Abstract— As cloud computing becomes increasingly pervasive, the data center energy consumption attributable to cloud computing is climbing, despite the clarion call of action to reduce consumption and reverse environmental effects. At the same time, the rising cost of energy — due to regulatory measures enforcing a “true cost” of energy coupled with finite natural resources rapidly diminishing, resulting in scarcity — is refocusing IT leaders on efficiency and total cost of ownership (TCO), particularly in the context of the world-wide financial crisis. We propose a “smart metering” approach that encompasses all the stakeholders in the cloud computing ecosystem to achieve these twin goals of “energy conservation” and “demand response”. As such, this paper introduces our initial thoughts on “smart metering” and various implications and implementation ideas related to it.

I. INTRODUCTION

Cloud computing fulfills the long-held dream of computing as a utility [1] and thus represents an inflection point in the geography of computation and IT services delivery. This paradigm marks a fundamental yet massive shift from the traditional “desktop-as-a-platform” to “internet-as-a-platform” model. To achieve the infinite scalability, guaranteed performance and nearly “always-on” availability demands, these computing platforms are typically deployed in clusters of massive number of servers hosted in dedicated data centers. Each data center houses a large number of heterogeneous components for computing, storage, and networking, together with an infrastructure to distribute power and provide cooling. As the demand for cloud-based services drastically increasing in recent times, the energy consumption attributable to these services by data centers has also been skyrocketing. Recent reports also highlight this growing concern with data center energy consumption and show how current trends could make energy the dominant factor in the Total Cost of Ownership (TCO) [7]. For example, the approximately 6000 data centers in the United States consumed roughly 61 billion kilowatt-hours (kWh) of energy or about 1.5 percent of the total US electricity consumption in 2006, according to an EPA report [2]. Estimates also indicate that by 2011, data center energy consumption could nearly double [5]. Figure 1 shows how power is becoming the new limiting factor in data center costs.

On the other hand, the biggest challenge facing the environment today is global warming, caused by carbon emissions. About 98 percent of CO2 emissions (or 87 percent of all CO2-equivalent emissions from all greenhouse gases) can be directly attributed to energy consumption, according to a report by the EIA [6]. The data centers, often dubbed as “the SUVs of the tech world”, thus contribute in a larger scale to the global warming phenomena.

Alarmed by the growing concerns over the rising data center energy costs and its effect on environment, the entire IT ecosystem is quickly realizing these impacts and is making every attempt to achieve energy efficiency. The focus is on energy efficiency opportunities such as reducing the computational power of data centers through efficient application management, increasing the efficiency of servers, cooling systems, power supplies and distribution. Now with the recession pressuring operation budgets, environmental concerns waxing, and energy prices and constraints increasing, we believe that the time is ripe for a different form of doing energy evangelism for data centers. We need to shed the traditional view of how environmentally sustainable data centers are looked at and metered. Our most used approaches - primarily focused on infrastructure optimization- may be too narrow to deal with the power and environmental challenges of tomorrow. This requires an ongoing and holistic approach such as Smart Metering, which is eliciting growing interest among regulatory, legal and advocacy groups and is continuing to garner important commitments from all parts of the IT ecosystem spanning energy suppliers, hardware manufacturers, data centers, cloud service providers, third party software vendors, IT organizations, governments and even end customers.

In this paper, we propose that smart metering helps mitigate the power related costs and risks in the longer run. For example, consider the simple fact that it is risky and expensive to avoid redundancy in data centers, especially in the cases of power delivery and provisioning. We certainly do not afford to lose the entire site owing to power failure or network access failure and yet outages and blackouts do happen if redundancy and geo-diversity is not built into the data center ecosystem. Therefore, it is particularly valid, and quite useful to ask about the sustainability of electric grids in responding to the unrealistic demands from mega cloud
data centers, in these times of raising energy prices, limited energy supplies and increasing energy consumption.

We explore this nature and potential of “Smart Metering” in this paper. As such, the key contribution of this paper is two fold. First, we introduce “Smart Metering” as the central theme for optimizing the cloud-based data center infrastructure usage and discuss the concepts underlying it.

Second, we also outline how this vision would infer efficient, open and API based metering approaches. We present our initial thoughts on an extensible architecture with provisioning for various pricing mechanisms.

The rest of the paper is organized as follows. In Section 2, we discuss the various concepts, definitions and trends underlying the cloud computing paradigm, data center energy efficiency and smart metering. Section 3 explains a high level conceptual description of our vision model encompassing the various ecosystem partners and stakeholders, possible implementation methods, the key benefits to be realized and lists various pricing mechanisms to monetize our model. We finally summarize and lay future directions in Section 4 before concluding.

II. CONCEPTS AND TRENDS

The emergence of embarrassingly distributed cloud-based applications and services brings plenty of opportunities to improve cost, scale, reliability and performance. We look at some of these trends, and try to define and understand related concepts in this section.

A. Emergence of Cloud and Cloud based Services

The emergence of cloud computing represents a fundamental yet massive shift in “computing” as we have known [23]. In other words, it is the next natural step in the evolution of computing in general, and IT services delivery in particular. The key idea of this paradigm is to provide a utility service, similar to a power grid, into which a user may plug-in regardless of location to access centrally located services. The metaphor is of an “electronic cloud” that follows the users as they change their physical locale. The term “Cloud Computing” refers to an on-line delivery and consumption model for business and customer services, ranging from IT services like Software-as-a-Service (SaaS), Storage or Server capacity as a service, and many “non-IT” services which are not “computing” tasks per se. All these services can be commonly referred as “X as a service (XaaS)” and share the fundamental approach of running server components elsewhere, over the Internet.

In engineering terms, this paradigm refers to providing services on virtual machines allocated on top of a large machine pool in large data centers, where as in business terms, the term means a method to address scalability and availability concerns for large scale applications. It builds on decades of research in virtualization, parallel and distributed systems, utility computing, and more recent advances in networking, web and software services. We recommend [18],[20-24] for the interested reader. Figure 2 shows a generic high level architecture of cloud computing and various equivalent services (adapted from [21] ) possible with it.

B. Data Center Energy Consumption, Costs and Efficiency

Cloud Services are made possible with massively built modern data centers, which consume and waste enormous amounts of electricity for four different reasons:

1. Energy consumption from physical servers, storage and network equipment has increased many a fold over the last few years.
2. Existing IT system implementations are demanding increasing amounts of energy to power large and larger solutions involving complex processing elements and dependencies.
3. New cloud-based services and applications place unpredictable demands for scale, elasticity and availability.
4. The high cost of cooling systems and low efficiencies of large power distribution units etc.

All these reasons and more translate to growing energy costs [4] [5][14][15]. Gartner Group estimates that energy costs may increase from 10% of the IT budget today to over 50% in the next few years [10]. Figure 1 compares the purchasing dollars spent on new servers with the power and cooling cost since 1996 and projects those numbers until 2010. The total power and cooling bill for US based servers stands at a whopping $14billion a year, and if current trends persist, this bill is going to rise to $50 billion by the end of the decade [12]. IDC projects the cost to power servers will exceed the cost of the servers by 2010 [11]. Forrester says servers would use about 30% of their peak electricity consumption while sitting idle 70% of the time.

Moreover, data centers are always built to suit peak load. There is always a huge gap in data center peak-to-average processing/energy consumption. This gap calls in for better fine-grained idling, faster power shut down/restoration and pervasive support in the computational ecosystem architecture to improve the efficiencies and remove diseconomies of power [37]. For this, we’ll list here some of the useful metrics used in measuring the data center efficiency. The Green Grid [3] defines two useful terms...
when looking at data center efficiency: Power Usage Effectiveness (PUE) and Data Center Infrastructure Effectiveness (DCiE).

- \[ \text{PUE} = \frac{\text{Total Power}}{\text{IT Equipment Power}} \]
- \[ \text{DCiE} = \frac{\text{IT Equipment Power}}{\text{Total Power}} \times 100 \]

PUE measures how much power goes directly to computing compared to ancillary services such as lighting and cooling. DCiE is the reciprocal of PUE and tells us what percentage of the power delivered to the facility actually gets delivered to the servers. A perfect score of 1 means no power goes to the extra costs; 1.5 means that half the power goes to ancillary services. Also, PUEs vary greatly. Very inefficient enterprise data centers often have a score of 2.0 and 3.0 and with excellent data centers report around 1.2 [5][7][37].

C. Environmental Implications

Apart from these technological implications, the growing environmental awareness in government and corporate circles on the impact of data center energy consumption is demanding for more proactive steps in this direction. Now, corporate executives are allocating time, energy and money to invest in environmental initiatives. Governments are allocating research and regulations, and drafting laws to address the efficiency of data centers and other critical components of IT infrastructure. Consumer advocates, policy makers and influential leaders are prompting IT organizations to significantly reduce the impact that computing and electronics make on the environment. This has led in creating environmental impact measurements such as carbon footprint, which is usually measured in tCO2eq (metrics tons of CO2 equivalent) based on the source of energy and amount consumed, manufacturing and logistics impact, as well as end-of-life impact (e-waste, environmental externalities etc) [6]. Most of the IT organizations are also measuring the ongoing energy costs and environmental impact of specific applications within the enterprise through the use of a set of Energy Usage Profiles (EUPs) [13][17][19].

D. New Architectural Decision Points

The need to reduce power consumption is thus obvious. As we have seen, energy consumption control is a much more complex architecture problem and thus requires a coordinated array of tactics, from capacity planning techniques to optimizing operational processes and facilities design [16]. This encompasses an ongoing and holistic plan of action for developing an environmentally sustainable solution that leverages - technological advances, process frameworks and strategic initiatives across the computational ecosystem, ranging from energy suppliers, data centers, and governance bodies to end-customers. Some of the attributes of such plan can include: encouraging IT reuse, reducing IT complexity, aligning all the stakeholders, optimize functional and non-functional quality goals, usage of market based metering schemes, provisioning power APIs from energy suppliers to name a few.

In this direction, we propose considering “Smart Grid” and “Smart Metering” as two useful constructs in driving down the associated operating costs for data centers and businesses [25-30]. Based on the implementation successes seen so far and the realized technical/business benefits accrued, we see “smart metering” based approaches becoming pervasive and imminent in the coming days.

![Smart Grid Diagram](image)

Figure 2: Generic Cloud Architecture and various services [21] possible with it

1) Smart Grid

In its most basic form, a smart grid refers to an effort to prod users (consumers) of electricity to change their behavior around variable electric rates or to pay vastly increased rates for the privilege of reliable electrical service during peak conditions, and reduced rates during off-peak conditions. The grid is designed so as to reduce demand during peak usage periods by making use of the technology advances in metering and communication protocols. In this way, it relies on the free-market capitalism ideals to level the load curve assuming that the customer will always act in response to the market signals.

2) Smart Metering

In our definitional framework, “smart meters” represent the next generation utility measurement devices with real-time sensors, power outage notification, and power quality monitoring and come only as part of the smart grid environment. Smart meters provide a economical way of measuring “when” and “where” of energy consumption, allowing price setting agencies (energy suppliers) to introduce differential pricing based on the time of day and the season.

III. OUR MODEL: SMART METERING THE CLOUDS

In the previous sections, we have seen how rising fuel prices and capacity costs together with shrinking reserve margins and green house gas emissions are setting the policy maker’s agenda today. The governance and compliance regulations are placing a renewed emphasis on the demand side of the equation to deal with these issues. We believe that both the energy efficiency and demand response plays a
vital role in meeting future energy needs and “smart grid” and “smart metering” approaches open up new vistas in this inevitable future with their multi-faceted, global concentrated effort to reduce energy consumption and promote sustainability. This not only benefits the utility company, but also allows the partners to monitor their daily energy usage and to optimize their energy usage.

And also with smart-meters, comes the real time pricing (RTP). RTP will improve the efficiency of energy consumption and investment and would lessen the potential harm from market fluctuations. RTP helps in getting reduced emissions from reduced peak demand. It reduces variance of electricity load and estimates the short-run impacts of a reduction in variance on emissions of So2, NOx and CO2.

A. Our Vision: Application of Smart Metering to Clouds

Our vision\textsuperscript{1} looks at a multi-faceted approach for meeting the energy needs of clouds, which comprises of:

- Information about energy costs as they are incurred and ideas on how to manage those costs.
- Codes and standards for new applications, hardware appliances, and data center buildings and industrial processes.
- Enabling technologies that control costs in real-time conditions such as circuit meters, power strip meters, plug meters, base board controller energy consumption metering, external thermal sensor metering, and internal server thermal metering.
- Rebates and financing for accelerating the adoption of smart end-use applications/technologies
- Smart rate design such as inclining block rates and dynamic pricing rates

B. Extensible Architecture

We hope to see an extensible architecture for cloud computing environments that encompasses the laid premises and also all the stakeholders of the data center computational ecosystem as it is fast becoming an imperative business necessity. In that architecture of various vendors with different metering standards, interfaces and protocols, we see an inherent need for \textit{sustainable}, \textit{efficient} and \textit{open standard-based} metering provisioning with environmental impact metrics. We outline one such architecture, first of its kind, in figure 3.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{architecture.png}
\caption{Our Extensible Architecture}
\end{figure}

The key features of our architecture include:

- \textbf{Smart Metering Interfaces}: These interfaces does metering based on how much is consumed and at what time of day in order to prompt the customers to adjust their consumption habits and be more responsive to the market clues. It may also check other factors including day of week, seasonality, economic booms and busts. As such, this layer comprises of two components - Proprietary energy API services and a Consumption communication layer.

- \textbf{Energy API services}. Most vendors have their own proprietary API model to interface with the metering devices. Because energy metering architectures differ with many data centers, larger organizations may have more than one proprietary interface environment. It is important to set up a reliable and open design standard that reaches across data centers and technologies.

- \textbf{Consumption communication Layer}. Because of diverse proprietary environmental metering interface systems in the open market, there is a need for a common consumption layer to interface with different metering and reporting systems.

- \textbf{Management Interfaces}: These interfaces control the business management functions such as billing, access to customer information system, outage management system through the consumption data aggregation layer.

\textsuperscript{1} http://www.ideationcloud.com/2009/03/smart-metering-in-the-clouds-im-writing-a-paper-on-converged-metering-solutions/
• **Consumption data aggregation Layer.** This is a common data collection repository designed for frequent updates. This area is the collection point for environmental data across data centers in compliance with government and other environment agencies norms and regulations.

![Figure 4: Smart-metering the clouds](image)

**C. Implementation Idea**

With this extensible architecture, one possible way of implementing this could be via Automatic Meter Reading (AMR) systems using the IEEE 802.15.4-compliant wireless networks or any other reliable communication mechanism [34] [35]. For example, an IEEE 802.15.4 network can operate in either the star topology or the peer-to-peer topology. If it is a peer-to-peer implementation, each node has the ability to do two-way communications and may relay or forward the data for the neighboring nodes within the transmit range, until the local grid.

**D. Pricing Model**

Pricing is important in our extensible architecture, as one of the goals of “smart-metering the clouds” is to take advantage of the differentiated pricing structures available from the energy suppliers side to improve the scalability requirements and to deal with the bursts in resource demands on data centers side. Charging variable prices can be useful for resource management as it can result in the diversion of demand from high-demand time periods to low-demand time periods[32] [36].

This kind of pricing mechanism becomes even more apparent if we are to do “smart-metering” to dynamically balance the data centers to conserve power through “instant provisioning” of infrastructure on-demand or “live migrate” entire VMs (OS + Applications) in a transparent and lower headed manner. We believe that this enables a new category of systems that can react better to changes in workloads at the aggregation level itself. Figure 4 shows one such possible scenario, which we hope will become a common use case in near future.

However, this poses several challenges. At present, service providers have inflexible pricing, generally limited to flat rates or tariffs based on usage thresholds, and customers are restricted to offerings from a single provider at a time. Charging fixed prices is also not fair to both the parties since different consumers have distinctive usage scenarios and demand specific QoS for various resource requests that can change any time [33]. In addition, the scenario becomes even more delicate with many providers having proprietary interfaces to their services, restricting the ability of customers to swap one provider for another.

Optionally, we can however look at multi-part tariffs, typically used in “Grid Economics”. These include pricing models such as two-part tariffs (fixed-membership fee, plus per-use fee), three-part tariffs (fixed fee, allowance level, per-use fee), flat-rate tariffs (fixed fee, unlimited use), linear tariffs, progressive tariffs (initial rate up to some level, then higher rate), etc.

One such approach that can be readily used is the Real Time Pricing (RTP). Some of its examples include spot price contract, time-of-use tariff (TOU) which has prices that vary by blocks of time within the day, and critical-peak pricing (CPP) [31][36]. The TOU and the CPP tariffs are more predictable for the consumers than the hourly spot price as they provide incentives for consumption adjustments. Other variants include a two-part real-time pricing (RTP), which offers a fixed price for an agreed volume and the spot-price for deviations from this volume.

**E. Key Advantages**

Smart metering allows “smart economic benefits” of cloud Computing such as elasticity and transference of risk, especially the risks of over provisioning (under utilization) and under provisioning (saturation). Apart from these simple diurnal patterns, this approach also helps in common cloud usage cases where certain services experience seasonal or other periodic demand variation (e-commerce peaks during Christmas Eve, and photo-sharing sites peak after holidays, weekdays etc) as well as some unexpected demand surges or bursts due to external events (news).

We list some of the key benefits here, which includes - Providing real-time feedback on energy consumption should help better management of energy behavior; ability to store consumption/billing details so as to effectively account for the “measure-as-you-go” model; achieve lower energy costs with existing rate designs and even greater impacts with dynamic pricing mechanisms based on seasonality, day and time among many others [30] [36].

**IV. SUMMARY AND FUTURE WORK**

Given the scale, cost and importance of emerging cloud service data centers, it is incumbent on us to rethink the way metering is done from a multi-faceted view point. We’ve shown how smart-metering can be part of a holistic & integrated solution encompassing the whole spectrum of players including energy suppliers, data centers, cloud service providers, third-party software vendors, and even
end-customers. In this way, smart-metering can be seen as the main information gathering device for the operation of a cloud utility services model and a missing link in the understanding of how energy is consumed. The other advantages include self-healing using real-time information and automatic controls to anticipate, detect and respond to system problems.

We propose that this shift represents a unique opportunity and the ecosystem partners should take advantage of this, and build applications that leverage the technological capabilities of the new infrastructure (platform) to create new solutions that require next-generation energy-efficiency and demand response. This will require open standards and protocols as proprietary communication protocols and closed systems will make such solutions difficult, if not impossible. The interfaces can be exposed as consumable APIs for remaining partners, thereby closely following the "democratization" principles. Consumers and businesses could choose those that best meet their needs and the competitive market place would lead to the emergence of better solutions. However, we see that some aspects of smart-grid and smart-metering are in nascent stages and can certainly benefit from the enhanced models/approaches in the fields of performance modeling, optimization and control theory.

This work is still evolving and we hope to present effective smart-metering pricing mechanisms to deal with cloud environments in our future work.

REFERENCES


