Современная электроника и искусственный интеллект

Часть 3. Новые процессорные модули ведущих производителей для систем искусственного интеллекта

Виктор Алексеев

Основное направление разработок новых аппаратных средств, предназначенных для искусственного интеллекта, связано с поисками возможного компромисса между увеличением вычислительных мощностей и энергопотреблением. Эксперты отмечают, что в настоящее время наблюдается отход от доминировавшей несколько последних лет тенденции наращивания вычислительных мощностей в центрах обработки данных. В противовес этому конструкторы стараются найти решения, позволяющие конструировать новые ИИ-модели более продвинутыми, но с меньшими затратами. В статье описаны новые электронные компоненты для моделей ИИ трёх лидирующих на этом рынке компаний: NVIDIA, AMD и Cerebras.

Введение

Если оценивать в общих чертах новинки аппаратной части систем искусственного интеллекта (АПИИ), то можно выделить две основные линии развития: оптимизацию энергопотребления и снижение затрат на разработку и эксплуатацию [1].

Лидеры индустрии производства АПИИ стараются усовершенствовать свои модели за счёт лучшей пропускной способности памяти и увеличенных скоростей обработки данных. В основном это достигается с помощью использования передовых технологий [2].

В то же время флагманские ускорительные модули, например NVIDIA (B200, B300), обладая огромной вычислительной мощностью, потребляют значительное количество энергии и не всегда оптимальны для решения специфических задач ИИ [3].

Несмотря на доминирование нескольких крупных компаний, рынок аппаратных решений ИИ остаётся динамичным и открытым для инноваций. Этот факт можно объяснить несколькими причинами. Лидеры рынка больше заинтересованы в расширении своей существующей экосистемы, чем в радикальных инновациях, потенциально подрывающих их основной бизнес. Разработка новых архитектур требует огромных инвестиций в НИОКР, что ограничивает скорость инноваций у крупных игроков. Поэтому такие концерны, как NVIDIA и AMD, фокусируются на высокодоходных сегментах центров обработки данных. При этом многие другие направления, например периферийные вычисления и специализированные решения, остаются недостаточно разработанными.

Всё вышеперечисленное создаёт прекрасные стартовые условия для многочисленных начинающих фирм, старающихся заполнить образовавшиеся ниши. Важно то, что такие небольшие фирмы стараются создавать решения «под ключ», включающие как аппаратное, так и программное обеспечение. Такой подход позволяет стартапам получить определённые преимущества по сравнению с универсальными решениями флагманских концернов этой индустрии [4].

В целом аппаратное обеспечение ИИ становится быстрее, дешевле и экономичнее. Новые электронные компоненты и блоки, появившиеся в 2024–2025 годах, значительно расширяют возможности ИИ, улучшая производительность, эффективность и компактность компьютеров и вычислительных центров.

Современная ситуация и неутешительные прогнозы вызывают необходимость создания нового поколения энергоэффективных электронных компонентов, предназначенных для искусственного интеллекта.

Новые ускорители графических процессоров корпорации NVIDIA

Корпорация NVIDIA на сегодняшний день считается мировым лидером в разработке и производстве ускорителей графических процессоров, мощных серверов и кластеров для

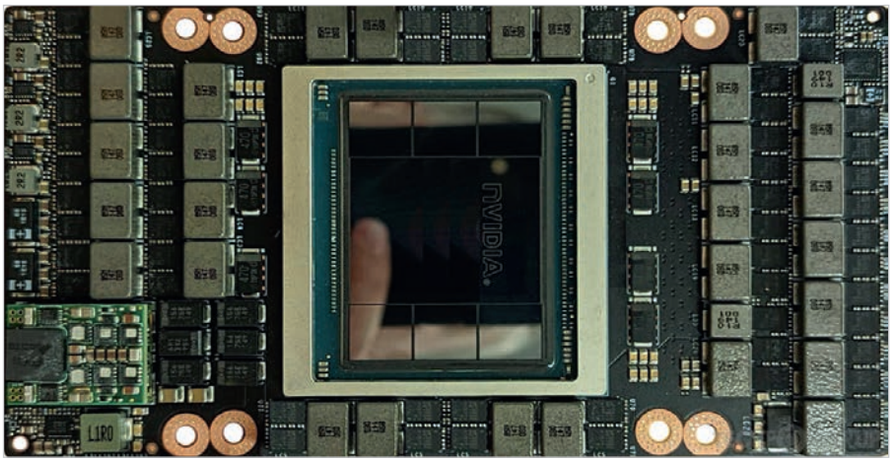

Рис. 1. Ускоритель графического процессора B100 выпускается в форм-факторе SXM5

ИИ дата-центров. О доминировании NVIDIA можно судить по объёму продаж в этом сегменте рынка. Рыночная капитализация (FY) на конец финансового 2025 года (май 2025 года) составляла около трёх с половиной триллионов долларов США. Эта цифра сравнима с бюджетами таких развитых стран, как, например, Канада. Причём по сравнению с 2022 годом рост FY составил примерно 230%, что напрямую связано со взрывным характером развития искусственного интеллекта [5].

Подробно историю развития этой фирмы и описание всей номенклатуры выпускаемой продукции можно найти, например, на сайте [6].

Ускоритель NVIDIA B100 AI B200 Accelerator компания NVIDIA представила в 2024 году в рамках рекламы следующего поколения микроархитектуры графических процессоров Next-Generation B200 Microarchitecture Blackwell [7].

Имеет смысл пояснить название этого устройства, которое чаще всего употребляют сами изготовители. Термин «B200 Accelerator» обозначает устройство, которое позволяет увеличить производительность графического процессора, используя возможности параллельной обработки в дополнение к центральному процессору CPU. В англоязычных статьях часто оставляют только слово «Accelerator». Однако лучше полностью переводить этот термин как «ускоритель графического процессора» – УГП.

Ускоритель графического процессора NVIDIA B100 представляет собой базовую версию архитектуры Blackwell, специально разработанную для организаций, которые хотят модернизировать свои существующие системы без значительных изменений в инфраструктуре энергоснабжения и охлаждения [8].

Ускоритель ГП B100 выпускается в форм-факторе SXM5, совместимом с предыдущей моделью H100 по интерфейсу и мощности потребления (700 Вт), что позволяет использовать его в качестве переходной модели на новую технологию (рис. 1).

УГП B100 совместим с предыдущей моделью H100.

При этом у B100 производительность (FP8 Tensor) 4500 Тфлопс значительно выше, чем у H100 (2000 Тфлопс для INT8 или 1000 Тфлопс для FP16 Tensor Core) [9].

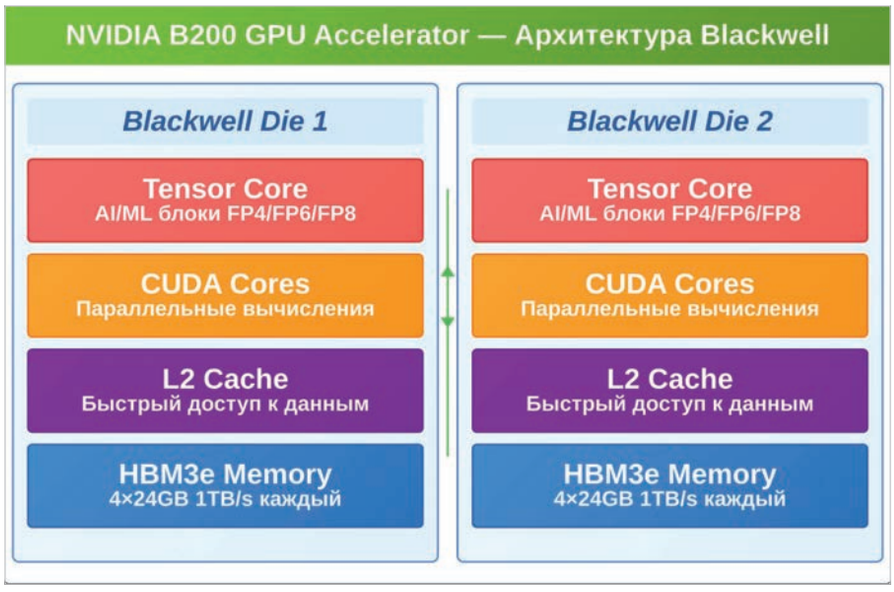

Рис. 2. Упрощённая схема архитектуры УГП «Blackwell B200 AI B200 Accelerator»

Специализированный форм-фактор Server PCI Express Module – SXM разработан NVIDIA для высокопроизводительных вычислений в центрах обработки данных. Это модуль с прямым разъёмом, обеспечивающий лучшую подачу питания, охлаждение и более высокую пропускную способность по сравнению с традиционными графическими PCIe-картами.

Ускоритель ГП обладает 192 Гбайт памяти HBM3e с пропускной способностью до 8 Тбайт/с, что в 2,4 раза превышает пропускную способность памяти H100. Эта существенная разница особенно важна для инференса больших языковых моделей, где производительность часто ограничена именно пропускной способностью памяти [10].

NVIDIA Blackwell B200 AI B200 Accelerator – новый ускоритель графического процессора, который поступил в коммерческую продажу в 2025 году. Полные физические параметры нового ускорителя графического процессора B200 не разглашаются в открытых источниках. Известно только, что этот УГП выполнен в стандарте 6-го поколения NVIDIA Server PCI Express Module – SXM6 [11].

Понять, насколько большой B200, можно по описанию предыдущей модели H100, которая имеет габариты 260×110×347 мм и вес 1,7 кг [12].

Новая архитектура Blackwell является преемником предыдущих версий NVIDIA Hopper и Ada Lovelace, однако отличается от них как по назначению, так и по техническому исполнению [13].

Несмотря на то что версия Hopper (H100/H200) также построена на 4-нм техпроцессе, она рассчитана на меньшее количество транзисторов: около 80 миллиардов в H100. Для более ясного представления о преимуществах Blackwell следует напомнить, что означают используемые в ИИ форматы данных. Числовые форматы FP16, FP8, FP4 определяют соответственно 16-битное, 8-битное и 4-битное числа с плавающей точкой.

Специальный формат BF16 – это упрощённая версия FP16, разработанная для задач ИИ, в которых можно отбрасывать мелкие детали. Другой особый ИИ-формат INT8 предназначен для работы с числами без дробной части.

Архитектура Blackwell поддерживает форматы FP16, FP8, FP4 и INT8, что даёт возможность ускорить вычисления с меньшей точностью, не критичной для ИИ. Так, например, Blackwell позволяет достичь 4-кратного прироста производительности для инференса LLM по сравнению с Hopper, оптимизированного для FP16 и BF16.

Главное архитектурное отличие архитектуры Blackwell, по сравнению с предыдущей версией NVIDIA Hopper H100, заключается в том, что B200 построен по схеме двух кристаллов (Dual-Die), соединённых высокоскоростным интерфейсом NV-HBI (NVIDIA High Bandwidth Interface), что позволяет им функционировать как единый ускоритель ГП.

Максимальная площадь кристалла для технологии TSMC 4NP составляет примерно 860 мм². В конструкции старой версии ускорителя ГП H100 была использована практически вся

доступная площадь (814 мм²) базового кристалла. Поэтому в ускорителях B100/B200 использовано два кристалла, позволивших удвоить вычислительную мощность.

Интерфейс NV-HBI с пропускной способностью 10 Тбайт/с обеспечивает полную когерентность кэша, единое адресное пространство памяти и прозрачность для программного обеспечения.

Упрощённая схема архитектуры УГП Blackwell B200 AI B200 Accelerator показана на рис. 2.

Технология Tensor Core NVIDIA предназначена для ускорения вычислений с использованием смешанной точности, что особенно важно в задачах искусственного интеллекта. По существу, Tensor Cores представляют собой специализированные вычислительные ядра в ускорителе B200, предназначенные специально для работы с ИИ.

Особенность Tensor Cores заключается в том, что они могут выполнять сразу много параллельных операций, например, одновременно умножать целые матрицы за один такт. Кроме того, эти ядра используются для операций с низкой точностью (FP16, FP8, FP4), обучения нейронных сетей, а также для инференса AI-моделей. По сравнению с предыдущими версиями УГП ядра в Blackwell Ultra Tensor Cores обеспечивают ускорение слоёв внимания (Attention-Layer) в два раза больше (для формата FP8) и в полтора раза больше для операций с плавающей запятой (FLOPS) [2].

Кроме ядер Tensor Cores, ответственных за сложные параллельные вычисления, в структуре Blackwell B200 существуют ядра CUDA Cores, выполняющие простые последовательные математические операции. Кроме обычной арифметики (сложение, умножение, деление) CUDA Cores используются для тривиальных графических вычислений, физической симуляции, криптографии, а также для любых других вычислительных задач, которые можно решать по шагам.

Высокоскоростная буферная L2 кэш-память (L2 Cache Memory) предназначена для ускорения доступа к часто используемым данным, что снижает задержки при обращении к основной памяти. Кроме того, она может быть использована как когерентный кэш между двумя кристаллами.

Кэш L2 в ускорителях графических процессоров NVIDIA является важным компонентом, повышающим производительность, особенно в задачах глубокого обучения.

В новых ускорителях B200 NVIDIA используется High Bandwidth Memory 3E Stacks (HBM3e), сверхбыстрая основная память. Эта память в конфигурации 8 стеков по 24 Гбайт каждый (всего 192 Гбайт) имеет пропускную способность 8 Тбайт/с (по 1 Тбайт/с на стек). В качестве интерфейса используется 1024-битная шина данных на каждый стек.

В ускорителях ГП B200 применяется технология ускорения работы моделей NVIDIA Transformer Engine TE 2.0 второго поколения, которая позволяет автоматически выбирать оптимальный формат. Важно то, что новая версия поддерживает форматы низкой точности, такие как FP6 и FP4, что даёт возможность в ряде случаев значительно повышать эффективность обработки данных. Эта технология предоставляет инструменты для повышения производительности глубокого обучения GLM/LLM. Например, в её комплекте есть библиотека с открытым исходным кодом NVIDIA TensorRT-LLM, позволяющая создавать и оптимизировать LLM, используя простой API на Python.

Основные технические характеристики новой модели B200 в сравнении с предыдущей версией H100 приведены в табл. 1.

Из других важных особенностей Blackwell B200 следует отметить систему безопасности Trusted Execution Environment (TEE-I/O), которая обеспечивает защиту ИИ-моделей и интеллектуальной собственности, поддерживая конфиденциальную вычислительную среду как на самом ускорителе ГП, так и на хостах [14].

Ускоритель графического процессора B200 является в настоящее время флагманской моделью, на основе которой NVIDIA производит серверные платы, законченные серверы, кластерные серверные стойки.

УГП Blackwell B200 выпускаются в форм-факторе SXM6, который рассчитан на более высокие значения мощности и пропускной способности. Этот модуль SXM6 с памятью 192 Гбайт HBM3e и высокой скоростью обмена через NVLink ориентирован на классические серверные платформы.

Данный стандарт NVIDIA SXM6 поддерживает как воздушное, так и жидкостное охлаждение.

NVIDIA B300 Blackwell Ultra – это следующая модель ускорителей графических процессоров, которую NVIDIA анонсировала в марте 2025 года, и ожидается эта модель к выпуску во второй половине 2025 года [15].

Официального детального технического описания для B300 от NVIDIA пока нет. На сайте NVIDIA есть только общая информация о продуктах B300, DGX B300 и GB300, но без точных технических спецификаций.

Поэтому ниже приведена только краткая информация, доступная на сегодняшний день на других сайтах партнёров NVIDIA.

Новый УГП B300 предназначен для инференса сложных задач, когда ИИ должен сначала «подумать», проанализировать несколько вариантов решения, а потом выбрать лучший.

Этот ускоритель ГП B300 изготовлен по той же технологии TSMC 4NP, что и рассмотренный выше B200. Однако

Таблица 1. Основные технические характеристики B200 и H100

Характеристика	NVIDIA B200	NVIDIA H100
Архитектура	Blackwell	Hopper
Год выпуска	2025	2022
Техпроцесс	TSMC 4NP (enhanced 4N)	TSMC 4N
Количество транзисторов	208 млрд	80 млрд
Конфигурация	Двухкристальная (Dual-Die)	Однокристальная
Форм-фактор	SXM6	SXM5
Память B200	192 Гбайт HBM3e	80 Гбайт HBM3
Пропускная способность памяти	До 8 Тбайт/с	3,35 Тбайт/с
Производительность FP32	80 Тфлопс	67 Тфлопс
Тензорная производительность FP32	2,2 Пфлопс	500 Тфлопс
Тензорная производительность FP16/BF16	4,5 Пфлопс	1 Пфлопс
Поддержка FP8	Да (10 Пфлопс)	Да (2 Пфлопс)
Поддержка FP4	Да (20 Пфлопс)	Нет
Максимальное энергопотребление	1000 Вт	700 Вт
Поколение NVLink	5-е поколение	4-е поколение
Transformer Engine	2-е поколение	1-е поколение

ряд изменений в архитектуре позволил заметно улучшить его параметры.

Прежде всего, использована новая 12-слойная HBM3E память B300 Blackwell Ultra объёмом 288 Гбайт. Это критично для работы с крупнейшими языковыми моделями, которые имеют триллионы параметров [16].

Кроме того, B300 может работать с производительностью до 15 петафлопс в формате FP4. Однако при этом увеличилась мощность до 1400 Вт (TDP) [17].

В отличие от B200, которые комплектуются ConnectX-7 (до 400 Гбит/с), B300 использует ConnectX-8, где NVIDIA интегрировала дополнительные функции в сетевой интерфейс, уменьшив таким образом зависимость от сторонних чипов (например, PCIe Retimer).

Ускоритель ГП B300 разработан в новом конструктиве, получившем название SXM Puck благодаря тому, что по форме напоминает шайбу (Puck). Он объединяет УГП, высокоскоростную память HBM3e и другие компоненты в компактный, высокоэффективный блок, который легко интегрируется в серверные системы.

Этот модуль позволяет легко собирать, заменять или обновлять УГП в серверах, как конструктор LEGO. Модуль можно также вставить в большую систему, например, NVIDIA HGX, DGX или GB300 NVL72. Такой подход значительно упрощает создание и ремонт мощных ИИ-кластеров.

Увеличение мощности потребовало использования в конструктиве SXM Puck оптимизированного жидкостного охлаждения. В этой новой схеме используются жидкостные охлаждающие пластины для прямого отвода тепла от B300, а также усовершенствованные разъёмы UQD (Universal Quick Disconnect), предназначенные для быстрого подключения системы охлаждения. Специальные стойки с перепроектированными каналами охлаждения позволяют разместить больше УГП в кластере, а также интегрировать модули в такие системы, как NVIDIA MGX.

Платформа HGX B200 представляет собой первую ступеньку в линейке систем масштабирования устройств на основе ускорителей ГП Nvidia. Однако сама по себе HGX B200 – это своего рода основа будущего устройства, в которой нет центрального процессора CPU, но есть посадочные места под них, а также под другую периферию.

Отсутствие в комплекте поставки центрального процессора объясняется тем, что это даёт заказчикам возможность подобрать CPU под свои конкретные задачи с необходимой именно им архитектурой.

Таблица 2. Варианты конфигураций HGX B200 с разными количествами УГП

B200, шт.	CPU	RAM (ГБ)	Память HBM3e (ГБ)	Количество необходимых линий PCIe 5.0
1	1x Intel Xeon 6960P (72 ядра)	250	180	16–32
1	1x AMD EPYC 9124 (16 ядер)	250	180	16–32
2	1x Intel Xeon 6952P (96 ядер)	~500	360	32–64
2	1x AMD EPYC 9224 (24 ядра)	~500	360	32–64
2	2x Intel Xeon 6952P (96 ядер каждый)	~500	360	64–128
2	2x AMD EPYC 9224 (24 ядра каждый)	~500	360	64–128
4	2x Intel Xeon 6979P (120 ядер каждый)	~750	720	64–128
4	2x AMD EPYC 9655 (96 ядер каждый)	~750	720	64–128
8	2x Intel Xeon 6980P (128 ядер каждый)	~1000	1440	128–256
8	2x AMD EPYC 9965 (192 ядра каждый)	~1000	1440	128–256

Поэтому партнёрские компании, такие как, например, Supermicro, ASUS, Lenovo или DataCrunch, добавляют CPU, память и другие необходимые комплектующие детали. Как правило, HGX B200 комплектуется CPU Intel Xeon 6 6900 или AMD EPYC серии 9004/9005 [18].

Наиболее распространённый и стандартный вариант платформы HGX B200, предназначенный для максимальной производительности в задачах ИИ, обучения моделей и инференса, включает восемь УГП NVIDIA B200.

Ключевой особенностью HGX B200 является использование NVLink 5.0 с пропускной способностью 1,8 Тбайт/с между ускорителями. Это позволяет всем восьми B200 функционировать как единый вычислительный комплекс, превосходящий по техническим возможностям традиционные структуры с PCIe-соединениями.

Платформа HGX B200 поддерживает воздушное или жидкостное охлаждение (например, в Lenovo ThinkSystem SR780a V3 с водяным охлаждением Lenovo Neptune). Конфигурация 8×B200 представляет собой оптимальный вариант для обучения больших языковых моделей (LLM), аналитики данных и выполнения сверхсложных глобальных вычислительных задач [19].

Платформа NVIDIA Blackwell HGX B200 заметно превосходит предшествовавшую модель NVIDIA HGX H100 по базовым параметрам:

- 15-кратное увеличение скорости инференса в реальном времени;
- 12-кратное снижение стоимости эксплуатации и энергопотребления;
- 3-кратное ускорение обучения больших языковых моделей благодаря движку Transformer Engine второго поколения с форматом FP8 [8].

Платформа HGX B200 поддерживает современные сетевые технологии NVIDIA, включая NVIDIA BlueField-3 B3140H VPI 400GbE и NVIDIA ConnectX-7 NDR OSFP400 InfiniBand. Кроме того, использование NVIDIA Quantum-2 InfiniBand и Spectrum-X Ethernet позволяет работать с сетевыми скоростями до 400 Гбит/с, обеспечивая эффективное масштабирование на уровне датацентра.

Для менее требовательных задач или компактных систем, где не нужна такая высокая производительность, разработаны варианты с меньшим количеством ускорителей графических процессоров (1, 2, 4) [20].

В табл. 2 приведены варианты конфигураций HGX B200 с разными количествами УГП для CPU серий Intel Xeon 6 6900 и AMD EPYC 9004/9005 [21].

Важно подчеркнуть, что выбор того или иного CPU определяется количеством линий PCIe 5.0, необходимых для поддержки 1, 2, 4 и 8 ускорителей графических процессоров B200.

Платформу NVIDIA HGX B300 на базе рассмотренного выше нового УГП Blackwell Ultra SXM NVIDIA планирует выпустить в коммерческую продажу во второй половине 2025 года. Детальной технической информации пока нет. На сайте NVIDIA приведены только сравнительные технические характеристики HGX B300 и HGX B200, которые показаны в табл. 3.

Использование нового ускорителя ГП позволило значительно улучшить такие параметры HGX B300, как FP4 Tensor Core, INT8 Tensor Core, FP64/FP64 Tensor Core, общую память, пропускную способность сети.

Для ускорения вычислений и генеративного ИИ, кроме NVLink, использованы высокоскоростные соединения InfiniBand, Spectrum-X Ethernet. Помимо того, HGX B300 оснащён NVIDIA BlueField-3 DPUs, который предназна-

чен для гипермасштабируемых облаков ИИ.

О своих разработках на базе нового HGX B300 сообщили партнёры NVIDIA. Так, 29 апреля 2025 года была опубликована информация о NVIDIA HGX B300 NVL16, содержащем шестнадцать B300 [22].

NVIDIA DGX B200 – это полностью готовый к использованию сервер, который включает ускорители графических процессоров Blackwell B200, управляющий процессор x86 и всё необходимое оборудование для работы с различными задачами искусственного интеллекта. Сервер NVIDIA DGX B200 построен на платформе HGX B200 с ускорителями графических процессоров Blackwell B200 [23].

Полностью готовый к работе сервер DGX B200 имеет в своём составе серверный блок HGX B200, процессор x86 CPU, а также всю необходимую периферию. На рис. 3 показан внешний вид материнской серверной платы DGX B200.

На рис. 3 цифрами показано следующее.

1. SSD-накопители формата M.2 для ОС сервера (сверху) и BlueField-3 DPU (Data Processing Unit) (снизу) – специальный процессор для обработки сетевого трафика.
2. Интерфейс для подключения дополнительных плат расширения (сетевые карты, накопители и др.), разъём PCIe-карт для слотов 2 и 4.
3. Высокоскоростные сетевые модули для соединения серверов в кластере, QSFP-трансиверы для внешней сети ConnectX-7.
4. Сетевая карта 100Gb Ethernet (сверху) и сетевое хранилище BlueField-3 DPU (снизу), обеспечивающие подключение к внешней сети и управление сетевым трафиком.
5. Дополнительные слоты для плат расширения, разъём PCIe-карт для слотов 1 и 3.
6. Оперативная память сервера, 32x 64 Гбайт или 128 Гбайт, модули DIMM общим объёмом до 4 Тбайт.
7. Специализированные сетевые контроллеры ConnectX-7 для высокоскоростной передачи данных между ускорителями ГП.
8. Центральная плата, которая соединяет все компоненты сервера и обеспечивает их питанием.
9. Специальные кабели DensiLink для подключения внутренних сетевых карт к внешним разъёмам сервера.
10. Два CPU-процессора, которые управляют всей системой и координируют работу ускорителей ГП (обычно 2x Intel Xeon 6980P или 2x AMD EPYC 9965).

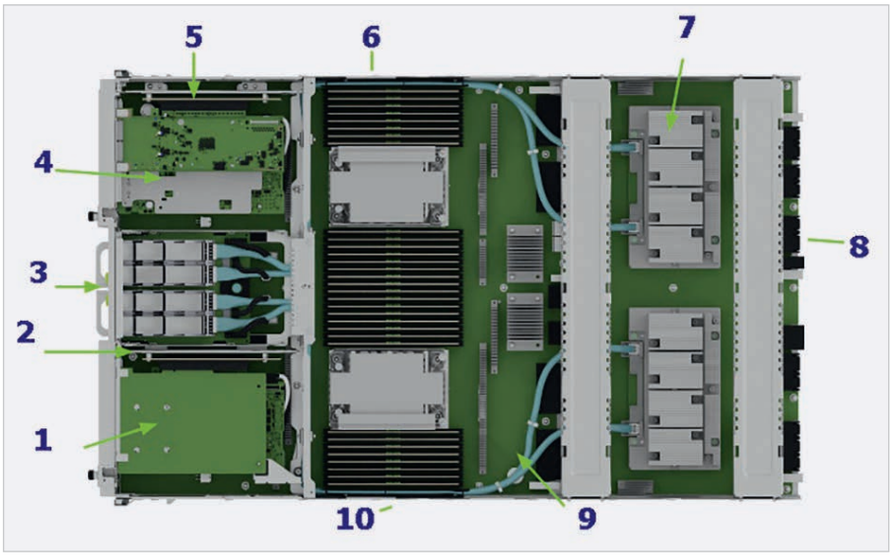

Рис. 3. Материнская плата сервера DGX B200 (пояснения приведены в тексте)

Основные технические характеристики NVIDIA DGX B200 приведены в табл. 4 [24].

В наиболее распространённом стандартном варианте – сервер, объединяющий восемь ускорителей ГП NVIDIA Blackwell B200, соединённых NVLink пятого поколения.

По сравнению с системами предыдущего поколения DGX B200 обеспечивает 3-кратное увеличение производительности обучения и 15-кратное ускорение инференса. Этот сервер DGX B200 имеет производительность до 72 петафлопс. Общий объём B200-памяти составляет 1440 Гбайт, что позволяет работать с моделями, содержащими триллионы параметров.

Важнейшим преимуществом DGX B200 является наличие встроенного полного программного стека NVIDIA AI Enterprise, включая NVIDIA Base Command и обширную экосистему поддержки третьих сторон.

Это превращает DGX B200 в готовое к использованию решение, не требующее дополнительной интеграции программного обеспечения [25].

В некоторых ситуациях полезной может оказаться поддержка современных форматов сжатия (LZ4, Snappy, Deflate), а также функция ускорения полного конвейера запросов к базам данных.

Из других важных функциональных возможностей DGX B200 следует отметить систему надёжности и диагностики RAS Engine (Reliability, Availability, Serviceability), которая обеспечивает: предиктивное управление,

Таблица 3. Сравнительные технические характеристики B300 и B200

Наименование	HGX B300	HGX B200
Форм-фактор	8x NVIDIA Blackwell Ultra SXM	8x NVIDIA Blackwell SXM
FP4 Tensor Core**	144 Пфлопс / 105 Пфлопс	144 Пфлопс / 72 Пфлопс
FP8/FP6 Tensor Core*	72 Пфлопс	72 Пфлопс
INT8 Tensor Core*	72 Пфлопс	72 Пфлопс
FP16/BF16 Tensor Core*	36 Пфлопс	36 Пфлопс
TF32 Tensor Core*	18 Пфлопс	18 Пфлопс
FP32	600 Тфлопс	600 Тфлопс
FP64/FP64 Tensor Core	10 Тфлопс	296 Тфлопс
Общая память	До 2,3 Тбайт	1,4 Тбайт
NVLink	Пятое поколение	Пятое поколение
NVIDIA NVSwitch™	NVLink 5 Switch	NVLink 5 Switch
Пропускная способность NVSwitch GPU-to-GPU	1,8 Тбайт/с	1,8 Тбайт/с
Общая пропускная способность NVLink	14,4 Тбайт/с	14,4 Тбайт/с
Пропускная способность сети	1,6 Тбайт/с	0,8 Тбайт/с
Увеличение производительности с учётом Attention Performance	2x	1x

**С разреженностью **С разреженностью / без разреженности*

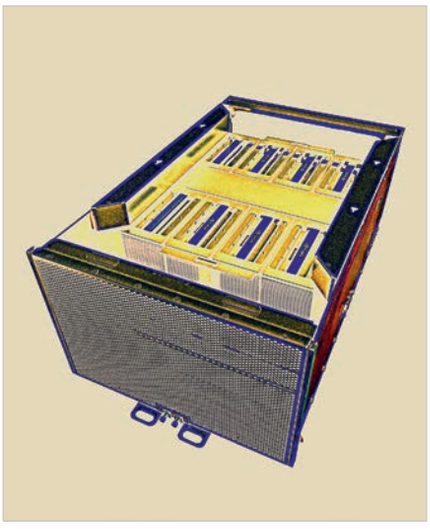

Рис. 4. Внешний вид сервера NVIDIA DGX B200 [27]SXM5

Рис. 5. Внешний вид NVIDIA DGX Station A100

мониторинг тысяч точек данных для предотвращения простоев, глубокую диагностическую информацию для планового обслуживания.

Практически сервер NVIDIA DGX B200 – это унифицированная платформа искусственного интеллекта, предназначенная для компаний любого размера. Внешний вид сервера NVIDIA DGX B200 показан на рис. 4.

Сервер DGX B300 на базе нового ускорителя ГП NVIDIA Blackwell Ultra GPUs B300 планируется начать выпускать во второй половине 2025 года. Технические характеристики DGX B300 приведены в табл. 4 [26].

NVIDIA DGX B300 – это мощная система, разработанная специально для задач искусственного интеллекта (ИИ), включая обучение, тонкую настройку и вывод (Inference) моделей. Она предназначена для предприятий, которым нужны высокопроизводительные решения для генеративного ИИ, аналитики данных и высокопроизводительных вычислений (HPC).

В настоящее время кроме DGX B200 NVIDIA выпускает целую линейку продуктов DGX, адаптированных для различных масштабов применения искусственного интеллекта.

Сервер выполнен в конструктиве 10U Rackmount. Габаритные размеры: 444×482×897 мм. Вес: 142 кг.

NVIDIA DGX Station представляет собой компактную настольную рабочую станцию, которая позволяет разработчикам и исследователям работать с проектами искусственного интеллекта без необходимости в серверной инфраструктуре. Такой подход позволяет использовать возможности датацентра прямо на рабочем столе [28].

DGX Station может служить как индивидуальная настольная система для одного пользователя, работающего с продвинутыми ИИ-моделями на локальных данных, так и в качестве централизованного вычислительного узла для нескольких членов команды. Система поддерживает технологию NVIDIA Multi-Instance B200 (MIG), позволяющую разделить ресурсы на семь изолированных экземпляров, каждый со своей высокоскоростной памятью, кэшем и вычислительными ядрами.

Особенно важно для образовательных учреждений то, что бюджетные ИИ-компьютеры, такие как NVIDIA DGX Station A100, предоставляют студентам и преподавателям доступ к передовым технологиям без необходимости строить дорогостоящую серверную инфраструктуру. Это делает изучение и разработку ИИ-приложений более доступными для начинающих пользователей. Внешний вид NVIDIA DGX Station A100 показан на рис. 5.

Последняя версия DGX Station разработана на базе суперчипа NVIDIA GB300 Grace Blackwell Ultra Desktop и обладает следующими характеристиками: 784 Гбайт единого системного пространства памяти и производительность до 20 петафлопс для ИИ-вычислений. Система использует жидкостное охлаждение для управления тепловыделением компонентов мощностью почти 1500 Вт, что позволяет поддерживать уровень шума ниже 35 дБ под нагрузкой.

GB200 Grace Blackwell Superchip представляет собой комбинированную плату ускорителя графического процессора, содержащую два B200 Tensor Core, соединённых с процессором NVIDIA Grace через NVLink (900 Гбайт/с) [29, 30].

В технических описаниях можно встретить наименование Bianca

Таблица 4. Основные технические характеристики серверов NVIDIA DGX B200 и DGX B300

Параметр	DGX B200	DGX B300
УГП	8x NVIDIA Blackwell B200	8x NVIDIA Blackwell Ultra GPUs
Память УГП	1,44 Тбайт, пропускная способность HBM3e 64 Тбайт/с	2,3 Тбайт
Производительность	72 петафлопс, FP8 обучение; 144 петафлопс, FP4 инференс	
Системы NVIDIA® NVSwitch™	2x	
Общая пропускная способность NVIDIA NVLink	14,4 Тбайт/с	
Суммарная потребляемая мощность	~14,3 кВт	
Центральный процессор	2x процессора Intel® Xeon® Platinum 8570	2x процессора Intel® Xeon® 6776P
Мониторинг и управление сетевыми устройствами	4x порта OSFP – 8 NVIDIA ConnectX-7 VPI, до 400 Гбит/с; NVIDIA InfiniBand/ Ethernet, 2x двухпортовых QSFP 112 NVIDIA BlueField-3 DPU, до 400 Гбит/с	8x портов OSFP – 8 NVIDIA ConnectX-8 VPI, до 800 Гбит/с; NVIDIA InfiniBand/ Ethernet, 2x портовый QSFP112 NVIDIA BlueField-3 DPU, до 400 Гбит/с NVIDIA
Подключение к сети Интернет	Встроенная сетевая карта с RJ45, 1 Гбит/с и контроллер базовой платы, 1 GbE RJ45 Host BMC, 1 Гбит/с, RJ45	
Носитель ОС	Два NVMe M.2 SSD объёмом 1,9 Тбайт каждый, управляемые контроллером BMC	
Встроенный накопитель	8 твердотельных SSD с интерфейсом U.2, ёмкость 3,84 Тбайт	8 твердотельных SSD с интерфейсом E1.S, ёмкость 3,84 Тбайт
Программное обеспечение	NVIDIA DGX OS; NVIDIA Mission Control; NVIDIA Base Command Manager / NVIDIA AI Enterprise; Red Hat Enterprise Linux, Rocky, Ubuntu	
Тип корпуса	10 RU	
Размеры системы	Высота: 444 мм, ширина: 482 мм, длина: 897 мм	Н/Д
Интервал рабочих температур	5–30°C	10–35°C

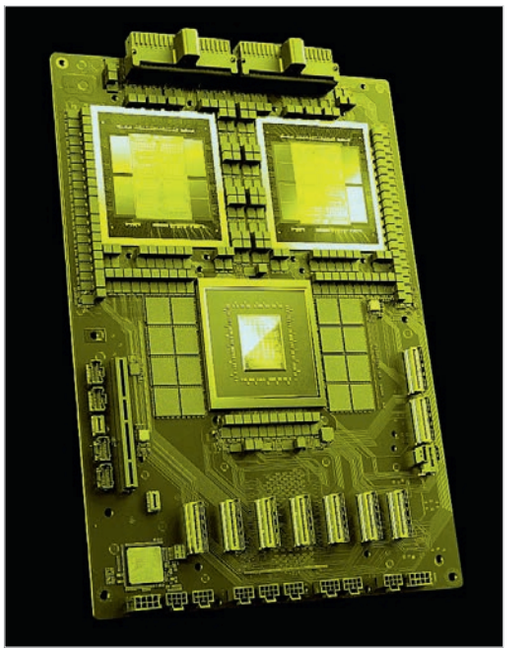

Рис. 6. Внешний вид комбинированной платы ускорителя графического процессора GB200 Grace Blackwell Superchip

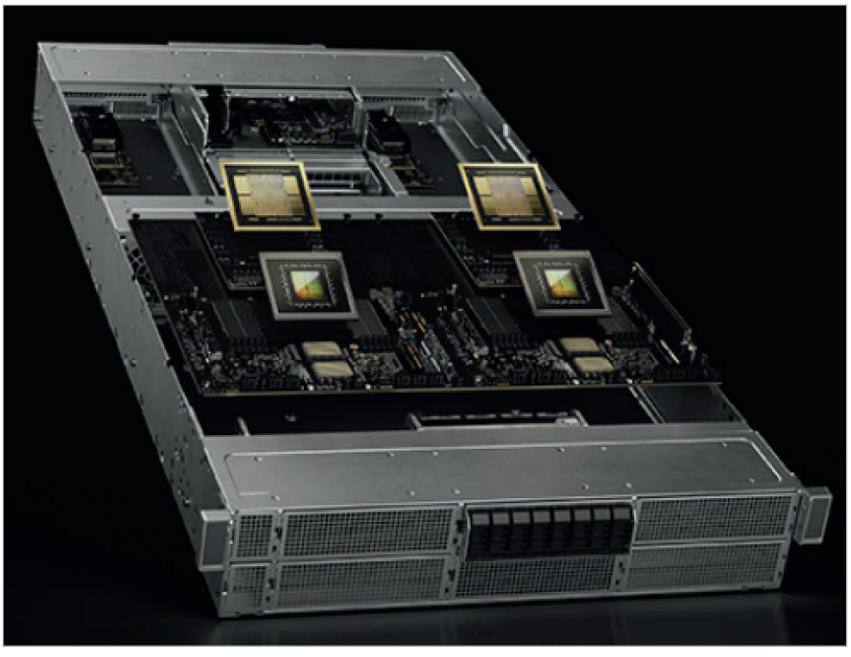

Рис. 7. Платформа GB200 NVL2 объединяет 2 ускорителя ГП B200 Blackwell и 2 процессора Grace [34]

Board, которое является внутрифирменным названием для модуля GB200 Grace Blackwell Superchip. Каждая плата Bianca содержит один процессор Grace CPU и два ускорителя графического процессора B200 [31].

Внешний вид комбинированной платы ускорителя графического процессора GB200 Grace Blackwell Superchip показан на рис. 6.

Семейство серверных процессоров NVIDIA Grace было специально разработано для дата-центров, ориентированных на обработку больших объёмов данных с высокой энергоэффективностью. Процессоры NVIDIA Grace представляют собой односокетные платформы, которые оптимизированы для масштабируемости и высокой пропускной способности. Они предназначены для сложных вычислительных задач, требующих высокой производительности.

Процессоры семейства Grace строятся на базе архитектуры Arm с ядрами Neoverse V2. В зависимости от конфигурации Grace CPU может содержать от 72 до 144 ядер. Например, в составе GB200 Superchip используется один чип Grace с 72 ядрами, а автономный Grace CPU Superchip объединяет два чипа Neoverse V2 общим количеством 144 ядра [32].

Каждый GB200 Superchip способен обеспечить 40 петафлопс разреженных операций FP4, что делает его идеальным решением для инференса сверхкрупных языковых моделей.

Суперчип создаёт единое когерентное адресное пространство памяти между CPU и B200, что кардинально упрощает программирование и повышает эффективность обмена данными. Общий объём памяти суперчипа составляет 864 Гбайт: 480 Гбайт LPDDR5x для CPU и 384 Гбайт HBM3e для B200 [33].

GB200 NVL2 – платформа для стандартных дата-центров, объединяющая 2 ускорителя Blackwell B200 и 2 процессора Grace (рис. 7).

Платформа использует масштабируемую одноузловую архитектуру NVIDIA MGX, что позволяет легко интегрировать её в существующую инфраструктуру дата-центров. Особенно эффективна NVL2 для задач поиска в векторных базах данных, генерации с расширенным поиском (RAG) и обработки данных.

Энергопотребление системы составляет примерно 2700 Вт, что делает её доступной для широкого круга дата-центров без необходимости кардинальной модернизации систем энергоснабжения и охлаждения.

GB200 NVL4 представляет собой одноплатное решение, объединяющее четыре ускорителя Blackwell B200 и два процессора Grace на единой материнской плате [35].

Система обеспечивает 768 Гбайт памяти HBM3E для GPU с суммарной пропускной способностью 32 Тбайт/с, а также 960 Гбайт памяти LPDDR5X для процессоров Grace. Все процессоры объединены через технологию NVLink, создавая единый домен с когерентной памятью объёмом 1,3 Тбайт.

По сравнению с предыдущим поколением GH200 NVL4 новая система обеспечивает 2,2-кратное увеличение производительности симуляций и 1,8-кратное ускорение обучения и инференса. Энергопотребление составляет 5400 Вт, что требует применения жидкостного охлаждения. Важной особенностью платформ NVL2 и NVL4 является соотношение процессоров Grace CPU к ускорителям B200, которое составляет 1:1 (2 CPU на 2 GPU в NVL2 и 2 CPU на 4 GPU в NVL4).

Такая архитектурная разница обусловлена различным назначением платформ: NVL2 и NVL4 ориентированы на задачи, требующие интенсивной предварительной обработки данных, сложной логики управления потоками и активного взаимодействия с внешними системами.

Увеличенное количество CPU-ядер обеспечивает:

- эффективную обработку множественных запросов пользователей;
- сложную предварительную обработку входных данных;
- управление векторными базами данных и поисковыми индексами;
- координацию работы в распределённых системах.

NVIDIA DGX GB200 NVL72 – это мощный суперкомпьютер, созданный спе-

Рис. 8. Дата-центр на базе DGX SuperPOD

циально для работы с искусственным интеллектом (ИИ), в состав которого входят 36 процессоров NVIDIA Grace и 72 ускорителя графических процессоров Blackwell B200, размещённых в серверной стойке. Они соединены между собой с помощью технологии NVLink, которая позволяет им работать как единое целое, как один гигантский процессор, содержащий суммарное количество ARM-ядер Neoverse V2, равное 2592 шт. Эта платформа может выполнять до 1,44 экзафлопс (1,44 квинтиллиона операций в секунду) вычислений для ИИ, что делает её одним из самых мощных решений для таких задач. Кроме того, суперкомпьютер имеет до 240 Тбайт быстрой памяти, что позволяет обучать и развёртывать даже самые большие GLM ИИ-модели.

Однако такое количество высокопроизводительных компонентов делает систему GB200 NVL72 очень энергоёмкой: энергопотребление стойки достигает 120 кВт. Поэтому вся система оснащена мощной разветвлённой системой жидкостного охлаждения.

Размеры GB200 NVL72 Rack составляют 600×1000×2236 мм. Общий вес составляет 1360 кг [36].

DGX SuperPOD – масштабируемые ИИ-суперкомпьютеры находятся на вершине линейки платформы NVIDIA DGX. Это полностью готовый к работе ИИ-кластер, который может масштабироваться до десятков тысяч УГП B200 для решения самых сложных задач обучения и вывода генеративных ИИ-моделей, содержащих триллионы параметров.

Платформа DGX SuperPOD построена по модульной архитектуре на основе масштабируемых единиц (SU), каждая из которых состоит из 32 систем DGX B200 или из восьми стоек DGX GB200 NVL72. Полностью протестированная система масштабируется до четырёх SU, но могут быть построены и более крупные развёртывания в зависимости от требований заказчика. Каждая масштабируемая единица DGX SuperPOD способна обеспечить 640 петафлопс ИИ-производительности при точности FP8 [37].

На рис. 8 показан дата-центр на базе DGX SuperPOD.

Новейшие версии DGX SuperPOD оснащены системами DGX GB300 с суперчипами NVIDIA Grace Blackwell Ultra. Их появление и подробное описание ожидается ближе к концу 2025 года.

Успех NVIDIA в области искусственного интеллекта обусловлен не только мощными графическими ускорителями, но и обширной программной экосистемой, которая делает эти вычислительные ресурсы доступными для широкого круга разработчиков.

CUDA (Compute Unified Device Architecture) является программной основой, которая обеспечивает прямой доступ к виртуальному набору инструкций УГП и параллельным вычислениям на рассмотренном выше оборудовании NVIDIA [38].

NVIDIA AI Enterprise представляет собой облачно-ориентированную программную платформу, обеспечивающую взаимодействие между облаком, дата-центром и периферией. ПО NVIDIA AI Enterprise включает в себя микросервисы NIM и NeMo для повышения производительности моделей и ускорения времени развёртывания генеративного ИИ [39].

NVIDIA AI Data Platform интегрирует корпоративные хранилища с NVIDIA-ускоренными вычислениями и программным обеспечением NVIDIA [40].

NVIDIA NIM – это набор предварительно созданных, оптимизированных микросервисов вывода для быстрого развёртывания новейших AI-моделей на любой NVIDIA-ускорительной инфраструктуре [41].

NVIDIA NeMo – представляет собой набор микросервисов, который обеспечивает комплексный набор функций для создания End-to-End платформ тонкой настройки, оценки и обслуживания больших языковых моделей [42].

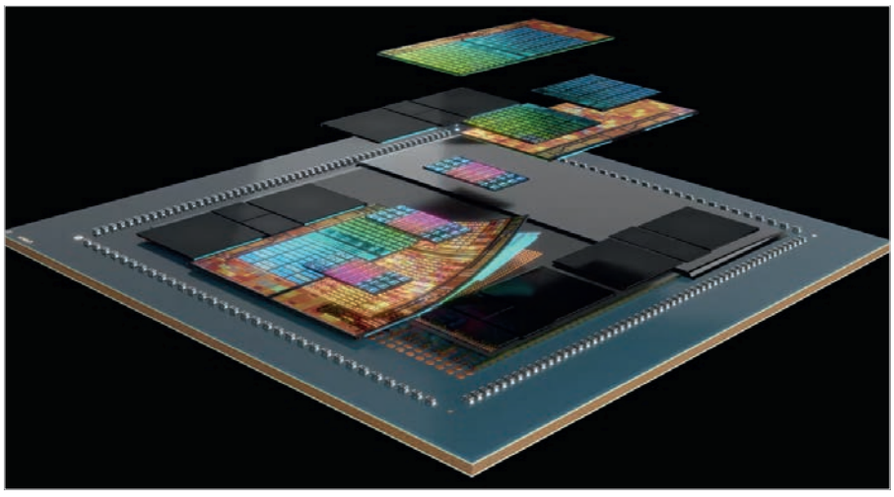

Рис. 9. Трёхмерная архитектура AMD Instinct MI300X

Кроме того, NVIDIA разработала множество приложений для разработчиков, таких, например, как Omniverse, Omniverse Cloud [43], DRIVE, DRIVE AGX [44], NVIDIA NGC, Project DIGITS (Linux) [45], AI Blueprints [46], Cosmos [47].

Программная экосистема NVIDIA представляет собой комплексное решение, охватывающее весь цикл разработки AI-приложений: от базовых вычислительных платформ, таких как CUDA, до специализированных отраслевых решений, таких как DRIVE и Omniverse.

Завершая обзор продукции NVIDIA, можно предположить, что, несмотря на ожесточённую конкурентную борьбу, эта корпорация по совокупности позиций будет оставаться в лидерах, по крайней мере, ещё несколько ближайших лет.

Ускоритель графического процессора AMD Instinct MI300X и технология упаковки 3.5D

Компания Advanced Micro Devices (AMD) может претендовать на роль крупного игрока на рынке процессоров для ИИ. Это одна из старейших компаний в области микропроцессоров, основанная ещё в 1969 году Джерри Сандерсом. В последние годы AMD сделала радикальный поворот в сторону искусственного интеллекта. Приобретение компании Xilinx за $49 млрд в 2022 году стало крупнейшей сделкой в истории полупроводниковой индустрии, а покупка Pensando за $1,9 млрд укрепила позиции в области сетевых технологий центров обработки данных.

В 2024 году AMD запустила в производство революционный для того времени высокопроизводительный ускорительный модуль MI300X, разработанный специально для решения задач, связанных со сверхбольшими вычислительными мощностями (High Performance Computing – HPC), в том числе для обучения и работы больших языковых моделей.

По итогам года этот процессор стал одним из наиболее успешных проектов, который достиг отметки в $1 млрд продаж быстрее любого другого продукта в истории AMD.

Последняя модель ускорительного модуля AMD Instinct™ MI300X GPU Accelerator представляет собой одну из наиболее совершенных на сегодняшний день конструкций ГПУ-ускорителей, разработанных когда-либо компанией AMD [48].

Имеет смысл пояснить название этого устройства, которое употребляет сам изготовитель. Термин GPU Accelerator обозначает устройство, которое позволяет увеличить производительность обработки данных, используя возможности параллельной обработки GPU в дополнение к центральному процессору CPU. В англоязычных статьях часто оставляют только слово Accelerator, которое лучше по смыслу переводить как «ускорительный модуль». Именно этот термин мы будем использовать в дальнейшем.

Ускорительный модуль AMD Instinct MI300X – это не просто «большая видеокарта», а сложнейшая трёхмерная структура взаимосвязанных микросхем, созданная по технологии 3D-стекинга (рис. 9). Эта схема получила такое внутрифирменное название AMD, как «Технология упаковки 3.5D», фиксируя внимание на комбинации 3D-стекирования GPU и I/O-кристаллов с помощью гибридного бондинга в сочетании со стандартной 2.5D-упаковкой (рис. 9).

Конструкция Instinct MI300X включает восемь графических вычислительных чипов (Accelerator Complex Die – XCD), уложенных в трёхмерные стеки. В состав XCD входят ядра и компоненты для обработки данных. Каждый XCD, созданный по технологии 5 нм TMSC, содержит 38 вычислительных блоков (Compute Units – CU), разработанных на основе архитектуры AMD CDNA 3. Всего в одном MI300X содержится 304 вычислительных блока (CU).

При этом конкретный XCD имеет свой выделенный кэш L2 4 МБ, обеспечивающий более быстрый доступ к часто используемым данным.

Кроме того, MI300X может быть оснащён общим 256 МБ AMD Infinity Cache™ для всех восьми XCD, что обеспечивает ещё один уровень кэширования для сокращения доступа к памяти вне чипа.

Четыре интегральные сборки на кристалле (IO Dies), выполняющие функции ввода/вывода, управляют памятью, взаимосвязью и маршрутизацией данных. Кроме того, они координируют работу всех XCD с использованием специальной сети связи AMD Infinity Fabric. Эта сеть фактически выполняет функцию «нервной системы» MI300X, которая позволяет всем частям ускорительного модуля мгновенно обмениваться информацией.

Технология AMD Infinity Fabric позволяет реализовать высокоскоростные связи с малой задержкой между отдельными элементами и всей системой в целом со скоростью до 896 Гбайт/с.

Взаимодействие XCD и IOD осуществляется с помощью усовершенствованной технологии 3D Stacking.

По существу, AMD Instinct MI300X представляет собой сложную структуру взаимосвязанных чипсетов, оптимизированных для высокой производительности и эффективного доступа к памяти, объединяющую восемь 12-слойных стеков памяти HBM3 с восемью 3D-стековыми 5-нм чипсетами XCD на четырёх магистральных сборках на кристалле – 6-нм IOD-кристаллах. Процессорный модуль Instinct MI300X использует восемь стеков внешней памяти HBM3, по 24 Гбайт каждый, что обеспечивает суммарный объём 192 Гбайт. Скорость доступа к памяти составляет 5,3 Тбайт/с.

Таким образом, XCD обеспечивают вычислительную мощность, IOD управляют вводом-выводом и памятью, а Infinity Fabric объединяет все эти элементы в общий вычислительный модуль [49, 50].

Ускорительный модуль AMD Instinct MI300X может работать в различных форматах с различной производительностью: FP8 – 2,6 петафлопс; FP16/BF16 – 1,3 петафлопс; FP32 – 163 терафлопс; FP64 – 82 терафлопса.

Значительно увеличить производительность и память можно за счёт платформы AMD Instinct MI300X Platform, которая состоит из восьми ускорительных модулей Instinct MI300X, размещённых в крэйте Universal Baseboard – UBB 2.0 (рис. 10).

В такой конфигурации восемь модулей Instinct MI300X работают в формате OAM (Open Accelerator Module), в котором каждый ускоритель напрямую соединён с другим. Этот режим поддерживает 7 связей Infinity Fabric на каждый ускоритель (128 Гбайт/с каждая) и 8 подключений PCIe Gen 5 для связи с сервером. Общая пропускная способность между всеми ускорителями составляет 896 Гбайт/с. Внутри каждого ускорителя 8 вычислительных чипов работают параллельно. На уровне платформы все 8 ускорителей видят друг друга как единую систему.

Основные характеристики платформы AMD Instinct MI300X Platform приведены в табл. 5.

Открытая программная экосистема ROCm 6 позволяет пользователям разрабатывать и адаптировать свои программы, используя такие популярные фреймворки, как PyTorch, TensorFlow, JAX.

Текущая версия открытого программного стека AMD ROCm 6.1 включает драйверы, инструменты разработки и API для программирования GPU от низкоуровневых ядер до конечных пользовательских приложений ROCm. Кроме того, она предоставляет расширенную поддержку нескольких GPU для создания масштабируемых AI-систем [53].

Платформа AMD Instinct MI300X Platform может обучать модели с триллионами параметров, распределяя обработку между всеми 2432 вычислительными блоками (304×8). Благодаря огромной параллельной вычислительной мощности эта платформа позволяет также создавать большие мультимодальные модели ИИ с генерацией текстов, изображений и видео в реальном времени.

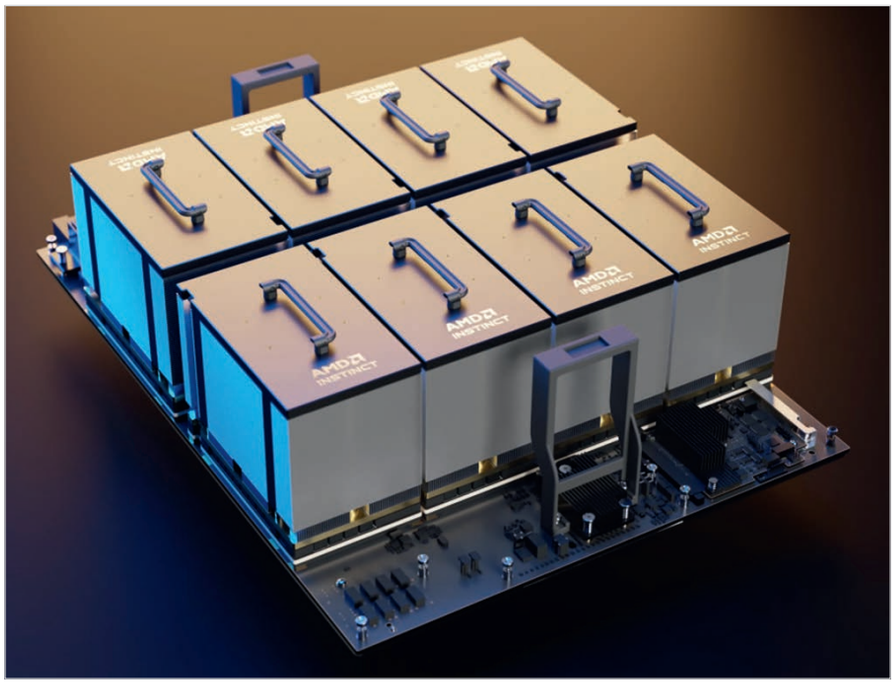

Рис. 10. Платформа AMD Instinct™ MI300X Platform [51]

Безусловно, нужно обратить внимание на цены продукции AMD. Один ускорительный модуль Instinct MI300X (192 Гбайт HBM3) обойдётся корпоративным заказчикам, таким как Microsoft, примерно по цене $10 000 за единицу.

Для других, менее важных корпоративных клиентов цена может составить около $15 000 за единицу Instinct MI300X [54].

Полная платформа Instinct MI300X Platform MI300X (8 ускорительных модулей + крэйт UBB 2.0) стоит ориентировочно в районе $34 000 [55].

Эти цифры позволяют говорить о том, что ускорительные модули AMD Instinct MI300X значительно дешевле аналогичной продукции того же класса B200 NVIDIA.

При этом AMD предлагает больше памяти (192 Гбайт против 80–192 Гбайт у конкурентов) и открытую программную экосистему ROCm 6.

Кроме того, Instinct MI300X имеет лучшие показатели по энергоэффективности (TBP – 750 Вт на ускорительный модуль).

Приведённые выше данные показывают, что AMD Instinct M1300X может составить серьёзную конкуренцию, отбирая у NVIDIA таких клиентов, как: Microsoft, Azure, Meta, Oracle Cloud, Google, Amazon, Hugging Face, которые начали интегрировать Instinct MI300X в свои инфраструктуры искусственного интеллекта [56, 57].

Семейство AMD Instinct MI300X/MI325X представляет собой серьёзную попытку бросить вызов доминированию NVIDIA в области ИИ-ускорителей. Благодаря революционной 3.5D-архитектуре, рекордному объёму памяти и стратегическим партнёрствам с ведущими технологическими компаниями AMD демонстрирует, что конкуренция в сфере аппаратного обеспечения искусственного интеллекта становится всё более интенсивной.

Особенно впечатляющими выглядят планы AMD в отношении новой модели MI355X с заявленным 35-кратным улучшением производительности инференса и поддержкой новых низкоточных форматов FP4/FP6.

Они будут оснащены 256 Гбайт памяти HBM3E и обеспечат пропускную способность памяти до 6 Тбайт/с. AMD видит большой потенциал в рынке

Таблица 5. Основные характеристики платформы AMD Instinct MI300X Platform [52]

Наименование	Значение
Количество ускорителей	8 штук в одной системе AMD MI300X Platform
Общий объём памяти	1,5 Тбайт HBM3
Скорость памяти	5,3 Тбайт/с
Производительность ИИ	42 Пфлопс (FP8)
Потребление энергии	750 Вт на ускоритель
Соединение между чипами	896 Гбайт/с
Подключение к серверу	PCIe Gen 5 x16

Рис. 11. Схема суперкомпьютера Cerebras CS-3 AI

ускорителей ИИ, который по прогнозам достигнет $500 млрд к 2028 году, что подчёркивает важность данного продукта для бизнеса [58].

Если эти обещания оправдаются, AMD может существенно изменить расклад сил на рынке ИИ-ускорителей, предоставив заказчикам реальную альтернативу решениям NVIDIA и простимулировав инновации во всей отрасли.

Cerebras WSE-3 – всё для ИИ на одном кристалле

По всей видимости, одной из наиболее впечатляющих новинок в области АПИИ стал Cerebras WSE-3, представленный в мае 2024 года. Этот процессор знаменует собой революцию в архитектуре процессоров для искусственного интеллекта.

Несмотря на то что название Cerebras Systems знакомо далеко не всем специалистам, это достаточно крупная американская компания с «хорошей родословной». Она была основана в 2015 году пятью инженерами из фирмы SeaMicro, которая была продана AMD в 2012 году. На сегодняшний день Cerebras Systems имеет офисы в Саннивейле, Сан-Диего, Торонто и Бангалоре (Индия) и насчитывает около 525 сотрудников. В 2025 году планируется завершение оформления IPO под тикером CBRS с целевой оценкой до $8 млрд. Примечательно то, что эта компания сильно зависит от одного крупного клиента G42 из Объединённых Арабских Эмиратов, который обеспечил 87% всей выручки за 2024 год.

Компания G42 является ведущей технологической группой арабского мира, специализирующейся на искусственном интеллекте (ИИ), облачных вычислениях, хранении данных и геопространственном интеллекте. Штабквартира находится в Абу-Даби [59].

Компания Cerebras Systems разработала Wafer-Scale Engine 3 WSE 3 – процессор, который кардинально переосмысливает подходы к созданию вычислительных систем для искусственного интеллекта. Этот чип представляет собой не просто очередное улучшение существующих технологий, а принципиально новую архитектурную концепцию.

Главная особенность WSE-3 заключается в его монолитной конструкции: весь процессор изготавливается из единой кремниевой пластины (Wafer) площадью 462 см², что в 57 раз больше флагманского GPU NVIDIA H100 [59]. Эта инновационная конструкция позволила создать чип рекордных размеров с беспрецедентными характеристиками.

В табл. 6 приведены сравнительные характеристики Wafer Scale Engine-3 и H100 [60].

Данные, приведённые в табл. 6, демонстрируют преимущества процессора Cerebras Wafer Scale Engine 3 по сравнению с ускорителем графического процессора Nvidia H100.

Процессор WSE-3, созданный с использованием новейшего 5-нм техпроцесса, является основой суперкомпьютера Cerebras CS-3 AI (рис. 11) [61].

Нужно отметить, что особое внимание в процессоре Cerebras WSE-3 уделено системе памяти, которая принципиально отличается от конкурентов. В процессоре используется 44 ГБ встроенной SRAM-памяти, распределённой непосредственно на кристалле рядом с вычислительными ядрами. Это в 880 раз больше, чем у NVIDIA H100.

В этом заключается коренное отличие системы памяти Cerebras от NVIDIA, использующей внешнюю HBM3E-память [64].

Дополнительно системы суперкомпьютерных блоков CS-3 на основе WSE-3 могут быть оснащены внешней памятью объёмом 1,5 Тбайт, 12 Тбайт или даже 1,2 Пбайт.

Монолитная конструкция на одной кремниевой пластине обеспечивает ключевое преимущество WSE-3, которое заключается в исключительной пропускной способности, составляющей 21 Пбайт в секунду для памяти (это в 7000 раз выше H100) и 214 Пбит в секунду для внутренних соединений (это в 3715 раз быстрее H100). В традиционных системах ИИ множество отдельных процессоров должны постоянно обмениваться данными через относительно медленные межсоединения. В WSE-3 все ядра находятся на одном кристалле и связаны высокоскоростными внутренними шинами,

Таблица 6. Сравнительные характеристики Cerebras Wafer Scale Engine-3 и NVIDIA H100

Характеристика	Cerebras WSE-3	NVIDIA H100	Преимущество Cerebras
Размер чипа	46 225 мм²	814 мм²	57x
Ядра	900 000	16 896 FP32 + 528 Tensor	52x
Память на чипе	44 Гбайт	0,05 Гбайт	880x
Пропускная способность памяти	21 Пбайт/с	0,003 Пбайт/с	7000x
Пропускная способность шины	214 Пбит/с	0,0576 Пбит/с	3715x

что устраняет эти узкие места [65]. Такая архитектура особенно эффективна для обучения больших языковых моделей, где требуется обработка огромных объёмов данных со множественными взаимосвязями между различными частями нейросети [66].

Для создания ещё более мощных систем Cerebras разработала технологию масштабирования, позволяющую объединять до 2048 суперкомпьютеров CS-3 в единые кластеры. Это обеспечивает линейное масштабирование производительности без потери эффективности.

Впечатляющим примером такого масштабирования является проект Cerebras Condor Galaxy, который представляет собой сеть взаимосвязанных суперкомпьютеров CS-3, разрабатываемых совместно с компанией из ОАЭ G42.

Целью проекта является создание крупнейшей и самой быстрой в мире инфраструктуры суперкомпьютеров, предназначенной для генеративного искусственного интеллекта.

Уже построены Condor Galaxy 1 и 2 с производительностью в четыре экзафлопс. Продолжается строительство Condor Galaxy 3, планируемая мощность которого восемь экзафлопс.

Система Condor Galaxy 1–2 спроектирована как облачный распределённый суперкомпьютер ИИ, позволяющий учёным в разных городах и странах использовать её ресурсы для крупномасштабных задач ИИ и научных задач. В настоящее время вычислительными мощностями Condor Galaxy 1–2 могут пользоваться учёные во многих городах США и в некоторых других зарубежных странах, в том числе и в режиме онлайн.

Condor Galaxy 3 (CG-3) будет состоять из 64 системных блоков CS-3, что позволит обучать модели с параметрами до 24 триллионов. Это в 10 раз больше, чем GPT-4 или Gemini [67].

В рамках создания третьей стадии проекта CG-3 компания Cerebras Systems добавляет шесть новых центров обработки данных ИИ в Северной Америке и Европе. Это увеличит пропускную способность вывода до более чем 40 миллионов токенов в секунду. Новые объекты будут созданы в Далласе, Миннеаполисе, Оклахома-Сити, Монреале, Нью-Йорке и городах Франции, при этом 85% от общей мощности будет находиться в Соединённых Штатах [68].

Следует обратить внимание на то, что суперкомпьютеры CS-3 на базе WSE-3 физически расположены в США (Санта-Клара, Даллас), но управляются совместно с G42 из ОАЭ. Это позволяет Cerebras и G42 контролировать доступ в соответствии с американским законодательством, при этом предоставляя ресурсы международным исследователям из дружественных стран [69].

Опыт использования системы Condor Galaxy уже показал её огромную значимость для развития ИИ, а также многих направлений мировой науки и техники, требующих больших вычислительных мощностей.

Системы CS-3 на базе WSE-3 от Cerebras представляют собой значительный скачок в скорости и эффективности обучения ИИ, особенно для больших языковых моделей, таких как Llama 2 (70 млрд параметров). Например, традиционное обучение Llama 2 требовало около 1,7 миллиона часов работы GPU, распределённых по тысячам высокопроизводительных GPU. На это уходило несколько недель рабочего времени в зависимости от размера кластера и эффективности [70].

Сегодня компактная конфигурация всего из четырёх систем CS-3 может точно обучить модель Llama 2 с семьюдесятью миллиардами параметров менее чем за день [71].

Впечатляющие результаты были получены в фундаментальных научных исследованиях. Например, исследователи Аргоннской национальной лаборатории получили премию Гордона Белла за исследования вариантов COVID-19, выполненные на кластере CS-2.

В мае 2024 года команда из Сандийской национальной лаборатории смоделировала взаимодействие 800 000 атомов, сократив год вычислений до двух дней [72].

По достоинству оценили возможности процессоров Cerebras некоторые ведущие производители электронных компонентов. Например, концерн Qualcomm заключил с Cerebras соглашение об обучении с помощью WSE-3 моделей ИИ, которые будут работать на процессорах Qualcomm AI 100 Ultra [73].

Также объявлено о сотрудничестве с Dell Technologies в области разработки новой прикладной инфраструктуры генеративного искусственного интеллекта [74].

Нужно также сказать несколько слов о стоимости этого оборудования. Согласно данным отраслевого издания Data Center Dynamics, компания Cerebras не раскрывает официальную стоимость своих чипов, однако, по оценкам экспертов, они стоят около $2–3 млн за единицу [75].

По информации Reuters, первый суперкомпьютер Condor Galaxy 1 с 32 узлами обошёлся партнёру G42 в $100 млн, что составляет $3,13 млн за узел, включая обслуживание. Полная программа развёртывания 9 кластеров может превысить $900 млн [76].

Аналитики The Next Platform оценивают стоимость полномасштабного кластера из 2048 систем CS-3 в районе $5–6 млрд [77].

Завершая этот короткий обзор новой продукции Cerebras, можно с уверенностью говорить о том, что WSE-3 представляет собой радикальный отход от традиционных подходов в проектировании процессоров. Успех WSE-3 может стать началом нового этапа в развитии специализированных процессоров для искусственного интеллекта.

Заключение

Анализ современного рынка аппаратного обеспечения для искусственного интеллекта показывает, что эта отрасль переживает период активных изменений. Рассмотренные в данной статье решения от NVIDIA, AMD и Cerebras представляют лишь наиболее яркие примеры инновационных подходов к созданию ИИ – ускорителей процессоров.

Благодаря экосистеме CUDA и универсальности решений сегодня на рынке ускорителей графических процессоров продолжает доминировать NVIDIA. Архитектура Blackwell (B200/B300) с поддержкой FP4 и революционной двухкристальной конструкцией устанавливает новые стандарты производительности для обучения и инференса больших языковых моделей.

Технология упаковки 3.5D, использованная AMD в своих ускорителях MI300X/MI325X, предлагает конкурентоспособную альтернативу с большим объёмом памяти (192 Гбайт) и открытой программной платформой ROCm, что особенно привлекательно для корпоративных заказчиков, стремящихся к независимости от монополии NVIDIA.

Таблица 7. Сравнительные характеристики современных ИИ-процессоров

Процессор	Тех. процесс	Производительность (FP16)	Память (HBM3/SRAM)	TDP	Применение	Экосистема ПО
Cerebras WSE-3	5 нм	125 Пфлопс	44 Гбайт SRAM, до 1,2 Пбайт внешней	~15 кВт	Обучение сверхмасштабных LLM	Cerebras Software Platform
NVIDIA B200	4 нм	~10 Пфлопс*	192 Гбайт HBM3e	1000 Вт	Универсальное (обучение и инференс)	CUDA, TensorRT
Intel Gaudi 3	5 нм	~2,5 Пфлопс	128 Гбайт HBM2e	600 Вт	Обучение и инференс LLM	Intel AI Suite, PyTorch
AMD MI325X	5 нм	1,3 Пфлопс	256 Гбайт HBM3e	750 Вт	Обучение и инференс в HPC и облаке	ROCm, PyTorch, TensorFlow
Google TPU v5	4 нм	~1 Пфлопс**	HBM3 (объём не раскрыт)	Н/Д	Инференс и обучение в TensorFlow	TensorFlow, JAX
Huawei Ascend 910C	7 нм	640 Тфлопс	128 Гбайт HBM3	400–600 Вт	Обучение и инференс LLM в дата-центрах	MindSpore, CANN
Graphcore IPU-POD256	7 нм	~350 Тфлопс	460 Гбайт In-Processor Memory	~15 кВт	Исследовательские задачи, обучение	Poplar SDK, PyTorch
Qualcomm AI 100 Ultra	5 нм	~75 Тфлопс	LPDDR5/DDR5	75–150 Вт	Edge Computing, локальный инференс	Qualcomm AI SDK
Apple M4 Max	3 нм	~40 Тфлопс	128 Гбайт унифицированной памяти	40–60 Вт	Локальный инференс на устройствах	Core ML, Metal
DeepSeek (RISC-V)	7 нм/5 нм	~16 Тфлопс***	HBM3/DDR5 (зависит от SoC)	50–150 Вт	Локальный инференс LLM	Собственная экосистема RISC-V

Примечания к таблице:

**NVIDIA B200: производительность в FP8 (20 Пфлопс) пересчитана в FP16 (~10 Пфлопс, так как FP8 обычно вдвое эффективнее) [14].*

***Google TPU v5: точные данные не раскрыты, оценка ~1 Пфлопс основана на двукратном превосходстве над TPU v4 (~500 ТФ FP16).*

****DeepSeek (RISC-V): оценка ~16 Тфлопс FP16 основана на 32 TOPS INT8 для Sophgo SG2380 (1 ТФ FP16 ≈ 2 TOPS INT8).*

Инновационный подход демонстрирует Cerebras с технологией Wafer-Scale процессоров WSE-3, показывая возможности кардинального переосмысления архитектуры для сверхмасштабных вычислительных задач.

Однако современный рынок ИИ-процессоров значительно шире представленных в статье решений. В табл. 7 приведён рейтинг ведущих мировых компаний с учётом, прежде всего, производительности их процессоров (таблица подготовлена с помощью ИИ Claude-4 Sonet, 2025).

В заключение можно сказать, что даже наш поверхностный обзор продукции ведущих производителей позволяет отметить несколько ключевых тенденций, определяющих направление развития аппаратного обеспечения для искусственного интеллекта. Прежде всего, это повышение энергоэффективности при росте производительности.

Например, такие процессоры, как Huawei Ascend 910C (TDP 400–600 Вт) и Apple M4 Max (TDP 40–60 Вт), демонстрируют тенденцию к снижению энергопотребления при высокой производительности. При этом Ascend 910C обеспечивает 640 Тфлопс (FP16), конкурируя с NVIDIA H100 при существенно меньшем энергопотреблении. Cerebras WSE-3, несмотря на высокий TDP (~15 кВт), показывает исключительную эффективность для сверхмасштабных задач благодаря производительности 125 ПФлопс (FP16) [78].

Другое быстроразвивающееся направление связано со специализацией оборудования под конкретные задачи ИИ.

Различные процессоры оптимизируются под специфические применения. В то время как NVIDIA B200 остаётся универсальным решением для обучения и инференса, Google TPU v5 специально оптимизирован для экосистемы TensorFlow. Идеально подходит для обучения моделей с триллионами параметров процессор Cerebras WSE-3. Китайский Huawei Ascend 910C поддерживает экосистему MindSpore для таких моделей, как Pangu. Процессоры Intel Gaudi 3 ориентированы на снижение стоимости в корпоративном сегменте [79].

Также заметно увеличивается интерес к открытым архитектурам после публикации информации о специализированных чипах RISC-V (SiFive P670, DeepSeek). Особенно заметным было появление на рынке ИИ-моделей DeepSeek R-1, которые позволили значительно снизить затраты на разработку и производство. Характерный пример показывает Google, интегрирующий RISC-V ядра (SiFive X280) в TPU v5, обеспечивая таким образом гибкость и универсальность системы программирования [80].

В ближайшие 2–3 года ожидается дальнейшая диверсификация рынка ИИ-процессоров. Конкуренция между ведущими производителями будет стимулировать инновации в области энергоэффективности, специализации архитектур и масштабируемости решений.

Литература

1. AI Index Report 2025. URL: https://hai.stanford.edu/ai-index/2025-ai-index-report.
2. NVIDIA Tensor Cores. URL: https://www.nvidia.com/en-us/data-center/tensor-cores/.
3. CXL 3.0. URL: https://computeexpresslink.org/wp-content/uploads/2025/02/CXL_Q1-2025-Webinar-Presentation_FINAL.pdf.
4. CEREBRAS SYSTEMS, INC. URL: https://f.hubspotusercontent30.net/hubfs/8968533/Cerebras-Systems-Overview.pdf.
5. Market cap. URL: https://companiesmarketcap.com/nvidia/marketcap/.
6. NVIDIA. URL: https://abrv.in/jukp.
7. NVIDIA Blackwell Platform Arrives to Power a New Era of Computing. URL: https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing.
8. NVIDIA Blackwell Architecture and B200/B100 Accelerators Announced. URL: https://www.anandtech.com/show/21310/nvidia-blackwell-architecture-and-b200b100-accelerators-announced-going-bigger-with-smaller-data.
9. NVIDIA Hopper Architecture In-Depth. URL: https://developer.nvidia.com/

blog/nvidia-hopper-architecture-in-depth/.
10. NVIDIA Blackwell B100, B200, GPU. URL: https://datacrunch.io/blog/nvidia-blackwell-b100-b200-gpu.
11. SXM (socket). URL: https://en.wikipedia.org/wiki/SXM_(socket).
12. NVIDIA-H100. URL: https://www.shi.com/product/46062270/NVIDIA-H100-GPU-computing-processor.
13. NVIDIA Blackwell vs NVIDIA Hopper: A Detailed Comparison. URL: https://www.nexgencloud.com/blog/performance-benchmarks/nvidia-blackwell-vs-nvidia-hopper-a-detailed-comparison.
14. NVIDIA B200. URL: https://www.nvidia.com/en-us/data-center/products/b200/.
15. NVIDIA announces Blackwell Ultra B300. URL: https://www.tomshardware.com/pc-components/gpus/nvidia-announces-blackwell-ultra-b300-1-5x-faster-than-b200-with-288gb-hbm3e-and-15-pflops-dense-fp4.
16. NVIDIA's Christmas Present: GB300 & B300. URL: https://semianalysis.com/2024/12/25/nvidias-christmas-present-gb300-b300-reasoning-inference-amazon-memory-supply-chain.
17. NVIDIA GB200 Delivered, and Here Comes the GB300. URL: https://www.fibermall.com/news/nvidia-here-comes-the-gb300.htm.
18. NVIDIA HGX Platform. URL: https://www.nvidia.com/en-us/data-center/hgx/.
19. Blackwell Platform Arrives to Power a New Era of Computing. URL: https://resources.nvidia.com/en-us-blackwell-architecture/datasheet?ncid=no-ncid.
20. NVIDIA® B200 SXM6. URL: https://datacrunch.io/b200.
21. Created with Grok 3 AI by xAI, 2025. URL: https://x.com/i/grok.
22. NVIDIA HGX B300 NVL16 information. www.fibermall.com.
23. NVIDIA DGX B200. URL: https://www.nvidia.com/en-us/data-center/dgx-b200/.
24. Datasheet DGX B200. URL: https://resources.nvidia.com/en-us-dgx-systems/dgx-b200-datasheet.
25. Introduction to NVIDIA DGX B200 Systems. URL: https://www.nvidia.com/en-us/data-center/dgx-b200/.
26. DGX-B300-datasheet. URL: https://resources.nvidia.com/en-us-dgx-systems/dgx-b300-datasheet.
27. NVIDIA DGX B200 User Guide. URL: https://docs.nvidia.com/dgx/dgxb200-user-guide/dgxb200-user-guide.pdf.
28. NVIDIA DGX Station. URL: https://www.nvidia.com/en-us/products/workstations/dgx-station/.
29. GB200. URL: https://www.nvidia.com/en-us/data-center/gb200-nvl72/.
30. GB200 Hardware Architecture. URL: https://semianalysis.com/2024/07/17/gb200-hardware-architecture-and-component/.
31. Blackwell B200. GB200 Grace Blackwell. URL: https://nvdam.widen.net/s/wwnsxrhm2w/blackwell-datasheet-3384703.
32. NVIDIA CPU Grace. URL: https://www.nvidia.com/en-us/data-center/grace-cpu.
33. NVIDIA GB200 NVL72 training. URL: https://training.continuumlabs.ai/infrastructure/servers-and-chips/nvidia-gb200-nvl72.
34. NVIDIA GB200 NVL2. URL: https://www.nvidia.com/en-us/data-center/gb200-nvl2/.
35. NVIDIA GB200 NVL4. URL: https://www.techpowerup.com/328966/nvidia-prepares-gb200-nvl4-four-blackwell-gpus-and-two-grace-cpus-in-a-5-400-w-server.
36. NVIDIA GB200 NVL72. URL: https://www.nvidia.com/en-us/data-center/gb200-nvl72/.
37. NVIDIA DGX SuperPOD. URL: https://www.nvidia.com/en-us/data-center/dgx-superpod/.
38. CUDA. URL: https://developer.nvidia.com/cuda-toolkit.
39. NVIDIA AI Enterprise. Software Platform. URL: https://www.nvidia.com/en-us/data-center/products/ai-enterprise/.
40. New Class of Enterprise Infrastructure. URL: https://gclnk.com/AnBA24Uv.
41. NIM for Developers. URL: https://developer.nvidia.com/nim.
42. Nemo-microservices. URL: https://gclnk.com/9p0wrZho.
43. Omniverse Platform. URL: https://developer.nvidia.com/omniverse.
44. Car Technology from NVIDIA. URL: https://www.nvidia.com/en-us/solutions/autonomous-vehicles/.
45. NVIDIA unveils robot training. URL: https://www.reuters.com/technology/ces-nvidia-ceo-set-take-stage-ces-just-after-shares-hit-record-high-2025-01-07/.
46. AI Agents: Built to Reason, Plan, Act. URL: https://www.nvidia.com/en-us/ai/.
47. NVIDIA Expands Omniverse. URL: https://nvidianews.nvidia.com/news/nvidia-expands-omniverse-with-generative-physical-ai.
48. Meet the New AMD Instinct™ MI325X Accelerators. URL: https://www.amd.com/en/products/accelerators/instinct.html?utm_campaign=instinct&utm_medium=redirect&utm_source=301.
49. AMD Instinct™ MI300 series microarchitecture. URL: https://rocm.docs.amd.com/en/latest/conceptual/gpu-arch/mi300.html.
50. AMD MI300X Accelerator Unpacked: Specs, Performance & More. URL: https://tensorwave.com/blog/mi300x-2.
51. AMD Instinct™ MI300X Platform. URL: https://shorturl.at/SxDcB.
52. AMD INSTINCT™ MI300X PLATFORM. URL: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300x-platform-data-sheet.pdf.
53. AMD ROCm™ 6 OPEN SOFTWARE FOR AI AND HPC WORKLOADS. URL: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/amd-rocm-6-brief.pdf.
54. IXBT.com (2024). URL: https://www.ixbt.com/news/2024/02/03/amd-instinct-mi300x-192-10-15-nvidia-h100.html.
55. Interpro Microsystems. AMD Instinct MI300X 192GB 8 OAM + UBB. URL: https://www.interpromicro.com/product/39955-amd-instinct-mi300x-192gb-8-oam-ubb.
56. AMD's MI300 AI accelerator sales drive 80 percent growth. URL: https://www.datacenterdynamics.com/en/news/amds-mi300-ai-accelerator-sales-drive-80-percent-growth-in-data-center-segment/.
57. Meta and Microsoft say they will buy AMD's new AI chip. URL: https://www.cnbc.com/2023/12/06/meta-and-microsoft-to-buy-amds-new-ai-chip-as-alternative-to-nvidia.html.
58. AMD Instinct™ MI325X Accelerators. URL: https://www.amd.com/en/products/accelerators/instinct/mi300/mi325x.html.
59. G42 partners with Cerebras. URL: https://www.mediaoffice.abudhabi/en/technology/g42-partners-with-cerebras-to-to-launch-worlds-largest-ai-training-supercomputer-network/.
60. Cerebras WSE-3 wafer-scale AI chip. URL: https://www.tweaktown.com/news/96843/cerebras-wse-3-wafer-scale-ai-chip-57x-bigger-than-largest-gpu-with-4-trillion-transistors/index.html.

61. Cerebras CS-3: the world's fastest and most scalable AI accelerator – Cerebras. URL: https://www.cerebras.ai/blog/cerebras-cs3.
62. Cerebras Unveils Six Data Centers. URL: https://www.datacenterfrontier.com/hyperscale/article/55273769/cerebras-unveils-six-data-centers-to-meet-accelerating-demand-for-ai-inference-at-scale.
63. Cerebras reveals revenue surge in US IPO filing. URL: https://economictimes.indiatimes.com/?back=1.
64. Cerebras WSE-3 AI Chip Launched 56x Larger than NVIDIA H100. URL: https://www.servethehome.com/cerebras-wse-3-ai-chip-launched-56x-larger-than-nvidia-h100-vertiv-supermicro-hpe-qualcomm/.
65. Cerebras Unveils Its Next Waferscale AI Chip. URL: https://spectrum.ieee.org/cerebras-chip-cs3.
66. Cerebras Systems Sets New Benchmark in AI. URL: https://www.unite.ai/cerebras-systems-sets-new-benchmark-in-ai-innovation-with-launch-of-the-fastest-ai-chip-ever/.
67. CONDOR GALAXY. URL: https://www.cerebras.ai/condor-galaxy.
68. Cerebras WSE3 Versus Nvidia B200. URL: https://www.nextbigfuture.com/2025/05/cerebras-wse3-versus-nvidia-b200.html.
69. Cerebras's 36 exaFLOPS Condor Galaxy AI super takes flight. URL: https://www.theregister.com/2023/07/20/cerebras_condor_galaxy_supercomputer.
70. Reddit discussion on Llama training. URL: https://www.reddit.com/r/LocalLLaMA/comments/1f4ecsl/its_insane_how_much_computer_meta_has_they_could/.
71. CEREBRAS TRAINS LLAMA MODELS TO LEAP OVER GPUS. URL: https://www.nextplatform.com/2024/10/25/cerebras-trains-llama-models-to-leap-over-gpus/.
72. Wikipedia – Cerebras. URL: https://en.wikipedia.org/wiki/Cerebras.
73. Cerebras Selects Qualcomm. URL: https://www.eetimes.com/cerebras-sells-100-million-ai-supercomputer-plans-8-more/.
74. Cerebras Enables Faster Training. URL: https://www.cerebras.ai/press-release/cerebras-enables-faster-training-of-industrys-leading-largest-ai-models.
75. Data Center Dynamics. Cerebras unveils four trillion-transistor giant chip. URL: https://www.datacenterdynamics.com/en/news/cerebras-unveils-four-trillion-transistor-giant-chip-targets-generative-ai/.
76. Cerebras Sells $100 Million AI Supercomputer Plans Eight More. URL: https://www.eetimes.com/cerebras-sells-100-million-ai-supercomputer-plans-8-more/.
77. The Next Platform. Cerebras goes hyperscale with third gen waferscale supercomputers. URL: https://www.nextplatform.com/2024/03/14/cerebras-goes-hyperscale-with-third-gen-waferscale-supercomputers/.
78. Ascend Computing. URL: https://e.huawei.com/en/products/servers/ascend.
79. Intel Gaudi 3 AI Accelerators. URL: https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.html.
80. RISC-V Core IP Portfolio. URL: https://www.sifive.com/risc-v-core-ip.

НОВОСТИ МИРА. ЧИТАЙТЕ НА ПОРТАЛЕ WWW.CTA.RU

30-ваттные модули DC/DC-преобразователей напряжения для применения в сетях промышленной электроники с 4-кратным изменением напряжений

Серия WINL30 DC/DC-преобразователей напряжения, выпускаемых под торговой маркой Wibbow, включает ряд 30-ваттных модулей с гальванической развязкой для установки на печатную плату в непосредственной близости от нагрузки.

Устройства выполнены с регулированием выходного напряжения методом широтно-импульсной модуляции с постоянной высокой рабочей частотой 300 кГц и синхронным выпрямлением.

Силовой низкопрофильный трансформатор реализован по планарной технологии из высокочастотного феррита, что обусловливает меньшие потери в сердечнике. Эти технические решения позволили получить высокий КПД преобразователей – до 90%, значение удельной мощности 3720 Вт/дм^3. Габаритные размеры модулей 32×21×12 мм.

Серия включает одно- и двухканальные модели, обеспечивающие наиболее востребованные напряжения и предназначенные для работы в сетях с изменением напряжения от 9 до 36 В (кратность диапазона входного напряжения 4:1). Модули содержат набор сервисных и защитных функций, необходимых для безопасной работы: защита от пониженного входного напряжения, перегрузки по току и короткого замыкания, перегрева, повышенного выходного напряжения; подстройка выходного напряжения; выключение внешним сигналом со стороны входа, что может использоваться при необходимости снижения энергопотребления. Встроенный помехоподавляющий фильтр на входе обеспечивает уменьшение уровня кондуктивных помех со стороны питающей сети. Модули стабильно функционируют при температурах от –40 до +105°C. Можно заказать модели для работы в диапазоне температур от –55 до +105°C (температурный класс М).

Использование DC/DC-преобразователей с расширенным диапазоном входного напряжения позволяет сократить номенклатуру потребляемой продукции, так как возможна установка одного и того же типа модулей в узлы аппаратуры, питающиеся от различающихся входных напряжений. Но при выборе блоков питания для каждого конкретного случая необходимо учитывать специфику схемотехники модулей с расширенным диапазоном входного напряжения, применение которых может быть оправдано только в тех случаях, когда не удаётся обойтись другими средствами.

Основные параметры 30-ваттных DC/DC-преобразователей серии WINL30 с расширенным диапазоном входного напряжения

Тип модуля	Выходное напряжение, В	Выходной ток, А	Пульсация выходного напряжения, мВ (от пика до пика)	КПД
WINL30-24WHS5P	5	6	<75	89
WINL30-24WHS12P	12	2,5	<150	90%
WINL30-24WHS15P	15	2	<150	90%
WINL30-24WHS24P	24	1,25	<180	90%
WINL30-24WHS48P	48	0,63	<480	90%
WINL30-24WHD5P	±5	±3	<400	89%
WINL30-24WHD12P	±12	±1,25	<120	89%
WINL30-24WHD15P	±15	±1	<120	89%

Компактные модули серии WINL30 гарантируют высокую стабильность рабочих характеристик и долговременную надёжность в жёстких условиях эксплуатации в промышленных сетях электроснабжения.