Матрица классов GPU по задачам

RTX PRO 6000, H100, H200, B200, B300: память, пропускная способность, мощность, связь между ускорителями. Данные на 06.10-25.10.2026.

Таблица А. Характеристики классов (на один ускоритель)

Класс	Память	Тип памяти	Пропускная способность памяти	Тепловой пакет (TDP)	Связь между ускорителями	Форма поставки
RTX PRO 6000 (серверное исполнение)	96 ГБ	GDDR7 с ECC	1,6 ТБ/с (1 597 ГБ/с)	до 600 Вт	Нет, ускорители общаются по PCIe Gen5	Карта PCIe в сервере 2U-5U
RTX PRO 6000 (рабочая станция)	96 ГБ	GDDR7 с ECC	1,79 ТБ/с	600 Вт	Нет	Карта PCIe в станции
RTX PRO 6000 (рабочая станция, версия 300 Вт)	96 ГБ	GDDR7 с ECC	1,79 ТБ/с	300 Вт	Нет	Карта PCIe
H100 (модуль на плате HGX)	80 ГБ	HBM3	3,35 ТБ/с	700 Вт	Прямая шина 4-го поколения, 900 ГБ/с	Модуль на плате HGX
H100 (карта PCIe, 94 ГБ)	94 ГБ	HBM3	3,9 ТБ/с	350-400 Вт	Мост, 600 ГБ/с	Карта PCIe
H200 (модуль на плате HGX)	141 ГБ	HBM3e	4,8 ТБ/с	до 700 Вт	Прямая шина 4-го поколения, 900 ГБ/с	Модуль на плате HGX
H200 (карта PCIe)	141 ГБ	HBM3e	4,8 ТБ/с	до 600 Вт	Мост на 2 или 4 карты, 900 ГБ/с на ускоритель	Карта PCIe
B200	180 ГБ (в системах HGX; в описаниях других систем встречается 192 ГБ)	HBM3e	8 ТБ/с	1 000 Вт в серверной плате (предел чипа до 1 200 Вт)	Прямая шина 5-го поколения, 1,8 ТБ/с	Модуль на плате HGX, только узлом на 8 ускорителей
B300	до 288 ГБ (в расчётах сайта 270 ГБ)	HBM3e	8 ТБ/с	1 100 Вт в воздушном узле 8U (предел чипа до 1 400 Вт)	Прямая шина 5-го поколения, 1,8 ТБ/с	Модуль на плате HGX, только узлом на 8 ускорителей

Год выхода: H100 2022 (карта PCIe 2023), H200 2023 (карта PCIe 2024), B200 2024, RTX PRO 6000 и B300 2025. H100 и H200 относятся к одному поколению, B200, B300 и RTX PRO 6000 к следующему.

Значения TDP у B200 и B300 относятся к серверной плате, предел чипа указан отдельно; у остальных классов: максимум по данным производителя.

Таблица Б. Узел на 8 ускорителей (системный уровень)

Класс	Память узла	Мощность узла	Высота	Охлаждение	Источник
RTX PRO 6000, 8 карт	768 ГБ	до 9,1 кВт пик	4U	Воздух	Страница каталога от 16.10.2026, спецификация производителя системы
H100, 8 ускорителей	640 ГБ	до 10,2 кВт (платформа из руководства производителя)	8U в каталоге	Воздух	Страницы /gpu/nvidia-h100/ и каталога
H200, 8 ускорителей	1,13 ТБ	до 10,2 кВт для платформы из руководства производителя; у других производителей на той же плате публикуются значения до 15 кВт	8U в каталоге	Воздух	Страницы /gpu/nvidia-h200/, каталога, калькулятора питания

Класс	Память узла	Мощность узла	Высота	Охлаждение	Источник
B200, 8 ускорителей	1,44 ТБ	до 14,3 кВт	10U (есть варианты 8U и 4U с жидкостным охлаждением)	Воздух; один вариант 4U: жидкость	Страница /gpu/nvidia-b200/, каталог от 25.10.2026
B300, 8 ускорителей	2,16 ТБ при 270 ГБ на ускоритель (страница класса также указывает до 2,3 ТБ)	14,5 кВт типовая, до 16,4 кВт пик	8U (вариант 4U с жидкостью)	Воздух; один вариант 4U: жидкость	Страница /gpu/nvidia-b300/, статья о серверной под узел

Масса узлов в источниках для большинства позиций не публикуется. Масса узла 8 × B300 8U: 119 кг без упаковки (статья сайта).

Таблица B. Задачи по классам

- Подходит: прямое утверждение страницы класса или решения.
- С оговоркой: подходит при условии, указанном в ячейке.
- Не подходит: прямое ограничение на странице.
- Вывод: вывод из чисел страниц, прямого утверждения нет.
- Нет данных: сайт не публикует.

Расчёты памяти на страницах классов сделаны для точности FP8, контекста 8 192 токена, одного запроса и запаса 20 %. Для вашего контекста и числа пользователей считайте в калькуляторе памяти.

Задача	RTX PRO 6000	H100	H200	B200	B300
Модель 27-32 млрд параметров	Подходит: умещается на одной карте с памятью 96 ГБ	С оговоркой: умещается в 80 ГБ при 8-битной точности и малом контексте. Вывод	Подходит. Вывод: 141 ГБ с запасом	Избыточно: класс поставляется узлом на 8 ускорителей	Избыточно: класс поставляется узлом на 8 ускорителей
Модель 70 млрд в FP8, один запрос	С оговоркой: около 88 ГБ, на одной карте в обрез; две карты дают запас	Не помещается на одну карту при 8-битной точности: нужны две карты или меньшая точность	Подходит: на одной карте около 62 % памяти	Избыточно: узел 8 ускорителей	Избыточно: узел 8 ускорителей
Модель 120 млрд в FP8 (около 141 ГБ)	С оговоркой: две карты (192 ГБ)	С оговоркой: две карты по 80 ГБ заполнены почти полностью (около 88 %), запас малый; надёжнее три. Вывод	С оговоркой: на одной карте 100 % памяти, на двух 50 %	Подходит в составе узла	Подходит в составе узла
Модель около 671 млрд в FP8 (около 822 ГБ по расчёту сайта)	Не помещается на 8 картах: 768 ГБ меньше 822 ГБ. Вывод	Не помещается на 8 ускорителях: 640 ГБ меньше 822 ГБ. Вывод	Подходит: узел из 8 карт, около 73 % памяти	Подходит: около 57 % памяти узла	Подходит: около 38 % памяти узла
Длинный контекст, где важны объём и полоса памяти	С оговоркой: полоса памяти ниже, чем у H200 примерно втрое (страница класса)	Нет данных: прямых утверждений на странице класса нет	Подходит: «инференс с длинным контекстом» по странице класса	Нет данных	Нет данных
Обучение одной модели на одном узле из 8 ускорителей	Не подходит: нет межкарточной шины, обучение на 8 картах нецелесообразно	Подходит: узел из 8 ускорителей	Подходит: узел из 8 ускорителей	Подходит: обучение и полное дообучение	Подходит: обучение и полное дообучение крупных моделей
Дообучение LoRA, 1-2 карты	Подходит: рабочие станции и LoRA	Нет данных	Нет данных	Не подходит для 1-2 карт: узел только на 8 ускорителей	Не подходит для 1-2 карт: узел только на 8 ускорителей

Задача	RTX PRO 6000	H100	H200	B200	B300
Рабочая станция или сервер на 2-4 карты	Подходит: серверы 2U-5U на 2-8 карт и станции	Карты PCIe в станциях и серверах на 2-4 карты по каталогу. Вывод	Карты PCIe в серверах на 2-4 карты по каталогу. Вывод	Не подходит: только узел на 8	Не подходит: только узел на 8
Вычисления двойной точности (научные расчёты)	Нет данных	Нет данных	Нет данных	Подходит: «научные расчёты двойной точности», 296 TFLOPS на узел	Не подходит: двойная точность ограничена, около 10 TFLOPS на узел
4-битный формат с плавающей точкой	Нет данных: на странице класса не указано	Не поддерживается	Не поддерживается	Упоминаются 8- и 4-битные вычисления	Подходит: в 4-битном формате заявлен быстрее B200 примерно в 1,5 раза
Расширение имеющегося парка по готовой спецификации	Нет данных	Подходит: «расширение парка H100», «закупка по готовой спецификации»	Для новых проектов страница H100 рекомендует H200 или RTX PRO 6000	Нет данных	Нет данных
Требование «не менее 140 ГБ на ускоритель» в ТЗ	Не подходит: 96 ГБ	Не подходит: 80 или 94 ГБ	Подходит: 141 ГБ	Подходит: 180 ГБ	Подходит: 270 ГБ и более
Нагрузка на площадку: мощность на стойку 20 кВт	2 сервера на стойку (до 9,1 кВт каждый)	1 узел (до 10,2 кВт); по платформе из руководства	1 узел	1 узел (до 14,3 кВт)	1 узел (до 16,4 кВт пик)
Видеоаналитика, голосовой ИИ, виртуальные рабочие места	Нет данных: страницы решений класс не называют	Нет данных	Нет данных	Нет данных	Нет данных

Расхождения между источниками

1. B300: память на ускоритель указана как до 288 ГБ и как 270 ГБ (расчёты сайта); на странице производителя платформы HGX указан общий объём 2,1 ТБ. Сайт использует 270 ГБ.
2. B200: 180 ГБ в системах HGX; 192 ГБ в описаниях других систем. Сайт использует 180 ГБ; на странице производителя платформы HGX указан общий объём 1,4 ТБ.
3. Мощность узла 8 × H100 или 8 × H200 до 10,2 кВт относится к платформе из руководства производителя; другие производители на той же плате публикуют до 15 кВт.
4. Пропускная способность памяти и TDP для B200 и B300 взяты со страниц сайта (первичный источник: блоги и руководства производителя); страница производителя платформы HGX их не публикует.
5. Общая память узла B300 в таблице каталога указана как 2,3 ТБ и 2,16 ТБ.

Что матрица не говорит

1. Она не ранжирует классы и не называет победителя: выбор определяют память, связь между ускорителями, площадка и закон закупки.
2. Она не содержит сумм, сроков поставки и наличия.
3. Она не заменяет расчёт памяти на вашу модель.

Что уточнит инженер

1. Реальную нагрузку: число одновременных запросов, длину контекста, точность модели. Матрица считает один запрос, FP8, 8 192 токена.
2. Скорость работы: публичных измерений скорости для ваших моделей на сайте нет, расчёт памяти даёт нижнюю границу по объёму.
3. Системную мощность, массу и высоту выбранной модели по паспорту.
4. Подходит ли площадка: мощность на стойку, охлаждение, два ввода (лист «Паспорт площадки»).
5. Режим закупки и пороговые значения памяти для ТЗ.

6. Статус наличия и срок: уточняются в КП на дату запроса.

Справочная сводка по открытым данным на 06.10-25.10.2026 (даты у каждой таблицы). Не является офертой (ст. 437 ГК РФ). Названия классов использованы в словесной форме, логотипы производителей не используются. ООО «ВИСТЛАН», Москва, с 2006 года. gpuservers.ru. Версия от 04.11.2026.