A Renewable Energy Driven Approach for Computational Sprinting

Haoran Cai, Qiang Cao, *Senior Member, IEEE*, and Hong Jiang, *Fellow, IEEE*

Abstract—Computational Sprinting, which allows a chip to exceed its power and thermal limits temporarily by turning on all processor cores and absorbing the extra heat dissipation with certain phase-changing materials, has proven to be an effective way to boost the computing performance for bursty workloads. However, extra power available for sprinting is constrained by existing power distribution infrastructures. Using batteries alone to provide the additional power to achieve performance target not only limits the effectiveness of sprinting, but also negatively impacts the lifetime of the batteries. Leveraging renewable power supply in a green data center provides an opportunity to make full use of Computational Sprinting. However, the intermittent nature of renewable energy, along with limited cooling capacity, makes it very challenging. In this paper, we propose GreenSprint, a renewable energy driven approach that enables a data center to boost its computing performance efficiently by conducting computational sprinting under the intermittent and time-varying nature of renewable energy supply. Three basic strategies are designed to determine the core count and frequency level for sprinting based on current power supply. Furthermore, we propose a *Hybrid* strategy that combines reinforcement learning to dynamically determine the optimal server setting, targeting at both the power provision safety and the quality of service. In consideration of practical cooling conditions, we also present a thermal-aware sprinting strategy *Hybrid-T*. Finally, we build an experimental prototype to evaluate GreenSprint on a cluster of 10 servers with a simulated solar power generator. The results show that renewable energy by itself can sustain different duration lengths of sprinting when its supply is sufficient and can improve performance by up to 4.8x for representative interactive applications. We also show the effectiveness of core-count and frequency scaling in the presence of varied renewable power and limited battery energy.

Index Terms—Computational sprinting, green data center, renewable energy, energy efficiency

1 Introduction

Computational sprinting has been widely explored in recent years [10], [11], [33], [35], [43]. Due to thermal constraints, some of the processor cores on a chip must be powered off most of the time, a phenomenon known as dark silicon [10], [34]. Many recent studies have demonstrated that computational sprinting, in which idle cores are activated and voltage and/or frequency are increased to allow the thermal constraints to be crossed for a short period of time by absorbing the extra heat dissipation with special phase-changing materials [11], [33], can effectively and significantly speed up application performance during workload bursts. For data centers with interactive workloads (e.g., search, forum, news), while workload bursts can be less frequent, the intensity of such bursts are usually much higher under a variety of circumstances [43], such as breaking news, online shopping big sales (e.g., the Black Friday after Thanksgiving), etc. As illustrated by Fig. 1, the diurnal workload pattern (dotted line) from a study of a Google data center [38] consists of several load spikes during the whole day with varying burst intensities and durations. There exists a great opportunity for leveraging computational sprinting to guarantee the quality of service in these cases.

However, many prior works manifested that today's data centers are already approaching the peak capacity of their power infrastructures [22], [39] which is similar to the thermal constraint at the chip level. The extra bursty power demand required by computational sprinting at the data-center level can induce serious power emergencies [43] as indicated in Fig. 1 by the red ovals when the demand exceeds the grid power capacity. Rejecting service requests due to power capacity cap may cause data centers to lose revenue and customers in the long term. Existing solutions to deal with bursty power demand mainly focus on battery-backed power system [18], [25] or one combined with overloading circuit breaker [11], [43]. However, using battery alone can be energy inefficient and even harms the lifetime of batteries due to the frequent charging/discharging activities [28]. An emerging solution to the power emergency problem is to leverage green energy sources to supplement grid power capacity. In such green data centers, the power-constrained grid can be used as backup for the renewable power supply or vice versa. Also, the renewable power solution can tackle the environmental challenges brought by power consumption and carbon emissions. Compared

- *H. Cai and Q. Cao are with Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System of Ministry of Education, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China. E-mail: {caihaoran, caoqiang}@hust.edu.cn.*
- *H. Jiang is with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019. E-mail: hong.jiang@uta.edu.*

Manuscript received 14 Aug. 2018; revised 10 Nov. 2018; accepted 23 Dec. 2018. Date of publication 28 Dec. 2018; date of current version 12 June 2019. (Corresponding author: Haoran Cai.)
Recommended for acceptance by S. He.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPDS.2018.2890230

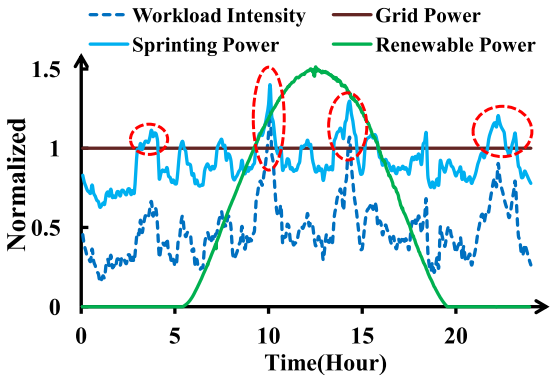


Fig. 1. Workload pattern for a Google data center [38] and scaled power demand of sprinting normalized to grid power.

with traditional data centers without green power supply, this solution avoids expensive capital expenditure and time-consuming construction cycle (ranging in the hundreds of millions USD and decades of years) of upgrading grid power infrastructures [12], [20], [21]. In a typical data center, the capital cost spent on provisioning the grid utility infrastructure is between $10-$25 for each watt [12]. Therefore, the renewable power solution brings us an opportunity to deal with burst power demand in a cost-efficient way.

Thus, in this paper, we first ask and then try to answer the following questions, *can we leverage green power supply to support computational sprinting in green data centers, and if so, how?* Although renewable energy is an attractive solution, it is also well known for its intermittent and variable nature which is also demonstrated in Fig. 1 for a typical solar power supply. Thus, directly conducting sprinting using renewable energy in a green data center, without appropriate control, can have negative impacts, for example, reducing the lifetime of batteries, degrading the performance of services leading to SLA violations, and even causing power failures in the data center. Hence, we consider the solution under three cases: (1) When the green power supply complementing the grid power can fully satisfy the bursty power requirement, we operate the sprinting with only green power and then charge the surplus green power into energy storage devices. (2) When the green power supply is insufficient for the power demand, the batteries, strategically charged either by renewable source or grid source during non-sprinting periods, discharge to make up for the power shortage. (3) When the green power is not available, certain power management knobs (e.g., server-level low power state) can be adopted to manage sprinting power to match the current power provision, potentially resulting in a performance degradation for some applications. Obviously, applying computational sprinting by leveraging green power to provision power bursts in power-constrained data centers can help significantly save the capital expense while improving application performance. Further, although the thermal constraint on chip has been discussed in prior work [33] and emerging heat sink materials or PCM can be used to increase the thermal dissipation capacity, the impact on renewable power system has not been well exploited. Specifically, even though the renewable power supply is sufficient, the maximal sprinting can cause chips overheating due to strict thermal constraint, resulting in performance degradation, sudden server outage and even being destroyed [7]. Therefore, the thermal aware design is inevitable because the actual thermal capacity determines the sprinting intensity and duration. As a new challenge, besides of conventional concerns on workloads and power provision, a green data center conducting computational sprinting must comprehensively consider the thermal capacity of server and current server status such as temperature, to determine optimal scheduling.

Based on the analysis above, we propose GreenSprint, a green data center based approach that exploits renewable energy to effectively and efficiently conduct computational sprinting. To the best of our knowledge, while many prior works have focused on supporting computational sprinting at the chip level or at the data center level by utilizing battery supply only, the issues of computational sprinting in the presence of green energy at the data center level and its cost-benefit trade-offs have not been well addressed in the literature. In exploring the design space of employing renewable energy for computational sprinting, this paper makes the following contributions: (1) We propose *GreenSprint*, a renewable energy driven approach that enables data center level sprinting by turning on more cores and/or boosting their voltage and frequency in the era of dark silicon, in order to handle occasional workload bursts. (2) We first present three basic strategies to determine the core count and the frequency level for sprinting based on the intermittent and time-varying renewable power supply. Moreover, we further propose a *Hybrid* strategy that combines reinforcement learning to determine the optimal server setting, targeting at both the power provision safety and the quality of service. (3) Concerning to the practical limitation of cooling capacity on chip, we present an extended strategy called *Hybrid-T*, which is the first thermal aware strategy for computational sprinting. (4) We develop an experimental prototype consisting of 10 servers, a simulated solar power generator, and a server-level battery provision to evaluate our approach. Using representative data center workloads, the evaluation shows that our solution can improve the average computing performance by up to 4.8x for SPECjbb, 4.1x for Web-Search, 4.7x for Memcached, and 2.6x for scientific computation workload as MCF under renewable power supply. (5) Based on the experimental results, we further analyze the interplay between renewable power, cooling capacity, battery energy, sprinting duration, workload characteristics. We draw several insightful observations to guide computational sprinting in green data centers.

2 Power Infrastructure in Green Data Centers

Considering the fact only part of the cores in the multicore servers in typical data centers are active due to the dark silicon phenomenon, it offers the potential to apply the computational sprinting technique to boost performance of applications with bursty workloads, particularly interactive workloads, by turning on additional cores and increasing their voltage and frequency. This is possible, however, only if thermal constraints at both the server/chip level and data center level can be temporarily stretched, i.e., with the necessary heat-absorbing materials and cooling equipment, as well as the power supply infrastructure is able to meet the

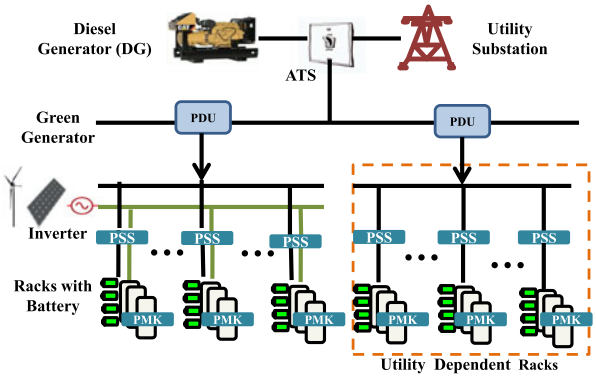


Fig. 2. The Power Infrastructure of GreenSprint.

bursty power demand of sprinting whenever the demand arises. Since our focus in this paper is on leveraging renewable energy in data centers to meet the bursty power demand of computational sprinting, in this section we will introduce the power infrastructure for computational sprinting in a green data center and make appropriate assumptions about the thermal constraints.

Overview. Fig. 2 depicts an architectural overview of an on-site green data center power hierarchy for computational sprinting, similar to prior works [22], [45]. To achieve the sprinting goal, we directly connect on-site renewable power supplies such as photovoltaic (PV) and wind to the power distribution unit (PDU) level to provide a dual-power supply of the grid and renewable power rather than integrating the renewable energy into the utility power. This can help mitigate the impacts of voltage transients, frequency distortions and harmonics. Compared with the centralized power integration, our distributed integration prevents PDUs becoming a power delivery bottleneck. The existing uniform centralized power provision mechanism forces all servers to obtain the same or similar renewable power capacity, significantly limiting the sprinting power supply. To this end, provisioning renewable energy on the PDU level allows us to apply computational sprinting in a data center on a per-rack basis. When we conduct sprinting, some servers can be powered only by the renewable energy with a separate green power bus while others still depend on the grid power. This can help greatly relieve the burden on the circuit breakers (CB) and the underlying constrained power infrastructure.

Renewable Power. The greatest challenge of powering sprinting with renewable energy is time-varying, intermittent nature of renewable power. Therefore, we employ a power-source selector (PSS) to adaptively switch among different power sources (i.e., green, battery and grid power). PSS performs switch tuning based on the discrepancy between the workload power demand and the green power supply. PSS is also configured to charge batteries when there is an excess of green power, and discharge them when green power is insufficient or unavailable. PSS can identify all switching parameters for the inverter and charge controllers of batteries to allow for a full control of every power source. Programmable power electronics circuitry can be used to implement PSS. However, since our experimental setup does not provide this functionality, our evaluations account for such control capability in the form of managing the sprinting decisions. As shown in Fig. 2, a server-level power management knob (PMK) receives the execution output from the PSS to control the power demand on a per-server basis. When renewable power and battery are not sufficient, PMK decreases the sprinting intensity to keep servers within the power budget by considering applications' diverse characteristics.

Battery. Energy storage devices must be deployed to support sprinting continuously when the renewable energy is insufficient. Prior works proposed to rely on uninterruptible power supply (UPS) devices when the utility power source suddenly fails [11], [43]. Since we connect the green power to the PDU level, we also leverage the distributed battery architecture shown in Fig. 2, which is widely employed by IT companies such as Google [40] (server-level battery) and Facebook [3] (rack-level) to smooth the supply of the renewable power. The former design achieves energy efficiency by bringing the AC distribution (green and grid) even closer to the IT load, before it is converted (we adopt this design in our solution). The distributed design can provide great scalability and avoid AC-DC-AC double conversion.

However, battery-based sprinting can have significant adversary effects on the battery lifetime because batte ries wear out under irregular charging and discharging regimes [27]. There are two main factors affecting the battery failure rate. First, discharging current, which has strong relationship with the sprinting intensity, indicates the capacity performance of battery. For example, while the rated capacity is 24Ah at a 20-hour discharging rate, the actual capacity drops to only 12Ah at a 12-min discharging rate. Second, a high depth of discharging (DoD) can degrade the battery lifetime. For the purpose of prolonging the battery lifetime, we cannot exhaust the energy of a battery. In our work, we model a server-level 12V value-regulated lead-acid battery (VRLA), similar to that used by Li et al. [27]. Batteries are characterized by their supply time as approximated by Peukert's Law (Peukert's exponent is 1.15 for LA battery [21]), which shows the time taken to drain a certain capacity for different power demands. We also assume DoD=40 percent in our setup, which translates to a lifetime of 1300 recharge cycles [21].

Thermal Concerns at the Chip Level. Another challenge is cooling at the server level. A chip multiprocessor's sprinting level depends on the cooling capacity of its thermal package and heat sink. In other words, the thermal constraint at the chip level directly determines the maximum duration of sprinting. In runtime, sprinting activities actually produce more heat into the ambient environment than in a normal server mode. Therefore, the server system needs to effectively remove the extra heat and prevents over-heating on chip-level. Fortunately, prior work [36] found an effective way to shape the thermal load of a data center. In that work, phase changing materials (PCM) is used to temporarily store the heat generated by servers and other equipment during peak loads, and release the heat into the ambient environment when there is excess cooling capacity during non-sprinting periods. Indeed, PCM can delay the onset of thermal limits by hours. In our work, we first assume that servers are equipped with such emerging thermal package for sprinting and the server can sustain different demands of heat dissipation resulting from computational sprinting. Then we further discuss about the thermal impact on sprinting detailed in Section 4.

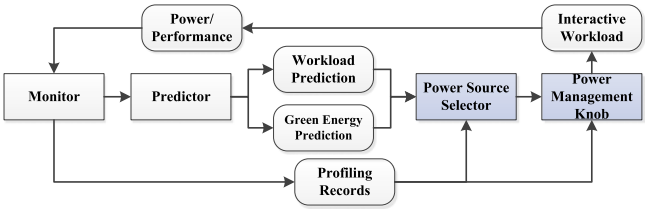


Fig. 3. The GreenSprint architecture.

3 Design of GreenSprint

In this section, we present GreenSprint, a renewable energy driven framework that enables data center sprinting by managing power sources and scheduling workloads. The GreenSprint framework is designed to assist the power source selector and power management knobs in their decision-making process. We present the architecture of GreenSprint in Fig. 3. The main components are *Monitor*, *Predictor*, *PSS*, and *PMK*. The *Monitor* collects the performance of workload (e.g., latency and throughput) and the power used (e.g., battery energy, renewable power, and server power). The *Predictor* predicts the workload intensity and the renewable energy production. The *PSS* takes these predictions and the available power to choose appropriate power sources. The *PMK* uses the profiling records and the power supply to manage sprinting activities for bursty workloads. In the following, we emphasize on the details of the *PSS* and *PMK*.

3.1 Power Source Selector

The sprinting power provision can come from renewable power or battery, depending on the decision made by a power source selector. Fig. 4 illustrates how the green power and battery energy are provided at the rack level. In each case, the duration of a power burst is divided into a series of discrete scheduling epochs ($T1$, $T2$, ..., Tn) which can be classified into the following three possible cases:

Case 1: Renewable power, (*RESupp*), is abundant and can be independently used for sprinting (from T1 to T2). The excess power beyond the sprinting needs can be used to charge the battery. In this period, power supply depends on renewable power. Sprinting starts from additional cores being activated and ends when the workload requests are finished or batteries join back in power supply.

Case 2: Renewable power is insufficient. Due to the time-varying, intermittent nature (e.g., weather condition, time of sprinting, etc.), the green power supply temporarily needs the supplement from other power sources, such as batteries. To this end, we employ battery power (*BattSupp*) to supplement the green power to sustain the sprinting (from T2 to T3) immediately. To make this work, power management knobs must work cooperatively with PSS. This case ends when green power supply becomes unavailable.

Case 3: The battery independently sustains the sprinting when the renewable power is unavailable (from T3 to T4). If the workload burst can be completed in this period, then we charge the battery with grid power in anticipation of future sprints. The worst case happens when there is no sufficient battery energy left, then overloading circuit breaker for the grid power may be the last resort to maintaining sprinting. Recharging is activated when battery depth of discharge

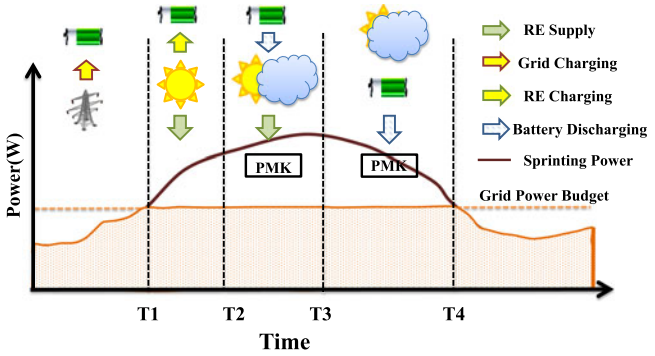


Fig. 4. An Illustration of Power Source Selector under Different Power Supply/Demand Scenarios.

reaches the set goal (40 percent DoD). To prevent tripping the CBs with too much power overload, we limit the total power of all the downstream branches under an upper bound. Again, due to limitations on batteries, PMK and PSS should cooperate with each other in case of a power emergency.

To help the PSS determine which case the power system should step in, GreenSprint makes a judgement on the relationship between *RESupp*, *BattSupp* and the power demand *LoadPower* in each scheduling epoch. In this design, we calculate *BattSupp* based on previous discharging activities. Specifically, the energy usage of battery during each epoch is recorded by a controller node in a cluster. The available capacity is derived from the maximum capacity and the used capacity. To properly capture *Peukert's* effect during discharging, we recalculate the remaining discharging time after each scheduling epoch.

For *RESupp*, we present a renewable energy prediction model for the *Predictor* with lower complexity and short time horizons. The prediction is based on the past power production records. In particular, we continuously calculate an exponentially weighted moving average (EWMA) of the past average power production in Equation (1):

$$RESupp(t) = \alpha * RESupp(t-1) + (1-\alpha) * Obs(t), \quad (1)$$

where *Obs(t)* is the observed power production and *RESupp (t)* is the predicted power supply for the next epoch (e.g., 5 minutes) in the current epoch t. α reflects a tradeoff between stability and responsiveness and it ranges from 0 to 1. When α varies, we find α=0.3 to be the most consistent, which weights the model more heavily towards current observed data. Note that most solar prediction algorithms are accurate when weather conditions are stable.

3.2 Power Management Knobs

The power management knob is introduced to manage power demand. When a workload burst occurs and there is an abundant renewable power supply, the workload requests can be quickly completed by sprinting without triggering the grid or battery power. Further, the surplus renewable energy is used to charge the battery. When the power source can no longer sustain the power demand, we finish sprinting by deactivating the additional active cores and setting the frequency to the lowest level, *Normal* mode. The case where the power supply is insufficient becomes more complicated. PMK should determine an appropriate sprinting intensity for

better energy efficiency. In what follows, we present four different power management strategies for computational sprinting in a green data center, namely, *Greedy*, *Parallel*, *Pacing* and *Hybrid*.

The most straightforward solution is to simply activate all cores and set the highest frequency to temporarily accommodate workload bursts. We call this the *Greedy* strategy because it needs aggressive power supply. This strategy does not assume any prediction of the future green energy production. It greedily tries to run with the maximal sprinting intensity, seeking maximum improvement on performance for each application. When the power supply is sufficient or the workload intensity is moderate, *Greedy* can strictly ensure the quality of service (QoS) for latency-critical applications, further reduce the response time of each request. For example, *Greedy* can achieve an average 270 ms latency for SPECjbb at 70 percent burst load intensity, while a best-efficiency policy (lower sprinting intensity and higher response time) can only provide 466 ms latency with a 500 ms latency constraint. Obviously, most service providers are willing to deploy *Greedy* that improves the user experience by maximal sprinting, which may bring additional revenue. However, when the power supply is insufficient, or the workload bursts are intensive, the service may be unavailable due the high power consumption of *Greedy*. When the batteries kick in, considering the burst duration, lower sprinting intensity may be a better choice because of lower power consumption, which leads to longer discharging time. For comparison, we develop the following three strategies.

During each sprinting interval, each server $j = 1, ..., n$ in a rack c can operate in a particular sprinting intensity $S_j \in S$. S is a two-dimensional set consisting of frequency level and core count. It is ordered from S_0, which is the *Normal* mode (e.g., 6 cores with 1.2 GHz in our testbed), to S_r, which is the maximum sprinting mode (e.g., 12 cores with 2.0 GHz). We also denote the intensity of a workload during this sprinting interval as $L_j \in L$, which can be any of the w levels, $L_1, ...,$ 'L_w, between the minimum and maximum intensity levels for a given application. The power demand for each sprinting epoch t not only depends on the workload intensity level $L_{j,t}$ being served on server j during epoch t but also on the server configuration S_j. We measure and collect the power demand (denoted as $LoadPower_j(L_{j,t}, S_{j,t})$) of an individual workload for each server settings S_j and workload intensity levels L_j with *a priori* knowledge using an exhaustive method on real servers.

Another two strategies we proposed are called *Parallel* and *Pacing*. *Parallel* scales only the core count while *Pacing* scales only the frequency levels at each time. For *Parallel* and *Pacing*, we intend to explore the impact of two power scaling techniques on the performance with the renewable power supply, which has not been well discussed in prior works. We use the EWMA prediction again to predict the workload intensity $L_{pre,t}$. Then the potential maximal power demand can be denoted as $LoadPower_j(L_{pre,t}, S_{r,t})$. Then there exists a power mismatch between power supply and demand, denoted as $M_t = \sum_{j=1}^{n} LoadPower_j(L_{pre,t}, S_{r,t})$- $PowerSupp_t$, where $PowerSupp_t$ is the sum of $RESupp(t)$ and $BattSupp(t)$ for the whole rack. The power reduction $P_{M,t} = \sum_{j=1}^{n} (LoadPower_j(L_{pre,t}, S_{r,t}) - LoadPower_j(L_{pre,t}, S_{j,t}))$ offered by scaling core count or frequency level for each epoch t. Therefore, we handle the power mismatch M_t by carefully managing power reduction $P_{M,t}$. Thus, we arrive at the following equation:

$$\forall t \in T : RESupp(t) + PBattSupp(t) + P_{M,t} = \sum_{j=1}^{n} LoadPower_j(L_{pre,t}, S_{r,t}). \quad (2)$$

To this end, an optimal setting S_j is achievable to maximize the overall performance $Perf_{j,t}$ in each epoch t for servers in rack c. We denote the optimization target as:

$$max \sum^{t \in T} \sum^{j \in c} Perf_{j,t}(L_{j,t}, S_{j,t}). \quad (3)$$

Parallel and *Pacing* solve the problem under constraints on service quality of service (QoS), renewable power production, and battery power supply (e.g., DoD, capacity), which is similar to some previous studies [18], [19], [45].

Finally, we present a *Hybrid* strategy, which combines both frequency and core count scaling in *Parallel* and *Pacing* with *reinforcement learning*. *Hybrid* tries to learn the optimal settings to achieve higher energy efficiency and strict QoS guarantee. In our work, we first formulate this problem as a Markov Decision Process (MDP). In an MDP, a decision-making process must learn the best course of action to maximize its total reward over time. At each discrete epoch, the system can observe its current *state*, c_t, and it must choose an *action*, a_t from a finite set of alternatives. Depending on the chosen action and current state, there is an unknown probability distribution controlling which state c_{t+1} it enters next and the reward r_t that it receives. The problem is to maximize the total discounted reward, $\sum_{t=0}^{T} \gamma^t r_t$, where γ is the discounting factor. γ should be positive and less than one, in order to reflect a preference for rewards in the near future.

In our power management problem, the *state* c_t indicates the current power supply $PowerSupp$ and workload intensity, measured during epoch t-1. Specifically, we quantize the power supply for each server, from the point of idle server power to the point of maximum sprinting power, into discrete sets by static *step* like the workload intensity level L. A small step improves the energy savings, but it tends to cause frequent changes in configuration for small changes in workload intensity and power supply. In our design, we empirically determine the step as 5 percent to improve energy efficiency. The *action* a_t, which is chosen depending on the state, is the combinations of core count and frequency levels, i.e., $a_t \in S$. The reward r_t is determined by the level of QoS relative to the target and the power consumption relative to the power demand.

Reinforcement learning is a type of unsupervised machine learning with a focus on online learning [31]. It solves an MDP by maintaining a *lookup table R(c,a)*, which is similar to another work [32]. The entry estimates the total discounted reward that will be received if the action a has been chosen based on the current state c. In our work, to reduce the complexity of the problem, we learn the initial values of lookup table from the profiling data collected by *Parallel* and *Pacing* using Algorithm 1. The power reward and QoS reward in reward mechanism are defined as R_{power} and R_{qos}

respectively. If R_{power} is greater than one, then the power demand has been satisfied by the power supply, it demonstrates that the server can be powered normally and the power demand has been well managed. In this case, if the QoS has been ensured, i.e., R_{qos} is greater than one, then we give a positive reward. A larger reward means sprinting can provide lower response time for each request. If the QoS can not been ensured, we add a negative reward. Finally, if R_{power} is less than one, then the power supply can not meet the demand due to sprinting, therefore the total reward is negative. Once the reward r_t has been calculated, line 15 updates the value of $R(c_t, a_t)$ in the lookup table. We empirically set the discounting factor γ as 0.9 to allow a balance between short-term and future rewards. The learning rate α, we used α=0.7 in our experiments, controls the rate at which the values of $R(c_t, a_t)$ are updated. A large value of α means that the algorithm learns quickly. *Hybrid* uses the lookup table to select the best action a_t, which is the one that gives the largest total reward; i.e., $a_t = \arg\max_{a \in S} R(c_t, a)$. In order to improve the decisions, we also continue to update the values in the lookup table. *Hybrid* has a simple algorithm implemented using *Python*, so its runtime overhead is negligible (<2 ms).

Algorithm 1. Reward Mechanism

1: // Calculate reward r_t based on epoch t and $t+1$
2: QoS_{target} and $QoS_{current}$ represent the target QoS of the workload and the current latency result. $PowerSupp$ and $PowerCurr$ are the power supply and the current power demand at time t.
3: R_{power} = $PowerSupp$ / $PowerCurr$ //*Power reward*
4: R_{qos} = QoS_{target} / $QoS_{current}$ //*QoS reward*
5: **If** $R_{power} > 1$ **Then**
6: **If** $R_{qos} > 1$ **Then**
7: $r_t = R_{power}+R_{qos}+1$
8: **Else**
9: $r_t = R_{power}-R_{qos}+1$
10: **EndIf**
11: **Else**
12: $r_t = -R_{power}-1$
13: **EndIf**
14: // Update the value of $R(c_t, a_t)$ in lookup table
15: $R(c_t, a_t)=R(c_t, a_t) + \alpha[r_t + \gamma max_{a_i \in S} R(c_{t+1}, a_i) - R(c_t, a_t)]$

3.3 Green Power Distribution

Since the renewable energy is provisioned at the rack level and each server is deployed with distributed battery (Section 2.2), we need to know how to power all the servers on the rack when current power supply is insufficient. In principle, there are two power distribution schemes that are feasible for our data center. In the first scheme, all the serves are kept active and operate with the same configuration with an equal but low average-power supply, referred as *Uniform*. In this case, all green-provisioned servers can sprint at a low performance level all the time. The second scheme, called *Packing*, powers off/on these green-provisioned servers one at a time once there is sufficient extra renewable power in a given scheduling epoch. It prefers to power a single server to achieve the best performance state in each scheduling epoch. If there is enough renewable power headroom for the next server to wake up, then this process repeats for the next server in the next scheduling epoch. Otherwise, when the headroom shrinks, we adjust the server into *save-state* (e.g., *Sleep or Hibernation*) as appropriate. To ensure data consistency, we use shared storage through a network file system (NFS) in the controller node.

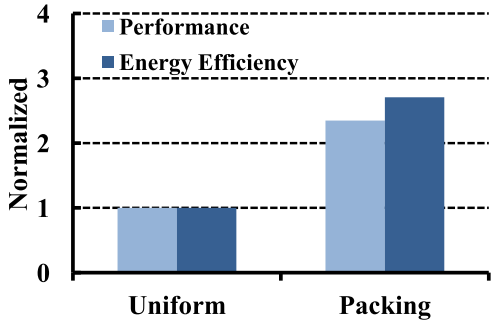


Fig. 5. The impact of power distribution schemes.

We evaluate these two schemes using 4 servers of the same type. The power budget is set as a constrained power supply that can only support 4 servers running in the *Normal* mode with the *Uniform* power distribution scheme. We generate as many SPECjbb workload requests as possible to simulate workload bursts. As Fig. 5 shows, *Packing* outperforms *Uniform* for both performance and energy efficiency (defined as jops/watt for SPECjbb) and achieves 2.3x and 2.7x overall improvement respectively. We find that *Packing* can very effectively exploit the power supply potential since this scheme leads to a linear relationship between energy efficiency and sprinting intensity. However, we expect the discrepancy to become smaller for future multi-server and multi-core systems because high idle power can be amortized by more cores, resulting in higher energy efficiency. Another important concern for renewable power is its utilization. *Uniform* needs higher *start* point of power supply due to the large fraction of server idle power, while *Packing* can provide service once one of the servers can be powered on. This can significantly improve utilization of green power, which is also a major target of computational sprinting. As a result of this analysis, we choose the *Packing* scheme to conduct green power distribution.

4 GreenSprint under Thermal Constraint

As we mentioned above, there are two critical challenges when conducting computational sprinting in data centers. The first important one is the power limitation of the data center power infrastructures. To mitigate the side effect of the power problem, we propose relative four strategies aforementioned. These strategies are effective with the assumption that the extra heat dissipation generated by sprinting can be effectively released by emerging heat sink materials or PCM with larger thermal capacitance, which have a high but still limited thermal capacity. In real implementation, especially for future many-core processors, we must take the thermal effect into considerations for specific power management. In this section, we will first analyze the thermal impact on sprinting activities. Then we propose a new strategy Hybrid-T based on Hybrid to manage computational sprinting under a thermal constraint.

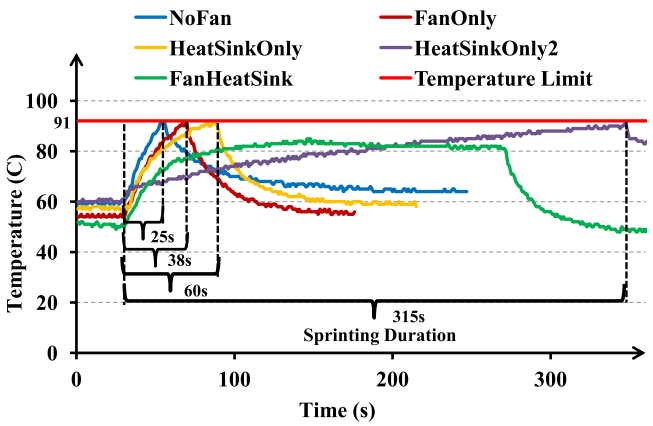


Fig. 6. The thermal impact on sprinting duration with Xeon E5-2620 processor.

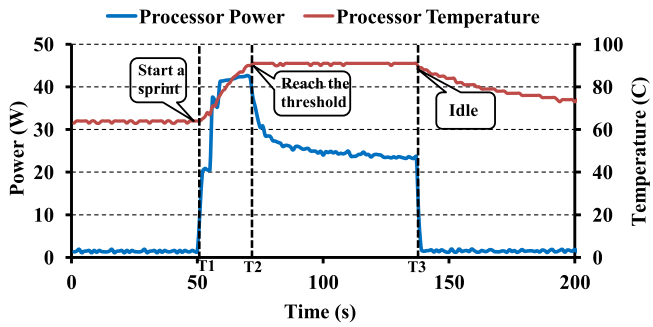


Fig. 7. The thermal impact on sprinting power.

4.1 Sprinting Duration

The most important impact by thermal constraint is sprinting duration that the server can support. In other words, the capability of thermal dissipation directly determines how long the sprinting can sustain before the chip temperature reaches an upper threshold. To analyze the impact of thermal constraint, we conduct a series of experiments on Xeon E5-2620 processor with six CPU cores. Its maximal sprinting frequency level is 2.0 GHz. We use *turbostat* command in Linux to collect the temperature and power consumption of the processor. Fig. 6 presents the chip temperature results under different cooling conditions running SPECjbb workload. We turn on all six cores on the processor, denoted as sprinting mode, and collect the temperature data. In this figure, NoFan represents the case where the chip has no additional cooling materials, such as fan and heat sink. For FanOnly and HeatSinkOnly, we place a fan and a heat sink separately above the chip to provide additional cooling capacity. HeatSinkOnly2 uses a different heat sink from HeatSinkOnly and shows a larger thermal capacity compared with HeatSinkOnly. Finally, FanHeatSink uses both a heat sink and a speed-variable fan, which can create different cooling conditions by adjusting fan speed. In our platform, once the temperature reaches the measured junction point of 91°C, the chip has to terminate the sprinting activity and decreases the frequency level continuously until the temperature returns to normal. As Fig. 6 shows, the sprinting durations for NoFan, FanOnly, HeatSinkOnly, and HeatSinkOnly2 are 25s, 38s, 60s, and 315s respectively. FanHeatSink in this figure presents a sustainable sprinting and never touches the temperature limit when we set the speed of the fan to the maximum. Note that, different processors have different default junction points, which can be obtained by *sensors* command. In this experiment, the junction point for Xeon E5-2620 processor is 91°C. We also conduct the same experiment on a server platform equipped with Intel Core i5-4460 processor, which has four CPU cores and a maximal frequency level of 3.2 GHz. The results show that this processor has a junction point of 100°C and its sprinting duration varies from 24s to 230s. As a result, we can conclude that different cooling conditions can significantly affect the sprinting durations, and further performance. If we define the sprinting duration as D, the cooling capacity of each server directly determines the fixed value of D during sprinting. Note that, the cumulative performance of sprinting is equal to *Throughput* * D, where *Throughput* represents burst intensity.

4.2 Sprinting Power

Fig. 7 presents the processor power and temperature results of a thermal unaware sprinting using Xeon E5-2620 processor. That is, the strategy does not have the knowledge of the temperature data, and take no actions to the temperature fluctuation. In this figure, the whole sprint starts from T1 and ends at T3. We can see that, after the sprint makes the temperature rise to the rated junction point rapidly at T2, the power of processor starts to decrease by degrees because the operating system forces the core to lower frequency level. This observation is important because, for renewable energy driven power management, once the power demand fluctuates during a scheduling epoch, the strategy can hardly determine appropriate power supply and a fixed sprinting intensity. What is worse, the whole server could become unstable, suddenly shut down, or even suffer component damage if the processor keeps overheating for a long time (e.g., the duration from T2 to T3 in Fig. 7) [7]. In this work, we mainly focus on the thermal constraint on the processors due to the *dark silicon* phenomenon and neglect the thermal impact of memory units. Actually, a prior research on DRAM thermal obtains an average 4.1 percent performance gain by reducing about 5.36°C of the DRAM chip temperature [29]. Therefore, the thermal impact of memory units on performance can be negligible compared with the performance gain from computational sprinting.

4.3 Thermal Aware Power Management

Based on the above analysis, we modify the *Hybrid* strategy and propose a new thermal aware strategy called *Hybrid-T*. The main target of this strategy is to maximize the sprinting performance under the thermal constraint with the green power supply. First of all, the Monitor in Fig. 3 should collect the temperature data of server processors continuously. The sampling interval of the temperature (e.g., 5 seconds) will be more fine-grained than the scheduling epoch in order to detect the thermal warning when the temperature reaches the junction point, and to prevent the processors from long overheating. The key part of *Hybrid-T* is to choose a proper sprinting intensity just like *Hybrid*. In order to reflect the thermal impact in our decision-making process, we modify the *Hybrid* strategy based on the following concerns.

First, the temperature state will be added into the set c_t. When *Hybrid-T* uses the lookup table to select the best action based on current state, the temperature will be considered. Specifically, the range of temperature will also be divided into a discrete set as the power supply and the workload intensity. We adopt a *conservative policy* here to control the temperature. That is, we set an upper threshold and spare about 5°C he'adroom between the junction point and this threshold to keep the processor running safely. The rationale of such design is to trigger the control operation before the measured temperature reaches the warning threshold. For example, we set 86°C as the upper threshold if the junction point of the processor is actually 91°C (e.g., Xeon E5-2620 processor). This control method can be found in many feedback systems in datacenters [30], [41].

Second, the reward mechanism will be revised. We introduce two new rewards into the total reward calculation as shown in Algorithm 2. They are temperature reward R_{temp} and cumulative work reward R_{cwork}. R_{cwork} is a positive reward, which can represent the overall performance improvement by sprinting activities during a scheduling epoch. If R_{temp} is larger than one, then the temperature of processor is under a safety threshold, which is appreciated for sprinting. However, when R_{temp} is less that one, it means that the temperature reaches the warning area. Therefore, we add a negative value into the total reward.

Algorithm 2. Reward Mechanism

1: // Calculate reward r_t based on epoch t and $t+1$
2: QoS_{target} and $QoS_{current}$ represent the target QoS of the workload and the current latency result. $PowerSupp$ and $PowerCurr$ are the power supply and the current power demand at time t.
3: R_{power} = $PowerSupp$ / $PowerCurr$ //Power reward
4: R_{qos} = QoS_{target} / $QoS_{current}$ //QoS reward
5: **If** $R_{power} > 1$ **Then**
6: **If** $R_{qos} > 1$ **Then**
7: r_t = R_{power}+R_{qos}+1
8: **Else**
9: r_t = R_{power}-R_{qos}+1
10: **EndIf**
11: **Else**
12: r_t = -R_{power}-1
13: **EndIf**
14: $Temp_{threshold}$ and $Temp_{current}$ represent the temperature threshold and the current temperature. $CWork_{processed}$ and $CL_{requested}$ represent the cumulative work processed and requested separately.
15: R_{temp} = $Temp_{threshold}$ / $Temp_{current}$ //Temperature reward
16: R_{cwork} = $CWork_{processed}$ / $CL_{requested}$ //Cumulative work reward
17: **If** $R_{temp} > 1$ **Then**
18: $r_t = r_t + R_{cwork}+R_{temp}$
19: **Else**
20: $r_t = r_t + R_{cwork}-R_{temp}$
21: **EndIf**
22: // Update the value of $R(c_t, a_t)$ in lookup table
23: $R(c_t, a_t)=R(c_t, a_t) + \alpha[r_t + \gamma max_{a_i \in S} R(c_{t+1}, a_i) - R(c_t, a_t)]$

Third, the scheduling epoch should be determined according to the sprinting duration D. When the processor can only sustain a sprinting less than five minutes, which is the default duration of each epoch, shorter sampling interval can help the system detect the condition of overheating. Further, the duration D is also determined by the heat dissipation capacity, such as additional fan, heat sink, and new phase changing materials.

Fourth, the last resort for overheating is terminating the sprinting activity. However, different from thermal unaware strategy, *Hybrid-T* will restart the sprinting once the signal of a normal temperature is detected. The duration used for cooling is called *Cooling Time*. Obviously, when the cooling capacity is larger, the heat can be dissipated more quickly resulting in a shorter *Cooling Time* according to Fig. 6.

TABLE 1
Options for Green Provision

Configurations	RE	Batt. (Server level)
RE-Batt	30% servers	10Ah
REOnly	30% servers	0
RE-SBatt	30% servers	3.2Ah
SRE-SBatt	20% servers	3.2Ah

5 Prototype Evaluation

Our scale-down experimental prototype of GreenSprint uses a cluster of $N = 10$ servers each with two 6-core 2.0 GHz Intel Xeon E5-2620 processors (i.e., 12 cores per server), 48 GB RAM and 1 Gbps Ethernet interface and run our applications hosted on the Ubuntu Linux OS. The cluster has an NFS storage volume shared by all the servers. The power consumption of each server is monitored by an external power meter [2]. Their idle power is around 76W. The dynamic power consumption can be modulated with 9 frequency states and sprinting scales the core count from 6 to 12. The temperature of processor is measured by *turbostat* in Linux. Also, *sensors* command can be used to monitor the temperature data. We choose *turbostat* because it can provide more specific data, such as processor power, temperature, and frequency level.

To simulate a data center with renewable energy provision, we randomly choose one of the renewable power production traces with one-week duration from NREL [6], including irradiation every minute, and replay the chosen trace on our prototype. We scale the solar power production to correspond the power source configuration (Table 1) to simulate the available renewable power output. In this table, for example, 'RE' represents renewable energy provision. 'S' represents *small* renewable energy and battery energy capacity provisions. In our setup, we consider a solar panel provisioned for a server j with 275W DC output (theoretical peak power), which is in line with the existing capacities in Grapesolar [4]. Hence, we can obtain the peak renewable power AC supply for a single solar panel that generates $Peak_{RE} * \alpha = 275 * 0.77 = 211.75W$. As shown in Fig. 8, for the *RE-Batt* configuration, we assume that 3 servers in our prototype are provided with a renewable energy system that is capable of supplying the maximum green power of 635.25 W. For the configuration with *SRE* that provides 3 servers with smaller renewable power supply, the maximum green power obtainable is 423.5W. We assume that each server in the cluster is equipped with a battery unit and the battery energy capacity

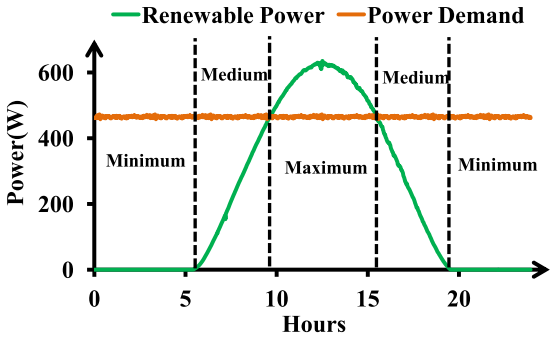


Fig. 8. The SPECjbb power profile as a function of the renewable energy availability over time.

is shown in Table 1. We use *cpufreq* to scale frequency and *taskset* to redirect workload threads to right cores.

Workloads and Strategies. We consider the following representative data center workloads (Table 2) that exhibit different performance characteristics with different peak power demands on renewable energy and batteries. These interactive applications are SPECjbb [5], an in-memory key-value store Memcached benchmark, Web-search from Cloudsuite [14], and Mcf from SPECCPU2006 [1]. We measure the maximal sprinting power demand of each application, yielding 155W for SPECjbb, 156W for Web-Search, 146W for Memcached, and 176W for Mcf. We also evaluate five strategies for comparison. They are *Normal*, *Greedy*, *Parallel*, *Pacing*, and *Hybrid*.

5.1 Performance of GreenSprint

We now compare the performance and power impact of GreenSprint against the baseline (i.e., without renewable energy). We first show representative results for SPECjbb. We generate the workload in the cluster until all 10 servers are fully utilized to produce a workload burst. As a result, the aggregate power draw of these servers can exceed the grid power budget. We statically set the sprinting to the highest intensity. For instance, when the workload saturates all 10 servers with 12 active cores, the aggregate power consumption hits 1550W. If the grid power infrastructure can support 10 servers to operate at *Normal* mode, then the power budget of the grid can be 1000W. From the renewable energy side, if renewable energy can supply 3 servers in the cluster (i.e., RE-Batt configuration), then the grid can conservatively support the other 7 servers sprinting at sub-optimal performance (e.g., 12 core-sprinting with 1.5GHz or 7 core-sprinting with 2GHz). As specified above, we inject the workloads to deliberately induce power burst durations of 10, 15, 30 and 60 minutes. We use the average throughput (jops) of whole cluster as our performance metric for SPECjbb. To find out the impact of renewable power supply, we mainly focus on the analysis of green-provisioned servers.

TABLE 2
Workload Description

Workloads	Memory Usage	Performance Metric
SPECjbb	10 GB	jops (99%-ile 500ms constrained)
Web-search	20 GB	ops (90%-ile 500ms constrained)
Memcached	20 GB	rps (95%-ile 10ms constrained)
Mcf	24 GB	ips (Completion time)

Fig. 8 shows the evolution of the aggregated peak power of the 3 green-provisioned servers running SPECjbb at given different levels of renewable energy *availability*. We see high variation of the renewable power production over time. We have evaluated such performance consequences for all the cases of *medium* availability over different power shortage durations. Moreover, we consider the *minimum* availability case for comparison where the sprinting goal can only be achieved by the batteries.

Impact of Renewable Energy Availability and Burst Duration. Fig. 9 presents the average performance of SPECjbb under different renewable energy availability and power burst durations using the RE-Batt configuration. As shown in this figure, for the *maximum* availability of renewable energy, three servers in the cluster can be directly powered by renewable energy with full-sprinting and the performance is always the best with 4.8x gains over *Normal*. Further, the surplus green power can be used to charge the battery for later use.

In the case of minimum and medium availability levels of renewable energy, the performance varies with different lengths of burst duration. For short bursts (10 minute duration), even when the renewable energy is unavailable, battery alone is able to completely handle the sprinting operation with maximal performance. For the durations of 15, 30, and 60 minutes, the performance varies significantly for different strategies. The performance improvement decreases relatively for longer burst durations, especially for the *minimum* availability (60 minute), in which the performance improvement drops to 1.8x for *Parallel*. Comparing with the 4.8x improvement with sufficient renewable power supply, battery-based sprinting is unsatisfactory. However, for *medium* availability, battery can supplement the green power to sustain the sprinting performance. For 60 minute durations, Sprinting can still provide up to 3.4x performance gains over *Normal*.

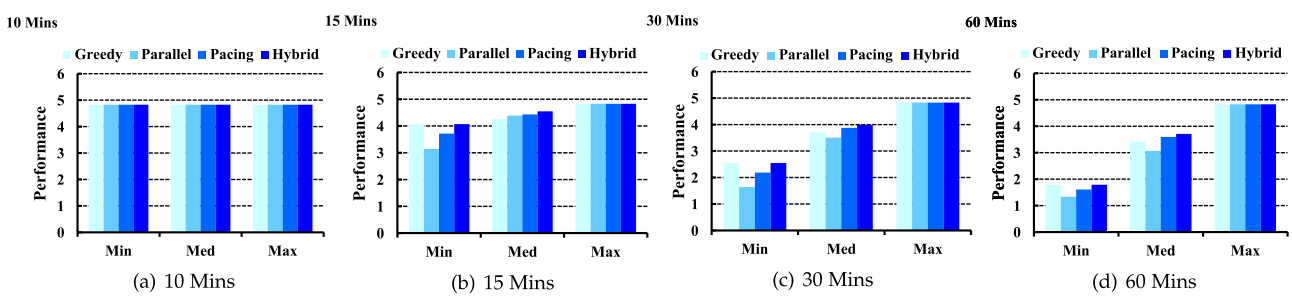


Fig. 9. Performance of GreenSprint with varying renewable energy availability and burst durations for SPECjbb using RE-Batt, normalized to *Normal*.

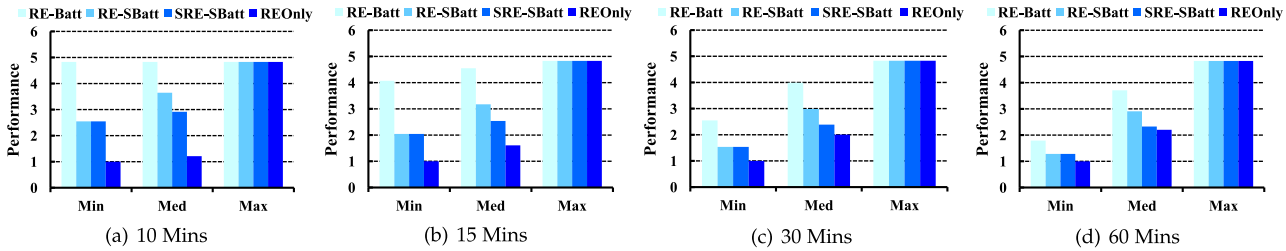


Fig. 10. Performance of GreenSprint for SPECjbb using different power configurations, normalized to *Normal*.

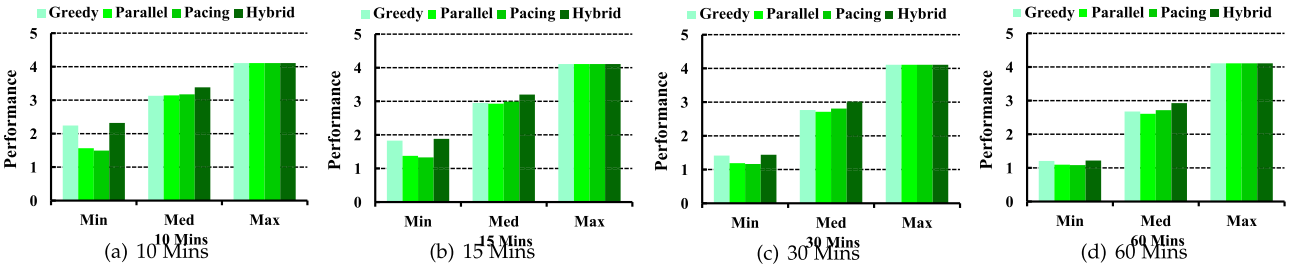


Fig. 11. Performance of GreenSprint for Web-Search with RE-SBatt, normalized to *Normal*.

Impact of Power Management Knobs. We adopt two techniques in power management knobs, scaling core counts (*Parallel*) and scaling frequency (*Pacing*). In these figures, *Pacing* slight outperforms *Parallel* in all cases. We attribute these results to higher energy efficiency of frequency scaling. Also, it demonstrates that decreasing the number of cores can still influence the performance after mitigating the oversubscription on cores. For the *Greedy* strategy with battery-based power supply, the system achieve the same results as *Hybrid* that always performs the best. This indicates that sprinting as much as possible for SPECjbb using battery receives much better energy efficiency. However, due to the higher start point of power needed to wake up servers, *Greedy* underperforms *Pacing* because it loses the opportunity to utilize the lower green power supply periods. *Hybrid* always performs the best because it always learns the optimal combinations of scaling core count and frequency for better performance.

5.2 Impact of Green Configurations

We also evaluate the other three green configurations (RE-SBatt, SRE-SBatt, and REOnly) using SPECjbb. We only show the result of the *Hybrid* strategy to examine the differences among these configurations in Fig. 10.

Impact of Renewable Energy. Configurations with RE-SBatt and SRE-SBatt show the difference of renewable power supply. When we use smaller green power, the performance degrades accordingly. However, powering two or three servers with green energy has a great effect on the cap-ex cost. Maximal performance can always be achieved during maximum green power supply. In the REOnly configuration, the performance results with minimum renewable energy availability are the same as the *Normal* mode because there is no power supply for sprinting, when all servers return to the *Normal* mode powered by the grid utility. With only renewable energy supply, GreenSprint significantly improves performance, from 2.2x (medium availability) to 4.8x (maximum availability) for the 60 minute long power burst.

Impact of Battery. Given the minimum renewable energy availability, in configurations with the battery (RE-SBatt and RE-Batt) and with the minimum renewable energy availability, performance impact for the REOnly configuration can be reduced since the battery can supply additional power. Therefore, battery is preferred as a complement to deal with bursty power demand when there is no green power supply. We notice that configurations with small battery capacity (RE-SBatt and SRE-SBatt) show less improvement than the REOnly configuration when sprinting lasts for 60 minutes or longer, because the battery is not able to sustain such long operations for the entire sprinting duration. For different capacity of battery, RE-Batt (10Ah) performs better than RE-SBatt (3.2Ah) for the minimum and medium green power availability and can sustain more than 10 minutes at the maximal power burst. This implies that we can significantly improve the performance, irrespective of renewable energy availability, by purchasing a larger capacity of battery.

5.3 Impact of Application Characteristics

We also evaluate another two applications, Web-Search and Memcached under the RE-SBatt configuration.

Web-Search. Web-Search is the query serving portion of a production web search service and has high memory footprint. It is fairly compute intensive due to scoring and sorting search hits. In Fig. 11, *Pacing* shows no more benefits than *Parallel*, relative to SPECjbb, and similar performance under varied conditions. Especially, when the green energy availability is *minimum*, lowering core count from 12 to 6 is slightly better in performance than decreasing frequency. Therefore, scaling core count can be a better choice in a battery-based power system for Web-Search because lowering chip frequency has a great impact on throughput. For longer durations, battery-based sprinting can barely achieve performance improvement over the *Normal* mode.

When the renewable energy supply is sufficient, GreenSprint can achieve 4.1x performance gain over the baseline. In face of reduction in green power, batteries can help sustain

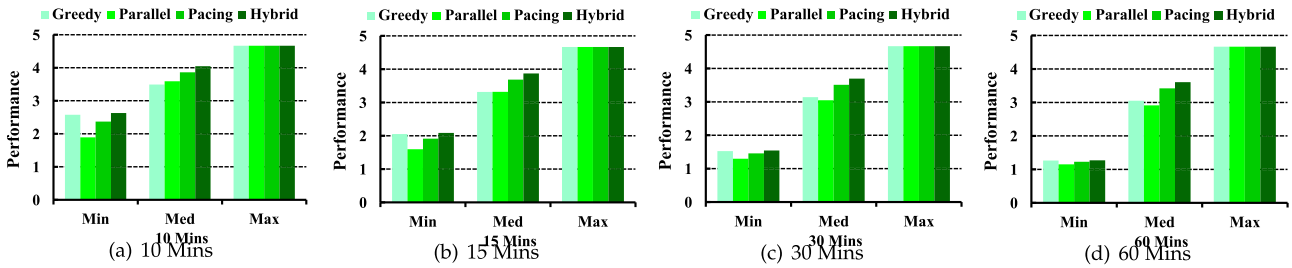


Fig. 12. Performance of GreenSprint for Memcached with RE-SBatt, normalized to *Normal*.

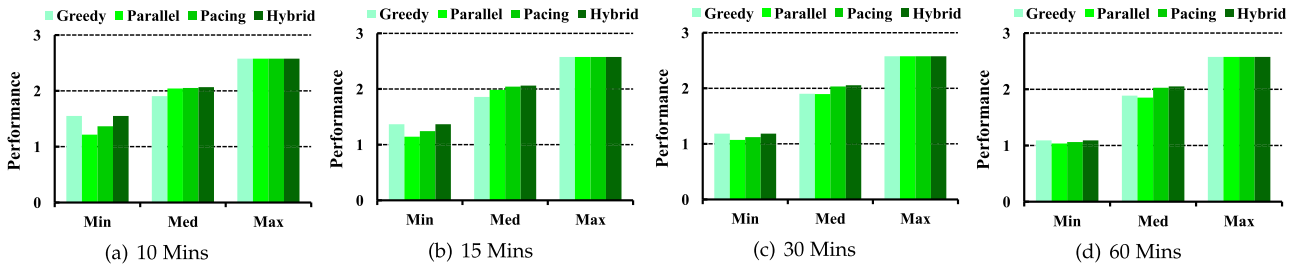


Fig. 13. Performance of GreenSprint for *MCF* with RE-SBatt, normalized to *Normal*.

sprinting. In the *medium* case, although scaling frequency can significantly affect the performance of Web-Search, it achieves higher energy efficiency than scaling core counts, resulting in better performance. Different from constrained energy capacity of a battery system, the green power is time-varying, resulting in dynamically adjusted server power demand.

Memcached. Memcached is used as a caching service in the back-ends and loads the memory with the necessary data from disk before handling client requests. In Fig. 12, the maximal performance improvement for Memcached is 4.7x, similar to Web-Search. For the *medium* and *maximum* green supply, the results show a similar trend to SPECjbb. *Pacing* performs better under different cases because of the characteristics of Memcached, i.e., less computation intensive and need more on parallelism. *Greedy* is no more beneficial for both Web-Search and Memcached under battery-based supply because the optimal energy efficiency points are not always achieved with the maximal sprinting intensity.

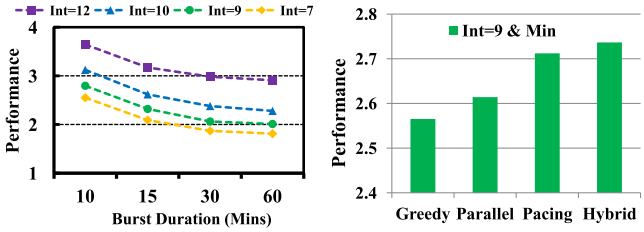
MCF. *MCF* from SPECCPU2006 is represented as a memory intensive scientific computation workload. Since each MCF instance only consumes 2 GB memory, we execute 12 MCF instances to increase its memory usage (a total of 24 GB) to emulate power emergencies of HPC applications. Note that, MCF has no strict latency constraint, so we use the metric *ips* to guide the scheduler to manage service performance. We collect the *ips* using performance monitor *perf* in Linux. To this end, when we use *Hybrid* strategy to evaluate this workload, we calculate the performance reward using *ips*. Fig. 13 shows the results of MCF. Compared with other three interactive workloads, the performance improvement of MCF is smaller, i.e., 2.6x under the best case. We contribute such difference to that the scientific computation workload is less dependent on computation parallelism than those interactive workloads. For all power supply conditions, certainly, the performance for *maximum* green power availability performs the best. When the green power supply is insufficient, battery is still a good supplement that can maintain the performance improvement for up to 2x.

5.4 Impact of Workload Burst Intensity

To evaluate the performance of GreenSprint for different burst intensities, we generate the workload of SPECjbb by several other burst patterns to draw power bursts. In Fig. 14, for example, 'Int=9' indicates the case that the peak burst load is the maximal processing capability of running workloads on 9 cores at 2.0 GHz. Fig. 14a shows the case of *RE-SBatt* configuration and *medium* availability with *Hybrid*. According to the results, the performance is much lower (from 3.6x to 2.6x) when the burst intensity decreases (from Int=12 to Int=7) for different burst durations. Obviously, sprinting may lose the advantage on performance when burst intensity is low. In the case 'Int=7', GreenSprint can only provide 2.6x-1.7x improvement with the duration from 10 minutes to 60 minutes compared with *Normal*. Fig. 14b presents the performance of four strategies with 'Int=9' and *minimum* availability. The duration is 10 minutes. In this case, *Greedy* performs the worst because, when the burst intensity becomes lower, maximal sprinting on 12 cores is less efficient than other strategies. Lower sprinting intensity, though higher response time under QoS constraint, can extend the discharging time of battery to achieve better overall throughput.

5.5 Impact of Thermal Constraint

In this section, we evaluate the impact of thermal constraint using SPECjbb. In order to show the effectiveness of *Hybrid-T*, we also present the results of both thermal unaware strategy *Hybrid* for comparison. In Hybrid, once the temperature of processors reaches the junction threshold, the sprinting activities will be terminated immediately to prevent the component from overheating. To build the configurations of different cooling capacity, we use a speed-variable fan and a heat sink for each processor. We adjust the speed of fan to create several cases that the sprinting can sustain from 1 minute to 15 minutes (e.g., 1 min, 2 min, 5 min, 10 min and 15 min). Note that, such experimental design can also simulate the cooling conditions and obtain the same effect of deploying real PCM.



(a) Impact on burst durations (b) Impact on strategies

Fig. 14. Performance impact of workload burst intensity, normalized to *Normal*.

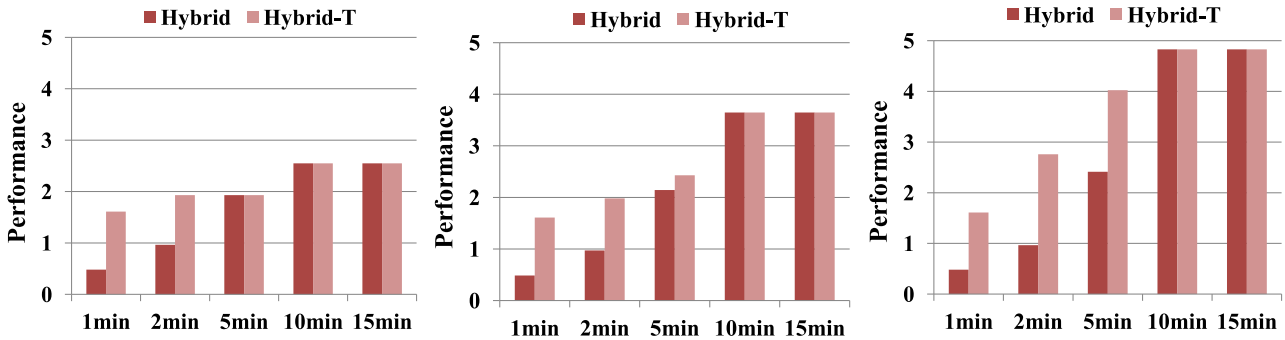
Fig. 15 shows the performance results for three different availabilities of renewable power supply under different heat dissipation capabilities. We emphasize on the burst duration of 10 minutes. In general, thermal aware strategy *Hybrid-T* performs better than thermal unaware strategy *Hybrid*. Specifically, first, when the cooling capacity can only sustain 1 minute and two minutes for all renewable power supply conditions, *Hybrid* underperforms even than *Normal*. For maximum power availability, the performance gain stays no more at the highest level. We attribute these negative effect on performance to the thermal constraint. That is, when the temperature of processors reaches the junction threshold, the sprinting activity is terminated simultaneously. Second, when the cooling capacity increases (e.g., 5 minutes), the overall performance increases too. Suffered by the limited battery capacity, *Hybrid* and *Hybrid-T* present similar performance results when the availability is minimum. Third, when the cooling capacity can sustain longer duration (e.g., 10 minutes and 15 minutes) than the burst duration (e.g., 10 minutes), the performance gain of *Hybrid* and *Hybrid-T* show no significant difference. Finally, both strategies perform better when the renewable power supply rises.

5.6 Summary of Observations

In summary, we find that: (1) Sprinting can significantly improve performance by activating more cores. (2) Despite of the intermittent nature, renewable energy can effectively support computational sprinting in power constrained data centers. (3) Batteries alone can achieve performance improvement for short sprinting durations. However, batteries are not appropriate for longer durations. (4) Renewable energy can supplement battery to reduce performance impact. (5) Compared with scaling core count for interactive applications, scaling frequency is more energy efficient in utilizing the battery energy. In other words, those interactive applications have high expectations of parallelism, so scaling core count is not the best choice to get better energy efficiency. (6) Sprinting in turn can increase the renewable power utilization due to higher power demand and energy efficiency. (7) Thermal aware strategy can significantly mitigate performance loss due to the limitation on cooling capacity.

5.7 TCO Consideration

Although we have considered sprinting activities with the support of renewable energy and battery that already exist in green data centers, we will illustrate whether the cost of additional green provision can be justified. Thus, we consider the revenue increase due to sprinting for two reasons. First, sprinting raises the power demand with renewable or battery energy. Second, performance improvement by sprinting can provide additional revenue for a cloud provider such as Google, where a large majority of its revenue comes from its data center operations. Conservatively, we assume its revenue is \$0.28/KW/min due to operations [37]. However, due to the implementation of renewable energy (PV panels in our design), we estimate the green-power capacity to be \$4.74/W [13]. In addition, we ensure that PV panels are amortized over 25 years (lifetime). We assume the cost of batteries to be \$50/KW/year [37]. Note that, the cost is amortized over a battery lifetime of 4 years. For the lifetime concern, first, we set the depth of discharge of the battery to be 40 percent to prolong the battery lifetime as aforementioned in Section 2. In other words, uncontrolled battery discharging activity can drain out the battery capacity easily, leading to less cycles of the battery. Second, GreenSprint leverages the renewable energy to supply the bursty power in the first priority, which reduces the battery usage compared with the power system only depended on battery energy. Therefore, GreenSprint will not specifically shorten the battery lifetime and increase additional battery cost. In addition, we include the cost of adding the wax (PCM) into the server cost, although the additional cost is almost negligible representing less than 0.1 percent of the server cost [36]. Subtracting the capital expenditure of green power and battery, we can then capture the revenue by sprinting as a function of total sprinting durations in a year. As shown in Fig. 16, all values to the right of the cross-over point (around 14 hours per year in this case) indicate profitable operations despite of the additional capital cost. This



(a) Minimum Availability (b) Medium Availability (c) Maximum Availability

Fig. 15. Analysis of thermal impact on performance for *SPECjbb* with RE-SBatt. The results are normalized to *Normal*.

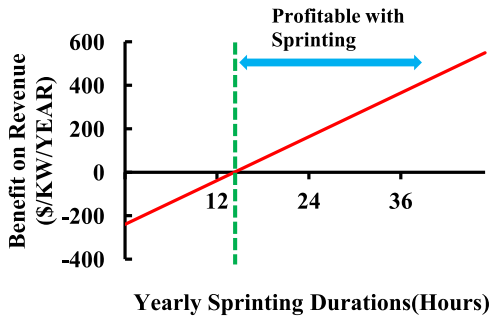


Fig. 16. POI with additional renewable energy, battery and cooling investment.

suggests that investing in additional green provision may be worthwhile when conducting as much computational sprinting in data centers as possible.

6 Related Work

Computational Sprinting. Although transistor density continues to increase, the chip cannot sustainably power up all cores due to thermal constraints. As a result, much of a chip must be powered off at any time [10]. While the cores of a chip cannot be all turned on in a sustainable way, recent studies have proposed a concept called Computational Sprinting [33], [35]. It allows a chip to temporarily exceed its power and thermal limits by activating additional cores and boost their voltage and frequency for a short time period and then immediately returns to normal operation to cool down. Prior work proposed a three-phase methodology that enables safe computational sprinting at data center level which combines data center circuit breakers, energy storage devices and thermal energy storage technologies to control sprinting activities [43]. However, it is no doubt that simply relying on batteries to supply additional power demand cannot fully exploit the potential of sprinting due to strict lifetime constraints. More importantly, the associated carbon footprint expansion still poses significant challenges for large data centers sprinting.

Data Center Power Provisioning. With the rapid adoption of cloud computing and tremendous data sets flushing into today's data center, the computing resources are increasing continually to their existing sites to support services(i.e., scaling out), which leads data centers to be power-constrained [22]. That is, today's data centers are already approaching the peak capacity of their power infrastructures. Li et al. proposed a real solar-power server-level prototype called Oasis [22] to scale out data center power capacity economically and sustainably. Further, they also proposed a framework called GreenWorks that considers renewable power variability, battery behavior, and over-clocking (e.g., Intel's Turbo Boost) [24]. Our work differs from Oasis and GreenWorks in the following three aspects: (1) GreenSprint exploits the sprinting technique to increase the processing capacity in a short period and leverages green power to deal with the power burst due to the sprinting. Oasis focuses on incremental solar power integration in case of the performance degradation. GreenWorks emphasizes on the hybrid renewable energy system and the solution to gracefully handle the power mismatch between intermittent renewable power supply and fluctuating server power demand. We emphasize the role of performance sprinting. (2) GreenSprint pays more attention to the power and performance impact of both core count and frequency scaling techniques. Oasis and GreenWorks dynamically adjusts its load processing speed in response to the energy supply only with frequency scaling. GreenSprint takes a wider spectrum of techniques into considerations. (3) Green energy and battery in GreenSprint are called upon only for occasional power bursts. In contrast, Oasis and GreenWorks have more frequent green and battery energy demands due to daily power fluctuations. (4) Compared with Oasis and GreenWorks, GreenSprint also considers the thermal constraint on servers and proposes a thermal aware power management.

The high cost of power provisioning and consumption in data centers draws attentions to underprovision the power infrastructure [9], [12], [17], [21], [37], [42], [45]. Different from power capping work, first, data center sprinting with renewable energy can provide more power to servers instead of throttling their power when they need it most. Second, sprinting in green data center is designed for temporarily boosting the computational capacity to achieve better performance, while existing works on data center provisioning are trying to save capital or operating expenses. Also, we must consider the intermittent nature of green power and energy storage device simultaneously. This characteristic of the green mechanism calls for power management schemes to handle the varied power supply and limited energy storage capacity. Green energy as a power source, as opposed to single energy storage supply, has the potential to provide a long-term power supply for data center sprinting.

Scheduling for Green Energy. While some researchers have studied the scheduling of batch jobs to maximize the use of renewable energy [15], [16], others have proposed to adapt the amount of batch processing dynamically to a mix of interactive and batch workloads in data centers [8]. GreenPar [19] schedules parallel high-performance applications in green data centers to maximize the green energy consumption and minimize the brown energy consumption, while respecting a performance service-level agreement. SolarCore [26] leverages per-core DVFS on multi-core systems to temporarily lower server power demand when solar power drops. Li et al. [23] proposed an architecture in which two sets of servers draw power respectively from two different sources (e.g., the grid or a wind farm). In their setup, energy source management entails migrating workloads between the two sets. In [27], Li et al. investigated the benefits of the load following mechanism in distributed generation powered data centers and tailor data center power demand for improving renewable energy utilization. Zhou et al. [44] proposed heterogeneous server system to improve energy efficiency for green data centers.

7 Conclusion

In this paper, we propose GreenSprint, a renewable energy driven framework that enables data center sprinting to handle occasional workload bursts. We present four strategies to address the challenge imposed by the intermittent and time-varying renewable power supply. We develop an experimental prototype to evaluate our approach. Using

representative data center workloads, the results show that our solution can improve the average computing performance significantly by a factor of 4.8x for SPECjbb, 4.1x for Web-Search, and 4.7x for Memcached with sufficient renewable power supply.

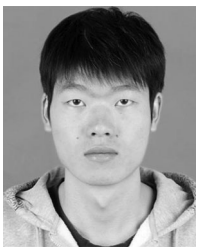
Acknowledgments

This work is supported in part by Nature Science Foundation of China under Grant No. 61872156 and No.61821003 , the Fundamental Research Funds for the Central Universities No. 2018KFYXKJC037, the US NSF under Grant No.CCF-1704504 and No.CCF-1629625, and Alibaba Group through Alibaba Innovative Research (AIR) Program. Qiang Cao is the corresponding author of this paper.

References

[1] SPECCPU. 2016. [Online]. Available: http://www.spec.org/cpu2006/
[2] Zhongyuanhuadian, Zh-101 portable electric power fault recorder and analyzer, 2009.
[3] Facebook, "Hacking conventional computing infrastructure," 2011. [Online]. Available: http://opencompute.org/
[4] Grapesolar, 2014. [Online]. Available: www.grapesolar.com/
[5] SPECJBB 2013:Java Business Benchmark, [2014]. [Online]. Available: http://www.spec.org/jbb2013/
[6] Measurement and instrumentation data center, 2015. [Online]. Available: http://www.nrel.gov/midc/
[7] Intel, "Computer is overheating warning signs," 2018. [Online]. Available: https://www.intel.com/content/www/us/en/support/articles/000005791/processors/intel-core-processors.html
[8] B. Aksanli, J. Venkatesh, L. Zhang, and T. Rosing, "Utilizing green energy prediction to schedule mixed batch and service jobs in data centers," in *Proc. 4th Workshop Power-Aware Comput. Syst.*, 2011, Art. no. 5.
[9] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Comput.*, vol. 40, no. 12, pp. 33–37, 2007.
[10] H. Esmaeilzadeh, E. Blem, R. St Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. 38th Annu. Int. Symp. Comput. Archit.*, 2011, pp. 365–376.
[11] S. Fan, S. M. Zahedi, and B. C. Lee, "The computational sprinting game," in *Proc. 21st Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2016, pp. 561–575.
[12] X. Fan, W. D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *Proc. Proc. 34th Annu. Int. Symp. Comput. Archit.*, 2007, pp. 13–23.
[13] D. Feldman, "Photovoltaic (PV) pricing trends: Historical, recent, and near-term projections," Joint Tech. Rep. DOE/GO-102012–3839, Nat. Renewable Energy Lab./Lawrence Berkeley Nat. Lab., Golden, CO, USA/Berkeley, CA, USA, 2012.
[14] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi, "Clearing the clouds: A study of emerging scale-out workloads on modern hardware," in *Proc. 17th Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2012, pp. 37–48.
[15] Í. Goiri, K. Le, M. E Haque, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenslot: scheduling energy consumption in green datacenters," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, 2011, pp. 1–11.
[16] Í. Goiri, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenhadoop: Leveraging green energy in data-processing frameworks," in *Proc. 7th ACM Eur. Conf. Comput. Syst.*, 2012, pp. 57–70.
[17] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and limitations of tapping into stored energy for datacenters," in *Proc. 38th Annu. Int. Symp. Comput. Archit.*, 2011, pp. 341–352.
[18] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters," in *Proc. 17th Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2012, pp. 75–86.
[19] M. E Haque, I. Goiri, R. Bianchini, and T. D. Nguyen, "Greenpar: Scheduling parallel high performance applications in green datacenters," in *Proc. 29th ACM Int. Conf. Supercomputing*, 2015, pp. 217–227.
[20] T. W. Keller, C. Lefurgy, M. Chen, and X. Wang, "Ship: A scalable hierarchical power control architecture for large-scale data centers," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 1, pp. 168–176, Jan. 2012.
[21] V. Kontorinis, L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen, and T. S. Rosing, "Managing distributed ups energy for effective power capping in data centers," in *Proc. 39th Annu. Int. Symp. Comput. Archit.*, 2012, pp. 488–499.
[22] C. Li, Y. Hu, R. Zhou, M. Liu, L. Liu, J. Yuan, and T. Li, "Enabling datacenter servers to scale out economically and sustainably," in *Proc. 46th Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2013, pp. 322–333.
[23] C. Li, A. Qouneh, and T. Li, "iSwitch: Coordinating and optimizing renewable energy powered server clusters," in *Proc. 39th Annu. Int. Symp. Comput. Archit.*, 2012, pp. 512–523.
[24] C. Li, R. Wang, D. Qian, and T. Li, "Managing server clusters on renewable energy mix," *ACM Trans. Auton. Adaptive Syst.*, vol. 11, 2016, Art. no. 1.
[25] C. Li, Z. Wang, X. Hou, H. Chen, X. Liang, and M. Guo, "Power attack defense: Securing battery-backed data centers," in *ISCA*, 2016.
[26] C. Li, W. Zhang, C.-B. Cho, and T. Li, "Solarcore: Solar energy driven multi-core architecture power management," in *Proc. IEEE 17th Int. Symp. High Perform. Comput. Archit.*, 2011, pp. 205–216.
[27] C. Li, R. Zhou, and T. Li, "Enabling distributed generation powered sustainable high-performance data center," in *Proc. IEEE 19th Int. Symp. High Perform. Comput. Archit.*, 2013, pp. 35–46.
[28] L. Liu, C. Li, H. Sun, Y. Hu, J. Gu, T. Li, J. Xin, and N. Zheng, "Heb: Deploying and managing hybrid energy buffers for improving datacenter efficiency and economy," in *Proc. ACM/IEEE 42nd Annu. Int. Symp. Comput. Archit.*, 2015, pp. 463–475.
[29] S. Liu, B. Leung, A. Neckar, S. O. Memik, G. Memik, and N. Hardavellas, "Hardware/software techniques for dram thermal management," in *Proc. IEEE 17th Int. Symp. High Perform. Comput. Archit.*, 2011, pp. 515–525.
[30] D. Lo, L. Cheng, R. Govindaraju, L. A. Barroso, and C. Kozyrakis, "Towards energy proportionality for large-scale latency-critical workloads," in *Proc. 41st Annu. Int. Symp. Comput. Archit.*, 2014, pp. 301–312.
[31] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[32] R. Nishtala, P. Carpenter, V. Petrucci, and X. Martorell, "Hipster: Hybrid task manager for latency-critical cloud workloads," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit.*, 2017, pp. 409–420.
[33] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. K. Martin, "Computational sprinting on a hardware/software testbed," in *Proc. ASPLOS*, 2013, pp. 155–166.
[34] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. K. Martin, "Utilizing dark silicon to save energy with computational sprinting," *IEEE Micro*, vol. 33, no. 5, pp. 20–28, Sep./Oct. 2013.
[35] A. Raghavan, Y. Luo, A. Chandawalla, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. K. Martin, "Computational sprinting," in *Proc. HPCA*, 2012, pp. 249–260.
[36] M. Skach, M. Arora, C.-H. Hsu, Q. Li, D. Tullsen, L. Tang, and J. Mars, "Thermal time shifting: Leveraging phase change materials to reduce cooling costs in warehouse-scale computers," in *Proc. ACM/IEEE 42nd Annu. Int. Symp. Comput. Archit.*, 2015, pp. 439–449.
[37] D. Wang, S. Govindan, A. Sivasubramaniam, A. Kansal, J. Liu, and B. Khessib, "Underprovisioning backup power infrastructure for datacenters," in *Proc. 19th Int. Conf. Archit. Support Program. Lang. Operat. Syst.*, 2014, pp. 177–192.
[38] D. Wang, C. Ren, A. Sivasubramaniam, B. Urgaonkar, and H. Fathy, "Energy storage in datacenters: what, where, and how much?, in *Proc. 12th ACM SIGMETRICS/PERFORMANCE Joint Int. Conf. Meas. Model. Comput. Syst.*, 2012, pp. 187–198.
[39] G. Wang, S. Wang, B. Luo, W. Shi, Y. Zhu, W. Yang, D. Hu, L. Huang, X. Jin, and W. Xu, "Increasing large-scale data center capacity by statistical power control," in *Proc. 11th Eur. Conf. Comput. Syst.*, 2016, Art. no. 8.

[40] W. Whitted, M. Sykora, K. Krieger, B. Jai, William Hamburgen, J. Clidaras, Donald L Beaty, and G. Aigner, "Data center uninterruptible power distribution architecture, U.S. Patent 7,560,831, Jul. 14, 2009.
[41] Q. Wu, Q. Deng, L. Ganesh, C.-H. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, and Y. J. Song, "Dynamo: Facebook's data center-wide power management system," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit.*, 2016, pp. 469–480.
[42] W. Zheng, K. Ma, and X. Wang, "Exploiting thermal energy storage to reduce data center capital and operating expenses," in., 2014, pp. 132–141.
[43] W. Zheng and X. Wang, "Data center sprinting: Enabling computational sprinting at the data center level," in *Proc. IEEE 35th Int. Conf. Distrib. Comput. Syst.*, 2015, pp. 175–184.
[44] X. Zhou, H. Cai, Q. Cao, H. Jiang, L. Tian, and C. Xie, "Greengear: Leveraging and managing server heterogeneity for improving energy efficiency in green data centers," in *Proc. Int. Conf. Supercomputing*, 2016, Art. no. 12.
[45] X. Zhou, Q. Cao, H. Jiang, and C. Xie, "Underprovisioning the grid power infrastructure for green datacenters," in *Proc. 29th ACM Int. Conf. Supercomputing*, 2015, pp. 229–240.



Haoran Cai received the BS degree in applied computer science and technology from Huazhong University of Science and Technology, in 2014 . He is currently working toward the PhD degree in the Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology. His research interests include computer architecture, green data center power management, and multi-core systems.

Qiang Cao received the BS degree in applied physics from Nanjing University, in 1997, the MS degree in computer technology, and the PhD degree in computer architecture from the Huazhong University of Science and Technology in 2000 and 2003, respectively. He is currently a full professor with the Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology. His research interests include computer architecture, large scale storage systems, and performance evaluation. He is a senior member of the China Computer Federation (CCF), IEEE, and ACM.

Hong Jiang received the BSc degree in computer engineering from Huazhong University of Science and Technology, Wuhan, China, in 1982, the MASc degree in computer engineering from the University of Toronto, Toronto, Canada, in 1987, and the PhD degree in computer science from the Texas A&M University, College Station, Texas, in 1991. He is currently a chair and Wendell H. Nedderman endowed professor with Computer Science and Engineering Department, University of Texas at Arlington. Prior to joining UTA, he served as a program director with National Science Foundation (2013.1-2015.8) and he was with University of Nebraska-Lincoln since 1991, where he was Willa Cather professor of computer science and engineering. He has graduated 16 PhD students who upon their graduations either landed academic tenure-track positions in PhD-granting US institutions or were employed by major US IT corporations. His present research interests include computer architecture, computer storage systems and parallel I/O, high-performance computing, big data computing, cloud computing, performance evaluation. He recently served as an associate editor of the *IEEE Transactions on Parallel and Distributed Systems*. He has more than 300 publications in major journals and international Conferences in these areas, including the *IEEE Transactions on Parallel and Distributed Systems*, the *IEEE Transactions on Computers*, the *Proceedings of IEEE*, the *ACM Transactions on Architecture and Code Optimization*, the *ACM Transactions on Storage*, the *Journal of Parallel and Distributed Computing*, ISCA, MICRO, USENIX ATC, FAST, EUROSYS, LISA, SIGMETRICS, ICDCS, IPDPS, MIDDLEWARE, OOPLAS, ECOOP, SC, ICS, HPDC, INFOCOM, ICPP, etc., and his research has been supported by NSF, DOD, the State of Texas and the State of Nebraska, and industry. He is a fellow of the IEEE, and member of the ACM.