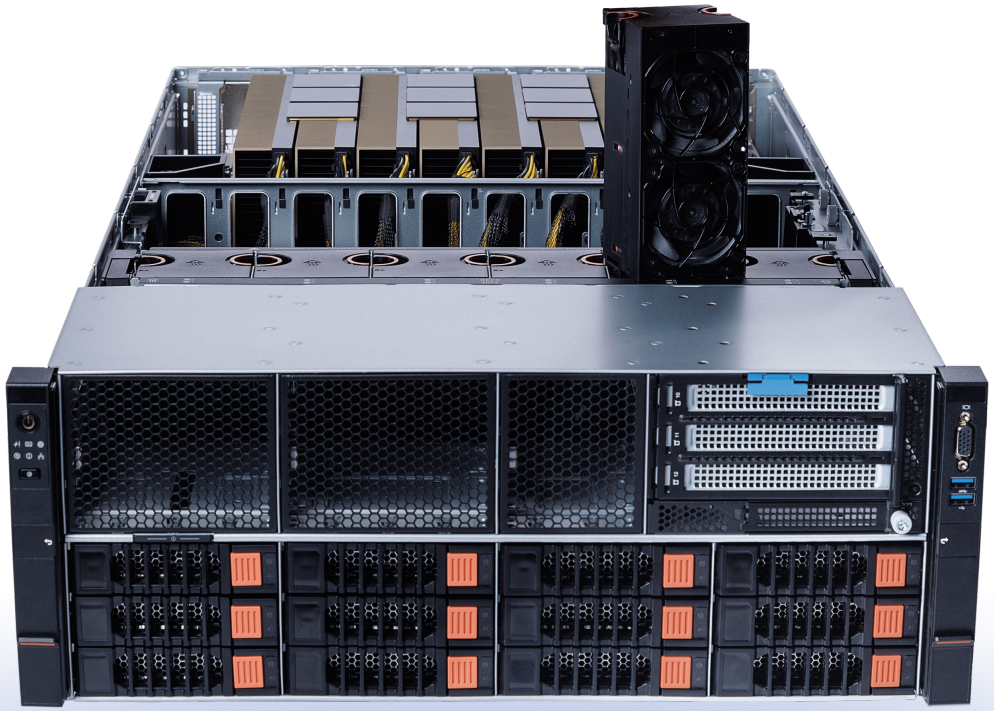
YADRO G4208P G3

Сервер для корпоративных ИИ-нагрузок, машинного обучения и глубокой аналитики



Основные характеристики сервера

- 8 GPU двойной ширины до 600 Вт[1]
- 4 БП по 3200 Вт, 2+2 или 3+1
- 12 двухроторных вентиляторов N+2



- 12 накопителей SAS/SATA/NVMe
- До 8 ТБ DDR5-5600 ECC
- Intel Xeon Scalable 4-го и 5-го поколений

[1] GPU по 600 Вт — при температуре окружающей среды до 30°C

CPU и память

2× Intel Xeon Scalable
4/5 поколения TDP до 350 Вт

32 DIMM слота
2S × 8 каналов × 2DPC

8 ТБ DDR5-5600 ECC
RDIMM/LRDIMM коррекция ошибок



Поддерживаемые ускорители

Флагман под LLM

H200 NVL

Большие модели, длинный контекст и высокая параллельность

Максимальные требования к питанию, охлаждению и инфраструктуре

Новое поколение

RTX PRO 6000 Blackwell

Универсальный ускоритель нового поколения для инференса, дообучения и смешанных нагрузок

Универсальный инференс

L40S | A30 | MetaX | C500

Рабочие конфигурации для инференса моделей разных классов и смешанных ИИ-нагрузок

Выбор ускорителя зависит от модели, программного стека и требований к производительности

Графика, 3D-визуализация

RTX 6000 Ada

RTX 4000 Ada

3D-визуализация, рендеринг, инженерная графика и профессиональные виртуальные рабочие места

Прикладные нагрузки и VDI

L4

A16

Видеоаналитика, компьютерное зрение, обработка видео и VDI — для сценариев, где не требуется максимальный класс GPU

Продолжаем валидацию ускорителей альтернативных производителей и расширяем пул конфигураций

Поддержка NVLink Bridge

Технология обеспечивает высокоскоростной прямой обмен между GPU и снижает зависимость обмена между ускорителями от PCIe

Где особенно важно

Крупные LLM, HPC-задачи и аналитика, где модель или данные распределяются между несколькими GPU

Что меняется

Обмен между GPU проходит напрямую, поэтому меньше зависит от узких мест системной шины

Результат

Более предсказуемая работа распределенных GPU-нагрузок и основа для масштабирования внутри сервера



Локальное хранение

12 отсеков LFF/SFF

8 × SAS/SATA + 4 × SAS/SATA/NVMe

Суперконденсатор

Защита кэша при отключении питания

RAID 0/1/10/5/50/6/60

RAID-контроллер с памятью до 8 ГБ

Горячая замена

Ускоряет восстановление при обслуживании



Возможности расширения

8 × PCIe 5.0 x16 FHFL

Задние слоты для GPU двойной ширины

1 × PCIe 5.0 x16 FHFL

Фронтальный слот, активно x8

2 × PCIe 5.0 x16 FHFL

Активно x8 в каждом разъеме для карт расширения

1 × OCP 3.0 PCIe 5.0 x8

Современные сетевые адаптеры и интеграция в ЦОД



Эксплуатация

Охлаждение

- 12 двухроторных вентиляторов 60 × 56 мм
- Резервирование N+2 и горячая замена

4 дополнительных вентилятора 80 × 80 мм — опциональная система охлаждения

01

Питание

- CRPS Platinum до 3200 Вт
- Резервирование 2+2 или 3+1

В зависимости от конфигурации

02

Климат

- 5–40°C при влажности 5–90%
- Отдельные конфигурации имеют более строгие ограничения

03

Габариты

447 × 844,32 × 174,8 мм
или 928,1 мм

С дополнительным охлаждением

04

Масса

До 65 кг, требуется учитывать монтаж и сервисный доступ

05

Обслуживание

Горячая замена дисков, вентиляторов и блоков питания

06

Управление и мониторинг

OpenBMC

BMC на основе OpenBMC
с корпоративным
функционалом YADRO

Redfish API

REST— интерфейс для
автоматизации, инвентаризации
и оркестрации оборудования

IPMI 2.0 / SNMP

Совместимость с классическими
инструментами управления,
мониторинга и алертинга

WebUI | CLI

Администрирование через
веб-интерфейс и командную
строку для инженерных команд

YADRO СУПРИМ

Централизованный контроль
парка оборудования YADRO

Серверы СХД Коммутаторы

Satellites

Инструменты диагностики
и сервисных операций
в инфраструктуре YADRO

Совместимость с корпоративным ПО

ОС	Виртуализация	Управление
Windows Server	РЕД-Виртуализация	OpenBMC
Ubuntu	zVirt	Redfish
РЕД ОС	moncloud	IPMI
Альт		SNMP
Astra Linux		YADRO СУПРИМ
SberLinux		VEGMAN SATELLITES

Сценарии применения

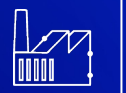
Отраслевое применение

Сервер YADRO G4208P G3 адресует отрасли, где ИИ переходит в продуктивные сервисы и критичные процессы

Финансы
- Антифрод
- Скоринг
- RAG по регламентам
- Ассистенты операторов



Промышленность
- Контроль качества
- Предиктивное обслуживание
- Цифровые двойники



Медицина
- Анализ снимков
- Обработка исследований
- Поддержка диагностики



Госсектор
- Поиск по архивам
- Обработка обращений
- Локальные ИИ-сервисы



Ритейл
- Рекомендации
- Персонализация
- Прогноз спроса
- Аналитика очередей



Телеком
- Оптимизация сети
- Речевая аналитика
- Предиктивная поддержка



ИТ и ISV
- MLOps
- AI-продукты
- API-сервисы
- GPU-as-a-Service



Образование и наука
- Исследовательские вычисления
- Обучение моделей
- Массивы данных



Сценарии применения

01



Инференс и промышленная эксплуатация ИИ-моделей

- Корпоративные ассистенты
- ИИ-сервисы
- RAG
- Др. нагрузки, переходящие из пилота в продуктив

02



Инженерное моделирование и научные расчеты

- Цифровые двойники
- Математическое моделирование
- HPC
- Вычислительно интенсивные расчеты

03



Рендеринг и 3D-визуализация

- Профессиональная графика
- Визуализация проектов
- 3D-рендеринг
- Удаленные рабочие места специалистов

04



Другие сценарии применения

- Компьютерное зрение и видеоаналитика
- Аналитика
- Прогнозирование
- Обучение и дообучение ИИ-моделей

Инференс и промышленная эксплуатация ИИ-моделей

Задачи

- «Выводим ИИ-ассистента из пилота в продуктив»
- «Хотим перенести сервис из облака в свой контур»
- «Растет число одновременных запросов»
- «Увеличивается TTFT / время до первого токена»
- «Нужно масштабировать инференс»
- «Надо запустить LLM / RAG»

Наше предложение

Рассчитать конфигурацию под конкретный профиль инференса

- Модель
- Число одновременных запросов
- Длину контекста
- Требования к отклику и росту нагрузки

Дополнительные выгоды

- **Программа валидации GenAI-нагрузок**
  Проверяем работу конкретной модели на целевой конфигурации до промышленного внедрения
- **Гетерогенная GPU-матрица**
  Подбираем ускорители под профиль нагрузки и снижаем зависимость от доступности одного класса GPU



Инженерное моделирование и научные расчеты

Задачи

- «Моделирование занимает часы или дни»
- «Нужно сократить цикл инженерных расчетов»
- «Строим цифровой двойник»
- «Строим HPC-кластер»
- «Нужен кластер для инженерного или математического моделирования»

Наше предложение

- **Подобрать конфигурацию под профиль расчета**
  С учетом FP64/FP32 и программного стека
- **Рассмотреть GPU разных классов и производителей**
  Под требования конкретной нагрузки

Дополнительные выгоды

- **Поддержка GPU разных классов**
  Подбираем ускорители под требования конкретной нагрузки
- **Альтернативные GPU**
  Поддержка ускорителей для задач, где критична высокая точность вычислений FP64/FP32
- **Реестр Минпромторга**
  Дополнительный аргумент для НИИ, госсектора и регулируемых проектов



Рендеринг и 3D-визуализация

Задачи

- «Рендер занимает слишком много времени»
- «Работаем с тяжелыми 3D-сценами»
- «Инженерам и дизайнерам нужны мощные графические рабочие места»
- «Нужен удаленный рабочий стол для архитектора, инженера или дизайнера»

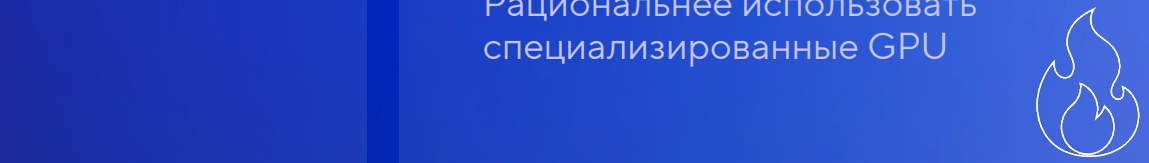
Наше предложение

Подобрать конфигурацию под графическую нагрузку и число виртуальных рабочих мест

С учетом требований к производительности и профиля пользователей

Дополнительные выгоды

- **Поддержка профессиональных графических GPU**
  RTX PRO 6000 Blackwell, RTX 6000 Ada и RTX 4000 Ada
- **RT-ядра для визуализации**
  Нужны для профессиональной графики. В HGX их нет
- **HGX избыточен для этого профиля нагрузки**
  Рациональнее использовать специализированные GPU



Разработка, обучение и дообучение ИИ-моделей

Задачи

- «Разрабатываем собственные модели»
- «Дообучаем модель на своих данных»
- «Эксперименты в облаке становятся дорогими»
- «Нужно сократить время обучения»
- «Нужно быстрее проверять гипотезы»

Наше предложение

- **Подобрать конфигурацию под обучение**
  С учетом размера модели, объема датасета, программного стека и целевого времени обучения
- **Заложить масштабирование от одного узла к кластеру**
  Если нагрузка будет расти

Дополнительные выгоды

- **Программа валидации GenAI-нагрузок**
  Проверяем работу конкретной модели на целевой конфигурации
- **Гетерогенная GPU-матрица**
  Подбираем ускорители под профиль нагрузки и снижаем зависимость от доступности HGX-платформ



Компьютерное зрение и видеоаналитика

Задачи

- «Десятки или сотни камер»
- «Нужно распознавание на объектах»
- «Автоматизируем контроль качества»
- «Нужно обрабатывать несколько видеопотоков одновременно»
- «Видео должно обрабатываться внутри собственного контура»

Наше предложение

- **Подобрать конфигурацию под профиль видеоаналитики**
  Число потоков, CV-модель, требования к скорости обработки и задержке
- **Подобрать класс GPU под масштаб нагрузки**
  Без избыточной конфигурации

Дополнительные выгоды

- **Поддержка GPU разных классов**
  Выбор ускорителя под объем потоков, модель и требования
- **Поддержка прикладных CV-нагрузок**
  Компьютерное зрение входит в пул исследуемых нагрузок платформы
- **Практический кейс в медиа**
  «Россия сегодня» использует G4208P G3 в том числе для поиска по фотобанку



Прогнозирование и аналитические модели

Задачи

- «Регулярно пересчитываем прогнозы»
- «Объем исторических данных растет»
- «Нужно быстрее проверять новые признаки и алгоритмы»
- «Расчеты занимают слишком много времени»
- «Нужно сократить время получения результата»

Наше предложение

- **Подобрать конфигурацию под аналитическую нагрузку**
  С учетом объема данных, модели, программного стека и требований ко времени расчета
- **Подобрать GPU под вычислительный профиль**
  В зависимости от характера алгоритмов и требуемой производительности

Дополнительные выгоды

- **Поддержка GPU разных классов**
  Ресурсы подбираются под конкретный вычислительный профиль
- **Высокая плотность вычислений**
  До 8 GPU и до 8 ТБ DDR5 для ресурсоемких аналитических задач



Подтвержденные результаты

GigaChat 3.1 Ultra на G4208P G3

GigaChat3-702B-A36B-preview

Что проверили

Проверяли на 1 и 2 серверах G4208P G3: профили ChatBot, Math & Logic и Summarization

Для масштаба: DeepSeek-R1 — 685B | Llama — 405B | Qwen3 — 235B

Подтвержден запуск модели

1 сервер

Стартовая точка для PoC, лабораторной проверки и сценариев с невысокой параллельностью

Подтвержден запас под рост нагрузки

2 сервера

throughput: 2,06–3,30× **TTFT: до –98,9%** **E2E: до –58,6%**

Измерено при высокой параллельной нагрузке

- **TTFT** — время от запроса до первого токена ответа
- **Throughput** — сколько выходных токенов система генерирует за секунду
- **E2E latency** — полное время от запроса до завершения ответа

Основные выводы

01 G4208P G3 проверен на инференсе крупной российской LLM ≈ на 700 млрд параметров

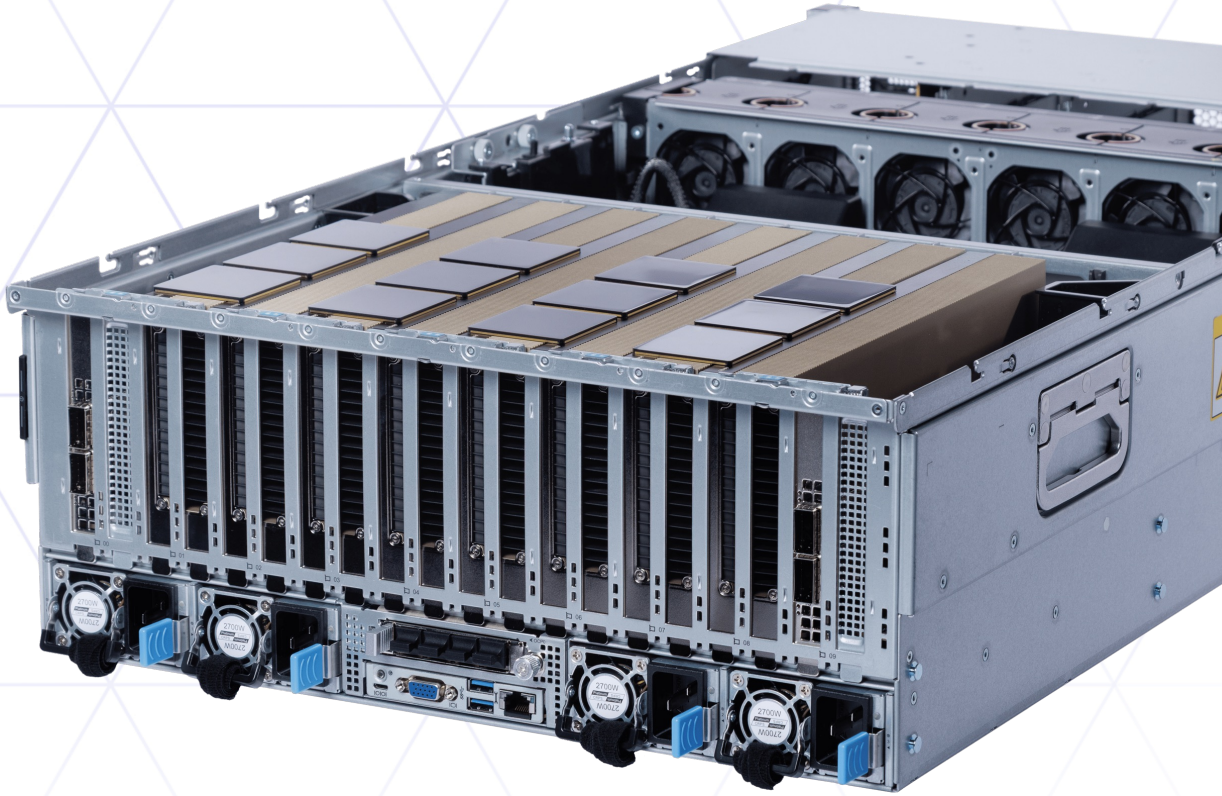
02 Начать можно с одного узла, а при росте числа пользователей масштабировать решение на несколько серверов

«Россия сегодня»

Локальная ИИ-инфраструктура на G4208P G3

Ссылка

Редакционные ИИ-сервисы работают внутри собственного ИТ-контура — с контролем данных и вычислительной инфраструктуры



Результаты проекта

- **Локальный контур**
  ИИ-инструменты используются без передачи чувствительных данных во внешние сервисы
- **Прикладные сценарии**
  Автоматическая транскрибация, обработка аудио и видео, субтитры и поиск по фотобанку более чем из 3,5 млн снимков
- **Промышленное внедрение**
  Работы начались в марте 2025 года, серверы введены в эксплуатацию в декабре 2025 года

Сервер валидирован Rubytech в составе Скала^р МИИ

Ссылка

Подтверждена работа G4208P G3 в составе комплексной ИИ-инфраструктуры с оптимизацией взаимодействия GPU, сети и СХД

Результаты валидации

- **Оптимизировано взаимодействие GPU и сети**
  Подобраны NUMA-компоновки GPU/NIC для снижения потерь производительности при межсерверном обмене
- **Подтверждена работа под высокой нагрузкой**
  Нагрузочные испытания охватили вычислительную, сетевую и дисковую подсистемы в составе единого решения
- **Подготовлена архитектура для масштабирования**
  Валидированные конфигурации рассчитаны на промышленное развертывание и дальнейшее наращивание ИИ-инфраструктуры



Для тех,
кто приобрел продукт

Состав POST-SALE программы поддержки клиентов

01

Сервис YADRO

02

СУПРИМ

03

Обновления и миграции

04

Расширение функциональности продуктов

05

TATLIN.SATELLITES

06

Расширенное обучение

Курсы по продукту для заказчиков

Рост технической компетентности специалистов

Минимизация операционных ошибок при работе с продуктами

Быстрое погружение в работу с продуктами YADRO

Практико-ориентированное обучение

Применение знаний в рабочих сценариях

Повышение эффективности работы команд

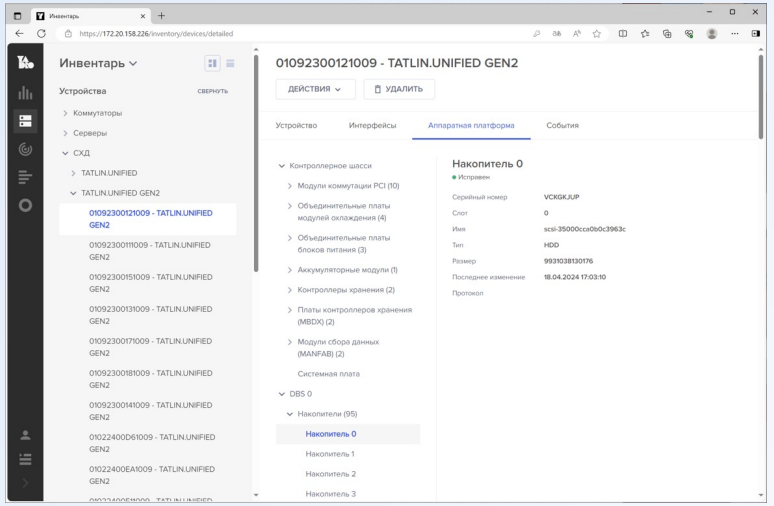
YADRO СУПРИМ

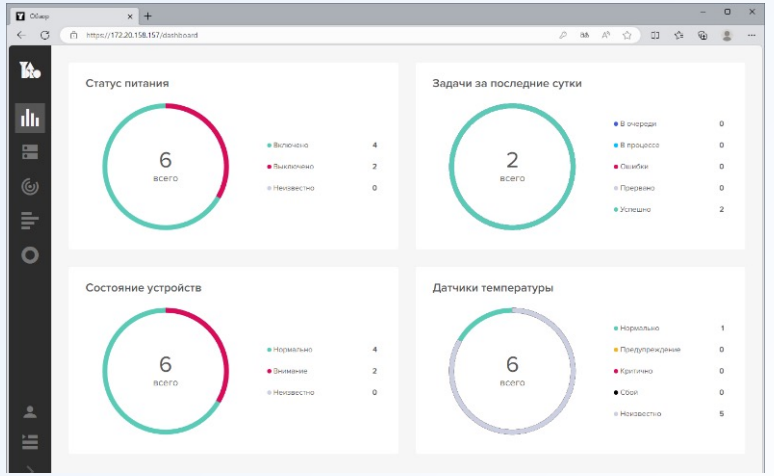
YADRO СУПРИМ

Программный пакет
для группового управления
серверами
YADRO в дата-центрах

Получить СУПРИМ

- Позволяет упростить процедуры ввода устройств в эксплуатацию, обслуживания и сбора информации о составляющих компонентах
- Доступен всем заказчикам вместе с серверами YADRO
- Распространяется в виде шаблона виртуальной машины





SATELLITES

Набор инструментов для бесшовной интеграции продуктов YADRO с существующими системами управления заказчика

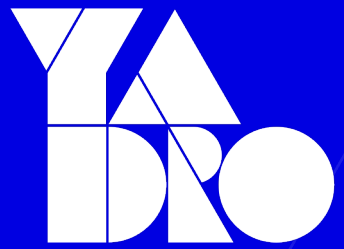
OpenStack, Zabbix, Grafana, Prometheus, K8S и другие готовые интеграционные инструменты

Собственный релизный цикл с частотой не менее раз в два месяца позволяет поддерживать максимальную актуальность

Доступен в виде комплимента для любого продукта YADRO



Что дальше?



3 шага к сотрудничеству

Шаг 1

Свяжитесь с нами

Получите консультацию по продукции компании

Шаг 2

Запросите демо

Запросите демонстрацию или обсудите проведение испытаний на доступном стенде

Шаг 3

Переходите на YADRO

- Подбор конфигурации
- Планирование пилотного или промышленного внедрения

+7 495 540 5055

sales@yadro.com

Что необходимо для квалификации запроса?

Какую задачу решаете?

- Инференс
- Обучение
- HPC
- Инженерное моделирование
- Компьютерное зрение
- Профессиональная графика
- VDI

01

Какой профиль нагрузки?

- Модель и ее размер
- Тип расчета
- Число видеопотоков
- Сложность 3D-сцен
- Другой ключевой параметр сценария

02

Какой масштаб и требования к производительности?

- Параллельные запросы
- Пользователи
- Время отклика или расчета
- Текущая и планируемая нагрузка

03

Есть ли ограничения у инфраструктуры?

- Температура входящего воздуха
- Доступная мощность на стойку
- Условия размещения

04

Есть ли у вас регуляторные требования?

- КИИ
- Реестр Минпромторга
- Состав решения

05

Какой программный стек вы используете?

- ОС
- Фреймворки
- Прикладное ПО
- Средства виртуализации
- Другие компоненты целевого контура

06

Спасибо!

Будем на связи

БУДУЩЕЕ
В НАШИХ РУКАХ

©2026 YADRO, все права защищены. YADRO®, VESNIN®, TATLIN®, VEGMAN® и KORNFELD® являются торговыми марками компании YADRO (или ее дочерних компаний), зарегистрированными на территории России и других стран.

Сведения, содержащиеся в данном документе, могут быть изменены без дополнительного извещения. Все гарантии, касающиеся товаров и услуг, реализуемых компанией YADRO, изложены в формулировках прямых гарантий, сопровождающих соответствующие товары и услуги. Никакая информация, приведенная в данном документе, не должна рассматриваться как дополнительная гарантия. Компания YADRO не несет ответственности за технические или редакторские ошибки либо пропуски в данном документе.

Компания YADRO придерживается высоких стандартов качества процессов разработки, производства и тестирования продуктов, однако в редких случаях это не исключает выявления дефектов в процессе эксплуатации. Мы продолжим совершенствовать свои процессы качества для предотвращения возникновения критических дефектов в дальнейшем.