New Course: Bachelor’s Seminar in Comp Arch

- Fall 2018
- 2 credit units

- **Rigorous seminar on fundamental and cutting-edge topics in computer architecture**

- Critical presentation, review, and discussion of seminal works in computer architecture
  - We will cover many ideas & issues, analyze their tradeoffs, perform critical thinking and brainstorming

- Participation, presentation, report and review writing
- Stay tuned for more information
Agenda for Today & Next Few Lectures

- Single-cycle Microarchitectures
- Multi-cycle and Microprogrammed Microarchitectures
- Pipelining
- Issues in Pipelining: Control & Data Dependence Handling, State Maintenance and Recovery, ...
- Out-of-Order Execution
- Other Execution Paradigms
Readings for Today


Other Approaches to Concurrency
(or Instruction Level Parallelism)
Approaches to (Instruction-Level) Concurrency

- Pipelining
- Out-of-order execution
- Dataflow (at the ISA level)
- Superscalar Execution
- VLIW
- Fine-Grained Multithreading
- SIMD Processing (Vector and array processors, GPUs)
- Decoupled Access Execute
- Systolic Arrays
SIMD Processing: Exploiting Regular (Data) Parallelism
Recall: Flynn’s Taxonomy of Computers


- **SISD**: Single instruction operates on single data element
- **SIMD**: Single instruction operates on multiple data elements
  - Array processor
  - Vector processor

- **MISD**: Multiple instructions operate on single data element
  - Closest form: systolic array processor, streaming processor

- **MIMD**: Multiple instructions operate on multiple data elements (multiple instruction streams)
  - Multiprocessor
  - Multithreaded processor
Recall: SIMD Processing

- Single instruction operates on multiple data elements
  - In time or in space
- Multiple processing elements

- Time-space duality
  - **Array processor**: Instruction operates on multiple data elements at the same time using different spaces
  - **Vector processor**: Instruction operates on multiple data elements in consecutive time steps using the same space
Recall: Array vs. Vector Processors

**ARRAY PROCESSOR**

- LD  VR ← A[3:0]
- ADD VR ← VR, 1
- MUL VR ← VR, 2
- ST  A[3:0] ← VR

**VECTOR PROCESSOR**

- LD
- ADD VR ← VR, 1
- MUL VR ← VR, 2
- ST  A[3:0] ← VR

**Instruction Stream**

- Same op @ same time
- Different ops @ time
- Different ops @ same space
- Same op @ space
Recall: Memory Banking

- Memory is divided into banks that can be accessed independently; banks share address and data buses (to minimize pin cost)
- Can start and complete one bank access per cycle
- Can sustain N parallel accesses if all N go to different banks
Some Issues

- Stride and banking
  - As long as they are relatively prime to each other and there are enough banks to cover bank access latency, we can sustain 1 element/cycle throughput

- Storage of a matrix
  - **Row major**: Consecutive elements in a row are laid out consecutively in memory
  - **Column major**: Consecutive elements in a column are laid out consecutively in memory
  - You need to change the stride when accessing a row versus column
Matrix Multiplication

- A and B, both in \textit{row-major order}

- \begin{align*}
    &A_0 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \\
    &6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11
\end{align*}

- \begin{align*}
    &B_0 \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \\
    &10 \quad 11 \quad 12 \quad 13 \quad 14 \quad 15 \quad 16 \quad 17 \quad 18 \quad 19 \\
    &20 \\
    &30 \\
    &40 \\
    &50
\end{align*}

- $A_{4x6} \times B_{6x10} \rightarrow C_{4x10}$

- Dot products of rows and columns of A and B

- \begin{itemize}
    \item \textbf{A}: Load $A_0$ into vector register $V_1$
      \begin{itemize}
        \item Each time, increment address by one to access the next column
        \item Accesses have a \textit{stride of 1}
      \end{itemize}
    \item \textbf{B}: Load $B_0$ into vector register $V_2$
      \begin{itemize}
        \item Each time, increment address by 10
        \item Accesses have a \textit{stride of 10}
      \end{itemize}
\end{itemize}

- Different strides can lead to bank conflicts

- How do we minimize them?
Minimizing Bank Conflicts

- More banks

- **Better data layout** to match the access pattern
  - Is this always possible?

- **Better mapping of address to bank**
  - E.g., randomized mapping
Recall: Questions (II)

- What if vector data is not stored in a strided fashion in memory? (irregular memory access to a vector)
  - Idea: Use indirection to combine/pack elements into vector registers
  - Called scatter/gather operations
Want to vectorize loops with indirect accesses:

for (i=0; i<N; i++)
    A[i] = B[i] + C[D[i]]

Indexed load instruction (Gather)

LV vD, rD  # Load indices in D vector
LVI vC, rC, vD  # Load indirect from rC base
LV vB, rB  # Load B vector
ADDV.D vA,vB,vC  # Do add
SV vA, rA  # Store result
Gather/Scatter Operations

- Gather/scatter operations often implemented in hardware to handle **sparse vectors (matrices)**
- **Vector loads and stores use an index vector which is added to the base register to generate the addresses**

**Scatter example**

<table>
<thead>
<tr>
<th>Index Vector</th>
<th>Data Vector (to Store)</th>
<th>Stored Vector (in Memory)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.14</td>
<td>Base+0 3.14</td>
</tr>
<tr>
<td>2</td>
<td>6.5</td>
<td>Base+1 X</td>
</tr>
<tr>
<td>6</td>
<td>71.2</td>
<td>Base+2 6.5</td>
</tr>
<tr>
<td>7</td>
<td>2.71</td>
<td>Base+3 X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Base+4 X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Base+5 X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Base+6 71.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Base+7 2.71</td>
</tr>
</tbody>
</table>
Array vs. Vector Processors, Revisited

- Array vs. vector processor distinction is a “purist’s” distinction

- Most “modern” SIMD processors are a combination of both
  - They exploit data parallelism in both time and space
  - GPUs are a prime example we will cover in a bit more detail
Recall: Array vs. Vector Processors

**ARRAY PROCESSOR**

- **Instruction Stream**
  - LD VR ← A[3:0]
  - ADD VR ← VR, 1
  - MUL VR ← VR, 2
  - ST A[3:0] ← VR

- **Time**
  - LD0
  - LD1
  - LD2
  - LD3
  - AD0
  - AD1
  - AD2
  - AD3
  - MU0
  - MU1
  - MU2
  - MU3
  - ST0
  - ST1
  - ST2
  - ST3

- **Space**
  - Same op @ same time
  - Different ops @ same space

**VECTOR PROCESSOR**

- **Instruction Stream**
  - LD0
  - LD1
  - AD0
  - LD2
  - AD1
  - MU0
  - LD3
  - AD2
  - MU1
  - AD3
  - MU2
  - ST0
  - MU3
  - ST1
  - ST2
  - ST3

- **Time**
  - LD0
  - LD1
  - AD0
  - LD2
  - AD1
  - MU0
  - LD3
  - AD2
  - MU1
  - AD3
  - MU2
  - ST0
  - MU3
  - ST1
  - ST2
  - ST3

- **Space**
  - Different ops @ time
  - Same op @ space
Vector Instruction Execution

VADD A,B → C

Execution using one pipelined functional unit


Execution using four pipelined functional units


Time

Space

Slide credit: Krste Asanovic
Vector Unit Structure

Partitioned Vector Registers

Elements 0, 4, 8, ...

Elements 1, 5, 9, ...

Elements 2, 6, 10, ...

Elements 3, 7, 11, ...

Memory Subsystem

Functional Unit

Lane

Slide credit: Krste Asanovic
Vector Instruction Level Parallelism

Can overlap execution of multiple vector instructions
- Example machine has 32 elements per vector register and 8 lanes
- Completes 24 operations/cycle while issuing 1 vector instruction/cycle

Load Unit  Multiply Unit  Add Unit

Slide credit: Krste Asanovic
Automatic Code Vectorization

Scalar Sequential Code

Vectorized Code

for (i=0; i < N; i++)
    C[i] = A[i] + B[i];

Vectorization is a compile-time reordering of operation sequencing
⇒ requires extensive loop dependence analysis

Slide credit: Krste Asanovic
Vector/SIMD Processing Summary

- Vector/SIMD machines are good at exploiting regular data-level parallelism
  - Same operation performed on many data elements
  - Improve performance, simplify design (no intra-vector dependencies)

- Performance improvement limited by vectorizability of code
  - Scalar operations limit vector machine performance
  - Remember Amdahl’s Law
  - CRAY-1 was the fastest SCALAR machine at its time!

- Many existing ISAs include (vector-like) SIMD operations
  - Intel MMX/SSEn/AVX, PowerPC AltiVec, ARM Advanced SIMD
SIMD Operations in Modern ISAs
**SIMD ISA Extensions**

- Single Instruction Multiple Data (SIMD) extension instructions
  - Single instruction acts on multiple pieces of data at once
  - Common application: graphics
  - Perform short arithmetic operations (also called *packed arithmetic*)
- For example: add four 8-bit numbers
- Must modify ALU to eliminate carries between 8-bit values

```
padd8 $s2, $s0, $s1
```

<table>
<thead>
<tr>
<th>Bit position</th>
<th>32</th>
<th>24</th>
<th>23</th>
<th>16</th>
<th>15</th>
<th>8</th>
<th>7</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s0</td>
<td>$a_3$</td>
<td>$a_2$</td>
<td>$a_1$</td>
<td>$a_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>$b_3$</td>
<td>$b_2$</td>
<td>$b_1$</td>
<td>$b_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s2</td>
<td>$a_3 + b_3$</td>
<td>$a_2 + b_2$</td>
<td>$a_1 + b_1$</td>
<td>$a_0 + b_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Intel Pentium MMX Operations

- Idea: One instruction operates on multiple data elements simultaneously
  - À la array processing (yet much more limited)
  - Designed with multimedia (graphics) operations in mind

No VLEN register

Opcode determines data type:
- 8 8-bit bytes
- 4 16-bit words
- 2 32-bit doublewords
- 1 64-bit quadword

Stride is always equal to 1.


Figure 1. MMX technology data types: packed byte (a), packed word (b), packed doubleword (c), and quadword (d).
MMX Example: Image Overlaying (I)

- Goal: Overlay the human in image 1 on top of the background in image 2

![Figure 8. Chroma keying: image overlay using a background color.](image)

```markdown
for (i=0; i<image_size; i++) {
  if (x[i] == Blue) new_image[i] = y[i];
  else new_image[i] = x[i];
}
```

![Figure 9. Generating the selection bit mask.](image)

**MMX Example: Image Overlaying (II)**

\[ Y = \text{Blossom image} \]
\[ X = \text{Woman's image} \]

---

**Figure 10. Using the mask with logical MMX instructions to perform a conditional select.**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movq mm3, mem1</td>
<td>/* Load eight pixels from woman's image */</td>
</tr>
<tr>
<td>Movq mm4, mem2</td>
<td>/* Load eight pixels from the blossom image */</td>
</tr>
<tr>
<td>Porsche mm1, mm3</td>
<td></td>
</tr>
<tr>
<td>Pand mm4, mm1</td>
<td></td>
</tr>
<tr>
<td>Pandn mm1, mm3</td>
<td></td>
</tr>
<tr>
<td>Por mm4, mm1</td>
<td></td>
</tr>
</tbody>
</table>

---

**Figure 11. MMX code sequence for performing a conditional select.**

GPUs (Graphics Processing Units)
GPUs are SIMD Engines Underneath

- The instruction pipeline operates like a SIMD pipeline (e.g., an array processor)

- However, the programming is done using threads, NOT SIMD instructions

- To understand this, let’s go back to our parallelizable code example

- But, before that, let’s distinguish between
  - Programming Model (Software)
  - Execution Model (Hardware)
Programming Model vs. Hardware Execution Model

- Programming Model refers to **how the programmer expresses the code**
  - E.g., Sequential (von Neumann), Data Parallel (SIMD), Dataflow, Multi-threaded (MIMD, SPMD), ...

- Execution Model refers to **how the hardware executes the code underneath**
  - E.g., Out-of-order execution, Vector processor, Array processor, Dataflow processor, Multiprocessor, Multithreaded processor, ...

- Execution Model can be very different from the Programming Model
  - E.g., von Neumann model implemented by an OoO processor
  - E.g., SPMD model implemented by a SIMD processor (a GPU)
How Can You Exploit Parallelism Here?

Scalar Sequential Code

for (i=0; i < N; i++)
C[i] = A[i] + B[i];

Let’s examine three programming options to exploit instruction-level parallelism present in this sequential code:

1. Sequential (SISD)
2. Data-Parallel (SIMD)
3. Multithreaded (MIMD/SPMD)
Prog. Model 1: Sequential (SISD)

Scalar Sequential Code

for (i=0; i < N; i++)
    C[i] = A[i] + B[i];

- Can be executed on a:
  - Pipelined processor
  - Out-of-order execution processor
    - Independent instructions executed when ready
    - Different iterations are present in the instruction window and can execute in parallel in multiple functional units
    - In other words, the loop is dynamically unrolled by the hardware
  - Superscalar or VLIW processor
    - Can fetch and execute multiple instructions per cycle
Prog. Model 2: Data Parallel (SIMD)

Scalar Sequential Code

Vector Instruction

Vectorized Code

for (i=0; i < N; i++)
    C[i] = A[i] + B[i];

Iter. 1

Iter. 2

Realization: Each iteration is independent

Idea: Programmer or compiler generates a SIMD instruction to execute the same instruction from all iterations across different data

Best executed by a SIMD processor (vector, array)
Program Model 3: Multithreaded

Scalar Sequential Code

```
for (i=0; i < N; i++)
    C[i] = A[i] + B[i];
```

Realization: Each iteration is independent

Idea: Programmer or compiler generates a thread to execute each iteration. Each thread does the same thing (but on different data)

Can be executed on a MIMD machine
Prog. Model 3: Multithreaded

```c
for (i=0; i < N; i++)
    C[i] = A[i] + B[i];
```

**Load, Add, Store**

Realization: Each iteration is independent

This particular model is also called:

- **SPMD**: Single Program Multiple Data
- Can be executed on a SIMT machine: Single Instruction Multiple Thread
A GPU is a SIMD (SIMT) Machine

- Except it is **not** programmed using SIMD instructions

- It is **programmed using threads** (SPMD programming model)
  - Each thread executes the same code but operates a different piece of data
  - Each thread has its own context (i.e., can be treated/restarted/executed independently)

- A set of threads executing the same instruction are dynamically grouped into a **warp (wavefront)** by the hardware
  - A warp is essentially a SIMD operation formed by hardware!
for (i=0; i < N; i++)
C[i] = A[i] + B[i];

Iter. 1
Iter. 2

Warp: A set of threads that execute the same instruction (i.e., at the same PC)

This particular model is also called:

SPMD: Single Program Multiple Data

A GPU executes it using the SIMT model:
Single Instruction Multiple Thread
Graphics Processing Units
SIMD not Exposed to Programmer (SIMT)
SIMD vs. SIMT Execution Model

- **SIMD**: A single *sequential instruction stream* of SIMD instructions → each instruction specifies multiple data inputs
  - [VLD, VLD, VADD, VST], VLEN

- **SIMT**: *Multiple instruction streams* of scalar instructions → threads grouped dynamically into warps
  - [LD, LD, ADD, ST], NumThreads

**Two Major SIMT Advantages:**
- **Can treat each thread separately** → i.e., can execute each thread independently (on any type of scalar pipeline) → MIMD processing
- **Can group threads into warps flexibly** → i.e., can group threads that are supposed to *truly* execute the same instruction → dynamically obtain and maximize benefits of SIMD processing
Multithreading of Warps

- Assume a warp consists of 32 threads
- If you have 32K iterations, and 1 iteration/thread → 1K warps
- Warps can be interleaved on the same pipeline → Fine grained multithreading of warps

```plaintext
for (i=0; i < N; i++)
    C[i] = A[i] + B[i];
```

- Iter. 20*32 + 1
- Iter. 20*32 + 2

![Diagram showing multithreading of warps]
Warps and Warp-Level FGMT

- **Warp**: A *set of threads that execute the same instruction* (on different data elements) → SIMT (Nvidia-speak)
- All threads run the same code
- Warp: The threads that run lengthwise in a woven fabric ...

![Diagram showing thread warps and SIMD pipeline](image)
High-Level View of a GPU
Latency Hiding via Warp-Level FGMT

- **Warp**: A set of threads that execute the same instruction (on different data elements)

- **Fine-grained multithreading**
  - One instruction per thread in pipeline at a time (No interlocking)
  - Interleave warp execution to hide latencies

- Register values of all threads stay in register file

- FGMT enables long latency tolerance
  - Millions of pixels

Slide credit: Tor Aamodt
Warp Execution (Recall the Slide)

32-thread warp executing $ADD \ A[tid], B[tid] \rightarrow C[tid]$

Execution using one pipelined functional unit

Execution using four pipelined functional units


A[27]  B[27]
A[22]  B[22]

C[2]
C[1]
C[0]

C[8]
C[4]
C[0]

C[9]
C[5]
C[1]

C[10]
C[6]
C[2]

C[11]
C[7]
C[3]

Slide credit: Krste Asanovic
SIMD Execution Unit Structure

Memory Subsystem

Functional Unit

Registers for each Thread

Lane

Registers for thread IDs 0, 4, 8, ...

Registers for thread IDs 1, 5, 9, ...

Registers for thread IDs 2, 6, 10, ...

Registers for thread IDs 3, 7, 11, ...

Slide credit: Krste Asanovic
Warp Instruction Level Parallelism

Can overlap execution of multiple instructions

- Example machine has **32 threads per warp and 8 lanes**
- Completes 24 operations/cycle while issuing 1 warp/cycle

Slide credit: Krste Asanovic
SIMT Memory Access

- Same instruction in different threads uses **thread id** to index and access different data elements

Let’s assume $N=16$, 4 threads per warp $\rightarrow$ 4 warps

Slide credit: Hyesoon Kim
Sample GPU SIMT Code (Simplified)

CPU code

```c
for (ii = 0; ii < 100000; ++ii) {
}
```

CUDA code

```c
// there are 100000 threads
__global__ void KernelFunction(...) {
    int tid = blockDim.x * blockIdx.x + threadIdx.x;
    int varA = aa[tid];
    int varB = bb[tid];
    C[tid] = varA + varB;
}
```
Sample GPU Program (Less Simplified)

**CPU Program**

```c
void add_matrix
( float *a, float *b, float *c, int N) {
    int index;
    for (int i = 0; i < N; ++i)
        for (int j = 0; j < N; ++j) {
            index = i + j*N;
            c[index] = a[index] + b[index];
        }
}

int main() {
    add_matrix(a, b, c, N);
}
```

**GPU Program**

```c
__global__ add_matrix
( float *a, float *b, float *c, int N) {
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = blockIdx.y * blockDim.y + threadIdx.y;
    int index = i + j*N;
    if (i < N && j < N)
        c[index] = a[index]+b[index];
}

int main() {
    dim3 dimBlock( blocksize, blocksize);
    dim3 dimGrid (N/blockDim.x, N/blockDim.y);
    add_matrix<<<dimGrid, dimBlock>>>( a, b, c, N);
}
```
Warp-based SIMD vs. Traditional SIMD

- Traditional **SIMD** contains a single thread
  - Sequential instruction execution; lock-step operations in a SIMD instruction
  - Programming model is SIMD (no extra threads) → SW needs to know vector length
  - ISA contains vector/SIMD instructions

- **Warp-based SIMD** consists of multiple scalar threads executing in a SIMD manner (i.e., same instruction executed by all threads)
  - Does not have to be lock step
  - Each thread can be treated individually (i.e., placed in a different warp) → programming model not SIMD
    - SW does not need to know vector length
    - Enables multithreading and flexible dynamic grouping of threads
  - ISA is scalar → SIMD operations can be formed dynamically
  - Essentially, it is **SPMD programming model implemented on SIMD hardware**
SPMD

- Single procedure/program, multiple data
  - This is a programming model rather than computer organization

- Each processing element executes the same procedure, except on different data elements
  - Procedures can synchronize at certain points in program, e.g. barriers

- Essentially, multiple instruction streams execute the same program
  - Each program/procedure 1) works on different data, 2) can execute a different control-flow path, at run-time
  - Many scientific applications are programmed this way and run on MIMD hardware (multiprocessors)
  - Modern GPUs programmed in a similar way on a SIMD hardware
SIMD vs. SIMT Execution Model

- **SIMD**: A single sequential instruction stream of SIMD instructions → each instruction specifies multiple data inputs
  - [VLD, VLD, VADD, VST], VLEN

- **SIMT**: Multiple instruction streams of scalar instructions → threads grouped dynamically into warps
  - [LD, LD, ADD, ST], NumThreads

- **Two Major SIMT Advantages:**
  - Can treat each thread separately → i.e., can execute each thread independently on any type of scalar pipeline → MIMD processing
  - Can group threads into warps flexibly → i.e., can group threads that are supposed to truly execute the same instruction → dynamically obtain and maximize benefits of SIMD processing
Threads Can Take Different Paths in Warp-based SIMD

- Each thread can have **conditional control flow instructions**
- Threads can execute different control flow paths
Control Flow Problem in GPUs/SIMT

- A GPU uses a SIMD pipeline to save area on control logic
  - Groups scalar threads into warps

- Branch divergence occurs when threads inside warps branch to different execution paths

This is the same as conditional/predicated/masked execution. Recall the Vector Mask and Masked Vector Operations?
Remember: Each Thread Is Independent

- Two Major SIMT Advantages:
  - Can treat each thread separately \(\rightarrow\) i.e., can execute each thread independently on any type of scalar pipeline \(\rightarrow\) MIMD processing
  - Can group threads into warps flexibly \(\rightarrow\) i.e., can group threads that are supposed to \textit{truly} execute the same instruction \(\rightarrow\) dynamically obtain and maximize benefits of SIMD processing

- If we have many threads
- We can find individual threads that are at the same PC
- And, group them together into a single warp dynamically
- This reduces “divergence” \(\rightarrow\) improves \textbf{SIMD utilization}
  - SIMD utilization: fraction of SIMD lanes executing a useful operation (i.e., executing an active thread)
Dynamic Warp Formation/Merging

- **Idea:** Dynamically merge threads executing the same instruction (after branch divergence)
- Form new warps from warps that are waiting
  - Enough threads branching to each path enables the creation of full new warps
Dynamic Warp Formation/Merging

- **Idea:** Dynamically merge threads executing the same instruction (after branch divergence)

Dynamic Warp Formation Example

A new warp created from scalar threads of both Warp x and y executing at Basic Block D

Execution of Warp x at Basic Block A
Execution of Warp y at Basic Block A

Baseline

Dynamic Warp Formation

Slide credit: Tor Aamodt
Hardware Constraints Limit Flexibility of Warp Grouping

Can you move any thread flexibly to any lane?

Slide credit: Krste Asanovic
We did not cover the following slides in lecture. These are for your preparation for the next lecture.
An Example GPU
NVIDIA GeForce GTX 285

- NVIDIA-speak:
  - 240 stream processors
  - “SIMT execution”

- Generic speak:
  - 30 cores
  - 8 SIMD functional units per core
NVIDIA GeForce GTX 285 “core”

- 64 KB of storage for thread contexts (registers)

- = SIMD functional unit, control shared across 8 units
- Orange = multiply-add
- Blue = multiply

- = instruction stream decode

- = execution context storage

Slide credit: Kayvon Fatahalian
NVIDIA GeForce GTX 285 “core”

- Groups of 32 threads share instruction stream (each group is a Warp)
- Up to 32 warps are simultaneously interleaved
- Up to 1024 thread contexts can be stored

64 KB of storage for thread contexts (registers)

Slide credit: Kayvon Fatahalian
NVIDIA GeForce GTX 285

30 cores on the GTX 285: 30,720 threads

Slide credit: Kayvon Fatahalian
Evolution of NVIDIA GPUs

GFLOPS

#Stream Processors

GFLOPS

Stream Processors

GFLOPS

0

1000

2000

3000

4000

5000

6000

7000

8000

9000

10000

11000

12000

13000

14000

15000

16000

0

1000

2000

3000

4000

5000

6000

7000

8000

9000

10000

11000

12000

13000

14000

15000

16000

GTX 285 (2009)

GTX 480 (2010)

GTX 780 (2013)

GTX 980 (2014)

P100 (2016)

V100 (2017)
NVIDIA V100

- NVIDIA-speak:
  - 5120 stream processors
  - “SIMT execution”

- Generic speak:
  - 80 cores
  - 64 SIMD functional units per core
  - Tensor cores for Machine Learning
NVIDIA V100 Block Diagram

80 cores on the V100
NVIDIA V100 Core

15.7 TFLOPS Single Precision
7.8 TFLOPS Double Precision
125 TFLOPS for Deep Learning (Tensor cores)

https://devblogs.nvidia.com/inside-volta/