NVIDIA L40S GPU: A Critical Analysis for Generative AI Workloads

Generative AI, a revolutionary field in artificial intelligence, focuses on creating new content, such as text, images, music, and even code. This rapidly evolving technology relies heavily on powerful GPUs to handle the complex computations involved in generating such content. NVIDIA, a leading manufacturer of GPUs, has introduced the L40S to address the growing demands of Generative AI workloads. This article provides a critical analysis of the NVIDIA L40S GPU's performance in this domain.

NVIDIA L40S Specifications and Features

The NVIDIA L40S, built on the NVIDIA Ada Lovelace architecture, is specifically designed for enterprise data centers seeking to accelerate their Generative AI applications. Here's a closer look at its key specifications and features 1:

Feature	Description
Architecture	NVIDIA Ada Lovelace Architecture
Process Size	4nm NVIDIA Custom Process (TSMC)
Transistors	76.3 Billion
Die Size	608.44 mm2
CUDA Cores	18176
Tensor Cores	568 (Gen 4)
RT Cores	142 (Gen 3)
GPU Memory	48 GB GDDR6 with ECC
Memory Interface	384-bit
Memory Bandwidth	1,024 GB/s
Display Connectors	4x DP 1.4a
Maximum Digital Resolution	4x 5K at 60 Hz, 2x 8K at 60 Hz, 4x 4K at 120 Hz (30-bit Color)
Form Factor	4.4" H x 10.5" L, Dual Slot
Thermal Solution	Passive
Maximum Power Consumption	300 W

vGPU Software Support	NVIDIA vApps, vPC, vWS
vGPU Profiles Supported	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 16 GB, 24 GB, 48 GB
Graphics APIs	DirectX 12 Ultimate, Shader Model 6.6, OpenGL 4.6, Vulkan 1.3
NVENC / NVDEC	3x ENC / 3x DEC (Includes AV1 Encode and Decode)
Compute APIs	CUDA 12.0, DirectCompute, OpenCL 3.0
NVIDIA 3D Vision and 3D Vision Pro	Support via Optional 3-pin mini-DIN Bracket
Frame Lock	Supported with optional NVIDIA Quadro Sync II
Power Connector	1x PCIe CEM5 16-pin
NEBS Ready	Level 3
Secure Boot with Root of Trust	Supported

The L40S boasts a substantial 48 GB of GDDR6 memory with Error Correction Code (ECC) to ensure data integrity. ECC is crucial for AI workloads, where even minor data errors can significantly impact the accuracy of results. The inclusion of 4th generation Tensor Cores, specialized processing units designed to accelerate AI operations, is a key feature. These Tensor Cores, combined with a dedicated FP8 Transformer Engine, contribute significantly to the L40S's performance in Generative AI tasks.

Furthermore, the L40S supports a wide range of vGPU profiles 2 (see the(https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/Virtual-GPU-Packaging-and-Licensing-Guide.pdf) for details). This allows for flexible allocation of GPU resources to multiple users, enabling efficient utilization in virtualized environments.

L40S Performance in Different AI Tasks

The NVIDIA L40S demonstrates its capabilities across various AI tasks, including training, fine-tuning, and inference. Here's a breakdown of its performance in each area:

Training

While the H100 reigns supreme in raw performance for training complex AI models from scratch 3, the L40S still delivers respectable performance, especially when considering its price-to-performance ratio 3. In generative AI model training, the L40S exhibits 1.2 times the performance of the A100 GPU when running Stable Diffusion 3. It can accelerate the training of foundational models with up to 175 billion parameters and enhance the fine-tuning and retraining of existing large-scale models 15.

Fine-tuning

Fine-tuning involves adjusting a pre-trained AI model to perform a specific task. The L40S excels in this area, leveraging its Ada Lovelace Tensor Cores and FP8 Transformer Engine 3. These architectural advancements enable efficient fine-tuning of large language models (LLMs) and other generative AI models 15.

Inference

Inference refers to using a trained AI model to make predictions or generate new content. The L40S truly shines in inference tasks, achieving up to 1.5x greater inference performance than the A100 GPU 3. This is attributed to its efficient use of FP8 precision for AI inference 3, which allows for faster processing and reduced memory consumption without significant loss of accuracy.

Benchmark Results for Generative AI Workloads

While comprehensive independent benchmarks are still emerging, NVIDIA's initial data suggests the L40S delivers substantial performance improvements compared to the previous generation A100 GPU. Notably, it offers:

- **Up to 1.2x faster generative AI inference performance** 4
- **Up to 1.7x faster training performance** 4

These gains are attributed to the advancements in the Ada Lovelace architecture, particularly the enhanced Tensor Cores and the FP8 Transformer Engine.

The L40S demonstrates its prowess in various AI tasks, including:

- **Image processing:** The L40S efficiently handles complex image processing tasks, such as image recognition, object detection, and image generation5.
- **Data aggregation:** Its high memory capacity and processing power enable efficient aggregation and analysis of large datasets, crucial for training and fine-tuning AI models5.
- **Generative AI, including Large Language Model (LLM) inference:** The L40S excels in running Generative AI models, particularly for inference tasks involving LLMs. This is facilitated by its FP8 Transformer Engine, which accelerates the processing of transformer networks, a key component of many LLMs5.

Reviews and Analyses

Reviews and analyses of the L40S consistently highlight its versatility and strong performance in Generative AI applications. Here are some key observations:

- **Balanced Performance:** The L40S strikes a balance between AI inference performance and 3D rendering capabilities, making it suitable for a diverse range of workloads. This versatility allows it to handle tasks ranging from running AI models to generating high-quality graphics and processing video content7.
- **Cost-Effectiveness:** For smaller inference tasks and AI experimentation, the L40S presents a cost-effective alternative to the higher-priced H100 and A100 GPUs. This makes it an attractive option for organizations with budget constraints or those looking to explore Generative AI without committing to top-of-the-line hardware7.
- **Suitability for Specific Use Cases:** The L40S is particularly well-suited for specific use cases within Generative AI. These include running small to medium-sized AI models, graphics-intensive applications that leverage its rendering capabilities, and multimodal models that combine different data types, such as text and images7.

However, it's important to acknowledge that the L40S may not be the optimal choice for all scenarios. For instance, training highly complex AI models from scratch, which often demand the highest levels of precision and computational power, might be better suited for the H1003.

Comparison to Other GPUs

The L40S competes with other GPUs in the data center market, including NVIDIA's A100 and H100, as well as AMD's Instinct MI series. Here's a comparative overview:

GPU	Architecture	Memory	Performance	Key Advantages
NVIDIA L40S	Ada Lovelace	48GB GDDR6	Up to 1.2x faster inference than A100	Versatile, cost-effective, strong inference performance
NVIDIA A100	Ampere	80GB HBM2e	Strong for training and double-precision	Higher memory capacity, mature software ecosystem
NVIDIA H100	Hopper	80GB HBM3	Top-tier performance, especially for training	Highest performance, advanced features like Transformer Engine

- **NVIDIA A100:** The L40S generally demonstrates better performance in Generative AI inference tasks and offers improved FP32 performance for general computing, making it suitable for a wider range of applications5. However, the A100 might be preferable for tasks that require higher memory capacity or involve extensive double-precision operations5.

- **NVIDIA H100:** As NVIDIA's flagship AI GPU, the H100 delivers the highest performance, particularly for complex model training3. However, the L40S provides a more balanced approach, offering competitive performance at a more attractive price point for many Generative AI workloads3.
- **AMD Instinct MI Series:** While a direct comparison necessitates further benchmarking, the L40S competes with AMD's MI series in the high-performance computing arena. The choice between them depends on specific workload requirements, software ecosystem considerations, and overall system compatibility.

Detailed Comparison to RTX 6000 Ada

Snippet 9 raises important points about combining the L40S with other GPUs, particularly the RTX 6000 Ada. While both GPUs offer strong performance, there are potential compatibility and cooling considerations when using them together.

- **Compatibility:** Mixing GPUs from different generations, such as the L40S (Ada Lovelace) and RTX 6000 Ada, may lead to compatibility challenges. This can limit the utilization of features specific to the latest generation and potentially cause software issues.
- **Cooling:** Combining a data center GPU like the L40S with a workstation GPU like the RTX 6000 Ada can create cooling challenges. Data center GPUs typically rely on the data center's cooling infrastructure, while workstation GPUs often have active cooling solutions. This difference in cooling requirements needs careful consideration when integrating these GPUs into the same system.
- **Multi-GPU Scaling:** It's crucial to understand that multiple GPUs do not function as a single, larger GPU. While parallel tasks like ray tracing can benefit from multiple GPUs, other workloads may not scale linearly. Software optimization and explicit GPU affinity settings might be required to fully leverage the capabilities of multiple GPUs.

OpenCL Benchmark Comparison

The OpenCL benchmark results from 10 provide valuable insights into the L40S's performance relative to a wide range of GPUs, including those from AMD and Intel. Here's a summarized table of the results:

Device	Score
AMD Instinct MI300X	346795
NVIDIA L40S	345962
NVIDIA H100 80GB HBM3	335247
NVIDIA L40	331157
NVIDIA RTX 6000 Ada Generation	323789
NVIDIA GeForce RTX 4090	317442

These results show that the L40S achieves a very high score in the OpenCL benchmark, surpassing the NVIDIA H100 and RTX 6000 Ada, and closely trailing the AMD Instinct MI300X. This highlights the L40S's strong general-purpose computing capabilities and its competitiveness against other leading GPUs in the market.

Power Consumption and Thermal Performance

The NVIDIA L40S has a maximum power consumption of 300W 2. It utilizes a passive cooling solution, meaning it relies on the data center's cooling infrastructure for heat dissipation. This design contributes to the overall energy efficiency of the data center by reducing the need for individual GPU fans and minimizing power consumption.

Pricing and Availability

The NVIDIA L40S is readily available through various channels, including NVIDIA partners and system integrators 11. The price of the L40S can vary based on the vendor, configuration, and specific market conditions. Based on available information, the price generally ranges from approximately $6,100 to $9,750 1.

Strengths and Weaknesses for Generative AI

Strengths:

- **Excellent Inference Performance:** The L40S excels in Generative AI inference tasks, delivering significant performance improvements over its predecessor, the A100. This makes it a strong choice for running pre-trained Generative AI models and deploying them for various applications.
- **Versatile:** Its ability to handle a wide range of workloads, including AI inference, training, graphics rendering, and video processing, makes it a versatile option for data centers with diverse needs.
- **Cost-Effective:** The L40S provides a compelling balance of performance and price, making it an attractive option for organizations seeking to optimize their investment in AI infrastructure.

- **Energy Efficient:** With its passive cooling solution and optimized architecture, the L40S contributes to a more energy-efficient data center environment.

Weaknesses:

- **Not Ideal for Complex Model Training:** While capable of training AI models, the L40S may not be the absolute best choice for training the most complex AI models that require the highest levels of performance and precision. In such cases, the H100 might be a more suitable option.

Conclusion

The NVIDIA L40S emerges as a powerful and versatile GPU well-suited for a wide range of Generative AI workloads. Its strengths lie in its excellent inference performance, versatility, cost-effectiveness, and energy efficiency. While it may not be the ultimate solution for training the most demanding AI models, it offers a compelling combination of features and performance for many Generative AI applications.

As Generative AI continues to advance and find applications in diverse fields, such as natural language processing, image generation, drug discovery, and more, the L40S is poised to play a crucial role in driving innovation and progress. Its ability to efficiently run Generative AI models, combined with its versatility and cost-effectiveness, makes it an attractive option for organizations seeking to harness the power of this transformative technology.

Works cited

1. NVIDIA L40 & L40S Enterprise 48GB — Vipera - Tomorrow's Technology Today, accessed January 9, 2025, https://viperatech.com/shop/nvidia-l40-l40s-48g/
2. NVIDIA L40 GPU for Data Center | NVIDIA, accessed January 9, 2025, https://www.nvidia.com/en-us/data-center/l40/
3. NVIDIA L40S GPU Overview: Characteristics, Performance, AI Use Cases - Gcore, accessed January 9, 2025, https://gcore.com/learning/nvidia-l40s-overview/
4. NVIDIA, Global Data Center System Manufacturers to Supercharge Generative AI and Industrial Digitalization, accessed January 9, 2025, https://nvidianews.nvidia.com/news/nvidia-global-data-center-system-manufacturers-to-supercharge-generative-ai-and-industrial-digitalization
5. Evaluating NVIDIA A100 and NVIDIA L40S: Which GPU Excels in AI ..., accessed January 9, 2025, https://medium.com/@GPUnet/evaluating-nvidia-a100-and-nvidia-l40s-which-gpu-excels-in-ai-and-graphics-tasks-1c26a8022f5f
6. Choosing the right GPU. NVIDIA L40S vs. A100 and H100 | by Veronica Nigro | mkinf, accessed January 9, 2025, https://medium.com/mkinf/choosing-the-right-gpu-05953d541d48
7. Everything you need to know about the NVIDIA L40S GPU - Blog, accessed January 9, 2025, https://blog.ori.co/nvidia-l40s-gpu-comprehensive-overview

8. NVIDIA L40S GPU Compared to A100 & H100 GPUs | Exxact Blog, accessed January 9, 2025, https://www.exxactcorp.com/blog/components/NVIDIA-L40S-GPU-Compared-to-A100-and-H100-Tensor-Core-GPU
9. Can we add all three GPU's GPU(A6000) + GPU (1x Nvidia ADA 6000 48 GB and 1x Nvidia L40s 48 GB) in same Motherboard, accessed January 9, 2025, https://forums.developer.nvidia.com/t/can-we-add-all-three-gpus-gpu-a6000-gpu-1x-nvidia-ada-6000-48-gb-and-1x-nvidia-l40s-48-gb-in-same-motherboard/293449
10. OpenCL Benchmarks - Geekbench Browser, accessed January 9, 2025, https://browser.geekbench.com/opencl-benchmarks
11. NVIDIA L40S 48GB PCIe Gen4 Passive GPU - ServerSupply.com, accessed January 9, 2025, https://www.serversupply.com/GPU/GDDR6/48GB/NVIDIA/L40S_395278.htm
12. nvidia l40s 48gb graphics card - GPU - ASA Computers, accessed January 9, 2025, https://www.asacomputers.com/nvidia-l40s-48gb-graphics-card.html
13. PNY NVIDIA Quadro L40S Graphic Card - 48 GB GDDR6 - NVL40STCGPU-KIT - CDW, accessed January 9, 2025, https://www.cdw.com/product/pny-nvidia-quadro-l40s-graphic-card-48-gb-gddr6/7582191
14. Tesla L40S 48GB AI HPC Graphics Accelerator PNY 900-2G133-0080-000 PG133G TCSL40SPCIE-PB - Amazon.com, accessed January 9, 2025, https://www.amazon.com/Graphics-Accelerator-900-2G133-0080-000-PG133G-TCSL40SPCIE-PB/dp/B0CLTDCZ82
15. QCT Delivers Unparalleled AI and Graphics Performance with the NVIDIA L40S GPU, accessed January 9, 2025, https://blog.qct.io/qct-delivers-unparalleled-ai-and-graphics-performance-with-the-nvidia-l40s-gpu/