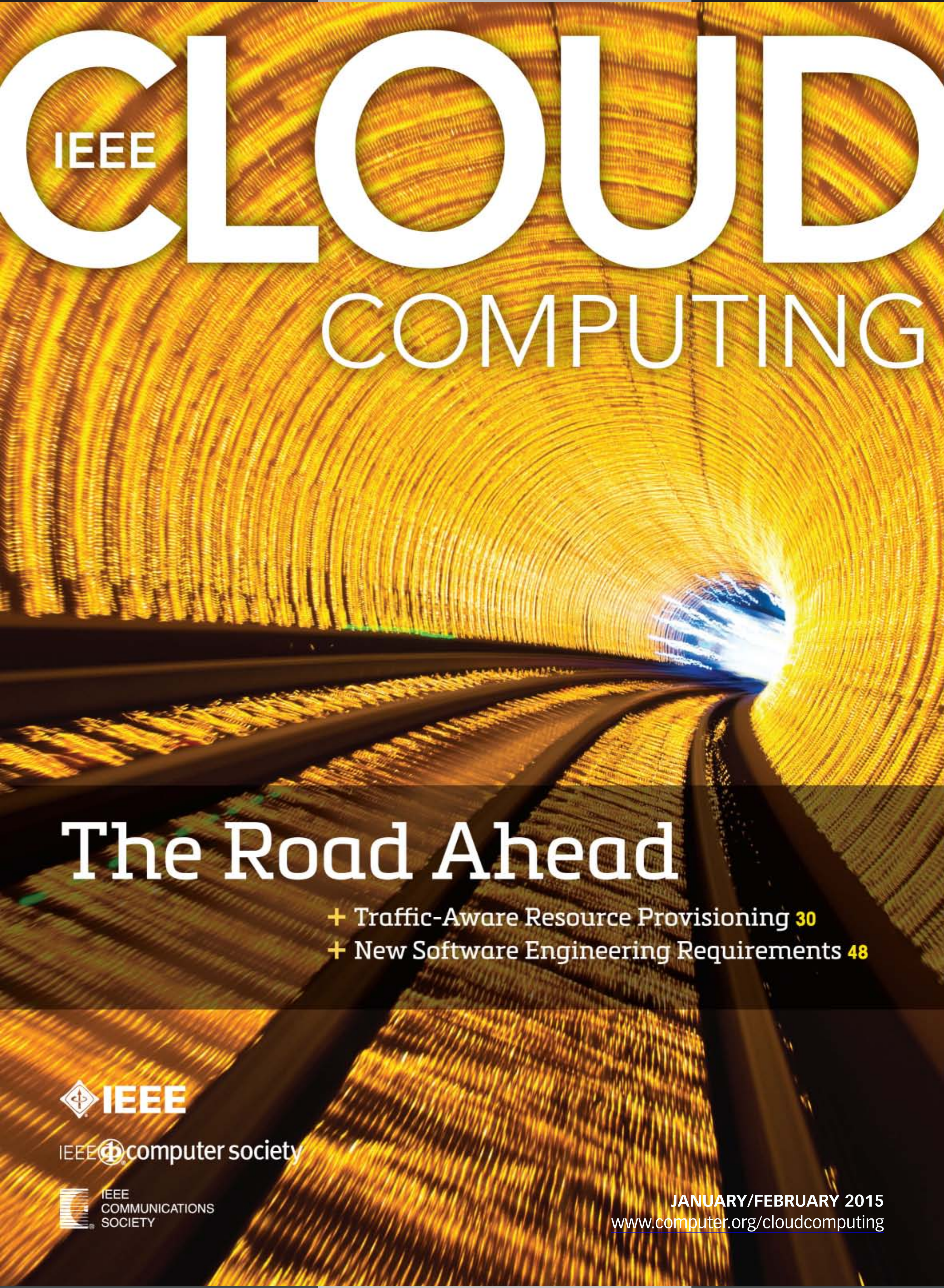

IEEE CLOUD COMPUTING

The Road Ahead

+ Traffic-Aware Resource Provisioning 30
+ New Software Engineering Requirements 48

IEEE

IEEE computer society

IEEE COMMUNICATIONS SOCIETY

JANUARY/FEBRUARY 2015
www.computer.org/cloudcomputing

CLOUD ECONOMICS COLUMN

Cloud Pricing and Markets

JOE WEINMAN

joeweinman@gmail.com

PRICING IS AN IMPORTANT MEANS BY WHICH CLOUD SERVICE PROVIDERS COMPETE. Price wars have gotten a lot of press, with providers lowering prices dozens of times over the last few years, nominally due to reductions in cost structure, but also in an attempt to gain share, receive publicity, and signal their aggressiveness in the market to competitors. However, although cloud *prices* are of interest, so are cloud *pricing models*, such as reserved instances, sustained-use pricing, and spot instances. It would be a mistake to consider pricing models to be an afterthought; they're a means of competitive differentiation as well as for creating value for both customers and providers.

Pay Per Use and Beyond

The initial model for cloud pricing was pay per use, in which the price was proportional to the product or the quantity of resources and time allocated—for example, 2 medium instances * $0.10 per hour/instance * 3 hours. Since then, various providers have introduced other pricing models. ProfitBricks introduced per-minute billing. Google introduced sustained-use pricing, whereby customers with more consistent use pay a lower unit cost than those with spiky demand, with such determination made after the fact.

Amazon Web Services (AWS) introduced reserved instances, which sound like reservations as for a hotel room, but are more akin to a discount for a volume commitment within a given time period, without implying prereservation of timeslots. As I predicted in 2007, dynamic pricing is a logical implication of clouds' perishable capacity.[1] In 2009, AWS also introduced spot instances, which, among other things, brought dynamic pricing to the cloud (see http://aws.amazon.com/ec2/purchasing-options/spot-instances). Such dynamic pricing, where the offering price varies over time, is well known in many industries, such as commodities, hotel rooms, airline tickets, and e-commerce. Amazon.com (the retailer, not the cloud) reportedly changes prices millions of times each day.[2] Firms do such things for many reasons, including yield management of perishable resources such as airline seats, hotel rooms, and computing resources; response to competitor pricing moves; and A/B demand testing. Yield management might be disguised so that price shifts don't match actual momentary capacity, while still promoting higher overall utilization. For example, Orna Agmon Ben-Yehuda and his colleagues claim that AWS prices reflect a "random reserve price that is not driven by supply and demand."[3]

Because AWS spot instances can be terminated at any time, the dynamic pricing at AWS isn't exactly like it is at Amazon.com, where delivery is still assured if the purchase is made at the agreed-upon price. Moreover, there is no predefined purchase price, but rather a bid representing an agreed-upon price limit, thus providing attributes of an auction as well: customers willing to pay more are more likely to have their workloads run. Perhaps counterintuitively, both providers and customers benefit from this multiplexing of different classes of service. Customers can save money by running their workloads at off-peak times, and providers can manage yields of perishable cloud resources by lowering the price to promote demand and drive better utilization, reducing idle capital.

In the future, no doubt some enterprising cloud providers will introduce additional innovations to the industry, such as transparency into prices for future services, as with plane tickets. Moreover, prices are also likely to vary by location, and the price at each location will be dynamic.

Different locations serviced by the same or different providers might have different prices at any given time due to a variety of reasons: higher value due to proximity to a given location, such as a stock exchange; different capacity utilization, due to aggregate customer behavior such as "follow the sun" cycles or statistical variation; or differences in power (and cooling) costs or in how power is priced (for example, breakered power, draw power, actual power, or bundled). Power is particularly of interest, because it's a large fraction of compute costs, and such power costs can fluctuate dramatically in the short term, even today. As electric grids become smart and increasingly use intermittent energy sources such as wind turbines, and electric demand response systems use pricing actions and Internet of Things connections to throttle demand, such location-dependent fluctuations will likely increase, barring breakthroughs in energy storage capabilities, distributed power generation, or transmission costs.

In such a world, cloud providers compete not only against competitors, substitutes (such as do-it-yourself in your own private cloud or a colocation facility), and between locations, but against their future selves if customers decide to defer purchase, anticipating lower future prices to run deferrable workloads.

The Impact of Dynamic Pricing

The existence of price differentials among providers, across time slots, and across locations will have several consequences: standards, segmentation, behavioral anomalies, hedging, and intermediaries.

Markets don't need standard offers or payment mechanisms to function. For example, a flea market might have one-of-a-kind items, and accept payments in not only cash or credit cards, and dollars or euros, but also in bartered exchanges of random goods and services. However, standardized offers—such as West Texas intermediate crude or #2 sweet corn—aid liquidity, as do standardized metrics for goods and services, such as 6fusion's Workload Allocation Cube (www.6fusion.com/technology/workload-allocation-cube), which fuses six dimensions of compute—memory, storage, processor, storage area network (SAN), local area network (LAN), and wide area network (WAN)—into a single figure of merit.

The IEEE Intercloud initiative has also begun to look at means for not just customers but also other service providers to acquire resources.[4] To do this requires elements such as a conversational substrate for message passing, and, for advertising services and resource availability, a resource description framework with an ontology including cloud pricing, enabling semantics for providers, offers, resources, and pricing models.[5] In the future, today's business customers may well also become providers, similar to "cogeneration," and it's long been anticipated that consumers will sell spare cycles as well.[6]

Segmentation of customers and workloads is a natural consequence of pricing model differences. Generally, consumers have differing degrees of "time consciousness," and can be divided into "time buyers" and "time sellers."[7] The first group is sensitive to delays, but less sensitive to price. They will pay for expedited shipping or pay higher airfares for a direct, nonstop flight. The second group is price conscious, and prefers to spend time rather than money. These customers are willing to spend an hour clipping coupons or take convoluted routes to avoid tolls. Cloud customers—and their applications—can be viewed the same way. Some applications must be run continuously, compliantly, and/or immediately, with very little price sensitivity, or, more likely, the revenue associated with the application greatly outweighs its cost. Other *deferrable* workloads can be delayed until the price is right, while some *discretionary* workloads might not be run at all.

Hedging will become more commonplace as customers, possibly with the aid of intermediaries, attempt to reduce business risk as they increasingly encounter computing environments where prices to execute business-critical workloads can fluctuate. This might include hedging cloud prices through derivatives such as cloud price futures and options, or through existing mechanisms for hedging other potentially relevant business risk components such as energy costs or currency fluctuations.

Behavioral anomalies are also relevant to dynamic pricing. Most customers exhibit a "flat-rate bias" (an irrational preference for flat rates over pay per use) due to "loss aversion," the excessive psychological weighting of losses over gains, as well as poor forecasting of demand and the "taxi-meter effect," in which panic over the size of the bill rises as your taxi sits in traffic. Some exhibit the reverse—a "pay-per-use bias."[8] The main point to note is that rational economic considerations sometimes take a back seat to cognitive biases.

Intermediaries are already arising. Examples include cloud service brokers and markets for trading cloud computing capacity and reserved instance commitments, and for trading derivatives. One key intermediary is a market in which a customer can select from multiple cloud providers to meet

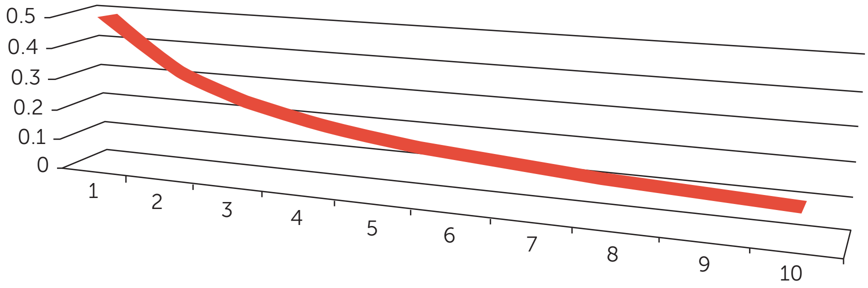

Figure 1. Expected value of best price given *n* providers with dynamic prices uniformly distributed on [0, 1].

a particular compute need. An interesting area of research is to quantify the value of such markets and the essential characteristics required to deliver such value.

The Value of a Market

The value of a cloud market to a customer depends on that customer's requirements and how many providers can meet those requirements within a constellation of feasible tradeoffs. It also depends on where the customer is in the purchase process. In attempting to discern viable service providers, a market (or advisory service) can reduce search costs; in facilitating payment, it can reduce (or increase) transaction costs; and in selecting a provider or location, it can reduce service costs.

In the real world, service comparisons are complicated by planned and unplanned differences between offers. Planned differences include datacenter power reliability through dual power grids, storage reliability through local and remote mirroring, and performance, due to storage network bandwidth. Unplanned differences, such as the degree of downtime due to major weather events or performance issues such as "noisy neighbors," also make service comparisons difficult. However, to keep the analysis simple, consider a market for a single, undifferentiated cloud resource such as compute or storage.

All other things—performance, availability, security, service portfolio, quality, support, and so on—being equal, if there is a clear price leader, the rational economic decision would be to select that provider.

Suppose, however, to make things interesting, that there are multiple providers, each of which dynamically varies prices within the same range. To keep the analysis simple, assume that for each provider, the cloud resource price fluctuates independently and uniformly from \$0.00 to \$1.00. If there is only one provider, the expected value of the price is clearly \$0.50, but what if there are n providers?

In the case of a uniform distribution on [0, 1], the theory of order statistics says that the expected value of the minimum value of n independent uniform values is simply $1/(n + 1)$. This means that the expected price that a customer can find in such a cloud market looks like Figure 1.

One key insight is that it doesn't take many providers for such a market to deliver value. Three providers deliver a 50 percent saving in expected price over just one, five deliver a two-thirds savings.

Caveats

However, there are also caveats.

Range and Distribution

When price ranges from \$0.00 to \$1.00, the degree of savings is notable. If the range was from \$100 to \$101, the savings would be inconsequential. Also, we assumed a uniform distribution, but a very tight normal distribution, say, would drive trivial benefits.

Independence

In the case of cloud pricing, this assumption might not hold. One reason for this is explicit collusion between providers to set prices, which is generally illegal, however. A second reason is a type of tacit collusion, which could actually be an emergent effect of the system dynamics of yield management algorithms. A first algorithm at a first provider raises prices based on higher demand. This drives some customers to the second provider, who in turn raises prices based on that higher demand. Or there can be feedback and amplification effects among pricing algorithms. The list price of a single copy of the genetics textbook *The Making of a Fly* (Wiley-Blackwell, 1992) famously shot from \$60 to more than \$23 million due to feedback cycle of interactions between two automated pricing algorithms.[9] A third reason is correlated demand and thus correlated utilization and thus correlated dynamic pricing. Pricing for airline seats during spring break and for New York City hotel rooms during the holidays is high; similarly, providers might charge more for clouds serving retailers during peak holiday shopping periods or enterprises for quarterly book close. There are also longer-term correlations, such as business demand due to macroeconomic cycles.

Switching Costs

Another practical issue is the cost to switch providers. It would not be economically wise to incur millions of dollars in data transfer costs to move a large data warehouse every hour to save 25 cents on processing. However, some applications, such as e-commerce front-ends and content delivery, might not incur such costs. Or, multiple providers may have a presence within a single provider-neutral datacenter.

Network Costs

There are also costs related to network architecture, performance, and pricing. Provider switching is facilitated when there is low latency, as in a multiprovider colocation facility, and low and/or flat rates for network transport, as with some direct connect or interavailability-zone pricing.

Application Capabilities

Most cloud applications are not architected to be dynamic-pricing aware. Dynamic markets may require application logic, cloud management logic, or access to micro-services to enable dynamic changes of providers or locations.

Game Theory and Customer Expectations

In a world of perishable capacity, selling resources at any price above marginal cost is rational. For example, for a transpacific flight that's about to take off, an airline could make, say, $5 by selling an empty seat for $20, assuming a marginal cost of $15 extra for jet fuel due to the added weight. However, implementing such a plan would destroy advance sales as customers queued up before flight time, and then not work anyway as the first customer would balk at being charged the entire cost of the flight. Similar logic—preserving pricing and customer perception of value—applies to cloud providers.

PRICING IS A FASCINATING DISCIPLINE; ONE IN WHICH CLOUD PROVIDERS CAN AND DO DIFFERENTIATE THEMSELVES. As dynamic pricing becomes more prevalent and transparent and Intercloud standards and other standardized metrics gain traction, related financial intermediaries and instruments are likely to rise in importance. ●●●

References

1. J. Weinman, "Evolution of Networked Computing Utilities," *Business Comm. Rev.*, Nov. 2007, pp. 36–44; www.joeweinman.com/Resources/WeinmanUtility.pdf.
2. M.J. Perry, "Dynamic Pricing: Amazon and Delta Airlines," blog, 16 Dec. 2013; www.aei.org/publication/dynamic-pricing-amazon-and-delta-airlines.
3. O.A. Ben-Yehuda et al., "Deconstructing Amazon EC2 Spot Instance Pricing," *Proc. IEEE 3rd Int'l Conf. Cloud Computing Technology and Science* (CloudCom 11), 2011, pp. 304–311.
4. J. Weinman, "What's Next for the Cloud? The Intercloud," Forbes.com, 8 Oct. 2013; http://www.forbes.com/sites/joeweinman/2013/10/08/whats-next-for-the-cloud-the-intercloud-2/.
5. B. Di Martino et al., "Towards an Ontology-based Intercloud Resource Catalogue: The IEEE P2302 Intercloud Approach for a Semantic Resource Exchange," to be published in *Proc. IEEE Int'l Workshop Cloud Computing Interclouds, Multiclouds, Federations, and Interoperability*, 2015.
6. O. Regev and N. Nisan, "The POPCORN Market—An Online Market for Computational Resources," *Proc. 1st Int'l Conf. Information and Computation Economies* (ICE 98), 1998, pp. 148–157.
7. M. Kleijnen, K. de Ruyter, and M. Wetzels, "An Assessment of Value Creation in Mobile Service Delivery and the Moderating Role of Time Consciousness," *J. Retailing*, vol. 83, no. 1, 2007, pp. 33–46.
8. A. Lambrecht and B. Skiera, "Paying Too Much and Being Happy about It: Existence, Causes and Consequences of Tariff-Choice Biases," *J. Marketing Research*, vol. 43, no. 2, 2008, pp. 212–223.
9. M. Eisen, "Amazon's $23,698,655.93 Book about Flies," blog, 22 Apr. 2011; www.michaeleisen.org/blog/?p=358.

JOE WEINMAN *is the chair of the IEEE Intercloud Testbed executive committee. 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.*