Introduction to Control Divergence

Goal

- Understand optimizations to improve the performance of kernels in the presence of control divergence
- Better understanding of constraints imposed by the microarchitecture
Reading


• Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," *IEEE/ACM International Symposium on Computer Architecture (ISCA)*, 2013

• V. Narasiman, et. al., "Improving GPU Performance via Large Warps and Two-Level Scheduling," MICRO 2011


Thread Block Compaction

W. Fung and T. Aamodt

HPCA 2011
**Goal**

- Overcome some of the disadvantages of dynamic warp formation
  - Impact of Scheduling
  - Breaking implicit synchronization
  - Reduction of memory coalescing opportunities

**DWF Pathologies: Starvation**

- **Majority Scheduling**
  - Best Performing
  - Prioritize largest group of threads with same PC

- **Starvation**
  - **LOWER** SIMD Efficiency!

- **Other Warp Scheduler?**
  - Tricky: Variable Memory Latency

```
B: if (K > 10)
C:     K = 10;
else
D:     K = 0;
E: B = C[tid.x] + K;
```

---

Wilson Fung, Tor Aamodt

Thread Block Compaction
DWF Pathologies: Extra Uncoalesced Accesses

- Coalesced Memory Access = Memory SIMD
  - 1st Order CUDA Programmer Optimization
- Not preserved by DWF

```
B = C[tid.x] + K;
```

**No DWF**

```
E 1 2 3 4
E 5 6 7 8
E 9 10 11 12
```

#Acc = 3

```
Memory
0x100
0x140
0x180
```

**With DWF**

```
E 1 2 7 12
E 9 6 3 8
E 5 10 11 4
```

#Acc = 9

```
Memory
0x100
0x140
0x180
```

L1 Cache Absorbs Redundant Memory Traffic

L1S Port Conflict

DWF Pathologies: Implicit Warp Sync.

- Some CUDA applications depend on the lockstep execution of "static warps"

```
if (tid < 32) {
    sdata[tid] += sdata[tid + 32];
    sdata[tid] += sdata[tid + 16];
    sdata[tid] += sdata[tid + 8];
    sdata[tid] += sdata[tid + 4];
    sdata[tid] += sdata[tid + 2];
    sdata[tid] += sdata[tid + 1];
}
```

**Performance Impact**


---

**Thread Block Compaction**

- **Block-wide Reconvergence Stack**
  - Thread Block 0
    - Warp 1
      - PC: E
      - RPC: C
      - Active Mask: 1111
      - AMask: 1100
    - Warp 2
      - PC: E
      - RPC: C
      - Active Mask: 1111
      - AMask: 1100

- Regroup threads within a block
- **Better Reconv. Stack: Likely Convergence**
  - Converge before Immediate Post-Dominator
- **Robust**
  - Avg. 22% speedup on divergent CUDA apps
  - No penalty on others
Observation

- Compute kernels usually contain divergent and non-divergent (coherent) code segments
- Coalesced memory access usually in coherent code segments
  - DWF no benefit there
Thread Block Compaction

- **Barrier @ Branch/reconverge pt.**
  - All avail. threads arrive at branch
  - Insensitive to warp scheduling

- **Run a thread block like a warp**
  - Whole block move between coherent/divergent code
  - Block-wide stack to track exec. paths reconvg.

- **Warp compaction**
  - Regrouping with all avail. threads
  - If no divergence, gives static warp arrangement

---

### Code Example

A: \( K = A[tid.x] \);  
B: if (K > 10)  
C: \( K = 10 \);  
else  
D: \( K = 0 \);  
E: \( B = C[tid.x] + K \);
Thread Block Compaction

- Barrier every basic block?! (Idle pipeline)
- Switch to warps from other thread blocks
  - Multiple thread blocks run on a core
  - Already done in most CUDA applications

High Level View

- DWF: warp broken down every cycle and threads in a warp shepherded into a new warp (LUT and warp pool)
- TBC: warps broken down at potentially divergent points and threads compacted across the thread block
Microarchitecture Modification

- Per-Warp Stack → Block-Wide Stack
- I-Buffer + TIDs → Warp Buffer
  - Store the dynamic warps
- New Unit: Thread Compactor
  - Translate activemask to compact dynamic warps
- More Detail in Paper

![Diagram of thread block compaction](image)

Wilson Fung, Tor Aamodt
Thread Block Compaction

---

Microarchitecture Modification (2)

Thread Compactor

- Convert *activemask* from block-wide stack to *thread IDs* in warp buffer
- Array of Priority-Encoder

Likely-Convergence

- Immediate Post-Dominator: Conservative
  - All paths from divergent branch must merge there
- Convergence can happen earlier
  - When any two of the paths merge
    ```c
    while (i < K) {
        X = data[i];
        A: if (X = 0)
        B: result[i] = Y;
        C: else if (X = 1)
        D: break;
        E: i++;
    }
    F: return result[i];
    ```
- Extended Recvg. Stack to exploit this
  - TBC: 30% speedup for Ray Tracing

Details in Paper

Likely-Convergence (2)

- NVIDIA uses break instruction for loop exits
  - That handles last example
- Our solution: Likely-Convergence Points

<table>
<thead>
<tr>
<th>PC</th>
<th>RPC</th>
<th>LPC</th>
<th>LPos</th>
<th>ActiveThds</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>--</td>
<td>--</td>
<td>1 2</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>E</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>F</td>
<td>E</td>
<td>1</td>
<td>3 4</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>E</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Convergence!

- This paper: only used to capture loop-breaks
Likely-Convergence (3)


Likely-Convergence (4)

- Applies to both per-warp stack (PDOM) and thread block compaction (TBC)
  - Enable more threads grouping for TBC
  - Side effect: Reduce stack usage in some case
Effect on Memory Traffic

- TBC does still generate some extra uncoalesced memory access

\[ \#\text{Acc} = 4 \]

2nd Acc will hit the L1 cache

No Change to Overall Memory Traffic In/out of a core

Normalized Memory Stalls

- Normalized Memory Stalls

<table>
<thead>
<tr>
<th>TBC</th>
<th>DWF</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
<td>150%</td>
</tr>
<tr>
<td>150%</td>
<td>150%</td>
<td>200%</td>
</tr>
<tr>
<td>250%</td>
<td>250%</td>
<td>250%</td>
</tr>
<tr>
<td>300%</td>
<td>300%</td>
<td>300%</td>
</tr>
</tbody>
</table>

Conclusion

- Thread Block Compaction
  - Addressed some key challenges of DWF
  - One significant step closer to reality

- Benefit from advancements on reconvergence stack
  - Likely-Convergence Point
  - Extensible: Integrate other stack-based proposals
CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures
M. Rhu and M. Erez
ISCA 2012

Goals

- Improve the performance of inter-warp compaction techniques
- Predict when branches diverge
  - Borrow philosophy from branch prediction
- Use prediction to apply compaction only when it is beneficial
Issues with Thread Block Compaction

- TBC: warps broken down at potentially divergent points and threads compacted across the thread block
- Barrier synchronization overhead cannot always be hidden
- When it works, it works well
- The presence of divergence does not imply utility of compaction

Divergence Behavior: A Closer Look

Example 1 Code that contains two potentially divergent branches.

```
/* - Copied from _global__ void __kernel_forwarded CUDA() kernel of BGRDF
  1 - float input_code[ ] designates a global memory region */
2 3 int bx = blockIdx.x;
4 int tx = threadIdx.x;
5 int ty = threadIdx.y;
6
7 //shared float input_node[index];
8 if (tx == 0) // Conditional branch that is divergent but compaction ineffective
9 input_node[ty] = input_code[index i];
10 __syncthreads();
11
12 for (i = 1; i < log2(HEIGHT); i++)
13 {
14    for (j = 0; j < log2(WIDTH/2); j++)
15       // Loop-end branch: Conditional branch that is non-divergent
```

Figure from M. Rhu and M. Erez, “CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures,” ISCA 2012
Compaction-Resistant Branches

if (tx == 0) // Conditional branch that is divergent but compaction-ineffective
    input_node[ty] = input_cuda[index_in];

Control flow graph

Figure from M. Rhu and M. Erez, “CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures,” ISCA 2012

Compaction-Resistant Branches(2)

for (i = 1; i < __log2(HEIGHT); i++) {
    ...
    // Loop-end branch: Conditional branch that is non-divergent
}

Control flow graph

Figure from M. Rhu and M. Erez, “CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures,” ISCA 2012
Impact of Ineffective Compaction

Threads shuffled around with no performance improvement

However, can lead to increased memory divergence!

Figure from M. Rhu and M. Erez, “CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures,” ISCA 2012

Insight

- Synchronization barrier delays cannot be amortized for some branches
  - Barrier restricts the parallelism available to hide latency
- Distinguish between productive and unproductive candidates

Goal: More parallelism, more latency hiding, more scheduling targets, retain compaction
Basic Idea

- Only stall and compact when there is a high probability of compaction success
- Otherwise allow warps to bypass the (implicit) barrier
- Compaction adequacy predictor!
  - Think branch prediction!

Example: TBC vs. CAPRI

Bypassing enables increased overlap of memory references

---

Figure from M. Rhu and M. Erez, “CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures,” ISCA 2012
No Divergence $\rightarrow$ No Stalling

(a) Because $W_4$ does not diverge upon arriving at PC=$BR_{38}$, $W_4$ is bypassed and increments WCnt of path B by one. UMask is updated to [011] to have $W_4$’s active mask not be considered for generating compacted warps.

Do not consider for warps on a control path compaction

On Divergence

(b) $W_4$ arrives at PC=$BR_{38}$ and is stalled upon divergence. WCnt of path A is decremented by one and UMask remains at [011] because the active mask for $W_4$ needs to be considered for compaction.

Diverge (TBC)  Decrement (TBC)

Update prediction table

Diverge, stall, initialize history, all others will now stall
History-Based Compaction

(c) WCnt becomes zero as Wt arrives at PC=BRt. The CTA-wide active masks at TOS is forwarded to WCU and compaction is initiated. Note that Wt’s active mask is not considered for compaction as UMask is [011].

All warps have arrived
Compact (except for w0)

- Diverge and history available, predict, update prediction
- Compact
- CAPT updated

Initialize & Proceed

(d) WCU generates two warps and increments WCnt by two. As the Predictor evaluates PC=BRt, to be compaction-ineffective, the history bit is reset. UMask is initialized back to [1111].

Compact and update total number of new warps
The Predictor

- Prediction uses active masks of all warps
  - Need to understand what could have happened
- Actual compaction only uses actual stalled warps
- Minimum provides the maximum compaction ability, i.e., #compacted warps
- Update history predictor accordingly
  - Learn from branch predictor

Figure from M. Rhu and M. Erez, "CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures," ISCA 2012

Behavior

(a) Divergent benchmarks

(b) Non-divergent benchmarks

Figure from M. Rhu and M. Erez, "CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures," ISCA 2012
Figure from M. Rhu and M. Erez, “CAPRI: Prediction of Compaction-Adequacy for Handling Control-Divergence in GPGPU Architectures,” ISCA 2012

(b) Normalized idle cycles accumulated across all cores.

- Idle cycle count helps us understand the negative effects of implicit barriers in TBC

Summary

- The synchronization overhead of thread block compaction can introduce performance degradation
- Some branches more divergence than others
- Apply TBC judiciously → predict when it is beneficial
- Effectively predict when the inter-warp compaction is effective.
Improving GPU Performance via Large Warps and Two-Level Scheduling
V. Narsiman et. al
MICRO 2011

Goals

- Improve performance of divergent code via compaction of threads within a warp
- Integrate warp scheduling optimization with intra-warp compaction
Resource Underutilization

32 warps, 32 threads per warp, single SM, RR warp scheduler

Due to control divergence

Due to memory divergence

Warp Scheduling and Locality Effects

- Opportunities for memory coalescing
- Potential for exposing memory stalls
- Exploit row buffer locality

- Degrades memory reference locality and row buffer locality
- Overlaps memory accesses
Challenges and Tradeoffs

- Hitting long latency operations at the same time vs. exploiting locality in the memory hierarchy
- Generating enough parallelism to hide memory latencies vs. starvation and last warp effects
- Can we do both \textit{\textbf{\textarrowinva}} intra-warp compaction + scheduling
  - Large Warp Microarchitecture
  - Two-Level Round Robin scheduling

Key Ideas

- Conventional design today \textarrowinva warp size = \#SIMD lanes
- Use large warps \textarrowinva multi-cycle issue of sub-warsps
  - Compact threads in a warp to form fully utilized sub-warsps
- 2-level scheduler to spread memory accesses in time
  - Reduce memory related stall cycles \textit{enough}
  - Maintain enough locality
Large Warps

Typical Operation

Warp size = 4

Warp size = SIMD width

SIMD Width = 4

Proposed Approach

Warp size = 16

• Large Warp
• Multi-cycle Issue

• Large warps converted to a sequence of sub-warsps

• Similar in spirit to Intel Gen strategy for intrawarp compaction

Impact on the Register File

Baseline Register File

Large Warp Active Mask Organization

Need separate decoders per bank
Sub-Warp Compaction

- Iteratively select one thread per column to create a packed sub-warp
- Dynamic generation of sub-warsps

Scheduling Constraints

- Next large warp cannot be scheduled until first sub-warp completes execution
  - Scoreboard checks for issue dependencies
    - Thread not available for packing into a sub-warp unless previous issue (sub-warp) has completed ➔ single bit status
    - Simple check
  - Re-fetch policy for conditional branches
    - Must wait till last sub-warp finishes
- Optimization for unconditional branch instructions
  - Do not create multiple sub-warsps
  - Sub-warping always completes in a single cycle
Effect of Control Divergence

- Note that divergence is unknown until all sub-warps execute
  - Divergence management just happens on large warp boundaries
  - Need to buffer sub-warps state, e.g., active masks

- The last warp effect
  - Cannot fetch the next instruction in a warp until all sub-warps issue
  - Trailing warp (warp divergence effect) can lead to many idle cycles

- Effect of the last thread
  - E.g., in data dependent loop iteration count across threads
  - Last thread can hold up reconvergence

A Round Robin Warp Scheduler

- Exploit inter-warp reference locality in the cache
- Exploit inter-warp reference locality in the DRAM row buffers
- However, need to maintain latency hiding
Two Level Round Robin Scheduler

Fetch Group 0

LW0  LW1
LW2  LW3

Fetch Group 1

LW4  LW5
LW6  LW7

Fetch Group 2

LW8  LW9
LW10 LW11

Fetch Group 3

LW12 LW13
LW14 LW15

Scheduler Behavior

- Need to set fetch group size carefully → tune to fill the pipeline
- Timeout on switching fetch groups to mitigate the last warp effect
Some Interesting Behaviors

- Large warp size/Fetch Group size effects
  - Tuning latency hiding → avoiding last warp effects
- Timeout on warp execution
  - 32K instructions (empirical number)
- What is the coherent vs. non-coherent regions and large warps?
- Variance in divergence patterns
  - Have larger negative impact

Summary

- Intra-warp compaction made feasible due to multi-cycle warp execution
  - Mismatch between warp size and SIMD width enables flexible intra-warp compaction
- Do not make warps too big → last thread effect begins to dominate
M. Rhu and M. Erez, “Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation,” 
ISCA 2013

Goals

• Understand the limitations of compaction techniques and proximity to ideal compaction

• Provide mechanisms to overcome these limitations and approach ideal compaction rates
Limitations of Compaction

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013 (63)

Mapping Threads to Lanes: Today

Linearization of thread IDs

Modulo assignment to lanes

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013 (64)
Data Dependent Branches

- Data dependent control flows less likely to produce lane conflicts

Programmatic Branches

- Programmatic branches can be correlated to lane assignments (with modulo assignment)
- Program variables that operate like constants across threads can produce correlated branching behaviors

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013
P-Branches vs. D-Branches

P-branches are the problem!

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013

Compaction Opportunities

- Lane reassignment can improve compaction opportunities

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013
Aligned Divergence

- Threads mapped to a lane tend to evaluate (programmatic) predicates the same way
  - Empirically, rarely exhibited for input, data dependent control flow behavior

- Compaction cannot help in the presence of lane conflicts

- Performance of compaction mechanisms depends on both divergence patterns and lane conflicts

- We need to understand impact of lane assignment

Impact of Lane Reassignment

Goal: Improve “compactability”

Figure from M. Rhu and M. Erez, “Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation,” ISCA 2013
Random Permutations

- Does not always work well → works well on average
- Better understanding of programs can lead to better permutations choices

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013

Mapping Threads to Lanes: New

What criteria do we use for lane assignment?
### Balanced Permutation

Each lane has a single instance of a logical thread from each warp.

#### Even Warps:
Permutation within a half warp

#### Odd Warps:
Swap upper and lower lanes

\[
\text{XOR}_{\text{evenWID}} = \frac{\text{evenWID}}{2}
\]

Logical TID of 0 in each warp is now assigned a different lane.

*Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013*
Characteristics

- **Vertical Balance**: Each lane only has logical TIDs of distinct threads in a warp.
- **Horizontal balance**: Logical TID x in all of the warps is bound to different lanes.
- This works when CTA have fewer than SIMD_Width warps: why?
- Note that random permutations achieve this only on average.

Impact on Memory Coalescing

- Modern GPUs do not require ordered requests.
- Coalescing can occur across a set of requests → specific lane assignments do not affect coalescing behavior.
- Increase is L1 miss rate offset by benefits of compaction.

Figure from M. Rhu and M. Erez, “Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation,” ISCA 2013
**Speedup of Compaction**

- Can improve the compaction rate of divergence due to the majority of programmatic branches

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013

---

**Compaction Rate vs. Utilization**

Distinguish between compaction rate and utilization!

Figure from M. Rhu and M. Erez, "Maximizing SIMD Resource Utilization on GPGPUs with SIMD Lane Permutation," ISCA 2013
Application of Balanced Permutation

- Permutation is applied when the warp is launched
- Maintained for the life of the warp
- Does not affect the baseline compaction mechanism
- Enable/disable SLP to preserve target specific, programmer implemented optimizations

Summary

- Structural hazards limit the performance improvements from inter-warp compaction
- Program behaviors produce correlated lane assignments today
- Remapping threads to lanes enables extension of compaction opportunities
Summary: Inter-Warp Compaction

Co-Design of applications, resource management, software, microarchitecture,