Ответы на вопросы

1. Попытайтесь дать асимптотическую оценку времени выполнения в зависимости от объёма кэша и пропускной способности шины. Есть ли способы ускорить вычисления в условиях ограничений?

Оценку временной сложности алгоритма предлагается вычислить по ниженаписанному псевдокоду. Обозначения:

- CACHE – размер КЭШ в числах (в алгоритме CACHE = 2049 чисел);
- SPEED – скорость загрузки данных из RAM относительно скорости выполнения операций матричного умножения с накоплением над блоками из КЭШ и обращения к КЭШ (исчисляется в разах);
- A_SIZE – размер матриц (не квадратная матрица будет превращена в квадратную путём добавления нулевых строк или столбов);
- BLOCK_NUM – количество блоков подстроки и подстолбца при передачи из RAM в КЭШ
- MAC_NUM – количество вычислителей;
- MAC_INPUTS_NUM – количество одновременно обрабатываемых чисел одним вычислителем.

```
# O(A_SIZE * A_SIZE * (A_SIZE / SPEED + 1 / SPEED))
for row_idx in range(A_SIZE):
    # O(A_SIZE * (A_SIZE / SPEED + 1 / SPEED))
    for col_idx in range(A_SIZE):
        # O(A_SIZE / SPEED)
        for subcol_idx in range(A_SIZE / CACHE):
            # O(CACHE / SPEED)
            for block_idx in range(BLOCK_NUM):
                # O(CACHE / (SPEED * BLOCK_NUM))
                # передача блока RAM -> КЭШ
                # и подсчёт умножения с накоплением
        # O(1 / SPEED)
        # передача результата из КЭШ в RAM
```

Примечания:

а) Истинно предположение о том, что скорость и количество вычислителей позволяют утверждать, что матричное умножение блоков подстроки и подстолбца происходит много быстрее, чем копирование аналогичного по размерам блока данных из RAM в КЭШ;

б) Истинно предположение о том, что вычислители работают параллельно с копированием данных из RAM в КЭШ;

в) Истинно предположение о том, что скорость передачи КЭШ -> RAM равна скорости передачи RAM -> КЭШ.

Оценка времени выполнения равна $O(A\_SIZE^2 * (A\_SIZE / SPEED + 1 / SPEED))$. Из предыдущего выражения следует, что:

- сложность данного алгоритма от количества элементов равна $O(A\_SIZE^3)$;
- сложность данного алгоритма от пропускной способности шины равна $O(1 / SPEED)$;
- сложность данного алгоритма не зависит от размера КЭШ.

Способы ускорить вычисления при наличии ограничений:

- производить загрузку данных RAM -> КЭШ и вычислять матричное умножение блоков данных из КЭШ параллельно;
- сохранять промежуточный накапливаемый результат матричного умножения подстрок и подстолбцов в КЭШ и посылать в RAM только конечный результат перемножения строки на столбец.

2. Какого оптимальное соотношение объёма КЭШ и массива вычислителей, при условии ограничения скорости доступа в RAM?

В процессе разработки алгоритма было установлено, что для предотвращения ошибок вида “index out of range” и ошибки затирания необработанных данных в КЭШ необходимо выполнения следующего неравенства:

$$CACHE \leq 1 + 2 * SPEED * MAC\_NUM * MAC\_INPUTS\_NUM.$$

Выделение дополнительной ячейки памяти позволяет хранить промежуточный результат вычислений. Двойной размер КЭШ позволяет производить загрузку и вычисления параллельно.

Оптимальное соотношения объёма КЭШ, массива вычислителей и скорости доступа в RAM достигается обращением вышеуказанного неравенства в равенство.

Оптимальное соотношение объёма КЭШ и массива вычислителей примерно равно $2 * SPEED * MAC\_INPUTS\_NUM$.

3. Какого оптимальное соотношение объёма КЭШ и скорости доступа в RAM, при условии ограничения размера массива MAC?

Оптимальное соотношение объёма КЭШ и скорости доступа в RAM примерно равно $2 * MAC\_NUM * MAC\_INPUTS\_NUM$.