Selectel TEAM

AMD vs NVIDIA — все еще грусть или уже не совсем?

Взгляд из Ноября 2024

Ефим Головин
MLOps-инженер

О Selectel любимом

Серверы и вычисления

- Выделенные серверы готовых и произвольных конфигураций
- Облачные серверы с моментальным запуском
- Серверы с GPU
- Экспериментальное железо

Облачная платформа

- Облачные базы данных
- Объектное и файловое хранилище
- Managed Kubernetes и Container Registry

Организация сети

- Сеть доставки контента (CDN)
- Балансировщики нагрузки
- Selectel Connect и сети L3 VPN
- Облачный DNS

Облако на базе VMware

- Публичное облако
- Частное облако
- Удаленные рабочие столы (VDI)

ML и обработка данных

- ML-платформа
- Платформа обработки данных
- Data Science & Analytical Virtual Machine

Безопасность

- Аттестованный сегмент ЦОД
- Соответствие 152-ФЗ
- Защита от DDoS и WAF

Единая панель управления и система биллинга

Документация к API и база знаний

Система управления ролями (IAM)

Базовая защита от DDoS по умолчанию

Техническая поддержка 24/7

План “А”

Часть I: Железо

Немного поговорим про архитектурные компоненты, память, энергопотребление, важные фичи и стоимость.

Часть II: Софт

Обсудим, как можно сопоставить программный стек компаний NVIDIA и AMD, поддерживаемые фреймворки etc.

Часть III: Эксперимент

Посмотрим на результаты запуска кода на сопоставимых GPU от NVIDIA и AMD.

Часть IV: Итоги

Подведем итоги, пообщаемся, пожелаем друг другу счастья и крепкого здоровья 😉

План “Ы”

- Вопрос I: Можно ли запустить обучение ML-моделей на AMD?
- Вопрос II: Можно ли запустить инференс ML-моделей на AMD?
- Вопрос III: Можно ли работать с AMD в Kubernetes?
- Вопрос IV: Можно ли работать с Distributed ML на AMD?
- Вопрос V: Дорого ли запустить ML-проект на AMD?
- НеВопрос VI: Итоги

Вопрос I: Можно ли запустить обучение ML-моделей на AMD?

Подсказка #1: MosaicML

All / Mosaic Research / Training LLMs with AMD MI250 GPUs and MosaicML

Training LLMs with AMD MI250 GPUs and MosaicML

by Abhi Venigalla

June 30, 2023 in Mosaic AI Research

Подсказка #2: OpenAI

OpenAI will start using AMD chips and could make its own AI hardware in 2026

Image: OpenAI

/ Reuters reports an updated hardware strategy to run ChatGPT and OpenAI's other projects involves using AMD chips via Microsoft Azure in addition to Nvidia.

By Umar Shakir, a news writer fond of the electric vehicle lifestyle and things that plug in via USB-C. He spent over 15 years in IT support before joining The Verge.

Oct 29, 2024, 10:05 PM GMT+3

0 Comments (0 New)

Подсказка #3: Fujitsu

Newsroom > AMD and Fujitsu to Begin Strategic Partnership to Develop More Sustainable Computing Infrastructure Intended to Accelerate Open-Source AI Initiatives

AMD and Fujitsu to Begin Strategic Partnership to Develop More Sustainable Computing Infrastructure Intended to Accelerate Open-Source AI Initiatives

Media Contacts

Shinichi Sunohara
AMD Japan Communications
+81 50 5530-3152
shinichi.sunohara@amd.com

Mitch Haws
AMD Investor Relations
+1 512-944-0790
mitch.haws@amd.com

Fujitsu Limited
Public and Investor Relations Division
Inquiries

Media Library

Find and download the latest AMD corporate and product logos, images, and b-roll footage.

Вопрос I.I: СЛОЖНО ли запустить обучение ML-моделей на AMD?

Ставим драйверы

AMD

```
sudo apt install "linux-headers-$(uname -r)" "linux-modules-extra-$(uname -r)"
```

```
sudo apt install amdgpu-dkms
```

```
sudo reboot
```

NVIDIA

```
sudo apt-get install linux-headers-$(uname -r)
```

```
sudo apt-get install cuda-drivers
```

```
sudo reboot
```

Ставим Docker

AMD

```
curl -sSL https://get.docker.com/ | sh
```

NVIDIA

```
curl -sSL https://get.docker.com/ | sh
```

```
apt-get install -y nvidia-container-toolkit
```

```
nvidia-ctk runtime configure --runtime=docker
```

```
systemctl restart docker
```

Запускаем Docker

AMD

Accessing GPUs in containers

In order to grant access to GPUs from within a container, run your container with the following options:

```
docker run --device /dev/kfd --device /dev/dri --security-opt seccomp=unconfined <image>
```

NVIDIA

Running a Sample Workload with Docker

After you install and configure the toolkit and install an NVIDIA GPU Driver, you can verify your installation by running a sample workload.

- Run a sample CUDA container:

```
sudo docker run --rm --runtime=nvidia --gpus all ubuntu nvidia-smi
```

Запускаем рандомный пример для PyTorch

```
for epoch in tqdm(range(10)):  # loop over the dataset multiple times
    running_loss = 0.0
    for i, data in enumerate(tqdm(trainloader), 0):
        # get the inputs; data is a list of [inputs, labels]
        inputs, labels = data[0].to(device), data[1].to(device)
        # zero the parameter gradients
        optimizer.zero_grad()
        # forward + backward + optimize
        outputs = net(inputs)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()
        # print statistics
        running_loss += loss.item()
    print(f'Epoch #{epoch + 1}; Epoch loss: {running_loss / 2000:.3f}')
    running_loss = 0.0
print('Finished Training')
```

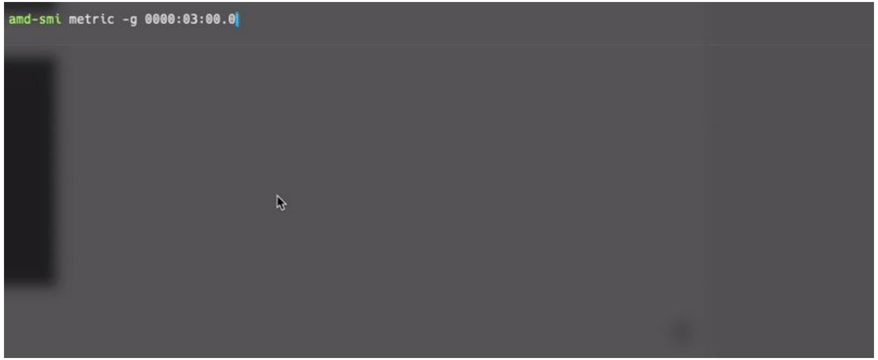
Вопрос I.II: СЛОЖНО ли мониторить обучение ML-моделей на AMD?

Есть несколько утилит

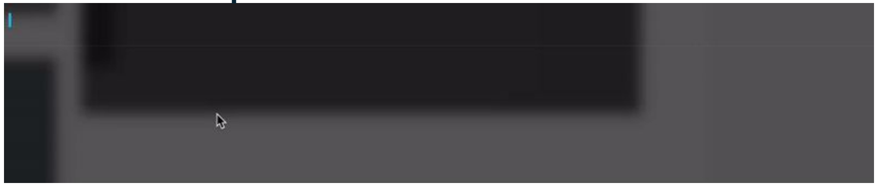
rocm-smi

rocminfo

amd-smi

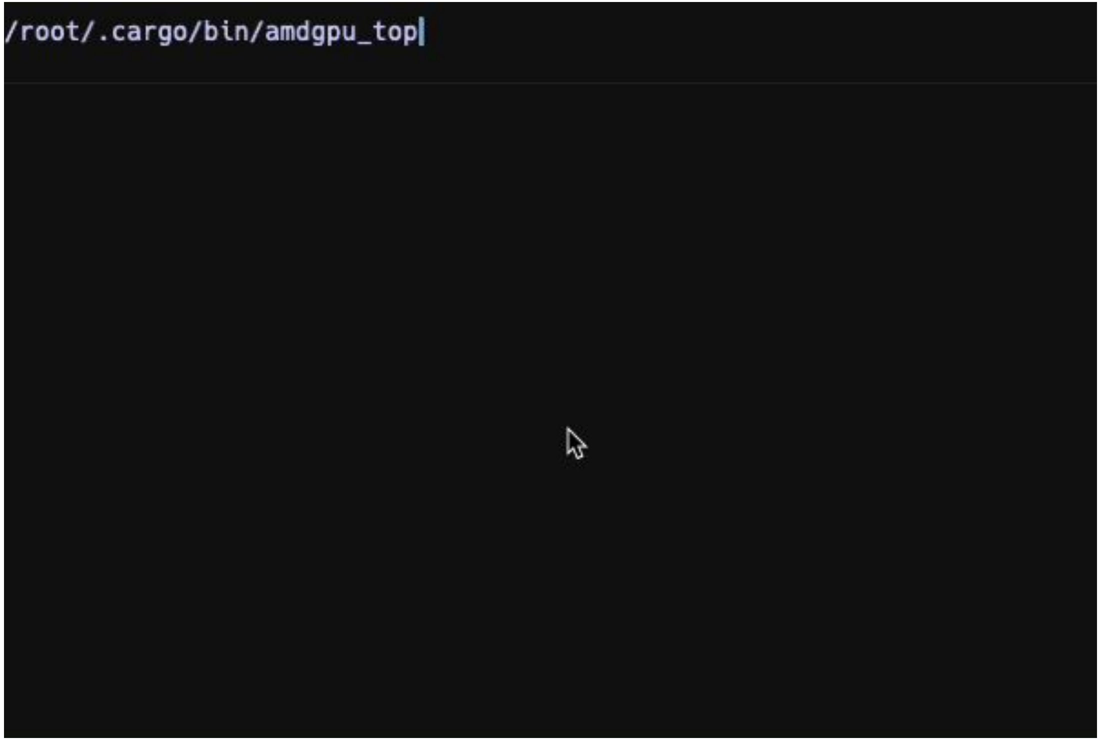


radeontop



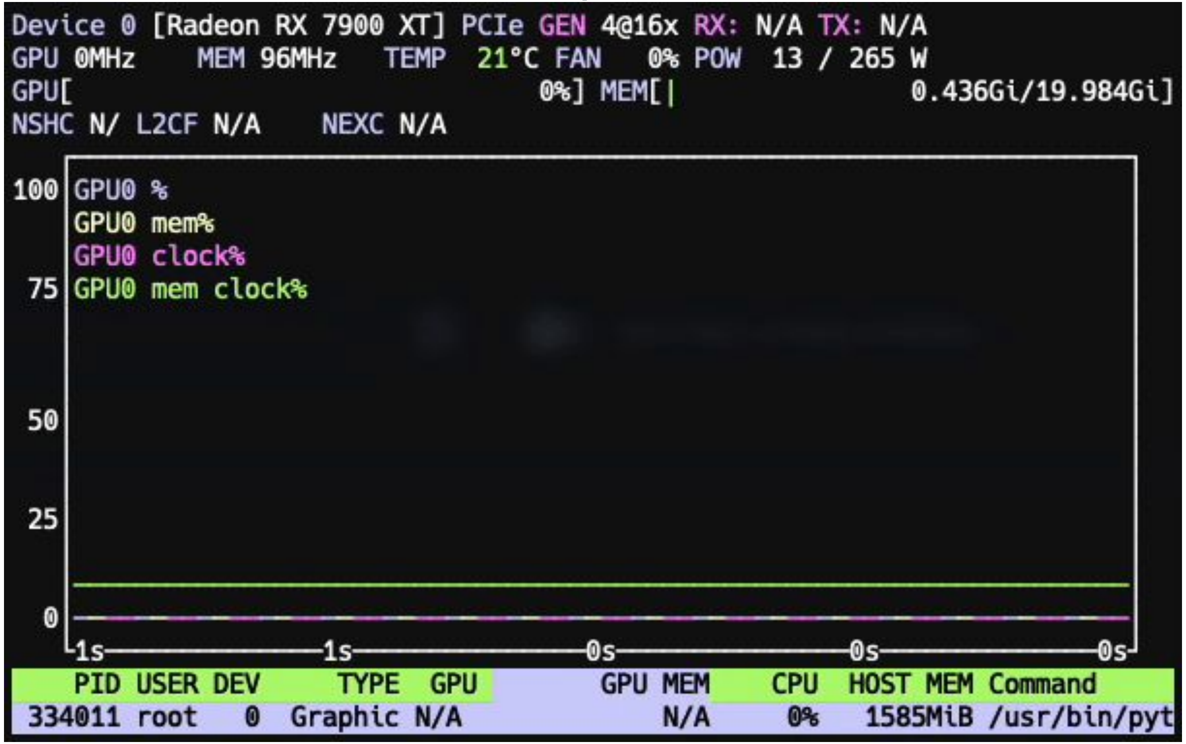
Есть более продвинутый вариант

amdgpu_top



А есть nvtop

nvtop



PyTorch Profiler

```python
model = models.resnet18().cuda()
inputs = torch.randn(5, 3, 224, 224).cuda()

with profile(activities=[
        ProfilerActivity.CPU, ProfilerActivity.CUDA], record_shapes=True) as prof:
    with record_function("model_inference"):
        model(inputs)

print(prof.key_averages().table(sort_by="cuda_time_total", row_limit=10))
```

```python
print(prof.key_averages(group_by_stack_n=5).table(sort_by='self_cpu_time_total'))
```

Name	Self CPU %	Self CPU	CPU total %	CPU total	CPU time avg
hipMemcpyWithStream	99.70%	508.341ms	99.70%	508.341ms	169.447ms
hipMalloc	0.21%	1.066ms	0.21%	1.066ms	1.066ms
aten::addmm	0.03%	155.018us	0.04%	187.037us	187.037us
aten::empty_strided	0.01%	52.459us	0.22%	1.120ms	559.976us
aten::copy_	0.01%	48.318us	99.66%	508.183ms	254.091ms
aten::_to_copy	0.01%	37.660us	99.89%	509.340ms	254.670ms
hipLaunchKernel	0.01%	34.089us	0.01%	34.089us	11.363us
aten::sum	0.01%	26.909us	0.01%	34.189us	34.189us
aten::t	0.01%	25.980us	0.01%	39.160us	39.160us
aten::mean	0.00%	15.580us	0.00%	21.730us	21.730us
aten::to	0.00%	14.420us	99.89%	509.355ms	169.785ms
aten::linear	0.00%	13.389us	0.05%	239.586us	239.586us
aten::_local_scalar_dense	0.00%	10.060us	0.04%	216.647us	216.647us
aten::as_strided	0.00%	8.320us	0.00%	8.320us	2.080us
detach	0.00%	8.289us	0.00%	8.289us	8.289us
aten::transpose	0.00%	6.870us	0.00%	13.180us	13.180us
aten::detach	0.00%	5.400us	0.00%	13.689us	13.689us
hipDeviceSynchronize	0.00%	4.730us	0.00%	4.730us	4.730us
hipExtModuleLaunchKernel	0.00%	4.510us	0.00%	4.510us	4.510us
aten::lift_fresh	0.00%	4.219us	0.00%	4.219us	4.219us
aten::expand	0.00%	3.850us	0.00%	4.590us	4.590us
aten::item	0.00%	3.740us	0.04%	220.387us	220.387us
hipStreamIsCapturing	0.00%	1.360us	0.00%	1.360us	1.360us
hipGetDevicePropertiesR0600	0.00%	0.990us	0.00%	0.990us	0.990us
aten::resolve_conj	0.00%	0.540us	0.00%	0.540us	0.540us
aten::resolve_neg	0.00%	0.190us	0.00%	0.190us	0.190us
void at::native::elementwise_kernel<128, 2, at::nati...	0.00%	0.000us	0.00%	0.000us	0.000us
[memory]	0.00%	0.000us	0.00%	0.000us	0.000us
Cijk_Alik_Bljk_SB_MT16x16x8_SN_1LDSB0_APM1_ABV0_ACED...	0.00%	0.000us	0.00%	0.000us	0.000us
void at::native::reduce_kernel<512, 1, at::native::R...	0.00%	0.000us	0.00%	0.000us	0.000us
void at::native::reduce_kernel<512, 1, at::native::R...	0.00%	0.000us	0.00%	0.000us	0.000us
Memcpy DtoH (Device -> Host)	0.00%	0.000us	0.00%	0.000us	0.000us
Memcpy HtoD (Host -> Device)	0.00%	0.000us	0.00%	0.000us	0.000us

Self CPU time total: 509.894ms
Self CUDA time total: 473.209ms

Вопрос I.III: Какие полезные фишки поддерживает AMD?

Как насчет, скажем, AMP?

Automatic mixed precision in PyTorch using AMD GPUs

As models increase in size, the time and memory needed to train them–and consequently, the cost–also increases. Therefore, any measures we take to reduce training time and memory usage can be highly beneficial. This is where Automatic Mixed Precision (AMP) comes in.

In this blog, we will discuss the basics of AMP, how it works, and how it can improve training efficiency on AMD GPUs.

Prerequisites

To run the code used in this blog, you will need the following:

- Hardware
    - AMD GPU - see the list of compatible GPUs

Как насчет, скажем, AMP?

Supported GPUs

The following table shows the supported AMD Instinct™ accelerators, and Radeon™ PRO and Radeon GPUs. If a GPU is not listed on this table, it's not officially supported by AMD.

Accelerators and GPUs listed in the following table support compute workloads (no display information or graphics). If you're using ROCm with AMD Radeon or Radeon Pro GPUs for graphics workloads, see the Use ROCm on Radeon GPU documentation to verify compatibility and system requirements.

AMD Instinct | AMD Radeon PRO | **AMD Radeon**

GPU	Architecture	LLVM target	Support
AMD Radeon RX 7900 XTX	RDNA3	gfx1100	✅
AMD Radeon RX 7900 XT	RDNA3	gfx1100	✅
AMD Radeon RX 7900 GRE	RDNA3	gfx1100	✅
AMD Radeon VII	GCN5.1	gfx906	⚠

Как насчет, скажем, AMP?

```
[1]: import os

[2]: os.environ["TORCH_BLAS_PREFER_HIPBLASLT"] = "0"

[3]: import gc
     import time
     import numpy as np
     import torch
     import matplotlib.pyplot as plt

[25]: torch.cuda.get_device_name()

[25]: 'Radeon RX 7900 XT'

[4]: def test_amp():
         """Test type casting of torch.autocast"""
         device = "cuda" if torch.cuda.is_available() else "cpu"

         # Create two vectors of size N
         x = torch.rand((1024, 1), device=device)
         y = torch.rand((1024, 1), device=device)
         print(f"Input dtypes:\n  x: {x.dtype}\n  y: {y.dtype}")

         # Perform operations with autocast enabled
```

Вопрос I.IV: А что там с клиентским кодом?

Попробуем запустить проект заказчика

Пришел заказчик:

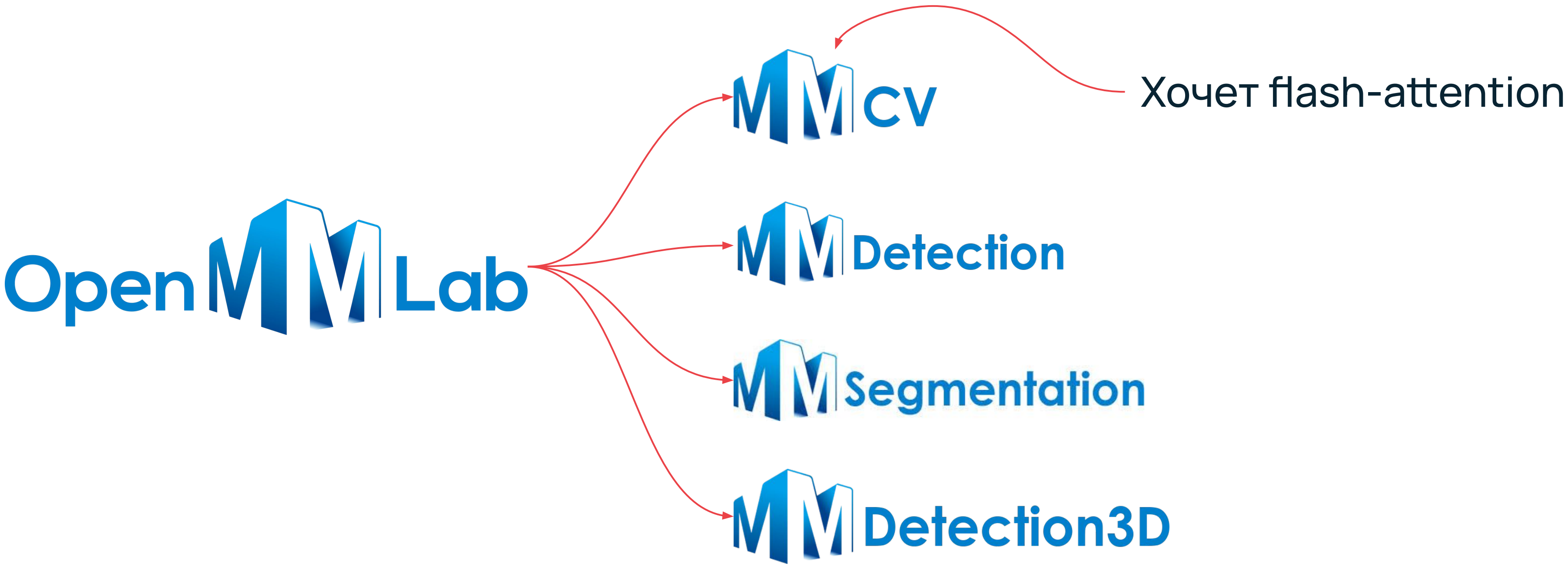
- Работает с CV-задачей;
- Есть наработанный пайплайн;
- Есть сформировавшийся тех. стек:
  - CUDA/cuDNN;
  - NCCL;
  - Docker;
  - PyTorch;
  - OpenMMLab библиотеки.

Попробуем запустить проект заказчика

В теории:

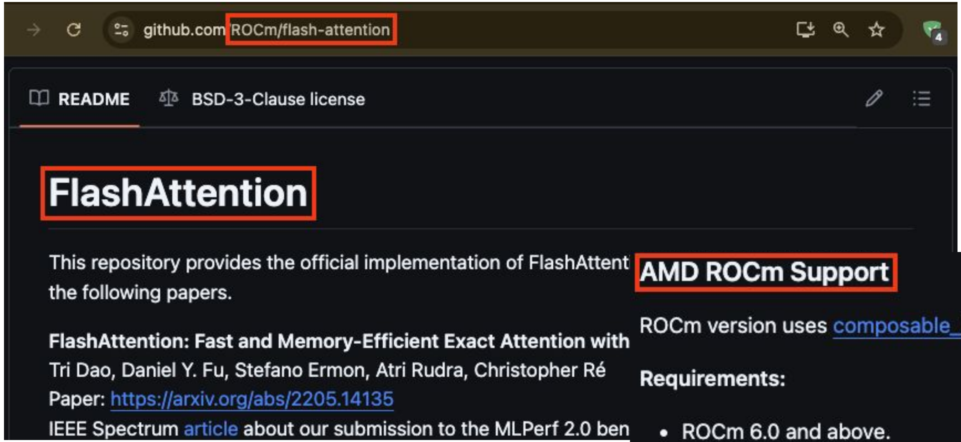
- Работает с CV-задачей;
- Есть наработанный пайплайн;
- Есть сформировавшийся тех. стек:
    - ROCm/MIOpen;
    - RCCL;
    - Docker;
    - PyTorch;
    - OpenMMLab библиотеки.

Все ли так же просто на практике?



А работает ли Flash-Attention на AMD?

И да ✅


github.com/ROCm/flash-attention

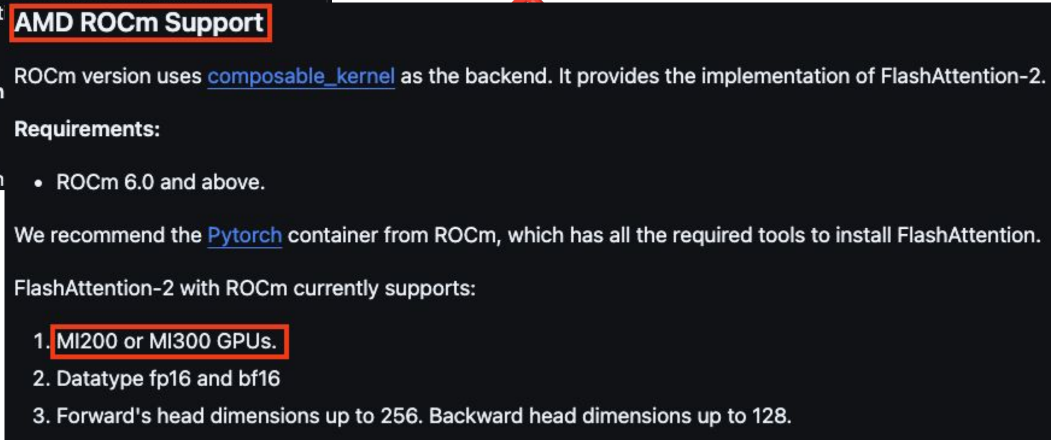
README BSD-3-Clause license

FlashAttention

This repository provides the official implementation of FlashAttent
the following papers.

FlashAttention: Fast and Memory-Efficient Exact Attention with
Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, Christopher Ré
Paper: https://arxiv.org/abs/2205.14135
IEEE Spectrum article about our submission to the MLPerf 2.0 ben

И нет ❌


AMD ROCm Support

ROCm version uses composable_kernel as the backend. It provides the implementation of FlashAttention-2.

Requirements:

- ROCm 6.0 and above.

We recommend the Pytorch container from ROCm, which has all the required tools to install FlashAttention.

FlashAttention-2 with ROCm currently supports:

1. MI200 or MI300 GPUs.
2. Datatype fp16 and bf16
3. Forward's head dimensions up to 256. Backward head dimensions up to 128.

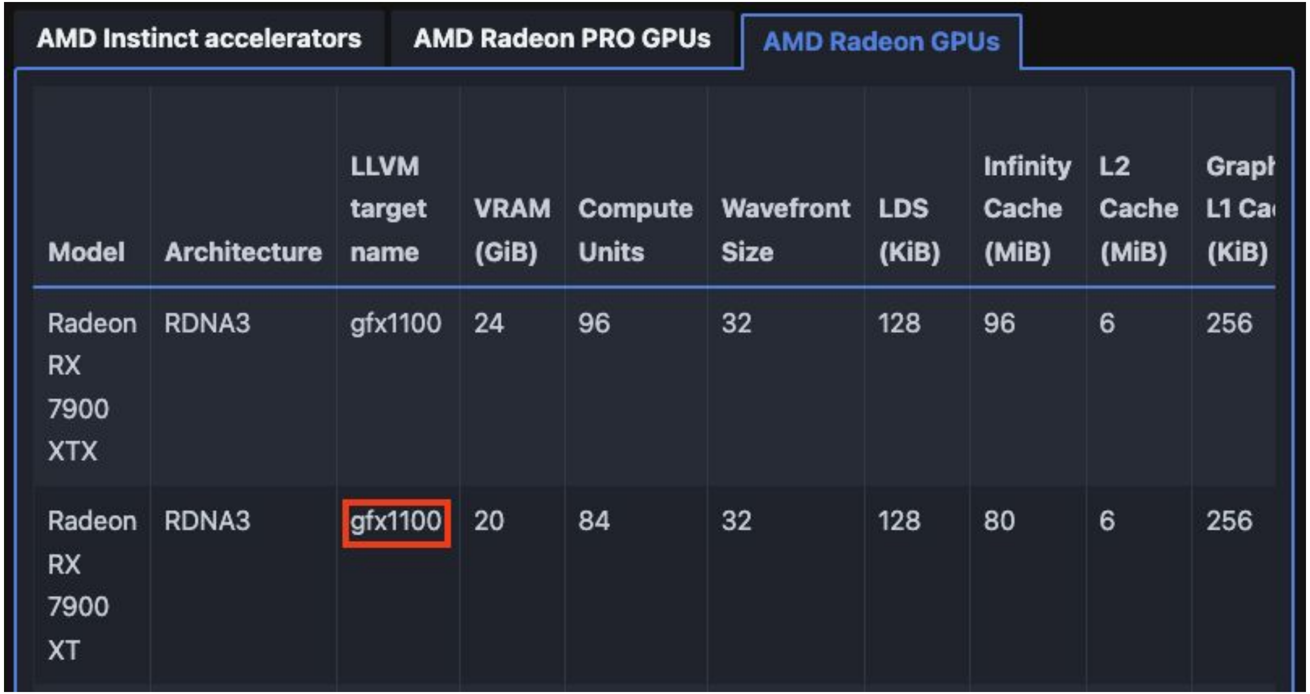
А работает ли Flash-Attention на AMD?

Hardware targets

CK library fully supports *gfx908* and *gfx90a* GPU architectures, while only some operators are supported for *gfx1030* devices. Check your hardware to determine the target GPU architecture.

GPU Target	AMD GPU
gfx908	Radeon Instinct MI100
gfx90a	Radeon Instinct MI210, MI250, MI250X
gfx1030	Radeon PRO V620, W6800, W6800X, W6800X Duo, W6900X, RX 6800, RX 6800 XT, RX 6900 XT, RX 6900 XTX, RX 6950 XT

А работает ли Flash-Attention на AMD?



AMD Instinct accelerators | AMD Radeon PRO GPUs | AMD Radeon GPUs

Model	Architecture	LLVM target name	VRAM (GiB)	Compute Units	Wavefront Size	LDS (KiB)	Infinity Cache (MiB)	L2 Cache (MiB)	Graph L1 Ca (KiB)
Radeon RX 7900 XTX	RDNA3	gfx1100	24	96	32	128	96	6	256
Radeon RX 7900 XT	RDNA3	gfx1100	20	84	32	128	80	6	256

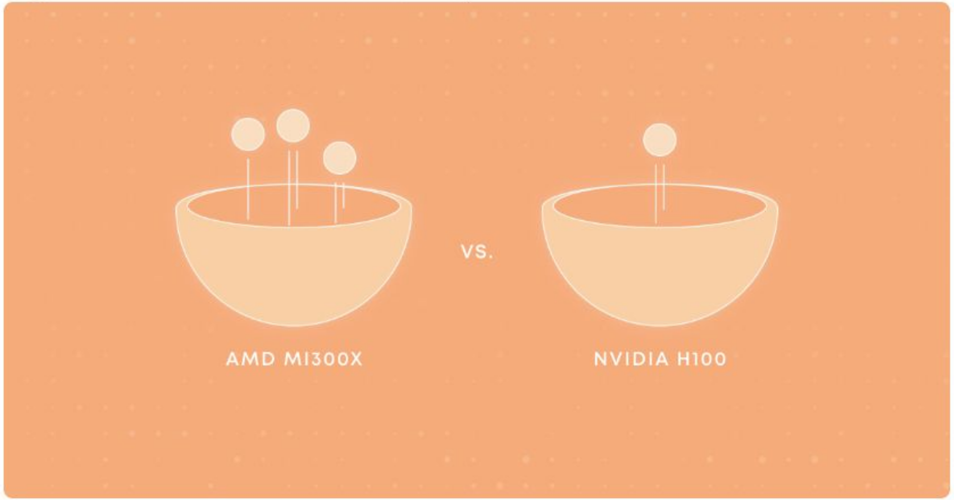
Так вот: В целом обучение запускать **МОЖНО**. Но Есть **ВОПРОСИКИ** 😏

Вопрос II: Можно ли запустить инференс ML-моделей на AMD?

Подсказка #1: Valohai

Valohai

Blog / AMD GPU Performance for LLM Inference: A Deep Dive



AMD GPU Performance for LLM Inference: A Deep Dive

by Eero Laaksonen | on October 31, 2024

Подсказка #2: mlc-ai

MACHINE LEARNING COMPILATION

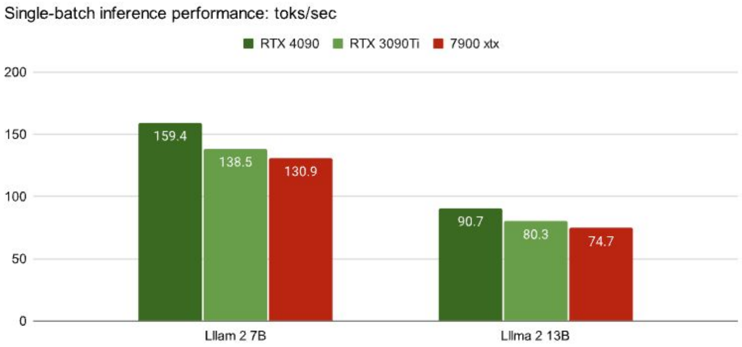
Home

Making AMD GPUs competitive for LLM inference

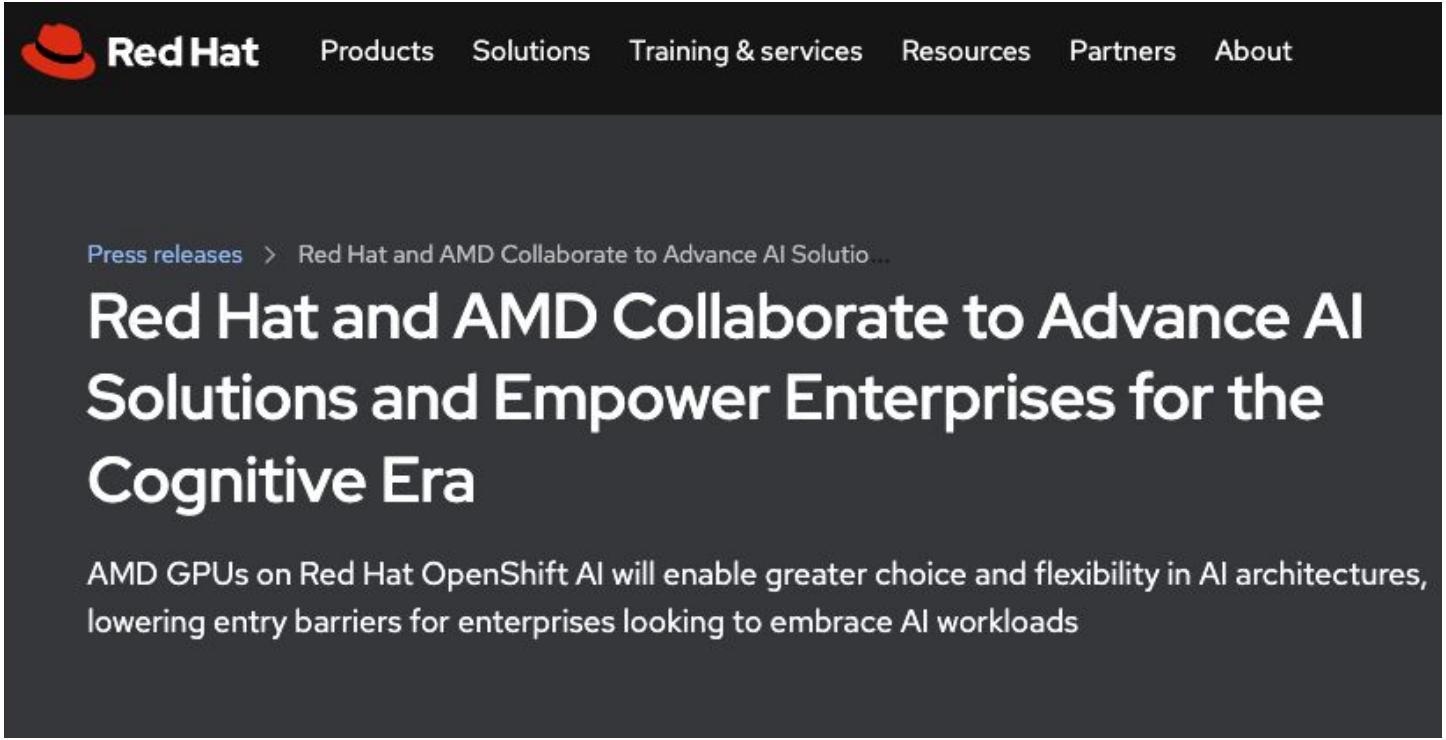
Aug 9, 2023 • MLC Community

TL;DR

MLC-LLM makes it possible to compile LLMs and deploy them on AMD GPUs using **ROCm** with competitive performance. More specifically, AMD Radeon™ RX 7900 XTX gives **80%** of the speed of NVIDIA® GeForce RTX™ 4090 and **94%** of the speed of NVIDIA® GeForce RTX™ 3090Ti for Llama2-7B/13B. Besides ROCm, our Vulkan support allows us to generalize LLM deployment to other AMD devices, for example, a SteamDeck with an AMD APU.



Подсказка #3: Red Hat



WSGI/ASGI servers

django

uWSGI

FastAPI

CherryPy

Flask

Inference servers

vLLM

TensorRT

AMD Inference Server

AMD

SELDON® ML SERVER

APHRODITE

ONNX RUNTIME

MACHINE LEARNING COMPILATION

Inference servers

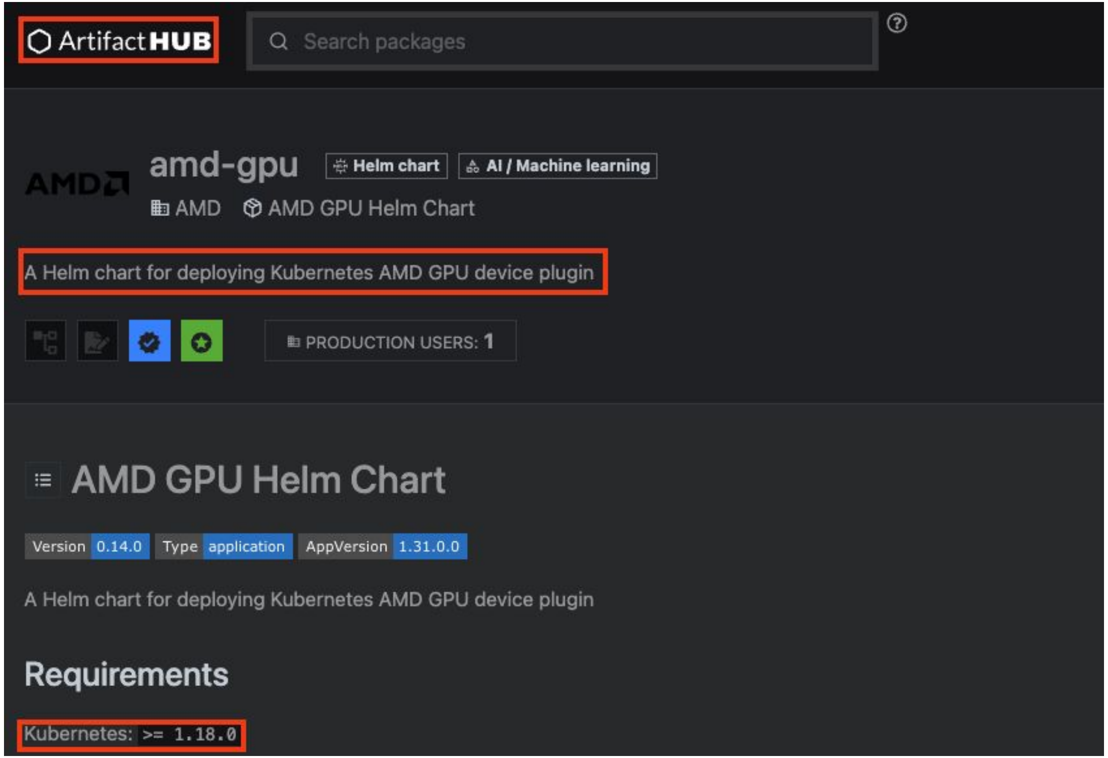
	NVIDIA SUPPORT	AMD SUPPORT
vLLM	✅	✅
Tensor-RT-LLM	✅	❌
Triton Inference Server	✅	❌
AMD Inference Server	❌	✅
ML Server	✅	✅
Ollama	✅	✅
Text Generation Inference	✅	✅
Aphrodite	✅	✅
TorchServe	✅	❓
ONNX-Runtime	✅	✅
Machine Learning Compilation	✅	✅

Так вот: Инференс запускать **МОЖНО**. Но

Что если мне надо масштабироваться?

Вопрос III: Можно ли работать с AMD в Kubernetes?

AMD GPU device plugin



AMD GPU operator

README Apache-2.0 license

AMD GPU Operator

The AMD GPU Operator uses the operator framework to enable the provisioning of AMD GPU in a Kubernetes cluster.

Components

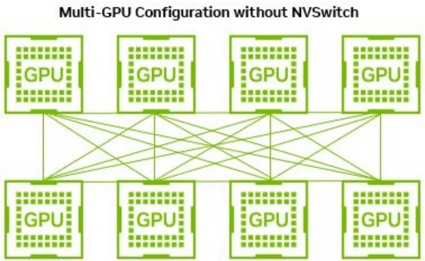
- AMD GPU device plugin

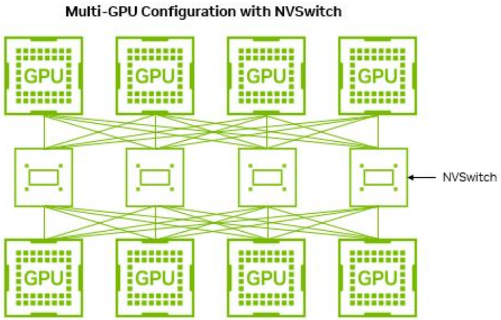
Так вот: В кубер можно ГИПОТЕТИЧЕСКИ заехать. Но Это неточно 😁

Вопрос IV: Можно ли работать с Distributed ML на AMD?

Коннект GPU-шек: NVIDIA vs AMD

NVLink/NVSwitch



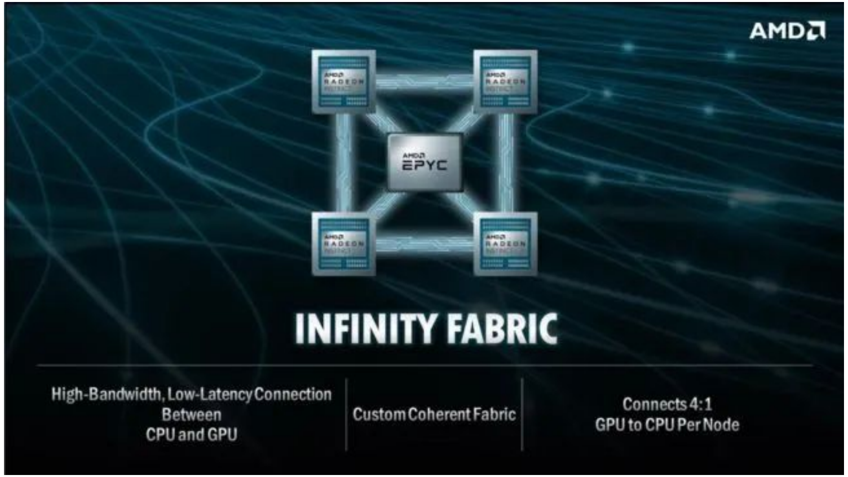


линк

линк

линк

Infinity Fabric



линк линк линк

Infinity Fabric

05-05-2024 06:06 AM

How to Utilize Multi-GPU Infinity Fabric Link in ML

We know that the Infinity Fabric (IF) Link (XGMI) Bridge can greatly improve the performance of Inter-GPU communication just like the NVLink. I'm actually a user who has two Radeon Pro VII with IF Link connected, and I'm sure that this question is the same for those who have four MI100 with IF Link connected. So, the main question is that how can we make use of the advantages of the Infinity Fabric Link in Machine Learning? For example, in PyTorch, can we utilize the high Inter-GPU bandwidth and the shared memory space offered by IF Link so that we can process bigger model and more efficiently? (So far specifically for running the model, I tried running stable diffusion, but after the memory of a single card is full, HIP gave me a OOM error, and the second card's memory usage was 0, I don't know if this is a bug, and whether AMD is aware of this.) I have no idea after searching the internet, and all materials I found is about the usage of NVLink. For the Infinity Fabric Link, I don't even know if PyTorch support the usage of this bridge. Can any dear developers, users or AMD officials share some information on this? Thank you so much!

ЛИНК

blakeblossom Journeyman III

05-17-2024 05:47 AM

Currently, PyTorch doesn't offer native support for Infinity Fabric Link specifically. However, you can still utilize IF Link's high bandwidth for distributed training with some additional configuration.

Infinity Fabric

February 14, 2023

Democratizing AI with PyTorch Foundation and ROCm™ support for PyTorch

KEY PYTORCH LIBRARIES SUPPORT ADDED

PyTorch ecosystem libraries like TorchText (Text classification), TorchRec (libraries for recommender systems - RecSys), TorchVision (Computer Vision), TorchAudio (audio and signal processing) are fully supported since ROCm 5.1 and upstreamed with PyTorch 1.12.

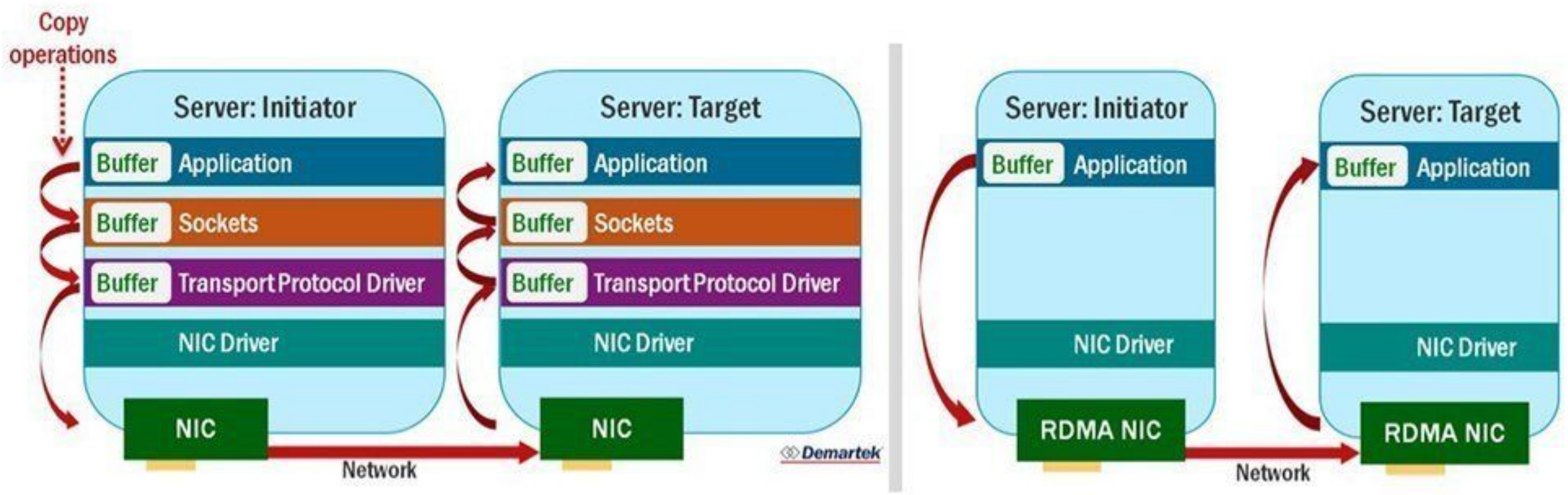
Key libraries provided with the ROCm software stack including MIOpen (Convolution models), RCCL (ROCm Collective Communications) and rocBLAS (BLAS for transformers) were further optimized to offer new potential efficiencies and higher performance.

MIOpen innovates on several fronts, such as implementing fusion to optimize for memory bandwidth and GPU launch overheads, providing an auto-tuning infrastructure to overcome the large design space of problem configurations, and implementing different algorithms to optimize convolutions for different filter and input sizes. MIOpen is one of the first libraries to publicly support the bfloat16 data-type for convolutions, allowing efficient training at lower precision maintaining expected accuracy.

RCCL (pronounced "Rickle") is a stand-alone library of standard collective communication routines for GPUs, implementing all-reduce, all-gather, reduce, broadcast, reduce-scatter, gather, scatter, and all-to-all. There is support for direct GPU-to-GPU send and receive operations. It has been optimized to achieve high bandwidth on platforms using PCIe®, Infinity Fabric™ (GPU to GPU) as well as networking using InfiniBand Verbs or TCP/IP sockets. RCCL supports an arbitrary number of GPUs installed in single or multiple nodes and can be used in either single- or multi-process (e.g., MPI) applications.

линк

Коннект машин по сети: RDMA



линк линк

Коннект машин по сети: RDMA

INFINIBAND℠ TRADE ASSOCIATION

RoCE

Our Members

- Akhetonics
- Alignment Engine
- Alphawave Semi
- AMD
- Amphenol
- Anritsu
- Applied Optoelectronics Inc., Ltd.
- Beijing Zitiao Technology
- BizLink Technology, Inc.
- Broadcom
- Bull SAS / Atos
- CIENA Corp
- Cisco Systems, Inc.
- Cloud Light Technology Limited
- Coherent Corp
- ConnPro Industries Inc.
- DreamBig Semiconductor Inc.
- Eoptolink Technology Inc., Ltd.
- Foxconn Interconnect Technology, Ltd.
- Fujitsu Limited
- Grovf LLC
- **Hewlett-Packard Enterprise**
- Hisense Broadband Multimedia Technologies Co., Ltd.
- Huawei Technologies Co., Ltd.
- **IBM**
- Infraeo
- InnoLight Technologies
- **Intel Corporation**
- Juniper Networks
- Keysight Technologies, Inc.
- LeapMind Inc.
- Marvell Technology Group
- MaxLinear
- Microsoft
- Molex, LLC
- NetApp
- **NVIDIA**
- Optomind Inc.
- Oracle America Inc.
- Parade Technologies
- Rivos Inc.
- Rohde & Schwarz
- Semtech EMEA
- Shanghai Yunsilicon Technology Co. Ltd.
- Shenzhen Jaguar Microsystems Co. Ltd.
- Siemon Company
- Siemens Industry Software, Inc.
- Software Forge, Inc.
- TE Connectivity
- Tenesix Inc.
- UNH InterOperability Lab
- Vcinity, Inc.
- Volex inc.
- Wilder Technologies
- Wuxi Stars Microsystem Technology Co., Ltd.
- Yamaichi Electronics USA
- Zitiao Network Technology Co., Ltd.

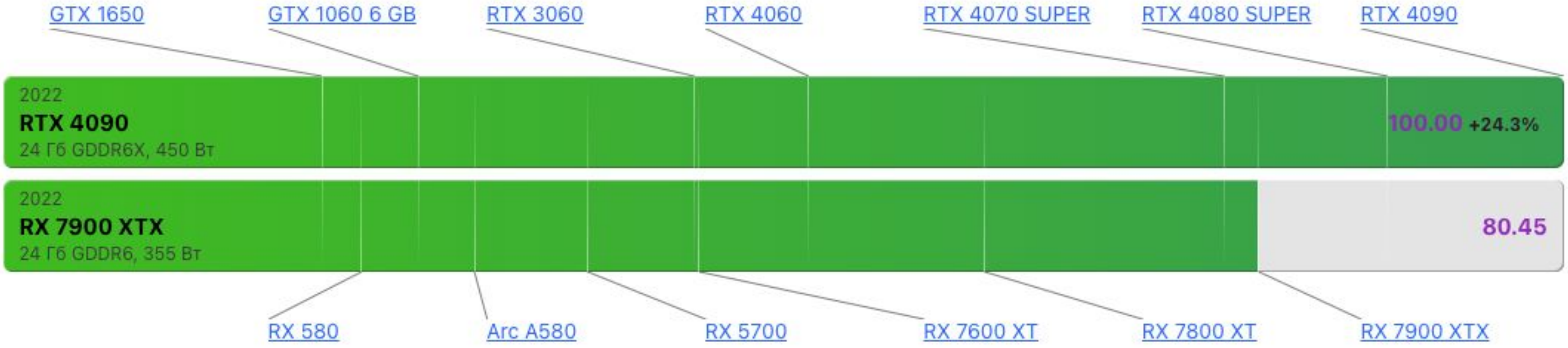
BOLD = Steering Committee Members

линк

Так вот: Distributed ML ГИПОТЕТИЧЕСКИ есть. Но Это неточно 😁

Вопрос V: Дорого ли запустить ML-проект на AMD?

NVIDIA RTX 4090 vs AMD RX 7900 XTX



NVIDIA RTX 4090 vs AMD RX 7900 XTX

	NVIDIA RTX A5000	AMD RX 7900 XT
Архитектура	Ada Lovelace (2022−2024)	RDNA 3.0 (2022−2024)
Тип	Десктопная	Десктопная
Дата выхода	20 сентября 2022	3 ноября 2022
Количество потоковых процессоров	16384	6144
Количество транзисторов	76,300 млн	57,700 млн
FP32 TFLOPS	82.58	61.39
FP16 TFLOPS	330	123
Объём памяти	24 ГБ	24 ГБ
Частота памяти	1313 МГц	2500 МГц
TDP	450W	320W

NVIDIA RTX 4090 vs AMD RX 7900 XTX

88 баллов

Видеокарта GIGABYTE GeForce RTX 4090 AERO OC [GV-N4090AERO OC-24GD]

62 · 6

249 999 ₽
от 24 370 ₽/ мес.

88 баллов

Видеокарта Sapphire AMD Radeon RX 7900 XTX PULSE OC [11322-02-20G]

90 · 7

119 999 ₽
от 11 698 ₽/ мес.

2x+

NVIDIA RTX 4090 vs AMD RX 7900 XTX

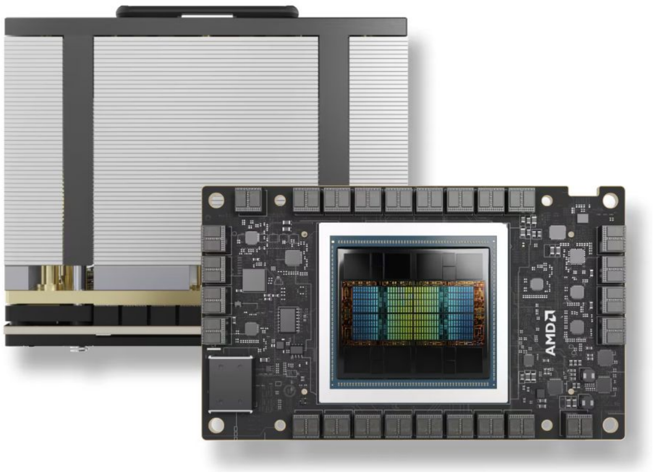
Benchmark with MLC Python Package

We benchmarked the Llama 2 7B and 13B with 4-bit quantization. And we measure the decoding performance by setting a single prompt token and generating 512 tokens. All the results are measured for single batch inference.

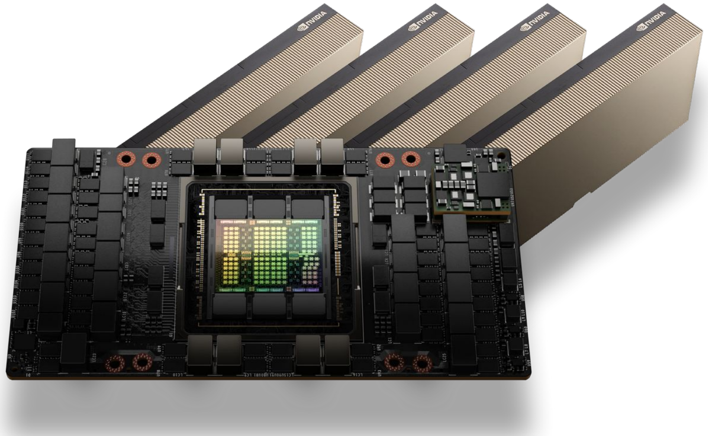
	AMD Radeon™ RX 7900 XTX	NVIDIA ® GeForce RTX™ 4090	NVIDIA ® GeForce RTX™ 3090 Ti
Llama 2 7B	130.9 toks/s	159.4 toks/s	138.5 toks/s
Llama 2 13B	74.7 toks/s	90.7 toks/s	80.3 toks/s

For single batch inference performance, it can reach 80% of the speed of NVIDIA 4090 with the release of ROCm 5.6.

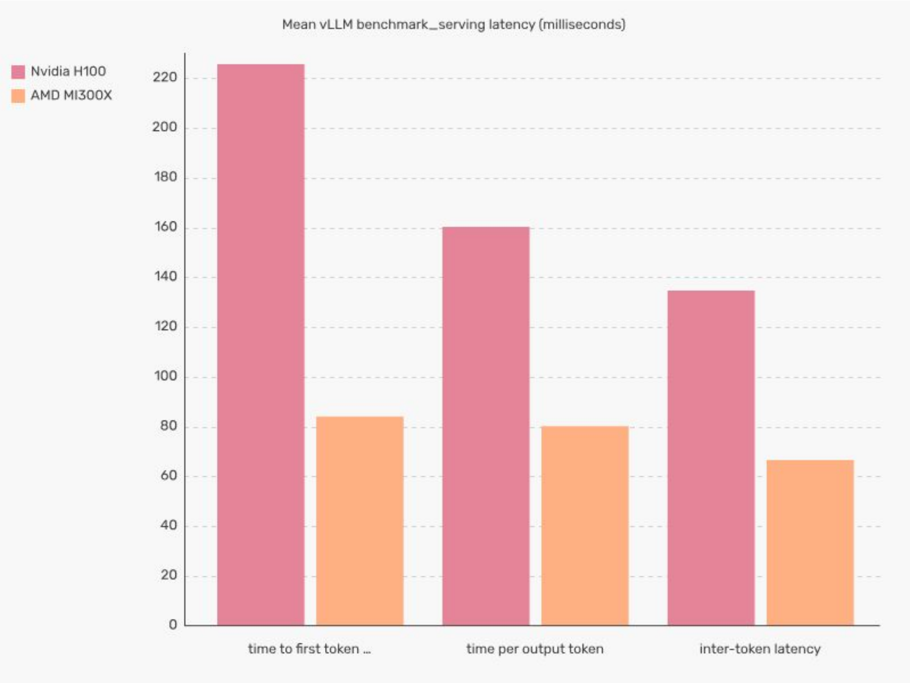
А как там поживают флагманы?

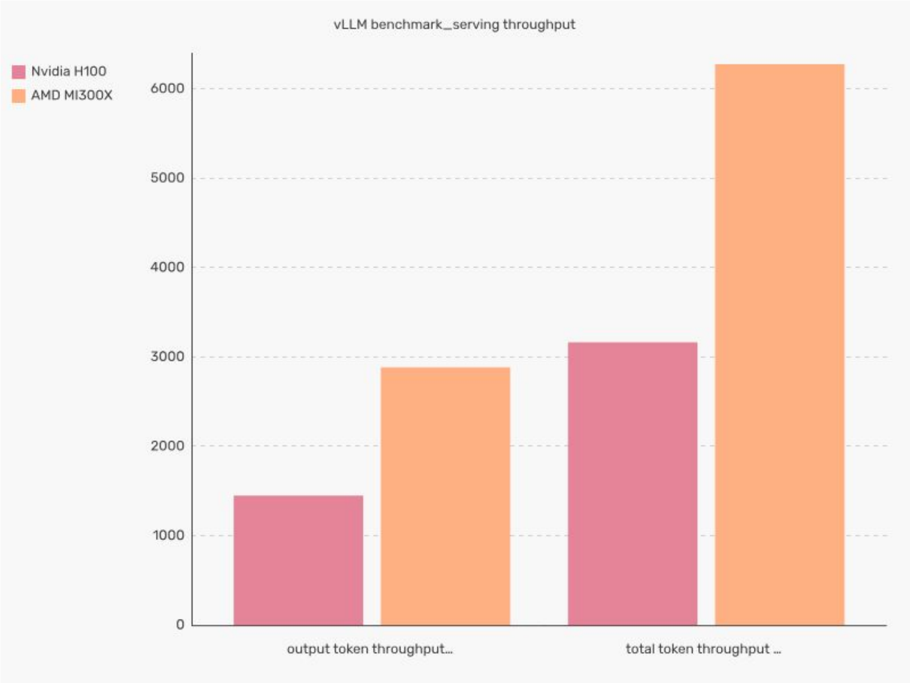


VS



А как там поживают флагманы?

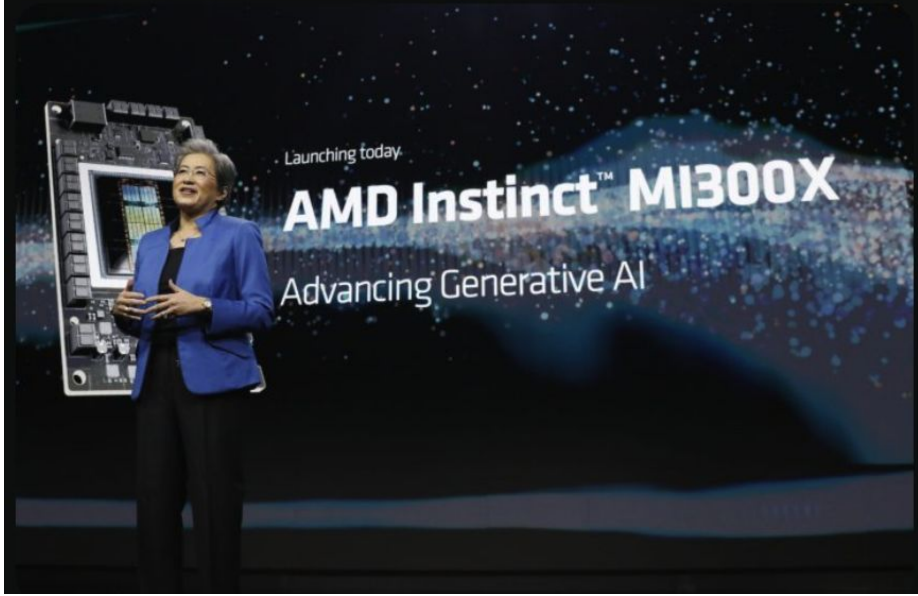




NVIDIA H100 PCIe vs AMD Instinct 300X

Samsung Purchases $20 Million Worth of AMD MI300X Data Center GPUs for AI Development

Sarfraz Khan · Oct 7, 2024 at 07:30 AM EDT · Copy Shortlink 40



NVIDIA H100 PCIe vs AMD Instinct 300X

Similar to most giants, Samsung Electronics is also working on developing its own AI and needs serious horsepower to power its systems. While companies like Meta and xAI went with NVIDIA H100 GPUs, Samsung is going through a more affordable route. Compared to NVIDIA's H100, which sells for $30000-$40000, the AMD MI300X costs several times less. The GPU, despite lacking behind NVIDIA's Hopper lineup in AI workloads, has been seen as a good alternative in terms of its pricing.

The AMD MI300X is said to have cost Samsung roughly $10000 per piece and is currently the flagship model from AMD in the Instinct family, released at the end of 2023. It brings 19456 Stream Processors, 304 Compute Units, and 192 GB HBM3(High Bandwidth Memor) memory for intensive workloads. More on the chip here. The GPU is hence, a much cost-effective solution for large-scale projects. An official in the semiconductor industry said,

НеВопрос VI: Итоги

Что же можно в итоге сказать?

- Запустить обучение ML-моделей на AMD МОЖНО?
- Запустить инференс ML-моделей на AMD МОЖНО
- Работать с AMD в Kubernetes МОЖНО?
- Работать с Distributed ML на AMD МОЖНО?
- Запустить ML-проект на AMD НЕ ДОРОЖЕ, чем на NVIDIA, НО ЕСТЬ НЮАНС!

ML-чик в Selectel любимом

MLечный путь

ML в Selectel

Мой канальчик

Selectel TEAM

Пожелаем же друг другу счастья и крепкого здоровья 😊

Ефим Головин
MLOps-инженер