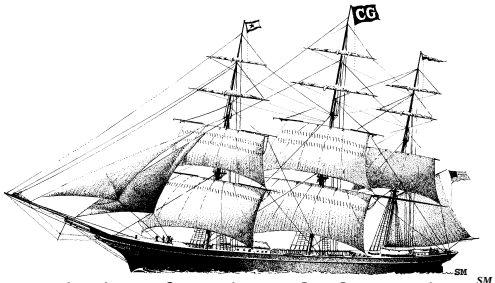
Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE FOR THE INFORMATION TECHNOLOGY BUYER



Navigating Information Technology Horizons℠

Published Since 2007 | Report #TCG2008050 | September 25, 2008

Greening the Data Center Requires Refocusing from "Bigger-Better-Sooner" to "Wise Use"

Analyst: Anne MacFarland

Management Summary

The more precisely we monitor and measure the Earth's environment, the more clearly we see the effects that our existence – and the things we do to optimize our existence – have on the planet. A century ago, scientists started addressing the problem of asbestos, arsenic, and other known toxins. A half century later, the U.S. E.P.A. and other governmental organizations institutionalized the demand for a solution. More recently, concern spread to something seemingly benign – carbon – that is not benign at the rate we are dumping it in our atmosphere. Today's escalating U.S. energy prices make controlling energy use a priority. Thus, ***green* is a convergence of two complementary shades – that of the environment and that of money.**

Best practices were easier to determine when resources were constrained and the process was local. Now that data centers suffer from a surfeit of plenty, with an onslaught of possible scenarios and sourcing options, *how to use resources wisely* becomes both imperative and hard to determine. The difficulty comes from technology's increasingly interwoven nature. The best infrastructure for an application depends on its use – and on what other applications it must feed or ask for information. Businesses are using more tiers of accessory or transformative applications as they attempt to address more markets and use information in multiple contexts (something that is second nature to human workers). The *efficiency* of business depends, not just on the integration of on-the-glass browser capabilities that underlies human productivity, but the coordination of data center components that support business process productively. In many cases, the *effectiveness* of the business depends not just on not making mistakes, but also on how a business leverages a wide range of technology for strategic advantage. Thus, **wise use and green use must accommodate many business variables and technical elements.**

Such an effort involves a substantial reassessment – first, of the status quo in all its messiness, and then of how doing things differently can bring about a permanent and substantial reduction of IT's dependency on polluting processes – while not involving a reduction in IT performance. There are a lot of proven strategies that can green any size data center. For a more extensive discussion about greening IT, please read on.

IN THIS ISSUE

Quantifying Greenness

If you can't quantify something, it is hard to discern how to change it. To produce a change, one must start with what one can measure and then derive useful metrics from those measurements. The more precisely we measure data center capabilities, the more effectively we can weigh our options. Modeling scenarios and employing charge-back for operation makes it easier to assess the tradeoffs of different strategies and sourcing options.

Work performed per Watt can become the unit of efficiency.[1] While this disadvantages information-intensive workloads that support strategy in favor of the brevity of transactions that document business as it is done, work per watt does represent – consistently and accurately – the cost of doing business the way you are doing it. It does so in a way that lets you compare options.

One must also define what is to be mitigated. In the financial kinds of green, it is *cost.* In the environmental green, it is *pollution* – and use of energy derived from polluting processes. These are inherently linked, due to the way we generate energy today. For some companies, however, it is more urgent to focus on reducing *energy consumption* by current operations, in order to free up energy capacity for IT expansion that is needed to get through the coming year.

Carbon has becomes the *metric of environmental atonement.* There is no position of absolute virtue – consumable energy is created by means that often involve fossil fuels and almost always involve some kind of environmental degradation, particularly at the scale at which data centers consume it.[2] Carbon is as artificial a metric as money is of value – but, as a metric, is a great way to compare and contrast – and improve.

Carbon credits is a concept developed to give energy squanderers time to reform their energy use strategies by paying for their excessive use, while rewarding companies who have reduced energy use by giving them a way of monetizing their "greater good" behavior.

Assessing in a Business Context

Business survival often mandates that doing less with technology is not a path to be taken. Instead, better use practices must be developed. To address the symptoms of low asset utilization, data centers are consolidating workloads on servers using virtual machines, using leaner provisioning of storage, and leveraging more disciplined configurations and templates to standardize operations. The symptoms of high cooling costs are addressable by a variety of strategies. These efforts mitigate the sprawl that occurred during the recent economic expansion and offer sizeable savings for most of today's data centers with relatively little effort. However, the rate of change in both business and the technology that supports business continues to accelerate. A new mode of thinking also is needed.

As part of any comprehensive assessment, the data center must consider the business model of the organization it supports. If meeting the demands of business can only be met by constant data center expansion, IT will be dancing a constant tradeoff between smaller form factors and newer equipment (to save costs and reduce energy use) and an increasing need to mitigate heat.

Many data centers aspire to attain, to some degree, a steady state of existence involving moderate growth. In a bounded organization – say, a university of a certain size, or a franchise or branch office serving a certain size population – this is a reasonable scenario. But, to make it work, **as new services are offered, old ones need to be retired. This takes either a benevolent business model (most aren't), limited ambitions (most don't want to be), or a ruthless discipline to *not do that anymore*.**

As we have found in sunset laws for some government projects, creating a retirement plan – for applications and systems as well as for physical infrastructure and data – is extremely useful. Archiving of data, done properly in a way that makes the data addressable independent of the application that generated it, can reduce application-licensing costs and cut server count.

Until turning things off is not seen as risky, the ability of the data center to constrain its power consumption will probably be insufficient. With copious bandwidth and sophisticated software, workloads can be moved and aggregated on fewer servers during lulls, and other equipment powered down – or off. Remember, *not doing* can often be accomplished by *doing differently*.

Measure more; assume less.

To optimize for a rate of change that underlies energy efficiency, you have to know more about what is happening and what your operational op-

[1] The unit of work must be defined consistently, but also within the context of the application (e.g., transaction throughput is a different kind of metric from analysis or transformation workloads).

[2] Widely distributed local power, where everybody runs the meter backwards, is as big a disruption to existing power services as the rise in user-generated content is proving to be to Internet services. It won't come easy.

Copyright © 2008 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

tions are. Characterize patterns of use. Identify underutilized assets – and leverage lifecycle TCO templates that can determine quickly whether it is worth upping the utilization in the short term. Find complementary workloads that can be co-located on a server advantageously.

For many businesses, success is measured in growth. While the soaring cost of energy in the United States, and impending governmental regulations worldwide, drive a need to control energy use, the fast, interactive pace of business must not be impaired. This makes green enough a moving target that requires not just discipline but strategic thinking.

Almost all strategies that underlie reducing the energy draw of technology while not reducing the benefit derive from a systems approach. Systems, by their nature, are designed for extensibility. They are built to support multiple stakeholders and multiple priorities, all of which are subject to change. This kind of thinking requires an up-front focus on what it is you want to do, quite apart from how you are going to do it. Ongoing assessments of the situation and feedback on operations are needed to address the problem as broadly as is necessary. A *whack-a-mole* approach will not do.

Greening Data Center Performance

It's a Good Time to Rethink

It is a great time to think about affecting the environment less, because the Data Center has a lot of levers to pull. There is *capacity*. There is *bandwidth*. There is *memory*. Moreover, there are many kinds of assets already at hand. Some – particularly the older hardware elements – just cost a lot, in both kinds of green, to use.

Focus on requirements and, as well, the various alternatives you have to meet them. With bandwidth and virtualization, your options may be greater than you think. Leveraging the heritage of grid for ongoing tasks where response time is not an issue, your desktops may be part of many solutions, particularly if they use low-power chips.

Many cost-effective "green" solutions are available today in how you address your processing needs, how you power and cool the processing environment, and how you provide both the performance and resilience that business requires. As part of a solution in using secure collaborative spaces to cut down on needless travel, technology can earn you carbon credits, not just consume them.

A New Definition of BIG and ENOUGH

The old kind of capacity planning – the requirements of *BIG* – focused on hardware redundancy. This was needed when hardware failed early, often, and unexpectedly. With better monitoring and autonomics, we have moved from fully-redundant IT environments to an *N+X* approach, where the sparing reflects anticipation of greater demand, not just of failure. Redundancy is accomplished in software by deploying multiple instances, as is needed, and by (1) using a load balancer to redirect requests around failure and (2) rapid deployment of an application clone to take up the slack. This is a new, more svelte concept of *big*.

In data center planning, what is *ENOUGH* has also become hard to determine. How do you predict the costs of operation of a particular piece of infrastructure when it depends on the operating system, the applications, and the usage patterns – which involve procurement decisions made by different stakeholders and management decisions made by different people? Looking at it from the other end, how do you determine the cost to support a business process, using a work-done-per-Watt metric, when exceptions to routine operations may involve unexpected costs? What is needed is a refocus from scenarios of theoretical numbers on energy consumption – and adding hardware as a solution for most problems – to a focus on far more precise measurements of what actually is going on. This is like moving from focusing on the probably cause of a case of sniffles to doing a throat culture and calculating which antibiotic, dose, and cadence will address the symptom. It does not replace old diagnostics, but it gives a fuller understanding of the situation. This systemic approach to greening the data center shows clear benefits in three areas: Design, Asset Utilization, and Support for Change.

Data Center Redesign

Redesign starts with an assessment of the status quo, enhanced by hardheaded assessments of what is needed. A majority of the data centers in use today are more than a decade old, and many are more than two decades. Like the schools built for the baby boomers after World War II, they have been adapted and extended in many unnatural ways. Many current usage requirements are at odds with their original design criteria.

Layout, Thermal and Cooling Issues

The use of computers has proliferated and the form factors have shrunk, generating the need to support more power, cables and cooling than the rooms – and often the buildings they are in – were designed for. Newer servers, like newer house-

Copyright © 2008 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

hold appliances, draw less power to produce equivalent performance. The reduced operating costs over the lifecycle of the equipment should be a part of TCO calculations – and may dictate moving workloads from older equipment. They also produce more heat, which may make traditional room air conditioners an insufficient strategy.

Design elements include the physicality and spatial layout, including vented tiles and cabling. The plenum of the raised floor is meant for ventilation, not for hiding cables. The vented tiles may be in the wrong places, if you have added equipment and changed form factors over the years.

Cooling requirements generated by more dense form factors are not effectively met by peripheral room-edge air conditioning[3]. The watchword is *Act Locally*. Computational fluid dynamics (no longer something that requires a supercomputer) can determine where the cooling should be located most effectively, as well as where enclosing equipment in a "pod" (discussed below) may be most effective.

Some of the design requirements of the data center have changed. Remote monitoring and management make human access to consoles, and, more generally, walking the aisles, always a potential risk, also less necessary. Higher computing densities make hot spots form faster, so more granular thermal monitoring probably will be needed be needed. More-efficient cooling strategies, such as chilled water, may be useful. (See Exhibit 1 on Water vs. Air, at the top of the next column.)

There are greater organizational benefits to be gained by extending the data center assessment to include the building, the campus, and even the national or global layout of properties. The full breadth of organizational operations can benefit from a system-level environmental assessment.

Physical Redesign Strategies

While dispersing the offenders to mitigate the problem may have worked in the past, mitigation is not a solution. The following new strategies are an antidote to laissez-faire sprawl.

- In the *cold-aisle, hot-aisle strategy*, equipment is arranged to vent to the hot aisle. This redesign cuts in half the volume of air to be cooled.
- More effective still, and counterintuitive to the old method of dispersing assets to defuse the cooling challenge, is the concept of *pods*, which uses metal covers and dedicated plenums, together with localized air conditioning, to make cooling still more closely targeted in scope and more efficient.
- Recently, several equipment vendors have offered *datacenter-in-a-box containers*. While many are targeted at highly-redundant, scale-out environments, such as service providers and niche businesses using only a few applications at scale, they are an extension of the above continuum. If they fulfill the promise of greater performance with lower price and energy usage, they may change sourcing discussions for businesses of all sizes.
- Leveraging cloud computing and software as a service (SaaS) is another way to avoid the need for data center growth and local energy consumption. To be consistent, an organization should consider its outsourcing as part of its total energy consumption. However, an outsourcer or SaaS provider may support deployment of a single application more efficiently than local deployment. This is particularly true for applications that are used irregularly.

Exhibit 1 – Water vs. Air

- Water absorbs many times more heat than air.
- Water holds a chill longer than air. It can be created asynchronously (say, at night when the air is cooler). When there is a power outage, air heats quickly. Chilled water persists, allowing recovery of functionality to be more prompt.
- Water obeys gravity. Air must be moved by fans, which use power and create heat.
- Both water and Air work most effectively when contained and directed. When contained, both approaches are safe.

Tools and Management

Many new tools enhance the environmental responsiveness of existing data center environments. Digital thermal sensors and metered power consumption at the box level let far more about the data center environment be both known and usefully aggregated. With the capabilities of very low-watt embedded systems (another recent development), thermal sensors can be linked to variable speed fans to throttle back the power demands of cooling when it is not needed. Assessment services from thermal experts, who know the trade-offs between different approaches, can be money well spent. Software exists to optimize the composition of racks for whatever cooling scenario you choose.

[3] Inefficient air conditioning can also remove moisture to an extent that is bad for electronic devices, necessitating air hydration that takes still more power.

Copyright © 2008 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

Accessory Systems

Uninterruptable Power Supplies (UPS) originally were intended to allow machines to shut down gracefully in times of power outages. With modern autonomics, the time needed to shut down is much shorter – and many UPS units are larger than is needed for that function. Used as an alternative to a generator as a stopgap during power outages, they are an extremely expensive alternative.

Think of how the circuit breakers (or fuse boxes) in your house are organized. They are sized to limit the power drawn to the rated capacity of the circuit's wiring. Proper sizing and use of infrastructure is a similar exercise.

Exporting the Problem and End of Life Issues

Reducing the hazardous materials in IT equipment has been going on for some time. European ROHS regulations have pushed technology vendors to minimize the use of toxic substances in technology components. Where the use of substances, like mercury, are hard to avoid, end of life recycling is key not just to keeping these substances out of landfills, but also to recovering them in a purity and volume that supports reuse.

Another part of addressing pollution is dealing properly with non-poisonous residue – such as heat. If there is a process to heat adjacent spaces with datacenter heat, that is good reuse – but if it is dumped into a larger business environment, heat mitigation will be less efficient and more costly. Exporting heat to become a facilities-level problem is not a solution and in fact makes it harder to solve. Instead, any green solution should expressly address both the data center's role in the larger challenge, and the end-of life strategies for IT equipment.

Enhanced Sharing

With the exception of mainframes, SMP, and grids, IT equipment and applications generally have not been designed to share resources opportunistically. Most applications have been developed in an environment that supports the assumption that anything that they can discover, they can use. So, this second area of environmental reform involves partitions, virtual machines, and other containers. It also involves messaging, file systems, authentication, scheduling, choreography, and the notion of declarative components that support the sharing that is the first step toward more domain-aware IT operations.

Server Consolidation and Networked Storage

Server Consolidation has been going on for some time to curb data center sprawl, leveraging partitions or virtual machines (discussed in section below on *Containers). Virtualization* lets a moving target of requests persist in the mind of the requester while the physicality (either servers or storage) that addresses those requests changes. The collocation of applications is supported by the virtualization of the resources that the applications share. If data latency is a problem, virtual I/O (also called virtual networks), can address the problem, using the same more granular approach to redundancy that was discussed in an earlier section on *BIG*. Virtualization strategies put the system (in the form of control elements guided by policies), rather than administrators, in charge of the sharing. When you add in instrumentation of the system by lightweight sensors, also discussed earlier, you start to leverage the operational benefits of tactical automation.

Pooling

A corollary of sharing is *pooling* – the ability to concatenate or federate assets (or access to assets) in order to provide a larger addressable and shareable domain. Some vendors address this pooling under the banner of *standardization* (often on their brand). Others propose functional groupings such as IBM's *ensembles*, or 3Par's *PcV*. Many are switching to terms like *orchestration* and *choreography* to address management of groupings instead of individual devices.

This pooling concept combines well with the modular data center approaches discussed earlier in the *Redesign* section. There is a trade-off between the consistency of pooling a monoculture, which is vulnerable to pandemic vulnerabilities and may be limited in scope, and the agility but increased grid-like overhead of pooled heterogeneity[4]. Industry standards, virtualization, and automated management allow you to get the best of both worlds – and also leverage everything you already have deployed.

In hardware, this is accomplished by the formalities of clustering (or, for mainframes, *Sysplex*), or by using the informalities of *tiering* and *load balancing* mentioned earlier in this paper. For information assets, tools such as *XQuery* (XML-based query language that works across

[4] For a data take on this trade off, see **The Clipper Group Navigator** entitled *The Data Side of Grid: The Role of Containers and a Single Name Space* dated April 19, 2005, and available at http://www.clipper.com/research/TCG2005022.pdf.

Copyright © 2008 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

federated data sources) and, more generally, *search*, allow you to address an aggregation of assets. Large and clustered file systems allow large (i.e., up to enterprise-wide) logical-level aggregations of data sources. The larger your asset pools, the more operations scale without requiring additional hardware. Pooling also supports optimized use of special-purpose components, to which appropriate workloads can be offloaded or by which they are optimized.[5]

Containers and other Agents of Segregation

Many application *containers* have come on the market over the past several years. They range from rigid and flexible partitions whose controls are in hardware management to virtual machines that are software constructs. They include both semi-permeable and self-defining containers, as well as jukeboxes and other container systems. They differ in their complexity and overhead.[6]

Any discussion of containers would not be complete without a discussion of system assets you may not think of as containers. Don't forget to leverage all your system assets. Today's bandwidth plus data and application mobility allow networked desktops, which now sport a lot of processing power, to be used for internal grid-style execution of routine, low-priority processes. Assets such as desktops, in this era of the browser, are often sparsely utilized. *Why not use the 97% of capacity that is idling … for other workloads?*

Support for Both Constant and Unanticipated Change

Business constantly changes, but it does not always grow. It may even divest itself of operations that are better done elsewhere. The same is true of the data center. Bandwidth and new vehicles of collaboration have increased sourcing options, and the economies of doing one particular thing very well have made sourcing options attractive.

Every data center should have a plan of how to degrade in response to a power curtailment. Many of these plans involve diverting transportable workloads and non-data intensive applications to remote sites. Remote data centers, particularly when the two are coordinated in an *active-active* fashion, may support operational resilience more cost effectively.

Conclusion

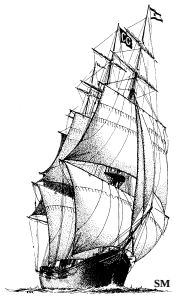
It is time to rethink how you will use and support the technology that has become ingrained in your business. This is particularly true for companies facing an increase in competition, or in data that must be retained.

Software-focused strategies for availability and scalability of operations, and current networking availability and capacities, favor the development of very large, very dense data centers. Internet irregularities favor the development of very distributed, very local capabilities (think of computing in delivery trucks and police cars). With telecommunications, organizations can leverage the best of both worlds.

Consider and measure your technology environment as a system, rather than as a collection of individual elements. Characterize its patterns of use. Look at your funnel of projects to be implemented not simply as a to-do list but as a design challenge, using energy costs and availability as the key constraints.

Leverage redundancy in software, application images, and data mobility as tools. Use retirement of assets and applications, outsourcing, and use of gridding workloads to free up space and energy capacities in the data center. You will find you your data center can use much less energy, and can reduce operational costs as well. Once you are on a more sustainable course of energy use, you can start looking at ways the data center.

Some businesses are looking at greening as a way to get through the foreseeable (short-term) future (perhaps while waiting for the next service-oriented "something"). Others are looking at it as a new parameter of operations that must become part of long-term strategy. The best way may be to think of it as an inevitable imperative, which casts IT operations as both a villain and a critical part of the solution. Plan well, measure copiously, and survive long enough to prosper.



[5] Familiar examples of this are appliances for encryption, deduplication, data compression, etc.

[6] For more about Virtual Machines, see the issue of Clipper Notes entitled *Virtual Machines, Three Things to Consider and Three Ways to Use Them*, dated February 28, 2007, and available at http://clipper.com/research/TCG2007029.pdf.

Copyright © 2008 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Data Strategies and Information Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **Clipper Notes,** and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.

Copyright © 2008 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.