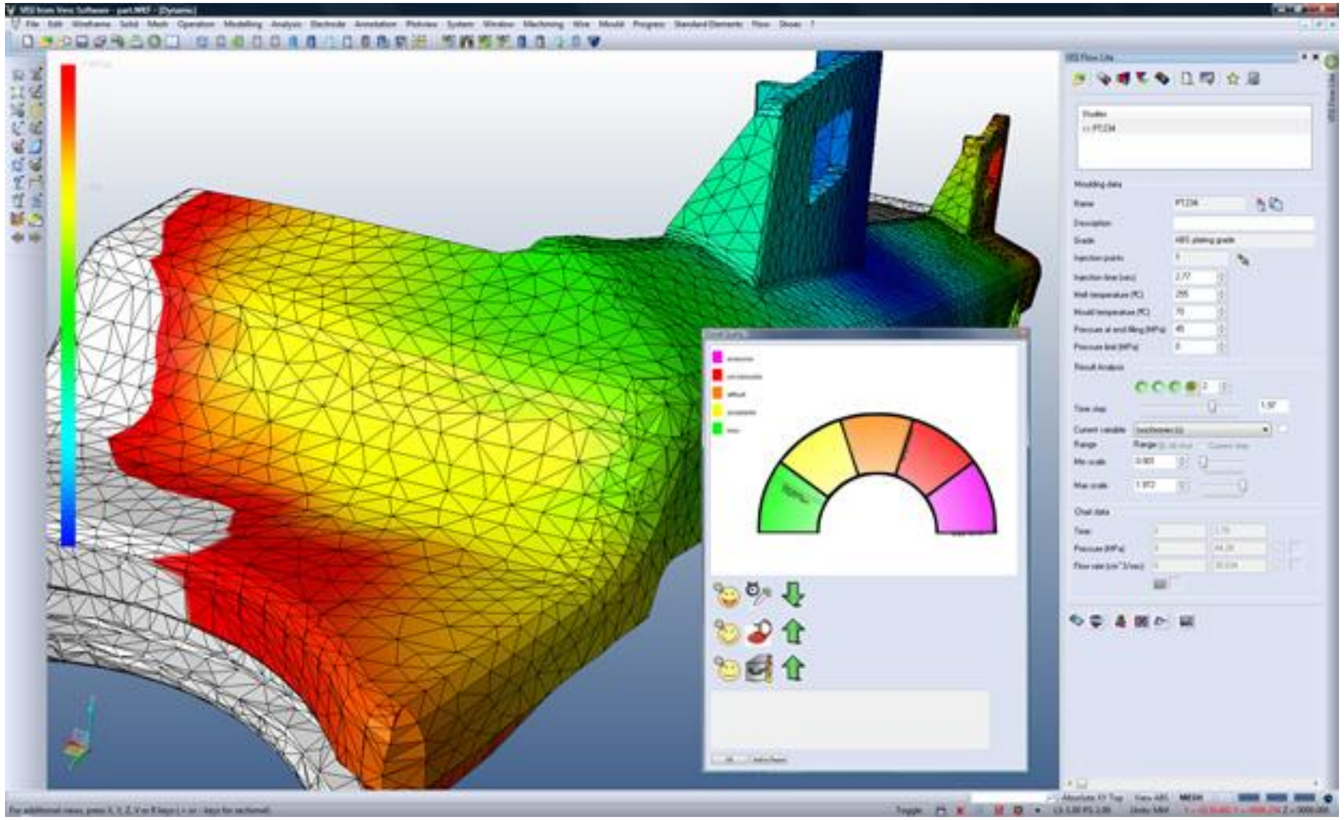


Image source: The Artful Science of Mold Simulation - Digital Engineering 24/7

Table of Contents

Measuring GPU Performance [4]

- Data Transfer Speed
- Read/Write Speed
- Computation Speed

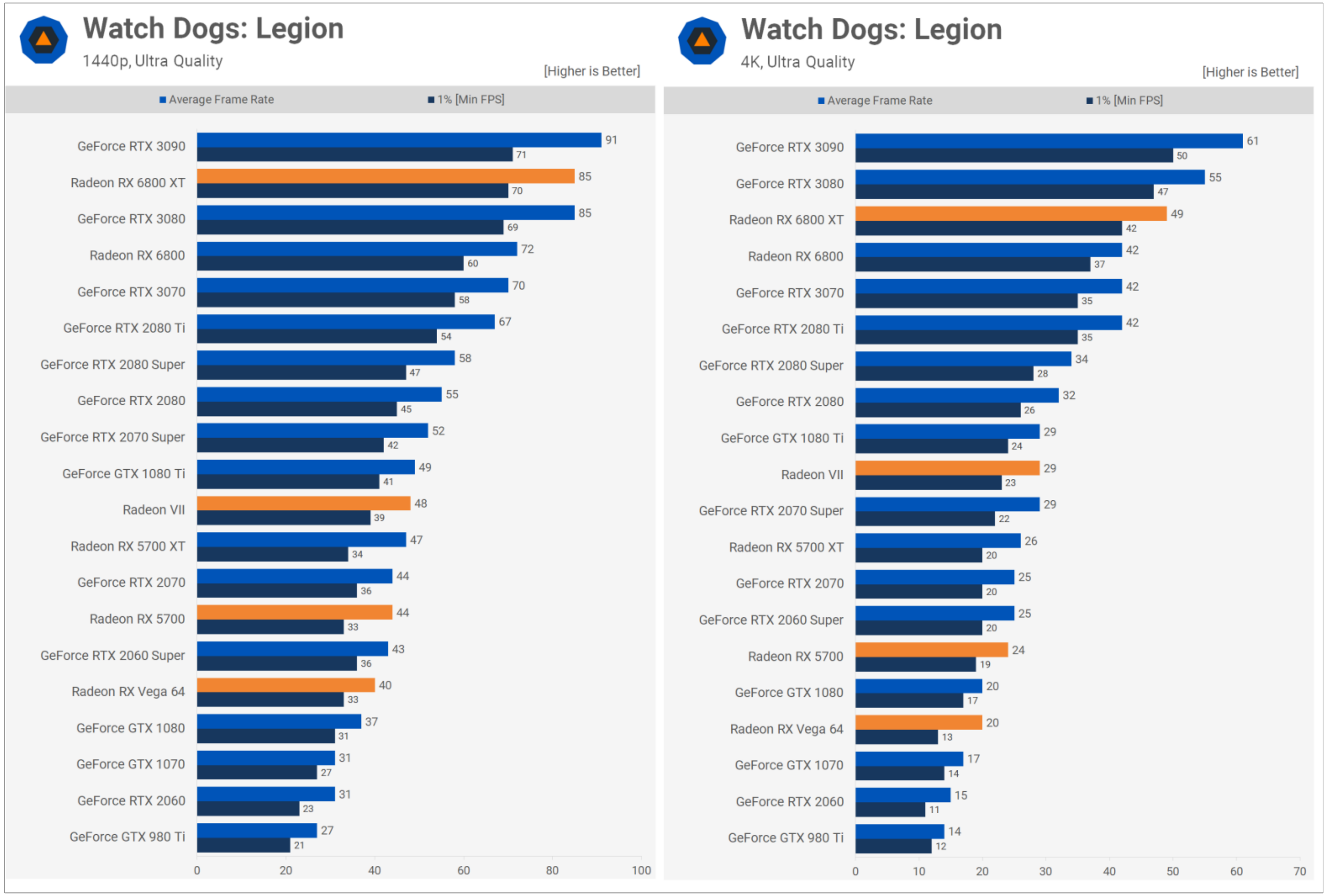


Image Source: [AMD Radeon RX 6800 Review - TechSpot]

Available GPU Benchmarks [13, 14]

- 3DMark
- Superposition
- Cinebench 2024
- FurMark
- In-game benchmarks
- PassMark Software
- MLCommons

Image source: Julian M. Kunkel – HPDA Slides

MLCommons Benchmark Categories [9]

- AI Safety Benchmarks
- MLPerf Training
- Scientific MLPerf Inference: Mobile
- Machine MLPerf Training: HPC
- Cryptocurrency MLPerf Inference: Tiny
- MLPerf Inference: Datacenter
- MLPerf Storage
- MLPerf Inference: Edge

Table of Contents

Intro - MLPerf™ HPC Overview [8]

- Benchmark Suite:
  - Climate Segmentation (CAM5+TECA)
  - Cosmological Parameter Prediction (CosmoFlow)
  - Catalyst Modeling (Open Catalyst 2020)
  - Protein Structure Prediction (OpenFold)

- Key Metrics:
  - Time to Solution(TTS)
  - Throughput(optional)

Reference: Benchmark MLPerf Training: HPC | MLCommons V2.0 Results

MLPerf™ HPC Overview [3]

- Data Handling
  - Data can start on any durable storage (excluding RAM) as of v3.0
- Submission Requirements
  - TTS in every submission
  - Power measurements optional but encouraged
- Minimum runs per benchmark

Benchmark	Min. Runs
DeepCAM	5
OpenCatalyst	5
CosmoFlow	10
OpenFold	10

Reference: Benchmark MLPerf Training: HPC | MLCommons V2.0 Results

Closed Division Vs Open Division

- Closed Division:
  - Standardized Settings
  - Restricted Hyperparameters and Optimizers

To create a level playing field

- Open Division:
  - Flexibility in Implementation
  - Unrestricted Hyperparameters and Optimizers

Encourages innovation and optimization

Reference: MLCommons: MLPerf™ HPC Training Rules | Github

Problems Benchmarking HPC

- Requires all available resources
- Access restrictions
- Other work needs to be put on hold
    - Creates backlog
- Scale Adjustment
- Compliance and Validation
- Documentation

Index of /project/dasrepo/cosmoflow-benchmark

Name	Last modified	Size	Description
Parent Directory		-	
cosmoUniverse_2019_05_4parE_tf_v2.tar	2021-03-17 08:11	1.6T	
cosmoUniverse_2019_05_4parE_tf_v2_mini.tar	2023-03-28 23:32	5.5G	

Reference: CosmoFlow Datasets (nersc.gov)

Table of Contents

MLPerf Training [3]

- Relatively small scaled
- Not optimized
- Works using docker file
- Cover various domains
- Frameworks include
  - TensorFlow
  - PyTorch
  - TorchRec

model	reference implementation	framework
resnet50v1.5	vision/classification_and_detection	tensorflow2
RetinaNet	vision/object detection	pytorch
3DUnet	vision/image segmentation	pytorch
Stable Diffusionv2	image generation	pytorch
BERT-large	language/nlp	tensorflow
GPT3	language/llm	paxml,megatron-lm
LLama2 70B-LoRA	language/LLM fine-tuning	pytorch
DLRMv2	recommendation	torchrec
RGAT	GNN	pytorch

Table source: MLPerf™ Training Reference Implementations v4

MLPerf Training (Benchmarks)

Area	Benchmark	Dataset	Quality Target	Reference Implementation Model	Latest Version Available
Vision	Image classification	ImageNet	75.90% classification	ResNet-50 v1.5	v4.0
Vision	Image segmentation (medical)	KiTS19	0.908 Mean DICE score	3D U-Net	v4.0
Vision	Object detection (light weight)	Open Images	34.0% mAP	RetinaNet	v4.0
Language	NLP	Wikipedia 2020/01/01	0.72 Mask-LM accuracy	BERT-large	v4.0
Language	LLM	C4	2.69 log perplexity	GPT3	v4.0
Language	LLM finetuning	GovRep r1/r2/r3	ROUGE score	Llama 2 70B	v4.0
Commerce	Recommendation	Criteo 4TB multi-hot	0.8032 AUC	DLRM-dcnv2	v4.0
Marketing, Art, Gaming	Image Generation	LAION-400M-filtered	FID<=90 and CLIP>=0.15	Stable Diffusionv2	v4.0

Table source: Benchmark MLPerf Training | MLCommons Version 2.0 Results

MLPerf Training (Benchmarks)

Area	Benchmark	Dataset	Quality Target	Reference Implementation Model	Latest Version Available
Marketing, Art, Gaming	Image Generation	LAION-400M-filtered	FID<=90 and CLIP>=0.15	Stable Diffusionv2	v4.0
Graph neural network	Graph neural network (GNN)*	IGBH-Full	72% classification accuracy	R-GAT	v4.0
Vision	Object detection (heavy weight)	COCO	0.377 Box min AP and 0.339 Mask min AP	Mask R-CNN	v3.1
Language	Speech recognition	LibriSpeech	0.058 Word Error Rate	RNN-T	v3.1
Commerce	Recommendation	1TB Click Logs	0.8025 AUC	DLRM	v2.1
Research	Reinforcement learning	Go	50% win rate vs. checkpoint	Mini Go (based on Alpha Go paper)	v2.1
Vision	Object detection (light weight)	COCO	23.0% mAP	SSD	v1.1
Language	Translation (recurrent)	WMT English-German	24.0 Sacre BLEU	NMT	v0.7
Language	Translation (non-recurrent)	WMT English-German	25.00 BLEU	Transformer	v0.7

Table source: Benchmark MLPerf Training | MLCommons Version 2.0 Results

Table of Contents

Challenges in GPU Performance Measurement

- Complexity of Applications
- Synchronization and Memory Transfers
- Variability in Workloads
- Hardware and Software Variations
- Benchmarking Limitations

Benchmarketing [10, 11, 12]

- Cherry-Picking Benchmarks
  - Products shown in the best light
- Over-Optimization for Benchmarks
- Misleading Benchmarking Practices
  - Outdated benchmarks
  - Inappropriate workloads
  - Unfair comparisons
- Lack of Transparency
  - Not provide information about their benchmarking methodologies

Benchmarketing – Example [5, 6]

- AMD Radeon RX 6000 series
  - Released in November 2020
  - Based on the new RDNA2 architecture
  - Performance benchmarked on SD2.1
  - Previous series (RX 5000) benchmarked on SD1.5
  - Promised a 1.65x performance per watt gain over RX5000
- Led to numerous controversies and AMD being publicly questioned

Benchmarketing – Example [15, 16, 17]

- Nvidia's GeForce RTX 4090 graphics cards
    - Released in October 2020
    - Melting wires in the 16 pin 12VHPWR power connector adapter
    - Approximately 20 consumers reported this
    - Lawsuit seeking class-action status and was filed by Lucas Genova
- The lawsuit was dismissed
- Potential settlements and reasons of dismissal undisclosed

GPU bottlenecks Prevention

- CPU-GPU Balance
- Memory Access Patterns
- Parallel Scalability
- Data Transfer Speeds
- VRAM Limitations
- GPU Utilization
- Hardware Compatibility

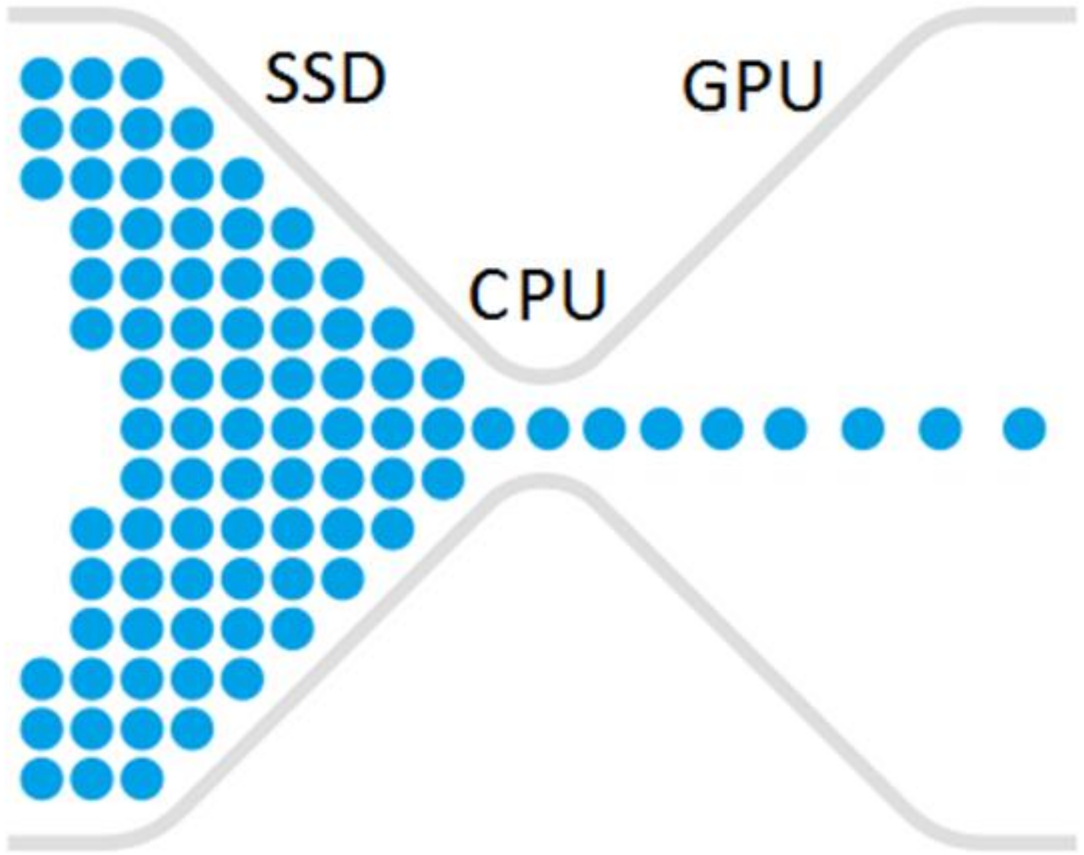


Image source: Intel i5 Bottlenecking GTX 1080 - Quora

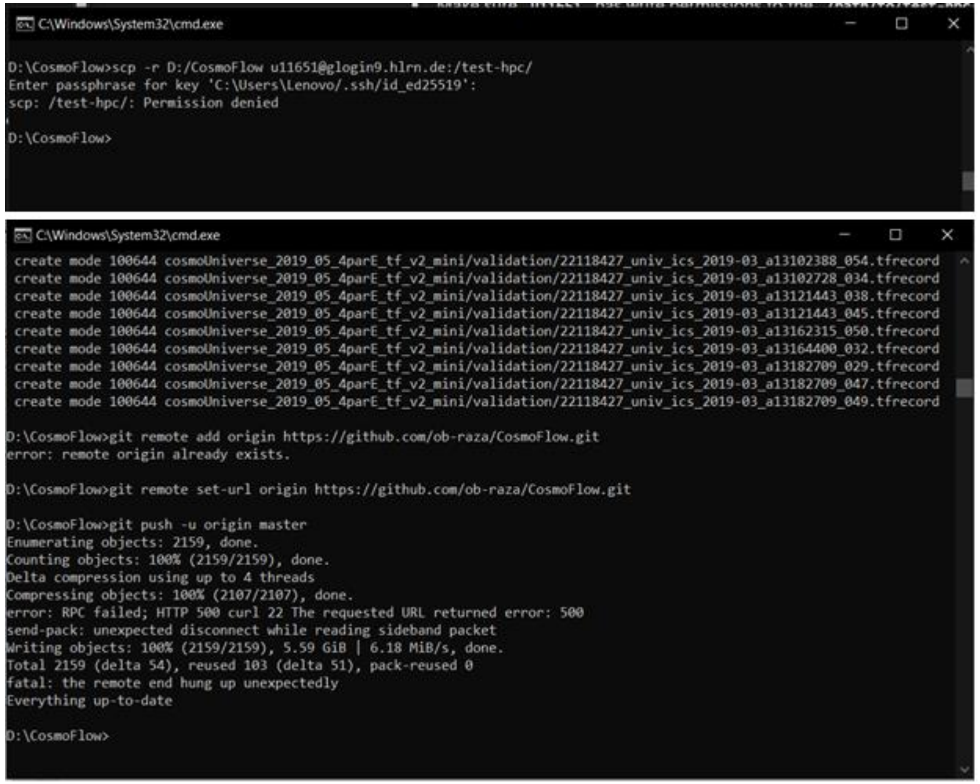
Table of Contents

Future Trends in GPU Technology [19]

- AI and Machine Learning Integration
- New GPU Architectures e.g Nvidia's Hopper architecture
- Ray Tracing Technology
- Enhanced VR and AR Experiences
- Energy Efficiency and Sustainability
- The Rise of Cloud Gaming
- Custom GPUs for Specific Workloads
- Advancements in Rendering Techniques

Implementation: Progress and Problems

- Progress:
  - Working grete: shared and grete:interactive
  - Working Slurm script
  - Acquired benchmarks and datasets
- Problems:
  - Resource allocation delays
  - Github data upload error
  - Direct data upload error



Conclusion and Futurework

- Benchmark analysis
- Conducting a more in depth literature review
- Practical implementation
    - Troubleshoot existing problems
    - Run MLPerf Training benchmarks on different GPUs
    - Discuss finding in the final report

References

[1] Benchmarking TPU, GPU, and CPU Platforms for Deep Learning [https://arxiv.org/abs/1907.10701]

[2] MLPerf Training Benchmark (last revised 2 Mar 2020 (this version, v3)) [https://arxiv.org/abs/1910.01500]

[3] MLPerfTM HPC: A Holistic Benchmark Suite for Scientific Machine Learning on HPC Systems [https://arxiv.org/pdf/2110.11466]

[4] Performance and Scalability of GPU-Based Convolutional Neural Networks [https://ieeexplore.ieee.org/document/5452452]

[5] https://www.tomshardware.com/reviews/gpu-hierarchy,4388.html

[6] https://www.videocardbenchmark.net/high_end_gpus.html

[7] https://mlcommons.org/benchmarks/training-hpc/

[8] https://mlcommons.org/benchmarks/training/

[9] https://github.com/mlcommons

[10] https://pythonspeed.com/articles/gpu-vs-cpu/

[11] https://www.pcgamer.com/hardware/processors/pay-no-attention-to-amds-horribly-misleading-benchmarks-for-its-new-ryzen-5000-xt-cpus/

[12] https://www.digitaltrends.com/computing/gpu-benchmarks-mislead-gpu-upgrade/

[13] https://www.gearprimer.com/technology/best-pc-benchmark-tools/

[14] https://www.tomshardware.com/pc-components/gpus/stable-diffusion-benchmarks

[15] https://www.theregister.com/2022/11/18/nvidia_flawsuit_4090/

[16] https://www.pcgamer.com/nvidia-hit-with-class-action-suit-over-melting-rtx-4090-gpu-adapters/

[17] https://uk.pcmag.com/graphics-cards/143891/nvidia-faces-class-action-lawsuit-over-melting-12vhpwr-cables

[18] The Peak Performance Analysis Method for Optimizing Any GPU Workload [https://developer.nvidia.com/blog/the-peak-performance-analysis-method-for-optimizing-any-gpu-workload]

[19] The Peak Performance Analysis Method for Optimizing Any GPU Workload [https://developer.nvidia.com/blog/the-peak-performance-analysis-method-for-optimizing-any-gpu-workload]