CUDA: Memory model

План семинара

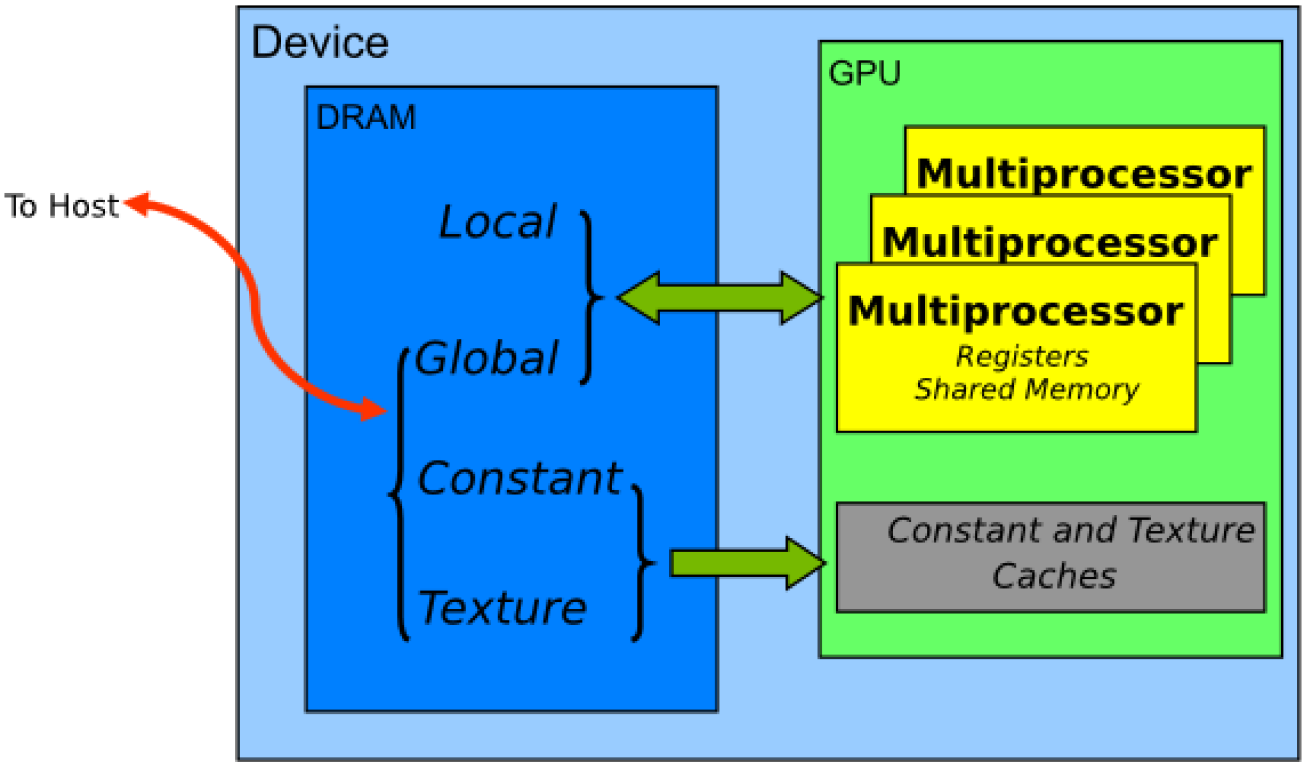
1. Память GPU
2. Global память
3. Shared память
4. Умножение матриц

Код этого семинара

Global и Shared память

Умножение матриц

Память GPU



Подробно:

https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html#device-memory-spaces

Принцип обращения к global памяти

Доступ к памяти должен быть слитым (**coalesced**).

Для CUDA CC 6.0+ это означает следующее:

> Все (параллельные) запросы доступа к глобальной памяти в данной инструкции будут объединены в такое число **32-байтных** транзакций, чтобы выполнить все запросы.

В предыдущих версиях CUDA CC размер транзакции бывал другим.
Подробнее:

https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html#coalesced-access-to-global-memory

Принцип обращения к global памяти

Пропускная способность в зависимости от **смещения**
при обращении из нескольких threads за **одно исполнение** kernel

Подробнее:

https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html#strided-accesses

Операция с векторами

```
__global__
void add(int n, float* x, float* y, float* z) {
    int tid = threadIdx.x + ILP * blockDim.x * blockIdx.x;
    for (int i = 0; i < ILP; ++i) {
        int current_tid = tid + i * blockDim.x;
        z[current_tid] = 2.0f * x[current_tid] + y[current_tid];
    }
}

__global__
void stupid_add(int n, float* x, float* y, float* z) {
    int index = blockIdx.x * blockDim.x + threadIdx.x;
    int tid = ILP * index;
    for (int i = 0; i < ILP; ++i) {
        int current_tid = tid + i;
        z[current_tid] = 2.0f * x[current_tid] + y[current_tid];
    }
}
```

Shared память

Для каждого **блока** доступна
особо быстрая (но небольшая) shared память.

Её суммарный размер зависит от конкретного device.

Декларация **внутри Kernel**:

```
__shared__ int N[42];
```

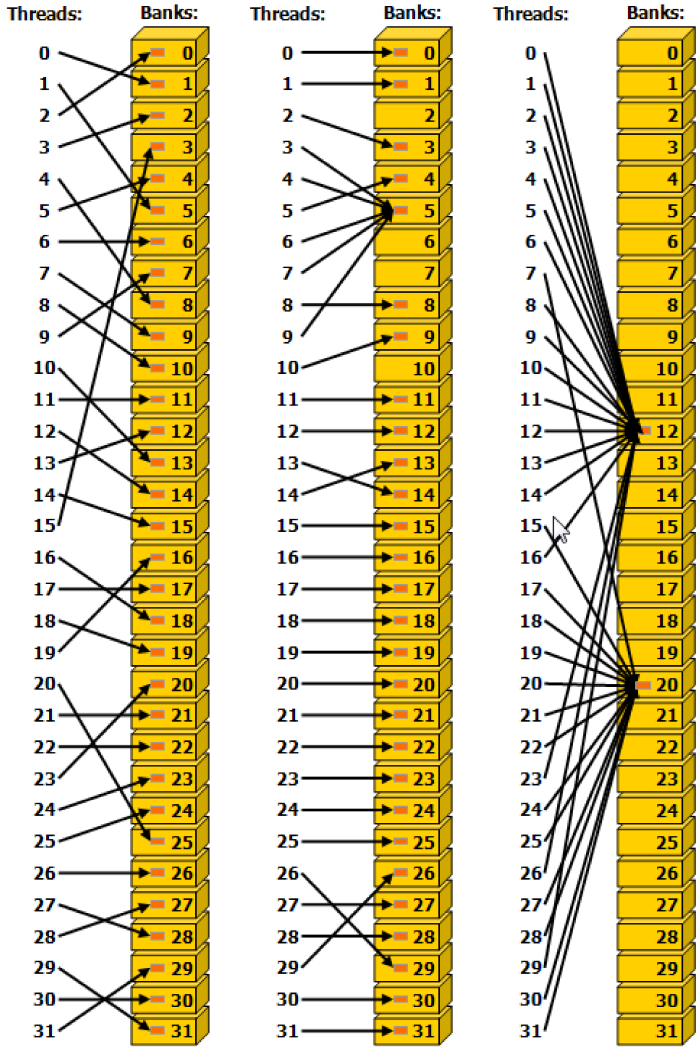
Доступ должен быть без гонок.

Пропускная способность одной ячейки shared памяти —
32 бита за инструкцию.

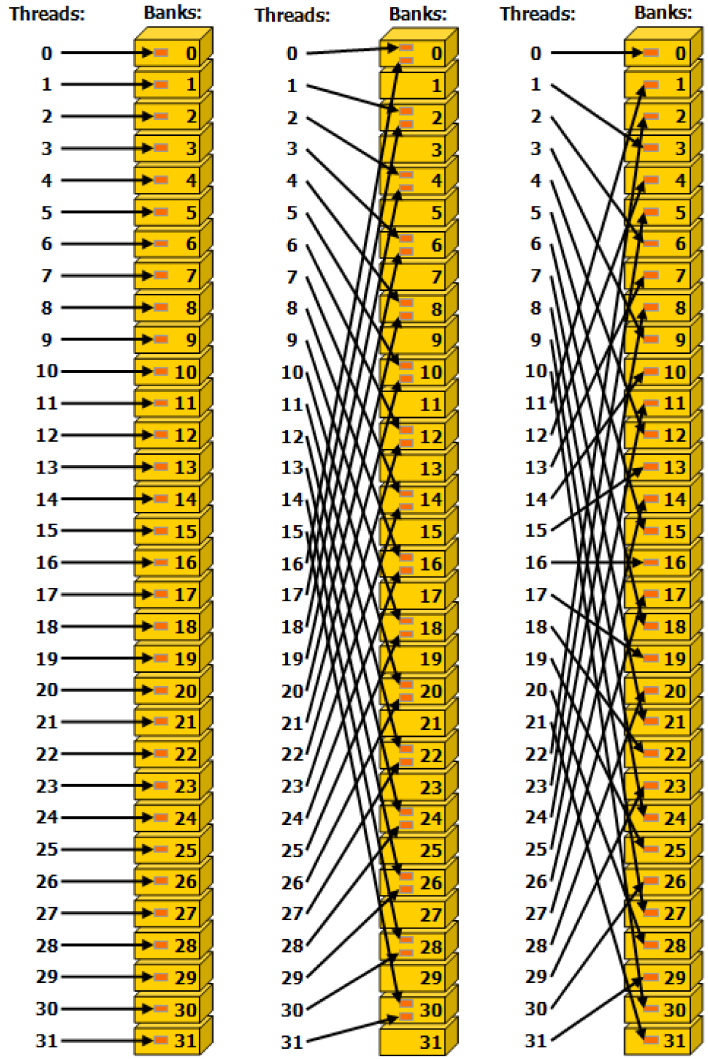
Подробно:

https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html#shared-memory

Принцип обращения к shared памяти



Плохо



Хорошо

Умножение матриц

https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html#shared-memory