XML Parsing: A Review

Rashmi P. Sonar, M. S. Ali

Abstract- ***A well formed and valid XML document is always a standard of efficient transmission of data for internet. To increase the performance of XML document, right choice of parser is always an important issue. In this paper we try to review some parsing techniques for XML documents so that researches related to XML can get new possibility of the parsing methods.***

Index Terms— XML parsing, DOM, SAX

I. INTRODUCTION

XML processing consist of four steps parsing, access, modification, and serialization. XML parsing is an important operation in reading a document with wellformedness and validating the document with Schema and DTD. Basically there are two approaches-

1) Document Object Model.(DOM)- DOM is tree based API, which can use XPath for finding information. In DOM, tokens are extract as object and build tree of object i.e nodes. DOM gives random access to XML document but it is expensive in terms of loading the tree into memory.

2) Simple API for XML (SAX)- SAX is a event based API. In SAX, tokens are extract as object and object creates the events for example string. This type of API reports parsing events directly to application. No random access is available in SAX.

Further XML parsers can also be classified on computation power as " heavy" and "light" parser. Heavy parsers are designed to provide advance computation power and Light parsers are designed for limited processing power and memory availability [1][2][3].

II. XML PARSING METHODS

We require parsing as one of the important part in XML data processing. The correctness of XML depends on welformedness and validation .In this section we are going to review some parsing techniques which are based on basic DOM and SAX model. These techniques allow to handle the XML data efficiently to increase the performance. Prefiltering technique is one of the method to increase performance ,In[4] author proposed a framework within the existing DOM and SAX model and can access the large XML documents based on the approximate execution of user's query. The special requirement of this type of technique is that it must guarantee that 100% call rate for maintaining the correctness of user application. This model of XML processing can be used where the data need infrequent updates. Constructing finite automata method is based on deterministic finite automaton from ordered schema. It translates XML schema to DFA and SOAP/XML messages, which can be used for high performance web services [5]. The table driven streaming XML parsing(TDX) method is a schema specific parsing method. It consists of three stages: specification, processing, code generation, and run-time processing. TDX is based on linguistic principles, where XML parsing and validation are in context free grammar rule. All the constraint of XML schema are being checked by grammar productions at run time. Application specific events are converted as semantic actions in tightly defined production rules. TDX tables are interchangeable easily whenever the XML schema is updated along with parsing and validation [6]. Table-driven permutation phrase grammar parsing is a schema specific parsing technique which is optimal in terms of time and space [7]. In the era of technology, computers will have more cores day by day with faster clock speed and softwares are going to rely on parallelism of task. In [8], the parallel XML parsing model is discussed along with experimental result on four core. Fig 1 shows the architecture of parallel parsing model

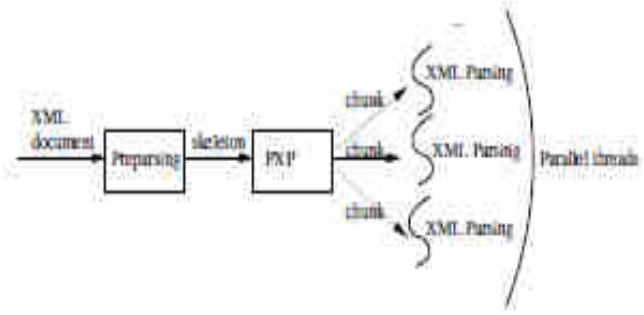

Fig1: PXP Archiecture

The approach focuses on DOM-style parsing, in which a tree data structure of document is created in memory. The procedure for parse and generating skeleton for XML document is called preparsing. Once the preparsing is done , define initial pass for logical structure of the XML document and Logical structure is then divides into chunks with proper XML grammar. Here dynamic partitioning balances the load during parsing and can apply to any irregular tree structure. PiXiMaL is the parallel processing library for large scale XML data files which takes advantage of Chip multi processor. In this approach, an effective scheme to parallelize the tokenization process along with DFA based parser which recognizes subsets of XML and converts DFA to NFA which can be applied to any subset of input. The output generate asa sequence of SAX events [9].

A Data Parallel Algorithm called ParDOM discusses the XML DOM parsing which supports map and sort operations. ParDOM consist of two phase,

1) Phase I- Chunk creation, where construction of partial DOM structure on chunks of XML document is created.
2) Phase II- Linking Partial DOM Trees- where the linking of partial structure is done.

The sequence of task in ParDOM is given in Figure 2.

Revised Version Manuscript Received on May 25, 2015.

Rashmi P. Sonar, Department of Computer Sc & Engineering, Prof Ram Meghe College of Engineering and Management Badnera, Amravati.

Dr. M. S. Ali, Department of Computer Sc & Engineering, Prof Ram Meghe College of Engineering and Management Badnera, Amravati.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

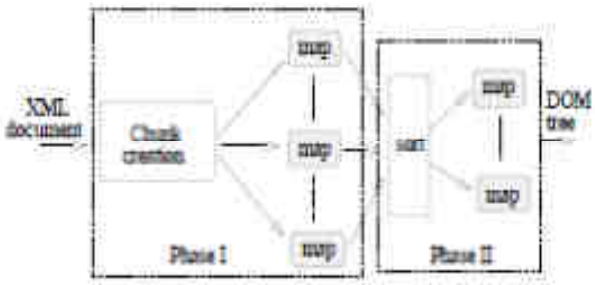

Fig 2: Sequence of task in ParDOM

When the technique is compared with PXP, it gives better scalable results on multicore processor along with wide variety of dataset with complex structure [10].

Hybrid Parallelism for XML SAX Parsing is basically parallel XML SAX parser with four stage software pipeline every sequential stage is followed by data parallel stage. The first stage address the ordering dependencies and the second stage the processed chunks of the data streams are available in data parallel style. In third stage, it is possible to address the data dependency in fast sequential stage and the fourth stage is again the data parallel stage with computationally intense data which be then process chunks of the incoming data stream independently. The is technique is basically useful to handle internet data dependencies[11]. In [12],a method of retrieving data from DBMS is defined ,DOM based technique is used to parse XML documents and read into memory along with tree structure creation. System will send a query to cache then XML query is send to DBMS so that DBMS will search and retrieve the data.

The main issue in XML processing is the structure of the document. By referring structure we can identify the entity in the result. If the redundancy of structure related processing is reduced, the performance of XML processing increases. Dong Zhou presents the concept of structure encoding and identify recurring structure. The processing optimization can be categorized as-

1) API Specialization
2) Data Structure Optimization
3) Structure-related optimization

Dong zhou suggested some approaches to quickly identifying recurring structure in mobile environment, including collision resistant hash function [13].

XML is known for interoperatibility and ease of use. In [14], author defines how XML and SOAP is useful in small devices with constructing the properties of XML. Experimental analysis with microcontroller board shows how the performance can achieve with strict resources and how XML can be useful to address this. EXDOM is a Embedded XML DOM Parser designed for data analysis on Networked Embedded System and developed using Java 2 Micro Edition. A set of optimization practices like classmerging, elimination of variables, or method *Inlining* can reduce code size or Heap usage with more availability for memory location for other task. The design of EXDOM based on memory reuse and introduced as a solution to XML processing in environments that provide limited memory and computing power [4].

SCBXP is hardware technique for XML processing on mobile networks along with limited memory resource. The architecture of SCBXP consist of –

1) Two dual-port memory modules—one is located first for loading stage and second for reading stage.
2) Four 8-bit FIFOs and their associated control module that are part of the aligning stage.
3) An XML aligning state machine that operates in the aligning stage.
4) A CAM that is the main part of the matching stage.
5) Five parsing state machines that act in the post matching stage.
6) A scheduler/writer module that operates in the scheduling /writing stage

The technique is implemented on FPGA with rate of 2bytes of XML data per clock cycle and ensures the fully well formed XML [15].

In [16], memory-side data loading in the parsing stage incurs a significant performance overhead, as much as the computation does. Here the focus is on computation acceleration of XML parsing. XML parsing from memory side can reduce cache misses upto 80%.

III. CONCLUSION

Due to flexibility and interoperatibilty of XML, efficient performance in XML parsing is always the key requirement in every area. We studied some XML parsing techniques with basic models DOM and SAX. This paper also focused on use XML for internet and embedded system along with parsing methods.

REFERENCES

[1] H.M.Deitel, P.J.Deitel, T.R.Nieto, T.M.Lin, and P.Sadhu, XML How to Program, 2nd ed, LPE,Pearson Education,2009.
[2] [online] http://www.fdi.ucm.es/profesor/jlsierra/e-learning/segunda-sesion/XMLParsingModels.pdf.
[3] Esther Mınguez Collado, M. Angeles Cavia Soto, Jose A. Perez Garc´a,Ivan M.Delamer, and Jose L.Martınez Lastra, "Embedded XML DOM Parser: An Approach for XML DataProcessing on Networked Embedded Systems with Real-Time," EURASIP Journal on Embedded Systems Volume 2008, 6 pages 2008.
[4] Chia-Hsin Huang and Tyng-Ruey Chuang, "Prefiltering Techniques for Efficient XML Document Processing," in Proc. DocEng'05, 2005.
[5] Robert A. van Engelen , "Constructing Finite State Automata for High Performance Web Services," in Proc. IEEE International Conference on Web Services, 2004
[6] Wei Zhang & Robert van Engelen,"A Table-Driven Streaming XML parsing Methodology for High-Performance Web Services," in Proc.ICWS'06,2006
[7] Wei Zhang and Robert A. van Engelen, "High-Performance XML Parsing and Validation with Permutation Phrase Grammar Parsers," in Proc. ICWS ,2008, p.101
[8] W. Lu , K. Chiu, Y. Pan, "A Parallel Approach to XML Parsing," in Proc. Grid06, 2006
[9] Michael R, Madhusudhan Govindaraju, "Parallel Processing of Large-Scale XML-Based Application Documents on Multi-core Architectures with PiXiMaL," in Proc. eScience '08, 2008, p. 261 - 268
[10] Bhavik Shah, Praveen R. Rao, and Bongki Moon and Mohan Rajagopalan, A Data Parallel Algorithm for XML DOM Parsing Lecture Notes in Computer Science, Springer Berlin Heidelberg 2009, Volume 5679.
[11] Yinfei Pan, Ying Zhang, Kenneth Chiu, "Hybrid Parallelism for XML SAX Parsing," in Proc. ICWS '08, 2008.
[12] Yusof Mohd Kamir, Mat Amin Mat Atar, "High Performance of DOM Technique in XML for Data Retrieval," in Proc. ICIMT '09, 2009, Page(s): 303 – 305.
[13] Dong Zhou, " Exploiting Structure Recurrence in XML Processing," in Proc.ICWE '08 ,Pages: 311-324.
[14] Johannes Helander, "Deeply Embedded XML Communication –Towards an Interoperable and Seamless World", in Proc EMSOFT'05, September 19–22, 2005, Jersey City, New Jersey, USA.
[15] Fadi El-Hassan, and Dan Ionescu, "SCBXP: An Efficient CAM-Based XML Parsing Technique in Hardware Environments," IEEE

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Transaction on Parallel and Distributed Systems, vol. 22, pp.1879-1887, Nov 2011

[16] Jie Tang, Shaoshan Liu, Chen Liu, Zhimin Gu, and Jean-Luc Gaudiot, "Acceleration of XMLParsing through Prefetching," IEEE TRANSACTIONS ON COMPUTERS, VOL. 62, NO. 8, AUGUST 2013

Ms. Rashmi P. Sonar, completed her ME in Computer Sc and Engineering in 2011 and now pursuing PhD. Currently, she is working with PRMCEAM, Badnera-Amravati

Dr. M. S. Ali, is the renowned personality in Sant Gadge Baba Amravati University region and working with PRMCEAM,Badnera-Amravati . He completed his PhD in 2006 and sharing his contribution in the field of Research..

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.