Little's Law assumptions: "But I still wanna use it!" The Goldilocks solution to sizing the system for non-steady-state dynamics

Alex Gilgur

Abstract

Little's Law is well known: number of concurrent users of a resource equals the product of arrival rate and the holding time of one user. Somewhat less known is the fact that Little's Law is based on the assumption of stationarity of the underlying processes. But at The Knee, all assumptions break down because this is where the system is not stationary anymore. Without going into an existential discussion of The Knee, this paper proposes a solution that allows an estimation of concurrency that is "not too low, not too high; just right" for non-stationary conditions.

Little's Law is well known: number of concurrent users (Q) of a resource during a period of time is equal to the product of average rate of arrival of users (X) and the average holding time of one user (R):

$$Q = X * R \qquad (1)$$

From this description, it follows that it doesn't matter what the duration of the time interval is, as long as the X and the R are measured in the same time interval. According to its creators ([1]):

> *"We require **stationarity** assumptions about the underlying stochastic processes…"*

But what if the underlying stochastic process is not stationary (Figure 1, $X > 50$)? For example, while the holding time (R) was nearly constant at lower workloads ($X < 50$), we arrive at a value of X where the response times will be increasing more rapidly than at lower workloads?

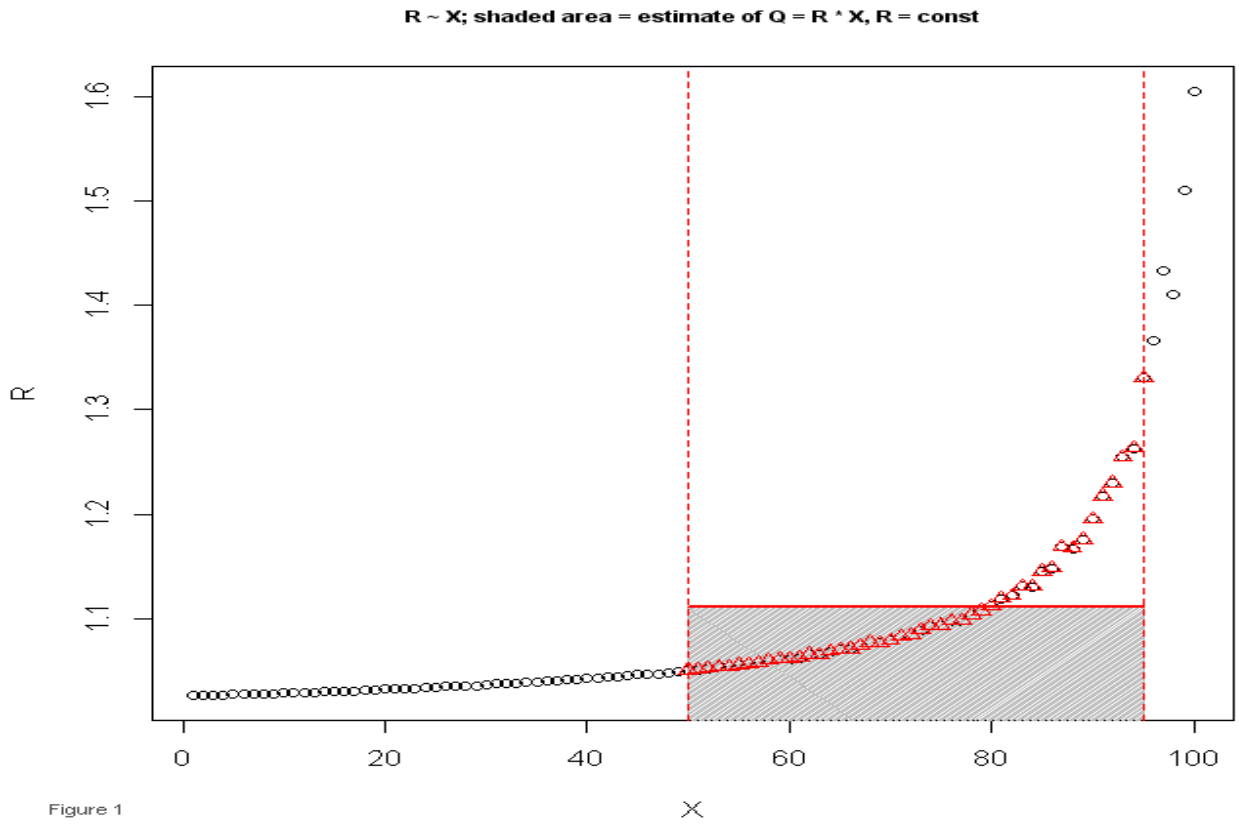

Figure 1: Illustration of Little's Law in an unstable system: a potential underestimation of concurrency.

In Figure 1, processing time (R) is constant up to X = 50, but at higher X, the response times start going up. We are interested in the period of time when X was between 50 and 95 arrivals per second.

The horizontal red line corresponds to applying Little's Law literally in a non-stationary system: area under the red line will correspond to the estimation of the number of concurrent sessions in this interval of arrival rates.

Why is it a bad idea in non-steady-state conditions?

Using Little's Law in this situation becomes a dangerous proposition, as all of a sudden, we lose the R as a valid average measure of holding time for this time interval: it will depend on the request arrival rate. In other words, on the knee, Little's Law assumption of stationarity of parameters is not valid, and technically we cannot use it anymore: the thick solid red line representing literal application of Little's Law in this scenario demonstrates that we will underestimate the number of concurrent sessions (users).

What to do?

One way to solve the problem would be to draw a straight line connecting the R(X1) and the R(X2) (Figure 2) and compute the area under that line. But that will cause a severe overestimation of the number of concurrent sessions and therefore a gross oversizing of the system if our purpose is capacity planning.

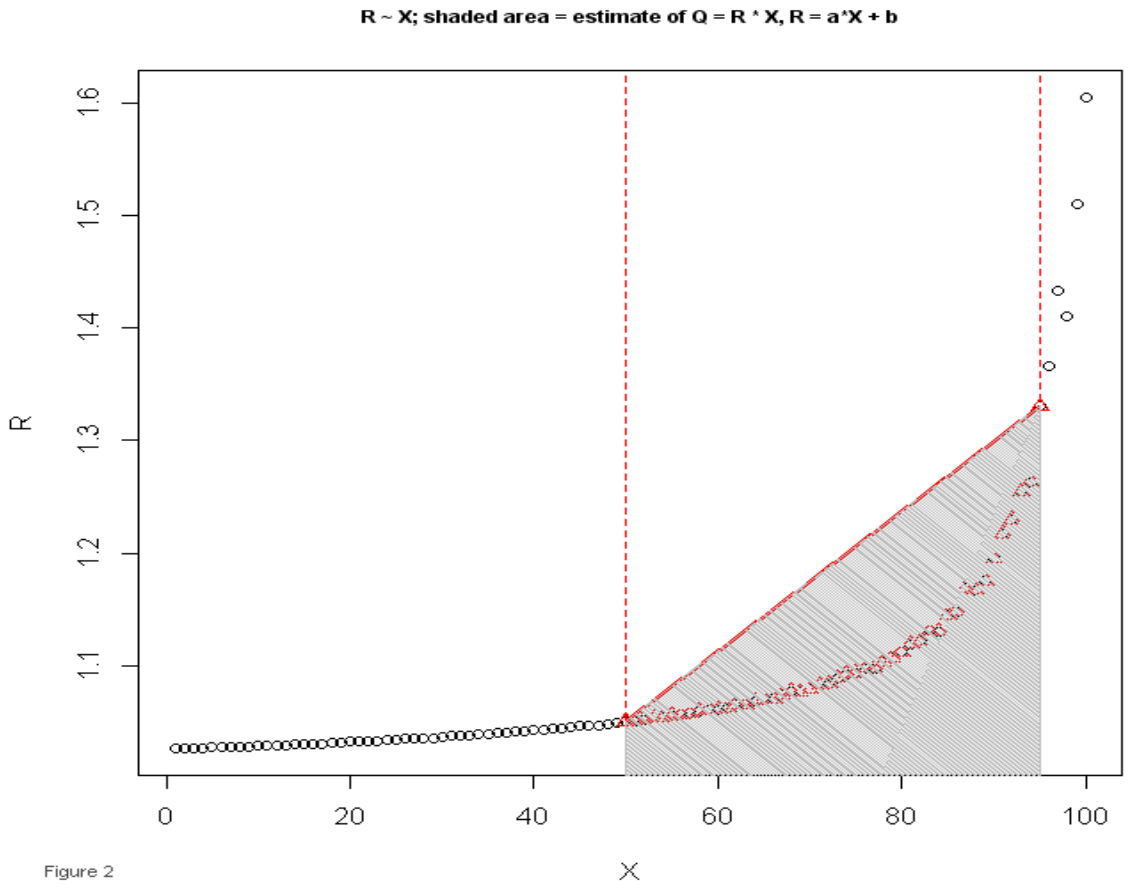

Figure 2: Illustration of Little's Law in an unstable system: an overestimation of concurrency.

The next logical iteration would be to break the big time interval where arrival rates vary widely (from X1 to X2) into intervals small enough that

$$R \approx const(X), X \in [X1, X2]$$

Then we sum them all up for the time interval of interest:

$$Q \approx \sum_{X1}^{X2} R(X) * \Delta X \qquad (3)$$

But that is equivalent to integrating R by dX:

$$Q = \int_{X1}^{X2} R(X)dX \qquad (4)$$

This approach (Figure 3, Eq. 4) will size the system “Not too big, not too little; just right”.

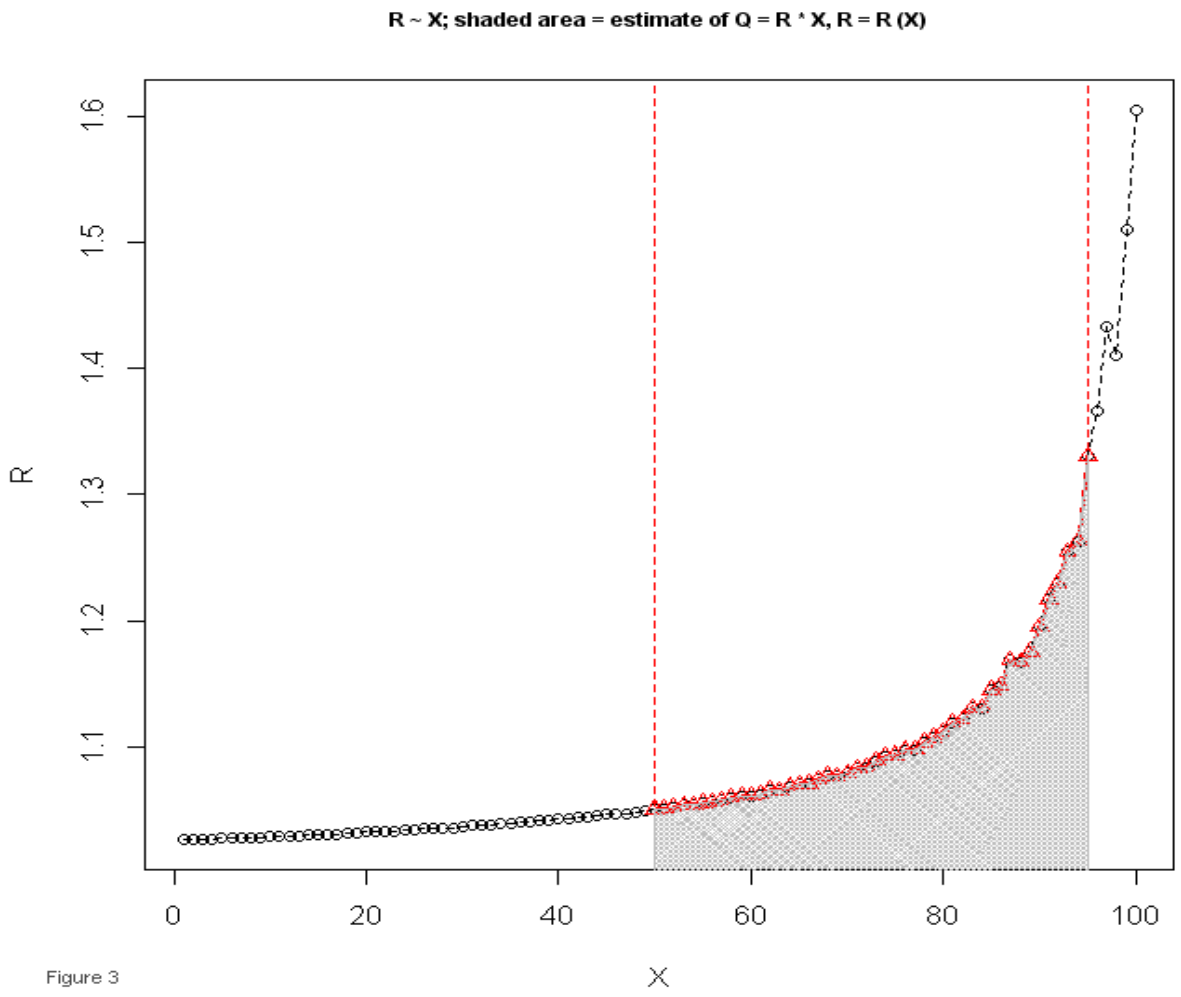

Figure 3: Concurrency measured as the area under the red segments in the R(X) curve

Alternatively, in any system – stationary or not – would-be-stationary response time can be computed as the derivative of concurrency by the throughput, and vice versa. Generalizing, it is safe to say that, if we have a theoretical or empirical relationship of two parts of the equation (4), we can easily obtain the third component by integrating (if looking for Q) or differentiating (if looking for R or for X).

And if we have time intervals that are too granular, and we only have the average (mean/median) R and the average (mean/median) X for these intervals?

If we have no insight into the system dynamics at a finer granularity, then there is nothing that can be done; we can only "hope for the best", assume that the process was stationary within each given time interval and apply Little's Law directly. After all, if the X did not vary much within that interval, we can approximate the integration (4) by the area under the {(X1, 0), (X1, R1), (X2, R2), (X2, 0)}.
In this case, using the Figure 2 approach will be preferred to Figure 1, as it will allow us to oversize the system sufficiently that it will be able to handle whatever micro-level perturbation may occur within that time interval.
If concurrency and throughput are given mathematically as functions of one another or as functions of a business metric (see more about that [2]), then we can get the service time as the ratio of these two functions. In this case, at steady-state conditions, response time will asymptotically be converging to the dQ/dX:

$$\lim_{steady} (R) = \frac{dQ(X)}{dX} \qquad (5)$$

Conclusions

In situations where sizing the system is dictated by the concurrent traffic (e.g., determining the number of connections to a database), and if we cannot allow ourselves to oversize it (and we can never afford to undersize it), we need to come up with a "Goldilocks" solution.

The knee is used in this paper merely as an example; the same approach can be applied to any source of instability – e.g., different traffic mixes coming into the system at different times of the day provides a plethora of other examples, which are outside the scope of this paper, but to which the same principle can be applied. Care must be taken to identify the intervals of X within which the system can be assumed to be in a quasi-steady state, but even when the system is unstable, we can still use Little's law.

References

1. John D.C. Little, Stephen C. Graves. Little's Law. Downloaded from http://web.mit.edu/sgraves/www/papers/Little%27s%20Law-Published.pdf on 02/01/2012

2. Alexander Gilgur, Josep Ferrandiz, Matthew Beason. Time-Series Analysis: Forecasting + Regression: And or Or? – *presented at the 38th International Computer Measurement Group (CMG12) Conference, December 2012, Las Vegas, NV.*