Does *One-Size-Fit-All* Suffice for Service Delivery Clients?

Shivali Agarwal, Renuka Sindhgatta, and Gargi B. Dasgupta

IBM Research India
{shivaaga,renuka.sr,gdasgupt}@in.ibm.com

Abstract. The traditional mode of delivering IT services has been through customer-specific teams. A dedicated team is assigned to address all (and only those) requirements that are specific to the customer. However, this way of organizing service delivery leads to inefficiencies due to inability to use expertise and available resources across teams in a flexible manner. To address some of these challenges, in recent times, there has been interest in shared delivery of services, where instead of having customer specific teams working in silos, there are cross-customer teams (shared resource pools) that can potentially service more than one customer. However, this gives rise to the question of what is the best way of grouping the shared resources across customer? Especially, with the large variations in the technical and domain skills required to address customer requirements, what should be the service delivery model for diverse customer workloads? Should it be customer-focused? Business domain focused? Or Technology focused? This paper simulates different delivery models in face of complex customer workload, diverse customer profiles, stringent service contracts, and evolving skills, with the goal of scientifically deriving principles of decision making for a suitable delivery model. Results show that workload arrival pattern, customer work profile combinations and domain skills, all play a significant role in the choice of delivery model. Specifically, the complementary nature of work arrivals and degree of overlapping skill requirements among customers play a crucial role in the choice of models. Interestingly, the impact of skill expertise level of resources is overshadowed by these two factors.

1 Introduction

Service-based economies and business models have gained significant importance over the years. The customers (a.k.a. clients) and service providers exchange value through service interactions with the goal of achieving their desired outcomes. Given the focus on the individual customer's value and uniqueness of the customer's needs, the service providers need to meet a large variety of expectations set by the customers with due diligence. At the same time, they need to continuously evolve better methods of operations to minimize cost of delivery in order to be competitive in the market. In this paper, we focus on how to organize IT (software) service delivery for diverse customer workloads under strict contractual agreements.

Services in software service industry are typically delivered by specialized Service Workers (SW) or human resources who are teamed together in order to serve the

S. Basu et al. (Eds.): ICSOC 2013, LNCS 8274, pp. 177–191, 2013.

Service Requests (SR) or work of the customer. The structure of this team and the flow of customer work across multiple teams define a Service Delivery Model (SDM). A service provider typically caters to multiple customers belonging to different industry domains that require multiple business functions, applications and technologies to be supported. For example, it is possible to service clients from banking, telecom and insurance domain at the same time by a service delivery organization. In spite of belonging to different verticals, customers may share common business functions like payroll, HR etc. Analogously, it is possible that all these functions for all the customers require common set of technical skills like Storage, Database, and Mainframes etc. In such situations, it becomes important to identify the optimal way of grouping customers and forming SW teams to service them such that service provider can minimize resource costs without compromising customer satisfaction.

A customer's work could be potentially mapped to one or more teams in accordance with one of the following service delivery models: (a) Customer focused (b) Business Function focused and (c) Technology-focused. Figure 1 shows a relationship among business functions, technologies and teams for each of the three models. The legend for technology, business and customer in the figure is as follows: technologies are denoted by colors, the business functions are denoted by the shape of the boxes and the customers are denoted by the different patters in the boxes. A customer has systems based on different technologies (Unix, Windows, Transaction Server, etc.) catering to different business functions (Payroll, Billing, Marketing,etc.). In the **Customer focused (CF)** SDM, all service interactions of a customer, across all business functions are served from single customer dedicated team. While this model is believed to have high customer satisfaction levels, the practical challenges involve scalability (since every new customer on-boarded now needs a dedicated team). In the **Business focused (BF)** model, business functions of multiple customers are served from the common pool. The resources in such a pool have the desired domain knowledge in addition to the required technical skills required to carry out the tasks. This model addresses the utilization issue of the dedicated scenario by supporting multiple customers with similar business functions and also maintains no fragmentation within the business function of a customer. However since business functions may map to different technologies, the common pool again requires expertise in multiple technologies, which results in higher labor costs. In **Technology-focused (TF)** SDM, multiple customers using similar technologies are grouped into a team which is served by highly skilled people in the relevant technologies. There are dedicated teams for each required technology in this model and it carries out work related to that technology from multiple business functions across different customers. In this model single skilled people are needed which is easier to hire and train. The drawback of this model is that customer work is split by technology and tends to get very fragmented. This may result in complex situations taking longer to resolve, as they traverse through the multiple teams, thereby causing customer dissatisfaction.

Given the choice of the types of SDM and their associated merits and de-merits, it becomes challenging for an organization to decide which model to adopt. The situation is further aggravated by the fact that various client specific factors, that may be static and dynamic, play a role in accentuating or diminishing the merits/de-merits of the SDMs. Section 2 of this paper, describes some of the key factors that impact SDM performance. A static one-time decision that is universally applied to all customers may not suffice for the design of a large-scale service provider. Especially with services business revenue being close to a billion USD for major providers, its success is strongly related to the trust and satisfaction of its existing customers. This necessitates a superior decision process regarding which customer workload, service contracts and skill distributions effectively map best to which SDM and optimize cost to the provider. In this paper, we aim to analyze the three SDMs from the perspective of performing highly diverse and complex clients' workload and focus on the multiple performance parameters of SLA, cost, throughput and utilization. The goal is to not only establish the best SDM under a subset of specific circumstances, but also understand the Pareto improvements that can be made to any SDM parameter. The simulation analysis presented here can be used by organizations to find the most appropriate delivery model for a client portfolio. It can also be used to find the appropriate customers groupings for a given SDM type.

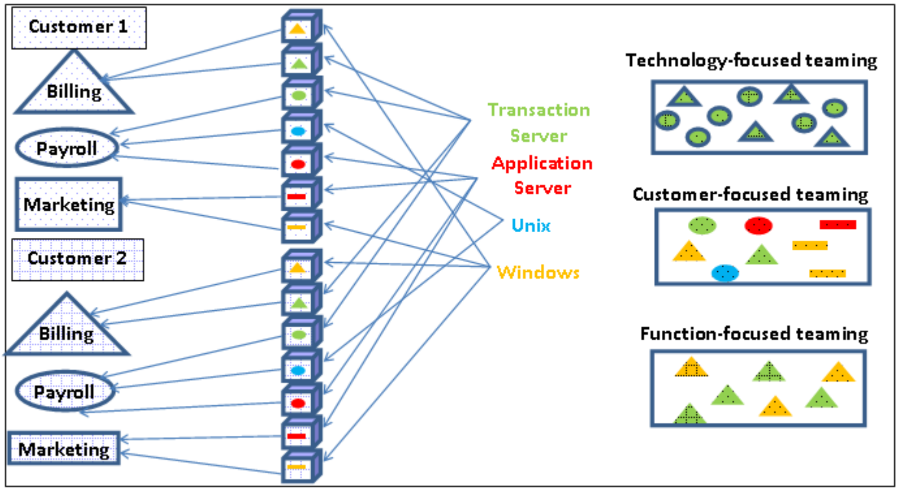


Fig. 1. Customer processes to SDM mapping

Rest of the paper is organized as follows: Section 2 describes the different factors specific to customers that affect the choice of SDM. Section 3 introduces our simulation model and the various operational parameters of interest. Section 4 presents the experimental analysis and section 5 presents a review of the related work.

2 Why One Model May Not Fit All

In this section, we describe the key factors that should be considered in choosing a service delivery model for diverse customer group. Each of these factors capture some aspect of the customer and its' workload. A combination of these factors defines the clients' work portfolio being serviced by the service provider. Different portfolios will typically suit different delivery models. Portfolios are dynamic in nature as existing customers can undergo changes and new customers may get on boarded. A service provider has to deal with different work portfolios at different points in time making it difficult to have a de-facto model because it can perform in a very sub-optimal manner for portfolios that it does not correspond to.

Customer Work Profiles - Work profile of a customer defines the nature of SRs that arrive in that customer's workload. It is a mapping of the customer's business functions to the technologies that are required to carry it out. The combined profiles of customers determine the required skills for the service delivery. Fig 2 provides the combined profiles samples that are studied in this paper. These are representative samples of the actual profiles and capture the key features relevant for simulation. The Type 1 profile in Fig 2 depicts a case where the provider is catering to three customers, C1, C2, C3 such that C1 has work that belongs to business functions of type B1 and B2. The business function B1 needs the technologies T1 and T2 both, while B2 requires T2 and T3 both. The label x1 and y1 denote the percentage of work of type B1 and B2 respectively. The work of C2 and C3 can be interpreted analogously. Each of the types illustrates different levels of overlap between the customer requirements. For example, C1 and C2 in Type 2 have a complete overlap in business domain and technology skill requirements (e.g. payroll and HR may both require Unix and DB2) but only a partial overlap in Type 1.

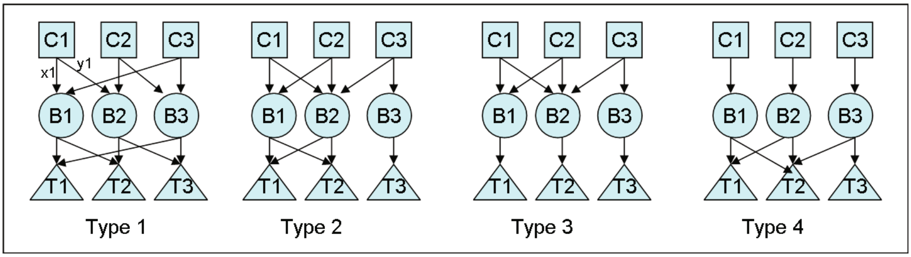


Fig. 2. Clients' Work Profile Sample Combinations

In some cases the combined profiles may look obviously tailored for a certain type of shared model, for example, because of the higher sharing of business functions in Type 3, it is intuitive that sharing of resources at business function level may benefit Type 3. This is less clear in Type 1 and Type 2, where it is possible that TF outperforms BF. A detailed analysis is required to develop the insight into the effect of different overlapping patterns that may occur in customer work profiles.

Workload Arrival Patters – It may be a myopic strategy to make decisions about shared delivery models solely based on customer work profiles, because the work arrival patterns also play a role in accentuating the benefits of sharing. The benefits of sharing will be visible most when customers have complementary workload arrival patterns. That is, the peaks and crests of one customer do not coincide with the others who are being serviced from the same pool. Then the question arises that how are the comparative performances of CF, TF and BF models in case of non-complementary workloads. It is also important to understand the role of overlapping business functions and technical skill requirements in the performance of the three SDMs in case of complementary arrivals.

Business Function Complexity – Some complex SRs may need deep domain knowledge and customer knowledge while the others may be relatively simple to handle. Consequently, the service times for a request involving a complex function will be different in the case where it is handled by a SW with less domain knowledge vs. one with high domain knowledge. As mentioned in section 1, resources in TF SDM will typically have lower level of domain and customer knowledge. This can potentially lead to SLA misses for service requests and thus skill levels become an important factor in choosing the model.

We resort to simulation based analysis for studying the interplay of these factors in mapping the class of portfolios that are best suitable for each of the three SDMs.

3 Formalizing the Service Delivery Model

We now formalize the SDM and present the framework that models the various customer and workload related factors. Each SDM is typically characterized by:

- A finite set of customers, denoted by C, to be supported.
- A finite set of W *S*ervice *W*orkers (SW).
- A finite set of teams consisting of a mutually exclusive subset of W,
  - denoted by CT, if *C*ustomer *F*ocused SDM
  - denoted by BT, if *B*usiness function *F*ocused SDM
  - denoted by TT, if *T*echnology *F*ocused SDM
- A finite set of domain skills, denoted by BD, with L levels in each skill.
- A finite set of technical skills, denoted by TD, with L levels in each skill.
- A finite set of skills pertaining to customer knowledge, denoted by CK.
- A finite set of priority levels, denoted by the set P.
- A finite set of service requests (SR) raised by the customers that arrive as work into the system.
- A map of service requests to required skills, defined by $SR \rightarrow CK \bigcup BD \bigcup TD$.
- A map of service workers to skills,
  - One-to-many map, defined by $W \rightarrow CK \bigcup BD \bigcup TD$, if CT
  - One-to-many map, defined by $W \rightarrow BD \bigcup TD$, if BT
  - One-to-one map, defined by $W \rightarrow TD$, if TT
- A finite set of Key Performance Indicators, denoted by KPI.

In CF model, each customer team has a dedicated set of SWs for each business function, such that they have the customer knowledge, business domain knowledge and are skilled in the required technologies for that function. In BF model, the SWs working in a team are shared across customers and are knowledgeable about the business domain handled by that team as well as skilled in the required technologies for servicing that business function. The workers may acquire customer knowledge in the process of servicing customers for a long period of time, In TF model, the SWs are skilled in a particular technology and may acquire domain and customer knowledge over a period of time by virtue of servicing multiple customers. It is possible to have delivery models that are a combination of CF, BF and TF but such models are outside the scope of this paper. The goal of this work is to fundamentally understand the suitability of specified models for specific type of workloads.

We next discuss the operational aspects like customer SLAs of the SR, service times and evolving skills of workers, and how SRs are dispatched to service workers. We also discuss the specifics of performance indicators.

3.1 Service Level Agreements

SLA constraints, given by the mapping $\gamma : C \times P \to (r_1, r_2), r_i \in \Re, i = 1,2$ is a map from each customer-priority pair to a pair of real numbers representing the SR resolution time deadline (time) and the percentage of all the SRs that must be resolved within this deadline within a month (pct). For example, $\gamma(Customer_1, P_1) = \langle 4,95 \rangle$, denotes that 95% of all SRs from customer$_1$ with priority P_1 in a month must be resolved within 4 hours. Note that SLAs are computed at the end of each month and hence the aggregate targets are applicable to all SRs that are closed within the month under consideration. Also the SLAs are on the entire SR itself, which means the targets apply to resolution across multiple domains.

3.2 Service Time

The time taken by a SW to complete an SR is stochastic and follows a lognormal distribution for a single skill, where the parameters of the distribution are learned by conducting time and motion exercises described in [6]. Service time distributions are characterized by the mapping $\tau : P \times D \to \langle \mu_{1,} \sigma_1 \rangle$, where $D = BD \cup TD \cup CK$ and μ_1, σ_1 are the mean and standard deviation parameters of the lognormal distribution and represent the longest time a worker usually takes to do this work. The distribution varies by the priority of a SR as well as the minimum skill-level required to service it. For complex work requiring multiple skills $(D_1, \ldots D_i)$ the total service time is an additive component of the individual work completions and follows a shifted lognormal distribution [16].

However with some learning in the environment and with repeated use of skills, these service times become lower according to a power law equation given by LFCM [13]. Also since complex work takes more time to complete, for the sake of maintaining throughput, it becomes imperative to assign some work to people skilled below the minimum skill-level. When lower skilled people (s_w) do higher skilled work (s_r), where $s_r > s_w$, the service times become higher. This increase in service time is obtained from an adaptation of the LFCM algorithm [17], where the service time $\mu_n(s_w, s_r)$ to finish the n^{th} repetition of work requiring skill s_r by worker with skill level s_w is given by:

$$\mu_n\left(s_w, s_r\right) = \mu_1 n^{-\beta\left(1-\log\left(1+\gamma/t_n\right)/\log n\right)} \tag{1}$$

where μ_1 is the mean service time to execute the higher skilled work for the *first* time, β is the learning factor, γ is the skill gap between levels s_w and s_r, t_n is the time spent by worker at level s_r. Higher the gap γ, and lower the time spent t_n, higher is μ_n. μ_1 represents the longest time to do this type of work, but with work repetitions, expertise is gained and μ_n decreases. In practice we bound the minimum value of μ_n at $\mu_{\min}$, which is the lowest service time work s_r can take. The parameters $\left\langle \mu_1, \beta, \gamma, \mu_{\min} \right\rangle$ are learned by conducting time and motion studies [6] in real SS to measure the exclusive time spent by a SW on a SR. As given by Eqn. (1), slower learning rates and bigger gaps in the skill required of a SR and skill possessed by a SW, both contribute to longer service times.

3.3 Dispatching

The Dispatcher is responsible for diagnosis of the faulty component(s) as well as work assignment to a suitable worker. During work assignment, SRs are assigned in order of their work priorities to SWs of the matching skill-level requirements. When matching skill levels are not available, higher or lower skilled SW may be utilized for servicing a SR. For fault diagnosis the dispatcher intercepts the SR to determine the most likely faulty component(s) and maps them to appropriate skill domains (from BD,TD). In case of CF model, it ensures that it maps to the right customer team as well. In TF model, a SR dispatched with $\{TD_{1,}TD_2\}$, needs to traverse through teams that support TD_1 and TD_2. When multiple domains of customers are supported, solving the fault-diagnosis without ambiguity is non-trivial [24] and may result in misroutes. We assume no misrouting in our model, without loss of generality.

3.4 Key Performance Indictors

Cost: The cost of delivery is directly related to the cost of the resources working in the teams. Let C_l be the base cost of the resource in TF model with single skill expertise at level *l*. The base cost is assumed to be higher for higher skilled people (i.e., $C_{l1} > C_{l2}, \forall l1 > l2$). In contrast BF/CF model has multi-skilled people who would need training for each additional skill. Let l_H be the highest skill level of a resource in this model. We assume that the base cost of a multi-skilled resource is dominated by the base cost of her highest expertise. She also has *N* additional skills, out of which n_i skills are at level l_i. Let the cost for training each skill to level l_i be given by δ_{l_i}.

Assuming a linear cost model of skills, the cost incurred for training a multi-skilled resource is given by:

$$C = C_{l_H} + \sum_i n_i * \delta_{li}, where \sum_i n_i = N \quad (2)$$

It can be seen that a resource in a BF/CF model is much more expensive than in the TF model.

Utilization: If a resource works for x hours out of available H hours, then the utilization is x/H. A SDM with higher utilization of SWs is indicative of good staffing.

Throughput: Ratio of the amount of work completed and the amount of incoming work is defined as the throughput. A model with higher throughput will typically lead to improved chances of SLA adherence.

4 Simulation Based Evaluation

In this section, we describe the simulation set up for SDM according to its definitions in Section 3 and present the experimental analysis.

Workload Parameters

- *Customer Work Profiles* : The workload is generated as per the customer work profiles given in Fig. 2. These are a very small scale representation of the actual clients' profiles but capture all the essential attributes required for simulation. The values of distributions, x1 and y1 are simulated with either of the two distributions: (i) uniform distribution, (ii) an extremely biased distribution where 90% work is of one business function type and 10% of the other.
- *Work Arrivals* : According to existing body of literature in the area of Service Delivery systems [6,8], work arrives into the system at a finite set of time intervals, denoted by T , where during each interval the arrivals stay stationary. Arrival rates are specified by the mapping $\alpha : C \times T \rightarrow \Re$, assuming that each of the SR arrival processes from the various customers C_i are independent and Poisson distributed with $\alpha(C_i, T_j)$ specifying the rate parameter. Customers can have complementary patterns of work arrival where peaks and troughs complement each other, or it can be amplifying workload with overlapping peaks or the work arrival can be a simple uniform pattern without much variation in time.

Simulation Parameters

- T contains one element for each hour of week. Hence, |T| = 168. Each time interval is one hour long.
- Priority Levels P = {P1, P2, P3, P4}, where, P1 > P2 > P3 > P4.
- Customer Skills CK={C1,C2,C3}, Business Domain Skills BD={B1,B2,B3}, Technology Skills TD={T1,T2,T3}

- *Skill Levels and Service Times*: We assume $L = 3$. The three different levels of expertise simulated are {Low, Medium, High}, where, High > Medium > Low. Each level of expertise has a least service time distribution $(\mu_{\min}, \sigma_{\min})$ associated with it, which characterizes the minimum time this work type could take. The estimates are obtained from real life, time and motion studies [6].
- *Learning Factor:* We assume a learning rate of $\beta = 0.1$ for each SW with high skill level, 0.08 for medium and 0.06 for low.
- *Transfer Time*: In case of work requiring multiple skills, the work gets handed over from one team to another. The teams could be geographically co-located (transfer time ~20min) or dispersed (> 20min).
- *Cost:* A blended rate (across skill levels) of 80K USD per SW and an additional cost of 10K USD per specialized skill or customer knowledge is assumed.

4.1 Experimental Analysis

We employ the AnyLogic Professional Discrete Event simulation toolkit [4] for the experiments. We simulate up to 40 weeks of simulation runs with the aforementioned parameters and dispatching as described in section 3.3. Measurements are taken at end of each week. No measurements are recorded during the warm up period of first four weeks. In steady state the parameters that were measured include:

- SLA measurements at each priority level
- Completion times of work in minutes (includes queue waiting times, transfer times, and service times)
- Throughput (work completed/week)
- Resource utilization (captures the busy-time of a resource)
- Number of resources that is an indication of cost

For all the above parameters the observation means and confidence intervals are reported. Whenever confidence intervals are wider, the number of weeks in simulation is increased and reported values in the paper are within $\pm 5\%$ confidence intervals. We seed the simulation with a good initial staffing solution from the Optimizer kit [15] which returned the optimal number of resources that can meet the contractual SLAs (we assume SLA adherence as a required condition for a model). Table 1 shows the distribution of work across priorities, the target resolution times and the percentage of SRs that need to be completed within the target resolution time. These values are defined based on our analysis of the real life data collected from projects at IBM.

Table 1. Work Distribution and SLA Target Times and Percentages

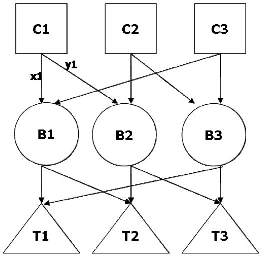
Priority of SR or Work	% Distribution	SLA Target Times (minutes)	% Meeting Target Time
1	10	240	90
2	20	480	90
3	40	720	100
4	30	1440	100

Simulation Results for Studying the Impact of Arrival Patterns and Skills on SDM

For the first set of experiments we take the work profiles with substantial overlap like Type 1 as shown in Fig. 2, and vary the workload arrival patterns. We have two type of arrival patterns, i) non-complementary workload for all customers and ii) complementary workload for customer C1 and C2 and uniform for C3. For the purpose of experiments, we differentiate resources with specialized customer knowledge (CK) and technical skills (TD). Specialized customer knowledge includes fair amount of knowledge of customer specific details of the business functions in addition to adequate relevant BD skills. The effect of skills is captured by differentiating the service times as described in section 3.2. We simulate non-complementary workload by simulating simultaneous peaks in the customer workload. In this case as shown in Table 2, we see that when all SDMs have equally skilled people with high customer knowledge (CK), then the optimal staffing required by CF, BF teams is very similar. Since the service workers have similar skills, the average completion times for work and the resource utilizations are comparable. We next simulate the environment where the skills of people in the SDMs vary. We assume in CF, people are highly well-versed with the customer domain while the people in BF and TF have comparatively lower customer knowledge. The results in Table 2 show the trend that with increasingly different levels of customer knowledge between the three models, CF increasingly tends to outperform the other two delivery models. Thus we conclude, without loss of generality, that the CF focused SDM is actually the best choice among all SDMs, when the workload arrivals offer no real benefit of work multiplexing especially when customer knowledge is an important part of the work environment.

Table 2. KPI Performance for Non-complementary, Type1 work profile

Non-complementary workload, Type 1 profile	Same Customer Knowledge			Specialized Customer Knowledge		
	CF(HI CK)	BF(HI CK)	TF(HI CK)	CF (HI CK)	BF (Med CK)	TF (Med CK)
Num Resources	105	107	120	105	108	124
Cost (USD)	9.4M (@80K/SW)	9.6M (@90K/SW)	9.6M (@90K/SW)	9.4M (@90K/SW)	9.7M (@90K/SW)	9.9M (@80K/SW)
Utilization %	49%	50%	44%	49%	56%	51%
Completion Time	141	120	130	141	130	170



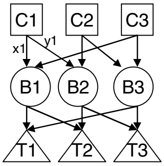
Type 1

We next simulate the scenario, when the workload arrivals are complementary. The customer knowledge is still assumed to be an important part of the environment, i.e., CF has a higher customer knowledge skill. However the fact that the workload peaks are now staggered and no longer happen simultaneously, changes the landscape of the results. Table 3 shows that in this case BF and TF show big improvements when compared to CF in terms of resource cost and utilization. The CF suffers from low resource utilization in this scenario. Both BF and TF perform well, with BF having the lowest cost, and completion times. The TF completion times are slightly higher, even though both have the same level of skills (medium CK). This can be explained due to the effect of transfer times on multi-skill work requirements in the TF model. Recall that the work coming to these teams require multiple skills for resolution. Workers in BF SDMs typically have multiple skills and work on tickets for a longer amount of time. In contrast the TF workers only work on their specific specialized skill and pass it on to the next expert resulting in higher completion time.

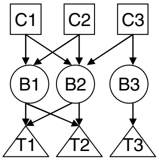
We conclude that when workloads have some complementary behavior, either BF or TF should be the SDM of choice; and the impact of complementary behavior overrides the requirement of customer knowledge in the SDM KPIs. The simulations for other profile types follow suite and for sake of brevity, results are not presented here.

Table 3. KPI Performance for Complementary, Type1 and 2 work

Complementary workload	Specialized Customer Knowledge Type 1 Profile			Specialized Customer Knowledge Type 2 Profile		
	CF (HI CK)	BF (Med CK)	TF (Med CK)	CF (HI CK)	BF (Med CK)	TF (Med CK)
Num Resources	115	103	124	115	100	112
Cost (USD)	9.9M (@90K/SW)	9.3M (@90K/SW)2	9.9M (@80K/SW)	9.9M (@90K/SW)	9.0M (@90K/SW)	8.9M (@80K/SW)
Utilization %	46%	59%	50%	46%	60%	50%
Completion Time	141	152	173	147	147	151



Type 1



Type 2

Simulation Results for Studying the Impact of Work Profiles on SDM

We simulate different work profiles on SDM that captures the level of sharing that can be achieved at the customer, business or the technical domains. We restrict to complementary workloads to study BF vs. TF. It was seen that BF clearly outperforms others in most KPIs, as shown in Table 3, where there is sufficient overlap between business functions of customers with complementary workloads. Note that this is true even though people in CF have higher skills and lower completion times when compared to BF.

We next simulate scenario for type 4 where the customers do not have many overlapping business functions, but a high technology overlap. In this case customers have a very diverse set of business functions but they all require common technologies. In this case, Table 4 shows that the TF model performs the best in terms of resource cost, completion time and utilization. This is when we assume that customer knowledge is still higher with the CF SDM. Since TF builds specialized skills, it is often believed that workers in TF are highly skilled in the technical areas of expertise. With this assumption we simulate the case where TF workers have higher skills in their individual domains but lower customer knowledge skill. In this scenario, while transfer times are reasonable (~10 to 20 minutes) TF is remains the best SDM.

The biggest drawback of the TF model in case of multi-skill work is the *hand-off* between teams, causing transfer delays. We next analyze the sensitivity of the TF performance with respect to transfer times. Fig. 3 and Fig. 4 show that TF performance degrades, both in terms of completion time and number of resources deteriorate as transfer times increase beyond 30 min. Hence, while TF is a good model for type 2 customer profiles, its sensitivity to transfer times needs to be considered. Especially when the teams are geographically distributed and transfer times are naturally higher due time-zone shifts, the benefit of using TF in a high technology overlap customer profile may be overridden.

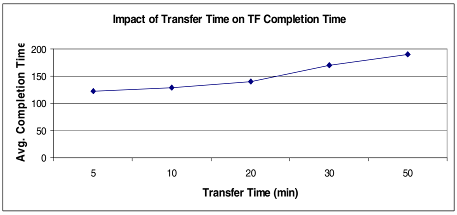


Fig. 3. Transfer time Vs Completion time (TF)

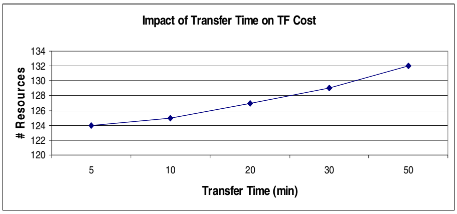
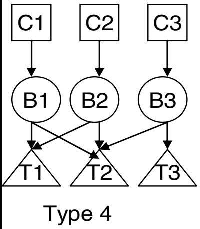


Fig. 4. Transfer time Vs Cost (TF)

Table 4. KPI Performance for Complementary, Type4 work

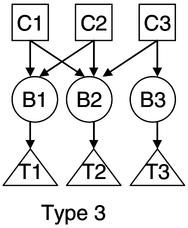
Complementary workload, Profile Map Type 4	Specialized Customer Knowledge			Specialized Customer Knowledge And Specialized Technical Domain Knowledge		
	CF (Hi CK)	BF (Med CK)	TF (Med CK)	CF (Hi CK + Med TD)	BF (Med CK + Med TD)	TF (Med CK + Hi TD)
Num Resources	110	110	120	110	110	116
Cost (USD)	9.9@(90K/SW)	9.9M@(90K/SW)	9.6M@(80K/SW)	9.9@(90K/SW)	9.9M@(90K/SW)	9.2M@(80K/SW)
Utilization %	49%	55%	52%	49%	55%	55%
Completion Time	157	155	154	157	155	150



In the last profile experiment, we simulate the profile type 3 when technology requirements are very distinct even though a lot of business function sharing is present. In this case, BF is the clear choice, even when customer skills are higher in CF and technology skills are higher in TF as show in Table 5.

Table 5. KPI Performance for complementary, Type3 work

Complementary workload, Profile Map Type 3	Specialized Customer Knowledge			Specialized Customer Knowledge And Specialized Technical Domain Knowledge		
	CF (Hi CK)	BF (Med CK)	TF (Med CK)	CF (Hi CK + Med TD)	BF (Med CK + Med TD)	TF (Med CK + Hi TD)
Num Resources	110	105	130	110	104	128
Cost (USD)	9.9@(90K/SW)	9.4M@(90K/SW)	10.4M@(80K/SW)	9.9@(90K/SW)	9.3M@(90K/SW)	10.2M@(80K/SW)
Utilization %	49%	55%	47%	49%	57%	50%
Completion Time	157	155	154	157	155	150



We conclude that customer profile maps are a very prominent factor in deciding the SDM of choice and has a bigger impact than skills (customer knowledge or technical domain expertise) on SDM KPIs. For most profiles, that have some degree of commonality in business functions, BF is the best SDM choice. When customers have diverse business requirements but common technology requirements, TF outperforms other models, as long as transfer times are reasonable (~20 min).

A note on Utilization: With respect to the utilization of resources, we realize that the averages may not accurately represent the distributions that have a lot of skew. Hence, we look at the distributions of the utilization across the three scenarios to draw conclusions on their utilization pattern. Fig 5 shows the box-plots for CF, BF and TF utilization, when sharing is high at both business and technology levels. The whiskers represent the min and max of the distributions while the '+' indicates the median. In the high sharing case for all distributions, the median lies in the centre of the box. In CF the box is equally placed between the whiskers indicating normally distributed utilization. The smaller inter-quartile range shows similar utilizations among the resources in TF, but varied in case of BF to CF. Overall Fig. 5 shows that in each scenario, peoples' utilization distributions are more symmetric. In Fig. 6 for a profile like type 4, where the business function sharing is low, the CF and BF utilizations are skewed. In CF, the median is closer to the first quartile value, indicating that more people are lowly utilized. For BF, the median is closer to the third quartile value indicating a left skew with a larger clustering of higher values in that section. However the median of the TF is centrally located indicating a more uniform distribution. This shows that from the utilization perspective, TF focused teams exhibit uniform utilization patterns irrespective of the amount of sharing available. BF shows uniform utilizations with high business function sharing customer profiles.

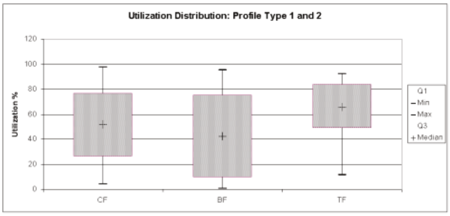


Fig. 5. Utilization Distribution: Type 1

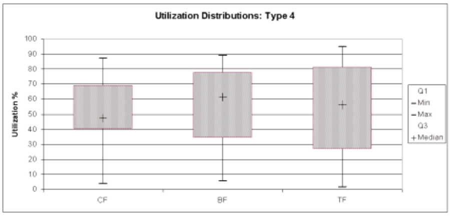


Fig. 6. Utilization Distribution: Type 4

5 Related Work

The concept of shared service has existed for a long time, for e.g., multiple departments within an organization shared services like HR, finance, IT etc. A recent study [23] of global service delivery centers revealed that shared services not only reduces costs, but also improves quality. There is also work on organizational design principles underlying an effective service delivery model [1,3,5] and resource hiring, cross-training in such models [20]. However, there is no work on generalizing the service delivery models and evaluating the pros and cons when presented with different kinds of workloads and work arrival patterns. Learning and forgetting curves in production and manufacturing industry [13] have received a lot of attention.

The service delivery work, being repetitive in nature can benefit from these results in modeling the effect of learning and forgetting on service times. This paper incorporates some of the manufacturing domain results. One of the interesting results in this body of work is [14] where the authors demonstrate that forgetting by workers in an establishment or line of production as a substantive characteristic of actual production processes is overstated and that although important and interesting, is not as influential as previous work for labor productivity has suggested. There is another line of work that studies the effects of task assignment on long term resource productivity. This is because the task assignment impacts mean learning rate, mean forgetting rate, mean prior expertise, variance of prior expertise etc and thus has a direct consequence on productivity. The work in [18] presents a heuristic approach for assigning work by taking into account all these factors. We have modeled productivity differences between various skill levels for the same skill type. How to staff, cross-train them and utilize multi-skill resources have also received adequate attention in the past and especially in the context of call-centers [7,9]. The work in [12] advocates that a flexible worker should process a task s/he is uniquely qualified for before helping others in shared tasks. This is advocated in work-in-process constrained flow-lines staffed with partially cross-trained workers with hierarchical skill sets. The effect of collaboration between teams has also been studied in work in [21] which proposes the concept of social compute unit. The work in [10] theorizes how task/team familiarity interact with team coordination complexity to influence team performance. They find that task and team familiarity are more substitutive than complementary in: Task familiarity improves performance more strongly when team familiarity is weak and vice versa. The work in [2] elaborates on the impact of high transfer times on SLAs and deals with minimizing the transfer times in the context of service tickets.

6 Conclusion

We conclude that when strict SLA adherence is a pre-requisite for service delivery, complementary nature of work arrivals and overlapping skill requirements of customers play a crucial role in the choice of SDM. Domain knowledge plays a role mainly in case of non-complementary workload. The business function focused model performs best or at par with others in most cases if costs are ignored. Technology focused model performs best for certain specific work profile combinations and is at par with BF in most cases if labor cost is the primary KPI. TF model also exhibits more uniform utilization, but suffers from high sensitivity to transfer times. Such kind of detailed analysis will give useful insights in choosing the delivery model. An interesting extension of this study would be to evaluate the conditions for hybrid SDMs.

References

1. Agarwal, S., Reddy, V.K., Sengupta, B., Bagheri, S., Ratakonda, K.: Organizing Shared Delivery Systems. In: Proc. of 2nd International Conference on Services in Emerging Markets, India (2011)
2. Agarwal, S., Sindhgatta, R., Sengupta, B.: SmartDispatch: enabling efficient ticket dispatch in an IT service environment. In: Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012 (2012)

3. Alter, S.: Service System Fundamentals: Work System, Value Chain, and Life Cycle. IBM Systems Journal 47(1), 71–85 (2008)
4. Anylogic Tutorial, How to build a combined agent based/system dynamics model in Anylogic. In: System Dynamics Conference (2008), `http://www.xjtek.com/anylogic/articles/13/`
5. Assembly Optimization: A Distinct Approach to Global Delivery, IBM GBS White Paper (2010)
6. Banerjee, D., Dasgupta, G.B., Desai, N.: Simulation-based evaluation of dispatching policies in service systems. In: Winter Simulation Conference (2011)
7. Cezik, M.T., L'Ecuyer, P.: Staffing multi-skill call centers via linear programming and simulation. Management Science Journal (2006)
8. Diao, Y., Heching, A., Northcutt, D., Stark, G.: Modeling a complex global service delivery system. In: Winter Simulation Conference 2011 (2011)
9. Easton, F.F.: Staffing, Cross-training, and Scheduling with Cross-trained Workers in Extended-hour Service Operations. Robert H. Brethen Operations Management Institute (2011) (manuscript)
10. Espinosa, J.A., Slaughter, S.A., Kraut, R.E., Herbsleb, J.D.: Familiarity, Complexity, and Team Performance in Geographically Distributed Software Development. Organization Science 18(4), 613–630 (2007)
11. Franzese, L.A., Fioroni, M.M., de Freitas Filho, P.J., Botter, R.C.: Comparison of Call Center Models. In: Proc. of the Conference on Winter Simulation (2009)
12. Gel, E.S., Hopp Wallace, J., Van Oyen, M.P.: Hierarchical cross-training in work-in-process-constrained systems. IIE Transactions, 39 (2007)
13. Jaber, M.Y., Bonney, M.: A comparative study of learning curves with forgetting. Applied Mathematical Modelling 21, 523–531 (1997)
14. Kleiner, M.M., Nickelsburg, J., Pilarski, A.: Organizational and Individual Learning and Forgetting. Industrial and Labour Relations Review 65(1) (2011)
15. Laguna, M.: Optimization of complex systems with optquest. OptQuest for Crystal Ball User Manual Decisioneering (1998)
16. Lo, C.F.: The Sum and Difference of Two Lognormal Random Variables. Journal of Applied Mathematics 2012, Article ID 838397, 13 pages (2012)
17. Narayanan, C.L., Dasgupta, G., Desai, N.: Learning to impart skills to service workers via challenging task assignments. IBM Technical Report (2012)
18. Nembhard, D.A.: Heuristic approach for assigning workers to tasks based on individual learning rates. Int. Journal Prod. Res. 39(9) (2001)
19. Ramaswamy, L., Banavar, G.: A Formal Model of Service Delivery. In: Proc. of the IEEE International Conference on Service Computing (2008)
20. Subramanian, D., An, L.: Optimal Resource Action Planning Analytics for Services Delivery Using Hiring, Contracting & Cross-Training of Various Skills. In: Proc. of IEEE SCC (2008)
21. Sengupta, B., Jain, A., Bhattacharya, K., Truong, H.-L., Dustdar, S.: Who do you call? Problem resolution through social compute units. In: Liu, C., Ludwig, H., Toumani, F., Yu, Q. (eds.) ICSOC 2012. LNCS, vol. 7636, pp. 48–62. Springer, Heidelberg (2012)
22. Spohrer, J., Maglio, P.P., Bailey, J., Gruhl, D.: Steps Toward a Science of Service Systems. IEEE Computer 40(1), 71–77 (2007)
23. Shared Services & Outsourcing Network (SSON) and The Hackett Group, Global service center benchmark study (2009)
24. Verma, A., Desai, N., Bhamidipaty, A., Jain, A.N., Barnes, S., Nallacherry, J., Roy, S.: Automated Optimal Dispatching of Service Requests. In: Proc. of the SRII Global Conference (2011)