ЦИФРОВЫЕ
ИНДУСТРИАЛЬНЫЕ
ТЕХНОЛОГИИ

Решения для аппаратного ускорения работы нейронных сетей

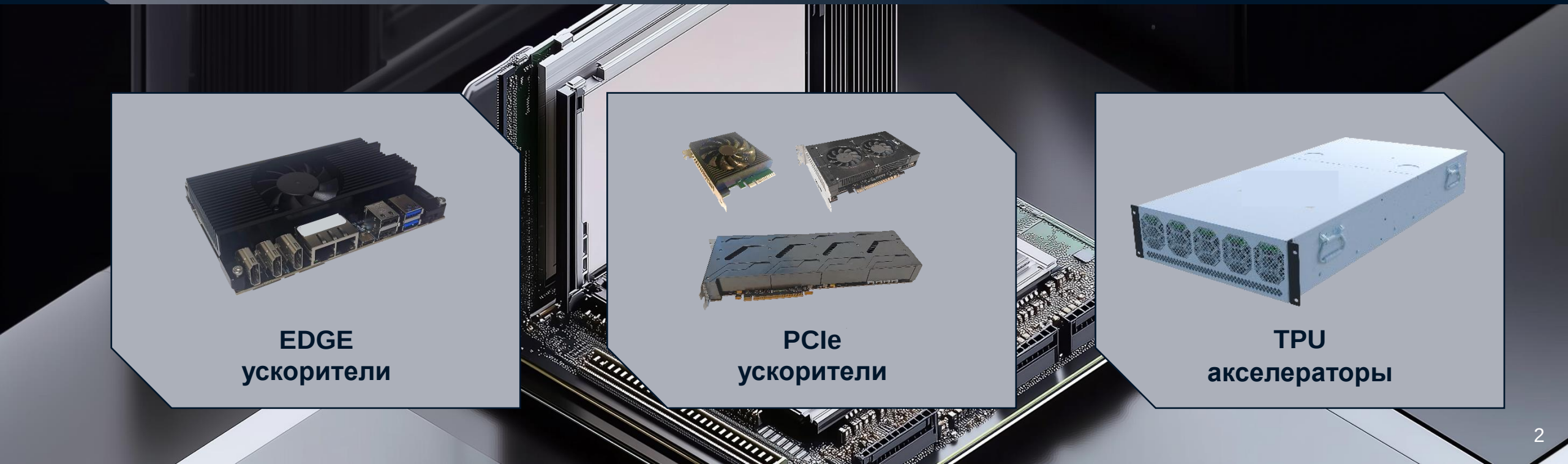
Каталог ИИ-решений

Отечественные вычислители для запуска нейронных сетей

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

Вычислители предназначены для запуска ИИ-приложений и применяются в области периферийных вычислений, туманных вычислений, а также в стандартной инфраструктуре центров обработки данных.

Созданы на базе отечественного процессора собственной архитектуры и полного набора необходимого программного обеспечения.



EDGE ускорители

PCIe ускорители

TPU акселераторы

Доверенный процессор

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ



ГИСП
В РЕЕСТРЕ РОССИЙСКОЙ ПРОДУКЦИИ

Тензорный процессор построен на базе сложнофункциональных блоков собственной разработки без применения иностранных IP-блоков

BGA1296
Корпус

750 МГц
Частота блока TPU

28 нм
Техпроцесс

Gen3, EndPoint, 8 линий
Контроллер PCIe

16 Мб
Встроенная память

2 канала 72bit до 32 Гб
Контроллер DDR4+ECC

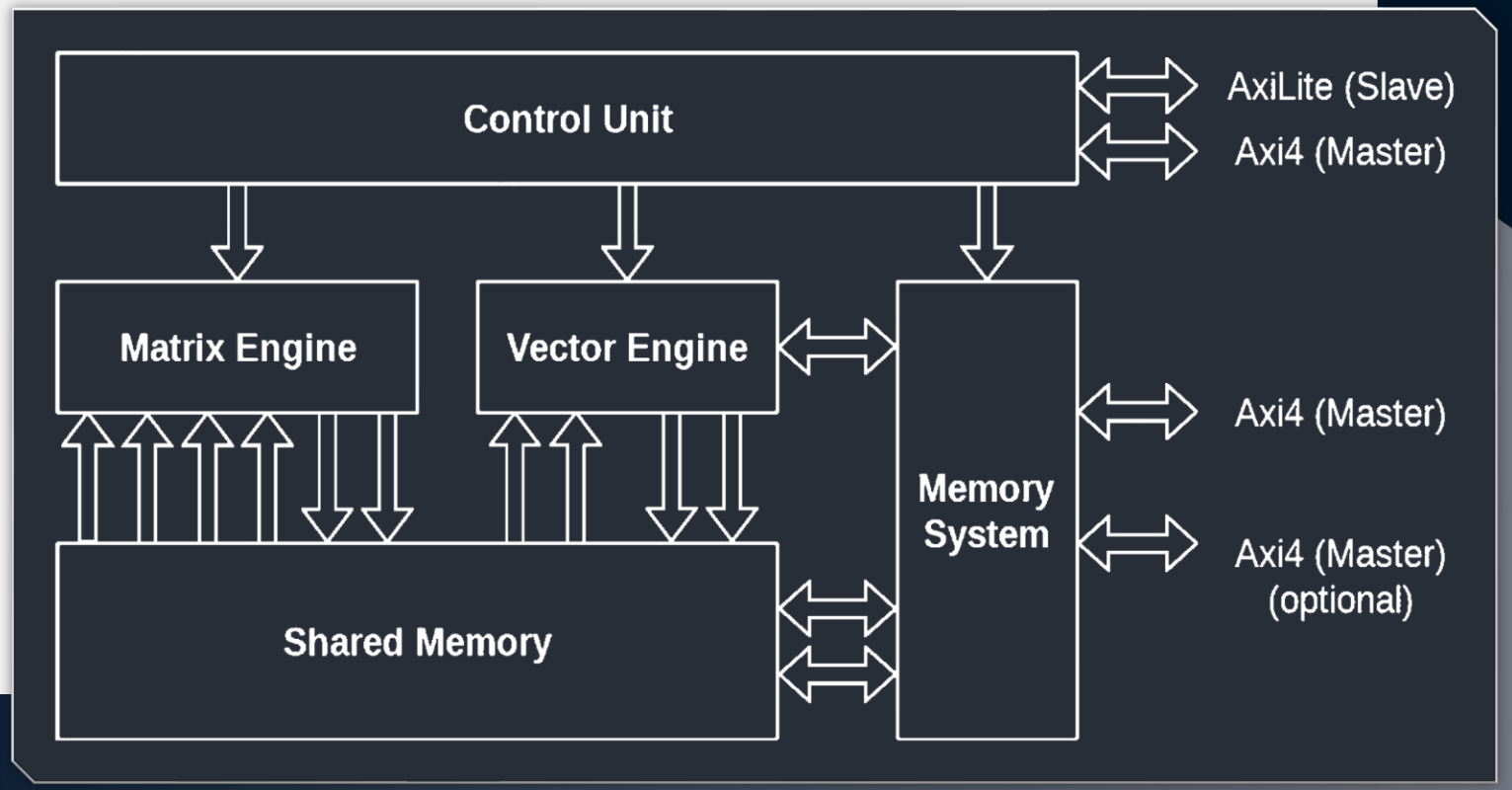
≤ 25 Вт
Энергопотребление

RISC-V
Управляющий контроллер

Собственное IP-ядро

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

- Полностью **российское** AXI_4 совместимое IP-ядро
- Расчет (inference) **сверточных** (CNN) и **рекуррентных** (RNN) нейронных сетей
- Производительность **24 TOPS (int8)** на систолическом массиве 128x128
- Вычислительных элементов **16 K**
- Максимальная рабочая частота **812 МГц**
- Тип данных **int8**
- Встроенная память до **16 МБ**



Линейка решений для инференса

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

Характеристики	P	D	Q	S	X
Внешний вид					
Производительность	24 (int8) ТОПС	48 (int8) ТОПС	96 (int8) ТОПС	960 (int8) ТОПС	24 (int8) ТОПС
ResNet-50 (batch 1)	460 FPS	920 FPS	1840 FPS	-	460 FPS
ResNet-50 (batch 8)	1046 FPS	2092 FPS	4184 FPS	-	1046 FPS
Частота TPU	до 812 МГц	до 812 МГц	до 812 МГц	до 812 МГц	до 812 МГц
Интерфейсы	PCIe 2.0 x8	PCIe 3.0 x16	PCIe 3.0 x16	2x PCIe x16 Gen4	мулти
Форм-фактор	PCIe x8, 1 слот	PCIe x16, 1 слот	PCIe x16, 1 слот	3U Rackmount	SBC3,5”
Энергопотребление	<25 Вт	<55 Вт	<105 Вт	<1150 Вт	<45 Вт



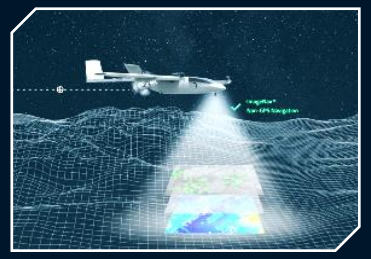
Робототехника



Медицина



Транспорт



Космос и ДЗЗ



Биометрия



Безопасность

Поддерживаемые сети

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

- densenet121
- efficientnet_b0
- efficientnet_b1
- efficientnet_b2
- efficientnet_b3
- inception_v1
- inception_v2
- inception_v3
- inception_v4
- lenet5
- mobilenet_v1
- mobilenet_v2
- nasnet_mobile
- pnasnet_large
- pnasnet_mobile
- resnet152
- resnet34
- resnet50_mlperf
- resnet50_v2
- ssd_mobilenet_v1
- ssd_mobilenet_v2
- scrfd 2.5g
- tiny_yolo2
- tiny_yolo3
- torch_densenet169
- torch_mobilenet_v2
- torch_resnet50
- vgg19
- xception
- YOLO2
- YOLO3
- YOLO4
- YOLO5s
- YOLO8
- YOLO11
- … и другие

Производительность*

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

Сети (выборочно)		Batch	P		Q	
			Throughput, FPS	Latency, ms	Throughput, FPS	Latency, ms
1	anti_spoof_mn3	1	1561,1	985,72	6244,4	985,72
2	inception_v3	1	287,1	4,77	1148,4	4,77
3	inception_v3	8	480,8	-	1923,2	-
4	lenet5	1	12519,7	128,90	50078,8	128,90
5	lenet5	8	84961,6	-	339846,4	-
6	mobilenet_v2	1	1201,7	1,44	4806,8	1,44
7	mobilenet_v2	8	2456,8	-	9827,2	-
8	resnet152	1	228,4	5,34	913,6	5,34
9	resnet50	1	432,9	2,99	1731,6	2,99
10	resnet50	8	1112,8	-	4451,2	-
11	ssd_mobilenet_v2	1	556,7	3,54	2226,8	3,54
12	ssd_mobilenet_v2	8	428	-	1712	-
13	yolo5s	1	106,1	23,71	424,4	23,71
14	yolov8n_pose	1	104,8	21,51	419,2	21,51

**Измерение выполнено по методике MLCommons MLPerf (Intel Core i7-10700K CPU @ 3.80GHz, Ubuntu 24.04, TPU Runtime 16.18.6, TPU Framework 2026.8.43*

Оригинальный Framework

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

1. **Разработка модели через стандартные фреймворки машинного обучения**
2. **Квантование модели**
3. **Формирование вычислительного графа**
4. **Расчет графа на TPU**



Оригинальный Framework

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ



Компьютерное зрение



Фото- и видеоаналитика



Промышленный интернет вещей



Автономная навигация и ADAS



Интеллектуальные камеры



Анализ мед. данных



Предиктивное техобслуживание



Распределенные вычисления



Умный город



Робототехника



Обработка данных



Умный ритейл

Ускорители для инференса нейронных сетей в режиме реального времени

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ

P
(1 чип)



D
(2 чипа)



Q
(4 чипа)



Ускорители **P**, **D** и **Q** построены на базе российского тензорного процессора и предназначены для выполнения инференса нейронных сетей в реальном времени с низкой задержкой в составе ПЭВМ, серверов и ПАК на их основе

Универсальный ускоритель HPS для высокоплотных вычислений

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ



Акселератор **HPS** построен на базе модулей ускорения HPQ, имеет пиковую производительность 960 TOPS и предназначен для выполнения инференса тяжелых нейронных сетей в реальном времени с низкой задержкой, обеспечивая потоковую передачу входных данных и результатов инференса.

ХАРАКТЕРИСТИКИ

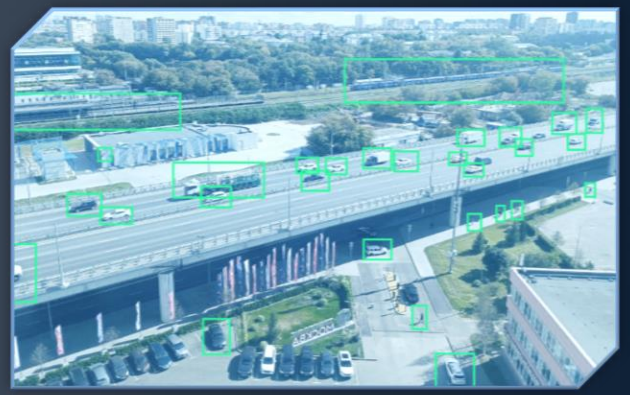
- ✓ **Сверхнизкая задержка при малом батче (текущий: 2,3 мс, возможность оптимизации до 1,5 мс для ResNet-50)**
- ✓ **Высокая эффективность на малом батче на CNN /Transformer моделях**
- ✓ **Нет сторонних IP в вычислителях (ARM, etc.)**
- ✓ **Оптимизаторы компилятора с элементами ИИ**
- ✓ **Программный стек собственной разработки**
- ✓ **Прямая интеграция с TensorFlow 2**
- ✓ **Интеграция с PyTorch через ONNX**
- ✓ **Пиковая производительность 960 TOPS (int8)**
- ✓ **Поддержка инференса более 40 нейронных сетей**
- ✓ **Форм-фактор: 3U Rackmount**
- ✓ **PCIe интерфейс: x32 Gen4**

Области применения

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ



Помощь в принятии решений



Детектирование и классификация объектов



Предиктивное техобслуживание



Распределенные вычисления



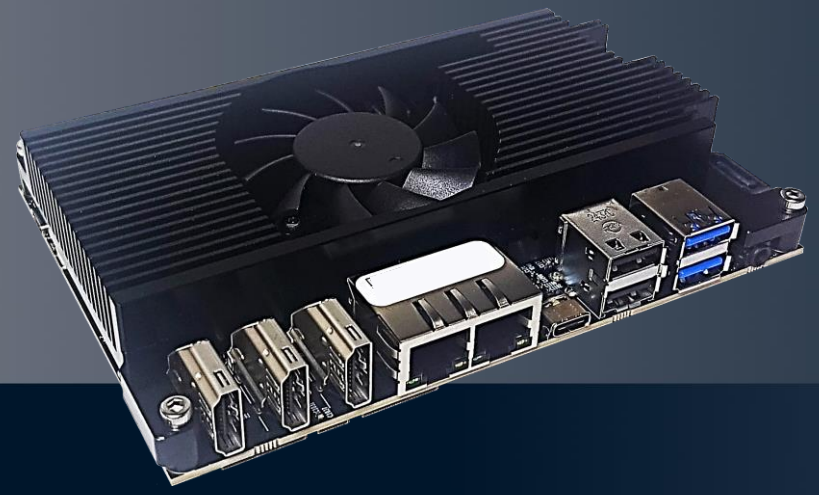
Аналитика и безопасность



Научные исследования

Ускоритель для инференса нейронных сетей в режиме реального времени

ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ



Модуль построен на базе тензорного процессора и встроенного процессора общего назначения с широким набором интерфейсов.

Ускоритель предназначен для запуска ИИ-приложений и применения в беспилотных системах, выполняющих задачи в области мониторинга, доставки, картографии, поисково-спасательных операций и др.

Производительность

до **24 TOPS** (int8) и 6 TOPS на одном модуле,

одновременный запуск **более 5 моделей** нейронных сетей

Гибкая интеграция

широкий набор интерфейсов обеспечивает управление и отладку модуля, подключение внешних устройств и различных компонентов, прием/передачу данных и пр.

Низкая задержка

~2,3 мс для ResNet-50 (batch=1),
при оптимизациях – **до 1,5 мс**

специализированные оптимизаторы в компиляторе (DNN Quant, Graph Optimizer)

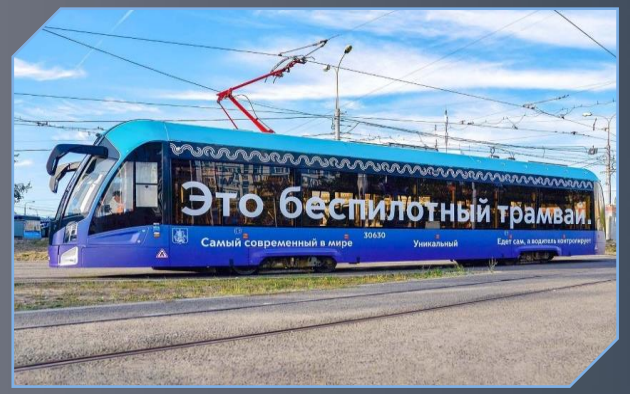
Программный стек собственной разработки

нативная интеграция с TensorFlow 2.x

поддержка **PyTorch** через **ONNX**

Области применения

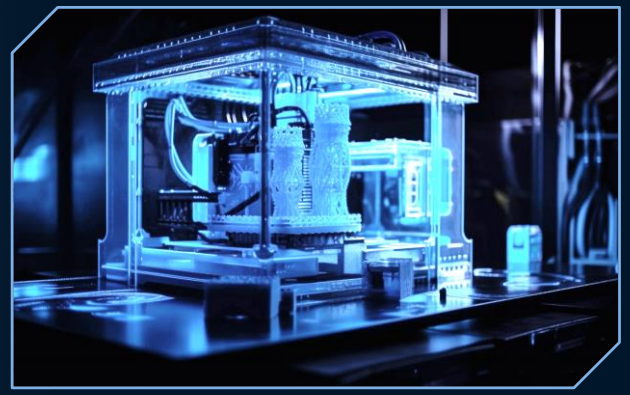
ЦИФРОВЫЕ ИНДУСТРИАЛЬНЫЕ ТЕХНОЛОГИИ ЦИТ



Беспилотный рельсовый транспорт



Автономный транспорт



Промышленные роботы и робототехника



Промышленный интернет вещей (IIoT)



Беспилотные летательные аппараты



Системы помощи водителю (ADAS)

ЦИФРОВЫЕ
ИНДУСТРИАЛЬНЫЕ
ТЕХНОЛОГИИ

Благодарим за внимание!

t.me/cit_gov

cit.gov.ru