Intercloudonomics: Quantifying the Value of the Intercloud

IN *CLOUDONOMICS*, I ADDRESSED TOTAL COST AND PERFORMANCE OPTIMIZATION FOR A CUSTOMER IN RELATION TO A CLOUD PROVIDER.[1] For example, all other things being equal, a hybrid cloud architecture can often lead to total cost savings in the presence of variable demand, even if the unit cost of the cloud services is priced at a premium. However, in addition to simple customer-cloud relationships, there are also cloud-cloud relationships. In the same way that the Internet is a set of interoperable, loosely coupled networks, the Intercloud is intended to be a set of interoperable, loosely coupled clouds. The Intercloud promises reduced complexity, optimized prices, reduced latency, enhanced reliability, and better capacity utilization. For example, a cloud provider can extend its geographic reach or become a virtual operator through footprint augmentation, that is, by leveraging physical resources such as compute, network, and storage in other geographic regions—from partners or competitors. Or, in another scenario, a cloud provider can enhance its reliability by failing over to or replicating data to a facility operated by another cloud provider. All of these benefits can be quantified.

JOE WEINMAN

joeweinman@gmail.com

What Is the Intercloud?

As used here, the Intercloud is an emerging, generic concept, not an offer from any particular cloud provider or enabling technology vendor, and not to be confused with a hybrid cloud (a combination of an enterprise datacenter and public cloud provider resources) or a multicloud (the use by a customer of multiple cloud providers). Broadly speaking, it's analogous to similar concepts from other industries.[2] Airlines, for example, will use capacity from other airlines, for reasons such as limited capacity (for example, rebooking a passenger from an overbooked flight to a competitor's), acting as a (mobile) virtual (airline network) operator (that is, "code-sharing"), or federating to extend a geographic footprint, as with the Star Alliance (United Airlines, Lufthansa, Air China, and so on). Similarly, in cellular telephony, some operators are mobile virtual network operators and most have international roaming agreements. The differences between the Intercloud, multiclouds, and hybrid clouds become clearer by analogy: a hybrid transportation solution might entail a combination of your own car and a public airline service; a multicloud might be akin to a passenger buying one ticket on one airline and a separate ticket on a different one; and the Intercloud is similar to purchasing a multi-leg trip from a

2325-6095/15/$31.00 © 2015 IEEE

single airline, say, United, and letting it worry about code-sharing and ticketing on alternate airlines.

IEEE is developing a set of Intercloud standards, such as P2301 and P2302,[3] for portability, interoperability, and federation among cloud providers. These promise benefits for both customers and cloud providers. For example, a customer is more likely to be able to acquire and pay for needed capacity, even if a preferred cloud service provider has a temporary “stock-out.” To enable this and related scenarios, emerging approaches[4] will allow cloud providers to advertise and acquire resources via a shared communications substrate and ontology (like clouds, airlines must be able to talk to each other and have a common understanding of seat classes, airports, departure and arrival times, and so on), and workloads must be portable (passengers must be able to move from one plane or airline to another).[5] Moreover, a control-plane layer needs to be able to orchestrate dynamic allocation of atomic resources across the Intercloud in real time: one free airline seat shouldn’t be allocated to two different customers, and neither should a compute resource.

Interface Complexity Benefits

If there are n cloud providers, a single uniform standard for user-to-cloud interfaces such as the Open Cloud Computing Interface (http://occi-wg.org) reduces the complexity of access by a factor of n. This can be viewed similarly to a traveler requiring multiple electricity adapters (or perhaps, electric appliances) if travelling to multiple regions of the world, but only needing a single one if there were a universal standard. As far as the Intercloud, a single standard for cloud-provider-to-cloud-provider interfacing would reduce the number of interfaces from $n - 1$ per cloud provider and thus perhaps $n(n - 1)$ individual or $n(n - 1)/2$ partner interface development efforts in total to only n such efforts, as Figure 1 shows.

Market Benefits

In today’s cloud market, there are a number of providers, and dynamic (that is, time-varying) pricing has emerged through mechanisms such as “spot instances” (https://aws.amazon.com/ec2/spot). Using the Intercloud and its mechanisms for advertising resources, service-level agreements (SLAs), and prices, customers could exploit dynamic pricing and market dynamics to lower their expected average cost of resources by dynamically “shopping around” for the lowest cost provider, as with Expedia or Travelocity, and migrating workloads via, say, live virtual server migration or containers. I addressed the quantification of these exact benefits in a prior column.[6] To recap briefly, the theory of order statistics teaches that if prices are random, independent, and identically distributed (IID) uniformly over the same range, say, [0, 1], the expected minimum price as the number of participants in a market grows to n cloud providers is $1/(n + 1)$, as Figure 2 shows (subject to caveats such as independence, distribution, and real price ranges). This implies that a market with even a few cloud providers can provide benefits.

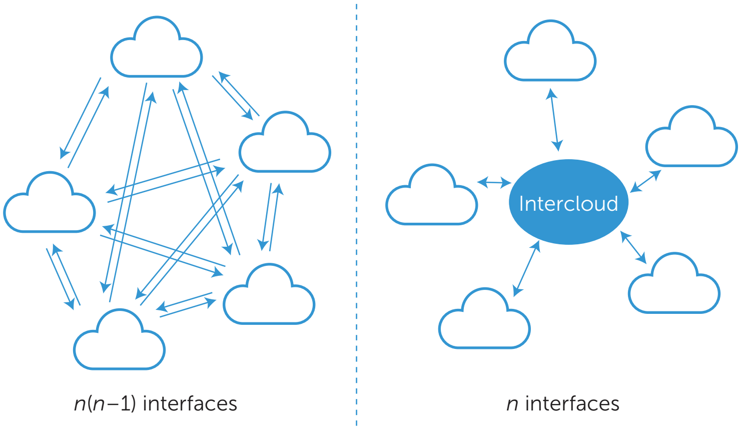

FIGURE 1. Reduction in interface complexity among n cloud service providers.

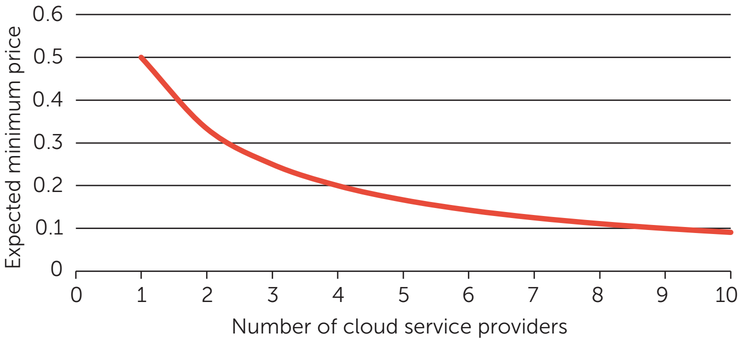

FIGURE 2. Expected minimum price is proportional to $1/(n + 1)$ for n cloud service providers with uniform IID prices on [0,1].

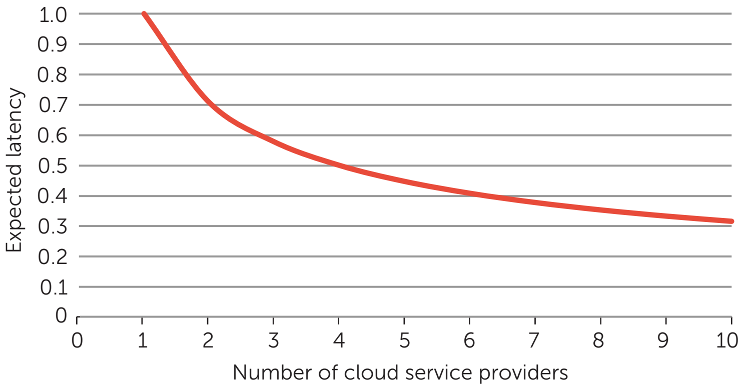

FIGURE 3. Expected planar latency proportional to $1/\sqrt{n}$ for n cloud service providers.

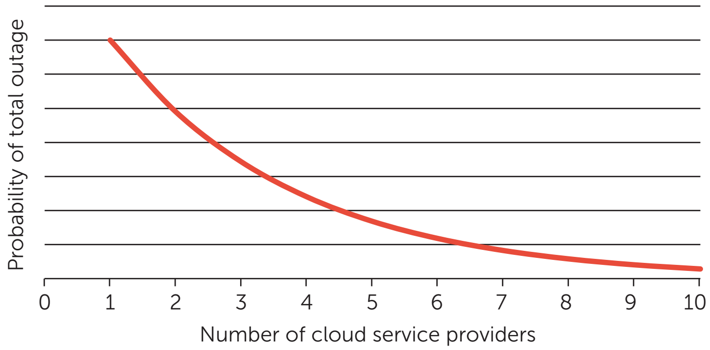

FIGURE 4. Probability of total Intercloud outage of $(1 - p)^n$ for n independent cloud service providers.

Footprint Augmentation

Any given cloud service provider has a given geographic footprint, such as, say, New York, Miami, and San Francisco. Because latency for interactive tasks can be reduced through geographic proximity, average and worst-case latency can be reduced by having resources closer to end users, such as through a content delivery network (cloud) or edge computing through a highly dispersed cloud.

One way to deploy such resources is to directly invest in facilities. Another is to augment one's own footprint by federating with one or more other providers. In the worst case (for geographic dispersion), the one or more other providers have exactly the same footprint. Having another facility in each of New York, Miami, and San Francisco won't reduce data transport latency due to propagation delay (although it might due to latency issues induced by insufficient bandwidth or server capacity).

However, in the best case, one cloud provider's geographic footprint will be optimally augmented by another's highly complementary footprint. For example, a single facility in New York serving a global audience won't have latency reduced much by one in New Jersey, but it certainly will if complemented by a facility in Singapore.

On a plane, the average and worst-case latencies are proportional to the inverse square root of the number of nodes, or facilities. This is because the area covered within a radius r is, of course, πr^2, so s service nodes can cover an area of $A \propto s\pi r^2$ (the exact equation depends on the degree of coverage overlap). Therefore, holding A constant, r, which is distance and is thus a reasonable proxy for latency, follows $r \propto 1/\sqrt{s}$. For a sphere, such as the planet Earth, this formula needs to be adjusted: for reasonable numbers of nodes this approximately holds; but given only one node, an additional (antipolar) node will halve latency, so the ratio is $r \propto 1/s$.[7] As the number of nodes on a sphere increases, the inverse square root law holds more closely.

If each of n cloud providers has k nodes, then we have $s = nk$ and the latency is proportional to $1/\sqrt{s} = 1/\sqrt{nk} = 1/(\sqrt{n}\sqrt{k})$ and since k is fixed, the latency is merely proportional to the inverse square root of the number of cloud providers, as Figure 3 shows.

Reliability

Experience has shown that a number of issues can befall cloud providers. These can occur at a single facility, as when four successive lightning strikes caused lost data at a Google datacenter in Belgium earlier this year,[8] or spread across multiple facilities, as when software issues caused a wide-scale outage at Amazon Web Services on Christmas Eve, 2012 (see https://aws.amazon.com/message/680587), and more recently in September 2015.[9] In fact, every major cloud provider has been down for either planned or unplanned reasons, or both.[10]

Let's assume that the probability of any given physical datacenter site being available is p (and thus the chance of it being down is $1 - p$). If site outages are independent, then if there are s sites, the probability that they are all down is $(1 - p)^s$; therefore, the probability that at least one site is still functional is $1 - (1 - p)^s$. If each of n cloud providers has k sites, then $s = nk$ and the probability that they're all down is $(1 - p)^s = (1 - p)^{nk} = (1 - q)^n$, where $q = 1 - (1 - p)^k$. Put differently, whether we're talking about site outages or system-wide cloud provider outages, the same generic curve, shown in Figure 4, applies. In practice, additional considerations such as software reliability, network reliability, and network capacity to support mirroring or replication will be important.

To be fair, a single (set of) common Intercloud protocol(s) in use across multiple providers could provide a means for enhancing reliability, but it could also be the foundation for anomalies and unintended consequences leading to Intercloud-wide outages, or introduce a new vulnerability that wouldn't exist across heterogeneous environments.

Demand at a Single Cloud Provider

To understand the benefits of capacity sharing among cloud providers, we must first understand the demand at any given cloud provider. Whether an internal private cloud or a public cloud, one of the essential characteristics of a cloud is that it utilizes a dynamically allocated shared resource pool to serve multiple workloads: different customers, business units, and/or application types, each with varying demand. A retailer, for example, might have peaks on Black Friday, Cyber Monday, and during its semi-annual private sale. A tax preparation firm may have peaks for early filers in mid-February and for late ones on 15 April. Another firm might do load testing every few weeks. Yet another might be a broker or bond trader specializing in the transportation sector. A news outlet might have peaks as important events occur such as elections, natural disasters, or celebrity situations.

The central limit theorem says that whenever an increasing number of (what, for computing workloads, for all intents and purposes are) random variables are added, the sum increasingly follows a normal distribution. So let's assume that the aggregate demand at that cloud provider is $D(t)$, a time-varying random variable with mean μ and variance σ^2, and thus standard deviation σ. You'll recall that a normal distribution follows the famous "bell-shaped" curve in its probability density function, as Figure 5 shows, and the larger the standard deviation, the "wider" that curve is.

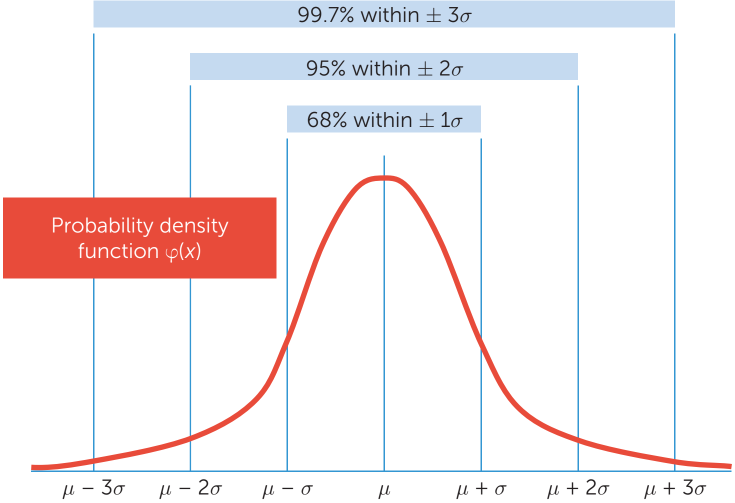

FIGURE 5. The normal distribution with mean μ and standard deviation σ.

What about capacity? A cloud provider can't deploy infinite capacity, so there's always a chance that the provider, or a given location, will have insufficient capacity due to either a single customer suddenly requiring explosive capacity, or a confluence of expected and unexpected peaks. For example, 14 February might represent a peak due to early tax filers, Valentine's Day flower and candy purchases, and the unexpected death of a celebrity. It's tempting to assume that computing is free and cloud providers will deploy "near-infinite" capacity, but it's perhaps more likely that eventually cloud providers will look like airlines, oriented toward maximizing profitability through maximal resource utilization.

Insufficient Capacity

For such a provider, how often will it suffer from insufficient capacity? To put it another way, what's the probability that a cloud provider with a given (fixed) capacity C will have sufficient capacity to meet all the customer demand given a (time-varying) demand function $D(t)$? Let's call this the sufficiency probability $S(D(t), C)$, which is the probability $p(D(t) < C)$.

When $D(t)$ is normally distributed with mean μ and standard deviation σ, we can benefit by expressing C in terms of μ and σ as well, namely as $C = \mu + k\sigma$. Our job is now easy, because S is now just the Φ function, that is, the cumulative distribution function (CDF) $\Phi(x)$ for a normally distributed

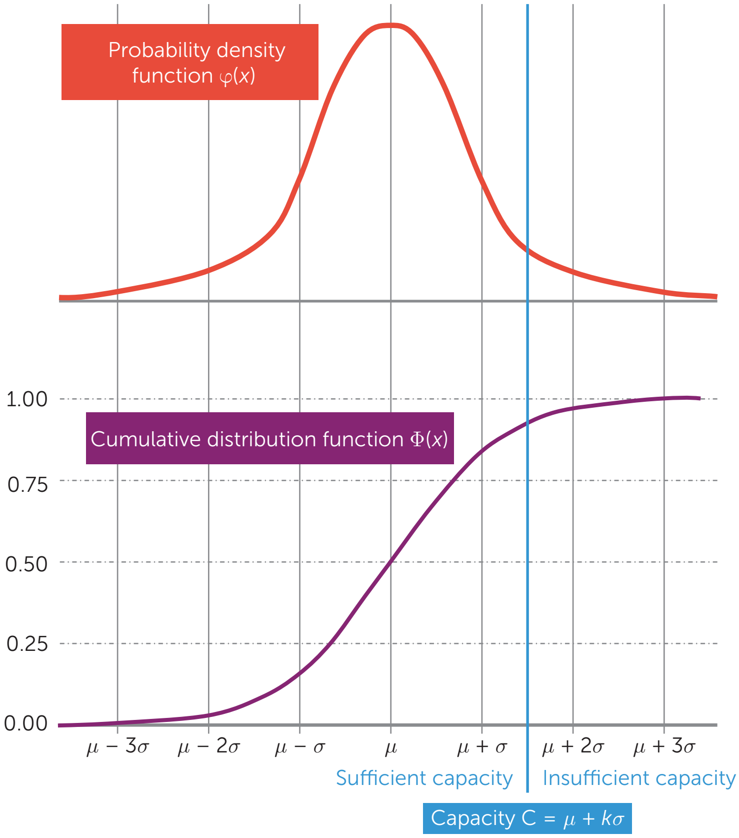

FIGURE 6. The probability density function $\varphi(x)$ and CDF $\Phi(x)$ for a normal distribution, and a capacity level $C = \mu + k\sigma$ set in terms of that distribution.

random variable with probability density $\varphi(x)$, as Figure 6 shows.

The function $\Phi(x)$ tells us the probability that the random variable is less than x (where x is the number of standard deviations). To normalize, we set $x = k = (C - \mu)/\sigma$. In other words, we can determine the portion of the bell curve that lies to the left of C, representing demand less than capacity, and the portion of the bell curve that lies to the right of C, representing how frequently demand exceeds capacity, as Figure 6 shows ($1 - \Phi(x)$ is often referred to as the Q function, or (right) tail probability, $Q(x)$).

Unfortunately, such a function isn't easy to evaluate, but fortunately, precalculated tables exist. Table 1 is a simplified one.

To make these abstractions more concrete, recall that about 68 percent of values for a normally distributed random variable will lie within one standard deviation of the mean, as shown in Figure 5. This means that about 32 percent lie outside the mean, leaving about 16 percent to be less than $\mu - (1)\sigma$ and 16 percent to be more than $\mu + (1)\sigma$. So, if we set capacity $C = \mu + (1)\sigma$, we'll have sufficient capacity about $\Phi(1) \cong 0.84 = 0.68 + 0.16$ of the time. Because about 95 percent of values lie within two standard deviations, if we set capacity to be $\mu + 2\sigma$, by similar logic, we would have sufficient capacity just over 97.5 percent of the time. If we set capacity at $\mu + 3\sigma$, we'd have sufficient capacity about 99.9 percent of the time, and so on.

Why Not Just Set Capacity at $\mu + 3\sigma$ or Higher?

To maximize the likelihood of serving customers (and capturing revenue), why wouldn't a cloud provider just set capacity 3 or 4 or more standard deviations above the mean, thus typically ensuring sufficient capacity? The reason is that having enough capacity to meet extremely rare spikes means that all the rest of the time (that is, virtually always) there's way too much capacity. This excess capacity can represent a large capital investment or operating lease commitment for equipment as well as, typically, for additional cost structure elements such as heating, ventilation, and air conditioning, floor space, insurance, and power. "Typically," because some technologies—say, intermittently powering down unused servers—can reduce costs such as power and cooling as well as extend the average life of equipment. Running with a poorly managed cost structure would not be a good strategy in today's highly competitive cloud marketplace, made increasingly competitive due to intermediaries.[11]

Consequently, a cloud provider faces a conundrum. High levels of capacity maximize revenue, but at a high cost; low levels reduce costs, but also reduce revenue. Therefore, the challenge is to find a balance between lost revenue and costs associated with fixed capacity, while recognizing that a resale model of other cloud providers' capacity is likely to have a different cost structure than one based on one's own capacity. In the same way that an individual customer can complement limited capacity with the elastic capacity offered by a cloud provider, so can a cloud provider.

The Case of the Intercloud

Let's expand the simple demand and capacity model to two or more clouds. At any given time, a cloud might have excess capacity (extra unsold airline seats or virtual machines) or exhibit insufficient capacity (that is, be overbooked). Most of this article focuses on how the Intercloud can address the latter case, but of course if cloud provider A sells capacity from cloud provider B, whether because it's a virtual operator, or because of capacity outages or spikes in demand, it solves A's under-capacity and B's overcapacity issues simultaneously.

Assuming we have n clouds, let the demand at each cloud be represented by $D_i(t)$ for $i = 1, 2, \ldots, n$. To keep things simple, let's assume that each cloud has independent, identically distributed demand, namely, normally distributed with mean μ and standard deviation σ. Keeping things simple, let's assume that each cloud hires the same capacity planning consulting firm and determines that $C = \mu + k\sigma$ is the optimal capacity; in other words, that $C_i = \mu + k\sigma$ for all $i = 1, 2, \ldots, n$.

Because the Intercloud enables capacity sharing, where provider 1 can provide needed capacity to provider 2 or vice versa, the aggregate capacity is now $C^+ = nC = n(\mu + k\sigma)$. The aggregate demand can be represented by

$$D^+(t) = \sum_{i=1}^{n} D_i(t).$$

For independent, normally distributed random variables, the sum is also normally distributed, the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances. Consequently, the mean of $D^+(t)$is $n\mu$ and, because the standard deviation is the square root of the variance, the standard deviation of $D^+(t)$ is $\sqrt{n\sigma^2}$, which is just $\sqrt{n}\sigma$.

Let us now ask how $S(D(t), C)$, the capacity sufficiency of a single cloud provider, compares with $S(D^+(t), C^+)$, the sufficiency of the Intercloud. We already know that $S(D(t), C) = \Phi(k)$, where $k = (C - \mu)/\sigma$. Similarly, we can say that $S(D^+(t), C^+) = \Phi(k^+)$, but we need to figure out what k^+ is. It follows that $C^+ = nC = n\mu + k^+\sqrt{n}\sigma$, and we recall that $C = \mu + k\sigma$, so $nC = n\mu + nk\sigma$. Therefore, $nC = n\mu + nk\sigma = n\mu + k^+\sqrt{n}\sigma$. Eliminating the $n\mu$ term and dividing both sides by $\sqrt{n}\sigma$ tells us that $k^+ = \sqrt{n}k$. Since all CDFs are nondecreasing, but the CDF for the normal distribution is monotonically increasing, and since it's always the case that $\sqrt{n} > 1$ whenever $n > 1$, $\Phi(k^+) > \Phi(k)$ and thus $S(D^+(t), C^+) > S(D(t), C)$ whenever there's an Intercloud with two or more cloud providers, that is, $n \geq 2$. To put it more simply, the Intercloud will have sufficient capacity more often than a single provider will.

Table 1. Values for the cumulative distribution function $\Phi(x)$ for a standard normal distribution.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.0	0.5000	1.0	0.8413	2.0	0.9772
0.1	0.5398	1.1	0.8643	2.1	0.9821
0.2	0.5793	1.2	0.8849	2.2	0.9861
0.3	0.6179	1.3	0.9032	2.3	0.9893
0.4	0.6554	1.4	0.9192	2.4	0.9918
0.5	0.6915	1.5	0.9332	2.5	0.9938
0.6	0.7257	1.6	0.9452	2.6	0.9953
0.7	0.7580	1.7	0.9554	2.7	0.9965
0.8	0.7881	1.8	0.9641	2.8	0.9974
0.9	0.8159	1.9	0.9713	2.9	0.9981

Some examples show this effect, using the reduced set of tabulated values of $\Phi(x)$ shown in Table 1. Suppose there are two providers ($n = 2$) and capacity is set at one standard deviation above the mean ($k = 1$). Then, $k^+ = \sqrt{n}k = \sqrt{2} \times 1$, so the probability of sufficient capacity rises from about 84 percent for a single provider to about 92 percent for two. If we increase the number of providers to four, probability rises to almost 98 percent. If we start with a capacity at two standard deviations above the mean, our probability of sufficiency is already close to 98 percent, but just one cloud partner takes us to about 99.8 percent $\left(k^+ = \sqrt{n}k = \sqrt{2} \times 2\right)$. A few percentage points or tenths of a percent might not seem like a lot, but we are talking about multibillion-dollar businesses that are headed to be multi-ten-billion-dollar ones.

Caveats

The assumptions that permit a straightforward mathematical analysis might not hold exactly in the real world. For example, a normal distribution can take on unboundedly large negative values, but demand can't be negative. Also, workloads might not be divisible into parts that can be distributed at all, much less across multiple geographically distributed entities, due to either the performance of the technical architecture or the reality of Intercloud data transport surcharges. It's an NP-complete problem, that is, computationally intractable, to divide up a number of workloads across multiple buckets (cloud providers or cloud locations) when the workloads are indivisible and vary in size.[12]

Also, cloud providers aren't at the mercy of individual or aggregate demand. They can attempt to shape demand by utilizing dynamic pricing, spot instances, or promotions to incent demand during low-demand periods, or utilizing surge pricing to disincent it. Or, like Google, they can accomplish both at the same time by effectively offering a discount for flatter demand patterns.[13] However, although this might flatten demand, there's still an argument to be made that demand is normally distributed, just with reduced variance, and thus there's still a role for the Intercloud.

Further Research

Ultimately, a provider doesn't just want to reduce the likelihood of being unable to serve customers, but to maximize profitability. A full model would consider the margin from utilizing others' capacity by selling at one price but acquiring the resources at one or more wholesale prices; the margin from selling one's own capacity to resellers; a variety of capacity levels, distributions, costs, and perhaps quality from different sized providers, and the cost of moving workloads and moving, replicating, or remotely accessing data associated with those workloads. Moreover, we don't just want to know the probability of insufficient capacity, but the lost revenue and/or profit, and how much is recaptured through the Intercloud. This requires utilizing the expected value of the tail distribution, i.e., Q-function, less the baseline capacity $E(Q(x)) - C = E(1 - \Phi(x)) - C$, which is similar to the "hazard rate," times the total actual capacity and the profit contribution from that capacity. There might also be game theoretic and option value considerations: does serving a given customer preclude serving a different customer who might have a greater willingness to pay a higher future price?

THE INTERCLOUD—AS WITH RELATED CONCEPTS FROM DOMAINS BEYOND COMPUTING—PROMISES A VARIETY OF BUSINESS BENEFITS, SUCH AS REDUCED LATENCY, ENHANCED RELIABILITY, AND GREATER ABILITY TO SERVICE CUSTOMER DEMAND. Even a simple analysis can quantify what these are under conditions of uncertainty.

Acknowledgments

Many thanks to Henry C.B. Chan, an associate professor at the Hong Kong Polytechnic University; Gerard Parr, professor and chair in telecommunications engineering at the University of Ulster Coleraine; Alexander Willner, a research assistant at Technische Universität Berlin; and H. N. "Raj" Nagaraja, Professor Emeritus of Bio-Statistics at Ohio State University, for thoughtful comments on a first draft of this article.

References

1. J. Weinman, *Cloudonomics: The Business Value of Cloud Computing*, Wiley, 2012.
2. J. Weinman, "What's Next for the Cloud? The Intercloud," *Forbes.com*, 8 Oct. 2013; www.forbes.com/sites/joeweinman/2013/10/08/whats-next-for-the-cloud-the-intercloud-2.
3. D. Bernstein, V. Deepak, and R. Chang, *IEEE P2302/D0.2 Draft Standard for Intercloud Interoperability and Federation (SIIF)*, IEEE P2302/D0.9, IEEE, 2015.
4. B. Di Martino, G. Cretella, and A. Esposito, "Advances in Applications Portability and Services Interoperability among Multiple Clouds," *IEEE Cloud Computing*, vol. 2, no. 2, 2015, pp. 22–28.
5. B. Di Martino et al., "Towards an Ontology-Based Intercloud Resource Catalogue—The IEEE P2302 Intercloud Approach for a Semantic Resource Exchange," *Proc. 2015 IEEE Int'l Conf. Cloud Eng.* (IC2E), 2015.
6. J. Weinman, "Cloud Pricing and Markets," *IEEE Cloud Computing*, vol. 2, no. 1, 2015, pp. 10–13.
7. J. Weinman, "As Time Goes By: The Law of Cloud Response Time," working paper, 2011; http://joeweinman.com/Resources/Joe_Weinman_As_Time_Goes_By.pdf.
8. D. Worth, "Lightning Strike Causes Google Data Centre Power Outage and Data Loss,"

V3.co.uk, 20 Aug. 2015. www.v3.co.uk/v3-uk/news/2422770/lightning-strike-causes-google-data-centre-power-outage-and-data-loss.

9. B. Butler, "3 Big Takeaways from Amazon's Latest Cloud Outage," *Network World*, 21 Sept. 2015; www.networkworld.com/article/2985128/cloud-computing/3-big-takeaways-from-amazon-s-latest-cloud-outage.html.
10. B. Butler, "Which Cloud Providers Had the Best Uptime Last Year?" *Network World*, 12 Jan. 2015; www.networkworld.com/article/2866950/cloud-computing/which-cloud-providers-had-the-best-uptime-last-year.html.
11. J. Mitchell, "What's the Best Way to Purchase Cloud Services?" *IEEE Cloud Computing*, vol. 2, no. 3, 2015, pp. 12–15.
12. J. Weinman, "Cloud Computing Is NP-Complete," working paper, 2011; http://joeweinman.com/Resources/Joe_Weinman_Cloud_Computing_Is_NP-Complete.pdf.
13. "Introducing Sustained Use Discounts—Automatically Pay Less for Sustained Workloads on Compute Engine," blog, 4 Apr. 2014; http://googlecloudplatform.blogspot.com/2014/04/introducing-sustained-use-discounts.html.

JOE WEINMAN *is a frequent keynoter and the author of* Cloudonomics *and* Digital Disciplines. *He also serves on the advisory boards of several technology companies. Weinman has a BS from Cornell University and an MS from the University of Wisconsin-Madison, both in computer science, and he has completed executive education at the International Institute for Management Development in Lausanne. Contact him at joeweinman@gmail.com.*