PreMatch: An Adaptive Cost-Effective Energy Scheduling System for Data Centers

Daping Li*, Jiguang Wan*, Nannan Zhao†, Duo Wen*, Chao Zhang*, Fei Wu* and Changsheng Xie*
*Wuhan National Laboratory for Optoelectronics, HUST, China
†Department of computer science, The Virginia Tech, USA
*{ldp,jgwan,wen-duo,zhang0chao0,wufei,cs_xie}@hust.edu.cn
†znannan1@vt.edu
corresponding author: Jiguang Wan.

***Abstract*—As data centers expand, increasingly growing traditional grid energy consumption and carbon dioxide emissions have caused considerable challenges. Therefore, many data centers have focused on renewable energy. However, such data centers fail to maintain high performance while trying to fully utilize renewable energy, as they cannot make a balance between the uncontrollable storage-based workload and variable renewable energy. This paper proposes PreMatch, a tiered caching storage system that considers both high-performance demands and renewable energy utilization. PreMatch deploys a Solid State Drive (SSD) cache and an Hard Disk Drive (HDD) cache for the disk-based massive storage system, which can provide a data transfer station while maintaining the reliability. We also design an adaptive energy scheduling scheme to make the active devices proportional to the dominant one of the green energy and workload. To make decisions in advance, we introduce Long Short-Term Memory (LSTM) neural network to forecast the information on workload and green energy. Experimental results show that the storage system using PreMatch can achieve the same performance as Workload-Driven Scheme (WDS), but consumes only half grid energy of WDS and has higher green energy utilization.**

***Index Terms*—Renewable energy, Data center, LSTM, SSD cache**

I. Introduction

With the rise of Internet and the emergence of various applications, the scale of the virtual world is expanding rapidly. And the virtual world is mainly built on massive data centers and consumes roughly 2% of the world's electricity. More seriously, the ever-increasing Internet data is driving the growth of the energy consumption of data centers rapidly. So making data centers work with cheaper and environmental-friendly renewable energy is meaningful. Using green energy can not only reduce the carbon footprint and slow down global warming, it also makes business sense. Over the last six years, the electricity cost of the wind and solar green energy came down so fast. A major new report from Bloomberg New Energy Finance [1] shows that the green energy electricity will be cheaper than most of the existing coal and gas plants in the world by 2027. In particular, Google has reached 100% renewable energy deployment in 2017 without traditional energy, consuming about 2.6 gigawatts of wind/solar energy to supply for both their data centers and offices [2].

However, not all the data center components have been prepared for the variable green energy, especially the storage system. As it consumes about 20% of the total energy [3], it is quite significant to reduce the traditional energy consumption of the storage system.

For the purpose of low traditional energy consumption, high green energy utilization and high performance, many studies have been done and can be divided into workload-driven and supply-oriented schemes.

Workload-driven schemes mainly tend to update the number of active devices according to workload variations [4, 5]. These schemes get satisfactory performance but could rarely leverage abundant green energy. Furthermore, some methods [6] use energy buffer units (such as batteries) to store the spare green energy and discharge when necessary. Thus, the green energy peak can be shifted to make up the green energy trough. However, various problems arise with the large-scale use of energy buffer units in data centers, such as environmental contamination, energy loss, and equipment cost.

Supply-oriented schemes usually update the number of active devices according to the green energy supply [4, 7–9]. These schemes delay most of the latency-insensitive tasks to match the green energy, or try to migrate workloads between devices. However, many online applications, such as web services, are latency-sensitive. The mismatch [10] between green energy supply and workload will have a significant impact on the performance, especially when the green energy trough meets the workload peak. We can use some high-performance devices as the cache of disk-based storage systems to improve the performance [11–13], but it is not easy to deal with the mismatch of green energy and workload.

Either way, effective schedulers should be aware of the variances in the workload intensity or the green energy supply. But the nonlinear workload and green energy supply are too complicated to be modeled analytically. Fortunately, one of the machine learning (ML) technologies–Long Short-Term Memory (LSTM) [14] can solve this problem. Because LSTM neural network is able to retain a lot of information from the historical data, and it has been demonstrated to be efficient in various fields that have a sequential nature [15–17].

In this paper, we propose an integrated storage system scheduling method for data centers called **PreMatch** (Prediction Match), which aims at obtaining significant grid energy saving with little performance degradation and high green

energy utilization. We use SSDs as a cache of our disk-based storage system to intercept requests and shift the storage workload, as SSDs [18] offer faster access speed and consume less energy than HDDs. Through the SSD cache, we can reduce the access to the underlying disks and offer opportunities to match the variation of green energy.

With consideration of the energy saving and storage performance, we design an adaptive cost-effective system-scheduling scheme, where the active devices will depend on the dominant one of green energy and workload dynamically. To avoid frequent fluctuations of performance, we also introduce the local and long-range variation trend. Meanwhile, to figure out the workload and green energy conditions, we employ the LSTM neural network.

The results show that the predicted data is accurate and our PreMatch with predicted information works almost as well as the PreMatch with real-world information. And the storage system using PreMatch can achieve the same performance as WDS, but consumes only half grid energy of WDS and has higher green energy utilization. Meanwhile, PreMatch works well when used in large-scale storage systems.

In summary, we make the following contributions:

- We leverage LSTM to predict both the workload and green energy information for system scheduling and introduce the local and long-range variation trend to avoid frequent fluctuations.
- We use an SSD cache to shift the workload for grid energy saving. Base on this architecture, we design an adaptive cost-effective energy scheduling scheme, which can adapt to the variation of workload and renewable energy and focus on the dominant one dynamically.
- We implemented PreMatch as a simulated system based on a block-level distributed storage called Sheepdog [19]. And the results show that PreMatch saves much grid energy with little performance degradation.

The rest of this paper is organized as follows: Section II, Section III and Section IV describe the design of PreMatch in general and in detail respectively, the implementation is introduced in Section V, and the experiments are illustrated in Section VI. Section VII discusses related work, and Section VIII gives a summary of our work.

II. Overall Design

A. Goal

PreMatch aims at maximizing the green energy utilization, minimizing the grid energy consumption and suffering little performance degradation.

To this end, the active devices are made proportional to the periodically dominant one of the green energy and workload. Thus when the workload is heavy, we can alleviate inactive device accesses by activate all the devices to avoid the latency penalty. When the workload is light, we can control the number of active devices to match the green energy variation and save grid energy.

However, even when the workload is light, latency-sensitive requests makes it hard for the storage system to match the variable green energy. To alleviate this problem, we employ an SSD cache. Through the SSD cache, we can divide the process of a request, whether latency sensitive or insensitive, into two stages: online SSD cache stage and offline underlying disk stage. Through the online stage, clients can obtain excellent performance and the peak workload can be partly shifted by the SSD cache with little energy consumption. Meanwhile, the underlying disks could be powered off to save energy during the online stage. In the other hand, the storage system can adjust the offline stage process to match the variation of green energy supply and workload, which will ultimately lead to high green energy utilization and much grid energy saving.

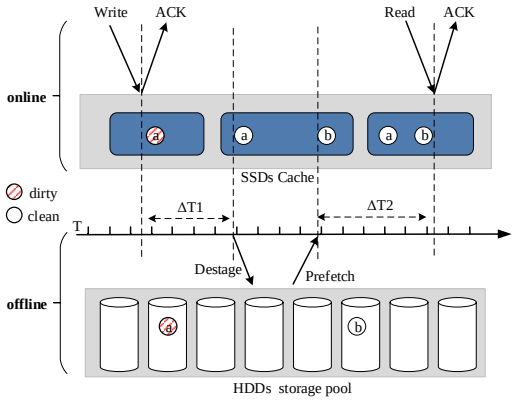

Fig. 1: The idea processes for write and read

Fig. 1 shows the idea processes for write and read requests. In Fig. 1, the write request is firstly processed by the SSD cache with an acknowledgement returned to the client, and the final offline write stage is delayed until the green energy supply is sufficient or the SSD cache is almost full. And the read request can be immediately processed by the SSD cache if the required data has been fetched from a underlying disk in advance.

Meanwhile, the specific green energy and workload conditions have significant effects on the total performance and grid energy saving. To make the storage system scheduling cost-effective, we employ LSTM to predict the future workload and green energy information. With the future information, we can schedule the storage system in advance and provide a good tradeoff between performance and energy consumption.

B. Architecture of PreMatch

As shown in Fig. 2, PreMatch comprises three parts: key information prediction, system scheduling, and storage architecture.

The key information prediction part trains the LSTM models by the historical data of workload and green energy, and utilizes the trained network models to predict the future information. System scheduling part mainly comprises the power control scheme and disk selection method. Combining the information from key information prediction and storage architecture parts, system scheduling part will figure out the proper number of active disks and decide the disk selection

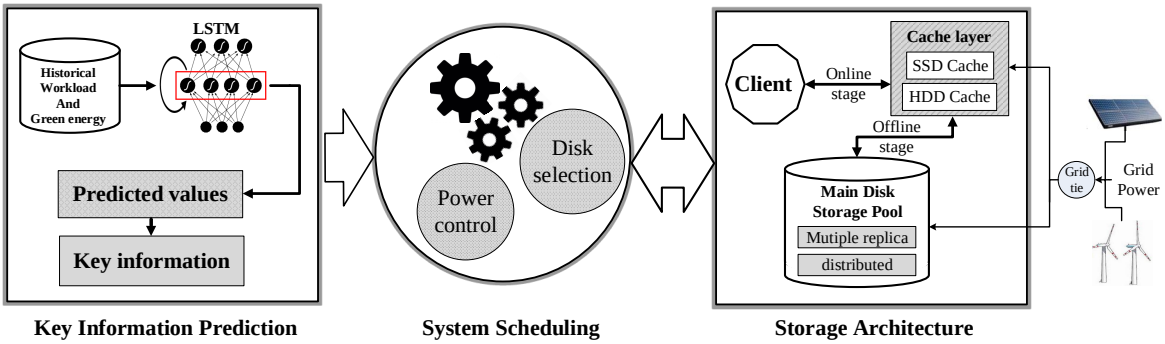

Fig. 2: The architecture of PreMatch

method. Finally, the storage architecture part will schedule the devices as the system scheduling part wishes.

Key information prediction and system scheduling will be described in section 4 and section 3 respectively. Next, we will explain the storage architecture in detail.

C. Storage architecture

As shown in Fig. 2, the storage architecture mainly comprises energy equipment, the client, a cache layer, and a main disk storage pool.

The energy equipment offers a mixture of solar power, wind power, and traditional grid power. We use the grid-tie [20] to synchronize the green energy and the grid power, and the dominating supply is the green power. Grid power is used as standby power and will be used when the green energy can't satisfy the minimum demand of the storage system.

The main disk storage pool is a distributed HDD storage system, and disks in the storage pool are named **P-disks**. In detail, the main storage pool is grouped into three replicas: the primary replica and two non-primary replicas. Access requests of the storage system are entirely served by the cache layer and the primary replica, and the non-primary replicas mainly exist to guarantee the data reliability of the primary replica. Thus the non-primary P-disks could be powered off to save energy when green energy is insufficient and storage workload is light, and the related reliability concerns are addressed by the cache layer.

The cache layer is composed of SSD cache and HDD cache. The SSD cache is used to store the frequently and recently accessed data. Through the SSD cache, we can divide the process of a request, whether latency sensitive or not, into two stages: online SSD cache stage and offline underlying disk stage. Through the online stage, clients can obtain excellent performance and the peak workload can be partly shifted by the SSD cache with little energy consumption. Meanwhile, we can control the execution time of offline stage to match the variation of a green energy supply, which can improve green energy utilization and save grid energy.

The HDD cache is logically divided into two zones: a RAID (Redundant Array of Independent Disks)-Log zone and an HDD-Remain zone. The HDD-Remain zone is used as a supplement of the SSD cache, as SSDs have a lower capacity/price rate and the hybrid cache layer will be more cost-effective. The RAID-Log zone is used to guarantee the reliability of the written data in the SSD cache and the HDD-Remain zone.

As shown in Fig. 3, The upcoming write requests from the client will be written into the SSD cache and the RAID-Log zone simultaneously. When the SSD cache is full, the newly written data will be stored in the HDD-Remain zone instead. For example, the coming object 7 is written into the HDD-Remain zone, as there is no place in the SSD cache. In the RAID-Log zone, we use RAID5 here, and the fixed data objects 3, 4, and 5 are stored in a stripe with the generated parity object P1.

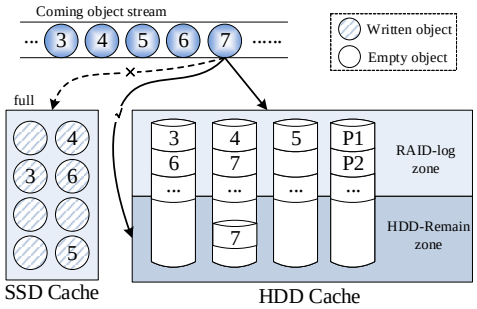

Fig. 3: The division of HDD cache

III. PreMatch Scheduling

Our system scheduling comprises the power control policy and the P-disk selection method.

We compartmentalize the whole action time into many continuous local decision-making cycles, and one long-range cycle consists of many local cycles. At the end of every local decision-making cycle, a new specific system scheduling method will be generated for the following local decision-making cycle.

A. Power Control Policy

In our policy, the SSD cache and the HDD cache will always be active regardless of the green energy conditions. Green energy will be utilized to supply the primary replica firstly, and then the non-primary replicas. When green energy can't satisfy the normal service, the grid energy will be used. Besides, all of the replications in the storage pool will be interchanged to be the primary replica in Round-Robin order for every long execution cycle (a week, a month, and so on).

For simplicity, we plan to make a quantitative classification on the workload. Considering the time-varying characteristics of the workload, we set a time window T_{range} (6 hours as an example) to narrow the range of measurement, which will be dynamically moved to contain the current local decision-making cycle.

TABLE I: The definition of variables

Variables	definition
$local(t+1)$	the following local decision-making cycle
$W_{local(t+1)}$	workload variation trend of $local(t+1)$
$G_{local(t+1)}$	green energy variation trend of $local(t+1)$
$W_{local\_mode(t+1)}$	workload mode of $local(t+1)$
$W_{disk(t+1)}$	number of **P-disks** workload needs in $local(t+1)$
$G_{disk(t+1)}$	number of **disks** green energy can supply in $local(t+1)$
G_{disk_last}	number of **P-disks** green energy can supply in the last local decision-making cycle
$Power_{one\text{-}disk}$	energy required to power on one disk
$E_{SSD+HDD}$	the energy to activate both the SSD and HDD cache
K	number of replicas
M	each replica contains M P-disks
$W_{long\text{-}range(t+1)}$	workload variation trend of the following long-range cycle
$G_{long\text{-}range(t+1)}$	green energy variation trend of the following long-range cycle
$N_{Prim\_disk(t+1)}$	the final number of active primary P-disks in $local(t+1)$
$N_{Non\_Prim\_disk(t+1)}$	the final number of active non-primary P-disks in $local(t+1)$

W_{max} and W_{min} represent the maximum and minimum workload (the request number) in the T_{range} period respectively. $W_{overall-avg}$ means the average workload for all of the historical workload.

Finally, the workload is divided into two modes by the request number: **Heavy mode** and **Light mode**.

- **Heavy mode.** Which means the top β percent of the workload, belongs to range $[MAX(W_{overall-avg}, W_{max} - \beta(W_{max} - W_{min})), \infty)$. We use $W_{overall-avg}$ as the minimum value limit for high mode workload recognition to avoid weak identification.
- **Light mode.** Which means the remain workload, belongs to range $[0, MAX(W_{overall-avg}, W_{max} - \beta(W_{max} - W_{min})))$.

To control the energy consumption of a storage system, the number of active devices is significantly important. In our PreMatch, the active P-disks are made proportional to the periodically dominant one of the green energy and workload dynamically. Thus when the workload is heavy in the following local decision-making cycle, we can alleviate inactive P-disk accesses by powering on all the primary P-disks in the primary replica to avoid latency penalty. Because it takes about 11 seconds to switch a disk from standby to ready [21]. When the workload is light, we can control the number of active P-disks to match the green energy variation to save grid energy. If inactive P-disk access happens, we will power on the accessed P-disk by grid energy during the local decision-making cycle.

However, when there are conflicts between the local and long-range variation trends, frequent switchings of P-disks will result in device wear and performance fluctuation. Therefore, considering the local and long-range variation trends of both workload and green energy, the optimum number of active P-disks will be described in algorithm 1.

To describe the power control policy clearly, we have defined some variables in Table I. And we use **Prim_disk** and **Non_Prim_disk** to represent primary P-disk and non-primary P-disk respectively.

Algorithm 1 Power Control Policy With Predicted Information

Require: $W_{local(t+1)}$, $W_{local\_mode(t+1)}$, $G_{local(t+1)}$, $W_{disk(t)}$, G_{disk_last}, K, M, $W_{long\text{-}range(t+1)}$, $G_{long\text{-}range(t+1)}$.

Ensure: $N_{Prim\_disk(t+1)}$, $N_{Non\_Prim\_disk(t+1)}$

The SSD cache is always on
The HDD cache will be on unless the green energy can supply all the storage devices and the HDD cache is clear of dirty data
if $W_{local(t+1)}$ is downward but $W_{long\text{-}range(t+1)}$ is upward **then**
  Ignore the local variation and keep the same number of active P-disks as the last local decision-making cycle
  $W_{disk(t+1)} \Leftarrow W_{disk(t)}$
else
  We will follow the workload value in the following local decision-making cycle
  if $W_{local\_mode(t+1)}$ belongs to $Heavymode$ **then**
    Open all of the M primary P-disks
    $W_{disk(t+1)} \Leftarrow M$
  else
    $W_{disk(t+1)} \Leftarrow 0$
  end if
end if
$E_{cache} \Leftarrow E_{SSD+HDD}/Power_{one\text{-}disk}$
if $G_{disk(t+1)} \geq E_{cache}$ **then**
  $G_{disk(t+1)} \Leftarrow MIN((G_{disk(t+1)} - E_{cache}), K * M)$
else
  $G_{disk(t+1)} \Leftarrow 0$
end if
if $G_{disk(t+1)} \geq W_{disk(t+1)}$ **then**
  if $G_{local(t+1)}$ is downward but the related $G_{long\text{-}range(t+1)}$ is upward **then**
    $N_{Prim\_disk(t+1)} \Leftarrow MAX(N_{Prim\_disk(t)}, W_{disk(t+1)})$
    We will power off the Non-primary P-disks to reply the decreased green energy
    $N_{Non\_Prim\_disk(t+1)} \Leftarrow N_{Non\_Prim\_disk(t)} - (G_{disk\_last} - G_{disk(t+1)})$
  else
    if $G_{local(t+1)}$ is upward but the related $G_{long\text{-}range(t+1)}$ is downward **then**
      $N_{Prim\_disk(t+1)} \Leftarrow MAX(N_{Prim\_disk(t)}, W_{disk(t+1)})$

if $W_{disk(t+1)} = M$ **then**
The increased green energy will be used to supply the possibly added primary P-disks for priority
$N_{tem} \Leftarrow G_{disk(t+1)} - G_{disk\_last} - (M - N_{Prim\_disk(t)})$
else
$N_{tem} \Leftarrow G_{disk(t+1)} - G_{disk\_last}$
end if
Then the remain added green energy will be used to supply the Non-primary P-disks
$N_{Non\_Prim\_disk(t+1)} \Leftarrow N_{Non\_Prim\_disk(t)} + N_{tem}$
else
$N_{Prim\_disk(t+1)} \Leftarrow MIN(G_{disk(t+1)}, M)$
$N_{Non\_Prim\_disk(t+1)} \Leftarrow G_{disk(t+1)} - N_{Prim\_disk(t+1)}$
end if
end if
else
$N_{Prim\_disk(t+1)} \Leftarrow M$
$N_{Non\_Prim\_disk(t+1)} \Leftarrow 0$
end if
$N_{Non\_Prim\_disk(t+1)} \Leftarrow MAX(N_{Non\_Prim\_disk(t)}, 0)$
$N_{Non\_Prim\_disk(t+1)} \Leftarrow MIN(N_{Non\_Prim\_disk(t)}, (K-1) * M)$

If the workload is downward in the local decision-making cycle but upward in the long-range cycle, we will ignore the local variation and keep the same number of active P-disks as the last local decision-making cycle. Otherwise, we will follow the workload mode in the local decision-making cycle.

When the local and long-range variation trends of the green energy supply conflict, we will turn to confirm the number of P-disks the green energy can supply and the workload needs in the local decision-making cycle. If workload needs more active primary P-disks than green energy can supply, we would power on enough primary P-disks as the workload needs. Otherwise, we will guarantee that the primary P-disks are not influenced by the local green energy variation. To this end, we will only power off the non-primary P-disks with the temporarily decreased green energy respectively. And the increased green energy will be used to supply the possibly added primary P-disks for priority.

B. P-disk Selection

Knowing the proper number of active P-disks, then we should decide the specific P-disk selection rules. Before that, we will make a short announcement. To guarantee the performance of the SSD cache, dirty data in the SSD cache and the HDD-Remain zone will only be synchronized to the primary replica. And the non-primary replicas can only get the up-to-date written data from the RAID-Log zone.

1) Metadata Management:

Firstly, to manage the required metadata of all the data in the storage system, we use two Least Recently Used (LRU) lists and a Hot Data Level (HotLev) list and mainly focus on the location and hotness information.

As shown in Fig. 4, the LRU1 list records the metadata entry of the recently accessed data, and the twice-visited metadata entry will be moved to the LRU2 list. The HotLev list manages the metadata entry by the hotness level. There are n (10 for example in our experiment) hotness levels, and each level contains a range of hotness value. And the hotness value will decay exponentially in time. The accessed metadata entry in the HotLev list will be raised to the head of the LRU1 list. Meanwhile, the metadata entry evicted by two LRU lists will be thoroughly moved to the HotLev list.

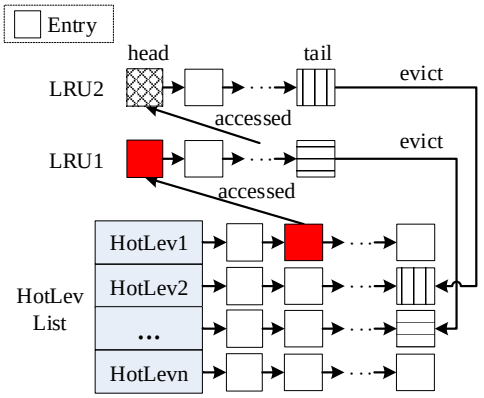

Fig. 4: The metadata information that we require

Each entry in Fig. 4 is designed as (obj_id, $P\text{-}disk\_id$, $hotness$, $location$, $dirty\_flag$). The obj_id represents the logical address, the $P\text{-}disk\_id$ represents the P-disk that the data belongs to, the $hotness$ means the accessed popularity, the $location$ shows where the data is stored (SSD cache, HDD-Reserved zone, RAID-Log zone or P-disks), and the $dirty\_flag$ shows that the data is dirty($dirty\_flag$=1) or clean($dirty\_flag$=0).

2) Primary P-disk Selection:

To select the cost-effective primary P-disks, we should record the real-time power state (on or off) and the sum of data that belongs to HotLev1 and HotLev2 for each primary P-disk ($HotLev(1+2)\_Num$). And we only record the data that exists in the primary P-disks.

First, we can order all the primary P-disks by the $HotLev(1+2)\_Num$ from largest to smallest. Then, we can spin up the inactive primary P-disk with the largest $HotLev(1+2)\_Num$ until the $N_{Prim\_disk(t+1)}$ equals zero. Finally, the remaining active primary P-disks should be spun down, before which we should prefetch the HotLev1 and HotLev2 data that only exists in these primary P-disks to the HDD cache.

3) Non-primary P-disk Selection:

As non-primary P-disks will be spun up only when the green energy is sufficient or the HDD cache is almost full of dirty data, it will be cost-effective to activate the non-primary P-disk that has the largest number of dirty data in the RAID-Log zone.

To select the cost-effective non-primary P-disks, we record the real-time information of the dirty data in the RAID-Log zone ($Dirty\_Num$). In detail, we should classify the dirty data by the non-primary P-disk ID. We only record the information for one non-primary replica, as we always spin up or down the related non-primary P-disks simultaneously.

First, we can order all non-primary P-disks by the $Dirty\_Num$ from largest to smallest. Then, we can spin up the inactive non-primary P-disk with the largest $Dirty\_Num$ until the $N_{Non\_Prim\_disk(t+1)}$ equals zero. Finally, The remain active non-primary P-disks should be spun down.

C. Destage and Prefetch

When part of the P-disks are spun up in the following local decision-making cycle, dirty data in the SSD cache and the HDD-Reserved zone will be synchronized to the active primary P-disks, and dirty data in the RAID-Log zone will be synchronized to the active non-primary P-disks. However, when the workload is heavy, the destage task to the primary P-disks will be temporarily stopped to guarantee the performance. Besides, when both of the SSD and the HDD cache are almost full of dirty data, the system will be forced to do the destage process. If green energy is insufficient, traditional grid energy will be utilized to support the destage.

Meanwhile, when the workload is light, part of the hottest data in the active primary P-disks will be prefetched to the cache. And there will also be data exchanges between the SSD cache and the HDD-Reserved zone, which can guarantee that the SSD cache maintains the hottest data. But to avoid data migration fluctuations, we prefetch at most one-twentieth (an example in this article) the size of SSD cache hottest data from the primary P-disks at once.

To relieve the destage impact on the normal cache access, we also split the destage task into many pieces and deploy these pieces in the whole local decision-making cycle.

IV. Key Information Prediction

It is clear that we need the local and long-range variation trends (upward or downward) of the workload and green energy, as well as the accurate value of the following local decision-making cycle. Indeed, both of the local and global variation trends can be obtained from the historical and following local and long-range cycle respectively. Thus, we only need to predict the accurate values of the following local and long-range cycle, and the upward and downward variation trend are labeled as 1 and 0 respectively.

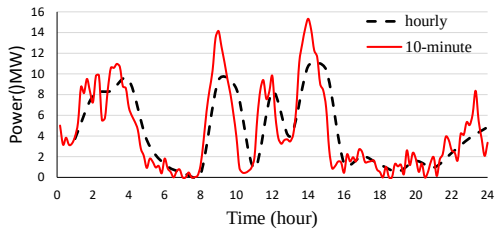

Fig. 5: Hourly and 10-minute time internal data curve of the green energy

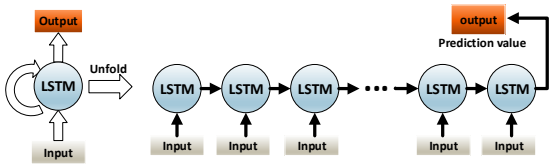

Fig. 6: LSTM neural network achitecture

In this paper, we use 10 minutes and 1 hour as the local decision-making cycle and the long-range cycle respectively. As shown in Fig. 5, we take the green energy for an example. There are many irregular fluctuations in the hourly time interval data curve of workload and green energy supply. When zooming in for 10-minute time interval observation, the data curve is more erratic.

To achieve accurate prediction, we eventually adopt the LSTM neural network architecture [14], which has been demonstrated to be efficient in various fields that have a sequential nature. The LSTM neural network is a recurrent neural network and is able to capture the long-term dependencies in the time-series data.

The LSTM structure is shown in Fig. 6 and comprises three layers: the input layer, LSTM cell layer, and output layer. We predict the accurate values of the following local and long-range cycle on the past 36 10-minute and 24 hourly time internal data respectively. And 36 and 24 are the look-back time steps. The prediction models for the workload and green energy are completely independent, and there will be four LSTM network models at last. For all of the LSTM network models, all of the LSTM cells in one model share the same weights and there are 10 hidden units in each LSTM cell.

V. Experiment Setup

With the limited experimental environment, we evaluate the effect of the PreMatch by using the simulation method. The system implementation module diagram is shown in Fig. 7.

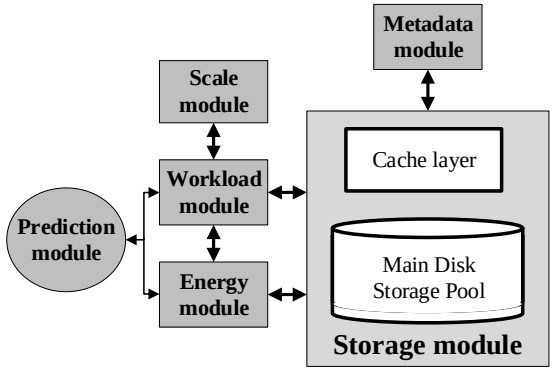

Fig. 7: Simulation module diagram

The Scale module is used to control the scale of the system. As the storage pool is designed as an n-way replicated distributed storage system, and the SSD cache is organized by consistent hashing, we can scale up and down the storage system easily.

The Energy module is used to adapt the renewable energy supply to different scales. Both of the wind and solar datasets

are collected from the National Renewable Energy Laboratory [22]. As the sunshine intensity and wind speed vary in different seasons, we randomly chose a hybrid week-long solar and wind trace in summer and winter respectively from Dodge City. The characteristics of two hybrid green energy traces (*Summer, Winter*) are listed in Table II, which shows the green power name, average power output, the ratio of low power period that is less than 1 MW and the total time. Besides, the two traces can be linearly scaled to match different storage scales.

TABLE II: The characteristics of green energy traces

ID	Avg(MW)	low-power ratio	note
Winter	15.17	42.66%	7 days
Winter3	14.24	41.9%	3 days of Winter
Summer	17.4	13.49%	7 days
Summer3	16.46	20.07%	3 days of Summer

In the Workload module, many clients are simulated, and each client will create some threads to dispatch the same trace, such as usr trace, to the Storage module at a time. The workload traces are collected from the MSR with a total of 36 different traces[23]. We chose two week-long traces, whose characteristics are listed in Table III. Because running a seven-day trace non-stop is infeasible, our experiment accelerates the test by a factor of 60. And the top 10 percent of the workload is defined as heavy mode.

TABLE III: The features of traces

Traces	Write Ratio	IOPS	Avg. Req
usr	59%	83.87	22.66KB
rsrch	91%	21.17	8.93KB

The Prediction module, using the LSTM neural network, is responsible for providing accurate future information for the Workload and Energy module respectively. All of the workload and green energy traces are divided into two parts: a four-day training dataset, and a three-day testing dataset. We train the LSTM networks with the training datasets, and then will evaluate our schemes with the testing datasets.

The Metadata module deals with the required metadata of all the data in the storage system.

The Storage module is responsible for the simulation of actual HDDs and SSDs. The HDD cache and the main storage pool use the same simulated disks. Simulated SSDs and disks parameters are generated from the real devices, which are shown in Table IV. In the simulation, we just provide the request access time calculated by the request type, size and device conditions. Moreover, we install the modified sheepdog [19] in three similar physical servers to simulate the primary replica server nodes and two non-primary replica server nodes respectively. The parameters of the real server are also shown in Table IV.

To evaluate our PreMatch, there are four configurations used for comparison.

- **Standard**: This scheme is an SSD-cache based data center without any power management policy.

TABLE IV: Hardware details

OS	Linux version 2.6.35.6-45.fc14.x86_64	
CPU	Intel (R) Xeon (R) CPU E5506@2.13GHz	
Hard Disk	Seagate ST2000DM008 2TB SATA 7200rpm	
SSD	SAMSUNG 850 EVO 120G SATA3	
Disk Parameters	Average Latency	6 ms
	Maximum data transfer rate	220 MB/Sec
	Idle Power	3.9 W
	Standby Power	0.3 W
	Active Power	5.1 W
SSD Parameters	Sequential Read/Write(up to)	520 MB/s
	Random Read (up to)	94000 IOPS
	Random Write (up to)	88000 IOPS
	Active Power	3.7W
	Idle Power	0.5W

- **WDS** (**W**orkLoad-**D**riven **S**cheme): We keep the cache and the primary replica active, and will power on any P-disk when the storage system desires.
- **PreMatch**: A standard data center with our novel system scheduling scheme. The predicted information is used in this mode, and both the SSD and HDD cache are always kept active.
- **PreMatch-T** (PreMatch **T**rue): The only difference with PreMatch is that we use the real-world information instead in this mode.

VI. Evaluation

In this section, we will keep the number of devices in the data center with a proportion of 1 SSD cache: 3 HDD cache: 30 P-disks. And we take the storage system with 30 P-disks for an example.

A. Accuracy of Prediction

Table V shows the prediction accuracy of the key characteristics on the workload and green energy traces. It is clear that most of the accuracy values of workload are above 95%. The accuracy of green energy in the local decision-making cycle is evaluated by the RMSE (Root Mean Squared Error), and it is small too. Fig. 8 shows the true and predicted data curve of the winter3 trace. We can see that the predicted curve is almost coincided with the true winters trace.

TABLE V: The prediction accuracy of the key characteristics

workload	local mode	local variation	long-range variation
usr	97.22%	95.35 %	97.14%
rsrch	97.47%	95.59%	95.71%
Green energy	RMSE of local value	local variation	long-range variation
Winter3	1.54	86.77%	80.28%
Summer3	1.84	80.26%	79.05 %

B. Energy Consumption Comparison

Fig. 9 shows the energy consumed by four configurations under various green energy and workload traces. The upper part of the splicing cylinder represents the grid energy actually used, and the lower part represents the green energy actually used. First, PreMatch can save more grid energy than other methods under various traces. The primary replica is always on in WDS, whereas PreMatch will only activate the primary

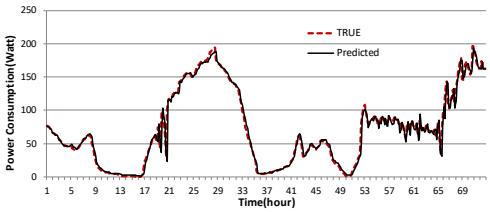

Fig. 8: The true and predicted data curve of the winter3 trace

replica when the workload is heavy or the green energy is sufficient. Therefore, much grid energy will be wasted in WDS. We can also see that PreMatch consumes more green energy than WDS, as much green energy will be wasted when the green energy supply exceeds the primary replica demands and the non-primary replicas needn't be powered on.

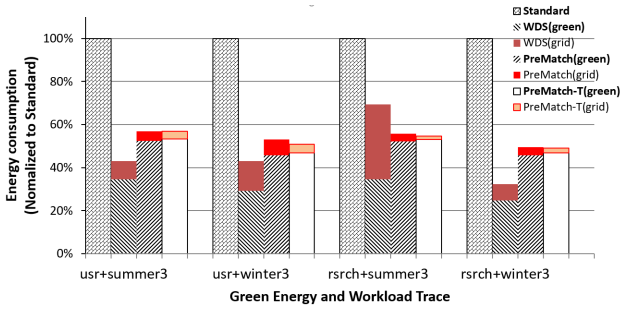

Fig. 9: Total energy consumption of four configurations under various traces

Under $summer3$ trace, PreMatch can reduce a little more grid energy, because $summer3$ has the smaller low ratio (20.07%) and higher average value (16.46MW) than $winter3$ trace as shown in Table II, thus the $summer3$ trace can provide more efficient green energy. And under $rsrch + summer3$, PreMatch delivers the highest grid energy saving, and reduce grid energy up to 98.5% when compared to the standard method. Meanwhile, PreMatch consumes only the half grid energy of WDS at most. Besides, under $usr + summer3$ and $rsrch + winter3$, we get the maximum utilization ratio (95.8%) and minimum utilization ratio (94.4%) respectively. In the WDS, the green energy will only be used when the workload demands, plenty of the green energy is wasted when the workload is light.

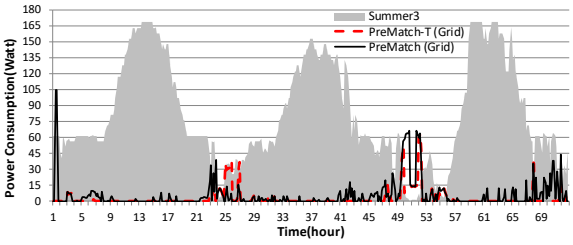

Fig. 10: Real-time grid energy consumption of PreMatch and PreMatch-T under $rsrch$ and $Summer3$

We can also see that PreMatch consumes more traditional grid energy than PreMatch-T, especially under the $rsrch$ and $Summer3$ traces. To make it clear, we introduce the Fig. 10, which describes the real-time grid energy consumption of PreMatch and PreMatch-T under the $rsrch$ and $Summer3$ traces. In Fig. 10, we can see that PreMatch consumes more grid energy than PreMatch-T when the green energy is unable to supply all of the P-disks. Because larger predicted value of green energy in the local decision-making cycle will bring more active P-disks, and the extra activated P-disks will be supported by the grid energy, such as the 17th and 61th hours. In addition, the light mode workload period in the real-world workload trace might be predicted as the heavy mode in PreMatch, which will also waste some grid energy when the green energy is insufficient (the 23th and 24th hours). Therefore, PreMatch consumes more grid energy than PreMatch-T.

C. Performance Comparison

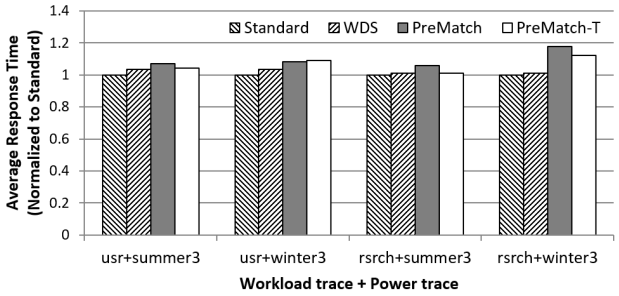

Fig. 11: The performance of four configurations under various green energy and workload traces

Fig. 11 shows the performance of four configurations under various power and workload traces. First, we can see that standard configuration has the best performance, as all of the replicas are active all the time. WDS performs only a little worse than the standard method, as the primary replica is always on and the performance will only be affected by the destage. PreMatch will only activate the primary replica when the workload is heavy or the green energy is sufficient, so PreMatch delivers the worst performance. And we can find that PreMatch performs a little better under $summer3$ than $winter3$. Because the $summer3$ trace can provide more efficient green energy than $winter3$.

In PreMatch, the SSD cache hit ratio is very high, so only a small fraction of data accesses will be redirected to the main storage pool when the workload is light. Furthermore, we will activate the whole primary replica when the workload is heavy, thus inactive P-disk access will not bother us. Therefore, considering the high green energy utilization ratio and significant grid energy saving, the performance degradation of PreMatch is acceptable and graceful.

We can also find that PreMatch-T always performs a little better than PreMatch in most cases. Because the wrong variation trend prediction in green energy will result in more conflicts between the local and long-range variation trends, which might lead to the performance degradation in PreMatch. As when variation trend conflicts happens in the green energy, we will guarantee that the primary P-disks are not influenced

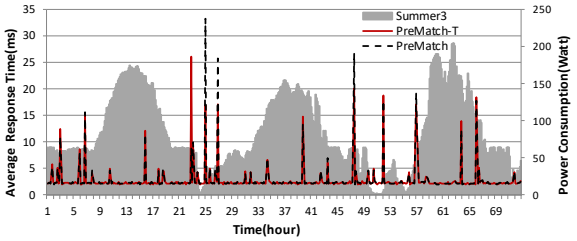

Fig. 12: Real-time performance of PreMatch and PreMatch-T under $rsrch$ and $Summer3$ traces

by the conflicts, and non-primary P-disks will be operated by the green energy part that changes.

Besides, the wrong predictions in the workload local mode will also give rise to performance degradation. Because several heavy workload modes might be recognized as light modes, and the primary replica could not be activated as the real-world traces expected when the green energy is insufficient, thus the performance will decline. To make it clear, the Fig. 12 is introduced, which shows the real-time performance of PreMatch and PreMatch-T under $rsrch$ and $Summer3$. For example, we can see that the PreMatch-T performs better than PreMatch at the 25th and 27th hours. The reason is that PreMatch-T knows the correct heavy workload modes, and consumes more grid energy than PreMatch at the 25th and 27th hours, which is shown in Fig. 10.

In conclusion, our PreMatch, using the predicted green energy and workload information, consumes a little more grid energy and performs a little worse than PreMatch-T, but the difference is acceptable. And not like the optimal PreMach-T, **our PreMatch is genuinely viable**.

D. Large Scale Analysis

To evaluate PreMatch in different scale sizes, we set the number of P-disks in the storage pool as 3000, 15000 and 30000. For simplicity, we take the tests under $usr+summer3$ as examples for comparison.

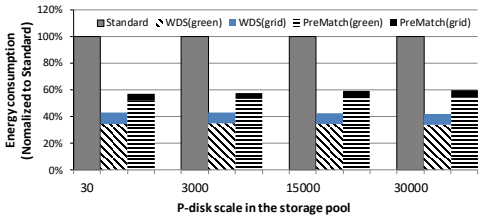

Fig. 13: Energy consumption of different storage scales under $usr + summer3$

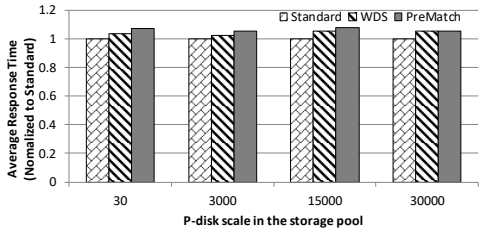

Fig. 14: Performance of different storage scales under $usr + summer3$

As shown in Fig. 13 and Fig. 14, The results of different scale sizes demonstrate the same regularity as our previous tests (30 P-disks). Meanwhile, with the rise of the storage scale, green energy utilization rate rises a little. Because the extra green energy that could not supply one device can be reused with an enlargement factor. In a word, though there are some limitations in our simulation tests, our PreMatch could match different storage scale sizes well.

VII. RELATED WORK

Green Energy utilization. Rutgers University and HP Labs have built the data centers partly powered by green energy to verify their methods [24, 25]. And several researchers [26] have studied the simulation method to reduce carbon emissions and cost. Chao Li et al. have developed sustainability, scale and power security [27, 28] studies on green data center. SolarCore [29] try to temporarily reduce the server energy consumption when solar power is low. Blink [9] uses a staggered blinking schedule to address the variability of wind and solar power. The energy storage devices [6] are always used as standby power to avoid lack of green energy, but bring in energy loss and equipment cost.

Workload-driven schemes. B. Aksanli, et al. have proposed two separate job arrival queues to process mixed workloads in data centers [30]. GreenSlot [7], GreenHadoop [4], and GreenSwitch [24] maximize the green energy usage by deferring a batch of latency-insensitive jobs. Li, et al. [8] migrated workloads across all the servers and GreenGear [31] leverages the heterogeneity to match green energy. GreenCassandra [32] saves grid energy by guaranteeing that one data replica is always on. Our PreMatch aims at saving grid energy by spinning down disks as many as possible when the workload is light and green energy is insufficient within a data center.

Usages of SSDs. Generally, SSD [18] can act as a cache for HDD or make up a hybrid storage system with HDD [11, 13, 33, 34]. Lazy-SSD [35] focuses on the endurance problem of the SSD cache, and design a caching algorithm to avoid cache pollution and preserve popular blocks in cache for a longer period of time. Article [36] put the popular data in the SSD and part of the disks to save energy. In the video area, article [37] use an SSD and a parity disk on S-RAID to optimize the random reads and writes, and address the sequential writes by part of the disks. Article [38] focuses on the reliability of the energy-saving hybrid storage system. In total, some of the energy-saving methods are orthogonal to our overall design, and could be used in our primary replica.

LSTM Neural Network. The LSTM neural network is one popular variation of the recurrent neural network, and it is able to capture the long term dependencies in the time-series data [14]. LSTM has been demonstrated to work well in various domains that have a time-series nature, and some researchers have studied the usage of LSTM in some complex applications [15–17]. Paper [17] investigates the effectiveness of using LSTM to do a prediction for constructing dynamic energy management algorithms in chip multiprocessors. And article [15] focuses on the forecasting of CPU usage

of machines in data centers. In article [16], a hierarchical framework was proposed for cloud resource allocation and power management, and the LSTM was used to provide the requests arrival interval.

VIII. Conclusion

When using green energy in HDD-based storage systems, the workload-driven schemes get satisfactory performance but could rarely leverage the abundant green energy, while supply-oriented schemes reduce the performance of latency-sensitive workloads. This paper proposes **PreMatch** to take into consideration both of the workload and green energy by LSTM. In detail, we utilize an SSD cache to shift part of the workload, and design an adaptive cost-effective scheme to maintain an available number of active HDDs to provide service and save grid energy according to the predicted information. Compared with the storage system using WDS, PreMatch can further improves 35% additional green energy utilization ratio and reduces at least 50% grid energy with little performance degradation. Besides, PreMatch can work well with large-scale storage systems.

Acknowledgement

This work was sponsored in part by the Creative Reaserch Group Project of NSFC No.61821003; the National Key Research and Development Program of China No.2018YFB1003305; the National Natural Science Foundation of China under Grant No.61472152, No.61872413, No.U1709220, No.61902137. We also greatly appreciate Peng Xu's discussions.

References

[1] G. Parkinson. (2016) Solar and wind energys stunning cost falls to continue. [Online]. Available: https://reneweconomy.com.au/solar-and-wind-energys-stunning-cost-falls-to-continue-25263/

[2] (2017) 100% renewable is just the beginning. [Online]. Available: https://sustainability.google/projects/announcement-100/

[3] A. Verma, R. Koller, L. Useche, and R. Rangaswami, "Srcmap: Energy proportional storage using dynamic consolidation," in *8th USENIX Conference on File and Storage Technologies, San Jose, CA, USA, February 23-26, 2010*, R. C. Burns and K. Keeton, Eds. USENIX, 2010, pp. 267–280.

[4] I. Goiri, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenhadoop: leveraging green energy in data-processing frameworks," in *European Conference on Computer Systems, Proceedings of the Seventh EuroSys Conference 2012, EuroSys '12, Bern, Switzerland, April 10-13, 2012*, 2012, pp. 57–70.

[5] J. Wan, X. Qu, N. Zhao, J. Wang, and C. Xie, "Thinraid: Thinning down raid array for energy conservation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 10, pp. 2903–2915, 2015.

[6] 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 *Proceedings of the 42nd Annual International Symposium on Computer Architecture, Portland, OR, USA, June 13-17, 2015*, 2015, pp. 463–475.

[7] I. Goiri, R. Beauchea, K. Le, T. D. Nguyen, M. E. Haque, J. Guitart, J. Torres, and R. Bianchini, "Greenslot: scheduling energy consumption in green datacenters," in *Conference on High Performance Computing Networking, Storage and Analysis, SC 2011, Seattle, WA, USA, November 12-18, 2011*, 2011, pp. 20:1–20:11.

[8] C. Li, A. Qouneh, and T. Li, "iswitch: Coordinating and optimizing renewable energy powered server clusters," in *39th International Symposium on Computer Architecture (ISCA 2012), June 9-13, 2012, Portland, OR, USA*, 2012, pp. 512–523.

[9] N. Sharma, S. K. Barker, D. E. Irwin, and P. J. Shenoy, "Blink: managing server clusters on intermittent power," in *Proceedings of the 16th International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2011, Newport Beach, CA, USA, March 5-11, 2011*, 2011, pp. 185–198.

[10] D. Li, X. Qu, J. Wan, J. Wang, Y. Xia, X. Zhuang, and C. Xie, "Workload scheduling for massive storage systems with arbitrary renewable supply," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 10, pp. 2373–2387, 2018.

[11] S. He, X. Sun, and B. Feng, "S4d-cache: Smart selective SSD cache for parallel I/O systems," in *IEEE 34th International Conference on Distributed Computing Systems, ICDCS 2014, Madrid, Spain, June 30 - July 3, 2014*, 2014, pp. 514–523.

[12] N. Liu, J. Cope, P. H. Carns, C. D. Carothers, R. B. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the role of burst buffers in leadership-class storage systems," in *IEEE 28th Symposium on Mass Storage Systems and Technologies, MSST 2012, April 16-20, 2012, Asilomar Conference Grounds, Pacific Grove, CA, USA*, 2012, pp. 1–11.

[13] T. Pritchett and M. Thottethodi, "Sievestore: a highly-selective, ensemble-level disk cache for cost-performance," in *37th International Symposium on Computer Architecture (ISCA 2010), June 19-23, 2010, Saint-Malo, France*, 2010, pp. 163–174.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] D. Janardhanan and E. Barrett, "CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models," in *12th International Conference for Internet Technology and Secured Transactions, ICITST 2017, Cambridge, United Kingdom, December 11-14, 2017*, 2017, pp. 55–60.

[16] N. Liu, Z. Li, J. Xu, Z. Xu, S. Lin, Q. Qiu, J. Tang, and Y. Wang, "A hierarchical framework of cloud resource

allocation and power management using deep reinforcement learning," in *37th IEEE International Conference on Distributed Computing Systems, ICDCS 2017, Atlanta, GA, USA, June 5-8, 2017*, 2017, pp. 372–382.

[17] M. G. Moghaddam, W. Guan, and C. Ababei, "Investigation of LSTM based prediction for dynamic energy management in chip multiprocessors," in *Eighth International Green and Sustainable Computing Conference, IGSC 2017, Orlando, FL, USA, October 23-25, 2017*, 2017, pp. 1–8.

[18] F. Chen, R. Lee, and X. Zhang, "Essential roles of exploiting internal parallelism of flash memory based solid state drives in high-speed data processing," in *17th International Conference on High-Performance Computer Architecture (HPCA-17 2011), February 12-16 2011, San Antonio, Texas, USA*, 2011, pp. 266–277.

[19] (2018) Sheepdog. [Online]. Available: http://sheepdog.github.io/sheepdog/

[20] N. Deng, C. Stewart, and J. Li, "Concentrating renewable energy in grid-tied datacenters," in *Proceedings of the 2011 IEEE International Symposium on Sustainable Systems and Technology*. IEEE, 2011, pp. 1–6.

[21] (2018) Disk parameters. [Online]. Available: https://www.seagate.com/internal-hard-drives/

[22] (2019) Grid modernization. [Online]. Available: https://www.nrel.gov/grid/

[23] (2017) Msr cambridge traces. [Online]. Available: http://iotta.snia.org/tracetypes/3

[24] I. Goiri, W. A. Katsak, K. Le, T. D. Nguyen, and R. Bianchini, "Parasol and greenswitch: managing datacenters powered by renewable energy," in *Architectural Support for Programming Languages and Operating Systems, ASPLOS '13, Houston, TX, USA - March 16 - 20, 2013*, 2013, pp. 51–64.

[25] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '12, London, United Kingdom, June 11-15, 2012*, 2012, pp. 175–186.

[26] M. Brown and J. Renau, "Rerack: power simulation for data centers with renewable energy generation," *SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 77–81, 2011.

[27] 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 *The 46th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-46, Davis, CA, USA, December 7-11, 2013*, 2013, pp. 322–333.

[28] C. Li, Z. Wang, X. Hou, H. Chen, X. Liang, and M. Guo, "Power attack defense: Securing battery-backed data centers," in *43rd ACM/IEEE Annual International Symposium on Computer Architecture, ISCA 2016, Seoul, South Korea, June 18-22, 2016*, 2016, pp. 493–505.

[29] C. Li, W. Zhang, C. Cho, and T. Li, "Solarcore: Solar energy driven multi-core architecture power management," in *17th International Conference on High-Performance Computer Architecture (HPCA-17 2011), February 12-16 2011, San Antonio, Texas, USA*, 2011, pp. 205–216.

[30] B. Aksanli, J. Venkatesh, L. E. Zhang, and T. Rosing, "Utilizing green energy prediction to schedule mixed batch and service jobs in data centers," *Operating Systems Review*, vol. 45, no. 3, pp. 53–57, 2011.

[31] 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 *Proceedings of the 2016 International Conference on Supercomputing, ICS 2016, Istanbul, Turkey, June 1-3, 2016*, 2016, pp. 12:1–12:14.

[32] W. A. Katsak, I. Goiri, R. Bianchini, and T. D. Nguyen, "Greencassandra: Using renewable energy in distributed structured storage systems," in *Sixth International Green and Sustainable Computing Conference, IGSC 2015, Las Vegas, NV, USA, December 14-16, 2015*, 2015, pp. 1–8.

[33] Q. Yang and J. Ren, "I-CASH: intelligently coupled array of SSD and HDD," in *17th International Conference on High-Performance Computer Architecture (HPCA-17 2011), February 12-16 2011, San Antonio, Texas, USA*, 2011, pp. 278–289.

[34] F. Chen, D. A. Koufaty, and X. Zhang, "Hystor: making the best use of solid state drives in high performance storage systems," in *Proceedings of the 25th International Conference on Supercomputing, 2011, Tucson, AZ, USA, May 31 - June 04, 2011*, 2011, pp. 22–32.

[35] S. Huang, Q. Wei, J. Chen, C. Chen, and D. Feng, "Improving flash-based disk cache with lazy adaptive replacement," in *IEEE 29th Symposium on Mass Storage Systems and Technologies, MSST 2013, May 6-10, 2013, Long Beach, CA, USA*. IEEE Computer Society, 2013, pp. 1–10.

[36] C. Ho, H. Chen, Y. Chang, Y. Chang, P. Huang, T. Kuo, and D. H. Du, "Energy-aware data placement strategy for ssd-assisted streaming video servers," in *IEEE Non-Volatile Memory Systems and Applications Symposium, NVMSA 2014, Chongqing, China, August 20-21, 2014*. IEEE, 2014, pp. 1–6.

[37] X. Yu, C. Zhang, Y. Xue, H. Zhu, Y. Li, and Y. Tan, "An extra-parity energy saving data layout for video surveillance," *Multimedia Tools Appl.*, vol. 77, no. 4, pp. 4563–4583, 2018.

[38] W. Felter, A. Hylick, and J. B. Carter, "Reliability-aware energy management for hybrid storage systems," in *IEEE 27th Symposium on Mass Storage Systems and Technologies, MSST 2011, Denver, Colorado, USA, May 23-27, 2011*, A. Brinkmann and D. Pease, Eds. IEEE Computer Society, 2011, pp. 1–13.