common performance engineering properties continue to hold. We briefly describe several of these.

Exploiting parallelism is challenging. While we were able to improve performance by up to an impressive 8.4x, this is still well below the potential speedup of 24x that one might have expected, moving from a single-threaded application to a multi-threaded implementation running on a 24 core server. In several cases, the performance actually decreases as additional cores are used, owing to contention amongst the various threads. Clearly, the opportunity remains to improve performance further.

Bottlenecks shift. When we began our work, the performance of the single-threaded application was limited by the use of a single CPU core. When we introduced a multi-threaded implementation, that bottleneck was eliminated, but new ones emerged. In our NFS experiments, the bottleneck then became the 100 Mb/s network link. We eliminated that by upgrading to a 1 Gb/s link. The bottleneck is now contention amongst the decompression and analysis threads. Addressing this bottleneck is left for future work.

6. CONCLUSION AND FUTURE WORK

In this paper we described how we used DataSeries as the online logging format for two popular open source applications, the Apache Web server and the Bro intrusion detection system. We modified the Webalizer tool to efficiently analyze Web server logs in DataSeries format. We also demonstated how to efficiently analyze the Bro Intrusion Detection System logs in DataSeries format. We quantified the benefits of storing and accessing information in DataSeries format relative to the default log format of the chosen applications.

Our experimental results showed significant benefits are possible from leveraging DataSeries and multicore servers. The sizes of the Apache logs decreased by up to 7x (2.6x for Bro), and the time to analyze the Apache logs decreased by 8x (3x for Bro). Our work verifies that DataSeries is beneficial for online logging applications, and that it facilitates efficient analysis. A motivating goal of our work is to have our initial implementations serve as examples that others can use to integrate DataSeries into the applications we examined, or into other applications that generate or analyze structured serial data. We have shared our results and source code with the developers of Bro, who are considering adding DataSeries to a future release of the IDS, as improving the efficiency of Bro is one of their key goals. As future work, we will perform a comparison of the performance of binary log formats and logging libraries with the best applicability areas for each. We intend to communicate our results and source code with the Apache and Webalizer groups, and to continue to search for opportunities to improve the effective use of IT infrastructure.

7. REFERENCES

[1] Analog: a free logfile analyzer. *[Online] Available: http://www.analog.cx/.*

[2] Apache HTTP server project. *[Online] Available: http://httpd.apache.org/download.cgi.*

[3] Apache logging module mod_log_config. *[Online] Available: http://httpd.apache.org/docs/2.0/mod/mod_log_config.html.*

[4] Awstats: a free logfile analyzer. *[Online] Available: http://awstats.sourceforge.net/.*

[5] Bro IDS Reference Manual: Analyzers and Events. *[Online] Available: http://www.bro-ids.org/wiki/index.php/Reference_Manual:_Analyzers_and_Events.*

[6] Bro IDS Reference Manual: Getting Started (the cf utility). *[Online] Available: http://www.bro-ids.org/wiki/index.php/Reference_Manual:_Getting_Started#The_cf_utility.*

[7] Bro intrusion detection system. *[Online] Available: http://www.bro-ids.org/download.html.*

[8] Bro Quick Start Guide. *[Online] Available: http://www.bro-ids.org/Bro-quick-start.pdf.*

[9] Capstats: a quick hack to get some NIC statistics. *[Online] Available: http://www.icir.org/robin/capstats/.*

[10] Conn.log connection summaries. *[Online] Available: http://tinyurl.com/bro-conns.*

[11] DataSeries technical report. *[Online] Available: http://tesla.hpl.hp.com/opensource/DataSeries-tr-snapshot.pdf.*

[12] How do you create a new Apache module? *[Online] Available: http://ivascucristian.com/how-do-you-create-a-new-apache-module.*

[13] Httperf: a tool for measuring web server performance. *[Online] Available: http://www.hpl.hp.com/research/linux/httperf/.*

[14] The Inline::CPP module: put C++ source code directly "inline" in a Perl script. *[Online] Available: http://search.cpan.org/ neilw/Inline-CPP-0.25/lib/Inline/CPP.pod.*

[15] Open Source software at Hewlett-Packard Laboratories. *[Online] Available: http://tesla.hpl.hp.com/opensource/.*

[16] RUBiS: an auction site prototype. *[Online] Available: http://rubis.ow2.org/.*

[17] RUBiSVA: a virtual appliance of the RUBiS benchmark. *[Online] Available: http://rubis.ow2.org/download/rubisva_v1.0.pdf.*

[18] Source files with modifications done within this work. *[Online] Available: http://www.sfu.ca/ sba70/files/dataseries/.*

[19] Tmpfs: a temporary file storage facility. *[Online] Available: http://en.wikipedia.org/wiki/Tmpfs.*

[20] Traces from the Internet Traffic Archive. *[Online] Available: http://ita.ee.lbl.gov/html/traces.html.*

[21] Webalizer: a free logfile analyzer. *[Online] Available: http://www.webalizer.org/.*

[22] The Webalizer: free web server log file analysis program. *[Online] Available: http://www.webalizer.org/download.html.*

[23] E. Anderson. Capture, conversion, and analysis of an intense NFS workload. In *FAST '09*, pages 139–152, 2009.

[24] E. Anderson, M. Arlitt, C. B. Morrey, III, and A. Veitch. DataSeries: an efficient, flexible data format for structured serial data. *SIGOPS Oper. Syst. Rev.*, 43(1):70–75, 2009.

[25] S. Blagodurov and M. Arlitt. Improving the efficiency of information collection and analysis in widely-used IT applications. *HPL Technical report [Online] Available: http://www.hpl.hp.com/techreports/2010/HPL-2010-164.html.*

[26] B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitzer, and A. Tantawi. An analytical model for multi-tier internet services and its applications. In *SIGMETRICS '05*, pages 291–302, 2005.

[27] T. Wood, L. Cherkasova, K. Ozonat, and P. Shenoy. Profiling and modeling resource usage of virtualized applications. In *Middleware '08*, pages 366–387, 2008.