The Nuances of Cloud Economics

ECONOMICS IS CENTRAL TO CLOUD COMPUTING. The cloud's financial and strategic benefits have been the catalysts for its explosive growth, and pay-per-use pricing, sometimes referred to as measured service,[1] is a core attribute. Some say that cost reduction and business agility are the two most important benefits of the cloud,[2] some argue that economies of scale from large providers are the main drivers of cloud benefits,[3] and others therefore conclude that eventually all IT should and will move to the cloud.[4] However, the theory and practice of cloud economics are considerably more nuanced, and encompass numerous challenges, ranging from the practical to the theoretical, across service architecture, statistics, behavioral economics, computing foundations, game theory, business strategy, and regulatory policy.

Private, Public, or Hybrid

Some consider "private cloud" to be a misnomer. However, many of the key criteria of cloud still apply, such as dynamic allocation of resources from a common pool and pay-per-use pricing through either a commercial transaction or chargeback to an internal customer. A basic question facing most IT shops today is whether to use their own datacenters, a public cloud provider, colocation facilities, or all of the above.

Organizations must consider many quantitative and qualitative criteria when making this decision, such as focusing leadership time on "core vs. context" issues—that is, those that are critical to developing competitive advantage versus those that aren't.[5] For example, a movie studio should focus on scripting, cinematography, and casting, not datacenter technology and operations. From a rational economic cost-optimization perspective, however, there are several key drivers.

The first driver is the loaded unit cost of a company's IT relative to the offered unit price of the cloud service provider. The *cost structure* for public cloud service providers might be better, but additional components can increase the offered *price*, such as a provider's profit, underutilized resources, taxes, and sales, general, and administrative expenses. If a company's IT shop is not cost-optimized or can't achieve scale and high utilization, the cloud might well offer a cost advantage, but for well-run organizations, the unit costs for public cloud services might actually be higher.

Moreover, there may be a performance penalty from running on shared resources and/or a general-purpose infrastructure, as opposed to one tuned to a specific application. Sometimes this will favor the enterprise, but it can also favor the cloud, for example, through shared access to quantum computing as a service.[6] A slightly lower unit cost might not create a net benefit if substantially more resource units are required.

Making all these calculations even more difficult is the fact that prices and technology aren't constant. Enterprises engage in cost-control initiatives; providers regularly reduce prices. Some argue that Moore's Law means that cloud providers have advantages,[7] but of course, enterprise computing infrastructure exploits the same "law." Faster technology refresh rates enhance price/performance, but require higher capital expenditures. Furthermore, some ser-

JOE WEINMAN

joeweinman@gmail.com

2325-6095/14/$31.00 © 2014 IEEE

vice providers have introduced dynamic pricing—for example, Amazon Web Services (AWS) spot instances[8]—or pricing based on usage patterns—for example, Google's "sustained use" pricing.[9]

However, even if the cloud's *unit* cost is higher, pay-per-use pricing can deliver a lower *total* cost in the presence of variable or intermittent demand. For relatively flat workloads, the cloud's on-demand, pay-per-use nature might not always offer benefits. But for "spiky" workloads—say, where demand peaks exist for online tax preparation companies on April 15 or for retailers on Black Friday or Cyber Monday—the cloud might offer total cost savings even if the unit price is higher.

There are a variety of other factors to consider, such as wasted instances or data transfer costs, but generally, all other things being equal, in the presence of variable demand, even if the public cloud has a higher performance-adjusted unit price, a hybrid architecture that includes a public cloud will be cost optimal.

Statistical Multiplexing Effects

To understand relative costs, you also can't just look at unit costs, but must also consider utilization. Operating, say, at 33 percent utilization means that there are two unused resources for every one that is used. The effective unit cost then triples, not unlike buying two additional peaches at the fruit stand for every one that you actually eat. In the case of the public cloud, this is factored into the offered price. In the case of dedicated resources, low utilization can be caused by poor resource management, but it can also be an inevitable result of spiky workloads in the presence of fixed capacity.

Both public and private clouds create utilization benefits through workload aggregation. When workloads are statistically independent, multiplexing them smooths aggregate demand: the troughs and peaks tend to cancel each other out. As a result of this smoothing, both private and public clouds can achieve better resource utilization than if the workloads were individually run on siloed resources. The impact of this effect is to reduce the effective cost of each resource. Interestingly, such utilization benefits are exhibited with the first additional workload. The implication: although more statistically multiplexed independent workloads always reduce penalties associated with under- or overcapacity, economic benefits can be achieved even by private clouds and small providers.[10]

Even if a public or hybrid cloud is theoretically cost-optimal, there are additional relevant costs to factor in.

Migration, Integration, and Management Costs

Even if a public or hybrid cloud is theoretically cost-optimal, there are additional relevant costs to factor in. There are "search" costs to find the right provider, although intermediaries such as brokers and markets can help. There might be costs for migrating applications to a cloud architecture. There might be technology costs—such as for network infrastructure, ongoing data transfer, or management tools—and people costs—such as for an organization to audit service provider Service Level Agreement (SLA) conformance. Then there might be soft costs, such as loss of productivity over potential job security concerns.[11]

There are additional migration and integration scenarios: migration from cloud back to datacenter, from cloud to cloud, and applications or infrastructure spanning two or more clouds. Providers would like to discourage customers from migration, but customers would like flexibility to avoid lock-in and help keep the market competitive.

On-Demand Provisioning

Having the right quantity of resources to match aggregate demand has clear economic benefits. Too many resources, and there is a loss commensurate with the opportunity cost of the capital deployed or excess expense. Too few, and the application will perform slowly or not at all, impacting key metrics such as revenue (for customer-facing applications), labor productivity (for employee ones), or time to market—for example, for cloud-based collaboration among partners. On-demand resources ensure the right quantity at the right time.

However, the value of on-demand resources varies based on the nature of the demand. If demand is completely flat, there is no real value to speedy provisioning except through second-order effects such, say, the option value

associated with the possibility that demand may increase in the future. However, if demand is growing exponentially, there is an exponential penalty associated with any fixed provisioning interval. Consequently, in such growth situations, which are typical for start-ups, a public cloud strategy makes sense because otherwise, increasing levels of revenue-generating customer demand will not be served, leading to economic loss.[12]

> The cloud can be part of a strategy to achieve competitive advantage.

User Experience Enhancement through Latency Reduction

Clouds—typically for infrastructure services and always for content delivery networks—are inherently geographically dispersed. Latency reduction for interactive workloads has business benefits: faster load times imply that more customers can visit more webpages and make more purchases or that employees can be more productive. However, additional service node build-outs have diminishing returns: halving latency essentially requires quadrupling the number of nodes. In early stages of geographic expansion, adding a few nodes can reduce worst-case roundtrip network latency from, say, 160 milliseconds to 5 or 10 milliseconds. At some point, though, reducing latency by a few microseconds would require billions of additional nodes. The tradeoffs between latency and investment strongly favor public providers, because dedicated distributed resources that might be substantially underutilized would be prohibitively expensive.[13]

Business Strategy

The cloud can be part of a strategy to achieve competitive advantage through better information-enabled processes, cloud-connected solutions, cloud-mediated customer relationships, or cloud-enabled accelerated innovation.[14] Although the benefits of cloud computing are often said to be cost reduction and business agility, there are many others, including revenue growth, business model innovation, and enhanced customer experience and relationships.

The Intercloud

Multiple independent networks eventually gave way to the Internet; multiple clouds might well become the Intercloud.[15] Cloud federation and interoperability will emerge at the infrastructure, platform, and (software) application layers. Because clouds are inherently commercial entities, means of advertising resource availability, enabling cloud markets, trading resources, bidding, and so forth will be essential. IEEE is currently working on vendor-independent standards for the Intercloud through the P2302 standard and the Intercloud Testbed initiative.[16] As an example of one interesting area, evolving conversational protocols and resource description frameworks to work with an ontology for pricing will be essential to enable seamless cloud-to-cloud interoperability and third-party intermediaries that can, for example, convert flat-rate to pay-per-use pricing or vice versa.

Human Behavior

Classical economics, in which individuals behave rationally and optimally, has been complemented, or perhaps even supplanted, by behavioral economics, in which humans often behave irrationally and emotionally.[17] Cloud computing and enterprise IT aren't immune from such cognitive biases. Everything from "free-to-play" games, which have in-app purchases, to "free tiers" of use provide real-world examples of the application of such behavioral economic insights to business and cloud strategy.

Computational Complexity

An important part of economics is the efficient functioning of markets, which match supply and demand. A traditional example is where price relates to the equilibrium between supply and demand. More recently, the Nobel Prize was awarded to Alvin Roth and Lloyd Shapley for developing efficient algorithms for "matching" supply with demand, even without pricing. Examples include kidney donors and recipients and marriage-minded singles.

However, efficient algorithms might not always exist to match the supply of cloud computing resources with workload demand. Computational complexity theory addresses the difficulty of solving certain types of problems. The Traveling Salesman Problem is probably one of the most famous. In this problem, a salesman must determine a "best" order in which to visit n cities to minimize, say, the distance traveled. Although this is easy for a few cities,

the number of possible city sequences grows quickly as cities are added, making the problem essentially intractable at only a few dozen cities.

Now consider a simple example in which there are a number of cloud computing facilities, and a number of customers with different-sized workloads that can be run in some, but not all of the facilities. It turns out that this simple "satisfiability" problem—that is, whether there are sufficient resources in the right locations to satisfy the demand—is NP-complete, and is therefore intractable given today's technology.[18] Related problems, such as whether a given number of cloud computing datacenters can "cover" a user population within a given latency constraint, are also computationally intractable.[19] The implication: optimal solutions for cloud computing can be theoretically unattainable in a reasonable time (although very good ones can be found in practice).

Game Theory and System Dynamics

Game theory is relevant to cloud computing as well. One example is a market with two competitors that offer different pricing plans—say, one with flat rate (for example, "unlimited" or "infinite" storage) and one with a pay-per-use model. The system dynamics of such a market suggest that light users will prefer to pay on a per-use basis, whereas heavy users will prefer flat-rate plans. As more heavy users migrate to flat-rate plans, the average consumption level increases, driving the flat-rate price higher. More users will then defect to the pay-per-use plan, creating a virtuous cycle in which the pay-per-use plan dominates the flat-rate one.[20] Theory can inform practice: Bitcasa, a cloud storage company, originally offered flat-rate "unlimited" storage for $99 per year; but they were forced to raise their price to $999 per year,[21] and ultimately discontinued the plan.[22]

Another application of game theory to cloud computing is in deducing the system dynamics and game theoretic considerations of interoperability. According to one analysis,[23] two smaller competitors, by supporting interoperability standards, can tip a competitive market away from a dominant player, by offering a differential value proposition of "no lock in." A noninteroperable provider—even a dominant player—has to then choose between offering greater cloud interoperability and abdicating the market.

Axiomatic Formulations

Although many of the conclusions presented here can be derived independently, a rich theoretical framework can provide a basis for everything from an analysis of economic tradeoffs to new computing models. Traditional models of computing include the Turing machine, in which a finite state automaton reads and writes an infinite tape comprising a finite symbol alphabet, and Petri nets, which are models for parallel processing. The Turing machine model, however, never anticipated the additional complexities of distributed computing—issues such as eventual consistency, partitioning, and latency. In addition, the Turing machine had no notion of the commercial elements inherent in the cloud, such as pay-per-use pricing.

However, an axiomatic formulation of cloud computing has been created,[24] which can formally specify notions such as geographic dispersion through a metric space, comprising points and a distance metric, and a model of commercial relationships, in which a price for resource allocation over time can be virtually any mathematical function in a function space—say, a constant for a lifetime subscription, a constant per time unit for flat rate, a constant times the integral of time and quantity for pay per use, and so forth.

Such models can be shown to be equivalent to the fundamental model of computation, the Turing machine, while also helping to explore the complexities of commercial models and distributed parallel processing.

IN SHORT, CLOUD COMPUTING IS NOT JUST AN INTERESTING, POTENTIALLY TRANSFORMATIONAL TECHNOLOGY OR OPERATIONS MODEL. It is a rich subject for exploring statistics, system dynamics, economics, behavioral economics, the foundations of mathematics and computing, complexity, regulatory policy, and business strategy, all in one.

References

1. P. Mell and T. Grance, *The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology*, NIST special publication 800-145, Sept. 2011; http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.
2. G. Garrison, S. Kim, and R. Wakefield, "Success Factors for Deploying Cloud Computing," *Comm. ACM*, vol. 55, no. 9, 2012, pp. 62–68.
3. D. Sholler and D. Scott, "Economies of Scale Are the Key to Cloud Computing Benefits," report G00158192, Gartner, 30 June 2008; https://www.gartner.com/doc/710610/economies-scale-key-cloud-computing.
4. N. Carr, *The Big Switch: Rewiring the World, From Edison to Google*, W.W. Norton, 2008.
5. J. Weinman, *Cloudonomics: The Business Value of Cloud Computing*, John Wiley & Sons, 2012.
6. R. Cortland, "D-Wave Aims to Bring

Quantum Computing to the Cloud," *IEEE Spectrum*, podcast, 9 Apr. 2014; http://spectrum.ieee.org/podcast/computing/hardware/dwave-aims-to-bring-quantum-computing-to-the-cloud.
7. G. O'Connor, "Moore's Law Gives Way to Bezos's Law," GigaOm, 19 Apr. 2014; https://gigaom.com/2014/04/19/moores-law-gives-way-to-bezoss-law.
8. S. Higginbotham, "Dynamic Pricing Comes to Amazon's Cloud," GigaOm, 14 Dec. 2009; https://gigaom.com/2009/12/14/dynamic-pricing-comes-to-amazons-cloud.
9. N. Joneja, "Introducing Sustained Use Discounts—Automatically Pay Less for Sustained Workloads on Compute Engine," Google Cloud Platform Blog, blog, 4 Apr. 2014; http://googlecloudplatform.blogspot.com/2014/04/introducing-sustained-use-discounts.html.
10. J. Weinman, "Smooth Operator: The Value of Demand Aggregation," working paper, 27 Feb. 2011; http://joeweinman.com/Resources/Joe_Weinman_Smooth_Operator_Demand_Aggregation.pdf.
11. O. Rana, "The Costs of Cloud Migration," *IEEE Cloud Computing*, vol. 1, no. 1, 2014. pp. 62–65.
12. J. Weinman, "Time Is Money: The Value of 'On-Demand,'" working paper, 7 Jan. 2011; http://joeweinman.com/Resources/Joe_Weinman_Time_Is_Money.pdf.
13. J. Weinman, "As Time Goes By: The Law of Cloud Response Time," working paper, 12 Apr. 2011; http://joeweinman.com/Resources/Joe_Weinman_As_Time_Goes_By.pdf.
14. J. Weinman, "Strategies for Thriving in the Networked Economy of Things," blog, 29 May 2014; http://blogs.sap.com/innovation/innovation/strategies-thriving-networked-economy-things-01252574.
15. D. Bernstein et al., "Blueprint for the Intercloud—Protocols and Formats for Cloud Computing Interoperability," *Proc. 4th Int'l Conf. Internet and Web Applications and Services*, 2009, pp. 328–336.
16. *IEEE Standard P2302—Standard for Intercloud Interoperability and Federation (SIIF)*, IEEE, http://standards.ieee.org/develop/project/2302.html.
17. D. Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions*, Deckle Edge, 2009.
18. J. Weinman, "Cloud Computing Is NP-Complete," working paper, 21 Feb. 2011; http://joeweinman.com/Resources/Joe_Weinman_Cloud_Computing_Is_NP-Complete.pdf.
19. N. Megiddo and K.J. Supowit, "On the Complexity of Some Common Geometric Location Problems," *SIAM J. Computing*, vol. 13, no. 1, 1984, pp. 182–196; http://theory.stanford.edu/~megiddo/pdf/complexity of common geometric location problems.pdf.
20. J. Weinman, "The Market for 'Melons': Quantity Uncertainty and the Market Mechanism," working paper, 6 Sept. 2010; http://joeweinman.com/Resources/Joe_Weinman_The_Market_For_Melons.pdf.
21. A. Santos, "Bitcasa's Infinite Cloud Storage Balloons to $999 a Year," Engadget, 19 Nov. 2013; www.engadget.com/2013/11/19/bitcasa-infinite-cloud-storage-price-change.
22. N. Lomas, "Bitcasa Ends Unlimited Storage," Techcrunch, 24 Oct. 2014; http://techcrunch.com/2014/10/24/bitcasa-no-unlimited.
23. O. Rogers, "Game of Clones: Should Cloud Providers Interoperate?" 451 Research, 8 Jan. 2014; https://451research.com/report-short?entityId=79812.
24. J. Weinman, "Axiomatic Cloud Theory," working paper, 29 July 2011; http://joeweinman.com/Resources/Joe_Weinman_Axiomatic_Cloud_Theory.pdf.

JOE WEINMAN *is the chair of the IEEE Intercloud Testbed executive committee and an analyst for GigaOm Research. He also serves on the advisory boards of several technology companies. He has been awarded 21 patents in areas such as homomorphic encryption, pseudoternary line coding, adaptive bandwidth schemes, Web search, and distributed storage and computing. and is the author of* Cloudonomics. *Weinman has BS and MS degrees in computer science from Cornell University and the University of Wisconsin-Madison, respectively, and has completed executive education at the International Institute for Management Development in Lausanne.*