INFORMATION
TECHNOLOGY JOURNAL

Time Analysis for BPMN Gateways Using Queuing Theory

Mahmoud Hasan Mahmoud Saheb

College of IT and Computer Engineering, Palestine Polytechnic University, Hebron, Palestine

Abstract: Currently Business Process Management connected directly with information technology, either by automating some of the tasks or by simulating the process for understanding and enhancing. Business process modeling is an important phase of business process management phases. One of the steps needed to enhance a business process model is analyzing the timing and the queuing behavior for each task in the process. Traditional methods as Critical Path Analysis (CPM) and Program Evaluation and Review Technique (PERT) cannot be used for such analysis. This study presents an analytical approach to analyze the business process models using queuing systems classifications for the basic BPMN2 gateways as M/M/1, M/M/c and M/D/1. Five types of the BPMN gateways have been analyzed depending on their specifications to define the inflow and the outflow and the probability for producing and consuming tokens. An algorithm for computing arrival rate and average waiting time and illustrative example for our approach has been presented.

Key words: BPMN, business process modeling, queuing, time analysis, simulation

INTRODUCTION

Business Process Modeling plays an important role in Business Process Management (BPM), to manage a process we have to model the current process "as-is" model then to model it as "to-be" model after analyzing and enhancing the "as is" model. Analyzing the current process may include the time analysis; the time needed to complete a process, the critical path and waiting time at each stage of the process. This is important for managing the resources to achieve the goals of the organization running the process. The combination of BPM and Service-oriented (SOA) architecture benefits information technology professionals and business users.

Enhancing the current process may not include automating the involved tasks, but in many cases we do that. Currently BPM connected directly with information technology, either by automating the task, or by simulating the process for understanding and enhancing. The current version of Business Process Model Notation (BPMN2.0) has been designed with full intention for automating the business processes; it includes a mapping of the business process model to Business Process Execution Language (BPEL) in its specifications.

It may appear that it is easy to use Critical Path Method (CPM) or Program Evaluation and Review Technique (PERT) for analyzing the timing and queuing behavior of the business process model after converting this model to single-source and single-sink graph which is not quite true. These methods cannot be directly implemented due to reasons related with the "tokens" concept in BPMN; gateways may consumes or produce tokens. Other reasons related with the probability of the inflow and outflow of these gateways. Simulation and heuristic methods has been used to analyze the stochastic networks because of the computation complexity involved in developing complete formula for such networks (Hashemin and Ghomi, 2012; Pritsker, 1966; Ibe, 2011).

In this study, we propose the using of queuing theory and the BPMN gateways behavior to analyze the Business Process Models which presented using BPMN. We have used an analytical approach to develop the simulation process by used the specifications of the BPMN gateways to define the probability of inflow and outflow of each gateway, a queuing model defined for each gateway according to the nature of the arrival and processing behavior and the number of services.

BUSINESS PROCESS MODELING AND MANAGEMENT

Business Process is a series of tasks for fulfilling an organization goal; this goal could be a service or a product (Rummler and Brache, 1995). Business Process Management (BPM) is a set of structured activates for handling and directing the organization business processes to achieve the goals of the organization. BPM's activities include business process identification, documentation, analysis and improvement. BPM will improve the competitive advantage of the organization in the daily changing market demands. The business process improvements may remove or add tasks, it may parallelize sequential tasks and it may automate some tasks.
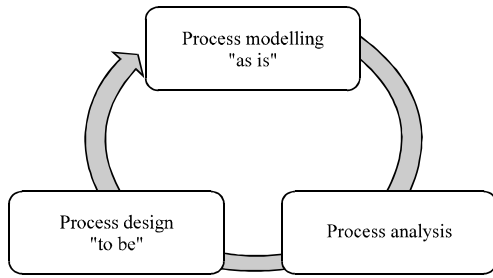
Fig. 1: BPM phases

Business process modeling is an activity used in BPM to identify and represent the business process tasks and their interactions. Many methods can be used for modeling including flowchart and UML activity diagrams.

The first phase in BPM, as presented in Fig. 1, is modeling the current business process "as is", analyzing the current process produces the modified "to be" model which will be "as is" model for the next stage of BPM.

Business process Modeling and Business Process management share the same acronym (BPM), in this study we use BPM for Business Process Management. Business Process modeling was first used in 1960s in systems engineering field (Brun and Garcia, 2000).

BPMN is a graphical language for representing the logical steps of business process model. The current version of BPMN is BPMN2.0, the final version of its specifications including syntax and semantics was announced in January 2011 and it is currently managed by the Object Management Group (OMG). The initial developer of BPMN was Business Process Management Initiative (BPMI) OMG (Object Management Group) 2011. BPMN supports only the modeling of business processes which is one of the enterprise models, its elements are tightly coupled with business processes. So, it cannot be used for modeling organizational structures, business communication, business cohesion, business strategy, maturity and responsibility. Although BPMN represents flow of data through massages and the association of tasks with data artifacts, but it cannot be used for data or dataflow modeling.

BPMN provides four core elements for modeling a business process (OMG (Object Management Group) 2011). These elements are events, activity, gateway and sequence flow. Table 1 summarizes these elements.

BPMN Gateways are flow control elements used to control how the process flows; only sequence flow will affect the flow of work and message flow can not affect the flow of work. BPMN gateways includes five types, these types are Exclusive, inclusive, Parallel, Event-Based and complex.

Table 1: Icons and definitions of BPMN2.0 basic elements

Name	Icon	Element definition
Event	◯	An event is represented with a circle and denotes something that happens during the course of the process, affecting the process flow and normally has a trigger or result
Activity	▢	Activities represent work or tasks carried out by members of the organization or automated internally or performed by an external system or user. Activities can be atomic or compound
Gateway	◇	Determines forking and merging of paths, depending on the conditions expressed
Flow	→	Shows the order of the activities to be performed

Table 2: BPMN Gateways

Icon	Name
◈ ◇	Exclusive gateway with and without marker
◉	Inclusive gateway
⊕	Parallel gateway
✳	Complex gateway
◎	Event-based gateway

Gateways are used for conversion and diversion of the flow within the process. Gateways may consume or generate tokens according to the gate specifications and the execution semantics, they are similar to Activities in this regard, but gate ways do not represent a work. We shod notice that gateways consume no time since they do not represent work, but they may cause a delay as waiting time. Tokens are abstract elements showing execution point in the process, in one process we may find many tokens for the same instance of a process indicating that many activities are running for the same process instance. Table 2 shows the graphical notation of different types of BPMN gateways.

TIME ANALYSIS METHODS

The path with longest average execution time from start to end of a workflow or business process model presented as directed single source single sink graph is called critical path. Any delay in any task on the critical path will delay the project completion (Chang *et al.*, 2002). Well known method for determining the critical path is called Critical Path Method (CPM).The maximum time amount which an activity can be delayed without delaying the overall project or process end is called slake, critical tasks have zero slake.

CPM can be used to compute the maximum needed time to complete a process, assuming that there is no waiting or queuing. So, using CPM will not give us the exact maximum time to complete a process if there is waiting time.

PERT models the activities as single source and single sink directed acyclic graph where nodes presents events (end or start of activities or gateways) and edges presents activities (tasks).

PERT and CPM has been traditional used to analyze critical path in project domain. But it is more complicated to shorten a critical path in workflow schema than in PERT chart (Li and Zhan, 2005) because they cannot support the selective and alternative structures of workflow models and business process models and the tokens concept in BPMN.

Graphical Evaluation and Review Technique (GERT) is a stochastic network analysis technique used in project management that allows for conditional and the probabilistic treatment of the logical relationships between the project's activities following randomly determined sequence of observations. GERT is similar to PERT but have the advantages of allowing for looping, branching and multiple project end results (Pritsker, 1966). In GERT branches of the project network can have a probabilistic chance of not having to be accomplished whereas in PERT/CPM all branches must be accomplished.

GERT has been developed to Q-GERT and then to SLAM-II. These systems are difficult to be used due to their computing complexity.

Martinez developed Stroboscope (Martinez and Ioannou, 1994) which is a simulation programming language and system for modeling complex processes in a variety of areas ranging from construction, to manufacturing, to hospital operations.

QUEUING MODELS

In business process model, many processes can be active at the same time, each process will be presented by one or more tokens and many tokens of different processes may be waiting to be served. This situation can be presented as queue or queues at each element of the business process model. The business process model can be presented as network of queues.

Little's theorem states that the average number of customers (L) can be determined from the following equation (Little, 2008):

$$L = \lambda \, \mu$$

where, λ is the average customer arrival rate and μ is the average service time for a customer. In our discussions customers represents tokens.

Assume customers arrive at the rate of 20 per hour and stay an average of 0.7 h. This means the average number of customers in the process at any time is 14:

$$L = 20 \times 0.7 = 14$$

The time and queuing analyze depend on the nature of the customer arrival (A), the process behavior (S) and the number of servers (n), this combination is denoted as A/S/n which vary for different systems. The symbols A and S can be any of the following:

- **M (Markov):** The arrival or service time is negative exponentially distributed. It can be applied for any system with a very large number of independent customers which can be approximated as a Poisson process
- **D (Deterministic):** All customers have the same value. The service time can be assumed to be same for all customers
- **G (General):** Any arbitrary probability distribution

Little's theorem can be applied to systems or subsystems; in our case this can be applied to the process or the tasks within the process. For an M/M/1 queue (Bose, 2002), there is one server with an exponential service rate μ. The arrival rate to the system is λ and $\lambda < \mu$ for steady state, the waiting area is infinite. When the queue is stable, we will observe busy and idle periods continuously alternating. We can derive that the average number of customers in the system (Ls), waiting and being served, as:

$$Ls = \frac{\lambda}{\mu - \lambda} \ \text{Or} \ \frac{\rho}{\rho - 1}$$

where, the utilization of the system is $\rho = \lambda/\mu$.

The average waiting time in the system (Ws) is Ws = Wq + average service time; Ws includes the waiting time in the queue plus the average service time to get the service and knowing that the average service time is $1/\mu$ and using Little's law $L = \lambda \, \mu$, then the average time a customer spends in the system is Ws = $1/\mu$-λ.

For M/D/1 the arrival rate is λ, the service rate μ is deterministic and there is one server. ρ, L, Wq and Ws are defined as the following (Brun and Garcia, 2000):

$$\rho = \frac{\lambda}{\mu}$$

$$L = \frac{\rho^2}{2(1-\rho)}$$

$$Wq = \frac{1}{2\mu}\left(\frac{\rho}{1-\rho}\right)$$

$$Wq = \frac{1}{2\mu}\left(\frac{2-\rho}{1-\rho}\right)$$

For an M/M/c queue (Bose, 2002), there are c servers with an exponential service rate μ. The arrival rate to the system is λ and $\lambda < \mu c$ for steady state, the waiting area is infinite.

For this queue the utilization of the system is:

$$\rho = \lambda/\mu e$$

P_0 = Probability that no one is being served:

$$P_0 = \frac{1}{1 + c\rho + \frac{c^2\rho^2}{2!} + \cdots + \frac{c^{c-1}\rho^{c-1}}{(c-1)!} + \frac{c^c\rho^c}{c!(1-\rho)!}}$$

Pn = Probability that n are in system including one being served:

$$Pn = \frac{\frac{P_0 C^n \rho^n}{n!} \quad n = 0.1.....c-1}{\frac{P_0 C^c \rho^n}{c!} \quad n = c, c+1}$$

Lq = Average number in system excluding what being served:

$$Lq = \frac{\frac{P_0 C^c \rho^{c+1}}{C!}}{C!(1-\rho)^2}$$

Wq = average waiting time:

$$Wq = \frac{Lq}{\lambda}$$

Ws = Average time in system including ones being served

Ws = $Wq + 1/\mu$

L = Average number in system including ones being served:

$$L = \lambda Ws = Lq + c\rho$$

Table 3 summarizes the average waiting time for M/D/1, M/M/1 and M/M/c queues models (Mannering and Washburn, 2012).

Business process model represents a network of tasks and from queuing point of view it is a network of queues with M nodes and λi is the external arrival rate into node i and Pi,j is the routing probability from node i to node j. The effective arrival λ' rates for node j can be obtained as follows (Bose, 2002; Yan and Veeraraghavan, 2004):

$$\lambda_j' = \lambda_j + \sum_{i=1}^{j}\lambda_j' P_{i,j}$$

If there is no external arrival entering the system, as in our case; one source one sink, then λ'_j:

$$\lambda_j' = \sum_{i=1}^{j}\lambda_j' P_{i,j}$$

Jackson's Theorem provides a general product-form solution for both feed both open and close queuing networks, one of its assumptions says (Ibe, 2011).

Once a customer is served at queue i, it joins each queue j with probability Pi,j or leave the system with probability:

$$1 - \sum_{j=1}^{k} Pi,j$$

We cannot use Jackson's Theorem because the sum of possibilities for all branches leaving the gate way, for some gateways, may be greater than 1; i.e., parallel gate way with n outflows has probability of 1 for each outflow according to its specifications in BPMN. For each inflow token it produces one token for each outflow.

Table 3: Queue models summarization

Queue model	Utilization rate	Average length of queue	Average time waiting in queue	Average time spent in system
M/D/1	$\rho = \frac{\lambda}{\mu}$	$L = \frac{\rho^2}{2(1-\rho)}$	$Ws = \frac{1}{2\mu}\left(\frac{2-\rho}{1-\rho}\right)$	$Wq = \frac{1}{2\mu}\left(\frac{\rho}{1-\rho}\right)$
M/M/1	$\rho = \frac{\lambda}{\mu}$	$L = \frac{\rho}{(1-\rho)}$	$Ws = \frac{1}{\mu-\lambda}$	$Wq = \frac{1}{\mu}\left(\frac{\lambda}{\mu-\lambda}\right)$
M/M/c	$\rho = \frac{\lambda}{\mu c}$	$L = \frac{P^0\rho^{c+1}}{c!c}\left[\frac{1}{(1-\rho/c)^2}\right]$	$Ws = \frac{\rho+\overline{Q}}{\lambda}$	$Wq = \frac{\rho+\overline{Q}}{\lambda} - \frac{1}{\mu}$

QUEUING MODELS FOR BPMN2.0 GATEWAYS

Queuing theory can be applied to systems or sub system; in our case we can apply it for task or process. Each gateway can be presented as one of the queuing models if it represents a queue and it will be represented as zero-time's task if it has no queue since; gateways do not present work, they presents splitting or joining gates. Each task can be presented as M/D/1, M/M/1 or as M/M/c according to its specifications. According to the gateway specifications which define how it deals with the inflow tokens and its outflow tokens we have determined queuing behavior and the probability of its outflow as shown in Table 4.

In Table 4 we are dealing with diversion or conversion gateways but not mixed gateways; i.e gateway used for conversion and diversion at the same time. And we have to note that:

- Column 2 in Table 4 shows that routing probability must be assigned for each split gateway out branch. For parallel gate way it is always 100% for each branch. For each inflow branch for each join gateway with waiting, probability must assign or it can be computed from the split gateways that forms the inflow branches
- Column 3 in Table 4 shows that Split Event-based gateway is the only split gateway that needs waiting before activating one of its outflows
- Columns 4 in Table 4 shows that all join gateways need waiting time except Exclusive gateways. Event-based gateway cannot be used for joining

For single source, single sink and positive Acyclic Directed Graph (ADG), we can use Algorithm 1 and Table 4 to compute the Ws and λ for each task. The algorithm starts by assigning λ for the start event, then moving breadth first to compute λ for each task. Table 4 shows the gateways that need routing probability computing and the gates that need waiting either for synchronizing the input tokens as parallel join gateway,

or waiting for event as event based gateway. Outflow routing probability should be provided from historical data, inflow probability will be computed.

We can compute λ for each inflow of a task in sequence of tasks is the same as the first task in the sequence. And we can compute λ for the first task in any sequence of tasks from the routing probability of the sequence and the λ of the gateway preceding the sequence.

BPMN gateways not appearing in Table 1 and 2 are not included in this algorithm.

At the STOROSCOPE can be used to analyze the network.

Algorithm 1: ()
```
Begin
    Convert BPM to directed acyclic graph;
    Assign λ for the start event;
    While not end event move breadth first
        Foreach split gateway
            Assign Routing Probability P to each outflow using Table-4
            Compute λ for each outflow using P
            If gateway is Event-Based then add dummy task with event average
            waiting time
        Foreach Task in the graph using breadth first
            Assign μ and c
            If S = M and n = 1 then
                Compute Ws using M/M/1 formula;
            If S= M and n>1 then
                Compute Ws using M/M/c formula;
            If S = D and n = 1 then
                Compute Ws using M/D/1 formula;
        For each join gateway with waiting
            MaxWS = Max(summation of Ws of tasks of each inflow tasks)
            If gateway type is Parallel-gateway
                For each inflow
                    Add dummy task with waiting time = MaxWs-summation
                    of Ws of tasks of that inflow
            If gateway type is Inclusive-gateway
                For each active inflow
                    Add dummy task with waiting time = MaxWs-summation of
                    Ws of tasks of that inflow
            If gateway type is Complex-gateway
                For each candidate inflow
                    Add dummy task with waiting time = MaxWs-summation of
                    Ws of tasks of that inflow
        endwhile
        call STOROSCOPE;
end
```

Table 4: BPMN Gateways

Icon	Name	Wait for diverge gateway (Split)	Wait for converge gateway (Join)	Need for split/join routing-probability (P)
	Exclusive gateway with and without marker	No wait tokens will split according to the condition selectivity in only one outflow	No wait when first inflow token arrives others tokens of the same process instance are ignored	Yes/Yes
	Inclusive gateway	No wait token split into one or more branches	Wait for all activated inflow branches	Yes/Yes
	Parallel gateway	No wait the same number of tokens will be generated for each outflow	Wait for all incoming flows before triggering the flow through all its outgoing sequence flows	100%/100% For all Branches
	Complex gateway	No wait a boolean condition is evaluated to determine whether a sequence flow receives a token	Wait for a specific number of incoming tokens; candidate N out of M join	Yes/Yes
	Event-based gateway	Wait for the first event	Not aplicable can not be used as join gateway	Yes/-

ILLUSTRATIVE EXAMPLE AND DISCUSSION

Some of the previous studies discussed the queuing behavior of the workflow gateways by simplifying the mathematical formulas for waiting time at these gateways. Other studies tried to find the critical path for workflows diagrams. This study analyzes the timing and tokens behavior according to the BPMN2.0 specifications, it also provide an algorithm for computing the waiting time for business process model.

Business process model defines a process as set of flow elements that consists of three types of flow nodes connected by sequence flows. Flow nodes may be: events, activities and gateways.

Each time a process receives a new start event, a new instance of that begins executing, so the process may have many process instances and each instance may be in different stage. A token is an imaginary object moving in the process, a token navigate from a source to the sink via a sequence flow (White and Miers, 2008). A flow object executes when it has token on one or more of its input flows; when a flow object starts to execute it takes tokens of its input flows, when a flow object has finished executing it offers a tokens on one or more of its output flows.

Figure 2 presents a business process model using BPMN; the model consists of two events, three gateways and four tasks. The time unit in our example is hour. Lambda in the start event indicates that the process instance rate is 20 per hour. Task 1 has one server or person; it can handle one process instance at a time and the service rate is 25 per hour. The exclusive gateway receives all the tokens produced from task 1 and passes the token of each process instance to Task 2 or 3 depending on the evaluation of cond-1, 70% of the tokens passes task-1 will be routed to Task 2, the other 30% will be routed to Task 3. All Tasks can handle one token at a time, but Task 2 can handle 3 tokens at the same time. The splitting parallel gateway produces one token to Task 4 and one token to Task 5 as outflow for each token received from Task 3, no waiting time at this gateway. The joining parallel gateway produces one token as outflow to the end event for every two tokens related to one process instance received from Task 4 and 5; note that there is waiting time on this parallel gateway.

Table 5 uses the formulas that have been developed in section 4 and Algorithm 1 to computing the average spent time for each token at each task (Ws) including waiting and processing time.

Using CPM analysis and Ws as duration for each task will produce Gant chart in Fig. 3. Noting that we have added two dummy tasks with Ws = 0; these tasks are B1 to join T4 and T5 and E to join T3 and B1. The critical path is T1, T3, T5, B1, E. CPM assumes that all tasks will be
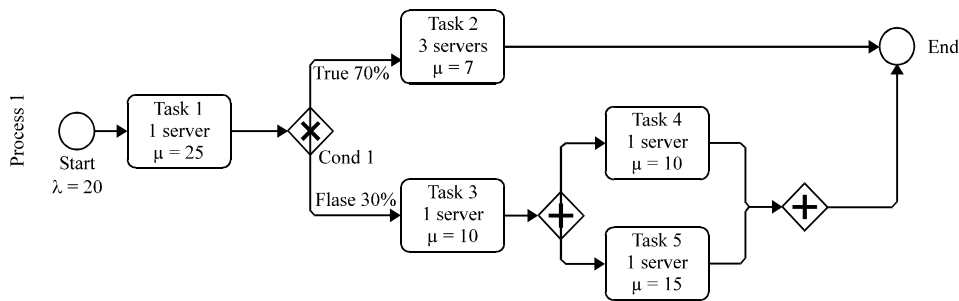

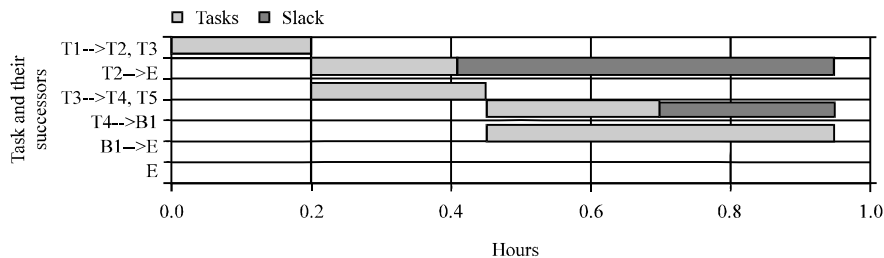Fig. 2: Sample BPM using BPMN2.0 modeled using BizAgi

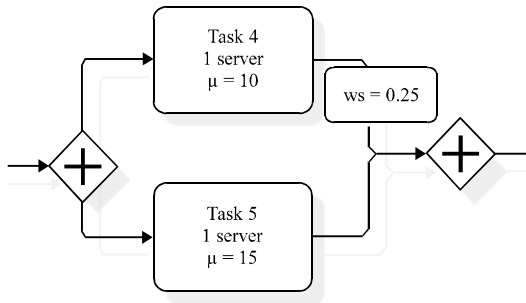Fig. 3: Gant chart showing the CPM analysis of Fig. 2

Fig. 4: Dummy task with Ws = 0.25

Table 5: Waiting time at each task of Fig. 2

Task	λ	μ	c	ρ	Ls	Lq	Wq	Ws
1	20	25	1	0.80	4.00	3.20	0.16	0.2
2	20*70% = 14	7	3	0.67	2.89	0.89	0.06	0.21
3	20*30% = 6	10	1	0.60	1.50	0.90	0.15	0.25
4	6	10	1	0.60	1.50	0.90	0.15	0.25
5	6	8	1	0.75	3.00	2.25	0.38	0.50

activated and completed to complete the process. This is not true for business processes; 30% of our process instances will be finished by completing T1 and then T2 and E with total average time is 0.41 h and 70% of our process instances will be finished by completing T1, T3, (T4+T5 parallel), B1 and E, with total average time is 0.95 h.

Algorithm 1 will produce the needed data to analyze and simulate the moving tokens in the business process. And it will add dummy tasks with waiting time according to the semantics of the BPMN gateways. Figure 4 shows the added dummy task with Ws = 0.25 h after Task-4 since it will be joined using parallel gateway.

CONCLUSION AND FUTURE WORKS

Although it is true that the assumptions of queuing theory may be too restrictive to be able to exactly model a real Business Process, but it can be used with specialized tools to simulate, analyze and visualize the tokens behavior of a process instance. Pert and CPM cannot be used to analyze Business Process because they cannot support the selective and alternative structures of workflow models and business process models and the tokens concept in BPMN. Business process model can be presented as network of queues but we cannot use Jackson's Theorem because the sum of possibilities for all branches leaving the gate way, for some gateways, may be greater than 1; i.e., parallel gate way with n outflows has probability of 1 for each outflow according to its specifications in BPMN. For each inflow token it produces one token for each outflow.

For the mentioned reasons and the computation complexity involved in developing a complete formula for such networks the simulation and heuristic methods has been used to analyze the stochastic networks.

In this study, we have analyzed five of the BPMN2.0 gateways; conversion diversion behavior, tokens production and consuming and waiting time for synchronization. An algorithm for computing Lambda and Ws for each task and gateway has been presented. And illustrative example showing the presented concepts has been presented.

Additional work should be conducted to include all the gateways elements and all the tasks types such as loop tasks, multiple instance tasks and compensation tasks.

REFERENCES

Bose, S., 2002. An Introduction to Queuing Systems. Plenum Publishers, Kluwer.

Brun, O. and J.M. Garcia, 2000. Analytical solution of finite capacity M/D/1 Queues. J. Applied Prob., 37: 1092-1098.

Chang, D.H., J.H. Son and M.H. Kim, 2002. Critical path identification in the context of a workflow. Inf. Soft. Technol., 44: 405-417.

Hashemin, S.S. and S.M.T.F. Ghomi, 2012. Constrained consumable resource allocation in alternative stochastic networks via multi-objective decision making. J. Indust. Eng. Int., Vol. 8, 10.1186/2251-712X-8-18

Ibe, O.C., 2011. Fundamentals of Stochastic Networks. John Wiley and Sons, Inc., New York, Pages: 312.

Li, H. and D. Zhan, 2005. Workflow timed critical path optimization. Nature Sci., 3: 65-74.

Little, J.D., 2008. Little's Law. Int. Series Operat. Res. Manage. Sci., Vol. 115.

Mannering, F.L. and S.S. Washburn, 2012. Principles of Highway Engineering and Traffic Analysis. 5th Edn., Chapter 5, Wiley, New York, USA., ISBN13: 9781118120149, Pages: 352.

Martinez, J.C. and P.G. Ioannou, 1994. General purpose simulation with stroboscope. Proceedings of the 26th conference on Winter Simulation Conference, December 11-14, 1994, San Diego, CA, USA, PP: 1159-1166.

OMG (Object Management Group), 2011. Business Process Model And Notation (BPMN) version 2.0. Retrieved December 20, 2012, http://www.omg.org/spec/BPMN/2.0/

Pritsker, A., 1966. Graphical evaluation and review technique. Rand Corporation, Rand Memorandum RM-4973-NASA, California.

Rummler, G.A. and A.P. Brache, 1995. Improving Performance: How to Manage the White Space on the Organisation Chart. Jossey-Bass, San Francisco, ISBN-10: 0787900907, pp: 256.

White, S.A. and D. Miers, 2008. BPMN Modeling and Reference Guide: Understanding and Using BPMN. Future Strategies Inc., Book Divisio, Lighthouse Point, USA., ISBN: 9780977752720, Pages: 226.

Yan, T. and M. Veeraraghavan, 2004. Network of Queues. Retrieved from http://www.ece.virginia.edu/mv/edu/715/lectures/QNet.pdf