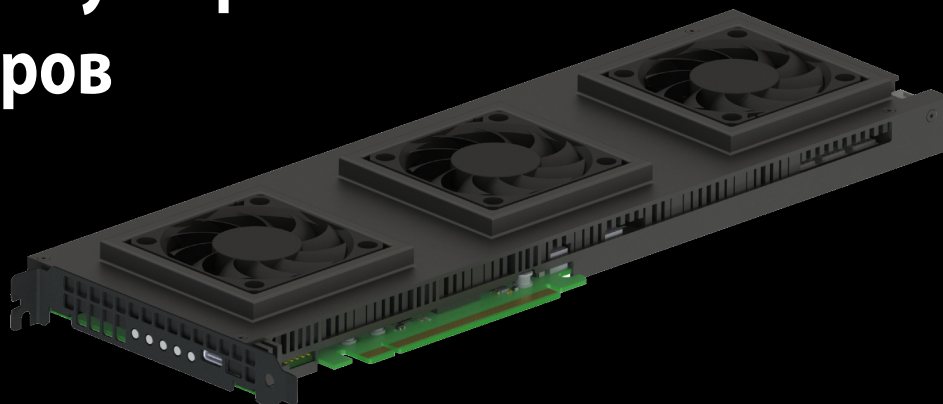


## Универсальный ускоритель для дата-центров



Модуль LinQ HPQ предназначен для запуска ИИ-приложений и применения в стандартной инфраструктуре центров обработки данных.

LinQ HPQ входит в линейку модулей ускорения LinQ стандарта PCI Express (PCIe). Скорость передачи данных до 8 Гбайт/с (8 линий) позволяет использовать модули LinQ HPQ для выполнения инференса нейронных сетей в реальном времени с низкой задержкой, обеспечивая потоковую передачу входных данных и результатов инференса.

LinQ HPQ построен на базе четырех тензорных процессоров собственной архитектуры LinQ H с тактовой частотой блока вычислителя 500-812 МГц и микроконтроллером DDR4+ECC до 64 Гб, что позволяет добиться максимальной производительности устройства при решении целевых задач.

## КЛЮЧЕВЫЕ 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 TF2
- Интеграция PyTorch через ONNX
- Пиковая производительность 96 TOPS (int8)
- Поддержка инференса нескольких моделей



## ОБЛАСТИ ПРИМЕНЕНИЯ



Машинное обучение  
как услуга (MLaaS)



Распределенные  
вычисления



Оптимизация  
энергопотребления



Городская  
безопасность



Персонализация  
контента



Исследование  
и разработка лекарств



Интеллектуальная  
видеоаналитика



Биометрия

## ПРОГРАММНЫЙ СТЕК

### ML Frameworks

- TensorFlow 2.x
- PyTorch с ONNX
- ONNX

### TPU Frameworks

- ONNX Converter
- DNN Quant
- TPU Compiler

### Applications

- Atlantic TPU
- TPU Cloud
- T3: TPU Testing Tools

### Platforms

- x86
- ARM
- Mips64
- E2K

### Analysis & Development

- DNN Model Zoo
- DNN Stat
- Performance Profiler

### TPU Runtime

- TPU Driver
- PyTPU
- LibTPU
- Real-time scheduler
- Multi-model orchestrator

## ТЕХНИЧЕСКАЯ СПЕЦИФИКАЦИЯ

### Основные параметры

- Форм-фактор: PCIe Full Length
- PCIe интерфейс: x16 PCIe Gen3
- Тактовая частота: 500-812 МГц (ПО управляема)
- Память: 64 Гб DDR4 ECC
- Потребление: 120/140 Вт (типовое/пиковое)
- Power Monitor
- Контроль теплового режима

### Надежность

- ECC защита памяти
- Watchdog timer
- Защита от перегрузки по току

### Производительность

- int8: 96 TOPS

## ЛИНЕЙКА УСКОРИТЕЛЕЙ LINQ

	LinQ HP	LinQ HPD	LinQ HPQ	LinQ HPS (Server)
Производительность	24 TOPS	48 TOPS	96 TOPS	До 960 TOPS
Форм-фактор	PCIe Half-Length	PCIe 3/4 Length	PCIe Full Length	3U Rackmount
Охлаждение	Active	Passive/Active	Passive/Active	Passive/Liquid
Память	16GB DDR4	32GB DDR4	64GB DDR4	до 640GB DDR4
PCIe	x8 Gen3	x16 Gen3	x16 Gen3	-
Потребление	25W	60W	120W	До 1000W
Применение	Edge/Embedded	Fog	Cloud	Data Center
Особенности	Низкая задержка	Низкая задержка	Высокая плотность	до 10x HPQ в шасси

Каждый продукт имеет  
отдельный подробный даташит  
с полными техническими характеристиками  
и сценариями применения

