Optimizing Energy Efficiencies in Cloud Data Center Resources with Availability Constraints

H.L. Phalachandra, Dinkar Sitaram

Abstract*: Cloud infrastructure Resources hosted in Data Centers, support the effective execution of Cloud computing applications. Given the increased adoption of the Cloud Computing Applications and the Businesses getting to be Data-driven, there is a huge increase in the number of Data Centers and the Size and amount of resources hosted in these Data Centers. These Data Center resources consume a significant amount of energy and this continuous scaling of the resources is leading to increased power consumption and a large carbon footprint. Given our fragile eco-system, optimization of the Data Center resources for energy conservation and thus the carbon footprint is the primary area of our focus. Businesses also need to satisfy QoS guarantees on Availability to their customers. Optimization towards Energy efficiencies may compromise on the Availability and thus may warrant a trade-off, and a need for them to be considered together. Although there have been numerous studies towards Energy efficiencies, most of them have been focused on only energy. In this paper, we initially segregate Optimization activities towards the Data Center resources like Compute, Network, and Storage. We then study the different control parameters or approaches which will lead to meeting the objectives of Energy Efficiencies, Availability and Energy Efficiency constrained with Availability. Thus, this will support the selection of approaches for the optimization of energy while meeting the QoS Availability requirement.

Keywords*: Availability, Data Center Resources, Energy Efficiency, Optimization, *QoS

I. INTRODUCTION

Data explosion in the last decade due to the availability and adoption of a plethora of devices and Applications creating data has led to this data permeating into our lives and Businesses. This data and its availability in this bustling data-driven economy are fast getting to be one of the major success factors for the Business. The hosting of these applications and the data in a scalable Data Center is now the prominent and preferred approach. This is leading to growth in the number, and the volume of resources hosted in these Data Center. These Data Center consume significant power, and according to Statistics in [1], the total energy consumed in countries in 2006 was to the tune of 61 billion KWH, which increased to 100 billion KWH in 2011 [2]. This is getting to be ~1.4% of the world's electricity consumption. This high energy-consuming Data Centers are increasing the Carbon footprint at the rate of 6% and is expected to be at 10-12% by 2020 [3], thus causing damage to our environment, which has been one of the major global concerns in the last few years.

Energy consumed in these Data Center is from the non-IT components like the Power sources, Power converters, Power distribution units, Cooling related components like the CRAC units, Chillers, lighting, and other infrastructure components and IT components like the Server, Networking and Storage. These Active IT resources of the Data Center viz. Server, Networking Equipment and Storage consume ~45% of the Data Center Energy [4].

Data which is one the major success factors for Businesses in our data-driven economy is hosted in the Cloud Data Center infrastructure, Availability considerations for the storage resources hosting this data is also a critical factor to be addressed while administering the Data Center.

Energy optimization and Availability of these Active components are areas of major focus for the Cloud and Data Center Research.

Several studies have been carried out towards the optimization of energy and Availability.

To get a holistic view of all the optimization activities towards these Energy and Availability objectives, we have segregated the work based on different Data Center resources viz. Compute, Network and Storage. Activities within each of these have been aggregated and focused as Surveys, based on their control parameters or approaches as below.

Summary of Data Center Resource Surveys

Compute Resource Management	
Control Parameter/Approach	Surveys
Resource Provisioning	[5],[6],
Scheduling	[7][8][9][10]
Load Balancing	[11][12][13]
Fault Tolerant	[14]
Energy Efficiencies	[15][16] [17][18][19][20][21]
Availability	[23]

Table- I: Compute Resource Optimization Studies

Network Resource Management	
Control Parameter/Approach	Surveys
Summary of Green Activities	[24][25]
Dynamic Resource provisioning	28
Traffic Consolidation	
Virtualization	[26],[27]
Network Latencies and adaptive link rate	[29]
Energy Efficiencies	[30][31][32][33][34][35]

Table- II: Network Resource Optimization Studies

Revised Manuscript Received on February 05, 2020.
***** Correspondence Author
H.L. Phalachandra*, CCBD, CSE, PESU, Bangalore, India. Email: phalachandra@pes.edu
Dinkar Sitaram, CCBD, CSE, PESU, Bangalore, India. Email: dinkars@pes.edu

Storage Resource Management	
Control Parameter/Approach	Surveys
Disk Characteristics	[36]
Caching	[37]
De-Duplication	[38]
Architectures	[39]
Energy	[36],[37],[38],[39]

Table- III: Storage Resource Optimization Studies

In the subsequent sections, we look at Optimizations done towards each of the Data Center resources viz. Compute in the granularity of VMs, Network, and Storage, around the few objectives of our focus, while considering the different control parameters. We also specifically explore work that has looked at energy optimization along with Availability considerations.

II. COMPUTE RESOURCE MANAGEMENT

Compute resources contribute significantly towards a Data Center energy consumption and the optimization activities would be around the VM, the granular view of the Compute.

A. Optimization Objectives

VM provisioning can be visualized as in Fig.1.

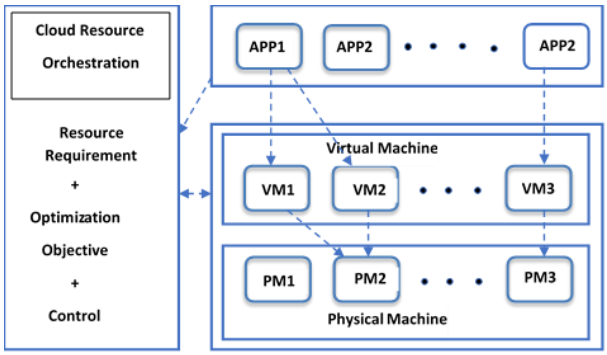

Fig. 1.VM Provisioning into Physical Machines

There can be several optimization objectives including Performance and Cost associated with VMs, but this study focuses on

1. Reduction in Power or Energy consumed
2. Meeting the QoS Availability expectation

Availability from a Data Center Compute perspective is typically supported through VM replication. Some of the other approaches that help supporting Availability are by scaling to the resource requirements of Applications, increasing Reliability by reducing failures or increasing MTBF, impacts of failures and recovery from failure by reducing MTTR or by building in resilience into the Compute infrastructure.

B. Control Parameters for Optimization and the Associated work

The following are the base parameters or approaches considered for driving the optimizations along with references towards the research work, associated with these. There a few more parameters like the VM Size i.e. in terms of pre-fixed sized or few custom sized pre-fixed sized VMs like with AWS v/s custom sizes or in terms of cost as with the work [40], [41], [42], [43], [44], [45] looking at pricing, brokering or Auctions to manage the trade-off between over-provisioning for peak and risk of performance and cost.

- ***Technologies and operating modes of components on the Physical Server***

  There are optimization approaches that choose technologies like CMOS which have different energy characteristics due to leakage power caused by leakage current [36]. There are also approaches as seen with [46],[47],[48],[49] which based on the Power computation below,

$$P_{(dynamic)} = \sigma \times C \times V \times f$$

σ – switching activity C – Physical capacitance
V – Supply Voltage f – Clock frequency

  orchestrate the parameters for Dynamic Power Management (DPM) considering the leakage short circuit current, switched capacitance and clock rates or dynamically scale to the voltage and frequency (DVFS) by estimating the total CPU frequency required for supporting the responsiveness, and computing the frequency and number of Servers needed for that [50].

- ***Locations of the Physical Servers hosting the VMs***

  There has been work on locating Data centers like the Ballengen's Kolos facility near the Arctic or the Data Center set up by Facebook in northern Sweden, in geographical locations which need minimal cooling [176] and leveraging this for energy efficiencies in terms of reduced need for cooling and also towards Disaster recovery [51], [52]. There have also been other approaches to move computing to the Edge like what is now called Edge Data Centers, closer to the users as part of Edge computing, where the performance and latency challenges are addressed [177],[178].

- ***Utilization of the Physical Servers***

  Servers are typically not evenly loaded and are not continuously in a utilized state but consume energy. Bringing down the active physical servers will decrease the energy consumed. There have work which has been explored towards achieving this, like with the initial static placements of VMs [53], or allocation based on optimal resources needed [54], or allocation and placement based on probabilistic prediction of the application resource needs [13], or deploying the VM initially using statistical assignment, and then migrating for optimization [55], or reliably assessing the resource needs using ML and then deploying the VMs [6] or by adjustment the VM allocations based on utilization [56]. This optimization could also be done by Reallocations or Dynamic placements of the VMs into the Physical Machines [57], [12].

- ***QoS Factors***

  Redundancies built through Replication are supported by provisioning and scheduling multiple copies of the VMs as needed to support the QoS Availability requirement.

  There are various approaches for VM placement which factor in Availability as seen with the Survey [23].

There are ones which, as a policy keep the availability in context and look to allocate resources to either scale horizontally or vertically as with [58],[59]. Approach as with [60] focuses on Resilience to support Availability using an Exact solution based on heuristics. The approach in [61] forms a failover group to support Availability by being aware of the component availability characteristic and their interdependencies. Some approaches as with [62] increase the fault tolerance of the VMs using mechanisms as Virtualization Fault Tolerance (VFT) by extending Xen and Nebula. The approach in [63] models Availability using Markov-based models to reduce the number of faults. Approaches as with [64] support Application Availability, after VM provisioning by keeping track of the health of the VM through Heartbeat and migration as needed. There are approaches like [65] define an Availability, Migration and Recovery policy for a VM, and look to support VM availability through migration.

- ***Orchestration of provisioning & Scheduling of VMs to the Physical Servers***

  All of the above orchestration approaches from the choice of the technologies and operating modes to the utilization of the Servers are looking indirectly for energy efficiencies. Approaches as with [66] look at focusing the workload to a small number of physical nodes to enhance Energy efficiencies [67] or using five diverse power management policies [68]. The work in [69] considers the high resource dynamics, latencies of taking the processors to low power states uses a meta-scheduler to map VMs to Servers using utilization based on the workload prediction. Mistral, a framework [70] has been used to control and orchestrate for efficiencies towards energy and performance. The approach in [71] estimates the energy consumed and schedules based on the same. Other work as with [72],[73],[74],[75],[76],[77],[78],[79],[80],[81],[82] orchestrate VMs to the Physical Machines with a focus on conservation of energy.

C. Energy-Efficient Optimization with Availability constraints

There have been approaches that have looked at scheduling VMs while keeping energy and Reliability as a constraint as with [83]. The approach in [84] characterizes workload data, and clusters the same for both user and VM requests, and orchestrates resources, by estimating the future workload and thus scaling for energy efficiencies and Availability.

Some other approaches have looked at enhancing the Availability by using Reliability, like the Availability and Maintainability (RAM) model to analyze the riskiness and the impact of interactions between different components and enabling identification and taking measures for energy efficiencies [85].

The approach in [86] uses the lowest energy cost with minimal deadline miss ratio (thereby increasing Availability) as a significant factor for migration for fault resolution. The work in [87] models support for Availability, by dividing Cloud Data Centers into a hierarchy of failure zones viz. a complete Data Center, a Zone or a Sector or an Aisle or a Rack or a Server within a Data Center which has the granular probability of failure. VM replicas are scheduled to different failure zones navigating a hierarchical tree based on the survivability of the failure zone and thus supporting Availability. Energy optimization is then done within the failure zone by the choice of the Physical machine with the lowest energy cost using Gravity Algorithm, an enhanced variant of the Hill Climbing Algorithm providing a Global Minima, and thus supporting Availability with optimization for energy efficiencies.

In the case of faults in the environment, there has been work to reduce the time for which the Data Center equipment would need to be down, by monitoring the events raised using a multilayer node event processing (MNEP) mechanism and thus increase its Availability [88].

Cooling and improper management of temperature will have an impact on the optimal functioning and operation of the Data Center resources. The work in [67] considers the temperature of these active devices and ensures that the cooling system is functioning appropriately for the optimal performance of the resources.

All of the above approaches look at QoS Availability as an additional constraint over the Energy efficiencies while scheduling VMs to PMs.

III. NETWORK RESOURCE MANAGEMENT

Data Center Networking typically has the goal of ensuring that latencies observed during data exchanges are acceptable and support the QoS requirements of the Applications. These goals are challenging due to the need for the network, to scale and be efficient in terms of energy and cost.

The following sections discuss the Optimization objectives and the control parameters for achieving these objectives and the research work towards them.

A. Optimization Objectives

Given the focus on the reduction of Energy and Availability of the Data Center Network resources, although there can be several other optimization objectives including cost, Performance (considered with Availability), this study focuses on

1. Reduction in Power or Energy consumed
2. Meeting the QoS Availability expectation

   Given that there is no standard articulation of Network Availability in a Data Center we define it as the ability to support expected throughput with acceptable latencies, even under a non-uniform volume of traffic, with provisioned Network Capacities, within the Reliability & Failure characteristics of the available Network Infrastructure.

B. Control Parameters for Optimization and the Associated work

The following parameters are considered for controlling the Network optimizations viz. Architecture, choice of technologies and operating modes, consolidation and balancing of the loads and virtualization which facilitates scaling, resiliency to faults and energy optimization.

These optimization approaches and the references to the research work associated with them are as below:

- *Architecture or topology of the Data Center Network*
  The choice of the Architecture has a bearing on the scalability, cost, fault tolerance and power consumption of the DCN.

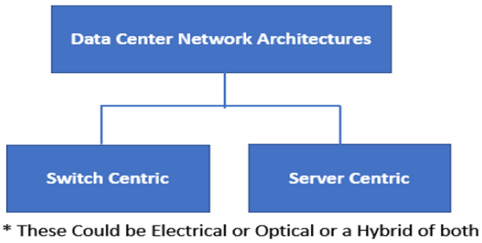

Fig. 2. High-Level Classification of NW Architecture

Different work towards the Architectures as in Fig.2. look to address some of the network issues and influence energy consumed as below:

- Switch centric, where the focus is interconnection, routing and connecting users to the cloud, and work as in [89],[90],[91],[92] addresses over-subscription, agility, server-to-server traffic flow load balancing, etc.
- Server centric Architectures where packet forwarding and routing forms the core of the Architecture as in [92], [93], [94]

These Architectures have different energy profiles and positively increases energy efficiencies in some specific scenarios. The Balanced Tree Switch Centric Architecture a variant of [89] is found to consumes the least power irrespective of the number of Servers.

- ***Technologies of the Network Components and their Operating modes***
  The choice of technology of the Network components, whether Electrical, Optical or Hybrid, for the network to send across packets, has a bearing on the energy consumed, bandwidth and the ability to send to higher distances are deliberated in [95],[96],[97],[98], [99],[100].

  Network devices like the hubs, switches, and routers have operating modes that conserve energy and devices are orchestrated as with [101], to move the devices into these states for the max amount of time. Techniques like DVS, DVFS have also been employed in conjunction with the VM computes to achieve the optimization goals as with [46],[47],[48],[49],[50],[102],[103].
  There have also been activities to route network traffic in a manner, which enables network devices to be moved to low power states as in [104],[105] for energy-efficiency. Network speeds or rates have been adopted through DVM or by shaping traffic into bursts, in [106], based on the load determined as optimal or practical (using history).

- ***Network Static and Dynamic load management***
  Network traffic tends to be bursty. Thus, factoring in the load, either based on historic patterns or dynamically based on the network state, will actively manage the power consumed of the network components and thus help towards the optimization objectives.
  There have been approaches [105],[107],[108],[109] which look to allocate computing resources and network paths simultaneously which can minimize energy consumption. There have also been approaches as with [110] where the load-based energy consumption profile is factored in for energy-aware routing. The work in [111], [112] look to selectively and transparently move idle devices to a low power state. There has been work as in [113], where VM placements are made with the awareness of traffic.

- ***Virtualization***
  There have been researched approaches that optimize energy efficiencies through VM migrations using Virtualization, where services are moved around transparently as if connected to the same switch, thus helping the migration of VMs [109]. Virtualization could be implemented as a software component or using additional Hardware like fabric managers while factoring in the Network load as seen in [89], [90], [114]. VM live migrations have also been implemented which factor in the network load added due to the migration, and ensure effective bandwidth utilization [66], [115].
  There have also been approaches like [89] where a special flat addressing scheme is used for separating the Server names and location making it location-independent addressing. Similar location-independent addressing is used while consolidating network load and traffic into a select cluster thus enabling lightly loaded devices to be moved to a low power state. [90] uses Pseudo MACs to handle issues related to VLANs, ACLs, Broadcast domains.
  Some approaches like [116], [117] look at traffic flow routing with energy reduction as a focus. There are also SDN based algorithmic approaches in [118],[119], which have been looked at for energy efficiencies.

- ***QoS factors***
  There has been work towards making the network to be Available, by supporting workload beyond the provisioned capacity, by migrating unmodified workloads to other Data Centers while retaining the networking configuration parameters as with [120].
  There have also been various approaches to support Availability by addressing non-uniformity of workloads and failures, like by balancing the load in various points on the network [121],[122] or by assessing the risk of failures and recovery times to support resilience [123], or by using different architectures [124]. There have been approaches that look to avoid congestion by Multipathing [125], or by dynamically reconfiguring and creating latency-sensitive paths based on the size of workflows as with [126],[127],[128] while keeping the planned latencies.
  An approach like in [124] considers different architectures to address the risk of failures and those like [129] use heterogeneous network service chaining to support network service availability.

C. Energy-Efficient Optimization with Availability constraints

There have been approaches that consider Network Energy Efficiencies with the QoS Availability constraint in a Data Center environment.

There's been work towards providing Availability by managing the bandwidth and the duration of the bandwidth of network paths used for communication, factoring in the energy efficiency and orchestrating the schedule of the data flow based on the heuristics of the communication pattern [130] at the time of VM Scheduling.

There have also been approaches which have looked to manage the responsiveness and Latency to ensure Availability during high workload, by using Multipathing and using the multiple paths for Availability through redundancy [131] or by using predictive ML-based Auto scaling mechanisms which manage responsiveness and latency aligned to the QoS expectation [132], or by avoiding congestion through per packet-based energy-aware segment routing and load balancing in SDNs while turning of links for energy efficiencies [133].

Some approaches have looked at the impacts of moving network devices into energy-conserving low power states [134] which typically may increase the failure rate of the devices and thus the Availability [135]. There are also work which have explored in Optical core networks, usage of optical components for energy efficiencies and the trade-off for acceptable failure rates for non-impact to Availability [136]. Some approaches have also looked at supporting Availability by avoiding blocking probability in the network as a trade-off to energy [137]. Given that the components can fail, some approaches have looked at graceful degradation of performance on component failure, keeping the energy as a constraint and supporting Availability through redundancy and failover [138].

IV. STORAGE RESOURCE OPTIMIZATION

Storage components are estimated to have ~27% influence on the performance and the energy consumption in a Data Center. The capacity needing to be supported by the Storage devices has been geometrically increasing and is expected to be around 2PB. This leads to challenges for supporting IO performance, Size, Energy, Reliability, Availability, Security, etc.

A. Optimization Objective

Several optimizing goals can exist on the Storage components of the Data Center to ensure that all the challenges are addressed effectively. We have chosen the following objectives as part of our study

1. Reduction in Power or Energy consumed
2. Meeting the QoS Availability expectation

   Given that there is no standard definition of Storage Availability in a Data Center, we define it as the ability to support storage capacity when needed, with read and write latencies meeting the expectations for the workload, with reliability in terms of resilience and recoverability of the data in case of errors/failures or Data corruption.

B. Control Parameters for Optimization and the Associated work

The following parameters are considered for controlling the Storage optimizations viz. choice of components and devices based on the technologies, or disk modes and states, or by using techniques like Caching, Load balancing, Tiering or Virtualization for IO performance and effective utilization of the disks, or by reducing the data footprint to be stored in the Data Center, or by optimizing on the capacity and energy consumed or by keeping the focus on Reliability and Availability by factoring in faults of the storage devices.

These optimization approaches and the references to the research work associated with them are as below

- ***Choice of storage components and devices based on technologies***

  There has been work on considering different kinds of HDDs in terms of form-factor, capacity, operational speed of disks, energy, protocols which they can support like the IDE, SATA, SAS, SCSI, FC with different design objectives and overheads, or technologies like SSDs, DRAMs, NVRAMs, etc. in [139],[140], [141] and [142].

  There have also been approaches that have looked at grouping disks of different technologies to build a disk hierarchy with low and high-power disks, and use this tiering as a mechanism for the optimization. [143], [144].

- ***Disk Mode and States:***

  Disks have modes where the speed of the disk spindles as with DRPM can be varied and they also have a low power consumption inactive state and a normal active state, which can be orchestrated for power efficiencies. Approaches as with [147], [148] have used DRPM for energy efficiencies. There have also been approaches that spin down disk adaptively for managing power distributed to the disk drive [139].

  Approaches have been proposed to keep the disks in the Inactive state, by keeping the data access to locations which need lower power [146] [149]. There are also approaches which based on the workload, offload the data to different permanent stores to reduce the spin down and spin of disks [157] or in disk arrays where data is concentrated to a few disks with Popular Data Concentration PDC [158] to enable moving disks to low power states. These have also been explored in RAID-based systems, where RAID data blocks have been grouped together and dynamically rearranged based on the workload to enable most disks to be power-saving modes as with [150]. There have also been approaches that manage the storage queue depth for power efficiencies keeping the performance in context as with [151].

- ***Cache Based Approaches***

  There have been several approaches that have looked at Caching as a mechanism for optimizing energy and performance.

Some of them have a small amount of NVcache built into the Disks which help with the performance as with [139]. The MAID approach [152], based on the workload/Application profile, uses a small number of the total available drives as a data cache for all the data and provides energy efficiencies by moving the rest of the disks to power-saving state. The work in [149] looks to lay out the data with power in consideration, into a cache disk based on the Application profile. There have been other optimization approaches using Cache, which populate the cache based on a prediction from historic traces [153] or structure the writes to the disks from the cache with energy into consideration [154].

There are also cache-aware algorithms which when working with RAIDs use techniques like TRAs to increase the hits to the cache, and thus reduce the need to access the disks for longer times and allow the disks to be in the spun-down state and conserve power [155] or use hierarchical caches as with [156] or use TRCs and TRDs (Transformable Reads on Cache's or Disk's) for energy conservation. The approach in [153] uses offline-online Power-aware Algorithms that address Cache misses and cache replacement using storage management policies and optimize on energy.

- ***IO Load Sharing***

  There have been approaches towards provisioning Storage IO resources based on the workload to ensure IO performance, by considering simple fairness, like proportional share allocation or approaches like the reward Scheduler which provides an incentive to processes which have better runtime characteristics [163], or a scheduling policy like vFair for sharing the IO load regardless of IO workload pattern [164], or a scheduling approach where a fair share is computed, and when bottlenecked, provide a share proportional to the fair share as in [165]. The approach in [166] looks to increase the load sharing capacity by using tiered storage consisting of SSDs and a technique like a reward scheduling which favors the clients whose IOs are less costly on the backend storage array. There are also approaches which use Hybrid disks and Hybrid Storage Algorithms for efficiencies [39].

- ***Capacity Optimizing Technologies***

  Given the increasing need for storage resources, usage of provisioned capacity has a significant bearing on all of the above optimizing parameters. Technologies like Delta snapshots, Thin provisioning, Advanced RAID, Data De-duplication and Compression [142] have been explored to reduce the data footprint and thus optimizing on the need for storage capacity and hence the energy in Data Centers.

- ***Energy Efficiencies***

  Several approaches have been considered in Data Centers which are focused on conserving energy consumed by the Storage devices. The orchestrations as seen with [146] to [158] above focus on energy efficiencies along with other control parameters.

  There are also approaches where load-based optimization for energy conservation in terms of multi-speed disks used in an environment with disk speeds are reduced based on the loads [167], switching off systems/disks in the cluster with optimization based on load balancing [162].

  There has also been work that looks at the stimulus responses of a disk [159] and models the dynamic power characteristics for a historic workload IO traces [160], which in turn is used for predicting the energy characteristics and optimization for energy efficiencies [161]. The approach with [168] assesses the data patterns and distributes the data onto a hybrid set of devices. These are also followed inside enterprise-class storage arrays which are hosted in Data Center, by implementing heuristics-based policies to drive the data into a heterogeneous set of disks like SSD and HDDs both through initial allocation and through automatic migration [169].

- ***Optimizations based on Availability***

  There have been approaches as with [170] where a local Storage Array in the Application environment is used to front a Cloud Storage array in a Cloud Datacenter, and pseudo availability is supported by using heartbeats to identify the accessibility of the Data Center Storage Array. There are approaches as in [171] where virtual storage is created within the storage device and in case of issues/errors/failures, the data is rebuilt within the device transparent to the Application and thus supporting Availability. The work in [172] uses a thin layer of Storage management and provides tolerance to failures by sharding and associating the shards with parity or error correction codes and thus supports recreation even in case of non-availability of the replications and a Hierarchical Storage policy, and data spreading policy has been used to tolerate failures and thus increasing the Availability.

C. Energy-Efficient Optimization with Availability constraints

There are approaches as with [162] which using the SSD Staged, Energy Efficient Object Storage Architecture, which uses a small SSD staging layer, complemented with niche Algorithms, and provides performance and energy enhancements, without compromising on Availability. The work in [174], caps the power of the storage device in a power-controlled mode, by adjusting the storage transaction queue depth for I/O performance, and thus supports Availability. The work by [37] considers technologies like SSDs, NVM and techniques like Caching and Tiering to increase Availability by considering performance, replication, reliability and energy awareness. There is also work by [175] which overprovisions and uses two data storage areas in memory with energy consideration and alleviates the failure characteristics of SSDs by minimizes the impact of wear and thus supports Availability.

V. RESULTS AND DISCUSSION

All the work discussed above based on the different control parameters or approaches can be summarized as in Table IV below.

Summary of the Research Work References			
	Considerations to Energy	Considerations to Availability	Considerations to Energy and Availability
Compute	[6],[12],[13],[36], [40] to [57], [66] to [82], [176] to [178]	[23],[58] to [65]	[67], [83] to [88]
Network	[46] to [50], [66], [89] to [119]	[120] to [129]	[130] to [138]
Storage	[39], [139] to [172]	[170] to [172]	[37],[162],[174],[175]

Table- IV: Summary of the Research Work References

Given that Availability is predominantly supported by Redundancy, supporting Availability in most scenarios increases the energy consumption in the Data Center. Thus, approaches that look to optimize energy efficiency should also simultaneously consider Availability, as there may be a need for a trade-off. As can be seen from Table IV above, a significant amount of work is focused only on the independent optimization goals of Energy and Availability, with very less focus on Energy along with Availability.

VI. CONCLUSION

There have been several diverse research approaches considering the energy efficiencies of Data Center Resources for specific contexts as seen above. These approaches ensure the adequate but optimal capacity of resources to be available in the Data Center to support the needs of the Applications, at a minimal cost, while supporting QoS requirements like Performance, Reliability, Availability, etc.

As seen in Table IV, most Energy Optimization activities in the Data Center have kept energy as the sole focus. There are also a few approaches that have only considered the Availability requirements. There have not been many studies that have factored in Energy and Availability requirements simultaneously, which at times may need a trade-off. So, using the interpretation of Availability as discussed above for each of the Data Center resources, this study identifies the need for considering approaches for Energy while factoring in Availability simultaneously across the different Data Center Resources.

REFERENCES

1. A. Naskos, A. Naskos, E. Stachtiari, P. Katsaros, and A. Gounaris, “Probabilistic Model Checking at Runtime for the Provisioning of Cloud Resources.”
2. U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, “A cost-aware elasticity provisioning system for the cloud,” in Proceedings - International Conference on Distributed Computing Systems, 2011, pp. 559–570.
3. S. Zaman and D. Grosu, “A Combinatorial Auction-Based Mechanism for Dynamic VM Provisioning and Allocation in Clouds,” in IEEE Transactions on Cloud Computing, 2013, vol. 1, no. 2, pp. 129–141.
4. G. F. Anastasi, E. Carlini, M. Coppola, and P. Dazzi, “QBROKAGE: A genetic approach for QoS cloud brokering,” in IEEE International Conference on Cloud Computing, CLOUD, 2014, pp. 304–311.
5. B. Jennings and R. Stadler, “Resource Management in Clouds: Survey and Research Challenges.”
6. T. le Duc, “Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey,” 2019.
7. Kumar, Mohit, et al. "A comprehensive survey for scheduling techniques in cloud computing." Journal of Network and Computer Applications 2019.
8. Z. Zoltán´ and Z. Mann, “Allocation of Virtual Machines in Cloud Data Centers-A Survey of Problem Models and Optimization Algorithms *.”
9. M. Masdari, F. Salehi, M. Jalali, and M. Bidaki, “A Survey of PSO-Based Scheduling Algorithms in Cloud Computing,” Journal of Network and Systems Management, vol. 25, no. 1, pp. 122–158, 2017.
10. W. J. et al. Li J, Yang S, “Research on Dynamic Virtual Machine Scheduling Strategy Based on Improved Genetic Algorithm.”
11. Kumar, Pawan, and Rakesh Kumar. "Issues and challenges of load balancing techniques in cloud computing: A survey." ACM Computing Surveys (CSUR) 51.6: 120. 2019
12. Mukati, Lalit, and Arvind Upadhyay. "A Survey on Static and Dynamic Load Balancing Algorithms in Cloud Computing." Available at SSRN 3365568, 2019.
13. S. K. Panda and P. K. Jana, “Load balanced task scheduling for cloud computing: a probabilistic approach,” Knowledge and Information Systems, vol. 61, no. 3, pp. 1607–1631, Dec. 2019.
14. Kathpal, Chesta, and Ritu Garg. "Survey on Fault-Tolerance-Aware Scheduling in Cloud Computing." In Information and Communication Technology for Competitive Strategies, pp. 275-283. Springer, Singapore, 2019.
15. Khattar, Nagma, Jagpreet Sidhu, and Jaiteg Singh. "Toward energy-efficient cloud computing: a survey of dynamic power management and heuristics-based optimization techniques." The Journal of Supercomputing: 1-61. 2019
16. G. Prasad Babu and A. K. Tiwari Associate Professor, “Energy Efficient Scheduling Algorithm for Cloud Computing Systems Based on Prediction Model,” Int. J. Advanced Networking and Applications, pp. 4013–4018, 2019.
17. D. Kapil, E. S. Pilli, and R. C. Joshi, Live Virtual Machine Migration Techniques: Survey and Research Challenges.
18. Ahmad, Raja Wasim, et al. "A survey on virtual machine migration and server consolidation frameworks for cloud data centers." Journal of Network and Computer Applications 52: 11-25. 2015
19. V. Medina and J. Manuel García, “A Survey of Migration Mechanisms of Virtual Machines,” vol. 46, 2014.
20. T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J.-M. Pierson, and A. v Vasilakos, “Cloud computing: A survey on energy efficiency,” ACM Computing Surveys, vol. 47, no. 2, pp. 1–36, 2015.
21. S. W. Han, S. D. Min, and H. M. Lee, “Energy-efficient VM scheduling for big data processing in cloud computing environments,” Journal of Ambient Intelligence and Humanized Computing, 2019.
22. Z. Tao et al., “A Survey of Virtual Machine Management in Edge Computing,” 2019.
23. Nabi, Mina, Maria Toeroe, and Ferhat Khendek. "Availability in the cloud: State of the art." Journal of Network and Computer Applications 60: 54-67. 2016
24. A. P. Bianzino, C. Chaudet, D. Rossi, and J.-L. Rougier, “A Survey of Green Networking Research.”
25. U. Samee, et al., “Green networks,” J Supercomput.
26. S. Nanda and T.-C. Chiueh, “A Survey on Virtualization Technologies.”
27. Chowdhury, NM Mosharaf Kabir, and Raouf Boutaba. "A survey of network virtualization." Computer Networks 54.5: 862-876. 2010
28. Ł. Budzisz et al., “Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: a Survey and an Outlook.”
29. K. Bilal et al., “A survey on Green communications using Adaptive Link Rate,” Cluster Computing, vol. 16, no. 3, pp. 575–589, Sep. 2013.
30. R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, “Energy efficiency in the future internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures,” IEEE Communications Surveys and Tutorials, vol. 13, no. 2, pp. 223–244, 2011.
31. G. Luigi Valentini et al., “An overview of energy efficiency techniques in cluster computing systems,” vol. 16, pp. 3–15, 2013.
32. S. Zeadally et al., “Energy-efficient networking: past, present, and future,” J Supercomput, vol. 62, pp. 1093–1118, 2012.
33. A. Hammadi and L. Mhamdi, “A survey on architectures and energy efficiency in Data Center Networks,” Computer Communications, vol. 40, pp. 1–21, 2014.
34. C. Fang, C., Yu, F.R., Huang, T., Liu, J. and Liu, Y. A survey of green information-centric networking: Research issues and challenges. IEEE Communications Surveys & Tutorials, 17(3), pp.1455-1472. 2015
35. Ge, Chang, Zhili Sun, and Ning Wang. "A survey of power-saving techniques on data centers and content delivery networks." IEEE Communications Surveys & Tutorials 15.3: 1334-1354. 2012
36. A. Beloglazov, R. Buyya, C. Lee, A. Zomaya, and Y. C. Lee, “A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud Computing Systems,” 2012.
37. M. Hoseinzadeh, “A Survey on Tiering and Caching in High-Performance Storage Systems,” Apr. 2019.
38. Neelaveni, P., and M. Vijayalakshmi. "A survey on deduplication in cloud storage." Asian J. Inf. Technol 13.6: 320-330. 2014

39. Niu, Junpeng, Jun Xu, and Lihua Xie. "Hybrid storage systems: a survey of architectures and algorithms." IEEE Access 6: 13385-13406. 2018
40. J. Tordsson, R. S. Montero, R. Moreno-Vozmediano, and I. M. Llorente, “Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers,” Future Generation Computer Systems, vol. 28, no. 2, pp. 358–367, Feb. 2012.
41. L. Tomás and J. Tordsson, “Improving cloud infrastructure utilization through overbooking,” ACM International Conference Proceeding Series, no. August, 2013.
42. T. Truong Huu and J. Montagnat, “Virtual resources allocation for workflow-based applications distribution on cloud infrastructure,” pp. 1–6, 2010.
43. Lucas-Simarro, Jose Luis, et al. "Scheduling strategies for optimal service deployment across multiple clouds." Future Generation Computer Systems 29.6: 1431-1441. 2013
44. Song, Zz, Xiaojing Zhang, and Clas Eriksson. "Data center energy and cost-saving evaluation." Energy Procedia 75: 1255-1260. 2015
45. X. Sui and H.-F. Leung, “An Adaptive Bidding Strategy in Multi-Round Combinatorial Auctions for Resource Allocation
46. Lorch, Jacob R., and Alan Jay Smith. "PACE: A new approach to dynamic voltage scaling." IEEE Transactions on Computers 53.7: 856-869. 2004.
47. V. Haldar, C. W. Probst, V. Venkatachalam, and M. Franz, “Virtual-Machine Driven Dynamic Voltage Scaling.”
48. D. Lago, E. R. M Madeira, and L. Fernando Bittencourt, “Power-Aware Virtual Machine Scheduling on Clouds Using Active Cooling Control and DVFS,” in Proceedings ACM 9th International Workshop on Middleware for Grids, Clouds and e-Science, 2011, p. 2
49. S. Lee and T. Sakurai, “Run-time Voltage Hopping for Low-power Real-time Systems,” 2000
50. J.S. Chase, D.C. Anderson, P.N. Thakar, A.M. Vahdat, R.P. Doyle, Managing energy and server resources in hosting centers, in: Proceedings of the 18th ACM Symposium on Operating Systems Principles, ACM, New York, NY, USA, pp. 103–116, 2011
51. A. Khosravi, S. Kumar Garg, and R. Buyya, “Energy and Carbon-Efficient Placement of Virtual Machines in Distributed Cloud Data Centers.”
52. B. Spinnewyn, R. Mennes, J. F. Botero, and S. Latré, “Resilient application placement for geo-distributed cloud networks,” Journal of Network and Computer Applications, vol. 85, pp. 14–31, 2017.
53. A. Moreau, et al., “From Data Center Resource Allocation to Control Theory and Back,” 2010.
54. Chang, Fangzhe, Jennifer Ren, and Ramesh Viswanathan. "Optimal resource allocation in clouds." 2010 IEEE 3rd International Conference on Cloud Computing. IEEE, 2010.
55. C. Mastroianni, M. Meo, and G. Papuzzo, “Self-economy in cloud data centers: Statistical assignment and migration of virtual machines,” in Lecture Notes in Computer Science, 2011, vol. 6852 LNCS, no. PART 1, pp. 407–418.
56. Mao, M., Li, J. and Humphrey, M., 2010, October, “Cloud Auto-Scaling with Deadline and Budget Constraints.” IEEE/ACM International Conference on Grid Computing
57. Y.B.Chen, et al., “Method and apparatus of dynamically allocating resources across multiple virtual machines,” 2013.
58. W. Wang and H. Chen, “An availability-aware virtual machine placement approach for dynamic scaling of cloud applications,” in IEEE 9th International Conference on Ubiquitous Intelligence and Computing, 2012, pp. 509–516.
59. J. Yang, C. Liu, Y. Shang, Z. Mao, and J. Chen, “Workload Predicting-Based Automatic Scaling in Service Clouds,” 2013.
60. S. Bart et al., “Resilient application placement for geo-distributed cloud networks Reference: Institutional repository IRUA Resilient Application Placement for Geo-Distributed Cloud Networks,” pp. 14–31, 2017.
61. M. Jammal, A. Kanso, and A. Shami, “CHASE: Component High Availability-Aware Scheduler in Cloud Computing Environment,” 2015.
62. C. T. Yang, J. C. Liu, C. H. Hsu, and W. L. Chou, “On improvement of cloud virtual machine availability with virtualization fault tolerance mechanism,” Journal of Supercomputing, vol. 69, no. 3, pp. 1103–1122, Sep. 2014.
63. D. Mani and A. Mahendran, “Availability modeling of a fault-tolerant cloud computing system,” International Journal of Intelligent Engineering and Systems, vol. 10, no. 1, pp. 154–165, 2017.
64. J. R. and M. S. S. S. C. Vijay, “Providing application high availability in highly-available virtual machine environments,” 2013.
65. L. A. 55. Chodroff, B.E., Edelstein, A.S., Harper, R.E., Kanaskar, M., Schildhauer, W.F., Schneider, M.S., and Tomek, “Controlling an availability policy for a virtual machine based on changes in a real-world environment,” 2011.
66. E. Pinheiro, R. Bianchini, E. v Carrera, and T. Heath, “Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems,” 2001.
67. C.-C. Lin, P. Liu, and J.-J. Wu, “Energy-efficient Virtual Machine Provision Algorithms for Cloud Systems,” 2011
68. R. Raghavendra, et al., No ‘‘power’’ struggles: coordinated multi-level power management for the data center, SIGARCH Computer Architecture News 36 (1) 48-59 2008
69. Jeyarani, Rajarathinam, N. Nagaveni, and R. Vasanth Ram. "Design and implementation of adaptive power-aware virtual machine provisioner (APA-VMP) using swarm intelligence." Future Generation Computer Systems 28.5: 811-821. 2012
70. Jung, Gueyoung, et al. "Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures." 2010 IEEE 30th International Conference on Distributed Computing Systems. IEEE, 2010.
71. N. Kim, J. Cho, and E. Seo, “Energy-credit scheduler: An energy-aware virtual machine scheduler for cloud systems,” Future Generation Computer Systems, vol. 32, pp. 128–137, 2014.
72. A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing,” Future Generation Computer Systems, vol. 28, pp. 755–768, 2012.
73. K. Le, J. Zhang, J. Meng, R. Bianchini, Y. Jaluria, and T. D. Nguyen, Reducing Electricity Cost Through Virtual Machine Placement in High-Performance Computing Clouds. 2011.
74. Y.-C. Lee et al., “Energy-efficient utilization of resources in Cloud computing systems,” Article in The Journal of Supercomputing, 2010.
75. B. Li, J. Li, J. Huai, T. Wo, Q. Li, and L. Zhong, “EnaCloud: An Energy-saving Application Live Placement Approach for Cloud Computing Environments,” 2009.
76. H. Liu, C.-Z. Xu, H. Jin, J. Gong, and X. Liao, Performance and Energy Modeling for Live Migration of Virtual Machines. 2011.
77. L. Liu et al., GreenCloud: A New Architecture for Green Data Center General Terms. 2009.
78. Kansal, Nidhi Jain, and Inderveer Chana. "An empirical evaluation of energy-aware load balancing technique for a cloud data center." Cluster Computing 21.2: 1311-1329. 2018
79. S. K. Mishra, D. Puthal, B. Sahoo, S. K. Jena, and M. S. Obaidat, “An adaptive task allocation technique for green cloud computing,” Journal of Supercomputing, vol. 74, no. 1, pp. 370–385, Jan. 2018.
80. F. Juarez, J. Ejarque, and R. M. Badia, “Dynamic energy-aware scheduling for parallel task-based application in cloud computing,” Future Generation Computer Systems, vol. 78, pp. 257–271, 2018.
81. S. Ismaeel and A. Miri, “Real-Time Energy-Conserving VM-Provisioning Framework for Cloud-Data Centers,” in IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0765–0771.
82. S. Ismaeel, R. Karim, and A. Miri, “Proactive dynamic virtual-machine consolidation for energy conservation in cloud data centers,” Journal of Cloud Computing, vol. 7, no. 1, Dec. 2018.
83. S. S. Gill et al., “Holistic resource management for sustainable and reliable cloud computing: An innovative solution to a global challenge,” The Journal of Systems & Software, vol. 155, pp. 104–129, 2019.
84. S. Ismaeel, A. Al-Khazraji, and A. Miri, “Energy-Consumption Clustering in Cloud Data Centre,” in 3rd MEC International Conference on Big Data and Smart City (ICBDSC), IEEE, 2016, pp. 1–6.
85. M. Kunbaz and J. Bieser, “An innovative approach for harmonizing availability and energy efficiency in data centers: A case study,” International Review of Applied Sciences and Engineering, vol. 10, no. 1, pp. 93–99, 2019.
86. S. M. Ghoreyshi, “Energy-efficient resource management of cloud datacenters under fault tolerance constraints,” 2013.
87. D. Sitaram, H. L. Phalachandra, S. Gautham, H. v. Swathi, and S. Tp, “Energy-efficient Data Center management under availability constraints,” 9th Annual IEEE International Systems Conference, SysCon 2015 - Proceedings, pp. 377–381, 2015.
88. V. Matko and B. Brezovec, “Improved Data Center Energy Efficiency and Availability with Multilayer Node Event Processing,” 2018.
89. A. Greenberg et al., “VL2: A scalable and flexible data center network,” Communications of the ACM, vol. 54, no. 3, pp. 95–104, Mar. 2011.
90. R. N. Mysore et al., PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric. 2009.

91. M. Al-Fares, A. Loukissas, and A. Vahdat, “A scalable, commodity data center network architecture,” ACM SIGCOMM Computer Communication Review, vol. 38, no. 4, p. 63, 2008.
92. A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, Towards a Next-Generation Data Center Architecture: Scalability and Commoditization. 2008.
93. C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers. 2008.
94. D. Li, C. Guo, H. Wu, K. Tan, Y. Zhang, and S. Lu, “FiConn: Using Backup Port for Server Interconnection in Data Centers.”
95. N. Farrington et al., Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. 2010.
96. C. Kachris and I. Tomkos, “A Survey on Optical Interconnects for Data Centers,” IEEE COMMUNICATIONS SURVEYS & TUTORIALS, vol. 14, no. 4.
97. H. J. Chao, K.-L. Deng, and Z. Jing, “PetaStar: A Petabit Photonic Packet Switch,” IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, vol. 21, no. 7, 2003.
98. A. Gladisch, “Power efficiency of optical versus electronic access networks Wireless Big Data for Smart Network Optimization and Resource Management View project,” 2014.
99. G. Wang et al., c-Through: Part-time Optics in Data Centers. 2010.
100. M Żal - Optical Switching and Networking, “Energy-efficient optical switching nodes based on banyan-type switching fabrics,” Elsevier, 2019.
101. M. Gupta and S. Singh, “Greening of the internet,” in Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM ’03, 2003, p. 19.
102. L. Shang, L.-S. Peh, and N. K. Jha, “Dynamic Voltage Scaling with Links for Power Optimization of Interconnection Networks.”
103. T. Horvath, T. Abdelzaher, K. Skadron, and X. Liu, “Dynamic Voltage Scaling in Multi-tier Web Servers with End-to-end Delay Control.”
104. N. Vasić and D. Kostić, “Energy-aware traffic engineering,” in Proceedings of the e-Energy 2010 - 1st Int’l Conf. on Energy-Efficient Computing and Networking, 2010, pp. 169–178.
105. Tomás, Luis, et al. "Network-aware meta-scheduling in advance with the autonomous self-tuning system." Future Generation Computer Systems 27.5: 486-497. 2011
106. S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, “Reducing Network Energy Consumption via Sleeping and Rate-Adaptation.”
107. E. K. M. Koseoglu, “Optimizing Energy Consumption in Cloud Data Center Resources,” Future Generation Computer Systems, vol. 4, pp. 576–589, 2010.
108. A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing,” Future Generation Computer Systems, vol. 28, pp. 755–768, 2012.
109. A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, “The Cost of a Cloud: Research Problems in Data Center Networks.”
110. C. Panarello, A. Lombardo, G. Schembra, and M. Mellia, “Energy-saving and network performance: a trade-off approach,” pp. 41–50, 2010.
111. M. Allman, K. Christensen, B. Nordman, and V. Paxson, “Enabling an Energy-Efficient Future Internet Through Selectively Connected End Systems *.”
112. K. Christensen, C. Gunaratne, B. N.-C., 2004, “The next frontier for communications networks: power management,” Elsevier.
113. X. Meng, V. Pappas, and L. Zhang, Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement.
114. A. Stage and T. Setzer, CLOUD: Network-Aware Migration Control and Scheduling of Differentiated Virtual Machine Workloads. 2009.
115. A. Beloglazov and R. Buyya, “Energy-Efficient Resource Management in Virtualized Cloud Data Centers.”
116. W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, “VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers,” Computer Networks, vol. 57, no. 1, pp. 179–196, 2013.
117. B. Addis, A. Capone, G. Carello, L. G. Gianoli, and B. Sansò, “Energy management through optimized routing and device powering for greener communication networks,” IEEE/ACM Transactions on Networking, vol. 22, no. 1, pp. 313–325, Feb. 2014.
118. P. Charalampou and E. D. Sykas, “An SDN Focused Approach for Energy-Aware Traffic Engineering in Data Centers.”
119. Kaljic, Enio, et al. "A Survey on Data Plane Flexibility and Programmability in Software-Defined Networking." IEEE Access 7: 47804-47840. 2019
120. T. Koponen et al., “Network Virtualization in Multi-tenant Datacenters Network Virtualization in Multi-tenant Datacenters,” p. 203, 2014.
121. J. P. G. Sterbenz et al., “Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines,” Computer Networks, vol. 54, pp. 1245–1265, 2010.
122. M. Alizadeh et al., “CONGA: Distributed Congestion-Aware Load Balancing for Datacenters,” 2014.
123. C. Meixner, Cloud Network Resiliency in Optical Networks. 2016.
124. M. Handley et al., “Re-architecting datacenter networks and stacks for low latency and high performance,” 2017.
125. C. Hernandez Benet, A. J. Kassler, T. Benson, and G. Pongracz, “MP-HULA: Multipath Transport Aware Load Balancing Using Programmable Data Planes,” vol. 18, 2018.
126. W. M. Mellette, R. Das, Y. Guo, R. Mcguinness, A. C. Snoeren, and G. Porter, “Expanding across time to deliver bandwidth efficiency and low latency,” 2019.
127. H. Almasi, H. Rezaei, M. U. Chaudhry, and B. Vamanan, “Pulser: Fast congestion response using explicit incast notifications for Data Center networks,” IEEE Workshop on Local and Metropolitan Area Networks, vol. 2019-July, 2019.
128. R. Wang, S. Mangiante, A. Davy, L. Shi, and B. Jennings, QoS-aware Multipathing in Datacenters Using Effective Bandwidth Estimation and SDN.
129. Yang, Hyunsik, Cong-Phuoc Hoang, and Younghan Kim. "Architecture for Virtual Network Function's High Availability in Hybrid Cloud Infrastructure." 2018 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN). IEEE, 2018.
130. A. Dalvandi, “Time aware virtual machine placement and routing for power efficiency in data centers,” 2016.
131. I. Bouras, R. Figueiredo, M. Poss, and F. Zhou, “Minimizing energy and link utilization in ISP backbone networks with multi-path routing: a bi-level approach,” Optimization Letters, 2019.
132. R. Moreno-Vozmediano, R. S. Montero, E. Huedo, and I. M. Llorente, “Efficient resource provisioning for elastic Cloud services based on machine learning techniques,” Journal of Cloud Computing, vol. 8, no. 1, Dec. 2019.
133. Ghuman, Karanjot Singh, and Amiya Nayak. "Per-packet based energy-aware segment routing approach for data center networks with SDN." 2017 24th International Conference on Telecommunications (ICT). IEEE, 2017.
134. L. Chiaraviglio, et al., “Lifetime-Aware Cloud Data Centers: Models and Performance Evaluation,” 2016.
135. Chiaraviglio, Luca, et al. "Is green networking beneficial in terms of device lifetime?." IEEE Communications Magazine 53.5: 232-240. 2015
136. P. Wiatr, J. Chen, P. Monti, and L. Wosinska, “Energy Efficiency and Reliability Tradeoff in Optical Core Networks,” 2014.
137. Wiatr, Pawel, Paolo Monti, and Lena Wosinska. "Power savings versus network performance in dynamically provisioned WDM networks." IEEE Communications Magazine 50.5: 48-55. 2012
138. Li, Zhenhua, and Yuanyuan Yang. "RRect: A novel server-centric data center network with high power efficiency and availability." IEEE Transactions on Cloud Computing 2018
139. T. Bisson, S. A. Brandt, and D. D. E. Long, “A Hybrid Disk-Aware Spin-Down Algorithm with I/O Subsystem Support.”
140. Y. Deng, “Exploiting the performance gains of modern disk drives by enhancing data locality,” Information Sciences, vol. 179, pp. 2494–2511, 2009.
141. Y. Deng, “What is the Future of Disk Drives, Death or Rebirth?” ACM Computing Surveys, vol. X, 2010.
142. A. G. Yoder, “Energy Efficient Storage Technologies for Data Centers,” in Workshop on Energy-Efficient Design, 2010.
143. Bohrer, Patrick J., et al. "Data storage on a multi-tiered disk system." U.S. Patent No. 6,925,529. 2 Aug. 2005.
144. D. Cohn and M. Kistler, “Multiple disk data storage systems for reducing power consumption,” 2007.
145. A. Hylick, R. Sohan, A. Rice, and B. Jones, “An Analysis of Hard Drive Energy Consumption.”
146. Y. Deng, “Exploiting the performance gains of modern disk drives by enhancing data locality,” Information Sciences, vol. 179, pp. 2494–2511, 2009.
147. Y. Kim, S. Gurumurthi, and A. Sivasubramaniam, “Understanding the performance-temperature interactions in disk I/O of server workloads,” Proceedings - International Symposium on High-Performance Computer

Architecture, vol. 2006, pp. 179–189, 2006.
148. “Saving Data center power by reducing hdd spin speed,” 2009.
149. S. W. Son, G. Chen, and M. Kandemir, “Disk Layout Optimization for Reducing Energy Consumption *.”
150. Otoo, Ekow, Doron Rotem, and Shih-Chiang Tsao. "Dynamic data reorganization for energy savings in disk storage systems." International Conference on Scientific and Statistical Database Management. Springer, Berlin, Heidelberg, 2010.
151. Khatib, Mohammed Ghiath, and Damien Cyril Daniel Le Moal. "Performance-aware power capping control of data storage devices." U.S. Patent Application 10/146,293, filed December 4, 2018.
152. D. Colarelli, D. Grunwald, and M. Neufeld, “The Case for Massive Arrays of Idle Disks (MAID),” 2002.
153. Q. Zhu, F. M. David, C. F. Devaraj, Z. Li, Y. Zhou, and P. Cao, “Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management.”
154. L. Ganesh, H. Weatherspoon, M. Balakrishnan, and K. Birman, “Optimizing Power Consumption in Large Scale Storage Systems.”
155. D. Li, P. Gu, H. Cai, and J. Wang, “EERAID: Energy Efficient Redundant and Inexpensive Disk Array.”
156. X. Yao and J. Wang, “RIMAC: A novel redundancy-based hierarchical cache architecture for energy-efficient, high-performance storage systems,” in Proceedings of the 2006 EuroSys Conference, 2006, pp. 249–262.
157. D. Narayanan, A. Donnelly, and A. Rowstron, “Write Off-Loading: Practical Power Management for Enterprise Storage.”
158. E. Pinheiro and R. Bianchini, “Energy Conservation Techniques for Disk Array-Based Servers,” 2004.
159. Zedlewski, John, et al. "Modeling Hard-Disk Power Consumption." FAST. Vol. 3. 2003.
160. J. S. Bucy, J. Schindler, S. W. Schlosser, and G. R. Ganger, “The DiskSim Simulation Environment Version 4.0 Reference Manual,” 2008.
161. Allalouf, M., Y. Arbitman, M. Factor, R. I. Kat, K. Meth, and D. Naor. "Storage modeling for power estimation, in proceedings of The Israeli Experimental Systems Conference (SYSTOR’09), May 4-6." Haifa, Israel. 2009
162. Sai Vishwas, Sameer Kulkarni, Phalachandra HL, Dinkar Sitaram, “SEA: SSD Staged Energy Efficient Object Storage System Architecture”, Elsevier Journal of Information Systems, submitted for publication.
163. A., D. K. and V. P. Elnably, “Reward scheduling for QoS in cloud applications,” 2012.
164. H. Lu, B. Saltaformaggio, R. Kompella, and D. Xu, “vFair: Latency-Aware Fair Storage Scheduling via Per-IO Cost-Based Differentiation.”
165. H. Wang and P. Varman, “Balancing Fairness and Efficiency in Tiered Storage Systems with Bottleneck-Aware Allocation,” Fast, pp. 229–242, 2014.
166. A. Elnably, H. Wang, A. Gulati, and P. Varman, “Efficient QoS for Multi-Tiered Storage Systems.”
167. E. v Carrera, E. Pinheiro, and R. Bianchini, “Conserving Disk Energy in Network Servers £,” 2002.
168. Wu, Lin, et al. "BOSS: An efficient data distribution strategy for object storage systems with hybrid devices." IEEE Access 5: 23979-23993. 2017
169. Kimmel, Jeffrey S., Steven R. Kleiman, and Steven C. Miller. "Hybrid media storage system architecture." U.S. Patent No. 9,134,917. 15 Sep. 2015.
170. M. Brunner, L. De, and H. De, “Locally providing cloud storage array services.,” 2016.
171. Cooper, Alastair, and Gordon D. Hutchison. "Controlling data storage in an array of storage devices." U.S. Patent No. 9,378,093. 28 Jun. 2016.
172. Sieklucki, Mark Robert, et al. "Computation refinement storage in a data storage system." U.S. Patent Application No. 10/198,319. 2019.
173. S. de Keyser, K., De Schrijver, F.J.L., and Blyweert, “Hierarchic Storage Policy for Distributed Object Storage Systems,” 2019.
174. I. Mohammed, G. Khatib, S. Jose, and U. S. Ci, “Performance-Aware Power Capping Control of Data Storage Devices,” 2018.
175. J. W., F. T. R., V. W. H., G. R. J., G. K., and G. M. A., S. T. L. Haines, “Data segregation in a storage device.,” 2016.
176. CBINSIGHTS Research Portal, “Future of Data Centers”, 2019
177. Shi, Weisong, et al. "Edge computing: Vision and challenges." IEEE Internet of Things Journal 3.5: 637-646, 2016
178. Shi, Weisong, and Schahram Dustdar. "The promise of edge computing." Computer 49.5: 78-81, 2016

AUTHORS PROFILE

Phalachandra HL received his Bachelor's degree from Karnataka University and Masters from BITS Pilani, India. His interests are in Cloud Computing, Energy efficiencies in Data Center, Storage Area Networks and Network Management. He has worked extensively in the industry for over 25+ years heading storage groups at EMC, HP and Research roles at Bharat Electronics. He currently teaches at PES University from the last 7 years.

Dr. Dinkar Sitaram received his Ph.D. in Computer Science from the University of Wisconsin Madison. Cloud & Hybrid Clouds, Speech Recognition and Intelligent Big Data systems are his areas of interest. The books ’Moving to the Cloud” and ’Multimedia Servers’ have been written by him. He has over 30 patents and 45 publications. An IBM Corporate Innovation Awardee was also the CTO of Novell Software, Andiamo Systems, and HP-STSD, India, before joining PES University.