Half-Latency Rule for Finding the Knee of the Latency Curve

Conference Paper *in* ACM SIGMETRICS Performance Evaluation Review · October 2015

DOI: 10.1145/2825236.2825248

1 author:

Naresh M. Patel
Google Inc.
21 PUBLICATIONS **441** CITATIONS

Half-Latency Rule for Finding the Knee of the Latency Curve

Naresh M. Patel
NetApp, Inc.

Abstract

Latency curves for computer systems typically rise rapidly beyond some threshold utilization and many mathematical methods have been suggested to find this *knee*, but none seem to match common practice. This paper proposes a trade-off metric called ATP (an alternative to Kleinrock's power metric) which generates the *half-latency rule* for calculating the location of the knee for a latency curve. Exact analysis with this approach applied to the simplest single-server queue results in an optimal server utilization of 71.5%, which is close to the 70% utilization used in practice. The *half-latency rule* also applies to practical situations that generate a discrete set of throughput and latency measurements. The discrete use cases include both production systems (for provisioning new work) or lab systems (for summarizing the entire latency curve into a single figure of merit for each workload and system configuration).

1. Introduction

Latency (or mean response time) versus throughput curves for queuing systems typically exhibit a *hockey stick*-like curve (Figure 1) and there has been an ongoing debate about the location of the *knee of the curve* where systems *should* operate for maximum efficiency [3, 4]. This optimal point (or knee) trades off some maximum throughput to deliver acceptable latency, so that further increases in throughput result in substantially higher latency. The open question is *what is acceptable latency* for a given latency versus throughput curve of a computer system.

System designers and service providers typically use a rough 70% utilization rule for sizing system capacity. In both cases, the standard power metric [1] is too conservative (yielding 50% utilization for the M/M/1 queue and even lower as variability increases) and other approaches (outlined in [3, 4, 5]) require arbitrary constants to be specified. Designers need a composite metric to make trade-offs between latency and throughput in a consistent way at load levels expected in production. For example, when evaluating a set of latency curves for several design options that positively or negatively impact either latency or throughput, a single *higher-is-better metric* makes it easier to determine whether or not a design change should be accepted. Service providers strive to simplify operations and prefer consistent latency as new workloads are added (ideally, constant latency up to saturation at 100%). In practice, they would settle for a throughput range $(0, x)$ over which users experience a *narrow latency range*, as new user workload is added incrementally, say in units of 1% of utilization. For simplicity, the monetary charge for the first unit of work is the same as that of the last accepted unit of work and so the latency expectations should also be *similar*. Thus the latency metric of interest is the average of all the latencies experienced across the operating throughput range $(0, x)$. We call this the Overall Response Time (ORT) defined as follows:

$$r(x) = \frac{a(x)}{x} = (1/x)\int_0^x w(u)du \tag{1}$$

where $w(x)$ is the latency at load level x and $a(x)$ is the area under $w(x)$ over the range $(0, x)$.

IFIP WG 7.3 Performance 2015, October 19-21, Sydney, Australia.

2. The ATP Metric and Properties

Naturally, the service provider also wants to maximize profits by adding enough units of work while maintaining acceptable ORT for their users. We propose a composite metric called Accelerated Throughput Power (ATP) over the throughput range $(0, x)$:

$$\text{ATP}(x) = \frac{x^\alpha}{r(x)} = \frac{x^{1+\alpha}}{a(x)} \tag{2}$$

By default we will use $\alpha = 1$ but we note that this is a general parameter for altering the relative importance of throughput. Maximizing ATP(x) leads to the *half-latency rule* which states that for a latency curve $w(x)$ and its corresponding ORT curve $r(x)$, the maximum of ATP(x) occurs when the ORT is half the latency (for $\alpha = 1$). In other words, the ATP knee occurs at $x = x^*$ where

$$r\left(x^*\right) = \frac{w\left(x^*\right)}{1+\alpha} = \frac{w\left(x^*\right)}{2}, \text{ when } \alpha = 1 \tag{3}$$

The *half-latency rule* makes it easy to compute the ATP knee for standard queuing systems (see Figure 1) as well as empirical data. Plotting the w-curve with the doubled r-curve and locating the point of intersection gives the ATP knee point.

For the M/G/1 queue with unit service time, we have $w(x) = 1 + \frac{kx}{1-x}$ where $k = \frac{1}{2}\left(1 + c^2\right)$ and c is the coefficient of variation of the service time. For $k = 1$, the exact solution can be derived by analysis:

k	0.5	1	2	3	4
x_{SP}	58.6%	50%	41.4%	36.6%	33.3%
L_{SP}	1	1	1	1	1
x^*	80%	71.5%	61.7%	55.7%	51.4%
L^*	2.42	2.51	2.60	2.65	2.69
w^*	3.02	3.51	4.22	4.77	5.23
ATP	0.531	0.407	0.292	0.234	0.197

Table 1: M/G/1 queuing metrics at the SP knee and the ATP knee as variability (k) increases

$$x^* = 1 - e^{\frac{1}{2}+W_{-1}\left(-\frac{1}{2\sqrt{e}}\right)} \simeq 71.5\% \quad (4)$$

where W_{-1} is the Lambert's W function (lower branch)

Figure 1 shows $w(x)$ (solid blue line) and $r(x)$ (red dashed line) for the M/M/1 queue. The standard power (SP) knee [1] is at $x = 50\%$ whereas the ATP knee is at $x^* = 71.5\%$ when $r(x)$ is half of $w(x)$ which occurs when the tangent line from the origin touches the $r(x)$ curve.

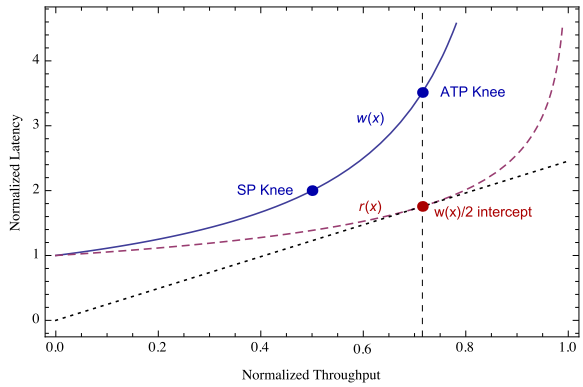

Figure 1: Latency curve $w(x)$ and ORT curve $r(x)$ showing the SP knee and ATP knee (at the half-latency, $r(x) = w(x)/2$)

Table 1 shows how both the SP knee ($x_{SP} = 1/(1+\sqrt{k})$) and ATP knee (x^*) values decrease as k, the service time variability increases from 0.5 to 4. This reduction matches the expectation that efficiency goes down as the service time becomes more variable. At $k = 1$, the $x^* = 71.5\%$ is close to the 70% value that is used by many practitioners whereas $x_{SP} = 50\%$. At $k = 4$, $x^* = 51.4\%$ whereas $x_{SP} = 33\%$.

The rows labeled L_{SP} and L^* show the mean queue length at the SP knee and the ATP knee. ATP mean queue lengths (2.51 at $k = 1$) increase slowly with higher variability (2.69 at $k = 1$) whereas the SP knee always has a mean queue length of one. This contrasts with w^* the mean waiting time which grows more rapidly than L^*. The row labeled ATP quantifies how much our composite metric decreases as service time becomes more variable: a four-fold increase in variability (k from 1 to 4) roughly halves the ATP metric.

3. Half-Latency Rule for Empirical Data

Suppose we have measurement data given as (throughput, latency) pairs, where the ith value is (x_i, w_i). With these values we can apply discrete integration to compute the area under the curve and thus the ORT. Now with the known set of (x_i, w_i) values for the w-curve, the ORT-curve (the discrete form of $r(x)$) is given by (throughput, ORT) pairs where the ith value is (x_i, r_i). Applying the *half-latency rule*, we can easily find j, the smallest index for which $2r_j - w_j$ is negative. The four coordinates enclosing the cross-over point are $(x_{j-1}, 2r_{j-1})$ and $(x_j, 2r_j)$ for the doubled r-curve and (x_{j-1}, w_{j-1}) and (x_j, w_j) for the w-curve. Using linear interpolation between points, the value of the cross-over point (x^*, w^*) of the doubled r-curve and the w-curve, and hence the ATP knee is given by:

$$x^* = \frac{(w_j - 2r_j)\, x_{j-1} + (2r_{j-1} - w_{j-1})\, x_j}{(w_j - w_{j-1}) - 2\,(r_j - r_{j-1})}$$

$$w^* = \frac{2r_{j-1}w_j - 2r_j w_{j-1}}{(w_j - w_{j-1}) - 2\,(r_j - r_{j-1})}$$

SPEC SFS2008 [2] benchmark results provide ten or more data points for many storage systems. We selected publications for three systems and computed ATP knees at 88%, 87%, and 79%, respectively for the fraction of peak reported operations per second. The corresponding normalized latencies were 2.7, 3.1, and 4.3, respectively. In this way, the ATP knee gives us a way to compare vendor systems independently of choices they may have made about how hard to drive their system.

4. Conclusion

We have defined a new ATP metric that can be maximized by using the *half-latency rule* to quantify the highest throughput level at which a queuing system can operate efficiently with respect to ORT. This crucial modification to the standard power metric changes the location of the maximum where queuing delays are higher but small changes in load level still do not change the latency significantly. In particular, for a M/M/1 queue, the ATP knee has a closed form solution near 71.5% which is near the subjective 70% threshold used in practice. The *half-latency rule* can be readily applied to empirical latency data, or applied more generally for making trade-offs for latency-like curves.

1. REFERENCES

[1] L. Kleinrock, *Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications.* in Conference Record, International Conference on Communications, Boston, Massachusetts, June 1979, p. 43.1.1-43.1.10.

[2] *SPECsfs2008 User's Guide. www.spec.org*

[3] M. Ley, *Does the Knee in a Queuing Curve Exist or is it just a Myth?.* Computer Measurement Group, MeasureIT, July 2009.

[4] N. J. Gunther,, *Mind Your Knees and Queues, Responding to Hyperbole with Hyperbolæ.* Computer Measurement Group, MeasureIT, August 2009.

[5] F.A. Gonzalez-Horta, et al, *Mathematical Model for the Optimal Utilization Percentile in M/M/1 Systems: A Contribution about Knees in Performance Curves.* ADAPTIVE 2011: pp 85-91