System Utilization: Keeping the Glass Half Full

Bruce McNutt
IBM Corporation

Abstract

This paper describes how the black box model, as introduced in earlier papers, helps to plan and operate systems that grow gracefully. This comes about through a better understanding of system utilization. Earlier papers showed that the black box model can provide automated utilization measurements. This paper applies the same model more broadly to the planning and operation of a system. Our goal is to grow the productivity of the system up to the level defined in the capacity planning process, with minor delays due to queueing and/or contention.

1 Introduction

Previous papers have shown that by concealing the details of the server configuration, it is possible to develop a simple "generic" model of system queueing, that fits well with a variety of more detailed queueing models [1]. Also, the resulting *black box* model provides a powerful technique for measuring the utilization of a system based upon its external behavior, even when traditional utilization measurements based upon cycle counts are not applicable [2].

Just as the black box model helps with the measurement of utilization, it also sheds light on the implications of utilization to system performance and capacity planning. The purpose of this paper is to explore more fully this conceptual side of the black box model.

Ideally, we would like the increase in system use that comes from growth to occur gracefully. When this happens, utilization is a half-full, rather than a half-empty glass. The productivity of the system grows until its use reaches the level initially defined during the capacity planning process. Meanwhile, the system continues to perform well, with relatively minor delays due to queueing and/or contention.

Unfortunately, the glass can sometimes become half empty. Systems can grow past their realistic limits. At that point, the use of a system for its intended purpose may become a struggle, due to severe delays.

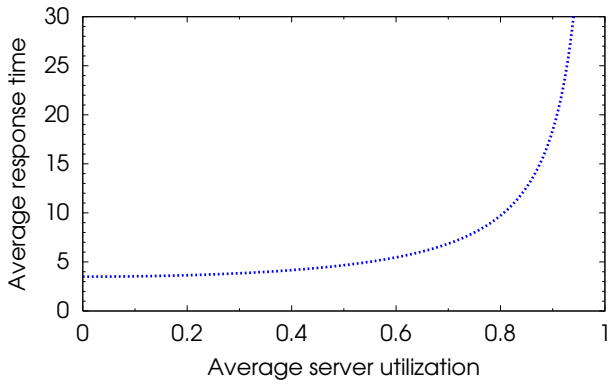
This paper describes how the black box model provides a quantitative mechanism to plan and operate systems so that the glass remains half full. Section 2 puts in place a simple, geometrically oriented definition of the desired operating region, together with multiple criteria that can be used to verify whether a given capacity plan, or a given operational environment, falls within that region. Sections 3 and 4 then follow up on the actual use of the proposed criteria in, respectively, capacity planning as well as day to day system management.

Although the black box model provides the underpinning of the proposed scheme, its value comes from the criteria provided in this paper for assessing utilization. We find that it is surprisingly easy to to decide whether the use of a system exceeds the level that can be accomplished gracefully. In the capacity planning process, this assessment is based upon the estimated system resources and processing requirements. For a running system, the same assessment is based upon live measurements.

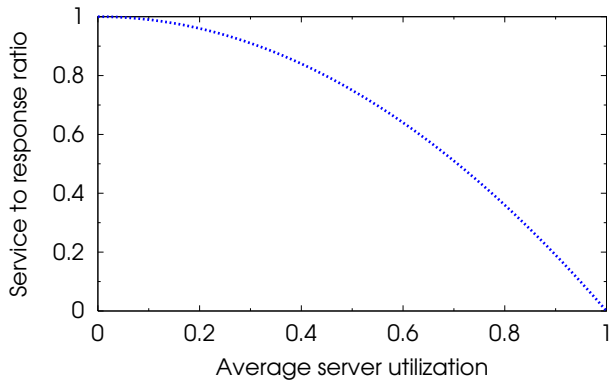
2 Region of Graceful System Growth

The mathematics of the black box model are tied closely to the *system map*, an idea first suggested by Allen [3]. Figure 1 presents an example of how the system map works. The example shows the behavior of the M/M/2 model, although the exact model chosen does not matter for the purposes of the present discussion.

Figure 1a presents the average response time R as a function of the system utilization ρ. This is a typical example of a so-called response time/throughput curve.



a. Response time/throughput curve.



b. System map.

Figure 1 *The same M/M/2 queueing model, presented in two ways.*

For very low system utilizations, little waiting occurs, and the response time is dominated by the time spent actually receiving service; that is,

$$R = Q + s$$

where the average queue time Q is small compared with the average service time s.

The drawback of Figure 1a comes at very high system utilizations. In this region of the figure, the time spent waiting for service becomes unbounded, so no matter what range is chosen for the vertical axis, it is impossible to show the entire curve.

The corresponding *system map*, as shown in Figure 1b, solves this problem by presenting the *ratio of service time to response time*. For the sake of having a convenient term, let us call this quantity the system map ratio:

$$M = \frac{s}{R} = \frac{s}{Q + s}$$

In the system map, this ratio gradually falls from unity (no waiting) to zero (infinite waiting); in this way, the entire curve can be shown, including the behavior as the system utilization approaches 100 percent.

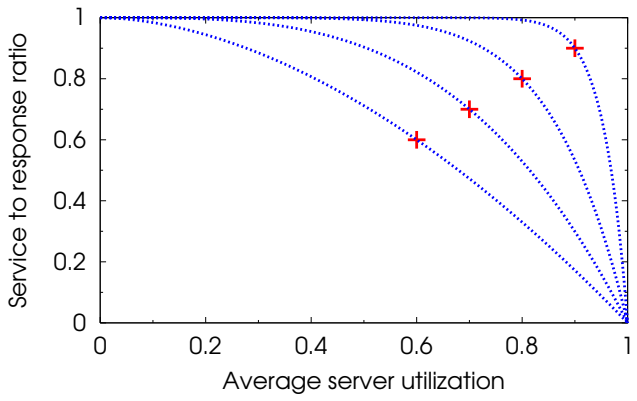


Figure 3 *Family of system maps produced by the black box model.*

As a thought exercise, imagine a blank system map. Suppose we know that a particular system's range of performance includes some point on that map – for example, any of the points marked "+" in Figure 3. How much can we conclude from this minimal amount of information?

Remarkably, it seems to be possible to conclude a great deal. A variety of different queueing models, if they are forced to produce a curve on the system map that passes through an identified point, tend to produce very similar results. Thus, given a point on the system map, it is possible to estimate fairly well the remainder of the map, *without needing to know all the details of the queueing mechanism*.

The easiest way to produce the needed estimate is to use a family of curves that is not a valid solution to any well-known queueing model. This family, given by the general formula

$$M = 1 - \rho^c \tag{1}$$

is, however, very simple mathematically. Figure 3 presents the inferred system map behavior, based upon (1).

The black box model is an approximate mathematical model that can be invoked to explain the success of (1). For this paper, we do not actually require the full black box model; instead, we begin with (1), and reason based upon that starting point.

Assume, then, that the system map of a given system is specified by (1) for some number of servers $c \geq 1$. The number of servers may require estimation, and is not necessarily assumed to be an integer. We now assert that the range of system operating conditions over which growth can occur in a graceful manner corresponds with the condition $M \geq \rho$ as shown on the system map. In geometric terms, the desired operating region is the portion of the system map at or above a diagonal line extending from the origin.

To see that this is the case, consider the specific utilization defined by $\rho_{knee} = (c+1)^{-1/c}$. This helpful point of reference is always close to, but no greater than, the utilization that occurs on the diagonal.

By (1), the corresponding system map ratio M_{knee} is given by $c/(c+1) >= \rho_{knee}$. Thus, the fraction of time spent in the queue, waiting for service, is given by $1 - M_{knee} = 1/(c+1)$. This quantity is equal to $\frac{1}{2}$ when $c = 1$, and becomes less and less significant as c increases.

Since $M_{knee} >= \rho_{knee}$ we may draw a vertical line from the point (ρ_{knee}, M_{knee}) downward, meeting the diagonal; then another line toward the right, meeting the curve again at a point such that $M = \rho_{knee}$. Clearly, this point is at or below the diagonal. In addition, the quantity $1 - \rho_{knee}$ has the same desirable properties as those just described at the end of the previous paragraph. This quantity is *also* equal to $\frac{1}{2}$ when $c = 1$, and becomes less and less significant as c increases.

The value of M that occurs at the diagonal can be bounded between the two calculations of M just given; therefore, the fraction of time spent in the queue, for an operating point along the diagonal of the system map, has the same desirable properties as it does for the two bounding cases. It is equal to $\frac{1}{2}$ when $c = 1$, and becomes less and less significant as c increases.

On the other hand, by the time the curve of performance as shown on the system map crosses the diagonal, it has already turned downward. Any further gains in utilization must come at the expense of significant additions to queue time. This situation calls to mind the old saying "quit while you are ahead", and explains the recommended geometry for defining when the utilization glass can be considered half full.

The core idea of this paper is that a surprising variety of mathematical criteria can be used to identify the desired operating region. They are equivalent to each other in the sense that they produce the same pass/fail grade when applied to a given set of operational conditions. However, each criterion uses different information, reducing the amount that must be known about the environment.

Let x represent the average rate of system requests per second, and N the average population of outstanding system requests. Note also that by Little's law, these quantities are related by $N = xR$. Then within the framework of the black box model, all five of the following criteria are equivalent:

$$M \geq \rho \tag{2}$$

$$\rho \leq \cdots (1 - (1 - (1 - \frac{c}{c+1})^{\frac{1}{c}})^{\frac{1}{c}})^{\frac{1}{c}} \cdots \tag{3}$$

$$N \leq c \tag{4}$$

$$1 - \left(\frac{xs}{c}\right)^c \geq \frac{xs}{c} \tag{5}$$

$$\left[1 - \frac{s}{R}\right]^{\frac{1}{N}} \leq \frac{s}{R} \tag{6}$$

Also, the utilization threshold values that (3) states in the form of a limit agree well with those of the closed expression

$$\rho \leq \frac{1}{4}\,\frac{c}{c+1} + \frac{\rho_{knee}}{2} + \frac{1}{4}(1 - \rho_{knee})^{1/c} \tag{7}$$

To three digit precision, the two forms (3) and (7) of the utilization threshold are interchangable.

As an example of how to reason about the black box model based upon a variety of criteria, consider again the point of reference that occurs when the utilization is given by $\rho_{knee} = (c+1)^{-1/c}$. As we observed in a previous paragraph, this point reflects conditions where the criterion (2) holds; but we might ask, how close is that criterion to failing?

The simplest way to answer becomes apparent by examining instead criterion (4). The value of N at the point of reference is given by $N_{knee} = (c+1)^{(c-1)/c}$. When c is an integer, that same value can be stated approximately as $N_{knee} \approx c - H_{c+2} + H_3$, where H_i represents the ith harmonic number. Keeping in mind that $c \geq 1$, we thus see that the value of N_{knee} is equal to that of c when $c = 1$, and falls below that of c for $c > 1$.

Turning now to the offset δ between N_{knee} and c, we have the relationship $\delta \approx H_{c+2} - H_3 << c$. For this reason, the value of N_{knee} always comprises by far the largest portion of c. It is fair to say that the point of reference corresponds to a case where the criterion (4) just barely passes. By extension, the same statement applies to the other four criteria as well.

The remaining sections of the paper exploit (5) and (6) by applying them respectively to the capacity planning and the system monitoring phases of the system management life cycle.

3 Capacity Planning

The capacity plan for a complex system proceeds by examining that system's most important resources. The black box model is intended to help with the analysis of resources that are used temporarily to process a particular request, such as processor cores or system ports. The requirement for a resource of this type is driven by the rate x of requests against it, as well as the average service time s that will be required to complete any single request, when running by itself.

Given an estimate of both of the quantities just identified, it is then necessary to assess how many individual units of the given resource should be configured. Ordinarily, system performance will suffer unless *more* units of the resource are provided than the average number required by the rate of requests. The criterion (5) provides a simple and effective way to address the question of *how many* additional units must be added to the capacity plan, beyond those called for directly by the estimated demand.

For example, consider the number of 8 Gbps fibre channel ports being provided with a Linux server. Based upon past levels of load, we estimate that the server's I/O rate when running its maximum intended level of application work is 50,000 I/Os per second, with an average transfer size of 16 KB. Assuming the nominal 8 Gbps fibre channel protocol, this implies that the number of ports in active use will be $(16/1024) \times 50,000/800 = 0.976$ on average. It is easy to guess that we ought to configure at least 2 ports, based upon this average level of port demand. But are two ports sufficient?

A simple way to address this question is to apply the criterion (5). By that criterion, 2 ports *are* sufficient, since $1 - (0.976/2)^2 = 0.762 > 0.976/2$ (the criterion passes).

On the other hand, in the analysis just presented, we assumed that the port can run at the full 8 Gbps permitted by the fiber channel protocol. Suppose, more conservatively, we wish to assume that the actual effective speed of the port is 75 percent of this theoretical maximum. This then implies that we need 1.30 ports on average. Applying the criterion (5), we now obtain $1 - (1.30/2)^2 = 0.577 < 1.30/2$ (the criterion fails).

We therefore arrive at the conclusion that two ports *may* be sufficient, but only if the performance of the port technology runs at close to the theoretical maximum. A sensible next step may therefore be to investigate further any performance data that is available for the affected type of port hardware. Alternately, we could choose to configure 4 ports. By (5) this will be sufficient even if the port is only 75 percent efficient, since $1 - (1.30/4)^4 = 0.989 > 1.30/4$ (the criterion passes).

4 System Monitoring

For many systems today, it is possible to directly measure the throughput x and response time R during any defined measurement interval throughout the day. In addition, the quantity $N = xR$ can be obtained using Little's Law. To apply criterion (6), however, we also require data for the average service time s.

One method to measure s is to instrument the system with the *General Purpose Utilization Monitor*, which performs measurements of s in each measurement period in addition to producing estimates of system utilization [2]. In this paper, however, we do not assume that instrumentation of that type is available. A reasonable alternative is to estimate s by considering a *shoulder* period in which the mix of applications is similar to that running during the daily peak. In a well chosen shoulder interval, queueing may be light enough so that the measurement of R can also be taken as the approximate value of s.

Assuming that a reasonable value for s can be identified, it is then possible to apply criterion (6) to assess based on measured data whether the utilization glass is half full.

It is important to note that the application of (6) does *not* require a value for c. Instead, this criterion provides an independent check on actual delivered concurrency of the system. If for any reason the value of N exceeds the effective system concurrency during some periods, then we should expect that the criterion (6) will fail in those periods.

For example, consider the same Linux system as before, configured with two ports. Using data from a shoulder period as well as a peak period, where the average transfer size was 16K in both cases, we conclude that the minimum time required to complete such a transfer is 26 microseconds, but under peak conditions the average response time is 35 microseconds. Also, the peak throughput is 40,000 I/Os per second, hence $N = 40,000 \times 0.000035 = 1.4$. Based on that information, the criterion (6) gives $[1-26/35]^{1/1.4} = 0.379 < 26/35$ (the criterion passes).

Although the system has been configured with two ports, this does not *necessarily* mean that it can always support two concurrent transfers without the use of a queue. For example, if both ports are in the same adapter, the processor provided in the adapter may not be able to accept interrupts from both ports at the same time. Also, under some conditions, the adapter may break a single host transfer into more than one physical transfer. The reverse case can also sometimes occur, in which multiple transfers are consolidated. The black box model cannot help with capturing such effects in detail, but it provides a way to gauge their impact. Effects of this kind may influence the effective concurrency of the system.

In the example just given, we applied (6) to a single measurement period. If, however, we observe the system through a large number of measurement periods, and assuming that we are able to capture wide swings in the load level, we can then assess the maximum value of N that is still capable of passing the criterion (6). That maximum value provides an independent check on the actual concurrency delivered by the configured pair of ports.

Suppose, then, that the two ports do deliver a concurrency of $c = 2$. Given that fact, we can then use the black box model to estimate the average port utilization. Graphically, the condition $c = 2$ allows us to draw a curve on the system map that belongs to the family given by (1). We then superimpose, on the same map, a straight line that corresponds to the condition $M = s/R = sx/N = \rho c/N$. To estimate the utilization, we find the utilization level at which the curve intersects with the straight line.

Conditions at the point of intersection are described by a polynomial equation of order c. Typically, the solution of an equation of that type is best described using bounds or a numeric approximation. In the case of the geometric problem just described in the previous paragraph, it is possible to state the utilizaton value at the point of intersection with more than enough numeric precision for the purpose of system monitoring.

To accomplish this, define the quantities $\beta = \sqrt{(c+1)/2}$ and $u = \max(N,\ N_{knee})$. Then the utilization at which the curve and the straight line meet each other is given within one percentage point by

$$\rho_0 = \frac{N}{u + \delta(N_{knee}/u)^{\beta} + (N/u)^{\beta^2}} \tag{8}$$

Also, if we now define the two related quantities $Z = \lfloor N/u \rfloor \in \{0,\ 1\}$ and $\eta = (N/c)^Z \in \{1,\ N/c\}$, then the same point of intersection can be obtained to three digit precision using

$$\rho \approx \frac{N}{N + c(1 - A^{c^{-Z}})/(1 - A^{c^{1-Z}})} \tag{9}$$

where

$$A = \frac{\eta cN - \rho_0^{c^{1-Z}} cN + \rho_0^{c^Z} N^2/\eta^2}{c^2 + N^2} \tag{10}$$

In these equations, the quantity A is a geometric construct that represents the closest point on the straight line to an identified point on the curve. By giving the value of ρ^{c^2} at that closest point, A provides a new estimate of the needed position along the curve either horizontally (for the upper portion of the system map) or vertically (for the lower portion). Also, the desired utilization is an attractive fixed point of the function

$$f(\rho) = \frac{N}{N + c(1-\rho)/(1-\rho^c)}$$

For that reason, (9) provides an effective way to apply the value of A.

Continuing the example of the two port system, we have already concluded that $c = 2$; this, in turn, implies that $N_{knee} = (c+1)^{(c-1)/c} = \sqrt{3} = 1.732$, so $\delta = c - N_{knee} = 0.268$. Also, we have measured $N = 1.4$, less than N_{knee}, so $u = N_{knee} = 1.732$. Finally, $\beta^2 = \frac{3}{2}$. Based upon these quantities, a rough value for the port utilization is given by $\rho_0 = 1.4/(1.732 + 0.268 \times 1 + (1.4/1.732)^{3/2}) = 0.513$.

Although a solution within one percentage point would appear to be good enough in the case of this example, the option also exists to compute the same quantity more precisely by plugging the estimate 0.513 into (9) and (10). This procedure yields the slightly adjusted estimate 0.515 for the utilization value.

Some room for *graceful* growth remains at the utilization level $\rho \approx 0.515$ just obtained. However, the opportunity for such growth extends only until we reach the condition $N = c = 2$. We can obtain the corresponding threshold for the utilization level either by plugging the assumption $N = 2$ into (9), or by applying (7). Using either approach, we find that growth should not continue past the utilization level given by $\rho_{at\_c} = 0.618$.

5 Conclusions

This paper follows up on earlier papers about the black box model, and uses that model to provide important guidelines for system management. In particular, we have examined the question of whether the utilization of a given, specific system is excessive. Within the framework of the black box model, this condition has a simple geometric interpretation.

The most important conclusion of the paper is that the black box model provides several distinct but equivalent mathematical tests for excessive system utilization. These tests can simplify both the capacity planning as well as the performance measurement phases of the system management life cycle.

The General Purpose Utilization Monitor, described in an earlier paper, can assist in the management of system utilization. It can take needed measurements, report system utilization and dynamically maintain an estimate of the concurrency of the system. This paper does not assume that a monitor with these functions is in use, but clarifies the underlying ideas behind the use of such a monitor.

References

[1] B. McNutt, "Waiting for a Black Box", CMG Proceedings, Nov. 2013.

[2] B. McNutt, "Meter Anything from Component to Cloud: A General-Purpose Utilization Monitor", CMG Proceedings, Nov. 2014

[3] A.O. Allen, *Probability, Statistics, and Queueing Theory*, Academic Press, 1978. See particularly pp. 221-223.