Revolutionizing Data Center Energy Efficiency

July 2008

James M. Kaplan
William Forrest
Noah Kindler

McKinsey&Company

Overview

Global data center capacity is growing rapidly, consuming more financial resources and emitting more greenhouse gases.

Data centers are driving rapid expense growth for many technology-intensive businesses. Five years ago, a typical data center cost $150 million to build. Today, a similarly sized data center costs several times that amount.

Until now, the principal job of data center managers has been to maintain data center stability and increase availability in support of rising business demand. With that goal largely achieved, managers should now apply a renewed business focus to their operations, to curb inefficiencies and check rising costs.

Despite the huge amounts of capital tied up in data centers, significant inefficiencies exist. Server utilization rarely exceeds 6 percent and facility utilization can be as low as 50 percent. With average data center costs now threatening to crowd out other technology investments, the matter has become a board level concern.

While data center inefficiency has been a factor for some time, significant organizational limitations have made it difficult to address this issue. Accountability for data center costs is fragment – in most organizations financial responsibility for the data center facility can fall under one department and operational responsibility under another. Decision-making is siloed, with business, application, systems and facility decisions made without sufficient consideration of upstream and downstream impacts. Costs too are fragmented, with the data center's portion of facilities cost often excluded from total cost of ownership (TCO) calculations.

With their enormous appetite for energy, today's data centers emit as much carbon dioxide as all of Argentina. Left on their current path, data center output will quadruple by the year 2020.

This report, part of McKinsey's ongoing body of research on data center management, outlines ways for organizations and the broader community to improve data center energy efficiency and address the twin challenges of rising data center spend and Greenhouse Gas (GHG) emissions. It introduces CADE (Corporate Average Datacenter Efficiency), a new industry standard efficiency measure developed by McKinsey, in conjunction with the Uptime Institure.
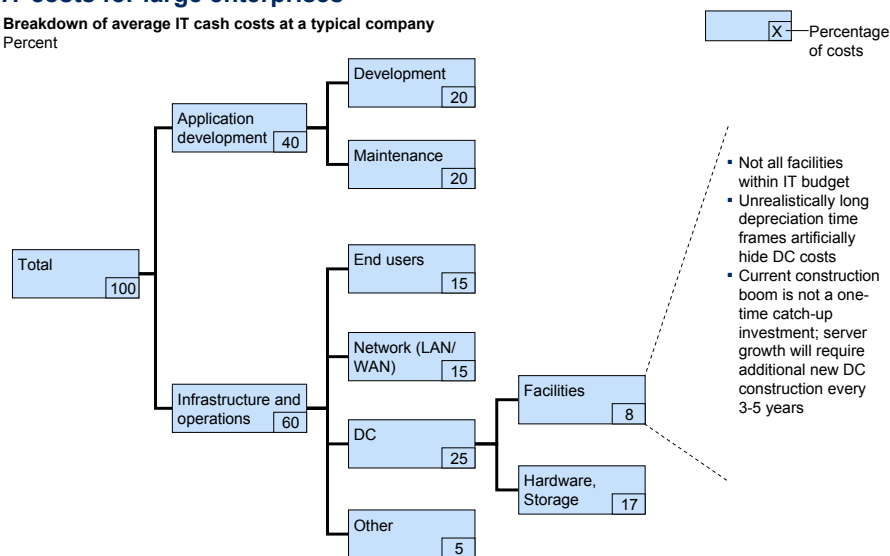
Data center inefficiency may threaten profitability

Demand for data center capacity is expected to grow at 10% CAGR over the next decade, as businesses automate more processes, perform complex analytics, store more granular customer data and employ rich media[1]. Capacity demand in emerging markets like China will rise especially quickly, as institutions there become larger and more sophisticated.

With this spike in capacity comes a corresponding escalation in data center IT costs. Today's data centers account for approximately 25% of the total corporate IT budget, when you take in account facilities, servers, storage and the labor to manage them.

Exhibit 1: Data center expenditures are typically a quarter or more of total IT costs for large enterprises

Breakdown of average IT cash costs at a typical company
Percent

- Not all facilities within IT budget
- Unrealistically long depreciation time frames artificially hide DC costs
- Current construction boom is not a one-time catch-up investment; server growth will require additional new DC construction every 3-5 years

SOURCE: McKinsey & Company

This share will grow as the number of servers, the amount of power consumed, and the unit cost of power all increase. In particular, facility costs are growing at 20% annually compared to overall IT spend of 6%, companies must pay careful attention to quantifying and managing data center costs.

The portion of the IT budget consumed by infrastructure and facilities is significantly reshaping the economics of many businesses. In information

---

1   Source: EPA 2007 Report to Congress

intensive businesses like investment banks, telecoms and business information data center costs are diverting capital from new product development, making some products and segments uneconomical and materially affecting margins. Without radical changes in operations, many companies with large data centers face reduced profitability.

Exhibit 2: For information intensive industries, data center cost growth threatens to impact profitability materially

Data center operating income impact
USD, Millions

Client Example

Growth drivers

- **40% transaction volume growth**
- **16% database record volume growth**
- **Aggressive new product development plans in multiple business units**
- **High regional demand in Asia**

2007	2008	2009	2010	2011
100	105	113	121	129
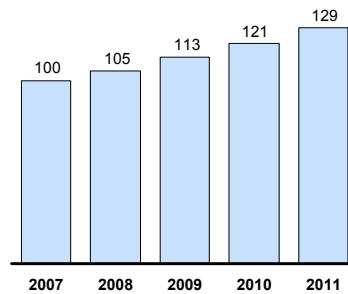
- **Data center cost**
  - Dramatically increasing as a percentage of revenue
  - Growing twice as rapidly as revenue
  - Significantly affects profitability for next two years

- **Ongoing data center build programs requiring huge amount of time and energy from senior technology team**

SOURCE: McKinsey & Company

Data center emissions are expected to quadruple by 2020

Due to their enormous energy consumption, data centers have a large and growing carbon footprint. Including the massive web farms at online service providers, data center electricity consumption is almost .2% of world production. Between 2000 and 2006 the amount of energy consumed by data centers around the world has doubled and today, the average data center consumes as much energy as 25,000 households[2].

 Data centers now drive more in carbon emissions than both Argentina and the Netherlands. By 2010 data centers in the US are expected to consume as much energy as 10 new major power plants.

---

2   Source: Financial Times; Gartner report 2007; Stanford University; AMD; Uptime Institute; McKinsey analysis

Exhibit 3: Data center emissions will quadruple by 2020, causing increased scrutiny from regulators, customers, boards and other stakeholders
Emissions from Data Centers worldwide
Mt CO2

11% CAGR	

340

80

2007   2020
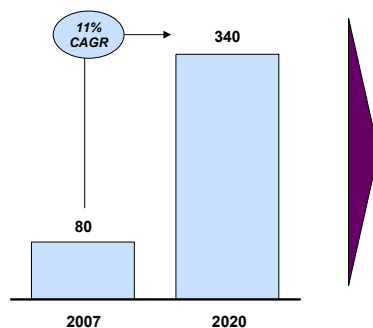
- US Public Law 109-431 requires EPA to submit a report on energy consumption of data centers to US congress

- EPA has advocated use of separate energy meters for large data centers and development of procurement standards

- The EU is developing a voluntary Code of Conduct for data centers proscribing energy efficiency best practices.

SOURCE: McKinsey & Company; IDC

This carbon footprint has begun to attract scrutiny from regulators, activists and corporate boards. In 2006, the US Congress passed a law requiring the Environmental Protection Agency (EPA) to submit a report on data center energy consumption[3]. The EPA also proposed that separate energy meters be used to monitor large data centers in an effort to spur development of procurement standards. Across the Atlantic, the European Union issued a voluntary Code of Conduct in 2007 prescribing energy efficiency best practices. The Green Grid, a global industry consortium focused on advancing energy efficiency, has also forwarded recommendations for improved metrics, standards and technologies to reduce data center carbon emissions. -

Local sub-optimization and inaccurate forecasting drive data center inefficiency

An institutions ability to build, manage and use data center capacity depends on multiple levels of decisions that span many parts of the organization. Business managers must make choices about what analytics they need or data they want to store. Application developers must design applications to meet business needs. Infrastructure managers must procure and configure servers to run applications. And facilities managers must build and fit out data centers house servers.

_____
3   US Public Law 109-431

Exhibit 4: Data center efficiency depends on a multi-level set of supply and demand decisions
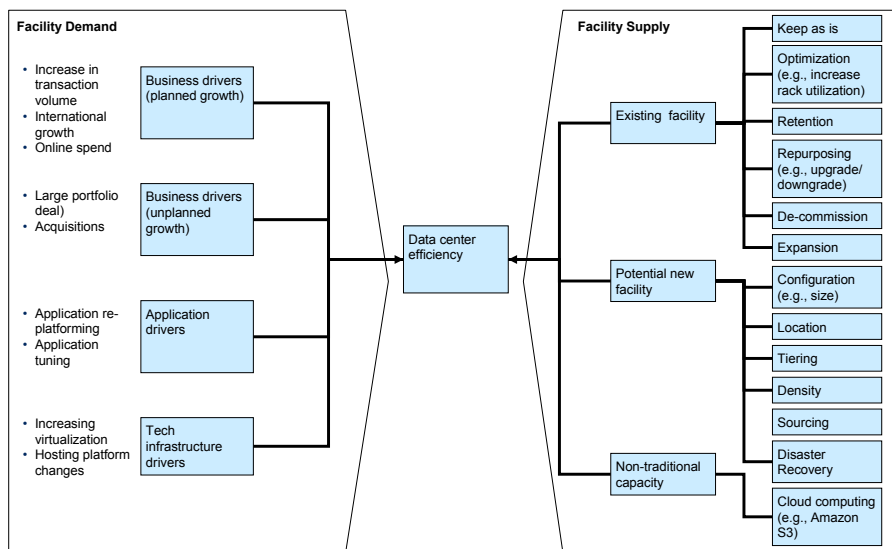
SOURCE: McKinsey & Company

Again and again we see two pervasive challenges in executing on this chain of decisions: location sub-optimization and inaccurate forecasting.

Location sub-optimization

Each step in the data center decision chain depends on input from other steps. How many iterations of a Monte Carlo analysis an equity derivatives desk should run or how much clinical imaging data a Pharma research group should store depends on the cost of processing and storage. How a development team should architect an application depends on what type of performance business users require and the cost of infrastructure. What type of server infrastructure managers specify depends on business requirements, application architectures and the cost of providing data center facilities. Decisions facility managers made about data center location, tiering and power density depend what type of servers infrastructure managers select.

All too often, decisions at each step of the chain fail to take into account impacts at other steps in the chain. Business managers choose to change transactions from overnight to real-time without understanding how this will impact the full set of data center costs. Developers decline to make investments in application tuning even when the payback in reduced server or data center capacity requirements far outweighs the effort required. Infrastructure managers select servers that have a low initial purchase price, but use data center power capacity

inefficiently, or they install applications on dedicated rather than virtual servers. And data center managers build large facilities to power densities or resiliency levels required by only a small number of systems.

Exhibit 5: At each step in the data center decision chain, there are sub-optimal decisions that drive up costs and environmental impact

Facility Demand		Facility Supply
Business decisions		
▪ Moving transactions from overnight to real-time		
▪ Running marginal analysis or scenarios		
▪ Specifying high availability for non-critical applications		**Existing facility decisions**
▪ Retaining multiple copies of historical data		▪ Making incremental investments facilities that should be decommissioned
Application decisions		
▪ Not investing in application tuning		
▪ Building non-parallel application architectures		**New facility decisions**
▪ Selecting commercial software packages that use resources inefficiently		▪ Building to power density not likely to be used fully by expected systems – especially by building uniformly to high power density
Infrastructure decisions		
▪ Hosting one application per server with low utilization		
▪ Limited use of highly efficient technologies like grid computing or MAID		

SOURCE: McKinsey & Company

Decisions like this expand data center footprints with large cost and environmental implications. Decisions about server management, like which servers to decommission and which to virtualize have a particular impact. Up to 30% of the servers housed within many data centers are functionally 'dead.' These are servers with less than three percent average daily utilization. They draw down power but serve a limited useful business purpose. Across the data center as a whole, average daily server utilization generally tops out at a low 6%, creating tremendous "waste" in terms of capital employed and energy used.

Inaccurate forecasting

As mentioned above the amount and type of data center facilities required depends on a set of business decisions which drive a set of application decisions which drive a set of infrastructure systems decisions. Two issues make forecasting data center requirements along this chain very difficult.

First is a timing disconnect – a data center takes two years to build and fit out, and it should last at least 10-12 years, without a major renovation. So

necessarily data center capacity must be built or procured in advance of business, application and infrastructure needs. And it will last far longer than typical business, application or infrastructure systems planning horizons. Who could have foreseen, for example, the application needs or server architectures of today in 1998?

Second are translation challenges. For most applications in most businesses, the relationship between business decisions, application architectures and data center capacity requirements is only incompletely understood. Many companies, for example, do not know whether a 50% increase in customer volume would require 25% or 100% more server and data center capacity. As a result, data center facilities can sit half empty, particularly just after construction. In other cases, companies find they complete one data center build program only to find, because of capacity constraints, they must launch a new one almost immediately.

Organizational limitations

Three organizational contribute greatly to challenges local sub-optimization and inaccurate forecasting in data center management.

First, accountability for data center facility management is often fragmented, with financial responsibility for critical facilities in real estate groups that lack the technical expertise to manage data center decisions closely. In many cases, infrastructure managers don't even see data center facility-related operating expenditures like electricity consumption.

Second, decisions get made in silos. Business leaders set product strategies with limited input from IT executives. Application developers consult with infrastructure managements late in the development lifecycle, long after key design decisions have been made and acted on. Infrastructure managers create technology roadmaps without consulting developers or data center facility teams.

Third, most organizations lack effective total cost of ownership information that's critical for business and technology decisions. The true cost of an incremental server, including labor, depreciation, maintenance and facilities. And organizations often cannot tie infrastructure to applications and applications to business drivers like products and customers, making close trade-offs between business value and technology cost all but impossible.

An integrated three-part solution for data center owners

Institutions can make achieve a step change in their data center efficiency by pulling three inter-locking levers: rationalize demand for data center capacity, optimize supply for data center capacity and put in place underlying organizational enablers.

When we began our research, we fully expected to find that building 'green,' energy efficient data centers would offer the best value for our clients. What we learned instead was that the 'greenest' and most cost-effective opportunities were in improving the efficiency of data centers that the clients already owned. By making substantive changes in how they manage their data center portfolio, organizations stand to make far and away the most significant gains in performance and cost savings.

Rationalize demand for data center capacity
Goal: Closely align data center capacity required with true business need

Multiple levels of demand contribute to data center requirements and must be addressed in order to eliminate excess.

1. Revisit business policies, including how much data gets retained, which analytics are performed, which transactions must be done in real-time and which applications real disaster recovery capabilities

Exhibit 6: Improve infrastructure asset utilization by improved solutions design, virtualization and application stacking

Client Example

- **Large environment (~20,000 servers) traditionally optimized for zero-downtime**

- **Pilot to prove concept, reduced installed capital by 80% in piloted applications**

- **Cleaned up asset inventory, set targets, and planned implementation for 20-25% reduction in installed capital**

- **Implement rigorous, Ehrlang-based process for server capacity management**
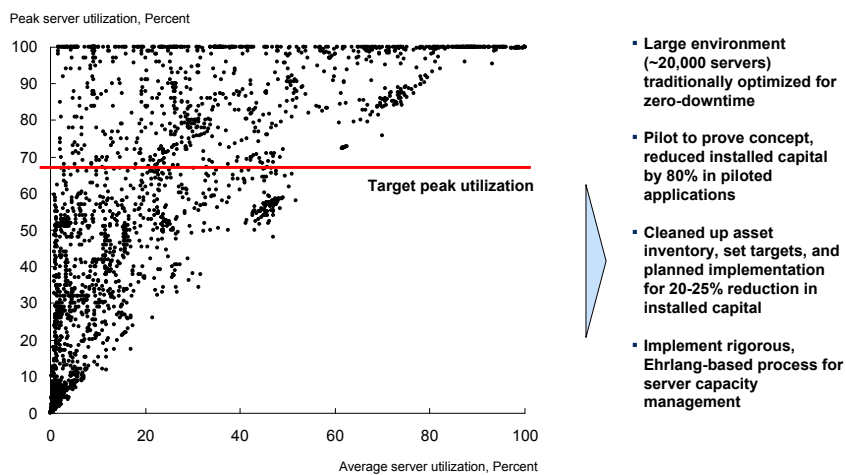
SOURCE: McKinsey & Company

2. Reduce infrastructure resource requirements by application tuning and re-architecture

3. Improve infrastructure asset utilization by improved solutions design, virtualization, application stacking and server de-commisioning

At the same, time there are multiple demand side levers to ensure that data centers provide capacity required with the greatest level of efficiency

4. Improve facility utilization by improved forecasting and capacity management

5. Optimize existing facilities (e.g. reduce cooling for hot aisles, restack equipment to improve space utilization

6. Optimize new facilities build out (e.g. design piping to use winter cooling)

Exhibit 7: Improve facility utilization by improved forecasting and capacity management

Client Example

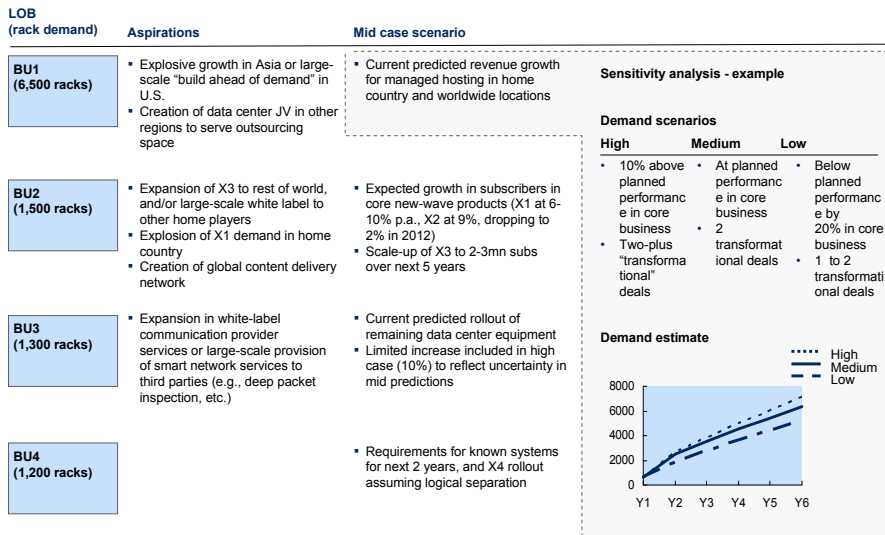| LOB (rack demand)	Aspirations	Mid case scenario	
BU1 (6,500 racks)	• Explosive growth in Asia or large-scale "build ahead of demand" in U.S. • Creation of data center JV in other regions to serve outsourcing space	• Current predicted revenue growth for managed hosting in home country and worldwide locations	**Sensitivity analysis - example**
BU2 (1,500 racks)	• Expansion of X3 to rest of world, and/or large-scale white label to other home players • Explosion of X1 demand in home country • Creation of global content delivery network	• Expected growth in subscribers in core new-wave products (X1 at 6-10% p.a., X2 at 9%, dropping to 2% in 2012) • Scale-up of X3 to 2-3mn subs over next 5 years	**Demand scenarios**
BU3 (1,300 racks)	• Expansion in white-label communication provider services or large-scale provision of smart network services to third parties (e.g., deep packet inspection, etc.)	• Current predicted rollout of remaining data center equipment • Limited increase included in high case (10%) to reflect uncertainty in mid predictions	**Demand estimate**
BU4 (1,200 racks)		• Requirements for known systems for next 2 years, and X4 rollout assuming logical separation	

Sensitivity analysis - example

Demand scenarios

High	Medium	Low
• 10% above planned performance in core business • Two-plus "transformational" deals	• At planned performance in core business • 2 transformational deals	• Below planned performance by 20% in core business • 1 to 2 transformational deals

Demand estimate

High ⋯⋯
Medium ——
Low – – –

8000
6000
4000
2000
0
Y1 Y2 Y3 Y4 Y5 Y6
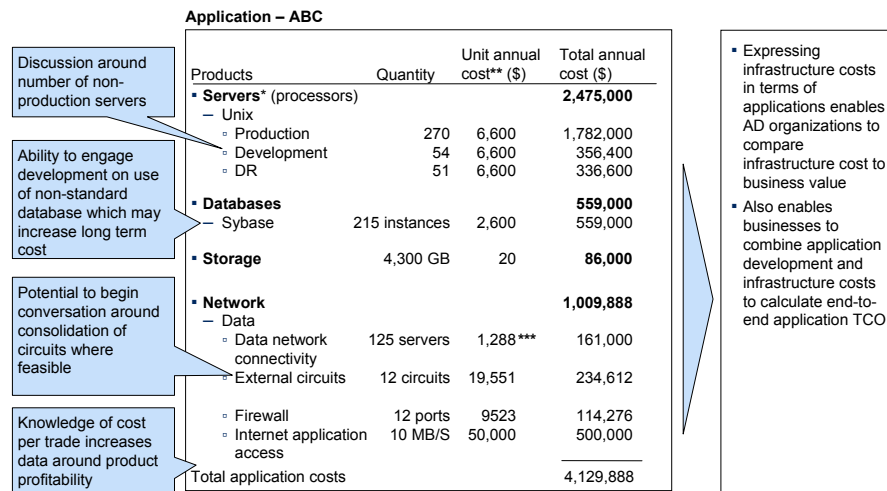
SOURCE: McKinsey & Company

Success in applying these demand and supply-side levers typically requires different organizational structures and management capabilities. There is a set of discrete actions institutions should undertake to enable data center efficiency improvement.

7. Invest in true application TCO that incorporates full lifecycle infrastructure and data center facility cost

8. Make technology leaders responsible and accountable for data center facilities, including decisions, capacity, cost and environmental impact

9. Mandate data center and infrastructure involvement in early in processes of green-lighting and designing applications

10. Implement metrics for data center energy efficiency

Exhibit 8: Invest in true application TCO that incorporates full lifecycle infrastructure and data center facility cost

Client Example

Application – ABC

Products	Quantity	Unit annual cost** ($)	Total annual cost ($)
▪ **Servers*** (processors)			**2,475,000**
— Unix			
▫ Production	270	6,600	1,782,000
▫ Development	54	6,600	356,400
▫ DR	51	6,600	336,600
▪ **Databases**			**559,000**
— Sybase	215 instances	2,600	559,000
▪ **Storage**	4,300 GB	20	**86,000**
▪ **Network**			**1,009,888**
— Data			
▫ Data network connectivity	125 servers	1,288 ***	161,000
▫ External circuits	12 circuits	19,551	234,612
▫ Firewall	12 ports	9523	114,276
▫ Internet application access	10 MB/S	50,000	500,000
Total application costs			4,129,888

Discussion around number of non-production servers

Ability to engage development on use of non-standard database which may increase long term cost

Potential to begin conversation around consolidation of circuits where feasible

Knowledge of cost per trade increases data around product profitability

- ▪ Expressing infrastructure costs in terms of applications enables AD organizations to compare infrastructure cost to business value
- ▪ Also enables businesses to combine application development and infrastructure costs to calculate end-to-end application TCO
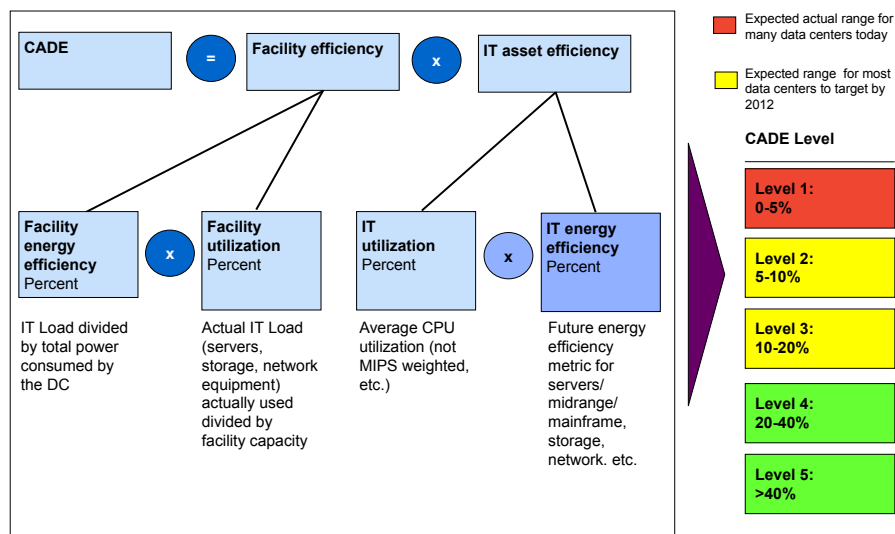
SOURCE: McKinsey & Company

Introducing CADE, a new standard for data center energy efficiency

With formal lines of responsibility in place, we believe a mandate to double energy efficiency by 2012 is the quickest and easiest way to improve an organization's return on assets and reduce GHG emissions. Such a goal we believe is not only achievable for most organizations but has the advantage of setting clear direction for the company while significantly lowering cost.

To achieve this doubling of energy efficiency, McKinsey proposes that CIOs, equipment manufacturers, and industry groups, in dialog with regulators, adopt a new Corporate Average Datacenter Efficiency (CADE) metric to measure the individual and combined energy efficiency of corporate, public sector and 3rd party hosted data centers.

Exhibit 9: Implement metrics for data center energy efficiency

SOURCE: McKinsey & Company; Uptime Institute

Similar to the automotive style "CAFE" standards, CADE measures data center efficiency across the entire corporate footprint[4]. Compared to other industry metrics, CADE is the first standard to combine facilities with IT, offering a much needed double lens.
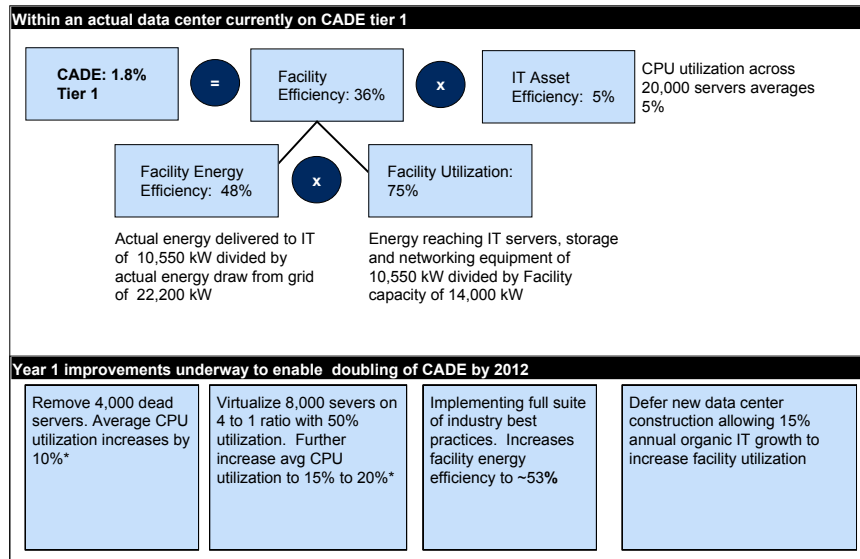
To measure how effectively the data center uses energy coming into the facility, CADE takes the amount of power consumed by IT, or the IT Load, and divides it by the total power consumed by the data center. To determine how fully the physical equipment installed at the facility level is being used, the CADE formula divides the IT Load by the facility's total capacity. This Facility Efficiency measure is then multiplied by the average CPU server utilization and yields the organization's CADE rating.

Each data center is measured independently with a weighted average value based upon installed facility capacity. Since data centers may draw upon different sources of energy, CADE can be used to determine the relative "cleanliness" of the company's GHG emissions.

---

4   The Corporate Average Fuel Economy (CAFE) regulations, first enacted by the US Congress in 1975, are federal regulations intended to improve the average fuel economy of cars and light trucks sold in the US.

Exhibit 10: Example improvements to CADE performance   Client Example

Within an actual data center currently on CADE tier 1

CADE: 1.8% Tier 1   =   Facility Efficiency: 36%   x   IT Asset Efficiency: 5%   CPU utilization across 20,000 servers averages 5%

Facility Energy Efficiency: 48%   x   Facility Utilization: 75%

Actual energy delivered to IT of 10,550 kW divided by actual energy draw from grid of 22,200 kW

Energy reaching IT servers, storage and networking equipment of 10,550 kW divided by Facility capacity of 14,000 kW

Year 1 improvements underway to enable doubling of CADE by 2012

| Remove 4,000 dead servers. Average CPU utilization increases by 10%* | Virtualize 8,000 severs on 4 to 1 ratio with 50% utilization. Further increase avg CPU utilization to 15% to 20%* | Implementing full suite of industry best practices. Increases facility energy efficiency to ~53% | Defer new data center construction allowing 15% annual organic IT growth to increase facility utilization |

SOURCE: McKinsey & Company; Uptime Institute

CADE confers a number of advantages. In contrast to other industry metrics, CIOs, CEOs and Boards now have a single, integrated metric that combines facility and IT energy efficiency levels to evaluate the total performance of their information factories. Just as the automotive industry can point to the miles per gallon their vehicles achieve, senior leadership can use CADE to reveal just how much efficiency their capital intensive data centers are driving. Most importantly, CADE is a metric that can propel tangible action.

To set targets for improvement, we have established five CADE Tiers as illustrated in Exhibit (X). Data centers operating at Level 1 have a CADE rating of 0-5% and are weakest from a point of view of efficiency. Centers operating at Level 5 are maximally efficient and have CADE ratings greater than 40%.

While most organizations are likely to fall within the lower bounds initially, leadership can use CADE tiering to establish performance goals. An organization currently at CADE Level 1, for instance, could seek to increase efficiency and move to Level 2 within 18 months. The ranges themselves will flux as more organizations seek to standardize upon them. Since the results are measurable, performance becomes easier to gauge across the organization as well as among individual data centers.

CADE has the potential to make a significant contribution to the data center community. Our hope is that industry, regulatory and government bodies will adopt CADE as a key metric in measuring data center efficiency and use performance against CADE standards as a means of developing appropriate incentives and penalties.

Conclusion

Data center inefficiency is a widespread and growing concern. Soaring costs and an ever-increasing environmental footprint are impacting the corporate investment landscape, threatening profitability, and inviting board and regulatory scrutiny. The consequences of inaction are high. Yet, our analysis shows that several immediate steps can be taken to arrest these negative developments. Data center owners and the industry at large should organize around our three-part solution and recommended best practices. We encourage the community to adopt our new CADE standard and commit to doubling efficiency by 2012. Through improved accountability, data centers can deliver on their high performance potential. We hope that industry can seize this opportunity.