УТВЕРЖДАЮ
Генеральный директор
ООО «НАЙС СОФТ ГРУПП»
_____ С.С. Беликов
«____» _____ 2026 г.

ПРОГРАММНЫЙ КОМПЛЕКС «НАЙС СОФТ ИИ СТУДИЯ»

версия 1.0.0

СИСТЕМНЫЕ ТРЕБОВАНИЯ

НССГ.10001-01 90 02-ЛУ

Лист утверждения

Красногорск
2026

Утвержден НССГ.10001-01 90 02-ЛУ

ООО «НАЙС СОФТ ГРУПП»

ПРОГРАММНЫЙ КОМПЛЕКС «НАЙС СОФТ ИИ СТУДИЯ»

Версия 1.0.0

СИСТЕМНЫЕ ТРЕБОВАНИЯ

НССГ.10001-01 90 02

Листов 51

Красногорск
2026

АННОТАЦИЯ

Настоящий документ устанавливает минимальные, рекомендуемые и проектные требования к вычислительной, программной и сетевой среде программного комплекса «Найс Софт ИИ Студия» версии 1.0.0. Требования предназначены для планирования закупки ресурсов, подготовки инфраструктуры, внедрения, квалификации совместимости и последующей эксплуатации.

Документ рассматривает базовый комплекс отдельно от локальной модели искусственного интеллекта. Студия может эксплуатироваться без локального графического ускорителя при использовании внешней либо управляемой модели. Для локального производственного вывода требования выбранной модели добавляются к требованиям базовых служб.

Основной, референсной и рекомендуемой операционной системой является НАЙС.ОС CLOUD. Заявленная совместимость распространяется на российские серверные операционные системы, удовлетворяющие требованиям настоящего документа и предоставляющие совместимую контейнерную среду Linux/OCI. Подтверждение конкретной версии ОС оформляется матрицей совместимости и протоколом испытаний.

ПРЕДУПРЕЖДЕНИЕ. Минимальные значения означают нижнюю границу поддерживаемого производственного запуска при типовой конфигурации. Для систем с высокой параллельностью, большим объёмом корпоративных данных, длинным контекстом или локальными крупными моделями выполняется отдельный расчёт ресурсов.



Рисунок 1 - Слои, формирующие итоговые системные требования

СОДЕРЖАНИЕ

1 Общие положения 6
2 Термины и уровни требований 7
3 Поддерживаемая архитектура вычислительной платформы 8
4 Требования к операционной системе 9
5 Требования к контейнерной среде 10
6 Минимальная конфигурация без локальной модели 11
7 Рекомендуемые профили серверов 12
8 Требования к процессору 13
9 Требования к оперативной памяти 14
10 Требования к дисковой подсистеме 15
11 Требования к файловой системе 16
12 Требования к сети 17
13 Сетевые порты 18
14 DNS, время и сертификаты 19
15 Требования к браузеру и рабочему месту 20
16 Требования к виртуализации 21
17 Требования к облачной инфраструктуре 22
18 Требования к локальному графическому ускорителю 23
19 Классы GPU и вычислительная совместимость 24
20 Требования локальных моделей 25
21 Контекст, кэш и параллельность 26
22 Работа без локального GPU 27
23 Корпоративные базы знаний и RAG 28
24 Требования к интеграциям 29
25 Требования к автоматизациям 30
26 Требования к резервному копированию 31
27 Требования к мониторингу и журналированию 32
28 Требования закрытого контура 33
29 Требования полностью изолированного контура 34
30 Требования информационной безопасности среды 35
31 Требования к масштабированию 36
32 Планирование производительности 37
33 Ограничения и неподдерживаемые конфигурации 38
34 Порядок подтверждения соответствия 39
35 Требования к испытательной среде 40
36 Нормативные и справочные сведения 41
Приложение А. Сводная таблица системных требований 42
Приложение Б. Профили серверов 43
Приложение В. Требования локальных моделей 44
Приложение Г. Формула расчёта хранилища 45

Приложение Д. Матрица сетевых требований 46
Приложение Е. Контрольный лист подготовки инфраструктуры 47
Приложение Ж. Контрольный лист GPU 48
Приложение И. Требования для российских ОС 49
Приложение К. Паспорт расчёта ресурсов 50

1 ОБЩИЕ ПОЛОЖЕНИЯ

1.1 Назначение

Системные требования определяют характеристики среды, при которых ИИ Студия 1.0.0 может быть установлена, запущена, проверена и эксплуатироваться с заявленными функциями. Документ используется совместно с руководством по установке, руководством администратора и матрицей совместимости.

1.2 Область применения

- локальные серверы и центры обработки данных;
- виртуальные машины;
- частные облака;
- публичные облака, включая Яндекс Облако;
- серверы с локальными графическими ускорителями;
- установки без локальной модели;
- закрытые и полностью изолированные контуры.

1.3 Разграничение документов

Документ	Отвечает на вопрос
Системные требования	Какие ресурсы и свойства среды необходимы?
Матрица совместимости	Какие конкретные ОС, браузеры, гипервизоры, GPU и другие продукты испытаны?
Руководство по установке	Как подготовить среду и установить продукт?
Руководство администратора	Как настроить и сопровождать установленный экземпляр?

2 ТЕРМИНЫ И УРОВНИ ТРЕБОВАНИЙ

Уровень	Смысл
Минимальный	Нижняя поддерживаемая граница для запуска указанного профиля при ограниченной нагрузке.
Рекомендуемый	Профиль, обеспечивающий эксплуатационный запас для типового корпоративного применения.
Расширенный	Профиль для повышенной нагрузки, больших баз знаний или множества параллельных операций.
Требование модели	Дополнительные ресурсы, зависящие от конкретного локального модельного артефакта.
Проектный	Значение рассчитывается для конкретной организации и подтверждается нагрузочными испытаниями.

ПРИМЕЧАНИЕ. Рекомендуемая конфигурация не является гарантией заданного времени ответа модели: скорость генерации зависит от модели, длины контекста, ускорителя, числа параллельных запросов и внешнего поставщика.

3 ПОДДЕРЖИВАЕМАЯ АРХИТЕКТУРА ВЫЧИСЛИТЕЛЬНОЙ ПЛАТФОРМЫ

Основной производственный профиль версии 1.0.0 рассчитан на 64-разрядную архитектуру x86_64 (AMD64). Процессор должен поддерживать стандартный современный набор команд x86-64, аппаратную виртуализацию при размещении в ВМ и механизмы, необходимые используемой серверной ОС и Docker.

Характеристика	Требование
Архитектура	x86_64 / AMD64
Режим процессора	64-разрядный
Виртуализация	Intel VT-x/AMD-V при использовании гипервизора
aarch64	Не заявляется базовым производственным профилем 1.0.0; возможна отдельная квалификация.
NUMA	Допускается; для многопроцессорных GPU-систем требуется проектная настройка.

4 ТРЕБОВАНИЯ К ОПЕРАЦИОННОЙ СИСТЕМЕ

Основная референсная платформа - НАЙС.ОС CLOUD. НАЙС.ОС включена в Единый реестр российского ПО: реестровая запись №30128 от 22.10.2025. Для ИИ Студии требуется серверная 64-разрядная Linux-среда, способная запускать поддерживаемый Docker Engine и предоставлять стандартные возможности ядра Linux, сетевых пространств имён, cgroups и файловых систем.

ИИ Студия заявляется совместимой со всеми российскими серверными операционными системами, которые удовлетворяют указанным требованиям. В эту категорию могут входить НАЙС.ОС CLOUD, Astra Linux, РЕД ОС, семейство «Альт», РОСА, МСВСфера, ОСнова, AlterOS, GosLinux и другие российские серверные Linux ОС. Конкретная версия и редакция получает статус «подтверждено» только после испытаний, отражённых в матрице совместимости.

Требование ОС	Обязательность
64-разрядное Linux-ядро	обязательно
Поддержка Docker Engine и Compose	обязательно
cgroups v2	рекомендуется; референсная конфигурация
Синхронизация времени	обязательно
Работающий DNS resolver	обязательно
Поддержка TLS 1.2/1.3	обязательно
SELinux/MAK	допускается; политики должны разрешать штатный запуск контейнеров и каталогов данных
Пакетный менеджер конкретного дистрибутива	не влияет на приложение после подготовки Docker-среды

ПРИМЕЧАНИЕ. Заявленная совместимость с семейством российских ОС не означает, что каждая историческая версия каждого дистрибутива прошла испытания. Для закупки и промышленного проекта следует использовать матрицу совместимости с конкретными версиями.

5 ТРЕБОВАНИЯ К КОНТЕЙНЕРНОЙ СРЕДЕ

- работающий Docker Engine промышленной ветви, поддерживаемой правообладателем ОС;
- подключаемый модуль Docker Compose v2;
- возможность создания внутренних bridge-сетей и постоянных томов/привязок каталогов;
- доступ к cgroups и namespaces;
- корректная работа DNS внутри контейнерных сетей;
- для локального GPU - доступ контейнеров к NVIDIA GPU через NVIDIA Container Toolkit/CDI или эквивалентный поддерживаемый механизм.

```
docker version
docker compose version
docker info
```

В производственной поставке сборка контейнеров на целевом сервере не требуется: используются заранее подготовленные и идентифицированные образы, определённые манифестом поставки.

6 МИНИМАЛЬНАЯ КОНФИГУРАЦИЯ БЕЗ ЛОКАЛЬНОЙ МОДЕЛИ

Ресурс	Минимум	Комментарий
Процессор	8 vCPU / 8 физических потоков	Для базовых служб, небольшого числа пользователей и внешней/управляемой модели.
Оперативная память	16 ГБ	Нижняя производственная граница. Подкачка не должна использоваться как замена ОЗУ.
Диск	120 ГБ SSD	Без локальных модельных артефактов и больших корпоративных коллекций.
Сеть	1 Гбит/с	Для локальной сети; внешняя пропускная способность зависит от интеграций и поставщика модели.
GPU	не требуется	При внешней или управляемой модели.
Архитектура	x86_64	Базовый профиль 1.0.0.

ПРИМЕЧАНИЕ. Конфигурации ниже 8 vCPU / 16 ГБ могут использоваться только для инженерных экспериментов по согласованию и не являются базовой производственной конфигурацией.

7 РЕКОМЕНДУЕМЫЕ ПРОФИЛИ СЕРВЕРОВ



Рисунок 2 - Рекомендуемые профили вычислительных ресурсов

Профиль	CPU	ОЗУ	Диск	Назначение
S1 Базовый	8 vCPU	16 ГБ	120 ГБ SSD	Небольшая организация, внешняя/управляемая модель.
S2 Типовой	16 vCPU	32 ГБ	250 ГБ SSD/NVMe	Типовая корпоративная эксплуатация, RAG, интеграции, автоматизации.
S3 Расширенный	24-32 vCPU	64 ГБ	500 ГБ+ NVMe	Большая база знаний, активные интеграции, повышенная параллельность.
S4 Локальный ИИ	S2/S3 + GPU	по модели	+ хранилище модели	Локальный производственный вывод.

8 ТРЕБОВАНИЯ К ПРОЦЕССОРУ

Процессор обслуживает веб-интерфейс, базы данных, векторный поиск, интеграционный шлюз, автоматизации, мониторинг и вспомогательные службы. Даже при локальном GPU процессорные ресурсы остаются необходимыми.

- не менее 8 vCPU для минимального профиля;
- 16 vCPU рекомендуется для типового корпоративного узла;
- 24-32 vCPU и более рекомендуется при высокой активности, интенсивных автоматизациях и крупных индексах;
- для многопроцессорных систем необходимо учитывать NUMA и близость GPU к CPU/PCIe при проектировании;
- переподписка vCPU гипервизором допускается только при сохранении требуемой фактической производительности.

9 ТРЕБОВАНИЯ К ОПЕРАТИВНОЙ ПАМЯТИ

Оперативная память используется контейнерными службами, PostgreSQL/совместимыми базами, Redis, поисковым и векторным контуром, индексаторами, автоматизациями и модельной средой. Для локальной модели требования каталога модели добавляются к эксплуатационному запасу базовой Студии.

Сценарий	ОЗУ
Минимальный без локальной модели	16 ГБ
Типовой без локальной модели	32 ГБ
Большая база знаний / высокая активность	64 ГБ и более
Локальная модель 4B класса	32 ГБ и более на весь сервер
Локальная модель 9B класса	64 ГБ и более на весь сервер
Локальная модель 27-30B класса	96-128 ГБ и более на весь сервер

ПРИМЕЧАНИЕ. Значение min_ram_gb в модельном каталоге относится к модели/среде выполнения и не должно трактоваться как полный объём ОЗУ всего сервера. Базовые службы Студии также потребляют память.

10 ТРЕБОВАНИЯ К ДИСКОВОЙ ПОДСИСТЕМЕ



Рисунок 3 - Состав дискового бюджета

Компонент	Ориентир
ОС и контейнерная среда	не менее 30 ГБ
Образы и служебный запас Студии	не менее 60 ГБ
Постоянные данные	по фактическому объёму файлов, индексов и истории
Временные операции/обновления	не менее 20-30 ГБ свободного резерва
Локальная модель	размер артефакта × 1,15 + минимум 5 ГБ на установку
Резервные копии	внешнее хранилище; объём определяется сроком хранения и числом копий

Для базового экземпляра без локальных моделей требуется не менее 120 ГБ SSD. Для типовой производственной установки рекомендуется 250 ГБ. При активной работе с документами и локальными моделями рекомендуется NVMe-хранилище от 500 ГБ.

10.1 Производительность диска

- SSD является минимально рекомендуемым классом носителя;
- NVMe рекомендуется для больших баз знаний, множества документов и локальных моделей;
- HDD допускается для внешнего архивного хранения резервных копий, но не рекомендуется для основных баз и векторного индекса;
- сетевое хранилище допускается только при предсказуемой задержке, устойчивом fsync и подтверждённых испытаниях.

11 ТРЕБОВАНИЯ К ФАЙЛОВОЙ СИСТЕМЕ

- поддержка POSIX-семантики, прав доступа, символических ссылок и блокировок;
- достаточная устойчивость к сбоям питания/хоста;
- корректная работа Docker overlay2 либо поддерживаемого драйвера хранения;
- не использовать сетевую файловую систему для критичных баз данных без подтверждённой совместимости;
- для Btrfs/XFS/ext4 допускаются штатные конфигурации ОС; конкретные параметры фиксируются в матрице совместимости при необходимости.

12 ТРЕБОВАНИЯ К СЕТИ

Показатель	Требование
Локальная сеть	1 Гбит/с минимум; 10 Гбит/с рекомендуется для больших файловых потоков/централизованного хранилища.
Задержка до пользователей	не нормируется жёстко; для комфортного интерфейса рекомендуется LAN/WAN с устойчивой связью без значительных потерь.
DNS	прямое и, при требованиях организации, обратное разрешение имени экземпляра.
NTP	устойчивая синхронизация времени.
MTU	единый согласованный MTU; при overlay/VPN необходимо исключить фрагментацию/PMTU-проблемы.
Прокси	допускается; должен поддерживать HTTPS и разрешённые назначения поставки.

13 СЕТЕВЫЕ ПОРТЫ

Порт	Направление	Назначение	Обязательность
8143/TCP HTTPS	входящий	Основной веб-доступ в типовой поставке	по умолчанию
443/TCP HTTPS	входящий	Внешний обратный посредник при публикации на стандартном порту	по архитектуре
80/TCP HTTP	входящий	ACME-проверка /.well-known/acme-challenge/	только при необходимости
22/TCP SSH	входящий	Администрирование ОС	по политике организации
443/TCP HTTPS	исходящий	Образы, обновления, внешние модели и интеграции	по выбранным функциям
53/UDP,TCP	исходящий/внутренний	DNS	обязательно через инфраструктуру организации
123/UDP	исходящий/внутренний	NTP	обязательно через инфраструктуру организации

Внутренние базы, кэши, векторные и административные службы не должны публиковаться во внешнюю сеть отдельными портами, если проектом явно не предусмотрено обратное.

14 DNS, ВРЕМЯ И СЕРТИФИКАТЫ

- стабильное DNS-имя для производственного экземпляра;
- синхронизация времени всех узлов с доверенным источником;
- TLS-сертификат, доверенный рабочими станциями пользователей;
- соответствие DNS-имени полям сертификата;
- в закрытом контуре - внутренний центр сертификации либо заранее распределённый локальный центр доверия.

ПРИМЕЧАНИЕ. Расхождение времени может нарушать проверку TLS, подписей, временных токенов и журналирование событий.

15 ТРЕБОВАНИЯ К БРАУЗЕРУ И РАБОЧЕМУ МЕСТУ

Клиентская часть работает через современный браузер. На рабочей станции не требуется отдельный толстый клиент ИИ Студии.

Параметр	Требование
Браузер	Современная поддерживаемая версия Chromium/Яндекс Браузера, Firefox/Firefox ESR или другого браузера с сопоставимой поддержкой веб-стандартов.
JavaScript	включён
Хранилище браузера/cookie	разрешено в объёме, необходимом для сеанса
WebSocket/SSE/потоковые ответы	не должны блокироваться корпоративным посредником
TLS	1.2 или 1.3
Разрешение	1366×768 минимум; 1920×1080 рекомендуется
Масштабирование	100-200 % должно оставаться работоспособным; фактическая доступность проверяется отдельно
ОС рабочего места	не зависит от серверной ОС; требуется поддерживаемый браузер и сетевой доступ

16 ТРЕБОВАНИЯ К ВИРТУАЛИЗАЦИИ

- виртуальная машина должна получать гарантируемые CPU/RAM и производительный постоянный диск;
- динамическое ballooning ОЗУ не должен снижать память ниже минимального профиля;
- время гостевой ОС должно быть синхронизировано и не конфликтовать с механизмами гипервизора;
- снимок ВМ не заменяет штатную резервную копию ИИ Студии;
- при GPU passthrough/vGPU ускоритель должен быть виден гостевой ОС и Docker;
- live migration GPU-ВМ допускается только если это поддерживает конкретная платформа и подтверждено испытанием.

17 ТРЕБОВАНИЯ К ОБЛАЧНОЙ ИНФРАСТРУКТУРЕ

- стабильный внутренний адрес или DNS;
- постоянный системный/данный диск требуемого класса IOPS;
- правила облачного межсетевого экрана только для необходимых портов;
- отдельная политика исходящих соединений;
- для GPU - профиль ВМ, поддерживаемый текущей средой выполнения;
- резервные копии должны экспортироваться за пределы единственной ВМ/диска.

В Яндекс Облаке каталог 1.0.0 содержит профили T4, T4i, A100 и новых GPU-платформ. Некоторые аппаратные варианты требуют проверки непосредственно на созданной ВМ, поскольку облачный поставщик может не публиковать все низкоуровневые характеристики.

18 ТРЕБОВАНИЯ К ЛОКАЛЬНОМУ ГРАФИЧЕСКОМУ УСКОРИТЕЛЮ

- NVIDIA GPU, совместимый с закреплённой средой vLLM;
- для основного производственного профиля вычислительная совместимость CUDA Compute Capability не ниже 7.5;
- достаточный объём VRAM для весов модели, кэша контекста и эксплуатационного резерва;
- драйвер NVIDIA, совместимый с контейнерной средой поставки;
- nvidia-smi должен успешно работать на хосте;
- контейнерная проверка GPU должна успешно видеть устройство;
- ECC, охлаждение, питание и аппаратная надёжность определяются требованиями серверной площадки.

```
nvidia-smi -L
nvidia-smi
docker run --rm --gpus all <проверочный-образ> nvidia-smi
```

19 КЛАССЫ GPU И ВЫЧИСЛИТЕЛЬНАЯ СОВМЕСТИМОСТЬ

Класс	VRAM	CC	Статус в базовой среде 1.0.0	Комментарий
NVIDIA T4	16 ГБ	7.5	поддерживаемый	Компактные/ квантизованные модели.
NVIDIA A100	80 ГБ	8.0	поддерживаемый	Основной класс для крупных корпоративных моделей.
Tesla V100	32 ГБ	7.0	не основной производственный профиль	Закреплённая среда требует CC 7.5+; отдельная квалификация возможна с иной средой.
GPU с CC 8.9+	различно	8.9+	поддерживается при совместимом драйвере	Необходим для некоторых FP8-профилей.
T4i/Gen2/V4 облачные	по профилю	проверяется на ВМ	проверка на конкретном узле	Статус зависит от фактического устройства и драйвера.

20 ТРЕБОВАНИЯ ЛОКАЛЬНЫХ МОДЕЛЕЙ

Ниже приведены представительные модели из каталога 1.0.0. Полный перечень может быть шире; авторитетными являются сведения модельного каталога поставки и результат автоматического анализа оборудования.

Модель	Артефакт, ГБ	Мин. VRAM, ГБ	Рек. VRAM, ГБ	Мин. RAM модели, ГБ	Рек. контекст	Мин. CC
Qwen3.5 4B	9.4	14	16	16	16384	7.5
Qwen3.5 9B	19.4	23	32	32	8192	7.5
Qwen3.8 27B	55.7	72	80	96	65536	7.5
Qwen3.8 27B · FP8	30.9	48	64	64	131072	8.9
Qwen3 Coder 30B-A3B	61.2	76	80	96	32768	7.5

20.1 Рекомендуемые полные профили узла

Класс модели	CPU	ОЗУ сервера	GPU	Диск сервера
4B	8-16 vCPU	32 ГБ+	16 ГБ VRAM	200 ГБ+
9B	16 vCPU	64 ГБ+	24-32 ГБ VRAM	250 ГБ+
27B BF16	24-32 vCPU	128 ГБ+	80 ГБ VRAM	350 ГБ+
27B FP8	24-32 vCPU	96 ГБ+	64 ГБ VRAM, CC 8.9+	300 ГБ+
30B Coder	32 vCPU	128 ГБ+	80 ГБ VRAM	400 ГБ+
Многопроцессорный профиль	48+ vCPU	256 ГБ+	2×80 ГБ и более	500 ГБ+

ПРИМЕЧАНИЕ. Эти профили предназначены для планирования целого узла, а не только процесса модели. Итоговая конфигурация подтверждается предварительной проверкой и нагрузочным испытанием.

21 КОНТЕКСТ, КЭШ И ПАРАЛЛЕЛЬНОСТЬ

Объём VRAM зависит не только от числа параметров модели. Значительную часть памяти может занимать кэш ключей/значений (KV cache), который растёт с длиной контекста и числом одновременно обслуживаемых последовательностей.

- увеличение максимального контекста повышает требования к VRAM;
- повышение числа параллельных запросов увеличивает резерв кэша и нагрузку на CPU/RAM;
- Tensor Parallel распределяет модель между несколькими GPU, но требует подходящей топологии и пропускной способности межсоединения;
- prefix caching может уменьшать повторные вычисления, но также использует память;
- производственный профиль должен оставлять резерв VRAM и не работать на постоянном пределе памяти.

22 РАБОТА БЕЗ ЛОКАЛЬНОГО GPU

GPU не является обязательным системным требованием для ИИ Студии в целом. Без GPU доступны веб-интерфейс, пользователи и роли, корпоративные знания, интеграции, автоматизации, мониторинг и работа с внешними/управляемыми моделями.

Функция	**Без GPU**
Веб-интерфейс	да
Пользователи/RBAC	да
RAG и базы знаний	да
Интеграции	да
Автоматизации	да
Внешние модели	да
Управляемые модели Яндекса	да, при настроенном доступе
Локальный производственный vLLM	нет
Диагностический CPU-режим	может применяться для отдельных тестовых сценариев; не является основным производственным профилем

23 КОРПОРАТИВНЫЕ БАЗЫ ЗНАНИЙ И RAG

При работе с большими коллекциями документов основными ограничивающими ресурсами становятся диск, IOPS, CPU и ОЗУ для разбора документов и построения/обслуживания индексов.

- для небольших коллекций достаточно типового профиля S2;
- при десятках/сотнях гигабайт документов рекомендуется NVMe и не менее 64 ГБ ОЗУ;
- объём векторного индекса зависит от числа фрагментов и размерности векторов и планируется отдельно;
- массовая первичная загрузка документов должна выполняться в контролируемое окно либо с ограничением параллельности;
- резервное копирование должно учитывать исходные файлы, метаданные и базы/индексы, определённые штатным механизмом продукта.

24 ТРЕБОВАНИЯ К ИНТЕГРАЦИЯМ

Тип интеграции	Необходимое условие
HTTP/HTTPS API	Маршрутизация и TLS до разрешённого узла; DNS; доверие к сертификату.
PostgreSQL	Сетевая доступность хоста/порта; отдельная учётная запись только чтения; TLS по политике.
Jira/Confluence/GitHub/GitLab	Доступ к корпоративному/внешнему API и пользовательские учётные данные.
S3/WebDAV/SFTP	Доступ к конечной точке; соответствующий порт; учётные данные пользователя.
Почта	Доступ к IMAP/SMTP либо согласованному протоколу.
Яндекс 360 / Яндекс Облако	Исходящий HTTPS до разрешённых сервисов; настроенные учётные данные.
Telegram	Исходящий HTTPS до Telegram API; только если сервис разрешён политикой организации.

25 ТРЕБОВАНИЯ К АВТОМАТИЗАЦИЯМ

- постоянно доступные внутренние базы/очереди автоматизаций;
- синхронизированное время для расписаний;
- достаточный CPU/RAM при большом числе одновременно исполняемых процессов;
- сетевая доступность разрешённых интеграций;
- учётные данные владельца автоматизации и соответствующие разрешения;
- для большого числа задач - отдельное нагрузочное испытание по реальному расписанию и длительности действий.

26 ТРЕБОВАНИЯ К РЕЗЕРВНОМУ КОПИРОВАНИЮ

Штатные резервные копии должны храниться не только на том же диске, что и рабочий экземпляр. Минимум одна актуальная копия и ключ восстановления должны быть вынесены за пределы единственного сервера/диска.

Объект	Требование
Локальный каталог backups/	может использоваться как промежуточное хранилище, но не единственная копия
Внешнее хранилище	обязательно для промышленной эксплуатации
Ключ восстановления	хранить отдельно от архива и сервера
Свободное место	должно позволять создать новую копию до удаления старых согласно политике
Проверка восстановления	выполнять периодически на отдельной среде

27 ТРЕБОВАНИЯ К МОНИТОРИНГУ И ЖУРНАЛИРОВАНИЮ

- свободное место для журналов и временных метрик;
- синхронное время всех компонентов;
- администратор должен иметь доступ к системной странице и журналам;
- для внешнего SIEM/журнального хранилища планируется отдельный канал и политика хранения;
- не следует хранить чувствительные секреты в журналах;
- порог свободного диска должен контролироваться до достижения критического уровня.

28 ТРЕБОВАНИЯ ЗАКРЫТОГО КОНТУРА

- внутренний DNS и NTP;
- внутренний реестр контейнерных образов либо контролируемый шлюз;
- внутреннее хранилище моделей при локальном выводе;
- внутренний центр сертификации либо распределённое доверие;
- маршруты только к разрешённым внутренним интеграциям;
- обновления доставляются через утверждённый внутренний процесс;
- внешние функции должны быть отключены либо явно разрешены.

29 ТРЕБОВАНИЯ ПОЛНОСТЬЮ ИЗОЛИРОВАННОГО КОНТУРА

В полностью изолированном контуре установка и эксплуатация не должны зависеть от Интернет-доступа. Поставочный носитель должен содержать все образы, модельные артефакты, контрольные суммы и документацию выбранного профиля.

- заранее подготовленный полный набор контейнерных образов;
- локальная модель и её зависимости, если требуется ИИ без внешнего сервиса;
- внутренние DNS/NTP;
- локальная инфраструктура сертификатов;
- внешнее офлайн-хранилище резервных копий;
- процесс переноса обновлений с проверкой подписи и контрольных сумм;
- запрет незаявленных сетевых зависимостей.

30 ТРЕБОВАНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ СРЕДЫ

- запрет удалённого входа root по SSH;
- администрирование ОС через отдельную учётную запись и sudo;
- межсетевой экран с разрешением только необходимых входящих соединений;
- TLS для пользовательского доступа;
- хранение секретов и ключей с ограниченными правами;
- регулярное резервное копирование;
- контроль обновлений и подписей;
- разделение пользовательских и административных ролей;
- при использовании внешней модели - явное разрешение передачи данных за пределы контура.

ПРИМЕЧАНИЕ. Настоящий документ описывает требования среды и не является сертификатом соответствия ФСТЭК России или ФСБ России. Необходимость сертифицированных средств защиты определяется категорией информационной системы заказчика.

31 ТРЕБОВАНИЯ К МАСШТАБИРОВАНИЮ

Вертикальное масштабирование выполняется увеличением CPU, ОЗУ, дисковой производительности и GPU. Горизонтальное масштабирование отдельных контуров и отказоустойчивая схема требуют проектного решения, поскольку состояние пользователей, данных, очередей, хранилищ и модельного маршрутизатора должно быть согласовано.

- для увеличения числа пользователей сначала оцениваются CPU/RAM и база данных;
- для локальной модели отдельно оцениваются VRAM, пропускная способность GPU и длина очереди;
- для RAG оцениваются скорость индексации и поиска;
- для автоматизаций оцениваются число параллельных запусков и длительность внешних действий;
- для нескольких узлов требуется проект балансировки, общего/реплицируемого состояния и резервного копирования.

32 ПЛАНИРОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ

Число зарегистрированных пользователей само по себе не определяет требуемый сервер. Ключевые показатели - одновременно активные пользователи, средняя длина контекста, частота загрузки документов, интенсивность интеграций и число параллельных генераций.

Планировочный класс	Одновременно активные пользователи*	Базовый профиль
Небольшой	до 25	S1/S2
Средний	25-100	S2
Крупный	100-300	S3
Более 300	проектный расчёт	нагрузочное испытание и масштабирование

* Значения являются ориентиром планирования веб- и служебной части, а не гарантией производительности модели. При локальном выводе максимальная параллельность определяется отдельно по модели и GPU.

33 ОГРАНИЧЕНИЯ И НЕПОДДЕРЖИВАЕМЫЕ КОНФИГУРАЦИИ

- 32-разрядные операционные системы;
- производственный локальный vLLM на GPU с вычислительной совместимостью ниже 7.5 в базовом профиле;
- хранение единственной резервной копии на том же рабочем диске;
- публикация внутренних баз данных и служебных портов в недоверенную сеть;
- использование временного/эфемерного диска для постоянных данных;
- неустойчивое системное время;
- развёртывание без достаточного свободного места для обновления и резервного копирования;
- замена штатного манифеста/образов произвольными неподтверждёнными версиями;
- использование swap как способа компенсировать постоянный дефицит ОЗУ.

34 ПОРЯДОК ПОДТВЕРЖДЕНИЯ СООТВЕТСТВИЯ

1. идентифицировать ОС, ядро, Docker, архитектуру и аппаратные ресурсы;
2. выполнить предварительную проверку ./nicesoft.sh check;
3. для GPU проверить nvidia-smi и доступ из контейнера;
4. проверить свободное дисковое место и производительность хранилища;
5. выполнить установку и ./nicesoft.sh health;
6. выполнить тестовый вход, запрос модели, RAG и необходимые интеграции;
7. создать и проверить резервную копию;
8. зафиксировать результаты в протоколе квалификации;
9. при подтверждении конкретной платформы обновить матрицу совместимости.

35 ТРЕБОВАНИЯ К ИСПЫТАТЕЛЬНОЙ СРЕДЕ

Испытательная среда должна воспроизводить значимые свойства промышленной: тип ОС, контейнерную среду, тип хранилища, сетевые ограничения, TLS, модель и GPU. Результаты испытаний на существенно меньшем стенде не должны автоматически переноситься на высоконагруженный промышленный профиль.

Испытание	Минимальный результат
Установка	завершается без критических ошибок
Проверка служб	health успешен
Вход пользователя	успешный сеанс через HTTPS
Модель	тестовый запрос успешен
RAG	документ загружается и используется в ответе
Интеграции	минимум одна выбранная для проекта интеграция проходит штатный тест
Резервирование	создание и проверка архива успешны
Восстановление	периодически подтверждается на отдельном стенде
Нагрузка	для производственных требований выполняется профильный тест

36 НОРМАТИВНЫЕ И СПРАВОЧНЫЕ СВЕДЕНИЯ

Источник	Назначение
ГОСТ 19.101-2024	Виды программ и программных документов.
ГОСТ 19.105-78	Общие требования к программным документам.
Официальный сайт ИИ Студии	https://studio.ncsgp.ru
Официальный сайт НАЙС.ОС	https://niceos.ru
Документация НАЙС.ОС	https://niceos.ru/docs
Реестр российского ПО: НАЙС.ОС CLOUD	https://reestr.digital.gov.ru/reestr/4119206/
Реестровая запись НАЙС.ОС CLOUD	№30128 от 22.10.2025

При изменении модельного каталога, контейнерной среды или требований конкретной поставки приоритет имеют подписанный манифест версии 1.0.0 и официальное извещение правообладателя, относящееся к этой поставке.

ПРИЛОЖЕНИЕ А. СВОДНАЯ ТАБЛИЦА СИСТЕМНЫХ ТРЕБОВАНИЙ

Область	Минимум	Рекомендуется
Архитектура	x86_64	x86_64
CPU	8 vCPU	16 vCPU; 24-32 для расширенного профиля
ОЗУ	16 ГБ	32 ГБ; 64 ГБ+ для расширенного профиля
Диск	120 ГБ SSD	250 ГБ SSD/NVMe; 500 ГБ+ для данных/моделей
Сеть	1 Гбит/с	1-10 Гбит/с по профилю
GPU	не нужен при внешней модели	NVIDIA CC 7.5+ для локального производственного vLLM
ОС	российская серверная Linux ОС, соответствующая требованиям	НАЙС.ОС CLOUD
Контейнеры	Docker Engine + Compose v2	штатная версия ОС/поставки
TLS	TLS 1.2+	TLS 1.3 при поддержке среды
Браузер	современный	актуальный Яндекс Браузер/Chromium/Firefox ESR
Резервирование	внешняя копия	политика с ротацией и проверкой восстановления

ПРИЛОЖЕНИЕ Б. ПРОФИЛИ СЕРВЕРОВ

Профиль	vCPU	ОЗУ	Диск	GPU	Применение
S1	8	16 ГБ	120 ГБ	нет	Внешняя/ управляемая модель, небольшая нагрузка
S2	16	32 ГБ	250 ГБ	нет/по модели	Типовая корпоративная установка
S3	24-32	64 ГБ	500 ГБ+	по модели	Большие базы знаний и высокая активность
G1	8-16	32 ГБ	200 ГБ+	16 ГБ VRAM	Локальная модель ~4B
G2	16	64 ГБ	250 ГБ+	24-32 ГБ VRAM	Локальная модель ~9B
G3	24-32	96-128 ГБ	300-400 ГБ+	64-80 ГБ VRAM	Локальная модель 27-30B
G4	48+	256 ГБ+	500 ГБ+	2×80 ГБ+	Многопроцессорный проектный профиль

ПРИЛОЖЕНИЕ В. ТРЕБОВАНИЯ ЛОКАЛЬНЫХ МОДЕЛЕЙ

Модель	Загрузка, ГБ	VRAM мин./рек.	RAM модели мин.	Контекст рек.	CC мин.
Qwen3.5 4B	9.4	14 / 16	16	16384	7.5
Qwen3.5 9B	19.4	23 / 32	32	8192	7.5
Qwen3.8 27B	55.7	72 / 80	96	65536	7.5
Qwen3.8 27B · FP8	30.9	48 / 64	64	131072	8.9
Qwen3 Coder 30B-A3B	61.2	76 / 80	96	32768	7.5

Данные таблицы получены из модельного каталога поставки, однако итоговый выбор выполняется мастером моделей по фактически обнаруженному оборудованию. Размеры и требования других моделей определяются их карточками.

ПРИЛОЖЕНИЕ Г. ФОРМУЛА РАСЧЁТА ХРАНИЛИЩА

Для предварительного планирования можно использовать следующую формулу:

```
D_итого = D_ОС + D_службы + D_данные + D_индексы + Σ(D_модель × 1,15 + 5 ГБ) +
D_свободный_резерв
```

Для базового узла рекомендуется принимать D_ОС + D_службы + D_свободный_резерв не менее 120 ГБ. Резервные копии в формулу рабочего диска не включаются, если для них предусмотрено отдельное внешнее хранилище.

Показатель	Как оценить
D_данные	суммарный объём загружаемых документов, пользовательских файлов, истории и служебных данных с коэффициентом роста
D_индексы	определяется числом фрагментов документов и параметрами векторизации
D_модель	размер конкретного артефакта из каталога
D_свободный_резерв	не менее 20-30 ГБ, а для крупных обновлений и моделей - больше

ПРИЛОЖЕНИЕ Д. МАТРИЦА СЕТЕВЫХ ТРЕБОВАНИЙ

Источник	Назначение	Протокол/порт	Нужен всегда
Рабочие станции	Веб-доступ	HTTPS 8143 или 443	да
Администратор	SSH	TCP 22	по политике
Сервер → DNS	Разрешение имён	53/UDP,TCP	да
Сервер → NTP	Синхронизация	123/UDP	да
Сервер → реестр/обновления	Получение поставки	443/TCP	нет в изолированном контуре
Сервер → модельный поставщик	Внешний ИИ	443/TCP	только при внешней модели
Сервер → интеграции	Jira/почта/S3 и др.	по сервису	только для включённых интеграций

ПРИЛОЖЕНИЕ Е. КОНТРОЛЬНЫЙ ЛИСТ ПОДГОТОВКИ ИНФРАСТРУКТУРЫ

☐ Выбрана российская серверная ОС и проверена её версия.

☐ Для референсного профиля используется НАЙС.ОС CLOUD.

☐ Архитектура x86_64.

☐ Выделено не менее 8 vCPU / 16 ГБ ОЗУ для базового профиля.

☐ Подготовлено не менее 120 ГБ SSD.

☐ Определён прогноз роста данных.

☐ DNS и NTP работают.

☐ Назначено стабильное имя/адрес экземпляра.

☐ Подготовлен TLS-сертификат или план его выпуска.

☐ Docker Engine и Compose доступны.

☐ Открыты только необходимые порты.

☐ Определён режим модели: локальная / внешняя / управляемая.

☐ Для локальной модели рассчитаны VRAM/RAM/диск.

☐ Подготовлено внешнее хранилище резервных копий.

☐ Для закрытого контура подготовлены внутренние источники образов/моделей.

ПРИЛОЖЕНИЕ Ж. КОНТРОЛЬНЫЙ ЛИСТ GPU

☐ GPU обнаруживается командой nvidia-smi.

☐ Вычислительная совместимость соответствует модели и среде выполнения.

☐ Объём VRAM не ниже минимального требования артефакта.

☐ Есть эксплуатационный резерв VRAM для кэша и параллельности.

☐ Драйвер совместим с контейнерной средой поставки.

☐ Docker видит GPU.

☐ Для нескольких GPU подтверждена топология и Tensor Parallel.

☐ Температура/питание/охлаждение соответствуют требованиям сервера.

☐ После перезагрузки GPU остаётся доступен контейнерной среде.

☐ Пробный запуск модели проходит проверку готовности.

ПРИЛОЖЕНИЕ И. ТРЕБОВАНИЯ ДЛЯ РОССИЙСКИХ ОС

Проверка	Критерий
Архитектура	x86_64
Linux/ядро	достаточно современное для поддерживаемого Docker Engine и cgroups
Docker	Engine запускается штатно; Compose v2 доступен
Сеть	bridge/NAT/DNS контейнеров работают согласно политике ОС
MAC/SELinux	политика не блокирует штатные каталоги/сокеты/сети либо настроены необходимые разрешения
Хранилище	overlay2 или иной поддерживаемый драйвер; постоянные каталоги доступны
Время	NTP/chrony/systemd-timesyncd или эквивалент настроен
TLS	системное хранилище доверия может содержать корпоративный CA
GPU	драйвер и NVIDIA container runtime доступны при локальной модели
Результат	./nicesoft.sh check и health проходят; функциональные проверки успешны

При выполнении этих требований ОС соответствует заявленной модели совместимости. Статус конкретной версии «подтверждено» присваивается после документированного испытания.

ПРИЛОЖЕНИЕ К. ПАСПОРТ РАСЧЁТА РЕСУРСОВ

Параметр	Значение проекта
Организация	
Площадка/контур	
ОС и версия	
Архитектура	x86_64
Количество зарегистрированных пользователей	
Оценка одновременно активных пользователей	
Режим модели	локальная / внешняя / управляемая
Модель/поставщик	
Максимальный контекст	
Параллельность генерации	
vCPU	
ОЗУ	
GPU / количество	
VRAM суммарно	
Рабочий диск	
Ожидаемый объём корпоративных документов	
Резерв свободного места	
Внешнее хранилище резервных копий	
Пропускная способность сети	
Требуемые внешние интеграции	
Результат нагрузочного испытания	
Заключение	

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

Изм.	Номера листов	Основание	Дата	Подпись