

ОГЛЯД ЕЛЕМЕНТНОЇ БАЗИ ДЛЯ РОЗРОБКИ АПАРАТНИХ ПРИСКОРЮВАЧІВ КЛАСИФІКАЦІЇ ТЕКСТІВ

Задачі класифікації є доволі поширеними в сучасному цифровому світі. Наприклад, набувають все більшої популярності різноманітні смарт-системи, які спрямовані на прийняття рішень виходячи з деякого набору вхідних даних. Перед прийняттям такого рішення ці дані підлягають аналізу, зокрема класифікації за певною ознакою. При досить високих обсягах вхідних даних стає критичним час їх обробки. Програмні рішення скорочення такого часу мають обмеження, пов'язані з устаткуванням. Тому пропонується способи скорочення часу на обробку даних, спрямованих на покращення програмних рішень, частково перенести на апаратну платформу і поєднати ці підходи в єдиній системі.

Це дослідження представлено на прикладі класифікації текстових даних. Процес класифікації текстів складається з двох основних етапів: формування простору ознак категорії та ідентифікації тексту. Кожен містить внутрішні компоненти. Таким чином, класифікація тексту є складним процесом, що вимагає значного часу, оскільки має велику кількість супутніх обчислень. У зв'язку з цим встановлено, що на кожному з етапів внутрішньої класифікації можна досягти певного скорочення часу виконання завдання.

В рамках даного дослідження на прикладі обробки та класифікації текстових даних розглянуто завдання прискорення обробки даних шляхом комплексного використання стандартного програмного забезпечення та програмовані логічні інтегральні схеми (ПЛІС). З метою обґрунтування вибору щодо елементної бази ПЛІС для реалізації прискорювача класифікації, є доцільним попереднє виконання огляду та аналізу сучасних апаратних платформ. Отже, розглянемо основні тенденції розвитку елементної бази, яку можна використовувати для апаратного прискорення процесів обробки інформації. На зміну універсальним обчисленням приходять гетерогенні: звичайні цифрові процесори (ЦП) доповнюються спеціалізованими процесорами або прискорювачами, що підвищують продуктивність та енергоефективність серверів при ресурсномістких робочих навантаженнях, які стають звичайним явищем у дата-центрі. В якості прискорювачів обчислень використовуються графічні процесори (GPU), програмовані користувачем вентиляльні матриці (FPGA) та інтегральні схеми спеціального призначення (ASIC). В ієрархії процесорів – від ЦП загального призначення до ASIC – існує компроміс між гнучкістю та ефективністю. Причому, як стверджують у Accenture [1], коли будь-яка конкретна програма

реалізується за допомогою спеціалізованих рішень, продуктивність може збільшитися на порядки.

GPU, FPGA та ASIC можуть використовуватися для прискорення задач віртуалізації та хмарних обчислень, 3D/2D-графіки, високопродуктивних обчислень (HPC), штучного інтелекту та машинного навчання, аналізу великих даних, віртуалізації мережевих функцій (NFV) та ін. Наприклад, їх можна з успіхом застосовувати у програмах машинного навчання, призначених для розпізнавання голосу та зображень чи аналізу тексту [2].

Комбінування універсальних процесорних ядр із спеціалізованими блоками обчислень за нейромережевими алгоритмами – актуальна тенденція. Спеціалізовані мікросхеми для систем штучного інтелекту – ще один новий ринок, що швидко розвивається [3].

Найпоширеніший тип апаратних прискорювачів – графічні процесори – часто використовуються для неграфічних завдань. Наприклад, система NVIDIA DGX-1 з графічним процесором Volta розроблена спеціально для підтримки програм штучного інтелекту, глибокого навчання та аналітики [4].

Більшість постачальників хмарних послуг, таких як Amazon, Microsoft та Google, пропонують своїм клієнтам використовувати GPU у хмарі. Це значно розширює коло компаній, які можуть дозволити собі виконувати розрахунки на GPU чи FPGA-прискорювачах у своїх проектах. Зокрема, прискорювачі на FPGA задіяні в системах різних типів, включаючи хмарні обчислення, що вбудовуються до платформ. На відміну від універсального та графічного процесорів, FPGA можна перепрограмувати відповідно до особливостей розв'язуваного завдання. Інші характерні риси FPGA – архітектура з паралельним виконанням безлічі векторних операцій, і навіть знижене енергоспоживання для одиницю обчислювальної потужності [5].

Intel впроваджує FPGA у свої сервери на базі Xeon, Microsoft розміщує FPGA на мережевих картах PCI Express у серверах, які встановлює у своїх системах цифрової обробки даних. Приклад збільшення обчислювальної потужності серверної системи за рахунок використання FPGA – анонсований Intel FPGA-прискорювач N3000, призначений для обробки мережевого трафіку з пропускнуою здатністю до 100 Гбіт/с. Він містить 9 Гбайт пам'яті DDR4, 144 Мбайт пам'яті QDR VI і два мережеві адаптери з підтримкою восьми 10-гігабітних та чотирьох 2,5-гігабітних з'єднань. N3000 може застосовуватись для задач віртуалізації мережевих функцій, а також в інфраструктурі мереж 5G [6].

Для Microsoft Azure та завдань глибокого навчання призначена система Microsoft Brainwave на основі FPGA. Як повідомляється, алгоритми машинного навчання Microsoft Bing на FPGA прискорюються у 40–100 разів [7].

В області FPGA свої рішення представили такі гіганти, як Amazon, Microsoft та Facebook: AWS Inferentia, Project Brainwave та Kings Canyon. Методи розробки додатків, управління ресурсами та масштабування систем на базі FPGA удосконалюються на даний момент та продовжуватимуть це робити і надалі.

Таким чином, зважаючи на розширення областей застосування FPGA для вирішення різних розрахункових задач, його використання для класифікації текстів є перспективним та дозволяє реалізовувати розв'язання цієї задачі різними алгоритмічними рішеннями. Спрощення процесів реалізації складних алгоритмів обробки текстів, які зараз зустрічаються здебільшого в програмному варіанті, на апаратній платформі дозволяє не тільки частково перенести функції програмного коду на кристал, а й розробити спеціалізоване апаратне рішення, здатне надати значні показники прискорення процесу класифікації із зменшенням вимог до енергоспоживання, що є приємним додатковим фактором.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Офіційний сайт компанії Accenture [Електронний ресурс]. Режим доступу: <https://www.accenture.com/pl-en> (Date of reference: 08.04.2024)
2. Wang Yujie Artificial-Intelligence integrated circuits: Comparison of GPU, FPGA and ASIC // Yujie Wang / Applied and Computational Engineering – 4(1), 2023 – P. 99-104. DOI:10.54254/2755-2721/4/20230358
3. FPGA vs. GPU vs. CPU – hardware options for AI applications [Electronic resource]. Access mode: <https://my.avnet.com/silica/resources/article/fpga-vs-gpu-vs-cpu-hardware-options-for-ai-applications/> (Date of reference: 08.04.2024)
4. CPU vs GPU vs TPU vs DPU vs QPU vs ASICs vs FPGA: Navigating the Labyrinth of Processing Units [Electronic resource]. Access mode: <https://www.linkedin.com/pulse/cpu-vs-gpu-tpu-dpu-qpu-ultimate-showdown-processing-units-dragutoiu> (Date of reference: 08.04.2024)
5. Amazon EC2 F1 Instances. Enable faster FPGA accelerator development and deployment in the cloud [Electronic resource]. Access mode: https://aws.amazon.com/ec2/instance-types/f1/?nc1=h_ls (Date of reference: 08.04.2024)
6. Intel® FPGA PAC N3000 [Electronic resource]. Access mode: <https://www.intel.com/content/www/us/en/products/sku/193920/intel-fpga-pac-n3000/specifications.html> (Date of reference: 08.04.2024)
7. Project Brainwave [Electronic resource]. Access mode: <https://www.microsoft.com/en-us/research/project/project-brainwave/> (Date of reference: 08.04.2024)