Алгоритм множення розріджених матриць на графічному процесорі

Матриці

$$A^R = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{mn} \end{bmatrix}$$

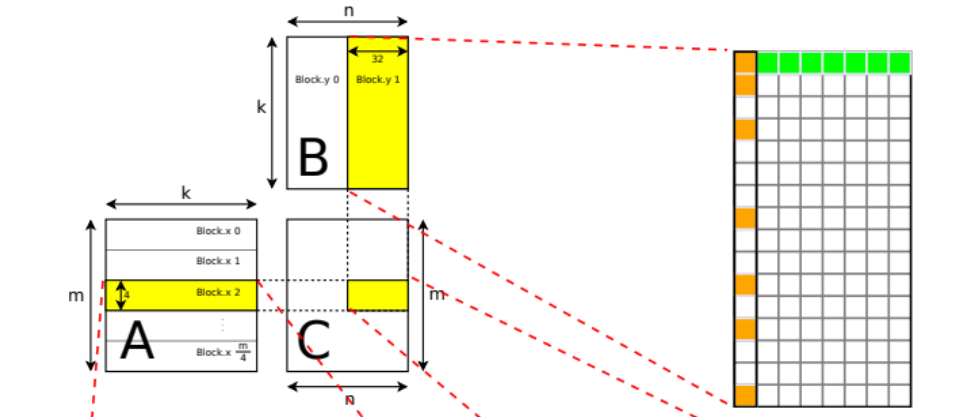
Множення матриць

$$c_{ij} = \sum_{r=1}^{1} a_{ir} b_{rj} \ (i = 1,2, \dots m; j = 1,2, \dots q).$$

Типи зберігання матриць

- Стиснене зберігання стрічкою(англ. Compressed sparse row – CSR, Compressed row storage - CRS, Yale format - Єльський формат) – досить складний спосіб зберігання розрідженої матриці. Саме цей формат зберігання використовується в цій роботі. Принцип такого зберігання полягає в тому, що ми використовуємо не одну структуру даних, а декілька(три, якщо точніше), тобто ми розбиваємо розріджену матрицю $Mn \times m$ три наступних масиви:
- a) Масив значень(array[value]) – масив розміру N, який зберігає в собі не нульові значення елементів матриці, які були взяті підряд із першої не нульової стрічки, потім із другої ненульової стрічки, потім із третьої і так далі;
- b) Масив індексів стовпчиків(array[columns]) – масив розміру N, який зберігає в собі номери стовпчиків, що відповідають номерам стовпчиків елементів розрідженої матриці, які зберігаються в масиві значень;
- c) Масив індексів рядків(array[rows]) – масив розміру N + 1(кількість рядків + 1). Для кожного індекса і зберігається кількість ненульових елементів в стрічках з першої до і – 1 стрічки включно. Також варто відмітити, що перший елемент масиву рядків завжди рівний нулю.

Множення розріджених матриць



(a) 2D block assignment of proposed row split SpMM

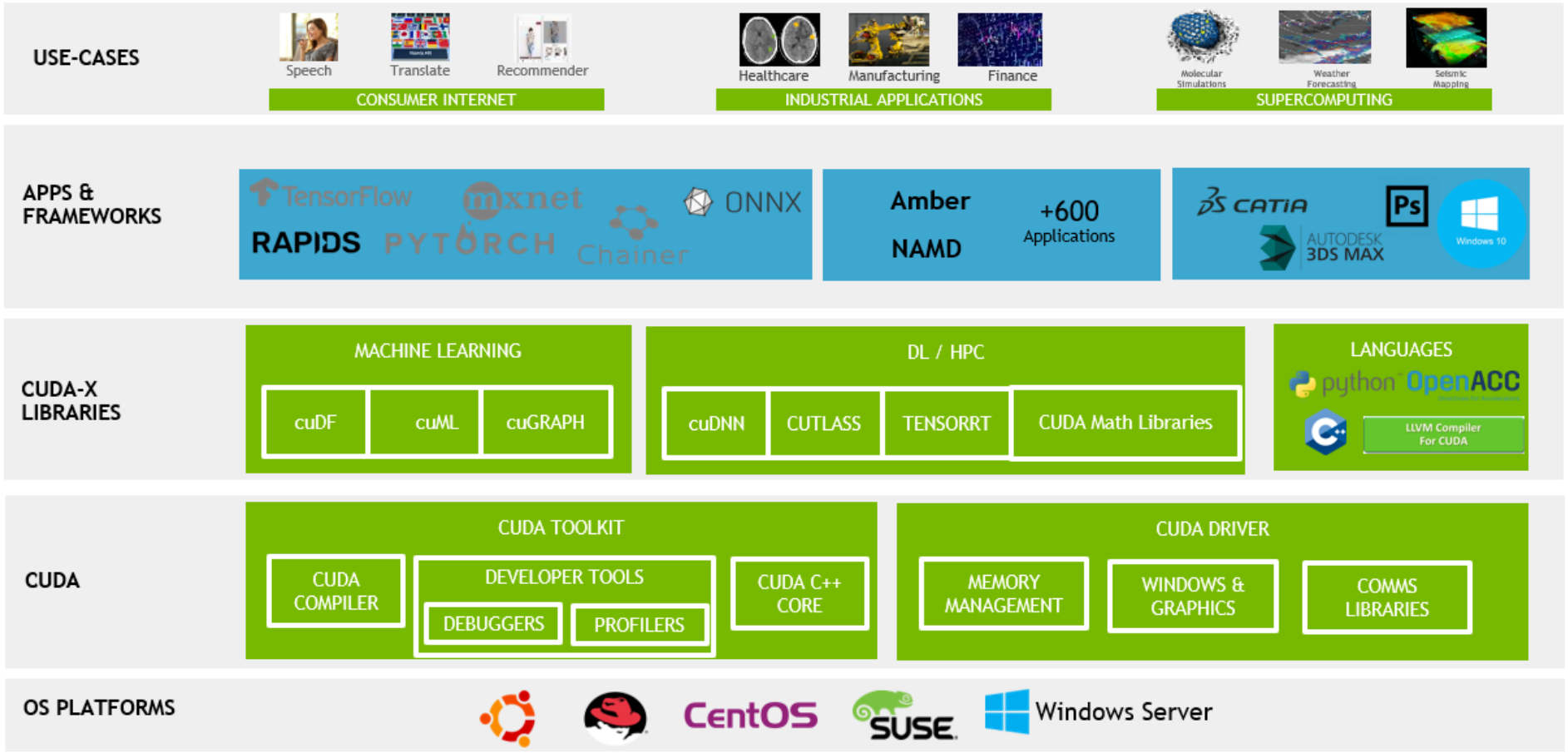
(b) Step 1: Load sparse matrix **A** column indices into registers. Zeroes are marked as such.

(c) Step 2: Load dense matrix **B** values into registers.

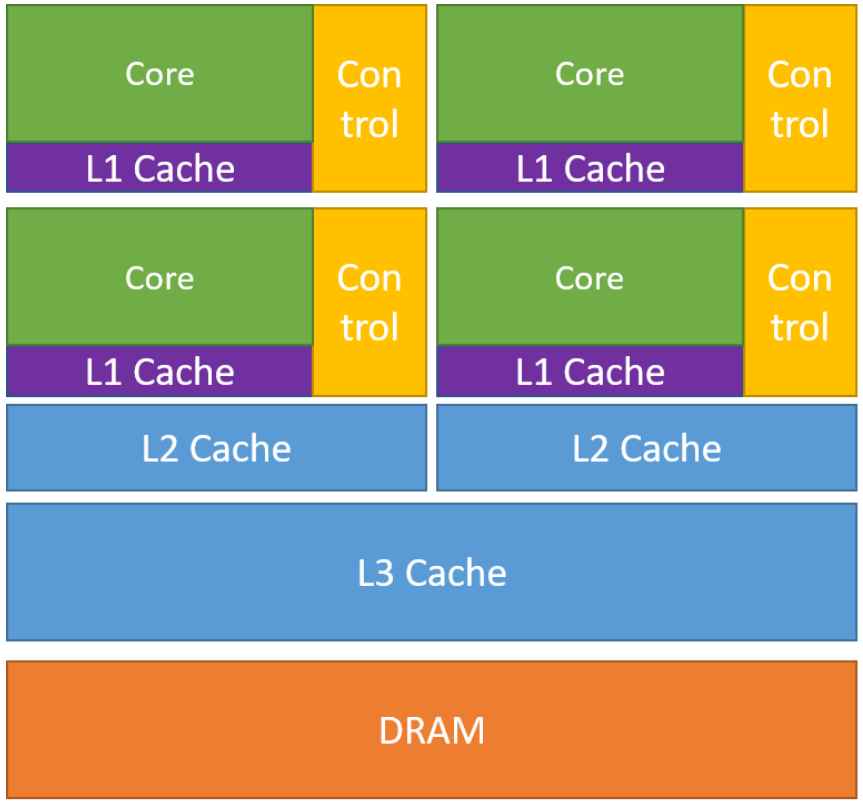
(d) Step 3: Write to dense matrix **C**.

Що таке CUDA

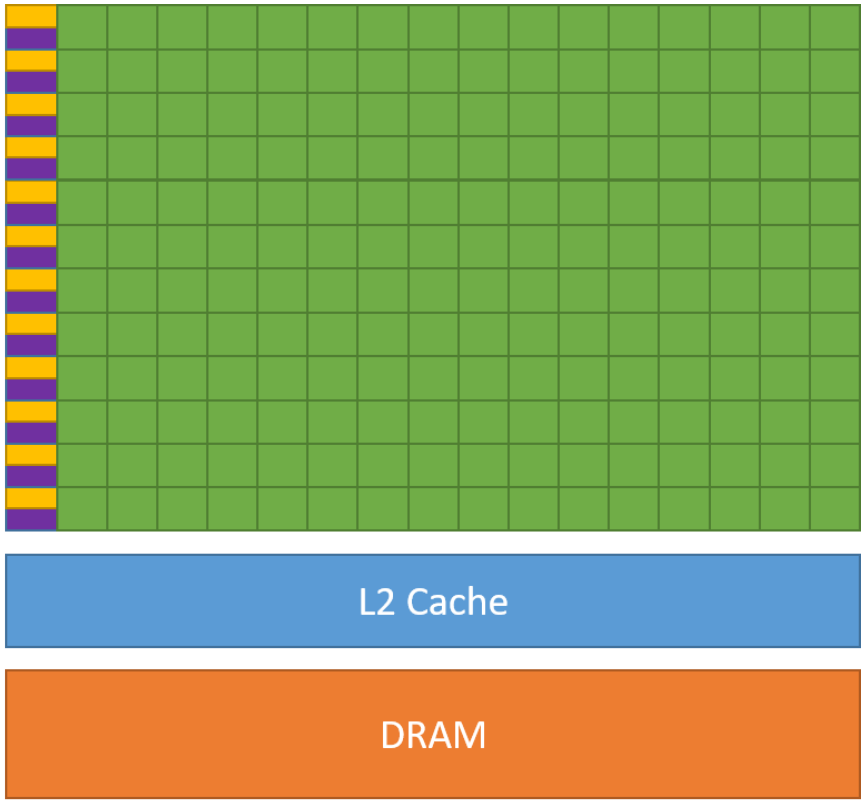
Структура CUDA



Порівняння CPU та GPU



CPU



GPU

Алгоритм множення розріджених матриць

- Перший алгоритм називається алгоритмом множення розрідженої матриці з розбиттям рядків(Row-splitting SpMM). Основний принцип полягає у призначенні кожного рядка різному потоці. Даний алгоритм множення розрідженої матриці є одним з найбільш поширених і найбільш простих алгоритмів множення розріджених матриць на графічному процесорі. Проте даний підхід має свої певні недоліки, але дана робота не про них.
- Наступний алгоритм являється алгоритмом множення розріджених матриць на основі злиття(Merge-based SpMM). Суть алгоритму на основі злиття полягає в явному і рівномірному розподіленні ненульових значень елементів між паралельними процесорами. Сам процес реалізується таким чином, що виконання виконується на основі двофазового розкладання: на першому етапі (PartitionSpmm) алгоритм злиття розподіляє роботу між потоками таким чином, що T роботи призначається кожному потоку, і на основі цього розподілу виводить початкові індекси кожного елементу. Після виконання відповідної координації, робота виконується на другому етапі. Виходячи з цієї реалізації, в алгоритмі реалізовуються наступні дизайнерські підходи: підхід в доступі до пам’яті, використання регістрів та накладні витрати на доступ до пам’яті.

Дякую за увагу