

# Percentile-Based Approach to Forecasting Workload Growth

Alexander Gilgur, C.Stephen Gunn, Douglas Browning, Xiaojun Di, Wei Chen, Rajesh Krishnaswamy  
(Google, Inc)

*"It's always the quiet ones." - Folk wisdom*

## Abstract

When forecasting resource workloads (traffic, CPU load, memory usage, etc.), we often extrapolate from the upper percentiles of data distributions. This works very well when the resource is far enough from its saturation point. However, when the resource utilization gets closer to the workload-carrying capacity of the resource, upper percentiles level off (the phenomenon is colloquially known as flat-topping or clipping), leading to underpredictions of future workload and potentially to undersized resources. This paper explains the phenomenon and proposes a new approach that can be used for making useful forecasts of workload when historical data for the forecast are collected from a resource approaching saturation.

## Workload

The workload on an IT resource (network node or link, CPU, disk, memory, etc.) is usually defined in terms of the number of commands (requests, jobs, packets, tasks,...) that are either being processed or sitting in the arrival queue (in some cases, the buffer for arrival queues is located on the sending element of the system; in such scenarios, it may be impossible for the resource in question to be aware of the pending workload).

Little's Law [[LTTL2011](#)], discovered, and expressed in stochastic terms, 40 years prior to John Little by A.K. Erlang, connects the workload, the arrival rate, and the service time in a very simple equation with unexpectedly complicated consequences:

$$W = X * T \tag{1}$$

where  $X$  = *arrival rate*;

$T$  = *service time (aka latency or response time)*;

$W$  = *workload*

The  $X$  and  $T$  describe two very different features of the system: the arrival rate ( $X$ ) characterizes demand, while latency ( $T$ ) characterizes the system's response to the workload.

As we collect throughput and latency data over time, we get two time series of measurements  $X(t)$  and  $T(t)$ , which together define a workload time series  $W(t)$ . Under low-arrival-rate conditions, the dependence of  $T(t)$  on  $X(t)$  can be treated as negligible. But when the resource approaches saturation, we observe the knee in the Receiver Operating Curve (ROC).

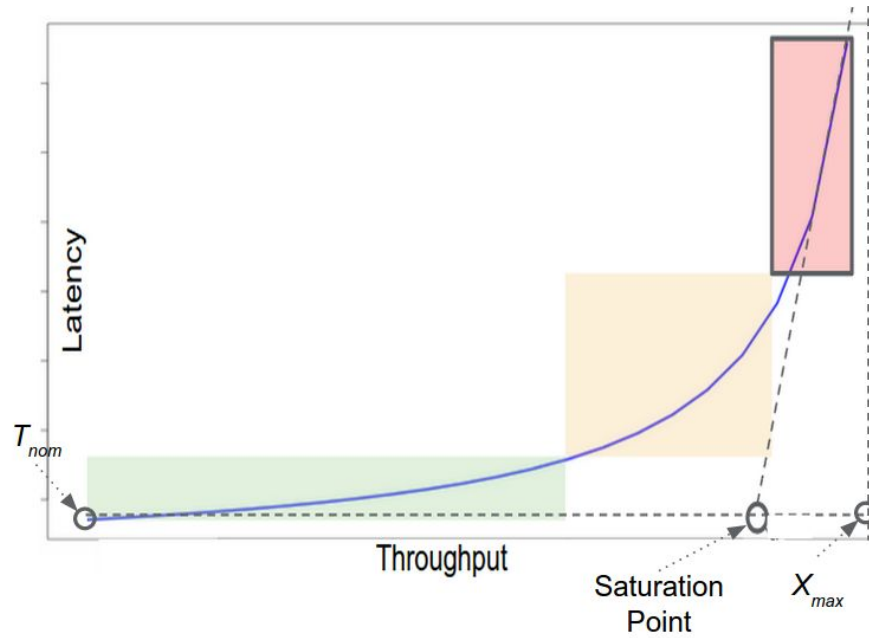


Figure 1. The Knee (an illustration)

At the point where green zone ends and yellow begins in Figure 1 (approaching “the Knee”), arrival rate and response time become significantly interdependent (see [\[FERR2012\(1\)\]](#), [\[GUNT2009\]](#), [\[GILG2013\]](#) and, for a truly rigorous discourse, [\[CHDY2014\]](#)).

This concept of knee behavior informs a number of practical considerations. One is that lead times for parts and capacity installation times impose the need for forecasting system behavior at a time far in the future. As economic forces often dictate seeking utilization levels “just below the knee”, the forecasting must often extrapolate histories of behavior below the knee into levels within and above the knee.

Little’s Law also allows us to express the holding capacity of an IT resource (maximum concurrency) as  $N = X_{max} * T_{nom}$ , where  $X_{max}$  = *bandwidth*, or *throughput capacity*, and  $T_{nom}$  = *nominal latency*. Nominal latency is latency observed under low load (when  $T(X)$  is nearly *const*) or calculated, e.g., as link length divided by the speed of light. In networking, holding capacity is known as BDP - Bandwidth-Delay Product; in a transactional system (e.g., a database; a telephone station, a cache register), it will be the maximum number of transactions that the system can hold at any given time without blocking.

## Problem Statement

### We have:

- An IT resource (e.g., network link) with a given holding capacity,  $N$ .
- Expected throughput for the element,  $X$ .
- Nominal latency (job holding time) for this element,  $T_{nom}$ .
- Historical data (time series) for Throughput,  $X(t)$  and Latency,  $T(t)$

### We need to:

Estimate when the element will reach its saturation point, usually with some built-in bias to address risk.

### Standard Approach:

1. Compute the historical workload,  $W(t)$ , using Little's Law (see [GILG2013] and [CHDY2014] for ways to deal with the high-workload conditions);
2. Get the 95th (or 99th, or 90th, or ... ) percentile of measurements on a suitable time step (usually weekly, to have sufficient data to isolate the top 5% and to accommodate the weekly and diurnal patterns often encountered in resource utilization data),  $W_{.95}(i)$  (where  $i$  = time interval over which the percentile is calculated);
3. Forecast (see, e.g., [MAKR1998], [ARMS2001]) the  $W_{.95}(i) \rightarrow \overline{W}_{.95}(i)$ ;
4. Add an overhead,  $\Delta(t)$ , to the forecasted  $\overline{W}_{.95}(i)$  value ([CRFT2006], [OSTR2011]) to create headroom for data variability.
5. Identify the earliest time when  $\overline{W}_{.95}(i) + \Delta(i) \geq N$ .

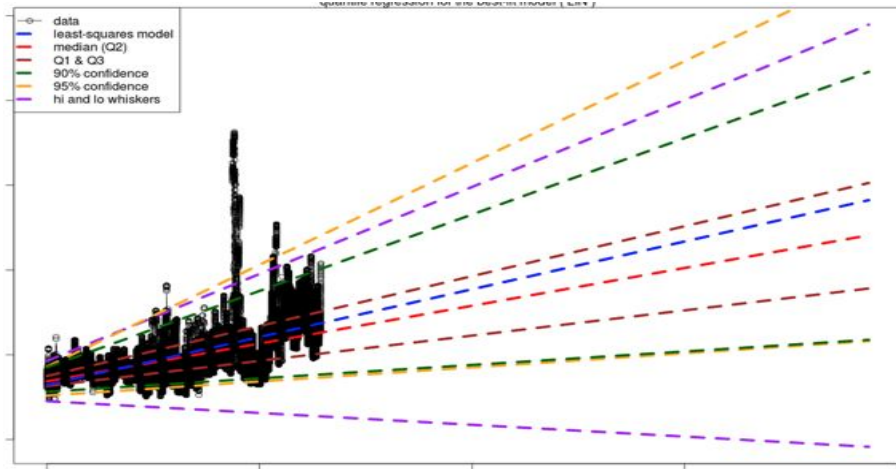
## Problem with Standard Approach

### Standard Approach Assumptions

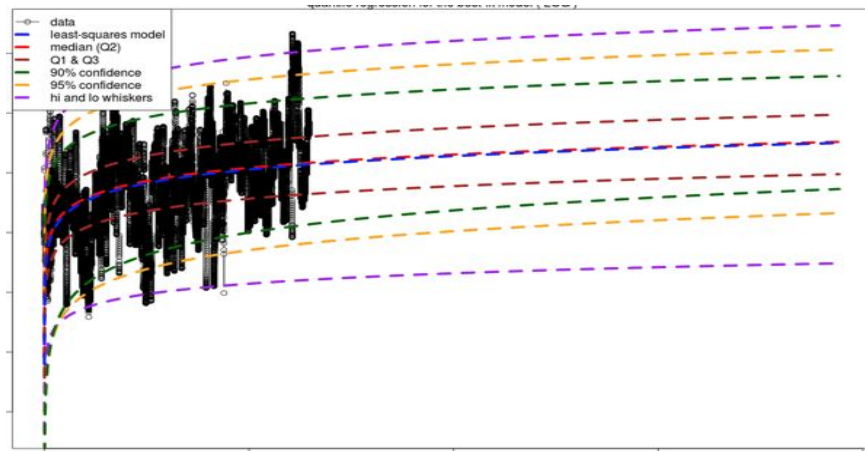
Usually the assumption in the standard approach is that latency will not change at higher throughput, which implies that throughput trajectory will be a good proxy for workload trajectory, and the workload forecast will be defined by that of the throughput:  $\overline{W}(t) = \overline{X}(t) * T$ . In addition, capacities of IT resources are typically measured in units of throughput (number of transactions per second; bits per second; integer operations per second; etc), which makes it convenient to measure workload as the rate of service of the arriving units of work. This creates a lot of confusion, but it is the current “state of the art”. With that in mind, illustrations below show throughput time series data.

### When the percentiles' trajectories behave “as expected”

The standard approach works well when the forecasted workload quantiles are nondecreasing, so that, for example,  $W_{.90}(i) > W_{.50}(i) > W_{.10}(i)$  and  $\overline{W}_{.90}(i) > \overline{W}_{.50}(i) > \overline{W}_{.10}(i)$ , or  $X_{.90}(i) > X_{.50}(i) > X_{.10}(i)$  and  $\overline{X}_{.90}(i) > \overline{X}_{.50}(i) > \overline{X}_{.10}(i)$  (see Figure 2).



(a)

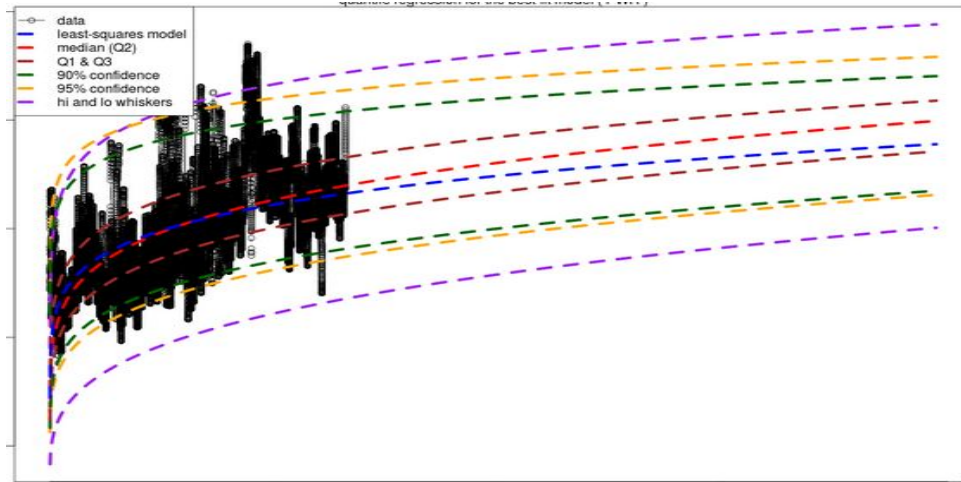


(b)

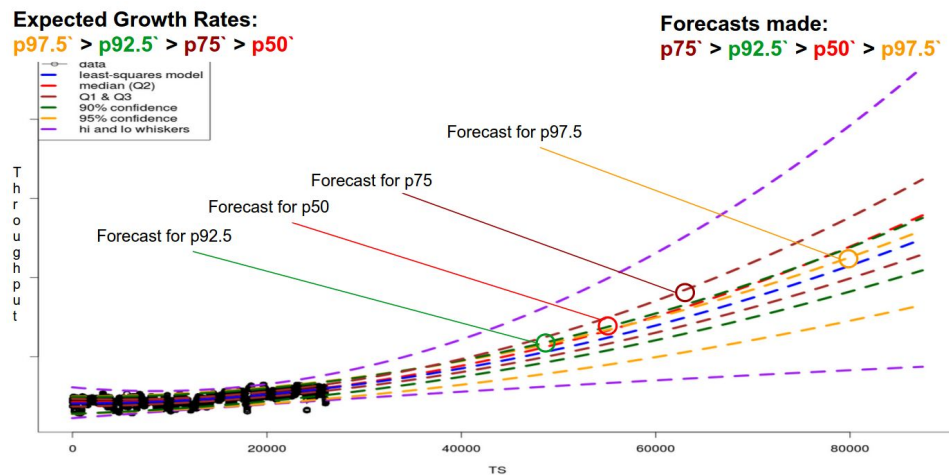
Figure 2: Examples of Throughput Time Series where the Standard Approach above works

Examples in Figure 2 show unconstrained throughput time series where trajectories of all percentiles are divergent (Figure 2a) and approximately parallel (Figure 2b). A close examination of Figure 2b reveals that the 5th and the 25th percentiles (third and fourth dashed lines from the bottom) appear to be converging, but their potential intersection is too far in the future to be material.

### When percentiles' trajectories converge



(a)



(b)

Figure 3: Throughput Time Series where percentiles' trajectories converge

Figure 3 shows examples with behavior different than Figure 2. In Figure 3a, 95th and 97.5th percentiles - second and third lines from the top - are converging to the 75th percentile. In Figure 3b percentiles' trajectories actually intersect, making the 3rd quartile higher than the 97.5th percentile (upper bound of the 95% confidence interval), and dropping the 97.5th percentile below the median. These lines reflect the growth rates. Their intersection merely means that they are converging very fast. Convergence, in turn, is important, because it points to saturation, as will be shown below.

In other words, the phenomenon does occur in practice, deserves explanation, and requires being dealt with.

Can it Be Explained?

Consider a resource-constrained system where a hidden or explicit feedback mechanism moderates the demand based on the workload, illustrated conceptually in Figure 4.

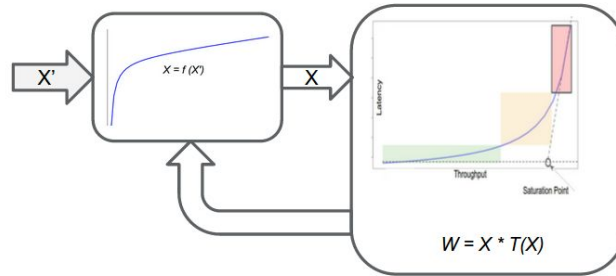


Figure 4: Workload Control System: a generalized view

If  $X$  (moderated demand) is below the “knee” (Figure 1), the mechanism will implement little or no reduction. At the knee, the latency ( $T$ ) grows quickly with growing  $X$ , leading to a disproportional increase of the workload ( $W$ ), as determined via Little’s Law. Thus  $W$  or a similar signal can be used when  $X'$  is large to effect an arrival rate  $X$  so that  $W$  does not exceed a target value.

Thus a congestion control mechanism like the one in Figure 4 seeks to ensure that

$$W = X * T \leq \alpha * N, \quad (2)$$

where  $\alpha =$  a coefficient,  $0 < \alpha < 1$ , and

$N =$  holding capacity of the connection (in units of work; e.g., packets)

Empirically, if  $W \leq \text{const}$ , and  $W = X * T(X)$ , upper percentiles of  $X$  will be dampened more than lower percentiles, especially when the demand is near the knee.

## Hyperbolic Intuition

As outlined in [GUNT2009] and independently in [FERR2014], the ROC curve near the knee (Figure 1) is approximated very closely by a hyperbolic function:

$$(X - L) * (T - H) = -A \quad (3)$$

Here  $L > 0$ ,  $H > 0$ , and  $A$  are parameters;  $A = f(\alpha * N)$ ;  $A > 0$ ;  $X =$  throughput;  $T =$  latency

This approximation follows from applying Little’s Law to a closed-loop queueing system. The slopes of the asymptotes are defined by Eq. (3) parameters, which, as demonstrated in [FERR2014], can be derived from known and measured parameters of the system.

For the open system, eq. (9) can be solved for  $T$  as :

$$T = H - \frac{A}{X-L} \quad (4a)$$

## Sensitivity Analysis

Sensitivity is calculated by taking first derivative:

$$\frac{dT}{dX} = \frac{A}{(X-L)^2} \quad (5)$$

Similarly, for throughput sensitivity to latency: in the open system

$$X = \frac{A}{H-T} + L \quad (4b)$$

Sensitivity is calculated by taking first derivative:

$$\frac{dX}{dT} = - \left[ - \frac{A}{(H-T)^2} \right] = \frac{A}{(H-T)^2} \quad (6)$$

## Note

*Eqs. (4a, 4b)* , as well as their “cleaner forms” (5, 6) demonstrate the asymmetrical relationship between throughput and latency in a closed system: higher throughput drives higher latency, but not vice versa; see the Interpretation section below.

Substitution of (5a) into (7) yields:

$$\frac{dX}{dT} = \frac{A}{(H-T)^2} = \frac{A}{\{H - [H - \frac{A}{X-L}]\}^2} = \frac{A}{\left[\frac{A}{(X-L)}\right]^2}$$

Finally, for a closed system:

$$\frac{dX}{dT} = \frac{(X-L)^2}{A} \quad (7)$$

Comparison of *Eq.(6)* and *Eq.(7)* confirms correctness of the derivation (3) - (7).

## Interpretation

As throughput increases, latency can only increase (*Eq. 4a*), whereas as latency increases, throughput can only decrease (*Eq. 4b*). Because  $A > 0$ , Equation (7) dictates that as we increase throughput in a closed-loop system, its upper percentiles must grow at a slower pace near the saturation point than lower percentiles; hence the patterns observed in Figure 3(a, b).

## “One should always generalize.” - Carl Jacobi

This discussion can be generalized by claiming (and proving, see below) that in a closed-loop system where  $X < X'$  , as throughput is approaching saturation point, its upper percentiles will grow at a slower pace than lower percentiles (compression of quantiles):

If  $X * T \leq \alpha * N$ , then

$$\lim_{X \rightarrow X_{saturation}} (\Delta_{X_P}) \leq 0 \quad (8)$$

where

$$\Delta_{X_P} = \left[ \frac{dX_P(t)}{dt} \right] - \left[ \frac{dX_{100\%-P}(t)}{dt} \right], \quad X_P = Pth \text{ percentile of } X; \quad 50\% < P < 100\%$$

The next section formalizes that empirical result mathematically.

## Quantile Compression Theorem

Here we provide some strong but reasonable assumptions where the empirical observation of compressed upper percentiles can arise, and formalize that result as a theorem and proof.

Let  $X(i)$  be a collection of throughput measurements over an interval of time. While it is not useful to speak of these measurements as drawn from a single distribution if there is a seasonal pattern, it is useful to speak of the expected value of each percentile. Thus for the  $a$ th percentile of  $X(i)$  we can write expected value  $E[X_a(i)]$  and similarly for  $X'(i)$  we can write  $E[X'_a(i)]$ . For convenience, we use  $Q$  to denote the natural logarithms of these expected values:  $Q_a(i) = \ln(E[X_a(i)])$  and  $Q'_a(i) = \ln(E[X'_a(i)])$ .

We assume that for any two time intervals  $i$  and  $j$  where the expected values of the  $a$ th percentiles of the unconstrained demand  $X'$  are scaled by some factor. For ease of derivations, we will set this factor to  $e^k$ . Thus for all percentiles  $a$

$$E[X'_a(j)] = e^k E[X'_a(i)] \quad (9).$$

For many resource-constrained systems (including data networks), the expected values of the quantiles are dominated by diurnal and weekly patterns that vary little with scale and time, and are well modeled by this assumption. Under conditions of demand growth over time, for  $j > i$  we will have  $k > 0$ , but growing demand is not a requirement for the theorem below.

We assume that the time scale of the dynamics of the system illustrated in Figure 4 are such that the expected values of  $X_a(i)$  and  $X'_a(i)$  can be related directly by a function that is dependent neither on the interval  $i$  nor the specific percentile  $a$ . For convenience, we write this function in terms of  $Q$  as  $Q = f(Q')$ , or in specific application as  $Q_a(i) = f(Q'_a(i))$ . This assumption is consistent with a system as shown in Figure 4 where the dynamical behavior dies out on a much faster timescale than the period of measurements, so that the measurements, or at least their expected values, can be treated according to a steady-state relationship that is purely a characteristic of the system.

For the system illustrated in Figure 4, we might expect  $f()$  to have a left asymptote that passes through the origin with a slope of unity, a right asymptote that is horizontal, and a reasonably smooth transition between the asymptotes. For our theorem, we apply more precise and general conditions consistent with these: that the derivative of  $f()$  is positive and monotonically decreasing. See Figure 4a.



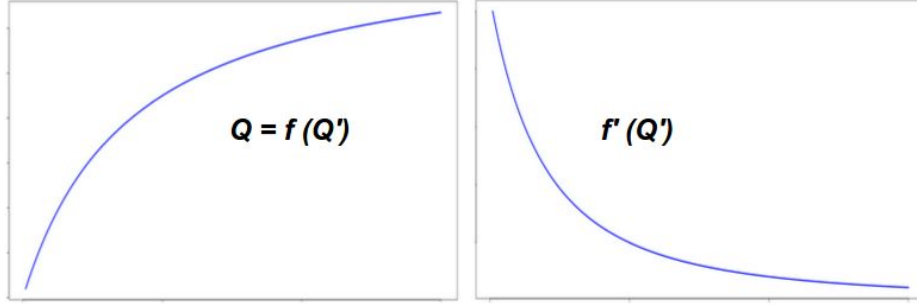


Fig. 4a: a form of the function  $f(Q')$  and its first derivative

Under these conditions, the following theorem specifies how a scaled increase in unconstrained demand produces lower percentiles that increase faster than upper percentiles.

### Theorem

Consider a resource constrained system with a moderated arrival rate where:

$X'_a(i)$  is the  $a$  th percentile of the unconstrained demands over a series of measurements in interval  $i$

$X_a(i)$  is the  $a$  th percentile of the moderated demands over the same series of measurements

The expected values of the percentiles of unconstrained and moderated demands in any measurement are related by a function  $f()$  so that  $Q = f(Q')$  where  $Q_a(i) = \ln(E[X_a(i)])$  and  $Q'_a(i) = \ln(E[X'_a(i)])$  and where the derivative of  $f()$  is positive and monotone decreasing.

Then if, in two intervals  $i$  and  $j$  the expected values scale by a common factor for all percentiles as  $E[X'_a(j)] = e^k E[X'_a(i)]$  with  $k > 0$ , then for any two percentiles  $a$  and  $b$  with  $b > a$ ,

$$\frac{E[X_b(j)]}{E[X_b(i)]} < \frac{E[X_a(j)]}{E[X_a(i)]} \quad (10)$$

### Proof

By the definition of  $f()$ ,

$$Q_b(j) - Q_a(j) = f(Q'_b(j)) - f(Q'_a(j)) \quad (11)$$

By taking the logarithm of both sides of the scaling relationship [might want to give it an equation number], we have

$$Q'_a(j) = Q'_a(i) + k$$

Then from elementary calculus and the characteristics of  $f()$ ,

$$Q_b(j) - Q_a(j) = \int_{Q'_a(j)}^{Q'_b(j)} f(y) dy = \int_{Q'_a(i)+k}^{Q'_b(i)+k} f(y) dy = \int_{Q'_a(i)}^{Q'_b(i)} f(y+k) dy < \int_{Q'_a(i)}^{Q'_b(i)} f(y) dy = Q_b(i) - Q_a(i).$$

So

$$Q_b(j) - Q_a(j) < Q_b(i) - Q_a(i) \quad (12)$$

Or, substituting and manipulating slightly,

$$\frac{E[X_b(j)]}{E[X_b(i)]} < \frac{E[X_a(j)]}{E[X_a(i)]} \quad (13)$$

QED

In words, as long as moderated demand  $X$  is related to unmoderated demand  $X'$  via a monotonically increasing damped function, when the system is approaching saturation, smaller percentiles of moderated demand grow on average faster than higher percentiles.  $\therefore$

## Applications

We have demonstrated that phenomena of “flat-topping” near resource saturation point need to be accounted for in capacity planning and performance engineering. Relationship (3) opens the way to a number of interesting approaches to, and applications of, analysis of resource-constrained system dynamics: the relative slowdown of growth in the upper bound is an indicator of the working point on the ROC curve getting closer to the saturation point.

## Resampling

One way to do so is to use [resampling](#) (jackknife or bootstrap):

1. Generate the bundle of lines representing the trajectory of all quantiles
2. Rebuild the distribution for each timestamp
3. Sample from the new distribution and obtain the 95th percentile for each timestamp.

Downsides of using resampling here:

- Resampling implementation is prohibitively slow and CPU-intensive.
- Resampling hides underlying problems with the system's dynamics.
- Resampling does not explain the “why” of the phenomenon.
- It introduces a resampling error due to approximation of the distribution at a future point.

## Congestion Detection

Throughput (being proportional to task arrival rate) is not normally distributed. In an unconstrained system, it is generally right-skewed (bulk of the data is on the left, or lower, side of the distribution, Figure 5a).

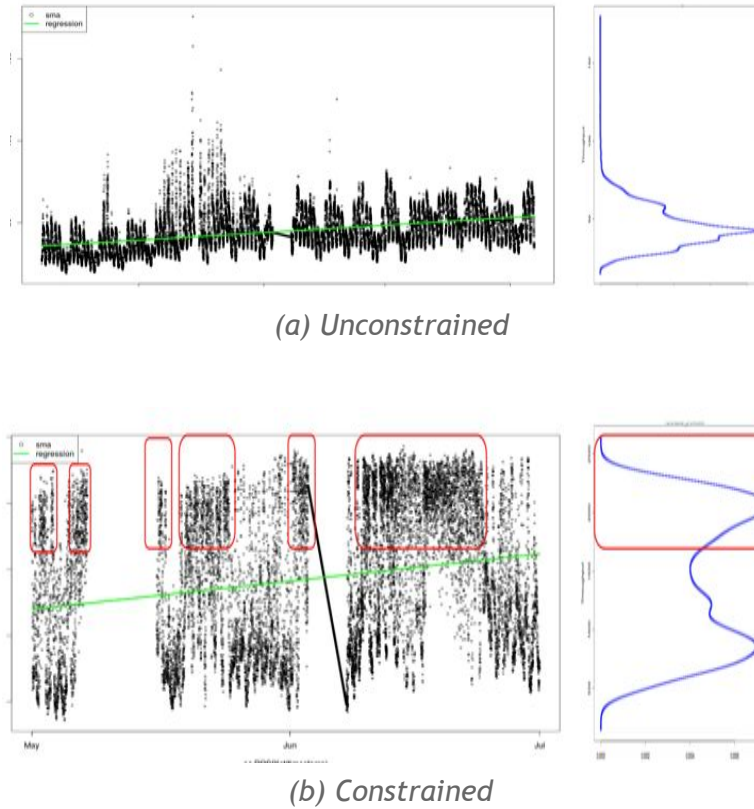


Figure 5: Throughput time series and distributions (the straight lines illustrate the use of linear interpolation to connect data samples)

As a corollary of the statement above, in a constrained closed-loop system the data can become bimodal and even right-skewed. (Figure 5b).

### Saturation Prediction

The percentiles' trajectories in Figure 3b point to a future saturation and possibly congestion. This statement is a direct corollary from the Statement (3) above. It leads to a very simple approach to congestion forecasting:

For the Throughput,  $X(t)$ , data:

1. Forecast the trajectories of two symmetrical far-away percentiles (e.g., first and third quartiles, 10th and 90th percentiles, etc.). Compute the distance between these two lines at each timestamp,  $D(t)$ .
2. Forecast the  $D(t)$  and find where  $D(t) = 0$ . This is the saturation point as found by these percentiles.

Following the same steps for multiple pairs of percentile lines will result in a distribution of the congestion point prediction, leading to a measure of prediction interval. In capacity planning, this will give the analyst an idea of how urgent it is to add capacity to a resource, and how much latitude there is.

### Forecasting Growth

If the growth rates of different percentiles are asymmetric - upper percentiles are unable to grow as fast as lower bound due to capacity constraint (reaching saturation point) - how much

capacity do we need to add to enable upper percentiles it to grow as fast or faster than lower percentiles? Because in capacity planning, we want to provision for the upper bound of the throughput distribution, it is a very relevant question.

If the throughput is growing, we can use an earlier time, when it was not constrained, to compute the skewness of the throughput distribution. Skewness, being the third standardized moment, is a property of the distribution that is distinct from the other moments (mean, variance, and kurtosis). It is fair to say that it is the property of the distribution itself and will be preserved unless the system becomes constrained.

An alternative measure of skewness is the Quartile (Bowley's) form, which defines it using only the three quartiles:

$$\text{quartile skewness} = \frac{Q_3 + Q_1 - 2 * Q_2}{Q_3 - Q_1} \quad (14)$$

where  $Q_3 = UB = \text{upper bound (p75)}$ ;  $Q_2 = M = \text{median}$ ;  $Q_1 = LB = \text{lower bound (p25)}$

### Forecasting Method for the Higher Percentiles based on Lower Percentiles

1. For each time interval (hour, or day, or week), compute history-based skewness<sup>1</sup>:

$$C = \text{median} \left[ \frac{UB(t) + LB(t) - 2 * M(t)}{UB(t) - LB(t)} \right] \quad (15)$$

where  $C = \text{estimate of quartile skewness for the time series}$ . It is natural to assume that the measured skewness (14) will vary from one time interval to the next; if we treat quartile skewness as stationary, we are dealing with a distribution of quartile skewness. We further assume that quartile skewness, or at least its median, can be treated as stationary. Stationarity will be lost during transition into and out of constrained state; however, such transitions tend to happen undetectably fast in data spans typically used in forecasting (hours or days in transition vs. months or years of historical data). Figure 6 is an illustration of quartile skewness of throughput for a network link over the course of 7 months.

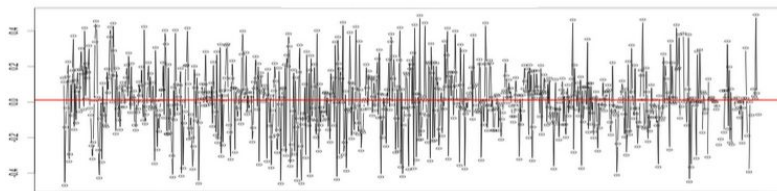


Figure 6. Daily Quartile Skewness for a typical resource

The median is used, rather than the mean, in order to reduce the influence of extreme values. It is computed over all historical-data intervals for which  $UB$ ,  $LB$ ,  $M$  have been computed.

<sup>1</sup> We had success with daily quartile skewness, but time interval choice depends on the data.

2. For each point  $\bar{t}$  of the forecast horizon, use quantile regression (see, e.g., [\[FERR2012\(2\)\]](#) for using quantile regression in capacity planning) to compute

$$\overline{UB}(\bar{t}) = \frac{2 * \bar{M}(\bar{t}) - \bar{LB}(\bar{t}) * (C + 1)}{1 - C} \quad (16)$$

where the bars designate future values:  $\bar{\xi}(\bar{t})$  is the forecasted value of  $\xi$  at time  $\bar{t}$ . An implementation of a forecasting algorithm based on Eq. (16) is shown in Figure 7.

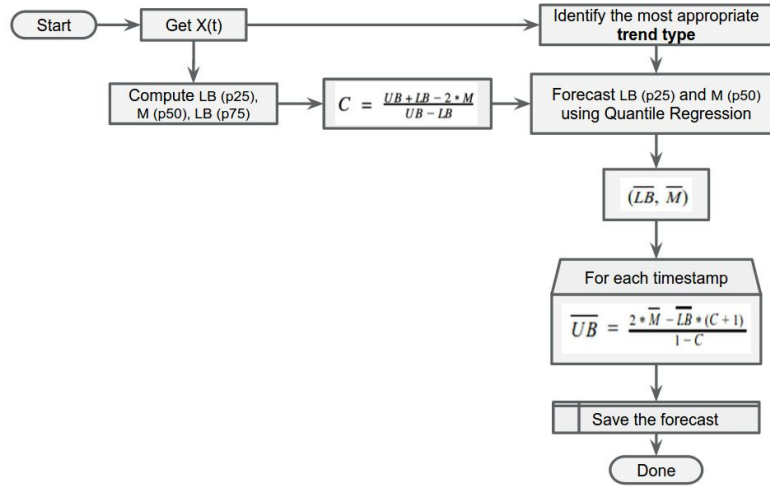
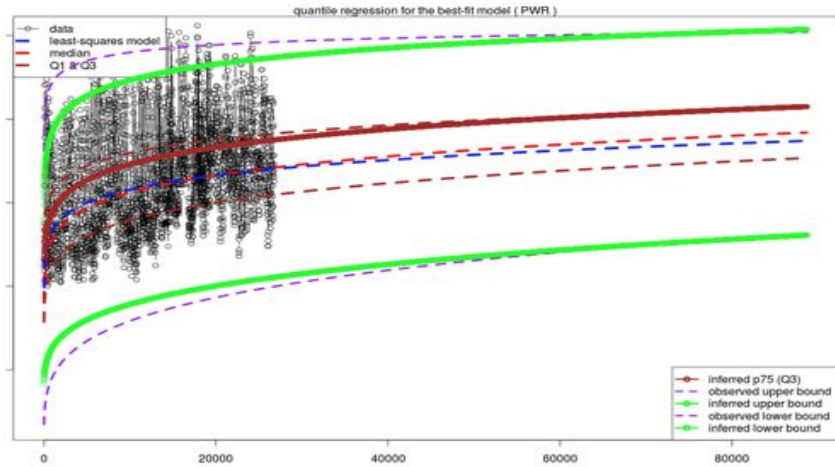


Figure 7: Inferring the forecast of the high percentiles of throughput distribution

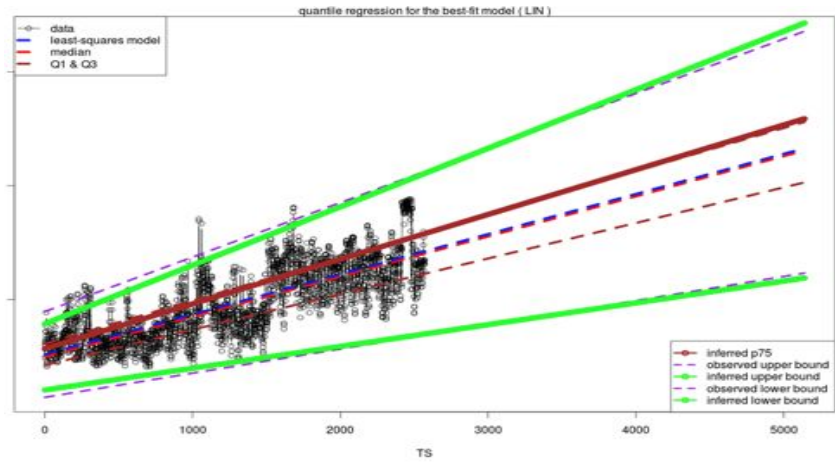
If the data were constrained in the historical time range used in forecasting, then the inferred line will come out same or below the directly computed line.

## Results

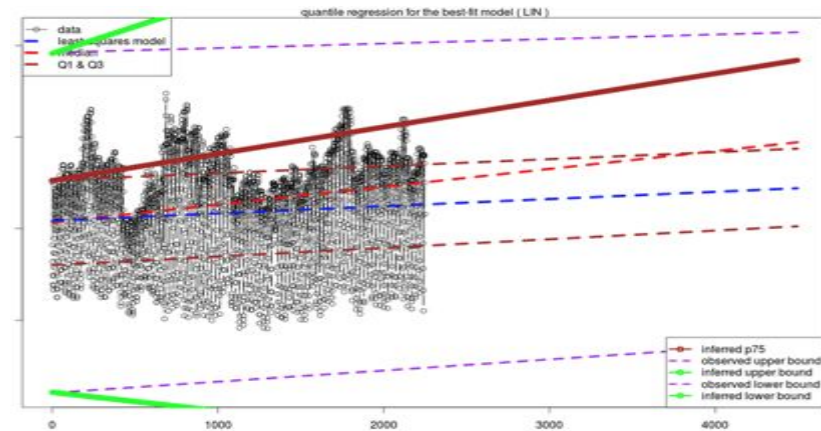
Figure 8 illustrates throughput data along with quantile-regression and inferred forecasts for network connections. The lines correspond to the first quartile (Q1), median (Q2), and third quartile (Q3), as well as inferred Q3 and forecasted and inferred upper and lower outlier boundaries (constructed using [Tukey's IQR method](#)) for the three possible scenarios.



(a) Unconstrained resource: Inferred outlier boundary (first solid thick line from the top) stays below the computed outlier boundary (first dashed thin line from the top).



(b): Slightly constrained resource: Inferred outlier boundary is close to the computed outlier boundary, but overtakes it at  $TS \sim 3000$ .



(c) Already congested

Figure 8: Inferred and calculated upper bounds and their relative positioning to other percentiles' trajectories

For a congested resource (Figure 8c), we see that the distribution is completely skewed to the left; the inferred outlier boundary is so steep that it is outside the frame of the picture; the inferred Q3 projection is going significantly higher and steeper than the forecasted Q3 projection, and the median line catches up to the computed Q3 projection at TS ~ 3900.

## A Use Case Example

Consider an enterprise having one or more ISP connections from their offices. The IT group needs to forecast the ISP requirement at least 6-12 months in advance to ensure on-time delivery. The throughput  $X(t)$  leaving the enterprise's interface is limited by the link's bandwidth (e.g., for an OC-3 -- 155 mbps -- link and packet sizes of 1000 bytes,  $X_{max}(t) \leq 19375 \text{ pps}$ ). One can analyze the observed hourly or daily boxplots of  $X(t)$  for the past year and estimate the quartile skewness using *Eq (15)*. Using methodology outlined in *Figure 7*, one can then obtain the forecast of the inferred 75th percentile of  $X(t)$  and use it to infer the upper boundary forecast. The latter can be converted to line bandwidth requirement of the ISP connection. Note that if at any point in the forecast horizon the inferred upper boundary projection exceeds 19375 pps, then this connection would require urgent attention of the IT team.

## Conclusion

When we forecast demand for an IT resource based on the 95th percentile, the information carried by the lower percentiles (95% of the data) remains unused, "the quiet ones".

On the other hand, we have demonstrated and proved mathematically that when the resource is already approaching its saturation point, the 95th-percentile approach can mislead capacity planners to undersizing the demand. Consequently, we will always be keeping ourselves busy upgrading capacity for such resources, which are often on critical path.

The method proposed in this paper allows detecting and predicting congestion and sizing resource based on the trajectory of the bulk of the flow (the quartiles, and in particular the first and second quartiles), which makes it possible to improve the efficiency of the capacity planners' and performance analysts' work.

## Acknowledgments

Authors express their sincere gratitude to Deepak Kakadia, Matt Mathis, Andrew McGregor, Harpreet Chadha, and Mahesh Kallahalla for making this paper possible and for reviewing it prior to submission.

## References

1. [GUNT2009] Mind Your Knees and Queues - Gunther. MeasureIT, Issue 62, 2009
2. [FERR2012(1)] A Note on Knee Detection. Ferrandiz and Gilgur. Las Vegas : CMG 2012
3. [GILG2013] Little's Law assumptions: "But I still wanna use it!" The Goldilocks solution to sizing the system for non-steady-state dynamics. Gilgur MeasureIT Issue 100, June 2013

4. [CHDY2014] Back to the Future of IT Resource Performance Modeling and Capacity Planning. Choudhury. Proceedings of the 2014 3rd International Conference on Educational and Information Technology (ICEIT2014). Toronto, Canada, 2014.
5. [CRFT2006] Utilization is Virtually Useless as a Metric! Cockcroft. International Conference of the Computer Measurement Group (CMG'06). Reno, NV, 2006.
6. [FERR2012(2)] Level of Service Based Capacity Planning. Ferrandiz & Gilgur. CMG '12 International Conference of Computer Measurement Group. Las Vegas, NV, 2012.
7. [FERR2014] Capacity Planning for QoS. Ferrandiz and Gilgur. Journal of Computer Resource Management. Issue 135 (Winter 2014). pp. 15-24
8. [OSTR2011] Minimizing System Lockup During Performance Spikes: Old and New Approaches to Resource Rationing. Ostermueller. Proceedings of the 37th International Conference of the Computer Measurement Group (CMG'11). Washington, DC.
9. [LTTL2011] Little's Law as Viewed on Its 50th Anniversary. Little. OPERATIONS RESEARCH Vol. 59, No. 3, May–June 2011, pp. 536–549
10. [MAKR1998] Forecasting: Methods and Applications. Makridakis, Wheelright, Hyndman. Wiley, 1998.
11. [ARMS2001] Principles of Forecasting: A Handbook for Researchers and Practitioners. J. Scott Armstrong (2001) Springer, 2001.