Special Topic: GPUs

COMP 520: Compiler Design

Alexander Krolik

alexander.krolik@mail.mcgill.ca

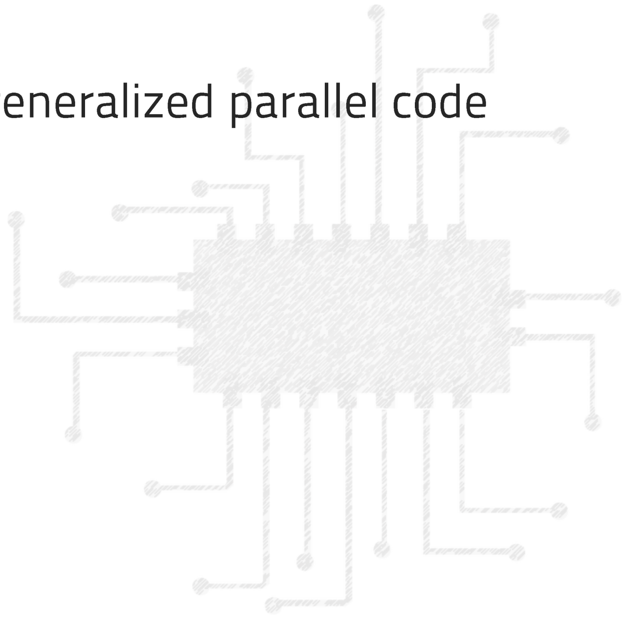
Introduction

What is a GPU?

- "Graphics Processing Unit"
- A specialized processor originally designed for graphics operations

What kind of code can they execute?

- **Historically**: only graphics code (OpenGL)
- **Currently**: GPGPUs (General-Purpose GPUs) execute generalized parallel code (OpenCL/CUDA)



GPU Architecture

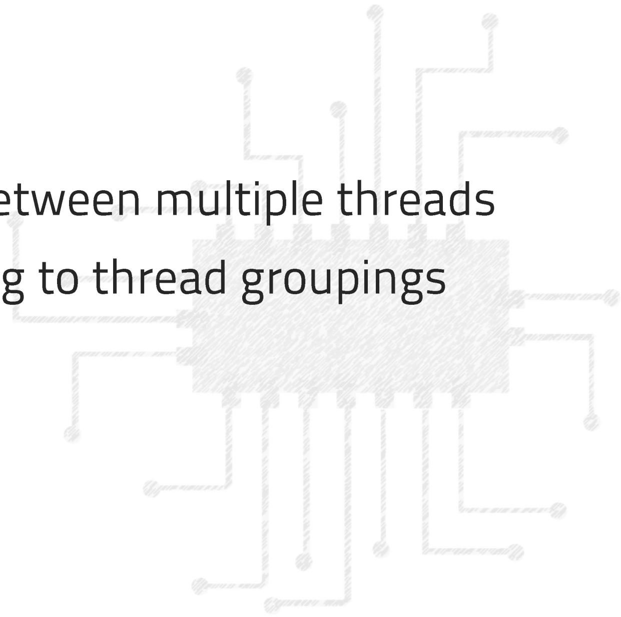
A modern GPU architecture is geared towards high degrees of parallelism

Execution

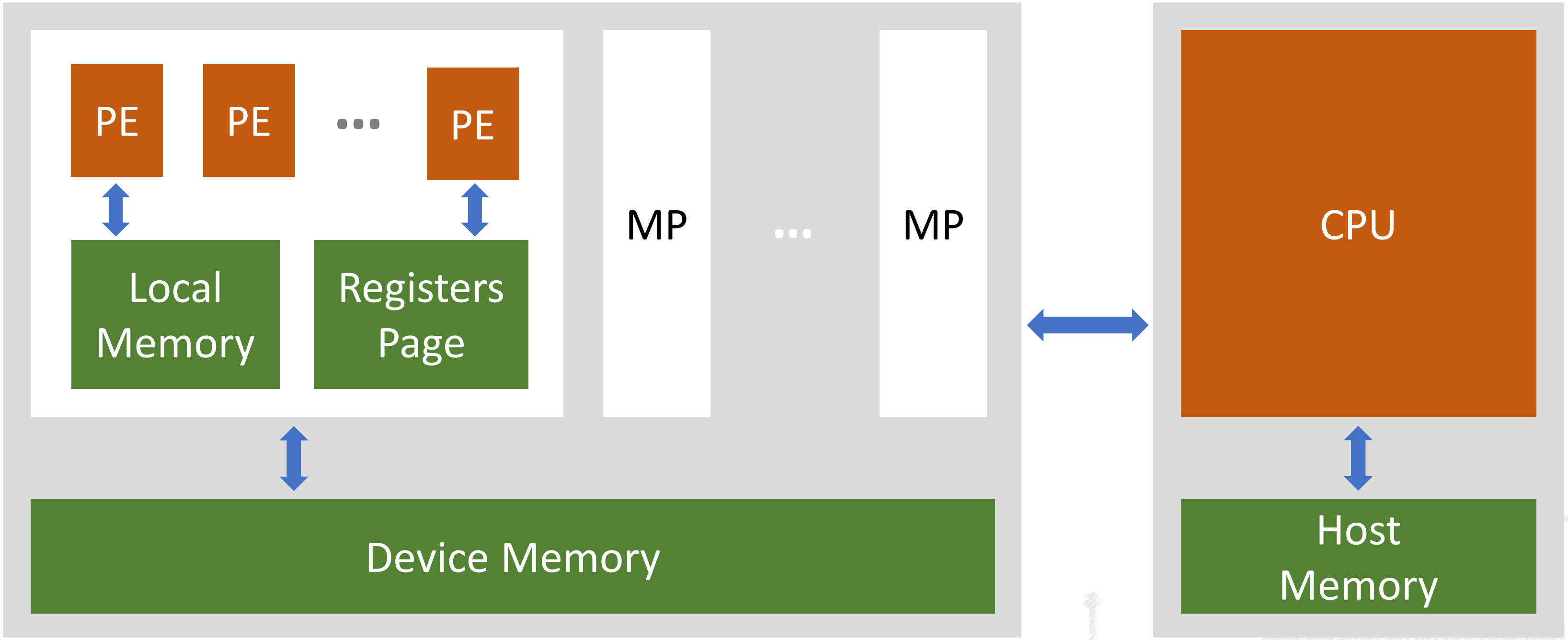
- Highly parallel, with thousands (and thousands of threads)
- Hierarchically parallel, with threads grouped at multiple levels

Memory

- High bandwidth, allowing fast concurrent accesses between multiple threads
- Hierarchical design, with multiple levels corresponding to thread groupings

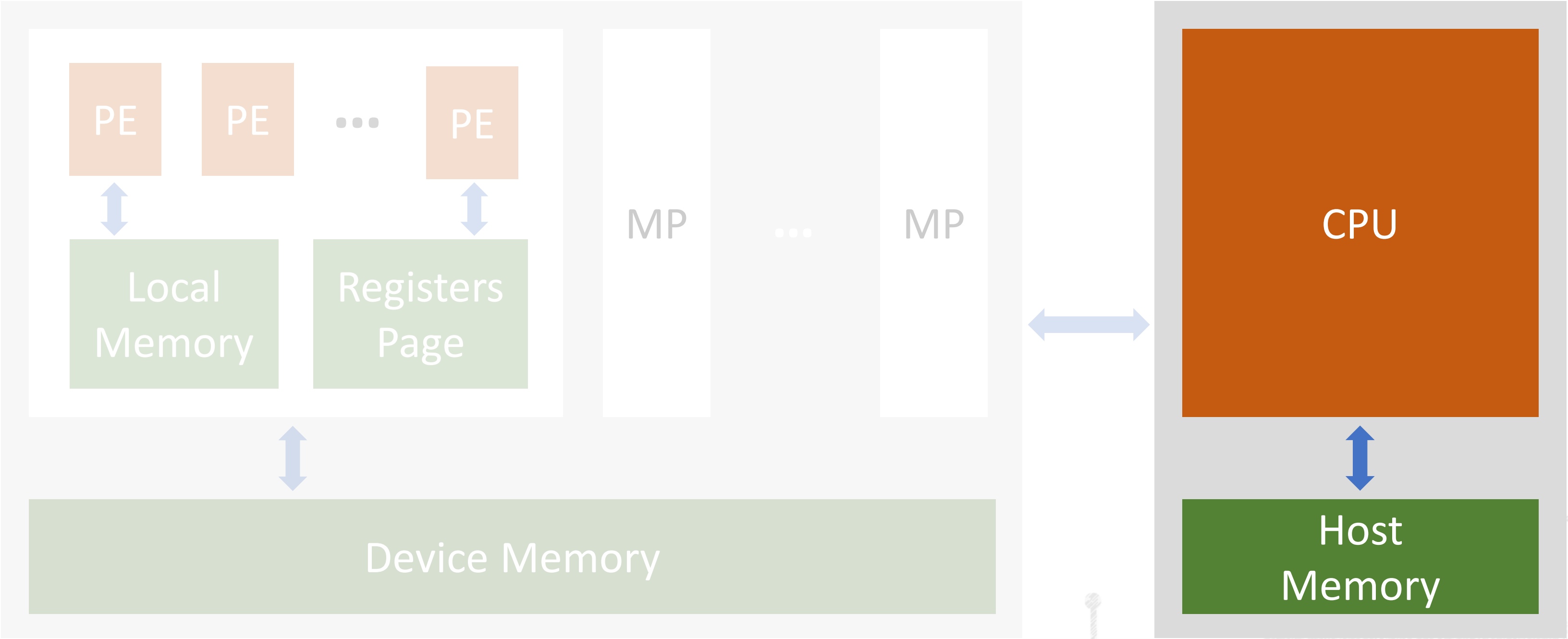


GPU Architecture



- 🟦 Connections
- 🟩 Memory
- 🟧 Processors
- ▨ Units/groupings

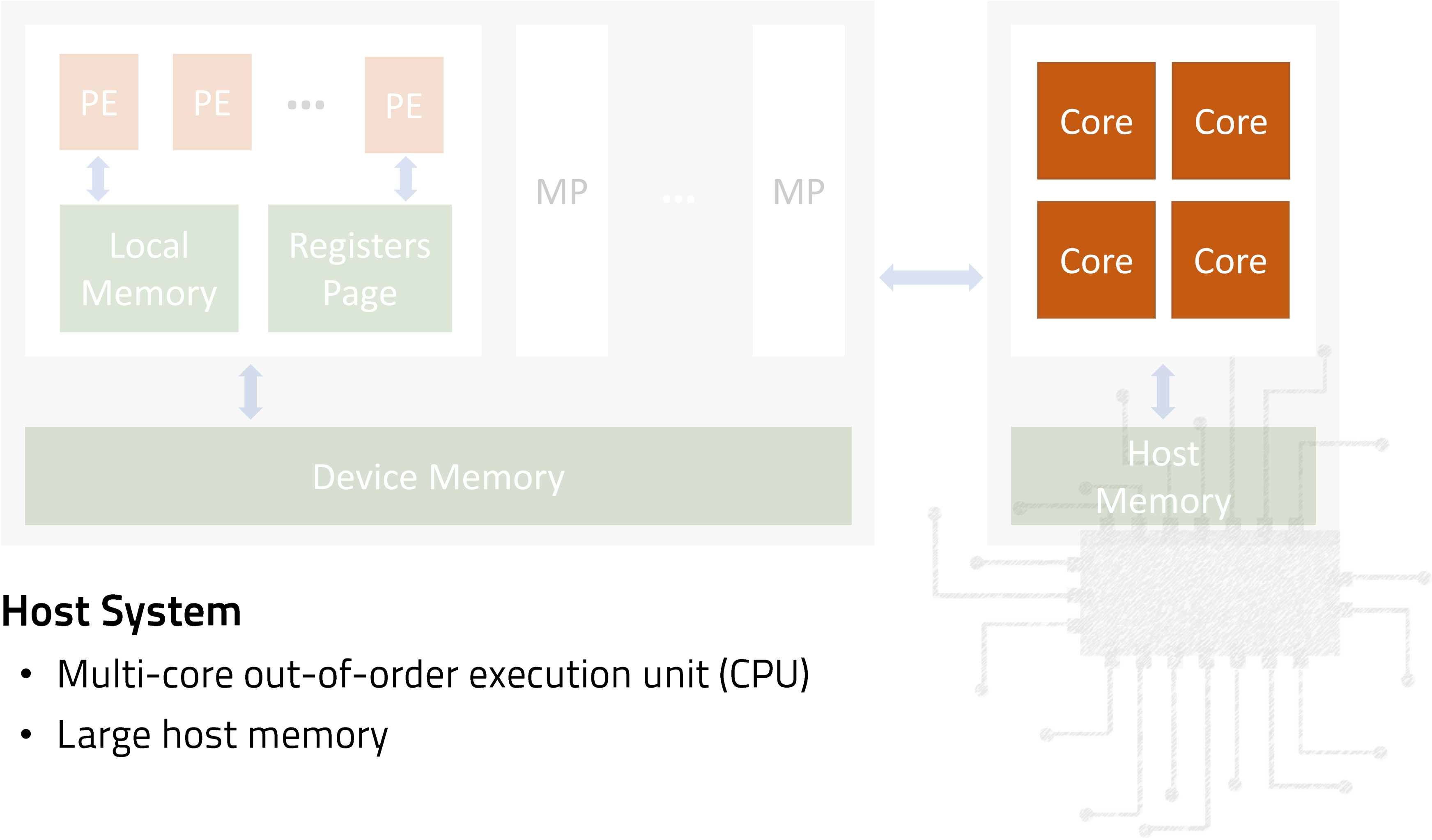
GPU Architecture



Host System

- Multi-core out-of-order execution unit (CPU)
- Large host memory

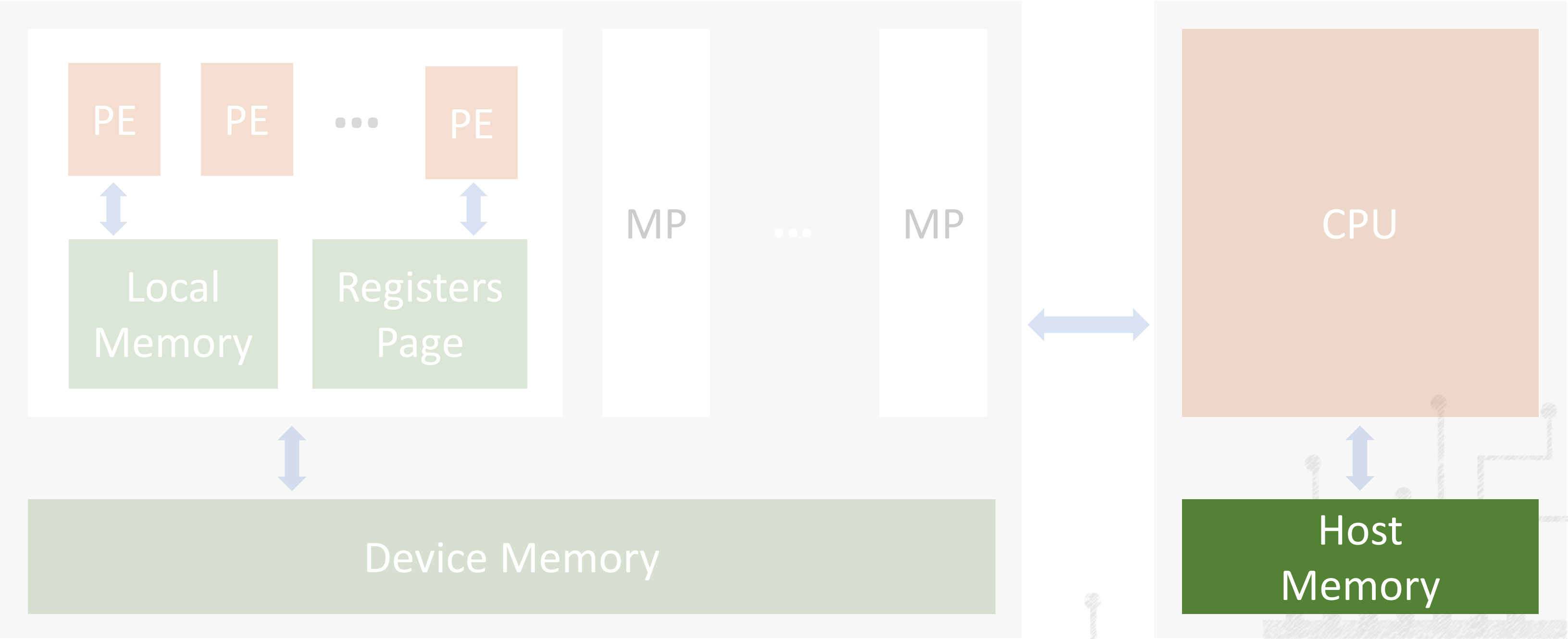
GPU Architecture



Host System

- Multi-core out-of-order execution unit (CPU)
- Large host memory

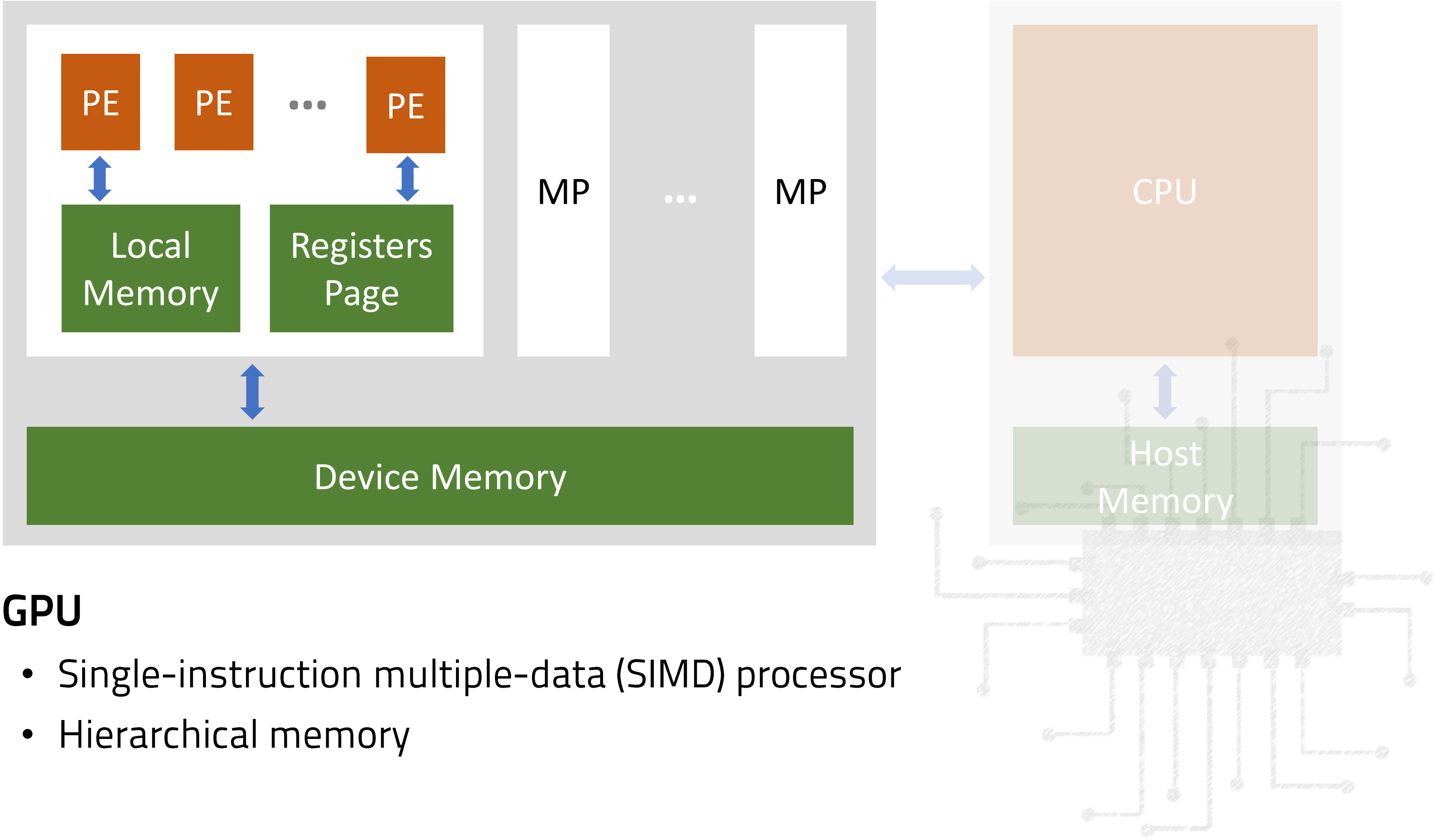
GPU Architecture



Host System

- Multi-core out-of-order execution unit (CPU)
- Large host memory

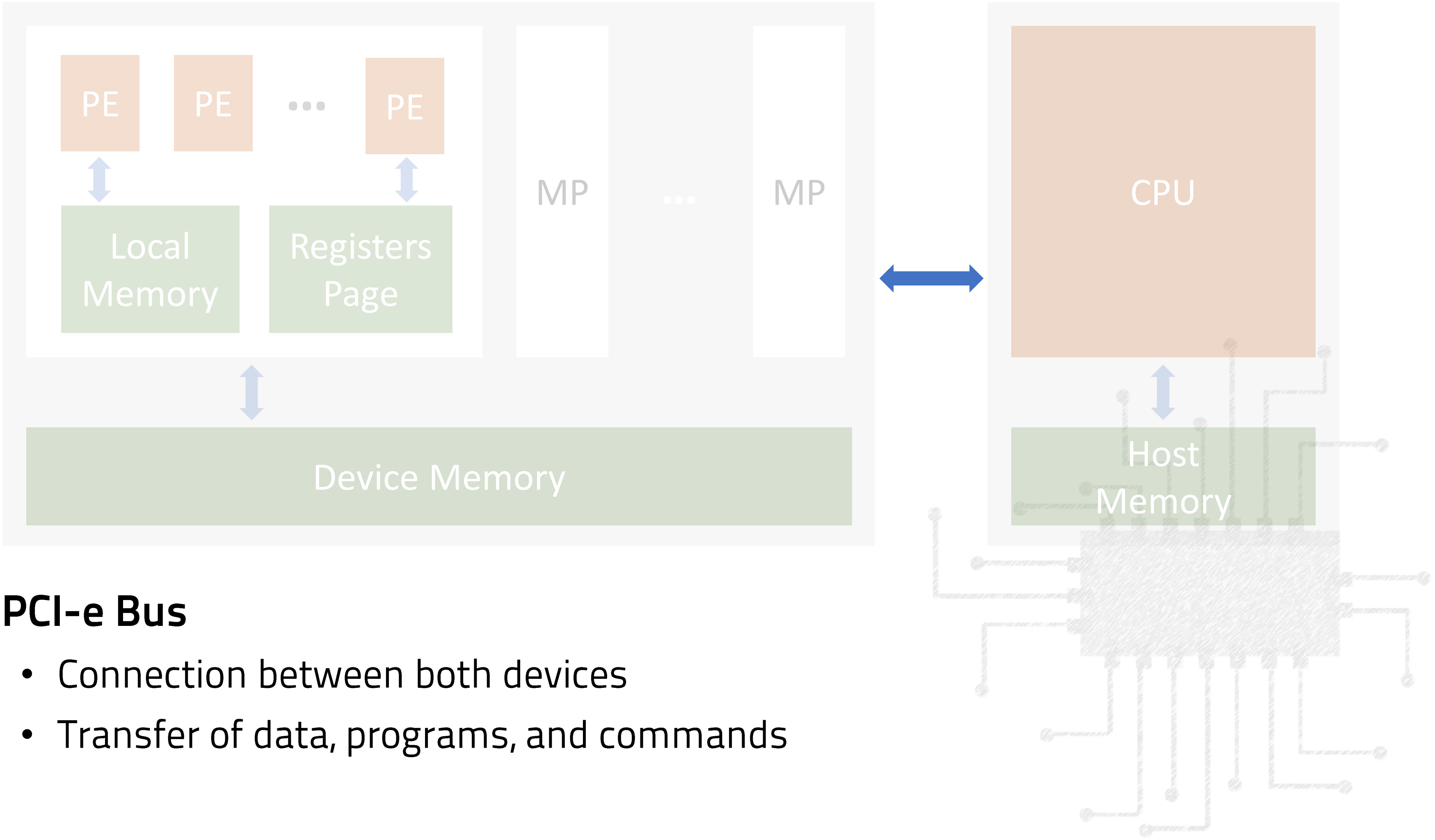
GPU Architecture



GPU

- Single-instruction multiple-data (SIMD) processor
- Hierarchical memory

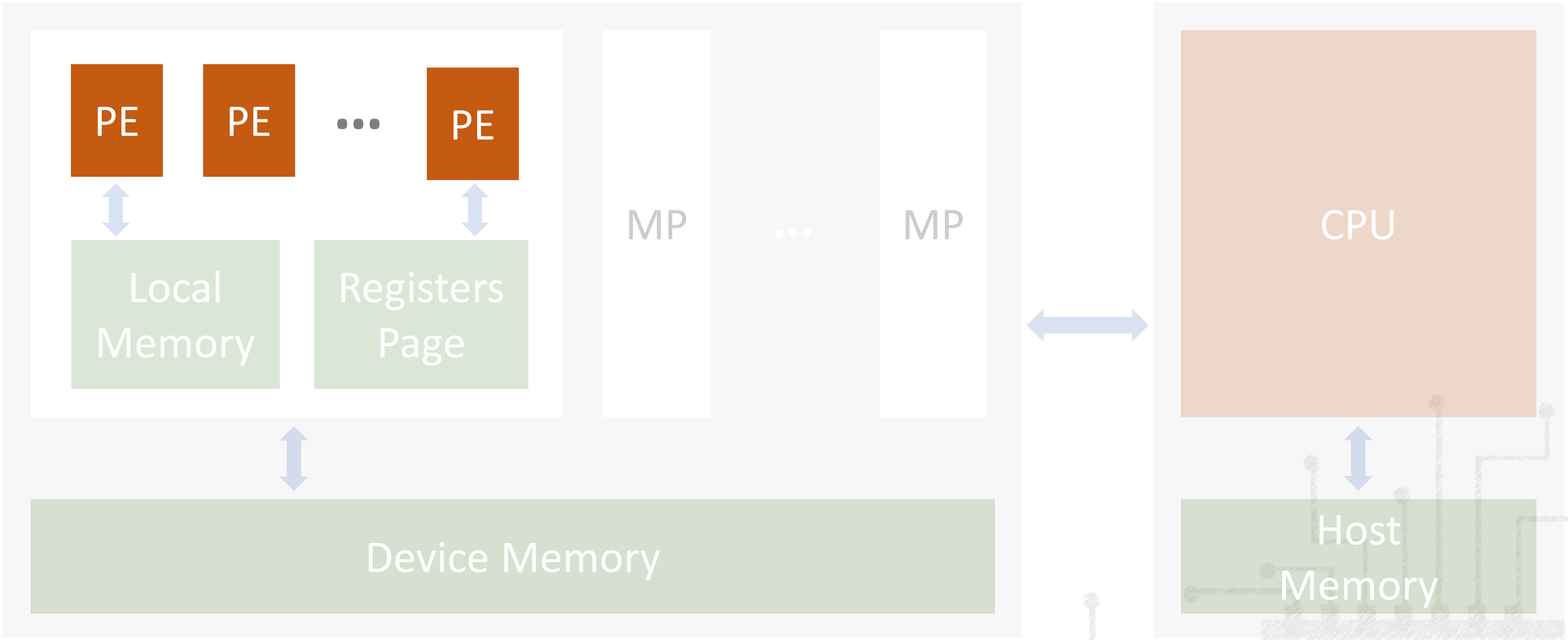
GPU Architecture



PCI-e Bus

- Connection between both devices
- Transfer of data, programs, and commands

GPU Architecture



PE

PE

...

PE

Local Memory

Registers Page

MP

...

MP

CPU

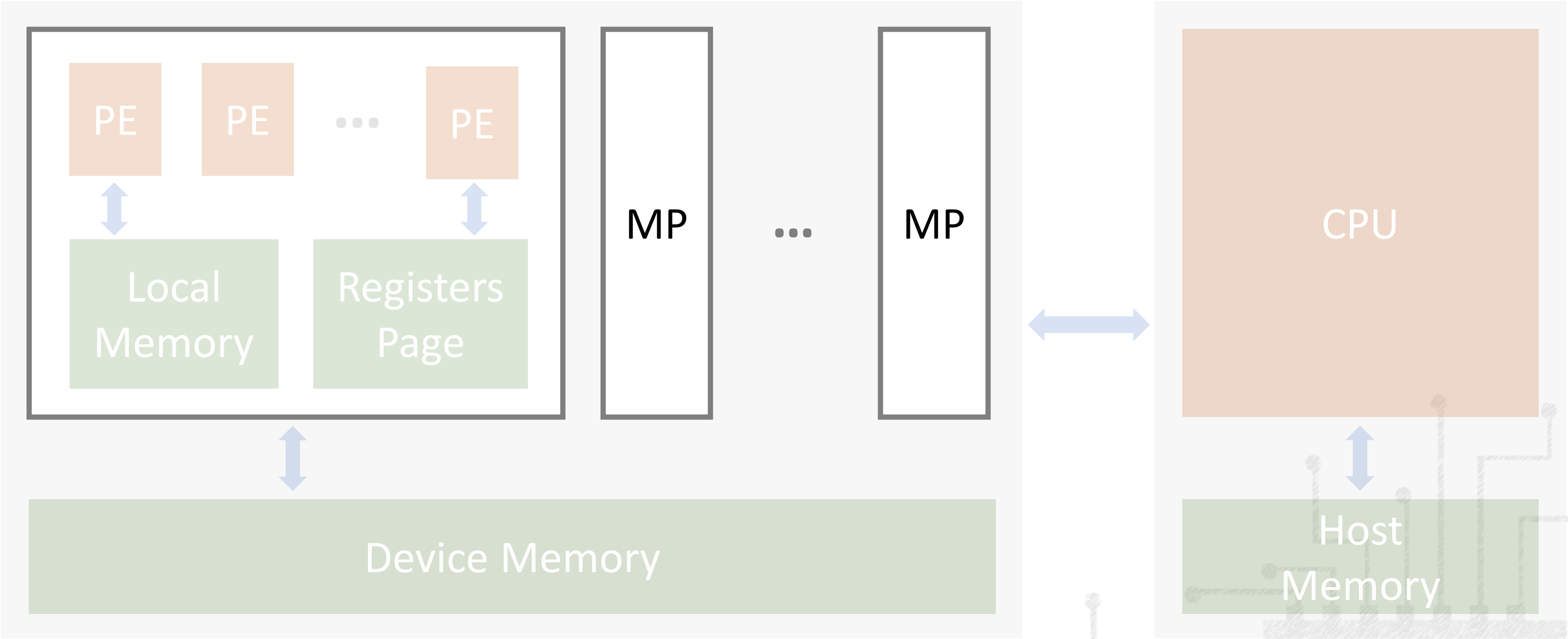
Device Memory

Host Memory

Processing Elements

- Processors on the GPU
- Each element computes the result for a single thread
- Slower clock speed than the CPU, but there are thousands !

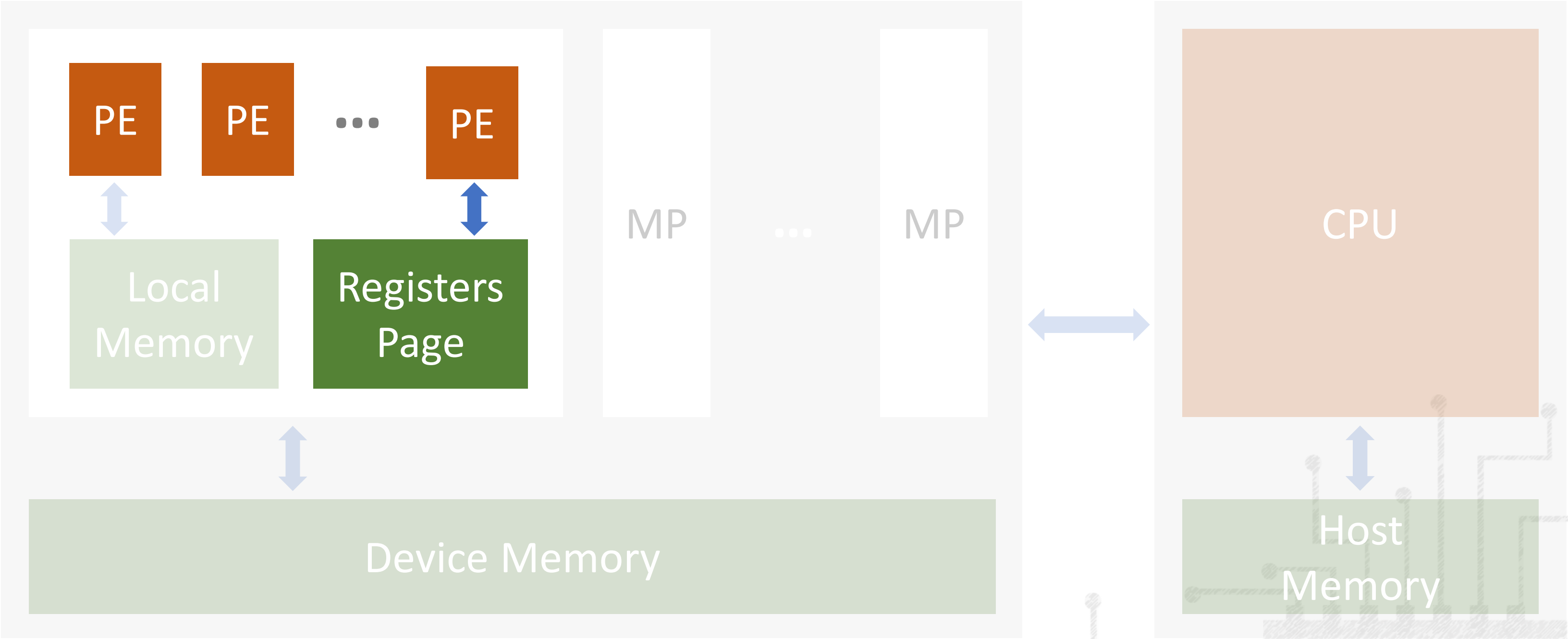
GPU Architecture



Multi-Processors

- Grouping of processing elements
- All PE within the same multi-processor execute in *lock-step*
- There is no guarantee on execution order between multi-processors

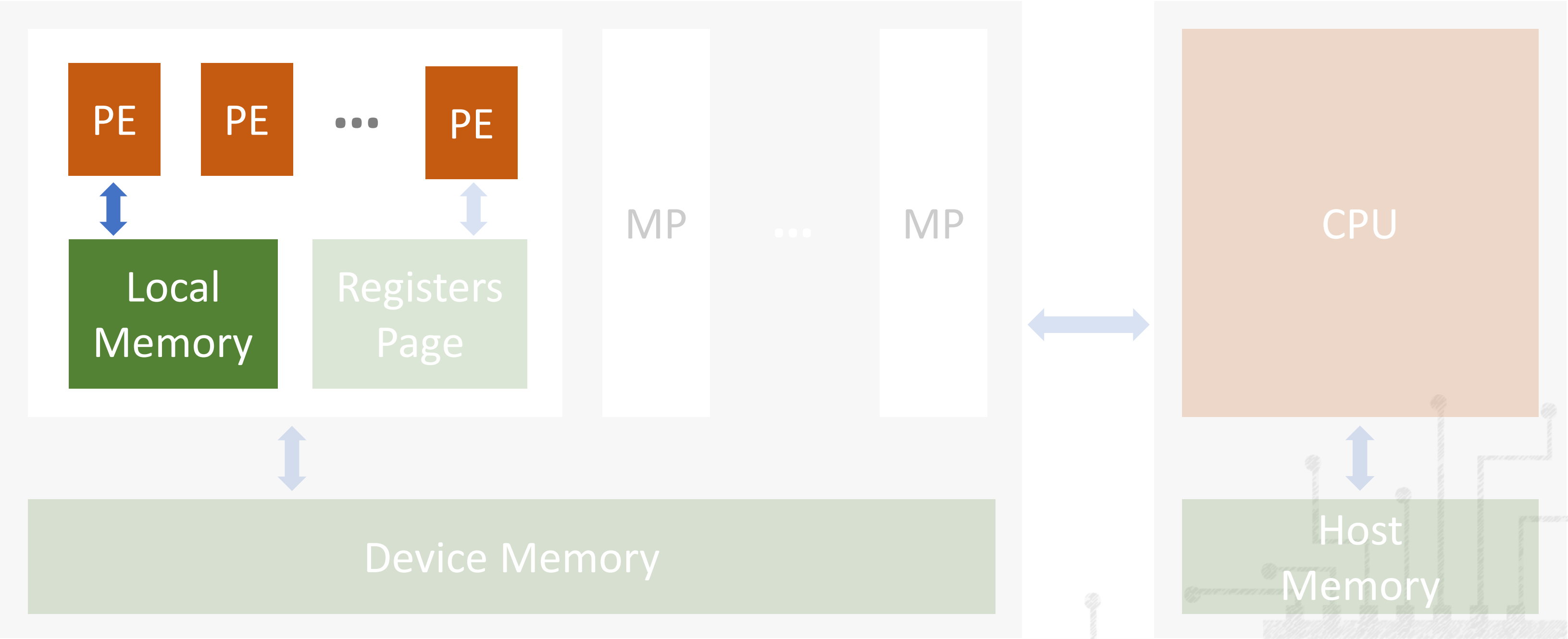
GPU Architecture



Registers Page

- Private memory for each processing element (no sharing permitted)
- Lowest latency (fastest) memory on the GPU

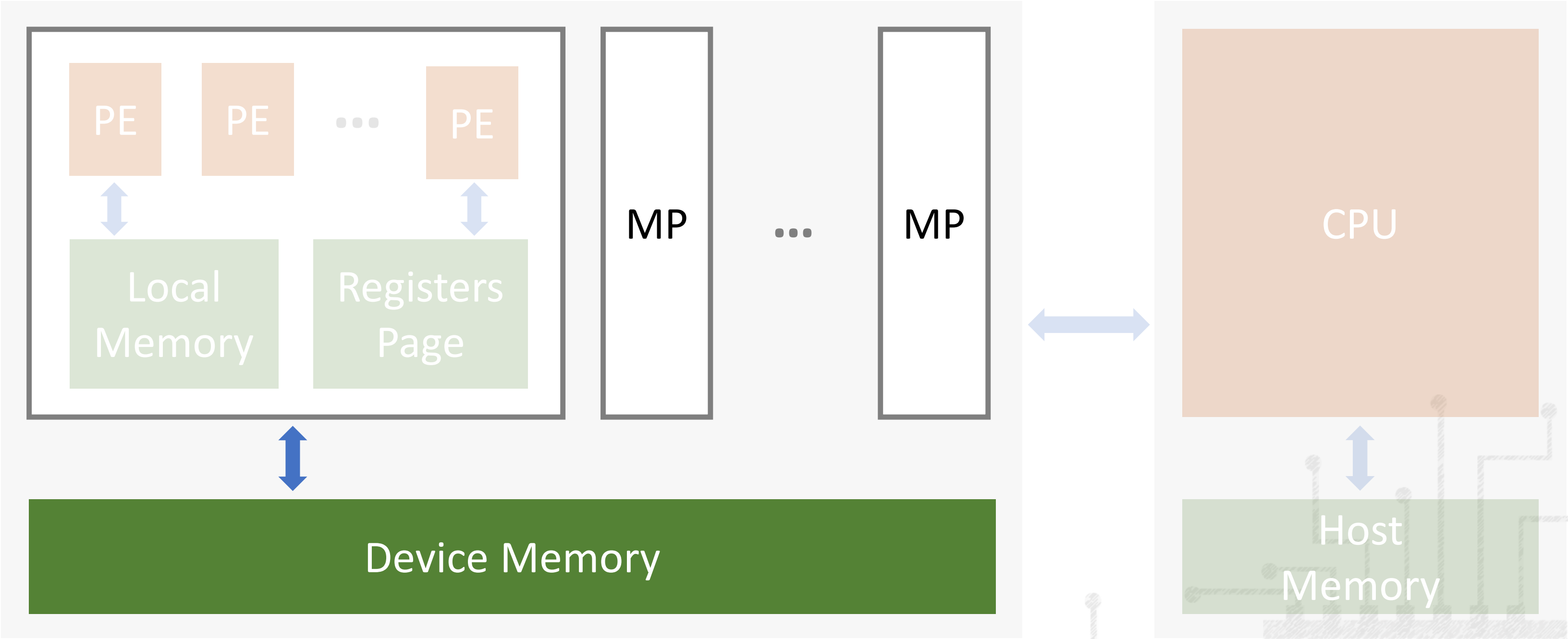
GPU Architecture



Local Memory

- Shared memory between processing elements *in the same multi-processor*
- Larger size but higher latency than registers

GPU Architecture



Device Memory

- Shared memory between all multi-processors
- Largest size but highest latency memory
- Persistent between GPU programs (kernels) – other memory is not

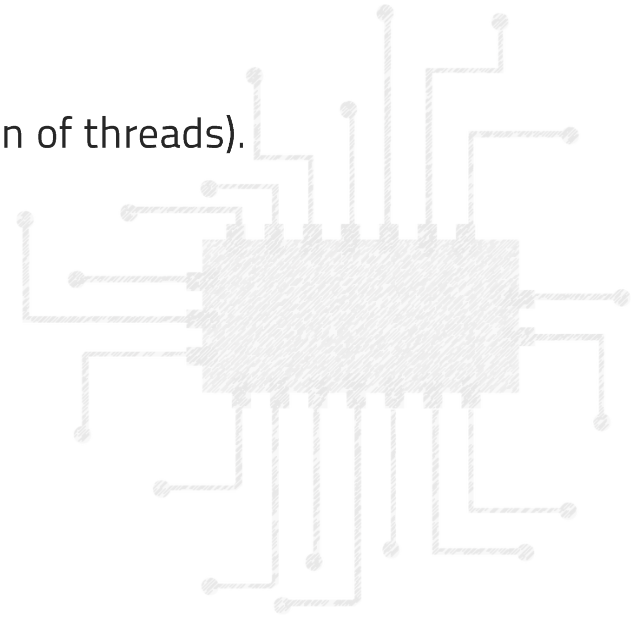
GPU Execution

GPUs are **highly** parallel devices, perfect for embarrassingly parallel code

Executing Code

A full GPU program consists of two code sections

- **Host** code that runs on the CPU
  - Compiles the program;
  - Transfers the data;
  - Specifies the thread geometry (number and organization of threads).
- **Kernel** (GPU code) executes the parallel section



GPU Thread Geometry

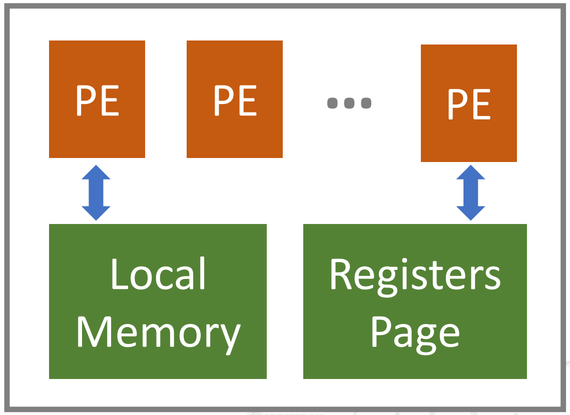
The host code specifies the thread geometry

- Number of threads; and
- Grouping of threads per multi-processor.

Thread Groups

Threads from the same group

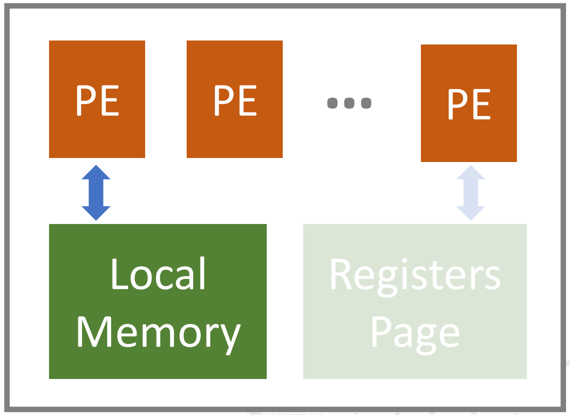
- Execute on the same multiprocessor; and
- Share the same local memory.



GPU Memory

Local Memory

- Shared by all threads in a group; but
- Is *not* synchronized automatically! !



Synchronization

Synchronization ensures that all threads in a group are at the same point in the kernel

- **Within a group**: memory barrier
- **Between groups**: impossible!

GPU Memory

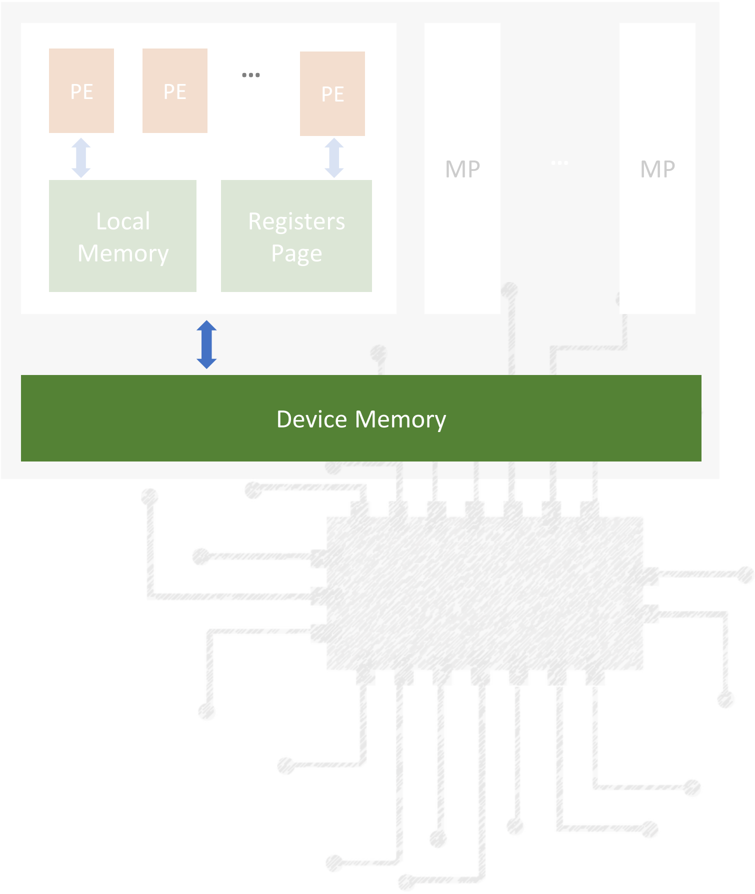
Device Memory

- Shared by all threads on the GPU; but
- Is *not* synchronized automatically ! !

Synchronization

Synchronization ensures that all threads in a group are at the same point in the kernel

- **Within a group**: memory barrier
- **Between groups**: impossible !

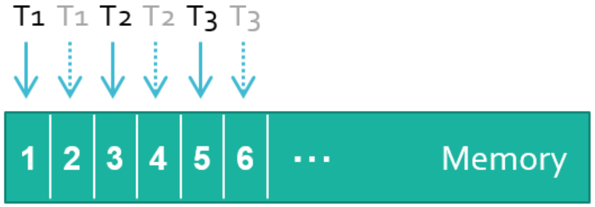


GPU Memory: Coalescing

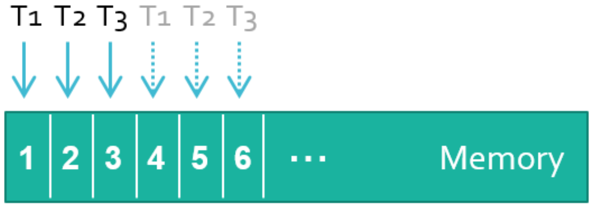
Optimizing GPU memory bandwidth is important for performance

Memory Coalescing

- Concurrent accesses to consecutive memory locations are merged into a single fetch
- **Pattern**: access **consecutive memory locations** from **consecutive threads**



Uncoalesced Access Pattern



Coalesced Access Pattern

Aggregation Functions

Idea: Group the values of multiple rows into a single value (fold)

0	2	4	2	3	2	3	0

Common Aggregate Functions:

COUNT	SUM	AVG	MAX	MIN
8	16	2	4	1

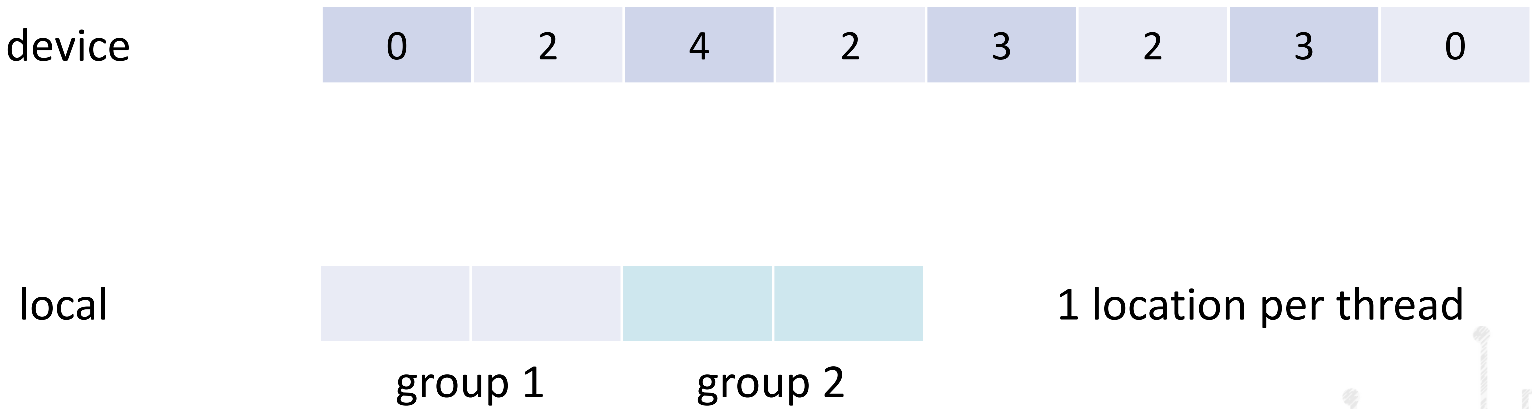
Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0

Parallel Reductions (SUM)

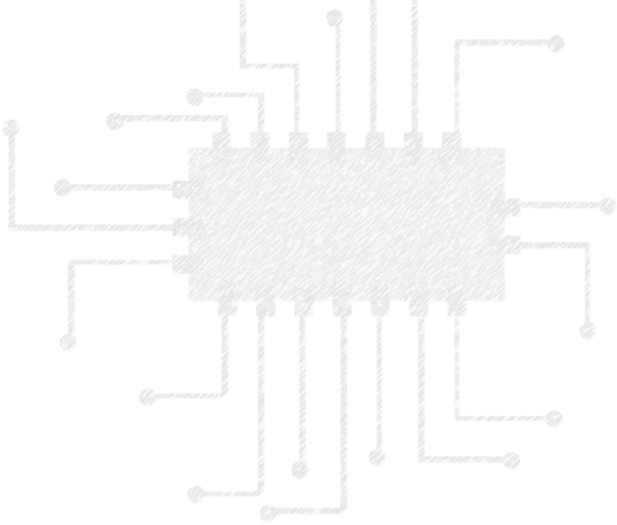
2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0

local				
	group 1		group 2	

1 location per thread



Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0

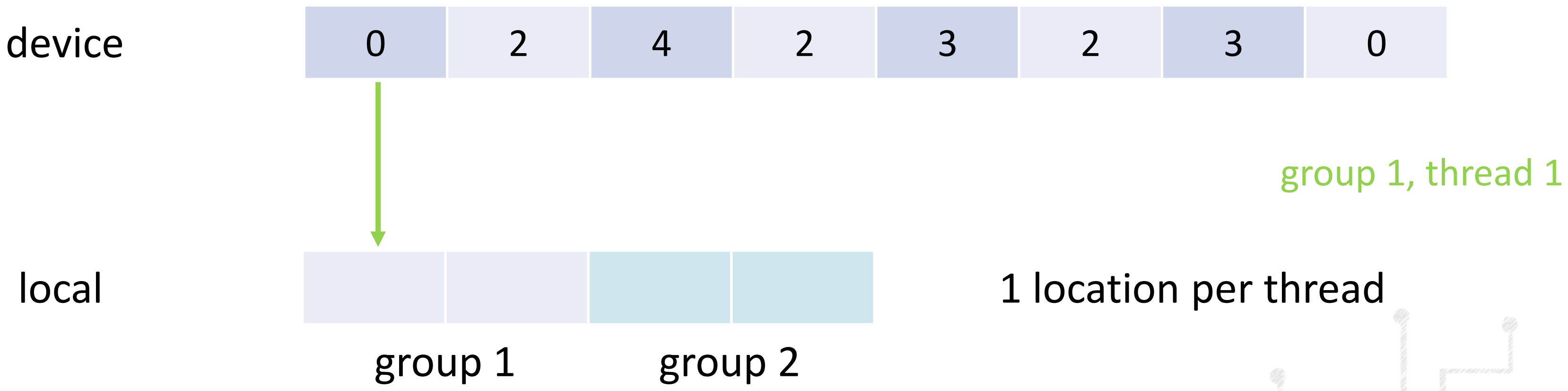
group 1, thread 1

local				
	group 1		group 2	

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

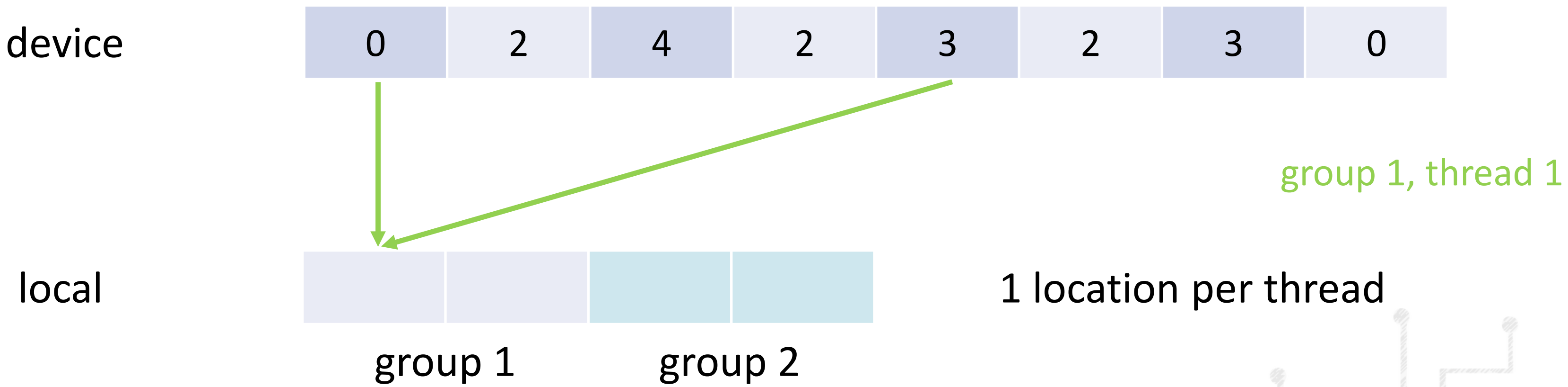


device	0	2	4	2	3	2	3	0
local								

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



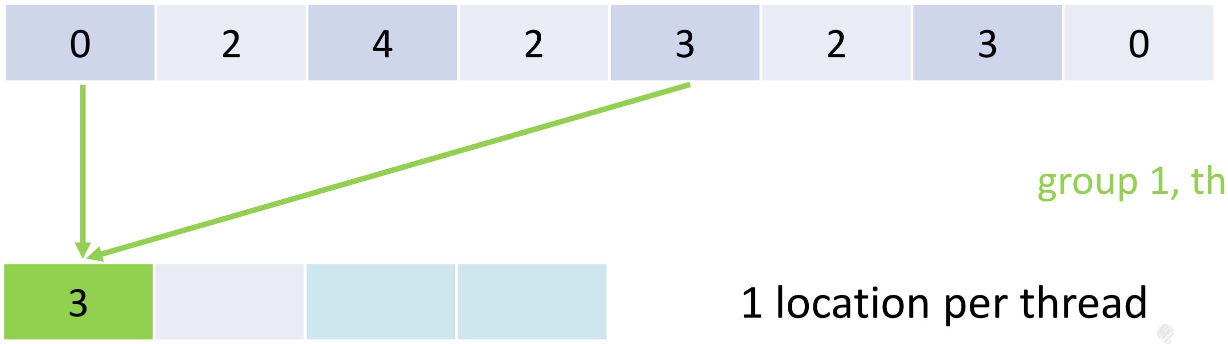
device	0	2	4	2	3	2	3	0

local				
	group 1		group 2	

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0

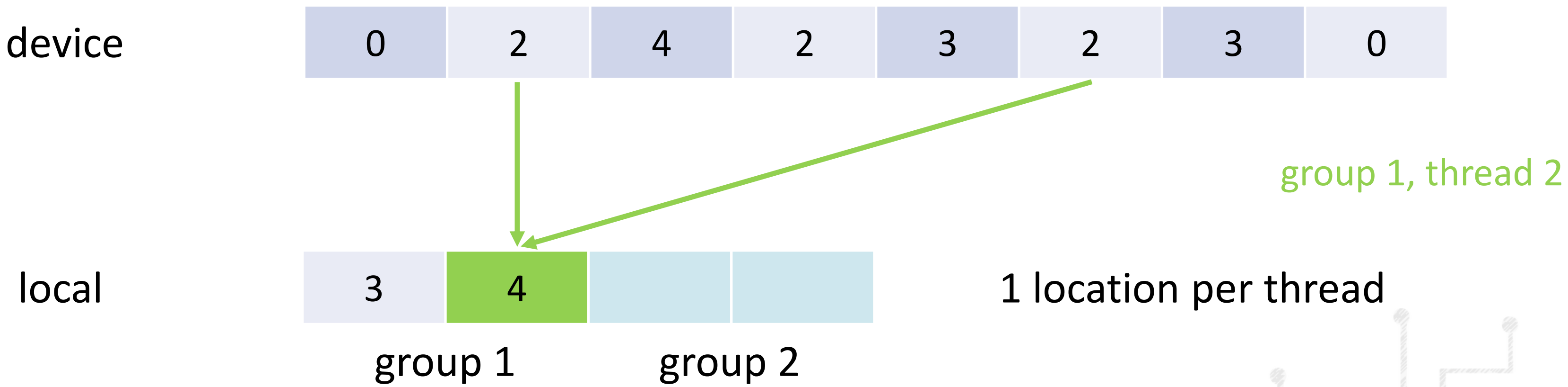
group 1, thread 1

local	3			
	group 1		group 2	

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device

0	2	4	2	3	2	3	0

group 1, thread 2

local

3	4		

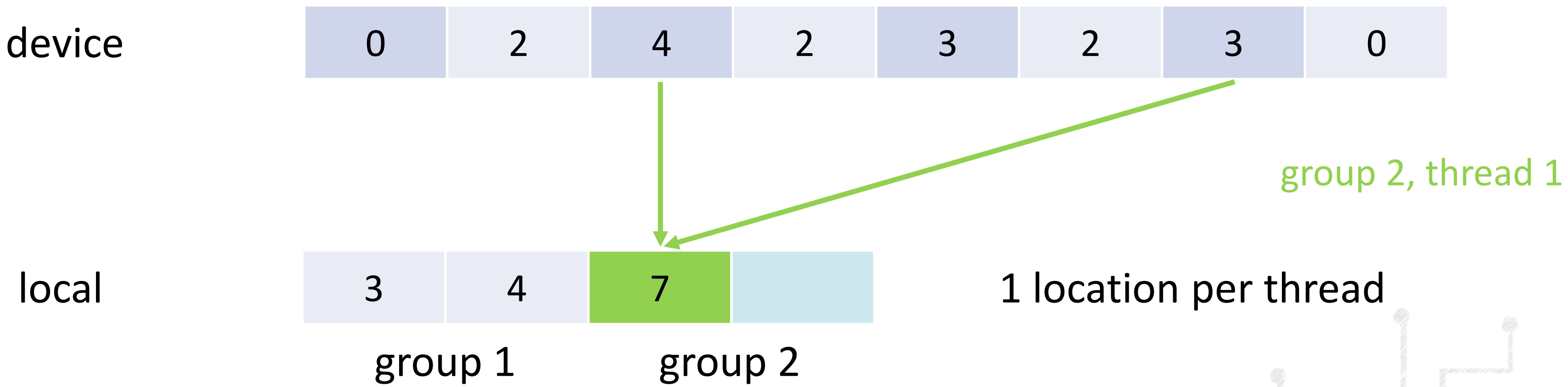
group 1

group 2

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device

0	2	4	2	3	2	3	0

group 2, thread 1

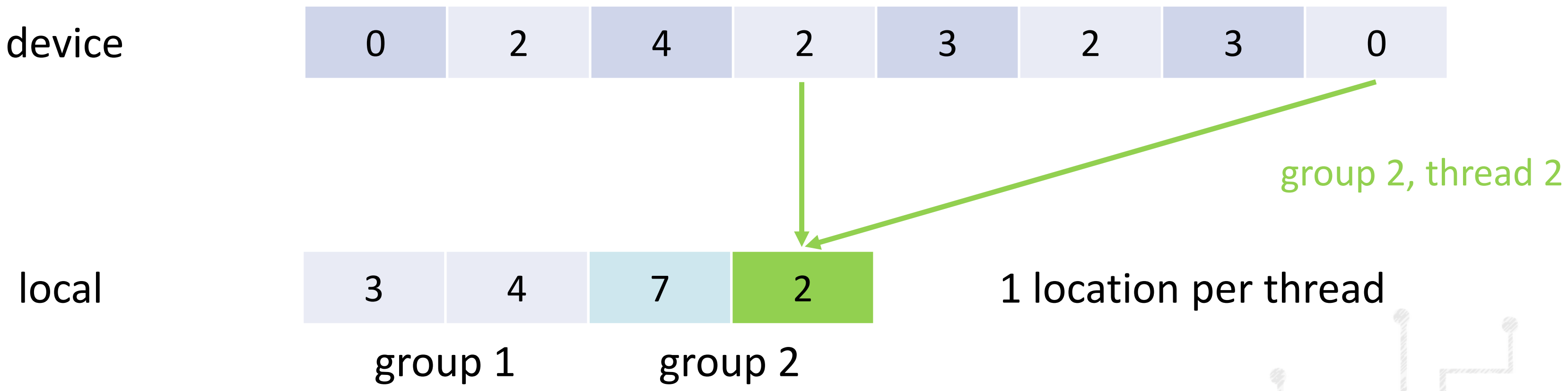
local

3	4	7	

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device							
0	2	4	2	3	2	3	0

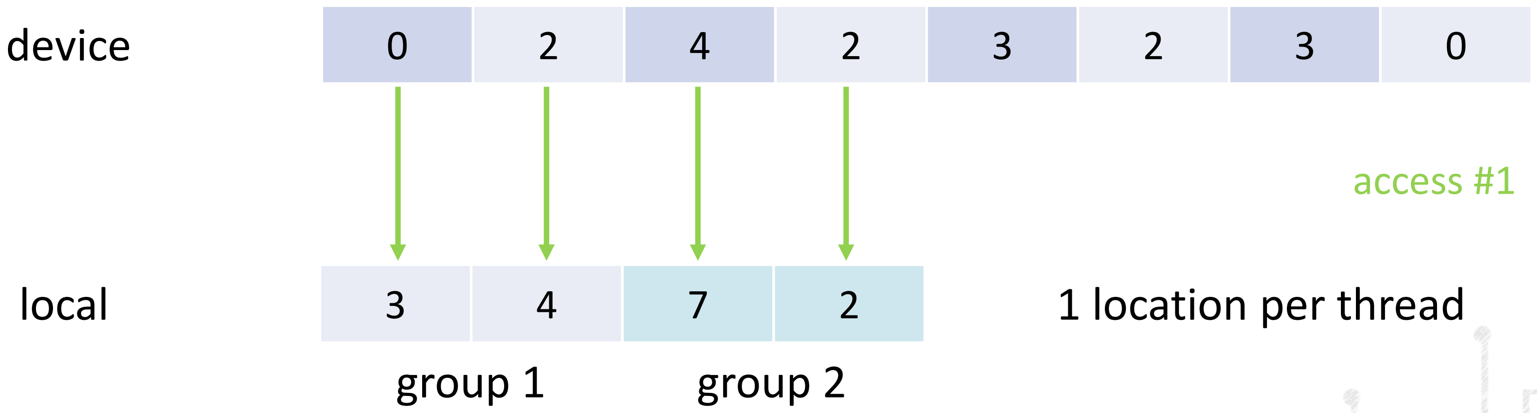
group 2, thread 2

local			
3	4	7	2
group 1		group 2	

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



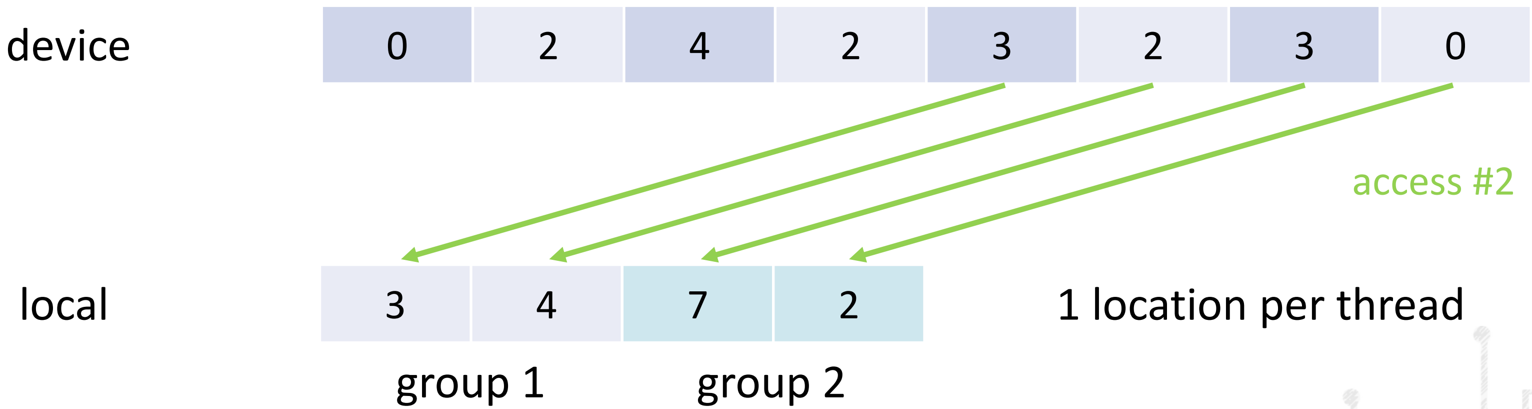
device	0	2	4	2	3	2	3	0
local	3	4	7	2				
	group 1		group 2					

access #1

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0

access #2

local	3	4	7	2
	group 1		group 2	

1 location per thread

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0
local	3	4	7	2				
	group 1		group 2					

1 location per thread

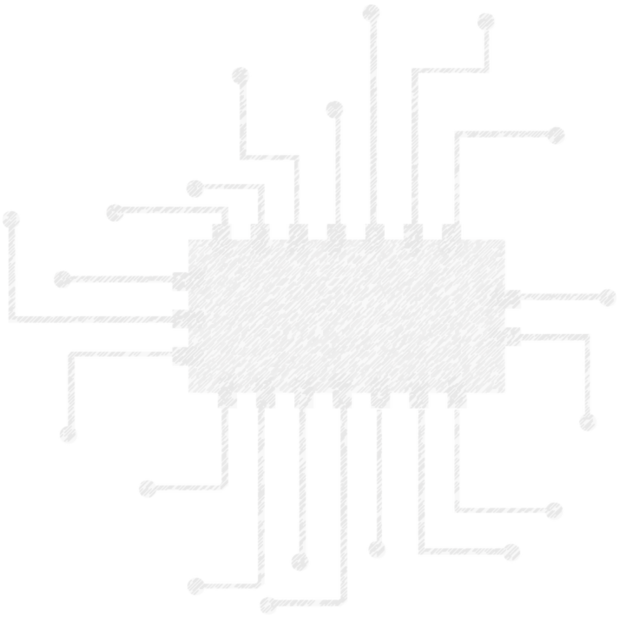
Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0
local	3	4	7	2				

1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)



Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0
local	3	4	7	2				

1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	3	4	9	2
	group 1		group 2	

Same local memory

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0

local	3	4	7	2

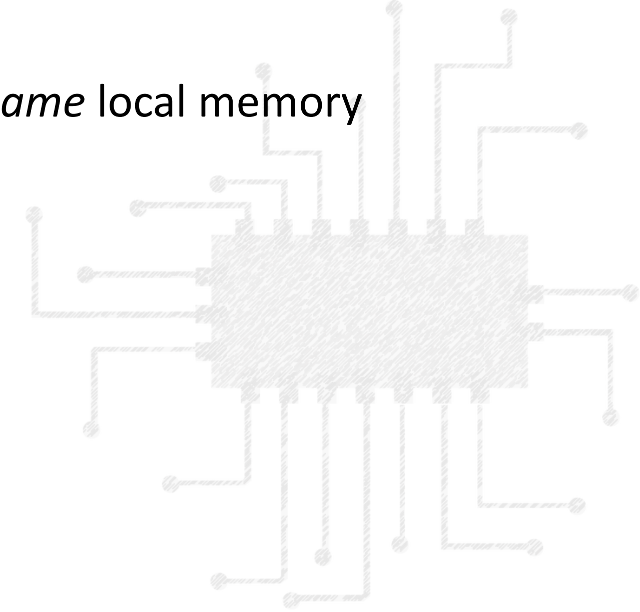
1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

group 1, thread 1

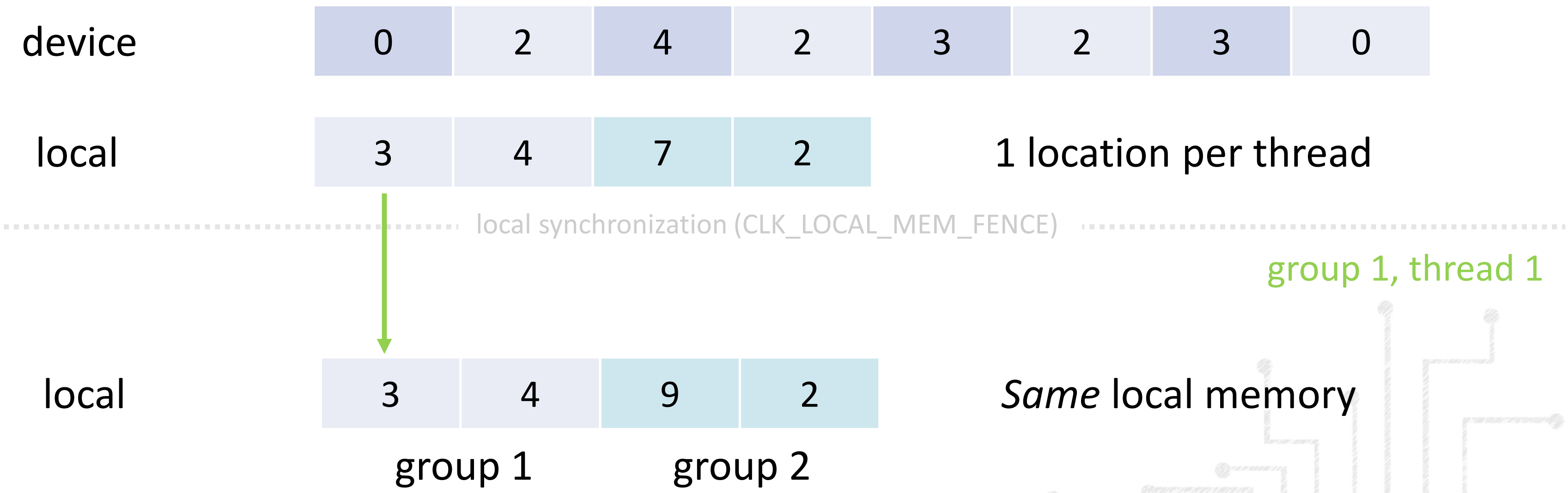
	group 1		group 2	
local	3	4	9	2

Same local memory



Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0

local	3	4	7	2	1 location per thread

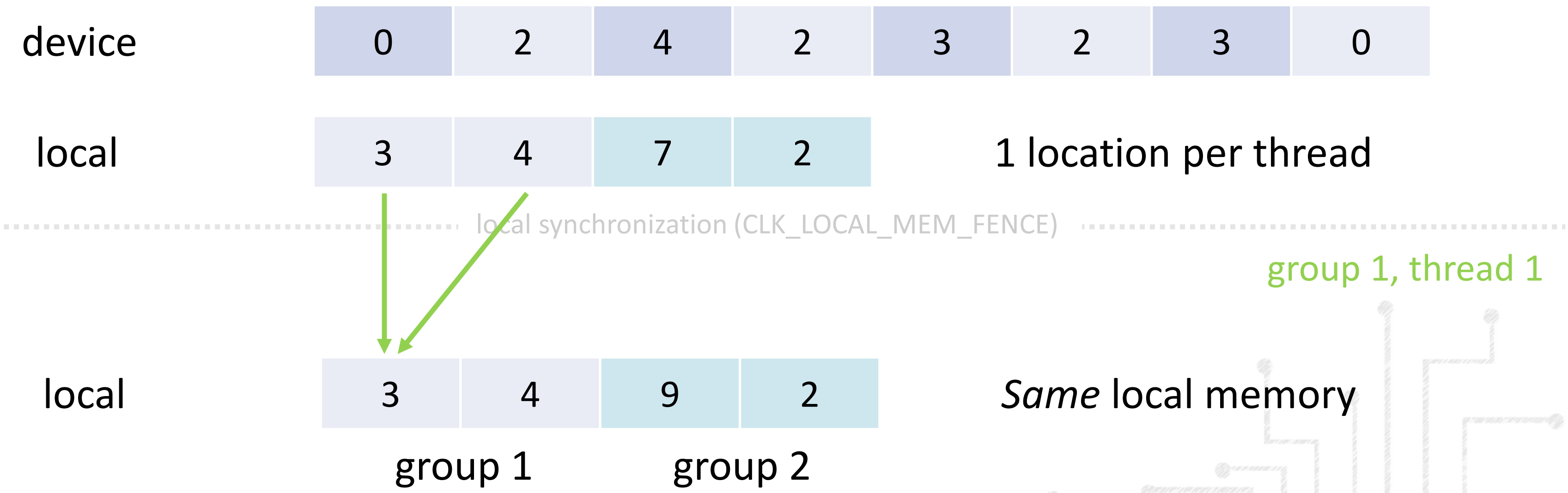
local synchronization (CLK_LOCAL_MEM_FENCE)

group 1, thread 1

local	3	4	9	2	*Same* local memory
	group 1		group 2		

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0
local	3	4	7	2				

1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

group 1, thread 1

local	3	4	9	2
	group 1		group 2	

Same local memory

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0
local	3	4	7	2				

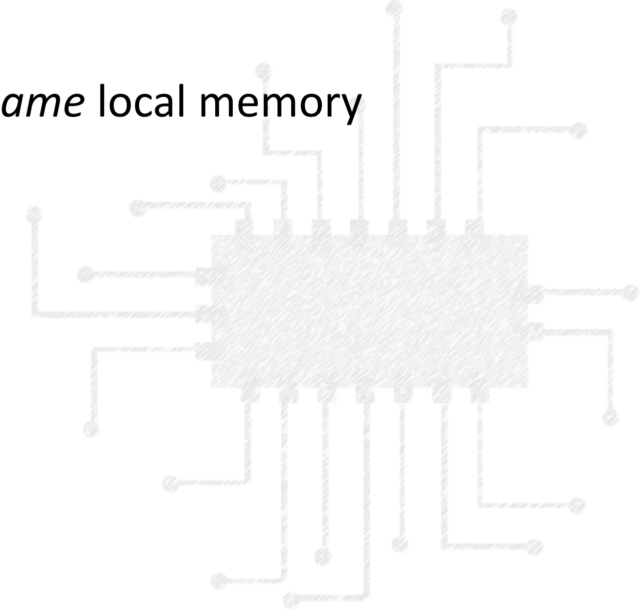
1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

group 1, thread 1

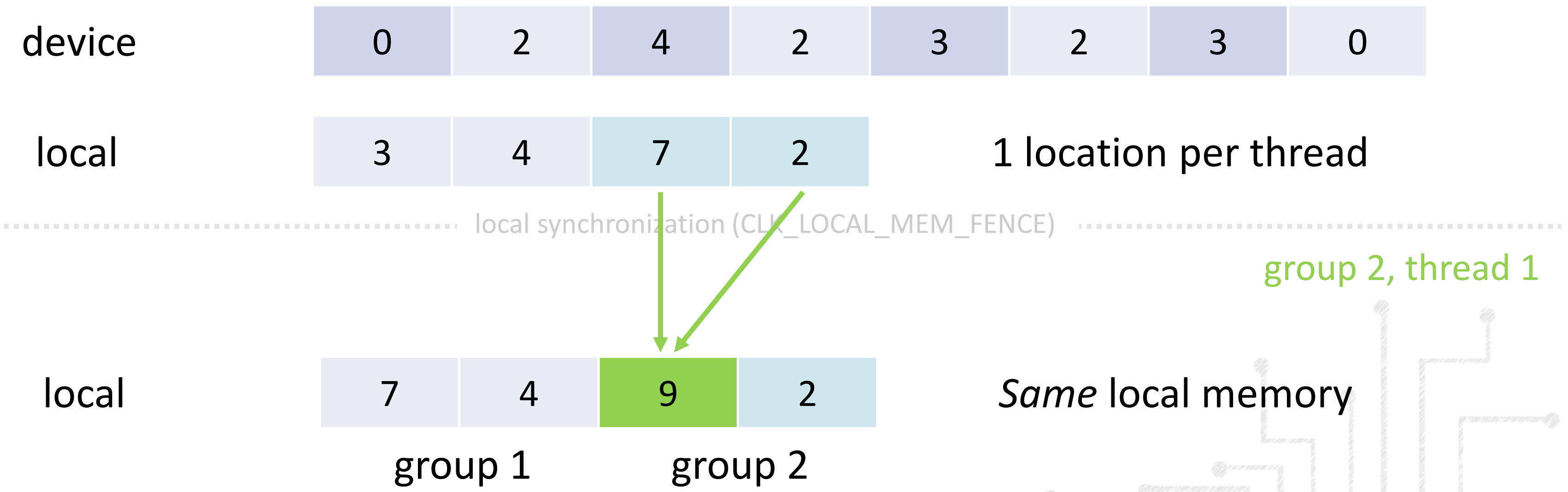
local	7	4	9	2
	group 1		group 2	

Same local memory



Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0
local	3	4	7	2				

1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

group 2, thread 1

local	7	4	9	2
	group 1		group 2	

Same local memory

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0
local	3	4	7	2				

1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2
	group 1		group 2	

Same local memory

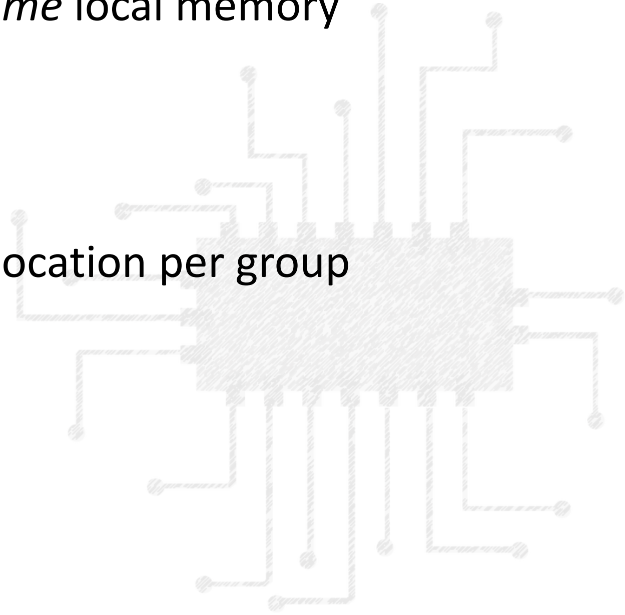
Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory
device					1 location per group



Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

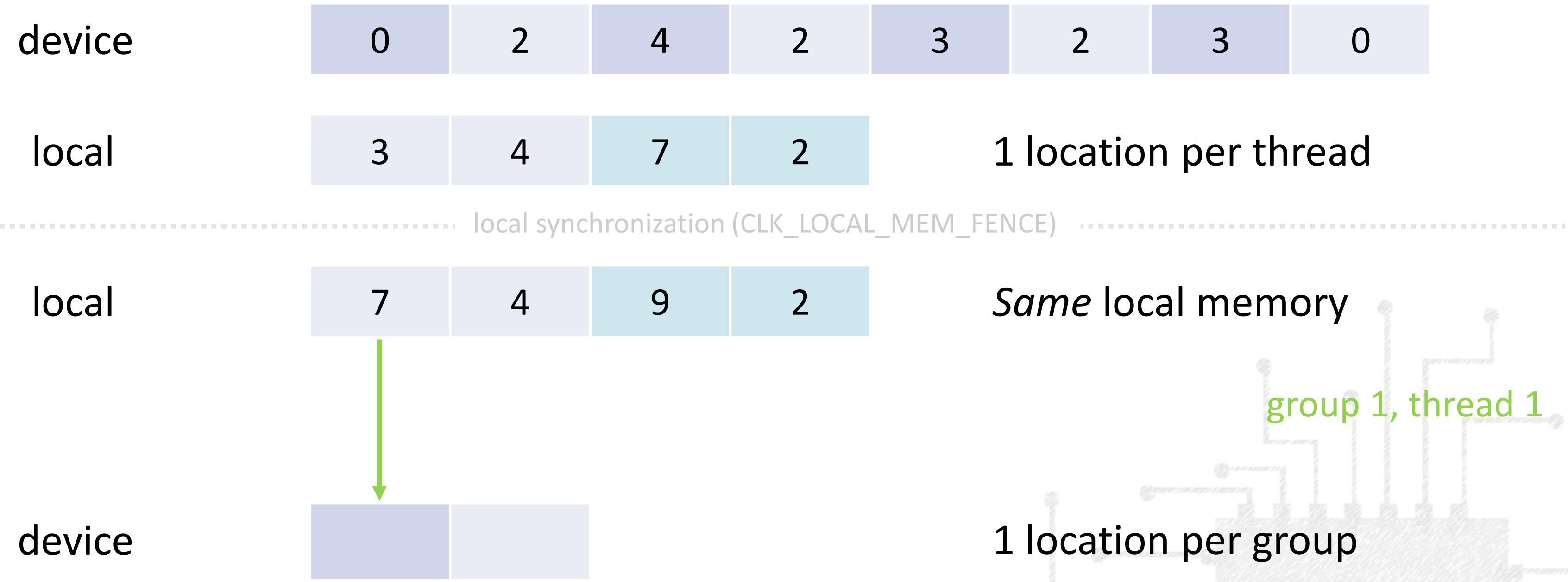
local	7	4	9	2	*Same* local memory

group 1, thread 1

device			1 location per group

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



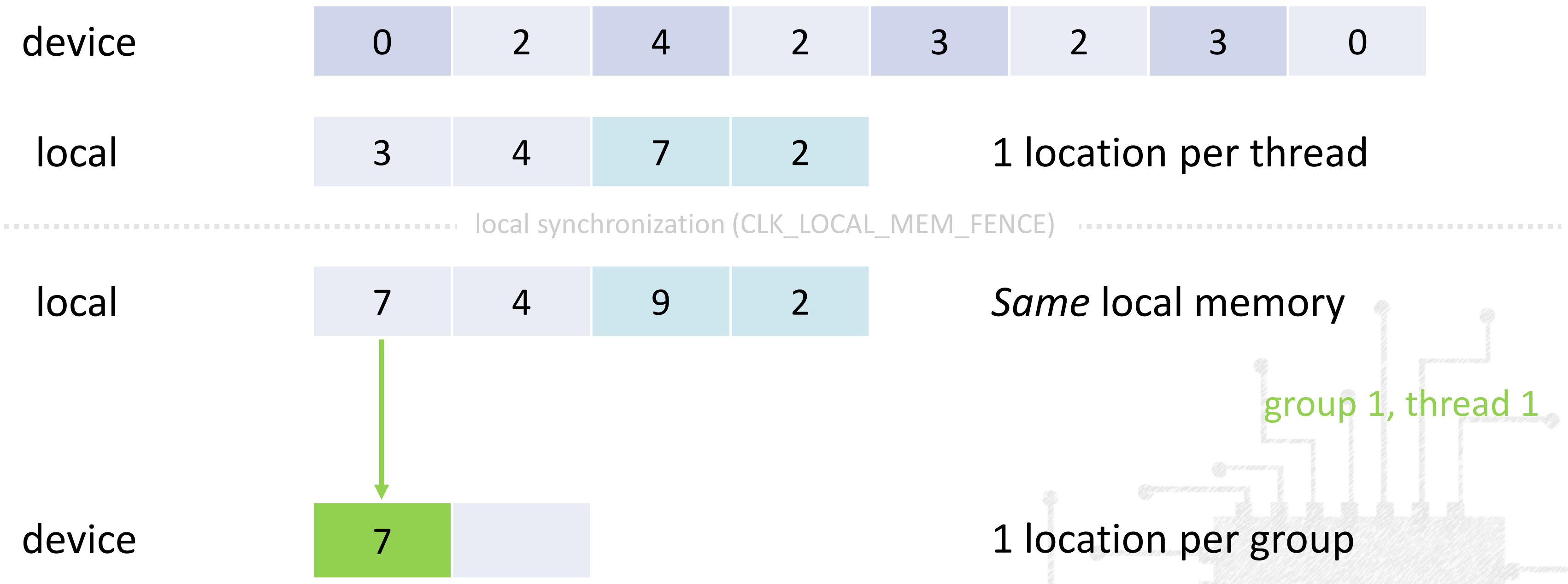
device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory
device					1 location per group

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory
device	7				1 location per group

group 1, thread 1

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

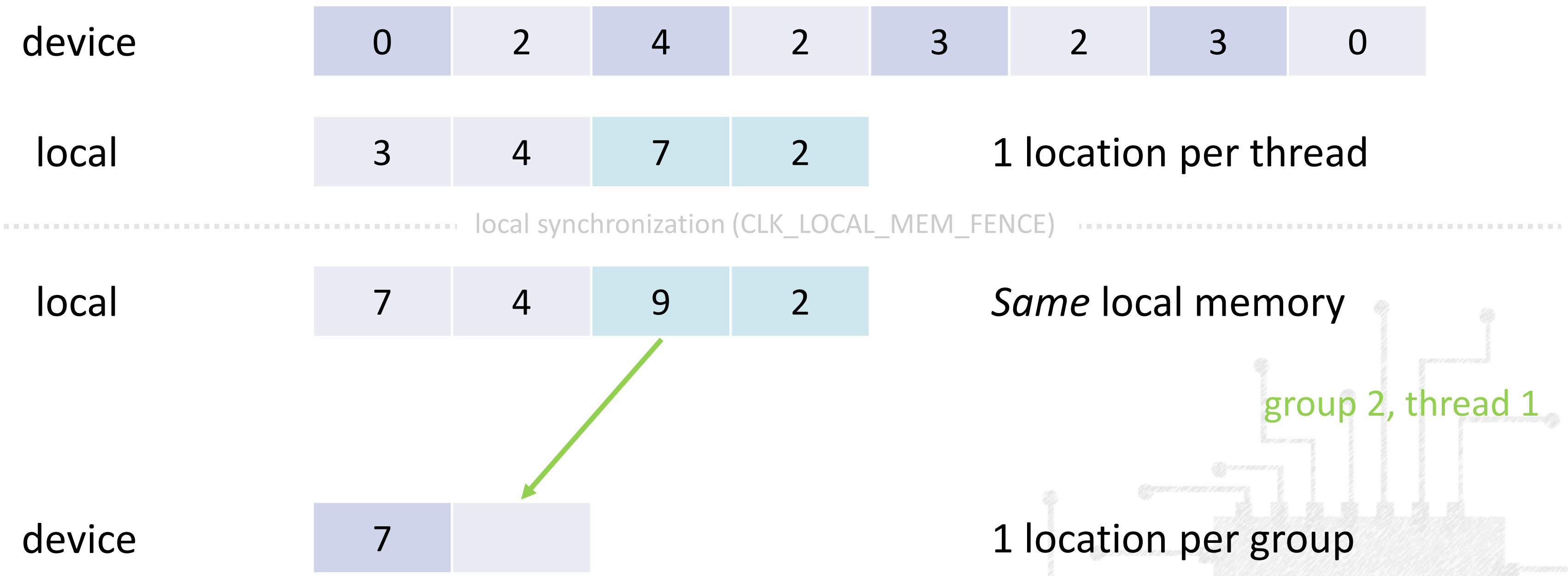
local	7	4	9	2	*Same* local memory

group 2, thread 1

device	7		1 location per group

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0

local	3	4	7	2	1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

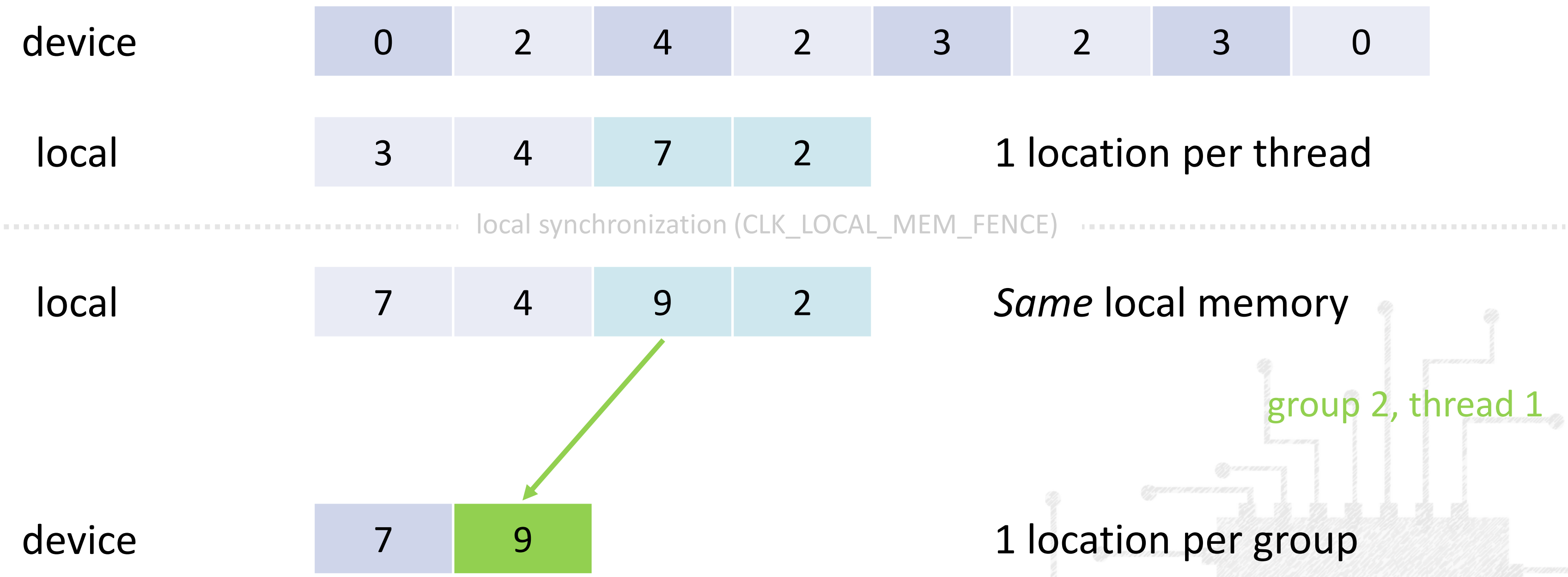
local	7	4	9	2	*Same* local memory

group 2, thread 1

device	7		1 location per group

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread
local	7	4	9	2					*Same* local memory
device	7	9							1 location per group

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

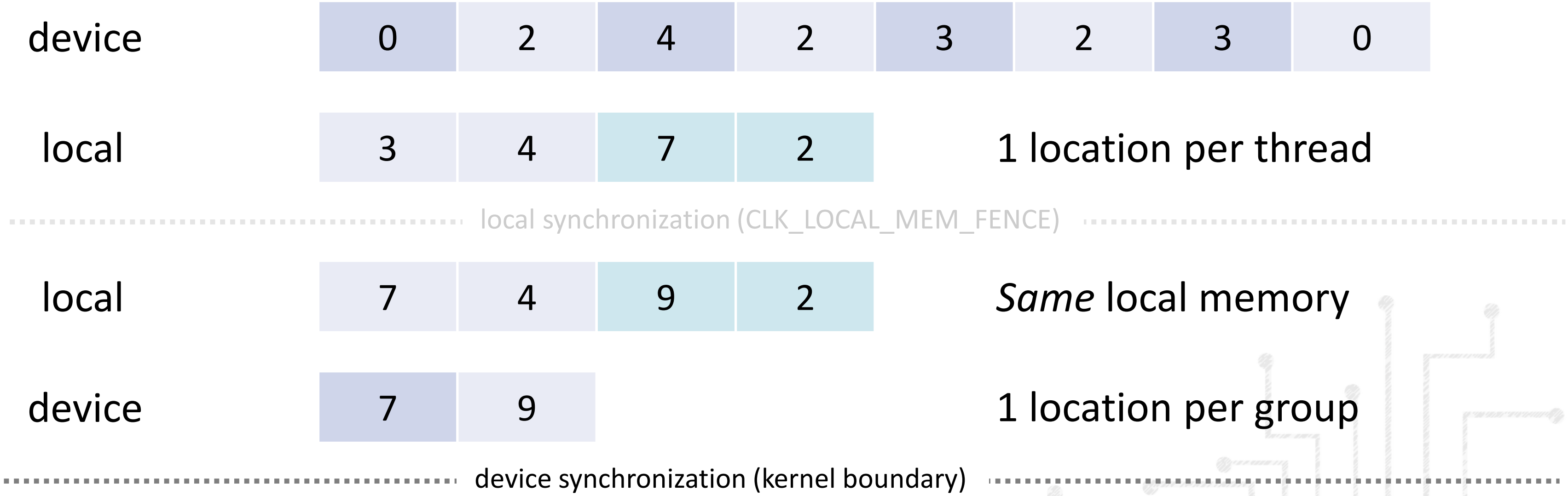
device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory
device	7	9			1 location per group

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



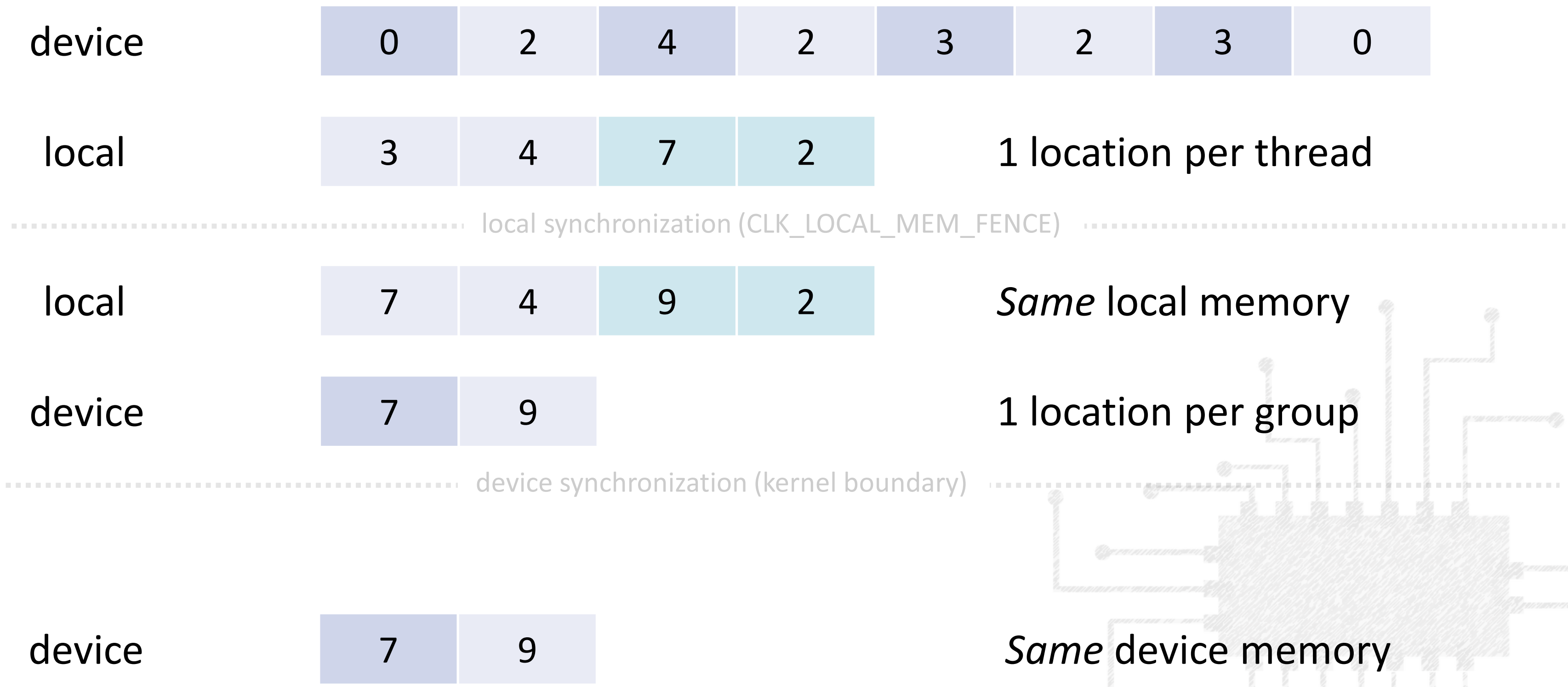
device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread
local	7	4	9	2					*Same* local memory
device	7	9							1 location per group

local synchronization (CLK_LOCAL_MEM_FENCE)

device synchronization (kernel boundary)

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device | 0 | 2 | 4 | 2 | 3 | 2 | 3 | 0

local | 3 | 4 | 7 | 2 — 1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local | 7 | 4 | 9 | 2 — *Same* local memory

device | 7 | 9 — 1 location per group

device synchronization (kernel boundary)

device | 7 | 9 — *Same* device memory

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

device	0	2	4	2	3	2	3	0

local	3	4	7	2	1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory

device	7	9	1 location per group

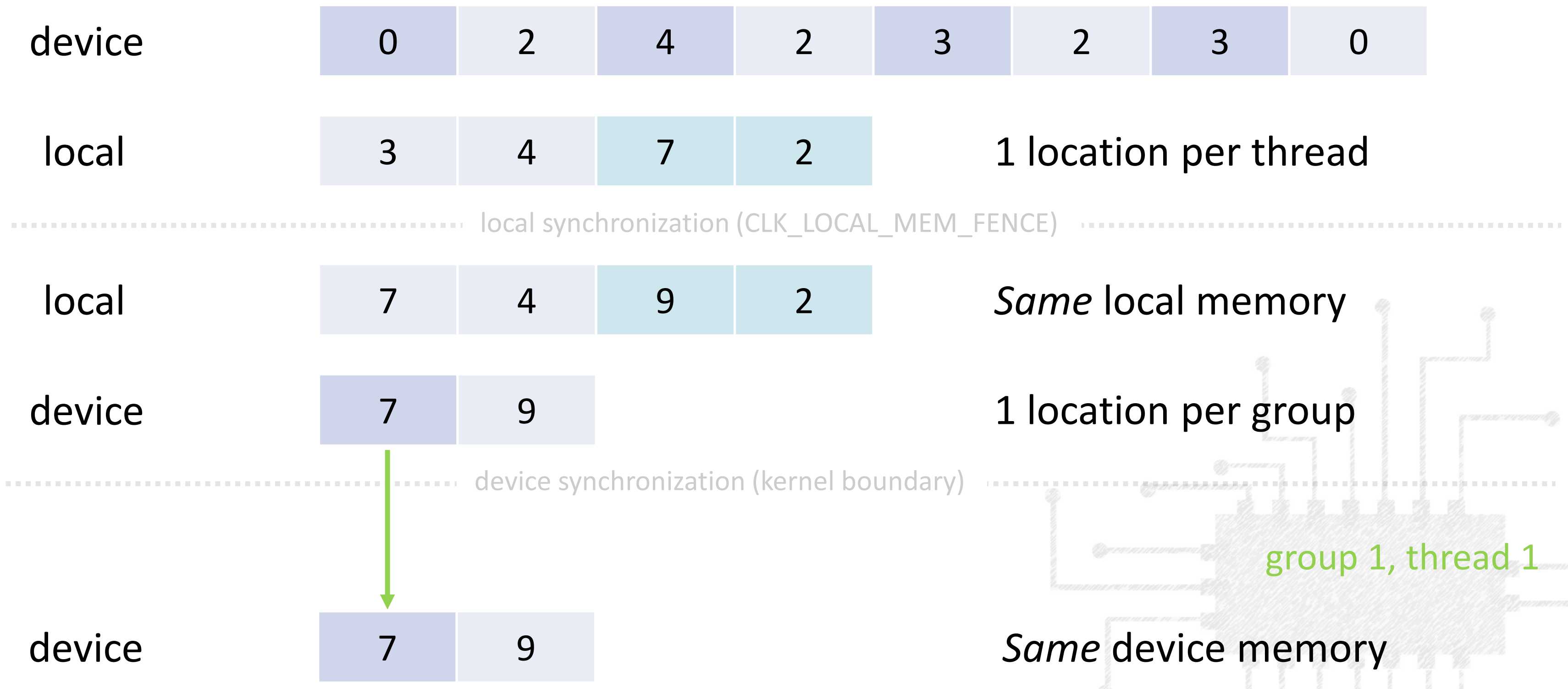
device synchronization (kernel boundary)

group 1, thread 1

device	7	9	*Same* device memory

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory
device	7	9			1 location per group

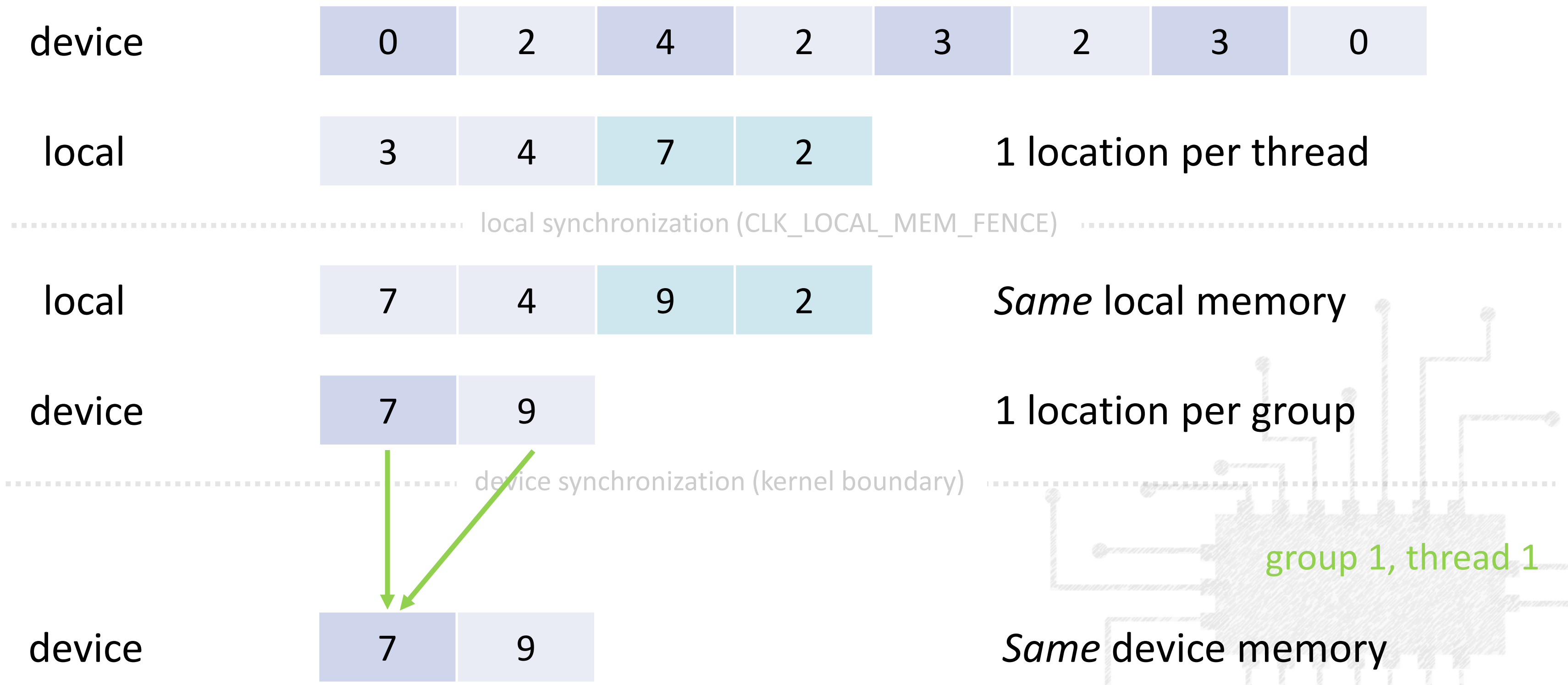
device synchronization (kernel boundary)

group 1, thread 1

device	7	9	*Same* device memory

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads

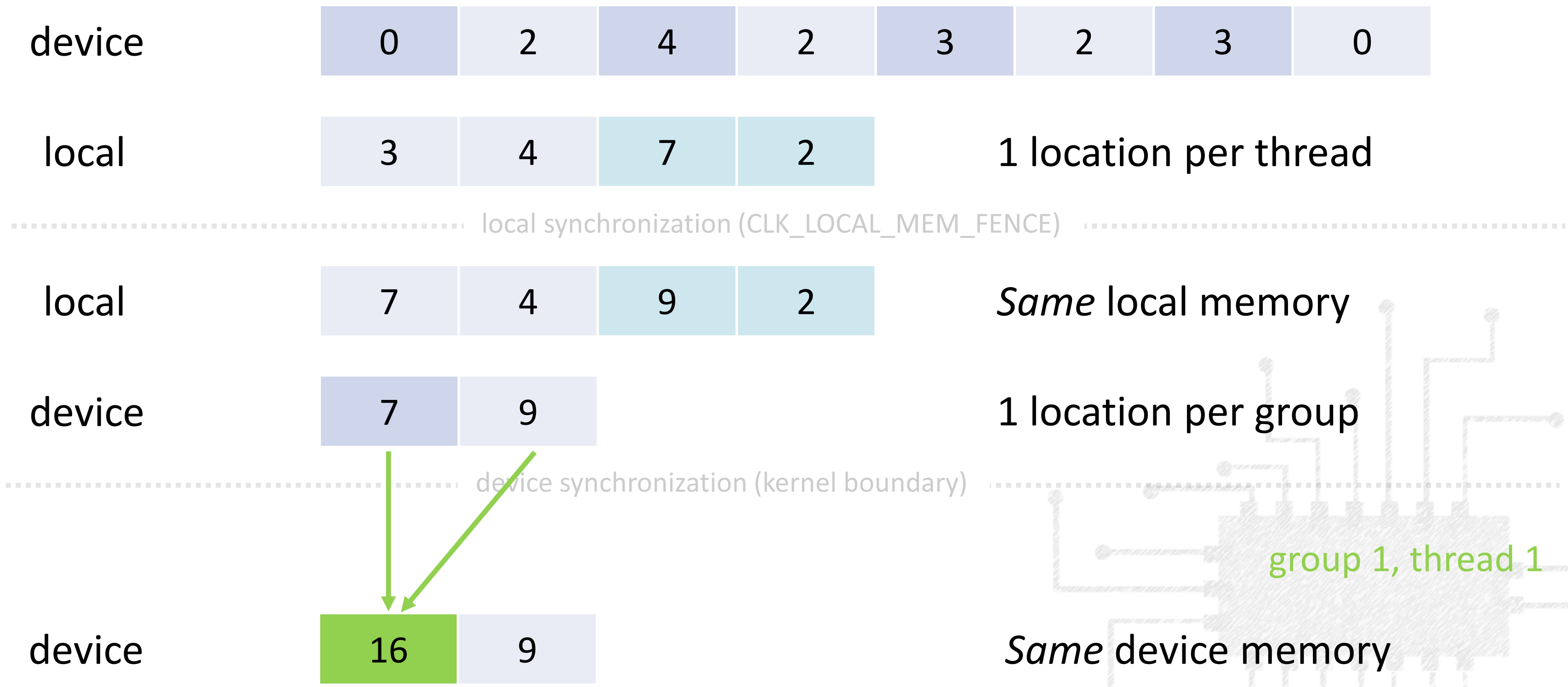


device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread
local	7	4	9	2					*Same* local memory
device	7	9							1 location per group
device	7	9							*Same* device memory

local synchronization (CLK_LOCAL_MEM_FENCE)
device synchronization (kernel boundary)
group 1, thread 1

Parallel Reductions (SUM)

2 thread groups, **2** threads/group = **4** threads



device	0	2	4	2	3	2	3	0	
local	3	4	7	2					1 location per thread

local synchronization (CLK_LOCAL_MEM_FENCE)

local	7	4	9	2	*Same* local memory
device	7	9			1 location per group

device synchronization (kernel boundary)

group 1, thread 1

device	16	9	*Same* device memory