

Seeing into a Public Cloud: Monitoring the Massachusetts Open Cloud

Ata Turk, Hao Chen, Ozan Tuncer, Hua Li, Qingqing Li, Orran Krieger, and Ayse K. Coskun

Boston University

Abstract

Cloud users today have little visibility into the performance characteristics, power consumption, and utilization of cloud resources; and the cloud has little visibility into user application performance requirements and critical metrics such as response time and throughput. This paper outlines new efforts to reduce the information gap between the cloud users and the cloud. We first present a scalable monitoring platform to collect and retain rich information on a regional public cloud. Second, we present two motivating use cases that leverage the collected information: (1) Participation in emerging smart grid demand response programs in order to reduce datacenter energy costs and stabilize power grid demands, (2) budgeting available power to applications via peak shaving. This work is done in the context of the Massachusetts Open Cloud (MOC), a new public cloud project that has a central goal of enabling cloud research.

1 Introduction

One of the main arguments for utilizing public clouds is that they are environments where developers with innovative ideas can try out their new services without investing in huge amounts of hardware [1]. However, public cloud providers do not expose performance characteristics, power consumption, and utilization of the cloud software and hardware resources to applications/users, and the public cloud has little visibility into the application requirements and critical metrics. As a result, optimization at any level requires reverse engineering techniques to discover information about the other levels.

A good example of reverse engineering is presented by Conley et al., who aimed to identify the best performing instance type for sorting 100TB of data on Amazon EC2 [2]. After evaluating 48 different available instance types and spending \$50,000 along the way, they were able to optimize their application. Unfortunately, it is not clear if their results are repeatable since the performance may well depend on the utilization of different components of the cloud infrastructure.

Our fundamental goal is to provide a monitoring infrastructure that can expose rich information about all layers of the cloud (facility, network, hardware, OS, middleware, application, and user layers [3]) to all the other layers. In this way, we hope to reduce the need for costly reverse engineering or coarse-grained assumptions. We believe that such detailed, multi-layered information is key to developing intelligent, realistic performance and energy optimization techniques. For example, to rapidly reduce datacenter power consumption while maintaining service level agreements (SLAs), we believe that it is essential to be able to identify candidate application components and their power/performance characteristics, and then determine the power saving technique to be applied to those components under SLA constraints.

One emerging real-life use case of this monitoring infrastructure is datacenter participation in smart grid programs. Today's power market provides incentives for electricity consumers to perform demand response, so as to help stabilize the power grid and integrate ever increasing renewable energy generation. Datacenters offer a unique opportunity to provide such demand response, as they are not only significant electricity consumers worldwide [4], but also can provide the necessary flexibility in their power consumption by leveraging power management techniques and workload scheduling policies. To achieve efficient participation, information such as power consumption and performance characteristics needs to be collected and used in power and workload control.

In this paper, we first present our design for a scalable cloud monitoring system that can collect, mediate, and expose data coming from multiple layers of the cloud. We then provide two motivating use cases that exercise our monitoring infrastructure and showcase the benefits of collecting and retaining information from different cloud layers. Information from the monitoring platform enables the cloud datacenter to efficiently perform *demand response* and *peak power shaving*, which not only help in stabilizing the power grid, but also enables significant energy and monetary savings.

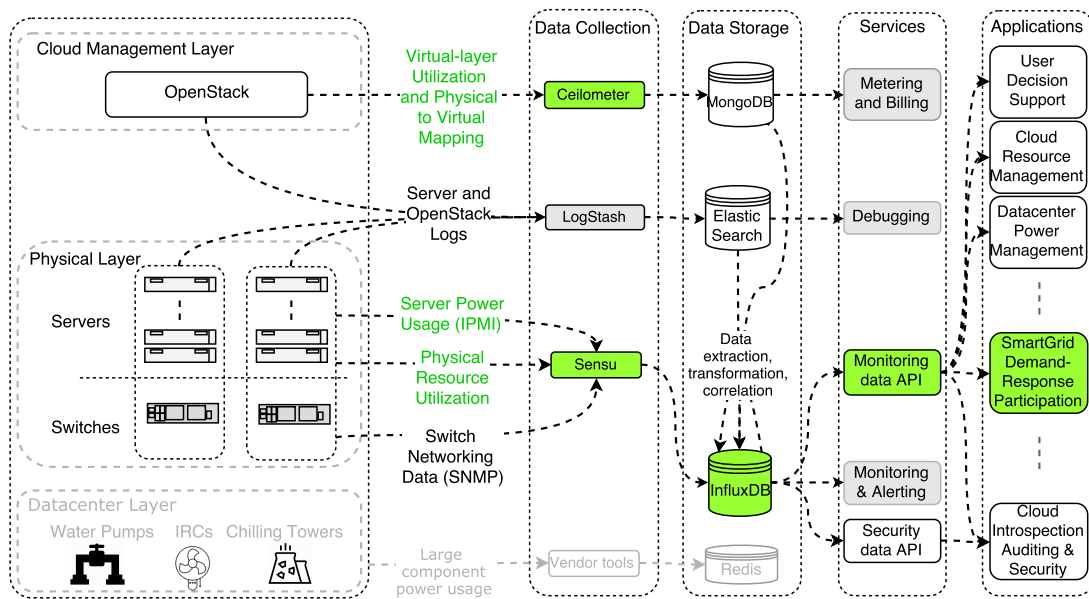


Figure 1: Architecture of the MOC Monitoring system. Green colored monitoring components support the energy market / smart grid demand response participation application described in Section 4.

2 Related Work

There are commercial solutions running on top of public clouds that monitor virtual cloud usage. For example, Amazon CloudWatch [5] is a service that monitors virtual resources of users such as Amazon EC2 instances. However, such services do not provide physical resource monitoring, and thus, lack a holistic view of the cloud.

A number of academic studies have been done to collect and retain information across different layers of the cloud and to provide a complete monitoring story [6, 7, 8, 9]. These studies, however, either focus on providing enhanced capabilities to cloud providers without exposing lower layer monitoring information to end users and researchers (which is essential for understanding resource usage or for designing power/performance management methods), or they focus on specific end applications. For example, PCMONS [6] does not collect the information needed to map virtual resources to physical resources, and GMonE [7] is not appropriate for a general purpose Infrastructure-as-a-Service (IaaS) cloud such as MOC as it requires installing agents on each VM and setting up a database for each user. The primary goal of our work, on the other hand, is to develop a platform that can be used by a wide variety of applications.

3 Monitoring the MOC

MOC monitoring system aims to scalably collect, store and serve integrated information from all layers of the cloud to further cloud research and to enable intelligent applications (such as power demand response participation) that reduce costs or improve service quality.

Proposed monitoring architecture assumes a standard IaaS cloud setup composed of switches, storage, and servers on the physical layer, and managed by OpenStack [10] on the virtual layer. The MOC is running in the Massachusetts Green High Performance Computing Center¹ (MGHPCC), a 15 megawatt datacenter.²

Figure 1 shows the components of the proposed monitoring architecture, which is composed of four layers:

- i. Data collection layer, where monitoring information from different layers are collected in a scalable and low-overhead manner;
- ii. Data retention & consolidation layer, where collected monitoring information is persisted and then consolidated in a time-series database (InfluxDB [11]);
- iii. Services layer, in which resides both IaaS services such as alerting and metering as well as privacy preserving API services to expose monitoring data to unprivileged users;
- iv. Advanced monitoring applications layer, where a wide variety of value-added services that utilize the monitoring data operate.

3.1 Collecting Monitoring Data

As seen in Fig. 1, the physical server resource utilization (e.g., CPU, memory, disk utilization), power metrics (e.g., consumed power, fan speeds, temperature sen-

¹<http://www.mghpcc.org>

²There is an infrastructure in place at MGHPCC that collects and retains the datacenter layer monitoring data (gray layer in Figure 1), but this is currently not integrated into our MOC monitoring infrastructure.

sors), and switch network utilization metrics (e.g., incoming/outgoing traffic on switch ports) are collected via Sensu [12]. Sensu periodically tracks resources/services and provides feedback or alert notifications to system managers. It harvests data by running agents on servers it monitors. The data collected by clients are filtered and passed to the Sensu server over a separate message bus (RabbitMQ). The virtual layer utilization information is collected using Ceilometer [13]. Ceilometer is an OpenStack project designed for enabling metering and billing of OpenStack IaaS clouds. It tracks the resource utilization of OpenStack users by monitoring notifications sent by OpenStack services or by running agents that poll the infrastructure. The power and performance data collected by Sensu and Ceilometer (highlighted with green in Fig. 1) are used in supporting the applications we present in Section 4.

Even though not utilized by the example applications we describe in this study, the `syslogs` of the individual compute servers and the logs of the OpenStack services are also collected by our monitoring system. We use LogStash, a log shipping system that collects logs from given sources, parses and stores them into a database in a structured form for this purpose.

Since our data collection services are stateless, scaling them is simply a function of running more collector processes and dividing the workload among them.

3.2 Data storage and consolidation

Ceilometer and LogStash are already coupled with scalable storage systems MongoDB [14] and Elasticsearch [15], respectively. Since Sensu is an alerting system, normally, data collected by Sensu is not stored in any database. We use the time-series database InfluxDB [11] for persisting the information collected by Sensu. InfluxDB also has clustering support for better scalability.

We also correlate and consolidate various types of data collected by Ceilometer and LogStash in InfluxDB. This consolidated database enables formulating complex queries, such as queries that expose the cross-layer state of the components of the datacenter they are utilizing to cloud users, or queries that enable cloud administrators to understand the impacts of changes made in the physical layer on virtual and application-layer performance.

3.3 Monitoring services

In our design, data collected from different layers can be used for performing various services of MOC. A metering service queries the MongoDB database populated by Ceilometer to create user cloud usage reports for given periods of time. The logging data collected by LogStash and indexed/served by Elasticsearch provides a keyword-based search interface including various filters to the MOC admins to be used for debugging.

Sensu goes through a set of checks and creates alerts if any anomalies in systems are observed.

In addition to the above mentioned fundamental services, proposed MOC monitoring system also provides two APIs that enable querying of the consolidated data in InfluxDB. The *Security API* provides detailed networking information as well as interaction and packet metadata information of users who opt-in to supply this data. The *Monitoring API* simply exposes all correlated performance, resource utilization and power data.

3.4 Applications

A unique feature of our monitoring system design is to offer transparency to cloud users and administrators for a number of applications. For example, we are working on providing privacy preserving APIs to end-users that expose the state of the physical resources that their VMs are hosted over. This will enable users to achieve and automate stable and repeatable performance. We are also working on correlating and exposing virtual layer performance and physical layer power utilization data to cloud administrators. This will enable participation of cloud datacenters in energy market demand-response programs. In the next section, we will present our initial analysis on the feasibility of this latter application.

4 Case Studies: Cloud Power Management Schemes and Demand Response Participation Enabled by Monitoring

Today's datacenters are capable of providing significant flexibility in power consumption owing to the advancements in dynamic power management techniques in servers as well as to the increasing diversity of cloud jobs ranging from latency-critical transactional jobs to delay-tolerant jobs. We believe that if we couple this flexibility with a systematic and accurate cloud power and performance monitoring infrastructure, cloud datacenters would be promising candidates for participating in emerging energy market demand response programs such as *regulation service reserves (RSR)* [16, 17] and *peak power shaving* [18, 19]. This participation not only can help stabilize the power grid and enable substantial electricity generation from renewables, but also could decrease rapidly growing energy costs of cloud datacenters due to the incentive credits offered by energy market operators for reserve provisioning.

Ability to monitor the cloud power and performance systematically and accurately is essential for participation in demand response programs, as real-time dynamic power control policies for response provisioning typically require feedback on power consumption and throughput. Our monitoring infrastructure accurately measures and collects this information, enabling the public cloud to join in the power market. We demonstrate

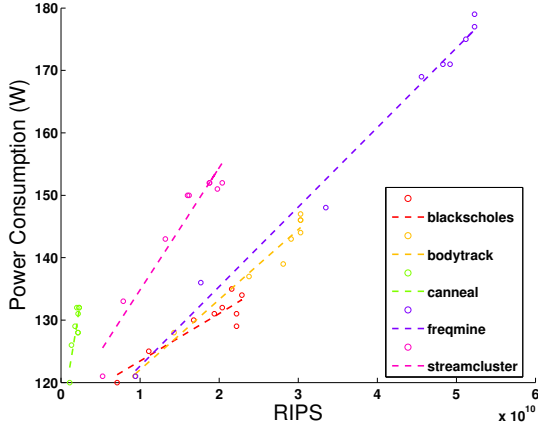


Figure 2: Power consumption vs. job RIPS when a varying number of vCPUs are allotted to the VM. Dots represent data points and dash lines represent the linear fit.

two demand response participation case studies: RSR provisioning and peak shaving. We simulate these two cases using data collected from our MOC monitoring infrastructure and by leveraging the dynamic control policies we proposed recently [17, 19]. These case studies indicate the feasibility of having an open cloud participate in demand response and serve as a proof-of-concept for deploying realistic dynamic power control policies.

4.1 Power and Performance Management on the MOC

A rich set of control knobs are available in a datacenter, including job scheduling or power capping of servers via voltage-frequency scaling and via changing resources allocated to each VM. In this work, we use the number of vCPUs allocated to a VM as an example, and also as a proxy to mimic a combination of various control knobs. Note that as we build upon the proposed infrastructure, a larger number of knobs and metrics will be available.

As the MOC is a virtualized platform, the number of vCPUs strongly impacts both power consumption and server throughput (assume $\#vCPUs \leq \#pCPUs$). In our experiments, we run PARSEC-2.1 benchmarks [20] on a VM, varying the number of allocated vCPUs (2, 4, 8, 12, 16, 20, 24, 28, 32) in each run. We measure the runtime as well as the total number of instructions of the jobs, and calculate Retired Instructions Per Second (RIPS), which is a widely used metric for evaluating performance. We place the VM on an isolated server and measure its power consumption through IPMI. The relation between RIPS and power consumption is shown in Figure 2.

The figure shows a linear fit between RIPS and power consumption with a mean square error of less than 3% across all jobs, which matches the results in previous studies (e.g., [21]). Note that neither RIPS nor power consumption increases further beyond a large number of vCPUs (e.g., $\#vCPUs \geq 20$), where another resource (e.g., memory) becomes the bottleneck.

4.2 Case Study 1: Regulation Service Reserve Provision in the Cloud

The increasing amount of intermittent renewable power sources leads to challenges in power grid stabilization. The demand-side power capacity reserve provisioning has emerged as a major sustainable solution for the integration of such power sources into the grid, and it has already been pursued by the largest US Independent System Operator (ISO), PJM, since 2006 [22].

For the data center participation in reserve provisioning, we target regulation service reserves (RSR) with hour-ahead signals [16, 23], as a typical datacenter is capable of regulating its power at the required reserve frequency (i.e., every few seconds) and the credit for providing such reserves is high. In RSR hour-ahead provision, each potential provider bids an average power consumption \bar{P} , and a reserve value R to the ISO for the hour ahead. After bidding is complete and prices settle, the provider is charged for its power consumption as $\Pi^E \bar{P} - \Pi^R R$ in the following hour, where Π^E and Π^R are market-cleared prices for energy consumed and reserve credited, respectively. In other words, the provider receives $\Pi^R R$ credits for providing R amount of reserves. Each RSR provider is obligated to regulate its dynamic power consumption, $P(t)$, to track an RSR signal, $y(t)$, that is broadcasted every 4 seconds by the ISO [22], such that $P(t) \approx \bar{P} + y(t)R$. The credit is reduced based on the tracking error, and the provider may lose the contract if the tracking error is too large. The signal $y(t)$ is generated every 4 seconds by ISO to control the power grid. The providers only know that $y(t)$ is a random variable between $[-1, 1]$ with an average of zero over long time intervals, and that $y(t)$ is updated in increments that do not exceed $\pm R/(\tau/4)$, where τ is within 100-300 seconds. $y(t)$ follows a well-behaved Markov model whose transition probabilities can be calculated in advance.

We simulate a recently proposed dynamic datacenter RSR participation policy [17] that tracks the power signal while guaranteeing reasonable performance, using data collected from our MOC monitoring infrastructure. The control knob, power consumption, and the job service rate (i.e., RIPS) have been studied in Section 4.1. Power consumption of a target MOC server in low-power (sleep) state is 40W, and waking up a server from sleep state takes 2 minutes. We assume that the cloud utilization is around 50%, which is a typical scenario in practice. Since the control policy is scalable with the size of the datacenter, we conduct experiments by assuming that 100 servers are participating in RSR.

Figure 3 provides results of RSR provisioning when servicing two workloads in the MOC: *bodytrack* and *freqmine*³. We select these two jobs because they have very different profiles in both power consumption and perfor-

³We simulate a homogeneous load case, where the same type of application arrives at the cloud over time in each experiment, as in the typical scenario of a cluster servicing a specific kind of load.

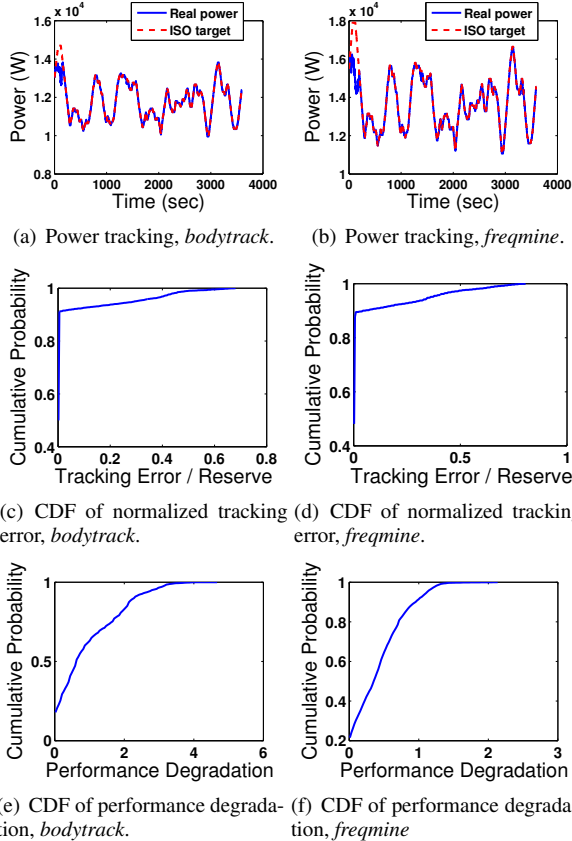


Figure 3: Results of RSR provisioning for *bodytrack* and *freqmine*. Figures 3(a) and 3(b) show the power tracking performance. The blue curve is the real power consumption and the red curve is the ISO target power (i.e., the RSR signal). Figures 3(c) and 3(d) are the cumulative distribution function (CDF) of the tracking error $\varepsilon(t)$, normalized by the reserve value R . Figures 3(e) and 3(f) are CDF of performance degradation of all serviced jobs.

performance: *freqmine* has a running time several times longer than *bodytrack* and higher peak power consumption.

Figures 3(a) and 3(b) are results of 1-hour power tracking of the cloud. We can see that our 100-server cloud is capable of tracking the RSR signal very accurately for both workloads. The only tracking gap appears when the magnitude of the RSR signal is too high and there is a lack of jobs to execute. Such a tracking gap can be filled either by running synthetic workloads or by leveraging energy storage devices such as an uninterruptible power supply (UPS).

The CDFs of the tracking errors $\varepsilon(t)$ normalized w.r.t. the reserve R are shown in Figures 3(c) and 3(d). The tracking error is close to 0 in more than 90% of time for both jobs, which further demonstrates the strong capacity of our cloud to track the RSR signal and provide reserves.

Figures 3(e) and 3(f) show the performance degradation, which is defined as the ratio of the increase in job servicing time due to participating in demand response to

the job servicing time without power capping and without any waiting in the job queue. More than 90% of the incoming jobs can be serviced with a degradation ratio smaller than 2.5 for *bodytrack*, while for *freqmine*, 90% of the degradation is close to 1. *Freqmine* has better performance than *bodytrack* because of its longer runtime, which makes the job waiting time in the queue relatively small, and thus the overall degradation is smaller.

Finally, we check the potential credits received by our cloud for RSR participation. Recall that the hourly electrical cost is $\Pi^E \bar{P} - \Pi^R R$, in which we have $\Pi^E \approx \Pi^R$ in today's market [22], and thus the monetary savings can be roughly estimated using R/\bar{P} . From the results, R/\bar{P} is 24.5% for *bodytrack*, and 30.5% for *freqmine*, both of which are highly promising. *Freqmine* has larger savings because it has a larger dynamic power range as mentioned before, which provides more flexibility for regulation. Overall, these results demonstrate promising profits of a cloud to provide RSR. Only scalable, multi-layer monitoring tools can realize this opportunity by providing detailed, correlated application performance and power data in real-time.

4.3 Case Study 2: Peak Power Shaving

The electricity bill of a datacenter is composed of: (1) the charge for the total energy consumed, (2) the charge for the peak power within the billing period (weekly, monthly, or annually) [18]. Market operators put extra charges on peak power to avoid power shortage during on-peak usage periods. In addition, as the power infrastructure and capacity for a datacenter has to be built based on the peak power requirements, reducing peak power also helps reduce this one-time infrastructure cost.

Peak shaving caps the power usage of a datacenter within an upper bound. It is usually implemented by using energy storage devices to modulate the power consumption during the off-peak and on-peak periods [18, 24]. Server-level dynamic power capping along with job scheduling techniques (e.g., load shedding and shifting) is another possible solution for peak shaving, when jobs serviced in the cloud are delay-tolerant. In this work, we focus on the latter solution as control policies in server power capping rely heavily on accurate monitoring.

We simulate a peak shaving policy [19] using the collected data from our monitoring infrastructure. Figure 4 shows results for 30% peak shaving; i.e., the new upper bound of the power consumption is limited to be 70% of the original peak. Note that the system may not have stable solutions (i.e., the length of job queue reaches infinity) if the shaved percentage is too large. Figure 4(a) and 4(b) show that leveraging the monitoring infrastructure, the policy enables our cloud to consume power at a level always below the peak power cap. Figure 4(c) shows that for *bodytrack*, close to 90% of the workload has no degradation at all, while more than 99% of the workload has a degradation ratio of less than 1, when the peak power is shaved by 30%. Similar results are

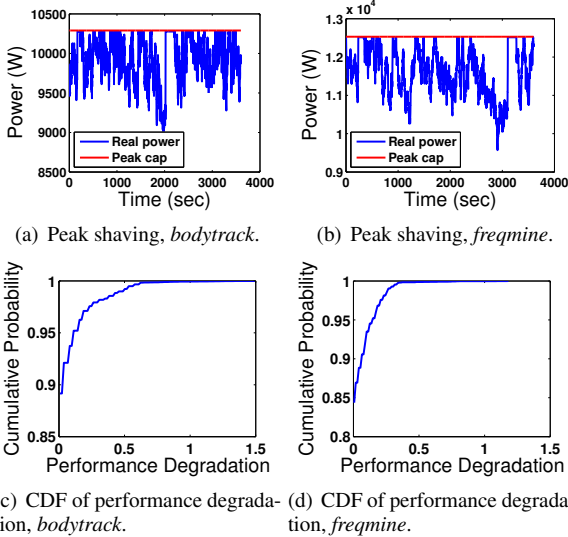


Figure 4: Results of peak shaving. Figure 4(a) and 4(b) show the real dynamic power in an hour capped by an upper limit (70% of the original peak) for *bodytrack* and *freqmine*, respectively. The blue curve is the real power consumption and the red curve is the power cap. Figure 4(c) and 4(d) are the CDF of performance degradation of all serviced jobs, for *bodytrack* and *freqmine*.

shown in Figure 4(d) for *freqmine*: more than 85% of the workload has no degradation at all, while more than 99% of the workload has a degradation ratio of less than 0.5. Overall, these results demonstrate that our cloud can strongly follow the peak power constraint while maintaining a low performance degradation, by leveraging the proposed policies and the data collected from the cloud.

5 Conclusion

We introduced the architecture for a scalable, multi-layer monitoring infrastructure for MOC, an open regional public cloud, and presented motivating examples that leverage the collected data for advanced performance and energy management. Our implementation is ongoing, yet the first results show substantial benefits of the proposed infrastructure for developing techniques that improve cloud sustainability. We believe our monitoring infrastructure will enable optimization of hardware and software components at all levels including application, OS, hypervisor/cloud, and physical infrastructure layers, and support a number of diverse research projects.

Acknowledgments

This work has been partially supported by Sandia National Laboratories.

References

[1] Michael Armbrust et al. Above the clouds: A Berkeley view of cloud computing. Technical Report UCB/ECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[2] Michael Conley, Amin Vahdat, and George Porter. Achieving cost-efficient, data-intensive computing in the cloud. In *ACM Symposium on Cloud Computing*, 2015.

[3] Jonathan Spring. Monitoring cloud computing by layer, part 1. *Security & Privacy, IEEE*, 9(2):66–68, 2011.

[4] H. Chen, B. Zhang, M. C. Caramanis, and A. K. Coskun. Data center optimal regulation service reserve provision with explicit modeling of quality of service dynamics. In *Decision and Control (CDC), 54th Annual Conference on*, pages 7207–7213, 2015.

[5] Cloudwatch. <https://aws.amazon.com/cloudwatch/>.

[6] De Chaves, Shirlei Aparecida, Rafael Brundo Uriarte, and Carlos Becker Westphall. Toward an architecture for monitoring private clouds. *Communications Magazine, IEEE*, 49(12):130–137, 2011.

[7] Jesús Montes, Alberto Sánchez, Bunjamin Memishi, María S Pérez, and Gabriel Antoniu. Gmone: A complete approach to cloud monitoring. *Future Generation Computer Systems*, 29(8):2026–2040, 2013.

[8] Javier Povedano-Molina, Jose M Lopez-Vega, Juan M Lopez-Soler, Antonio Corradi, and Luca Foschini. Dargos: A highly adaptable and scalable monitoring architecture for multi-tenant clouds. *Future Generation Computer Systems*, 29(8):2041–2056, 2013.

[9] JM Alcaraz Calero and Juan Gutierrez Aguado. Monpaas: An adaptive monitoring platform for a service for cloud computing infrastructures and services. *Services Computing, IEEE Transactions on*, 8(1):65–78, 2015.

[10] Omar Sefraoui, Mohammed Aissaoui, and Mohsine Eleudj. OpenStack: toward an open-source solution for cloud computing. *Intl. Journal of Computer Applications*, 55(3):38–42, 2012.

[11] InfluxDB. <https://influxdb.com/docs/v0.9/introduction/overview.html>.

[12] Sensu. <https://sensuapp.org/docs/0.20/overview>.

[13] Ceilometer. <http://docs.openstack.org/developer/ceilometer/>.

[14] MongoDB. <http://docs.mongodb.org/manual/>.

[15] Elasticsearch. <https://www.elastic.co/products/elasticsearch>.

[16] A. L. Ott. Experience with PJM market operation, system design, and implementation. *Power Systems, IEEE Trans. on*, 18(2):528–534, 2003.

[17] H. Chen, M. C. Caramanis, and A. K. Coskun. The data center as a grid load stabilizer. In *Design Automation Conference, 19th Asia and South Pacific (ASP-DAC)*, pages 105–112, 2014.

[18] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar. Benefits and limitations of tapping into stored energy for datacenters. In *Computer Architecture (ISCA), 38th Annual International Symposium on*, pages 341–351. IEEE, 2011.

[19] H. Chen, M. C. Caramanis, and A. K. Coskun. Reducing the data center electricity costs through participation in smart grid programs. In *IEEE International Green Computing Conference (IGCC)*, pages 1–10, 2014.

[20] B. Christian. Benchmarking modern multiprocessors. *Ph.D.Thesis. Princeton University*, 2011.

[21] H. Chen, C. Hankendi, M. C. Caramanis, and A. K. Coskun. Dynamic server power capping for enabling data center participation in power markets. In *Intl. Conf. on Computer-Aided Design (ICCAD)*, 2013.

[22] PJM (2013). market-based regulation [online]. <http://pjm.com/markets-and-operations/ancillary-services/mkt-basedregulation.aspx>.

[23] *PJM Manual 12: Balancing Operations*, www.pjm.com, 2012.

[24] B. Aksanli, E. Pettis, and T. Rosing. Architecting efficient peak power shaving using batteries in data centers. In *IEEE Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 242–253, 2013.