SAPLING 2012
A Speculative Parallel DFA Membership Test for Multicore, SIMD and Cloud Computing Environments

Yousun Ko*

November 20, 2012

Abstract

This work presents techniques to parallelize membership tests for Deterministic Finite Automata (DFAs). Our method searches arbitrary regular expressions by matching multiple bytes in parallel using speculation. We partition the input string into chunks, match chunks in parallel, and combine the matching results. Our parallel matching algorithm exploits structural DFA properties to minimize the speculative overhead. Unlike previous approaches, our speculation is *failure-free*, i.e., (1) sequential semantics are maintained, and (2) speed-downs are avoided altogether. On architectures with a SIMD gather-operation for indexed memory loads, our matching operation is fully vectorized. The proposed load-balancing scheme uses an off-line profiling step to determine the matching capacity of each participating processor. Based on matching capacities, DFA matches are load-balanced on inhomogeneous parallel architectures such as cloud computing environments.

We evaluated our speculative DFA membership test for a representative set of benchmarks from the Perl-compatible Regular Expression (PCRE) library [1] and the PROSITE [2] protein database. Evaluation was conducted on a 4 CPU (40 cores) shared-memory node of the Intel Manycore Testing Lab (Intel MTL), on the Intel AVX2 SDE simulator for 8-way fully vectorized SIMD execution, and on a 20-node (288 cores) cluster on the Amazon EC2 computing cloud. Obtained speedups are on the order of $\mathcal{O}(1+\frac{|P|-1}{|Q|\cdot\gamma})$, where $|P|$ denotes the number of processors or SIMD units, $|Q|$ denotes the number of DFA states, and $0 < \gamma \leq 1$ represents a statically computed DFA property. For all observed cases, we found that $0.16 < \gamma < 0.47$. Actual speedups range from 1.6x to 38.2x for up to 512 states for PCRE, and between 1.2x and 13.9x for up to 766 states for PROSITE on a 40-core MTL node. Not taking communication costs into account, speedups on the EC2 computing cloud range from 5.2x to 173.9x for PCRE, and from 2.2x to 98x for PROSITE. Including communication costs, EC2 speedups range from 5.1x to 71x for PCRE, and between 2.1x and 51.3x for PROSITE protein patterns. Speedups of our C-based DFA matcher over the Perl-based ScanProsite scan tool [3] range from 410.8x to 7781.3x on a 40-core MTL node.

*Dept. of Computer Science, Yonsei University, Korea. School of IT, The University of Sydney, Australia.

References

[1] Perl Compatible Regular Expression Library Web Site. `http://www.pcre.org`, retrieved Aug. 2012.

[2] PROSITE Web Site. `http://prosite.expasy.org`, retrieved Aug. 2012.

[3] ScanProsite Web Site. `http://prosite.expasy.org/scanprosite`, retrieved Aug. 2012.