An Overview of Operational Laws

Prepared by

Mohamed Mohamed Mohamed Faragallah
Computer Programs Operator
Port-Said Container & Cargo Handling Company

This paper provides an overview of operational laws. It addresses the concepts of operational quantities, operational variables and the analytical methods. Generally, these concepts are addressed to be used in constructing or analyzing some kinds of optimization problems. These problems are especially constructed to conduct the optimization processes for the operational systems which look like closed queueing networks. Basically, operational quantities are sampled during observing the studied system in a finite period of time. Thereafter, the other values of operational laws are derived from operational quantities. Analytical methods including asymptotic bounds, bottleneck analysis and mean value analysis are performed to analysis operational systems. Asymptotic bounds are used to represent both of optimistic and pessimistic bounds of systems' throughput and response time. Bottleneck analysis is used to guide decision makers toward the bottleneck device in their systems. Mean value analysis is used to provide the systems' performance indicators with each additional customer in the system.

وتقدم هذه الورقة لمحة عامة عن القوانين التشغيلية. و تتناول مفاهيم كل من الكميات التشغيلية والمتغيرات التشغيلية والطرق التحليلية. وعموما يتم استعراض هذه المفاهيم نظرا لاهميتها في بناء أو تحليل بعض أنواع مشاكل تحقيق الامثلية للنظم التشغيلية التي تبدو وكأنها شبكات الطابور المغلقة. فى البداية يتم معاينه الكميات التشغيلية أثناء مراقبة النظام المدروس في فترة زمنية محددة. بعد ذلك يتم استنباط المفاهيم الأخرى المذكورة من نتائج مراقبة الكميات التشغيلية. وتشمل الأساليب التحليلية تحليل حدود التقارب وتحليل عنق الزجاجة و كذلك تحليل القيمة المتوسطة. يستخدم تحليل حدود التقارب فى تمثيل كل من الحدود المتفائلة والمتشائمة للانتاجية وايضا وقت الاستجابة في الانظمة التشغيلية التى لها شكل شبكات الطابور المغلقة. يستخدم تحليل عنق الزجاجة لتوجيه صناع القرار نحو جهاز عنق الزجاجة ان وجد في أنظمتهم. كذلك تحليل القيمة المتوسطة لتوقع مؤشرات أداء النظم مع دخول كل عميل إضافي الى النظام.

Operational Quantities

Operational quantities, in queueing network systems, are measured during a finite observation period. Hence, assuming that a queuing network system is observed during a finite period of (T) seconds, the following operational quantities can been collected for each device $i = 1, 2, \ldots, k$ as shown in the following table,

(T)	Length of the observation period.
(A)	Number of arrivals which enter the queueing network system during the observation period.
(B)	Time that is consumed by a facility within a queuing network in providing service during the observation period $(B \leq T)$.
(C_0)	Number of customers that have been serviced during the observation period.

Operational Variables

Operational variables are considered important elements in operational analysis study. The concept of operational variables is considered wider than the concept of basic quantities. Operational variables terms carry either basic quantities or the values which are predicted by using laws derived from the basic quantities. Operational variables are computed under two main assumptions. (1) The studied system is in operational equilibrium state. So, the flow within the system is in balance state. The number of arrivals (A) equals or approaches the number of completions (C_0). (2) the arrival rate of the requests of a device in the system has to be independent of the length of the queues at the remaining devices. Thus, its interactions only occur via its own queue. According to the pervious assumptions, the relations among operational variables can be listed as follows,

Arrival rate is the number of arrivals per time $= \frac{A}{T}$

Throughput rate is the number of completions per time $= \frac{C_0}{T}$

Utilization rate is the busy time within observation period $= \frac{B}{T}$

Mean service time the busy time divided by completions$= \frac{B}{C_0}$

Operational Laws

Operational laws have been described as the simple equations that are derived from direct measurement of the queueing network systems. They are used to describe the average behavior of the facilities in such

systems. These laws are very general and don't need any assumptions about the behavior of the parameters which characterize the queuing network systems. These parameters are the probability distributions of either inter-arrival or service times.

1- Utilization Law (μ)

Utilization law measures the utilizing extent of a certain resource, facility, during the working time. It can be defined as the fraction of time in which a resource is busy during the observation period. It can be also defined as the throughput rate (x_i) multiplied by the service time (s_i) at device (i),

$$\therefore Utilization\ \mu_i = \frac{C_i \times B_i}{C_i \times T} = \frac{C_i}{T} \times \frac{B_i}{C_i} = x_i s_i$$

2- Visit Ratio (v)

Visit ratio expresses the mean number of jobs per a request on a certain device within queuing network system. It can be calculated individually from the job flow balance equation. Visit ratio is considered the proportion of a facility's productivity to whole network productivity. Accordingly, visit ratio can be modeled as,

$$v_i = \frac{x_i}{X}$$

On the other hand, visit ratio (v_i) can be also defined as the mean number of completions which are serviced at device(i) in a queueing network system. Therefore, according to such definition the visit ratio can be modeled as,

$$v_i = \frac{c_i}{C_0}$$

3- Forced Flow Law (Throughput)

Forced flow law states that the throughput of different components in a queueing network system is proportional to the number of times that each component needs to handle a request. Therefore, the flow in all parts of a system must be consistent where the number of arrivals equals the number of completions or approach it. Exactly, throughput of the facility (i) is considered the result of multiplying system's throughput by its visit ratio,

$$x_i = X v_i$$

4- Mean Response Time Law (R)

Mean response time for the facility (i) is defined as the average time which is accumulated on device (i) until it completes one request. Therefore, the mean response time for the facility (i) is,

$$R_i = \frac{B_i}{c_i}$$

But, in queueing network system level, the response time is generally defined as the average number of seconds that are consumed by one request from the moment it enters to the moment it leaves. Thus, the mean response time law can be,

$$R = \frac{T}{C_0}$$

5- Little's Law

Little's law is a very simple and general relationship where just stationary assumptions about the underlying stochastic processes are required regardless any other considerations. Little's law, (named after J.D.C. little 1961), confirms that the mean response time per a customer's visit to a queueing network system is the mean of queue length (N) divided by the output rate $\left(R = \frac{N}{X}\right)$. Likewise, if the system is in steady state, the number of arrivals must equal the number completions. Therefore, the pending number of customers (N) in such system must equal the multiplying of the throughput and the average response time which can be denoted as,

$$N = X \times R$$

On the other hand, in queuing networks with (N) customers, the queues (Q) in front of facilities increase in parallel with the customers' numbers in queuing network. The queue length in front of facility (i) is affected by many variables such as the throughput of the system, numbers of customers, facility's response time, facility's service time and facility's visit ratio. Thus, according to little's law the queue in front of a facility (i) can be calculated,

$$Q_i(N) = R_i(N)x_i(N)$$

Where

$$x_i(N) = X(N)v_i$$
$$R_i(N) = R(N)v_i$$

7- General Response Time Law

General response time law is used to calculate the subsystem's response time (R_i), facility level, or for whole network queuing system per a job. Basically, the general response time (R) for an open or closed network system or (R_i) for subsystems depends mainly upon applying little's law on the system. Applying the little's law on the (i^{th}) resource, the mean number of customers $Q_i(N)$ and the response time of the system *(R)* are,

$$Q_i(N) = x_i R_i, \qquad R = \sum_{i=1}^{M} R_i v_i$$

8- Interactive Response Time Law

Interactive response time law addresses the response time of delay centers $R(N)$ in parallel with the response time of queuing servers. Accordingly, in closed queueing network system, customers *(N)* are forced to stay (Z) seconds in delay centers until submitting a new transaction. So, the interactive response time is,

$$R(N) = \frac{N}{X(N)} - Z, or\ R(N) + Z = \frac{N}{X(N)}$$

$$\therefore X(N) = \frac{N}{R(N) + Z}$$

Definitely, the mentioned response time includes (1) the time that customer waits in front of each server in queues, (2) the time consumed in taking service and (3) delay time in delay centers.

9- Demand Time (D)

Service demand time at the facility (i) in a queueing network system is the period of time required to serve one customer at this facility. Thus, service demand time at facility (i) is,

$$D_i = v_i s_i$$

Bottleneck Device

The term of bottleneck device is used to signify a facility that slows down material flow and makes the system in a choking case. Exclusively, bottlenecks can be categorized into two kinds, hard or soft, where the hard bottleneck is the facility that completely limits the output of a system. Particularly, in queuing network system the bottleneck device is the device which has the highest demand time and consequently has the highest utilization.

Bottleneck Analysis

Bottleneck device, within a network queuing system, is the device which has the greatest service demand (D_i) value and its service demand is denoted (D_{max}). The bottleneck device is the resource that has the highest utilization rate. The determination of it is required as it is always the reason for limiting the performance of the queueing network system. Importantly, the delay centers cannot be absolutely a bottleneck device.

Asymptotic Bounds

Asymptotic bounds are used to present for both the optimistic and the pessimistic bounds of the system's throughput and the response time in queuing network systems. So, they are derived considering the extreme conditions of light and heavy workloads. Thus, they have the general relations $[R_{pes} \leq R \leq R_{opt}]$ *and* $[X_{pes} \leq X \leq X_{opt}]$ which specify the bounds of the performance measures for a queueing network system. The system throughput asymptotic bounds can be represented using the following interval.

$$\frac{N}{ND+Z} \leq X(N) \leq min\left\{\frac{1}{D_{max}}, \frac{N}{D+z}\right\}$$

Likewise, asymptotic response time bounds represent the optimistic and pessimistic response times. Response time bounds can be represented using the following interval.

$$\therefore max\{D, ND_{max} - Z\} \leq R(N) \leq ND$$

Mean value analysis

Mean value analysis is an analytical iteration algorithm which was invented by Reiser and Lavenberg (1980, 1982). It is used to analyze the behavior of closed queuing network systems without using the actual probability of arrival time or processing time distributions. Basically, mean value analysis is based upon the relation between the mean waiting time (response time) and the mean queue size in a closed queueing network system with one customer at least.

Fundamentally, the algorithm starts with one customer and increases one customers in each new iteration. Recently, mean value analysis equations were supported by Little's equation which is applied to the whole queuing network system and to each service center. Fortunately, the objective of the mean value analysis is to calculate all mean values of the variables of the closed network queuing system with the growth of customers' numbers such as response time, throughput, and queues lengths. Moreover, utilization, and bounds of the closed network systems can be computed or predicted using the mean value analysis algorithm.

1- Arrival Instant Theorem

The arrival instant theorem states that the average number of customers found upon arrival by a customer at (j^{th}) queue is $Q_{j(N-1)}$ in a closed queueing network and thus the number of customers will become (N) customer. So, the mean response time equals the waiting time in queues

plus the service time per each customer $R_j = S_j[1 + Q_{j(N-1)}]$ at device (j).

2- Mean Value Analysis Assumptions

The main assumptions of the mean value analysis can be summarized in the following points.

1. The routes of the customers among different nodes is predetermined and can be deterministic or probabilistic routing.
2. Service times at nodes are probabilistic.
3. Equilibrium of the system.
4. No Blocking; any device can conduct service whenever jobs reach it as there are no priorities among devices in the system.
5. All devices in the system can be considered as queueing facilities or delay centers according to the institutions working conditions.
6. Each queueing server is considered as a $M/M/1$ queueing system.

3- Mean Value Analysis Algorithm

Given a closed queueing network that has (N) customers and (M) facilities, the response time at each server can be formulated as,

$$R_i(N) = \begin{cases} S_i\left(1 + Q_{i(N-1)}\right), & N <> 1 \\ S_i \quad , & N = 1 \end{cases}$$

Hence, according to arrival instant theorem, $Q_{i(N-1)}$ is the mean queue length at (i^{th}) facility in the closed network queueing system which has (N) customer where $[N \neq 1]$. But, the response time of each facility with one considered user $(N = 1)$ can be easily calculated as it equals its service time.

The mean queuing length refers to the number of customers in front of the server plus the customer who is taking service. Therefore, given the response times at each individual device,

1. the system response time using the general response time law is,

$$R(N) = \sum_{i=1}^{m} R_i v_i$$

2. The system throughput using the interactive response time law is,

$$X(N) = \frac{N}{R(N) + Z}$$

3. The facilities' throughputs are measured as,

$$x_i(N) = X(N)v_i$$

4. The facilities' queue lengths with (N) customers in the closed network queueing system are measured using Little's law as,

$$Q_i(N) = x_i(N)R_i(N) = X(N)\, v_i\, R_i(N)$$

5. Response time equation for delay centers (i) is simply,

$$R_i(N) = S_i = Z$$

4- The Pseudocode of Mean Value Analysis

For instance, assuming that there are (M) queueing facilities, (i) as the facility number, and (Z) as the delay time in delay centers are considered and given the previous equations, the mean value analysis algorithm can be stated as follow,

1. Step one initialization

   Given $S_i = service\ time\ \forall\ i = \{1, 2, \dots\}$

   Given $v_i = visit\ ratio \quad \forall\ i = \{1, 2, \dots\}$

2. Step two mean value analysis algorithm

   For $N = 1$ to M

   1. If $N = 1$ then

      $R_i = S_i$

      Else

      $R_i = S_i\,(1 + Q_{i(N-1)})$

      End if
   2. $R(N) = \sum_i^M R_i v_i$
   3. $X(N) = \frac{N}{R(N)+Z}$
   4. $x_i(N) = X(N)\, v_i$
   5. $Q_i(N) = x_i(N)R_i(N)$
   6. Exit when stopping criteria is satisfied

   Next (N)

3. Step three showing outputs

Indeed, stopping criteria can be one condition or more which are stated by the decision maker who tends to stop iterations when the algorithm achieves his aims such as a certain level of throughput, customers, queues length or any else according to the working conditions of his institution.

References

[1] Allen, A. O. (2014). *Probability, statistics, and queueing theory*. Academic Press.

[2] Bhat, U. N. (2015). *An introduction to queueing theory: modeling and analysis in applications*. Birkhäuser.

[3] Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (2006). *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons.

[4] Denning, P. J. (1991). *Queueing in networks of computers*. Nasa Ames Research Center: Research Institute for Advanced Computer Science, Agreement number NCC 2-387.

[5] Denning, P. J., & Buzen, J. P. (1978). The operational analysis of queueing network models. *ACM Computing Surveys (CSUR)*, *10*(3), 225-261.

[6] Hillston, J. (2009). *Performance modelling: operational laws*. The University of Edinburgh, Scotland. Website: http://www.inf.ed.ac.uk/teaching/courses/pm/Note2.pdf

[7] Jain, R. (2008). *Operational laws*. Washington university. Website: http://www.cse.wustl.edu/~jain/cse567-08/ftp/k_33ol.pdf

[8] Jain, R. (2012). *Mean value analysis*. Washington University. Mini course.

[9] Lazowska, E. D., Zahorjan, J., Graham, G. S., & Sevcik, K. C. (1984). *Quantitative system performance: computer system analysis using queueing network models*. Prentice-Hall, Inc..

[10] Lazowska, E. D., Zahorjan, J., Graham, G. S., & Sevcik, K. C. (1984). *Quantitative system performance: computer system analysis using queueing network models*. Prentice-Hall, Inc..

[11] Parsaei, H. R., & Mital, A. (Eds.). (1992). *Economics of advanced manufacturing systems*. New York, NY: Chapman & Hall.

[12] Sivalingam, K. (2012). *Operational laws & Bottleneck analysis*. Performance evaluation of computer systems. Indian Institute of Technology Madras: Department of computer science and engineering: video course: mod-01 lec-24. Website: https://www.youtube.com/watch?v=qruxbs09hgi.