

# COORDINATING GLOBAL SERVICE DELIVERY IN THE PRESENCE OF UNCERTAINTY

Lav R. Varshney and Daniel V. Oppenheim

IBM Thomas J. Watson Research Center, USA,  
{lrvvarshn, music}@us.ibm.com

## ABSTRACT

Formal coordination mechanisms are of growing importance as service delivery becomes more globalized and informal mechanisms are no longer effective. Further it is becoming apparent that business environments, communication among distributed teams, and work performance are all subject to endogenous and exogenous uncertainty. This paper first provides a decomposition of service delivery into atomic service requests and then describes a stochastic model of an atomic service request that incorporates several forms of uncertainty to help support coordination. Reasoning inside the closed deductive system of the stochastic model will lead to an understanding of optimal structures and fundamental limits in coordinating global service delivery.

## INTRODUCTION

Whether engaged in designing a physical system like an airplane or building an information system like customer relationship management software, organizations providing informational services are becoming more and more globalized with an increasing degree of workforce specialization. Unfortunately project failures, excessive delays, and significant financial losses have been observed in many global service delivery projects. Traditional project management techniques that worked well with co-located teams (Mintzberg 1989) do not scale well to a global workforce (Gumm 2006). As such, it is important to find principled methodologies for structuring, coordinating, and supporting distributed service work.

Traditional approaches to studying distributed service delivery have used deterministic models of the business environment, of communication among people, and of the work itself. This has led to standardized communication protocols and encapsulations of service work, as well as coordination mechanisms that deal with work interdependencies based on business process and business entity lifecycles (Desai et al. 2009), (Oppenheim et al. 2011), (Leymann and Roller 2000), (Nigam and Caswell 2003). The notion of value co-creation in services is also derived from a deterministic paradigm (Spohrer and Maglio 2008).

The world, however, is uncertain and unpredictable. Since service systems are not closed, the environment may exert random exogenous perturbations on them. For example, unpredictable changes in market conditions such as demand or competition may necessitate changes in service schedule or requirements. Moreover, communication among people/machines within service systems may be carried out over noisy communication channels. Furthermore, the human provision of informational services may be faulty or erroneous; indeed software development teams have a propensity for producing software with bugs. Stochastic models are necessary to fully understand the limits of distributed service work and to develop methodologies for optimizing its organization and coordination in the presence of these uncertainties.

Here we decompose global service delivery into a set of interconnected atomic service requests, with service work encapsulated into work packets, similar to work packages in manufacturing. Each service request is modeled as a stochastic system, incorporating environmental uncertainty, noisy communication channels between service systems, and noisy information processing by service providers.

The primary goal of the present theoretical research program is to determine fundamental limits in global service delivery and provide insight into when distribution of service work is better than other organizational forms. A secondary goal is to understand organizational principles that are optimal for distributed service provision, e.g. to understand conditions when it is better to have loose coupling between service systems and when it is better to be very prescriptive using business process methods.

This paper is a first step in the broader research goal. We use control-theoretic and (Shannon) information-theoretic models as inspiration to develop a mathematical model of interconnected service requests. One basic result that arises from the model is that noise in doing work need not be explicitly considered, but can be incorporated into communication noise. General approaches to obtaining some basic limit theorems on global service delivery are suggested by analogy with other systems. Lower limits on coordination costs under uncertainty obtained from the model will eventually inform the shift towards more globalized and specialized coordination-intensive service structures, answering important business questions.

## THE NEED FOR COORDINATION

Performing work in service engagements can often be very complicated. As is known from classical economic theory, there are two fundamental

and opposing requirements in the performance of work: *division of labor* and *coordination*. In large global service delivery organizations, the need for coordination is particularly exacerbated since labor is so divided and physically distributed.

Several mechanisms of coordination have been developed in classical organization theory, such as mutual adjustment through informal communication, direct supervision, and standardization of norms, skills/knowledge, processes, and outputs (Mintzberg 1989). A general theory of coordination has grown out of these and related ideas (Malone and Crowston 1994). Unfortunately, most of these coordination mechanisms are very difficult to implement within globally distributed enterprises and well-near impossible in cross-enterprise collaboration—an emerging way of doing business (Oppenheim et al. 2011).

Wiredu (2006) has put forth four major interacting coordination issues that plague global service delivery:

- *Interdependencies* that arise from distributed work processes,
- *Conflicts of interest* that arise due to distributed work teams with localized incentives,
- *Technology representation problems* that arise from distributed technologies with localized standards, and
- *Uncertainties and equivocalities* that arise due to geographically and organizationally distributed information.

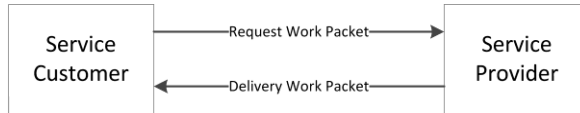
For the first issue, it has been recognized that interdependencies among different pieces of work can be represented with directed acyclic graphs or finite state machines, leading to work flow models, business entity lifecycles, business process models, etc. There are several planning and scheduling algorithms that can be built on top of these models to effectively carry out interdependent work.

The second issue, coordinating people with different costs and benefits associated with different pieces of work, is a central problem in game theory. In the game-theoretic study of coordination, it has been found that Nash equilibria in typical deterministic coordination games are not Pareto optimal, but incur deadweight loss. On the other hand, correlated equilibria may be Pareto optimal; there are several ways of establishing conditions for correlated equilibria or of imposing central coordination that improve upon totally free markets (Vlassis 2007).

This paper discusses potential solutions to the third and fourth of these problems: representation and uncertainty.

MODULARIZATION OF SERVICE WORK INTO WORK PACKETS

Figure 1: An atomic service request with formalized work packets



Technology representation problems can be addressed through the introduction of globally standardized formalisms for service provision and for the doing of work.

As an example, IBM's Application Assembly Optimization (AAO) approach (IBM 2009) applies factory floor assembly and automation principles to distributed software development and management; other companies are using similar factory models (Upton and Fuller 2005). The central construct is the *work packet*, a one-stop source for all of the information necessary to deliver a well-defined component of work, which is used to specify, transport, and deliver each work order. A work packet includes artifacts required by workers to do work, information on work flow, instructions, and performance metrics to be measured. Because the work packet is a self-contained unit with clearly defined predecessor and successor dependencies, the internal technological representations used by individual workers or teams of workers are suppressed for the coordination task (Bernardini et al. 2008), (Chaar et al. 2008).

Besides resolving technology representation issues through standardization, another advantage of delivering services through formalized work packets is *modularity*. Large service engagements can be decomposed into several smaller work packets, and conversely several work packets can be combined into larger work packets. These work packets can also be delegated and reassigned to other service providers without any global impact since they are self-contained and explicitly include dependency relationships. Since work packets of any level of specificity can be decomposed, delegated, and reassigned, they can all be treated identically as service requests.

Any service engagement can be thought of as comprising several atomic service requests formalized through work packets. A single atomic service request is depicted in Figure 1. A complete service engagement would have a (perhaps hierarchical) network of atomic service requests, each dealt with in the same manner (due to uniformity of work packets).

## UNCERTAINTY

Delivering services is subject to various forms of uncertainty. This inherent uncertainty is becoming more noticeable as inefficiencies are

being squeezed out of service organizations. As has been noted, “after years of optimizing supply chains, outsourcing, automation, and stripping costs and inefficiencies out of the back office, most employees spend very little of their day working on regularized activities. What they do is they manage exceptions to processes. Even in the most mundane workplaces like a call center, people are constantly wrestling with new problems” (Tapscott and Williams 2006). Three sources of randomness are of particular interest in the context of coordinating global service delivery.

The first thing to note is that the business environment in which a organization operates exerts random exogenous perturbations on that organization. There may be unpredictable changes in market conditions which necessitate changes in project requirements, such as schedule, cost, and technical requirements. There may also be unpredictable developments in science, engineering, or technology that necessitate a change in the underlying product being developed. So as to accomplish work effectively, a business organization should remain agile and flexible in the presence of these random environmental effects.

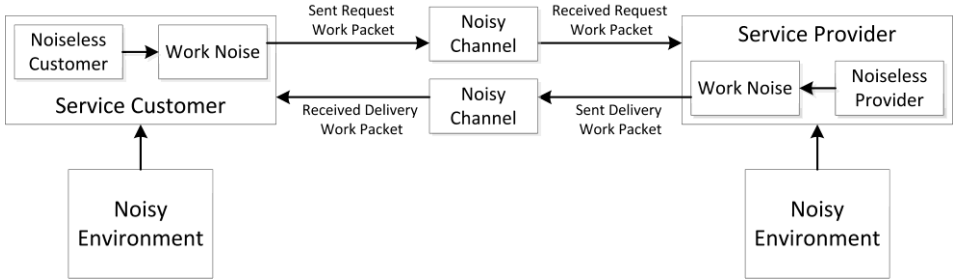
Although work packet communication protocols between people or groups may be standardized to establish a common language for work, communication may still be noisy. One form of communication noise may arise due to imperfectly defined protocols with latent ambiguities, but more importantly, human workers may not describe identical work in identical ways. Furthermore, the work packets may be incompletely or incorrectly interpreted by human workers. A service system should either be robust to miscommunication or have ways of rectifying miscommunication through redundancy or feedback.

Finally, the work carried out by a worker, group, or center may have faults or errors introduced at random. That is, even if the work packet is complete, meets the standard, and is correctly produced and interpreted, the work itself may be incorrect due to human error. Indeed software development teams have a propensity for producing software with bugs. As with communication in the presence of noise, a service system should be robust to work carried out in the presence of noise.

## A STOCHASTIC MODEL OF A SERVICE REQUEST

Having established the need to consider uncertainty in global service delivery, this section introduces a formal mathematical model of an atomic service request, subject to noise in the environment, communication, and doing of work. One can think of environmental noises as perturbing the service customer and perturbing

Figure 2: A schematic view of the fundamental cycle of a service request under uncertainty



the service provider. The customer and provider can be thought of as noiseless agents followed by work noise. Finally, the communication noise affects the flow of work packets between the customer and the provider. Figure 2 depicts the several places where noise impacts an atomic service request.

More precisely, a work packet is thought of as a message drawn from a discrete alphabet  $\Omega$  of possible services that can be performed. Due to standardization, the same alphabet is used by all parties. This alphabet  $\Omega$  is used for any random variable introduced in the sequel.

The noiseless customer initially generates an intended work packet that specifies the work to be done in the service request; the sent request work packet is modeled as a random variable  $W$ . One can also model the work packet generation procedure by the noiseless customer as non-random. This  $W$  is passed through a stochastic kernel representing work noise to yield the sent request work packet.

After receiving a possibly corrupted version of the request work packet through the stochastic kernel representing the noisy channel, denoted by the random variable  $\hat{W}$ , the noiseless service provider performs some transformation  $f$  to produce the results of the service request work. This is specified by another random variable  $Y$ , the sent delivery work packet. The noisy environment on the provider side also enters as an input to the work transformation function. Letting the provider-side environmental noise be the random variable  $M$ , the work function is  $Y = f(\hat{W}; M)$ .

This delivery work packet  $Y$  is passed through a stochastic kernel representing work noise and then through a stochastic kernel representing noisy communication to the service customer to yield the received delivery work packet random variable  $\hat{Y}$ .

The service customer is able to compare the requested work to the delivered work and determine the quality, using a service request quality

Figure 3: There is no essential loss of generality by combining work noise and communication noise into a common interaction noise

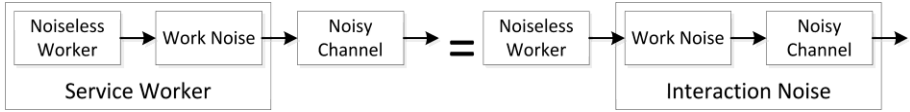
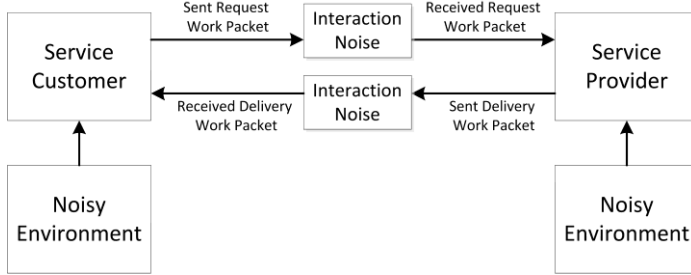


Figure 4: Noise combining yields a schematic view of a service request with external perturbations and noisy interaction



criterion  $q$ . The way that the noisy environment enters the picture on the customer side is as an input to the service request quality criterion. Letting the customer-side environmental noise be the random variable  $N$ , the real-valued quality criterion  $q(W, \hat{Y}; N) : \Omega^3 \rightarrow R$  is parameterized by the external environmental conditions and measures how well the requested work and the delivered work match up to each other.

## SYSTEM OPTIMIZATION IN THE PRESENCE OF UNCERTAINTY

The previous section specified a stochastic model of a service request. It also specified how to measure quality of service through the quality criterion  $q$  and the work function  $f$  that can be changed through system design. In order to determine fundamental limits and properties of optimal systems, we need to perform mathematical optimization.

It can be noted that without loss of essential generality, work noise and communication noise can be combined into a single form of noise, which we term *interaction noise*. This basic result follows from a noise combining argument depicted in Figure 3 that is an application of mathematical systems theory and more particularly the properties of stochastic kernels (Varshney 2011). With noise combining, a service request is as depicted in Figure 4. The transition probability assignments  $p_{\hat{W}|W}(\hat{W}|W)$  and  $p_{\hat{Y}|Y}(\hat{Y}|Y)$  specify the interaction noises.

Maximizing  $q$  through the choice of the work function  $f$  determines the optimal method of service provision in the presence of uncertainty and also the best possible system performance  $q^*$  in the presence of

uncertainty. One may possibly optimize the ensemble of work the service customer chooses to be done by the particular service provider being partnered with, i.e. the probability distribution  $p_w(w)$ .

In some sense, the service request can be thought of as the customer trying to control the provider and the provider trying to control the customer. In Figure 4, we can observe similarity between a service request under uncertainty and several kinds of systems studied in the literature. These include feedback control systems (Hellerstein et al. 2004), communication systems with feedback (Steinbach et al. 2011), interacting automata systems (Richoux and Verghese 2011), and basic perception-action loops in psychology/neurobiology (Klyubin et al. 2007).

Optimal strategies for these other systems are known and they can be modified for our purposes. In particular, extant mathematical tools from stochastic control theory and information theory, as well as new tools developed specifically for this problem, can hopefully give the optimizing work law  $f(Y;M)$  and optimal performance  $q^*(W,\hat{Y};M)$ . It remains to do so.

If there is no uncertainty, it stands to reason that the ultimate optimal performance should be possible for any service request system. When there is inherent and external noise in the system, the optimal performance  $q^*(W,\hat{Y};M)$  may be lower than the ultimate limit.

Looking at networks of stochastic service requests with work packet formalisms allows study of larger service engagements. Composing a global service delivery engagement from several optimal work laws for atomic service requests  $f$  and a stochastic scheduling mechanism that structures and coordinates flows among them will be a good coordination scheme for global service delivery in the presence of uncertainty.

## CONCLUSION

Modeling uncertainty is important for fully understanding the nature of distributed work and coordination in human organizations. Indeed, ignorance may lead to optimizations that are fragile to risk and uncertainty. In fact, this insight is leading to the use of stochastic optimization rather than traditional optimization in business analytics (Davenport and Harris 2007).

From the formal model we have specified, it should be possible to determine the fundamental limits of global service delivery as well as the structural properties and coordination mechanisms of optimal systems. Knowing fundamental limits in the presence of uncertainty could be used to answer questions of business interest.

Perhaps the most important question to be answered from such a theory is essentially a theory of the firm (Alchian and Demsetz 1972), (Grant 1996): *Under which circumstances is the cost of managing/coordinating resources low relative to the cost of allocating resources through markets?* For example, one might be able to say that coordination costs are sufficiently low with a coordination hub (Oppenheim et al. 2011), that a single firm doing distributed work is better than a market.

A second question to be answered from such a theory is one of organizational structure itself: *Under what conditions is it better to have loose coupling between agents and under what conditions is it better to be very prescriptive?* For example, one might be able to say that when there is almost no uncertainty, then business process management is a fine strategy since there is no need for agility or responsiveness; on the other hand, when the level of noise is high, then loose coupling is better.

Finding answers to such questions of coordination cost and structure can have major impacts on service globalization, specialization, and value (Spohrer and Maglio 2008). As noted by Malone and Crowston (1994), a major effect of reducing coordination costs may be to encourage a shift toward the use of more coordination-intensive structures. Optimizing rules and structures for distributed work can facilitate so-called *adhocracies*. Adhocracies would be very flexible organizations, with shifting project teams and decentralized networks of relatively autonomous entrepreneurial groups: fulfilling visions of globally integrated enterprises (Palmisano 2006).

## REFERENCES

- Alchian, A. A., and Demsetz, H. (1972) "Production, information costs, and economic organization," *Am. Econ. Rev.*, 62(5) 777–795
- Bernardini, F., Chaar, J. K., Chee, Y.-M., Huchel, J. P., Jobson, T. A., Oppenheim, D. V., and Ratakonda, K. C., (2008) "Staged Automated Validation of Work Packets Inputs and Deliverables in a Software Factory," U.S. Patent Application US 2009/0300586
- Chaar, J. K., Hamid, A. A., Harishankar, R., Huchel, J. P., Jobson, T. A., Oppenheim, D. V., and Ratakonda, K. C., (2008) "Work Packet Delegation in a Software Factory," U.S. Patent Application US2010/0031226
- Davenport, T. H., and Harris, J. G., (2007) *Competing on Analytics: The New Science of Winning*, Harvard Business School Press
- Desai, N., Chopra, A. K., and Singh, M. P., (2009) "Amoeba: A methodology for modeling and evolving cross-organizational business processes," *ACM Trans. Softw. Eng. Methodol.*, 19(2) 6

- Grant, R. M., "Toward a knowledge-based theory of the firm," *Strategic Manage. J.*, 17 109–122
- Gumm, D. C., (2006) "Distribution dimensions in software development projects: A taxonomy," *IEEE Software* 23(5) 45–51
- Hellerstein, J. L., Diao, Y., Parekh, S., and Tilbury, D. M., (2004) *Feedback Control of Computing Systems*, Wiley, Hoboken, NJ
- IBM Global Business Services White Paper (2009) *Application assembly optimization: A new approach to global delivery.*
- Klyubin, A. S., Polani, D., and Nehaniv, C. L., (2007) "Representations of Space and Time in the Maximization of Information Flow in the Perception-Action Loop," *Neural Comput.*, 19(9) 2387–2432
- Leymann, F., and Roller, D., (2000) *Production Workflow: Concepts and Techniques*, Prentice Hall, Upper Saddle River, NJ
- Malone, T. W., and Crowston, K., (1994) "The interdisciplinary study of coordination," *ACM Comput. Surv.*, 26(1) 87–119
- Mintzberg, H. (1989) *Mintzberg on Management*, Free Press, New York
- Nigam, A., and Caswell, N. S., "Business artifacts: An approach to operational specification," *IBM Syst. J.*, 42(3) 428–445
- Oppenheim, D. V., Bagheri, S., Chee, Y.-M., and Ratakonda, K., (2011), "Agility of enterprise operations across distributed organizations: a model of cross enterprise collaboration," in *Proc. SRII Global Conf.*
- Palmisano, S. J., (2006) "The globally integrated enterprise," *Foreign Affairs*, 85(3) 127–136
- Richoux, W. J., and Verghese, G. C., "A generalized influence model for networked stochastic automata," *IEEE Trans. Syst., Man, Cybernetics—Part A: Systems and Humans*, 41(1) 10–23
- Spohrer, J., and Maglio, P. P., (2008) "The emergence of service science: Toward systematic service innovations to accelerate co-creation of value," *Prod. Oper. Manag.*, 17(3) 238–246
- Steinbach, E., Hirche, S., Kammerl, J., Vittorias, I., and Chaudhari, R., (2011) "Haptic data compression and communication," *IEEE Signal Processing Mag.*, 28(1) 87–96
- Tapscott, D., and Williams, A. D., (2006) *Wikinomics: How Mass Collaboration Changes Everything*, Portfolio Penguin, New York
- Upton, D. M., and Fuller, V. A., (2005) *Wipro technologies: The factory model*, Harvard Business School: 9-606-021
- Varshney, L. R., (2011) "Performance of LDPC Codes Under Faulty Iterative Decoding," *IEEE Trans. Information Theory*, to appear.
- Vlassis, N., (2007) *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*, Morgan & Claypool Publishers
- Wiredu, G. O., (2006) "A framework for the analysis of coordination in global software development," in *Proc. 2006 Int. Wksp. Global Softw. Dev. Practitioner.*