УДК 004.31
В. Васьков, С. Лупенко
(Тернопільський національний технічний університет імені Івана Пулюя)

ПЕРЕВАГИ ВИКОРИСТАННЯ ТЕНЗОРНОГО ПРОЦЕСОРА ДЛЯ РОБОТИ З НЕЙРОННИМИ МЕРЕЖАМИ

UDC 004.31
V. Vaskov, S. Lupenko
(Ternopil Ivan Puluj National Technical University, Ukraine)

BENEFITS OF USING TENSOR PROCESSOR TO WORK WITH NEURAL NETWORKS

Тензорний процесор (TPU) є інтегральною схемою специфічного застосування (ASIC), що призначений для прискорення розрахунків штучного інтелекту, й був розроблений компанією Google для машинного навчання нейронних мереж [1].

Свою назву процесори отримали від бібліотеки програмного забезпечення TensorFlow. Основне призначення TPU полягає в прискоренні алгоритмів штучного інтелекту.

TPU ASIC побудований на 28 нм процесі, працює на частоті 700 МГц і споживає 40 Вт під час роботи. TPU підключається до слоту через шину PCIe Gen3 x16, яка забезпечує ефективну пропускну здатність 12.5 GB/s.

В середньому тензорний процесор в 15-30 разів швидше здійснює обчислення, в порівнянні із традиційними серверними CPU і GPU. Продуктивність у розрахунку на ватт у TPU у 25-80 разів вища, ніж у центрального і графічного чіпів [2].

Програмованість була ще однією важливою метою дизайну для TPU. TPU не призначений для запуску тільки одного типу моделі нейронної мережі. Замість цього він розроблений таким чином, щоб бути достатньо гнучким для прискорення обчислень, необхідних для запуску багатьох різних моделей нейронних мереж.

Більшість сучасних CPU побудовані з використанням архітектури Reduced Instruction Set Computer (RISC). У RISC основна увага приділяється визначенню простих інструкцій (наприклад, завантаження, зберігання, додавання та множення), які зазвичай використовуються більшістю додатків, а потім виконують ці інструкції якомога швидше. Архітектура Complex Instruction Set Computer (CISC) була обрана як основа набору інструкцій TPU. Архітектура CISC фокусується на реалізації високорівневих інструкцій, які виконують більш складні завдання (такі як обчислення багаторазового множення і додавання) з кожною інструкцією.

TPU включає наступні обчислювальні ресурси:

– матричний множник (MXU): 65,536 8-бітових одиниць множення та додавання для операцій матриці;

– єдиний буфер (UB): 24 МБ SRAM, які працюють як регістри;

– активаційний блок (AU): функції активації [2].

Дизайн TPU є строго мінімальним і детермінованим, оскільки він повинен виконувати лише одне завдання за один раз: прогнозування нейронної мережі.

З TPU, можемо легко оцінити, скільки часу потрібно для запуску нейронної мережі та прогнозування. Це дозволяє працювати з максимальною пропускною спроможністю даного чіпа [1].

Література

1. Cloud Tensor Processing Units (TPUs) [Електронний ресурс]. – 2019. – Режим доступу до ресурсу: https://cloud.google.com/tpu/docs/tpus.
2. Sato K. An in-depth look at Google's first Tensor Processing Unit (TPU) [Електронний ресурс] / K. Sato, C. Young, D. Patterson // Google Cloud Blog. – 2017. – Режим доступу до ресурсу: https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu.