The Allocated Active Blocks per SM (AABS)

- Equal importance to the occupancy.
- Directly affects the occupancy (utilization).
- The table shows the hardware limits (Ampere arch).
- I will only highlight four key rows for our purpose.
- The first row :Max Thread blocks/SM.

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB

The Allocated Active Blocks per SM (AABS)

- The first row :Max Thread blocks/SM.
- No more than 32 blocks per SM can run concurrently.
- The sec row :Max warps/SM.
- Max warps that can be executed simultaneously on a single SM.
- Regardless of the number of blocks allocated per SM
  - The summation of all warps in these blocks cannot exceed 64
- Warp limit controls how many active blocks can be allocated per SM

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB

The Allocated Active Blocks per SM (AABS)

- Example:
  - 32 blocks per SM .
  - Each block contains 4 warps.
  - The total warps would be 128 warps.

Solution

- The compiler will intervene with the following logic
- Decrease the Allocated blocks per SM to 16.
- With this adjustment, we didn't exceed the warp limit.

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB

The Allocated Active Blocks per SM (AABS)

- Term “Blocks limit warps”.
- The relation between the Occupancy and the AABS.
- We can assign millions blocks / GPU.
- Those blocks will be distributed across 108 SMs.
- We can assign more than 1000 blocks per SM.
- No more 32 blocks can run concurrently.

```
Section: Occupancy
------------------------------ ----------- ------------
Metric Name                    Metric Unit Metric Value
------------------------------ ----------- ------------
Block Limit SM                       block           16
Block Limit Registers                block           42
Block Limit Shared Mem               block           16
Block Limit Warps                    block           16
Theoretical Active Warps per SM       warp           48
Theoretical Occupancy                    %          100
Achieved Occupancy                       %        74.00
Achieved Active Warps Per SM          warp        35.52
------------------------------ ----------- ------------
```

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB

The Allocated Active Blocks per SM (AABS)

- The third row :max registers/SM.
- No more than 65536 registers/SM.
- Example:
  - One block per SM with a size of 1024 threads.
  - Each thread requires 100 registers.
  - More than 100,000 registers/SM are required.
  - The limit is 65536. (This is a problem).
  - We can’t decrease the number of blocks per SM.
  - Decrease the number of threads per block.
  - Otherwise: Local memory will be used.
  - Local memory causes performance degradation.
  - Because Local memory is lower than registers.

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB

The Allocated Active Blocks per SM (AABS)

- The third row :max registers/SM.
- No more than 65536 registers/SM.
- Another Example:
  - We need to calculate the AABS.
  - Consider the block size is 512 threads
  - We can control the block size value.
  - assume that each thread requires 16 registers .
  - We can’t control the allocated registers per thread.
  - registers needed per block 512*16
  - Total required are 8,192 registers.
  - Required registers < the register limit.
  - Divide both:

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB

The Allocated Active Blocks per SM (AABS)

- Final Example:
  - The block warp limit is 4 blocks.
  - The block register limit is 8 blocks.
  - the block limit shared memory is 16 blocks.
- The compiler will choose the minimum.

GPU Features	NVIDIA A100
GPU Codename	GA100
GPU Architecture	NVIDIA Ampere
Compute Capability	8.0
Threads / Warp	32
Max Warps / SM	64
Max Threads / SM	2048
Max Thread Blocks / SM	32
Max 32-bit Registers / SM	65536
Max Registers / Block	65536
Max Registers / Thread	255
Max Thread Block Size	1024
FP32 Cores / SM	64
Ratio of SM Registers to FP32 Cores	1024
Shared Memory Size / SM	Configurable up to 164 KB