ВикипедиЯ

Hopper (микроархитектура)

Материал из Википедии — свободной энциклопедии

Hopper — микроархитектура профессиональных графических процессоров класса Server/Datacenter представленная в марте 2022 года, и разработанная корпорацией NVIDIA Corporation в качестве преемника микроархитектуры Ampere. Она названа в честь Грейс Мюррей Хоппер (англ. *Grace Murray Hopper*) — американской учёной в области информатики и контр-адмирала Военно-морских сил США, которая была одной из первых программистов компьютера Марк I.

Микроархитектура Hopper с тензорными ядрами была анонсирована в конце марта 2022 года и впервые появилась в ускорителе GPGPU-вычислений уровня дата-центра NVIDIA H100 с 80 Гбайт HBM3 памяти, который содержит порядка 80 млрд транзисторов. Ускорители NVIDIA H100 например используются в HPC-серверах Nvidia DGX H100 для машинного обучения систем искусственного интеллекта[1][2].

Nvidia Hopper

Кодовое имя	GH100
Дата выпуска	Март 2022 (NVIDIA H100)
Производители	TSMC (4 нм)
Тип памяти	HBM3

← Ampere (consumer, professional), Volta (professional) | Blackwell →

Не существует массовых видеокарт десктопного уровня серии GeForce на базе микроархитектуры Hopper. В сентябре же 2022 года были представлены графические ускорители десктопного уровня серии GeForce RTX 40 с упрощённой микроархитектурой Ada Lovelace, названной в честь математика Ады Лавлейс, которая также пришли на смену микроархитектуры Ampere[3].

Содержание

Технические подробности

Архитектурные усовершенствования микроархитектуры Hopper включают следующее:

- *CUDA Compute Capability* 9.0
- Память с высокой пропускной способностью 3-го поколения (HBM3).
- *NVLink 4.0*: шина с высокой пропускной способностью между центральным процессором и графическим процессором, а также между несколькими графическими процессорами. Обеспечивает гораздо более высокие скорости передачи, чем те, которые достижимы при использовании PCI Express; обеспечивает скорость 50 Гбайт/с на один канал и до 900 Гбайт/с (18 × 50 Гбайт/с) на один GPU.
- Тензорные ядра: Тензорное ядро — это объект, который умножает две матрицы FP16 4×4, а затем добавляет к результату третью матрицу FP16 или FP32 с помощью операций умножения примесей и получает результат *FP32*, который при необходимости можно понизить до результатов *FP16*. Тензорные ядра предназначены для ускорения обучения нейронных сетей.

Спецификации

Сравнительная таблица GP100, GV100, GA100 и GH100[4][5]

GPU features	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA A100	NVIDIA H100
GPU codename	GP100	GV100	GA100	GH100
GPU architecture	NVIDIA Pascal	NVIDIA Volta	NVIDIA Ampere	NVIDIA Hopper
Compute capability	6.0	7.0	8.0	9.0
Threads / warp	32	32	32	32
Max warps / SM	64	64	64	64
Max threads / SM	2048	2048	2048	2048
Max thread blocks / SM	32	32	32	32
Max Thread Blocks / Thread Block Clusters	N/A	N/A	N/A	16
Max 32-bit registers / SM	65536	65536	65536	65536
Max registers / block	65536	65536	65536	65536
Max registers / thread	255	255	255	255
Max thread block size	1024	1024	1024	1024
FP32 cores / SM	64	64	64	128
Ratio of SM registers to FP32 cores	1024	1024	1024	512
Shared Memory Size / SM	64 KB	Configurable up to 96 KB	Configurable up to 164 KB	Configurable up to 228 KB

Матрица сравнения поддержания точности вычислений[6][7]

	Supported CUDA Core Precisions									Supported Tensor Core Precisions								
	FP8	**FP16**	**FP32**	**FP64**	**INT1**	**INT4**	**INT8**	**TF32**	**BF16**	**FP8**	**FP16**	**FP32**	**FP64**	**INT1**	**INT4**	**INT8**	**TF32**	**BF16**
NVIDIA Tesla P4	Нет	Нет	Да	Да	Нет	Нет	Да	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет
NVIDIA P100	Нет	Да	Да	Да	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Нет
NVIDIA Volta	Нет	Да	Да	Да	Нет	Нет	Да	Нет	Нет	Нет	Да	Нет	Нет	Нет	Нет	Нет	Нет	Нет
NVIDIA Turing	Нет	Да	Да	Да	Нет	Нет	Да	Нет	Нет	Нет	Да	Нет	Нет	Да	Да	Да	Нет	Нет
NVIDIA A100	Нет	Да	Да	Да	Нет	Нет	Да	Нет	Да	Нет	Да	Нет	Да	Да	Да	Да	Да	Да
NVIDIA H100	Нет	Да	Да	Да	Нет	Нет	Да	Нет	Да	Да	Да	Нет	Да	Нет	Нет	Да	Да	Да

Обозначение:

- FPnn: floating point with nn bits
- INTn: integer with n bits
- INT1: binary
- TF32: TensorFloat32
- BF16: bfloat16

Сравнение мощностей декодирования

Видео	**H.264 decode (1080p30)**	**H.265 (HEVC) decode (1080p30)**	**VP9 decode (1080p30)**
V100	16	22	22
A100	75	157	108
H100	170	340	260

Изображение/сек[5]	**JPEG 4:4:4 decode(1080p)**	**JPEG 4:2:0 decode(1080p)**
A100	1490	2950
H100	3310	6350

GPGPU-ускорители

Ускорители GPGPU-вычислений с тензорными ядрами, в которых используются чипы с микроархитектурой Hopper:

- NVIDIA H100 — с середины 2022 года;
- NVIDIA GH200 Grace Hopper — с середины 2023 года.

Примечания

1. NVIDIA анонсировала 4-нм ускорители Hopper H100 и самый быстрый в мире ИИ-суперкомпьютер EOS на базе DGX H100 (https://servernews.ru/1062434) (рус.). ServerNews. (22 марта 2022). Дата обращения: 21 сентября 2023. Архивировано (https://web.archive.org/web/20230920191819/https://servernews.ru/1062434) 20 сентября 2023 года.

2. Представлен ускоритель вычислений NVIDIA H100 на новейшей архитектуре Hopper и с памятью HBM3 (https://3dnews.ru/1062506/publikatsiya-1062506) (рус.). 3DNews. (22 марта 2022). Дата обращения: 18 сентября 2023. Архивировано (https://web.archive.org/web/20231125192152/https://3dnews.ru/1062506/publikatsiya-1062506) 25 ноября 2023 года.
3. NVIDIA представила GeForce RTX 4090 и две GeForce RTX 4080 — ускорители нового поколения с ценой от $899 (https://3dnews.ru/1074553/nvidia-predstavila-geforce-rtx-4090-i-dve-geforce-rtx-4080-uskoriteli-novogo-pokoleniya-stoimostyu-ot-899) (рус.). 3DNews. (20 сентября 2022). Дата обращения: 21 сентября 2023. Архивировано (https://web.archive.org/web/20221014214202/https://3dnews.ru/1074553/nvidia-predstavila-geforce-rtx-4090-i-dve-geforce-rtx-4080-uskoriteli-novogo-pokoleniya-stoimostyu-ot-899) 14 октября 2022 года.
4. NVIDIA A100 Tensor Core GPU Architecture (https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf) . *www.nvidia.com*. Дата обращения: 18 сентября 2020. Архивировано (https://web.archive.org/web/20210215205655/https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf) 15 февраля 2021 года.
5. NVIDIA H100 Tensor Core GPU Architecture Whitepaper (https://nvdam.widen.net/s/9bz6dw7dqr/gtc22-whitepaper-hopper). *NVIDIA*.
6. NVIDIA Tensor Cores: Versatility for HPC & AI (https://www.nvidia.com/en-us/data-center/tensor-cores/). *NVIDIA*. Дата обращения: 22 сентября 2023. Архивировано (https://web.archive.org/web/20230921054929/https://www.nvidia.com/en-us/data-center/tensor-cores/) 21 сентября 2023 года.
7. Abstract (https://docs.nvidia.com/deeplearning/tensorrt/support-matrix/index.html). *docs.nvidia.com*. Дата обращения: 22 сентября 2023. Архивировано (https://web.archive.org/web/20230922162747/https://docs.nvidia.com/deeplearning/tensorrt/support-matrix/index.html) 22 сентября 2023 года.

Ссылки

- Архитектура NVIDIA Hopper (https://www.nvidia.com/ru-ru/data-center/technologies/hopper-architecture/) (рус.). Официальный сайт NVIDIA Corporation. Дата обращения: 21 сентября 2023.
- GPU NVIDIA H100 с тензорными ядрами (https://www.nvidia.com/ru-ru/data-center/h100/) (рус.). Официальный сайт NVIDIA Corporation. Дата обращения: 21 сентября 2023.
