To Cloud Or Not To Cloud? Musings On Costs and Viability.

Yao Chen, Radu Sion
Stony Brook Computer Science
{yaochen,sion}@cs.stonybrook.edu

ABSTRACT

In this paper we aim to understand the types of applications for which cloud computing is economically tenable, i.e., for which the cost savings associated with cloud placement outweigh any associated deployment costs.

We discover two scenarios. In an (i) "unified client" scenario, once cloud-hosted, applications are meant to be accessible only to a single cloud customer (or small set of associates). It then becomes important to ensure that the cost savings (mainly computation-related) can offset the often significant client-cloud distance (network costs etc). In a (ii) "multi-client" setting on the other hand, outsourced applications serve numerous different thir-parties. We show that then clouds begin to act similarly in nature to content-distribution networks – by comparison, their better *network integration* is simply too good to pass on, when compared to locally hosting the applications (and incurring associated network costs).

Ultimately, we hope this work will constitue a first step in an objective evaluation of the technological side of costs of outsourcing and computing in general.

1. Introduction

As computing becomes embedded in the very fabric of our society, the exponential growth and advances in cheap, high-speed communication infrastructures allow for unprecedented levels of global information exchange and interaction. As a result, new market forces emerge that propel toward a fundamental, cost-efficient paradigm shift in the way computing is deployed and delivered: computing outsourcing.

Computing outsourcing promises to minimize client-side management overheads and benefit from a service provider's global expertise consolidation and bulk pricing. More recently, first storage and then computation outsourcing has been commoditized through the emergence of globally-sized enterprises such as Google, Yahoo, Amazon, and Sun which started offering increasingly complex storage and computation outsourcing "cloud" services. CPU cycles have become consumer merchandise.

Current clouds seem to be extremely well suited and cost-effective for personal and small enterprise clients that increasingly outsource data-driven web-based retail and end-user interfaces and minimize their in-house computing management footprints. Yet, despite the associated buzz, clouds have been somewhat less successful in attracting medium to large size corporations. Cloud migration faces an array of technological (cybersecurity, scalability, transparency etc.), regulatory (shift of liability, compliance), social (novel cloud-supported interaction paradigms) and economical (cost of migration, computing as a utility) challenges that lie on the path of successful large-scale adoption.

So far, the end-to-end viability of cloud computing has mostly not been explored. Is a remotely hosted computing cycle in a large data center indeed cheaper than performing it locally *when considering the end-to-end bottom-line*? It seems the markets have spoken and the increasing number of service providers can be viewed as testimony that this indeed is the case. Yet by what margins? And what are the features of suitable applications for cloud deployment? As the migration from in-house data centers to the clouds is non-trivial and fraught with potentially large costs, asking these questions is essential.

Moreover a certain skepticism regarding large-scale commercial deployment can be increasingly heard. For example, just recently, Whitfield Diffie, one of the fathers of modern cyber-security, when asked about the prospects of achieving cloud computing security against malicious insiders said: "The whole point of cloud computing is economy: [yet, regarding security for example,] current techniques would more than undo the economy gained by the outsourcing and show little sign of becoming practical" [26]. This is particularly worrisome, since security is one of the main challenges of achieving successful cloud infrastructures.

To understand the viability of clouds, here we provide a cost model for computing in different environments and derive the dollar cost of primitives such as CPU cycles, storage and network transfers. Using the model, we then evaluate cloud outsourcing end-to-end and derive a threshold principle defining when outsourcing indeed is economically viable, i.e., when computing-related savings outweigh the costs of networking. We then evaluate the footprints and types of applications most suited for cloud deployment.

2. Cost Models

To reach the granularity of compute cycles we explore first the cost of running computing at different levels. We chose environments of increasing size: home, small, mid-size and large size data centers. The boundaries between these setups are often dynamic and the main reason we're using them is to help differentiate a set of key parameters.

2.1 Levels

Home Users (H). We include this scenario as a baseline for a simple home setup containing several computers. This could correspond to individuals with spare time to maintain a small set of computers, or a very small home-based enterprise with no staffing overheads.

Small Enterprises (S). We consider here any scenario involving an infrastructure of up to 1000 servers run in-house in a commercial enterprise. Small enterprises however can not afford custom hardware, efficient power-distribution, and cooling or dedicated buildings among others. More importantly, due to their nature, small enterprises cannot be run at high utilization as they would be usually under the incidence of business cycles and its associated peak loads.

Mid-size Enterprises (M). We consider here setups of up to 10,000 servers, run by a corporation, often in its own dedicated data center(s). They are not fully global, yet could feature several centers across one or two time zones, allowing increased independence from

Parameters	**H**	**S**	**M**	**L**
CPU utilization	5-8%	10-12%	15-20%	40-56%
server:admin ratio	N.A.	100-140	140-200	800-1000
Space (sqft/month)	N.A.	$0.5	$0.5	$0.25
PUE	N.A.	2-2.5	1.6-2	1.2-1.5

Figure 1: Sample key parameters.

local load cycles as well as the ability to handle daily peaks better by shifting loads across timezones. All the above results ultimately in increased utilization (20-25% est.) and overall efficiency.

Large Enterprises/Clouds (L). Clouds and large enterprises run over 10,000 servers, cross multiple time-zones, often literally at a global level, with large data centers distributed across all continents and often in tens to hundreds of countries. Especially in cloud setups, high speed networks allow global-wide distribution and integration of load from thousands of individual points of load. This in turn flattens the 24-hour overall load curve and allows for efficient peak handling and comparably high utilization factors (50-60% est. [15]). Cloud providers have the clout to ask vendors for custom designed hardware and power supply components [15, 18]. Moreover, these providers run the most efficient infrastructures, and often are at the forefront of innovation. Finally, clouds have access to bulk-pricing for network service from large ISPs, often one order of magnitude cheaper than mid-size enterprises [15].

2.2 Factors

We now consider the cost factors that come into play across all of the above levels. These can be divided into a set of inter-dependent vectors, including: hardware (servers, networking gear), building (floor space leasing), energy (running hardware and cooling), service (administration, staffing, software maintenance), and network service. Other breakdown layouts of these factors are possible [4].

Server Hardware. Hardware costs include servers, racks, power equipment, network equipment, cooling equipment etc. We will discuss network equipment later. We note that these costs drop with time, likely even by the time this goes to print. For example, while many of the current documented mid-size deployments use single or multi-CPU System-X blade servers at around $1-2000 each [17] and large data centers deploy custom setups at about $3000 for 4 CPUs, near-future developments could yield important changes. In one documented instance, e.g., Amazon is working with Rackable Systems to deliver an under $700 AMD-based 6 CPU board dubbed CEMS (Cooperative Expendable Micro-Slice Servers) V3. We will be conservative and empirically assume home PC prices of around $750/CPU, small and mid-size enterprise costs of around $1000/CPU (for 2 CPU blades) and cloud-level costs of no more than $500/CPU.

Energy. Energy in data centers does not only include power, computing and networking hardware but the entire support infrastructure, including cooling, physical security, and overall facilities. A simple rough way to infer power costs is by estimating the Power Usage Efficiency (PUE) of the data center. The PUE is a metric defined by the GreenGrid Consortium to evaluate the energy efficiency of a data center [13] (PUE = Total Power Usage / IT Equipment Power Usage). We will assume 1.2-1.5 PUE for large enterprises, 1.6-2 PUE for mid-size enterprises and 2-2.5 for small enterprises [20]. Costs of electricity are relatively uniform and documented [3].

Service. Evaluating the staffing requirements for data centers is an extremely complex endeavor as it involves a number of components such as software development and management, hardware repair, maintenance of cooling, building, network and power services [4].

Analytical approaches are challenged by the sparsity of available relevant supporting data sets. We deployed a set of commonly accepted rule of thumb values that have been empirically developed and validate well [16]: the server to administrator ratio varies from 2:1 up to experimental 2500:1 values [16] due to different degrees of automation and data management. In deployment, small to mid-size data centers feature a ratio of 100-140:1 whereas cloud level centers can go up to 1000:1 [12, 15].

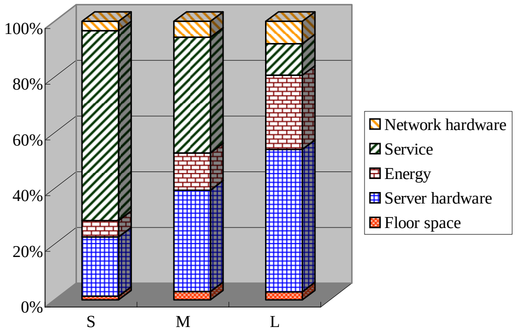

Figure 2: Impact of cost factors.

Network Hardware. Internal network infrastructure costs can be estimated by evaluating the number of required switches and routers. The design of scalable large economy network topology with high inter-node bandwidth for data centers is an ever ongoing research problem [21]. We base our results on some of the latest state of the art research, deploying fat tree interconnect structures. Fat trees have been shown to offer significantly lower overall hardware costs with good overall connectivity factors.

Floor Space. Floor space costs vary wildly, by location and use. While office space can be had for up to tens of dollars/sqft/month in Manhattan, data center space can be had at much lower rates, being as low as $0.1/sqft/month [8, 9, 22]. While small to mid-size enterprises usually have data centers near their location (thus sometimes incurring office-level pricing), large companies such as Google and Microsoft tend to build data centers on owned land, in less populated place where the per sqft price can be brought down much lower, often amortized to zero over time [10, 19].

We also note that floor surface is directly related to power consumption and cooling with designs supporting anywhere from 40 to 250 watt/sqft [11]. Thus, the overall power requirements (driven by CPUs) impact directly the required space.

3. Costs

Armed with knowledge of the above factors, we now estimate the cost of basic computing primitives.

3.1 CPU Cycles

Symbol	Definition
N_s, N_w	number of servers,switches
α	administrator : server ratio
β	watt per sq ft
λ_s, λ_w	server,switch price
λ_p, λ_f	personnel,floor cost/sec
λ_e	electricity price/(watt·sec)
μ	CPU utilization
ν	CPU frequency
τ_s, τ_w	servers,switches lifespan (5 years)
w_p, w_i	server power at peak,idle

Figure 3: Notations for (1).

We start by evaluating the amortized dollar cost of a CPU cycle in equation (1). See notations in Figure 3 and various setups' parameters in Figure 1.

The results are depicted in Figure 5, costs ranging from 0.45 picocents/cycle in very large cloud settings all the way to (S), the costliest environment, where a cycle costs up to 27 picocents (*1 US picocent* = $1 $\times 10^{-14}$). We validate our results by exploring the pricing of the main cloud providers (Figure 4). The prices lie surprisingly close to each other and to our estimates, ranging from 0.93 to 2.36 picocents/cycle. The difference in cost is due to the fact that

$$
\begin{aligned}
CycleCost &= \frac{Server + Energy + Service + Network + Floor}{Total\ Cycles} \\
&= \frac{\lambda_s \cdot N_s/\tau_s + (w_p \cdot \mu + w_i \cdot (1-\mu)) \cdot PUE \cdot \lambda_e + \frac{N_s}{\alpha} \cdot \lambda_p + \lambda_w \cdot N_w/\tau_w + \lambda_f \cdot \frac{(w_p \cdot \mu + w_i \cdot (1-\mu)) \cdot PUE}{\beta}}{\mu \cdot \nu \cdot N_s}
\end{aligned} \tag{1}
$$

these points include not only CPUs but also intra-cloud networking, instance-specific disk storage and cloud providers' profit.

3.2 Network Service

Provider	Picocents
Amazon EC2	0.93 - 2.36
Google AppEngine	up to 2.31
Microsoft Azure	up to 1.96

Figure 4: CPU cycle costs.

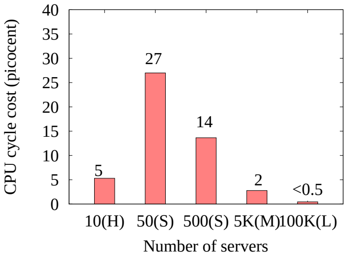

Figure 5: CPU cycle costs

Published numbers place network service costs for large data centers at around $13/ Mbps/ month and for mid-size setups at $95/ Mbps/ month [15] for *guaranteed* bandwidth. Similar pricing schemes have been quoted to us by network providers [6]. Home user and small enterprise pricing usually benefits from economies of scale and numbers are readily available, e.g., Optimum Online provides 15/5 Mbps internet connection for small business starting at $44.9/month. We note however that the quoted bandwidth is not guaranteed and refers only to the hop connecting the client to the provider. Figure 6 summarizes these costs.

	H, S	M	L
monthly	$44.90	$95	$13
bandwidth (d/u)	15/5 Mbps	per 1Mbps	per 1Mbps
dedicated	No	Yes	Yes
picocent/bit	115/345	3665	500

Figure 6: Summarized network service costs [15, 23].

The end-to-end cost of network transfer includes the cost on both communicating parties and the CPU overheads of transferring a bit from one application layer to another (a minimum about 20 CPU cycles per 32 bit data). Moreover, for reliable networking (e.g., TCP/IP) we need to also factor in the additional traffic and spent CPU cycles (e.g., SYN, SYN/ACK, ACK, for connection establishment, ACKs for sent data, window management, routing, re-transmissions, etc). If we assume a 1% TCP re-transmission rate, 1 ACK packet for every two data packets, it costs more than 900 picocents to transfer one bit reliably in the $S \rightarrow L$ scenario. We summarize the per bit transfer cost in other scenarios in Figure 8.

Per bit transfer cost	
(H, S) → Cloud	900
(M) → Cloud	4,500

Figure 8: Transfer costs.

Moreover, if the applications are not optimized to fully utilize payloads these costs could be much higher, e.g., if only a 32 bit value payload is sent, it would incur upwards of 10,000 picocents per bit.

3.3 Storage

Simply storing bits on disks has become truly cheap. Increased hardware reliability (with mean time between failures rated routinely above a million hours even for consumer markets) and economies of scale resulted in extreme drops in the costs of disks. Figure 7 shows the costs of ownership and operation of a representative sample (by no means exhaustive) set of commonly available consumer-level disks (numbers were obtained in November 2009 from numerous online sources, including the disk vendors' sites, price search engines and independent online hardware discussion sites). Costs incorporate energy and amortized acquisition components. Energy is dominating at 60-70% of the total cost. We note that actual observed MTBF are often up to about 3.4 times lower than advertised [25]. We considered this in computing the values in Figure 7.

In terms of amortized acquisition costs, the Seagate Barracuda provides the best price/hardware/MTBF ratio at 7.67 picocents/bit/year. We observe that hardware constitutes only a small percentage of the overall costs, e.g., for the Maxtor, the amortized hardware acquisition being only 12.16% of the overall ownership cost. And it holds across all considered (H,S,M,L) levels due to the fact that the existence of a critical mass of disk consumer level buyers results in economies of scale pricing available for everybody.

This leads to the insight that, if storage power and maintenance has been already factored in, then, for most scenarios direct storage hardware costs are very small and *can be mostly ignored when evaluating network and CPU intensive protocols.* Naturally this does not hold if the main costs include long-term data at rest with little or no computation and networking. But, as soon as data gets transferred or processed, direct storage costs become negligible.

4. To Or Not To

The insights gained above in the costs of computation, network and storage enable us to explore the viability of the outsourcing endeavor.

We start by noting that it is easy to find scenarios for which it does *not* make sense to outsource to clouds from a strict cost-centric perspective. For example, the CPU cycle costs in Figure 5 immediately show that it is not profitable to outsource personal workloads (H) to small (S) enterprises (we denote this $H \rightarrow S$) as it would naturally incur additional network bandwidth and CPU cycle costs are much higher for (S).

Yet, what about the other options, $\{H \rightarrow M$, $H \rightarrow L$, $S \rightarrow M$, $S \rightarrow L$, $M \rightarrow L\}$?

The answer in each of these cases is highly dependent on the type of applications outsourced. Basically, there are three main services the cloud provides: storage, networking and computation. The costs of these three primitives behave differently across computing environments of different scale, thus their outsourcing costs are different. Often the relation between these primitives in a application determines its outsoucing saving. In the following, we explore applications of different types in two outsourcing scenarios (single-client outsourcing and multi-client outsourcing).

4.1 Single-Client Model

One of the simplest computation outsourcing scenarios involves clients shifting their *own* CPU-intensive applications onto clouds, to save costs. Later *these same clients* (or delegates thereof) will access these cloud-hosted applications for their own use. An example of this are *large corporations considering migrating internal data centers to clouds*.

Naturally, this is feasible when the savings outweigh the outsourcing overhead costs. In general, outsourcing a computation load from environment a to environment b is economically justified when

$$
\begin{aligned}
&Savings = Cycles \times c_a - Cycles \times c_b - Trans_{a \rightarrow b} \geq 0 \\
&\Leftrightarrow Cycles \geq \frac{Trans_{a \rightarrow b}}{c_a - c_b}
\end{aligned} \tag{2}
$$

where $Cycles$ is the number of CPU cycles needed per bit data, and c_x denotes the CPU cycle cost for environment $X \in \{H, S, M, L\}$.

Disk	cap.	price	Adj. MTBF	amort. acq.	power	power2	power3	power cost	total cost	acq. %	avg. seek	avg. seek4	power5	read cost
	(GB)	(USD)	(mil.hrs)	(pcent/bit/yr)	seek (W)	idle (W)	(W)	(pcent/bit/yr)	(pcent/bit/yr)		time (ms)	cost (pcents)	read (W)	(pcent/bit)
Maxtor Diamond Max	500	53	0.35	32.89	13.6	8.10	10.85	237.62	270.50	12.16	9.00	377542	11.16	0.03
Hitachi Deskstar 7k500	500	67	0.29	49.89	15	9.60	12.30	269.37	319.26	15.63	8.50	407953		
Hitachi Ultrastar A7K1000	1024	153	0.35	46.36	14	9.00	11.50	122.97	169.33	27.38	8.20	417631		
WD Caviar GP Low Power	1024	103	0.29	37.45	7.5	4.00	5.75	61.49	98.93	37.85	8.90	271994	7.40	0.02
Seagate Barracuda 7200.10	750	63	0.35	26.06	12.6	9.30	10.95	159.87	185.93	14.02	9.25	369615	13.00	0.06
WD Caviar SE16	500	62	N/A		8.77	8.40	8.59	188.01			9.90		8.77	0.04
Samsung SSD	32	269	0.29	3129.65	1	1.00	1.00	342.19	3471.83	90.14	1.70	47912	0.5	0.0017
Intel SSD X18-M	80	389	0.35	1508.59	0.15	0.06	0.11	14.37	1522.96	99.06			0.15	0.0002
Intel SSD X25-M	160	765	0.35	1483.38	0.15	0.06	0.11	7.19	1490.57	99.52			0.15	0.0002

Figure 7: Magnetic disk storage costs.

We call this the *first minimal CPU-intensive requirement criterion* (we will also call this the "first outsourcing criterion"):

> **First outsourcing criterion:**
>
> For an application accessed mainly by clients in environment a, outsourcing it from a to another environment b is economically justified iff. its computation load exceeds $\frac{Trans_{a\rightarrow b}}{c_a - c_b}$ compute cycles per transferred input bit.

To illustrate, consider a 32 bit item in the $S \rightarrow L$ case. We know from Section 3.2, that the cost of reliably transferring 32 bits can be anywhere 28,000 and 320,000 picocents depending on the nature of the connection and whether connection establishment costs are amortized across multiple sends. For consistency, we disregard for now any application-specific costs, such as the existence of results and their transfer costs. As a lower bound, we get

$$Cycles \geq \frac{Trans_{S\rightarrow L}}{c_S - c_L} \in (1000, 12000).$$

In other words, *if the task at hand requires anywhere less than 1,000 CPU cycles (in the most optimized possible case) per 32 bits of input data, it is not profitable to outsource from a home setting to a large cloud.*

Moreover, $1,000$ turns out to also be a lower bound across all outsourcing options as can be seen in Figure 9. For $H \rightarrow L$, we have anywhere between $Cycles > 6400$ and $Cycles > 71,000$. For $M \rightarrow L$, due to the much higher network costs of (M), 32 bit transfers can cost anywhere between 144,000 and 1,615,000 picocents, which results in anywhere between $Cycles > 96,100$ and $Cycles > 1,070,000$.

Applications which are well suited in such CPU-intensive outsourcing may include highly scientific computations [24], which usually consume large amounts of CPU. We note that recently Mathworks seems to have tapped this niche, by adding a parallel toolbox in Matlab which enables users to do parallel computing on the Amazon Elastic Compute Cloud [2].

We note that the above *minimal CPU-intensive requirement* criterion specifically refers to network costs that cannot be amortized over multiple transactions, hence the wording "per *transferred* input bit". Yet, often applications involve significant amounts of already cloud-hosted data inputs, and in such cases, the criterion simply refers to any data that is transfered to/from the cloud.

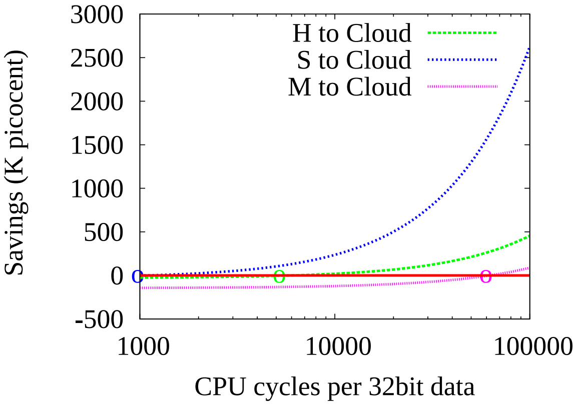

Figure 9: Cost savings of outsourcing per 32bit data from $S \rightarrow L$, $H \rightarrow L$, $M \rightarrow L$ with increasing application computation load. The lower bounds on the numbers of CPU cycles needed to justify cloud outsourcing are 1,000, 6,400, and 96,100 respectively.

Simple Storage. Overall, the CPU-intensive requirement of the criterion suggests that purely storage-centric applications are not good candidates for unified-client outsourcing in the cloud. This indeed seems to hold for simple storage outsourcing in which a single data customer places data remotely for future access. For the $S \rightarrow L$ scenario, the amortized cost of storing a bit reliably either locally *or remotely* is under 9 picocents/month (including power). Network transfer however, is of at least 900 picocents per accessed bit, a cost that is not amortized and two orders of magnitude higher.

Thus, from a pure technological cost-centric point of view, it is simply not effective to store data remotely. Depending on the application network footprint, *amortized outsourced storage costs can be upwards of 2+ orders of magnitude higher than local storage.*

Searchable Storage and Databases. Scenarios where outsourcing of data becomes viable include any data processing mechanisms that allow the amortization of networked data transfer over multiple queries to the data set.

Consider for example a searchable outsourced database of size n which allows queries of certain search selectivity s (search results are of size $n * s * S_r$, where S_r is the size of a single result) to be submitted. In this case, the intuition dictates that outsourcing is profitable for a CPU-intensive search process (e.g., for a large database size) and a high selectivity (very low s). For illustration, if searching involves a binary index (O($\log n$) CPU cycles), and a comparison takes $C_{compare} = 3$ cycles, we have

$$Savings = \log n \times C_{compare} \times (c_a - c_b)$$
$$Cost_{trans} = nsS_rTrans_{a\rightarrow b},$$

and, for cost viability, we want

$$\log n \times C_{compare} \times (c_a - c_b) \geq nsBTrans_{a\rightarrow b}$$
$$\Leftrightarrow s \leq \frac{\log n \times C_{compare} \times (c_a - c_b)}{nS_rTrans_{a\rightarrow b}}$$

In the $S \rightarrow L$ scenario, for a database of $n = 10^9$ keywords and $S_r = 32$ bits, this results in $s \leq 8.3 \times 10^{-11}$. And s will be even lower when database size grows.

4.2 Multi-Client Model

Yet, paradoxically, despite the above conclusion, storage outsourcing seems to be thriving. Just recently, Smugmug, a paid digital photo sharing website, announced $1M savings a year by outsourcing storage to Amazon S3 [1].

This can be explained as follows. The core storage costs coupled with the lack of a intense-enough CPU load, indeed do not justify outsourcing for a unified client scenario. Yet, web-based enterprises such as Smugmug, by their very nature provide services to third party clients and thus also require mechanisms to handle their clients' remote access, e.g., through often CPU-intensive web interfaces supported by web servers running on actual CPUs. This can increase the per-bit CPU footprint significantly. Moreover, network service pricing for mid-size enterprises can be up to one order of magnitude higher than for clouds, as can be seen in Figure 6 – and in effect, clouds can afford to also operate as an efficient content distribution (CDN) service.

Overall, the case for cloud feasibility becomes more complicated in multi-client scenarios. The outsourcing criterion needs to be updated as a function also of the different network service deals of the two environments. Then, outsourcing is economically tenable when

$$Cycles \times c_a - Cycles \times c_b + (Trans_{c\rightarrow a} - Trans_{c\rightarrow b}) \geq 0 \tag{3}$$

where c is the environment from which the majority of client accesses are coming to the outsourced application (Figure 10).

Then, the outsourcing criterion can be rewritten into a more complete ("second outsourcing criterion") form as follows:

Second outsourcing criterion:

For an application that resides in environment a, whose accesses come mainly from clients in environment c, outsourcing it from a to another environment b is economically justified iff.

its computation load exceeds $\frac{Trans_{c\rightarrow b} - Trans_{c\rightarrow a}}{c_a - c_b}$ compute cycles per transferred input bit – for $c_a \geq c_b$ and $Trans_{c\rightarrow a} \leq Trans_{c\rightarrow b}$, or,

its computation footprint is lower than $\frac{Trans_{c\rightarrow a} - Trans_{c\rightarrow b}}{c_b - c_a}$ compute cycles per transferred input bit – for $c_a \leq c_b$ and $Trans_{c\rightarrow a} \geq Trans_{c\rightarrow b}$

We can better understand equation (3) by detailing the following four cases:

- (i) $c_a \geq c_b$ and $Trans_{c\rightarrow a} \geq Trans_{c\rightarrow b}$, in this case, savings are constantly positive, yielding no *CPU intensive requirement*;
- (ii) $c_a \leq c_b$ and $Trans_{c\rightarrow a} \leq Trans_{c\rightarrow b}$,no savings can be achieved (constantly negative);
- (iii) $c_a \geq c_b$ and $Trans_{c\rightarrow a} \leq Trans_{c\rightarrow b}$, then $Cycles \geq \frac{Trans_{c\rightarrow b} - Trans_{c\rightarrow a}}{c_a - c_b}$
- (iv) $c_a \leq c_b$ and $Trans_{c\rightarrow a} \geq Trans_{c\rightarrow b}$, in this case, $Cycles \leq \frac{Trans_{c\rightarrow a} - Trans_{c\rightarrow b}}{c_b - c_a}$, this unusual case corresponds to an *upper bound* on the amount of computation an application can have before outsourcing becomes counter-productive;

Note, that, given today's cost points, *outsourcing from smaller scale to larger scale environments is always profitable* (as can be seen also in Figure 10) and falls into case (i) (except for $H \rightarrow S$).

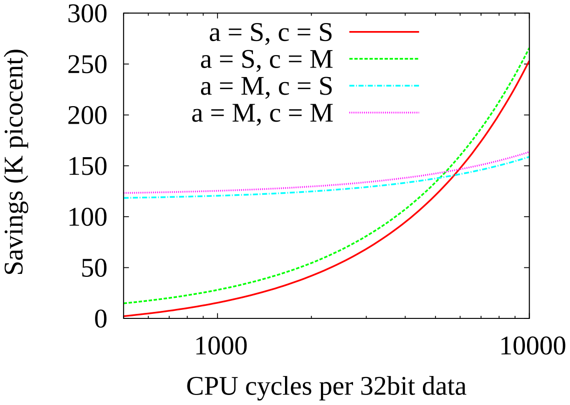

Figure 10: Illustration of the cost savings of outsourcing per 32 bit of data from $a \in \{S, M\}$ to $b = L$ with $c \in \{S, M\}$ – with increasing computation load – according to equation (3) (corresponding to the second outsourcing criterion). It can be seen that for today's pricing points, it is *always* economically desirable to outsource to clouds when third party clients are involved.

However, for completeness, the equation also covers cases when outsourcing occurs from larger to smaller scale environments, as in (iv). One illustrative instance of this is a large enterprise placing smaller data centers strategically closer to targeted clients. Although CPU cycles will cost more in these smaller data centers, this kind of outsourcing can effectively take advantage of its associated network proximity.

This illustrates another point of feasibility for clouds: content distribution for applications with numerous (often geographically dispersed) clients. This is not only profitable because of the better network service deals that clouds get from major ISPs, but also due to their on-demand scalability promise etc., which is outside of the scope of this paper.

	Amazon	Microsoft	Google
Data-in	1164	1164	1164
Data-out - first 10TB/mo	1979	1746	1396
next 40TB/mo	1513	1746	1396
next 100TB/mo	1280	1746	1396
next 150TB/mo	1164	1746	1396
intra-cloud/same region	0	0	0
intra-cloud/inter-region	116	N/A	N/A

Figure 11: Inter- and intra-cloud network transfer pricing (picocent).

For multi-client applications such as content distribution or data processing, it is important to consider also intra-cloud communication as well as the actual profit-including pricing of bit transfers in/out of clouds. For example, at the time of this writing, clouds charge 1164 picocents per incoming bit, roughly double than what they are paying to ISPs. Figure 11 illustrates these pricing points.

5. Related Work

There are numerous discussions about economics in cloud computing online, e.g., J. Hamilton's blog [14] provides in-depth analysis of cloud infrastructure cost and power efficiency. However, very few papers have been published to systematically quantify the viability of cloud computing. Here we discuss two technical reports related to this topic.

Armbrust et al. [5] define the term "cloud computing", and point out three essential aspects: the illusion of infinite computing resources available on demand, the elimination of an up-front commitment by its users, and the ability to pay for use of computing resources on a short-term basis as needed. The paper also discusses basic cost considerations for cloud deployment of a set of considered applications.

McKinsey [7] shows that cloud computing is most attractive for small and medium sized enterprises, but there are significant hurdles to its adoption by large enterprises; also, current offerings are not cost effective compared to large enterprise data centers. They evaluate the Total Cost of Assets (TCA) of a typical data center[1] at $45 per CPU equivalent – most EC2 options are comparatively costlier. The main flaw of the report is the lack of accurate consideration of network service costs (significant as we saw) as well as any of the other important factors discussed above.

It is important to note also that the above reports consider retail prices as a baseline for computing cost. In our paper, we show that it is important to explore deeper and isolate the technological discourse from market fluctuations and other factors of the moment.

Furthermore, by just looking at the prices, we lose important information we need in quantifying viability. For example, cost breakdowns vary in the cloud and smaller environments – this makes a difference when evaluating deployment feasibility of different applications (e.g., network vs. CPU vs. storage intensive applications).

6. Conclusion

We mused on the feasibility of cloud computing from a technological cost-centric point of view. We started by giving a cost model for computation, storage and networking in different environments. We saw that CPU cycles cost no less than 0.45 picocents, a bit cannot be transferred without paying at least 900 picocents, and stored a year without a pocket setback of at least 100 picocents. We validated the cost model with today's pricing points of clouds.

We determine two "outsourcing criteria", defining the boundary condition of cloud migration viability. The "first outsourcing criterion" considers unified client applications and postulates that, from a technological cost-centric perspective, outsourcing them *is profitable for computation intensive tasks*, specifically, *when its (mostly computation-related) cost savings are sufficient to offset client-cloud network distances*. This happens today for unified client applications requiring *no less than 1000 CPU cycles per each 32 bits of client-cloud transferred input*.

In the case of applications with third-party clients, the feasibility equation changes dramatically. The "second outsourcing criterion" postulates that, *for today's pricing points, it always makes sense to outsource to larger scale environments!* This is mainly because of the dominating costs of networking, and the fact that in the unified client model, the comparison baseline would not include any networking costs (as the data would be accessed locally).

We investigated here only the purely cost-centric technological aspect of clouds, yet a large number of other enablers of cloud migration are to be considered including costs of opportunity, on-demand pay-as-you-go scalability, transparency, increased availability and global distribution, among others.

7. References

[1] Amazon s3: Show me the money. http://blogs.smugmug.com/don/2006/11/10/amazon-s3-show-me-the-money/.

[2] Parallel computing with matlab on amazon elastic compute cloud (ec2). http://www.mathworks.com/programs/techkits/ec2-paper.html.

[3] Energy Information Administration. "average retail price of electricity to ultimate customers by end-use sector, by state". Online at http://www.eia.doe.gov/cneaf/electricity/epm/table5_6_a.html.

[4] APC. Determining total cost of ownership for data center and network room infrastructure. Online at http://www.apcmedia.com/salestools/CMRP-5T9PQG_R3_EN.pdf.

[5] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. Above the clouds: A berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[6] Personal Communication. "inquiry with optimum lightpath costumer service, november, 2008".

[7] McKinsey & Company. "clearing the cloud".

[8] Katherine Conrad. Data centers hot once again in the bay area. Online at http://findarticles.com/p/articles/mi_qn4176/is_20070401/ai_n18782997.

[9] Qwest Communications Corp. Space furniture rental. Online at http://www.qwest.com/about/policy/docs/qcc/documents/WO-sfr-Amd52_11100%6.pdf.

[10] Economie. Google wil energie eemshaven heeft het. Online at http://www.trouw.nl/nieuws/economie/article1247225.ece, Dec. 2007.

[11] Janice Fetzer. Internet data centers:end user & developer requirements. Online at http://www.utilityeda.com/Summer2006/Mares.pdf.

[12] Albert Greenberg, James Hamilton, David A. Maltz, and Parveen Patel. The cost of a cloud: Research problems in data center networks. In *SIGCOM Computer Communications Review*, 2009.

[13] The Green Grid. Green grid metrics: Describing data center power efficiency. Online at http://www.thegreengrid.org/gg_content/Green_Grid_Metrics_WP.pdf.

[14] James Hamilton. Perspectives Blog. Online at http://mvdirona.com/jrh/work/.

[15] James Hamilton. Internet-scale service efficiency. Large Scale Distributed Systems & Middleware (LADIS 2008),, 2008.

[16] James Hamilton. On designing and deploying internet-scale services. Technical report, Windows Live Services Platform, Microsoft, 2008.

[17] IBM. IBM blade servers. Online at http://www-03.ibm.com/systems/bladecenter/hardware/servers/.

[18] Om Malik. And now google is making its own 10-gigabit switches. Online at http://gigaom.com/2007/11/18/google-making-its-own-10gig-switches/, 2007.

[19] Om Malik. Googlenet going global. Online at http://gigaom.com/2007/09/21/googlenet-going-global/, 2007.

[20] Rich Miller. Microsoft: Pue of 1.22 for data center containers. Online at http://www.datacenterknowledge.com/archives/2008/10/20/microsoft-pue-of%-122-for-data-center-containers/.

[21] Al-Fares Mohammad, Loukissas Alexander, and Vahdat Amin. A scalable, commodity data center network architecture. *SIGCOMM Comput. Commun. Rev.*, 38(4):63–74, 2008.

[22] Department of Administration. Records management fact sheet 13. Online at http://www.doa.state.wi.us/facts_view.asp?factid=68&locid=2.

[23] Optimum. Optimum online plans. Online at http://www.buyoptimum.com.

[24] J. J. Rehr, J. P. Gardner, M. Prange, L. Svec, and F. Vila. Scientific computing in the cloud, 2008.

[25] Bianca Schroeder and Garth A. Gibson. Disk failures in the real world: what does an mttf of 1,000,000 hours mean to you? In *FAST '07: Proceedings of the 5th USENIX conference on File and Storage Technologies*, Berkeley, CA, USA, 2007. USENIX Association.

[26] Whitfield Diffie. How Secure Is Cloud Computing? Online at http://www.technologyreview.com/computing/23951/, November 2009.

[1] A typical data center: 10% utilization, $20M/MW for facility, $.1kW-hour, $14K/Server (2 CPU, 4 core)