Future Generation Computer Systems 37 (2014) 60–63

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Editorial

Innovative methods and algorithms for advanced data-intensive computing

Advanced data-intensive computing represents an active area of research that spans across a significant number of research topics ranging from traditional *parallel* and *distributed* computing to recent *Grid* and *Cloud computing*. All these high-performance paradigms share a common emphasis that focuses on the issue of *effectively and efficiently representing, managing and distributing large-size and large-scale data* that populate their internal layers. This conveys to the well-known term "*data-intensive computing*", which represents an emerging challenge in next-generation computing systems (e.g., [13,12]).

Data-intensive computing has recently been of great interest for the research community, mainly driven by modern research initiatives such as *big data management*, *analytics over large-scale data*, *very-large scientific data management*, *social network data management*, and so forth. There exists a wide range of application scenarios where data-intensive computing is relevant: scientific data management, bio-medical data management, sensor and stream data management, environmental data management, and so forth.

Managing these kinds of data poses critical and still-unsolved issues, mostly represented by the enormous size of data and the exponential scaling-up of data over growing-in-size inputs and requirements (e.g., [10,8]). A reliable solution to these issues comes from the usage of advanced computational paradigms, like *MapReduce*, and infrastructures, like *Clouds*. A necessary step towards the successfully achievement of this goal is represented by the need for *innovative methods and algorithms for advanced data-intensive computing*, as classical proposals appeared in traditional areas like parallel and distributed computing are clearly inadequate to cope with the requirements dictated by modern data-intensive scenarios.

With these goals in mind, this special issue on "*Innovative Methods and Algorithms for Advanced Data-Intensive Computing*" of *Future Generation Computer Systems* focuses on theoretical as well as practical aspects of innovative models and algorithms for advanced data-intensive computing on high-performance computational infrastructures like Grids and Clouds. The special issue contains twelve papers, which have gone through several rigorous review rounds before being accepted for final inclusion. Some of the contributions have been invited for submission as the best papers of the *11th LNCS International Conference on Algorithms and Architectures for Parallel Processing* (ICA3PP 2011), held in Melbourne, Australia, during October 24–26, 2011, leaded by the Editor. Papers of this special issue are briefly described next.

The first paper, titled "*Architectural Investigation of Matrix Data Layout on Multicore Processors*" [7], by Minwoo Kima and Won Woo Roa, recognizes that many practical applications include *matrix operations as essential procedures*. As a consequence, recent studies of matrix operations rely on *parallel processing* to reduce any calculation delays. Nevertheless, since these operations are highly data-intensive, several studies have investigated issues of *work distribution* and *data access latency* in order to accelerate algorithms. Despite this, previous studies have not considered hardware architectural features adequately, although they greatly affect the performance of matrix operations. In order to fulfill this gap, the research proposed in this paper considers *architectural characteristics that affect the performance of matrix operations on real multicore processors*. As testing, authors make use of *matrix multiplication*, *LU decomposition*, and *Cholesky factorization*, which are well-known data-intensive mathematical algorithms in various fields. Authors argue that applications only access matrices in a particular direction, and propose that *the canonical data layout is the optimal matrix data layout compared with the block data layout*. In addition, *a suitable tiling algorithm* is exploited to increase the temporal data locality in multi-level caches and to balance the workload as evenly as possible in multicore environments. Experimental results show that applications using the canonical data layout with tiling have relevantly-higher performance compared with applications executed with the block data layout.

The second paper, titled "*Exploiting Fine-Grain Parallelism in the H.264 Deblocking Filter by Operation Reordering*" [14], by Tsung-Hsi Weng and Chung-Ping Chung, focuses the attention on the specific application scenario represented by the *H.264 video compression standard*, where, as authors correctly recognize, *the deblocking filtering contributes about one-third of all the computation in the decoder*. With *many-core architectures* becoming the future trend of system design, authors argue that computation time can be reduced if the *deblocking appropriately distributes its operations to multiple processing elements*. With this idea in mind, authors propose using *a 4-pixel-long boundary as the basis for analyzing and exploiting possible parallelism*. Compared with the two-dimensional wave-front method order for deblocking both 1920×1080- and 1080×1920-pixel frames, the proposed design exhibits speedups by 1.92 and 2.44 times, respectively, given an unlimited number of processing elements. Compared with another previous proposal by the same authors, it gains speedups by 1.25 and 1.13 times,

http://dx.doi.org/10.1016/j.future.2014.04.004
0167-739X/

respectively. In addition, as the frame size grows, the proposed approach requires only extra time that is proportional to the square root of the frame size increase (keeping the same width to height ratio), pushing the boundary of practical real-time deblocking of increasingly larger video sizes. Authors complement their analytical contribution by means of several experiments that confirm the benefits of their proposed deblocking filtering strategy.

The third paper, titled "*Fault-tolerant Routing based on Approximate Directed Routable Probabilities for Hypercubes*" [4], by Dinh Thuy Duong and Keiichi Kaneko, proposes a fundamental study and related research contribution on so-called *hypercubes*, which are among the most popular *topologies for interconnection networks* that constitute the basis of parallel processing systems (of variegate topologies). In this investigated scientific context, authors propose *two new fault-tolerant routing algorithms for hypercubes based on approximate directed routable probabilities*. In the delineated modeling, probabilities represent ability of routing toward any node located at a specific distance and are calculated by considering from which direction the message has been received. Also, each node chooses one of its neighbor nodes to forward the message by comparing the approximate directed routable probabilities. The effectiveness of proposed algorithms is finally assessed via an extensive corpus of experiments.

The fourth paper, titled "*Courier: Multi-Dimensional QoS Guarantees for Consolidated Storage System*" [16], by Quan Zhang, Dan Feng and Fang Wang, introduces the novel concept of *multi-dimensional Quality of Servic*e (QoS), a derivation of the standard QoS concept particularly focused to *storage functionalities* which pursues the idea of *data centers that consolidate resources and provide storage services for multiplex applications*, where cost and complexity of data management are increasing more and more. However, satisfying performance targets for each workload is challenging, because of I/O characteristics of workloads usually varies widely and capability of storage systems changes significantly. In order to cope with this challenge, authors propose *a novel QoS scheduler, called Courier, whose main goal is to maintain satisfactory performance for applications even in the highly-volatile scenario described above*. Courier dynamically alternates between a *feedback-based latency controller* and a *reward budget-based scheduling* in order to achieve per-application performance requirement. The feedback-based controller is employed to estimate request service times and adjust the scheduling strategy dynamically. Based on the estimation, Courier can identify time-critical requests from throughput-sensitive requests, and schedule applications with time-critical requests preferentially to avoid latency violations. In addition to this, Courier rewards well-behavior application with more budget as to maintain high storage utilization while providing performance guarantees. In order to complement their main contribution, authors evaluate the effectiveness of the proposed approach using synthetic and real-life workloads, and the results show that Courier has good ability to achieve per-application performance targets.

The fifth paper, titled "*Secure and Energy-Efficient Data Aggregation with Malicious Aggregator Identification in Wireless Sensor Networks*" [9], by Hongjuan Li, Keqiu Li, Wenyu Qu and Ivan Stojmenovic, explores *the problem of supporting data aggregation in Wireless Sensor Networks with security and energy-efficiency features*. The final goal of issues above consists in reducing the communication overhead and prolonging the network lifetime. However, authors argue that an *adversary* may compromise some sensor nodes, and use them to forge false values as the aggregation result. Previous secure data aggregation schemes have tackled this problem from different angles. The goal of those algorithms is to ensure that the *Base Station* (BS) does not accept any forged aggregation results. But none of them have tried to detect the nodes that inject into the network bogus aggregation results. Moreover, most of them usually have a communication overhead that is (at best) logarithmic per node. Starting from this evidence, the paper proposes *a secure and energy-efficient data aggregation scheme that can detect the malicious nodes with a constant per node communication overhead*. In the proposed solution, all aggregation results are signed with the *private keys* of *aggregators*, so that they cannot be altered by others. Nodes on each link additionally use their pairwise shared key for *secure communications*. Each node receives the aggregation results from its parent (sent by the parent of its parent) and its siblings (via its parent node), and verifies the aggregation result of the parent node. As a proof of the quality of the proposed research, authors finally provide theoretical analysis on energy consumption and communication overhead, plus a comparison-based simulation study over random data aggregation trees.

The sixth paper, titled "*Mining Constrained Frequent Itemsets from Distributed Uncertain Data*" [2], by Alfredo Cuzzocrea, Carson Kai-Sang Leung and Richard Kyle MacKinnon, considers the application scenario represented by *large amounts of streaming data generated from various sources*, such as sensor data from environmental surveillance networks. In this context, a relevant problem is represented by *the issue of supporting distributed frequent itemset mining*, which is, indeed, relevant for a plethora of applications. Despite this, many existing distributed frequent itemset mining algorithms do not allow users to express the itemsets to be mined according to their intention via the use of *constraints*. As a consequence, these unconstrained mining algorithms can yield numerous itemsets that are not interesting to users. Moreover, due to inherited measurement inaccuracies and/or network latencies, the data are often riddled with *uncertainty*. At the convergence of both challenges (i.e., constraints, and uncertainty), *a strident call for both constrained mining and uncertain data mining arises*, with also relevant use cases (e.g., wireless sensor networks of uncertain data). In order to tackle this main challenge, authors propose *a tree-based system for mining frequent itemsets that satisfy user-defined constraints from distributed environments with uncertainty*. The analytical contribution is complemented by means of a comprehensive experimental evaluation and analysis whose results clearly support authors' claims.

The seventh paper, titled "*Mining Network Data for Intrusion Detection through Combining SVM with Ant Colony*" [5], by Wenying Feng, Qinglei Zhang, Gongzhu Hu and Jimmy Xiangji Huang, introduces *Combining Support Vectors with Ant Colony* (CSVAC), a *new machine-learning-based data classification algorithm that is applied to network intrusion detection*. The basic task of this proposal consists in classifying network activities as *normal* or *abnormal* by inspecting the network log modelled as connection records, *while minimizing misclassification*. As authors correctly recognize, although different classification models have been developed for network intrusion detection, each of them is characterized by strengths and weaknesses, including the most commonly-applied *Support Vector Machine* (SVM) method and the *Clustering based on Self-Organized Ant Colony Network* (CSOACN). CSVAC combines SVM with CSOACN in order to take the advantages of both while avoiding their weaknesses. The proposed algorithm is evaluated using a standard benchmark KDD99 data set, and deriving experiments show that it outperforms SVM alone or CSOACN alone in terms of both classification rate and run-time efficiency.

The eighth paper, titled "*A Green Energy-Efficient Scheduling Algorithm Using DVFS Technique for Cloud Datacenters*" [15], by Chia-Ming Wu, Ruay-Shiung Chang and Hsin-Yu Chan, starts from the evidence that *Information and Communication Technology* (ICT) profoundly impacts on environment because of its large amount of CO_2 emissions. As a consequence, during the past years, the research field of "green" and low power consumption networking infrastructures has been of great importance for