Hybrid Cloud Economics

PROGNOSTICATORS HAVE BEEN ARGUING OVER THE FUTURE OF IT. One rarely hears comments any longer about cloud computing along the lines of those attributed to Oracle CEO Larry Ellison in 2008, "What is it? It's complete gibberish. It's insane. When is this idiocy going to stop?"[1] However, some argue that the future of IT is 100 percent public cloud for all workloads and all companies; others argue for a hybrid future. Additional scenarios include bimodal IT,[2] multiclouds, and the Intercloud,[3] not to mention additional elements such as microcells, fog computing,[4] and colocation facilities. Sound economic and mathematical analysis shows that a rational outcome depends on underlying assumptions. Under assumptions that appear likely to hold given today's technologies, a multicloud or single public cloud environment is most likely for small businesses, including startups, whereas a hybrid cloud future is likely for large enterprises. A quantitative total cost analysis depends on several key factors: the relative unit cost of resources, performance differences, usage patterns, and uncertainties regarding future use. There are also qualitative factors that affect decisions: the desired focus of management attention, core competencies, and cognitive biases and human motivational drivers. Because of these complexities, a single solution is unlikely to fit all requirements.

JOE WEINMAN

joeweinman@gmail.com

Private, Public, Hybrid, and Other Clouds

Some argue that "private cloud" is a misnomer or marketing hype. However, the term encapsulates the notion of firm-owned resources (hence the "private") coupled with the notion that they're dynamically allocated across multiple workloads (the "cloud"). Instead of ownership, such resources might represent a fixed cost (such as financing or operating leases). A public cloud, represented by Amazon Web Services (AWS), Microsoft Azure, Google Cloud, and the like, services multiple workloads from multiple customers on a pay-per-use basis. Of course, this simple definition has grown more complex, thanks to reserved instances (which are really volume commitments with a variety of payment options including prepaid), spot instances (dynamically priced services), sustained-use pricing, and more. A true hybrid cloud should represent a seamless integration of private and public clouds, not just a random collection of clouds and disconnected uses.

For the purposes of economic analysis, private and public clouds might be better referred to as fixed-cost versus usage-based-cost clouds, respectively.

Private, public, and hybrid clouds generally refer to computing and storage. However, they appear in many industry domains: transportation, real estate, electricity, and capital. For example, for transportation, a company might own, or have fixed costs for (I'll stop repeatedly mentioning that option), a private cloud of vehicles (a car pool, delivery trucks, or installation vans, say) together with a public cloud (rental cars, trucks, or taxis) to service multiple workloads (visiting clients, traveling between locations, delivering packages). A firm or individual might own an office building or house yet stay in a hotel for various lodging and space workloads. Or a firm or individual might own a generator—that is, a private cloud with generating capacity dynamically allocated to multiple workloads such as lighting, refrigeration, and air conditioning—and be serviced by an electric utility. Firms or individuals might utilize their own private cloud of capital (equity/savings), or

2325-6095/16/$33.00 © 2016 IEEE

acquire capital on a pay-per-use basis (the interest rate) from a public cloud of capital (debt, a credit card, or a credit facility, say).

For a true, seamless hybrid cloud, workloads must be able to move from private to public, public to private, or run in both simultaneously. In the case of electricity, a transfer switch is used to switch workloads between the private cloud (a home generator) and the public cloud (the electric utility). In the case of capital, financial transfers over financial networks including SWIFT (a type of Intercloud) are used to execute transfers between accounts or pay off loans. In the case of data storage, data can be transferred across networks via mirroring, streaming data can be captured across a device gateway, or snapshots or replication can be used. And in the case of computing, workloads can be scaled and migrated via one or more technologies such as orchestration, virtual machine migration, containers, HTTP redirect, or microservices such as AWS Lambda functions.

In addition to the pure private, pure public, and hybrid scenarios, a company might use a multicloud—that is, use multiple clouds for various services, such as running applications in both AWS and Azure, or using both Salesforce.com and Zuora. Some consider a hybrid cloud a type of multicloud, others restrict "multicloud" to multiple public cloud providers. Cloud providers might evolve to use the Intercloud—that is, rely on each other for resources and services, the way airlines code-share; federate into alliances such as the Star Alliance; or rebook passengers from overbooked or cancelled flights.[5]

The Economics of Private, Public, and Hybrid

There's a widely held belief that public cloud providers must offer lower unit prices than any do-it-yourself approach could achieve, but the truth is more complicated. Moreover, it is a logical impossibility (see the sidebar), since cloud providers are themselves enterprises and it would imply scenarios such as that Google would need to use, say, AWS for its infrastructure while at the same time AWS would need to use, say, the Google Cloud, and both would somehow save money.

That said, a cloud provider might offer services or resources at a lower unit price than the unit cost a given individual or company can provide internally. For a small or medium business, or even a large enterprise that doesn't run efficiently, a public cloud might be the best or even the only feasible solution. On the other hand, large enterprises might have sufficient scale, IT infrastructure competence, and workload diversity or deferability to achieve cost structures on par with those of leading cloud providers, and be further advantaged because the correct comparison isn't between a cloud provider's cost and that of an enterprise, but between a cloud provider's price and the cost for an enterprise to do it itself. Elements contributing to the difference between a cloud provider's cost and the price of its services include sales, general and administrative costs, costs related to marketing its products and services, its profit margin, uncollectibles, legal expenses, and so forth.[6]

However, a number of savvy IT organizations have achieved documented savings by moving out of a public cloud. Zynga notably moved out of AWS and achieved a threefold savings. Instagram began to utilize a public cloud when its managed

LOGICAL PROOF THAT CLOUD PROVIDERS CAN'T ALWAYS COST LESS

Proof outline: Order the unit prices for cloud providers and unit costs for enterprises, such that P_i is the unit price (external sale price or internal cost) of a particular resource or service from the ith firm (either provider or enterprise), and such that $P_1 \leq P_2 \leq P_3 \leq \ldots \leq P_n$. Since a cloud provider can always offer lower prices than a do-it-yourself approach, some firm can offer lower prices than P_1. Without loss of generality, let it be firm 2 at price P_2. Then $P_2 < P_1$ but from the ordering $P_1 \leq P_2$, which is impossible, hence the premise is false.

services provider couldn't provision servers quickly enough. However, after its acquisition, Facebook enjoyed a fourfold savings by migrating Instagram applications to Facebook's internal IT organization. Evernote also enjoyed substantial savings by moving to private resources.

Making all the calculations even more challenging is the difficulty of comparing a given offering across cloud providers, constant price cuts, and a constant stream of new offers, including instance types (such as the recently introduced AWS X1 and T2.nano). And, before comparing with external providers, an enterprise must determine internal unit costs, which can be challenging. The data might not exist, and even if it does, allocating costs to applications can be challenging. How much corporate overhead should be allocated to a database that's used by three different applications with varying demands and different usage trends?

Moreover, the actual costs incurred when using a cloud provider are application-runtime dependent and based on a constellation of resources and services whose use can vary based on demand.

The Fundamental Equation of Cloud Economics

Assuming one can and does arrive at usable unit price/cost data (which we'll now just refer to as unit cost—that is, the cost to the customer), there are several possibilities.

If the unit cost from a cloud provider is lower than that of the enterprise, then, all things being equal (security, performance, management time, and attention), it's best to use the cloud provider.

If the unit cost from a cloud provider is identical (or roughly comparable, given margin of error, cost variability, and so on), and all other things are equal, then it's probably still best to use the cloud provider, because using the cloud provider can enhance agility to respond to new opportunities or threats and resilience in the face of unforeseen events, ranging from smoking hole disasters to unforeseen demand or business downturns.

If the unit cost from a cloud provider is higher than that of the enterprise, things are trickier than they seem, because the true choice is not between a higher and lower cost, but between a lower unit cost for a fixed-cost resource or service and a higher unit cost for a pay-per-use resource or service. Although many people seem fixated on the idea that cloud providers must offer lower cost due to economies of scale, they forget a key value driver: pay-per-use.

To state things another way, it might not matter much that the unit cost of a cloud resource or service is higher than that of a dedicated resource or service when used, because such a premium might be more than compensated for by the fact that, unlike a resource or service with a fixed cost, the cloud resource or service is free when not used.

The key determinants are the relative portion of use versus that of nonuse, which can be captured by the peak-to-average ratio of resource demand—some time-varying function $d(t)$—and how it compares to the relative cost of using a cloud provider versus doing it with a fixed-cost infrastructure. Let peak demand be represented by P (that is, $P = \max(d(t))$) and average demand by A (that is, $A = \mu(d(t))$), and let the unit cost (per unit time) of the pay-per-use cloud provider be C and the unit cost (per unit time) of the fixed-cost infrastructure be F. Let the time period of interest be T. Because the cost to use the pay-per-use cloud provider is the average demand level A times the pay-per-use rate C times the billing period or planning horizon T, and the cost to use the fixed cost infrastructure—assumed to be built to peak demand, although in practice overbuilt to have a safety margin—is the peak demand P times the unit cost of fixed-cost infrastructure F also times T, it makes sense to use the cloud whenever $A \times C \times T < P \times F \times T$. To put it another way, the cloud is a better solution than dedicated infrastructure whenever $P/A > C/F$. In other words, the cloud is the right choice whenever the peak-to-average ratio is higher than the premium paid for a pay-per-use service, which I call the *utility premium*.

Basic Analysis of Hybrid Cloud Economics

Using cloud resources or services is preferable over dedicated enterprise infrastructure whenever demand is spikier than the cloud is pricey. But those aren't the only two options. The third option is to use a hybrid cloud. Elsewhere, I show that if demand is continuous and not constant, and if there is a premium associated with cloud resources, then a hybrid architecture is always the best choice, no matter how "pricey" cloud resources or services are.[7] (The

proof relies on finding a "short enough" duration of demand by slicing off a snippet of the peak demand to make pay-per-use less expensive for that portion of the demand, even given the premium.) Here, I'll merely show that, given P, A, C, and F, where $P > A$ (that is, demand isn't constant) and $C > F$ (that is, cloud unit costs are higher), a hybrid is cheaper than either a pure cloud solution or a pure do-it-yourself solution as long as there is always nonzero demand. As we've seen, the cost of a pure do-it-yourself solution built to peak demand would be $P \times F \times T$. The cost of a pure cloud solution—driven by average demand—would be $A \times C \times T$.

Suppose we use our (assumed to be less expensive) infrastructure to address at least a small $\in$ (epsilon) of demand for the duration of T. Then we incur $\in \times F \times T$ for that demand, but we reduce our cloud spend to $(A - \in) \times C \times T$. The new total cost is clearly $\in \times F \times T + (A - \in) \times C \times T$, which is clearly $\in \times F \times T + (A \times C \times T) - (\in \times C \times T)$, which can be rewritten as $[(A \times C \times T)] + [(F - C)(\in \times T)]$. But since $C > F$, we know that $F - C$ is negative—in other words, the hybrid solution with the $\in$ of demand moved out of the cloud and into dedicated resources is cheaper, as Figure 1 shows.

An optimal balance (there might be an infinite number of minimal cost solutions, depending on the exact demand function) depends not just on P, A, C, and T, but on the exact nature of the demand function. The reason is that for "small enough" values of , given the assumption that there is always some demand, the quantity of demand shifted out of the cloud ($\in \times T$) leads to efficient utilization of fixed resources. At higher levels, though, if there is a t where $d(t) < \in$, we're back to the situation of overbuilt, and thus underutilized, fixed resources, leading to a cost structure that's higher than necessary.

Caveats

The analysis presented here overviews the cost tradeoffs between cloud services and owned/dedicated/fixed-cost infrastructure. However, there are many caveats. One is the relative cost of the two pure strategies. As noted earlier, if the cloud offers a unit price lower than the best unit cost you can achieve, a pure cloud strategy will be better than a pure do-it-yourself or hybrid approach. In this case, moving from a cloud to a hybrid solution has the unfortunate property of shifting some of the workload to a higher cost structure.

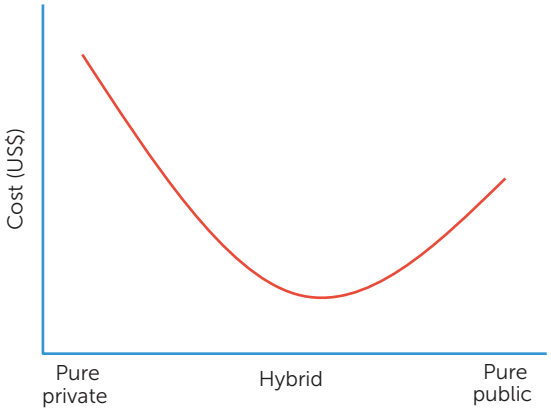

FIGURE 1. Cloud unit cost higher than fixed cost resources, variable nonzero demand.

Even if you're running a relatively efficient IT shop, though, the argument in favor of hybrids may fall short due to additional costs not captured by the simple equations above. For example, there might be additional costs for autoscaling, monitoring, management, or orchestration tools. In the best case, the cloud environment exactly matches the enterprise one, due to identical or sufficiently compatible stacks and/or technologies such as open virtualization formats or containers. If it doesn't, you might need to factor in additional costs to run a hybrid, such as developer training, overlapping development efforts, and further testing.

Another key factor is performance differences. A straight unit cost comparison assumes that the same number of units is required whether the workload is run in the cloud or on premises. However, there might be performance issues in the cloud, due to shared resources and the "noisy neighbor" problem, where heavy use of those shared resources leads to reduced efficiencies for your workload. There might also be performance advantages to having a customized infrastructure built to your application's requirements.

The reduction to practice of your hybrid architecture can create additional costs. For example, data might need to be replicated between your datacenter or a colocation facility and the cloud, leading to duplicate data storage and additional data network

or transport costs. These costs ultimately depend on the pattern you're using, a variety of which David Linthicum covers in his "Cloud Tidbits" column.[8] Making things even more complex, not only do these architectural patterns need to address security concerns, they also must address legal and data sovereignty requirements, also addressed in this issue.[9]

Finally, a focus on mere costs can obscure the true ultimate value of the cloud, including increased business agility and business resilience, cloud-based and cloud-enabled business strategies,[10] and user experience improvements due to latency reduction through the cloud's geographically dispersed architecture.[11]

HYBRID CLOUDS CAN OFFER ECONOMIC BENEFITS, EVEN WHEN—IN FACT, PARTICULARLY WHEN—THE UNIT COST OF PUBLIC CLOUD SERVICES AND RESOURCES IS HIGHER THAN THAT OF PRIVATE DEDICATED RESOURCES, A SCENARIO THAT SOME REASONABLY SIZED, WELL-RUN IT SHOPS CAN AND HAVE ACHIEVED. Such benefits arise in the presence of variable demand, in other words, for many if not most real-world computing workloads. However, additional costs, such as for hybrid cloud management, data transfer, or development complexity, must be considered.

References

1. D. Farber, "Oracle's Ellison Nails Cloud Computing," *CNET*, 26 Sept. 2008; www.cnet.com/news/oracles-ellison-nails-cloud-computing.
2. B. Golden, "What Gartner's Bimodal IT Model Means to Enterprise CIOs," *CIO.com*, 27 Jan. 2015; www.cio.com/article/2875803/cio-role/what-gartner-s-bimodal-it-model-means-to-enterprise-cios.html.
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. Cisco, "Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are," white paper, 2015; www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf.
5. J. Weinman, "Intercloudonomics: Quantifying the Value of the Intercloud," *IEEE Cloud Computing*, vol. 2, no. 5, 2015, pp. 40–47.
6. J. Weinman, *Cloudonomics: The Business Value of Cloud Computing*, John Wiley & Sons, 2012.
7. J. Weinman, "Mathematical Proof of the Inevitability of Cloud Computing," working paper, 2011. http://joeweinman.com/Resources/Joe_Weinman_Inevitability_Of_Cloud.pdf.
8. D. Linthicum, "Emerging Hybrid Cloud Patterns," *IEEE Cloud Computing*, vol. 3, no. 1, 2016, pp. 88–91.
9. C. Esposito, A. Castiglione, and K.-K. R Choo, "Encryption-Based Solution for Data Sovereignty in Federated Clouds," *IEEE Cloud Computing*, vol. 3, no. 1, 2016, pp. 12–17.
10. J. Weinman, "The Strategic Value of the Cloud, *IEEE Cloud Computing*, vol. 2, no. 5, 2015, pp. 66–70.
11. J. Weinman, "The Cloud and the Economics of the User and Customer Experience," *IEEE Cloud Computing*, vol. 2, no. 6, 2015, pp. 74–78.

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 in computer science from Cornell University and an MS in computer science from the University of Wisconsin-Madison, and has completed executive education at the International Institute for Management Development in Lausanne. He has been awarded 22 patents. Contact him at joeweinman@gmail.com.*