Parallel Scanning with Bitstream Addition: An XML Case Study

Robert D. Cameron *, Ehsan Amiri, Kenneth S. Herdy, Dan Lin, Thomas C. Shermer, and Fred P. Popowich

Simon Fraser University, Surrey, BC, Canada
{cameron, eamiri, ksherdy, lindanl, shermer, popowich}@cs.sfu.ca

Abstract. A parallel scanning method using the concept of bitstream addition is introduced and studied in application to the problem of XML parsing and well-formedness checking. On processors supporting W-bit addition operations, the method can perform up to W finite state transitions per instruction. The method is based on the concept of parallel bitstream technology, in which parallel streams of bits are formed such that each stream comprises bits in one-to-one correspondence with the character code units of a source data stream. Parsing routines are initially prototyped in Python using its native support for unbounded integers to represent arbitrary-length bitstreams. A compiler then translates the Python code into low-level C-based implementations. These low-level implementations take advantage of the SIMD (single-instruction multiple-data) capabilities of commodity processors to yield a dramatic speed-up over traditional alternatives employing byte-at-a-time parsing.

Keywords: SIMD text processing, parallel bitstream technology, XML, parsing

1 Introduction

Traditional byte-at-a-time parsing technology is increasingly mismatched to the capabilities of modern processors. Current commodity processors generally possess 64-bit general purpose registers as well as 128-bit SIMD registers, with 256-bit registers now appearing. General purpose processing on graphics processors can make available 512-bit or wider registers. Parsing models based on the traditional loading and processing of 8 bits at a time would seem to be greatly underutilizing processor resources.

Unfortunately, parsing is hard to parallelize. Indeed, in their seminal report outlining the landscape of parallel computing research, researchers from Berkeley identified the finite state machine methods underlying parsing and lexical processing as the hardest of the "13 dwarves" to parallelize, concluding at one point that "nothing helps." [1] SIMD methods, in particular, would seem to be ill-suited to parsing, because textual data streams are seldom organized in

*

convenient 16-byte blocks, tending to consist instead of variable-length items in generally unpredictable patterns. Nevertheless, there have been some notable works such as that of Scarpazza in applying the multicore and SIMD capabilities of the Cell/BE processor to regular expression matching [13] Intel has also signalled the importance of accelerated string processing to its customers through the introduction of new string processing instructions in the SSE 4.2 instruction set extension, demonstrating how those features may be used to advantage in activities such as XML parsing [11].

Our research has been exploring a promising alternative approach, however, based on the concept of *parallel bit streams* [3–5]. In this approach, byte streams are first sliced into eight basis bit streams, one for each bit position within the byte. Bit stream i thus comprises the ith bit of each byte. Using 128-bit SIMD registers, then, bitwise logic operations on these basis bit streams allows byte classification operations to be carried out in parallel for 128 bytes at a time. For example, consider a character class bit stream `[<]` using a 1 bit to mark the position of opening angle brackets in a byte stream. This stream may computed as logical combination of the basis bit streams using only seven bitwise logical operations per 128 bytes.

Based on this approach, our prior work has shown how parallel bit streams may be used to accelerate XML parsing by further taking advantage of processor *bit scan* instructions, commonly found within commodity processors[5]. On current Intel or AMD processors, for example, these instructions allow one to determine the position of the first 1 bit in a group of 64 in a single instruction. Using these techniques, our Parabix 1 parser demonstrated offered considerable accelaration of XML parsing in statistics gathering [5] as well as GML to SVG conversion [8].

In this paper, we further increase the parallelism in our methods by introducing a new parallel scanning primitive using bitstream addition. In essence, multiple 1 bits in a marker stream identify current scanning positions for multiple instances of a particular syntactic context within a byte stream. These multiple marker positions may each be independently advanced in parallel using addition and masking. The net result is a new scanning primitive that allows multiple instances of syntactic elements to be parsed simultaneously. For example, in dense XML markup, one might find several instances of particular types of markup tags within a given 64-byte block of text; parallel addition on 64-bit words allows all such instances to be processed at once.

Other efforts to accelerate XML parsing include the use of custom XML chips [12], FPGAs [7], and multithread/multicore speedups based on fast preparsing [14].

The remainder of this paper is organized as follows. Section 2 reviews the basics of parallel bitstream technology and introduces our new parallel scanning primitive. Section 3 illustrates how this primitive may be used in the lexical processing of XML references including the parallel identification of errors. Section 4 goes on to consider the more complex task of XML tag processing that goes beyond mere tokenization. Building on these methods, Section 5 de-

scribes how to construct a complete solution to the problem of XML parsing and well-formedness checking, in order to gauge the applicability and power of the techniques. Section 5 then considers the translation of high-level operations on unbounded bitstreams into equivalent low-level code using SIMD intrinsics in the C programming language. Performance results are presented in section 7, comparing the results of our generated implementations in comparison with a number of existing alternatives. The paper concludes with comments on the current status of the work and directions for further research.

2 The Parallel Bitstream Method

2.1 Fundamentals

A bitstream is simply a sequence of 0s and 1s, where there is one such bit in the bitstream for each character in a source data stream. For parsing, and other text processing tasks, we need to consider multiple properties of characters at different stages during the parsing process. A bitstream can be associated with each of these properties, and hence there will be multiple (parallel) bitstreams associated with a source data stream of characters [3, 4].

The starting point for bitstream methods are *basis* bitstreams and their use in determining *character-class* bitstreams. The kth basis bitstream B_k consists of the kth bit (0-based, starting at the LSB) of each character in the source data stream; thus each B_k is dependent on the encoding of the source characters (ASCII, UTF-8, UTF-16, etc.). Given these basis bit streams, it is then possible to combine them using bitwise logic in order to compute character-class bit streams, that is, streams that identify the positions at which characters belonging to a particular class occur. For example, the character class bitstream $D =$ `[0-9]` marks with 1s the positions at which decimal digits occur. These bitstreams are illustrated in Figure 1, for an example source data stream consisting of digits and hyphens. This figure also illustrates some of our conventions for figures: the left triangle $\triangleleft$ after "source data" indicates that all streams are read from right

```
source data ◁ ----173942---654----1----49731----321--
     B7       ..................................
     B6       ..................................
     B5       11111111111111111111111111111111111111
     B4       ....111111...111....1....11111....111..
     B3       1111...1..111...1111.1111.1...1111...11
     B2       1111.1..1.1111111111.11111.1..1111...11
     B1       .....11..1...1.............11.....11...
     B0       11111111..111.1.111111111.111111111.111
   [0-9]      ....111111...111....1....11111....111..
```

Fig. 1. Basis and Character-Class Bitstreams

to left (i.e., they are in little-endian notation). We also use hyphens in the input stream represent any character that is not relevant to a character class under consideration, so that relevant characters stand out. Furthermore, the 0 bits in the bitstreams are represented by periods, so that the 1 bits stand out.

Transposition of source data to basis bit streams and calculation of character-class streams in this way is an overhead on parallel bit stream applications, in general. However, using the SIMD capabilities of current commodity processors, these operations are quite fast, with an amortized overhead of about 1 CPU cycle per byte for transposition and less than 1 CPU cycle per byte for all the character classes needed for XML parsing [5]. Improved instruction sets using parallel extract operations or inductive doubling techniques may further reduce this overhead significantly [6, 9].

Beyond the bitwise logic needed for character class determination, we also need *upshifting* to deal with sequential combination. The upshift $n(S)$ of a bitstream S is obtained by shifting the bits in S one position forward, then placing a 0 bit in the starting position of the bitstream; n is meant to be mnemonic of "next". In $n(S)$, the last bit of S may be eliminated or retained for error-testing purposes.

2.2 A Parallel Scanning Primitive

In this section, we introduce the principal new feature of the paper, a parallel scanning method based on bitstream addition. Key to this method is the concept of *marker* bitstreams. Marker bitstreams are used to represent positions of interest in the scanning or parsing of a source data stream. The appearance of a 1 at a position in a marker bitstream could, for example, denote the starting position of an XML tag in the data stream. In general, the set of bit positions in a marker bitstream may be considered to be the current parsing positions of multiple parses taking place in parallel throughout the source data stream.

Figure 2 illustrates the basic concept underlying parallel parsing with bitstream addition. As with the previous figures, all streams are shown in little-endian representation, with streams reading from right-to-left. The first row shows a source data stream that includes three spans of digits, 13840, 1139845, and 127, with other nondigit characters shown as hyphens. The second row specifies the parsing problem using a marker bitstream M_0 to mark three initial marker positions at the start of each span of digits. The parallel parsing task is to move each of the three markers forward through the corresponding spans of digits to the immediately following positions.

The third row of Figure 2 shows the derived character-class bitstream D identifying positions of all digits in the source stream. The fourth row then illustrates the key concept: marker movement is achieved by binary addition of the marker and character class bitstreams. As a marker 1 bit is combined using binary addition to a span of 1s, each 1 in the span becomes 0, generating a carry to add to the next position to the left. For each span, the process terminates at the left end of the span, generating a 1 bit in the immediately following position. In this way, binary addition produces the marker bitstream M_1, with each of the

source data $\lhd$	`--721----5489311-----04831------`
M_0	`....1.........1.........1......`
D = `[0..9]`	`..111....1111111.....11111......`
$M_1 = M_0 + D$	`.1......1..........1..........`

Fig. 2. Bitstream addition

three markers moved independently through their respective spans of digits to the position at the end.

However, the simple addition technique shown in Figure 2 does not account for digits in the source stream that do not play a role in a particular scanning operation. Figure 3 shows an example and how this may be resolved. The source data stream is again shown in row 1, and the marker bitstream defining the initial marker positions for the the parallel parsing tasks shown in row 2. Row 3 again contains the character class bitstream for digits D. Row 4 shows the result of bitstream addition, in which marker bits are advanced, but additional bits not involved in the scan operation are included in the result. However, these are easily removed in row 5, by masking off any bits from the digit bitstream; these can never be marker positions resulting from a scan.

source data $\lhd$	`--134--31--59127---3--3474--`
M_0	`....1.........1.........1..`
D = `[0..9]`	`..111..11..11111...1..1111..`
$M_1 = M_0 + D$	`.1.....11.1....1...1.1......`
$M_2 = (M_0 + D) \wedge \neg D$	`.1........1..........1......`

Fig. 3. Parallel Scan Using Addition and Mask

The addition and masking technique allows matching of the regular expression `[0-9]*` for any reasonable (conflict-free) set of initial markers specified in M_0. A conflict occurs when a span from one marker would run into another marker position. However, such conflicts do not occur with the normal methods of marker bitstream formation, in which unique syntactic features of the input stream are used to specify the initial marker positions.

In the remainder of this paper, the notation $s(M, C)$ denotes the operation to scan from an initial set of marker positions M through the spans of characters belonging to a character class C found at each position.

$$s(M, C) = (M_0 + C) \wedge \neg C$$

3 XML Scanning and Parsing

We now consider how the parallel scanning primitive can be applied to the following problems in scanning and parsing of XML structures: (1) parallel scanning of XML decimal character references, and (2) parallel parsing of XML start tags. The grammar of these structures is shown in Figure 4.

```
   DecRef ::= '&#' Digit+ ';'
    Digit ::= [0-9]
    STag ::= '<' Name (WS Attribute)* WS? '>'
Attribute ::= Name WS? '=' WS? AttValue
AttValue ::= '"' [^<"]* '"' | "'" [^<']* "'"
```

Fig. 4. XML Grammar: Decimal Character References and Start Tags

Figure 5 shows the parallel parsing of decimal references together with error checking. The source data includes four instances of potential decimal references beginning with the `&` character. Of these, only the first one is legal according to the decimal reference syntax, the other three instances are in error. These references may be parsed in parallel as follows. The starting marker bitstream M_0 is formed from the `[&]` character-class bitstream as shown in the second row. The next row shows the result of the marker advance operation $n(M_0)$ to produce the new marker bitstream M_1. At this point, a hash mark is required, so the first error bitstream E_0 is formed using a bitwise "and" operation combined with negation, to indicate violations of this condition. Marker bitstream M_2 is then defined as those positions immediately following any M_1 positions not in error. In the following row, the condition that at least one digit is required is checked to produce error bitstream E_1. A parallel scan operation is then applied through the digit sequences as shown in the next row to produce marker bitstream M_3. The final error bitstream E_2 is produced to identify any references without a

source data ▷	`-&#978;-&9;--&#;--&#13!-`
M_0	`.1.....1....1....1.....`
$M_1 = n(M_0)$	`..1.....1....1....1....`
$E_0 = M_1 \wedge \neg$`[#]`	`.........1.............`
$M_2 = n(M_1 \wedge \neg E_0)$	`...1..........1....1...`
$E_1 = M_2 \wedge \neg D$	`..............1........`
$M_3 = s(M_2 \wedge \neg E_1, D)$	`......1..............1.`
$E_2 = M_3 \wedge \neg$`[;]`	`......................1.`
$M_4 = M_3 \wedge \neg E_2$	`......1.................`
$E = E_0 \mid E_1 \mid E_2$	`.........1.....1......1.`

Fig. 5. Parsing Decimal References

closing semicolon. In the penultimate row, the final marker bitstream M_4 marks the positions of all fully-checked decimal references, while the last row defines a unified error bitstream E indicating the positions of all detected errors.

One question that may arise is: how are marker bitstreams initialized? In general, this is an important problem, and dependent on the task at hand. In the XML parsing context, we rely on an important property of well-formed XML: after an initial filtering pass to identify XML comments, processing instructions and CDATA sections, every remaining `<` in the file must be the initial character of a start, end or empty element tag, and every remaining `&` must be the initial character of a general entity or character reference. These assumptions permit easy creation of marker bitstreams for XML tags and XML references.

The parsing of XML start tags is a richer problem, involving sequential structure of attribute-value pairs as shown in Figure 4. Using the bitstream addition technique, our method is to start with the opening angle bracket of all tags as the initial marker bitstream for parsing the tags in parallel, advance through the element name and then use an iterative process to move through attribute-value pairs.

Figure 6 illustrates the parallel parsing of three XML start tags. The figure omits determination of error bitstreams, processing of single-quoted attribute values and handling of empty element tags, for simplicity. In this figure, the first four rows show the source data and three character class bitstreams: N for characters permitted in XML names, W for whitespace characters, and Q for characters permitted within a double-quoted attribute value string.

The parsing process is illustrated in the remaining rows of the figure. Each successive row shows the set of parsing markers as they advance in parallel using bitwise logic and addition. Overall, the sets of marker transitions can be divided into three groups.

The first group M_0 through $M_{0,8}$ shows the initiation of parsing for each of the tags through the opening angle brackets and the element names, up to the first attribute name, if present. Note that the there are no attribute names in the final tag shown, so the corresponding marker becomes zeroed out at the closing angle bracket. Since $M_{0,8}$ is not all 0s, the parsing continues.

The second group of marker transitions $M_{1,1}$ through $M_{1,8}$ deal with the parallel parsing of the first attribute-value pair of the remaining tags. After these operations, there are no more attributes in the first tag, so its corresponding marker becomes zeroed out. However, $M_{1,8}$ is not all 0s, as the second tags still has an unparsed attribute-value pair. Thus, the parsing continues.

The third group of marker transitions $M_{2,1}$ through $M_{2,8}$ deal with the parsing of the second attribute-value pair of this tag. The final transition to $M_{2,8}$ shows the zeroing out of all remaining markers once two iterations of attribute-value processing have taken place. Since $M_{2,8}$ is all 0s, start tag parsing stops.

The implementation of start tag processing uses a while loop that terminates when the set of active markers becomes zero, i.e. when some $M_{k,8} = 0$. Considered as an iteration over unbounded bitstreams, all start tags in the document are processed in parallel, using a number of iterations equal to the maximum

```
source data ▷                --<e a= "137">---<el2 a="17" a2="3379">---<x>--
N = name chars               11.1.1...111..111.111.1..11..11..1111..111.1.11
W = white space              ....1..1.............1......1..................
Q = ¬["">]                   11.11111.111.1111.111111.11.1111.1111.1111.1111

M0                           ..1.............1.........................1....
M1 = n(M0)                   ...1.............1.........................1...
M0,7 = s(M1, N)              ....1.............1.........................1..
M0,8 = s(M0,7, W) ∧ ¬[>]     .....1.............1.......................

M1,1 = s(M0,8                ......1.............1.......................
M1,2 = s(M1,1, W)∧[=]        ......1.............1.......................
M1,3 = n(M1,2)               .......1.............1......................
M1,4 = s(1, 3, W)∧["]        ........1.............1.....................
M1,5 = n(M1,4)               .........1.............1....................
M1,6 = s(M1,5, Q)∧["]        ............1.............1..................
M1,7 = n(M1,6)               .............1.............1.................
M1,8 = s(M1,7, W) ∧ ¬[>]     ............................1.................

M2,1 = s(M1,8, N)            ..............................1..............
M2,2 = s(M2,1, W)∧[=]        ..............................1..............
M2,3 = n(M2,2)               ...............................1.............
M2,4 = s(M2,3, W)∧["]        ...............................1.............
M2,5 = n(M2,4)               ................................1............
M2,6 = s(M2,5, Q)∧["]        ....................................1........
M2,7 = n(M2,6)               .....................................1.......
M2,8 = s(M2,7, W) ∧ ¬[>]     ........................................
```

Fig. 6. Start Tag Parsing

number of attribute-value pairs in any one tag in the document. However, in block-by-block processing, the cost of iteration is considerably reduced; the iteration for each block only requires as many steps as there are attribute-value pairs overlapping the block.

Following the pattern shown here, the remaining syntactic features of XML markup can similarly be parsed with bitstream based methods. One complexity is that the parsing of comments, CDATA sections and processing instructions must be performed first to determine those regions of text within which ordinary XML markups are not parsed (i.e., within each of these types of construct. This is handled by first performance the parsing of these structures and then forming a *mask bitstream*, that is, a stream that identifies spans of text to be excluded from parsing (comment and CDATA interiors, parameter text to processing instructions).

4 XML Well-Formedness

In this section, we consider the full application of the parsing techniques of the previous section to the problem of XML well-formedness checking [2]. This application is useful as a well-defined and commercially significant example to assess the overall applicability of parallel bit stream techniques. To what extent can the well-formedness requirements of XML be completely discharged using parallel bitstream techniques? Are those techniques worthwhile in every instance, or do better alternatives exist for certain requirements? For those requirements that cannot be fully implemented using parallel bitstream technology alone, what preprocessing support can be offered by parallel bit stream technology to the discharge of these requirements in other ways? We address each of these questions in this section, and look not only at the question of well-formedness, but also at the identification of error positions in documents that are not well-formed.

Most of the requirements of XML well-formedness checking can be implemented using two particular types of computed bitstream: *error bitstreams*, introduced in the previous section, and *error-check bitstreams*. Recall that an error bitstream stream is a stream marking the location of definite errors in accordance with a particular requirement. For example, the E_0, E_1, and E_2 bitstreams as computed during parsing of decimal character references in Figure 5 are error bitstreams. One bits mark definite errors and zero bits mark the absence of error according to the requirement. Thus the complete absence of errors according to the requirements listed may be determined by forming the bitwise logical "or" of these bitstreams and confirming that the resulting value is zero. An error check bitstream is one that marks potential errors to be further checked in some fashion during post-bitstream processing. An example is the bitstream marking the start positions of CDATA sections. This is a useful information stream computed during bitstream processing to identify opening `<![` sequences, but also marks positions to subsequently check for the complete opening delimiter `<![CDATA[` at each position.

In typical documents, most of these error-check streams will be quite sparse or even zero. Many of the error conditions could actually be fully implemented using bitstream techniques, but at the cost of a number of additional logical and shift operations. In general, however, the conditions are easier and more efficient to check one-at-a-time using multibyte comparisons on the original source data stream. With very sparse streams, it is very unlikely that multiple instances occur within any given block, thus eliminating the benefit of parallel evaluation of the logic.

The requirement for name checking merits comment. XML names may use a wide range of Unicode character values. It is too expensive to check every instance of an XML name against the full range of possible values. However, it is possible and quite inexpensive to use parallel bitstream techniques to verify that any ASCII characters within a name are indeed legal name start characters or name characters. Furthermore, the characters that may legally follow a name in XML are confined to the ASCII range. This makes it useful to define

a name scan character class to include all the legal ASCII characters for names as well as all non-ASCII characters. A namecheck character class bitstream will then be defined to identify nonASCII characters found within namescans. In most documents this bitstream will be all 0s; even in documents with substantial internationalized content, the tag and attribute names used to define the document schema tend to be confined to the ASCII repertoire. In the case that this bitstream is nonempty, the positions of all 1 bits in this bitstream denote characters that need to be individually validated.

Attribute names within a single XML start tag or empty element tag must be unique. This requirement could be implemented using one of several different approaches. Standard approaches include: sequential search, symbol lookup, and Bloom filters [7].

In general, the use of error-check bitstreams is a straightforward, convenient and reasonably efficient mechanism for checking the well-formedness requirements.

Except for empty element tags, XML tags come in pairs with names that must be matched. To discharge this requirement, we form a bitstream consisting of the disjunction of three bitstreams formed during parsing: the bitstream marking the positions of start or empty tags (which have a common initial structure), the bitstream marking tags that end using the empty tag syntax ("`/>`"), and the bitstream marking the occurrences of end tags. In post-bitstream processing, we iterate through this computed bitstream and match tags using an iterative stack-based approach.

An XML document consists of a single root element with recursively defined structure together with material in the document prolog and epilogs. Verifying this top-level structure and the structure of the prolog and epilog material is not well suited to parallel bitstream techniques, in particular, nor to any form of parallelism, in general. In essence, the prolog and epilog materials occur once per document instance Thus the requirements to check this top-level structure for well-formedness are relatively minor, with an overhead that is quite low for sufficiently sized files.

Overall, parallel bitstream techniques are quite well-suited to verification problems such as XML well-formedness checking. Many of the character validation and syntax checking requirements can be conveniently and efficiently implemented using error streams. Other requirements are also supported by the computation of error-check streams for simple post-bitstream processing or composite stream over which iterative stack-based procedures can be defined for checking recursive syntax.

5 Compilation to Block-Based Processing

While a Python implementation of the techniques described in the previous section works on unbounded bitstreams, a corresponding C implementation needs to process an input stream in blocks of size equal to the SIMD register width of the processor it runs on. So, to convert Python code into C, the key question

becomes how to transfer information from one block to the next one. The answer lies in the use of *carry bits*, the collection of carries resulting from bitwise additions.

In fact, in the methods we have outlined, all the the information flow between blocks for parallel bit stream calculations can be modeled using carry bits. The parallel scanning primitive uses only addition and bitwise logic. Since the logic operations do not require information flow accross block boundaries, the information flow is entirely accounted by the carry. Carry bits can also be used to capture the information flow associated with upshift operations, which move information forward one position in the file. In essence, an upshift by one position for a bitstream is equivalent to the addition of the stream to itself; the bit shifted out in an upshift is in this case equivalent to the carry generated by the additon. The only other information flow requirement in the calculation of parallel bit streams occurs with the bitstream subtractions that are used to calculate span streams. In this case, the information flow is based on borrows generated, which can be handled in the same way as carries.

Properly determining, initializing and inserting carry bits into a block-by-block implementation of parallel bit stream code is a task too tedious for manual implementation. We have thus developed compiler technology to automatically insert declarations, initializations and carry save/restore operations into appropriate locations when translating Python operations on unbounded bit streams into the equivalent low-level C code implemented on a block-by-block bases. Our current compiler toolkit is capable of inserting carry logic using a variety of strategies, including both simulated carry bit processing with SIMD registers, as well as carry-flag processing using the processor general purpose registers and ALU. Details are beyond the scope of this paper, but are described in the on-line source code repository at parabix.costar.sfu.ca.

6 Performance Results

In this section, we compare the performance of our `xmlwf` implementation using the Parabix2 technology described above with several other implementations. These include the original `xmlwf` distributed as an example application of the `expat` XML parser, implementations based on the widely used Xerces open source parser using both SAX and DOM interfaces, and an implementation using our prior Parabix 1 technology with bit scan operations.

Table 1 shows the document characteristics of the XML instances selected for this performance study, including both document-oriented and data-oriented XML files. The jawiki.xml and dewiki.xml XML files are document-oriented XML instances of Wikimedia books, written in Japanese and German, respectively. The remaining files are data-oriented. The roads.gml file is an instance of Geography Markup Language (GML), a modeling language for geographic information systems as well as an open interchange format for geographic transactions on the Internet [10]. The po.xml file is an example of purchase order data, while the soap.xml file contains a large SOAP message. Markup density

is defined as the ratio of the total markup contained within an XML file to the total XML document size. This metric is reported for each document.

File Name	dewiki.xml	jawiki.xml	roads.gml	po.xml	soap.xml
File Type	document	document	data	data	data
File Size (kB)	66240	7343	11584	76450	2717
Markup Item Count	406792	74882	280724	4634110	18004
Attribute Count	18808	3529	160416	463397	30001
Avg. Attribute Size	8	8	6	5	9
Markup Density	0.07	0.13	0.57	0.76	0.87

Table 1. XML Document Characteristics

Table 2 shows performance measurements for the various `xmlwf` implementations applied to the test suite. Measurements are made on a single core of an Intel Core 2 system running a stock 64-bit Ubuntu 10.10 operating system, with all applications compiled with llvm-gcc 4.4.5 optimization level 3. Measurements are reported in CPU cycles per input byte of the XML data files in each case. The first row shows the performance of the Xerces C parser using the tree-building DOM interface. Note that the performance varies considerably depending on markup density. Note also that the DOM tree construction overhead is substantial and unnecessary for XML well-formedness checking. Using the event-based SAX interface to Xerces gives much better results as shown in the second row. The third row shows the best performance of our byte-at-a-time parsers, using the original `xmlwf` based on expat.

The remaining rows of Table 2 show performance of parallel bit stream implementations. The first row shows the performance of our Parabix 1 implementation using bit scan instructions. While showing a substantial speed-up over the byte-at-a-time parsers in every case, note also that the performance advantage increases with increasing markup density, as expected. The last two rows show different versions of the `xmlwf` implemented based on the Parabix 2 technology as discussed in this paper. They differ in the carry handling strategy, with the "simd_add" row referring to carry computations performed with simulated calculation of propagated and generated carries using SIMD operations, while the "adc64" row refers to an implementation directly employing the processor carry flags and add-with-carry instructions on 64-bit general registers. In both cases, the overall performance is quite impressive, with the increased parallelism of parallel bit scans clearly paying off in improved performance for dense markup.

7 Conclusion

In application to the problem of XML parsing and well-formedness checking, the method of parallel parsing with bitstream addition is effective and efficient.

Parser Class	Parser	dewiki.xml	jawiki.xml	roads.gml	po.xml	soap.xml
Byte-at-a-time	Xerces (DOM)	37.921	40.559	72.78	105.497	125.929
	Xerces (SAX)	19.829	24.883	33.435	46.891	57.119
	expat	12.639	16.535	32.717	42.982	51.468
Parallel Bit Stream	Parabix1	8.313	9.335	13.345	16.136	19.047
	Parabix2 (simd_add)	6.103	6.445	8.034	8.685	9.53
	Parabix2 (adc64)	5.123	5.996	6.852	7.648	8.275

Table 2. Parser Performance (Cycles Per Byte)

Using only bitstream addition and bitwise logic, it is possible to handle all of the character validation, lexical recognition and parsing problems except for the recursive aspects of start and end tag matching. Error checking is elegantly supported through the use of error streams that eliminate separate if-statements to check for errors with each byte. The techniques are generally very efficient particularly when markup density is high. However, for some conditions that occur rarely and/or require complex combinations of upshifting and logic, it may be better to define simpler error-check streams that require limited postprocessing using byte matching techniques.

The techniques have been implemented and assessed for present-day commodity processors employing current SIMD technology. As processor advances see improved instruction sets and increases in width of SIMD registers, the relative advantages of the techniques over traditional byte-at-a-time sequential parsing methods is likely to increase substantially. Of particular benefit to this method, instruction set modifications that provide for more convenient carry propagation for long bitstream arithmetic would be most welcome.

A significant challenge to the application of these techniques is the difficulty of programming. The method of prototyping on unbounded bitstreams has proven to be of significant value in our work. Using the prototyping language as input to a bitstream compiler has also proven effective in generating high-performance code. Nevertheless, direct programming with bitstreams is still a specialized skill; our future research includes developing yet higher level tools to generate efficient bitstream implementations from grammars, regular expressions and other text processing formalisms.

References

1. Krste Asanovic, Ras Bodik, Bryan Christopher Catanzaro, Joseph James Gebis, Parry Husbands, Kurt Keutzer, David A. Patterson, William Lester Plishker, John Shalf, Samuel Webb Williams, and Katherine A. Yelick. The landscape of parallel computing research: A view from berkeley. Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley, Dec 2006.
2. Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, and Franois Yergeau. Extensible markup language (XML) 1.0 (fifth edition). W3C Recommendation, 2008.

3. Rob Cameron, Ken Herdy, and Ehsan Amiri. Parallel bit stream technology as a foundation for XML parsing performance. In *International Symposium on Processing XML Efficiently: Overcoming Limits on Space, Time, or Bandwidth*, August 2009.
4. Robert D. Cameron. A Case Study in SIMD Text Processing with Parallel Bit Streams. In *ACM Symposium on Principles and Practice of Parallel Programming (PPoPP)*, Salt Lake City, Utah, 2008.
5. Robert D. Cameron, Kenneth S. Herdy, and Dan Lin. High performance XML parsing using parallel bit stream technology. In *CASCON '08: Proceedings of the 2008 conference of the center for advanced studies on collaborative research*, pages 222–235, New York, NY, USA, 2008. ACM.
6. Robert D. Cameron and Dan Lin. Architectural support for swar text processing with parallel bit streams: the inductive doubling principle. In *ASPLOS '09: Proceeding of the 14th international conference on Architectural support for programming languages and operating systems*, pages 337–348, New York, NY, USA, 2009. ACM.
7. Zefu Dai, Nick Ni, and Jianwen Zhu. A 1 cycle-per-byte XML parsing accelerator. In *FPGA '10: Proceedings of the 18th Annual ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, pages 199–208, New York, NY, USA, 2010. ACM.
8. Kenneth S. Herdy, David S. Burggraf, and Robert D. Cameron. High performance GML to SVG transformation for the visual presentation of geographic data in web-based mapping systems. In *Proceedings of SVG Open 2008*, August 2008.
9. Yedidya Hilewitz and Ruby B. Lee. Fast bit compression and expansion with parallel extract and parallel deposit instructions. In *ASAP '06: Proceedings of the IEEE 17th International Conference on Application-specific Systems, Architectures and Processors*, pages 65–72, Washington, DC, USA, 2006. IEEE Computer Society.
10. Ron Lake, David Burggraf, Milan Trninic, and Laurie Rae. *Geography Mark-Up Language: Foundation for the Geo-Web*, pages 3–4. John Wiley & Sons, Inc., 2004.
11. Zhai Lei. XML parsing accelerator with intel streaming SIMD extensions 4 (intel SSE4). http://software.intel.com/en-us/articles/xml-parsing-accelerator-with-intel-streaming-simd-extensions-4-intel-sse4/, 2008.
12. Michael Leventhal and Eric Lemoine. The XML chip at 6 years. In *International Symposium on Processing XML Efficiently: Overcoming Limits on Space, Time, or Bandwidth*, August 2009.
13. Daniele Paolo Scarpazza and Gregory F. Russell. High-performance regular expression scanning on the cell/b.e. processor. In *Proceedings of the 23rd international conference on Supercomputing*, ICS '09, pages 14–25, New York, NY, USA, 2009. ACM.
14. Ying Zhang, Yinfei Pan, and Kenneth Chiu. Speculative p-dfas for parallel xml parsing. In *High Performance Computing (HiPC), 2009 International Conference on*, pages 388–397, December 2009.