Labor Cost Reduction with Cloud: An End-to-End View

Murthy Devarakonda, Purnendu Gupta, Chunqiang Tang
IBM Thomas J. Watson Research Center
Yorktown Heights, NY (USA)

Abstract—This paper presents a new, compelling argument for a larger potential to reduce labor cost by adopting cloud computing in enterprise IT. The additional cost reduction is possible in the labor intensive activities of customer-specific customization of cloud services, tools support, and customer-facing service management. These essential service delivery processes are excellent candidates for standardization and automation using the cloud computing principles. We present examples of such processes and potential for optimization in these processes.

I. INTRODUCTION

For several years now, cloud based services have been touted as a panacea for cost sensitive enterprise customers. For the cloud owners and operators, the fundamental basis for cost reduction has been twofold: 1) economies of scale, in essence spreading the fixed costs, both capital expenses and operating expenses, over a greater number of units leading to a lower transaction costs or unit costs, and 2) elimination of under-utilization, taking advantage of the inherent flexibility of a virtualized system for utilization optimization.

For the cloud users, the benefits are a little different. They may find the pay-as-you-go cost model of a public cloud and the illusion of elastic resources quite compelling [1]. For those cloud users who see issues with a public cloud, such as the security and connectivity to the existing applications and data, a private cloud can be a good solution.

No matter whether we take the cloud provider’s view or the cloud user’s view, the question is what are all the cost optimization opportunities with a cloud? Especially, how does cloud affect the end-to-end IT cost?

The early targets of cost reduction have focused on hardware and software, followed by consumables like power, cooling, data center overheads, e.g., real estate costs, facilities engineering, and system maintenance labor [2]. A large body of applied and academic research exists around optimizing each one of these cost items. However, as those sources of cost reduction are identified and subsequently plateau out following a predictable path of diminishing returns, the attention must shift to the end-to-end cost optimization for cloud based solutions.

What we mean by the end to end cost of a cloud based solution is that managed IT solutions in enterprises involve activities way beyond just providing a virtual machine (VM) as in a typical public cloud. For example, even after a VM is provisioned, key additional activities are needed, such as setting up an application in the VM, monitoring the application, and periodically applying security patches to the VM. The end to end cost model, therefore, is a holistic view of all cost components in delivering an IT solution. It is important to consider the holistic view in understanding the benefits of cloud, because a narrow view would only yield small return on investment, thus discouraging significant investment.

According to the Gartner 2011 IT Spend and Staffing Report [3] – the single largest category of customer spend on IT, about 41%, has been on labor and labor-enablement (e.g. office space) costs. As consolidation and standardization shrinks the hardware (and to some extent software) spend, the labor costs segment could emerge as an even larger share of the pie and should be the next biggest target of cost reduction. Therefore, it makes sense to focus on the end-to-end labor cost optimization for cloud and this article focuses on the impact of cloud on this largest segment of the end-to-end IT delivery cost.

Cloud, as a transformative paradigm, has the potential to pervade all spheres of labor activity with sustainable efficiency gains unlike a classical technology centric transformation. For example, replacing one hardware platform with another or replacing for-fee software with open source software may have limited impact. But, the broader impact of cloud stems from the fact that cloud enables and forces a standardized and simplified solution, impacting a large spectrum of activities that are needed to run and manage an IT solution. This becomes even more as enterprises go beyond low hanging fruit towards legacy applications [9].

II. A HOLISTIC VIEW OF LABOR COST

We start with an examination of a more complete view of labor-centered activities in delivering IT service. The activities can be seen as three concentric spheres of value added labor with progressively higher service maturity and sophistication (see Figure 1). Seeing the labor as concentric spheres is more than an accident of schematic depiction. Each layer adds value to its inner spheres of activities and brings value closer to the point where the business derives economic benefit from IT.
A. Core Technology Services

This is the innermost sphere and is bare bones labor directly associated with “lights on” IT operations and support activity that encompasses all day-to-day direct operational work, e.g., incident and problem management. This is where predominantly the hands-on work done by the system administrators is accounted for. In a typical IT environment, monitoring tools or even human operators monitoring the systems create “tickets” or work orders based on observed incidents, and service support staff or system administrators act on the incident tickets. These may include functional problems (such as back up failure), performance problems (such as unacceptably high CPU utilization), and access control issues (user id and password). The activities also include routine actions such as obtaining approvals for patches and applying patches to various software elements in the stack.

B. Tools and Integration Services

The middle sphere encompasses labor associated with creating and maintaining tools and services that glue together core technologies into the practices of the IT provider (or of the enterprise IT if IT is in-sourced), and thus prepares the core technologies for business consumption and management insights. Consider for instance asset management. Well managed IT providers and enterprises have a standardized way to record, monitor and manage hardware and software assets including software licenses, which are implemented using a standard set of tools. Labor is involved in setting up and maintaining the tools. Other examples of tool-related labor include setting up and supporting tools needed for problem and incident ticket tracking, performance and capacity management, identity and access management, security, service desk, and reporting. We note that the activities related to resolving problems or managing assets are distinctly different from activities needed to set up and maintain the tools needed for these tasks.

Another distinct component of the activities is the integration of a standardized offering to meet specific business requirements, preferably through a set of add on options or at the worst using a customization layer around the standard offering. Concretely, this translates to matching the functions of off-the-shelf cloud management software with the specific business needs of the end users. Consider for example the disaster recovery needs or patch application process (timing and frequency) requirements of the business. The off-the-shelf cloud management may have some support or a default process for these and it is necessary to bridge the gap to the business requirements. There will be a distinct labor cost associated with delivering such add-ons, which can be far more substantial than one would assume.

C. Service and Business Management

The outermost sphere encompasses activities connecting IT management with the IT consuming business entity. This is a critical function as it wraps the entire IT offering as a managed service. It helps monitor, and continuously optimize and align the IT service outcomes with business goals in partnership with the IT consuming business entity. For IT service providers the business entity would be the all-important customer and there can be nothing more significant than well managed relationship with the customer. Even if the IT is in-sourced, it is important for the IT department to manage relationship with the lines of business based on well-defined business and operational outcomes.

Specifically, these activities include reporting on SLA (service level agreement) metrics on a periodic basis, understanding customer satisfaction (on the SLA metrics performance as well as on non-measurable aspects), and handling requests for add on services within the scope of overall agreement. These are often called service and business management and these activities contribute to the perceived service quality. As a result, significant labor is expended here. We will describe further details of these activities later in the article.

D. Discussion

Among the three components, cost reduction opportunities for the Core Technology Services [4] have been the main focus in the past and the findings led the way to developing techniques which can achieve at least some of the cost reduction in practice [5]. The approaches employed have ranged from process modeling and work-breakdown analysis from data sources such as the ticket (work order) repositories and workflow systems, generally aiming for productivity gains through full or partial automation of the activity and process steps using scripts, automation tools, best practices, standardization and process redundancy removal.

However, the activities defined in the remaining two components have largely remained untouched by previous cost reduction initiatives. Few analyses, such as [10][11], considered some but not the costs in tooling and business management. We believe the key reason for this is the
difficulty of quantifying labor effort in these categories vis-à-vis those mentioned above. Simply put, we tend to measure what is easily measurable and avoid the ones that are not. Most labor costs within these effort categories cannot be captured through objective means like a work order or ticket history nor is there use of standard operating procedures, pre-defined steps or outcomes that lend themselves easily to an activity breakdown type analysis suited for standard optimization techniques.

Besides these uneven efforts to reduce labor cost, it is also interesting that the impact of technology transformation is typically felt most in the Core Technology Services layer. For instance, virtualization has the most impact on the Core Technology Services, but had little or no impact on the other two components. Unfortunately, as we show in the next section, these last two categories of work may form the largest share of labor cost.

We believe cloud has the unique potential to translate well into the outermost sphere of service and business management and can have a positive impact, because of its transformative capability on how IT is provided and consumed. So far we have characterized the labor cost components from our experience as a part of an IT outsourcing provider, although it generalizes to in-sourced IT delivery as well. A broader industry view based on an industry analyst report [6] (after a reasonable terminology mapping), shows that this is indeed an industry-wide pattern even though we are the first to partition it in this way.

III. LABOR COST DISTRIBUTION – AN IT SERVICE PROVIDER PERSPECTIVE

We conducted an internal analysis at IBM for a few IT outsourced accounts [7] and estimated labor cost distribution as shown in Table 1.

<table>
<thead>
<tr>
<th>Labor Cost Category</th>
<th>Component of Total Labor Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service and Business Management</td>
<td>30-40 %</td>
</tr>
<tr>
<td>Tools and Integration Services</td>
<td>20-30 %</td>
</tr>
<tr>
<td>Core Technology Services</td>
<td>30-40 %</td>
</tr>
</tbody>
</table>

In Table 1, we show a range of values for each category because outsourcing contracts vary in the scope and terms, and they use dissimilar accounting practices for labor costs. Nevertheless, it can be seen that a non-trivial portion of total IT outsourcing labor cost lies in the Tools and Integration Services and Service and Business Management categories.

A Gartner 2011 study [8] of a typical CIO’s office staff breakdown indicates that activities we categorize as Service and Business Management (i.e. Business Process, Relationship, Financial, Human Resources, and Governance management) add up to 42% of total staff, which is comparable to our own internal analysis.

We use a concrete example to illustrate the three categories of labor cost in running and managing an application. Consider the DayTrader application (https://cwiki.apache.org/ GMOxDOC10/day-trader.html) whose architecture is shown in Figure 2. It is a three-tiered
Web application modeled after highly simplified operations of an online stock brokerage. DayTrader implementation uses J2EE components such as servlets, JSP files, enterprise beans, message-driven beans and JDBC to provide user services such as login/logout, stock quotes, buy, sell, and account details.

The run-manage functions are needed whether the application runs in a cloud or not, however, if the application were hosted on a cloud there would be opportunities for labor avoidance.

For the sake of simplicity, we assume that the DayTrader is virtualized, i.e. its components are hosted in virtual machines using one of several industry standard virtualization technologies, such as KVM.

- **Patch management.** When an OS security patch is released, the system administrator (SA) needs to find a change window for applying the patch and create a change request in the change management system. Note that when to apply the patch depends on the severity level of the patch. If the change request is approved, during the change window, the SA shuts down DayTrader, applies the patch, reboots the virtual machine (VM), and performs end-to-end testing to assess if the patch breaks DayTrader. If it works fine, the SA closes the change request. Otherwise, the SA creates an incident ticket to request recovering the VM. Either the same SA or another SA can handle the incident ticket.

- **Change management.** Suppose more data are needed for tuning DayTrader under the current workload. The SA goes through a formal change process to reconfigure WebSphere Application Server (WAS) to generate more detailed logging data. The SA finds a change window and creates a change request. If it is approved, the SA configures WAS with more detailed logging, reboots DayTrader, and closes the change request.

- **Incident management.** Suppose it is reported, either by an automated monitoring tool or by an SA, that the file system where DB2 stores its data is almost full. The SA creates an incident ticket before addressing the issue. The SA adds more storage from SAN, extends the file system, and confirms that sufficient free space now is available. Finally, the SA closes the incident ticket.

Managing the DayTrader application also involves labor to integrate the application into an IT provider’s or an enterprise’s own standardized management processes, for example, to record, monitor, and manage hardware and software licenses. Below are a few examples.

- **Monitoring and reporting tools.** By default, the host platform (whether it is cloud or otherwise) performs basic monitoring for hypervisor and guest OS, e.g., CPU utilization, network traffic, and storage space. Reporting tools are also needed to summarize the monitoring data into information that can be easily consumed by SA. Labor is involved in setting up, maintaining, and using the monitoring and reporting tools.

- **Backup and restore tools.** Data backup need be performed periodically and/or on demand to guard against failure and to satisfy auditing requirements. For example, if DayTrader’s database is corrupted, it must restore data from the latest backup to minimize loss. Backup and restore tools make the job easier, but labor is involved in setting up, maintaining, and using those tools.

- **Patch management tools.** Because patches are released frequently, patch management is one of the most labor intensive activities in IT management. It is possible to fully automate the patching process described above, from finding a change window and creating a change request, to applying the patch and performing post-patch testing. However, the patching process is complex, and hence labor is needed in developing and maintaining the patching tools.

Finally, managing the DayTrader application involves labor in the category of business and service management services. These services typically encompass all of an outsourcing contract rather than an individual application. With that in mind, let us consider a few examples.

- **Financial and Profit & Loss Management.** Profit and loss management includes financial budgeting and forecasting, billing, managing account receivables, and accounting for overall expenses in running and managing the application. These are critical business insights into the financial health of the contract for the client and service provider’s executive management. Traditionally, each contract is managed separately—a practice that has its origins in the early days of IT outsourcing when highly customized, large contracts are common than an exception. As a result, these services are delivered by a dedicated team using a customized approach (assuming that the entire contract contains running and managing the DayTrader application—in reality, of course, there may be multiple applications under one service contract).

- **Sourcing and Demand Management.** An engineering analogy to the P & L management is sourcing and demand management, where the future demand for the DayTrader application is estimated taking into account client future business plans and expected usage growth. The future demand drives specific requirements for future computing, networking, and storage resource needs. This plan takes into account possible subcontracts, lead time for the fulfillment process, and contingencies and risks involved. Lastly the plan goes through the approval processes within the IT provider as well as at the client end.
• **Contract Governance and Delivery Operations.** Contract governance includes policy adherence, exception handling, contract amendments, and liaison with legal and senior management. Delivery operations include performance metrics review, reporting, ad hoc reports, service assurance, escalation management, and SLA management (such as an availability metric). It also includes providing a Single Point of Contact (SPOC) for the client and ensuring that the client is aware of and able to leverage the provider’s centers of excellence and practices to manage unforeseen issues in service delivery such as Disaster Recovery. Another activity is responding to Request For Services (RFS), which clients use to buy additional services within the contract terms – for example, creating a replica of the application for testing purposes.

IV. OPPORTUNITY FOR EFFICIENCY GAINS

When an application is hosted in a cloud environment there are opportunities for cost reduction using elimination, standardization, and automation levers. Elimination refers to not providing a function because it is deemed not necessary for the service delivered or because of unbundling of services. Standardization refers to reducing variations of a cloud-based offering thus providing opportunities for simplification of service design and the process for delivering the service. Finally, automation is the cost reduction using software and/or hardware implementation for traditionally labor oriented activities. Below we provide a few examples of how these cost reduction levers apply to the three categories of services we identified in the previous sections.

A. Potential efficiency gains in core technology services

As mentioned earlier, the core technology services area has benefited from significant cost reduction efforts, however, there are still some opportunities in this area. Patch management is one such opportunity.

Operating systems, middleware, and applications need to be regularly patched to guard against newly found vulnerabilities or to provide additional functionality. In the non-enterprise space, updates are typically handled by turning on the auto-update feature of the operating system or middleware, which can apply the patches as they are released by vendors. However, these mechanisms for automatically updating are problematic in an enterprise environment. First, a system administrator (SA) needs to assess the impact of newly released patches before rolling them out to their IT infrastructure. Such pre-assessment is clearly not possible using the automatic update feature. Second, an SA needs to have a consistent view of their IT infrastructure, including the vulnerability assessment which cannot be achieved by the automatic updates of vendors. Lastly, an SA needs to assess the post patch impact, including failures of existing applications running on their IT infrastructure. All these changes should be recorded in a change management system for audit and recovery from failure purposes, i.e., involving integration with other management tools.

Figure 3 shows an example of a manual patch management process. Patch process standardization is the first step towards labor cost reduction, as it enables process elimination and automation. Suppose, all customers sign up for the same patching process, where all OS security patches by default must be applied but each customer has the freedom of using a flexible policy language to specify when to apply the patches, depending on patch severity, VM category, and each customer’s pre-defined recurring infrastructure maintenance window. Putting VMs under different categories (e.g., test, development, and production) allows rolling out patches in a staged fashion (e.g., first week covering test VMs and third week covering production VMs) so that a problematic patch is stopped before propagating to the more critical production system.

In this standardized process, the steps of “initiate change” and “approve change” can be eliminated, because all OS security patches are pre-approved and require no further per-patch manual approval. Many other steps can be automated by using a patch automation engine to 1) pull patch notification from patch advisory, 2) calculate a change window according to each customer’s policy, 3) schedule the patch, 4) shut down the application before patching, 5) deploy the patch to VMs, and 6) create an incident ticket in case of patch failure. Without sufficient process standardization, some automation steps would be difficult. For example, if customers’ patching policies are so diverse that they cannot be captured in a common policy language, then it would be impossible to code an algorithm to automate the process of finding a change window for deploying a patch.

Cloud presents opportunities for labor cost reduction in patch management, but also poses additional challenges, because cloud is a large, multi-tenant environment, with a deep vertical stack due to virtualization. These challenges need be handled carefully in automation. One example is that virtual machines can be in different time zones. Since virtual machines run on hypervisors, the scheduling of patches on VMs can interfere with the maintenance of the data center physical infrastructure. Another example of challenge is that, unlike physical machines, a customer can take snapshots of virtual machines, and restore the virtual machine from snapshots. Restoring a virtual machine from a snapshot is challenging from patching’s perspective, because a compliant VM being restored from an earlier snapshot may immediately become non-compliant.
B. Potential efficiency gains in tooling and integration services

Cloud also brings opportunities for cost reduction in tooling and integration services. Consider monitoring and performance reporting as example, which are perhaps the most common IT management function. In a traditional environment, it is common that different customers use different monitoring software suites, due to historical reasons. Service providers supporting these customers need to deploy and maintain a diverse set of monitoring tools, incurring a high cost. A multi-tenant cloud forces many customers to converge on a single monitoring suite, eliminating the cost for supporting a diverse set of monitoring tools.

Cloud can also automate the process of setting up the monitoring tool. For example, setting up the monitoring server and its relay topology is a one-time effort whose cost is amortized over all VMs owned by different customers. The monitoring agent can be embedded in the cloud’s VM image templates ahead of time. When a new VM is created, the monitoring agent is automatically activated and connects to the monitoring server, eliminating the need for manually installing and configuring the monitoring agent.

Even with a standardized monitoring solution in cloud, it is still possible that a customer requests customized monitoring, e.g., for an SAP application that is specific to a particular customer. In this case, an additional monitoring agent can be installed (e.g., ITM agent for SAP), but it still connects to the shared monitoring infrastructure, and hence incurs marginal additional cost. In another case, even with the same monitoring tool, some customer may request customized monitoring reports with special content or format. The self-service concept of cloud can be leveraged here to reduce cost. The cloud portal can be integrated with customizable reporting software such, which allows an end user to easily compose her own custom reports through self-service.

C. Potential efficiency gains in service and business management

Perhaps the most significant opportunity for cost reduction exists in service and business management area, because a cloud-based solution has the ability to transform one-off, customized, high-touch activities into a streamlined, standardized, online solution with minimum labor involvement. Some processes may be wholly rendered redundant in a cloud-based on-demand environment and there may be new opportunities for efficiency gains introduced by automation. For example, for the provider, a large part of the Demand Management function for compute resources is rendered redundant in cloud. Demand data still has to be gathered for provisioning and revenue forecasting purposes by the provider and as a cost budgeting input for the service consumer. However instead of having to poll each client specific data and consolidating them, using one-off solutions, this information can be centrally mined from the cloud resource metering tools and then reused for both automated demand fulfillment and forecasting purposes. In fact this information can be provided as a standard service for managing demand to multiple enterprise customers so that they can receive first hand views when they like and even create custom reports based on changing needs in a self-service manner.
On the Business Management (Financial and P&L) front, some activities such as client billing and account budgeting and forecasting become candidates for automation using the cloud metering service and the scope for human error and consequent reconciliation effort should reduce significantly. Cloud metering service provides data for an application quantifying usage of cloud resources such as network bandwidth and CPU cycles. From this data, cloud provider software can automatically produce billing reports using the sales agreement (e.g. cost per 1Gbit of network bandwidth usage) in effect for the application. Using the same data, if made available to the cloud consumer, cloud consumer can automatically develop budgets according to business guidelines. Even high value activities such as portfolio management will be aided by greater transparency and objectivity of data and predictability of the business model as a whole. On the other hand, issues like outstanding Account Receivables management will likely continue to remain highly a variable process per client. The opportunity to centralize or streamline this (for example through client’s vendor management) and enforce through contractual terms will largely depend on the client organizational structure, culture and business relationship.

Further, Delivery operations functions such as canned reporting and large parts of SLA management processes (including service credits, consequence reward/penalty models dictated by gain sharing schemes) can be fully automated and packaged as a tiered, value added service. SLA management tools which provide a standard features while allowing contract driven customizations already exist on the market. In cloud, this can also be deployed as a bundled service and re-used across multiple customer engagements. Now obviously, this will not resolve one-off needs like ad-hoc reporting and other highly contextual exercises with unique outcomes such as root cause analysis, remediation and dispute reconciliation. These will continue to demand manual intervention and collaboration on both the client and provider ends.

V. Summary

The key observation of this paper is that cloud as a transformative technology can have an impact on multiple components of the end to end IT service delivery. So far labor cost reduction is primarily focused on system administration tasks that we call core technology services. There are two other components - tools and integration, and service and business management - which are also excellent opportunities for labor cost reduction. Our observations indicate that the labor cost in these areas may be as much as, if not more than, in the obvious system administration tasks. As enterprises embrace cloud for its elastic and pay-as-you-go economic model, the service providers whether they are internal or outsourcers can benefit from an end-to-end view of labor cost.

REFERENCES

[4] Jay E. Pultz, 10 Key Actions to Reduce IT Infrastructure and Operations Cost Structure, Gartner Research
[7] IBM Research internal data