Do Clouds Compute?
A Framework for Estimating the Value of Cloud Computing

Markus Klems, Jens Nimis, Stefan Tai
FZI Forschungszentrum Informatik Karlsruhe, Germany
{ klems, nimis, tai } @fzi.de

1 Introduction

On-demand provisioning of scalable and reliable compute services, along with a cost model that charges consumers based on actual service usage, has been an objective in distributed computing research and industry for a while. Cloud Computing promises to deliver on this objective: building on compute and storage virtualization technologies, consumers are able to rent infrastructure "in the Cloud"as needed, deploy applications and store data, and access them via Web protocols on a pay-per-use basis.

In addition to the technological challenges of Cloud Computing there is a need for an appropriate, competitive pricing model for infrastructure-as-a-service. The acceptance of Cloud Computing depends on the ability to implement a model for value co-creation. In this paper, we discuss the need for valuation of Cloud Computing, identify key components, and structure these components in a framework. The framework assists decision makers in estimating Cloud Computing costs and to compare these costs to conventional IT solutions.

2 Objective

The main purpose of our paper is to present a basic framework for estimating value and determine benefits from Cloud Computing as an alternative to conventional IT infrastructure, such as privately owned and managed IT hardware. Our effort is motivated by the rise of Cloud Computing providers and the question when it is profitable for a business to use hardware resources "in the Cloud". More and more companies already embrace Cloud Computing services as part of their IT infrastructure [1]. However, there is no guide to tell when outsourcing into the Cloud is the way to go and in which cases it does not make sense to do so. With our work we want to give an overview of economic and technical aspects that a valuation approach to Cloud Computing must take into consideration.

Valuation is an economic discipline about estimating the value of projects

and enterprises [2]. Corporate management relies on valuation methods in order to make reasonable investment decisions. Although the basic methods are rather simple, like Discounted Cash Flow (DCF) analysis, the difficulties lie in appropriate application to real world cases.

Within the scope of our paper we are not going to cover specific valuation methods. Instead, we present a generic framework that serves for cost comparison analysis between hardware resources "in the Cloud"and a reference model, such as purchasing and installing IT hardware. The result of such a comparison shows the value of Cloud Computing associated with a specific project and measured in terms of opportunity costs. In later work the framework must be fleshed out with metrics, such as project free cash flows, EBITDA, or other suitable economic indicators. Existing cost models, such as Gartner's TCO seem promising candidates for the design of a reference model [3].

3 Approach

A systematic, dedicated approach to Cloud Computing valuation is urgently needed. Previous work from related fields, like Grid Computing, does not consider all aspects relevant to Cloud Computing and can thus not be directly applied. Previous approaches tend to mix business objectives with technological requirements. Moreover, the role of demand behavior and the consequences it poses on IT requirements needs to be evaluated in a new light. Most important, it is only possible to value the benefit from Cloud Computing if compared to alternative solutions. We believe that a structured framework will be helpful to clarify which general business scenarios Cloud Computing addresses.

Figure 1 illustrates our framework for estimating the value of Cloud Computing. In the following, we describe in more detail the valuation steps suggested with the framework.

3.1 Business Scenario

Cloud Computing offers three basic types of services over the Internet: virtualized hardware resources in form of storage capacity and processing power, plus data transfer volume. Since Cloud Computing is based on the idea of Internet-centric computing, access to remotely located storage and processors must be encompassed with sufficient data transfer capacities.

The business scenario must specify the business domain (internal processes, B2B, B2C, or other), key business objectives (cost efficiency, no SLA violations, short time to market, etc.), the demand behavior (seasonal, temporary spikes, etc.) and technical requirements that follow from business objectives and demand behavior (scalability, high availability, reliability, ubiquitous access, security, short deployment cycles, etc.).

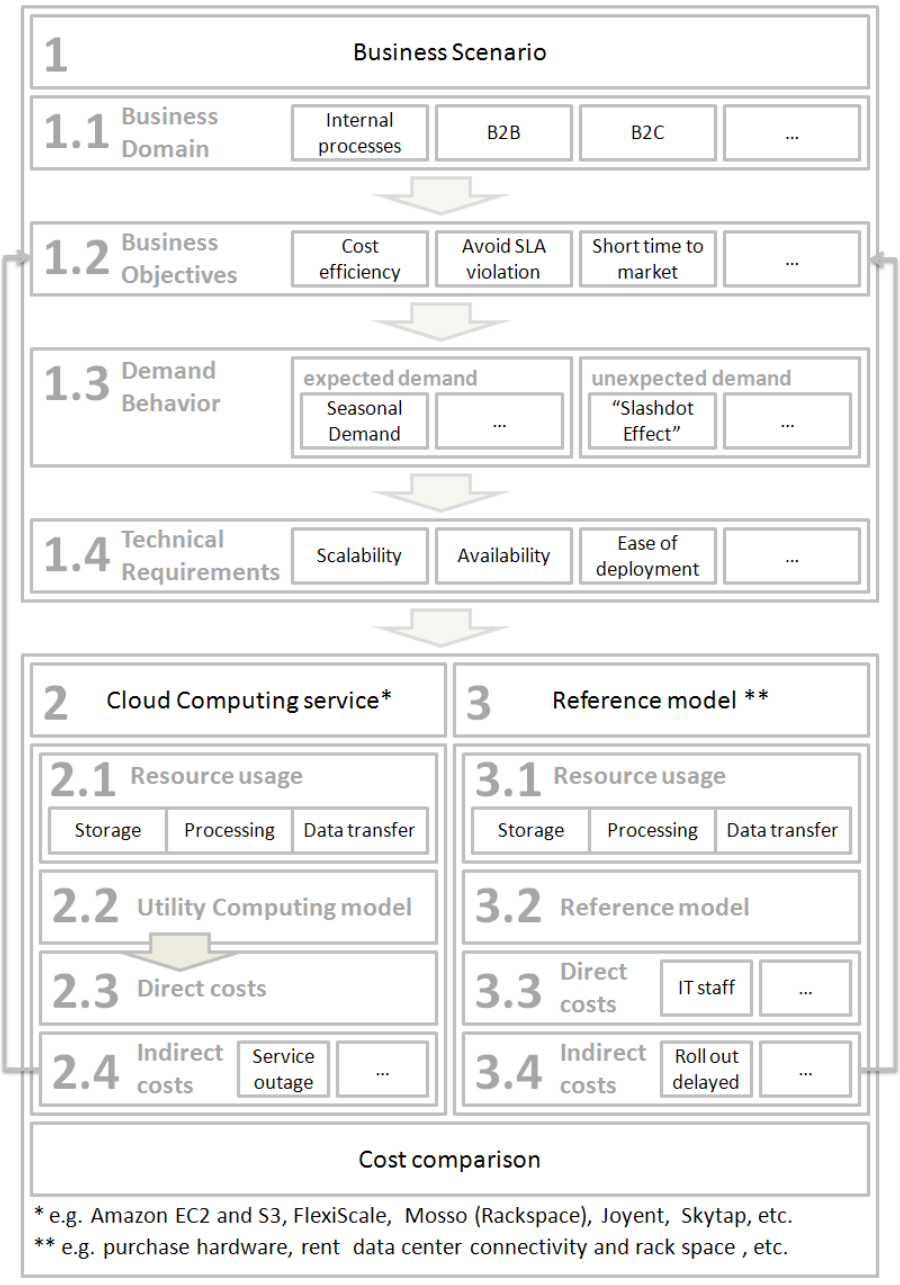

Figure 1: A framework for estimating the value of Cloud Computing

3.1.1 Business Domain

IT resources are not ends in themselves but serve specific business objectives. Organizations can benefit from Grid Computing and Cloud Computing in different domains: internal business processes, collaboration with business partners and for customer-faced services (compare to [14]).

3.1.2 Business Objectives

On a high level the typical business benefits mentioned in the context of Cloud Computing are high responsiveness to varying, unpredictable demand behavior and shorter time to market. The IBM High Performance on Demand Solutions group has identified Cloud Computing as an infrastructure for fostering company-internal innovation processes [4]. The U.S. Defense Information Systems Agency explores Cloud Computing with the focus on rapid deployment processes, and as a provisionable and scalable standard environment [5].

3.1.3 Demand Behavior

Services and applications in the Web can be divided into two disjoint categories: services that deal with somewhat predictable demand behavior and those that must handle unexpected demand volumes respectively. Services from the first category must be built on top of a scalable infrastructure in order to adapt to changing demand volumes. The second category is even more challenging, since increase and decrease in demand cannot be forecasted at all and sometimes occurs within minutes or even seconds.

Traditionally, the IT operations department of an organization must master the difficulties involved in scaling corporate infrastructure up or down. In practice it is impossible to constantly fully utilize available server capacities, which is why there is always a tradeoff between resource over-utilization, resulting in glaring usability effects and possible SLA violations, and under-utilization, leading to negative financial performance [6]. The IT department dimensions the infrastructure according to expected demand volumes and in a way such that enough space for business growth is left. Moreover, emergency situations, like server outages and demand spikes must be addressed and dealt with. Associated with under- and over-utilization is the notion of opportunity costs. The opportunity costs of under-utilization are measured in units of wasted compute resources, such as idle running servers. The opportunity costs of over-utilization are the costs of losing customers or being sued as a consequence of a temporary server outage.

Expected Demand: Seasonal Demand

An online retail store is a typical service that suffers from seasonal demand spikes. During Christmas the retail store usually faces much higher demand volumes than over the rest of the year. The IT infrastructure must be dimensioned such that it can handle even the highest demand peaks in December.

Expected Demand: Temporary Effect

Some services and applications are short-lived and targeted to single or seldom events, such as Websites for the Olympic Games 2008 in Beijing. As seen with seasonal demand spikes, the increase and decrease of demand volume is somewhat predictable. However, the service only exists for a comparably short period of time, during which it experiences heavy traffic loads. After the event, the demand will decrease to a constant low level and the service be shut down eventually.

Expected Demand: Batch Processing

The third category of expected demand scenarios are batch processing jobs. In this case the demand volume is usually known beforehand and does not need to be estimated.

Unexpected demand: Temporary Effect

This scenario is similar to the "expected temporary effect", except for one major difference: the demand behavior cannot be predicted at all or only short time in advance. A typical example for this scenario is a Web start-up company that becomes popular over night because it was featured on a news network. Many people simultaneously rush to the Website of the start-up company, causing significant traffic load and eventually bringing down the servers. Named after two famous news sharing Websites this phenomenon is known as "Slashdot effect"or "Digg effect".

3.1.4 Technical Requirements

Business objectives are put into practice with IT support and thus translate into specific IT requirements. For example, unpredictable demand behavior translates to the need for scalability and high availability even in the face of significant traffic spikes; time to market is directly correlated with deployment times.

3.2 Costs of Cloud Computing

After having modeled a business scenario and the estimated demand volumes, it is now time to calculate the costs of a Cloud Computing setting that can fulfill the scenario's requirements, such as scalability and high availability.

A central point besides the scenario properties mentioned in section 3.1.3 is the question: how much storage capacity and processing power is needed in order to cope with demand and how much data transfer will be used? The numbers might either be fixed and already known beforehand or are unknown and must be estimated.

In a next step a Utility Computing model needs to define compute units and thus provides a metric to convert and compare computing resources between the Cloud and alternative infrastructure services. Usually the Cloud Computing provider defines the Utility Computing model, associated with a pricing scheme, such as Amazon EC2 Compute Units (ECU). The vendor-specific model can be converted into a more generic Utility Computing unit, such as FLOPS, I/O operations, and the like. This might be necessary when comparing Cloud

Computing offers of different vendors. Since Cloud Computing providers charge money for their services based on the Utility Computing model, these pricing schemes can be used in order to determine the direct costs of the Cloud Computing scenario. Indirect costs comprise soft factors, such as learning to use tools and gain experience with Cloud Computing technology.

3.3 Costs of the Reference IT Infrastructure Service

The valuation of Cloud Computing services must take into account its costs as well as the cash flows resulting from the underlying business model. Within the context of our valuation approach we focus on a cost comparison between infrastructure in the Cloud and a reference infrastructure service. Reaching or failing to reach business objectives has an impact on cash flows and can therefore be measured in terms of monetary opportunity costs.

The reference IT infrastructure service might be conventional IT infrastructure (SME or big business), a hosted service, a Grid Computing service, or something else. This reference model can be arbitrarily complex and detailed, as long as it computes the estimated resource usage in a similar manner as in the Cloud Computing scenario of section 3.2. The resource usage will not in all cases be the same as in the Cloud Computing scenario. Some tasks might e.g. be computed locally, thus saving data transfer. Other differences could result from a totally different approach that must be taken in order to fulfill the business objectives defined in the business scenario.

In the case of privately owned IT infrastructure, cost models, such as Gartner's TCO [3], provide a good tool for calculations [8]. The cost model should comprise direct costs, such as Capital Expenditures for the facility, energy and cooling infrastructure, cables, servers, and so on. Moreover, there are Operational Expenditures which must be taken into account, such as energy, network fees and IT employees. Indirect costs comprise costs from failing to meet business objectives, e.g. time to market, customer satisfaction or Quality of Service related Service Level Agreements. There is no easy way to measure how this can be done and will vary from case to case. More sophisticated TCO models must be developed to mitigate this shortcoming. One approach might be to compare cash flow streams that result from failing to deliver certain business objectives, such as short time to market. If the introduction of a service offering is delayed due to slow deployment processes, the resulting deficit can be calculated as a discounted cash flow.

When all direct and indirect costs have been taken into account, the total costs of the reference IT infrastructure service can be calculated by summing up. Finally, costs of the Cloud Computing scenario and the reference model scenario can be compared.

4 Evaluation and Discussion

Early adopters of Cloud Computing technologies are IT engineers who work on Web-scale projects, such as the New York Times TimesMachine [9]. Start-ups with high scalability requirements turn to Cloud Computing providers, such as Amazon EC2, in order to roll out Web-scale services with comparative low entry costs [7]. These and other examples show that scalability, low market barriers and rapid deployment are among the most important drivers of Cloud Computing.

4.1 New York Times TimesMachine

In autumn 2007 New York Times senior software engineer Derek Gottfrid worked on a project named TimesMachine. The service should provide access to any New York Times issue since 1851, adding up to a bulk of 11 million articles which had to be served in the form of PDF files. Previously Gottfrid and his colleagues had implemented a solution that generated the PDF files dynamically from already scanned TIFF images of the New York Times articles. This approach worked well, but when traffic volumes were about to increase significantly it would be better to serve pre-generated static PDF files.

Faced with the challenge to convert 4 Terabyte of source data into PDF, Derek Gottfrid decided to make use of Amazon's Web Services Elastic Compute Cloud (EC2) and Simple Storage Service (S3). He uploaded the source data to S3 and started a Hadoop cluster of customized EC2 Amazon Machine Images (AMIs). With 100 EC2 AMIs running in parallel he could complete the task of reading the source data from S3, converting it to PDF and storing it back to S3 within 36 hours.

How does this use case fit in our framework?

Gottfrid's approach was motivated by the simplicity with which the one-time task could be accomplished if performed "in the Cloud". No up-front costs were involved, except for insignificant expenditures when experimenting if the endeavor was feasible at all. Due to the simplicity of the approach and the low costs involved, his superiors agreed without imposing bureaucratic obstacles.

Another key driver was to cut short deployment times and thereby time to market. The alternative to Amazon EC2 and S3 would have been to ask for permission to purchase commodity hardware, install it and finally run the tasks - a process that very likely would have taken several weeks or even months. After process execution, the extra hardware would have to be sold or used in another context.

This use case is a good example for a one-time batch-processing job that can be performed in a Grid Computing or Cloud Computing environment. From the backend engineer's point of view it is favorable to be able getting started without much configuration overhead as only the task result is relevant. The data storage and processing volume is known beforehand and no measures have to be taken to guarantee scalability, availability, or the like.

In a comparative study researchers from the CERN-based EGEE project argue that Clouds differ from Grids in that they served different usage patterns. While Grids were mostly used for short-term job executions, clouds usually supported long-lived services [10]. We agree that usage patterns are an important differentiator between Clouds and Grids, however, the TimesMachine use case shows that this not a question of service lifetime. Clouds are well-suited to serve short-lived usage scenarios, such as batch-processes or situational Mash-up services.

4.2 Major League Baseball

MLB Advanced Media is the company that develops and maintains the Major League Baseball Web sites. During the 2007 season, director of operations Ryan Nelson received the request to implement a chat product as an additional service to the Web site [11]. He was told that the chat had to go online as soon as possible. However, the company's data center in Manhattan did not leave much free storage capacity and processing power.

Since there was no time to order and install new machines, Nelson decided to call the Cloud Computing provider Joyent. He arranged for 10 virtual machines in a development cluster and another 20 machines for production mode. Nelson's team developed and tested the chat for about 2 months and then launched the new product. When the playoffs and World Series started, more resources were needed. Another 15 virtual machines and additional RAM solved the problem.

Ryan Nelson points out two major advantages of this approach. First, the company gains flexibility to try out new products quickly and turn them off if they are not a success. In this context, the ability to scale down shows to be equally important as scaling up. Furthermore, Nelson's team can better respond to seasonal demand spikes which are typical for Web sites about sports events.

5 Related Work

Various economic aspects of outsourcing storage capacities and processing power have been covered by previous work in distributed computing and grid computing [12], [13], [14], [15]. However, the methods and business models introduced for Grid Computing do not consider all economic drivers which we identified relevant for Cloud Computing, such as pushing for short time to market in the context of organization inertia or low entry barriers for start-up companies.

With a rule of thumb calculation Jim Gray points to the opportunity costs of distributed computing in the Internet as opposed to local computations, i.e. in LAN clusters [12]. In his scenario $1 USD equals 1 GB sent over WAN or alternatively eight hours CPU processing time. Gray reasons that except for highly processing-intensive applications outsourcing computing tasks into a distributed environment does not pay off because network traffic fees outnumber savings in processing power. Calculating the tradeoff between basic computing services can be useful to get a general idea of the economies involved. This method can

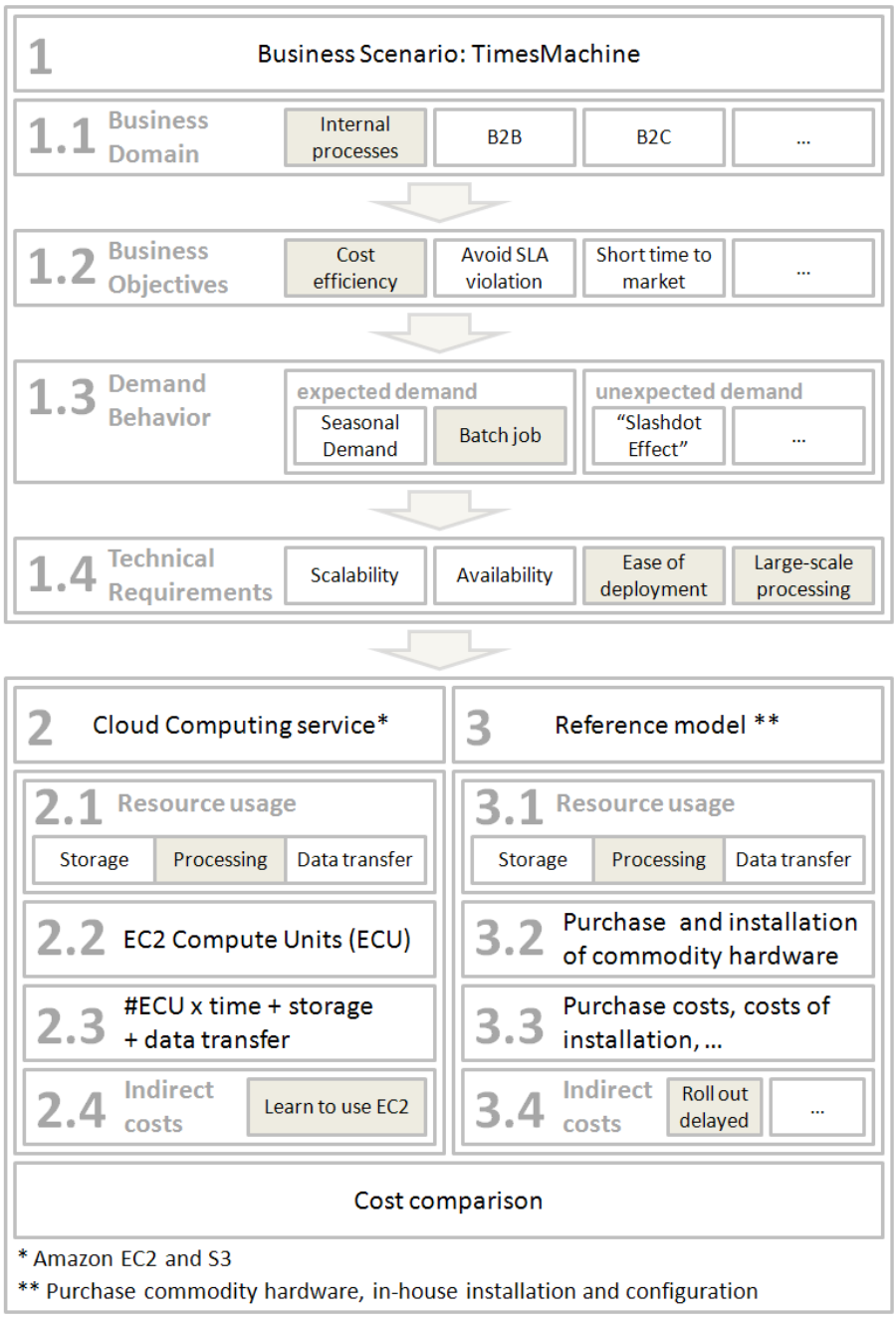

Figure 2: Use Case: New York Time TimesMachine

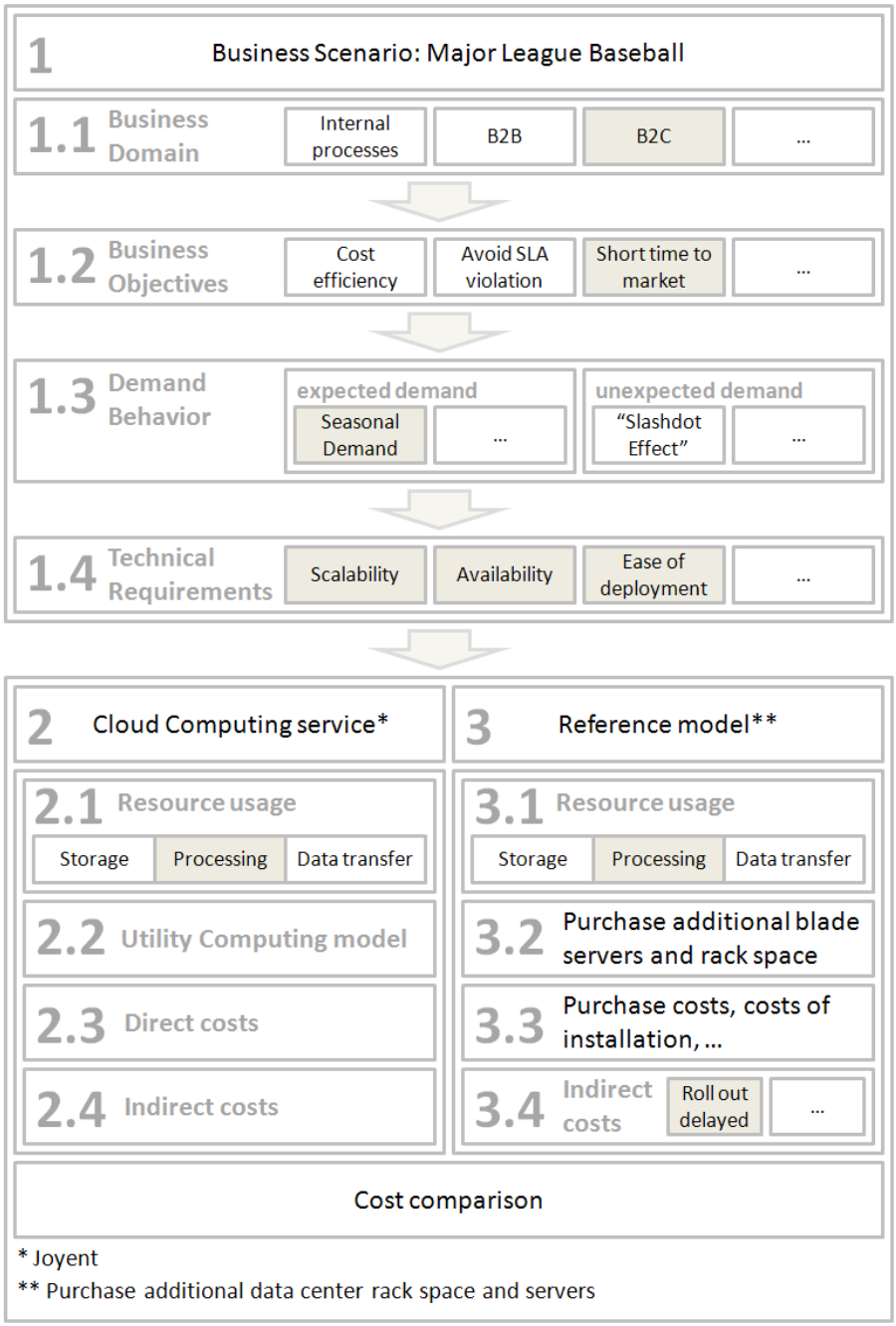

Figure 3: Use Case: Major League Baseball

easily be applied to the pricing schemes of Cloud Computing providers. For $1 USD the Web Service Amazon EC2 offers around 6 GB data transfer or 10 hours CPU processing [1]. However, this sort of calculation only makes sense if placed in a broader context. Whether or not computing services can be performed locally depends on the underlying business objective. It might for example be necessary to process data in a distributed environment in order to enable online collaboration.

George Thanos, et al evaluate the adoption of Grid Computing technology for business purposes in a more comprehensive way [14]. The authors shed light on general business objectives and economic issues associated with Grid Computing, such as economies of scale and scope, network externalities, market barriers, etc. In particular, the explanations regarding the economic rationale behind complementing privately owned IT infrastructure with utility computing services point out important aspects that are also valid for our valuation model. Cloud Computing is heavily based on the notion of Utility Computing where large-scale data centers play the role of a utility that delivers computing services on a pay-per-use basis. The business scenarios described by Thanos, et al only partially apply to those we can observe in Cloud Computing. Important benefits associated with Cloud Computing, such as shorter time to market and responsiveness to highly varying demand, are not covered. These business objectives bring technological challenges that Cloud Computing explicitly addresses, such as scalability and high availability in the face of unpredictable short-term demand peaks.

6 Conclusion and Future Work

Cloud Computing is an emerging trend of provisioning scalable and reliable services over the Internet as computing utilities. Early adopters of Cloud Computing services, such as start-up companies engaged in Web-scale projects, intuitively embrace the opportunity to rely on massively scalable IT infrastructure from providers like Amazon. However, there is no systematic, dedicated approach to measure the benefit from Cloud Computing that could serve as a guide for decision makers to tell when outsourcing IT resources into the Cloud makes sense.

We have addressed this problem and developed a valuation framework that serves as a starting point for future work. Our framework provides a step-by-step guide to determine the benefits from Cloud Computing, from describing a business scenario to comparing Cloud Computing services with a reference IT solution. We identify key components: business domain, objectives, demand behavior and technical requirements. Based on business objectives and technical requirements, the costs of a Cloud Computing service, as well as the costs of a reference IT solution, can be calculated and compared. Well-known use cases of

[1] According to the Amazon Web Service pricing in July 2008 one GB of outgoing traffic costs $0.17 for the first 10 TB per month. Running a small AMI instance with the compute capacity of a 1.0-1.2 GHz 2007 Xeon or Opteron processor for one hour costs $0.10 USD.

Cloud Computing adopters serve as a means to discuss and evaluate the validity of our framework.

In future work, we will identify and analyze concrete valuation methods that can be applied within the context of our framework. Furthermore, it is necessary to evaluate cost models that might serve as a template for estimating direct and indirect costs, a key challenge that we have only mentioned.

References

[1] *Amazon Web Services: Customer Case Studies*, http://www.amazon.com/Success-Stories-AWS-home-page/b/ref=sc_fe_l_1?ie=UTF8&node=182241011&no=3440661

[2] Titman, S.,Martin, J.: *Valuation. The Art & Science of Corporate Investment Decisions*, Addison-Wesley (2007)

[3] *Gartner TCO*, http://amt.gartner.com/TCO/index.htm

[4] Chiu, W..: *From Cloud Computing to the New Enterprise Data Center*, IBM High Performance On Demand Solutions (2008)

[5] *Pentagon's IT Unit Seeks to Adopt Cloud Computing*, New York Times, http://www.nytimes.com/idg/IDG_852573C400693880002574890080F9EF.html?ref=technology

[6] Schlossnagle, T.: *Scalable Internet Architectures*, Sams Publishing (2006)

[7] *PowerSet Use Case*, http://www.amazon.com/b?ie=UTF8&node=331766011&me=A36L942TSJ2AJA

[8] Koomey, J.: *A Simple Model for Determining True Total Cost of Ownership for Data Centers*, Uptime Institute (2007)

[9] *New York Times TimesMachine use case*, http://open.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/

[10] Bégin, M.: *An EGEE Comparative Study: Grids and Clouds - Evolution or Revolution?*, CERN Enabling Grids for E-Science (2008)

[11] *Major League Baseball use case*, http://www.networkworld.com/news/2007/121007-your-take-mlb.html

[12] Gray, J.: *Distributed Computing Economics. Microsoft Research Technical Report: MSRTR- 2003-24*, Microsoft Research (2003)

[13] Buyya, R.,Stockinger, H.,Giddy, J.,Abramson, D.: *Economic Models for Management of Resources in Grid Computing*, ITCom (2001)

[14] Thanos, G., Courcoubetis, C., Stamoulis, G.: *Adopting the Grid for Business Purposes: The Main Objectives and the Associated Economic Issues*, Grid Economics and Business Models: 4th International Workshop, GECON (2007)

[15] Hwang, J.,Park, J.: *Decision Factors of Enterprises for Adopting Grid Computing*, Grid Economics and Business Models: 4th International Workshop, GECON (2007)