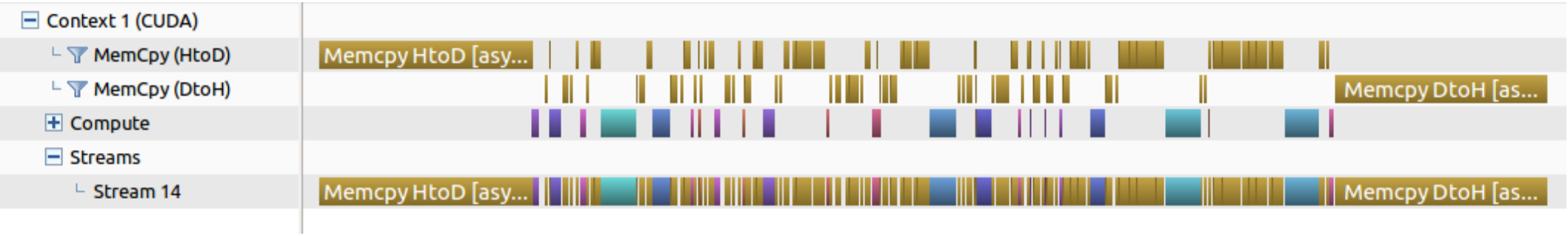
Performance Considerations for GPGPU computing

Performance Considerations

- PCIe transfers
- SIMT model
- Latency hiding

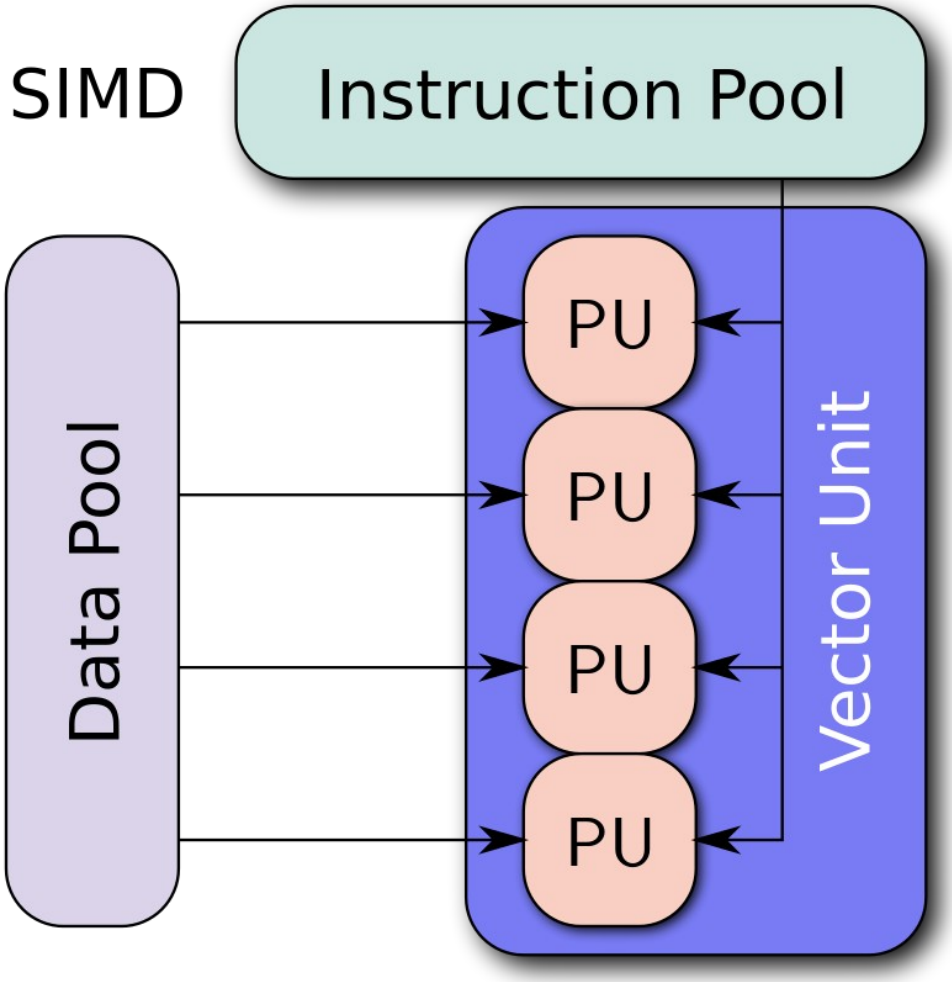
PCIe transfers

- Execution time = time for Computations + time for data transfers
- Slow* PCIe transfers degrade performance
- PCIe 3.0 : 32 GBps Bi-directional
- GPU may have only one copy engine!



SIMT model

- Single Instruction, Multiple Threads

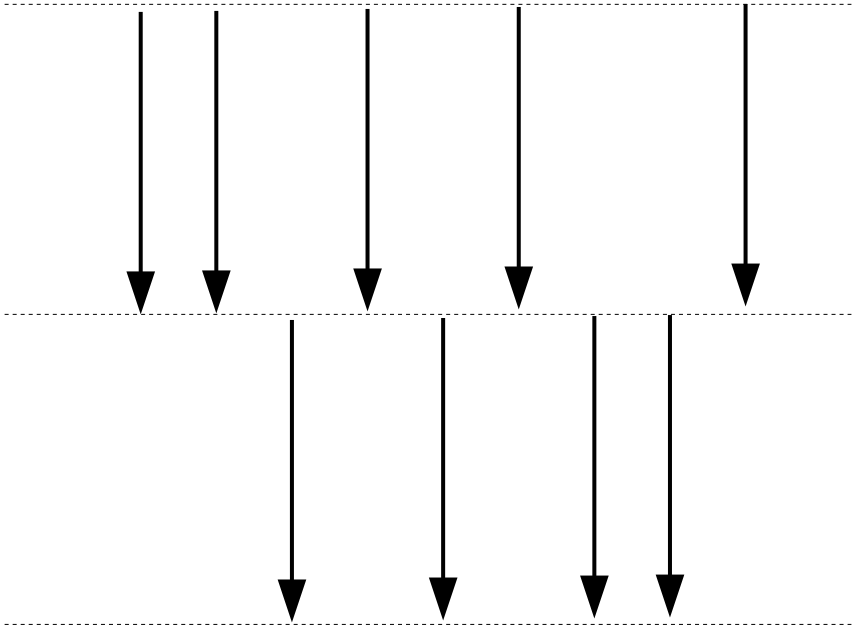


SIMT model issues

```
if (some_condition)
{
    . . .
    . . .
}
else
{
    . . .
    . . .
}
```

SIMT model issues

```
if (some_condition)
{
    . . .
    . . .
}
else
{
    . . .
    . . .
}
```



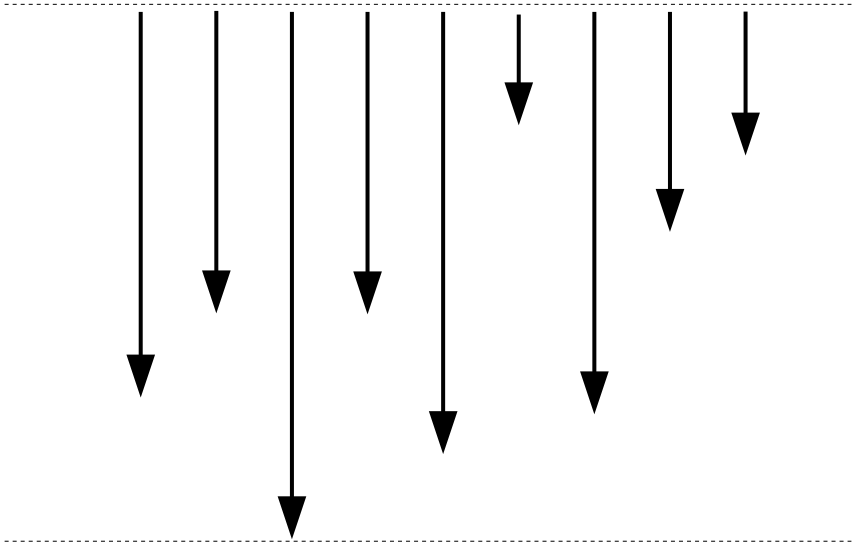
SIMT Model issues

```
for (“n” number of threads)
{
    for (“m” iterations)
    {

     }
}
```

What if “m” is not same for all the “n” threads?

SIMT Model issues

```
for (“n” number of threads)
{
    for (“m” iterations)
    {

     }
}
```



What if “m” is not same for all the “n” threads?

Latency Hiding

- CPUs are optimized for Single thread performance where as GPUs use large number of threads to amortize cost of time consuming operations (such as memory access etc.)
- No of threads for parallel application
  - CPU : should be **<=** number of cores
  - GPU : should be **>>** number of cores