**Физические оценки для реализации аппаратного ускорителя операции векторно-матричного умножения на основе фотоники**

**Байсланова Л.Р.1**

1студентка

Московский государственный университет имени М.В.Ломоносова,физический факультет, Москва, РоссияE–mail: *bayslanovalr@my.msu.ru*

Одна из главных проблем ИИ — ограниченность вычислительных ресурсов, так как обучение современных моделей может занимать недели [1]. Производительность увеличивается с количеством параметров [3], что требует роста вычислительных мощностей для дальнейшего прогресса.

Проблемы высокого энергопотребления и масштабируемости нейронных сетей становятся более актуальными в области, связанной с ИИ. Одно из решений —технологии, альтернативные классической микроэлектронике, в том числе, фотонные вычисления.

Устройства, построенные на базе фотоники позволяют производить вычисления, связанные с задачами ИИ, быстрее и с большей энергоэффективностью, чем устройства в основе которых лежит микроэлектроника [5]. Данные устройства также обладают потенциалом для упрощения производственного процесса, поскольку их изготовление не требует применения литографии с высоким разрешением на уровне единиц или десятков нанометров. Системы, в основе которых лежит фотоника, обладают высокой параллельностью, что позволяет быстро, относительно микроэлектроники, производить вычисления матриц больших размерностей [5].

Однако у фотонных вычислителей есть ограничения, включая способность фотоники работать с меньшими размерностями по сравнению с электроникой. Поэтому важной задачей является оценка разрядности данных, которую может обеспечить такой аппаратный ускоритель. Для оценки разрядности нужно учитывать оптические потери и уровень шумов, зависящих от длины оптического пути. Предлагается оценить эти параметры на основе потерь на наиболее длинном пути распространения света. Зная потери в цепи и характеристики шума компонентов (например, лазера и детектора), можно вычислить отношение сигнал/шум и нелинейные искажения (SINAD).

Это позволяет определить эффективное количество бит (ENOB), которое оценивает глубину битности системы. ENOB зависит от SINAD и рассчитывается по формуле$: ENOB=\frac{SINAD-1.76 dB}{6.02}$ [4].

В работе рассматриваются 2 архитектуры сопроцессора на основе интегральной фотоники, с помощью которых, можно производить векторно-матричное умножение: SVD-архитектура и Crossbar архитектура.

*Рис1. Схема Crossbar и SVD на фотонно-интегральной схеме.*

В архитектуре Crossbar векторные входные данные умножаются на элементы матрицы, после чего результаты суммируются посредством интерференции световых волн [6]. Архитектуры на основе сингулярного разложения матрицы (SVD) используют разложение целевой матрицы любых вещественных значений на 2 унитарные и 1 диагональную матрицы для выполнения вычислений [2]. Такую архитектуру также называют архитектурой Клементса в честь автора статьи, который впервые ее предложил. Потери в схеме складываются из cуммы потерь на каждом элементе схемы на самом длинном пути.

Максимальный размер ядра зависит не только от архитектуры, но и от используемых компонентов, таких как фазовые модуляторы, материал волноводов, ЦАП/АЦП и др. Для значений вектора с размерностью *N*>32, обрабатываемых за один такт фотонным сопроцессором, увеличение оптических потерь становится ограничивающим фактором, снижающим разрядность. Для улучшения масштабируемости и производительности необходимо совершенствование архитектуры и компонентов.



*Рис 2*. *Зависимость эффективного числа бит (ENOB) от размера ядра (core size) для архитектур Клементса (Clements) и кроссбар (crossbar).*

Таким образом, современные задачи в области машинного обучения и ИИ требуют значительных вычислительных мощностей и энергии, что делает проблему энергозатрат и масштабируемости особенно актуальной. Архитектуры, такие как SVD и Crossbar, показывают перспективы в умножении матриц, однако их эффективность ограничена потерями в системе, которые растут с увеличением размера матриц. Оптимизация компонентов фотонной интегральной схемы и оптимизация архитектуры необходимы для повышения производительности и точности вычислений.

Работа выполнена при поддержке Некоммерческого фонда развития науки и образования "Интеллект" (Договор пожертвования №03/03-2023/ФФ-ЛАБ от 30 марта 2023 г.)

**Литература**

1. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.
2. G. Giamougiannis et al., "Universal Linear Optics Revisited: New Perspectives for Neuromorphic Computing With Silicon Photonics," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 29, no. 2: Optical Computing, pp. 1-16, March-April 2023, Art no. 6200116, doi: 10.1109/JSTQE.2022.3228318.
3. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
4. Kester W. Understand SINAD, ENOB, SNR, THD, THD+ N, and SFDR so you don’t get lost in the noise floor //MT-003 tutorial. – 2009. – С. 01-08.
5. McMahon, P. L. (2023). The physics of optical computing. Nature Reviews Physics, 5(12). <https://doi.org/10.1038/s42254-023-00645-5>
6. S. Ning et al., "Photonic-Electronic Integrated Circuits for High-Performance Computing and AI Accelerators," in Journal of Lightwave Technology, vol. 42, no. 22, pp. 7834-7859, 15 Nov.15, 2024, doi: 10.1109/JLT.2024.3427716.