The Economics of Networking and the Cloud

THE NETWORK HAS A COMPLEX RELATIONSHIP WITH THE CLOUD. On the one hand, networking is a cost driver, whether for owned private cloud network infrastructure and management, or for public cloud data transfer charges or fixed network connections. However, networking is fundamental to the cloud, enabling key benefits such as statistical multiplexing of workloads from a broad set of customers into pooled, shared resources, and content delivery to a global user base. In other words, whether we want to pay for networks or not, we can't have a cloud without them. Optimizing infrastructure requires understanding all pricing elements from relevant potential cloud providers, application architecture, I/O patterns, workload characteristics, and demand variability.

Networks and the CLOUD

I've defined the cloud according to five characteristics that form a convenient acronym, CLOUD[1]:

JOE WEINMAN

joeweinman@gmail.com

- *Common*, or as the National Institute of Standards and Technology (NIST) would say, "a shared pool of configurable computing resources";
- *Location-independent*, which NIST refers to as "ubiquitous, convenient" access;
- *Online*, or "broad network access" as NIST calls it;
- *Utility*, that is, a pay-per-use charging model, which NIST calls "measured service"; and
- *On-demand* resources, that is, "on-demand self-service" with resources "that can be rapidly provisioned and released."[2]

The online criterion is a key enabler of the other characteristics. A pool of resources can't be dynamically shared unless different workloads are multiplexed into those resources over the time domain, the space domain, or both. A cloud computing datacenter is like a hotel, with different workloads from different customers checking into the dynamically shared pool of rooms. The whole concept doesn't work without networks, though—the highways and side roads that enable those customers to get to the hotel. Similarly, location independence—that is, ubiquitous, convenient access from a global set of users, which could be people, or increasingly, things—requires network connectivity between those endpoints and either a single hyperscale datacenter or, better yet, multiple geographically dispersed datacenters or other compute locations such as edge microcells.

Pay-per-use charging is at the heart of the cloud concept, although it's increasingly being complemented by a variety of payment mechanisms including reserved instances, sustained-use pricing, and dynamic pricing. However, pay-per-use charging would be an economic disaster without dynamic resource allocation. After all, one without the other would mean that resources—which entail capital expenditures or fixed operating leases on the part of a provider—would be statically allocated to a single customer, who might choose not to use them, thus not generating any return. This would be like reserving 50 seats at a restaurant and never showing up. Night after night. So, networking is also key to pay per use. And finally, on-demand resources require a local network to control resource allocation, monitor and manage resources, and securely carve up subnetworks so resources can be securely added to an existing pool.

2325-6095/16/$33.00 © 2016 IEEE

The bottom line is that networking is essential for the cloud.

However, networks enable these other criteria, and thus services and their benefits, but entail a cost. For example, a hybrid cloud comprising enterprise datacenter resources and public cloud provider resources could be used for cloudbursting during periods of peak demand, or for business continuity, with images and data for recovery in the event of a smoking hole disaster transferred to and maintained at one or more cloud providers. However, the costs of maintaining fixed network resources or utilizing pay-per-use data transport services to get data into or out of the cloud can dramatically shift the financials of such services and use cases. To put it bluntly, it doesn't help to save a nickel or two on servers or storage if it's going to cost an extra million dollars for networks.

Let's start by considering hybrid economics in the presence of a network.

Hybrid Economics

My January/February 2016 column addressed hybrid cloud economics in depth.[3] Briefly, if a public cloud provider has lower unit prices (their prices, your costs) than you can achieve yourself, it makes sense to go "all-in" with the cloud. Many IT shops will find this to be the case. However, for larger, well-run IT shops, cloud services' unit costs might be higher than an internal, do-it-yourself approach. In such a case, the variability of demand can drive a situation where the public cloud should still be a component of the overall solution. This isn't only because unit costs matter, but is also due to the cloud's "pay-per-use" nature. Dedicated resources cost money whether you use them or not, but pay-per-use cloud services only cost money when you use them. By analogy, an expensive rental car might still save you money if you only need it for a few days. The key insight is that the car might cost a lot when you rent it, but it costs you nothing when you don't. Consequently, the proportion of use to nonuse is as important as the price differential, that is, the "utility premium." Of course, the assumption is that you need to incur costs for transportation because of ultimate value generated, say, getting to work or school.

To make this clear, suppose one member of a married couple needs a car every day to go to work, but the other only needs transport on Saturdays to shop for the week. Moreover, suppose that leasing a car costs $300 per month, but renting a car costs $50 per day. A rental car is then 5 times as expensive as a dedicated solution (since $300 per month is roughly $10 per day). A fully dedicated solution would cost $140 per week, since two cars would be needed for seven days each week ($10 ∗ 7 ∗ 2). A pure cloud solution—only renting cars—would cost $400 per week ($50 per day times [7 + 1] days). But a hybrid solution is cost-optimal, at a total cost of $120 per week: $70 per week for the one dedicated car (the leased one) used every day, and $50 for the one rental car used only one day per week.

Hybrid Economics with Networking

Similar logic applies to any mix of dedicated and on-demand, pay-per-use resources. However, a comprehensive picture ultimately depends on any ancillary costs, such as for management and orchestration, and for networking. In the car example, the math changes if the less-traveled spouse needs to take a $15 cab ride from home to the rental car location and another one back. Then the dedicated solution would cost less at $140 per week, compared to the hybrid solution at $120 + (2 ∗ $15).

Of course, not all networking costs vary between pure and hybrid architectures. It might well be, for example, that user transactions cost the same whether they're load balanced to the enterprise datacenter or to the cloud. A solution's total cost will vary depending on the pattern of workload variability, the unit prices of cloud providers versus the unit costs of internal approaches, capacity selections, and therefore utilization and network costs. Moreover, these network costs ultimately depend on the nature of the application and its architecture. For example, a pure stateless application can easily be cloudburst across both dedicated enterprise resources and on-demand elastic cloud resources. Each user transaction is self-contained, so all the dataflows are between users and the cloud or between users and the datacenter, as Figure 1 shows. In such a case, assuming no difference in networking costs, only the total cost of the compute and storage resources matters in determining the optimal architecture based on total costs.

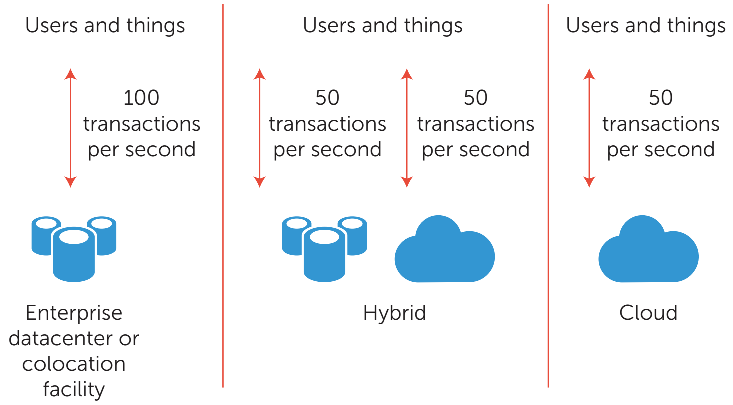
Users and things
100 transactions per second
Enterprise datacenter or colocation facility
Users and things
50 transactions per second
50 transactions per second
Hybrid
Users and things
50 transactions per second
Cloud

FIGURE 1. A stateless application with all traffic flowing between users and compute resources.

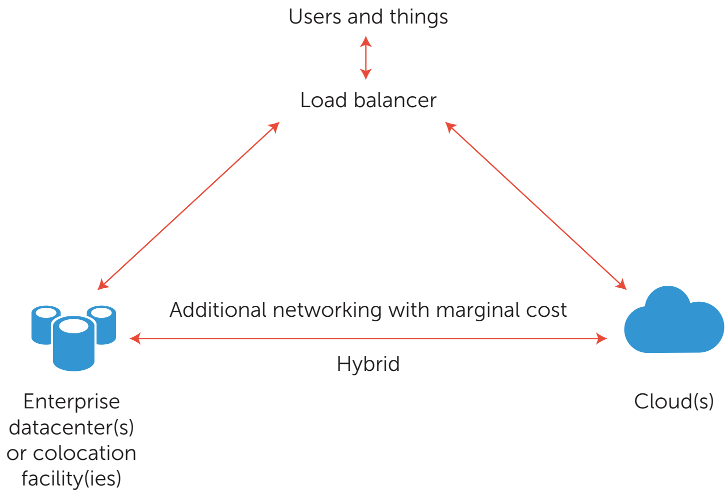
Users and things
Load balancer
Additional networking with marginal cost
Hybrid
Enterprise datacenter(s) or colocation facility(ies)
Cloud(s)

FIGURE 2. Hybrids can drive additional network costs.

At the other extreme, consider a very large database that's heavily accessed during the daytime, but hardly accessed at all at night. A very costly hybrid architecture might use an approach of spinning up cloud resources in the morning, copying the data over to the cloud, and then releasing the cloud resources in the evening after sending any updates back to the enterprise datacenters, as Figure 2 shows. Data transfer costs might more than outweigh the slight savings in compute and maybe storage that such a hybrid would achieve, just like cab fares damage the optimal lease/rent car strategy. This can tip the balance of different architectural options. For example, putting dedicated equipment in a colocation facility that also has pay-per-use cloud resources might lead to higher fixed storage and compute costs but lower pay-per-use network costs.

Real-world architectures create a mix of data transfer costs. For example, a movie streaming service might use two availability zones within a region, two regions in a given geo for reliability, linkages to a content delivery network (CDN), which might be part of the same or from a different cloud provider, and various load balancers and IP address schemes. It might back the data up to a competing cloud provider to protect against provider-wide outages, use a service such as a dedicated private line from its datacenter to a carrier-neutral colocation/interconnection facility, and then use a service such as Amazon Web Services (AWS) Direct Connect to get to a cloud provider. Depending on the provider, the service might incur charges at any or all of these points. For example, Table 1 shows what AWS charges to transfer a terabyte of data per month at the time of this writing (http://calculator.s3.amazonaws.com/index.html).

Several points are worth making. First, what might seem like arbitrary technical choices, such as how one uses IP addresses and sets up an application across availability zones or regions, will impact the total cost. Second, current pricing incents data to move to the cloud, and disincents leaving it, a smart move to accelerate cloud adoption, but one that might cause some level of regret in the future, especially since there is no guarantee that egress pricing won't increase. Third, there are numerous architectural options. At a high level, the three generic options—datacenter, hybrid, or cloud—lead to key differences in traffic levels from users or things to the compute environment, and data transfer or other networking costs between or among compute and storage nodes. Additional variations of these generic options, involving colocation and interconnection, direct connect, and so on, can also drive differences.

Also, one might observe that the monthly costs shown in the table are trivial. However, the question ultimately becomes how much traffic is incurring these charges, how prices from one cloud provider compare to another, and how they compare to a do-it-yourself approach. A video uploading and distribution service will have a different traffic and cost profile from that of a smart electric meter reading service. But numbers can add up: in one analysis, a midsized customer was reportedly able to save more than $3 million per year by moving off the cloud into a dedicated solution.[4] Obviously, as they say, mileage may vary.

Of course, the devil is in the details. For hybrid architectures that require some degree of data integration, costs can arise in a number of ways. For example, a large database might exist in an enterprise datacenter. If compute resources are spun up in the cloud, the question is how the cloud-based application component will access the data in the database. One approach might be a remote procedure call to an enterprise datacenter application or microservice, which returns the data. An advantage there is that the costs of data networking are commensurate with the (presumed revenues or benefits associated with the) number of user transactions. A related approach would be a remote database query or object or file read directly to a database or file server.

Another approach is to replicate the data from the datacenter to the cloud during the start of a cloudburst, and then synchronize updates bidirectionally. This incurs a large initial cost for the transfer, which might be done over a network or using physical disk shipping, as with AWS Snowball. After that, the cost advantages or disadvantages depend on the pattern of I/Os. If there are only a few user or thing transactions, the quantity of revenue-generating transactions won't recover a large upfront cost. On the other hand, if the application is I/O intensive—in particular, read intensive—it might be cheaper in the long run to replicate the data to the cloud up front.

THE NETWORK IS A CRITICAL ENABLING FOUNDATION FOR CLOUD COMPUTING, INCLUDING A VARIETY OF SCENARIOS SUCH AS HYBRID CLOUDS AND CLOUDBURSTING. However, network costs can shift the economics and therefore attractiveness of various cloud computing use cases. Consequently, an optimal decision requires an intimate understanding of unit costs for the enterprise datacenter, cloud provider prices, application architecture, I/O patterns, workload characteristics, and demand variability.

Table 1. Amazon Web Services data transfer charges.

Service (1,000 Gigabytes transferred)	Monthly charge
Interregion data transfer	$20.00
(Cloud) data transfer out	$88.65
(Cloud) data transfer in	Free
Virtual private cloud peering within an availability zone	$10.00
Intraregion data transfer	$10.00
Public IP address-based data transfer	$10.00

References

1. J. Weinman, *Cloudonomics: The Business Value of Cloud Computing*, John Wiley & Sons, 2012.
2. P. Mell and T. Grance, *The NIST Definition of Cloud Computing*, NIST Special Publication 800-145, National Institute of Standards and Technology, 2011; http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.
3. J. Weinman, "Hybrid Cloud Economics," *IEEE Cloud Computing*, vol. 3, no. 1, 2016, pp. 18–22.
4. S. Eschweiler, "Why Switching to AWS May Cost You a Fortune," blog, Hivelocity, 11 June 2015; www.hivelocity.net/blog/AWS-bandwidth-expensive.

JOE WEINMAN *is a frequent global 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.*