GPU

Topics

- GPU hardware and its components
- Theoretical performance
- Actual performance measurement
- Application models

GPU Model

- Integrated GPUs

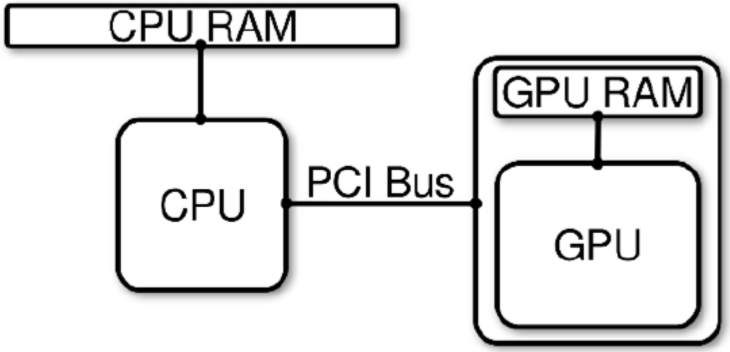
A graphics processor engine that is contained on the CPU

- Dedicated GPUS

A GPU on a separated peripheral card

Dedicated GPU Hardware

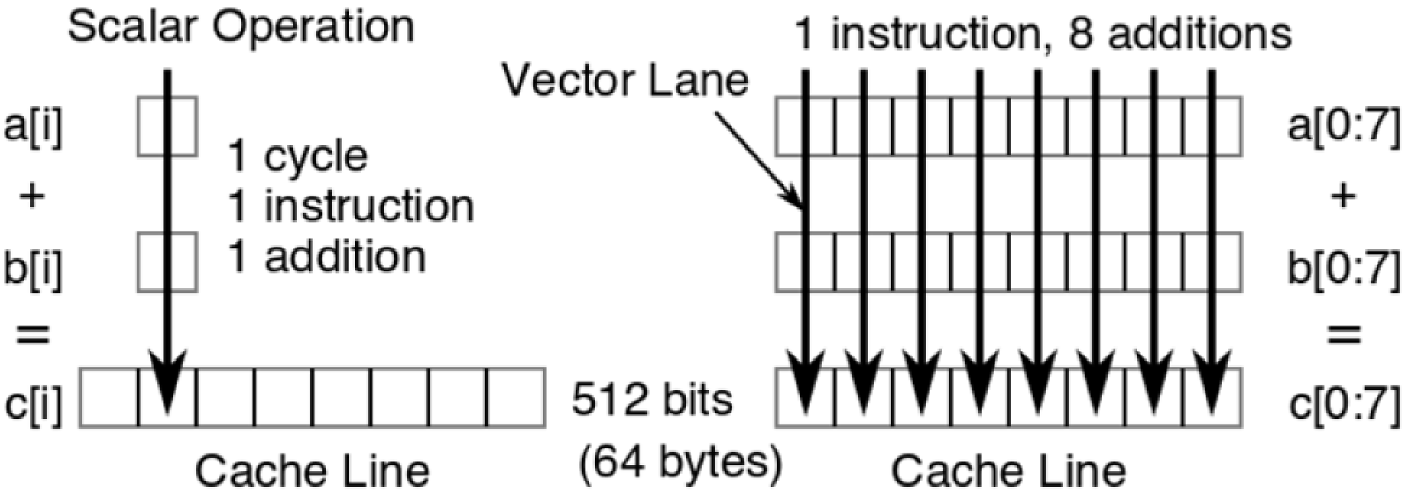
- Block diagram of GPU accelerated model
  - – CPU
  - – CPU Ram
  - – GPU
  - – GPU Ram
  - – PCI Bus



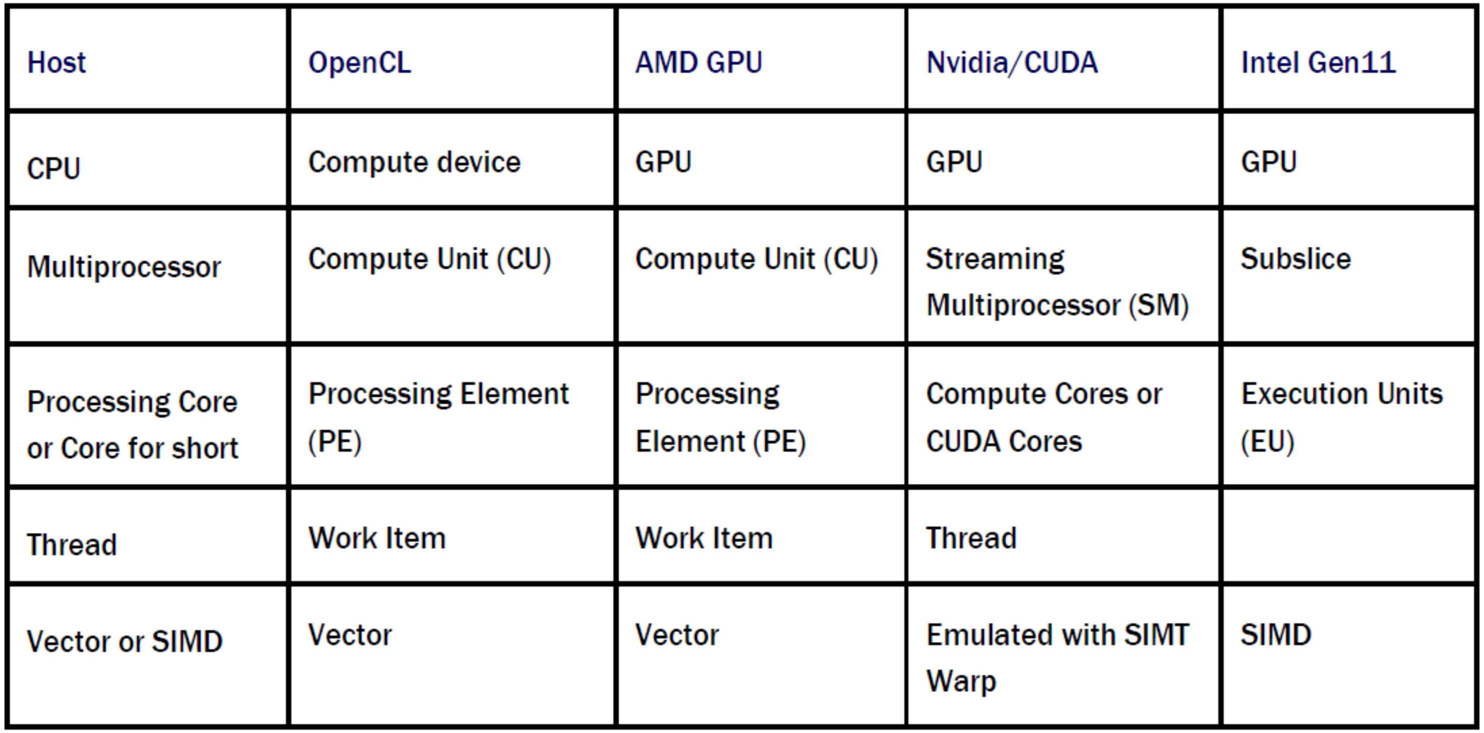
SIMD and Thread engine

- Single Instruction Multiple Data (SIMD)
- Thread engine
  - Large number of threads
  - Zero switching time
  - Latency hiding

SIMD Architecture



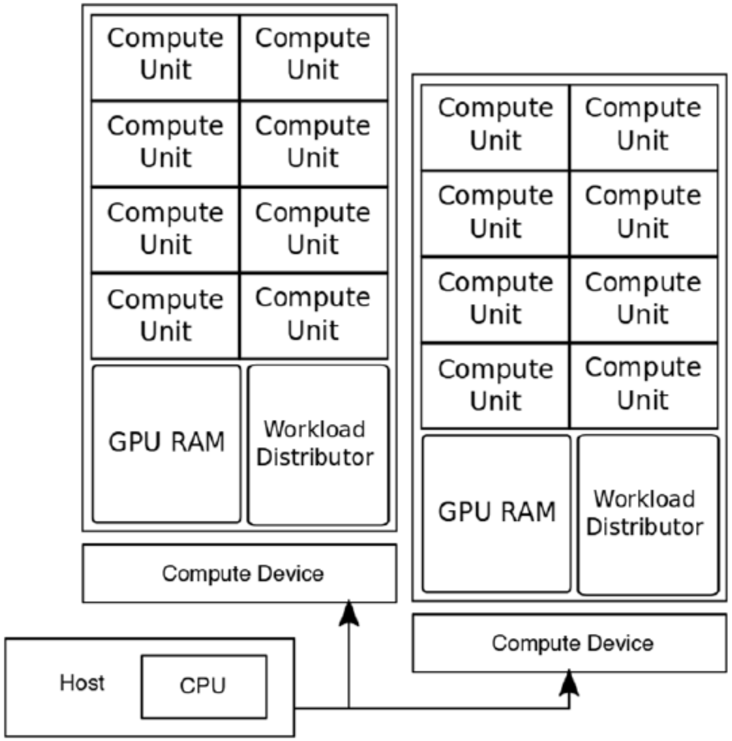
Hardware terminology



Host	OpenCL	AMD GPU	Nvidia/CUDA	Intel Gen11
CPU	Compute device	GPU	GPU	GPU
Multiprocessor	Compute Unit (CU)	Compute Unit (CU)	Streaming Multiprocessor (SM)	Subslice
Processing Core or Core for short	Processing Element (PE)	Processing Element (PE)	Compute Cores or CUDA Cores	Execution Units (EU)
Thread	Work Item	Work Item	Thread	
Vector or SIMD	Vector	Vector	Emulated with SIMT Warp	SIMD

Computing Unit

- It is term agreed by OpenCL standard
- Nvidia calls it Streaming multi processors(SMs)



Block diagram of a Computing Unit

- Each compute unit contains multiple Processing Elements (PEs)
- Each PE, it is composed of

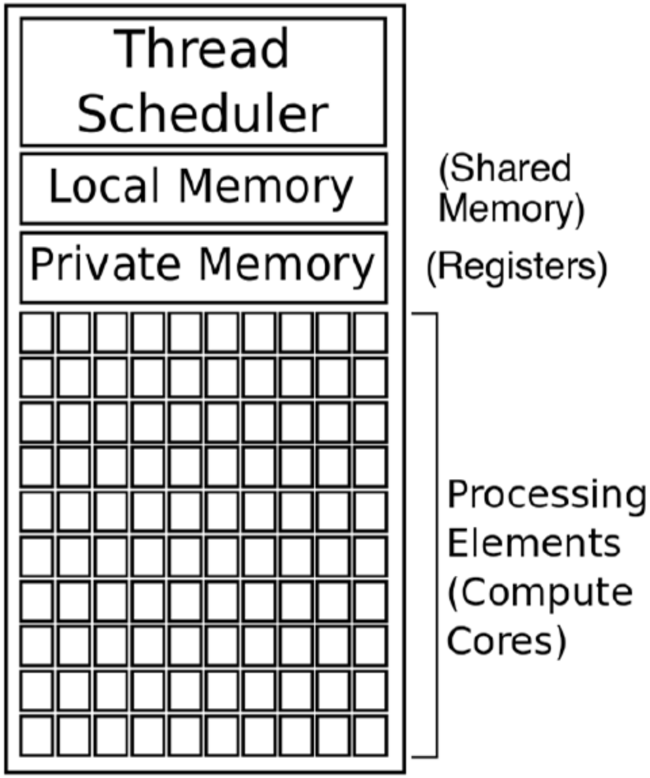
many functional units

Referred to as SIMT, SIMD,

or Vector operations by

ganging processing elements

together.



Thread Scheduler

Local Memory

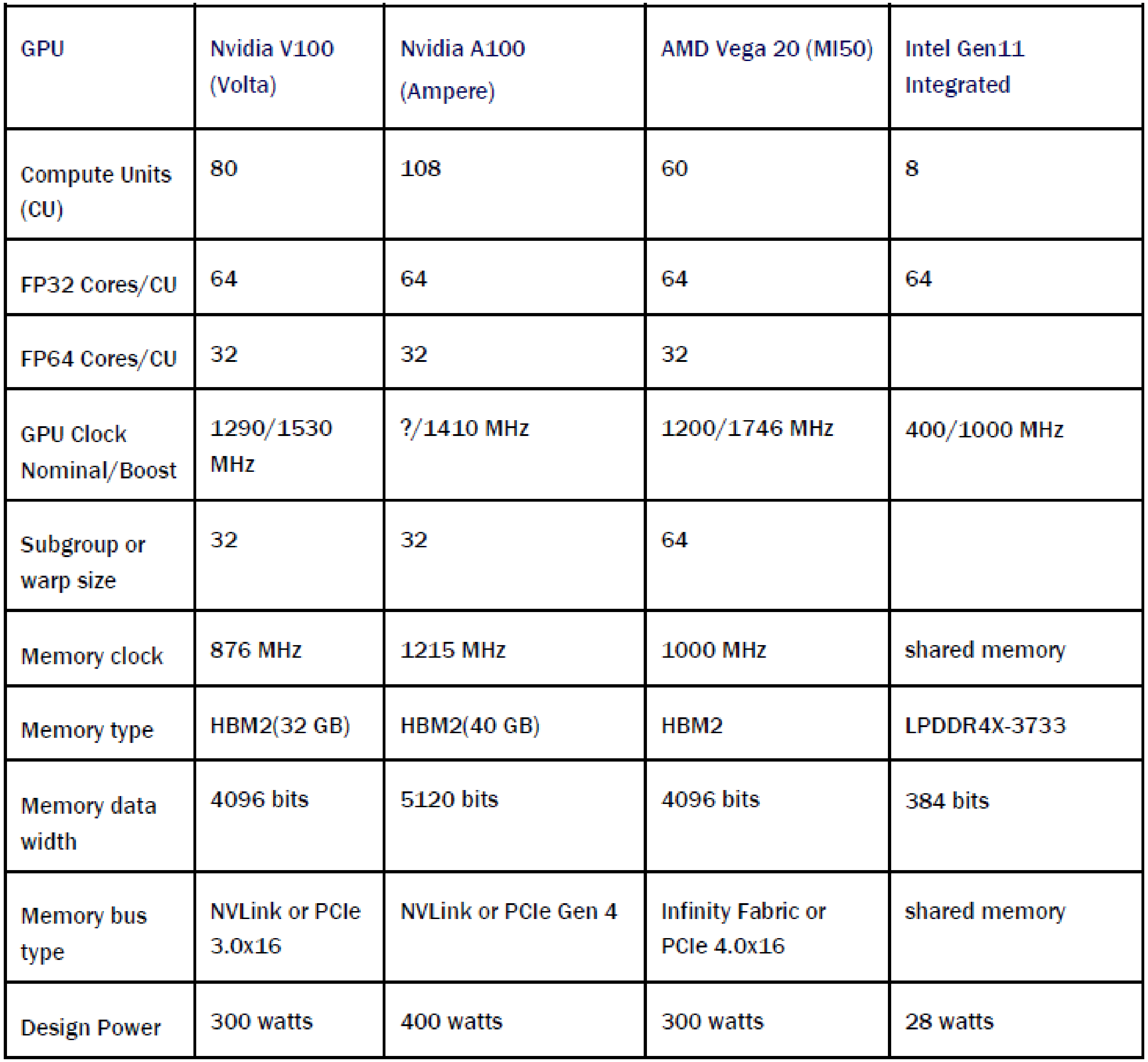
(Shared Memory)

Private Memory

(Registers)

Processing Elements (Compute Cores)

Hardware specification



GPU	Nvidia V100 (Volta)	Nvidia A100 (Ampere)	AMD Vega 20 (MI50)	Intel Gen11 Integrated
Compute Units (CU)	80	108	60	8
FP32 Cores/CU	64	64	64	64
FP64 Cores/CU	32	32	32	
GPU Clock Nominal/Boost	1290/1530 MHz	?/1410 MHz	1200/1746 MHz	400/1000 MHz
Subgroup or warp size	32	32	64	
Memory clock	876 MHz	1215 MHz	1000 MHz	shared memory
Memory type	HBM2(32 GB)	HBM2(40 GB)	HBM2	LPDDR4X-3733
Memory data width	4096 bits	5120 bits	4096 bits	384 bits
Memory bus type	NVLink or PCIe 3.0x16	NVLink or PCIe Gen 4	Infinity Fabric or PCIe 4.0x16	shared memory
Design Power	300 watts	400 watts	300 watts	28 watts

Theoretical Peak Flops

Peak Theoretical Flops (GFlops/s) =Clock rate MHz×Compute Units×Processing units ×Flops/cycle

- Nvidia v100: Single precision

  2 x 1530 x 80 x 64/ 10^6 = 15.6 TFlops
- Nvidia v100: Double precision

  2 x 1530 x 80 x 32/ 10^6 = 7.8 Tflops
- AMD Vega 20: Single precision

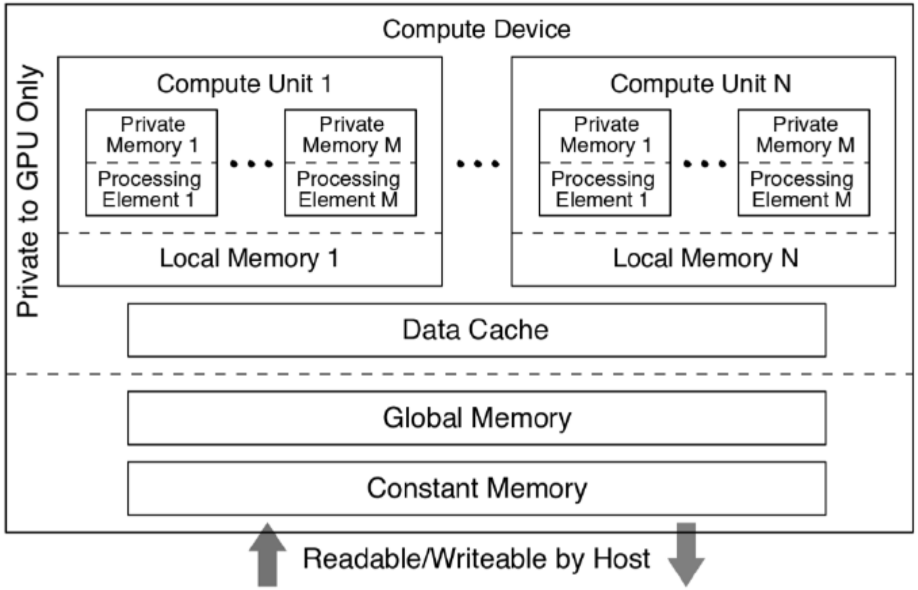
  2 x 1746 x 60 x 64/ 10^6 = 13.4 TFlops
- AMD Vega 20: Double precision

  2 x 1746 x 60 x 32/ 10^6 = 6.7 Tflops

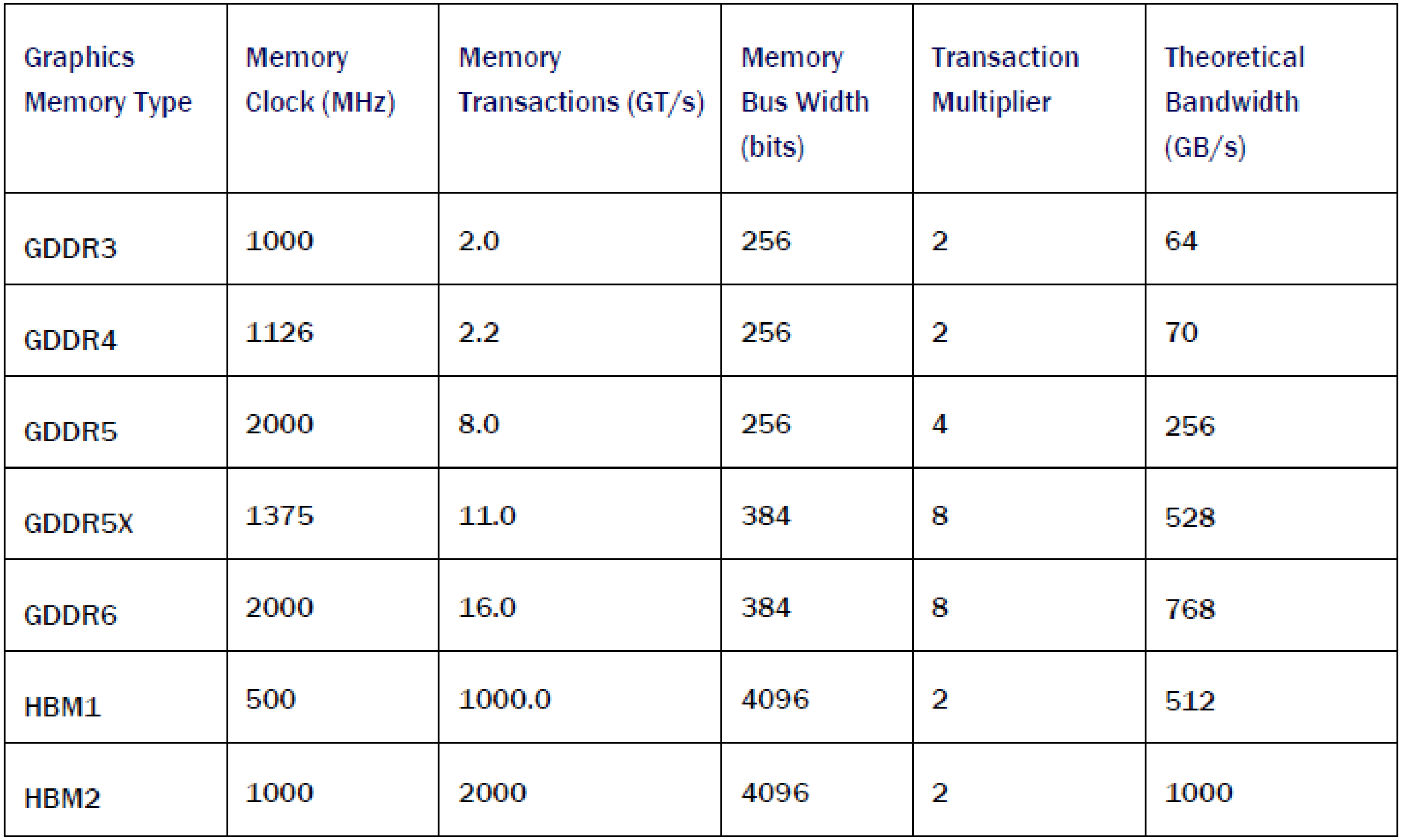
Fused Multiply and Add (=2 operations)

GPU Memory space

- Register memory (private memory)
- Local memory
- Constant memory
- Global memory



Calculating Peak Memory Bandwidth



Graphics Memory Type	Memory Clock (MHz)	Memory Transactions (GT/s)	Memory Bus Width (bits)	Transaction Multiplier	Theoretical Bandwidth (GB/s)
GDDR3	1000	2.0	256	2	64
GDDR4	1126	2.2	256	2	70
GDDR5	2000	8.0	256	4	256
GDDR5X	1375	11.0	384	8	528
GDDR6	2000	16.0	384	8	768
HBM1	500	1000.0	4096	2	512
HBM2	1000	2000	4096	2	1000

Theoretical memory bandwidth calculation

Theoretical bandwidth = Memory clock rate (GHz) * Memory bus (bits) * (1 byte/8bits) * transaction multiplier

Theoretical bandwidth = Memory Transaction rate (GHz) * Memory bus (bits) * (1 byte/8bits)

Theoretical memory bandwidth

- Nvidia V100

0.876 * 4096 * 1/8 * 2 = 897 GB/s

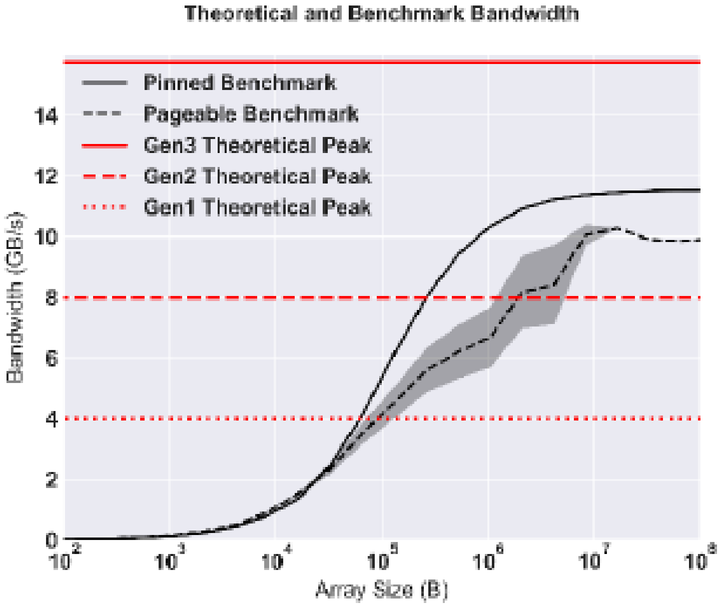
- AMD Radeon Vega20

1.00 * 4096 * 1/8 *2 = 1024 GB/s

Pinned / Pageable Memory

Pinned memory:
memory that cannot be page-out from ram and thus can be directly sent to the GPU

Pageable memory:
shared memory that can be paged-out to disk



The PCI bus

- CPU to GPU Data transfer overhead

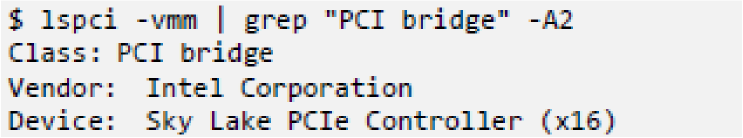
- The current version of the PCI bus is called PCI Express (PCIe)

- It has been revised from “generations” from 1.0 to 6.0

Theoretical bandwidth of the PCI bus

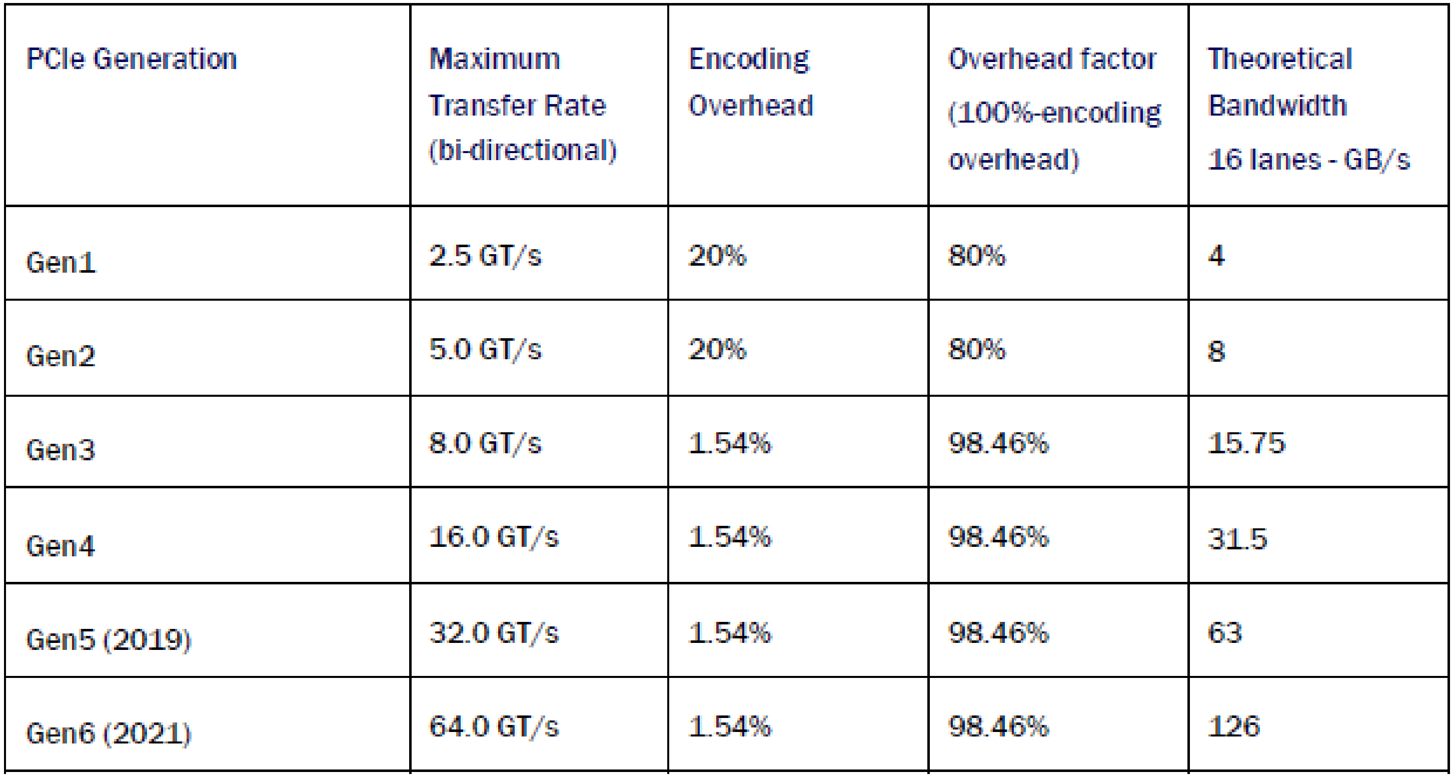
Theoretical Bandwidth (GB/s) = lanes x Transfer rate GT/s x Overhead factor (GB/GT) x byte/8bits

Command to check PCIe information



```
$ lspci -vmm | grep "PCI bridge" -A2
Class: PCI bridge
Vendor:  Intel Corporation
Device:  Sky Lake PCIe Controller (x16)
```

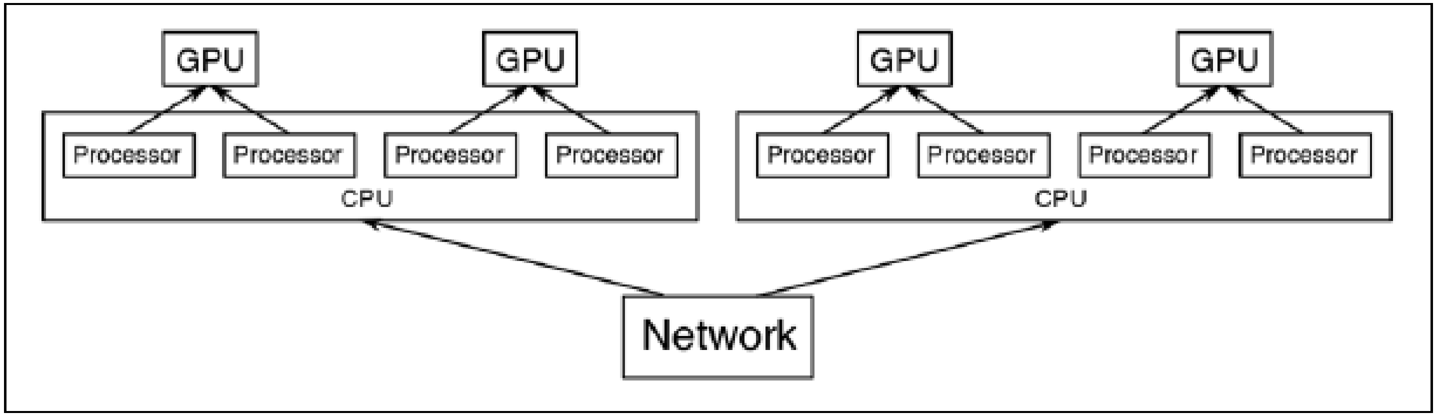
PCIe specification



PCIe Generation	Maximum Transfer Rate (bi-directional)	Encoding Overhead	Overhead factor (100%-encoding overhead)	Theoretical Bandwidth 16 lanes - GB/s
Gen1	2.5 GT/s	20%	80%	4
Gen2	5.0 GT/s	20%	80%	8
Gen3	8.0 GT/s	1.54%	98.46%	15.75
Gen4	16.0 GT/s	1.54%	98.46%	31.5
Gen5 (2019)	32.0 GT/s	1.54%	98.46%	63
Gen6 (2021)	64.0 GT/s	1.54%	98.46%	126

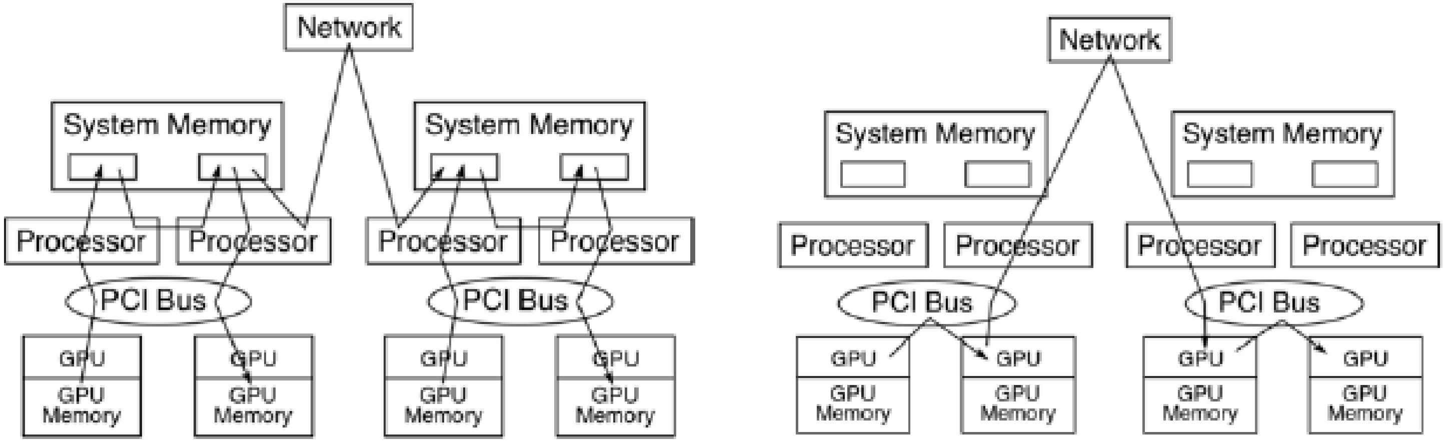
Multi-GPU Platform

- To further improve performance of the system, multi-GPU are sometimes used together



GPU Direct

- CUDA adds the capability for the GPU to send data in a message
- AMD has a similar capability called DirectGMA

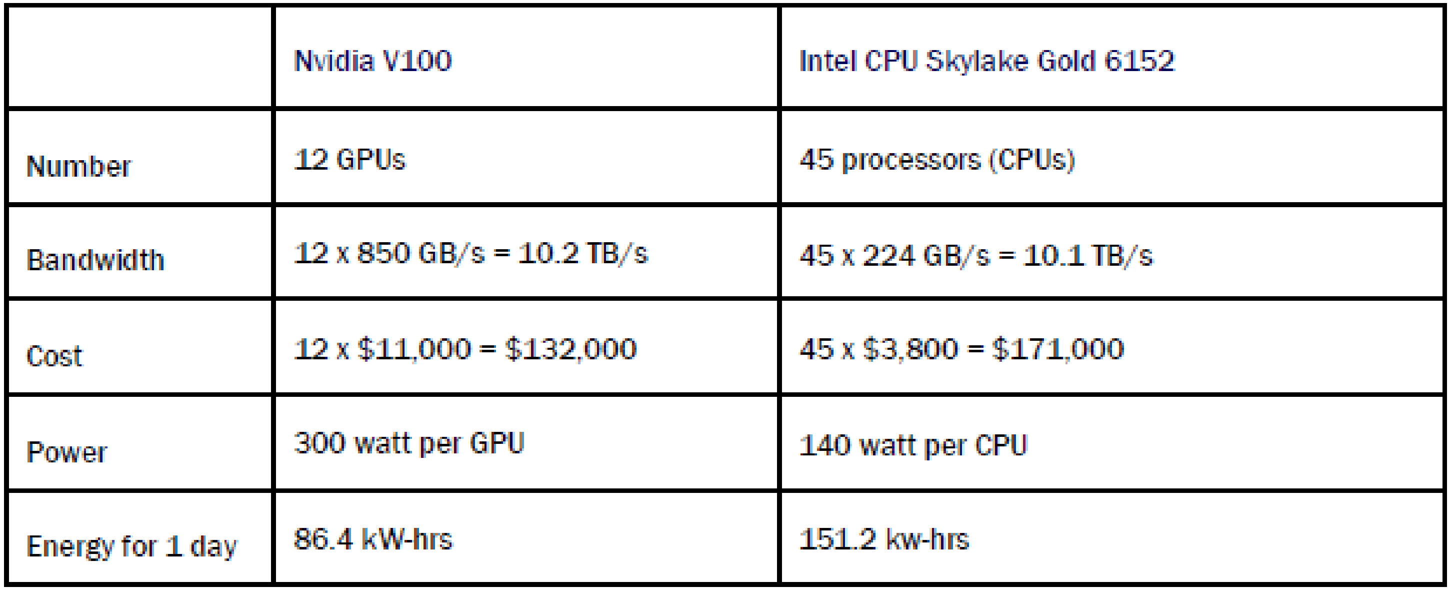


Energy consumption

- The estimation of Energy consumption

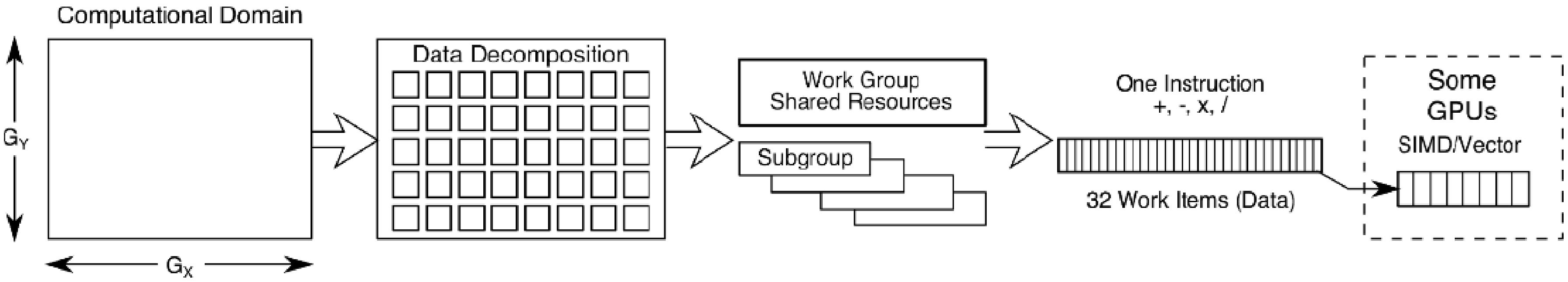
$$\mathit{Energy} = (\mathit{N\ Processors}) \times (\mathit{R\ Watts/Processor}) \times (\mathit{T\ hours})$$

Example of energy calculation



	Nvidia V100	Intel CPU Skylake Gold 6152
Number	12 GPUs	45 processors (CPUs)
Bandwidth	12 x 850 GB/s = 10.2 TB/s	45 x 224 GB/s = 10.1 TB/s
Cost	12 x $11,000 = $132,000	45 x $3,800 = $171,000
Power	300 watt per GPU	140 watt per CPU
Energy for 1 day	86.4 kW-hrs	151.2 kw-hrs

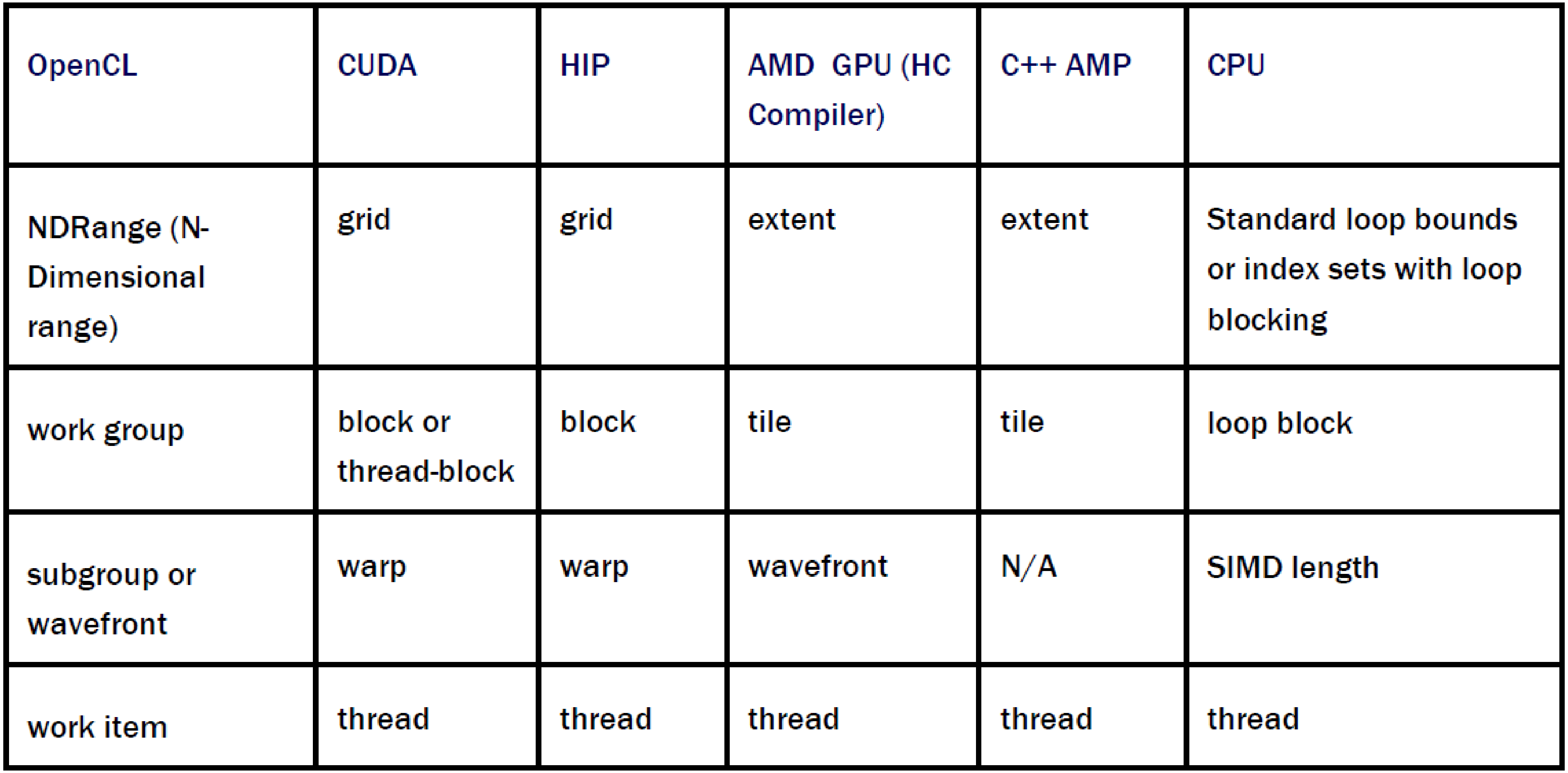
GPU Programming Model



Programming model

- Data decomposition
- Chunk-sized work for processing with some shared, local memory
- Operating on multiple data items with a single instruction
- Vectorization (on some GPUs)

Programming abstractions



OpenCL	CUDA	HIP	AMD GPU (HC Compiler)	C++ AMP	CPU
NDRange (N-Dimensional range)	grid	grid	extent	extent	Standard loop bounds or index sets with loop blocking
work group	block or thread-block	block	tile	tile	loop block
subgroup or wavefront	warp	warp	wavefront	N/A	SIMD length
work item	thread	thread	thread	thread	thread

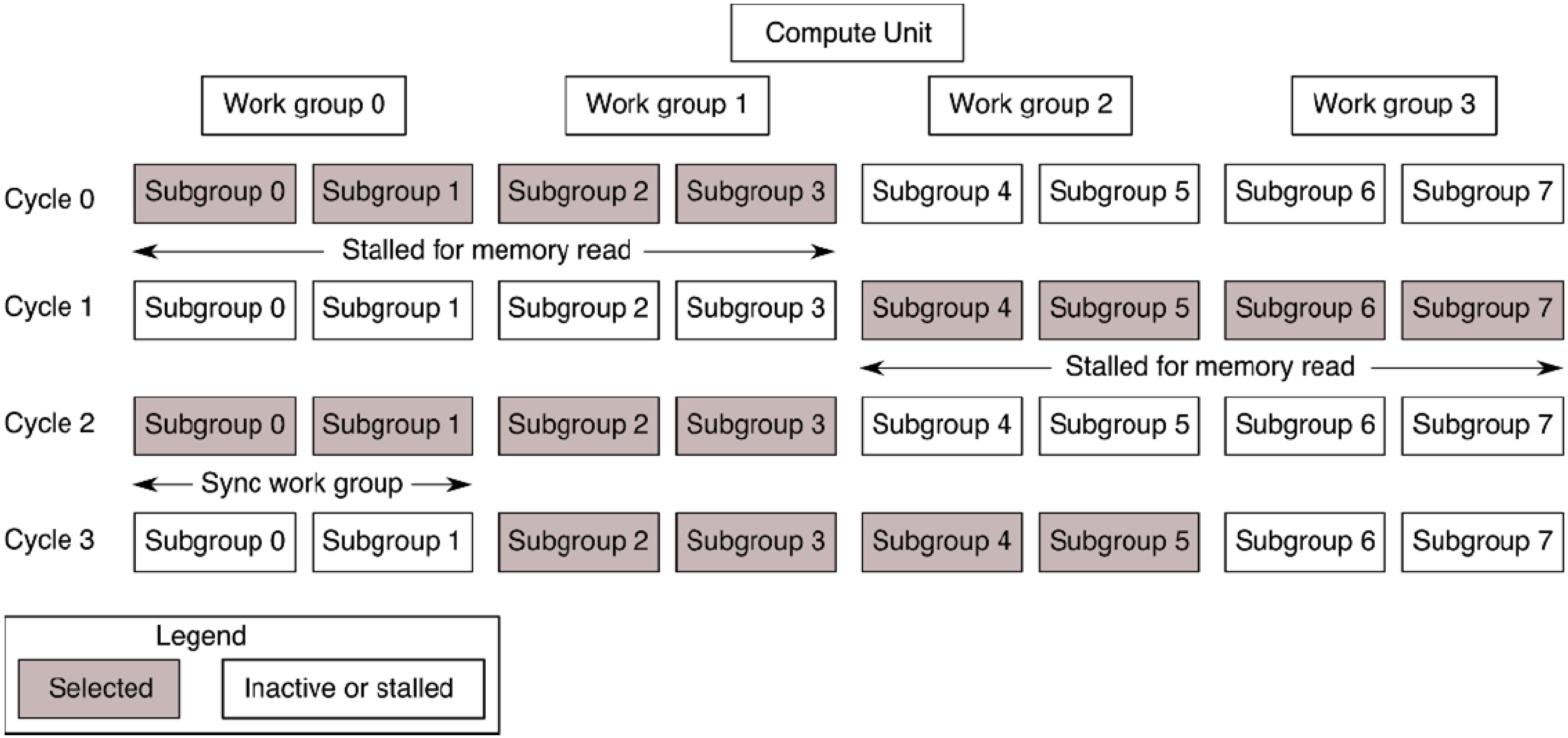
Work model

From the task, it can be broken down into workgroup.

Each workgroup will be composed of multiple subgroups or wavefronts

Each subgroup is composed of multiple work item

Data decomposition



Compute Unit

	Work group 0		Work group 1		Work group 2		Work group 3	
Cycle 0	▨ Subgroup 0	▨ Subgroup 1	▨ Subgroup 2	▨ Subgroup 3	□ Subgroup 4	□ Subgroup 5	□ Subgroup 6	□ Subgroup 7
	← Stalled for memory read →							
Cycle 1	□ Subgroup 0	□ Subgroup 1	□ Subgroup 2	□ Subgroup 3	▨ Subgroup 4	▨ Subgroup 5	▨ Subgroup 6	▨ Subgroup 7
					← Stalled for memory read →			
Cycle 2	▨ Subgroup 0	▨ Subgroup 1	▨ Subgroup 2	▨ Subgroup 3	□ Subgroup 4	□ Subgroup 5	□ Subgroup 6	□ Subgroup 7
	← Sync work group →							
Cycle 3	□ Subgroup 0	□ Subgroup 1	▨ Subgroup 2	▨ Subgroup 3	▨ Subgroup 4	▨ Subgroup 5	□ Subgroup 6	□ Subgroup 7

Legend

▨ Selected

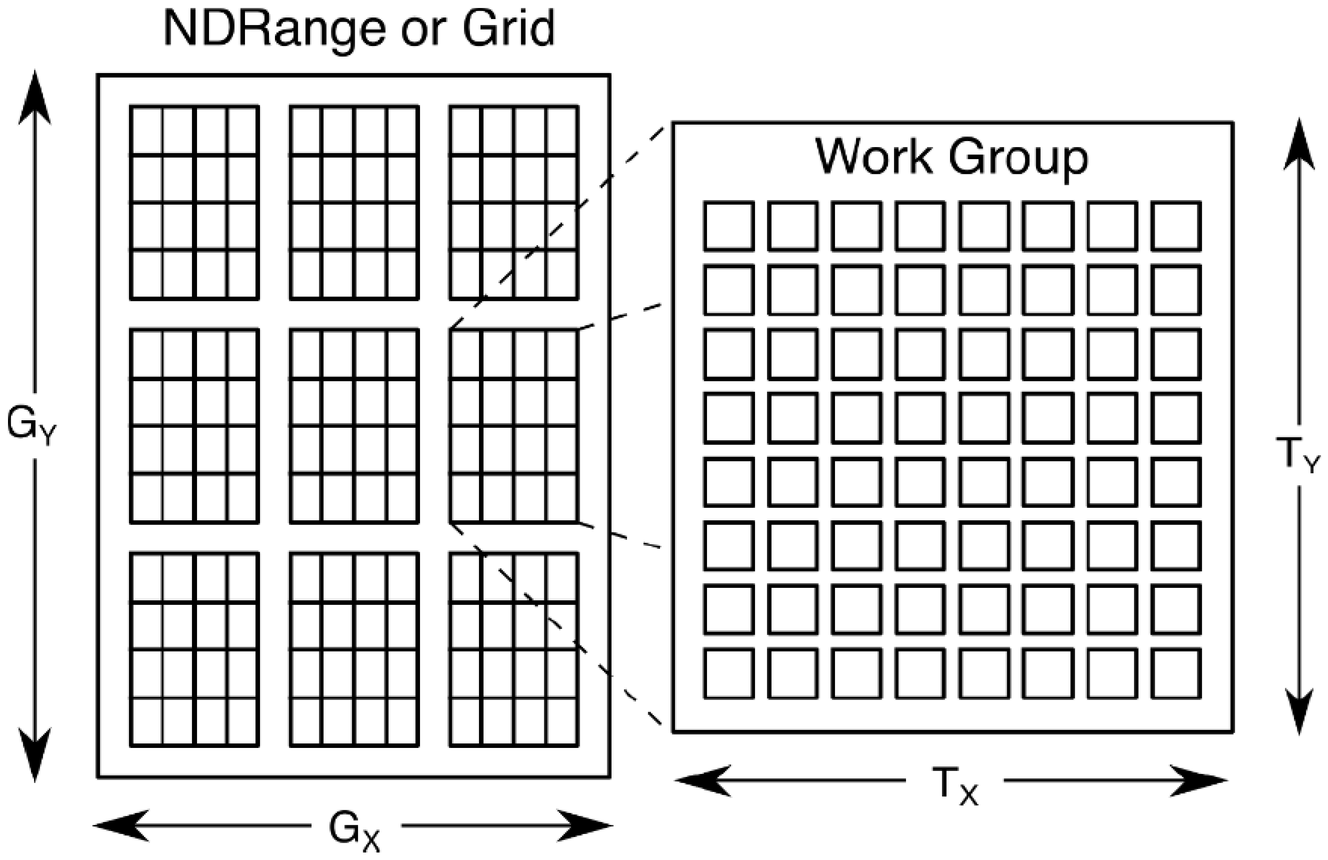
□ Inactive or stalled

GPU subgraph limit

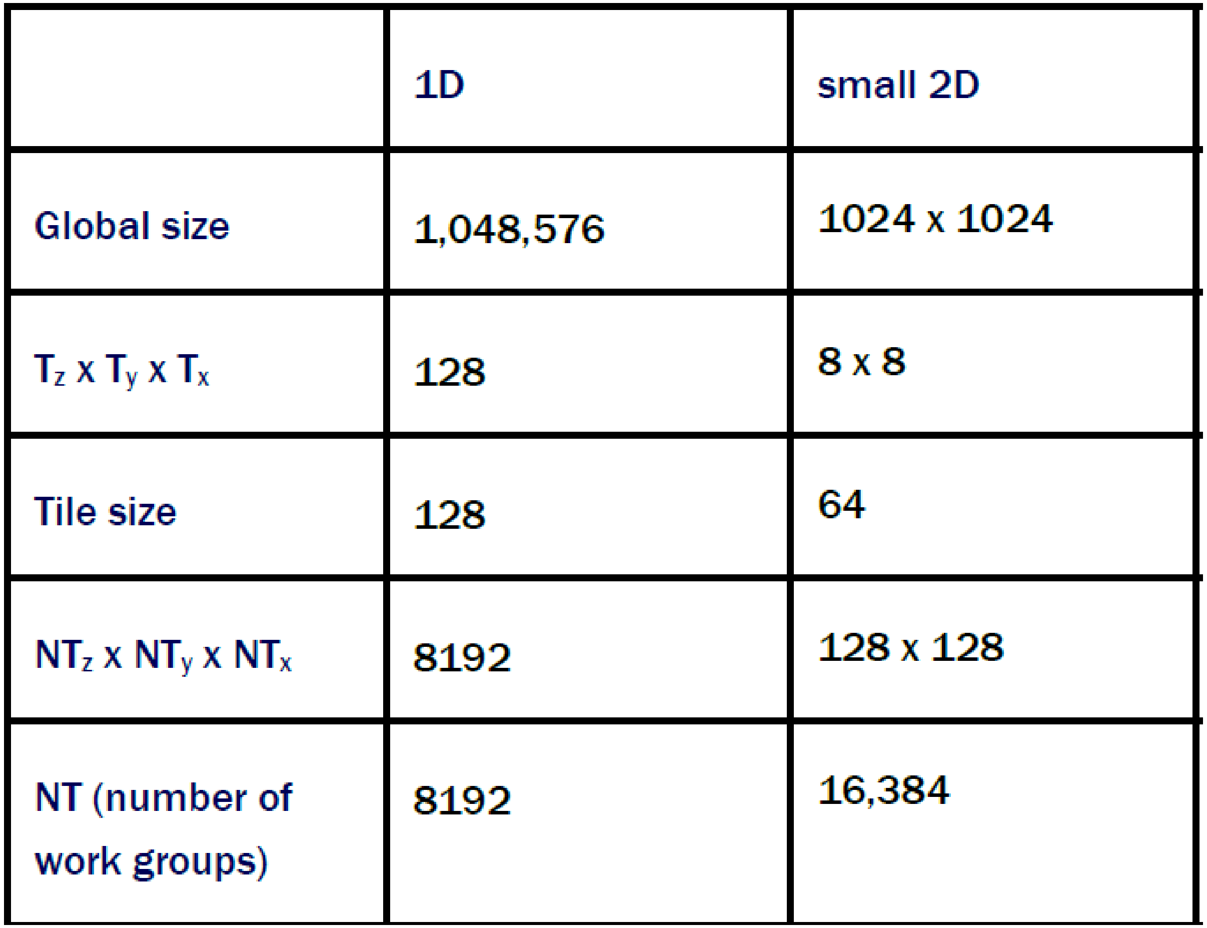
	Nvidia Volta and Ampere	AMD MI50
Active number of subgroups per compute unit	64	40
Active number of work groups per compute unit	32	40
Selected subgroups for execution per compute unit	4	4
Subgroup (warp or wavefront) size	32	64

Example

- Suppose we have data decomposition of a 1024 x 1024



Workgroup setup

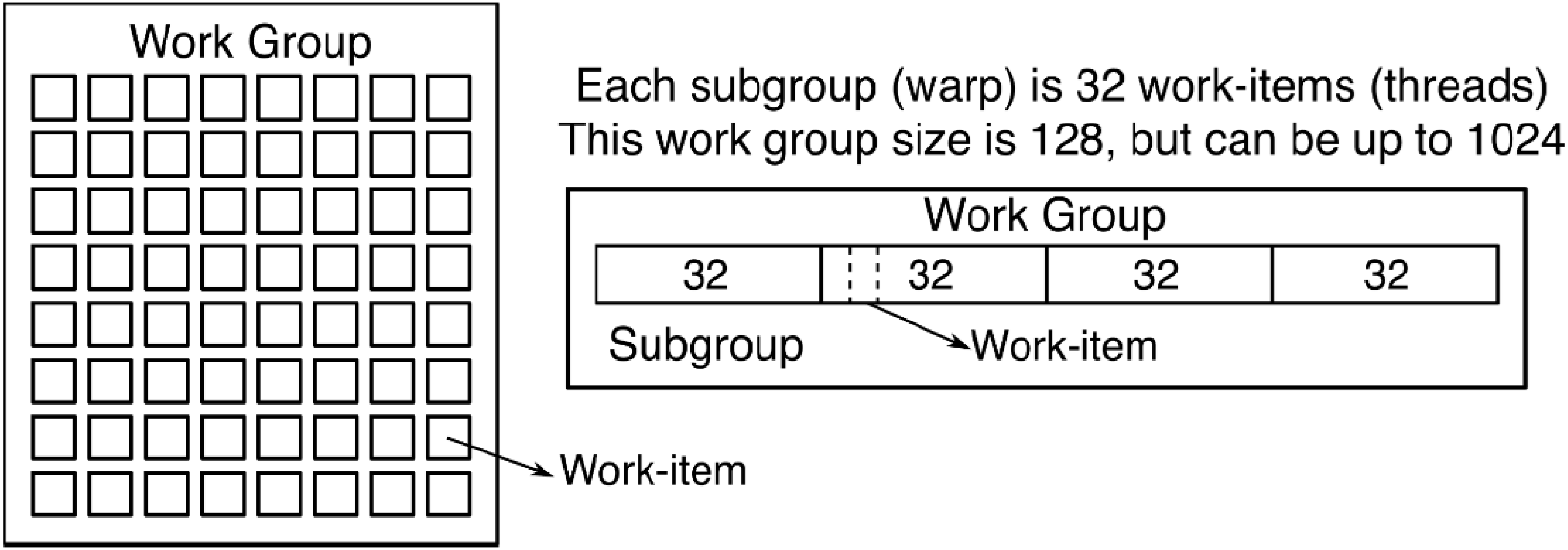


	1D	small 2D
Global size	1,048,576	1024 x 1024
T_z x T_y x T_x	128	8 x 8
Tile size	128	64
NT_z x NT_y x NT_x	8192	128 x 128
NT (number of work groups)	8192	16,384

Workgroup and subgroup

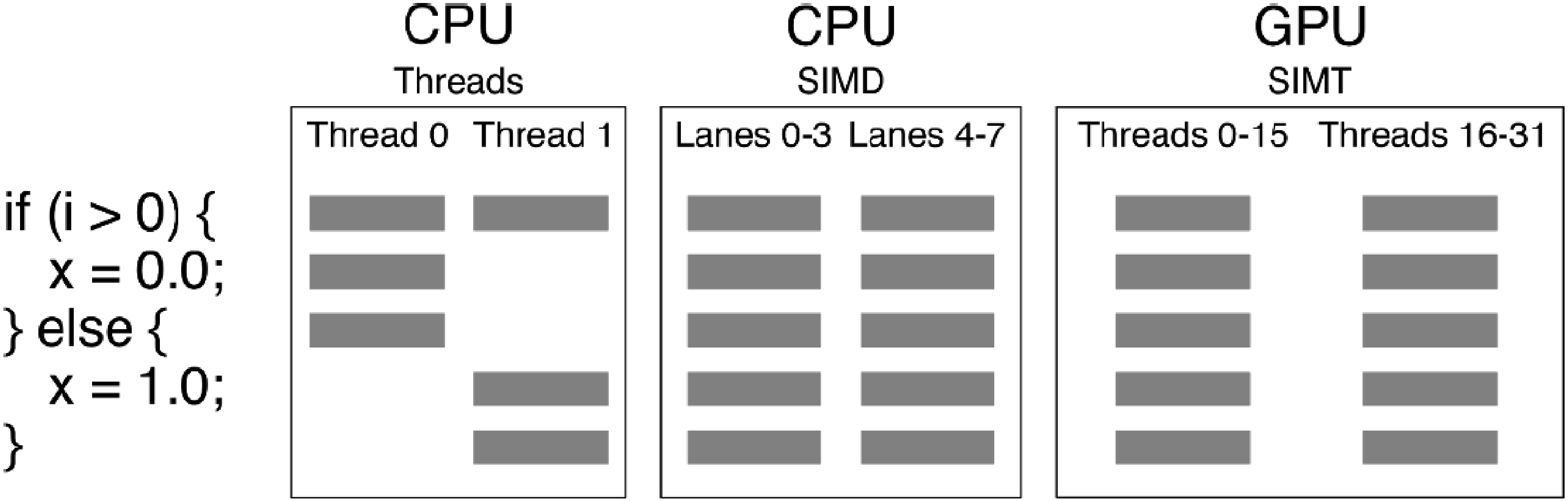
Characteristics of work groups on GPUs are:

- Cycles through processing each subgroup
- Has local memory and other resources shared within the group
- Can synchronize within a work group or a subgroup



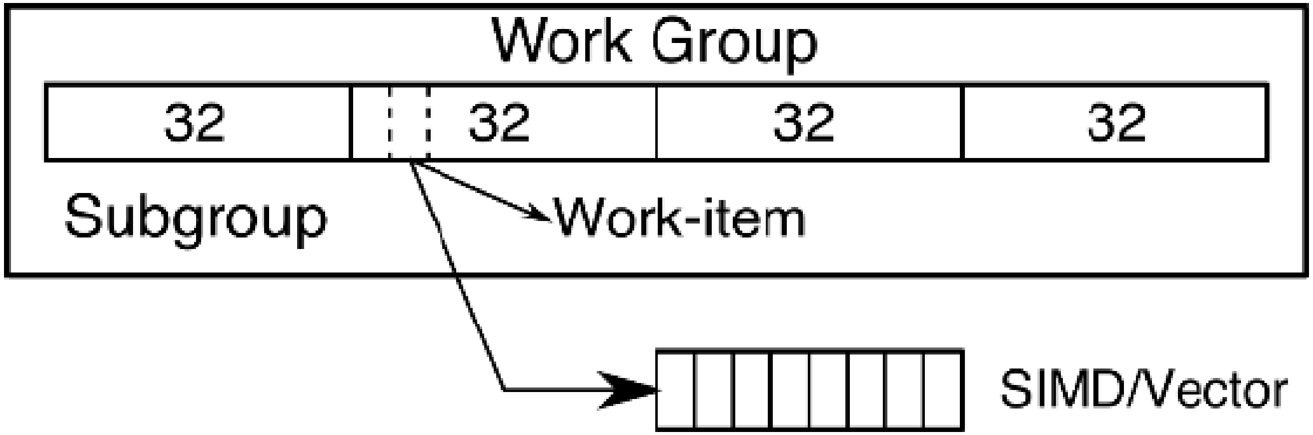
Each subgroup (warp) is 32 work-items (threads)
This work group size is 128, but can be up to 1024

Threads, SIMD, SIMT



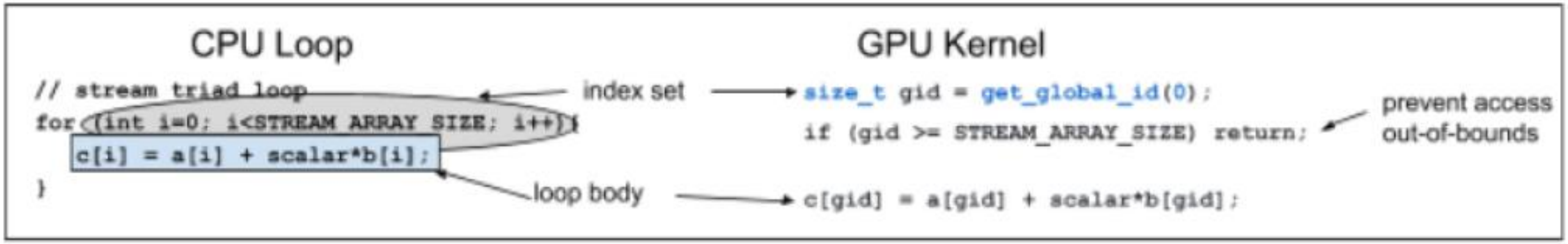
```
if (i > 0) {
   x = 0.0;
} else {
   x = 1.0;
}
```

Work Item



Loop and Kernel Code

- CPU code vs. GPU code



CPU Loop

```
// stream triad loop
for (int i=0; i<STREAM_ARRAY_SIZE; i++){
   c[i] = a[i] + scalar*b[i];
}
```

GPU Kernel

```
size_t gid = get_global_id(0);
if (gid >= STREAM_ARRAY_SIZE) return;

c[gid] = a[gid] + scalar*b[gid];
```

Questions?