Sustainable Ecosystems: Enabled by Supply and Demand Management

Chandrakant D. Patel

Hewlett Packard Laboratories, Palo Alto, CA 94304, USA
chandrakant.patel@hp.com

Abstract. Continued population growth, coupled with increased per capita consumption of resources, poses a challenge to the quality of life of current and future generations. We cannot expect to meet the future needs of society simply by extending existing infrastructures. The necessary transformation can be enabled by a sustainable IT ecosystem made up of billions of service-oriented client devices and thousands of data centers. The IT ecosystem, with data centers at its core and pervasive measurement at the edges, will need to be seamlessly integrated into future communities to enable need-based provisioning of critical resources. Such a transformation requires a systemic approach based on supply and demand of resources. A supply side perspective necessitates using local resources of available energy, alongside design and management that minimizes the energy required to extract, manufacture, mitigate waste, transport, operate and reclaim components. The demand side perspective requires provisioning resources based on the needs of the user by using flexible building blocks, pervasive sensing, communications, knowledge discovery and policy-based control. This paper presents a systemic framework for supply-demand management in IT—in particular, on building sustainable data centers—and suggests how the approach can be extended to manage resources at the scale of urban infrastructures.

Keywords: available energy, exergy, energy, data center, IT, sustainable, ecosystems, sustainability.

1 Introduction

1.1 Motivation

Environmental sustainability has gained great mindshare. Actions and behaviors are often classified as either "green" or "not green" using a variety of metrics. Many of today's "green" actions are based on products that are already built but classified as "environmentally friendly" based on greenhouse gas emission and energy consumption in use phase. Such compliance-time thinking lacks a sustainability framework that could holistically address global challenges associated with resource consumption.

These resource consumption challenges will stem from various drivers. The world population is expected to reach 9 billion by 2050 [1]. How do we deal with the increasing strain that the economic growth is placing on our dwindling natural

M.K. Aguilera et al. (Eds.): ICDCN 2011, LNCS 6522, pp. 12–28, 2011.

resources? Can we expect to meet the needs of society by solely relying on replicating and extending the existing physical infrastructure to cope with economic and population growth? Indeed, anecdotal evidence of the strain that society is placing on the supply side—the resources used for goods and services—is apparent: rising prices for critical materials, such as copper and steel; the dramatic reduction in output of the Pemex Canatrell oil field in Mexico, one of the largest in the world; and limitations in city scale waste disposal. Furthermore, a rise in the price of fuel has led to inflationary impact that could threaten the quality of life of billions. Thus, depletion of limited natural resources and increases in the cost of basic goods necessitates new business models and infrastructures that are designed, built and operated using the least possible amount of appropriate materials and energy. The supply side must be considered together with the societal demand for resources.

This paper presents a holistic framework for sustainable design and management. Unlike past work that has mostly focused on operational energy considerations of devices, this contribution weaves lifecycle implications into a broader supply-demand framework. The following are the salient contributions:

- Use of available energy (also called exergy) from 2nd law of thermodynamics as a metric for quantifying sustainability.
- Formulation of a supply-demand framework based on available energy.
- Application of this framework to IT, in particular, to data centers.
- Extension of the framework to other ecosystems such as cities.

1.2 Role of the IT Ecosystem

Consider the information technology (IT) ecosystem made up of billions of service oriented client devices, thousands of data centers and digital print factories. As shown in Figure 1, the IT ecosystem and other human managed ecosystems such as transportation, waste management, power delivery, industrial systems, etc. draw from a pool of available energy. In this context, IT has the opportunity to change existing business models and deliver a net positive impact with respect to consumption of available energy. To do so, sustainability of the IT ecosystem itself must be addressed holistically. Given a sustainable IT ecosystem, imagine the scale of impact when billons in growth economies like India utilize IT services to conduct transactions such as purchasing railway tickets, banking, availing healthcare, government services, etc. As the billions board the IT bus, and shun other business models, such as ones that require the use of physical transportation means like an auto-rickshaw to go to the train station to buy tickets, the net reduction in the consumption of available energy can be significant. Indeed, when one overlays a scenario where everything will be delivered as a service, a picture emerges of billions of end users utilizing trillions of applications through a cloud of networked data centers. However, to reach the desired price point where such services will be feasible—especially in emerging economies, where Internet access is desired at approximately US $1 per month—the total cost-of-ownership (TCO) of the physical infrastructure that supports the cloud will need to be revisited. There are about 81 million Internet connections in India [2]. There has been progress in reducing the cost of access devices [3], but the cost to avail services still needs to be addressed. In this regard, without addressing the cost of data centers—the foundation for services to the masses—scaling to billions of users is not possible.

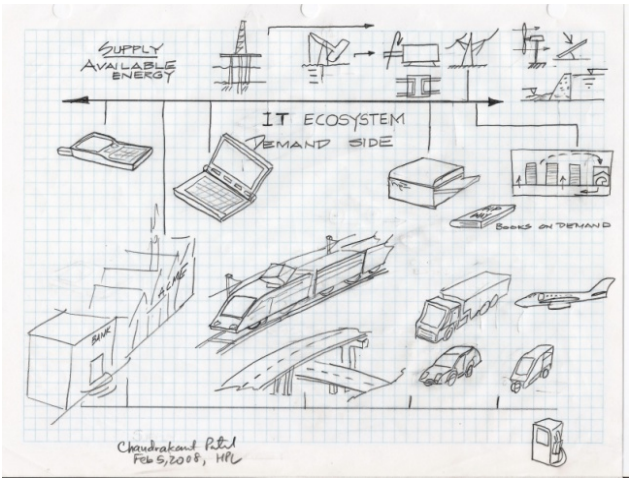

Fig. 1. Consumption of Available Energy

With respect to data centers, prior work has shown that a significant fraction of the TCO comes from the recurring energy consumed in the operation of the data center, and from the burdened capital expenditures associated with the supporting physical infrastructure [4]. The burdened cost of power and cooling, inclusive of redundancy, is estimated to be 25% to 30% of the total cost of ownership in typical enterprise data centers [4]. These power and cooling infrastructure costs may match, or even exceed, the cost of the IT hardware within the data center. Thus, including the cost of IT hardware, over half of the TCO in a typical data center is associated with design and management of the physical infrastructure. For Internet services providers, with thinner layers of software and licensing costs, the physical infrastructure could be responsible for as much as 75% of the TCO. Conventional approaches in building data centers with multiple levels of redundancies and excessive material—an "always-on" mantra with no regard to service level agreement, and lack of dynamic provisioning of resources—leads to excessive over provisioning and cost. Therefore, cost reduction requires an end to end approach that delivers least materials, least energy data centers. Indeed, contrary to the oft held view of sustainability as "paying more to be green", minimizing the overall lifecycle available energy consumption and thereby building sustainable data centers leads to lowest cost data centers.

2 Available Energy or Exergy as a Metric

2.1 Exergy

IT and other ecosystems draw from a pool of available energy as shown in Figure 1. Available energy, also called exergy, refers to energy that is available for performing work [5]. While energy refers to the quantity of energy, exergy quantifies the useful portion (or "quality") of energy. As an example, in a vehicle, the combustion of a

given mass of fuel such as diesel results in propulsion of vehicle (useful work done), dissipation of heat energy and a waste stream of exhaust gases at a given temperature. From the first law of thermodynamics, the quantity of energy was conserved in the combustion process as the sum of the energy in the products equals that in the fuel. However, from the 2nd law of thermodynamics, the usefulness of energy was destroyed since there is not much useful work that can be harnessed from the waste streams e.g. exhaust gases. One can also state that the combustion of fuel resulted in increase of entropy or disorder in the universe – going from a more ordered state in fuel to less ordered state in waste streams. As all processes result in increase in entropy, and consequent destruction of exergy due to entropy generation, minimizing the destruction of exergy is an important sustainability consideration. From a holistic supply-demand point of view, one can say that we are drawing from a finite pool of available energy, and minimizing destruction of available energy is key for future generations to enjoy the same quality of life as the current generation. With respect to making the most of available energy, it is also important to understand and avail opportunities in extracting available energy from waste streams.

Indeed, it is instructive to examine the combustion example further to understand the exergy content of waste streams. Classical thermodynamics dictates the upper limit of the work, *A*, that could be recovered from a heat source, Q (in Joules), at temperature T_j (in Kelvins) emitting to a reservoir at ground state temperature T_a as:

$$A = \left(1 - \frac{T_a}{T_j}\right) Q \tag{1}$$

For example, with reference to equation 1, relative to a temperature of 298 K (25 ^{o}C), 1 joule of heat energy at 773 K (500 ^{o}C)—such as exhaust gases from a gas turbine—can give 0.614 joules of available energy. Therefore, a waste stream at this high temperature has good availability (61%) that can be harvested. By the same token, the same joule at 323 K (50 ^{o}C)—such as exhaust air from a high power server—can only give 0.077 joules of work. While this determines the theoretical maximum Carnot work that can be availed with a perfect reversible engine, the actual work is much less due to irreversible losses such as friction. Stated simply, the 2nd law of thermodynamics places a limit on the amount of energy that can be converted from one form to another. Similarly, laws of thermodynamics can be applied to other conversion means e.g. electrochemical reactions in fuel cells to estimate the portion of reaction enthalpy that can be converted to electricity [6].

Traditional methods of design involve the completion of an energy balance based on the conservation theory of the first law of thermodynamics. Such a balance can provide the information necessary to reduce thermal losses or enhance heat recovery, but an energy analysis fails to account for degradation in the quality of energy due to irreversibilities predicted by the second law of thermodynamics. Thus, an approach based on the second law of thermodynamics is necessary for analyzing available energy or exergy consumption across the lifecycle of a product—from "cradle to cradle". Furthermore, it can also be used to create the analytics necessary to run operations that minimize destruction of exergy and create inference analytics that can enable need-based provisioning of resources. Lastly, exergy analysis is important to

determine the value of the waste stream and tie it to an appropriate process that can make the most of it. For example, the value of converting exhaust heat energy to electrical energy using a thermo-electric conversion process may apply in some cases, but not in others when one takes into account the exergy requirement to build and operate the thermo-electric conversion means.

2.2 Exergy Chain in IT

Electrical energy is produced from conversion of energy from one form to another — a common chain starts with converting the chemical energy in the fuel to thermal energy from the combustion of fuel to mechanical energy in a rotating physical device to electrical energy from a magnetically based dynamo. Alternatively, available energy in water—such as potential energy at a given height in a dam—can be converted to mechanical energy and to electrical energy. The electrical energy is 100% available. However, as electrical energy is transmitted and distributed from the source to the point of use, losses along the way in transmission and distribution lead to destruction of availability. The source of power for most data centers (i.e., thermal power station) operates at an efficiency in the neighborhood of 35% to 60% [7]. Transmission and distribution losses can range from 5% to 12%. System level efficiency in the data center power delivery infrastructures (i.e., from building to chip) can range from 60% to 85% depending on the component efficiency and load. Around 80% is typical for a fully loaded state-of-the-art data center. Overall, out of every watt generated at the source, only about 0.3 W to 0.4 W is used for computation. If the generation cycle itself as well as overhead of the data center infrastructure (i.e., cooling) is taken into account, the coal-to-chip power delivery efficiency will be around 5% to 12%.

In addition to consumption of exergy in operation, the material within the data center has exergy embedded in it. The embedded exergy stems from exergy required to extract, manufacture, mitigate waste, and reclaim the material. Exergy is also embedded in IT as result of direct use of water (for cooling) and indirect use of water (for production of parts, electricity, etc). Water too can be represented using exergy. As an example, assuming nature desalinates water and there is sufficient fresh water available from natural cycle, one can represent exergy embedded in water as a result of distribution (exergy required to pump) and treatment (exergy required to treat waste water). On average, treatment and distribution of a million gallons of surface water requires 1.5 MWh of electrical energy. Similarly, treatment of a million gallons of waste water consumes 2.5 MWh of electrical energy [8].

3 Supply Side and Demand Side Management

3.1 Architectural Framework

In order to build sustainable ecosystems, the following systemic framework articulates the management of supply and demand side of available energy based on the needs of the users.

- On the *supply side:*
  - minimizing the exergy required to extract, manufacture, mitigate-waste, transport, operate and reclaim components;
  - design and management using local sources of available energy
    - to minimize the destruction of exergy in transmission and distribution, e.g., dissipation in transmission; and,
    - take advantage of exergy in the waste streams, e.g., exhaust heat from turbine.
- On the *demand side:*
  - minimizing the consumption of exergy by provisioning resources based on the needs of the user by using flexible building blocks, pervasive sensing, communications, knowledge discovery and policy based control.

Sustainable ecosystems, given the supply and demand side definition above, are then built on delivering to the needs of the user. The needs of the user are derived from the service level agreement (SLA), decomposed into lower level metrics that can be applied in the field to enable integrated management of supply and demand.

The balance of the paper steps through the framework by examining lifetime exergy consumption in IT, evolving a supply-demand framework for data centers and closing by extending the framework to other ecosystems.

3.2 Quantifying Lifetime Exergy Consumption

As noted earlier, exergy or available energy stemming from the second law of thermodynamics fuses information about materials and energy use into a single meaningful measure. It estimates the maximum work in Joules that could theoretically have been extracted from a given amount of material or energy. By equating a given system in terms of its lifetime exergy consumption, it becomes possible to remove dependencies on the type of material or the type of energy (heat, electricity, etc) consumed. Therefore, given a lifecycle of a product, as shown in Figure 2, one can now create an abstract information plane that can be commonly applied across any arbitrary infrastructure. Lifecycle design then implies inputting the entire supply chain from "cradle to cradle" to account for exergy consumed in extraction, manufacturing, waste mitigation, transportation, operation and reclamation. From a supply side perspective, designers can focus on minimizing lifetime energy consumption through de-materialization, material choices, transportation, process choices, etc. across the lifecycle. The design toolkit requires a repository of available energy consumption data for various materials and processes.

With respect to IT, the following provides an overview of the salient "hotspots" discerned using an exergy based lifetime analysis [9]:

- For service oriented access devices such as laptops, given typical residential usage pattern, the lifetime operational exergy consumption is 20-30% of the total exergy consumed while the rest is embedded (exergy consumed in extraction, manufacturing, transportation, reclamation).
  - Of the 70-80% of the embedded lifetime exergy consumption, display is a big component.

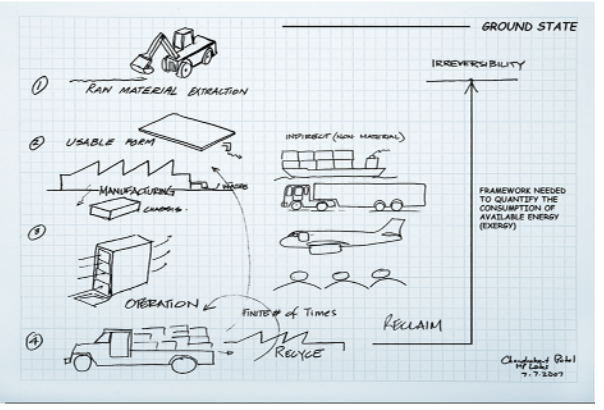

Fig. 2. Lifecycle of a product

- For data centers, for a given server, the lifetime operational exergy consumption is about 60% to 80% of the total lifetime exergy consumption [10].
    - The large operational component stems from high electricity consumption in the IT equipment and the data center level cooling infrastructure [11][12].

From a strategic perspective, for handhelds, laptops and other forms of access devices, reducing the embedded exergy is critical. And, in order to minimize embedded exergy, least exergy process and material innovations are important. As an example, innovations in display technologies can reduce the embedded footprint of laptops and handhelds. On the other hand, for data centers, it is important to devise an architecture that focuses on minimizing electricity (100% available energy) consumed in operation.

Next section presents a supply-demand based architecture for peak exergetic efficiency associated with synthesis and operation of a data center. A sustainable data center—built on lifetime exergy considerations, flexible and configurable resource micro-grids, pervasive sensing, communications and aggregation of sensed data, knowledge discovery and policy based autonomous control—is proposed.

3.3 Synthesis of a Sustainable Data Center Using Lifetime Exergy Analysis

Business goals drive the synthesis of a data center. For example, assuming a million users subscribing to a service at US $1/month, the expected revenue would be US $12 million per year. Correspondingly, it may be desirable to limit the infrastructure (excluding software licenses, personnel) TCO to be $1/5^{th}$ of that amount, or roughly US $2.4 million per year. A simple understanding of impact of power can be had by estimating the cost implications in areas where low cost utility power is not available, and diesel generators are used as a primary source. For example, if the data center supporting the million users consumes 1 MW of power at 1 W per user or 8.76 million KWh per year, the cost of just powering with diesel at approximately $0.25 per KWh will be about $2.2 million per year. Thus, growth economies strained on the

resource side and reliant on local power generation with diesel will be at a great disadvantage. Having understood such constraints, the data center must meet the target total cost of ownership (TCO) and uptime based on the service level agreements for a variety of workloads.

Data center synthesis can be enabled by using a library of IT and facility templates to create a variety of design options. Given the variety of supply-demand design options, the key areas of analysis for sustainability and cost become:

1. Lifecycle analysis to evaluate each IT-Facility template and systematically dematerialize to drive towards least lifetime embedded exergy design and lowest capital outlay, e.g., systematically reduce the number of IT, power and cooling units, remove excessive material in the physical design, etc.
   - Overall exergy analysis should also consider exergy in waste streams, and locality of the data center to avail supply side resources (power and cooling).
2. Performance modeling toolkit to determine the performance of the ensemble and estimate consumption of exergy during operation.
3. Reliability modeling toolkit to discover and design for various levels of uptime within the data center.
4. Performance modeling toolkit to determine the ability to meet the SLAs for a given IT-Facility template.
5. TCO modeling to estimate the deviation from the target TCO.

Combining all the key elements noted above enables a structure for analysis for a set of applicable data center design templates for a given business need. The data center can be benchmarked in terms of performance per Joules of lifetime available energy or exergy destroyed. The lifetime exergy can be incorporated in a total cost of ownership model that includes software, personnel and license to determine the total cost of ownership of a rack [4] and used to price a cloud business model such as "infrastructure as a service".

3.4 Demand Side Management of Sustainable Data Centers

On the demand side, it is instructive to trace the energy flow in a data center. Electrical energy, all of which is available to do useful work, is transferred to IT equipment. Most of the available electrical energy drawn by the IT hardware is dissipated as heat energy, while useful work is availed through information processing. The amount of useful work is not proportional to the power consumed by the IT hardware. Even in idle mode IT hardware typically consumes more than 50% of its maximum power consumption [32]. As noted in [32], it is important to devise energy-proportional machines. However, it is also important to increase the utilization of IT hardware and reduce the total amount of required hardware. [31] presents such a resource management architecture. A common approach to increasing utilization is executing applications in virtual machines and consolidating the virtual machines onto fewer larger servers [33]. As shown in [15] workload consolidation has the potential to reduce the IT power demand significantly.

Next, additional exergy is used to actively transfer the heat energy from chip to the external ambient. All of the exergy delivered to the cooling equipment to remove heat is not used to affect the heat transfer. While a fraction of the electrical energy provided to a blower or pump is converted to flow work (product of pressure in N/m^2 and volume flow in m^3/s), and likewise a portion of the electrical energy applied to a compressor is converted to thermodynamic work (to reduce temperature of the data center), the bulk of the exergy provided is destroyed due to irreversibility. Therefore, in order to build an exergy efficient system, the mantra for demand side management in data center becomes one of allocation of IT (compute, networking and storage), power and cooling resources based on the need with the following salient considerations:

- Decompose SLAs to Service Level Objectives
    - based on the SLOs, allocate appropriate IT resources while meeting the performance and uptime requirements [28][29][30];
    - account for spatial and temporal efficiencies and redundancies associated with thermo-fluids behavior in a given data center based on heat loads and cooling boundary conditions[14].
- Consolidate workloads while taking into account the spatial and temporal efficiencies noted above, e.g., place critical workloads in "gold zones" of data centers which have inherent redundancies due to intersection of fluid flows from multiple air conditioning units and turn off or scale back power to IT equipment not in use [13][14][31].
- Enable the data center cooling equipment to scale based on the heat load distribution in the data center [11].

Dynamic implementation of the key points described above can result in better utilization of resources, reduction of active redundant components and reduction in electrical power consumption by half [13][15]. As an example, a data center designed for 1 MW of power at maximum IT load can run up to 80% capacity with workload consolidation and dynamic control. The balance of 200 KW can be used for cooling and other support equipment. Indeed, besides availing failover margin by operating at 80%, the coefficient of performance of the power and cooling ensemble is often optimal at about 80% loading given the efficiency curves of UPSs, and mechanical equipment such as blowers, compressors, etc.

3.5 Coefficient of Performance of the Ensemble

Figure 3 shows a schematic of energy transfer in a typical air-cooled data center through flow and thermodynamic processes. Heat is transferred from the heat sinks on a variety of chips—microprocessors, memory, etc—to the cooling fluid in the system e.g. driven by fans, air as a coolant enters the system and undergoes a temperature rise based on the mass flow and is exhausted out into the room. Fluid streams from different servers undergo mixing and other thermodynamic and flow processes in the exhaust area of the racks. As an example, for air cooled servers and racks, the dominant irreversibilities that lead to destruction of exergy arise from cold and hot air streams mixing and mechanical inefficiency of air moving devices. These streams (or some fraction of them) flow back to the modular computer room air conditioning units

(CRACs) and transfer heat to the chilled water (or refrigerant) in the cooling coils. Heat transferred to the chilled water at the cooling coils is transported to the chillers through a hydronics network. The coolant in the hydronics network, water in this case, undergoes a pressure drop and heat transfer until it loses heat to the expanding refrigerant in the evaporator coils of the chiller. The heat extracted by the chiller is dissipated through the cooling tower. Work is added at each stage to change the flow and thermodynamic state of the fluid. While this example shows a chilled water infrastructure, the problem definition and analysis can be extended to other forms of cooling infrastructure.

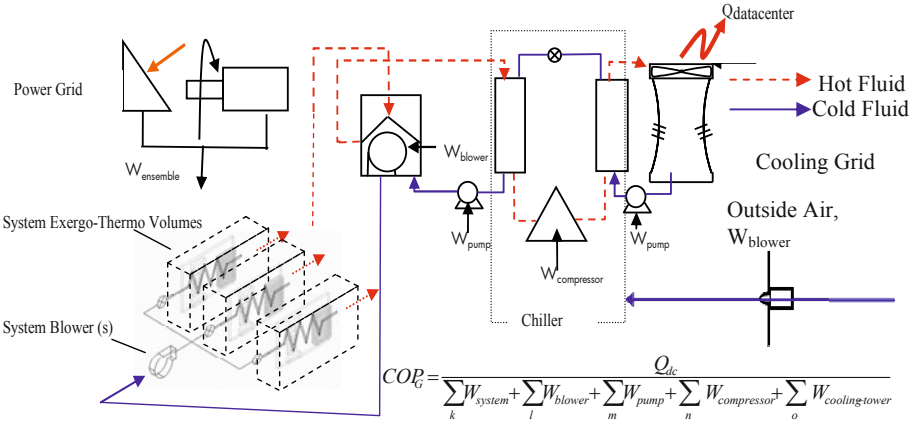

Fig. 3. Energy Flow in the IT stack – Supply and Demand Side

Development of a performance model at each stage in the heat flow path can enable efficient cooling equipment design, and provide a holistic view of operational exergy consumption from chips to the cooling tower. The performance model should be agnostic and be applicable to an ensemble of components for any environmental control infrastructure. [16] proposes such an operational metric to quantify the performance of the ensemble from chips to cooling tower. The metric, called coefficient of performance of the ensemble, COP_G, builds on the thermodynamic metric called coefficient of performance [16]. Maximizing the coefficient of performance of the ensemble leads to minimization of exergy required to operate the cooling equipment.

In Figure 3, the systems—such as processor, networking and storage blades—are modeled as "exergo-thermo-volumes" (ETV), an abstraction to represent lifetime exergy consumption of the IT building blocks and their cooling performance [17][18]. The thermo-volumes portion represents the coolant volume flow and resistance to the flow, characterized by volume flow ($\dot{V}$) and pressure drop (ΔP) respectively, to affect heat energy removal from the ETVs. The product of pressure drop (ΔP in N/m^2) and volume flow ($\dot{V}$ in m^3/s) determines the *flow work* required to move a given coolant (air here) through the given IT building block represented as an ETV. The minimum coolant volume flow ($\dot{V}$) for a given temperature rise required through the

ETV, shown by a dashed line in Fig 3, can be determined from the energy equation (Eq. 2).

$$\dot{Q} = \dot{m}\, C_p \left(T_{out} - T_{in}\right) \quad \text{where } \dot{m} = \rho \dot{V} \tag{2}$$

where $\dot{Q}$ is the heat dissipated in Watts, $\dot{m}$ is the mass flow in kg/s, ρ is density in kg/m^3 of the coolant (air in this example), C_p is specific heat capacity of air (J/kg-K) and T_{out} and T_{in} represent inlet and outlet temperatures of air.

As shown in Equation 3, the electrical power (100% available energy) required by the blower (W_b) is the ratio of the calculated *flow work* to the blower wire to air efficiency, ζ_b. The blower characteristic curves show the efficiency (ζ_b) and are important to understand the optimal capacity at the ensemble level.

$$W_b = \frac{(\Delta P_{etv} \times \dot{V}_{etv})}{\zeta_b} \tag{3}$$

Total heat load of the datacenter is assumed to be a direct summation of the power delivered to the computational equipment via UPS and PDUs. Extending the coefficient of performance (COP) to encompass power required by cooling resources in form of flow and thermodynamic work, the ratio of total heat load to the power consumed by the cooling infrastructure is defined as:

$$\begin{aligned} \text{COP} &= \frac{\text{Total Heat Dissipation}}{(\text{Flow Work} + \text{Thermodynamic Work}) \text{of Cooling system}} \\ &= \frac{\text{Heat Extracted by Air Conditioners}}{\text{Net Work Input}} \end{aligned} \tag{4}$$

The ensemble COP is then represented as shown below. The reader is referred to [16] for details.

$$COP_G = \frac{Q_{datacenter}}{\sum_k W_{system} + \sum_l W_{blower} + \sum_m W_{pump} + W_{compressor} + W_{coolingtower}} \tag{5}$$

3.6 Supply Side of a Sustainable Data Center Design

The supply side motivation to design and manage the data center using local sources of available energy is intended to minimize the destruction of exergy in distribution. Besides reducing exergy loss in distribution, a local micro-grid [19] can take advantage of resources that might otherwise remain unutilized and also present an opportunity to use the exergy in waste streams. Figure 4 shows a power grid with solar and methane based generator at a dairy farm. The methane is produced by anaerobic digestion of manure from dairy cows [20]. The use of biogas from manure is well known and has been used all over the world [7]. The advantage of co-location [20]

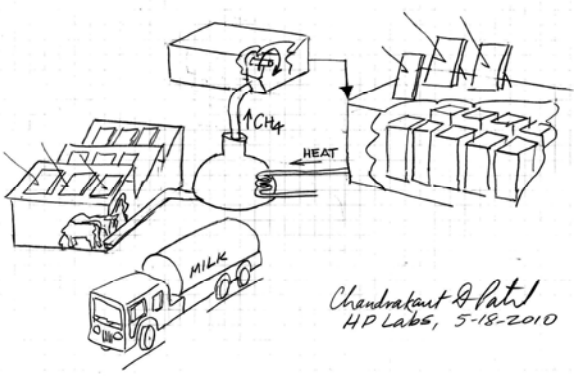

Fig. 4. Biogas and solar electric

stems from the use of heat energy exhausted by the server racks—one Joule of which has a maximum theoretical available energy of 0.077 W at 323 K (50 ^{o}C)—to enhance methane production as shown in Figure 4. The hot water from the data center is circulated through the "soup" in the digester. Furthermore, [20] also suggests the use of available energy in the exhaust stream of the electric generator to drive an adsorption refrigeration cycle to cool the data center.

Thus, multiple local options for power—wind, sun, biogas and natural gas—sourced locally can power a data center. And, high and low grade exergy in waste streams such as exhaust gases, can be utilized to drive other systems. Indeed, cooling for the data center ought to follow the same principles —a cooling grid made up of local sources. A cooling grid can be made up of ground coupled loops to dissipate heat into the ground, and use outside air (Figure 3) when at appropriate temperature and humidity to cool the data center.

3.7 Integrated Supply-Demand Management of Sustainable Data Centers

Figure 5 shows the architectural framework for integrated supply-demand management of a data center. The key components of the data center—IT (compute, networking and storage), power and cooling—have five key horizontal elements. The foundational design elements of the data center are lifecycle design using exergy as a measure and flexible micro-grids of power, cooling and IT building blocks. The micro-grids enable the integrated manager the ability to choose between multiple supply side sources of power, multiple supply side sources of cooling and multiple types of IT hardware and software. The flexibility in power and cooling provides the ability to set power levels of IT systems and the ability to vary cooling (speed of the blowers, etc). The design flexibility in IT comes from intelligent scheduling framework, multiple power states and virtualization [15]. On this design foundation, the management layers are sensing and aggregation, knowledge discovery and policy based control.

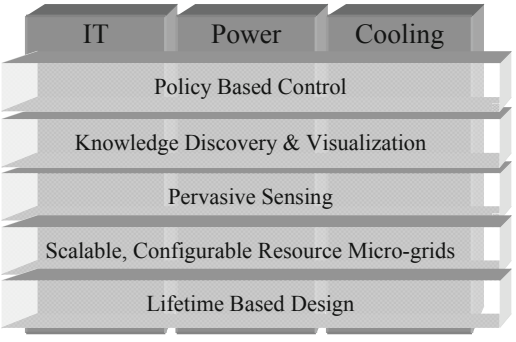

Fig. 5. Architectural framework for a Sustainable Data Center

At runtime, the integrated IT-Facility manager maintains the run-time status of the IT and facility elements of the datacenter. A lower level facility manager collects physical, environmental and process data from racks, room, chillers, power distribution components, etc. Based on the higher level requirements passed down by the integrated manager, the facility management system creates low-level SLAs for operation of power and cooling devices e.g. translating a high level energy efficiency goals to lower level temperature and utilization levels for facility elements to guarantee SLAs. The integrated manager has a knowledge discovery and visualization module that has data analytics for monitoring lifetime reliability, availability and downtimes for preventive maintenance. It has modules that provide insights that are otherwise not apparent at runtime e.g. temporal data mining of facility historical data.

As an example, data mining techniques have been explored for more efficient operation of an ensemble of chillers [23][34]. In [23], operational patterns (or motifs) are mined in historical data pertaining to an ensemble of water and air cooled chillers. These patterns are characterized in terms of their COP_G, thus allowing comparison in terms of their operational energy efficiency.

At the control level in Figure 5, a variety of approaches can be taken. As an example, in one approach, the cooling controller maintains dynamic control of the facility infrastructure (includes CRACs, UPS, Chillers, supply side power etc.) at levels determined by the facility manager to optimize the coefficient of performance of the COP_G, e.g., for providing requisite air flow to the racks to maintain the temperature at the inlet of the racks between 25 ^oC and 30 ^oC [11][12][14][21]. While exercising dynamic cooling control through the facility manager, the controller also provides information to the IT manager to consolidate workloads to optimize the performance of the data center, e.g., the racks are ranked based on thermo-fluids efficiency at a given time, and the ranking is used in workload placement [15]. The IT equipment not in use is scaled down by the integrated manager [15]. Furthermore, in order to reduce the redundancy in the data center, working in conjunction with IT and Facility managers, the integrated manager uses virtualization and power scaling as a flexibility to mitigate failures e.g. air conditioner failures [12][14][15].

Based on past work, the total energy consumed—in power and cooling—with these demand management techniques would be half of that of state of the art designs. Coupling the demand side management with the supply side options from the local

power grid, and the local cooling grid, opens up a completely new approach to integrated supply-demand side management [24] and lead to a "net zero" data center.

4 Applying Supply-Demand Framework to Other Ecosystems

In previous sections, the supply-demand framework was applied in devising least lifetime exergy data centers. Sustainable IT can now become IT for sustainability to enable need-based provisioning of resources—power, water, waste, transportation, etc—at scales of cities and can thus deliver a net positive impact by reducing the consumption and depletion of precious Joules of available energy.

Akin to the sustainable data center, the foundation of the "Sustainable City" or "City 2.0" is comprehensive lifecycle design [25][26]. Unlike previous generations, where cities were built predominantly focusing on cost and functionality desired by inhabitants, sustainable cities will require a comprehensive life-cycle view, where systems are designed not just for operation but for optimality across resource extraction, manufacturing and transport, operation, and end-of-life. The next distinction within sustainable cities will arise in the supply-side resource pool. The inhabitants of sustainable cities are expected to desire on-demand, just-in-time access to resources at affordable costs. Instead of following a centralized production model with large distribution and transmission networks, a more distributed model is proposed: augmentation of existing centralized infrastructure with local resource micro-grids. As shown for the sustainable data center, there is an opportunity to exploit the resources locally available to create local supply side grids made up of multiple local sources, e.g., power generation by photo-voltaic cells on roof tops, utilization of exergy available in waste stream of cities such as municipal waste and sewage together with natural gas fired turbines with full utilization of waste stream from the turbine, etc. Similarly, for other key verticals such as water, there is an opportunity to leverage the past experience to build water micro-grids using local sources, e.g., harvesting rain water to charge local man-made reservoirs and underground aquifers. Indeed, past examples such as the Amber Fort in Jaipur, Rajasthan, India show such considerations in arid regions [27].

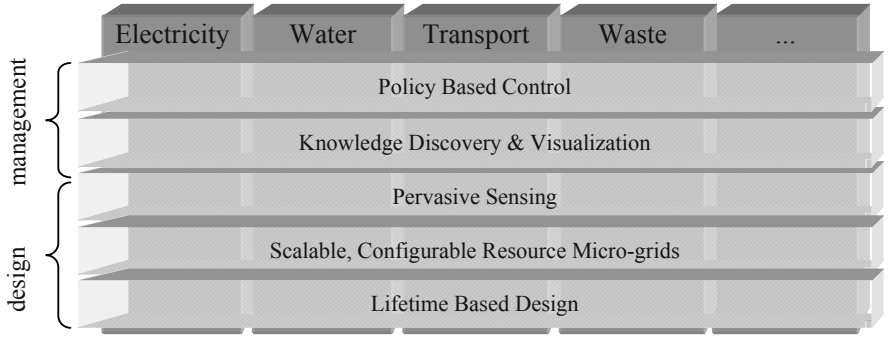

Fig. 6. Architectural Framework for a Sustainable City

As shown in Figure 6, having constructed lifecycle based physical infrastructures consisting of configurable resource micro-grids, the next key element is a pervasive sensing layer. Such a sensing infrastructure can generate data streams pertaining to the current supply and demand of resources emanating from disparate geographical regions, their operational characteristics, performance and sustainability metrics, and the availability of transmission paths between the different micro-grids. The great strides made in building high density, small lifecycle footprint IT storage can enable archival storage of the aggregated data about the state of each micro-grid. Sophisticated data analysis and knowledge discovery methods can be applied to both streaming and archival data to infer trends and patterns, with the goal of transforming the operational state of the systems towards least exergy operations. The data analysis can also enable construction of models using advanced statistical and machine learning techniques for optimization, control and fault detection. Intelligent visualization techniques can provide high-level indicators of the 'health' of each system being monitored. The analysis can enable end of life replacement decisions, e.g., when to replace pumps in water distribution system to make the most of the lifecycle. Lastly, while a challenging task—given the flexible and configurable resource pools, pervasive sensing, data aggregation and knowledge discovery mechanisms—an opportunity exists to devise policy based control system. As an example, given a sustainability policy, upstream and downstream pumps in a water micro-grid can operate to maintain a balance of demand and supply.

5 Summary and Conclusions

This paper presented a supply-demand framework to enable sustainable ecosystems and suggested that a sustainable IT ecosystem, built using such a framework, can enable IT to drive sustainability in other human managed ecosystems.

With respect to the IT ecosystem, an architecture for a sustainable data center composed of three key components—IT, power and cooling—and five key design and management elements stripped across the verticals was presented (Figure 5).The key elements—lifecycle design, scalable and configurable resource micro-grids, sensing, knowledge discovery and policy based control—enable the supply-demand management of the key components. Next, as shown by Figure 6, an architecture for "sustainable cities" built on the same principles by integrating IT elements across key city verticals such as power, water, waste, transport, etc. was presented. One hopes that cities are able to incorporate the key elements of the proposed architecture along one vertical, and when sufficient number of verticals have been addressed, a unified city scale architecture can be achieved.

The instantiation of the sustainability framework and the architecture for data centers and cities will require a multi-disciplinary workforce. Given the need to develop the human capital, the specific call to action at this venue is to:

- Leverage the past, and return to "old school" core engineering, to build the foundational elements of the supply-demand architecture using lifecycle design and supply side design principles.

- Create a multi-disciplinary curriculum composed of various fields of engineering, e.g., melding of computer science and mechanical engineering to scale the supply-demand side management of power, etc.
    - The curriculum also requires social and economic tracks as sustainability in its broadest definition is defined by economic, social and environmental spheres—the triple bottom line—and requires us to operate at the intersection of these spheres.

References

1. United Nations Population Division, http://www.un.org/esa/population
2. Aguiar, M., Boutenko, V., Michael, D., Rastogi, V., Subramanian, A., Zhou, Y.: The Internet's New Billion. Boston Consulting Group Report (2010)
3. Chopra, A.: $35 Computer Taps India's Huge Low-Income Market. Christian Science Monitor (2010)
4. Patel, C., Shah, A.: Cost Model for Planning, Development and Operation of a Data Center, HP Laboratories Technical Report, HPL-2005-107R1, Palo Alto, CA (2005)
5. Moran, M.J.: Availability Analysis: A Guide to Efficient Energy Use. Prentice-Hall, Englewood Cliffs (1982)
6. Barbir, F.: PEM Fuel Cells, pp. 18–25. Elsevier Academic Press (2005)
7. Rao, S., Parulekar, B.B.: Energy Technology. Khanna Publishers (2005)
8. Sharma, R., Shah, A., Bash, C.E., Patel, C.D., Christian, T.: Water Efficiency Management in Datacenters. In: International Conference on Water Scarcity, Global Changes and Groundwater Management Responses, Irvine, CA (2008)
9. Shah, A., Patel, C., Carey, V.: Exergy-Based Metrics for Environmentally Sustainable Design. In: 4th International Exergy, Energy and Environment Symposium, Sharjah (2009)
10. Hannemann, C., Carey, V., Shah, A., Patel, C.: Life-cycle Exergy Consumption of an Enterprise Server. International Journal of Exergy 7(4), 439–453 (2010)
11. Patel, C.D, Bash, C.E., Sharma, R., Friedrich, R.: Smart Cooling of Data Centers. In: ASME IPACK, Maui, Hawaii (2003)
12. Patel, C., Sharma, R., Bash, C., Beitelmal, A.: Thermal Considerations in Data Center Design, In: IEEE-Itherm, San Diego (2002)
13. Bash, C.E., Patel, C.D., Sharma, R.K.: Dynamic Thermal Management of Air Cooled Data Centers. In: IEEE-Itherm (2006)
14. Beitelmal, A., Patel, C.D.: Thermo-fluids Provisioning of a High Density Data Center. HP Labs External Technical Report, HPL-2004-146R1 (2004)
15. Chen, Y., Gmach, D., Hyser, C., Wang, Z., Bash, C., Hoover, C., Singhal, S.: Integrated Management of Application Performance, Power and Cooling in Data Centers. In: 12th IEEE/IFIP Network Operations and Management Symposium (NOMS), Osaka (2010)
16. Patel, C.D., Sharma, R.K., Bash, C.E., Beitelmal, M.: Energy Flow in the Information Technology Stack: Introducing the Coefficient of Performance of the Ensemble. In: ASME International Mechanical Engineering Congress & Exposition, Chicago, Illinois (2006)
17. Shah, A., Patel, C.D.: Exergo-Thermo-Volumes: An Approach for Environmentally Sustainable Thermal Management of Energy Conversion Devices. Journal of Energy Resource Technology, Special Issue on Energy Efficiency, Sources and Sustainability 2 (2010)
18. Shah, A., Patel, C.: Designing Environmentally Sustainable Systems using Exergo-Thermo-Volumes. International Journal of Energy Research 33 (2009)

19. Sharma, R., Bash, C.E., Marwah, M., Christian, T., Patel, C.D.: MICROGRIDS: A new approach to supply-side design of data centers. In: IMECE 2009: Lake Buena Vista, FL (2009)
20. Sharma, R., Christian, T., Arlitt, M., Bash, C., Patel, C.: Design of Farm Waste-driven Supply Side Infrastructure for Data Centers, In: ASME-Energy Sustainability (2010)
21. Sharma, R., Bash, C.E., Patel, C.D., Friedrich, R.S., Chase, J.: Balance of Power: Dynamic Thermal Management of Internet Data Centers. IEEE Computer (2003)
22. Marwah, M., Sharma, R.K., Patel, C.D., Shih, R., Bhatia, V., Rajkumar, V., Mekanapurath, M., Velayudhan, S.: Data Analysis, Visualization and Knowledge Discovery in Sustainable Data Centers. In: Computer 2009, Bangalore (2009)
23. Patnaik, D., Marwah, M., Sharma, R., Ramakrishna, N.: Sustainable Operation and Management of Data Center Chillers using Temporal Data Mining, In: ACM KDD (2009)
24. Gmach, D., Rolia, J., Bash, C., Chen, Y., Christian, T., Shah, A., Sharma, R., Wang, Z.: Capacity Planning and Power Management to Exploit Sustainable Energy. In: 6th International Conference on Network and Service Management, Niagara Falls, Canada (2010)
25. Bash, C., Christian, T., Marwah, M., Patel, C., Shah, A., Sharma, R.: City 2.0: Leveraging Information Technology to Build a New Generation of Cities. Silicon Valley Engineering Council (SVEC) Journal 1(1), 1–6 (2009)
26. Hoover, C.E., Sharma, R., Watson, B., Charles, S.K., Shah, A., Patel, C.D., Marwah, M., Christian, T., Bash, C.E.: Sustainable IT Ecosystems: Enabling Next-Generation Cities, HP Laboratories Technical Report, HPL-2010-73 (2010)
27. Water harvesting, http://megphed.gov.in/knowledge/RainwaterHarvest/Chap2.pdf
28. Cunha, I., Almeida, J., Almeida, V., Santos, M.: Self-adaptive capacity management for multi-tier virtualized environments. In: 10th IFIP/IEEE IM (2007)
29. Khana, G., Beaty, K., Kar, G., Kochut, A.: Application performance management in virtualized server environments. In: IEEE/IFIP NOMS (2006)
30. Gmach, D., Rolia, J., Cherkasova, L., Kemper, A.: Capacity management and demand prediction for next generation data centers. IEEE ICWS Salt Lake City (2007)
31. Kephart, J., Chan, H., Das, R., Levine, D., Tesauro, G., Rawson, F., Lefurgy, C.: Coordinating multiple autonomic managers to achieve specified power-performance tradeoffs. In: 4th IEEE Int. Conf. on Autonomic Computing, ICAC (2007)
32. Barroso, L.A., Hölzle, U.: The case for energy-proportional computing. IEEE Computer 40(12), 33–37 (2007)
33. Andrzejak, A., Arlitt, M., Rolia, J.: Bounding the Resource Savings of Utility Computing Models, HP Laboratories Technical Report HPL-2002-339 (2002)
34. Patnaik, D., Marwah, M., Sharma, R., Ramakrishnan, N.: Data Mining for Modeling Chiller Systems in Data Centers, IDA (2010)