AI-сервер гетерогенный вычислительный

ПАЛАДИН-ML

12 сопроцессоров
2 x Xeon Scalable

24xDDR4
1+1 БП

Назначение и применение:

AI-сервер гетерогенный вычислительный «Паладин-ML» предназначен для:

- Задач выполнения (инференс) нейросетевых моделей;
- Высокопроизводительных векторно-матричных вычислений, требующих операций двойной точности (FP64).

Паладин-ML / Вид спереди

Паладин-ML / Вид сзади

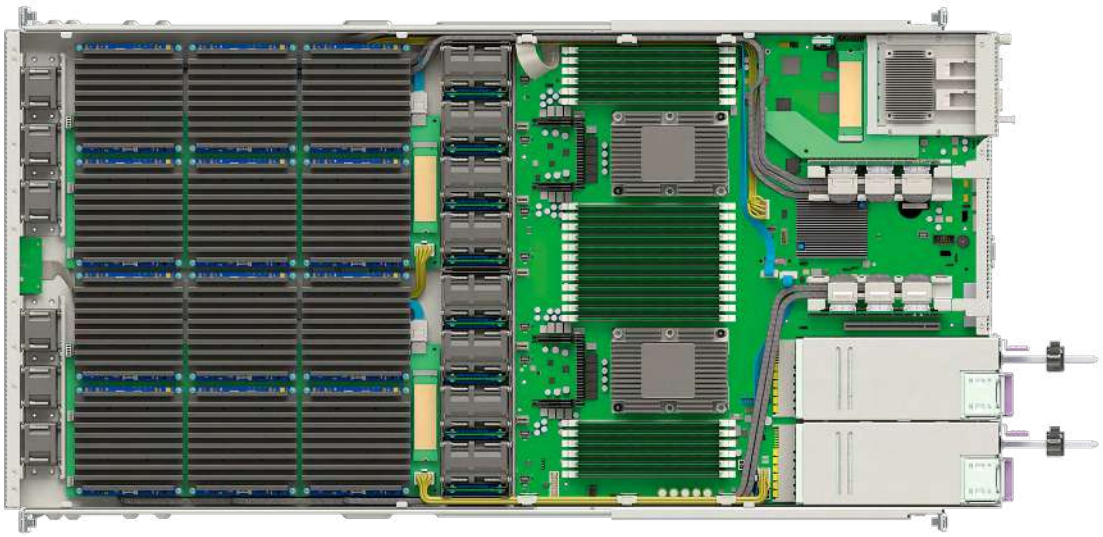
Платформа построена на базе материнской платы НИКА.469555.001 Паладин-X01.

Архитектура гетерогенного вычислителя:

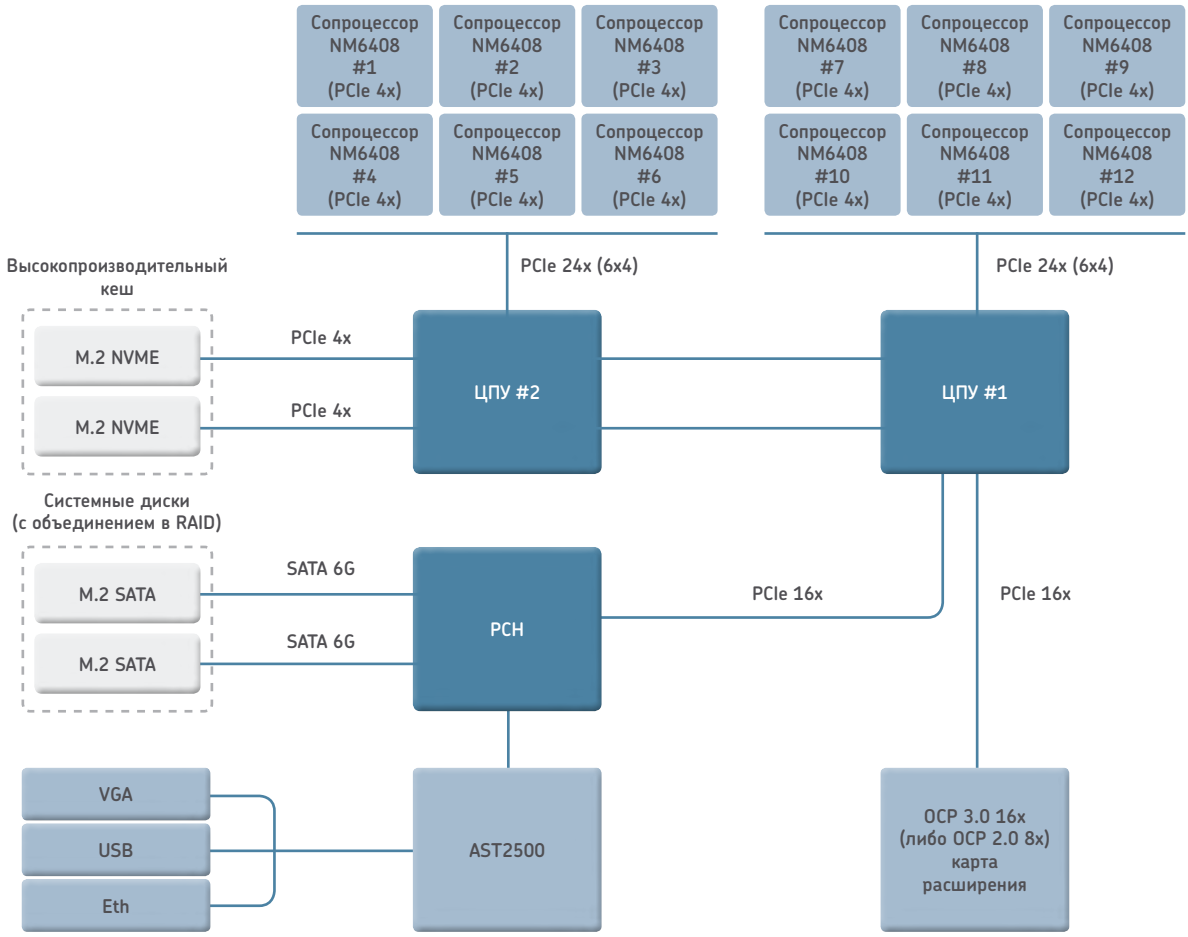
В качестве векторно-матричных сопроцессоров вычислителя используются процессоры NM6408 NeuroMatrix.

Суммарно вычислитель обеспечивает установку 12 шт. модулей сопроцессоров. Подключение каждого сопроцессора выполняется PCIe 4x линиями к основным процессорам Scalable-2. Доступная производительность на операциях FP64 за счет сопроцессоров – 1,53 ТФлоп.

Паладин-ML / Внутреннее устройство



Паладин-ML / Вид сверху



Операционная система устанавливается на M2. SATA SSD (2 шт, объединяемые через PCH в RAID). Для работы высокопроизводительного кеша промежуточного буферизирования данных для обработки и вычисленных результатов предусмотрены 2xNVME PCIe 4x диска.
Также доступна загрузка операционной системы через PXE, в таком случае системные M.2 SATA не устанавливаются (для загрузки ОС) либо в случае установки могут быть использованы как дополнительный дисковый кеш.

Внешние сетевые подключения

Вычислитель дает возможности установки следующих сетевых карт для подключения к внешней высокоскоростной сети (интерконнекту):

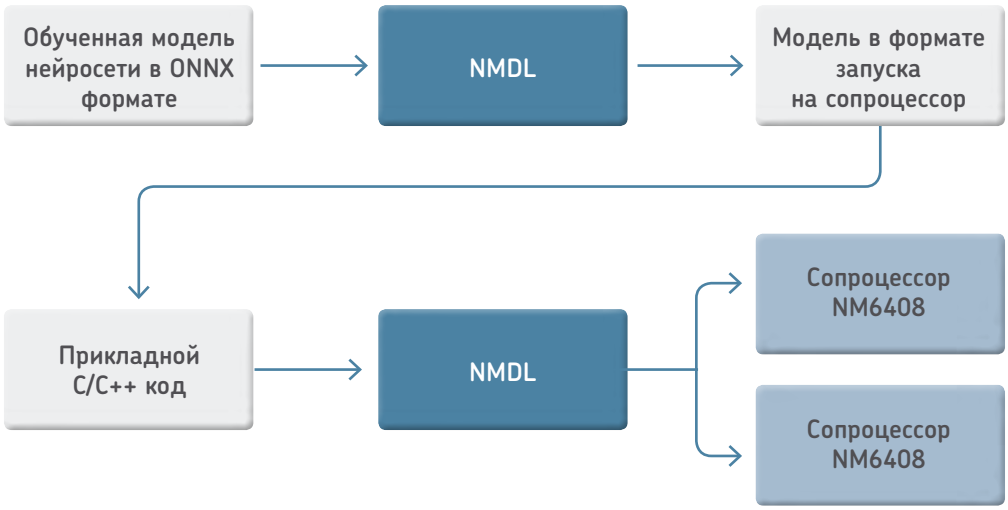
- OCP 3.0 (PCIe 16x): 2x10G, 4x10G, 2x25G, 2x40G, 2x100G
- OCP 2.0 (PCIe 8x): 2x10G, 2x25G

Характеристики

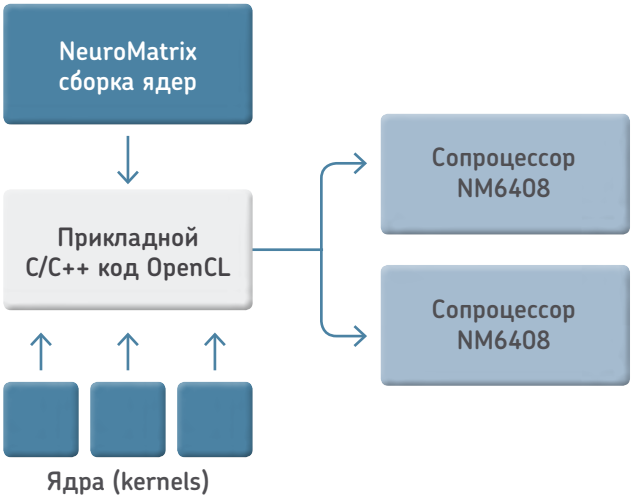
Вычислительные ресурсы	
Материнская плата	НИКА.469555.001 Паладин-X01
Процессор	Intel Xeon Scalable-2 с TDP до 185 Вт
ОЗУ	DDR4, до 24 шт.
Максимальный объем	3 Тб
Чипсет	Intel C624
Графический контроллер	Дискретный 2D на основе AST2500: макс. разрешение 1920 × 1200 @60Hz
Карты расширения	
OCP 3.0	1 шт. PCIe 16x либо 1 шт. PCIe 8x
Либо OCP 2.0	1
Дисковая подсистема	
M.2 SSD на материнской плате (системные диски)	2
M.2 PCIe 4x NVMe на объединительных платах (высокопроизводительный кеш)	2
Интегрированные интерфейсы	
1 Gbe Ethernet, портов	2 на тыльной панели
1Gbe BMC	1 на тыльной панели
USB 3.0	3 на тыльной панели
VGA	1 на задней панели
Электропитание и охлаждение	
Номинальная мощность, Вт	CRPS 1+1 БП, 1200 Вт
Напряжение	220/48В
Системные вентиляторы	8 шт. основных системных 6 шт. малооборотных вспомогательных
Габариты и масса	
Монтажный размер, U	1
Габариты, мм	43,5 x 438 x 917
Эксплуатационные параметры	Температура +5...35°C, Давление 630...800 мм рт ст

Интерфейсы взаимодействия с вычислительными ядрами

Запуск обученных моделей нейросетей выполняется с предварительным проеобразованием в ONNX-формат, который через библиотеку NMDL (NeuroMatrix® DeepLearning комплект программных средств для разработки и реализации глубоких нейронных сетей) преобразовывается в формат для запуска на сопроцессорах. Полученный формат использу-ется C/C++ кодом для запуска моделей на установленных в вычислитель сопроцессорах.



Второй способ взаимодействия доступен через стандартный интерфейс OpenCL.



Электропитание:

- 1+1 CRPS блоки питания мощностью по 1200 Вт включительно;
- Сервер комплектуется блоком питания 220В или 48В по желанию Заказчика.

Удобство обслуживания и эксплуатации:

- Установка в типовые 19’ шкафы (1000 мм);
- Направляющая для укладки кабелей, кабельные застёжки на блоках питания;
- Фирменная система удаленного управления оборудованием с мобильных устройств.

Полноценный монтажный комплект и фирменная транспортная упаковка «НТ»:

- Стандартные рельсы частичного выдвижения;
- Фирменная транспортная упаковка «НОРСИ-ТРАНС».

Возможности по построению специализированных вычислительных кластеров:

- Организация общего управления кластером, распределения задач, очередей и приоритетов обработки средствами Slurm;
- Интегрированный мониторинг оборудования и вычислительных задач.

ЗАО «НОРСИ-ТРАНС»
127015, Москва,
ул. Б. Новодмитровская, д. 12, стр. 15
+7 (495) 748-74-83
www.servers.norsi-trans.ru
servers@norsi-trans.ru