IMPLEMENTATION OF THE BLAS LEVEL 3 ON THE AP1000: PRELIMINARY REPORT

PETER E. STRAZDINS AND RICHARD P. BRENT

Abstract

The Basic Linear Algebra Subprogram (BLAS) library is widely used in many supercomputing applications, and is used to implement more extensive linear algebra subroutine libraries, such as LINPACK and LAPACK. The use of BLAS aids in the clarity, portability and maintenance of mathematical software. To take advantage of the high degree of parallelism of architectures such as the Fujitsu AP1000, BLAS level 3 routines (matrix-matrix operations) are proposed. These routines are not I/O bound; for $n \times n$ matrices, the number of arithmetic operations is $O(n^3)$, whereas the number of I/O operations is only $O(n^2)$.

We are concerned with implementing BLAS level 3 (BLAS-3) for single precision matrices on the AP1000, with emphasis on obtaining the highest possible performance without significantly sacrificing numerical stability. This paper discusses the techniques used to achieve this goal, together with the underlying issues.

The most important techniques are the use of software pipelining and loop unrolling for writing optimized assembler inner loops for matrix inner and outer products, which are able to operate at more than 90% and 70%, respectively, of the AP1000's theoretical peak performance.

The efficiency of cell communication using wormhole routing on the AP1000, especially the row/column broadcast, enables a sustained performance of 80% to 90% of the theoretical peak for all the BLAS-3 routines. It also means that many variations (using different communication schemes) for matrix multiplication have more or less equivalent performance. However, for future versions of the AP1000, optimizing communication must still be considered.

Techniques for improving the performance for large matrices (partitioning, to improve cache utilization) and for small matrices (minimizing communication) are employed. The latter have been developed for general rectangular AP1000 configurations.

Comments

Only the Abstract is given here. The full paper appeared as [7]. An even more "preliminary" report appeared as [6]. For related work see [1, 3, 5].

References

[1] R. P. Brent, "Parallel algorithms in linear algebra", *Proceedings Second NEC Research Symposium* (Tsukuba, Japan, August 1991), invited paper, to appear. Available as Report TR-CS-91-06, CSL, ANU, August 1991, 17 pp. rpb128.

[2] R. P. Brent (editor), *Proceedings of the Second Fujitsu-ANU CAP Workshop*, Department of Computer Science, Australian National University, November 1991, 254 pp. rpb129.

1991 *Mathematics Subject Classification.* Primary 65Y05; Secondary 65F05, 65Y10.

Key words and phrases. BLAS, basic linear algebra subroutines, parallel computation, distributed memory, local memory, MIMD, Fujitsu AP1000, Gaussian elimination, partial pivoting, matrix multiplication.

rpb131a typeset using $\mathcal{A}\mathcal{M}\mathcal{S}$-LaTeX.

[3] R. P. Brent, "The Linpack benchmark on the AP 1000", *Proc. Frontiers '92* (McLean, Virginia, October 1992), IEEE Press, 1992, 128–135. ISBN 0-8186-2772-7. rpb130.

[4] R. P. Brent and M. Ishii (editors), *Proceedings of the First Fujitsu-ANU CAP Workshop*, Fujitsu Research Laboratories, Kawasaki, Japan, November 1990, 205 pp. rpb123.

[5] R. P. Brent and P. E. Strazdins, "Implementation of the BLAS level 3 and Linpack benchmark on the AP 1000", *Fujitsu Scientific and Technical Journal* 29, 1 (March 1993), 61–70. rpb136.

[6] P. E. Strazdins and R. P. Brent, "Implementing BLAS level 3 on the CAP-II", in [4, paper 12, pp. 1–9]. rpb121.

[7] P. E. Strazdins and R. P. Brent, "The implementation of BLAS level 3 on the AP 1000: Preliminary report", in [2, H1–H17]. rpb131.

Department of Computer Science and Computer Sciences Laboratory, Australian National University, Canberra, ACT 0200

E-mail address: {peter,rpb}@cs.anu.edu.au