

Control Theory for Service Systems

Yixin Diao

IBM Thomas J. Watson Research Center

Hawthorne, New York 10532

Email: diao@us.ibm.com

Abstract—In recent years the trend towards service business is progressing rapidly throughout the world economics. Responding to the demands for service innovation and service productivity improvement, service science is emerging as a new research field that applies multi-disciplinary studies to service systems. In this paper we propose using control theory as a means to design service management solutions that are robust to system dynamics and uncertainty. We further present several control problems and challenges that are fundamental to service systems, and demonstrate how control technologies can be applied through case studies.

I. INTRODUCTION

The service business is becoming increasingly important throughout the world economics. Ranging from service organizations (such as retail, transportation, and health care) to service functions (e.g., marketing, logistics, customer service), the service business accounts for more than 80 percent of the U.S. gross domestic product and employ a large and growing share of workforce [1]. The trend towards services is also progressing rapidly in the developing world and the service sector is becoming the central industrial sector replacing agriculture and manufacturing.

Compared to manufacturing industries, however, the operation of service sector is heavily relying on experience and intuition, and the service productivity is inferior. As pointed out by National Academy of Engineering, one of the major challenges to services industries is “the adaptation and application of systems and industrial engineering concepts, methodologies, and quality-control processes to service functions and businesses” [1]. In response to the demands for service innovation and service productivity improvement, a new concept called “service science” has emerged [2]. Service science aims to increase the productivity of the service industry through applying scientific means and methods. By regarding services as a system associating people, business, technology, and society, service science addresses service system management problems through multi-disciplinary studies and needs to answer questions such as how to model the service systems and how to adaptively change with business strategy.

The complexity and importance of developing service science for service system management has attracted research efforts of a theoretical as well as an applied nature. In particular, control theory is emerging to provide promising technologies for designing robust service management solutions that take into account system dynamics and uncertainty. Control theory has been widely used in manufacturing industries. In the last

several years, there have been many examples of applying control theory to computing systems, including impact on commercial products [3] [4]. However, the application of control theory to service systems is very rare. The focus of this paper is on the control problems and challenges that are fundamental to service systems. We further identify several case studies for potential control applications.

The remainder of this paper is organized as follows. Section II discusses control problems and challenges in service systems. Section III presents several case studies. Our conclusions are contained in Section IV.

II. CONTROL PROBLEMS AND CHALLENGES

The service business is huge and diverse. Providing services in fast food restaurants, supporting end users through off-shore call centers, delivering IT services and infrastructure support from service providers, working with corporate customers to automate and streamline business processes are all examples of services. While how to classify services and what are the service systems to be studied are still questions often discussed at conferences concerning service science, the focus of this paper is not on what are services, but on how services can be managed and improved from the perspective of control theory.

Control problems in service systems arise from two desires that are happening in service business. First, the linkage between technology and business is desired to be even closer. For example, the objective of IT system operation and management are not just limited to response time, throughput, or availability, but to support and meet business objectives measured through Key Performance Indicators (KPIs) such as labor cost, average revenue per customer, or customer attrition. In one way, enforcing business objectives can be viewed as a regulation problem where a controller manipulates the system resource to achieve the desired effect on business objectives. Alternatively, enforcing business objectives is also an optimization problem where the service providers seek to minimize the cost needed to meet the objectives.

Second, instead of experience and intuition, quantitative analysis is desired now more than ever in making business decisions, improving work processes, and satisfying staffing needs. For example, given that large amount of information can be collected and measured, how to predict and assess the effects and risks of service investments? Providing quantitative analysis requires analytic system modeling that not only characterizes system dynamics including workload changes and

human behaviors, but is able to be robust to various unknown factors existing in business environment.

The foregoing desires in service systems result in several opportunities and challenges in applying control technologies to improve service productivity.

- 1) Modeling service systems: The service system is a complex system. A service model needs to include and quantify the relationships between the business objectives, the service delivery process, the people performing service tasks, and the service requests with different natures. Modeling service systems is not an easy task. The desired key performance indicators may not be measurable and need to be substituted by other performance indicators. Quantify the effect of people and relating them to the system model is important but highly subjective. In addition, the service system is subject to technology, social, economic, and environmental changes. Although Business Process Modeling (BPM) is a useful technology to model service system [5], it cannot provide quantitative insight that is needed for service predictability. Thus, rigorous mathematical models are required to capture the quantitative aspects of service systems, and control theoretical models are especially appealing since they model the dynamic nature of the service systems which are tightly coupled and governed by inherent feedback. Such models can not only assist service system design through providing valuable insights on stability, controllability, and observerability, but provide quantitative support to service management through what-if simulation and prediction.
- 2) Controlling service resources: Similar to IT systems where system resource such as CPU and memory can be controlled and reallocated for better performance, in service systems people are the resource and control variables to be managed. However, unlike physical devices, human resources are more difficult to control since they contain complex dynamics. The service personnel (e.g., system administrators, call center staff) have their own preference, and can respond to management commands with different attitudes, speeds, and flexibility. The control systems need either to consider the people dynamics explicitly, to predict and observe through close feedback, or to be robust enough to any unknown effects. Besides the control variables, there are also challenges in service systems associated with control objectives. The reference variables are not pre-defined and can be selected from a set of business objectives. The reference values are also subject to negotiation and can be changed. These flexibility and variability raise the questions on what are the best control objectives, and result in adding another dimension to controller design. Finally, similar to enterprise server systems where the end-to-end response time objective may need to be decomposed to a controllable level, transforming business problems to control problems also requires extensive, non-traditional control design efforts that need deep understanding of

both the business process and the control theoretical framework.

- 3) Monitoring service quality: Total quality control is one of the powerful tools in manufacturing industries [6]. Through applying statistical process control (SPC) methods to monitor product specifications [7], potential quality problems can be detected early in the process and remedied quickly to avoid quality erosion in final products. While quality is equally important in service industry, the service process and service result are typically intangible and vaguely defined between the service providers and the customers. Besides the difficulties in defining the quality metrics and obtaining the quality measurements (note that in services people are typically the subject to be measured), there are also challenges associated with applying existing manufacturing-rooted quality control approaches to services, since service practitioners generally do not have such knowledge and background.

Regarding the scope of control technologies, note that the applications of control theory to computing systems are typically centered around classical control theory such as PID controllers as well as different control strategies (e.g., adaptive control, optimal control). However, the diversity of service systems requires applying control theory in a much broader sense. Different control research fields and cross-disciplinary studies (e.g., discrete event systems, machine learning techniques) can all contribute for design and analysis of service systems with a view towards controlling them for improved productivity.

III. CASE STUDIES

In this section we discuss several case studies on applying control technologies to service systems. We first use IT service management as an example to illustrate the process and techniques for modeling service systems. Then, we show how to formulate a service technician dispatching problem as a control problem and solve it through feedback control design. Finally, we describe how to monitor service quality and variability using multivariate statistical process control approaches.

A. Quantifying the Complexity of IT Management

In the area of IT service management, service providers are increasingly looking to improve the efficiency of their IT processes (e.g., change management, patch management, release management) in order to contain and even reduce the labor cost of service management. Figure 1 shows an example (partial) workflow for a change management process, which is used to accept a change request and updates the corresponding configuration items (CIs). As shown in this example, a service system is a complex system that contains service requests with different natures (e.g., a change request can result in either the creation of new CIs or the modification of existing CIs), a service process with a sequence of activities (e.g., authorize and validate the changes against policies, issue queries to extract other needed CIs to accommodate the request), the

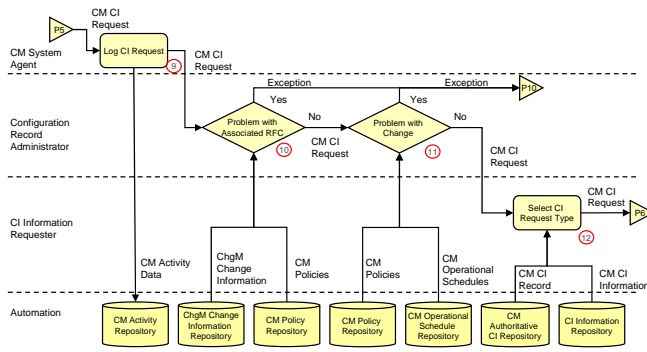


Fig. 1. An example (partial) workflow for a change management process.

people performing service tasks (e.g., change management system agent, configuration record administrator), and the business objectives (e.g., reduce the labor cost, improve the quality of request handling).

While many service process transformation and automation solutions exist for creating standard and reusable process components, one of the key challenges in process transformation is to identify the opportunities and bottlenecks for process reengineering, and subsequently to quantitatively evaluate the efficiency improvement from the transformed process. This is especially valuable before the new process has been deployed, in order to support the decisions for process transformation. One means of addressing the above challenge is to define a set of internal metrics for quantifying the complexity and human cost of carrying out IT service management processes, and to regard complexity as a surrogate for potential labor cost and human-error-induced problems. By correlating complexity metrics and KPIs, this gives us a quantitative approach to evaluate the cost and benefits of service processes.

From the perspective of control theory, the IT service management system can be modeled through a control block diagram, as shown in Figure 2. The service process is the target system to be controlled, which is not well known since not all the process details can be properly documented and modeled. The service process is also subject to two types of changes: workload and process adjustment. Examples of process adjustment include adding or removing activities in a process, assigning service management roles from people with different skill levels, and routing service requests. All of them can be designed and reengineered as desired. On the other hand, the workload is modeled as disturbance in a control system which may vary rapidly or slowly with respect to the volume of service requests, the number of servers in an enterprise, or the geographical locations where the service is performed.

In a service system, the controller stands for the service owner or process designer that are interested in transforming the service processes, and monitoring service performance and quality, and achieving the business objectives (from both the customer and the service provider). The above activities compose the macro feedback loop in service systems, even

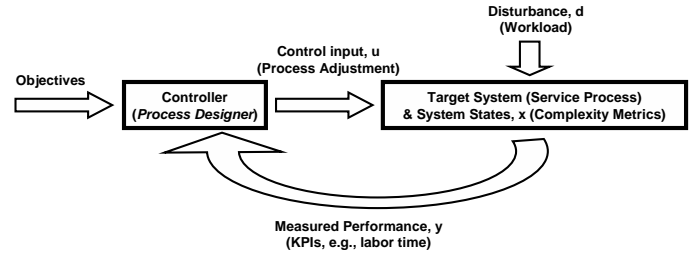


Fig. 2. Control block diagram of the IT service management system.

if various micro feedback loops may exist inside the target system (service process) when service personnel is handling the service requests.

Using the state space model, the IT service management system is expressed as

$$x(k+1) = f(x(k), u(k), d(k)) \quad (1)$$

$$y(k) = g(x(k), u(k), d(k)) \quad (2)$$

where $x(k)$ denotes the complexity metrics as state variables, $u(k)$ denotes process adjustment as control variables, $d(k)$ denotes disturbances such as workload changes, and $y(k)$ denotes the Key Performance Indicators (KPIs) as the measured output. While function $f(\cdot)$ defines the dynamics of the service process and is useful to reveal insights for process adjustment and controller design, the focus of our current work is to acquire function which shows how to infer the KPIs.

As a first step, we assume the workload remains constant and the process adjustment indirectly affects KPIs through the complexity metrics. Furthermore, we assume a linear relationship between them since linear models tend to be more robust, especially when a large number of model inputs are involved and the modeling noise is significant due to the low quality of model inputs. Thus, function $g(\cdot)$ is simplified to

$$y = b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (3)$$

See [8] for more modeling details and how the linear assumption is verified by comparing to nonlinear alternatives. One of central aspects of the above prediction model is to have a framework for measuring IT system management complexity. As defined in [9] and [10], this framework relies on categorical classification of individual complexities, and is briefly summarized as follows:

- *Execution Complexity* refers to the complexity involved in performing the tasks that make up the service process, typically characterized by the number of tasks, the context switches between tasks, the number of roles involved in a task, and their degree of automation. *Decision Complexity*, a sub-category of execution complexity, quantifies decision making according to the number of branches in the decision, the degree of guidance, the impact of the decision, and the visibility of the impact.
- *Coordination Complexity* represents the complexity resulting from coordinating between multiple roles, either

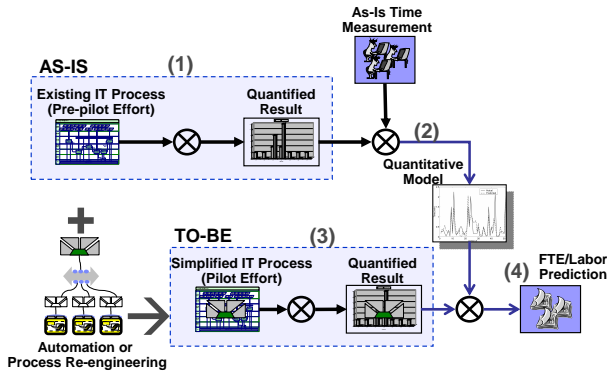


Fig. 3. Analysis flow for modeling IT service management and predicting labor cost.

within a task or between tasks, and depend on how business items are transferred and processed.

- *Business Item Complexity* addresses the complexity involved in providing data into the service process, and is quantified through how the values were obtained.
- *Memory Complexity* takes into account the number of business items that must be remembered, the length of time they must be retained in memory, and how many intervening items were stored in memory between uses of a remembered business item.

The process of using the above prediction model to quantify the labor cost in IT service management are shown in Figure 3, which is composed of four steps. (1) Collecting complexity metrics of the existing (*as-is*) process from the above four dimensions. (2) Identifying the model parameters (b_i) through correlating complexity metrics with time measurement. (3) Capturing the transformed (*to-be*) process with corresponding complexity metrics. (4) Predicting the *to-be* labor time through the quantitative model.

Note that the prediction model defined in Equation (3) is actually a hybrid model, which is based on a qualitative description of service complexity parameters together with quantitative KPI data of the corresponding process. For example, for characterizing the execution complexity three categories are defined based on the degree of automation: automatic, tool-assisted, and manual. A complexity score derived from such categorization does not directly reveal the actual labor reduction or cost savings, so that quantitative calibration is required to relate the complexity metrics to KPIs. We found such a hybrid approach is more practical than a pure qualitative or quantitative approach, since extensive quantitative process measurements (e.g., time to complete a specific task) are usually difficult to obtain in the field, and qualitative descriptions from the service process are much easier to obtain especially in the service design phase. Furthermore, the resulting model is reasonably accurate for the purpose of service investment prediction, and is able to explain 60-70% of the variability in the service data. Figure 4 illustrates the labor time modeling results with 51 action steps, where each action

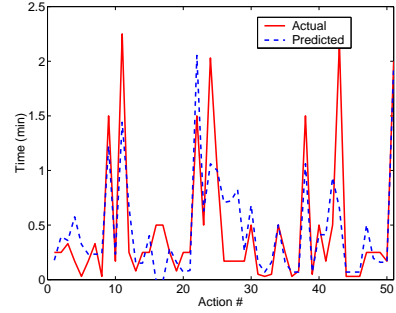


Fig. 4. Illustrative modeling results with data from IBM Tivoli Composite Application Manger configuration process.

is corresponding to one step in the configuration process with various system inputs provided by the system admin.

To make the prediction model further applicable, we need to include workload characteristics (e.g., request volume and severity) and people skills into the model, but not simply treat them as disturbance or noise. Note that these variables are external to the process, and thus do not belong to the same category as the process-inherent complexity metrics. Therefore, nonlinear models may be needed to reflect their different effects to the model output (KPIs).

To this extent, we consider a type of nonlinear models with spatially localized model architecture. As opposed to learning with global functions (e.g., multilayer perceptron neural networks with sigmoidal activation functions), localized information processing (e.g., radial basis function neural networks) can be more effective[11], [12]. In particular, we model the IT service management system using the spatially localized model which takes the form of

$$y = \frac{\sum_{i=1}^M b_i(x) R_i(w)}{\sum_{i=1}^M R_i(w)} \quad (4)$$

$$b_i(x) = b_{i,1}x_1 + b_{i,2}x_2 + \dots + b_{i,n}x_n \quad (5)$$

$$R_i(w) = \exp\left(-\frac{1}{2}(w - c_i)^\top D_i(w - c_i)\right) \quad (6)$$

where y is the output of the service system, that is, the Key Performance Indicators (KPIs). There are M local linear models $b_i(x)$, $i = 1, 2, \dots, M$ corresponding to M “receptive field units” $R_i(w)$. One receptive field unit represents one service condition under which the workload and people skills are specified and the KPIs are measured. Note that while the complexity metrics (x_1, x_2, \dots, x_n) are process-inherent and remain the same for different service conditions, the resulting KPIs can be different which leads to $b_i(x)$ functions with different parameters.

The receptive field functions $R_i(w)$ typically use Gaussian-shaped functions for analytical convenience. The vector w holds the service conditions (including workload characteristics and people skills), c_i parameterizes the centers of the receptive fields in the service condition space, and D_i determines the shapes (or relative widths) of the receptive fields. When the service condition matches the receptive field center (i.e.,

$w = c_i$, $R_i(w) = 1$ and $y = b_i(x)$, given all the receptive fields are narrowly defined so that $R_j(w) = 0$ for all $j \neq i$. Otherwise, for new service conditions never measured before, the values of KPIs will be nonlinearly interpolated between different receptive fields.

B. Dispatching Consultants in Service Transition

Service transition is an important phase that transition activities are undertaken to move the users from the previous service to new service. For major service transition where a large number of heavily-used services need to be replaced with new applications and interfaces, on-site support service may also be required. Dispatching service support technicians can be challenging in such situations, especially when a number of customer sites are involved and each with different needs and speeds to adapt to new services.

Generally, the consultant dispatching problem can be formulated as a constrained optimization problem: Let u_1, \dots, u_N be the number of consultants assigned for N customer sites. The optimization problem is to minimize the total service down time J , subject to the constraint of the total available consultants U . That is,

$$\min J = \sum_{i=1}^N x_i(u_i) \quad (7)$$

$$\sum_{i=1}^N u_i = U \quad (8)$$

where x_i denotes the service down time for site i . If the function between the service down time and the number of consultants is known in advance, the above constrained optimization problem can be solved using various optimization techniques [13].

However, this function is generally unknown due to site to site difference and lack of accurate knowledge on how users respond to new services. Furthermore, the function is also changing over time when the user becomes acquainted to the new services. Even if we may be able to build an online model estimator and solve the optimization problem subsequently, such an approach is sensitive to model uncertainty and noise. The modeling inaccuracy will propagate through the optimization process without attenuation, and lead to inaccurate consultant dispatches—a control type referred to as open-loop control in control theory.

We address this problem with a closed-loop feedback control approach, which uses a similar control paradigm as in managing the database memory [4] [14]. According to the first order Karush-Kuhn-Tucker necessary conditions, minimizing the total service down time is equivalent to equalizing the partial derivatives of the total service down time with respect to the number of consultants for each site. That is, to equalize $\frac{\partial J}{\partial u_i} = \frac{dx_i}{du_i}$. This objective can be achieved through designing a set of feedback controllers, one for each site. The control variable is the number of consultant u_i , the system output is the derivative of the service down time $y_i = \frac{dx_i}{du_i}$, and the

control reference is the average service down time derivatives from all sites. The controller is taking the form of

$$u_i(k+1) = u_i(k) - \frac{1-p}{b_i(k)} \left(y_i(k) - \frac{1}{N} \sum_{j=1}^N y_j(k) \right) \quad (9)$$

where p denotes the pole location, and $b_i(k)$ denotes the second derivative of service down time with respect to the number of consultants, which is obtained online to reflect the user behavior changes.

Below, we briefly summarize the steps on how the feedback control algorithm can be applied to dispatch the service consultants.

- 1) Initialize: Assign the consultants to customer sites several times, each with a different setting, i.e., $u_i(0), u_i(1), u_i(2), \dots, u_i(k)$ for $i = 1, 2, \dots, N$;
- 2) Evaluate: Measure the corresponding service down time $x_i(0), x_i(1), x_i(2), \dots, x_i(k)$, and compute the first derivative $y_i(k)$ and second derivative $b_i(k)$;
- 3) Navigate: Use the control law in Equation (9) to calculate and dispatch consultants for next iteration $u_i(k+1)$.
- 4) Test for completion: If the difference between $u_i(k+1)$ and $u_i(k)$ is sufficiently small, then the solution is converged; otherwise, $k = k+1$ and go to Step 2.

C. Monitoring Service Quality and Variability

Service variability is a service quality characteristic that affects the customer's perceptions of services. Service variability creates the uncertainty to customers with regard to the service quality and value. One of the challenges in monitoring service variability is to handle multi-dimensionality where the states of service are characterized by multiple variables which are often correlated to each other. The data dimension makes individually checking each variable not only cumbersome (or even infeasible due to high dimensionality), but inappropriate since the cross-data relationships are not being considered. Furthermore, since service variables are often correlated, various statistical approaches such as chi-square statistic or Hotelling's T-square statistic are not applicable, which require the variables to be independently distributed [15].

The above problems are addressed in statistical process control using multivariate projection methods such as Principle Component Analysis (PCA) [16]. Using orthogonal linear transformation, PCA transforms the data to a new coordinate system where each coordinated (*principal component*) is orthogonal to each other and ordered according to the variance it captures. By focusing the first few principal components and their arrangement patterns, the process analyst can identify possible quality degradation in an early stage.

We illustrate the use of PCA with the IT service management example discussed in Section III-A. The status of each service step is characterized by eight complexity metrics [8]: (1) Context Switch Distance, (2) Base Parameter Complexity, (3) Parameter Source Complexity, (4) Parameter Source Context Distance, (5) Memory Size, (6) Memory Additions, (7) Memory Latency, and (8) Memory Depth. Note that we

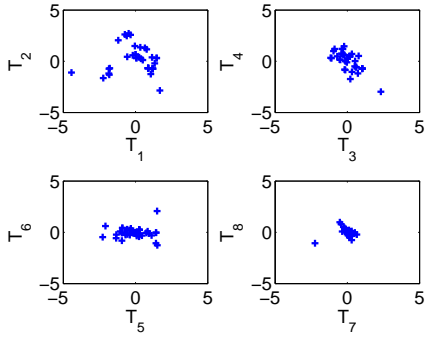


Fig. 5. Principle component analysis with data from IBM Tivoli Composite Application Manger configuration process.

exclude Base Execution Complexity and Parameter Adaptation Indicator because their values remain constant in our example process.

We first normalize the complexity metrics so that they all have zero mean and unit standard deviation. Then, we apply PCA to transform the complexity metrics. The principal components are shown in Figure 5 (where T_1 stands for the first principal component, T_2 for the second principal component, and so on). The standard deviations of these principal components are also given as follows: [1.8625 1.1700 0.9947 0.9432 0.8957 0.4636 0.3903 0.3372]. From both Figure 5 and the above standard deviations we can see that the first principal component (T_1) has captured the greatest variability in the data. This variability was originally represented in multiple correlated complexity metrics, as shown in Figure 6 that illustrates the contributions of complexity metrics to the first four principal components. Also note that the next four principal components (T_2 to T_5) are also worthy to monitor, while the rest three principal components may be negligible.

IV. CONCLUSIONS

In recent years the trend towards service business is progressing rapidly throughout the world economics. In this paper we proposed using control theory as a means to design service management solutions that are robust to system dynamics and uncertainty. We further presented the opportunities and challenges in applying control theory to service systems, and illustrated the applications of control technologies through several case studies.

Our future work involves exploring and validating the proposed models and controllers in more technical depth, and also establish business cases for service systems. For example, we need to evaluate the performance of the nonlinear model presented in Section III-A in a complex workload setting, and study how different control actuators (such as assigning workload and directing workflow to people with different skills) affect the service process performance. We also want to compare the effectiveness of the control algorithm in Section III-B with other optimization or feedforward approaches to demonstrate how a control theoretical approach can better

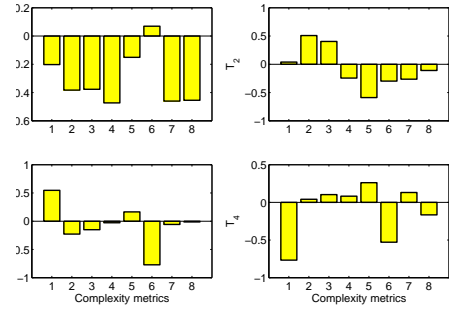


Fig. 6. Contributions of complexity metrics.

handle system uncertainty and dynamics (including the dynamics introduced by the controller itself). Furthermore, we want to use statistical process control technologies as a starting point for system monitoring, and close the control loop by developing resource controllers to improve service quality.

REFERENCES

- [1] "Impact of academic research on industrial performance." National Academy of Engineering, 2003.
- [2] L. D. Paulson, "Services science: A new field for today's economy," *Computer*, vol. 39, no. 8, pp. 18–21, 2006.
- [3] S. S. Parekh, K. R. Rose, J. L. Hellerstein, S. Lightstone, M. Huras, and V. Chang, "Managing the performance impact of administrative utilities," in *Proceedings of the 14th International Workshop on Distributed Systems: Operations and Management, Heidelberg, Germany*, pp. 130–142, 2003.
- [4] Y. Diao, J. L. Hellerstein, A. Storm, M. Surendra, S. Lightstone, S. Parekh, and C. Garcia-Arellano, "Using MIMO linear control for load balancing in computing systems," in *Proceedings of the American Control Conference, Boston, MA, USA*, pp. 2045–2050, 2004.
- [5] M. Havey, *Essential Business Process Modeling*. O'Reilly Media, 2005.
- [6] A. Feigenbaum, *Total Quality Control*. McGraw-Hill Education, 3rd ed., 1983.
- [7] J. S. Oakland, *Statistical Process Control*. Butterworth-Heinemann, 5th ed., 2003.
- [8] Y. Diao, A. Keller, S. Parekh, and V. V. Marinov, "Predicting labor cost through IT management complexity metrics," in *Proceedings of the 10th IFIP/IEEE International Symposium on Integrated Management, Munich, Germany*, 2007.
- [9] A. B. Brown, A. Keller, and J. L. Hellerstein, "A model of configuration complexity and its application to a change management system," in *Proceedings of the 9th IFIP/IEEE International Symposium on Integrated Management, Nice, France*, 2005.
- [10] Y. Diao and A. Keller, "Quantifying the complexity of IT service management processes," in *Proceedings of the 17th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, Dublin, Ireland*, pp. 61–73, 2006.
- [11] S. Schaal and C. G. Atkeson, "Constructive incremental learning from only local information," *Neural Computation*, vol. 10, pp. 2047–2084, 1998.
- [12] J. Spooner, R. Ordóñez, M. Maggiore, and K. Passino, *Stable Adaptive Control and Estimation for Nonlinear Systems*. NY: John Wiley and Sons, 2002.
- [13] D. G. Luenberger, *Linear and nonlinear programming*. Addison-Wesley, Reading, MA, 1984.
- [14] Y. Diao, C. W. Wu, J. L. Hellerstein, A. J. Storm, M. Surendra, S. Lightstone, S. Parekh, C. Garcia-Arellano, M. Carroll, L. Chu, and J. Colaco, "Comparative studies of load balancing with control and optimization techniques," in *Proceedings of the American Control Conference, Portland, OR, USA*, pp. 1484–1490, 2005.
- [15] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press, 1979.
- [16] I. Jolliffe, *Principal Component Analysis*. Springer, 2nd ed., 2002.