

ТЕХНИЧЕСКАЯ ДОКУМЕНТАЦИЯ WARPA

Серверы для искусственного интеллекта (AI) и машинного обучения (ML)

Документ: Полное техническое описание и руководство по эксплуатации — AI-серия

Версия: 3.0 (AI Edition)

Дата: Март 2026

Продукты: WARPA Deep (основной), WARPA Render (для AI-визуализации), WARPA Code (для ML-разработки)

Производитель: WARPA (российский бренд) / Jieda (ODM-сборка, Китай)

Партнёр по GPU: NVIDIA (официальный партнёр по сборке)

Стандарты: ГОСТ Р 70627-2023, IPC-A-610 Class 2, CE, FCC Class A, RoHS,
REACH, IEC 62368-1, NVIDIA-Certified Server стандарты

ОГЛАВЛЕНИЕ (AI-серия)

1. Общая информация о AI-серверах WARPA
2. Продуктовая линейка AI: общая сводная таблица
3. Раздел А: Инференс и разработка (WARPA Code для ML)
 - 3.1 WARPA Code ML Edition
4. Раздел Б: GPU-рендеринг и AI-визуализация (WARPA Render)
 - 4.1 WARPA Render AI Edition
5. Раздел В: Тренировка нейросетей (WARPA Deep) – ОСНОВНОЙ
 - 5.1 WARPA Deep Entry (Inference, LLaMA 7-13B)
 - 5.2 WARPA Deep Main (тренировка средних моделей)
 - 5.3 WARPA Deep Flagship (Foundation Models, H100 SXM)
6. Общие технические стандарты для AI-серверов
7. Условия эксплуатации AI-серверов
8. Требования к электропитанию и тепловыделению (AI)
9. Совместимость с AI-фреймворками и ОС
10. Инструкция по начальной настройке (AI-сервер)
11. Руководство по обслуживанию GPU и NVLink
12. Схемы подключения и чертежи (AI-серверы)
13. Коды ошибок GPU и диагностика NVIDIA
14. Гарантия и техническая поддержка (AI-серия)
15. Перечень документов в комплекте поставки (AI)

1. ОБЩАЯ ИНФОРМАЦИЯ О AI-СЕРВЕРАХ WARPA

WARPA AI-серия — это специализированные серверы для задач искусственного интеллекта: от инференса небольших моделей до тренировки Foundation Models с использованием NVIDIA H100 SXM.

Ключевые особенности AI-серии:

- Официальная поддержка NVIDIA (NVLink, InfiniBand, SXM-платформа)
- Жидкостное охлаждение для флагманских моделей
- Прямые поставки GPU без посредников
- Полная совместимость с PyTorch, TensorFlow, JAX, NeMo
- Готовые конфигурации под LLaMA, GPT, Stable Diffusion
- Сертификация NVIDIA-Certified Server (для Deep Flagship)

2. ПРОДУКТОВАЯ ЛИНЕЙКА AI: ОБЩАЯ СВОДНАЯ ТАБЛИЦА

#	Продукт	Назначение	GPU (макс)	Форм-фактор	Целевая аудитория
1	WARPA Code ML	Разработка, инференс (малые модели)	1-2x NVIDIA RTX A4000 / T1000	1U / 2U	ML-инженеры, Data Science
2	WARPA Render AI	AI-визуализация, Stable Diffusion, рендеринг	4x NVIDIA RTX 6000 Ada	4U	Дизайн-студии, AI-арт

3	WARPA Deep Entry	Инференс LLaMA 7-13B, компьютер ное зрение	4x NVIDIA L40S (48GB)	4U	AI-стартапы, инференс
4	WARPA Deep Main	Тренировка средних моделей (A100)	8x NVIDIA A100 80GB PCIe	4U/8U	Исследовате льские центры
5	WARPA Deep Flagship	Foundation Models (H100 SXM)	8x NVIDIA H100 SXM5	SXM-п латфо рма	Крупные корпорации, ЦОД

РАЗДЕЛ А: ИНФЕРЕНС И РАЗРАБОТКА (WARPA CODE ML EDITION)

3.1 WARPA CODE ML

3.1.1 Техническое описание

Назначение: Разработка ML-моделей, инференс небольших нейросетей, компьютерное зрение (пограничные вычисления), Data Science.

Целевая аудитория: ML-инженеры, дата-сайентисты, стартапы на этапе прототипирования.

Ключевые особенности:

- Компактный форм-фактор (1U/2U)
- Оптимизация под PyTorch и TensorFlow
- Поддержка NVIDIA T1000 / A4000

3.1.2 Технические характеристики

Параметр	Entry	Main
Форм-фактор	1U	2U
Процессор	Intel Xeon E-2336 (6 ядер, 2.9 ГГц)	Intel Xeon Silver 4310 (12 ядер, 2.1 ГГц)
ОЗУ	32 ГБ DDR4 ECC	64 ГБ DDR4
Диски	2x 480 ГБ SATA SSD RAID1	2x 480 ГБ SATA (система) + 2x 1.92 ТБ NVMe (данные)
GPU	1x NVIDIA T400 / T1000 (4-8 ГБ)	1x NVIDIA RTX A4000 (16 ГБ)

Блок питания	500 Вт	800 Вт (резерв)
--------------	--------	-----------------

Сеть	2x 1GbE	2x 10GbE SFP+
------	---------	---------------

РАЗДЕЛ Б: GPU-РЕНДЕРИНГ И AI-ВИЗУАЛИЗАЦИЯ (WARPA RENDER AI EDITION)

4.1 WARPA RENDER AI

4.1.1 Техническое описание

Назначение: Генерация изображений (Stable Diffusion, DALL-E подобные), нейросетевой рендеринг, AI-арт, VFX с ML.

Целевая аудитория: Дизайн-студии, AI-художники, продакшн-команды.

4.1.2 Технические характеристики

Параметр	Entry	Main	Flagship
----------	-------	------	----------

Форм-фактор	4U (усиленный)	4U (GPU-каркас)	4U (макс. плотность)
-------------	----------------	-----------------	----------------------

Процессор	Intel Xeon W5-3435X (16 ядер)	Intel Xeon W7-3455 (24 ядра)	2x Xeon Gold 6430 (64 ядра)
ОЗУ	64 ГБ DDR5 ECC	128 ГБ DDR5 ECC	256 ГБ DDR5 ECC
GPU	2x NVIDIA RTX 4080 (16 ГБ)	4x NVIDIA RTX A5000 (24 ГБ)	4x NVIDIA RTX 6000 Ada (48 ГБ)
Диски	1 ТБ NVMe	2 ТБ NVMe + 4 ТБ HDD	4 ТБ NVMe
БП	1600 Вт Titanium	2200 Вт (2+2)	3000 Вт (3+1)

РАЗДЕЛ В: ТРЕНИРОВКА НЕЙРОСЕТЕЙ (WARPA DEEP) – ОСНОВНОЙ

5.1 WARPA DEEP ENTRY (Inference)

5.1.1 Техническое описание

Назначение: Инференс LLM (LLaMA 7B-13B), компьютерное зрение, обработка естественного языка, рекомендательные системы.

Типовые задачи:

- Запуск LLaMA 2/3 (7B, 13B) в продуктивной среде
- Обработка видео в реальном времени (CV)
- Тонирование BERT-подобных моделей

5.1.2 Технические характеристики

Параметр	Значение
Форм-фактор	4U, GPU-сервер (усиленный)
Процессоры	2x Intel Xeon Silver 4410Y (12 ядер, 2.0 ГГц, TDP 150 Вт)
ОЗУ	128 ГБ DDR5 ECC (8x16 ГБ), макс. 2 ТБ
GPU	4x NVIDIA L40S (48 ГБ VRAM, PCIe Gen5, 300 Вт)
Интерконнект	PCIe Gen5 x16 (каждый GPU)
Диски	2x 1.92 ТБ NVMe (RAID1 для системы и данных)
Сеть	2x 100GbE (Ethernet)

Блок питания	2000 Вт (2+2, Titanium)
--------------	-------------------------

Охлаждение	Воздушное усиленное (8x 80mm вентиляторов с PWM)
------------	--

Уровень шума	≤ 65 дБ (при 100% нагрузке)
--------------	-----------------------------

5.1.3 Производительность (ориентировочная)

Модель	Токенов/сек (инференс)
--------	------------------------

LLaMA 2 7B	~1800
------------	-------

LLaMA 2 13B	~950
-------------	------

Stable Diffusion (генерация)	12-15 изображений/сек
------------------------------	-----------------------

5.2 WARPA DEEP MAIN (тренировка средних моделей)

5.2.1 Техническое описание

Назначение: Тренировка средних моделей (LLaMA 30B-70B), дообучение (fine-tuning), компьютерное зрение высокого разрешения.

Типовые задачи:

- Тренировка LLaMA 2 70B (с оптимизациями)
- Fine-tuning Stable Diffusion
- Кластеризация A100 для научных расчётов

5.2.2 Технические характеристики

Параметр	Значение
Форм-фактор	4U / 8U (расширенный), поддержка SXM (опционально)
Процессоры	2x Intel Xeon Gold 6438Y+ (32 ядра, 2.0 ГГц, TDP 225 Вт)
ОЗУ	256 ГБ DDR5 ECC (16x16 ГБ), макс. 4 ТБ
GPU	8x NVIDIA A100 80 ГБ PCIe (или 8x A100 SXM4)
Интерконнект	NVLink Bridge (600 ГБ/с между парами GPU) + PCIe Gen4

Диски	4x 3.84 ТБ NVMe U.2 (RAID10)
-------	------------------------------

Сеть	2x 100GbE + опционально 4x InfiniBand HDR (200 Гбит/с)
------	--

Блок питания	4000 Вт (4+4, Titanium)
--------------	-------------------------

Охлаждение	Воздушное усиленное (с реверсивными вентиляторами)
------------	--

Уровень шума	≤ 72 дБ
--------------	---------

5.2.3 NVLink Bridge конфигурация

Пара GPU	Пропускная способность
----------	------------------------

GPU 0-1	600 ГБ/с
---------	----------

GPU 2-3	600 ГБ/с
---------	----------

GPU 4-5	600 ГБ/с
---------	----------

GPU 6-7

600 ГБ/с

5.3 WARPA DEEP FLAGSHIP (Foundation Models, H100 SXM)

5.3.1 Техническое описание

Назначение: Тренировка Foundation Models (GPT-4 подобные, LLaMA 3 400B+), крупные научные симуляции, HPC + AI.

Требования к среде:

- Обязательное жидкостное охлаждение
- Отдельное помещение с контролем пыли (ISO 8)
- Питание 3 фазы, 400В (или 2 фазы, 220В с перегрузкой)

5.3.2 Технические характеристики

Параметр

Значение

Форм-фактор

SXM-платформа (NVIDIA HGX-совместимая, 8U)

Процессоры

2x Intel Xeon Platinum 8480+ (56 ядер, 2.0 ГГц, TDP 350 Вт)

ОЗУ	512 ГБ DDR5 ECC (32x16 ГБ), макс. 8 ТБ
-----	--

GPU	8x NVIDIA H100 SXM5 (80 ГБ HBM3, 3.35 ТБ/с)
-----	---

Интерконнект GPU	NVLink Switch System (полная связность 900 ГБ/с)
------------------	--

Диски	8x 7.68 ТБ NVMe U.2 (RAID10)
-------	------------------------------

Сеть	8x InfiniBand NDR (400 Гбит/с каждый)
------	---------------------------------------

Блок питания	6000 Вт (6+2, жидкостное охлаждение)
--------------	--------------------------------------

Охлаждение	Жидкостное (прямой контакт с GPU, холодные пластины)
------------	--

Уровень шума	≤ 55 дБ (благодаря жидкостному охлаждению)
--------------	--

MTBF	150 000 часов
------	---------------

5.3.3 NVLink Switch топология

- Полная связность между всеми 8 GPU
- Пропускная способность: 900 ГБ/с (любой GPU → любой GPU)
- Задержка: < 200 нс

5.3.4 Требования к жидкостному охлаждению

Параметр	Значение
Тип охлаждения	Прямой контакт (cold plate)
Температура жидкости на входе	+20...+30°C
Расход жидкости	15 л/мин (на систему)
Тепловая мощность отвода	до 6 кВт
Соединения	QD (quick disconnect) – металлические, без разливов
Жидкость	Деионизированная вода + ингибиторы коррозии

6. ОБЩИЕ ТЕХНИЧЕСКИЕ СТАНДАРТЫ ДЛЯ AI-СЕРВЕРОВ

Стандарт	Описание
NVIDIA-Certified Server	Для Deep Flagship (совместимость с NVIDIA NGC)
ГОСТ Р 70627-2023	ЦОД. Инженерные системы.
IPC-A-610 Class 2	Качество сборки
CE / FCC Class A	ЭМС и безопасность
RoHS / REACH	Экологические стандарты
IEC 62368-1	Безопасность ИТ-оборудования
NVLink 2.0 / 3.0	Интерконнект GPU
InfiniBand NDR	Сеть высокой пропускной способности

7. УСЛОВИЯ ЭКСПЛУАТАЦИИ AI-СЕРВЕРОВ

Параметр	Рабочий диапазон	Хранение
Температура	от +15°C до +30°C (рекомендуется +22°C)	от -40°C до +70°C
Влажность	20% – 80% (без конденсата)	5% – 95%
Высота	до 2 000 м (для H100 – обязательно)	до 12 000 м
Запылённость	ISO 8 (класс 100 000) или чище	–
Вибрация	0.2 Grms	1.5 Grms

8. ТРЕБОВАНИЯ К ЭЛЕКТРОПИТАНИЮ И ТЕПЛОВЫДЕЛЕНИЮ (AI)

Продукт	Макс. мощность (Вт)	Тепловыделение (BTU/ч)	Напряже ние	Резерви рование
<hr/>				

Code ML	600	2 047	220В	Опц.
Render AI	2 500	8 530	220В	3+1
Deep Entry	2 000	6 824	220В	2+2
Deep Main	4 000	13 648	220В / 380В	4+4
Deep Flagship	6 000 (жидкость)	20 472	380В (3 фазы)	6+2

9. СОВМЕСТИМОСТЬ С AI-ФРЕЙМВОРКАМИ И ОС

9.1 AI-фреймворки (сертифицировано)

Фреймворк	Версии	Поддержка GPU
PyTorch	2.0, 2.1, 2.2	CUDA 11.8 / 12.1

TensorFlow	2.13, 2.14	CUDA 11.8
JAX	0.4.14+	CUDA 12.0
NVIDIA NeMo	1.23+	H100/A100
Megatron-LM	Последняя	H100 (NVLink Switch)
DeepSpeed	0.10+	H100/A100
Hugging Face Transformers	4.36+	Все GPU

9.2 Операционные системы (для AI)

ОС	Версии	GPU-драйвер
Ubuntu Server	20.04, 22.04 LTS	NVIDIA Driver 535+
Rocky Linux	8, 9	NVIDIA Driver 535+

Debian	11, 12	NVIDIA Driver 525+
--------	--------	--------------------

Astra Linux (для РФ)	Special Edition 1.8	NVIDIA Driver (адаптированный)
----------------------	---------------------	-----------------------------------

10. ИНСТРУКЦИЯ ПО НАЧАЛЬНОЙ НАСТРОЙКЕ (AI-СЕРВЕР)

10.1 Распаковка и установка (Deep Flagship)

1. Распакуйте сервер на усиленной тележке (вес >50 кг).
2. Установите в стойку (8U) с помощью направляющих для жидкостных систем.
3. Подключите жидкостное охлаждение к внешнему чиллеру (QD-разъёмы).
4. Подключите питание (3 фазы, 380В, не менее 30А).
5. Подключите InfiniBand кабели к топ-оф-рейк коммутаторам.

10.2 Установка NVIDIA драйверов и CUDA

```
bash
```

```
# Для Ubuntu 22.04
```

```
sudo apt update
```

```
sudo apt install nvidia-driver-535 nvidia-cuda-toolkit
```

```
# Проверка GPU
```

```
nvidia-smi
```

10.3 Запуск контейнера с PyTorch (NGC)

```
bash
```

```
docker pull nvcr.io/nvidia/pytorch:23.12-py3
```

```
docker run --gpus all -it nvcr.io/nvidia/pytorch:23.12-py3
```

11. РУКОВОДСТВО ПО ОБСЛУЖИВАНИЮ GPU И NVLINK

11.1 Замена GPU (PCIe модели)

1. Выключите сервер, отключите питание.
2. Снимите крышку корпуса.
3. Отсоедините питание GPU (8-pin / 12VHPWR).
4. Нажмите на защёлку PCIe слота, извлеките GPU.
5. Установите новый GPU до щелчка.
6. Подключите питание, закройте крышку.

11.2 Диагностика NVLink

```
bash
```

```
nvidia-smi nvlink --status
```

Ожидаемый вывод (Deep Main с NVLink Bridge):

```
text
```

```
GPU 0: NVLink version 2.0, link status: All links active
```

```
GPU 1: NVLink version 2.0, link status: All links active
```

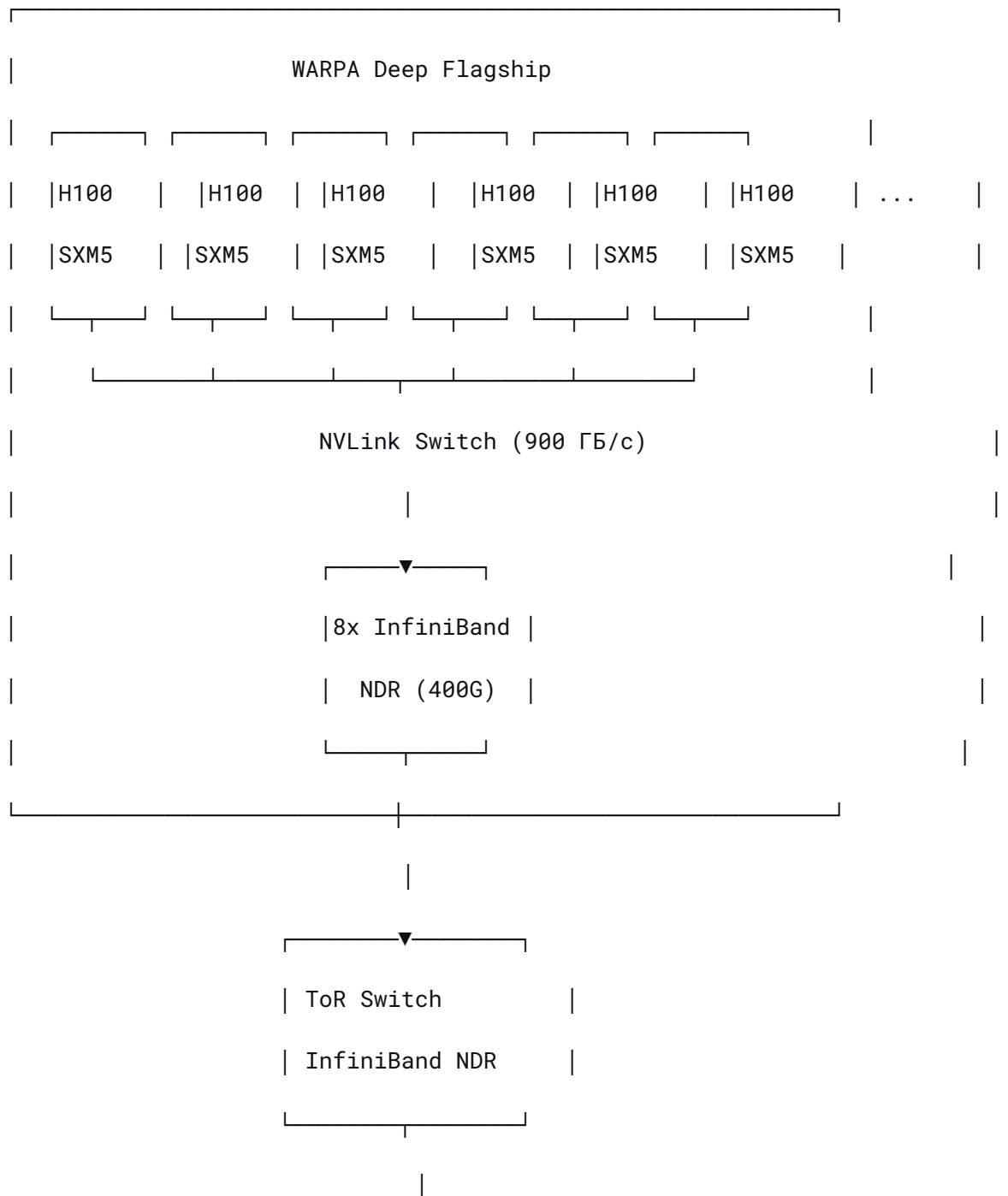
11.3 Замена вентиляторов (воздушные модели)

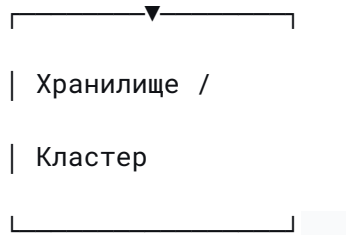
- Вентиляторы горячей замены только в Deep Main (специальные отсеки).
- В Deep Entry и Render AI – замена только при выключенном питании.

12. СХЕМЫ ПОДКЛЮЧЕНИЯ И ЧЕРТЕЖИ (AI-СЕРВЕРЫ)

12.1 Схема подключения Deep Flagship к InfiniBand

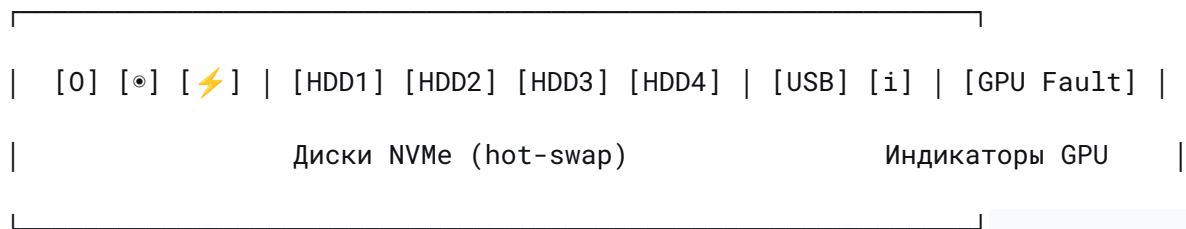
text





12.2 Вид спереди (Deer Main, 4U)

text



13. КОДЫ ОШИБОК GPU И ДИАГНОСТИКА NVIDIA

13.1 Типичные ошибки NVIDIA (nvidia-smi)

Ошибка	Причина	Решение
<code>GPU is lost</code>	Перегрев / сбой питания	Проверьте охлаждение, БП
<code>NVLink error</code>	Физическое повреждение кабеля	Замените NVLink Bridge

ECC error

Сбой памяти GPU

Замените GPU (по гарантии)

Driver not loaded

Конфликт драйверов

Переустановите драйвер

13.2 Световые индикаторы GPU (L40S / A100 / H100)

Цвет

Состояние

Зелёный

Нормальная работа

Мигающий зелёный

Активность вычислений

Жёлтый

Перегрев / снижение тактов

Красный

Критическая ошибка

14. ГАРАНТИЯ И ТЕХНИЧЕСКАЯ ПОДДЕРЖКА (AI-СЕРИЯ)

Тип гарантии

Срок

Особенности

Базовая (GPU + сервер)	36 месяцев	Замена GPU в течение 10 дней
------------------------	------------	------------------------------

Расширенная для H100	60 месяцев	Включая жидкостную систему
----------------------	------------	----------------------------

NVIDIA Enterprise Support	Опционально	Прямая поддержка NVIDIA
---------------------------	-------------	-------------------------

Контакты AI-поддержки:

- Телефон: 8 (961) 960-46-90
- Email: pavel@warpa.ru
- Портал: <https://warpa.ru>

15. ПЕРЕЧЕНЬ ДОКУМЕНТОВ В КОМПЛЕКТЕ ПОСТАВКИ (AI)

Документ	Формат
Quick Start Guide (AI Edition)	Бумажный
Гарантийный талон (с серийными номерами GPU)	Бумажный

NVIDIA NGC Ready сертификат

Бумажный

Паспорт жидкостной системы (для Flagship)

Бумажный

Полная техническая документация AI (PDF)

USB

Конец документации WARPA AI-серия