

ТЕХНИЧЕСКИЙ ДОКУМЕНТ ДЛЯ ИТ-КОМИТЕТА

Отечественный стек для ИИ 2026: что реально доступно

Инфраструктура для on-premise LLM в закрытом контуре: GPU-альтернативы, российские серверы, программный стек, гибридные сценарии и матрица проверки перед закупкой.

GPU-альтернативы

Серверы и реестр

Софт-стек

Гибрид · дорожная карта

Матрица проверки + чек-лист

Все характеристики оборудования и оценки производительности — **на дату редакции**. Цифры вендоров отделены от независимых оценок; реестры и правила закупок меняются ежеквартально — сверяйте на дату извещения. Любую цифру, от которой зависит решение, подтверждайте бенчмарком на своей задаче.

Три ограничения, на пересечении которых живёт решение

Стоимость самого быстрого ускорителя видно в прайсе. Стоимость того, что он не пройдёт по реестру или останется без сервиса через два года, — нет. Именно вторая половина уравнения чаще определяет, какую конфигурацию организация реально сможет эксплуатировать.

У коммерческой компании выбор упирается в цену и сроки. У КИИ-объекта или госзаказчика поверх этого ложатся ещё два независимых слоя, и оптимум по одному легко проваливается по другому.

Ограничение	Что проверяем	Чем оборачивается провал
Техника	Производительность под профиль нагрузки, объём и пропускная способность памяти ускорителя, топология узла	Модель не влезает или работает медленно; переплата за неиспользуемую мощность
Реестр и закупка	Запись в реестре Минпромторга (ГИСП), баллы локализации под действующую редакцию правил, приоритет отечественного на торгах	Позиция отклоняется на торгах или признаётся иностранной
Эксплуатация	Сервис и ЗИП в РФ, гарантия при параллельном импорте, документация, совместимость с отечественной ОС	Через 2–3 года оборудование превращается в неподдерживаемый риск

GPU-ускорители: что реально доступно

Базовая точка отсчёта — ускорители NVIDIA (H100 и новее) с экосистемой CUDA. Не из-за маркетинга, а потому что вокруг них за пятнадцать с лишним лет сложился зрелый софтверный слой — cuDNN, NCCL, TensorRT, глубокая интеграция с PyTorch. Любая альтернатива сравнивается именно с этой базой, и чаще проигрывает не в цифрах, а в предсказуемости.

Вариант	Память (ориентир)	Производительность	Софт и зрелость	Назначение и риски
NVIDIA H100 / новее	80–96 ГБ HBM	Промышленный эталон	CUDA, PyTorch «из коробки»; зрелость высокая	Обучение + inference. Риски: цена, гарантия при параллельном импорте, доступность
Huawei Ascend 910C	до ~128 ГБ HBM (в Atlas)	~60–70% H100 на inference (независимые оценки); вендор заявляет выше	CANN + torch_npu, MindSpore; зрелость средняя	Прежде всего inference. Риски: незрелый софт, стабильность torch_npu, поддержка
Ascend 910B	до 64 ГБ HBM	Ниже 910C	CANN + torch_npu; зрелость средняя	Inference. Риски те же, поколение старше
Moore Threads / lluvatar	зависит от модели	требует проверки	свои стеки, частичный PyTorch; зрелость низкая для LLM	Нишевый inference. Риски: доступность, поддержка, софт

Вариант	Память (ориентир)	Производительность	Софт и зрелость	Назначение и риски
МЦСТ «Эльбрус» (CPU)	системная RAM	не GPU-класс	LCC, нативная сборка; своя экосистема	CPU-нагрузки, не тяжёлый LLM. Не заменяет GPU под LLM

Диапазоны и оценки — на дату редакции; по неамериканским ускорителям источники расходятся в разы. Подтверждать бенчмарком на своей модели.

Частое заблуждение. «Эльбрус» — универсальный процессор (VLIW), а не GPU-ускоритель. На нём запускают отдельные нейросетевые задачи, но обучение и высоконагруженный inference LLM — не его класс. Для доверенных CPU-нагрузок в госсегменте полезен; как замену GPU под LLM закладывать не стоит. «Эльбрус-Б» — проект с горизонтом 2027+, в критический путь закупки на 2026 год не встаёт.

Практический ориентир: объём памяти ускорителя часто важнее пиковых TFLOPS. Для inference LLM решает, сколько параметров и какой длины контекст физически помещается в карту и с какой скоростью до них добираться. Отсюда интерес к платформам с большим суммарным HBM в узле — на них влезает крупная модель без дробления по картам.



СЕРВЕРЫ · ПРОГРАММНЫЙ СТЕК

Серверные платформы: где импортозамещение работает

С серверами ситуация заметно здоровее, чем с ускорителями. Российские вендоры выпускают платформы из реестра Минпромторга, в том числе шасси под несколько GPU — встречаются модели с поддержкой до восьми полноразмерных ускорителей, а материнские платы собственной разработки уже попадают в реестр как продукция с российским уровнем локализации.

Нюанс для отчётности: «российский сервер» — это чаще всего про шасси, материнскую плату, BMC, прошивку и локальную сборку. Процессоры обычно Intel Xeon или AMD EPYC; отечественные CPU (Baikal-S, «Эльбрус») в GPU-серверах пока встречаются реже. Сам ускоритель в большинстве случаев остаётся импортным. Реестровая запись подтверждает происхождение платформы, но не делает импортный GPU внутри неё отечественным.

Питание и охлаждение

GPU-узел на 6–8 карт — десятки киловатт (у плотных inference-платформ до ~15 кВт). Типовой срыв: узел подобрали по производительности и реестру, а серверная не держит плотность по питанию — монтаж уезжает на квартал под реконструкцию электрики.

Топология и межсоединения

Сколько GPU в узле и как связаны — PCIe, аналоги NVLink, внешние PCIe-фабрики. Хватает ли пропускной способности под вашу модель.

Сервис и гарантия

На шасси у российского вендора это решается лучше, чем на самом GPU. Проверять нужно контур целиком, вместе с ускорителем — включая условия гарантии при параллельном импорте.

Σ Программный стек: настоящая граница импортозамещения

Про железо спорят на совещаниях, а срывает сроки обычно софт. Базовый стек индустрии — PyTorch поверх CUDA; за пятнадцать лет он оброс библиотеками и отлажен так, что «просто работает». С этим «просто работает» и сравнивается любая альтернатива. Для Ascend она существует: CANN (аналог CUDA от Huawei) и адаптер torch_npu, запускающий PyTorch-модели на NPU через PrivateUse1. Huawei в 2025 открыл исходники CANN — шаг верный, но зрелость отстаёт: адаптер живёт вне основного дерева PyTorch, изменения в ядре не тестируются на совместимость автоматически, часть моделей требует ручной адаптации.

Слой стека	Статус импортозамещения	Комментарий
Обучение/inference на NVIDIA (CUDA+PyTorch)	Импортное, но зрелое	Эталон предсказуемости
Inference на Ascend (CANN+torch_npu)	Доступно, зрелость средняя	Работает на типовых открытых моделях; кастом требует усилий
Inference-движки (vLLM, TGI и аналоги)	Открытый софт	Заменять нечего — стек и так открытый
Контейнеризация (Docker, Kubernetes)	Открытый софт на росс. Linux	Живёт на Astra Linux, РЕД ОС и др.
Мониторинг (Prometheus, Grafana, Zabbix)	Открытый софт	Проблемы импортозамещения нет

Предел импортозамещения проходит именно здесь — по софту под неNVIDIA-ускорители. Железо купить можно; заставить его стабильно крутить вашу конкретную нагрузку — отдельная работа со своим бюджетом и сроками. Её и планировать надо как проект, а не как строчку «установка и настройка» в смете.



ГИБРИД · СТОИМОСТЬ

◆ Гибридная архитектура: четыре рабочих сценария

Раз «чистого» отечественного стека под тяжёлый ИИ в 2026 году нет, разумная архитектура — гибридная.

1 NVIDIA под обучение, альтернативы под inference

Тяжёлый fine-tuning и обучение — на NVIDIA; высоконагруженный inference обученной модели — на Ascend или отечественных ускорителях. Разделение по типу нагрузки, а не по вендору.

2 Российское шасси + доступные GPU

Платформа из реестра (баллы локализации, проходит по закупке) плюс реально доступные ускорители. Честно проходит и по реестру, и по технике.

3 On-premise inference + внешние эксперименты

Чувствительные данные — только в закрытом контуре; ресурсоёмкие эксперименты на обезличенных данных — во внешней инфраструктуре. Экономит CAPEX на пиковые нагрузки.

4 Постепенная миграция на отечественный стек

Начать с гибрида, переносить нагрузки по мере созревания альтернатив и их софта. Многолетняя программа с контрольными точками, а не «переключение».

Схема в закрытом контуре. Внешняя рамка — **закрытый контур КИИ**. Внутри: **отечественный сервер из реестра Минпромторга**, в узле два блока ускорителей — **NVIDIA · обучение** и **Ascend / отечественные · inference**. Поверх — **открытый софт-стек на российском Linux**: PyTorch, vLLM/TGI, Kubernetes, Prometheus/Grafana. Гибрид позволяет запустить ИИ в контуре сейчас, набрать баллы локализации на уровне платформы и не захлопнуть дверь для миграции.



Структура расчёта стоимости

Готовых цен здесь нет намеренно: рынок GPU волатилен, курс и каналы поставки меняют смету на десятки процентов за квартал. Полезнее структура, по которой считать под свой проект.

Статья	Что входит	Где чаще недосчитывают
CAPEX	Ускорители, серверные платформы, сеть, СХД, стойки	Модернизация электрики и охлаждения ЦОДа
ОРЕХ	Электроэнергия, охлаждение, аренда/амортизация ЦОДа	Энергопотребление GPU-узлов — заметная статья
Поддержка и сервис	Контракты на обслуживание, ЗИП	Риски параллельного импорта по гарантии
Лицензии и софт	В основном открытый софт	Время инженеров на портирование под железо
Команда эксплуатации	MLOps/инфраструктурные инженеры	Самая дефицитная и недооценённая статья
Риск поставки	Резерв на срыв/удорожание поставки ускорителей	Альтернативную конфигурацию закладывают редко

Опаснее всего для сметы не строчка с GPU, хотя она самая крупная. Бюджет чаще срывают три невидимые в прайсе вещи: инфраструктура ЦОДа, стоимость портирования софта под nVIDIA-железо и люди на эксплуатацию.



Матрица проверки позиции оборудования

По каждой позиции до внесения в ТЗ полезно заполнить короткую матрицу. Позиция, проходящая по производительности, но проваливающаяся по любому столбцу, для КИИ-проекта нерабочая.

Критерий	Вопрос	Статус
Реестр	Есть ли актуальная запись в ГИСП под эту модель и на какой срок?	☐ да / ☐ нет
Баллы	Набирает ли позиция минимальный порог локализации под действующую редакцию правил?	☐ да / ☐ нет
Сервис	Есть ли сервис, ЗИП и гарантия на территории РФ?	☐ да / ☐ нет
ОС	Подтверждена ли совместимость с целевой ОС протоколами, а не маркетингом?	☐ да / ☐ нет
Софт	Прогнан ли бенчмарк вашего стека и модели на этом железе?	☐ да / ☐ нет



Чек-лист перед закупкой

☐ Определён профиль нагрузки: обучение / fine-tuning / inference и их пропорции

☐ Под профиль подобран ускоритель, а не наоборот

☐ Проверена актуальная реестровая запись (ГИСП) на дату извещения

☐ Рассчитаны баллы локализации под действующую редакцию правил

☐ Подтверждён сервис и ЗИП на территории РФ

☐ Совместимость с целевой ОС подтверждена протоколами

☐ Прогнан бенчмарк софт-стека на вашей модели

☐ Оценены питание и охлаждение ЦОДа под реальный TDP узла

☐ Заложен риск поставки и альтернативная конфигурация

☐ Посчитана команда эксплуатации, а не только железо

☐ В ТЗ разведены «российское шасси» и «российская инфраструктура целиком»

☐ Согласован путь миграции на случай, если альтернативы созреют

→ Дорожная карта 2026–2028

● Краткосрочно · 2026

Гибридная архитектура. Российское шасси из реестра плюс доступные ускорители, разделение нагрузок обучение/inference, запуск в закрытом контуре сейчас — не дожидаясь «идеального» стека.

● Среднесрочно · 2026–2027

Стандартизация платформ и софт-обвязки, накопление собственных бенчмарков под альтернативные ускорители, отработка портирования inference на неNVIDIA-железо на некритичных нагрузках.

● Долгосрочно · 2027–2028

Перенос нагрузок на зрелые отечественные и доступные альтернативы — по мере подтверждения производительности, стабильности и сервиса реальными пилотами, а не анонсами.

Собрать конфигурацию под вашу инфраструктуру

Если вы прикидываете инфраструктуру под свой профиль нагрузки и требования КИИ, AZONE-AI может провести экспресс-аудит вводных, проверить кандидатов по реестру и совместимости и предложить архитектуру пилота в закрытом контуре — без обязательств с вашей стороны на этом этапе.

Обсудить вводные

azoneai.ru/contact · услуга «Развёртывание on-premise LLM» · лицензии ФСТЭК, ФСБ, МО РФ