Parallel Computing 40 (2014) 559–573

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Review

A survey of power and energy efficient techniques for high performance numerical linear algebra operations

Li Tan, Shashank Kothapalli, Longxiang Chen, Omar Hussaini, Ryan Bissiri, Zizhong Chen *

Department of Computer Science and Engineering, University of California, Riverside, United States

ARTICLE INFO

Article history:
Received 12 December 2013
Received in revised form 2 August 2014
Accepted 4 September 2014
Available online 16 September 2014

Keywords:
Power and energy
Performance
Power management
Supercomputers
Numerical linear algebra
DVFS

ABSTRACT

Extreme scale supercomputers available before the end of this decade are expected to have 100 million to 1 billion computing cores. The power and energy efficiency issue has become one of the primary concerns of extreme scale high performance scientific computing. This paper surveys the research on saving power and energy for numerical linear algebra algorithms in high performance scientific computing on supercomputers around the world. We first stress the significance of numerical linear algebra algorithms in high performance scientific computing nowadays, followed by a background introduction on widely used numerical linear algebra algorithms and software libraries and benchmarks. We summarize commonly deployed power management techniques for reducing power and energy consumption in high performance computing systems by presenting power and energy models and two fundamental types of power management techniques: static and dynamic. Further, we review the research on saving power and energy for high performance numerical linear algebra algorithms from four aspects: profiling, trading off performance, static saving, and dynamic saving, and summarize state-of-the-art techniques for achieving power and energy efficiency in each category individually. Finally, we discuss potential directions of future work and summarize the paper.

Contents

* Corresponding author.
E-mail address: chen@cs.ucr.edu (Z. Chen).

http://dx.doi.org/10.1016/j.parco.2014.09.001
0167-8191/

1. Introduction

In this era of pervasive high performance computing, supercomputers are of ever-growing computation capability and network bandwidth. Given the rapidly climbing power bills, achieving power and energy efficiency of supercomputers has become a prime concern and also been considered as a challenging issue. As a fundamental integrant of high performance computing, numerical linear algebra algorithms including Cholesky, LU, and QR factorizations, serve as backbone for most scientific algorithms. Generally numerical linear algebra algorithms have been widely employed for solving a system of linear equations in high performance scientific applications running on supercomputers around the world, ranked by the TOP500 list [1]. To name a few, for the purpose of benchmarking, HPL [2] is a portable high performance software package that solves a dense linear system via LU factorization on distributed-memory architectures; NPB [3] is a set of benchmarks for evaluating performance of supercomputers, where highly parallel implementations of numerical linear algebra algorithms such as conjugate gradient method and LU factorization are included. For the purpose of scientific computing, ScaLAPACK [4] and DPLASMA [5] are two extensively used high performance and scalable numerical linear algebra software libraries for distributed-memory multicore systems, where routines of Cholesky, LU, and QR factorizations are provided as standard functionality. Moreover, with regard to software products, MATLAB [6] is a commercialized computing software developed by MathWorks for performing numerical calculations, where matrix factorizations are implemented in terms of easy-to-use user commands. MKL [7] is a commercialized software library developed by Intel for optimized linear algebra routines for scientific computing, including highly tuned routines of matrix factorizations, sparse solvers, and fast Fourier transforms. The open source libraries include ATLAS [8], HPL [2], etc, which are extensively used for solving and benchmarking. Numerical linear algebra algorithms are also generally adopted in many other areas of high performance scientific computing, including computer graphics, quantum mechanics, game theory, and economics.

With the growing severity of power and energy consumption on high performance computing systems nowadays in terms of operating costs and system reliability [9,10], in particular given the fact that power supply for large-scale data centers and clusters are usually limited within a power cap, reducing power and energy costs has been deemed as a critical issue in high performance computing. The Green500 list [11], ranks the top 500 supercomputers around the world by energy efficiency in a six-month cycle, which indicates the trend that supercomputers with a high ratio of performance-power (FLOPS per Watt) are favored nowadays. Motivated by the pressing and ever-growing demands of power and energy issues nowadays, people have proposed numerous solutions to maximize the use of the deployed power capacity of data centers [12,13]. As general purpose fundamental operations required for solving a system of linear equations, numerical linear algebra algorithms are extensively adopted in a large body of high performance scientific applications for different purposes as presented above. Consequently attempts of decreasing power and energy costs of running numerical linear algebra algorithms is beneficial to lowering energy consumption of executions of the applications that employ these algorithms ultimately.

Various holistic hardware and software solutions have been proposed for mitigating power and energy costs of running such applications on supercomputers around the world, and substantial power and energy savings have been achieved for different types of architectures. In this paper, we survey common power management strategies for high performance computing systems, with a focus on state-of-the-art techniques for achieving power and energy efficiency of high performance scientific applications where numerical linear algebra algorithms are widely employed.

The remainder of the paper is organized as follows. Section 2 introduces high performance numerical linear algebra and related software libraries and benchmarks. We present commonly used power management strategies for saving power and energy for high performance computing systems in general in Section 3, and provide details of state-of-the-art techniques for power and energy efficient numerical linear algebra in Section 4. Section 5 summarizes the paper.