Lecture 3: CUDA Programming

Zheng (Eddy) Zhang
Rutgers University
Fall 2017, 9/19/2017
Lecture 3A: GPU Programming
Review: Performance Hazard I — Control Divergence

- Thread Divergence Caused by Control Flows

```
A[ ]:
0 0 6 0 2 4 1
```

```
tid: 0 1 2 3 4 5 6 7
```

How about this?

```
if (A[tid]) {
    do some work;
} else {
    do nothing;
}
```

OR this?

```
for (i=0;i<A[tid]; i++) {
    ...
}
```

May degrade performance by up to warp_size times.
(warp size = 32 in modern GPUs)

Optional Reading: Zhang et al. "On-the-Fly Elimination of Dynamic Irregularities for GPU Computing", ASPLOS’11
Review: Performance Hazard II — Global Memory Coalescing

• Data is fetched as a contiguous memory chunk (a cache block/line)
  - A cache block/line is typically 128 byte
  - The purpose is to utilize the wide DRAM burst for maximum memory level parallelism (MLP)

• A thread warp is ready when all its threads’ operands are ready
  - Coalescing data operands for a thread warp into as few contiguous memory chunks as possible
  - Assume we have a memory access pattern for a thread \( ... = A[P[tid]]; \)

\[
\begin{align*}
P[] &= \{0, 1, 2, 3, 4, 5, 6, 7\} \\
tid: \ &\begin{array}{cccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\end{array} \\
A[]: \ &\begin{array}{cccccccc}
\color{green}{\cdot} & \color{green}{\cdot} & \color{green}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} \\
\end{array} \\
\end{align*}
\]

Coalesced Accesses

\[
\begin{align*}
P[] &= \{0, 5, 1, 7, 4, 3, 6, 2\} \\
tid: \ &\begin{array}{cccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\end{array} \\
A[]: \ &\begin{array}{cccccccc}
\color{green}{\cdot} & \color{green}{\cdot} & \color{green}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} & \color{white}{\cdot} \\
\end{array} \\
\end{align*}
\]

Non-Coalesced Accesses

\[
\begin{align*}
a \text{ cache block.} \\
\end{align*}
\]
Review: Performance Hazard III — Shared Memory Bank Conflicts

• **Shared Memory**
  - Typically 16 or 32 banks
  - Successive 32-bit words are assigned to successive banks
  - A fixed stride access may cause bank conflicts
  - Maximizing bank level parallelism is important

• **Reduce Bank Conflicts**
  - Padding, might waste some shared memory space.
  - Data layout transformation, i.e., array of struct to structor of array.
  - Thread-data remapping, i.e., the tasks that are mapped to different banks are executed at the same time.
Review: Performance Hazard III — Shared Memory Bank Conflicts

- No Bank Conflicts

- No Bank Conflicts
Review: Performance Hazard III — Shared Memory Bank Conflicts

- 2-way Bank Conflicts

- 8-way Bank Conflicts
Case Study 2 — Parallel Reduction

• **What is a reduction operation?**
  - A binary operation and a roll-up operation
  - Let $A = [a_0, a_1, a_2, a_3, \ldots, a_{(n-1)}]$, let $\oplus$ be an associative binary operator with identify element $I$

  \[
  \text{reduce}(\oplus, A) = [a_0 \oplus a_1 \oplus a_2 \oplus a_3 \ldots a_{(n-1)}]
  \]

  ```
  int sum = 0;
  for ( int i = 0; i < N; i++ )
      sum = sum + A[i];
  ```

  ![Sequential](sequential.png)
  ![Parallel](parallel.png)

Sequential

Parallel
Case Study 2 — Parallel Reduction

• First Implementation

```c
tid = threadIdx.x;

for ( s=1; s < blockDim.x; s *= 2) {
    if ( tid % (2*s) == 0 ) {
        sdata[ tid ] += sdata[ tid + s ];
    }

    __syncthreads();
}
```

Any GPU performance hazard here?

Thread Divergence
Case Study 2 — Parallel Reduction

• First Implementation

```c
int tid = threadIdx.x;

for (int s=1; s < blockDim.x; s *= 2) {
    if (tid % (2*s) == 0) {
        sdata[tid] += sdata[tid + s];
    }
    __syncthreads();
}
```

Any GPU performance hazard here?

Thread Divergence

Reducing 16 Million Elements

Implementation 1:

Performance: Throughput = 1.5325 GB/s, Time = 0.04379 s

Implementation 2:

Performance: Throughput = 1.9286 GB/s, Time = 0.03480 s

Implementation 3:

Performance: Throughput = 2.6062 GB/s, Time = 0.02575 s
Case Study 2 — Parallel Reduction

- Reduce Thread Divergence

**Thread-Task Mapping Changed!**
Case Study 2 — Parallel Reduction

• Reduce Thread Divergence

First Version

```c
int tid = threadIdx.x;
for (s=1; s < blockDim.x; s *= 2) {
    int index = 2 * s * tid;
    if (index < blockDim.x) {
        sdata[index] += sdata[index + s];
    }
}
__syncthreads();
```

Second Version

```c
int tid = threadIdx.x;
for (s=1; s < blockDim.x; s *= 2) {
    int index = 2 * s * tid;
    if (index < blockDim.x) {
        sdata[index] += sdata[index + s];
    }
}
__syncthreads();
```

Thread-Task Mapping Changed!
Case Study 2 — Parallel Reduction

- Implementation 2

```c
    tid = threadIdx.x;
    for (s=1; s < blockDim.x; s *= 2) {
        int index = 2 * s * tid;
        if (index < blockDim.x) {
            sdata[index] += sdata[index + s];
        }
    }
    __syncthreads();
```

Reducing 16 Million Elements

Implementation 1:

**Performance:** Throughput = 1.5325 GB/s, Time = 0.04379 s

Implementation 2:

**Performance:** Throughput = 1.9286 GB/s, Time = 0.03480 s

Potential Shared Memory Bank Conflicts
Case Study 2 — Parallel Reduction

• Reduce Shared Memory Bank Conflicts

Let Logically Contiguous Thread Access Contiguous Shared Memory Locations
Case Study 2 — Parallel Reduction

- Third Implementation (utilize implicit intra-warp synchronization)

```c
int tid = threadIdx.x;
for (s = blockDim.x/2; s>0; s >>=1) {
    if (tid < s) {
        sdata[tid] += sdata[tid+s];
    }
}
__syncthreads();
```

Reducing 16 Million Elements

- **Implementation 1:**
  - **Performance:** Throughput = 1.5325 GB/s, Time = 0.04379 s

- **Implementation 2:**
  - **Performance:** Throughput = 1.9286 GB/s, Time = 0.03480 s

- **Implementation 3:**
  - **Performance:** Throughput = 2.6062 GB/s, Time = 0.02575 s

Code: /ilab/users/zz124/cs515_2017/samples/6_Advanced/reduction
Review: Synchronization Primitive

- **Within a thread block**
  - Barrier: `__syncthreads()`
  - All computation by threads in the thread block before the barrier complete before any computation by threads after the barrier begins
  - Barriers are a conservative way to express dependencies
  - Barriers divide computation into phases
  - In conditional code, the condition must be uniform across the block

- **Across thread blocks**
  - Implicit synchronization at the end of a GPU kernel
  - A barrier in the middle of a GPU kernel can be implemented using atomic compare-and-swap (CAS) and memory fence operations

*Optional Reading:* Xiao et al. “Inter-block GPU communication via fast barrier synchronization”, IPDPS’10
Scan

• Let $A = [a_0, a_1, a_2, a_3, \ldots, a_{n-1}]$
• Let $\oplus$ be an associative binary operator with identify element $I$

$$\text{scan\_inclusive}(\oplus, A) = [a_0, a_0 \oplus a_1, a_0 \oplus a_1 \oplus a_2, \ldots]$$
$$\text{scan\_exclusive}(\oplus, A) = [I, a_0, a_0 \oplus a_1, a_0 \oplus a_1 \oplus a_2, \ldots]$$

A prefix sum is $\text{scan\_inclusive} \text{ scan\_inclusive}(\oplus, A)$
Prefix Sum

• Let $A = [a_0, a_1, a_2, a_3, \ldots, a_{n-1}]$
• Let the binary operator $\oplus$ be $+$, $\text{prefix}_\text{sum}(\oplus, A) = [a_0, a_0 \oplus a_1, a_0 \oplus a_1 \oplus a_2, \ldots]$ 

Sequential Code:
```c
int PrefixSum[N];
PrefixSum[0] = 0;
for (i = 1; i < N; i++)
    PrefixSum[i] = PrefixSum[i-1] + A[i];
```

<table>
<thead>
<tr>
<th>$A$</th>
<th>$3$</th>
<th>$1$</th>
<th>$7$</th>
<th>$0$</th>
<th>$4$</th>
<th>$1$</th>
<th>$6$</th>
<th>$3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Prefix}_\text{Sum}(+, A)$</td>
<td>$3$</td>
<td>$4$</td>
<td>$11$</td>
<td>$11$</td>
<td>$15$</td>
<td>$16$</td>
<td>$22$</td>
<td>$25$</td>
</tr>
</tbody>
</table>
How Would You Parallelize Scan?

• Recall the parallel reduction algorithm
How Would You Parallelize Scan?

- Recall the parallel reduction algorithm
How Would You Parallelize Scan?

• Let $P = \text{scan}\_\text{exclusive}(\oplus, A)$
How Would You Parallelize Scan?

• Let $P = \text{scan\_exclusive}(⊕, A)$
How Would You Parallelize Scan?

• Let \( P = \text{scan\_exclusive}(\oplus, A) \)

\[
14 = 8 + 4 + 2 \\
p_{14} = a_{0-7} + a_{8-11} + a_{12-13}
\]
How Would You Parallelize Scan?

• Let $P = \text{scan\_exclusive}(\oplus, A)$

\[
14 = 8 + 4 + 2 \\
p_{14} = a_{0-7} + a_{8-11} + a_{12-13}
\]
Parallel Scan (Work Efficient Version)

- Let \( P = \text{scan}\_\text{exclusive}(\oplus, A) \)

\[
\begin{array}{ccccccccccccccc}
\text{A[ ]} & a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 & a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
\text{A'[ ]} & a_0 & a_{0-1} & a_2 & a_{0-3} & a_4 & a_{4-5} & a_6 & a_{6-7} & a_8 & a_{8-9} & a_{10} & a_{10-11} & a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
\end{array}
\]

The Up-Sweep (Reduce) Phase
Parallel Scan Code

• Let \( P = \text{scan}\_\text{exclusive}(\oplus, A) \)

\[\begin{array}{c}
\text{A[ ]} \\
\begin{array}{cccc}
a_0 & a_1 & a_2 & a_3 \\
a_0 & a_0-1 & a_2 & a_3 \\
a_0 & a_0-1 & a_2 & a_3 \\
a_0 & a_0-1 & a_2 & a_3 \\
\end{array}
\end{array}\]

\[\begin{array}{cccc}
a_4 & a_4-5 & a_6 & a_6-7 \\
a_4 & a_4-5 & a_6 & a_6-7 \\
a_4 & a_4-5 & a_6 & a_6-7 \\
a_4 & a_4-5 & a_6 & a_6-7 \\
\end{array}\]

\[\begin{array}{cccc}
a_8 & a_8-9 & a_{10} & a_{10-11} \\
a_8 & a_8-9 & a_{10} & a_{10-11} \\
a_8 & a_8-9 & a_{10} & a_{10-11} \\
a_8 & a_8-9 & a_{10} & a_{10-11} \\
\end{array}\]

\[\begin{array}{cccc}
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
a_{12} & a_{12-13} & a_{14} & a_{12-15} \\
\end{array}\]

\[\begin{array}{cccc}
a_{14} & a_{14-15} & a_{15} & a_{15} \\
a_{12} & a_{14} & a_{14-15} & a_{15} \\
a_{12} & a_{14} & a_{14-15} & a_{15} \\
a_{12} & a_{14} & a_{14-15} & a_{15} \\
\end{array}\]

\[\begin{array}{cccc}
a_0 & a_0-1 & a_2 & a_3 \\
a_0 & a_0-1 & a_2 & a_3 \\
a_0 & a_0-1 & a_2 & a_3 \\
a_0 & a_0-1 & a_2 & a_3 \\
\end{array}\]

\[\begin{array}{cccc}
a_4 & a_4-5 & a_6 & a_6-7 \\
a_4 & a_4-5 & a_6 & a_6-7 \\
\end{array}\]

\[\begin{array}{cccc}
a_8 & a_8-9 & a_{10} & a_{10-11} \\
\end{array}\]

\[\begin{array}{cccc}
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
\end{array}\]

\[\begin{array}{cccc}
a_{14} & a_{14-15} & a_{15} & a_{15} \\
\end{array}\]

\[\begin{array}{cccc}
a_0 & a_0-1 & a_2 & a_3 \\
\end{array}\]

\[\begin{array}{cccc}
a_4 & a_4-5 & a_6 & a_6-7 \\
\end{array}\]

\[\begin{array}{cccc}
a_8 & a_8-9 & a_{10} & a_{10-11} \\
\end{array}\]

\[\begin{array}{cccc}
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
\end{array}\]

\[\begin{array}{cccc}
a_{14} & a_{14-15} & a_{15} & a_{15} \\
\end{array}\]

\[\begin{array}{cccc}
a_0 & a_0-1 & a_2 & a_3 \\
\end{array}\]

\[\begin{array}{cccc}
a_4 & a_4-5 & a_6 & a_6-7 \\
\end{array}\]

\[\begin{array}{cccc}
a_8 & a_8-9 & a_{10} & a_{10-11} \\
\end{array}\]

\[\begin{array}{cccc}
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
\end{array}\]

\[\begin{array}{cccc}
a_{14} & a_{14-15} & a_{15} & a_{15} \\
\end{array}\]

\[\begin{array}{cccc}
a_0 & a_0-1 & a_2 & a_3 \\
\end{array}\]

\[\begin{array}{cccc}
a_4 & a_4-5 & a_6 & a_6-7 \\
\end{array}\]

\[\begin{array}{cccc}
a_8 & a_8-9 & a_{10} & a_{10-11} \\
\end{array}\]

\[\begin{array}{cccc}
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
\end{array}\]

\[\begin{array}{cccc}
a_{14} & a_{14-15} & a_{15} & a_{15} \\
\end{array}\]

\[\begin{array}{cccc}
a_0 & a_0-1 & a_2 & a_3 \\
\end{array}\]

\[\begin{array}{cccc}
a_4 & a_4-5 & a_6 & a_6-7 \\
\end{array}\]

\[\begin{array}{cccc}
a_8 & a_8-9 & a_{10} & a_{10-11} \\
\end{array}\]

\[\begin{array}{cccc}
a_{12} & a_{12-13} & a_{14} & a_{14-15} \\
\end{array}\]

\[\begin{array}{cccc}
a_{14} & a_{14-15} & a_{15} & a_{15} \\
\end{array}\]

The Up-Sweep Pseudocode:

\[
\text{for } d=0 \text{ to } (\log_2 n - 1) \{ \\
\quad \text{if } ( \text{tid} \% (2^{d+1}) == 0 ) \{ \\
\quad\quad a[\text{tid} + 2^{d+1} - 1] += a[\text{tid} + 2^d - 1] \\
\quad\} \\
\quad \text{barrier\_synchronization();} \\
\}
\]

The Up-Sweep (Reduce) Phase
Parallel Scan Code

- Let $P = \text{scan\_exclusive}(+, A)$

<table>
<thead>
<tr>
<th>A[ ]</th>
<th>a₀</th>
<th>a₁</th>
<th>a₂</th>
<th>a₃</th>
<th>a₄</th>
<th>a₅</th>
<th>a₆</th>
<th>a₇</th>
<th>a₈</th>
<th>a₉</th>
<th>a₁₀</th>
<th>a₁₁</th>
<th>a₁₂</th>
<th>a₁₃</th>
<th>a₁₄</th>
<th>a₁₅</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A'[ ]</th>
<th>a₀</th>
<th>a₀₋₁</th>
<th>a₂</th>
<th>a₀₋₃</th>
<th>a₄</th>
<th>a₄₋₅</th>
<th>a₆</th>
<th>a₆₋₇</th>
<th>a₈</th>
<th>a₈₋₉</th>
<th>a₁₀</th>
<th>a₁₀₋₁₁</th>
<th>a₁₂</th>
<th>a₁₂₋₁₃</th>
<th>a₁₄</th>
<th>a₁₄₋₁₅</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Asymptotic Complexity?

The Up-Sweep (Reduce) Phase
Parallel Scan Code

- Let $P = \text{scan\_exclusive}(\oplus, A)$

**Asymptotic Complexity?** $O(n)$

The Up-Sweep (Reduce) Phase
Parallel Scan (work efficient version)

• Let $P = \text{scan\_exclusive}(\oplus, A)$

<table>
<thead>
<tr>
<th>$A^[*]$</th>
<th>$a_0$</th>
<th>$a_{0-1}$</th>
<th>$a_2$</th>
<th>$a_{0-3}$</th>
<th>$a_4$</th>
<th>$a_{4-5}$</th>
<th>$a_6$</th>
<th>$a_{0-7}$</th>
<th>$a_8$</th>
<th>$a_{8-9}$</th>
<th>$a_{10}$</th>
<th>$a_{8-11}$</th>
<th>$a_{12}$</th>
<th>$a_{12-13}$</th>
<th>$a_{14}$</th>
<th>$a_{0-15}$</th>
</tr>
</thead>
</table>

$P[*]$

The Down-Sweep Phase
Parallel Scan (work efficient version)

• Let $P = \text{scan\_exclusive}(\oplus, A)$
Parallel Scan Code (work efficient version)

- Let $P = \text{scan}\_\text{exclusive}(\oplus, A)$

**Down-sweep Pseudocode:**

$$x[n-1] \leftarrow 0$$

for $d=(\log_2 n - 1)$ to 0 {
  if ( tid % $2^{d+1}$ == 0 ) {
    tmp $\leftarrow a[tid + 2^d - 1]$
    a[tid + 2^d - 1] $\leftarrow a[tid + 2^{d+1} - 1]$
    a[tid + 2^{d+1} - 1] $\leftarrow$ tmp + a[tid + 2^{d+1} - 1]
  }
  barrier\_synchronization();
}
Parallel Scan Code (work efficient version)

• Let $P = \text{scan\_exclusive}(\oplus, A)$

\[ \begin{array}{cccccccccccccccc}
A[i] & a_0 & a_{0-1} & a_2 & a_{0-3} & a_4 & a_{4-5} & a_6 & a_{0-7} & a_8 & a_{8-9} & a_{10} & a_{8-11} & a_{12} & a_{12-13} & a_{14} & a_{0-15} \\
& \downarrow & & & & & & & & & & & & & & & \downarrow \\
\text{Asymptotic Complexity?} & a_0 & a_{0-1} & a_2 & a_{0-3} & a_4 & a_{4-5} & a_6 & a_{0-7} & a_8 & a_{8-9} & a_{10} & a_{8-11} & a_{12} & a_{12-13} & a_{14} & a_{0-7} \\
& \downarrow & & & & & & & & & & & & & & & \downarrow \\
\text{Asymptotic Complexity?} & a_0 & a_{0-1} & a_2 & 0 & a_4 & a_{0-3} & a_6 & a_{0-5} & a_8 & a_{0-7} & a_{10} & a_{0-9} & a_{12} & a_{0-11} & a_{14} & a_{0-11} \\
& \downarrow & & & & & & & & & & & & & & & \downarrow \\
\text{Asymptotic Complexity?} & a_0 & 0 & a_2 & a_{0-1} & a_4 & a_{0-3} & a_6 & a_{0-5} & a_8 & a_{0-7} & a_{10} & a_{0-9} & a_{12} & a_{0-11} & a_{14} & a_{0-13} \\
& \downarrow & & & & & & & & & & & & & & & \downarrow \\
P[i] & 0 & a_0 & a_{0-1} & a_{0-2} & a_{0-3} & a_4 & a_{0-4} & a_{0-5} & a_6 & a_{0-6} & a_{0-7} & a_{0-8} & a_{0-9} & a_{0-10} & a_{0-11} & a_{0-12} & a_{0-13} & a_{0-14} \\
\end{array} \]
Parallel Scan Code (work efficient version)

• Let $P = \text{scan\_exclusive}(\oplus, A)$

\[ A' \]

\[
\begin{array}{cccccccccccccccc}
  a_0 & a_{0-1} & a_2 & a_{0-3} & a_4 & a_{4-5} & a_6 & a_{0-7} & a_8 & a_{8-9} & a_{10} & a_{8-11} & a_{12} & a_{12-13} & a_{14} & a_{0-15} \\
\end{array}
\]

Asymptotic Complexity?

\[ O(n) \]

The Down-Sweep Phase
Simple Scan (when less work than parallelism)

• Let $P = \text{scan}\_\text{inclusive}(\oplus, A)$
Simple Scan (when less work than parallelism)

• Let $P = \text{scan\_inclusive}(\oplus, A)$

\[
\begin{array}{cccccccccccccccc}
A[ ] & a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 & a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\
& a_0 & a_{0-1} & a_{1-2} & a_{2-3} & a_{3-4} & a_{4-5} & a_{5-6} & a_{6-7} & a_{7-8} & a_{8-9} & a_{9-10} & a_{10-11} & a_{11-12} & a_{12-13} & a_{13-14} & a_{14-15} \\
& a_0 & a_{0-1} & a_{0-2} & a_{0-3} & a_{1-4} & a_{2-5} & a_{3-6} & a_{4-7} & a_{5-8} & a_{6-9} & a_{7-10} & a_{8-11} & a_{9-12} & a_{10-13} & a_{11-14} & a_{12-15} \\
& a_0 & a_{0-1} & a_{0-2} & a_{0-3} & a_{0-4} & a_{0-5} & a_{0-6} & a_{0-7} & a_{1-8} & a_{2-9} & a_{3-10} & a_{4-11} & a_{5-12} & a_{6-13} & a_{7-14} & a_{8-15} \\
& a_0 & a_{0-1} & a_{0-2} & a_{0-3} & a_{0-4} & a_{0-5} & a_{0-6} & a_{0-7} & a_{0-8} & a_{0-9} & a_{0-10} & a_{0-11} & a_{0-12} & a_{0-13} & a_{0-14} & a_{0-15} \\
\end{array}
\]

\[
\begin{array}{cccccccccccccccc}
P[ ] & a_0 & a_{0-1} & a_{0-2} & a_{0-3} & a_{0-4} & a_{0-5} & a_{0-6} & a_{0-7} & a_{0-8} & a_{0-9} & a_{0-10} & a_{0-11} & a_{0-12} & a_{0-13} & a_{0-14} & a_{0-15} \\
\end{array}
\]

One Phase
Simple Scan Code

- Let $P = \text{scan}_{\text{inclusive}}(\oplus, A)$

**One-pass Simple Scan Pseudocode:**

```plaintext
for d = 0 to ($\log_2 n - 1$) {
    if (tid $\geq 2^{d+1} - 1$) {
        tmp ← a[tid - 2^d] + a[tid]
    }
    barrier_synchronization();
    if (tid $\geq 2^{d+1} - 1$) {
        a[tid] = tmp;
    }
    barrier_synchronization();
}
```
Simple Scan Code

• Let $P = \text{scan}_{\text{inclusive}}(\oplus, A)$
Simple Scan Code

• Let $P = \text{scan}\_\text{inclusive}(\oplus, A)$

Asymptotic Complexity? $O(n\log(n))$
Question

• Assume work-efficient scan is performed in shared memory

  - What are the potential approaches we can use to reduce bank conflicts?
Segmented Scan

• Let $A = [[a_0, a_1, a_2], [a_3, a_4], [a_5, a_6, a_7, a_8], \ldots ]$
• Let $\oplus = +$
• $\text{segmented\_scan\_inclusive}(\oplus, A) =$
  
  $[[a_0, a_0 \oplus a_1, a_0 \oplus a_1 \oplus a_2], [a_3, a_3 \oplus a_4],$
  
  $[a_5, a_5 \oplus a_6, a_5 \oplus a_6 \oplus a_7, a_5 \oplus a_6 \oplus a_7 \oplus a_8], \ldots ]$
Segmented Scan

• Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$
Segmented Scan

- Let \( P = \text{segmented} \_\text{scan} \_\text{inclusive}(\oplus, A) \)
Segmented Scan

- Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$
Segmented Scan

- Let \( P = \text{segmented\_scan\_inclusive}(\oplus, A) \)
Segmented Scan

• Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$

---

**One Phase**
Segmented Scan

• Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$
Segmented Scan

• Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$

<table>
<thead>
<tr>
<th>A[]</th>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$a_7$</th>
<th>$a_8$</th>
<th>$a_9$</th>
<th>$a_{10}$</th>
<th>$a_{11}$</th>
<th>$a_{12}$</th>
<th>$a_{13}$</th>
<th>$a_{14}$</th>
<th>$a_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>$a_0$</td>
<td>$a_0 - 1$</td>
<td>$a_0 - 2$</td>
<td>$a_0 - 3$</td>
<td>$a_0 - 4$</td>
<td>$a_0 - 5$</td>
<td>$a_0 - 6$</td>
<td>$a_0 - 7$</td>
<td>$a_0 - 8$</td>
<td>$a_0 - 9$</td>
<td>$a_0 - 10$</td>
<td>$a_0 - 11$</td>
<td>$a_0 - 12$</td>
<td>$a_0 - 13$</td>
<td>$a_0 - 14$</td>
<td>$a_0 - 15$</td>
</tr>
<tr>
<td>$a_0$</td>
<td>$a_0$</td>
<td>$a_0 - 1$</td>
<td>$a_0 - 2$</td>
<td>$a_0 - 3$</td>
<td>$a_0 - 4$</td>
<td>$a_0 - 5$</td>
<td>$a_0 - 6$</td>
<td>$a_0 - 7$</td>
<td>$a_0 - 8$</td>
<td>$a_0 - 9$</td>
<td>$a_0 - 10$</td>
<td>$a_0 - 11$</td>
<td>$a_0 - 12$</td>
<td>$a_0 - 13$</td>
<td>$a_0 - 14$</td>
<td>$a_0 - 15$</td>
</tr>
<tr>
<td>$a_0$</td>
<td>$a_0$</td>
<td>$a_0 - 1$</td>
<td>$a_0 - 2$</td>
<td>$a_0 - 3$</td>
<td>$a_0 - 4$</td>
<td>$a_0 - 5$</td>
<td>$a_0 - 6$</td>
<td>$a_0 - 7$</td>
<td>$a_0 - 8$</td>
<td>$a_0 - 9$</td>
<td>$a_0 - 10$</td>
<td>$a_0 - 11$</td>
<td>$a_0 - 12$</td>
<td>$a_0 - 13$</td>
<td>$a_0 - 14$</td>
<td>$a_0 - 15$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P[]</th>
<th>$a_0$</th>
<th>$a_0 - 1$</th>
<th>$a_0 - 2$</th>
<th>$a_0 - 3$</th>
<th>$a_0 - 4$</th>
<th>$a_0 - 5$</th>
<th>$a_0 - 6$</th>
<th>$a_0 - 7$</th>
<th>$a_0 - 8$</th>
<th>$a_0 - 9$</th>
<th>$a_0 - 10$</th>
<th>$a_0 - 11$</th>
<th>$a_0 - 12$</th>
<th>$a_0 - 13$</th>
<th>$a_0 - 14$</th>
<th>$a_0 - 15$</th>
</tr>
</thead>
</table>

One Phase
Segmented Scan

- Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$

\[
\begin{array}{ccccccccccccccc}
A[] & a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 & a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\
& \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
& a_0 & a_{0-1} & a_{1-2} & a_3 & a_{3-4} & a_5 & a_{5-6} & a_{6-7} & a_{7-8} & a_9 & a_{9-10} & a_{11} & a_{11-12} & a_{12-13} & a_{13-14} & a_{14-15} \\
& \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
& a_0 & a_{0-1} & a_{0-2} & a_3 & a_{3-4} & a_5 & a_{5-6} & a_{5-7} & a_{5-8} & a_9 & a_{9-10} & a_{11} & a_{11-12} & a_{11-13} & a_{11-14} & a_{12-15} \\
& \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
& a_0 & a_{0-1} & a_{0-2} & a_3 & a_{3-4} & a_5 & a_{5-6} & a_{5-7} & a_{5-8} & a_9 & a_{9-10} & a_{11} & a_{11-12} & a_{11-13} & a_{11-14} & a_{11-15} \\
\end{array}
\]

$P[]$

\[
\begin{array}{ccccccccccccccc}
& a_0 & a_{0-1} & a_{0-2} & a_3 & a_{3-4} & a_5 & a_{5-6} & a_{5-7} & a_{5-8} & a_9 & a_{9-10} & a_{11} & a_{11-12} & a_{11-13} & a_{11-14} & a_{11-15} \\
\end{array}
\]

One Phase
Segmented Scan

Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$
Segmented Scan

• Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$
Segmented Scan Code

- Let $P = \text{segmented\_scan\_inclusive}(\oplus, A)$

```plaintext
One-pass Segmented Scan Pseudocode:

```for d= 0 to $(\log_2 n - 1)$ {
    if ( tid >= $2^{d+1} - 1$ && flag[tid] == flag[tid - $2^d$]) {
        tmp ← a[tid - $2^d$] + a[tid]
    }
    barrier\_synchronization();
} if ( tid >= $2^{d+1} - 1$ && flag[tid] == flag[tid - $2^d$]) {
    a[tid] = tmp;
} barrier\_synchronization();
P = A;
```
Segmented Scan for the Work Efficient Version

• Any idea?
Segmented Scan for the Work Efficient Version

- Any idea?

A[]

Up-Sweep Phase
Segmented Scan for the Work Efficient Version)

• Any idea?

The Down-Sweep Phase
Segmented Scan for the Work Efficient Version)

• Any idea?

Extra-Credit Assignment

Design your own algorithm for the segmented work-efficient scan
1. Specify the task that every thread does
2. Specify the data every thread works on
3. Is there any GPU-specific performance hazard in your algorithm?
4. If yes to question 3, how are you going to resolve it?

The Down-Sweep Phase
Hierarchical Scan

- Perform Scan in Local Chunk of Data first and Merge the Results later

64-element scan local_scan_1

64-element scan local_scan_2

64-element scan local_scan_3

64-element scan local_scan_4

Add base to all elements in each local scan

Get the last element of each local scan

Perform scan again
Hierarchical Segmented Scan

- Perform Scan in Local Chunk of Data first and Merge the Results later

64-element scan
local_seg_scan_1

64-element scan
local_seg_scan_2

64-element scan
local_seg_scan_3

64-element scan
local_seg_scan_4

Get the last element of each local segmented scan

If the base is in the same segment as the element in local segmented scan, add the base to the first segment in each local result

local_seg_scan_1

cond. add base to local_seg_scan_2

cond. add base to local_seg_scan_3

cond. add base to local_seg_scan_32

Perform segmented scan again
An Example of Parallel Segmented Scan

• Sparse Matrix Vector Multiplication (SpMV)

Let $A = \begin{bmatrix} 3 & x_0 \\ 0 & x_2 \\ 2 & x_1 \\ 0 & 4 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 2 & 6 & \cdots & 8 \end{bmatrix}$,

\[
\begin{bmatrix}
y_0 \\
y_1 \\
y_2 \\
\vdots \\
y_{n-1}
\end{bmatrix} = \begin{bmatrix} 3 & 0 & 1 & \cdots & 0 \\ 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 2 & 6 & \cdots & 8 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix}
\]

and $\oplus = +$, for SpMV, $y = \text{segmented}_{\text{scan}}(\oplus, A)$

• Project 1 requires a segmented scan at the warp level
Lecture 3B: Lexical and Syntax Analysis
Scanner and Parser collaborate to check the syntax of the input program.

1. Scanner maps stream of characters into words (words are the fundamental unit of syntax)

   Scanner produces a stream of tokens for the parser
   Token is <part of speech, word>
   “Part of speech” is a unit in the grammar

2. Parser maps stream of words into a sentence in a grammatical model of the input language
The Compiler’s Front End

- **Scanner** looks at every character
  - Converts stream of characters to stream of tokens, or classified words: `<part of speech, word>`
  - Efficiency & scalability matter
  Scanner is only part of compiler that looks at every character

- **Parser** looks at every token
  - Determines if the stream of tokens forms a sentence in the source language
  - Fits tokens to some syntactic model, or grammar, for the source language
How can we automate the construction of scanner & parsers?

• **Scanner**
  - Specify syntax with regular expressions (REs)
  - Construct finite automaton & scanner from the regular expression

• **Parser**
  - Specify syntax with context-free grammars (CFGs)
  - Construct push-down automaton & parser from the CFG
Example

Suppose that we need to recognize the word “not”:

**How would we look for it?**
- Calling some routine?
  - for example, `fscanf()`, the Scanner class in Java, or a regex library (Python, Java, C++)
- In a compiler, we would build a recognizer (not just calling a routine)

**Implementation idea**
- The spelling is ‘n’ followed by ‘o’ followed by ‘t’
- The code shown to the right, is as simple as the description
- Cost is $O(1)$ per character
- Structure allows for precise error messages

**Pseudocode:**

```plaintext
c ← next character
If c = ‘n’ then {
    c ← next character
    if c = ‘o’ then {
        c ← next character
        if c = ‘t’
            then return <NOT,“not”>
        else report error
    }
    else report error
} else report error
```
Automata

We can represent this code as an automaton

\[
c \gets \text{next character}\n\]
\[
\text{If } c = \text{‘n’} \text{ then } \{
\text{c } \gets \text{next character}\n\text{if } c = \text{‘o’} \text{ then } \{
\text{c } \gets \text{next character}\n\text{if } c = \text{‘t’} \text{ then return } <\text{NOT, “not”}>\n\text{else report error}\n\}
\text{else report error}\n\}
\text{else report error}\n\]
Automata

To recognize a larger set of keywords is (relatively) easy

For example, the set of words \{ if, int, not, then, to, write \}
Specifying An Automaton

**We list the words**
Each word’s spelling is unique and concise
We can separate them with the | symbol, read as “or”

**The specification:**

```
if int not then to write
```

is a simple regular expression for the language accepted by our automaton
Specifying An Automaton

• Every Regular Expression corresponds to an automaton
  - We can construct, automatically, an automaton from the RE
  - We can construct, automatically, an RE from any automaton
  - We can convert an automaton easily and directly into code
  - The automaton and the code both have $O(1)$ cost per input character

An RE specification leads directly to an efficient scanner
Unsigned Integers

In the example, each word had one spelling

What about a single part of speech with many spellings?

Consider specifying an unsigned integer
  • An unsigned integer is any string of digits from the set [0…9]
  • We might want to specify that no leading zeros are allowed

is a simple regular expression for the language accepted by our automaton

The Automata

With Leading Zeros

Without Leading Zeros

How do we write the corresponding RE?

Is “0001” allowed? Or “00”?
Unsigned Integers

- The Automata

- We need a notation to represent that cyclic edge in the automata
- The Kleene Closure
  - $x^*$ represents zero or more instances of ‘x’
- The Positive Closure
  - $x^+$ represents one or more instances of ‘x’

$[0-9]^+$ for leading zeros; 0 | $[1-9] [0-9]^*$ for without leading zeros
Unsigned Integers

• The RE corresponds to an automaton and an implementation

The automaton and the code can be generated automatically

```
c ← next character
n ← 0
if c = ‘0’
    then return <CONSTANT,n>
else if (‘1’ ≤ c ≤ ‘9’) then {
    n ← atoi(c)
    c ← next character
    while (‘0’ ≤ c ≤ ‘9’) {
        t ← atoi(c)
        n ← n * 10 + t
    }
    return <CONSTANT,n>
}
else report error
```
Regular Expression

• A formal definition

Regular Expressions over an Alphabet $\Sigma$

If $x \in \Sigma$, then $x$ is an RE denoting the set $\{ x \}$ or the language $L = \{ x \}$

If $x$ and $y$ are REs then

- $xy$ is an RE denoting $L(x)L(y) = \{ pq \mid p \in L(x) \text{ and } q \in L(y) \}$
- $x \mid y$ is an RE denoting $L(x) \cup L(y)$
- $x^*$ is an RE denoting $L(x)^* = \bigcup_{0 \leq k < \infty} L(x)^k$ (Kleene Closure)
  
  **Set of all strings that are zero or more concatenations of $x$**

- $x^+$ is an RE denoting $L(x)^+ = \bigcup_{1 \leq k < \infty} L(x)^k$ (Positive Closure)
  
  **Set of all strings that are one or more concatenations of $x$**

- $\varepsilon$ is an RE denoting the empty set
Regular Expression

- How do these operators help?

Regular Expressions over an Alphabet Σ
If x is in Σ, then x is an RE denoting the set \{ x \} or the language \( L = \{ x \} \)

*The spelling of any letter in the alphabet is an RE*

If x and y are REs then
- \( xy \) is an RE denoting \( L(x)L(y) = \{ pq \mid p \in L(x) \text{ and } q \in L(y) \} \)

*If we concatenate letters, the result is an RE (spelling of words)*
- \( x \mid y \) is an RE denoting \( L(x) \cup L(y) \)

*Any finite list of words can be written as an RE, ( \( w_0 \mid w_1 \mid w_2 \mid \ldots \mid w_n \) )*
- \( x^\ast \) is an RE denoting \( L(x)^\ast = \cup_{0 \leq k < \infty} L(x)^k \)
- \( x^+ \) is an RE denoting \( L(x)^+ = \cup_{1 \leq k < \infty} L(x)^k \)

*We can use closure to write finite descriptions of infinite, but countable, sets*

\( \varepsilon \) is an RE denoting the empty set

*In practice, the empty string is often useful*
Regular Expression

- Let the notation [0-9] be shorthand for (0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9)

**RE Examples:**

Positive integer  \[[0-9] [0-9]^*\]

or  \[0-9]^*

No leading zeros  0 | [1-9] [0-9]^*

Algol-style Identifier  \((\{a-z\}|\{A-Z\}) (\{a-z\}|\{A-Z\}|[0-9])^*\)

Decimal number  0 | [1-9] [0-9]^* . [0-9]^*

Real number  \(((0 \ | \ [1-9] [0-9]^*) \ | \ (0 \ | \ [1-9] [0-9]^* . [0-9]^*) \ E \ [0-9] [0-9]^*)\)

Each of these REs corresponds to an automaton and an implementation.
From RE to Scanner

• We can use results from automata theory to construct scanners directly from REs

  - There are several ways to perform this construction
  - Classic approach is a two-step method.
    - 1. Build automata for each piece of the RE using a simple template-driven method
       Build a specific variation on an automaton that has transitions on ε and non-deterministic choice (multiple transitions from a state on the same symbol)
       This construction is called “Thompson’s construction”
    - 2. Convert the newly built automaton into a deterministic automaton
       Deterministic automaton has no ε-transitions and all choices are single-valued
       This construction is called the “subset construction”

  - Given the deterministic automaton, we can run a minimization algorithm on it to reduce the number of states
    - Minimization is a space optimization. Both the original automaton and the minimal one take \( O(1) \) time per character
Thompson’s Construction

• The Key Idea

- For each RE symbols and operator, we have a small template
- Build them, in precedence order, and join them with ε-transitions

**Precedence in REs:** Parentheses, then closure, then concatenation, then alternation
Thompson’s Construction

• Let’s build an NFA for \( a (b \mid c)^* \)

1. \( a, b, \& c \)

2. \( b \mid c \)

3. \( (b \mid c)' \)

4. \( a (b \mid c)' \)

**Precedence in REs:** Parentheses, then closure, then concatenation, then alternation
Subset Construction

• The Concept
  - Build a simpler automaton (no $\varepsilon$-transitions, no multi-valued transitions) that simulates the behavior of the more complex automaton
  - Each state in the new automaton represents a set of states in the original

![Diagram of NFA and DFA]

<table>
<thead>
<tr>
<th>DFA</th>
<th>NFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_0$</td>
<td>$n_0$</td>
</tr>
<tr>
<td>$d_1$</td>
<td>$n_1 , n_2 , n_3 , n_4 , n_6 , n_9$</td>
</tr>
<tr>
<td>$d_2$</td>
<td>$n_5 , n_8 , n_9 , n_3 , n_4 , n_6$</td>
</tr>
<tr>
<td>$d_3$</td>
<td>$n_7 , n_8 , n_9 , n_3 , n_4 , n_6$</td>
</tr>
</tbody>
</table>
Minimization

- DFA minimization algorithms work by discovering states that are equivalent in their contexts and replacing multiple states with a single one.

- Minimization reduces the number of states, but does not change the costs.

<table>
<thead>
<tr>
<th>Minimal DFA State</th>
<th>$s_0$</th>
<th>$s_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original DFA State</td>
<td>$d_0$</td>
<td>$d_1,d_2,d_3$</td>
</tr>
</tbody>
</table>
Implementing an Automaton

• A common strategy is to simulate the DFA’s execution
  - Skeleton parser + a table that encores the automaton
  - The scanner generator constructs the table
  - The skeleton parser does not change (much)

```plaintext
state ← s₀
char ← NextChar()
while (char ≠ EOF) {
    state ← δ[state,char]
    char ← NextChar()
}
if (state is a final state)
    then report success
else report an error
```

Simple Skeleton Scanner

Transition table for our minimal DFA

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₀</td>
<td>S₁</td>
<td>Sₑ</td>
<td>Sₑ</td>
</tr>
<tr>
<td>S₁</td>
<td>Sₑ</td>
<td>S₁</td>
<td>S₁</td>
</tr>
</tbody>
</table>
Automatic Scanner Construction

• **Scanner Generator**
  - Tasks in a specification written as a collection of regular expressions
  - Combines them into one RE using alternation ("|")
  - Builds the minimal automaton ( § 2.4, 2.6.2 EAC)
  - Emits the tables to drive a skeleton scanner (§ 2.5 EAC)
Automatic Scanner Construction

- **Scanner Generator**
  - As alternative, the generator can produce code rather than tables
  - Direct-coded scanners are ugly, but often faster table-driven scanners
  - Other than speed, the two are equivalent
Summary

• **Automating Scanner Construction**
  - Write down the REs for the input language and connect them with “|”
  - Build a big, simple NFA
  - Build the DFA that simulates the NFA
  - Minimize the number of states in the DFA
  - Generate an implementation from the minimal DFA

• **Scanner Generators**
  - lex, flex, jflex, and such all work along the lines
  - Algorithms are well-known and well-understood
  - Key issues are: finding longest match rather than first match, and engineering the interface to the parser
  - You can build your own scanner generator

“Find the longest match” is where $O(1)$ per character can turn into $O(n)$. More next lecture.
Reading

• “Parallel Prefix Sum (Scan) with CUDA”, GPU Gem3, Chapter 39, by Mark Harris, Shubhabrata Sengupta, and John Owens. 
• “Performance Optimization” by Paulius Micikevicius, NVIDIA 
• CUDA C/C++ Basics 
• Loop unrolling for GPU Computing 
• CUDA Programming Guide, Chapter 1-5.
• Book: Engineering a Compiler (EAC)
  • Chapter 2.2
  • Chapter 2.4
  • Chapter 2.5
  • Chapter 2.6