The Real Cost of a CPU Hour

Edward Walker, *University of Texas at Austin*

IT organizations can now outsource computer hardware by leasing CPU time through cloud computing services. A proposed modeling tool can quantitatively compare the cost of leasing CPU time from these online services to that of purchasing and using a server cluster of equivalent capability.

The acquisition and installation of IT server clusters is growing significantly worldwide.[1] Partly fueling this growth is society's reliance on the Internet for collaboration, entertainment, and communication; the adoption by IT organizations of the concept of software as a service; and the use of simulation in scientific discovery and product development. At the same time, commercial companies like Amazon.com, IBM, and Google are now letting any organization, indeed anyone with a credit card, purchase time on servers in their data centers through online Web services.

With these "cloud computing" services, IT organizations can lease only the compute time they require for their computational needs, rather than purchase a server cluster. Since August 2006, for example, Amazon.com has offered the Elastic Compute Cloud (http://aws.amazon.com/ec2) Web service. For $0.10 per CPU hour, organizations can purchase virtual machine instances with root access, running on a compute node in an Amazon.com data center.

What does $0.10 per CPU hour mean? Is this cost fair? How does this cost compare to that of purchasing a server cluster? Providing quantifiable answers to these questions is of intrinsic value to computational scientists, IT organizations, and equipment-granting agencies.

To allow consumers to make rational choices, a tool is needed to quantitatively compare the cost of purchasing a cluster versus the cost of leasing CPU time from an open competitive market. My proposed model computes the real cost of a CPU hour when performance depreciation (the time value of a CPU) is taken into account. Deriving this cost enables a quantitative comparison of the investment choices of leasing online compute time versus purchasing a server cluster.

Understanding this tradeoff through modeling lets IT organizations justify purchase decisions quantitatively. In addition, funding agencies can objectively evaluate alternative equipment proposals, and policymakers can fashion guidelines that ensure the most efficient outcome for all. Most importantly, if consumers understand the real cost of a CPU hour, an efficient market will eventually reflect this real cost, enabling a fair, competitive market for the benefit of all consumers.

CPU LEASING IMPACT

Acquiring assets is a central challenge in any organization, and deciding whether to lease or buy equipment isn't a new dilemma in economic finance. In fact, more than four decades of research describes models for rationally making equipment lease-or-buy decisions.[2-4] An often used decision model calculates the *net present value* (NPV) of

0018-9162/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society

CALCULATING A PURCHASED SERVER CLUSTER'S NET PRESENT VALUE

Robert Johnson and Wilbur Lewellen's seminal paper[1] first suggested modeling the lease-or-buy decision as a capital budgeting problem. They recast the solution as a comparison of the net present value of potential cash flows from two alternative investment strategies—leasing versus buying. In analyzing the buy investment strategy in particular, the authors presented the following formula that required taking taxation's impact:

$$NPV = \sum_{T=1}^{Y} \frac{(P_T - L_T) - t(P_T - L_T - D_T)}{(1+k)^T} + \frac{S - t_g(S-B)}{(1+k)^T} - A,$$

where P_T = cash revenue expected from the use of the asset at year T; L_T = pretax cash cost required to operate the asset at year T; D_T = depreciation charge for year T; k = cost of capital; A = cash purchase price of the asset; S = expected salvage value of the asset at the end of its life; B = expected book value of the asset at the end of useful life; t = corporate income tax; and t_g = tax rate for gain or loss on disposal of the asset.

The formula comprises three terms. The first term approximates the net after-tax operating profit. The second term approximates the after-tax proceeds from asset salvage after its retirement. Finally, the third term incurs the asset's initial purchase cost. This formulation (and its many simplified forms) is now generally regarded as the standard method for calculating the NPV from a purchased asset when facing a lease-or-buy decision.[2]

The analysis I present uses this formulation for calculating the NPV from a purchased asset. However, I also make two assumptions in my calculations throughout the article. First, I assume the server cluster has no salvage value after its retirement because only a small market exists for used CPU equipment. Second, I assume the server cluster's operational life has no expected cash revenue (or profit). I also apply this assumption uniformly to calculate the NPV from the leasing investment. I make this second assumption to accommodate scenarios in which the cash revenue is difficult to estimate, such as with a server cluster for research and education.

The net effect from these assumptions is that with the first assumption, I never apply the second term in the NPV calculation. With the second assumption, I always compute the NPV of the operating loss, so I never apply the taxation factor in the first term. Because the taxation term doesn't apply, I can't include the depreciation factor because it's simply an accounting tax shield; no profit means no applicable tax. Specifically, Equation 2 in the main text represents the simplified version of the NPV formula (with $P_T = 0$):

$$NPV(\text{purchase}) = \sum_{T=1}^{Y} \left(\frac{(-L_T)}{(1+k)^T} - A\right)$$

References

1. R.W. Johnson and W.G. Lewellen, "Analysis of Lease-or-Buy Decision," ***J. Finance***, vol. 27, no. 4, 1972, pp. 815-823.
2. G.B. Harwood and R.H. Hermanson, "Lease-or-Buy Decisions," ***J. Accountancy***, vol. 142, no. 3, 1976, pp. 83-87.

cash flow over time, taking into account the *time value of money*,[5] when buying or leasing equipment for a fixed duration. The investment strategy that results in the higher NPV is the rational choice. However, current NPV models also make two implicit assumptions: The lease terms are relatively long, allowing few opportunities to reevaluate competing products, and the acquired equipment in both the lease and buy cases has similar capabilities for the duration of its use.

NPV models don't account for a CPU nonmonetary performance depreciation. Consequently, there are few systematic methods for organizations to make intelligent decisions on leasing or buying CPUs for server clusters in the presence of online cloud services. When, and if, current models account for equipment depreciation, they model it as a *tax shield*, requiring a profit to realize. The "Calculating a Purchased Server Cluster's Net Present Value" sidebar explains this in more detail. This is difficult for many organizations to model, especially if the profit from operating a server cluster is difficult to quantify, such as a server cluster dedicated for research and education.

Figure 1 shows the relative difference in term lengths and depreciation rates when leasing assets for long-term acquisition. For example, in purchasing or leasing a car for long-term use, the traditional lease-or-buy models assume that the lease term is significant (spanning many months) and the performance benefit from re-leasing a new car every few months is negligible. These assumptions preclude the desire to frequently reevaluate the term arrangement to adopt the latest technology, which is really unnecessary in a slowly changing technology landscape like the auto industry. However, the situation is significantly different in the CPU lease market that cloud services offer, in which short-term leases are available, allowing for greater consumer mobility to pick the fastest CPU service.

A SUSTAINABLE IT DEVELOPMENT

Numerous factors suggest that the emergence of commercial online IT leasing services is a sustainable development in the IT industry. Some of these factors include the broad applicability of the market to a large portion of the IT community, the efficiencies it introduces in obviating the need to purchase and maintain small-to-medium size server clusters, and the competitive advantage it provides to businesses intent on avoiding rapid technology obsolescence.

Figure 2 shows the total number of volume servers installed in the US from 2000 to 2006[6] and the proportion installed in small-to-medium size clusters (SMCs). Borrowing from the IDC site infrastructure category definition,[7] I define an SMC as belonging to the "server closet," "server room," or "localized data center" categories, covering clus-

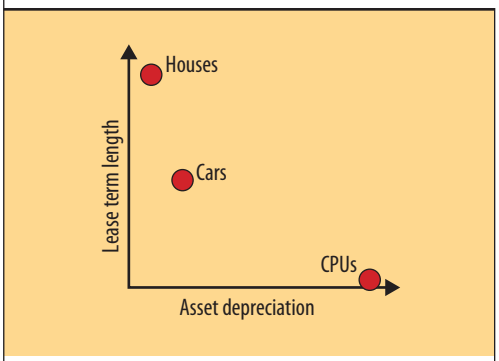

Figure 1. Lease terms versus depreciation. CPU performance depreciates much more rapidly than that of other types of assets, such as houses and cars, discouraging long-term lease arrangements.

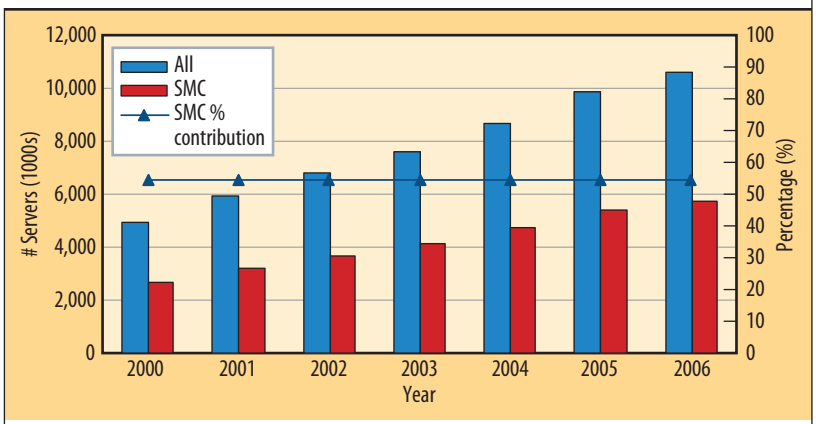

Figure 2. Small-to-medium size clusters. SMCs constitute a substantial portion of the total number of servers installed in the US.

ters of tens to hundreds of servers that occupy less than 1,000 sq. ft. of machine floor space.

From the data, it's apparent that a significant number of volume servers are installed in SMCs, and their contribution to the overall server installation base is increasing proportionally with its growth. These SMCs are expected to derive the most benefit from online IT leasing services, because these sites have few existing infrastructure investments (power distribution equipment, special cooling units, data center buildings, and so on). Hence, there's little entrenched resistance to converting their operations to a lease-based economy. Furthermore, SMCs have moderate computational workloads that should be offloadable to the shared clusters an online CPU leasing service offers.

Also, SMCs that exhibit poor long-term resource use would benefit from an online leasing service. These underutilized server clusters typically support a small group of users with only periodic computational demands. Thus, unlike large systems maintained for diverse communities, these smaller clusters exhibit periods of peak use interspersed with periods of little use. During low-use periods, the server cluster continues to consume space, power, and system administration effort.

Finally, organizations that use short-term compute time leases can benefit from constantly evaluating the online market and choosing the most up-to-date technology. This can give an organization a tremendous advantage over competitors tied to aging server clusters. Benefits of using a faster CPU service include faster product design, increased transaction volumes, and richer software interfaces.

THE TIME VALUE OF A CPU

An important foundational principle in finance is the time value of money.[5] This principle basically states that an investor always prefers to receive some fixed amount of money today rather than in the future. Hence, a popular technique for making lease-or-buy decisions involves comparing investment cash flows at their present value by discounting future cash consumption with a rate of interest. This interest is usually the cost of capital—that is, the approximate associated risk of raising the required capital for the investment.[8] Equation 1 shows the present value (PV) calculation for a future value (FV) in year T, with k representing the cost of capital.

$$PV = \frac{FV}{(1+k)^T} \tag{1}$$

Using Equation 1, the NPV of an investment with annual amortized cash flow (profit – cost) C_T for Y years is calculated as follows:

$$NPV = \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T} \tag{2}$$

In my model, I propose that the time value of a CPU also factors into the NPV calculation to accurately evaluate the lease-or-buy decision for server clusters in the presence of short-term CPU leases. As is widely known in the industry, Moore's law describes an important time trend for IT equipment.[9] The law states that integrated circuit transistors are expected to double approximately every two years. Its generalization to the microprocessor industry is that CPU performance is also expected to double every two years. Indeed, not withstanding the recent development in multicore technology, the microprocessor's CPU performance has doubled every two years since its invention.

Now, if Moore's law still describes the domineering CPU depreciation trend, the future capacity (FC) of a T-year-old CPU can be discounted to its present capacity (PC) through the biennial halving of CPU performance as follows:

$$PC = \frac{FC}{(\sqrt{2})^T} \tag{3}$$

Furthermore, assuming the total useful capacity (TC) represents the expected CPU hours users consume

annually in the cluster (for example, a 512-CPU cluster with 40 percent utilization provides a TC of 512 × 365 × 24 × 0.4 CPU hours per year), using Equation 3, I can similarly calculate a cluster's net present capacity (NPC) over an operational life span of Y years as follows:

$$NPC = TC \times \sum_{T=0}^{Y-1} \left(\frac{1}{\sqrt{2}}\right)^T \Rightarrow NPC = TC \times \frac{1-\left(\frac{1}{\sqrt{2}}\right)^Y}{1-\frac{1}{\sqrt{2}}} \quad (4)$$

This assumes $TC = TCPU \times H \times \mu$, where $TCPU$ is the total CPU cores the server cluster provides, H is the expected number of operational hours the server provides annually, and μ is the expected server utilization.

Finally, I propose that the NPV for a CPU hour, or the real cost of a CPU hour, is calculated as follows:

$$R = NPV/NPC \quad (5)$$

Purchase case

Starting with Equation 5, and substituting with Equations 2 and 4, I can derive the real cost of a CPU hour, from purchasing a server cluster:

$$R\ (\text{purchase}) = \frac{\left(1-\frac{1}{\sqrt{2}}\right) \times \sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{\left(1-\left(\frac{1}{\sqrt{2}}\right)^Y\right) \times TC} \quad (6)$$

Lease case

Calculating the real cost of a CPU hour from leasing is also quite simple. I use Equation 5 and substitute with the NPV Equation 2. However, I also replace the NPC calculator with Equation 7, in which I assume no depreciation in the computational capacity because the lessee can always acquire the latest IT capacity from a competitive market over the operational life span of Y years.

$$NPC = Y \times TC \quad (7)$$

I next modify Equation 5 with this new NPC calculator to derive the real cost of a CPU hour from leasing:

$$R\ (\text{lease}) = \frac{\sum_{T=0}^{Y-1} \frac{C_T}{(1+k)^T}}{Y \times TC} \quad (8)$$

Purchase-upgrade case

A variation of the purchase case is also a viable investment option: purchasing a cluster and upgrading it annually with the newest CPU to avoid incurring the performance degradation cost. In this scenario, I assume the annualized operating cost includes repurchasing new CPUs, which I also assume to be approximately equivalent to the server cluster's original purchase price.[7] With these assumptions, I can modify the NPV calculator to Equation 9, where A is the server cluster's original purchase cost.

$$NPV = C_0 + \sum_{T=1}^{Y-1} \frac{C_T - A}{(1+k)^T} \quad (9)$$

Now, because I'm annually upgrading the purchased server cluster, the same CPU performance degradation of an aging cluster isn't a factor. Thus, similar to the leasing case, I can apply the NPC calculator in Equation 7. The real cost of a CPU hour from purchasing and annually upgrading the cluster therefore calculates as:

$$R\ (\text{purchase-upgrade}) = \frac{C_0 + \sum_{T=1}^{Y-1} \frac{C_T - A}{(1+k)^T}}{Y \times TC} \quad (10)$$

NSF TRACK 2 HPC CLUSTER

My first example examines the first high-performance computing system acquisition under the NSF Track 2 initiative. The NSF awarded the Texas Advanced Computing Center $59 million to purchase and operate Ranger, the Sun Constellation server cluster, for the computational science research community.[10] The system has more than 15,000 AMD Barcelona quad-core processors, providing more than 60,000 CPU cores for scientific computing. The grant breaks down approximately into a $30 million hardware acquisition cost in year 0 (C_0) and a $7 million annualized-operations cost (including power, cooling, and support personnel) for years 1 to 4 (C_T where T = [1,4]).

I compare the cost of purchasing this server cluster for the five years of its operational life span to the two alternative investment strategies of leasing the required capacity from the online market (assuming such a service exists) or upgrading the cluster annually to avoid loss of computational capacity.

In my calculations, I assume a cost of capital of 5 percent; an operational cluster life span of five years; and a cluster with the computational capacity of 60,000 CPU cores. I also assume the cluster is unavailable for at least one day a week annually (required for preventive maintenance), with 99 percent operational reliability and 100 percent CPU utilization. Thus, $TCPU$ = 60,000 CPUs and $TC = TCPU \times H \times \mu$ = 60,000 × ((365 – 52) × 24) × (0.99 × 1.0) = 440 million CPU hours annually.

I can then calculate the real cost of a CPU in purchasing the Ranger system with Equation 6; in leasing the equivalent compute time at $0.10 per CPU hour with Equation 8; and in purchasing and annually upgrading an equivalent cluster with Equation 10.

R (purchase) = \$0.045 cents per CPU hour

R (lease) = \$0.092 cents per CPU hour

R (purchase-upgrade) = \$0.07 cents per CPU hour

Thus, the Ranger system proves a good investment compared with investment strategies involving either leasing the computational time from a hypothetical online service at \$0.10 per CPU hour or purchasing an equivalent server cluster and upgrading it annually. In fact, the benefit becomes even clearer when I plot a CPU hour's real cost over a range of operation life spans for Ranger, as Figure 3a shows. Here, it's clear that the cluster must be operational for at least 16 years with the lease option or 12 years with the purchase-and-upgrade option before these alternative investment strategies become attractive.

Figure 3a further shows that a three-year investment commitment is the optimal term length for the purchase case because it results in the lowest cost per CPU hour for NSF. Finally, even compared to the investment strategy that upgrades the CPU hardware annually, the leasing strategy is still the inferior investment option.

A COMPUTE BLADE RACK CLUSTER

In my second example, I calculate the real cost of a CPU hour in purchasing a fully populated 44 1U compute blade rack. Such a server cluster fits in a server room and uses the HVAC system to handle cooling.

A typical 1U compute blade (circa 2008) cost about \$2,000 for a base model with two dual-core processors. Thus, for a 44-blade rack, the estimated equipment cost is \$2,000 × 44 = \$88,000 (I don't calculate the additional cost of networking and disk storage for now). Currently, a fully populated blade rack requires up to 20-25 kW (assume 20 kW) of power to operate (P_1), with an additional 20-25 kW (assume 20 kW) required for cooling and power conversion (P_2).[6] So, assuming an electric utility cost of υ and server utilization of μ, the estimated annualized recurring cost for the purchased cluster is $\upsilon \times (P_1 + P_2) \times (365 \times 24) \times \mu = \$350{,}400\upsilon\mu$. This translates into

$C_0 = \$88{,}000 + \$350{,}400\upsilon\mu$

$C_T = \$350{,}400\upsilon\mu, \forall\ T > 1$

$TCPU = 44 \times 4 = 176$ CPUs

$TC = TCPU \times H \times \mu$

$= 176 \times (365 \times 24) \times \mu$

$= 1{,}541{,}760\mu$ CPU hours annually

Assuming an electric utility cost of υ = \$0.07 kWh and server utilization at μ = 90 percent, I can now calculate the real cost of a CPU hour in purchasing the server cluster with Equation 6; in leasing the required CPU time at a lease fee of \$0.10 per CPU hour with Equation 8; and in purchasing a server cluster and upgrading it annually

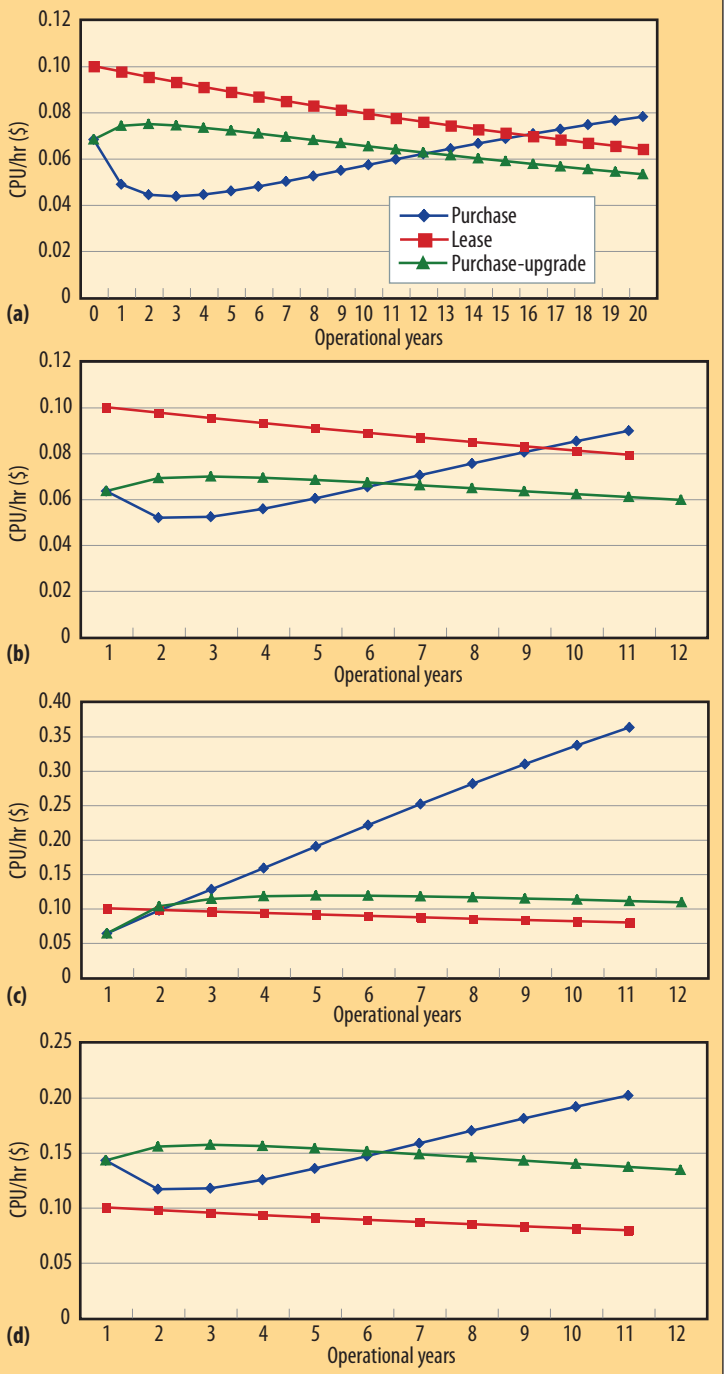

Figure 3. Comparing investment alternatives. (a) The NSF Track 2 Ranger high-performance computing system. (b) High utilization (μ = 90 percent) and low utility cost (υ = \$0.07 kilowatt hours). (c) High utilization (μ = 90 percent) and high utility cost (υ = \$0.4 kWh). (d) Low utilization (μ = 40 percent) and low utility cost (υ = \$0.07 kWh).

with Equation 10. Figure 3b shows the calculated values for increasing anticipated operational life spans. The purchase case is indeed cheaper than the lease case if the cluster's expected operation life is within 10 years. After that, it becomes cheaper to lease the required time at a rate of $0.10 per CPU hour.

Other hypothetical scenarios, such as the capital investment alternatives if the recurring electric utility cost increases to $0.40 kWh, produce different results. Figure 3c shows the real costs for the investment scenarios for increasing operational life spans. If the cluster's expected operational life span is longer than two years, it becomes cheaper to lease CPUs from an online service. More interestingly, the purchase-upgrade investment scenario also becomes an attractive option compared with the purchase case at all of the cluster's expected life spans and is only marginally inferior to the lease case.

Finally, I examine the investment alternatives if utilization is expected to be low. Figure 3d shows the real costs of the investment scenarios for increasing operational life spans when the expected utilization is 40 percent. In this situation, leasing is always the preferred option. For all anticipated operational life spans, the lease option incurs the least real cost per CPU hour for the organization.

LEASE PRICE VOLATILITY

Throughout my exposition, I haven't considered price volatility in the online CPU market. My model assumes that the lease price is stable throughout the cluster's anticipated life span. Thus, it's deficient in the face of heavy CPU price volatility. In such a case, I can still employ my approach by using option valuation models[11] to price an anticipated future CPU lease price. The fee used in calculating a lease case's NPV incorporates the expected lease price, plus the risk-neutral premium associated with an option to purchase the lease at that anticipated strike price (also commonly referred to as a call option).

This risk-neutral premium essentially adds a cost to the risk of making an assumption about the expected lease price in my analysis for a volatile market. However, the nascent CPU lease market doesn't yet exhibit this price volatility, so I'll leave further analysis of this extension to a future time.

A modeling tool can help organizations make better-informed lease-or-buy decisions for server clusters. However, the model described here is accurate only if two essential conditions exist. First, the online leasing market must be open, thriving, and competitive. For this to happen, barriers of entry into the market for lessors must be low. Moreover, plenty of potential investment capital must be available to incentivize lessors to set up the required enterprise-level data centers to support these services.

In a competitive market, lessees will have a variety of choices and the opportunity to switch between online services to exploit the fastest computation services. Equipment-granting agencies can therefore help stimulate this market by providing vouchers for acquisition grants instead of requiring traditional purchasing proposals. This approach would let consumers rationally choose the type and form of asset acquisition strategy most appropriate over the term of the project.

Second, lessees must not experience (suffer from) any kind of technology "lock-in." For example, the lessor might require applications to be developed in a closed-source proprietary runtime environment, prohibiting user-developed applications from running on other leasing services. This restriction could result in lessees becoming reluctant to move their applications to a new leasing service to avoid incurring a switching cost. However, in a rational economic universe, in which an open, competitive leasing market exists, lessees will in aggregate choose vendors who don't require this technology lock-in.

It's important to note that these two conditions don't yet exist. However, to support their development, the online leasing market must become more transparent in terms of its value. If IT organizations, equipment-granting agencies, and computational scientists understand the CPU hour's real cost, rational choices can be made, fair prices will prevail, and economies of scale will result. The modeling tool introduced in this article promotes this market transparency. C

Acknowledgments

This article is based on work supported by National Science Foundation grants 0721931 and 0503697.

References

1. L. Cohen, S.L. Josselyn, and H. Nguyen, *Worldwide Server Installed Base 2007-2011 Forecast*, Int'l Data Group, doc. no. 207044, May 2007.
2. R.F. Vancil, "Lease or Borrow: New Method of Analysis," *Harvard Business Rev.*, vol. 39, no. 5, 1961, pp. 122-136.
3. R.W. Johnson and W.G. Lewellen, "Analysis of Lease-or-Buy Decision," *J. Finance*, vol. 27, no. 4, 1972, pp. 815-823.
4. G.B. Harwood, and R.H. Hermanson, "Lease-or-Buy Decisions," *J. Accountancy*, vol. 142, no. 3, 1976, pp. 83-87.
5. E. von Bohm-Bawerk, *Capital and Interest: A Critical History of Economical Theory*, Macmillan and Co., 1890.
6. US Environmental Protection Agency, Energy Star Program, *Report to Congress on Server and*

Data Center Energy Efficiency: Public Law 109-431, Aug. 2007; www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf.

7. M. Bailey et al., *The Data Center of the Future*, Int'l Data Group, doc. no. 06C4799, Apr. 2007.
8. W.G. Lewellen, *The Cost of Capital*, Wadsworth Publishing, 1970.
9. G. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, 1965; http://download.intel.com/research/silicon/moores-paper.pdf.
10. Nat'l Science Foundation, "National Science Foundation Awards Texas Advanced Computing Center $59 Million for High-Performance Computing," press release 06-137; www.nsf.gov/news/news_summ.jsp?cntn_id=108051&org=NSF&from=news.
11. J. Cox, S. Ross, and M. Rubinstein, "Options Pricing: A Simplified Approach," *J. Financial Economics*, vol. 7, 1979, pp. 229-263.

Edward Walker *is a research scientist with the Texas Advanced Computing Center at the University of Texas at Austin. His research interests include fault-tolerant distributed systems, parallel programming languages, and operating systems. Walker received a PhD in computer science from the University of York, UK. He is a member of the IEEE and the ACM. Contact him at ewalker@tacc.utexas.edu.*