

Hyperthread, GPUs, etc

More performance

- HT
- SIMD
- GPU

Idle function units

- While fetch ~ 4 , often IPC 1-2
 - What then?

Simultaneous Multithreading (SMT)

- **Hyperthreading (Intel term)**
- **Take a single core, add a little logic and make it run multiple contexts**
 - **Round-robin fetch**
 - **Duplicate dispatch / register mapping / TLB**

SIMD

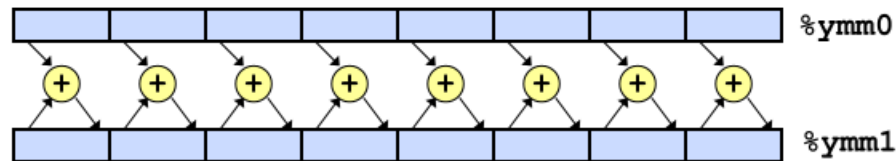
- Single instruction multiple data

Carnegie Mellon

SIMD Operations

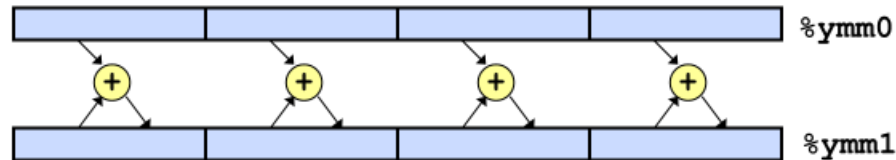
- SIMD Operations: Single Precision

`vaddsd %ymm0, %ymm1, %ymm1`



- SIMD Operations: Double Precision

`vaddpd %ymm0, %ymm1, %ymm1`



Bryant and O'Hallaron, Computer Systems: A Programmer's Perspective, Third Edition

48

Using Vector Instructions

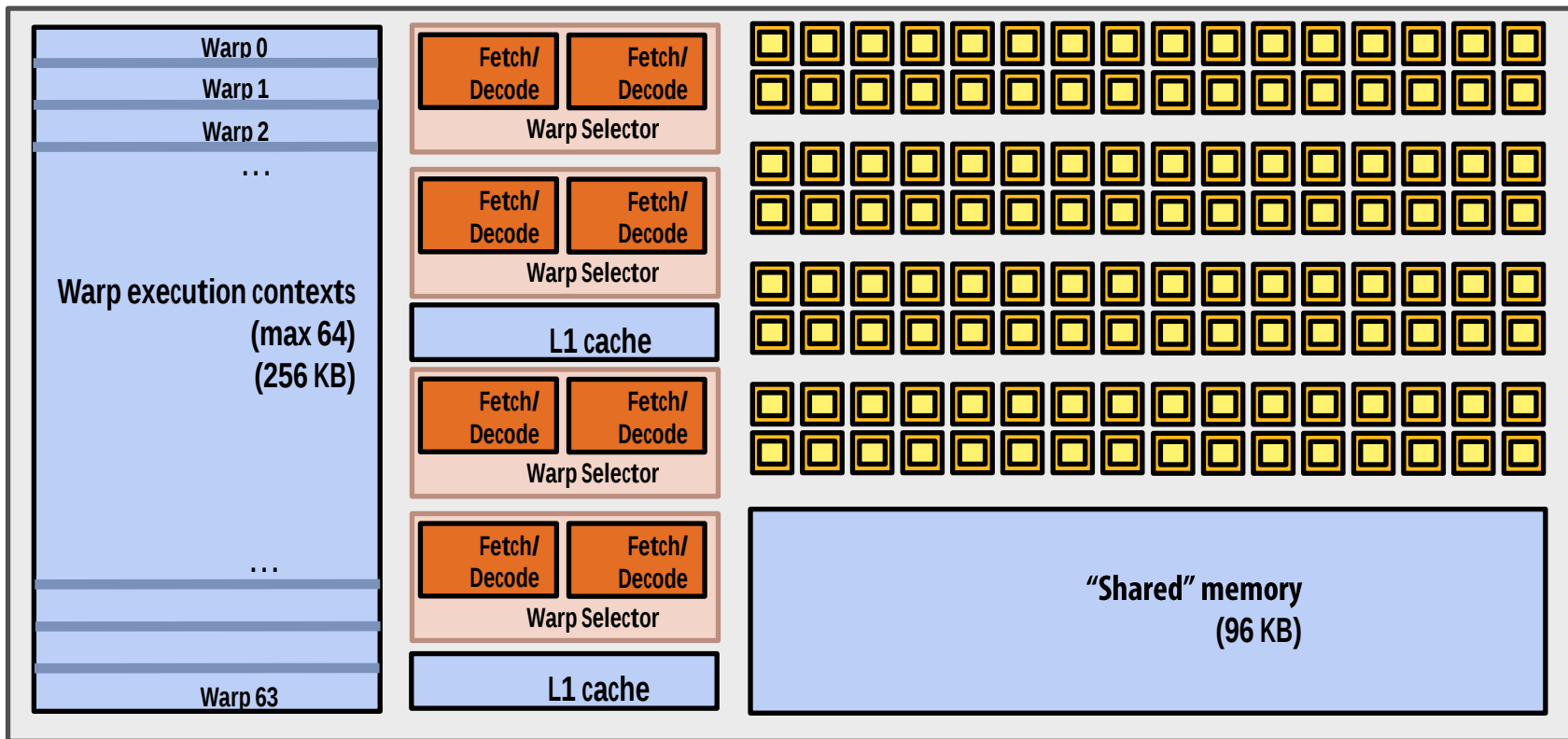
Method	Integer		Double FP	
Operation	Add	Mult	Add	Mult
Scalar Best	0.54	1.01	1.01	0.52
Vector Best	0.06	0.24	0.25	0.16
<i>Latency Bound</i>	<i>0.50</i>	<i>3.00</i>	<i>3.00</i>	<i>5.00</i>
<i>Throughput Bound</i>	<i>0.50</i>	<i>1.00</i>	<i>1.00</i>	<i>0.50</i>
<i>Vec Throughput Bound</i>	<i>0.06</i>	<i>0.12</i>	<i>0.25</i>	<i>0.12</i>

- Make use of AVX Instructions
 - Parallel operations on multiple data elements
 - See Web Aside OPT:SIMD on CS:APP web page

GPU

NVIDIA GTX 980 (2014)

This is one NVIDIA Maxwell GM204 architecture SMM unit (one “core”)



 = SIMD functional unit,
control shared across 32 units
(1 MUL-ADD per clock)

SMM resource limits:

- Max warp execution contexts: 64
(2,048 total CUDA threads)
- 96 KB of shared memory