Breaking Cyclic-Multithreading Parallelization with XML Parsing

Simone Campanoni  
Harvard University  
Cambridge, USA  
xan@eecs.harvard.edu  

Svilen Kanev  
Harvard University  
Cambridge, USA  
skanev@eecs.harvard.edu  

Kevin Brownell  
Harvard University  
Cambridge, USA  
brownell@eecs.harvard.edu  

Gu-Yeon Wei  
Harvard University  
Cambridge, USA  
guyeon@eecs.harvard.edu  

David Brooks  
Harvard University  
Cambridge, USA  
dbrooks@eecs.harvard.edu  

1. INTRODUCTION

HELIX-RC, a modern re-evaluation of the cyclic-multithreading (CMT) compiler technique [6], extracts threads from sequential code automatically. As a CMT approach, HELIX-RC gains performance by running iterations of the same loop on different cores in a multicore. It successfully boosts performance for SPEC CINT benchmarks previously considered unparallelizable (Table 1), assuming 16 Intel Atom cores with a special-purpose interconnect [3]. However, this paper shows there are workloads with different characteristics, which even idealized CMT cannot parallelize.

We identify how to overcome an inherent limitation of CMT for these workloads. CMT techniques only run iterations of a single loop in parallel at any given time. We propose exploiting parallelism not only within a single loop, but also among multiple loops. We call this execution model Multiple CMT (MCMT), and show that it is crucial for auto-parallelizing a broader class of workloads.

To highlight the need for MCMT, we target a workload that is naturally hard for CMT – parsing XML-structured data. XML parsing has been characterized as highly input-dependent, with a complex callgraph structure [4]. The particular implementation that we used (libxml2 [11]) adds further characteristics that complicate blindly applying a CMT approach. First, it spends a significant amount of execution in recursive calls or in non-natural loops, limiting the application scope of CMT (which exploits natural loop iterations). Second, it spends a rather small fraction of execution in small loops (which are preferred by HELIX-RC) - libxml2: 15%, SPEC CINT on average: 90%.

The rest of the paper starts by describing the importance of XML parsing and the code characteristics of libxml2. Then, after showing that HELIX-RC comes short for this library, we demonstrate that even idealized CMT-like techniques are not effective. On the other hand, an MCMT prototype speeds up the parsing task by up to 3.9× on 4 cores.

2. XML PARSING

XML has established itself as the de facto standard for exchanging semi-structured data in an interoperable and standardized way, leading to several efforts to speed up parsing by manual parallelization [7, 12, 13], or even by hardware acceleration [8]. Mobile browsers also spend considerable resources on the very similar task of HTML parsing and DOM tree construction. As an example, on the Exynos 5410 SoC, Chromium spends 17% of its execution time and 16% of its energy on this task [15].

In order to understand why CMT performs poorly on XML parsing, the rest of this section takes a closer look at the parser implementation in libxml2 (Figure 1). We find code properties that limit the amount of parallelism between iterations of the same loop – recursion, non-natural loops and loops with a single iteration.

Code characteristics.

The parser has a very commonly executed recursive cycle (between the functions xmlParseElement and xmlParseContent) for inspecting XML subelements. It builds up significant execution time without any loops whatsoever, limiting the amount of parallelism that loop-centric CMT can extract. Furthermore, not all hot loops are natural, and easily identifiable by a compiler. Such non-natural loops are the result of heavy manual optimization with goto statements (mostly in the function xmlParseStartTag2), some of which jump in the middle of a loop body. These jumps violate the property that a compiler uses to identify a loop: the header of the loop must predominate the loop body. Similarly to recursion, this renders CMT ineffective.

The balance between the functions described above (and hence the quality of CMT parallelization) is strongly dependent on the shape of the input XML tree\footnote{Similar dependences have been observed for manual paral-}.
Program speedup

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>LOC</th>
<th>HELIX-RC Speedup</th>
<th>Parallel loop coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFP2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>177.mesa</td>
<td>42,491</td>
<td>15.1x</td>
<td>(99.9%)</td>
</tr>
<tr>
<td>179.art</td>
<td>1,036</td>
<td>10.5x</td>
<td>99%</td>
</tr>
<tr>
<td>183.equake</td>
<td>1,042</td>
<td>10.1x</td>
<td>99%</td>
</tr>
<tr>
<td>188.annmp</td>
<td>9,805</td>
<td>12.5x</td>
<td>99%</td>
</tr>
<tr>
<td>CNT2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>164.gzip</td>
<td>5,630</td>
<td>6.0x</td>
<td>98.2%</td>
</tr>
<tr>
<td>175.vpr</td>
<td>11,300</td>
<td>6.1x</td>
<td>99%</td>
</tr>
<tr>
<td>181.mcf</td>
<td>1,482</td>
<td>8.7x</td>
<td>99%</td>
</tr>
<tr>
<td>197.parser</td>
<td>7,763</td>
<td>7.3x</td>
<td>98.7%</td>
</tr>
<tr>
<td>256.bzips2</td>
<td>3,235</td>
<td>12.0x</td>
<td>99%</td>
</tr>
<tr>
<td>300.twolf</td>
<td>17,875</td>
<td>7.6x</td>
<td>99%</td>
</tr>
<tr>
<td>libxl2</td>
<td>170,893</td>
<td>1.02x</td>
<td>48%</td>
</tr>
</tbody>
</table>

Table 1: Code complexity (approximated by # lines of code) and achieved speedup between SPEC2000 and libxl2.

4. BEYOND CMT

To overcome the inherent limitation of CMT, Multiple CMT (MCMT) distributes iterations of multiple loops among cores. MCMT runs multiple loops concurrently on different cores by spreading each loop’s iterations on a different core subset. A prototype runtime that implements MCMT achieves speedup between 3.0 and 3.9× on libxml2 parsing, assuming a four-core platform (Figure 3, third column).

In more detail, MCMT includes a runtime which starts executing non-loop code serially. Once it encounters a natural loop, it dispatches its iterations among a subset of cores. Without waiting for these iterations to complete, the runtime continues executing subsequent code. If this code contains another loop, its iterations are also dispatched on different cores, concurrently with those of the first loop. For correctness, data and control dependences are satisfied through normal synchronization. Finally, we optimistically assume zero overhead for the dispatch decisions.

While MCMT has potential, it requires run time support for accurate data dependence information. This is confirmed by the large difference between the first two sets of bars in Figure 3 – using static and dynamic data dependence graph (DDG). This is expected because obtaining high accuracy on data dependence analysis is a challenge on a large codebase. Moreover, because MCMT explicitly targets multiple loops, its DDG analysis cannot look into single loops in isolation, and must potentially address the entirety of a large codebase. This makes static analyses impractical for codes the size of libxml2 (Table 1). Hence the need for a run time approach potentially based on speculative multithreading [10].

One last optimization for MCMT allows it to capture recursion better. In this more general model, iterations of different invocations of the same loop can also run in parallel. The effect of this optimization is stronger on more recursive inputs – Figure 3, the gap between the second and the third column.
5. REFERENCES


