Тенденции в графических ускорителях для высокопроизводительных вычислений

А. С. Шмелёв[1]

[1]МСЦ РАН филиал ФГУ ФНЦ НИИСИ РАН, Москва, Россия, guest8993@rambler.ru

Аннотация. Графические карты, построенные на основе большого количества простых и однотипных исполнительных устройств и обладающие высокой пиковой производительностью уже давно используются в области высокопроизводительных вычислений в качестве ускорителей вычислений. В настоящее время выпускаются отдельные продукты, ориентированные на применение в вычислительных центрах. В данной работе приводится обзор современных ускорителей для вычислительных центров, приведены их показатели производительности, а также приведены анонсы перспективных ускорителей вычислений и показаны тенденции в данной области.

Ключевые слова: высокопроизводительные вычисления, графические ускорители, ускорители вычислений, суперЭВМ, память с высокой пропускной способностью.

1. Введение

Все современные и ожидаемые в ближайшей перспективе процессоры общего назначения имеют архитектуру управления потоком команд. Порядок выполнения программ задаётся счётчиком команд, а для выдачи на выполнение нескольких команд в такт используется достаточно сложная и ресурсоёмкая аппаратура (системы выдачи команд не по порядку, позволяющие искать готовые к выполнению команды в окне до 200 команд, схемы предсказания переходов и т.д.), но и она не позволяет выдавать более 4-6 команд – такой темп выдачи держится уже много лет. Тактовая частота процессорных ядер тоже уже практически не увеличивается, и увеличение производительности процессоров достигается за счёт роста числа процессорных ядер и использования векторной обработки. Для построения высокопроизводительных вычислительных систем требуется использование очень большого количества процессорных ядер, что накладывает серьёзные ограничения на классы задач, которые могут эффективно использовать столь большое число процессоров и осложняет программирование. В настоящее время для построения суперкомпьютеров в дополнение к процессорам общего назначения широко применяются графические ускорители, которые на кодах с высоким уровнем однородного параллелизма позволяют достичь ускорения до нескольких десятков раз.

2. Графические ускорители

Основу практически всех графических ускорителей уже много лет составляют относительно простые однотипные исполнительные устройства, объединённые в группы с набором управляющей логики. У фирмы NVIDIA эти группы называются потоковыми мультипроцессорами Streaming Multiprocessor (SM). Все исполнительные устройства в пределах SM выполняют одну и ту же программу, причем в каждом ИУ может выполняться в режиме разделения времени несколько копий этой программы с разными данными. Современные ускорители поддерживают до нескольких тысяч нитей на SM, называемых тредами (threads). Переключение между тредами в каждом такте позволяет увеличить время между двумя последовательными командами в одном треде и компенсировать задержки вычисления результата в любом ИУ за исключением выборки данных из глобальной памяти, реализованной на динамической памяти. Как правило, один тред вычисляет один элемент в массиве результата, и все SM выполняют одну и ту же программу параллельной обработки данных. Последние поколения ускорителей вдобавок к массиву обычных ИУ имеют в своём составе и специальные ядра для работы с многомерными массивами (на практике матрицами) – тензорами. Такие ядра не столь универсальны в программировании, в сравнении с обычными, но на совмещённых операциях умножения и сложения (Fused Multiply Add) небольших матриц производительность этих ядер может значительно превосходить производительность всех классических ИУ ускорителя. Такие ядра используются как в области высокопроизводительных вычислений, так и в области машинного обучения.

Графические ускорители обладали огромным количеством процессорных ядер и огромной пиковой производительностью, начиная со времён их использования в качестве ускорителей вычислений [1,2], однако, на вычислительных задачах, максимальная реальная производительность графических ускорителей (GPU), как правило,

не превышает 50% от пиковой производительности. Одним из основных узких мест, затрудняющих повышение производительности связки процессор + ускоритель является недостаточная пропускная способность памяти. Из-за больших накладных расходов на пересылку данных между памятью центрального процессора (CPU) и памятью GPU для достижения максимальной производительности в GPU нужно чтобы обрабатываемый массив содержал тысячи элементов. А на малых длинах векторов, порядка пары сотен элементов, ускорения при выполнении программы на GPU вообще может не быть [3].

Для нивелирования этой проблемы в последних поколениях графических ускорителей, ориентированных на установку в вычислительных центрах, разработчики перешли с ранее традиционной для графических ускорителей памяти GDDR на память High Bandwidth Memory (HBM). В HBM с помощью сквозных кремниевых межсоединений (through-silicon vias (TSVs)) и микроконтактных выводов (microbumps) объединяются несколько кристаллов динамической памяти (включая опциональную базовую схему контроллера памяти) с дополнительным базовым кристаллом, который может включать в себя буферную схему и тестовую логику. За счёт этого удаётся обеспечить более высокую пропускную способность при меньшем расходе энергии и существенно меньших размерах по сравнению с классическими вариантами с DDR и GDDR [4], вместе с тем в HBM используется более широка шина данных, чем в GDDR. Так, в памяти стандарта HBM2E предусмотрено до 12 кристаллов памяти в вертикальном стеке общим объёмом до 24ГБ и шиной шириной 1024 бит с пропускной способностью 307ГБ/с на стек [5].

3. Обзор современных и перспективных ускорителей для HPC

На данный момент имеются три основных производителя, выпускающих ускорители для сферы высокопроизводительных вычислений, это компания Intel со своей линейкой продуктов Data Center GPU Max, NVIDIA с линейкой Data Center GPUs (до 2020 называлась Tesla), и AMD с линейкой Instinct.

Компания Intel выпустила в 2023 году ускоритель для вычислительных центров Data Center GPU Max 1550 на микроархитектуре с кодовым названием Ponte Vecchio [6], который содержит 128 X^e ядер. Каждое такое ядро содержит в себе 8 векторных исполнительных устройств, с длиной вектора 512 бит, и 8 матричных вычислителей, с размером матрицы 4096 бит, а также кэш первого уровня размером 512КБ и 8-и уровневый систолический массив. Производительность одного такого векторного арифметико-логического устройства (АЛУ) - 256 ФЛОП/такт над числами двойной точности (FP64), такая же производительность и в случае одинарной точности (FP32), либо удвоенная над числами половинной точности (FP16). Матричный же вычислитель не поддерживает операции над числами двойной точности, но может показать производительность 4096 операций в такт над числами половинной точности (FP16) или (BF16), 2048 операций в такт над TF32 и 8192 над INT8, т.е. матричный вычислитель более нацелен на задачи искусственного интеллекта и машинного обучения, нежели на высокопроизводительные вычисления. Шестнадцать таких ядер вместе с 16 устройствами трассировки лучей объединяются в один слайс (slice). Такие слайсы по 4 штуки вместе с кэш второго уровня и четырьмя контроллерами памяти HBM объединяются в стек и на кристалле находятся два таких стека [6,7]. Этот ускоритель имеет теоретическую производительность 52.43 ТФЛОПС на числах с двойной точностью при частоте 1600МГц (базовая частота 900МГц) и пропускную способность к 128ГБайтам памяти стандарта HBM2e в 3276.8 ГБ/с через шину шириной 8Кбит [8]. Также имеется возможность объединения до 8 GPU посредством высокоскоростных линий связи X^e Link и встроенного полного коммутатора. С процессором ускорители общаются через шину PCI-Express.

Более десяти тысяч вычислительных узлов из шести таких ускорителей, совместно с двумя процессорами серии Intel Xeon Max, установлены в суперкомпьютер Aurora, имеющий пиковую производительность более 2х ЭФЛОПС [9]. Надо отметить, что подобные процессоры имеют встроенную память HBM2e в размере 64ГБ, которая может использоваться и в качестве кэш памяти, и в качестве оперативной [10].

У компании AMD наиболее производительным ускорителем на данный момент является Instinct MI250X, выпущенный в ноябре 2021 года на архитектуре CDNA 2 [11]. В данном ускорителе объединены два кристалла по 110 вычислительных блоков каждый. В каждом таком вычислительном блоке имеется по 4 матричных вычислителя и 64 шейдерных ядра (4xSIMD16)[11], вместе с планировщиком задач и 16КБ кэш 1-ого уровня. На кристалле также имеется 8МБ кэш второго уровня (суммарно 16МБ на ускоритель). Все АЛУ поддерживают вычисления с двойной точностью, обеспечивая ускорителю производительность в 47.9 ТФЛОПС при пиковой частоте 1700МГц (базо-

вая частота 1000МГц). Такую же пиковую производительность InstinctMI250X имеет на числах с одинарной точностью (FP32), а на половинной FP16, bfloat16, а также INT8 пиковая производительность 383 ТФЛОПС. AMD также заявляет матричную производительность в 95.7 ТФЛОПС (FP64/FP32). Ускоритель имеет два блока по 64ГБ (один кристалл на архитектуре CDNA 2 может адресовать до 64ГБ) памяти HBM2e с 4КБитной шиной к каждому и суммарную пропускную способность в 3276.8 ГБ/с. Ускоритель имеет 8 каналов связи InfinityFabric (16бит на дуплексный канал, 50Гбит/с в каждом направлении) для связи с другими ускорителями/процессорами или шиной PCI-ExpressGen4 [11]. Наиболее эффективно эти линии связи могут использоваться в связки со специально адаптированными процессорами AMD EPYC, поддерживающими эту технологию, все остальные процессоры будут связываться по PCI-Express.

Подобные ускорители установлены во многих суперкомпьютерах, в частности в машинах HPE Cray EX235a, которые в разных вариациях, в зависимости от количества установленных вычислительных узлов, занимают верхние строчки в списках GREEN500, TOP500 и HPCG [12] (на данный момент установка Frontier из Ок-Риджской национальной лаборатории занимает первое место с производительностью на тесте Linpack 1,194.00 ПФЛОПС) [13].

В планах у AMD стоит выпуск ускорителей следующего поколения серии MI300. На момент написания статьи известно лишь, что это будет не просто ускоритель, а самодостаточный вычислительный узел, включающий в себя 24 универсальных ядра Zen 4 и неизвестное пока количество ядер графического ускорителя на архитектуре CDNA 3 с поддержкой памяти стандарта HBM3. Особенности архитектуры CDNA3 и количество ядер графического ускорителя компания AMD пока не раскрывает, известно лишь, что будет две версии ускорителя: Instinct MI300A с 24 ядрами Zen4, 128ГБ памяти HBM3 с общей памятью CPU+GPU и Instinct MI300X со 192ГБ памяти HBM3. Вычислительные узлы поколения MI300 уже устанавливаются в суперкомпьютер ElCapitan для Ливерморской национальной лаборатории, с ожидаемой производительностью свыше двух ЭФЛОПС. Запуск суперкомпьютера перенесли с 2023 на 2024 год, но до сих пор ожидается, что он займёт первую строчку в рейтинге TOP500 [14].

Компания NVIDIA стояла у истоков использования графических ускорителей в суперкомпьютерах и во многих машинах из первого десятка в списках TOP500 и GREEN500 используются ускорители этой фирмы разных поколений (в частности, суперкомпьютер Lenovo ThinkSystem SR670 V2 (Henri) с ускорителями H100 PCI-еанимает первое место в списке GREEN500) [15]. В 2022 году компания выпустила ускорители на микроархитектуре Hopper (H100), и в 22-23 годах на микроархитектуре с кодовым названием Ada Lovelace (L4 и L40).

Потоковый мультироцессор(SM) ускорителя H100 состоит из четырёх блоков, в каждом из которых имеется 16 АЛУ для арифметики над числами двойной точности, 32 для FP32, 16 для INT32, тензорному ядру четвёртого поколения, способному работать с форматами чисел FP8,FP16, BF16, TF32, FP64 и INT8, а также кэш команд уровня L0, варп-планировщик [16] и регистровый файл на 64КБ. Все четыре модуля имеют выход на раздельный кэш для команд и данных уровня L1 и модуль Tensor Memory Accelerator (TMA), для передачи больших блоков данных и многомерных тензоров между разделённой памятью и глобальной [17]. Пара таких мультипроцессоров объединяется в кластер обработки текстур (TPC). Теоретически, архитектура предусматривает до 72 TPC на кристалле, но существует два варианта ускорителей, с несколько разнящимися характеристиками. В варианте для PCIe 5-ого поколения с 57 TPC с пятью стеками памяти HBM2e общим объёмом 80ГБ и в варианте SXM5 с 66 TPC и пятью стеками памяти HBM3 такого же объёма. Примечательно, что только 2 TPC из всего массива способны работать с графикой, т.е. архитектура спроектирована именно для вычислительных центров. Оба ускорителя имеют по 50МБ кэш второго уровняи поддерживают каналыдуплексные каналы связи по 25ГБ/с в каждом направлении NVLink 4-ого поколения (600ГБ/с у версии PCI-e и 900ГБ/с у SMX5) и PCI-e пятого поколения. Для H100 SMX заявлена производительность 33,5 ТФЛОПС (FP64) и 66,9 ТФЛОПС (FP64 на тензорах) при пропускной способности к памяти в 3,35ТБ/с, а для H100 PCI-e 25,6 ТФЛОПС и 51,2 ТФЛОПС при 2ТБ/с соответственно [18].

Ускоритель для вычислительных центров L40 на самом большом кристалле (AD102) семейства Ada Lovelace имеет 18176 CUDA ядер против 16896 у H100 SMX. При этом производительность его на числах FP64 всего 1414 ГФЛОПС. Это объясняется разностью в организации потоковых мультипроцессоров – там всего по два АЛУ для FP64 на SM, и те, как заявляет сама NVIDIA, оставлены для поддержки кода с командами над FP64 [19]. Ускоритель имеет 576 тензорных ядер 4-ого поколения и 48Гбайт памяти стандарта GDDR6 с суммарной пропускной способностью 864 ГБ/с. Судя по более низ-

ким показателям производительности над числами с двойной точностью, ускорители этого семейства (речь о продуктах, предназначенных для вычислительных центров, а не о традиционных графических ускорителях GeForce) направлены более на графические приложения и задачи машинного обучения, чем на высокопроизводительные вычисления.

Таблица 1. Пиковая производительность графических ускорителей в ТФЛОПС

	NVidia H100 SXM5	**AMD MI250X**	**Intel X^e**
FP64	33.5	47.9	52.43
FP32	66.9	47.9	52.43
FP16	133.8	383	52.43
INT8	н/д	383	1664
Tensor FP64	66.9	95.7	Не поддерживает
Tensor FP32	494.7	95.7	419.4 (формат TF32)
Tensor FP16	989.4	383	838.9
Tensor INT8	1978.9	383	1677.7

Как видно из таблицы 1, нет однозначно более производительного ускорителя по всем параметрам. Также видно, что фирма NVIDIA утрачивает лидирующие позиции среди ускорителей для суперкомпьютерных центров, и простым увеличением пиковой производительности одного лишь ускорителя это теперь вряд ли будет легко исправить, т.к. Intel и AMD производят ещё и процессоры общего назначения, т.е. они могут применять, и применяют, в своих связках CPU+GPU дополнительные технологии, позволяющие повысить производительность вычислительного узла в целом (как то быстрые каналы связи, позволяющие повысить производительность связки из нескольких CPU и нескольких GPU). Пока ещё на верхних строчках рейтингов TOP500 и GREEN500 находится много установок с ускорителями NVIDIA, но если не будет предпринято ничего кардинального, эта ситуация изменится в худшую для NVIDIA сторону. Получается, что все три фирмы продвигают свою экосистему для вычислений, состоящую из средств программирования (CUDA, OpenMP/ROC и DC++/IntelOneAPI), ускорителей и процессоров общего назначения, только у фирмы NVIDIA нет своего процессора общего назначения, а конкурировать на два фронта будет значительно сложнее. Компания NVIDIA решила пойти по тому же пути, что и AMD, представив следующее поколение (с кодовым названием Grace Hopper) не просто ускорителей, но полноценных вычислительных узлов, сочетающих в себе 72 ядра процессоров общего назначения на архитектуре Arm Neoverse V2 и ядер своих графических ускорителей на уже известной архитектуре Hopper с поддержкой памяти стандартов HBM3 и HBM3e [20,21]. В этом вычислительном узле обещано уравнять пропускную способность GPU-GPU и CPU-GPU (по 900ГБ/с) для задач с большой загрузкой как GPU, так и CPU, также заявлена поддержка общего адресного пространства у CPU и GPU, что облегчит программирование, сократит количество пересылок данных и позволит повысить реальную производительность. Кроме того, заявлена поддержка ускорения кода на языках C++, Fortran и Python.

Как мы видим, имеется тенденция на увеличение пропускной способности к памяти – многие ускорители поддерживают память HBM (у Intel поддерживает также процессор Xeon Max), выпускаются специализированные модули (OAM,SXM) с компактным размещением ускорителей и памяти для сокращения линий связи, уже применяется память стандарта HBM3, объявлена поддержка памяти HBM3e, имеющей пропускную способность 8Гбит/с на контакт. В будущем использование памяти HBM-PIM, разработку которой анонсировала компания Samsung [22], позволит сократить количество обращений к памяти, за счёт организации простейших арифметических операций в самой памяти. Пока предполагается использовать данную технологию в задачах машинного обучения [23]. Идея наделить память возможностью самостоятельно выполнять простейшие арифметические действия не нова и предлагалась в своё время в ИППМ РАН [24] при разработке процессора с архитектурой управления потоком данных и, на самом деле, может найти применение и в ускорителях вычислений в машинах с традиционной архитектурой для снижения количества обращений в память, ускорения синхронизации и т.д., но пока о таких планах компаний-производителей мне не известно.

4. Заключение

Хочется отметить, что основной тенденцией развития ускорителей для суперкомпьютерных центров является повышение эффективности работы с памятью. Уже ушли от использования чисто графических ускорителей для построения суперкомпьютеров. Более того, на смену про-

стой связке CPU+GPU приходят специализированные вычислительные узлы, совмещающие в себе память с высокой пропускной способностью, универсальные вычислительные ядра и простые ядра ускорителей. И это происходит не столько ввиду технических ограничений, хотя имеются плюсы в близком расположении логики, организующей вычисления, основных вычислительных мощностей и памяти, сколько ввиду причин экономических – производители получают конкурентное преимущество, предлагая сразу законченную экосистему из программных и аппаратных решений.

Работа была выполнена в МСЦ РАН в рамках государственного задания по теме FNEF-2022-0016. В исследованиях использовался суперкомпьютер МВС-10П.

Graphics Accelerators for High-Performance Computing

A. S. Shmelev

Abstract. GPUs, built from a large number of relatively simple an similar execution units and having high peak performance, have been used long time in high-performance computing as accelerators. Nowday there are GPUs that were specially designed for use in supercomputer centers, and not as graphics processors.

Current trends in their development are presented.

Keywords: High performance computing (HPC), Graphics processing unit (GPU), Accelerated Processing Unit (APU), High Bandwidth Memory (HBM)

Литература

1. Lal Shimpi, Anand; Wilson, Derek (June 16, 2008). "Lots More Compute, a Leetle More Texturing - Nvidia's 1.4 Billion Transistor GPU: GT200 Arrives as the GeForce GTX 280 & 260". Anandtech.com. Retrieved December 11, 2015. https://www.anandtech.com/show/2549/3

2. Radeon R5xx Acceleration Revision 1.1 February 22, 2008 © 2008 Advanced Micro Devices, Inc.

3. D.B. Kirk, W.W. Hwu. Programming Massively Parallel Processors. A Hands-on Approach, Morgan Kaufmann Publishers, 2010, 258 p.

4. HBM: MemorySolutionforBandwidth-HungryProcessors, JoonyoungKimandYounsuKim, SKhynix // Hot Chips 26, August 2014

5. JEDEC Updates Groundbreaking High Bandwidth Memory (HBM) Standard (Пресс релиз). JEDEC. 2018-12-17.

6. Intel Data Center GPU Max Series Technical Overview https://www.intel.com/content/www/us/en/developer/articles/technical/intel-data-center-gpu-max-series-overview.html Обращение сентябрь 2023.

7. Cutress, Ian (August 24, 2021). "Hot Chips 2021 Live Blog: Graphics (Intel, AMD, Google, Xilinx)". AnandTech. Retrieved August 29, 2021.

8. Intel® Data Center GPU Max 1550 Specifications https://www.intel.com/content/www/us/en/products/sku/232873/intel-data-center-gpu-max-1550/specifications.html Обращение сентябрь 2023.

9. Argonne installs final components of Aurora supercomputer. Пресс релиз 22.06.2023 Jim Collins https://www.anl.gov/article/argonne-installs-final-components-of-aurora-supercomputer

10. Intel Xeon MAX 9480 Deep-Dive 64GB HBM2e Onboard Like a GPU or AI Accelerator. Patrick Kennedy 12.09.2023 https://www.servethehome.com/intel-xeon-max-9480-deep-dive-intel-has-64gb-hbm2e-onboard-like-a-gpu-or-ai-accelerator/3/

11. INTRODUCING AMD CDNA™ 2 ARCHITECTURE © 2021 Advanced Micro Device https://www.amd.com/system/files/documents/amd-cdna2-white-paper.pdf

12. https://www.top500.org/lists/ Обращение сентябрь 2023.

13. https://top500.org/lists/top500/2023/06/ Обращение сентябрь 2023.

14. https://www.anandtech.com/show/18946/el-capitan-installation-begins-first-apu-exascale-system-shaping-up-for-2024 Обращение сентябрь 2023

15. https://top500.org/lists/green500/2023/06/ Обращение сентябрь 2023.

16. Константин Калгин. Аппаратная архитектура CUDA. Новосибирский национальный иссле-

довательский государственный университет, Новосибирск, 2013. https://ssd.sscc.ru/sites/default/files/content/attach/332/cuda-3-arch.pdf

17. NVIDIA H100 Tensor Core GPU Architecture whitepaper V1.04 © 2023 NVIDIA Corporation. https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

18. NVIDIA H100 Tensor Core GPU Datasheet © 2023 NVIDIA Corporation and affiliates. https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

19. NVIDIA Ada Lovelace Professional GPU Architecture Whitepaper v1.1 © 2023 NVIDIA Corporation. https://images.nvidia.com/aem-dam/en-zz/Solutions/technologies/NVIDIA-ADA-GPU-PROVIZ-Architecture-Whitepaper_1.1.pdf

20. NVIDIA GH200 Grace Hopper Superchip Architecture Whitepaper V1.11 © 2023 NVIDIA Corporation. https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

21. NVIDIA Grace Hopper Superchip Data Sheet © 2023 NVIDIA Corporation. https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip

22. Samsung Develops Industry's First High Bandwidth Memory with AI Processing Power (Пресс Релиз 17.02.2021) https://news.samsung.com/global/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power

23. Samsung Processing in Memory Technology at Hot Chips 2023 By Patrick Kennedy August 28, 2023 https://www.servethehome.com/samsung-processing-in-memory-technology-at-hot-chips-2023/

24. Яхонтов Д.Е., Левченко Н.Н., Окунев А.С. Принципы работы блока специальных операций модуля ассоциативной памяти параллельной потоковой вычислительной системы ППВС // Материалы Международной научно-технической конференции «Суперкомпьютерные технологии: разработка, программирование, применение» (СКТ-2010), Таганрог – Москва, 2010, Т. 1, С. 166-170.