Mathematical Proof of the Inevitability of Cloud Computing[1]

January 8, 2011

Joe Weinman[2]

Permalink: http://www.JoeWeinman.com/Resources/Joe_Weinman_Inevitability_Of_Cloud.pdf

Abstract

Cloud computing represents a new model and underlying technology for IT. However, the value of cloud computing may be abstracted as the value of any on-demand utility: rental cars, taxi cabs, hotel rooms, or the like.

In a companion paper[3], the value of "on-demand" resource provisioning is quantified.

Here, the value of "utility"—i.e., pay-per-use with a linear tariff—is quantified, and three major conclusions are presented based on the nature of the offered demand and the relative cost—or "utility premium"—of the utility vs. fixed resources on a unit cost basis.

If utility resource unit costs are lower than fixed resource unit costs, that is, the utility premium is less than unity, then all demand should be resourced via the utility. For Information Technology, this would mean that cloud computing is the preferred strategy for running applications.

One might assume that if utility resource unit costs are higher than fixed resource unit costs, the reverse would be true, but it is not, due to the fact that utility resources—unlike fixed resources—are not paid for when not used. Consequently, if the peak to average ratio of the demand is higher than the utility premium, it is less expensive to fully resource such demand with a utility than with fixed resources.

Finally, if the "duration" of the peak is short enough relative to the total time period, namely less than the inverse of the utility premium, a hybrid architecture is cost optimal.

This implies that for most real-world workloads, all other things being equal, leveraging a mix of on-premises and cloud-based capacity is likely to reduce cost vs. a premises-only approach.

---

[1] Originally posted at http://cloudonomics.wordpress.com/2009/11/30/mathematical-proof-of-the-inevitability-of-cloud-computing/ by the author on November 30, 2009 (appropriately, Cyber Monday). This version has been reformatted as a pdf for readability with a new abstract and notation slightly revised to match the companion piece.[3]

[2] Joe Weinman leads Communications, Media and Entertainment Industry Solutions for Hewlett-Packard. The views expressed herein are his own. Contact information is at http://www.joeweinman.com/contact.htm

[3] "Time is Money: The Value of 'On-Demand'", http://www.JoeWeinman.com/Resources/Joe_Weinman_Time_Is_Money.pdf

1. Introduction

In the emerging business model and technology known as cloud computing, there has been discussion regarding whether a private solution, a cloud-based utility service, or a mix of the two is optimal. My analysis examines the conditions under which dedicated capacity, on-demand capacity, or a hybrid of the two are lowest cost. The analysis applies not just to cloud computing, but also to similar decisions, e.g.: buy a house or rent it; rent a house or stay in a hotel; buy a car or rent it; rent a car or take a taxi; and so forth.

To jump right to the punchline(s), **a pay-per-use solution obviously makes sense if the unit cost of cloud services is lower than dedicated, owned capacity.** And, in many cases, clouds provide this cost advantage.

Counterintuitively, though, **a pure cloud solution also makes sense even if its unit cost is higher**, as long as the peak-to-average ratio of the demand curve is higher than the cost differential between on-demand and dedicated capacity. In other words, even if cloud services cost, say, twice as much, a pure cloud solution makes sense for those demand curves where the peak-to-average ratio is two-to-one or higher. This is very often the case across a variety of industries. The reason for this is that the fixed capacity dedicated solution must be built to peak, whereas the cost of the on-demand pay-per-use solution is proportional to the average.

Also important and not obvious, **leveraging pay-per-use pricing, either in a wholly on-demand solution or a hybrid with dedicated capacity turns out to make sense any time there is a peak of "short enough" duration.** Specifically, if the percentage of time spent at peak is less than the inverse of the utility premium, using a cloud or other pay-per-use utility for at least part of the solution makes sense. For example, even if the cost of cloud services were, say, four times as much as owned capacity, they still make sense as part of the solution if peak demand only occurs one-quarter of the time or less.

In practice, this means that cloud services should be widely adopted, since absolute peaks rarely last that long. For example, today, Cyber Monday, represents peak demand for many etailers. It is a peak whose duration is only one-three-hundred-sixty-fifth of the time. Online flower services who reach peaks around Valentine's Day and Mother's day have a peak duration of only one one-hundred eightieth of the time. While retailers experience most of their business during one month of the year, there are busy days and slow days even during those peaks. "Peak" is actually a fractal concept, so if cloud resources can be provisioned, deprovisioned, and billed on an hourly basis or by the minute, then instead of peak month or peak day we need to look at peak hours or peak minutes, in which case the conclusions are even more compelling.

I look at the optimal cost solutions between dedicated capacity, which is paid for whether it is used or not, and pay-per-use utilities. My assumptions for this analysis are that pay-per-use capacity is 1) paid for when used and not paid for when not used; 2) the cost for such capacity does not depend on the time of request or use; 3) the unit cost for on-demand or dedicated capacity does not depend on the quantity of resources requested; 4) there are no additional relevant costs needed for the analysis; 5) all demand must be served without delay.

These are assumptions which may or may not correspond to reality. For example, with respect to assumption (1), most pay-per-use pricing mechanisms offered today are pure. However, in many domains there are membership fees, non-refundable deposits, option fees, or reservation fees where one may end up paying even if the capacity is not used. Assumption (2) may not hold due to the time value of money, or to the extent that dynamic pricing exists in the industry under consideration. A (pay-per-use) hotel room may cost $79 on Tuesday but $799 the subsequent Saturday night. Assumption (3) may not hold due to quantity discounts or, conversely, due to the service provider using yield management techniques to charge less when provider capacity is underutilized or more as provider capacity nears 100% utilization Assumption (4) may or may not apply based on the nature of the application and marginal costs to link the dedicated resources to on-demand resources vs. if they were all dedicated or all on-demand. As an example, there may be wide-area network bandwidth costs to link an enterprise data center to a cloud service provider's location. Finally, assumption (5) actually says two things. One, that we must serve all demand, not just a limited portion, and two, that we don't have the ability to defer demand until there is sufficient capacity available. Serving all demand makes sense, because presumably the cost to serve the demand is greatly exceeded by the revenue or value of serving it. Otherwise, the lowest cost solution is zero dedicated and zero utility resources; in other words, just shut down the business. In some cases we can defer demand, e.g., scheduling elective surgery or waiting for a restaurant table to open up. However, most tasks today seem to require nearly real-time response, whether it's web search, streaming a video, buying or selling stocks, communicating, collaborating, or microblogging.

It is tempting to view this analysis as relating to "private enterprise data centers" vs. "cloud service providers," but strictly speaking this is not true. For example, the dedicated capacity may be viewed as owned resources in a co-location facility, managed servers or storage with fixed capacity under a long term lease or managed services contract, or even "reserved instances." By "dedicated" we really mean "fixed for the time period under consideration." For this reason, I will use the terms "pay-per-use" or "utility" rather than "cloud" except when providing colloquial interpretations.

Let the demand D for resources during the interval 0 to T be a function of time

$$D(t), 0 \leq \mathrm{t} \leq T$$

This demand can be characterized by mean $\mu(D)$, which we shall simply call A, and a peak or maximum $\max(D)$ which we shall simply call P. Needless to say, based on the definitions of

mean and maximum, $A \leq P$**.** For example, the average demand A might be for 9 CPU cores, with a peak P of 31 CPU cores.

Let the unit cost per unit time of fixed capacity be c_r, and let U be the utility premium. By utility premium, I mean the multiplier for utility (pay-per-use) capacity vs. fixed. The unit cost of on-demand capacity is then $U \times c_r$. For example, c_r might be $\$2.00$ per core hour for fixed capacity. If on-demand capacity costs $3.00 per core hour, then U would be 1.5, i.e., there is a 50% premium for on-demand capacity.

To be slightly more precise, because on-demand capacity is assumed to be pure pay-per-use, in contrast to fixed capacity which is paid for whether or not it is used, there is a premium when the capacity is used, and a 100% discount when the capacity is not used. As stated above, this assumption may not be valid in all cases.

> If $U = 1$, then fixed capacity and on-demand capacity cost the same.
>
> If $U < 1$, then pay-per-use resources (e.g., the cloud) are cheaper on a unit-cost basis. It has been argued that economies of scale and statistics of scale[4] can make cloud providers' unit costs lower.
>
> If $U > 1$, then pay-per-use resources (e.g., the cloud) are assessed to be more expensive on a unit-cost basis, as at least one study claims[5]. Even under these unit cost assumptions, a pure utility or hybrid solution may be less expensive in terms of total cost, as we shall see.

Thanks to assumption (2), we can rearrange the demand curve to be monotonically non-decreasing, i.e., in ascending order, to help illustrate the points. In practical terms, this means that, for a site supporting special events, like concert or movie ticket sales, if they have a peak during 3 days each month, we can just treat it as if this peak occurred for 36 days at the end of the year. This reordering doesn't impact mean, max, or any of the calculations below, but makes it easier to understand the proofs. In the real world, such an assumption may not be the case. Continuously growing, or at least non-decreasing, demand may be suitable for resourcing via fixed capacity.

Finally, it should be noted that thanks to assumptions (2) and (3), the cost of providing utility capacity to meet the demand D is just the utility premium U times the base cost c_r times the arithmetic mean A times the duration of time T. In other words, if the price of a hotel room doesn't vary based on day or quantity–and we ignore the time value of money–then renting 8 rooms on one night and 2 rooms the next night costs the same as renting 5 rooms for two nights. This is because

[4] http://gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/
[5] http://gigaom.com/2009/04/21/why-mckinseys-cloud-report-missed-the-mark/

$$\sum_{i=0}^{T} U \times c_r \times D(t_i) = U \times c_r \times \sum_{i=0}^{T} D(t_i) = U \times c_r \times A \times T$$

2. Scenario: Utility Unit Costs are Lower Than Fixed Unit Costs

Proposition 1: If $U < 1$, that is, the utility premium is less than unity, a pure pay-per-use solution costs less than a pure dedicated solution.

Proof: The cost of the pay-per-use solution is $A \times U \times c_r \times T$. The cost of a dedicated solution built to peak is $P \times c_r \times T$. Since $A \leq P$ and $U < 1$,

$$A \times U \times c_r \times T \leq P \times U \times c_r \times T < P \times 1 \times c_r \times T = P \times c_r \times T$$

Therefore, the pay-per-use solution costs less than the dedicated solution. ■

Colloquially, the cloud total cost is advantaged due to only paying for resources when needed, as well as paying less for those resources when used.

3. Scenario: Flat Demand, Utility Unit Costs are Identical to Fixed Unit Costs

Proposition 2: If $U = 1$, that is, the utility premium is unity, and $A = P$, that is demand is flat, then a pure pay-per-use solution costs the same as a pure dedicated solution built to peak.

Proof: The cost of the pay-per-use solution is $A \times U \times c_r \times T$. The cost of a dedicated solution built to peak is $P \times c_r \times T$. Since $A = P$ and $U = 1$,

$$A \times U \times c_r \times T = P \times U \times c_r \times T = P \times 1 \times c_r \times T = P \times c_r \times T$$

Therefore, the pay-per-use solution costs the same as the dedicated solution. ■

In other words, if there is no difference between unit costs, and there is no variability in demand, it doesn't matter which strategy you use. Of course, this assumes that your demand is predictable and that there is no financial risk, neither of which is typically the case. Even if you believed this to be true, all other things being equal, you might prefer the cloud solution due to demand forecasting risk and due to financial risk, e.g., residual values being lower than projected or changes in tax laws.

4. Scenario: Variable Demand, Utility Unit Costs are Identical to Fixed Unit Costs

Proposition 3: If $U = 1$ and demand is not flat, that is, $A < P$ then a pure pay-per-use solution costs less than a pure dedicated solution.

Proof: The cost of the pay-per-use solution is $A \times U \times c_r \times T$. The cost of a dedicated solution built to peak is $P \times c_r \times T$. Since $A < P$ and $U = 1$,

$$A \times U \times c_r \times T < P \times U \times c_r \times T = P \times 1 \times c_r \times T = P \times c_r \times T$$

Therefore, the pay-per-use solution costs less than the dedicated solution.■

5. Scenario: Variable Demand, Utility Unit Costs Greater Than Fixed Unit Costs but Premium Lower Than Peak-to-Average Ratio

Interestingly, even if the unit cost of the pay-per-use utility is higher than the dedicated capacity, the total cost may be lower if the demand curve is "spiky" enough.

Proposition 4: Even if the utility premium U is greater than 1, if it is less than the peak-to-average ratio $\frac{P}{A}$, that is, $1 < U < \frac{P}{A}$, then a pure pay-per-use solution costs less than a pure dedicated solution.

Proof: Again, the cost of the pay-per-use solution is $A \times U \times c_r \times T$. The cost of a dedicated solution built to peak is $P \times c_r \times T$. Since $U < \frac{P}{A}$

$$A \times U \times c_r \times T < A \times \frac{P}{A} \times c_r \times T = P \times c_r \times T$$

Therefore, the pay-per-use solution costs less than the dedicated solution.■

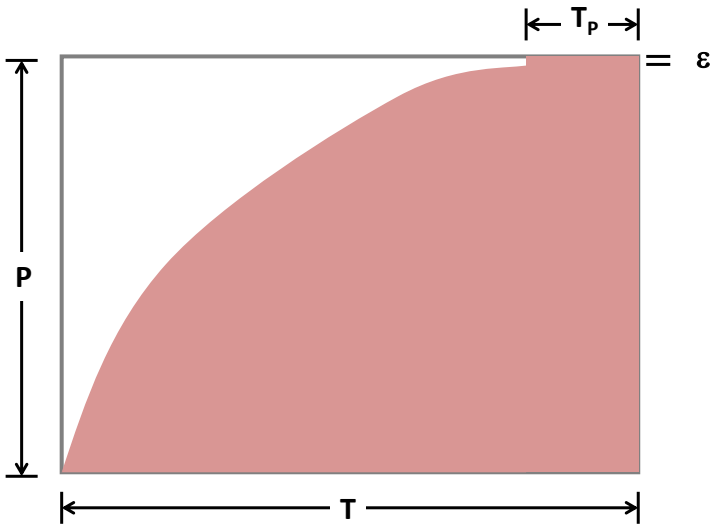
In other words, as I point out in my First Law of Cloudonomics[6], even if a utility costs more (on a unit cost basis), the total cost can be lower than a dedicated solution, because of the savings when resources are not needed due to variations in demand. The more "spiky" the demand is, the higher rate one might be willing to pay for the utility. For example, if one needs a car every

[6] http://gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/

day for years, it makes sense to own / finance / lease it, for a rate of, say, ten dollars a day. If one needs a car for only a few days, it makes sense to rent it, even though the rate might be, say, fifty dollars a day. And if one needs a car for only a few minutes, it makes sense to grab a taxi, even though paying a dollar a minute works out to an equivalent rate of over a thousand dollars per day.

6. Scenario: Total Peak Duration Less Than Inverse of Utility Premium

Let us define the total duration of the peak of the demand $D(t)$ to be T_P. That is, even if there are multiple periods when $D(t)$ is at peak, we sum them up to get T_P. This turns out to be an important criterion for determining the value of hybrid clouds.



Proposition 5: If the utility premium U is greater than 1, and $(T_P/T) < (1/U)$, that is, the percentage duration of the peak is less than the inverse of the utility premium, then a hybrid solution costs less than a dedicated solution.

Proof: Consider the cost of a hybrid solution consisting of $P - \varepsilon$ dedicated resources with any overflow handled on demand by pay-per-use capacity. Because utility resources are only required to handle the ε worth of demand, and this demand only occurs for a duration of T_P of time, the total cost to solution the demand is:

$$[(P-\varepsilon)\times T\times c_r]+[\varepsilon\times T_P\ \times c_r\times U]$$

However, since $T_P/T<1/U$, multiplying both sides by $T\times U$ we see that $T_P\times U<T$. But then,

$$[\varepsilon\times T_P\times c_r\times U]\ <\ [\varepsilon\times T\times c_r]$$

Which provides the inequality we need, namely that the total cost of the hybrid solution is

$$[(P-\varepsilon)\times T\times c_r]+[\varepsilon\times T_P\times c_r\times U]<[(P-\varepsilon)\times T\times c_r]+[\varepsilon\times T\times c_r]$$

Since $[(P-\varepsilon)\times T\times c_r]+[\varepsilon\times T\times c_r]=P\times T\times c_r$ is the cost of dedicated capacity, the total cost of the hybrid solution is less than the cost of dedicated capacity. ■

Note that $P-\ \varepsilon$ is not necessarily an *optimal* solution, it just helps to demonstrate that there is a cheaper way to do things than using dedicated resources when the peak is sufficiently short-lived. To find an optimal solution we would need to know more about the characteristics of the underlying demand curve, as we shall see below.

7. Scenario: "Long Enough" Non-Zero Demand

Conversely, let us define the total duration of non-zero demand to be T_{NZ}. That is, even if there are multiple periods when $D(t)$ is greater than zero, we sum up their durations to get T_{NZ}. This turns out to be an important criterion for determining when a hybrid architecture beats a pure cloud.

Proposition 6: If the utility premium is greater than unity and the percentage duration of non-zero demand is greater than the inverse of the utility premium, i.e.,

$$\frac{T_{NZ}}{T}>\frac{1}{U}<\ 1$$

then a hybrid solution costs less than a pure pay-per-use solution.

Proof: This proof is the mirror image of the prior one. Consider the cost of a hybrid solution consisting of ε dedicated resources with the remainder addressed by on-demand resources. The cost of serving this extra remaining demand doesn't change between the pure pay-per-use and the proposed hybrid solution, so we need only consider the differential between using a dedicated solution and a utility solution for this first ε of demand.

The cost of serving this demand with utility resources is $\varepsilon \times T_{NZ} \times U \times c_r$. The cost of serving the demand with dedicated resources is $\varepsilon \times T \times c_r$.

Since $T_{NZ}/T > 1/U$, equivalently $1/U < T_{NZ}/T$, so multiplying both sides by $T \times U$ gives us the inequality $T < T_{NZ} \times U$. Then

$$\varepsilon \times T \times c_r < \varepsilon \times T_{NZ} \times U \times c_r$$

Therefore, a hybrid solution costs less than the pure utility. ■

In other words, if there is usually some baseline demand and utilities are somewhat costly, you may as well serve the typical baseline demand with the cheaper dedicated resources and save the on-demand resources for the variable portion of the demand. As Jens Lapinski put it when commenting[7] on one of my articles, a good rule of thumb is to "own the base, and rent the spike."

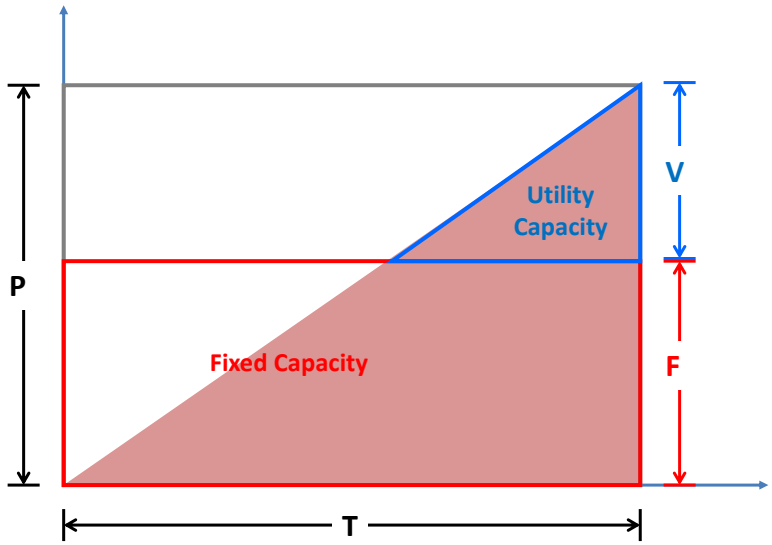
8. Optimal Hybrid Solution for Uniformly Distributed Demand Scenario

Knowing that, under the right conditions, a cost-optimal solution may be a hybrid cloud does not tell us what balance of dedicated and on-demand resources achieves the optimum balance. For that, we will solve a specific example first, then argue for the general condition.

Proposition 7: Let $D(t)$ be uniformly distributed with peak P and the utility premium $U > 1$. Then the optimal hybrid solution consists of P/U on-demand capacity and $P - (P/U)$ dedicated resources.

Proof: Let the fixed capacity be F and any demand over this amount be served by variable, on-demand pay-per-use capacity V, where $F + V = P$.

[7] http://gigaom.com/2009/04/21/why-mckinseys-cloud-report-missed-the-mark/#comments



The total cost of the solution is then the sum of the fixed cost plus the on-demand cost. The fixed cost is just $F \times T \times c_r$. The variable cost is based on the size of the triangle, which has height V. The base of the triangle is based on the proportion between V and P, namely $(V/P) \times T$. The cost, which is based on the area of the triangle, is

$$½V \times (V/P) \times T \times U \times c_r$$

so the total cost is:

$$[F \times T \times c_r] + [½V \times \frac{V}{P} \times T \times U \times c_r]$$

T and c_r are common, so this is just:

$$[T \times c_r] \times [F + ½V \times \frac{V}{P} \times U]$$

Substituting $(P - V)$ for F and simplifying terms, the total cost is

$$[T \times c_r] \times [(P - V) + ½V^2/P \times U]$$

The minimum occurs when the slope / derivative is zero. To solve this, it helps to remember that the derivative of a constant is zero, the derivative of a sum is the sum of the derivatives (as long as they exist), the derivative of x^n is nx^{n-1}, and the derivative of a constant times a function is the constant times the derivative of the function. T and c_r

are non-zero constants, so we take the derivative with respect to V and set it to zero, getting a minimum at:

$$0 = [T \times c_r] \times [0 - 1 + V \times U/P]$$

So,

$$0/[T \times c_r] = [0 - 1 + V \times U/P]$$

Then

$$0 = 0 - 1 + V \times U/P$$

Or, simplifying

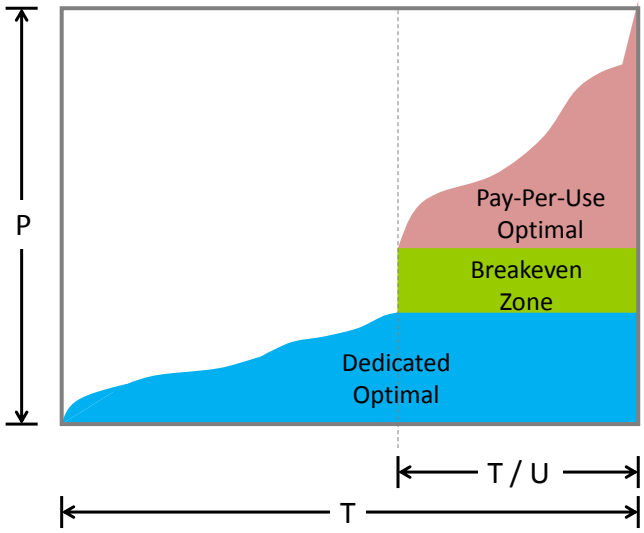
$$1 = V \times U/P$$

So the minimum occurs when $V/P = 1/U$. ■

In other words, for uniformly distributed demand, the percentage of resources that should be on-demand is the inverse of the utility premium. If there is no premium, all resources should be on-demand, if the utility premium is 2, half the resources should be on-demand, if the utility premium is 4, a quarter of the resources should be on-demand, and so forth.

It turns out that this points the way to finding an optimal hybrid solution for any demand curve. Utility simulation models[8] can be used to determine where the optimal solution lies, but the key insight is that if there is a lot of use, one may as well use dedicated resources, whereas if there is infrequent use, one should use a pay-per-use strategy. The break-even point occurs where the cost of a dedicated solution equals the cost of a pay-per-use solution, which is when the percentage of use is $1/U$. For a fixed solution, the cost to service a sliver of demand for ε resources enduring for a period T/U would be $\varepsilon \times c_r \times T$, whereas for a pay-per-use solution, the cost would be $\varepsilon \times T/U \times U \times c_r$, which is of course the same. This also means that there may not be a single optimum, but a range of optimal solutions that are equal cost because while there is a break-even point, for some curves, a "break-even zone" of a quantity of resources with the same duration can exist, and any of those resources can be assigned to dedicated or pay-per-use fulfillment without impacting the total cost.

[8] http://www.complexmodels.com/



These last few propositions show the value of hybrid resourcing strategies. If there is a short enough period of peak demand, rather than use only dedicated resources it makes sense to slice at least that out of the total solution and use on-demand pay-per-use resources to serve it. On the other hand, if there is a long enough duration of non-zero demand, you may as well use dedicated resources to serve that baseline.

9. Conclusion

So, these are the criteria for determining when pure clouds, pure dedicated solutions, or hybrid dedicated and pay-per-use solutions may be cost-optimal. The analysis above is oversimplified, since it assumes that there are no additional (marginal) costs for hybrid solutions. Whether there are or not ultimately depends on the nature of the application and the architecture implementation and cost structure, as I discuss in 4 ½ Ways to Deal with Data During Cloudbursts[9].

While, strictly speaking, this isn't proof of the inevitability of cloud computing, I've used reasonably rigorous math to determine the conditions under which cloud computing is relevant. And, because these conditions are so easily met given the demand fluctuations and price differentials seen in the real world, it means that cloud computing should be at least a part of virtually every enterprise's IT strategy.

[9] http://gigaom.com/2009/07/19/4-12-ways-to-deal-with-data-during-cloudbursts/

