CASE STUDY OF SHARED-MEMORY PARALLELIZATION

Jackson Marusarz - Intel
Agenda

• Shared Memory Review
• Performance implications of shared memory hardware
  • Data sharing
  • Contested accesses
• Performance implications of shared memory software
  • Data races
  • Deadlocks
  • Poor synchronization
  • Static thread scheduling
  • Scalability
  • Load imbalance
  • Oversubscription
  • Lock contention
Shared Memory Review
Performance Implications of Shared-Memory Hardware

- Modern CPUs have a shared address space for all the cores
- Need to maintain correctness, as different cores work on the same data
- Hardware protocols maintain coherency, but can have performance impacts

https://people.cs.pitt.edu/~xianeizhang/notes/cache.html
Performance Implications of Shared-Memory Hardware

- Metrics available in Intel® VTune™ Amplifier Microarchitecture Exploration

<table>
<thead>
<tr>
<th>Function / Call Stack</th>
<th>L2 Bound</th>
<th>Contested Accesses</th>
<th>L3 Bound</th>
<th>Data Sharing</th>
<th>L3 Latency</th>
<th>SQ Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>grid_intersect</td>
<td>4.9%</td>
<td>0.6%</td>
<td>97.8%</td>
<td>33.1%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>sphere_intersect</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>grid_bounds_intersect</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>30.9%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
Data Sharing

**Why:** Sharing clean data (read sharing) among cores (at L2 level) has a penalty at least the first time due to coherency

**What Now:** If this metric is highlighted for your hotspot, locate the source code line(s) that is generating HITs by viewing the source.

- Look for the MEM_LOAD_L3_HIT RETIRED.XSNP_HIT_PS event which will tag to the next instruction after the one that generated the HIT.
- Use knowledge of the code to determine if real or false sharing is taking place. Make appropriate fixes:
  - For real sharing, reduce sharing requirements
  - For false sharing, pad variables to cache line boundaries
Contested Accesses

**Why:** Sharing modified data among cores (at L2 level) can raise the latency of data access

**What Now:** If this metric is highlighted for your hotspot, locate the source code line(s) that is generating HITMs by viewing the source.

- Look for the MEM_LOAD_L3_HIT RETIRED.XSNP_HITM_PS event which will tag to the next instruction after the one that generated the HITM.
- Use knowledge of the code to determine if real or false sharing is taking place. Make appropriate fixes:
  - For real sharing, reduce sharing requirements
  - For false sharing, pad variables to cache line boundaries
Performance Implications of Shared-Memory Software

- Data races
- Deadlocks
- Poor synchronization
- Static thread scheduling
- Scalability
- Load imbalance
- Oversubscription
- Lock contention
Data Races - SSSP

```
do {
    auto old_dst_dist = dst_data.load();
    auto new_dst_dist = src_data.load() + w;
    if (new_dst_dist < old_dst_dist) {
        dst_data = new_dst_dist;
        swapped = true;
    }
} while (!swapped);
```

Application may:
- Crash Immediately
- Hang
- Run but give incorrect results
- Run and give correct results
- Run correctly 99 times but crash once (usually once you ship it to customers)

Non-determinism is always a concern in parallel programming. It may depend on how the OS decides to schedule threads.
Data Races - SSSP

```c
do {
    auto old_dst_dist = dst_data.load();
    auto new_dst_dist = src_data.load() + w;
    if (new_dst_dist < old_dst_dist) {
        dst_data = new_dst_dist;
        swapped = true;
    }
} while (!swapped);
```

**Shared variables unprotected**

```c
do {
    pthread_mutex_lock(&swap_mutex);
    auto old_dst_dist = dst_data.load();
    auto new_dst_dist = src_data.load() + w;
    if (new_dst_dist < old_dst_dist) {
        dst_data = new_dst_dist;
        swapped = true;
    }
    pthread_mutex_unlock(&swap_mutex);
} while (!swapped);
```

**Add a critical section**
Threading Problems - Deadlock

```c
CRITICAL_SECTION cs1;
CRITICAL_SECTION cs2;
int x = 0;
int y = 0;
InitializeCriticalSection(&cs1); // Allocation Site (cs1)
InitializeCriticalSection(&cs2); // Allocation Site (cs2)
```

**Thread #1**

```c
EnterCriticalSection(&cs1);
x++;
EnterCriticalSection(&cs2);
y++;
LeaveCriticalSection(&cs2);
LeaveCriticalSection(&cs1);
```

**Thread #2**

```c
EnterCriticalSection(&cs2);
y++;
EnterCriticalSection(&cs1);
x++;
LeaveCriticalSection(&cs1);
LeaveCriticalSection(&cs2);
```

**Deadlock**

1. EnterCriticalSection(&cs1); in thread #1
2. EnterCriticalSection(&cs2); in thread #2
do {
    auto old_dst_dist = dst_data.load();
    auto new_dst_dist = src_data.load() + w;
    if (new_dst_dist < old_dst_dist) {
        pthread_mutex_lock(&swap_mutex);
        &dst_data = new_dst_dist;
        pthread_mutex_unlock(&swap_mutex);
    } else {
        swapped = true;
    }
} while (!swapped);
do {
  auto old_dst_dist = dst_data.load();
  auto new_dst_dist = src_data.load() + w;

  if (new_dst_dist < old_dst_dist) {
    swapped = std::atomic_compare_exchange_weak(&dst_data, &old_dst_dist, new_dst_dist);
    changed |= swapped;
  } else {
    swapped = true;
  }
} while (!swapped);
```
do {
    auto old_dst_dist = dst_data.load();
    auto new_dst_dist = src_data.load() + w;

    if (new_dst_dist < old_dst_dist) {
        swapped =
        std::atomic_compare_exchange_weak(&dst_data, &old_dst_dist, new_dst_dist);
    changed |= swapped;
    }
else {
    swapped = true;
    }
} while (!swapped);
```
Static Thread Scheduling - SSSP

Hardcoding thread counts or relying on inputs can have performance impacts

```c
src_node = std::atoi(argv[2]) - 1;
initialize_sssp();

#ifdef SP2018_CS377P_PARALLEL
    num_threads = std::atoi(argv[3]);
#endif
    compute_sssp();
```

* I know this was used to teach concepts

```c
NUM_THREADS = 4;
pthread_t threads[NUM_THREADS];
int rc;
long t;
int chunk = limit/NUM_THREADS;
for(t=0;t<NUM_THREADS;t++){
    range *r = new range();
    ...
Static Thread Scheduling - SSSP

Hardcoding thread counts or relying on inputs can have performance impacts

```cpp
src_node = std::atoi(argv[2]) - 1;
initialize_sssp();

#ifdef SP2018_CS377P_PARALLEL
    num_threads = std::atoi(argv[3]);
#endif
compute_sssp();
```

*I know this was used to teach concepts*

```
...  
NUM_THREADS = 4;  NUM_THREADS = get_num_procs();
pthread_t threads[NUM_THREADS];
int rc;
long t;
int chunk = limit/NUM_THREADS;
for(t=0; t<NUM_THREADS; t++){
    range *r = new range();
...  
```

Use dynamic processor identification or scalable runtime library like OpenMP or Threading Building Blocks

```cpp
#pragma omp parallel for
```

Copyright © 2018, Intel Corporation. All rights reserved. *Other names and brands may be claimed as the property of others.*
Scalability is Not a Given - SSSP

Optimized Time:
- **Elapsed Time**: 22.947s
  - **CPU Time**: 21.470s
  - Total Thread Count: 3
  - Paused Time: 1.457s

Optimized Time:
- **Elapsed Time**: 15.791s
  - **CPU Time**: 25.830s
  - Total Thread Count: 4

Optimized Time:
- **Elapsed Time**: 10.978s
  - **CPU Time**: 30.200s
  - Total Thread Count: 6

Optimized Time:
- **Elapsed Time**: 7.239s
  - **CPU Time**: 34.140s
  - Total Thread Count: 10

Optimized Time:
- **Elapsed Time**: 6.124s
  - **CPU Time**: 49.720s
  - Total Thread Count: 18
  - Paused Time: 1.253s

Graph:
- **SSSP**
  - Time vs. Threads
  - Data points showing a decrease in time as the number of threads increases.
Scalability is Not a Given - SSSP

**Strong Scaling** - Solution time scales with the number of processors for a fixed *total* problem size.

**Weak Scaling** - Solution time scales with the number of processors for a fixed *total* problem size *per processor* - i.e. scales if the problem size also scales

**Serial Time** will always be a limiting factor as well.
Load Balancing

- Work should be divided among threads evenly
- What is even?
  - Loop Iterations? Elements to process?
  - Intelligent parallelism uses dynamic workload balancing
    - Work stealing and/or dynamic chunking

Should we divide work by subgraphs?
Load Balancing – Work Stealing

- Task starting
- Local task pool
- Execution
Example – Calculating Prime Numbers

```c
int _tmain(int argc, _TCHAR* argv[])
{
    DWORD msBegin = timeGetTime();

    #pragma omp parallel for
    for(int p = 3; p <= limit; p += 2) {
        if(IsPrime(p)) Tick();
    }
    DWORD msDuration = timeGetTime() - msBegin;

    printf("MS: %d\n", msDuration);
    printf("primes = %d\n", primes);
    return primes != correctCount;
}
```

- Is 7 prime?
- Is 76853341 prime?

Static Scheduling/Chunking:
- Check 1-10000
- Check 10001-20000
- Check 20001-30000
- ...
Example – Calculating Prime Numbers

```c
int _tmain(int argc, _TCHAR* argv[])
{
    DWORD msBegin = timeGetTime();

    //pragma omp parallel for
    for(int p = 3; p <= limit; p += 2) {
        if (IsPrime(p)) Tick();
    }

    DWORD msDuration = timeGetTime() - msBegin;

    printf("MS: %d
", msDuration);
    printf("primes = %d
", primes);
    return primes != correctCount;
}
```

- OpenMP uses Static Scheduling
- Load Imbalance

**CPU Usage Histogram**
This histogram displays a percentage of the wall time the specific number of CPUs

**Optimization Notice**
Copyright © 2018, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.
Example – Calculating Prime Numbers

```c
int _tmain(int argc, _TCHAR* argv[])
{
    DWORD msBegin = timeGetTime();

    #pragma omp parallel for schedule (dynamic, 1000)
    for(int p = 3; p <= limit; p += 2) {
        if (IsPrime(p)) Tick();
    }
    DWORD msDuration = timeGetTime() - msBegin;

    printf("MS: %d\n", msDuration);
    printf("primes = %d\n", primes);
    return primes != correctCount;
}
```

Switch to Dynamic Scheduling

More Balanced Threads

CPU Usage Histogram

This histogram displays a percentage of the wall time the specific number of CPUs

Simultaneously Utilized Logical CPUs

Optimization Notice

Copyright © 2018, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.
Oversubscription

Common Causes:
- Nested Parallelism
- Manual Threading
- Library Usage
Lock Contention

- Acquiring and releasing a lock isn't free – it has overhead
- Threads waiting for a lock also impacts performance
- How do we balance these?

Imagine an array that multiple threads read and write
Lock Contention

1 Shared Lock?

Dense lock contention and stalls
Lock Contention

1 Lock per Element?

Lots of locking overhead

What can we do?

• Adjust lock granularity
• Using lock free or thread safe data structures
  • `tbb::atomic<int> primes;`
  • `tbb::concurrent_vector<int> all_primes;`
• Local storage and reductions
Summary

• Programming for shared memory is difficult
• Correctness and performance issues are unique
• Issues are from hardware and software
  • Data sharing
  • Contested accesses
  • Deadlock
  • Data races
  • Poor synchronization
  • Static thread scheduling
  • Scalability
  • Load imbalance
  • Oversubscription
  • Lock contention
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Copyright © 2018, Intel Corporation. All rights reserved. Intel, Pentium, Xeon, Xeon Phi, Core, VTune, Cilk, and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804