



Metered IT: the path to utility computing

By Paul Miller

This research was underwritten by 6fusion.

Table of contents

Table of contents	2
EXECUTIVE SUMMARY	4
THE VALUE OF MEASUREMENT	5
What can we measure?	8
Allocation vs. consumption	9
UTILITY COMPUTING	11
Measuring consistently	13
The Workload Allocation Cube	14
The Service Measurement Index	16
CloudCmp	17
CloudHarmony	18
Clouddorado	18
U.S. National Institute of Standards and Technology	18
Standard Performance Evaluation Corporation's Open Systems Group	19
PROS AND CONS OF UTILITY COMPUTING	19
Pros of utility computing	19
Compare and contrast	20
Cons of utility computing	21
FUTURE POSSIBILITIES	23
A common measure?	23
Compute: the new commodity	24
KEY TAKEAWAYS	26





ABOUT PAUL MILLER	27
ABOUT GIGAOM PRO	27



Executive summary

Agility is increasingly recognized as one of the main advantages of cloud computing, replacing earlier justifications such as cost reduction and reduced environmental impact. But an important aspect of agility is choice: the choice to run computing jobs in house, in a private cloud, or on public cloud services from the likes of Amazon, Rackspace, and a growing number of other providers.

To exercise choice, customers require information and the ability to compare the costs and benefits of competing solutions. Current pricing models for most cloud solutions make conducting meaningful comparisons difficult and reduce the ease with which customers can select the best infrastructure for different computing jobs.

This report explores opportunities for accurately measuring computing resources and their use, simplifying the comparison of competing cloud offerings and opening the door to charging models based more closely on actual consumption.

As [Amazon CTO Werner Vogels writes on his blog](#),

“Many of our customers come to AWS with a reduction of TCO and other cost savings in mind but after using AWS for a while most of them will claim that agility is the even more [significant] benefit for them. Agility has many different faces within an enterprise: Operational advantages such as setting up infrastructure in minutes rather than months, completing massive computational projects with a large number of resources quickly, and scaling architecture up and down to provide the needed IT resources only when you need them, deliver targeted IT solutions fast for individual business units – these deliver a ‘return on agility.’ The return on agility delivers business value



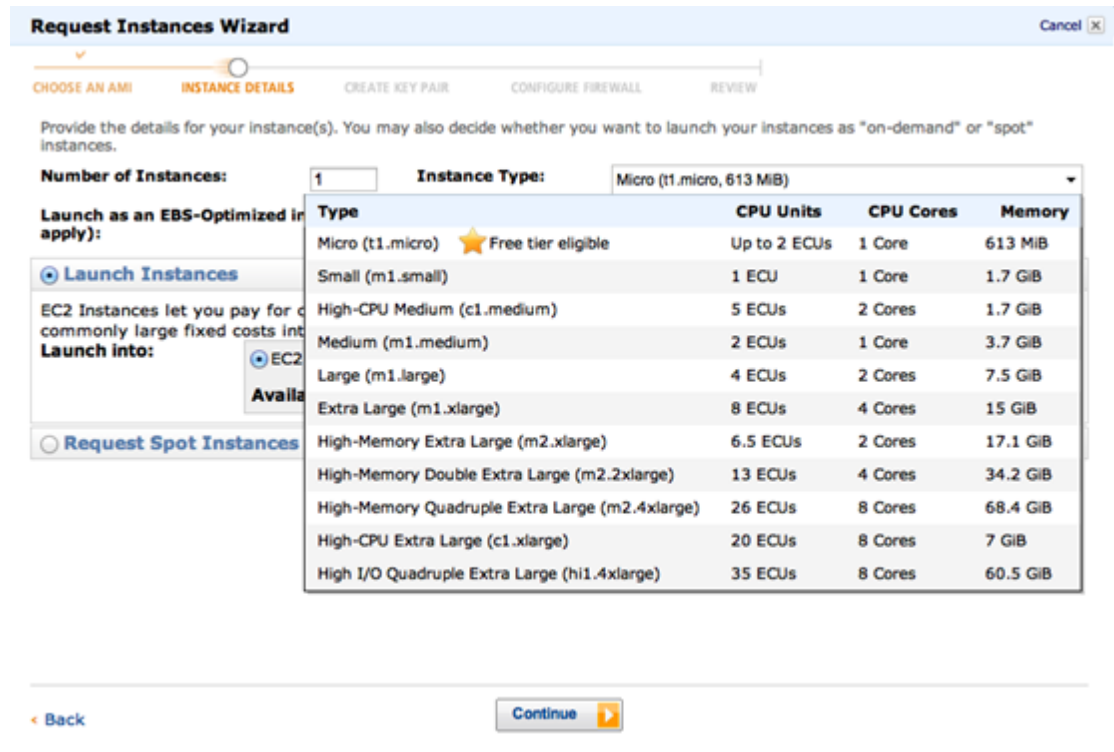
by allowing you [to] adapt rapidly, while remaining focused on your core competencies rather than [be] distracted by operating IT infrastructure.”

The value of measurement

Solutions providers in the IT industry promote their products by means of apparently quantifiable attributes: speed, efficiency, capacity, cost, and more. Providers of cloud infrastructure also do this, as anyone who has tried to evaluate competing cloud-infrastructure offerings can attest to. But what are the relative merits of Amazon’s [high-memory extra-large instance](#) compared to a [15 GB Rackspace cloud server](#) or a Dell server configured and physically deployed in your own data center?



Figure 1. Choosing a new virtual machine from Amazon Web Services



Request Instances Wizard Cancel X

CHOOSE AN AMI **INSTANCE DETAILS** CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

Provide the details for your instance(s). You may also decide whether you want to launch your instances as "on-demand" or "spot" instances.

Number of Instances: Instance Type:

Launch as an EBS-Optimized instance (if applicable):

Launch into: EC2 Request Spot Instances

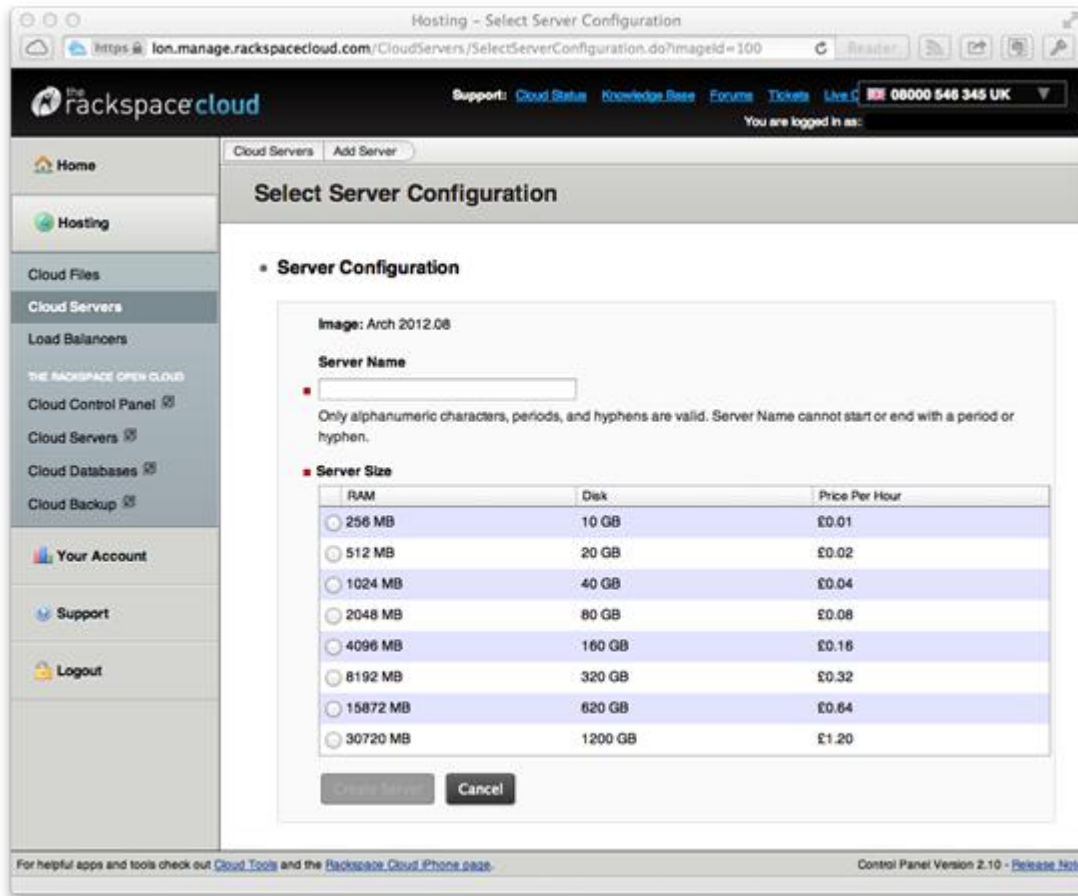
Type	CPU Units	CPU Cores	Memory
Micro (t1.micro) ★ Free tier eligible	Up to 2 ECUs	1 Core	613 MiB
Small (m1.small)	1 ECU	1 Core	1.7 GiB
High-CPU Medium (c1.medium)	5 ECUs	2 Cores	1.7 GiB
Medium (m1.medium)	2 ECUs	1 Core	3.7 GiB
Large (m1.large)	4 ECUs	2 Cores	7.5 GiB
Extra Large (m1.xlarge)	8 ECUs	4 Cores	15 GiB
High-Memory Extra Large (m2.xlarge)	6.5 ECUs	2 Cores	17.1 GiB
High-Memory Double Extra Large (m2.2xlarge)	13 ECUs	4 Cores	34.2 GiB
High-Memory Quadruple Extra Large (m2.4xlarge)	26 ECUs	8 Cores	68.4 GiB
High-CPU Extra Large (c1.xlarge)	20 ECUs	8 Cores	7 GiB
High I/O Quadruple Extra Large (hi1.4xlarge)	35 ECUs	8 Cores	60.5 GiB

[Back](#) Continue ▶

Source: Amazon Web Services



Figure 2. Choosing a new virtual machine from Rackspace



Source: Rackspace Cloud

The basic units of CPU, memory, storage, and bandwidth are broadly consistent and mostly understood, but they are combined and interconnected in ways that make comparisons difficult. More (or faster) memory from one provider may have a greater impact on actual performance than the more powerful CPU offered by the competition, for example. Equally, the apparently low up-front cost of one provider's solution may be offset by the extra staff effort required to configure and maintain the provider's servers.



This complexity is not new or unique to cloud computing. But compute jobs are increasingly able to break free from the static confines of corporate data centers and to run wherever makes the most business sense at the time, which is a clear requirement for being able to compare, contrast, and make informed business decisions. Equal to this requirement is a pressing need for tracking all the costs associated with running and supporting a computing task more realistically, enabling real comparisons among internal IT costs (which include factors such as the cost of support staff and network bandwidth) and the apparently lower costs of cloud providers (which typically do not include those extra factors).

The following pages look at the role shared metrics will play in enabling the emergence of a utility market for computing resources — something that should ultimately benefit both providers and consumers of computing power.

What can we measure?

Computing specifications are filled with numbers and provide the perhaps false impression that meaningful — that is, actionable — insight is always gained from the metrics being recorded. Some of the most significant variables that can be observed and recorded include:

Power consumption	Background power	Power cost
CPU speed	CPU utilization	CPU cost
Memory capacity	Memory speed	Memory utilization
Memory cost	Storage capacity	Storage speed
Storage utilization	Storage cost	LAN capacity
LAN speed	LAN cost	WAN capacity



WAN speed
Staff effort

WAN cost
Staff cost

Resource contention
Facility costs

In the vast majority of these cases, the measurement will not result in a single value. For example, senior managers cost more per hour than technicians. Flash storage will be faster, less capacious, and more expensive than the disks in the archival storage array. A cloud provider will often waive the cost of transferring data over the WAN into the data center but retain (or inflate) the cost of transferring data back out again.

It is possible to record values for each of these variables and for a host of additional factors affecting the performance and operation of everything from an individual virtual machine to an entire data center. While onetime measurement is interesting, ongoing measurement, which requires a standard and consistent measure, provides far greater insight and value. We can only extract full insight from these numbers when we also clearly understand how they relate and how the underlying processes affect the timely and cost-effective completion of real computing tasks.

Allocation vs. consumption

Typically users procure cloud providers (and, in an enterprise data center, pools of virtual machines) according to an allocation model: A user selects, for example, a Linux virtual machine with a dual-core 2 GHz CPU, 8 GB of memory, and 128 GB of storage. The user is then charged for that machine, even though usage may only actually consume 50 percent of the available memory and 10 percent of the available storage.



Tim McGuire, the manager of Messaging, Collaboration, and Workgroup Services at the University of North Carolina at Chapel Hill, illustrates the scale of this problem

with reference to the VMware cluster he runs. Physical machines were typically dual 6-core devices with 48 GB of memory, divided up into a number of virtual machines that users could request. According to McGuire:

“The problem with allocation-based chargeback models is that everyone buys up the memory. We then have a memory shortage and a glut of idle CPUs. So we don’t get the real benefit of slicing up the full cost of the entire cluster. We can’t easily charge for simple percentages of each element in the cluster, because we don’t then recoup our costs reliably.”

The alternative is a consumption-based model that charges customers for the resources they actually consume. This is far closer to the commonly understood model of the way in which a utility should behave, but it is likely to result in less-predictable (although possibly lower) costs for the customer.



Utility computing

For years commentators such as writer Nick Carr have compared the emergence of cloud computing to the rise of the shared electricity generation and transmission networks that became today's ubiquitous power utilities. In his 2008 book *The Big Switch*, Carr repeatedly draws upon examples from late 19th and early 20th century power generation to illustrate computing's shift to the cloud. He documents the transition from a time when factories and businesses generated their own power, and he suggests that today's move from enterprise data centers to the cloud is similar. Just as electricity generation became a utility, he claims, now computing is becoming one.

The notion that computing (especially but not solely in the cloud) is becoming, or has become, a utility is compelling and pervasive. But what are the characteristics of a "utility," and can we see them in today's computing models?

Writing in the *IBM Systems Journal* (PDF) back in 2004, Michael Rappa identified six characteristics he believed were common to utility services, from water and power to radio, television, and internet access:

- **Necessity.** The extent to which customers depend upon the service on a daily basis
- **Reliability.** The presumption that, according to Rappa, "temporary or intermittent loss of service may cause more than a trivial inconvenience"
- **Usability.** Simplicity at the point of use; for example, users do not need to know how electricity powers lights at the flick of a switch



- **Utilization rates.** Coping with peaks and troughs in customer demand, using for example, innovative pricing models that incentivize an even spread of demand
- **Scalability.** Benefits of economies of scale, with larger providers typically realizing lower unit costs that can be passed on to customers
- **Service exclusivity.** Government intervention that encourages the emergence of a monopolistic provider may be a benefit when utilities have significant setup costs or a particular requirement for scale

Rappa goes on to suggest that a business model for the provision of utilities is “based on metering usage and constitutes a ‘pay as you go’ approach. Unlike subscription services, metered services are based on actual usage rates.”

This principle of metering and payment for real usage has been an important aspect of the growth in public cloud services such as Amazon’s and Rackspace’s. Even inside corporate data centers, the rise of the private cloud is creating opportunities to charge more-realistic IT costs to the individuals, teams, and departments actually using the infrastructure. The growing tendency is to focus on a proportion of the physical computer’s resources — allocated to powering a virtual machine — for the time that it is actually required to run a particular task rather than purchasing and managing whole physical computers. Customers (whether users of public cloud resources or internal consumers of computing facilities provided by an enterprise IT organization) typically expect to be able to buy — and pay for — only the computing they need and actually use. Providers typically absorb the cost of any unused resources, incentivizing them to utilize their resources wherever possible.

Although Rappa does not enumerate on it, one other factor is crucial to the emergence of a market that supports utilities: the ability to compare providers and the means to move somewhat freely from one to another. Customers increasingly expect to be able



to compare costs and capabilities and then move with relatively little difficulty from one provider to another as circumstances change. With power, water, telephony, and other utilities, consumers can fairly easily compare providers and switch if they choose. With data center IT, standardization around a relatively small set of operating systems (Windows and a few Linux distributions) and virtual machine hypervisors (Xen, KVM) create a few technical barriers to moving from one virtual machine to another. An Ubuntu Linux or Microsoft Windows virtual machine running inside a private data center mostly looks the same and works the same as an Ubuntu Linux or Microsoft Windows virtual machine running in any number of public and private cloud providers. The underlying hardware is largely abstracted away, leaving the developer or application to interact with a familiar and predictable environment.

The biggest barrier to movement is not really technical but financial. As we explored above, given the wide variation in the way capabilities are described and priced, meaningfully comparing one provider with another becomes almost impossible. Without that meaningful comparison, it's not as easy to assess different offerings in the market. Rather than shopping around or promiscuously adopting multiple cloud providers, the far greater tendency is simply to continue using the familiar provider, whether it offers the best environment for a specific job or not.

Measuring consistently

Moving from the current IT landscape into one in which units of computing can be bought and sold according to models that would be recognizable to anyone familiar with other utilities requires a key element: a consistent means of measuring. Those measurements will typically encompass capability (of an IT resource), utilization (of that resource by a customer), and requirement (for specific attributes such as



storage capacity or network bandwidth) to deliver a holistic view encompassing supply and demand.

Despite a widespread recognition that consistent metrics would be of value to the customer, little evidence suggests that much of the work in this area has proceeded far enough to deliver real benefits. Standards bodies such as NIST and groups like SPEC are among those grappling with the problem, so far without much success.

Figure 3. Monitoring real consumption of various resources as they change over time



Source: 6fusion

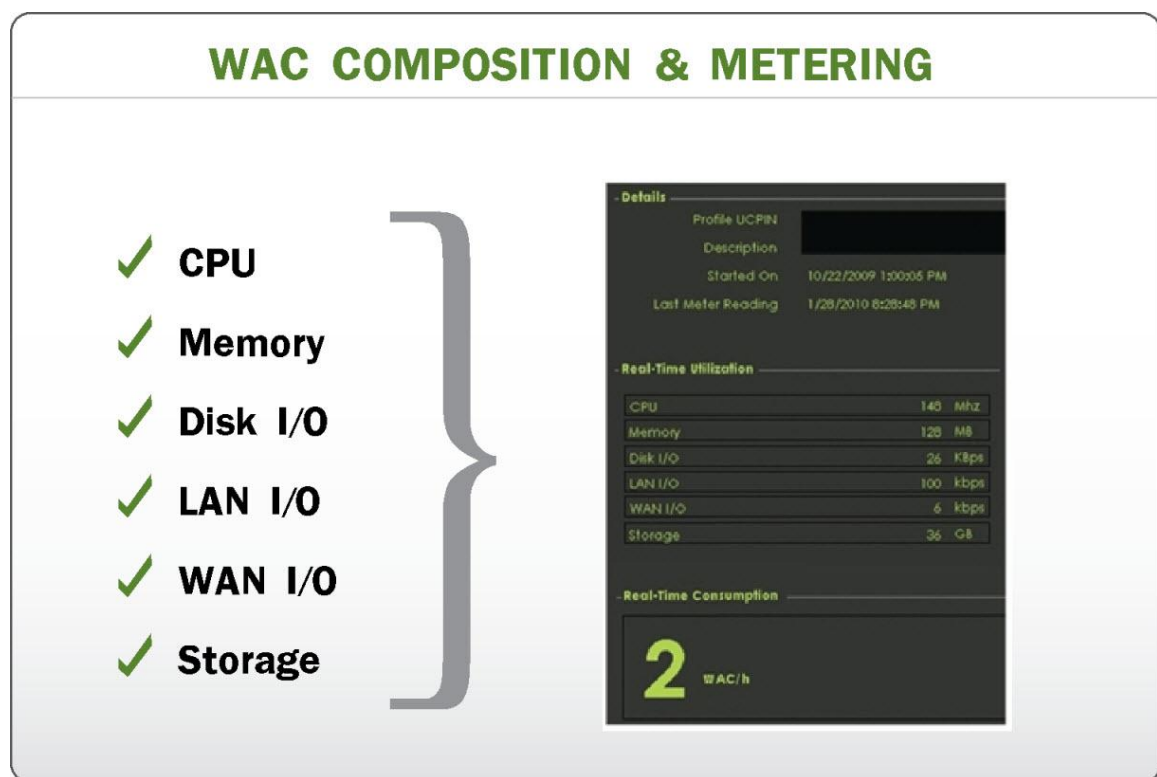
The Workload Allocation Cube

One attempt to identify and standardize a set of relevant measures is 6fusion's Workload Allocation Cube (WAC). According to the company,



“The WAC algorithm and costing methodology dynamically blends the critical compute resources required to operate practically every x86 based software application, yielding a single unit of measurement. It’s called a cube because a cube has six sides and there are six compute resources that comprise a single WAC unit: CPU, Memory, Storage, Disk I/O, LAN I/O, and WAN I/O.”

Figure 4. 6fusion’s Workload Allocation Cube



Source: 6fusion

Rather than consider all the variables we discussed earlier in this report, the WAC uses a proprietary algorithm that blends just six of the most significant variables, resulting



in a single value so that comparing alternative computing configurations from competing providers becomes far easier.

At the University of North Carolina, McGuire used the WAC to better understand the utilization of the university's own VMware cluster. This initially resulted in a realization that the basic hardware configuration did not really suit actual usage, leading to a shift from buying computers with 48 GB of memory to buying those with 72 GB of memory. The university is now also moving from its existing allocation-based pricing model to a consumption-based charging model.

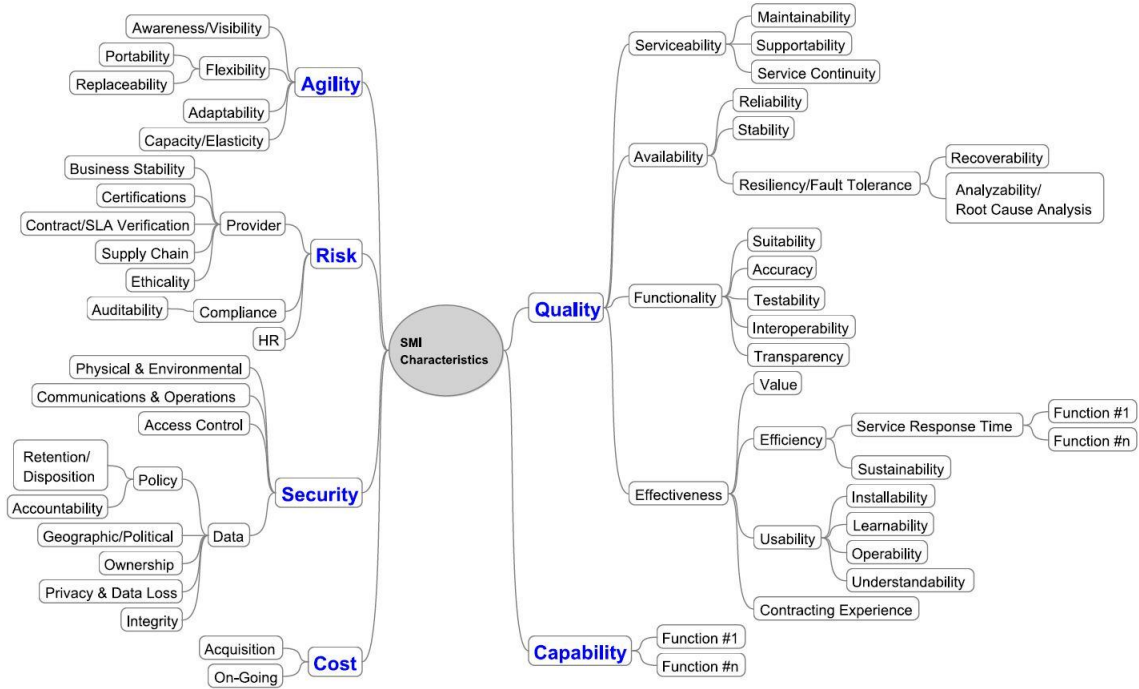
The Service Measurement Index

Sponsored by CA and led by a group at Carnegie Mellon's Silicon Valley campus, the Service Measurement Index (SMI) is **described as** a "set of business-relevant Key Performance Indicators (KPI) that provide a standardized method for measuring and comparing a business service regardless of whether that service is internally provided or sourced from an outside company."

The SMI groups a large number of KPIs (including measurable attributes such as "availability" and less tangible measures including "learnability") into six wide-ranging areas. The methodology is comprehensive, but it appears to be intended primarily as a tool to support the detailed desk-based evaluation of competing technology platforms rather than something that will deliver a single metric to drive ad hoc decision making.



Figure 5. The Service Measurement Index, Cloud Commons



Source: Jeff Abbott, CA Technologies

CloudCmp

Described in a [2010 paper from Duke University](#), CloudCmp comprised a set of metrics intended to support the consistent comparison of specific compute jobs running on different public cloud infrastructures. The research methodology sought to account for technical differences on each of the clouds to arrive at realistic comparisons for operations such as uploading and downloading data. At the time of writing, the [project website](#) no longer appears to be available.



CloudHarmony

CloudHarmony offers a wide range of more than **100 benchmarking tests**, each designed to test a particular aspect of system performance. These include general assessments of performance as well as very specific measures of attributes such as the write speed of system RAM. Once selected, tests can be run against various cloud system providers and the results can be compared. For application developers with very specific requirements, tests such as those offered by CloudHarmony can be important in ensuring that they select the best cloud provider for their needs.

Clouorado

Clouorado offers a simple online tool with which users can select a desired specification (CPU, RAM, storage) and view the prices charged by various public cloud providers to deliver that requirement. Like many of the companies and tools discussed in this research report, Clouorado is unable to provide direct comparisons with the on-premise cost of providing the same service from a private data center.

U.S. National Institute of Standards and Technology

The U.S. National Institute of Standards and Technology (NIST) supports an ongoing program of cloud-standardization work that has already delivered a well-regarded **definition** of cloud computing. The group's road map includes work on defining common metrics for cloud computing, which are likely to follow an approach similar to the one described above for SMI. Work is at an early stage, with little publicly available at present.



Standard Performance Evaluation Corporation's Open Systems Group

The Standard Performance Evaluation Corporation (SPEC) develops IT benchmarks across a range of areas, including CPU performance, graphics, and so on. SPEC's Open Systems Group includes a cloud subcommittee that has been established to develop benchmarks capable of measuring "the performance and scalability of cloud computing services." A report from April 2012 suggests the group is currently at an early stage in its work.

Pros and cons of utility computing

With the means to reliably measure available resources — and their utilization — considering available computing resources as a utility begins to be feasible. But what might this mean for the management, procurement, and use of these resources in the future?

Pros of utility computing

A utility model of computing delivers numerous benefits to customers. Many of these benefits are the same as those attributed to cloud computing in general, including:

- A move from capital expenditure (capex) to operating expenditure (opex)
- Rapid provisioning of computing
- Multitenancy
- Improved utilization of computing capacity
- Scalability and elasticity
- Agility



With the ability to measure and report cost and utilization metrics, additional advantages become apparent, including:

- Transparency of cost
- Metering and reporting
- The ability to compare and contrast different providers and to make informed decisions about what jobs cost and where they might best be run

Compare and contrast

This ability to compare and contrast the true cost and value of different computing solutions is a compelling piece of utility computing's promise. Assigning a single metric (such as 6fusion's WAC) to available computing resources and different tasks means that accurately managing resources and considering the cost of competing solutions is possible. At the University of North Carolina at Chapel Hill, McGuire recognizes the value of being able to evaluate different IT solutions according to a common metric. "If a WAC of compute in my VMware cluster costs x , a WAC of compute in my Xen environment might cost $0.8x$."

McGuire and his customers are able to compare very different solutions far more easily, and they can extend this ability off-campus so they can consider public cloud solutions such as those from Amazon. McGuire notes, however, that the UNC faculty tends to believe that his on-campus services are expensive. This is partly because they compare his charges to "the basic price of an EC2 instance on the Amazon website," without factoring in additional charges for getting data in and out of Amazon. The WAC helps make these comparisons more accurate, but it currently fails to consider



the cost of some factors outside the virtual machine (VM) itself — most significantly, staff costs.

Brent Eubanks, the executive VP of Strategic Services at Broadcloud, is another fan of metering models such as 6fusion's WAC because of the transparency it brings to costing. Eubanks suggests that customer demand for transparency may be outpacing the industry's ability to offer it. In disaster situations, for example, Eubanks suggests that the United States' Federal Emergency Management Agency (FEMA) has a requirement to access and use additional computing capacity. The amount of computing required will typically have been preapproved, and contracts will have been signed in advance of need. But the best data center to service a hurricane in Louisiana is probably not the same data center that would be used to cope with an earthquake in California or a twister in Kansas. With each data center offering different hardware, software, and pricing, FEMA's budgeting would be greatly simplified if it could procure a recognizable – and transferrable – unit of computing such as the WAC.

Cons of utility computing

The single biggest concern with adopting the rigorous metering of infrastructure lies in creating a misleading sense of accuracy. If we believe a measure such as the WAC contains every important variable in the decision-making process, then the measure will inevitably become the sole factor used in making decisions. Today, however, models such as the WAC are not comprehensive. They are part of a decision-making process that still requires us to consider other factors.

At UNC, McGuire points to variables outside the virtual machine itself, such as costs for power and staff time. At Broadcloud, Eubanks notes that the company has gone as far as to create a further metric known as the “Broadcloud kilowatt-hour” (BKWh),





which is based on the WAC and factors in additional metrics of value for its customers such as security and compliance.

When metering and metrics lead to consumption-based charging, an additional concern is that budgeting may become more complex. Rather than allocating a fixed (and large) sum to procuring physical hardware or a fixed (smaller) sum to procuring a number of virtual machines from a provider, consumption-based charging will typically lead to far less predictable usage patterns. Costs may well be lower than they would be under the other models (as you only pay for what you actually use, not what you think you'll use ahead of time), but planning and budgeting will be more complex.



Future possibilities

The choices for running computing jobs continue to grow as new cloud providers emerge and moving tasks from private data centers to the cloud and back again becomes increasingly feasible. The technical barriers continue to fall, pushed by open-source cloud projects such as [OpenStack](#), a [collaboration between Eucalyptus and Amazon](#), and the continued product innovation from companies such as [VMware](#).

A common measure?

Two significant remaining barriers to the routine movement of computing tasks are the lack of comparability among public cloud providers and the lack of comparability among internal capabilities and external providers. Without easier ways to compare and contrast costs and capabilities, the barrier to entry may well be too high.

Models such as 6fusion's WAC point to one way in which the capabilities of individual providers can be exposed in a more transparent fashion. 6fusion's algorithms are proprietary and only concerned with the six key variables governed by the WAC. This model enables differentiation among the various providers to the company's [iNode Network](#), and it can also be used to provide insight into local data centers. Internal IT offerings such as the VMware service at UNC Chapel Hill use the model for managing their finite resources more efficiently. Cloud providers such as Broadcloud use — and extend — the model to compare jobs running on different infrastructure accurately.

With more customer demand for portability, could there be an opportunity for the IT industry to describe its products in a more consistent fashion?



Compute: the new commodity

As computing capacity becomes more transparently described, the opportunity arises for it to be treated as a commodity that may be freely bought and sold. Wikipedia's [explanation of the term](#) includes a telling section:

“[The term] is used to describe a class of goods for which there is demand, but which is supplied without qualitative differentiation across a market. A commodity has full or partial fungibility; that is, **the market treats it as equivalent or nearly so no matter who produces it.** ‘From the taste of wheat it is not possible to tell who produced it, a Russian serf, a French peasant or an English capitalist.’ Petroleum and copper are examples of such commodities. The price of copper is universal, and fluctuates daily based on global supply and demand. Items such as stereo systems, on the other hand, have many aspects of product differentiation, such as the brand, the user interface, the perceived quality etc.” (Boldface added for emphasis)

Today's confusing mix of product offerings and specifications is a long way from being “without qualitative differentiation,” but wider acceptance of a metric like the WAC would bring that possibility within reach. With the required computing capacity bought from the market, placing many computing jobs anywhere would be feasible.

Spot markets for buying and selling excess computing capacity have so far failed to live up to [the promise attributed to them by services such as SpotCloud](#). Amazon took a [significant step](#) in September by extending its existing concept of “[reserved instances](#)” (Amazon virtual machines, secured at a lower price than normal by entering into long-term contracts) to allow unwanted instances to be bought and sold in a new [marketplace](#). It is currently too soon to fully understand the ways in which buyers and



sellers will adapt to this new model. For example, large users could possibly offload unwanted capacity at a loss (recouping a few cents per unused machine is better than nothing), skewing the economics of Amazon's existing pricing regime in unanticipated ways.

Amazon's move is interesting and introduces a far larger customer base to the idea of buying and selling spare compute capacity. But it's still a closed system, concerned only with trading machines that exist inside Amazon's data centers. The larger opportunity is to extend similar models so they encompass other providers too.

Unfortunately, differentiation currently suits most cloud infrastructure providers better than easy comparison. A badly tuned metric reduces differently nuanced product offerings to simple numbers, and those simple numbers can then be translated into a bald statement of price (dollar per unit of compute) across an entire market. If customers respond to that by favoring the cheapest providers of the compute resources they require, a complex competitive landscape is in danger of becoming a race to the bottom that squeezes profit — and innovation — from the market.

6fusion's existing iNode Network would suggest that price is not the only motivator. Amazon's new marketplace has the potential to reinforce that message and to reassure cloud-infrastructure providers that the opportunities created by offering their services for trade on an open market far outweigh the risks of a price squeeze.



Key takeaways

While the number of cloud-infrastructure providers is growing, they all describe their offerings using different terminology and different metrics.

There is latent demand for a means to realistically compare and contrast the costs and capabilities of different computing infrastructures, including on-premise data centers and various cloud-infrastructure providers.

A common set of metrics would enable a comparison of competing cloud offerings.

Once cloud offerings can be compared, customers will be able to select the best service for specific tasks.

As a market for generic compute resources emerges, opportunities arise for excess capacity to be bought and sold as a commodity.





About Paul Miller

Paul Miller works at the interface between the worlds of cloud computing and the semantic web, providing the insights that enable you to exploit the next wave as we approach the World Wide Database. At GigaOM Pro, he was the curator for the cloud channel during 2011. He routinely serves as a moderator for GigaOM Pro webinars.

About GigaOM Pro

GigaOM Pro gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports, and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, GigaOM Pro addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: pro.gigaom.com

