Cluster Comput
DOI 10.1007/s10586-011-0171-x

An overview of energy efficiency techniques in cluster computing systems

Giorgio Luigi Valentini · Walter Lassonde · Samee Ullah Khan · Nasro Min-Allah · Sajjad A. Madani · Juan Li · Limin Zhang · Lizhe Wang · Nasir Ghani · Joanna Kolodziej · Hongxiang Li · Albert Y. Zomaya · Cheng-Zhong Xu · Pavan Balaji · Abhinav Vishnu · Fredric Pinel · Johnatan E. Pecero · Dzmitry Kliazovich · Pascal Bouvry

Received: 19 July 2011 / Accepted: 22 July 2011

Abstract Two major constraints demand more consideration for energy efficiency in cluster computing: (a) operational costs, and (b) system reliability. Increasing energy efficiency in cluster systems will reduce energy consumption, excess heat, lower operational costs, and improve system reliability. Based on the energy-power relationship, and the fact that energy consumption can be reduced with strategic power management, we focus in this survey on the characteristic of two main power management technologies: (a) static power management (SPM) systems that utilize low-power components to save the energy, and (b) dynamic power management (DPM) systems that utilize software and power-scalable components to optimize the energy consumption. We present the current state of the art in both of the SPM and DPM techniques, citing representative

G.L. Valentini · W. Lassonde · S.U. Khan (✉) · J. Li · L. Zhang
NDSU-CIIT Green Computing and Communications Laboratory, Department of Electrical and Computer Engineering, North Dakota State University, Fargo, ND 58108-6050, USA
e-mail: samee.khan@ndsu.edu

G.L. Valentini
e-mail: giorgio.valentini@ndsu.edu

W. Lassonde
e-mail: walter.lassonde@ndsu.edu

J. Li
e-mail: juan.li@ndsu.edu

L. Zhang
e-mail: limin.zhang@ndsu.edu

N. Min-Allah · S.A. Madani
COMSATS Institute of Information Technology, Islamabad, Pakistan

N. Min-Allah
e-mail: nasar@comsats.edu.pk

S.A. Madani
e-mail: madani@ciit.net.pk

L. Wang
Indiana University, Bloomington, IN, USA
e-mail: wanglizh@indiana.edu

N. Ghani
University of New Mexico, Albuquerque, NM, USA
e-mail: nghani@ece.unm.edu

J. Kolodziej
University of Bielsko-Biala, 43300 Bielsko-Biala, Poland
e-mail: jkolodziej@ath.bielsko.pl

H. Li
University of Louisville, Louisville, KY, USA
e-mail: h.li@louisville.edu

A.Y. Zomaya
University of Sydney, Sydney, NSW 2006, Australia
e-mail: albert.zomaya@sydney.edu.au

C.-Z. Xu
Wayne State University, Detroit, MI, USA
e-mail: czxu@wayne.edu

P. Balaji
Argonne National Laboratory, Argonne, IL, USA
e-mail: balaji@mcs.anl.gov

A. Vishnu
Pacific Northwest National Laboratory, Richland, WA, USA
e-mail: abhinav.vishnu@pnl.gov

G.L. Valentini · F. Pinel · J.E. Pecero · D. Kliazovich · P. Bouvry
University of Luxembourg, Luxembourg L1359, Luxembourg

F. Pinel
e-mail: fredric.pinel@uni.lu

J.E. Pecero
e-mail: johnatan.pecero@uni.lu

Published online: 10 September 2011

examples. The survey is concluded with a brief discussion and some assumptions about the possible future directions that could be explored to improve the energy efficiency in cluster computing.

Keywords Cluster computing · Energy efficiency · Power management · Survey

Acronyms

CMOS	Complementary Metal-oxide-Semiconductor
CPU	Central Processing Unit
CPU MISER	CPU Management Infra-Structure for Energy Reduction
DFS	Dynamic Frequency Scaling
DPM	Dynamic Power Management
DVS	Dynamic Voltage Scaling
DVFS	Dynamic Voltage and Frequency Scaling
FAWN	Fast Array of Wimpy Nodes
GCA	Grand Challenge Applications
HA	High Availability
HPC	High-Performance Computing
IT	Information Technology
LB	Load Balancing
Memory MISER	Memory Management Infra-Structure for Energy Reduction
NASA	National Aeronautics and Space Administration
NAS	NASA Advanced Supercomputing
NPB	NAS Division Parallel Benchmarks
PART system	Power-aware Run-time System
PID controller	Proportional-Integral-Derivative controller
PSC	Power-Scalable Components
SDRAM	Synchronous Dynamic Random Access Memory
SPM	Static Power Management
VLAN	Virtual Local-area Network

1 Introduction and motivation

Over the last two decades, the increases of the demand for computing performance and the energy-efficiency have not kept up. Since 1992, the performance of supercomputers has grown 10,000-fold against the 300-fold of the performance per Watt [10]. The need for computing power capable of solving the grand challenge applications (GCAs) of modern scientific research continues to push the development of technology with ever-higher levels of performance. Nowadays, cluster systems are a competitive alternative to traditional supercomputing systems in high-performance computing (HPC) applications [4, 17, 33], offering the same yield/performance levels at lower cost [37].

The advantage of a cluster system lies in the ability on handling large and extremely complex computations on more than one computer, working on the same problem or part thereof, simultaneously [46]. A cluster system consists of a group of independent computers linked together by high-speed networks. The complexities of the underlying system are hided from the user through the middleware. As a result, the user only perceives a single system instead of the all architecture [5].

There are three main paradigms in distributed computing system: (a) cluster computing, (b) grid computing, and (c) cloud computing.

Cluster computing can be described as the integration of more than one off-the-shelf commodity computer and resources incorporated through hardware, networks, and software to create a single system image. In traditional approaches the terms HPC and cluster computing referred to the same type of computing environment. Today, the definition of cluster computing has been extended beyond the definition of parallel computing to include high availability (HA) clusters and load-balancing (LB) clusters.

HA clusters (also referred as failover clusters in the literature) are primarily implemented with the purpose of improving the availability of the provided services. HA clusters operate having redundant nodes, used to provide service when system components fail. Because the minimum requirement for redundancy relies on at least two elements, HA clusters are more commonly composed by two nodes. Implementing the redundancy of cluster components, HA clusters attempt to eliminate single points of failure.

LB clusters are multiple computers linked together to share computational workload or function, behaving as a single virtual computer. Requests initiated from the user are managed by, and distributed among, all the standalone computers forming the cluster. The result is a balanced computational work among different machines, directly improving the performance of the cluster systems.

More generally, in a cluster computing system each computer is referred to as a node. Each node may have different features as single processor or multiprocessor architecture. Usually each node in the cluster is limited to a single switch or collection of switches operating at Layer 2 and within one virtual local-area network (VLAN). A cluster computing network is a dedicated network.

Grid computing is defined as system of computer resources from multiple administrative domains working together to reach a common goal. As a main difference, grids

D. Kliazovich
e-mail: dzmitry.kliazovich@uni.lu

P. Bouvry
e-mail: pascal.bouvry@uni.lu

computing are more loosely coupled, heterogeneous, and geographically dispersed than cluster computing systems.

Cloud computing is a new internet-based supplement, consumption, and delivery model to provide real-time, on demand, self-provisioned IT services to business users. The main difference from cluster computing is that clouds often take the form of web-based tools and applications, normally accessed through an internet browser.

Grid computing and cloud computing have as a common factor that both of the systems can embed cluster computing system.

While thermal and energy management are the main issues of grid computing system due to aggregation of computing, networking, and storage hardware, the energy consumption required to transport the data from and to the user constitute the major issue of cloud computing system.

Resulting from excessive power consumption and increased component density, large amounts of heat, along with greater than ever demand for electricity, point to two major areas of concern in cluster computing: (a) operational costs and (b) system reliability [15, 27, 42]. With rising energy costs, operational costs increase, making energy efficiency more lucrative than ever. Energy-awareness can improve reliability of cluster systems by decreasing the amount of heat in the system. Computing at higher temperatures is more error-prone than computing at moderate temperatures, in fact, component failure rates double with every 18°F (or 10°C) increase in temperature, according to the Arrhenius' equation [6, 15, 20, 27], defined by the following formula:

$$k = A * e\frac{-E_a}{(R + T_a)}, \tag{1}$$

where k is the rate constant, A is the pre-exponential factor, E_a is the activation energy, R is the gas constant, and T_a is the absolute temperature.

The main aim of our survey is to provide an overview of the recent research results in energy-efficient cluster computing. Energy efficiency in a cluster system can be enhanced at three different levels [2]: (a) energy-efficient applications, (b) power-aware resource management, and (c) efficiency of hardware. All these three levels must be addressed to develop a green IT cluster system. The total amount of energy (E) consumed in a cluster system in a time interval (T) is defined as a product of the time (T) and the average system power (P) consumed in the interval (T), i.e.:

$$E = P * T. \tag{2}$$

The energy consumption can be reduced if either the average power consumption or the time intervals T are reduced. For example, minimizing the time interval (T) the energy consumption (E) is limited. That is, below a certain (minimum) value of T the result on E cannot be further improved. Usually, the limitation of E results from the mapping of the applications to individual cluster system architectures (e.g. as a result of scalability or system bottlenecks). Therefore, in terms of energy efficiency, the power management approaches becomes increasingly important.

In Fig. 1 we present a simple classification of the energy-aware cluster computing systems. There are two main categories of the power management in cluster computing: (a) Static Power Management (SPM), and (b) Dynamic Power Management (DPM) [2]. SPM technologies use low-power energy-efficient hardware equipment (e.g. CPUs and power supplies) to reduce energy usage and peak power consumption. DPM techniques are based on the knowledge of current resource utilization and application workloads to reduce energy usage. It follows from (2), that there is no guarantee that DPM techniques, although they may improve the system efficiency, will (always) decrease the peak power consumption. Specifically, if the time interval (T) and the energy consumption (E) are reduced although the power consumed (P) is increased, the energy-efficiency is improved even though the possible increase of the peak power consumption.

The remainder of our paper is as follows. In Sect. 2, we explore the SPM methods that have been developed. In Sect. 3, we discuss the various DPM techniques. We conclude our work in Sect. 4.

2 Static power management

The main aim of power management approaches to energy optimization in computational clusters is to create the cluster systems by using low-power components and keeping the system at the acceptable level of performance. In the last decades the low-power technologies were effective for mobile and handheld systems [14]. Recently they are successfully applied in cluster computing. In most of the cluster systems, CPUs may consume 35–50% of a cluster nodes' total power [9, 29], which makes them the most energy-absorbing components of the system. The other such energy expensive components are memory modules [40]. Low-power memory and CPU components of the cluster system can effectively support the energy-aware (static power) management.

Green Destiny [45] and IBM Blue Gene/L [3], Blue Gene/P [23], and Blue Gene/Q [24] are the most popular examples of HPC machines, which are composed of the low-power modules, using less energy than traditional supercomputers. The Green500 [39] provides a list of the most energy-efficient supercomputers in the world, where Blue Gene/Q is currently ranked as the most energy-efficient HPC [38]. However, beside the awards, many cluster workloads do not scale as the number of cluster nodes increases [19], and the low-power (modules) technology is based on commodity

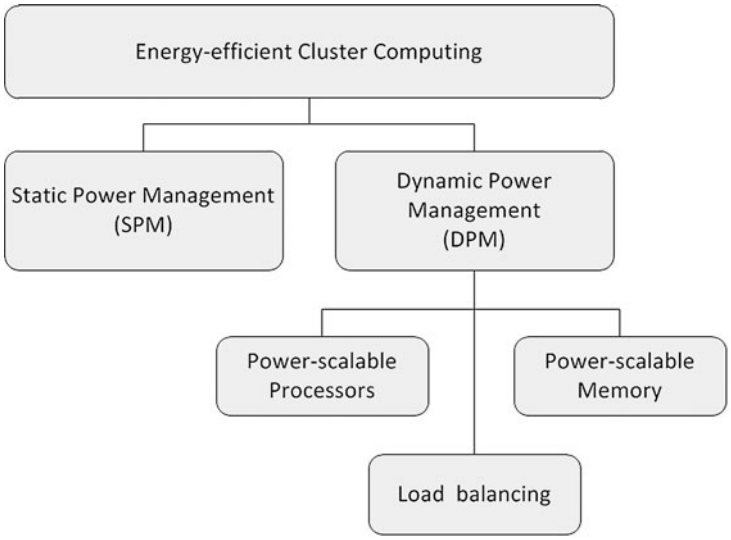


Fig. 1 Power management of cluster system

parts as non-commodity parts, which tend to be expensive [15, 19].

In recent years, research on low-power components in cluster systems has continued to attract scientific investigators interested in energy-efficient cluster computing. Anderson et al. in [1] define fast array of wimpy nodes (FAWN) system as a novel cluster architecture for low-power data intensive computing. FAWN combines low-power CPUs with small amounts of local flash storage, and balances computation and I/O capabilities in order to provide efficient parallel data access on a large-scale. In [43] FAWN has been experimentally evaluated on various workloads. The results suggest that overall, lower frequency nodes are more efficient than conventional high-performance CPUs. The FAWN architecture seems to be unfeasible to solving problems that cannot be parallelized or whose working set size cannot be further divided to be assigned into the available memory of the smaller nodes. The authors claim that a promising path in today's energy-efficient computing should tend us for the acceptation of: "*tight constraints on per-node performance, cache, and memory capacity, together with using algorithms that scale to an order of magnitude more processing elements. While many data-intensive workloads may fit this model nearly out-of-the-box, others may require substantial algorithmic and implementation changes.*"

Caulfield et al. define in [7] "Gordon" architecture, which is an example of a low-power data-centric system. By utilizing low-power processors, flash memory, and data-centric programming systems, Gordon reduces power consumption and improves performance specifically for data-centric applications. Results of the experiments presented in [7] suggest that Gordon may be able to out-perform disk-based clusters by 1.5 to 2.5 times more performance per Watt. We summarize in Table 1, the properties of the presented SPM techniques under the energy savings criterions and highlight briefly their limitations.

Table 1 Summary of SPM techniques

Ref.	Technique	Energy Savings	Limitations
[1, 43]	Low-power CPUs with small amounts of local flash storage (FAWN)	FAWN clusters can perform about 350 key-value queries per joule (two orders of magnitude more than a disk-based system)	May be an effective solution in node-scalable systems and rather ineffective (high energy consumption) in non-commodity parts
[7]	Low-power CPUs, flash memory, and data-centric programming systems (Gordon)	Gordon clusters are 1.5× faster and deliver 2.5× more performance per watt, when compared to disk-based systems	May be an effective solution in node-scalable systems and rather ineffective (high energy consumption) in non-commodity

3 Dynamic power management

DPM techniques include all the methods that facilitate the run-time adaptation of a cluster system according to current resource requirements and other dynamic characteristics of the cluster system states [2]. As illustrated in Fig. 1, there are two main approaches to DPM: (a) using the software and power-scalable components in dynamical adjustment of the power consumption in the cluster system [2, 8, 11–16, 18, 20–22, 26, 28–30, 32, 36, 40, 41, 44] and (b) load balancing techniques [34, 35]. In this section we briefly survey the DPM techniques, starting from software and power-scalable components, and present the main idea of load balancing.

3.1 Software and power-scalable components

Power-scalable components of the cluster system, such as high-performance dynamic voltage scaling (DVS) modules, allow to modulate the voltage supply of CPUs [25, 31], which can reduce the energy consumption in the case of decreasing the operating power. Power management is becoming more common in cluster hardware components, such as disk drives, memory banks, and network cards [14]. Cluster systems that utilize power-scalable components are called power-aware clusters [15]. Currently, the main researches on power-aware technology are focused on CPUs and memory banks due to the large share of the consumed system power. We will divide our subsequent discussion of power-aware clusters according to the two dominant power-scalable components: (a) memory, and (b) processors.

3.1.1 Power-scalable memory

Memory Management Infra-Structure for Energy Reduction (Memory MISER) developed by Tolentino et al. in [40] is an efficient solution for dynamic power-scalable memory management in cluster systems. Memory MISER utilizes a modified Linux kernel and a daemon implementation of a PID controller to on-line and off-line memory scaling in the system operating mode. Memory MISER was experimentally tested on a server with 8 processors, and 32 GB of SDRAM per processor, on parallel and sequential applications. The results of the experiments presented in [40] show that Memory MISER may reach 70% reduction in memory energy consumption and 30% reduction in total system energy consumption with a performance degradation of less than 1%. The achievements of Memory MISER in terms of energy efficiency on cluster systems warrant further investigation on power-scalable memory management.

3.1.2 Power-scalable processors

Power-scalable processors become one of the most promising energy-efficient off-the-shelf technology for modern cluster systems. The dynamic power of a CMOS circuit can be reduced by implementing the following voltage and frequency scaling modules: (a) dynamic voltage scaling module (DVS) [11, 14, 15, 18, 36], (b) dynamic frequency scaling module (DFS) [13], and (c) dynamic voltage and frequency scaling module (DVFS) [8, 12, 16, 20–22, 26, 28–30, 32, 44]. The total power of a processor (P_t) can be expressed as a sum of a dynamic power (P_d) and the static/leakage power (P_s), i.e.:

$$P_t = P_d + P_s. \tag{3}$$

The dynamic power consumption of a CMOS-based processor is proportional to the percentage of active gates (A), clock frequency (f), total capacitive load (C), and voltage (V) squared [15].

$$P_d \approx A * f * C * V^2. \tag{4}$$

The static power consumption is a result of energy leakage and it is calculated even if the CPU is idle [8]. DVFS techniques allow to change the voltage and frequency supply at the cluster node's to satisfy the computational requirements specified for the applications [41]. From the experiments, traditional clusters may only achieve 5–10% of the peak performance while executing scientific applications [14]. The low performance comes from the (cluster) workload that significantly differs among the executed applications.

Following the terminology and notation introduced in [12], each voltage and frequency level can be defined as a "gear", in the literature sometimes replaced with the term "p-state" or "power mode" [15, 30]. During slack periods (low operations power mode in certain computations, communications, or memory bottlenecks), significant savings can be achieved by reducing processor supply voltage and frequency (i.e. using DVFS module) [14, 30].

To exploit processor slack periods, various DVS, DFS, and DVFS scheduling strategies are used. In [22], S. Huang and W. Feng divided DVS, DFS, and DVFS scheduling strategies into two categories: (a) off-line trace-based schedulers, and (b) on-line/run-time profiling-based schedulers. Off-line trace-based scheduling is a requirement of a prior knowledge of application workload. However, once application workloads are known, the application can be decompose into phases and an appropriate gear selected for each node for the execution of each phase. The resolution of gear selection is referred to as the granularity. As a result, a fine-grained scheduler has more gear options than a coarse-grained scheduler. The granularity and the amount of time spent within and between gears determines the cost of the system performance (i.e. delay) and possible energy saving rate. Developing energy-aware on-line/run-time schedulers remain still a challenging task, mainly because of the requirements of the high precision in the prediction of the effects of gears in the following phases of the application without any prior characteristics of its computational complexity [22]. The main benefit of using the run-time schedulers is their ability to change gears during the execution of an application that has not been yet compiled. In the following paragraphs we highlight the state-of-the-art power-aware scheduling research.

Ge et al. propose in [14] an off-line DVFS scheduling algorithm that utilizes a weighted energy-delay product to improve the cluster energy efficiency. The weighed approach of the energy-delay product is user-driven. That is, through the weighting factors the priority can be given to the energy saving or the performance, one at the expense of the other. The

algorithms were tested on a 16-node Centrino-based cluster. The results of the experiments show that energy savings may reach 30% in average with less than 5% performance reduction. Off-line scheduling is effective if cluster planners have access to applications before their execution in the system. The authors also show in [14] that off-line techniques may be a good reference methodology in a comprehensive experiment analysis of run-time techniques.

In [16] Ge et al. present CPU MISER, a run-time DVFS scheduling system. CPU MISER is capable of providing fine-grained, performance-directed DVFS power management for a generic power-aware cluster. CPU MISER allows the limits of acceptable performance loss to be defined by the user. When tested on the NASA Advanced Supercomputing (NAS) Division Parallel Benchmarks (NPB), CPU MISER demonstrated that up to 20% energy savings was possible with a corresponding 4% performance loss.

Huang et al. present in [22] a run-time DVFS scheduling algorithm *eco*, and the associate system implementation *ecod*. According to [22], "*ecod manages application performance and power consumption in real time based on an accurate measurement of CPU stall cycles due to off-chip activities.*" The ecod system was tested on a cluster system and the results showed a 6% reduction in performance loss and a 3% increase in energy savings.

The authors in [20] propose a run-time DVFS scheduling algorithm, called the β-algorithm, which is capable of transparently and automatically reducing power consumption while maintaining a specified level of performance. Hsu et al. in [20] refer to their implementation of state-of-the-art techniques as the power-aware run-time (PART) system. When PART was tested on the NPB, cluster system energy reduction was as high as 25%, while performance degradation was tightly maintained between 3–5%.

Experiments with the DVFS of processors in parallel sparse applications have been conducted in [8]. Chen et al recorded a cluster system's energy savings and performance penalty for various voltage and frequency settings. The results showed that the rate of change of energy savings to allowable performance penalty is the highest (i.e. more energy is saved per amount of performance sacrificed) at low allowable performance penalties and tends to decrease with increasing allowable performance penalty until the energy savings converge. The reason for energy savings convergence is found to be a result of a DVFS processor technology limitation. Processor voltage has a minimum value below which it will not operate correctly.

Using software algorithms and power-scalable processors in DVFS implementation seems to be an effective solution for improving energy efficiency in cluster computing [14, 16, 20, 22]. However, one of the drawbacks of the DVS approach to power management may be the complexity of the run-time DVS scheduling implementations, and the inherent limitation imposed by the minimum processor voltage that guarantee the correctness of the operations performed. Finally, the selection of the most appropriate gear result complicated because there is an energy cost for the gear transition and the future application workload is unknown [18].

3.2 Load balancing

The main aim of load balancing (LB) methodology [34, 35] is to distribute the workload across the computing cluster to achieve optimal resource utilization, minimize the response time, and avoid overload of the system. As the result some nodes in the system can be switched to the stand-by mode or just switched off. Although the energy can be saved at low-power mode or inactive nodes, the overall system performance can be adversely impacted, which may be a reason of increasing the system energy utilization. Therefore LB methodologies can be characterized as a tradeoff between power supply and system performance.

Pinheiro et al. in [35] propose a dynamical cluster operational mode controller as develop an approach to save energy in cluster systems by dynamically turning cluster nodes on and off in a way that efficiently matches load demand. The method was implemented in two popular types of cluster-based systems: a network server and an operating system (OS) for clustered cycle servers. The results show that the technique can save energy by taking advantage of periods of light load in cluster-based systems. Load balancing finds limited application because light loads are the exception rather than the rule. Cluster planners try to schedule applications that take advantage of the cluster workload capacity, limiting the light loads.

In Table 2 we compare the DPM techniques from the literature under the criterion of their effectiveness in the reduction of energy consumption and highlight their main drawbacks and limitations.

4 Discussion

High operational costs and reduced cluster system reliability resulting from excessive heat are the major barriers to sustainable growth in computing power. The problem of energy efficiency in cluster computing remains challenging mainly because of the variety of applications that need to be processed on cluster systems and a continued demand for high performance. The main methods for increasing energy-efficiency in cluster computing are: (a) the SPM technique of using low power embedded CPUs coupled with flash storage and (b) the DPM technique of using software and power-scalable components to dynamically adjust cluster power consumption, especially in the form

Table 2 Summary of an experimental evaluation of various DPM techniques

Ref.	Approach	Technique	Energy Savings			Limitations
			Research Centre	Energy Improvement	Performance Drop	
[8]	PSC	DVFS control in parallel sparse applications using two different scheduling algorithms	N/A	N/A	N/A	Minimum processor voltage and high complexity of the scheduling process
[11]	PSC	OS-based automatic episode detection mechanism for controlling the DVS	(Advanced Computer Architecture Lab)—The University of Michigan	10%–75% of application-dependent desktop processor energy savings	Some application spends almost the whole time at the lowest performance setting allowed in each model	Minimum processor voltage and high complexity of the scheduling process
[12]	PSC	Novel scheduling heuristic to DVFS	Department of Computer Science—North Carolina State University	With utilization of the Crusoe Cluster • 16% reduction in energy consumption with a 1% increase in execution time compare with no DFS • 9% reduction in energy consumption compared with a single-gear solutions	Reduced gears have a large time penalty on CPU bound. The result is a longer completion time and little or no energy savings	Minimum processor voltage and high complexity of the scheduling process
[13]	PSC	Energy-performance tradeoff using DFS	Department of Computer Science—North Carolina State University	10% energy savings on one node with a 1% increase in execution time—tested on single Athlon-64 processor in NAS	Decreasing CPU speed of a largely CPU bound program may result in slower execution and higher energy consumption	Minimum processor voltage and high complexity of the scheduling process
[14]	PSC	Weighted energy-delay product for controlling the DVS	Department of Computer Science and Engineering—University of South Carolina	Application-dependent system—25% energy saving rate on a 16-node Intel Centrino cluster with performance impact 2%; tested with NPB	Low stability in the energy reduction, the energy saving rate depends on the type of application, workload, the type of the system, and DVS strategy	Minimum processor voltage and high complexity of the scheduling process
[15]	PSC	CPUSPEED daemon, internal and external DVS scheduling	(Scalable Performance Laboratory), Department of Computer Science and Engineering—University of South Carolina	Application-dependent system—36% energy saving rate on NEMO, a 16-node DVS based cluster, with no negative impact on performance	Low stability in the energy reduction, the energy saving rate depends on the type of application, workload, the type of the system, and DVS strategy	Minimum processor voltage and high complexity of the scheduling process
[16]	PSC	CPU MISER-system-wide, application-independent, fine-grain DVFS scheduler	(Center for High-End Computing Systems), Department of Computer Science—Virginia Tech	20% energy saving rate on the NPB; users can define the constraints of the system performance loss rate in CPU MISER for most of applications	N/A	Minimum processor voltage and high complexity of the scheduling process

Table 2 (*Continued*)

Ref.	Approach	Technique	Energy Savings			Limitations
			Research Centre	Energy Improvement	Performance Drop	
[18]	PSC	PowerWatch with an internal optimization module for monitoring the DVS scheduling	Graduate School of Systems & Information Engineering—University of Tsukuba	40% energy saving rate in EDP with less than 5% performance on impact	N/A	Minimum processor voltage and high complexity of the scheduling process
[20]	PSC	Power-aware run-time (PART) system with β-algorithm as a run-time DVFS scheduler	(Los Alamos National Laboratory)—Los Alamos	20% energy saving rate on the NAS and 25% on the NPB with 3–5% performance impact	3–5% of performance degradation tightly controlled by the PART system	Minimum processor voltage and high complexity of the scheduling process
[21, 22]	PSC	ECOD—with a power-aware eco-friendly algorithm 'eco' for controlling the DVFS	Department of Computer Science—Virginia Tech	11%—an overall energy saving rate on the NPB	5.1% of the average loss in system performance	Minimum processor voltage and high complexity of the scheduling process
[26]	PSC	DVFS system—Jitter, an adaptive system for just-in-time performance scaling	Department of Computer Science—North Carolina State University	8%—reduction in energy consumption in Aztecwith 2.6% increasing of the execution time	As little as 2% time penalty, on a unbalanced program	Minimum processor voltage and high complexity of the scheduling process
[44]	PSC	Three schedulers for DVFS control	N/A	N/A	N/A	Minimum processor voltage and high complexity of the scheduling process
[28]	PSC	DVFS scheduling modules	N/A	N/A	N/A	Minimum processor voltage and high complexity of the scheduling process
[29]	PSC	Run-time scheduling algorithm for controlling the DVFS	N/A	N/A	N/A	Minimum processor voltage and high complexity of the scheduling process
[30]	PSC	An embedded run-time system that selects gears based on an energy-delay product for controlling the DVFS	Department of Computer Science—North Carolina State University	12% (in average) reduction of energy consumption in NAS with 2.1% increasing of the execution time	In practice, reducing the p-state can increase time	Minimum processor voltage and high complexity of the scheduling process
[32]	PSC	Energy-performance tradeoff in DVFS	Department of Computer Science—North Carolina State University	20% reduction in energy consumption with 3% increasing of the execution time (in a single node)—application independent approach	Reduced gears have a large time penalty on CPU bound. The result is a longer completion time and little or no energy savings	Minimum processor voltage and high complexity of the scheduling process
[34, 35]	LB	Just few nodes are utilized by the system—the rest are idle and can be switched to the 'sleep' mode or deactivated	Department of Computer Science—Rutgers University	Load-dependent system; significant savings during periods of light load	The throughput can be excessively sacrificed in favor of power and energy savings	Light loads are the exception rather than the rule

Table 2 (*Continued*)

Ref.	Approach	Technique	Energy Savings			Limitations
			Research Centre	Energy Improvement	Performance Drop	
[36]	PSC	TADVS—a scheduling module for controlling the DVS	N/A	N/A	N/A	Minimum processor voltage and high complexity of the scheduling process
[40]	PSC	Memory MISER—a modified Linux kernel and a daemon implementation of a PID controller to on-line and off-line memory scaling in the system operating mode	Department of Computer Science—Virginia Tech	70% memory energy savings and 30% total energy savings with less than 1% performance impact	N/A	High complexity of the scheduling process

of DVFS. The potential drawback of current SPM techniques is that improving energy efficiency by using low-power components has proven to be expensive. DPM techniques have shown promise for improving energy efficiency; however, designing power-aware schedulers is not trivial. Energy savings vary significantly with application, workload, cluster system, and scheduling strategy [15]. Because DVFS technology is inherently limited, it appears that future work in cluster DPM will involve a combination of power-scalable components, each with its own management scheme.

M.Y. Lim et al. provided a foundation for future work in [30], where they suggested a multi-component approach that uses DVFS to conserve processor power consumption during communication and computational phases, and uses load balancing techniques to on-line and off-line memory while running, further reducing power consumption. Over the last decade, attempts to improve energy efficiency in cluster computing have revealed a trend toward using software and power-scalable components in synthesis to reduce cluster power consumption while limiting performance loss. It is possible that future work will lead to one or more novel DPM systems that can interact with and manage a variety of power-scalable components during the execution of applications, thereby greatly improving the energy efficiency of cluster computing.

References

1. Andersen, D.G., Franklin, J., Kaminsky, M., Phanishayee, A., Tan, L., Vasudevan, V.: FAWN: A fast array of wimpy nodes. In: Proc. of the 22nd ACM Symposium on Operating Systems Principles (SOSP), Big Sky, MT (2009)
2. Beloglazov, A., Buyya, R., Lee, Y.C., Zomaya, A.: A taxonomy and survey of energy-efficient data centers and cloud computing systems. In: Zelkowitz, M. (ed.) Advances in Computers. Elsevier, Amsterdam (2011). ISBN 13:978-0-12-012141-0
3. Blue Gene/LTeam: An overview of the BlueGene/L supercomputer. In: Supercomputing 2002 Technical Papers (2002)
4. Buyya, R. (ed.): High Performance Cluster Computing: Architectures and Systems. Prentice-Hall, New York (1999)
5. Buyya, R., Cortes, T., Jin, H.: Single system image. Int. J. High Perform. Comput. Appl. **15**(2), 124–135 (2001)
6. Cameron, K.W., Ge, R., Feng, X.: High-performance, power-aware distributed computing for scientific applications. Computer **38**(11), 40–47 (2005)
7. Caulfield, A.M., Grupp, L.M., Swanson, S.: Gordon: using flash memory to build fast, power-efficient clusters for data-intensive applications. In: Proc. of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '09) (2009)
8. Chen, G., Malkowski, K., Kandemir, M., Raghavan, P.: Reducing power with performance constraints for parallel sparse applications. In: Proc. of the 19th IEEE International Parallel and Distributed Processing Symposium, p. 231a. IEEE Comput. Soc., Los Alamitos (2005)
9. Feller, E., Morin, C., Leprince, D.: State of the art of power saving in clusters and results from the EDF case study. Institut National de Recherche en Informatique et en Automatique (INRIA) (2010)
10. Feng, W., Cameron, K.: The green500 list: Encouraging sustainable supercomputing. Computer **40**(12), 50–55 (2007)
11. Flautner, K., Reinhardt, S., Mudge, T.: Automatic performance setting for dynamic voltage scaling. Wirel. Netw. **8**(5), 507–520 (2002)
12. Freeh, V.W., Pan, F., Kappiah, N., Lowenthal, D.K.: Using multiple energy gears in MPI programs on a power-scalable cluster. In: Proc. of 10th ACM Symp. Principles and Practice of Parallel Programming (PPoPP), pp. 164–173. ACM, New York (2005)
13. Freeh, V.W., Pan, F., Kappiah, N., Lowenthal, D.K., Springer, R.: Exploring the energy-time tradeoff in MPI programs on a power-scalable cluster. In: Proc. of Parallel and Distributed Processing Symposium, vol. 01 (2005)
14. Ge, R., Feng, X., Cameron, K.W.: Improvement of power-performance efficiency for high-end computing. In: Proc. of the 1st Workshop on High-Performance, Power-Aware Computing (2005), 8 pp.
15. Ge, R., Feng, X., Cameron, K.W.: Performance constrained distributed DVS scheduling for scientific applications on power-

aware clusters. In: Proc. of Supercomputing Conference, p. 34 (2005)
16. Ge, R., Feng, X., Feng, W., Cameron, K.W.: CPU MISER: a performance-directed, run-time system for power-aware clusters. In: Proc. of International Conference on Parallel Processing (ICPP07), p. 18 (2007)
17. Gropp, W., Lusk, E., Sterling, T. (eds.): Beowulf cluster computing with Linux, 2nd edn. MIT Press, Cambridge (2003)
18. Hotta, Y., Sato, M., Kimura, H., Matsuoka, S., Boku, T., Takahashi, D.: Profile-based optimization of power performance by using dynamic voltage scaling on a PC cluster. In: Proc. of the 20th IEEE International Parallel and Distributed Processing Symposium (IPDPS) (2006), 8 pp.
19. Hsu, C., Feng, W.: A feasibility analysis of power awareness in commodity-based high-performance clusters. In: IEEE International Conference on Cluster Computing, pp. 1–10 (2005)
20. Hsu, C., Feng, W.: A power-aware run-time system for high-performance computing. In: Proc. of ACM/IEEE SC Conference, p. 1. IEEE Comput. Soc., Los Alamitos (2005)
21. Huang, S., Feng, W.: A workload-aware, eco-friendly daemon for cluster computing. Technical Report, Computer Science, Virginia Tech (2008)
22. Huang, S., Feng, W.: Energy-efficient cluster computing via accurate workload characterization. In: Proc. of the 9th IEEE/ACM International Symposium Cluster Computing and the Grid, pp. 68–75 (2009)
23. IBM: Blue Gene/P. http://www-03.ibm.com/press/us/en/pressrelease/21791.wss. Accessed: July 2011
24. IBM: Blue Gene/Q. http://www-03.ibm.com/press/us/en/pressrelease/33586.wss. Accessed: July 2011
25. Intel Developer's manual: Intel 80200 Processor Based on Intel XScale Microarchitecture. Intel Press (1989)
26. Kappiah, N., Freeh, V.W., Lowenthal, D.K.: Just in time dynamic voltage scaling: exploiting inter-node slack to save energy in MPI programs. In: Proc. of ACM/IEEE Conference Supercomputing, p. 33 (2005)
27. Kim, K.H., Buyya, R., Kim, J.: Power aware scheduling of bag-of-tasks applications with deadline constraints on DVS-enabled clusters. In: Proc. of CCGRID, pp. 541–548 (2007)
28. Li, K.: Performance analysis of power-aware task scheduling algorithms on multiprocessor computers with dynamic voltage and speed. IEEE Trans. Parallel Distrib. Syst. **19**(11), 1484–1497 (2008)
29. Lim, M.Y., Freeh, V.W.: Determining the minimum energy consumption using dynamic voltage and frequency scaling. In: Proc. of the 3rd Workshop on High-Performance, Power-Aware Computing, pp. 1–8 (2007)
30. Lim, M.Y., Freeh, V.W., Lowenthal, D.K.: Adaptive, transparent frequency and voltage scaling of communication phases in MPI programs. In: Proc. of ACM/IEEE Supercomputing, p. 14 (2006)
31. Mobile AMD Duron Processor Model 7 Data Sheet. AMD (2001)
32. Pan, F., Freeh, V.W., Smith, D.M.: Exploring the energy-time tradeoff in high performance computing. In: Proc. of Parallel and Distributed Processing Symposium (2005)
33. Pfister, G.F.: In Search of Clusters, 2nd edn. Prentice-Hall, New York (1998)
34. Pinheiro, E., Bianchini, R., Carrera, E.V., Heath, T.: Load balancing and unbalancing for power and performance in cluster-based systems. In: Proc. of Workshop on Compilers and Operating Systems for Low Power (2001)
35. Pinheiro, E., Bianchini, R., Carrera, E.V., Heath, T.: Dynamic cluster reconfiguration for power and performance. In: Proc. of Workshop on Compilers and Operating Systems for Low Power, pp. 75–93 (2003)
36. Ruan, X., Qin, X., Zong, Z., Bellam, K., Nijim, M.: An energy-efficient scheduling algorithm using dynamic voltage scaling for parallel applications on clusters. In: Proc. of the 16th IEEE International Conference on Computer Communications and Networks, Honolulu, Hawaii, pp. 735–740 (2007)
37. Smith, S.E.: What is cluster computing? O. Wallace (ed.). Copyright 2003–2011. http://www.wisegeek.com/what-is-cluster-computing.htm
38. The Green500 list (June 2011). http://www.green500.org/lists/2011/06/top/list.php. Accessed: July 2011
39. The Green500. http://www.green500.org. Accessed: July 2011
40. Tolentino, M.E., Turner, J., Cameron, K.W.: Memory-miser: a performance-constrained runtime system for power-scalable clusters. In: Proc. of International Conference Computing Frontiers, pp. 237–246 (2007)
41. US EPA: Report to congress on server and data center energy efficiency. Technical report (2007)
42. Vasić, N., Barisits, M., Salzgeber, V., Kostic, D.: Making cluster applications energy-aware. In: ACDC. Proc. of the 1st Workshop on Automated Control for Datacenters and Clouds, pp. 37–42 (2009)
43. Vasudevan, V., Andersen, D.G., Kaminsky, M., Tan, L., Franklin, J., Moraru, I.: Energy-efficient cluster computing with FAWN: Workloads and implications. In: Proc. of e-Energy, Passau, Germany (2010)
44. von Laszewski, G., Wang, L., Younge, A.J., He, X.: Power-aware scheduling of virtual machines in DVFS-enabled clusters. In: Proc. of IEEE International Conference on Cluster Computing and Workshops, pp. 1–10 (2009)
45. Warren, M.S., Weigle, E.H., Feng, W.-C.: High-density computing: a 240-processor beowulf in one cubic meter. In: Proc. of IEEE/ACM SC2002, Baltimore, Maryland, pp. 1–11 (2002)
46. Yeo, C., Buyya, R.: A taxonomy of market-based resource management systems for utility-driven cluster computing. Softw. Pract. Exp. **36**, 1381–1419 (2006)

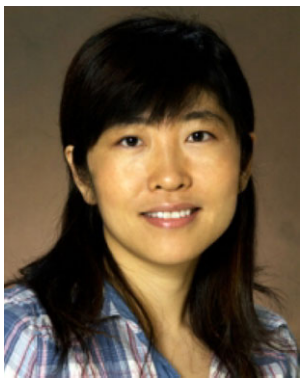
Giorgio Luigi Valentini is currently enrolled on a Master degree from the University of Luxembourg, G.D.L., working within the GreenIT project. He received a BEngCS degree from the University of Luxembourg, G.D.L. in 2009.

Walter Lassonde received his Undergraduate degree in Electrical Engineering in 2010 and he's currently enrolled on a Master degree from the North Dakota State University. His research interest are distributed systems and green cloud computing.

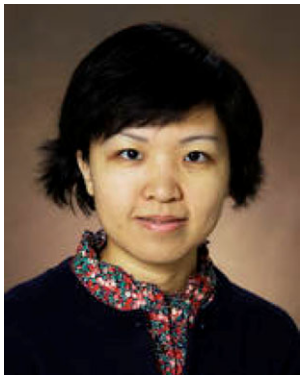
Samee Ullah Khan is Assistant Professor of Electrical and Computer Engineering at the North Dakota State University, Fargo, ND, USA. Prof. Khan has extensively worked on the general topic of resource allocation in autonomous heterogeneous distributed computing systems. As of recent, he has been actively conducting cutting-edge research on energy-efficient computations and communications. A total of 109 (journal: 39, conference: 50, book chapter: 12, editorial: 5, technical report: 3) publications are attributed to his name. For more information, please visit: http://sameekhan.org/.

Nasro Min-Allah is an Associate Professor and head of the Computer Science Department at COMSATS Institute of Information Technology, Pakistan. He received his Undergraduate and Master degrees in electronics and information technology in 1998 and 2001, respectively. He obtained a PhD in real-time systems from the graduate university of the Chinese Academy of Sciences, P.R China in 2008. His main research is focused on scheduling theory, green computing, and fault tolerant real-time systems. He is the recipient of the two most prestigious awards, i) CIIT Golden Medallion for Innovation (CIMI-2009), and Best Mobile Innovation in Pakistan (BMIP-2011).

Sajjad A. Madani is associate professor at the department of computer science at COMSATS Institute of information technology (CIIT), Abbottabad. From 2005 to 2008 he was a guest researcher with the institute of computer technology where he did his PhD research. He has done MS in Computer Sciences from Lahore University of Management Sciences (LUMS), Pakistan with excellent academic standing. He has also done BSc Civil Engineering from UET Peshawar and was awarded a gold medal for his outstanding performance in academics. His areas of interest include low power wireless sensor network and application of industrial informatics to electrical energy networks.



Juan Li is an assistant professor at the Computer Science Department of the North Dakota State University, Fargo, ND, USA. She received a B.S. degree from Beijing Jiaotong University, Beijing, China, in July 1997, and a Ph.D. degree from the University of British Columbia, Vancouver, Canada, in May 2008. Dr. Li's major research interest lies in distributed systems, including P2P networks, grid and cloud computing, mobile ad hoc network, social networking, and semantic web technologies.



Limin Zhang is an assistant professor of Management Information Systems in the Accounting, Finance, and Information Systems Department at North Dakota State University. She receives her PhD in Management Information Systems from the University of Arizona. Her research interests include context-based Web search, virtual team collaboration, and using technology to support competitive intelligence and decision-making. Her papers have been published in Review of Business Information Systems, International Journal of Management and Information Systems, and various conference proceedings. Her teaching interests include database design, Web system development, and information technology management.

Lizhe Wang is a Principal Research Engineer at School of Informatics and Computing, Indiana University. Dr. Lizhe Wang received his Bachelor of Engineering with honors and Master of Engineering both from Tsinghua University, P.R. China and his Doctor of Engineering with magna cum laude from University Karlsruhe (now Karlsruhe Institute of Technology), Germany. Dr. Lizhe Wang serves as an editor of Journal of Cluster Computing and Journal of Cloud Computing, Springer.

Nasir Ghani is an Associate Professor and Associate Chair of the Electrical and Computer Engineering Department at the University of New Mexico, USA. Currently he is involved in a wide range of research activities in the high-speed networking and services areas. Dr. Ghani is a Senior Member of the IEEE. He has also chaired several symposia for *IEEE GLOBECOM*, *IEEE ICC*, and *IEEE ICCCN*, as well as several workshops for *IEEE INFOCOM*. He also has extensive industrial experience totaling over 8 years and received the Ph.D. degree in Computer Engineering from the University of Waterloo.

Joanna Kolodziej graduated in Mathematics from the Jagiellonian University in Cracow in 1992, where she also obtained the Ph.D. in Computer Science in 2004. She joined the Department of Mathematics and Computer Science of the University of Bielsko-Biała as an Assistant Professor in 1997. She has served and is currently serving as PC Co-Chair, General Co-Chair and IPC member of several international conferences and workshops. Dr Kołodziej is Managing Editor of IJSSC Journal and serves as a EB member and guest editor of several peer-reviewed international journals.

Hongxiang Li is currently an assistant professor in the Electrical and Computer Engineering Department of University of Louisville. Prior to that, he was an assistant professor at North Dakota State University. He received a B.S. degree from Xi'an Jiaotong University, China in 2000, a M.S. degree from Ohio University in 2004, and a Ph.D. degree from University of Washington, Seattle in 2008, all in electrical engineering. His general research interests are mobile communications and wireless networks.

Albert Y. Zomaya is currently the *Chair Professor of High Performance Computing & Networking* in the School of Information Technologies, The University of Sydney. He is the author/co-author of seven books, more than 400 papers, and the editor of nine books and 11 conference proceedings. He is the Editor in Chief of the *IEEE Transactions on Computers* and serves as an associate editor for 19 leading journals. Professor Zomaya is the recipient of the *IEEE TCPP Outstanding Service Award* and the *IEEE TCSC Medal for Excellence in Scalable Computing*, both in 2011.

Cheng-Zhong Xu is a Professor of Electrical and Computer Engineering at Wayne State University and the Director of the Cloud and Internet Computing Laboratory (CIC). His main research interests are networked computing systems with an emphasis on resource management. Dr. Xu obtained BSc. and MSc. degrees from Nanjing University in 1986, and 1989, respectively, and a Ph.D. degree from the University of Hong Kong in 1993, all in Computer Science and Engineering. Dr. Xu is a recipient of "President's Awards for Excellence in Teaching" of Wayne State University in 2002 and "Career Development Chair Award" in 2003.

Pavan Balaji holds a joint appointment as an Assistant Computer Scientist at the Argonne National Laboratory and as a research fellow of the Computation Institute at the University of Chicago. He had received his Ph.D. from the Computer Science and Engineering department at the Ohio State University. His research interests include high-speed interconnects, efficient protocol stacks, parallel programming models and middleware for communication and I/O, and job scheduling and resource management. He has received several awards for his research activities, several best paper and other awards. He is a member of the IEEE and ACM.

Abhinav Vishnu is a member of the HPC group at the Pacific Northwest National Laboratory. His current research interests include designing scalable, energy efficient and fault tolerant programming models on high speed interconnects. He received his PhD from The Ohio State University in 2007 in scalable communication runtime systems for high performance networks. Dr. Vishnu is a recipient of IBM PhD fellowship for his research on runtime systems on IBM proprietary interconnection networks.

Fredric Pinel received his Master of Information Technology and Innovation from the Facultés Universitaires Notre Dame de la Paix, Belgium in 2005, and his Master in Electronics, Telecom and Hyperfrequencies from Esigelec, France in 1991. He is currently pursuing his Ph.D. degree in Computer Science from the University of Luxembourg. His research interests are complex systems, energy-efficiency and heuristics.

Johnatan E. Pecero received his Ph.D. degree in Computer Science from the Grenoble Institute of Technology (INPG) in 2008. His research at the LIG Laboratory focused on Scheduling with disturbances on New computing platforms. He is currently a postdoctoral research fellow under the Aide à la Formation Recherche (AFR) grant scheme of the Fonds National de la Recherche Luxembourg (FNR) with the Computer Science and Communications (CSC) research unit of the Faculty of Sci-

ences, Technology and Communications of Luxembourg University. His current research interests include green and robust scheduling, scheduling strategies for heterogeneous and distributed computing systems, nature-inspired optimization techniques.

Dzmitry Kliazovich is an AFR Research Fellow at the Faculty of Science, Technology, and Communication of the University of Luxembourg. He holds an award-winning Ph.D. in Information and Telecommunication Technologies from the University of Trento, Italy. Dr. Kliazovich is a holder of several scientific awards. His work on energy-efficient scheduling in cloud computing environments received Best Paper Award at the IEEE/ACM International Conference on Green Computing and Communications (GreenCom) in 2010. His main research activities are in the field of energy efficient communications and next-generation networking.

Pascal Bouvry earned his undergraduate degree in Economical & Social Sciences and his Master degree in Computer Science with distinction ('91) from the University of Namur, Belgium . He went on to obtain his Ph.D. degree ('94) in Computer Science with great distinction at the University of Grenoble (INPG), France. His research at the IMAG laboratory focussed on Mapping and scheduling task graphs onto Distributed Memory Parallel Computers. Dr. Bouvry is currently heading the Computer Science and Communications (CSC) research unit of the Faculty of Sciences, Technology and Communications of Luxembourg University, and serving as Professor. Pascal Bouvry is also treasurer & member of the administration board of CRP-Tudor, and member of various scientific committees and technical workgroups.