GPU Memories
Traditional Program Structure

- CPU threads and GPU kernels
  - Sequential or modestly parallel sections on CPU
  - Massively parallel sections on GPU

Serial Code (host)

Parallel Kernel (device)
KernelA<<< nBlk, nThr >>>(args);

Serial Code (host)

Parallel Kernel (device)
KernelB<<< nBlk, nThr >>>(args);
Memory Hierarchy in CUDA Programs

Grid (Device)

Block (0, 0)
- Shared memory
- Registers
  - Thread (0, 0)
  - Thread (1, 0)

Block (1, 0)
- Shared memory
- Registers
  - Thread (0, 0)
  - Thread (1, 0)

Global / Texture & Surface memory

Constant memory

Host
Tiled Matrix Multiplication (II)

- **Tiled implementation** operates on submatrices (tiles or blocks) that fit fast memories (cache, scratchpad, RF)

```c
#define A(i,j) matrix_A[i * P + j]
#define B(i,j) matrix_B[i * N + j]
#define C(i,j) matrix_C[i * N + j]

for (I = 0; I < M; I += tile_dim){
  for (J = 0; J < N; J += tile_dim){
    Set_to_zero(&C(I, J)); // Set to zero
    for (K = 0; K < P; K += tile_dim)
      Multiply_tiles(&C(I, J), &A(I, K), &B(K, J));
  }
}
```

Multiply small submatrices (tiles or blocks) of size tile_dim x tile_dim

---


Kirk & Hwu, "Chapter 5 - Performance considerations," in "Programming Massively Parallel Processors (Third Edition)", 2017. [DOI](https://doi.org/10.1016/B978-0-12-811986-0.00005-4)
Tiled Matrix-Matrix Multiplication (V)

__shared__ float A_s[TILE_DIM][TILE_DIM];
__shared__ float B_s[TILE_DIM][TILE_DIM];

unsigned int row = blockIdx.y * blockDim.y + threadIdx.y;
unsigned int col = blockIdx.x * blockDim.x + threadIdx.x;

float sum = 0.0f;

for(unsigned int tile = 0; tile < N/TILE_DIM; ++tile) {

    // Load tile to shared memory
    A_s[threadIdx.y][threadIdx.x] = A[row*N + tile*TILE_DIM + threadIdx.x];
    B_s[threadIdx.y][threadIdx.x] = B[(tile*TILE_DIM + threadIdx.y)*N + col];

    __syncthreads();

    // Compute with tile
    for(unsigned int i = 0; i < TILE_DIM; ++i) {
        sum += A_s[threadIdx.y][i]*B_s[i][threadIdx.x];
    }

    __syncthreads();

}

C[row*N + col] = sum;
Performance Considerations
Performance Considerations

- Main bottlenecks
  - CPU-GPU data transfers
  - Global memory access

- Memory access
  - Latency hiding
    - Occupancy
  - Memory coalescing
  - Data reuse
    - Shared memory usage

- SIMD (Warp) Utilization: Divergence

- Other considerations
  - Atomic operations: Serialization
  - Data transfers between CPU and GPU
    - Overlap of communication and computation
Memory Access
Latency Hiding via Warp-Level FGMT

- **Warp**: A set of threads that execute the same instruction (on different data elements)

- **Fine-grained multithreading**
  - One instruction per thread in pipeline at a time (No interlocking)
  - Interleave warp execution to hide latencies

- Register values of all threads stay in register file

- FGMT enables long latency tolerance
  - Millions of pixels

---

Slide credit: Tor Aamodt
Latency Hiding and Occupancy

- **FGMT** can hide long latency operations (e.g., memory accesses)
- **Occupancy**: ratio of active warps to the maximum number of warps per GPU core
Occupancy

- GPU core, a.k.a. SM, resources (typical values)
  - Maximum number of warps per SM (64)
  - Maximum number of blocks per SM (32)
  - Register usage (256KB)
  - Shared memory usage (64KB)

- Occupancy calculation
  - Number of threads per block (defined by the programmer)
  - Registers per thread (known at compile time)
  - Shared memory per block (defined by the programmer)
CUDA Occupancy Calculator

Click Here for detailed instructions on how to use this occupancy calculator.
For more information on NVIDIA CUDA, visit http://developer.nvidia.com/cuda

Your chosen resource usage is indicated by the red triangle on the graphs. The other data points represent the range of possible block sizes, register counts, and shared memory allocation.

---

### CUDA Occupancy Calculator

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V |
| 1 | CUDA Occupancy Calculator | | | | | | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | | | | | | | | |
| 3 | Just follow steps 1, 2, and 3 below (or click here for help) | | | | | | | | | | | | | | | | | | | | | |
| 4 | 1) Select Compute Capability (click) | $8.6$ | | | | | | | | | | | | | | | | | | | | |
| 5 | 2) Select Shared Memory Size Config (bytes) | $6536$ | | | | | | | | | | | | | | | | | | | | |
| 6 | 3) Select CUDA version | $11.1$ | | | | | | | | | | | | | | | | | | | | |
| 7 | 8) Enter your resource usage: | | | | | | | | | | | | | | | | | | | | |
| 8 | Threads Per Block | $256$ | | | | | | | | | | | | | | | | | | | | |
| 9 | Registers Per Thread | $32$ | | | | | | | | | | | | | | | | | | | | |
| 10 | Local Shared Memory Per Block (bytes) | $256$ | | | | | | | | | | | | | | | | | | | | |
| 11 | (Don’t edit anything below this line) | | | | | | | | | | | | | | | | | | | | |
| 12 | 1) GPU Occupancy Data is displayed here and in the graphs: | | | | | | | | | | | | | | | | | | | | |
| 13 | Active Threads per Multiprocessor | $1536$ | | | | | | | | | | | | | | | | | | | | |
| 14 | Active Wares per Multiprocessor | $46$ | | | | | | | | | | | | | | | | | | | | |
| 15 | Active Thread Blocks per Multiprocessor | $6$ | | | | | | | | | | | | | | | | | | | | |
| 16 | Occupancy of Each Multiprocessor | $100\%$ | | | | | | | | | | | | | | | | | | | | |
| 17 | 25 | Physical Limits for GPU Compute Capability: | $8.6$ | | | | | | | | | | | | | | | | | | | | |
| 18 | Threads per Warp | $32$ | | | | | | | | | | | | | | | | | | | | |
| 19 | Max Wares per Multiprocessor | $46$ | | | | | | | | | | | | | | | | | | | | |
| 20 | Max Thread Blocks per Multiprocessor | $16$ | | | | | | | | | | | | | | | | | | | | |
| 21 | Max Threads per Multiprocessor | $1536$ | | | | | | | | | | | | | | | | | | | | |
| 22 | Maximum Thread Block Size | $1024$ | | | | | | | | | | | | | | | | | | | | |
| 23 | Registers per Multiprocessor | $6536$ | | | | | | | | | | | | | | | | | | | | |
| 24 | Max Registers per Thread Block | $6536$ | | | | | | | | | | | | | | | | | | | | |
| 25 | Max Registers per Thread | $256$ | | | | | | | | | | | | | | | | | | | | |
| 26 | Shared Memory per Multiprocessor (bytes) | $6536$ | | | | | | | | | | | | | | | | | | | | |
| 27 | Shared Memory per Block | $6536$ | | | | | | | | | | | | | | | | | | | | |
| 28 | Register allocation unit size | $32$ | | | | | | | | | | | | | | | | | | | | |
| 29 | Register allocation granularity | warp | | | | | | | | | | | | | | | | | | | | |
| 30 | Shared Memory allocation unit size | $256$ | | | | | | | | | | | | | | | | | | | | |
| 31 | Warp allocation granularity | $128$ | | | | | | | | | | | | | | | | | | | | |
| 32 | Shared Memory Per Block (bytes) (CUDA runtime use) | $1024$ | | | | | | | | | | | | | | | | | | | | |
| 33 | Allocated Resources | Per Block | Limit Per SM | Limit of Blocks Per SM | | | | | | | | | | | | | | | | | | | | |
| 34 | Wares | $9$ | $48$ | $9$ | | | | | | | | | | | | | | | | | | | | |
| 35 | Registers | $54$ | $64$ | $8$ | | | | | | | | | | | | | | | | | | | | |
| 36 | Shared Memory (Bytes) | $2592$ | $65536$ | $32$ | | | | | | | | | | | | | | | | | | | | |
| 37 | Max. SM is an abbreviation for Maximum Multiprocessor |
| 38 | Max. Wares/SM = $46$ |
| 39 | Max. Thread Blocks/SM = $16$ |
| 40 | Occupancy = $48/48 = 100\%$ |
| 41 | Maximum Thread Blocks Per Multiprocessor Blocks/SM * Wares/Block = Wares/SM |
| 42 | Limited by Max. Wares or Max. Blocks per Multiprocessor |
| 43 | Limited by Register per Multiprocessor |
| 44 | Limited by Shared Memory per Multiprocessor |
| 45 | Physical Max Wares/SM = $48$ |
| 46 | CUDA Occupancy Calculator |
| 47 | Version: | $11.1$ | | | | | | | | | | | | | | | | | | | | |

---

[Deprecated] CUDA Occupancy Calculator

The CUDA Occupancy Calculator allows you to compute the multiprocessor occupancy of a GPU by a given CUDA kernel.

[Deprecated] Excel based Occupancy Calculator is deprecated. Occupancy calculator is available in Nsight Compute. Please refer to Nsight Compute Occupancy Calculator documentation for more details on usage.

Overview

The CUDA Occupancy Calculator allows you to compute the multiprocessor occupancy of a GPU by a given CUDA kernel. The multiprocessor occupancy is the ratio of active warps to the maximum number of warps supported on a multiprocessor of the GPU. Each multiprocessor on the device has a set of N registers available for use by CUDA program threads. These registers are a shared resource that are allocated among the thread blocks executing on a multiprocessor.

The CUDA compiler attempts to minimize register usage to maximize the number of thread blocks that can be active in the machine simultaneously. If a program tries to launch a kernel for which the registers used per thread times the thread block size is greater than N, the launch will fail.

Click CUDA Occupancy Calculator[XLS] to download the spreadsheet.
CUDA Occupancy Calculator (III)

9. Occupancy Calculator

NVIDIA Nsight Compute provides an Occupancy Calculator that allows you to compute the multiprocessor occupancy of a GPU for a given CUDA kernel. It offers feature parity to the CUDA Occupancy Calculator spreadsheet.

The Occupancy Calculator can be opened directly from a profile report or as a new activity. The occupancy calculator data can be saved to a file using File > Save. By default, the file uses the .ncu-occ extension. The occupancy calculator file can be opened using File > Open File.

1. Launching from the Connection Dialog

Select the Occupancy Calculator activity from the connection dialog. You can optionally specify an occupancy calculator data file, which is used to initialize the calculator with the data from the saved file. Click the Launch button to open the Occupancy Calculator.

https://docs.nvidia.com/nsight-compute/NsightCompute/index.html#occupancy-calculator
Memory Layout of a Matrix in C

\[
\begin{array}{cccc}
M_{0,0} & M_{1,0} & M_{2,0} & M_{3,0} \\
M_{0,1} & M_{1,1} & M_{2,1} & M_{3,1} \\
M_{0,2} & M_{1,2} & M_{2,2} & M_{3,2} \\
M_{0,3} & M_{1,3} & M_{2,3} & M_{3,3} \\
\end{array}
\]
The DRAM Subsystem
The Top-Down View
DRAM Subsystem Organization

- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column
The DRAM Subsystem

“Channel”

DIMM (Dual in-line memory module)

Processor

Memory channel

Memory channel
Breaking down a DIMM (module)

DIMM (Dual in-line memory module)

- **Side view**
- **Front of DIMM**
- **Back of DIMM**

**Rank 0:** collection of 8 chips

**Rank 1**
Breaking down a Rank

Rank 0

Chip 0

Chip 1

... Chip 7

Data <0:63>

<0:7> <8:15> <56:63>
Breaking down a Chip

Chip 0

8 banks

<0:7>
Inside a DRAM Chip

- Subarray (2D Array of DRAM Cells)
- Sense Amplifiers
- Row Buffer
- DRAM Bank
- DRAM Chips
- DRAM Module
- Bitline
- Wordline
- Access Transistor
- Storage Capacitor
- DRAM Cells
DRAM Cell Operation

1. ACTIVATE (ACT)
2. READ/WRITE
3. PRECHARGE (PRE)
DRAM Cell Operation - ACTIVATE

1. Raise wordline
2. Capacitor charge is restored charge with bitline
3. Enable sense amplifier
4. Amplify deviation in the bitline
5. $\frac{1}{2}V_{DD} + \delta$
6. Row buffer stores the cell value
DRAM Cell Operation – READ/WRITE

1. ACTIVATE (ACT)
2. READ/WRITE
3. PRECHARGE (PRE)

Read/Write the value latched in sense amplifier
1. Lower wordline
2. Precharge bitline for next access
3. Disable sense amplifier

1. ACTIVATE (ACT)
2. READ/WRITE
3. PRECHARGE (PRE)
Access Address:
(Row 0, Column 0)
(Row 0, Column 1)
(Row 0, Column 85)
(Row 1, Column 0)

Row address 0 -> Row decoder -> Rows -> Column mux -> Data

Columns

Row buffer

CONFLICT!
DRAM Burst

- Accessing data in different bursts (rows)
  - Need to access the array again

  Timeline: [Diagram showing different burst access times]

- Accessing data in the same burst (row)
  - No need to access the array again, just the multiplexer

  Timeline: [Diagram showing faster access times]

- Accessing data in the same burst is faster than accessing data in different bursts
Recall: Memory Banking

- Memory is divided into banks that can be accessed independently; banks share address and data buses (to minimize pin cost)
- Can start and complete one bank access per cycle
- Can sustain N concurrent accesses if all N go to different banks

![Memory Banking Diagram]

Picture credit: Derek Chiou
Multiple Banks (Interleaving) and Channels

- Multiple banks
  - Enable concurrent DRAM accesses
  - Bits in address determine which bank an address resides in

- Multiple independent channels serve the same purpose
  - But they are even better because they have separate data buses
  - Increased bus bandwidth

- Enabling more concurrency requires reducing
  - Bank conflicts
  - Channel conflicts

- How to select/randomize bank/channel indices in address?
  - Lower order bits have more entropy
  - Randomizing hash functions (XOR of different address bits)
Latency Hiding with Multiple Banks

- With one bank, time still wasted in between bursts

- Latency can be hidden by having multiple banks

- Need many threads to simultaneously access memory to keep all banks busy
  - Achieved with having high occupancy in GPU cores (SMs)
    - Similar idea to hiding pipeline latency in the core

Slide credit: Izzat El Hajj
Memory Coalescing (I)

- When threads in the same warp access consecutive memory locations in the same burst, the accesses can be combined and served by one burst
  - One DRAM transaction is needed
  - Known as memory coalescing

- If threads in the same warp access locations not in the same burst, accesses cannot be combined
  - Multiple transactions are needed
  - Takes longer to service data to the warp
  - Sometimes called memory divergence
When accessing global memory, we want to make sure that concurrent threads access nearby memory locations.

Peak bandwidth utilization occurs when all threads in a warp access one cache line (or several consecutive cache lines).

---

Slide credit: Hwu & Kirk
Uncoalesced Memory Accesses

Access direction in Kernel code

Time Period 1

Time Period 2

Slide credit: Hwu & Kirk
Coalesced Memory Accesses

Access direction in Kernel code

Time Period 1
\[ T_1 \quad T_2 \quad T_3 \quad T_4 \]

Time Period 2
\[ T_1 \quad T_2 \quad T_3 \quad T_4 \]

...
AoS vs. SoA

- Array of Structures vs. Structure of Arrays

Structure of Arrays (SoA)

```
struct foo{
    float a[8];
    float b[8];
    float c[8];
    int d[8];
} A;
```

Array of Structures (AoS)

```
struct foo{
    float a;
    float b;
    float c;
    int d;
} A[8];
```
CPUs Prefer AoS, GPUs Prefer SoA

- Linear and strided accesses

![Graph of Throughput vs Stride for GPU and CPU](image)

**AMD Kaveri A10-7850K**

Sung+, “DL: A data layout transformation system for heterogeneous computing,” INPAR 2012

Use Shared Memory to Improve Coalescing

Original Access Pattern

Tiled Access Pattern

Copy into scratchpad memory

Perform multiplication with scratchpad values

Slide credit: Hwu & Kirk
Data Reuse

- Same memory locations accessed by neighboring threads

```c
for (int i = 0; i < 3; i++){
    for (int j = 0; j < 3; j++){
        sum += gauss[i][j] * Image[(i+row-1)*width + (j+col-1)];
    }
}
```
To take advantage of data reuse, we divide the input into tiles that can be loaded into shared memory.

```c
__shared__ int l_data[(L_SIZE+2)*(L_SIZE+2)];
...
Load tile into shared memory
__syncthreads();
for (int i = 0; i < 3; i++){
    for (int j = 0; j < 3; j++){
        sum += gauss[i][j] * l_data[(i+l_row-1)*(L_SIZE+2)+j+l_col-1];
    }
}
```
Shared Memory

- Shared memory is an **interleaved (banked) memory**
  - Each bank can service one address per cycle

- Typically, 32 banks in NVIDIA GPUs
  - Successive 32-bit words are assigned to successive banks
    - Bank = Address % 32

- Bank conflicts are **only possible within a warp**
  - No bank conflicts between different warps
Shared Memory Bank Conflicts (I)

- **Bank conflict free**

  ![Diagram of bank and thread connections]

  **Linear addressing**: stride = 1

  **Random addressing 1:1**

  Slide credit: Hwu & Kirk
N-way bank conflicts

2-way bank conflict: stride = 2
8-way bank conflict: stride = 8
Use Shared Memory to Improve Coalescing

Original Access Pattern

Tiled Access Pattern

Copy into scratchpad memory

Perform multiplication with scratchpad values

Slide credit: Hwu & Kirk
Reducing Shared Memory Bank Conflicts

- Bank conflicts are only possible within a warp
  - No bank conflicts between different warps

- If strided accesses are needed, some optimization techniques can help
  - Padding
  - Randomized mapping
  - Hash functions
SIMD Utilization
Threads Can Take Different Paths in Warp-based SIMD

- Each thread can have **conditional control flow instructions**
- Threads can execute different control flow paths

Slide credit: Tor Aamodt
Control Flow Problem in GPUs/SIMT

- A GPU uses a SIMD pipeline to save area on control logic
  - Groups scalar threads into warps

- Branch divergence occurs when threads inside warps branch to different execution paths

This is the same as conditional/predicated/masked execution. Recall the Vector Mask and Masked Vector Operations?
Intra-warp divergence

Compute(threadIdx.x);
if (threadIdx.x % 2 == 0){
  Do_this(threadIdx.x);
}
else{
  Do_that(threadIdx.x);
}
Increasing SIMD Utilization

- **Divergence-free** execution

```
Compute(threadIdx.x);
if (threadIdx.x < 32){
    Do_this(threadIdx.x * 2);
}
else{
    Do_that((threadIdx.x%32)*2+1);
}
```
Vector Reduction: Naïve Mapping (I)

Slide credit: Hwu & Kirk
Program with low SIMD utilization

__shared__ float partialSum[]

unsigned int t = threadIdx.x;

for (int stride = 1; stride < blockDim.x; stride *= 2) {

    __syncthreads();

    if (t % (2*stride) == 0)
        partialSum[t] += partialSum[t + stride];

}
Divergence-Free Mapping (I)

- All active threads belong to the same warp

Slide credit: Hwu & Kirk
Divergence-Free Mapping (II)

- Program with high SIMD utilization

```c
__shared__ float partialSum[]

unsigned int t = threadIdx.x;

for (int stride = blockDim.x; stride > 0; stride >>= 1){
    __syncthreads();
    if (t < stride)
        partialSum[t] += partialSum[t + stride];
}
```
Atomic Operations
Atomic Operations (I)

- CUDA provides **atomic instructions** on shared memory and global memory
  - They perform **read-modify-write** operations atomically

- Arithmetic functions
  - Add, sub, max, min, exch, inc, dec, CAS
    ```c
    int atomicAdd(int*, int);
    ```
    - Pointer to shared memory or global memory
    - Value to add
    - Return value (old value)

- Bitwise functions
  - And, or, xor

- Datatypes: int, uint, ull, float (half, single, double)*)

*) Datatypes for different atomic operations in [https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#atomic-functions](https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#atomic-functions)
Atomic Operations (II)

- Atomic operations serialize the execution if there are atomic conflicts

```
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
```

No atomic conflict = concurrent updates

```
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
```

Atomic conflict = serialized updates
Uses of Atomic Operations

- **Computation**
  - Atomics on an array that will be the output of the kernel
  - Example
    - Histogram, reduction

- **Synchronization**
  - Atomics on memory locations that are used for synchronization or coordination
  - Example
    - Counters, locks, flags...

- Use them to prevent **data races** when more than one thread need to update the same memory location
Image Histogram

- Histograms are widely used in image processing
  - Some computation before voting in the histogram may be needed

```c
For (each pixel i in image I){
    Pixel = I[i]  // Read pixel
    Pixel’ = Computation(Pixel)  // Optional computation
    Histogram[Pixel’]++  // Vote in histogram bin
}
```

- Parallel threads frequently incur atomic conflicts in image histogram computation
7 versions in CUDA samples: Tree-based reduction in shared memory

- Version 0: No whole warps active
- Version 1: Contiguous threads, but many bank conflicts
- Version 2: No bank conflicts
- Version 3: First level of reduction when reading from global memory
- Version 4: Warp shuffle or unrolling of final warp
- Version 5: Warp shuffle or complete unrolling
- Version 6: Multiple elements per thread sequentially

https://docs.nvidia.com/cuda/cuda-samples/index.html#cuda-parallel-reduction
3 new versions of reduction based on 3 previous versions

- Version 0: No whole warps active
- Version 3: First level of reduction when reading from global memory
- Version 6: Multiple elements per thread sequentially

New versions 7, 8, and 9

- Replace the for loop (tree-based reduction) with one shared memory atomic operation per thread
Asynchronous Data Transfers between CPU and GPU
CUDA Streams

- **CUDA streams** (command queues in OpenCL)
- Sequence of operations that are performed in order
  - 1. Data transfer CPU-GPU
  - 2. Kernel execution
    - D input data instances, B blocks
    - #Streams: (D / #Streams) data instances, (B / #Streams) blocks
  - 3. Data transfer GPU-CPU
Asynchronous Transfers between CPU & GPU

- **Computation** divided into \#Streams
  - D input data instances, B blocks
  - \#Streams
    - D/\#Streams data instances
    - B/\#Streams blocks

- **Estimates**
  \[ t_E + \frac{t_T}{\#Streams} \]
  \[ t_T + \frac{t_E}{\#Streams} \]
  
  \[ t_E \geq t_T \text{ (dominant kernel)} \]
  \[ t_T > t_E \text{ (dominant transfers)} \]
Overlap of Data Transfers and Kernel Execution

```c
// Create streams
int number_ofStreams = 32;
cudaStream_t stream[number_ofStreams]; // Stream declaration
for(int i = 0; i < number_ofStreams; ++i)
    cudaStreamCreate(&stream[i]); // Stream creation

// CPU-GPU data transfers
for (int i = 0; i < number_of_streams; ++i)
cudaMemcpyAsync(inputDevPtr + i * size, hostPtr + i * size, size,
    cudaMemcpyHostToDevice, stream[i]);

// Kernel launches
for (int i = 0; i < number_of_streams; ++i)
MyKernel<<<num_blocks / number_of_streams, num_threads, 0, stream[i]>>>(
    outputDevPtr + i * size, inputDevPtr + i * size, size);

// GPU-CPU data transfers
for (int i = 0; i < number_of_streams; ++i)
cudaMemcpyAsync(hostPtr + i * size, outputDevPtr + i * size, size,
    cudaMemcpyDeviceToHost, stream[i]);

cudaDeviceSynchronize(); // Explicit synchronization

// Destroy streams
for (int i = 0; i < number_of_streams; ++i)
cudaStreamDestroy(stream[i]); // Stream destruction
```


Check CUDA programming guide
https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#streams
Use Case: Video Processing

- Applications with independent computation on different data instances can benefit from asynchronous transfers.
- For instance, video processing.

Video Processing: Performance Results (I)

- 256-bin histogram calculation

Video Processing: Performance Results (II)

- RGB-to-grayscale conversion

```
```
Performance Considerations

- Main bottlenecks
  - CPU-GPU data transfers
  - Global memory access

- Memory access
  - Latency hiding
    - Occupancy
  - Memory coalescing
  - Data reuse
    - Shared memory usage

- SIMD (Warp) Utilization: Divergence

- Other considerations
  - Atomic operations: Serialization
  - Data transfers between CPU and GPU
    - Overlap of communication and computation
Recommended Readings (I)

  - Chapter 5: Performance considerations
  - Chapter 18 - Programming a heterogeneous computing cluster, Section 18.5
Recommended Readings (II)

  - Chapter 6 - Performance considerations
  - Chapter 20 - Programming a heterogeneous computing cluster, Section 20.5
P&S Heterogeneous Systems
GPU Performance Considerations

Dr. Juan Gómez Luna
Prof. Onur Mutlu
ETH Zürich
Fall 2022
31 October 2022