

SMARC-МОДУЛИ АТРОНИК КАК АЛЬТЕРНАТИВА NVIDIA JETSON

АЛЕКСЕЙ МЕДВЕДЕВ

В статье рассмотрены архитектуры центральных процессоров для ускорения работы с искусственными нейронными сетями. Приведены примеры отечественных вычислительных модулей и блоков для решения задач машинного зрения, видеоаналитики и оптической навигации.

ВВЕДЕНИЕ

В настоящее время одной из самых популярных встраиваемых аппаратных платформ для задач искусственного интеллекта (ИИ), машинного зрения и видеоаналитики являются компьютерные модули Nvidia семейства Jetson [1].

Nvidia Jetson представляет собой линейку встраиваемых компьютерных модулей (SOM, System on module) на базе графических процессоров специально разработанных для взаимодействия с системами искусственного интеллекта и Edge Computing

Популярность Nvidia Jetson обусловлена высокой производительностью, легкостью использования и поддержкой сообществом разработчиков.

Несмотря на свою популярность, технологии Nvidia имеют некоторые сложности, связанные с международной политической обстановкой и санкциями, введенными против некоторых стран. В условиях ограниченного доступа к технологиям Nvidia разработчики и компании вынуждены искать альтернативные решения, такие как платформы на базе ARM-процессоров с интегрированными ядрами-ускорителями — например, RockChip, Nailo, НТЦ «Модуль», LinQ.

Каждая из этих альтернатив имеет свои сильные и слабые стороны, поэтому выбор зависит от требований к производительности, стоимости, энергопотреблению и специфике применения.

GPU, NPU, TPU

По своей архитектуре современные процессоры являются системами на кристалле (SoC — System on Chip), объединяя на одном кристалле несколько компонентов вычислительной системы. Это позволяет уменьшить количество отдельных микросхем и сделать устройство более компактным и экономичным с точки зрения потребления энергии.

В процессорах, предназначенных для работы с нейронными сетями и обработкой видео, помимо ядер центрального процессора (CPU, Central Processing Unit), присутствуют специализированные ядра-процессоров для ускорения некоторых операций.

К таким ядрам относятся GPU, TPU и NPU — три типа процессорных архитектур, наиболее пригодных для выполнения различных задач в области параллельных вычислений, связанных с обработкой графической информации, ускорением нейронных сетей и машинным обучением.

Графический процессор (GPU, Graphics Processing Unit) разрабатывался изначально для ускорения обработки графики и рендеринга изображений. Архитектура включает множество ядер, способных выполнять параллельные вычисления, что делает GPU оптимальным для параллельной обработки больших массивов данных, как это необходимо в графических приложениях и в задачах машинного обучения.

GPU-процессоры могут содержать тысячи простых ядер, что позволяет им обрабатывать большие потоки данных.

Высокий уровень параллелизма позволяет GPU-процессорам эффективно справляться с задачами, которые могут быть разбиты на множество мелких подзадач, таких как обработка изображений, инференс и обучение нейронных сетей.

Еще один тип специализированного процессора — тензорный процессор (TPU, Tensor Processing Unit), разработанный для реализации операций, характерных для ускорения и обучения нейронных сетей. Он использует систолические массивы, обеспечивая быстрое выполнение высокопроизводительных операций умножения и сложения матриц.

Нейронный процессор (NPU, Neural Processing Unit) — тип специализи-

рованного аппаратного ускорителя, который предназначен для обучения и инференса моделей глубокого или машинного обучения, имитирующих нейронные сети человеческого мозга. NPU оптимизированы для математических операций, таких как умножение матриц и свертки для задач, связанных с искусственными нейронными сетями. Обычно они используются с центральным процессором (ЦП) для обеспечения дополнительной вычислительной мощности.

В отличие от универсального GPU, процессоры NPU и TPU ориентированы для ускорения операций с нейронными сетями и решения задач искусственного интеллекта. И NPU, и TPU оптимизированы для математических операций, таких как умножение матриц и свертки.

По назначению и принципу работы NPU и TPU очень схожи. Довольно часто термин NPU выступает как общее название для акселераторов нейросетей.

Между NPU и TPU есть некоторые различия. Одно из ключевых различий заключается в том, что TPU специально разработаны для ускорения задач глубокого обучения, в то время как NPU могут ускорять более широкий спектр алгоритмов, в том числе машинного обучения.

С точки зрения производительности NPU и TPU являются высокоэффективными и мощными ресурсами для ускорения алгоритмов обработки нейронных сетей. Однако TPU могут иметь небольшое преимущество в производительности из-за их особой оптимизации для задач глубокого обучения. Также стоит отметить, что конкретная производительность NPU или TPU будет зависеть от его конструкции и реализации.

В качестве примера на рис. 1 представлена структура ядра TPU-процессоров одного из китайских производителей. TPU разработан с несколькими вычислительными

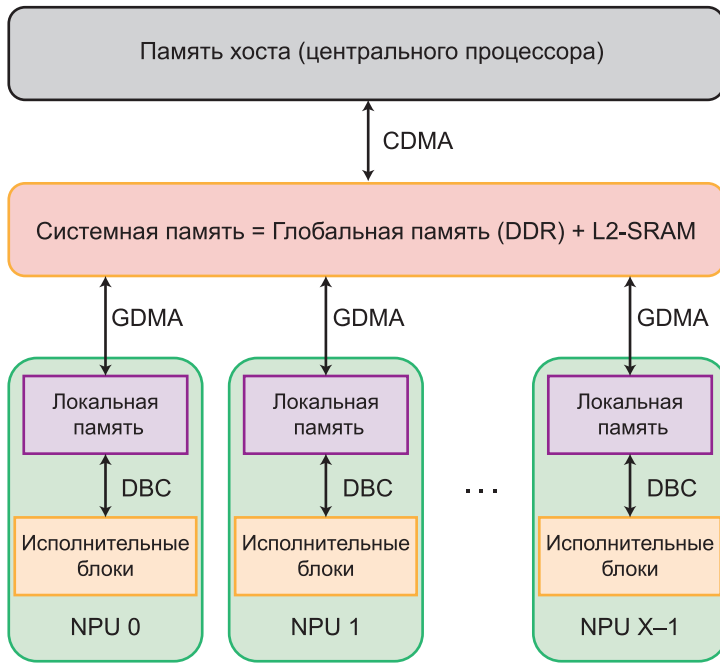


Рис. 1. ◀
Архитектура TPU-процессора

- DMA: Прямой доступ к памяти
- CDMA: Канал прямого доступа к памяти
- GDMA: Глобальный прямой доступ к памяти
- DBC: Контроллер широкополосных данных
- NPU: Блок обработки алгоритмов нейронных сетей

ядрами, каждое из которых называется NPU. TPU основан на архитектуре Single Instruction Multiple Data (одиночный поток команд, множественный поток данных, ОКМД) и имеет многоядерную конструкцию. Он состоит из DBC (контроллера широкополосных данных) и GDMA (глобального прямого доступа к памяти).

В этой архитектуре TPU выполняет вычисления по принципу SIMD, то есть в любой момент времени все NPU выполняют одну и ту же вычислительную инструкцию, но каждый NPU работает с разными данными.

TPU оптимален для выполнения крупномасштабных задач глубокого обучения, особенно в сценариях, требующих высокой пропускной способности и низкой задержки.

КИТАЙСКИЕ ПРОИЗВОДИТЕЛИ

Процессоры ряда китайских производителей могут рассматриваться как альтернативы Nvidia для устройств в области обработки нейронных сетей, видеоаналитики и компьютерного зрения. Одними из самых популярных в России являются процессоры RockChip.

RockChip — это китайский производитель процессоров, который предлагает широкий спектр чипов для встраиваемых систем, мобильных устройств и IoT-решений. Некоторые из моделей, такие как RK3568, RK3588, оснащены NPU для ускорения задач ИИ.

Среди процессоров RockChip можно выделить RK3588. Модули на базе данного ЦП могут представлять собой

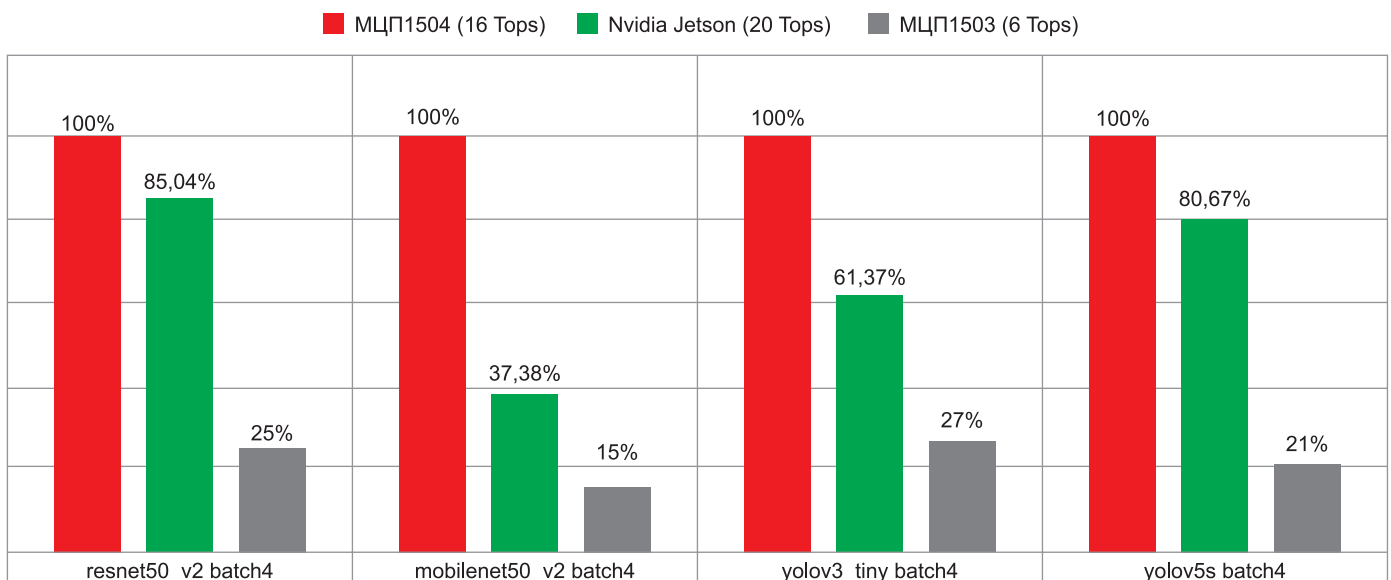
интересные альтернативы платформам Nvidia Jetson, особенно для задач, связанных с искусственным интеллектом и обработкой данных.

ПРИМЕРЫ МОДУЛЕЙ И БЛОКОВ

В настоящее время на российском рынке достаточно широко представлены встраиваемые вычислители на базе процессоров RockChip. В качестве примера рассмотрим компьютерные модули производства НПК «АТРОНИК».

В номенклатуре компании имеются встраиваемые компьютерные модули на процессорах архитектуры ARM со встроенными ядрами NPU/TPU. В таблице приведен сравнительный анализ модулей SMARC производства НПК «АТРОНИК» [2].

Рис. 2. ▼
Сравнение производительности компьютерных модулей с аналогами от Nvidia



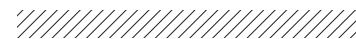


ТАБЛИЦА. МОДУЛИ SMARC ПРОИЗВОДСТВА НПК «АТРОНИК»




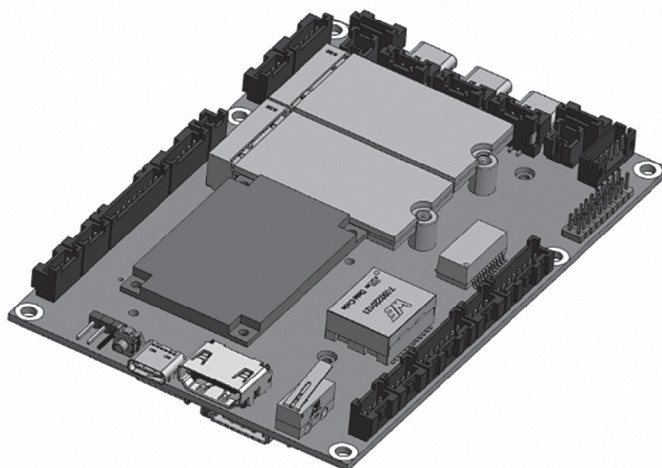
Наименование	МЦП1502	МЦП1503	МЦП1504
Внешний вид			
Форм-фактор	SMARC 2.1	SMARC 2.1	SMARC 2.1
Центральный процессор	4 ядра ARM Cortex-A55 1,4 ГГц	8 ядер (4×Cortex-A76 + 4×Cortex-A55), 2 ГГц	8 ядер Cortex-A53 1,6 ГГц
Ускоритель	NPU, 1Tops	NPU 6Tops	TPU, 16Tops
Объем ОЗУ	DDR4 4 Гбайт с ECC	LPDDR4 16 Гбайт	LPDDR4 16 Гбайт
Объем ПЗУ (eMMC)	32 Гбайт	64/128 Гбайт	64/128 Гбайт
Производительности Tops (INT8)	0,8	6	16
HDMI	v.2.0	v.2.1	v.2.1
Потребляемая мощность, Вт	До 7	До 20	До 20
Диапазон рабочей температуры, °C	-40...+85	-40...+85	-40...+85
Поддерживаемые операционные системы	Linux Ubuntu, AstraLinux, ЗОСРВ Нейтрино	Linux Ubuntu, AstraLinux, AltLinux	Linux

РИС. 3. ►
Бортовая доверенная
вычислительная платформа
производства
НПК «Атроник»



Далее представлено несколько примеров вычислительных устройств, разработанных на базе данных компьютерных модулей.

Бортовая доверенная вычислительная платформа, (рис. 3), представляет собой встраиваемый вычислитель для создания устройств видеоаналитики на основе нейронных сетей с возможностью криптографической защиты данных и каналов управления. Платформа может использоваться в качестве интеллектуального вычислителя различных робототехнических комплектов и беспилотных транспортных средств, а также на стационарных наземных объектах в составе интеллектуальных видеокамер и многопоточных AI EDGE серверов.

Бортовой компьютер (рис. 4) системы оптической навигации (СОН) — это система, которая позволяет определять местоположение и ориентацию беспилотного воздушного судна (БВС) в пространстве, автономно осуществлять навигацию, выполнять миссии и задачи в условиях отсутствия сигналов ГНСС с помощью методов визуальной одометрии. Бортовой компьютер обрабатывает входящую информацию и выдает координаты расположения БВС на местности.

Интеллектуальная IP-видеокамера со встроенной аналитикой и средствами кибербезопасности (рис. 5) обеспечивает установку и исполнение нейронных сетей пользователя, непрерывное кибербезопасное видеонаблюдение за объектом в усло-

РИС. 4. ▼
Бортовой компьютер
(НПК «Атроник») для БЛА



Дублированные коммуникационные интерфейсы (Ethernet, CAN, RS-232/422/485), система коррекции ошибок памяти (ECC), промышленный температурный диапазон

эксплуатации обеспечивают эффективное использование модулей НПК «Атроник» при создании надежных компьютерных систем.

На рис. 2 проведено сравнение производительности представленных компьютерных модулей с аналогами от Nvidia при работе с популярными наборами нейронных сетей.

На базе представленных компьютерных модулей могут быть созданы ИИ-видеосерверы, интеллектуальные камеры видеонаблюдения, системы интеллектуального мониторинга и управления беспилотным транспортом и другие высокопроизводительные вычислительные решения с низким энергопотреблением. Благодаря поддержке режима сопроцессора модули могут использоваться в качестве внешнего нейросетевого ускорителя.

виях размещения видеокамеры вне защищенного периметра и в условиях нестабильных энергообеспечения и каналов связи. Фиксирует юридически значимые события и добавляет электронную цифровую подпись на зафиксированные кадры.

ЗАКЛЮЧЕНИЕ

Выбор между Nvidia Jetson и его аналогами зависит от конкретных требований проекта, включая производительность, поддержку фреймворков, энергопотребление и бюджет. Если нужна высокая производительность для сложных AI-задач, Nvidia Jetson может быть лучшим выбором. Если же необходимы доступные решения для менее требовательных приложений, RK3588 и другие процессоры с интегрированными NPU/TPU-ускорителями могут стать хорошими альтернативами. ●



РИС. 5. ◀ Интеллектуальная киберзащитная IP-видеокамера производства НПК «Атроник»

ЛИТЕРАТУРА

1. NVIDIA Celebrates 1 Million Jetson Developers Worldwide at GTC. www.blogs.nvidia.com/blog/million-jetson-developers-gtc/
2. Медведев А. В. Компьютерные модули формата SMARC от НПК «Атроник» // Control Engineering Россия. 2023. № 4.