УДК: 004.89

Худайберидева Г. Б., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Кожухов Д. А., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Пименкова А. А., студент-бакалавр кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

НЕЙРО-АППАРАТНЫЕ СИСТЕМЫ НА КРИСТАЛЛЕ (NEUSOC) ДЛЯ МИКРО-LLM: ИНТЕГРАЦИЯ СПЕЦИАЛИЗИРОВАННЫХ АКСЕЛЕРАТОРОВ В МЕДЛЕННЫЕ ПРОМЫШЛЕННЫЕ МК

***Аннотация**: Предложена концепция Нейро-Аппаратных Систем на Кристалле (NeuSoC), направленная на эффективное исполнение микроскопических языковых моделей (Микро-LLM) на промышленных микроконтроллерах (МК) с ограниченными вычислительными ресурсами и частотой. В отличие от подходов, требующих высокопроизводительных центральных процессоров, NeuSoC интегрирует специализированные, сверхэнергоэффективные аппаратные акселераторы напрямую в кристалл существующих МК, выступая в роли специализированной периферии (аналогично SPI/I2C). Статья детализирует архитектуру таких акселераторов, фокусируясь на блоках для матричных умножений 8-bit, функций активации (Softmax) и операций внимания. Рассматривается взаимодействие акселераторов с основным ядром МК через стандартизированные интерфейсы и вопросы компиляции моделей под гетерогенную систему NeuSoC. Показана принципиальная возможность*

значительного ускорения вывода Микро-LLM при сохранении крайне низкого энергопотребления.

***Ключевые слова**: Нейро-Аппаратные Акселераторы, Система на Кристалле, Микро-LLM, Микроконтроллеры, Энергоэффективность, TinyML, Аппаратная Ускорение, Гетерогенные Вычисления, Специализированные Процессоры.*

Khudaiberideva G. B.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Kozhukhov D. A.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Pimenkova A. A.
bachelor's student at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

NEURO-HARDWARE SYSTEMS ON A CHIP (NEUSOC) FOR MICRO-LLM: INTEGRATION OF SPECIALIZED ACCELERATORS INTO SLOW INDUSTRIAL MC

***Annotation**: The concept of Neuro-Hardware Systems on a Chip (NeuSoC) is proposed, aimed at the efficient execution of microscopic language models (Micro-LLM) on industrial microcontrollers (MC) with limited computing resources and frequency. Unlike approaches that require high-performance central processors, NeuSoC integrates*

specialized, super-energy-efficient hardware accelerators directly into the chip of existing MC, acting as specialized peripherals (similar to SPI/I2C). The article details the architecture of such accelerators, focusing on blocks for 8-bit matrix multiplications, activation functions (Softmax), and attention operations. The interaction of accelerators with the main core of the MC through standardized interfaces and the issues of compiling models for the heterogeneous NeuSoC system is considered. The fundamental possibility of significantly speeding up Micro-LLM output while maintaining extremely low power consumption is shown.

***Keywords**: Neuro-Hardware Accelerators, System on a Chip, Micro-LLM, Microcontrollers, Energy Efficiency, TinyML, Hardware Acceleration, Heterogeneous Computing, Specialized Processors.*

Введение

Распространение алгоритмов глубокого обучения, в частности языковых моделей, в область периферийных и встраиваемых устройств сталкивается с фундаментальным ограничением вычислительной мощности доступных промышленных микроконтроллеров. Существующие МК, доминирующие в промышленной автоматике, IoT-датчиках и управляющих системах, характеризуются низкими тактовыми частотами (десятки-сотни МГц) и ограниченными ресурсами памяти. Традиционные подходы к ускорению нейронных сетей, часто предполагающие наличие мощных центральных или графических процессоров, неприменимы в данном контексте. Альтернативные решения на базе FPGA или внешних нейроускорителей увеличивают стоимость, сложность системы и энергопотребление, что неприемлемо для массовых промышленных применений. Возникает потребность в принципиально новых архитектурных решениях, позволяющих привнести возможности Микро-LLM в существующий парк медленных МК без их полной замены. Концепция NeuSoC предлагает путь решения через интеграцию специализированных нейроаппаратных блоков непосредственно в кристалл МК.

Архитектура Специализированных Акселераторов NeuSoC

Ядро концепции NeuSoC составляет набор специализированных, сверхоптимизированных аппаратных блоков, спроектированных для эффективного выполнения ключевых операций, доминирующих в вычислительном графе Микро-LLM. Каждый акселератор представляет собой законченный автономный модуль с выделенной локальной памятью (регистровым файлом или SRAM-буфером) и управляющим автоматом.

Матричный Акселератор (8-bit). Данный блок реализует операцию умножения матриц (GEMM) с 8-битными целочисленными (INT8) операндами. Архитектура строится вокруг массива систолических процессорных элементов (PE), организованных для максимальной утилизации данных. Каждый PE содержит несколько 8-битных умножителей-накопителей (MAC). Ключевой особенностью является минимизация перемещения данных: входные векторы и матрицы весов загружаются в локальные буферы акселератора большими блоками. Аппаратная поддержка включает насыщение и сдвиг для масштабирования результатов. Использование систолической организации позволяет достичь высокой степени параллелизма при фиксированной тактовой частоте МК.

Акселератор Функций Активации (Softmax, GELU, ReLU). Данный блок специализирован на вычислении нелинейных функций, наиболее вычислительно сложной из которых является Softmax. Реализация Softmax включает аппаратное вычисление экспоненты (например, с использованием CORDIC или таблиц LUT с интерполяцией), суммирование и деление. Блок проектируется с учетом низкой точности INT8 и возможностью конвейеризации для обработки векторов активаций. Поддержка других функций (GELU, ReLU, Sigmoid) реализуется через переконфигурируемую логику или отдельные простые схемы в рамках модуля. Критически важна оптимизация энергопотребления за счет отключения неиспользуемых частей схемы.

Акселератор Операций Внимания. Операция внимания (Attention) является ключевой для LLM, но вычислительно интенсивна. Акселератор внимания NeuSoC разбивает операцию на этапы: вычисление Q*K^T (с использованием матричного акселератора), масштабирование, применение маски (если требуется), вычисление Softmax (с использованием акселератора функций) и итоговое взвешенное суммирование V (снова матричный акселератор). Специализация заключается в оптимизированной передаче промежуточных результатов (Q, K, V) между внутренними блоками акселератора, минимизируя обращения к системной памяти. Аппаратная поддержка масок и масштабирования интегрирована непосредственно в схему.

Взаимодействие с Основным Ядром МК и Системная Архитектура

Интеграция блоков NeuSoC в существующий МК осуществляется по модели периферийного устройства. Акселераторы подключаются к системной шине МК (например, AHB или APB в ARM Cortex-M) через стандартизированный интерфейс, аналогичный интерфейсам SPI или I2C контроллеров, но оптимизированный для передачи блоками данных.

Каждый акселератор предоставляет набор регистров управления и статуса, отображаемых в адресное пространство МК. Программное обеспечение на основном ядре (CPU) инициирует операции путем:

1. Конфигурации параметров акселератора (размеры данных, адреса и т.д.) через запись в регистры.
2. Записи входных данных в выделенный буфер памяти акселератора (через DMA или программный ввод-вывод).
3. Запуска вычисления установкой бита запуска в регистре управления.
4. Ожидания завершения операции (по прерыванию или флагу статуса).
5. Считывания результатов из выходного буфера акселератора.

Аппаратная поддержка прямого доступа к памяти (DMA) критически важна для минимизации нагрузки на CPU и повышения общей пропускной способности системы.

Основная сложность заключается в разделении данных между CPU и акселераторами. NeuSoC предполагает использование единого адресного пространства. Акселераторы работают с данными в своих локальных буферах. Задача CPU – обеспечить когерентность данных между системной памятью (часто внешней SPI/SRAM) и локальными буферами акселераторов через явные операции копирования (с использованием DMA). Кэши CPU требуют осторожного управления (инвалидация/очистка) при обмене данными с акселераторами. Предлагается модель программирования с явным управлением памятью.

Компиляция Моделей для Гетерогенной Архитектуры NeuSoC

Эффективное использование NeuSoC требует специализированного компилятора, способного разбить вычислительный граф Микро-LLM на части, исполняемые на CPU и на специализированных акселераторах.

Компилятор анализирует граф модели, идентифицируя операторы, которые могут быть отображены на доступные аппаратные акселераторы (GEMM, Softmax, Attention). Остальные операторы (например, управляющие структуры, специфические функции) остаются для исполнения на CPU. Компилятор генерирует последовательность вызовов драйверов акселераторов и операций на CPU, обеспечивая корректную передачу данных между ними.

Ключевая задача – минимизация перемещения данных между системной памятью и локальными буферами акселераторов. Компилятор должен планировать выполнение таким образом, чтобы максимизировать повторное использование данных внутри акселератора до их выгрузки. Применяются техники: группировка последовательных операций одного типа, планирование с учетом размера буферов акселераторов,

минимизация числа запусков акселераторов за счет агрегации данных. Генерация кода для CPU включает вызовы низкоуровневых драйверов акселераторов и управление DMA-трансферами..

Заключение

Концепция NeuSoC предлагает путь для внедрения возможностей Микро-LLM в обширный класс промышленных микроконтроллеров, изначально не предназначенных для выполнения сложных задач ИИ. Интеграция специализированных аппаратных акселераторов непосредственно в кристалл МК в качестве периферийных блоков позволяет достичь значительного ускорения ключевых операций нейронных сетей при сохранении общего энергетического бюджета системы. Архитектура акселераторов, ориентированная на операции 8-bit GEMM, Softmax и Attention, напрямую соответствует вычислительному профилю Микро-LLM. Взаимодействие через стандартизированный интерфейс упрощает интеграцию в существующие платформы.

Разработка эффективного компилятора, способного выполнять разбиение графа модели и оптимизировать обмен данными в гетерогенной среде CPU + NeuSoC, является необходимым условием практической реализации концепции. При этом сохраняется возможность исполнения кода на основном ядре МК, обеспечивая гибкость. Основными преимуществами подхода являются потенциально высокая энергоэффективность, достигаемая за счет специализации аппаратуры, сохранение низкой стоимости системы за счет отсутствия внешних компонентов и совместимость с парком существующих промышленных МК. Реализация NeuSoC требует решения задач по аппаратному дизайну энергоэффективных блоков, разработке низкоуровневых драйверов и интеллектуального компилятора. Предложенная архитектура открывает перспективы для создания нового класса промышленных

интеллектуальных устройств с расширенными возможностями обработки естественного языка непосредственно на периферии сети.

СПИСОК ЛИТЕРАТУРЫ:

1. Horowitz, M. Computing's energy problem (and what we can do about it) / M. Horowitz // IEEE International Solid-State Circuits Conference (ISSCC). — 2014.
2. Lin, J. MCUNet: Tiny Deep Learning on IoT Devices / J. Lin, W.-M. Chen, Y. Lin [et al.] // Advances in Neural Information Processing Systems (NeurIPS). — 2020.
3. Banbury, C. Benchmarking TinyML Systems: Challenges and Direction / C. Banbury, V. J. Reddi, P. Torelli [et al.] // IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). — 2021.
4. Davies, M. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning / M. Davies, N. Srinivasa, T.-H. Lin [et al.] // IEEE Micro. — 2018. — Vol. 38, no. 1. — P. 82–99.
5. Judd, P. Stripes: Bit-Serial Deep Neural Network Computing / P. Judd, J. Albericio, A. Moshovos // IEEE/ACM International Symposium on Microarchitecture (MICRO). — 2016.
6. Parashar, A. SCNN: An Accelerator for Compressed-Sparse Convolutional Neural Networks / A. Parashar, M. Rhu, A. Mukkara [et al.] // ACM/IEEE International Symposium on Computer Architecture (ISCA). — 2017.
7. Chen, Y. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks / Y. Chen, T. Krishna, J. S. Emer, V. Sze // IEEE Journal of Solid-State Circuits (JSSC). — 2017. — Vol. 52, no. 1. — P. 127–138.

8. Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems (NeurIPS). — 2017.
9. Dally, W. J. Domain-specific hardware accelerators / W. J. Dally // Communications of the ACM. — 2020. — Vol. 63, no. 6. — P. 58–65.
10. Hennessy, J. L. Computer Architecture: A Quantitative Approach / J. L. Hennessy, D. A. Patterson. — 6th ed. — San Francisco : Morgan Kaufmann, 2019. — 928 p. — (Глава по Domain-Specific Architectures).
11. Что такое системы на кристалле (SoC): Принципы построения и применения // Электронные компоненты и системы. — 2022. — URL: https://www.eds-soft.ru
12. Шахнов, В. А. Проектирование цифровых систем на СБИС / В. А. Шахнов. — М. : Горячая линия — Телеком, 2010. — 424 с.
13. Mittal, S. A Survey on Modeling and Improving Reliability of DNN Algorithms and Accelerators / S. Mittal // Journal of Systems Architecture. — 2021. — Vol. 118. — P. 102188. — DOI: 10.1016/j.sysarc.2021.102188.
14. Warden, P. TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers / P. Warden, D. Situnayake. — Sebastopol : O'Reilly Media, 2019. — 500 p.
15. Макаров, С. Б. Микроконтроллеры ARM Cortex-M. Архитектура, программирование, разработка устройств / С. Б. Макаров, Ю. А. Первицкий. — СПб. : БХВ-Петербург, 2015. — 450 с. —
16. Chen, T. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning / T. Chen, T. Moreau, Z. Jiang [et al.] // USENIX Symposium on Operating Systems Design and Implementation (OSDI). — 2018. — P. 578–594.