The Curse of P90: An Elegant Way to Overcome It Without Magic

Alexander Gilgur, Tyler Price, Alexander Nikolaidis, Steve Politis
(Facebook)

“As we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don’t know we don’t know.”

Donald Rumsfeld

Abstract

Over the decades of development of methodologies and metrics for IT capacity planning and performance analysis, percentile terminology has become the *lingua franca* of the field. It makes sense: percentiles are easy to interpret, not sensitive to outliers, and directly usable for approximating the distribution of the variable being measured for stochastic simulations. However, depending on which percentile is used, we can miss important information, like multimodality of the metric’s distribution. Another, less obvious, downside of relying on percentiles comes into play when we size infrastructure for a high percentile of demand (e.g., p90). Given that it takes time to order, manufacture, receive, and install infrastructure, this means that we need to answer the statistically nontrivial question, “what will this percentile of demand be a few years from now?” This paper discusses the issues that arise in answering it and proposes an elegant way of resolving them.

Introduction and Problem Statement

It has been discussed at CMG, Sigmetrics, in books (e.g. [BUZN2015]) and other venues. Percentiles are not good for capacity planning – and they are not useful for performance analysis either.

Percentile measurements are based on counters and therefore do not tell us by how much the data may exceed the p90 (or p95, or p75) line, which puts under question the validity of using them in capacity and performance analytics. Even if we have a bimodal distribution (e.g., in [GIGB2015], the authors describe the workload becoming bimodal under congestion), percentile measurements will not treat it any differently from a unimodal distribution.

On the other hand, percentiles are very attractive for reporting. Indeed, they are easily interpretable at any level of understanding of statistics; they are insensitive to outliers; and they can be meaningfully estimated for any distribution, whereas mean and standard deviation are not as versatile: there are distributions (e.g., Pareto and Cauchy are often observed in the IT world) where these statistics are not even defined.

Additionally, if we do not use percentiles, the list of our options is limited: we either assume a distribution that we would like to work with (e.g., Poisson for throughput or Exponential for latencies), or we fall into the trap of sampling: by focusing on the means of sufficient number of samples, we cause the data we work with to follow the normal (Gaussian) distribution by virtue of the Central Limit Theorem (CLT). If we go with the former, we need to justify the assumptions – which is often a non-trivial task. If we go with the latter, we have to be very careful about conclusions drawn from data distorted by CLT.

In short, we must use percentiles, and we must make them work.

Measuring Demand

Demand is usually measured as the workload on the system in terms that make sense for the problem being solved.

E.g., if we are monitoring network traffic, demand can be the throughput (number of packets per second, or bits per second) on a network link between two nodes. Then we can alert to any "hot" links and reroute the traffic based on network topology and routing policies. This is the general idea behind the Software-Defined Networks (SDNs), as well as some TCP protocols.

On the other hand, capacity planning for a network requires a different definition of demand: we cannot be sure whether any "hot" links observed in historical data were due to poor routing or by design. If it was by design (e.g., we want to protect downstream network elements from overloading by routing more traffic through alternative routes), we want to perpetuate it in the plans. If it was due to poor routing, we should forecast demand between the initial source and destination endpoints (A-Z demand) and then simulate the network to predict the throughput for each link of the network.

For non-network resources, we can monitor load on the CPU, memory usage, I/O speed, storage utilization, latency, or CPU utilization. Parenthetically, while [CRFT2006] explained why it is not a good idea, [BAHÖ2007] proposed using utilization as the signal for power consumption. While intuitive, their paper does not show how accurately it predicts the upper tails (near 100%) of CPU utilization and power capacity utilization. Over the last 10 years, there have been a number of pubilications explaining that for VM and in general for multiprocessing systems with statistical multiplexing (e.g., hyperthreading), utilization will overestimate the actual load that the machine sees, whereas for network it is most likely to underestimate the actual load that a particular link or circuit sees. In 2016, [DOGI2016] proposed a new very promising metric; however, work still needs to be done in this direction.

In general, demand for an IT resource can be inferred from a causal model or from historical data. For example, if we know how many business transactions we expect at the end of the year, we can ask ourselves how much workload we should expect at that time and build a model connecting the workload with the transaction counts. However, often the demand variability is so high that the model's predictive ability is limited. In other cases, the interactions of components are so complicated that it makes unfeasible to go into these depths and analysts often resort to time-series analysis, which will miss the effect of changes in demand due to changes in the underlying causal variables.

Sources of Demand Variability

Discussion of sources of noise and variability in IT hardware demand has been in the literature for quite some time. It is important to understand that, even when demand is measured in bits (packets, queries, instructions) per second, demand and measured traffic are two different aspects of the way we describe the behavior of our systems. Demand can only be inferred from measured data. Therefore, there already is an uncertainty baked into it. Specifically [GIEC2014] describe the following sources of stochastic behavior of network demand:

(1) Pure stochastic noise
(2) Temporal Non-Stationarity
(3) Spatial Non-Stationarity
(4) QoS Requirements and activities related to the effort of meeting them
(5) Uncertainty inherent in forecasting demand.
(6) Variations in latency and packet loss can lead to variations in demand, especially with protocols inducing retransmission (e.g., TCP): indeed, when lost transmissions require re-sending of the packets, this increases the load on the system.

Tracking Demand Variability

The adage, "A lot can happen in 60 seconds", is especially true when we are talking about capacity usage dynamics. It is even more so when it comes to the "busy-hour" traffic (BHT): it is often hard to find 60 minutes per day of historical data when the system was consistently stationary while being highly utilized.

To add complexity to the issue, when forecasting demand from measurements on a system that is approaching saturation, high percentiles tend to slow down, while lower percentiles are growing (the "Quantile Compression" effect [GIGB2015]), leading to unrealistic predictions of "no growth" right when the resource is a bottleneck.

Conversely, in a system that is not constrained, it is the upper percentiles of demand that behave most erratically and therefore contain more stochastic noise than any other part of the distribution (Figure 1).

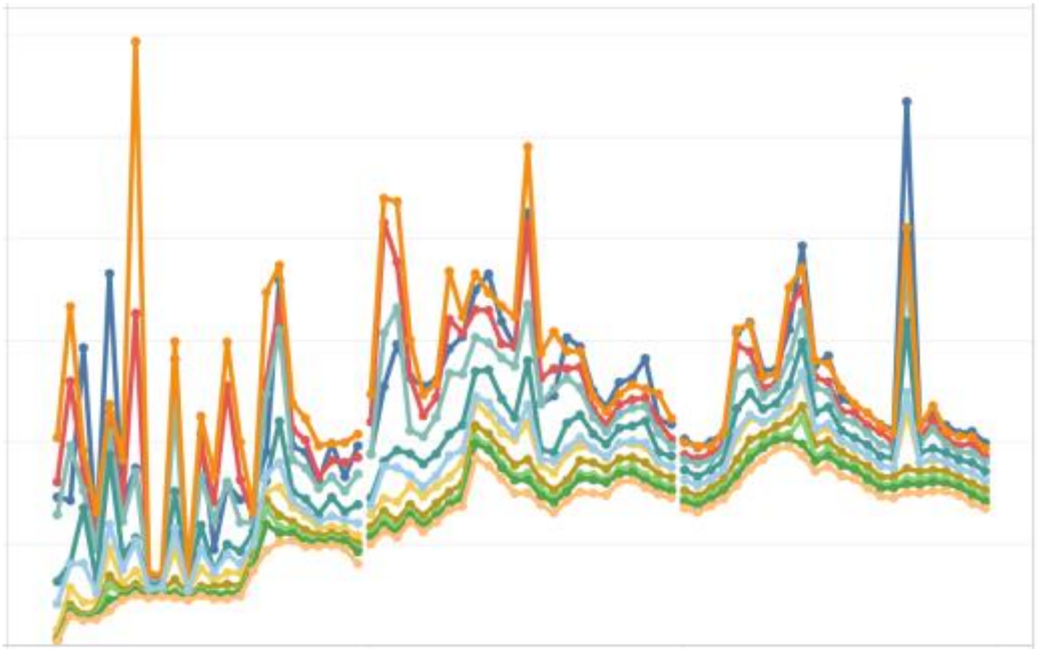

Figure 1: an illustration of inter-metro demand observed on a large backbone network. The graphs show hourly variability over the course of 3 days. Each color corresponds to an hourly measure of demand (min, max, p05, p10, p25, median, mean, p75, p90, p95, p99, and upper bound)

While these phenomena – quantile compression in a constrained system and erratic behavior in an unconstrained one – impose the need to revise the wisdom of using p90 (or p95) in forecasting demand, the alternatives are not better. E.g., we frequently see long-haul network demand where p90 exceeds the upper bound (outlier boundary) of the unconstrained data. In such cases, we do not want to use outlier boundaries as the measure for which we are forecasting.

Generally speaking, directly forecasting upper ranges of the demand data is not conducive to good predictions.

A brief discussion of solutions for dealing with demand variability is in Appendix A1. A way to use Monte-Carlo methodology for dealing with variable demand was presented in [GIEC2014].

Proposal

The closer we get to the bulk of the distribution, the better we will be at identifying the important growth patterns and the more accurate we will be at predicting the system demand.

We propose forecasting $p75$ and a measure of uncertainty in the data to get us to the $p90$ (or $p95$, $p99$, etc.)

In a more rigorous notation:

$$P_{\beta}^{Fcst}(t) = CI_{\alpha}^{Fcst}(P_{\gamma}(t)) \qquad (1)$$

where
t is the time (or point index for discrete data)
α is the confidence level;
γ is the quantile of historical data used in the forecast;
β is the quantile of demand for which we are sizing the resource;
$P_\gamma(t)$ is the γ – quantile of the time series at time t (in historical data or in forecast);
$CI_\alpha^{Fcst}(P_\gamma(t))$ is α- confidence interval of the forecast of γ quantile of the historical data.

Note that, if $\beta > 0.5$, then for Eq. (1) to work, we need $\gamma < \beta$.

Discussion

In forecasting, the prediction interval is determined as

$$P_\alpha(t) = Y(t) + f(\Delta, \alpha, t) \qquad (2)$$

where
t is the time (or point index for discrete data)
Δ is a measure of the forecasting model's uncertainty;
α is the confidence level;
$P_\alpha(t)$ is the upper bound of the forecast at the confidence level of α. It
$Y(t)$ is the point forecast;

As explained in [MWHH1998], upper bound at confidence interval α is the prediction of the α-quantile of the time series; in other words, Eq. (1) says that, at any point in time, we have $100 * (1-\alpha)\%$ chance of having a data point above the point $P_\alpha(t)$ – or $100 * \alpha\%$ chance of having a data point below the point $P_\alpha(t)$.

We say that β is the probability that future historical data points will be below the line we are forecasting. In other words, we need to calculate the α such that Equation (1) is true.

Note that Equation (1) does not imply any specific forecasting model or technique: it will be true for time-series forecasts (ARIMA and ETS), linear (least-squares), quantile, and any other regression technique.

Formally, for a process variable X (e.g., traffic on a link, or network demand, or CPU load), we can have three possible events:

$$A: X(t) \geq P_\gamma(t) \qquad (3a)$$

$$B: P_\beta(t) > X(t) \geq P_\gamma(t) \qquad (3b)$$

$$C: X(t) \geq P_\beta(t) \qquad (3c)$$

We want (3c) to be true and (3b) false; in other words, we want to find

$$\alpha \colon P_\alpha(t) = CI_\alpha^{Fcst}\left(P_\gamma(t)\right) = P_\beta(t)$$

E.g., if $\gamma = 0.75;\ \beta = 0.9$, we need to compute the uncertainty level at which our model's prediction made on $p75$ will correspond to the $p90$.

In probabilistic terms, we need to answer the question: what is the probability of event $C = B|A$?

Assumptions

1. Data distribution within the time intervals for which the $P_\gamma(t)$ is taken is independent of the distribution between these time intervals.
2. $1 \geq \beta > \gamma$

Proof

From Bayes theorem, for conditional probability:

$P(C) = P(A) * P(C|A)$ (3)

Then

$P(C|A) = P(C)/P(A)$ (4)

But we know that

$P(A) = 1 - \gamma;$ (5a)

$P(C) = 1 - \beta$ (5b)

Finally,

$\alpha = P(C|A) = \frac{P(C)}{P(A)} = \frac{1-\beta}{1-\gamma}$ (6)

Eq. (6) is what we set out to prove. ∴

Within the framework of our example, we have:

$\alpha = \frac{1-0.9}{1-0.75} = 0.6$ (6a)

To obtain a prediction of the 90th percentile of the time series, we need to use the 60% confidence interval of the forecast of the 75th percentile of the time series.

Note that Assumption 2 is critical for (6) to be true: indeed, $1 \geq \alpha = \mathrm{P}\ (\mathrm{C}|\mathrm{A}) \geq 0$ requires Assumption 2 to be true.

Example-Based Validation

To validate our proposal, we turn to an example. Suppose that we have a time series of hourly $p75$ of demand for a resource, and we want to prove that at 60% confidence level, we will underpredict the demand no more than 10% of the time.

If confidence level is 60%, then we need to demonstrate that

$$\Pr\{X < p75(X) + p60(p75(X))\} = 0.9 \qquad (7)$$

The proof is below:

$$\begin{aligned}
&Prob\,[\,X\ <\ p75(X)\ +\ p60(p75(X))\,]\ =\\
&\quad 1 - Prob\,[\,X\ >\ p75(X)\ +\ p60(p75(X))\,]\ =\\
&\quad 1 - Prob\,[\,X\ >\ p75(X)\,] * Prob\,[\,X\ >\ p75(X)\ +\ p60(p75(X))\,|\,X\ >\\
&\quad p75(X)\,]\ =\\
&\quad 1 - 0.25 * Prob\,[\,X\ -\ p75(X)\ >\ p60(p75(X))\,|\,X\ >\ p75(X)\,]\\
&\quad =\ 1 - 0.25 * Prob\,[\,residual\ of\ p75(X)\ greater\ than\ 60\%\ confidence\ interval\,]\\
&\quad =\ 1 - 0.25 * 0.4\\
&\quad =\ 1 - 0.1\\
&\quad =\ 0.9 \qquad (8)
\end{aligned}$$

Real-Life Examples

Forecasting p90 and p75

If we were to forecast p90 with zero uncertainty interval, we would have the grey line. The forecast is essentially flat. Adding uncertainty to the forecast tends to add a growing prediction interval (see e.g., [MWHH1998]). This growth reflects the fact that the farther away we are from the last historical point of known value, the less certain we are of the prediction made using the forecasting model. However, by applying Eq. (6), we calculate that forecast of p90 with the 60% confidence interval would bring us to the equivalent of a p96 line (Figure 2).

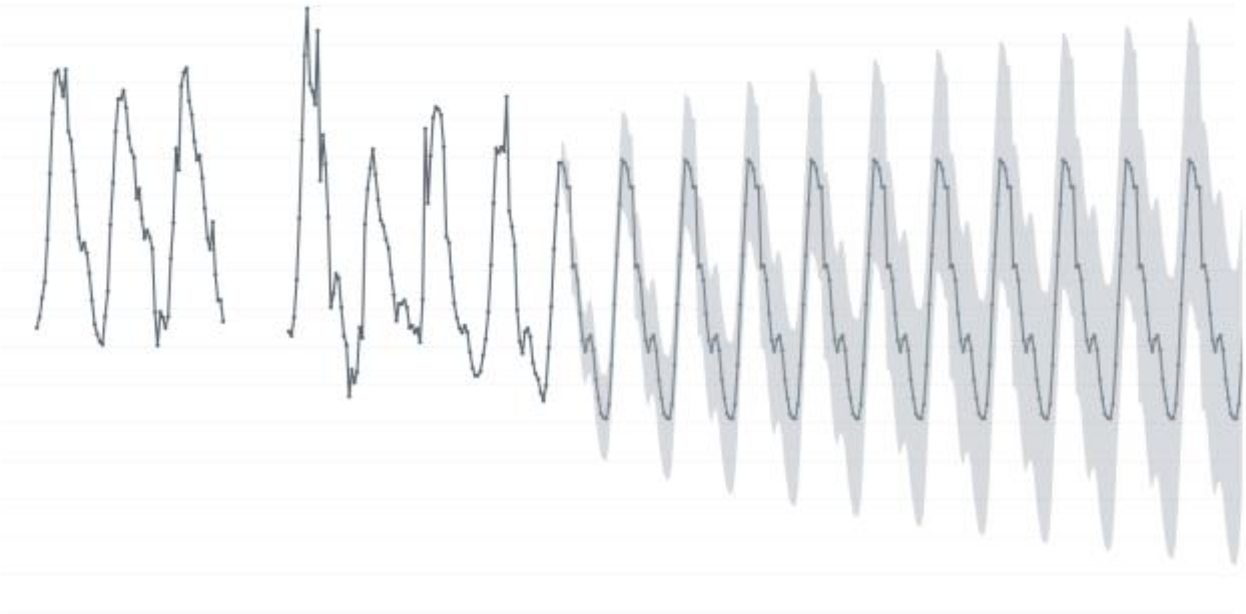

Figure 2: time series of the p90 of demand (live network data collected from the same source as in Figure 1): historical data and forecast with its 60% prediction interval.

On the other hand, applying Eq. (6) to the p75 line reveals that p75 line with the 60% prediction interval will produce the equivalent of the trajectory of p90 (Figure 3):

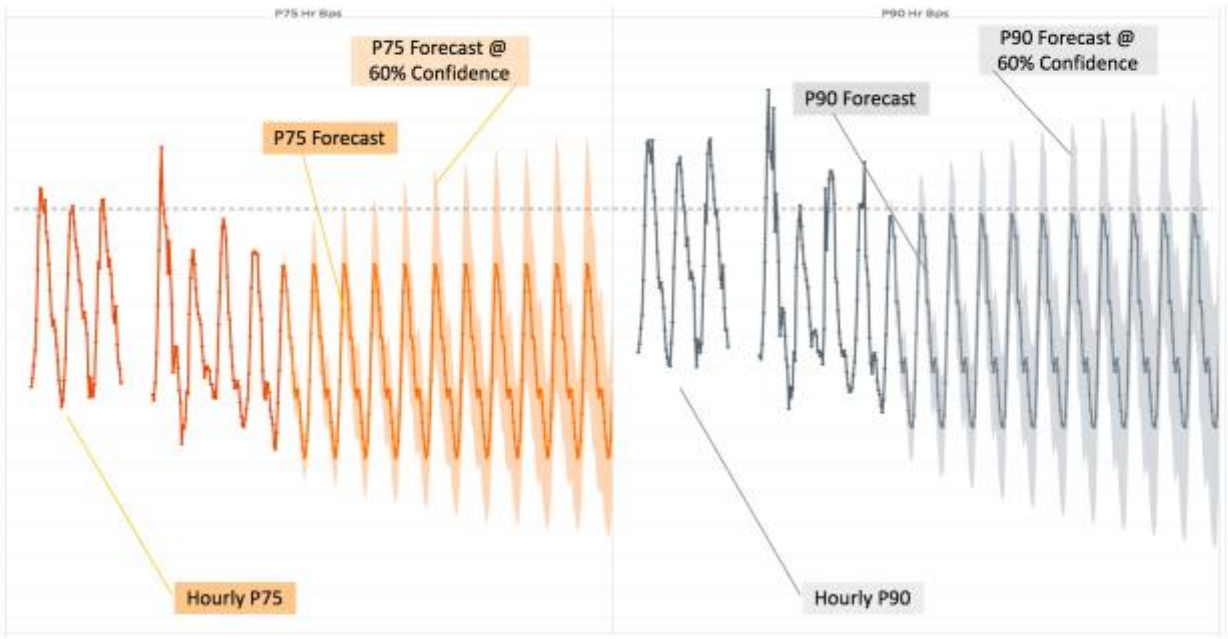

Figure 3: a time series of demand (live network data collected from the same source as in Figure 1) showing hourly p75 (red) and p90 (grey) lines. The forecasts of p75 and p90, as well as their 60% prediction intervals' ranges, are shown as orange and light-grey lines and shaded areas, respectively.

As we can see, the p90 point forecast (grey line) corresponds to the early values of the p75 forecast with a 60% prediction interval (orange shaded area). The growing uncertainty in prediction dictates that the forecasted value will grow, even if p75 (or p90) shows no growth in the historical data.

When the Demand Is Constrained

In the scenario of constrained demand, where quantile compression takes place, we are more likely to see growth in the lower percentiles than in the higher percentiles. Again, forecasting p75 will be advantageous in this case, as not only do we see that the system under consideration is working under constraint (p75 growth is faster than p90 growth), but also that using p90 prediction made based on p90 of the demand data would lead to under-provisioning the system (see Figure 4).

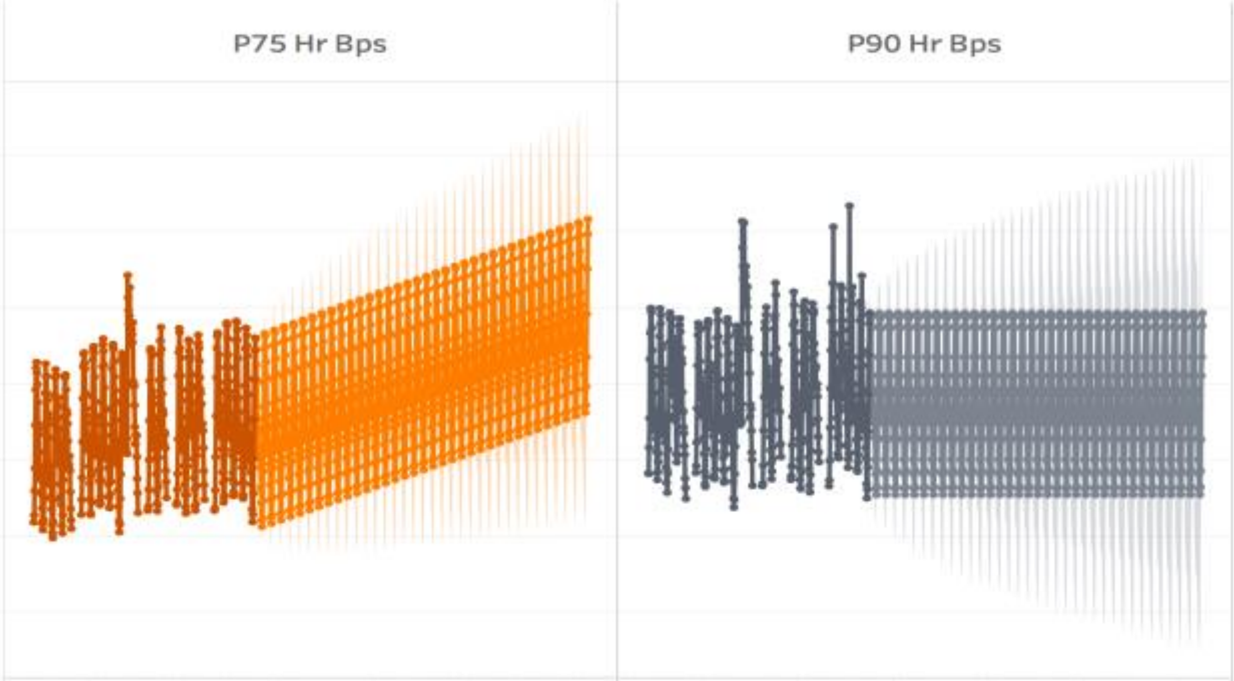

Figure 4: a time series of constrained demand on an inter-metro backbone network showing hourly p75 (red) and p90 (grey) lines. The forecasts of p75 and p90, as well as their 60% prediction intervals' ranges, are shown as orange and light-grey lines and shaded areas, respectively.

Side Note: Outlier Boundaries

For forecasting outlier boundaries $B_{upper} = p75 + 1.5 * (p75 - p25)$; $B_{lower} = p25 - 1.5 * (p75 - p25)$, this approach will not work because outlier boundaries do not represent percentiles. They do, however, reflect the degree of randomness in the data and can be forecasted directly.

Conclusion

We have proved mathematically, and demonstrated on examples, that using 60% prediction interval on top of the p75 forecast is a better way to forecast p90 of resource demand than using directly p90 of the historical data. Combined with the calculation of unconstrained p75 (see [GIGB2015]), we can forecast the trajectory of p90 of unconstrained demand, even if historically our resource was constrained.

We have also developed a mathematically rigorous way to compute forecast of any percentile of demand: e.g., for p95, we can use $100\% * \left(1 - \frac{1-0.95}{1-0.75}\right) = 80\%$ prediction interval of p75 (see Eq. 6).

We derived this methodology to overcome issues arising from forecasting time series, but the proof provided in this paper is generic enough to be applicable to any stochastic modeling methodology, from time series analysis to causal models to Monte-Carlo simulations. The power of this method is in the fact that it accounts for both the intrinsic uncertainty in data and the uncertainty of model prediction, allowing demand analysts and capacity planners to continually improve the quality of their models.

Appendix A1. Dealing with Demand Variability

Stationary Demand

When the demand (or its growth) is stationary, we can deal with its wide variability by studying the behavior of different percentiles of the data. Purely statistical methods have been recommended by, e.g., [KOEN2001]. Alternatively, when we know that we are dealing with a mix of demand from different sources, we can split the data into clusters using, e.g., the Gaussian-Mix-Model (GMM) method based on demand level, trend, and/or seasonality.

Non-Stationary Demand

Stationary demand for an IT resource is a rarity. One way to handle non-stationary demand is to identify the drivers of such behavior and to work with causal models. [GIFE2014] proposes a way to work around such non-stationary behavior by fitting the causal model and reformulating the question being answered by the capacity planner: instead of “what will my demand be in N years”, to answer the question, “what will my demand be in N years when its explanatory variable reaches the value of X”.

This method has been proven and implemented for application performance monitoring and demand forecasting. Similarly, a causal model allows the analysts to have a better prediction of future behavior of the demand metric.

Demand Mixtures

A frequent cause of non-stationary demand is blending different sources into one destination. Even when done with the best intentions, it throws stationarity off and leaves the capacity planners with a bigger problem than they are typically prepared to handle.

The Palm-Khintchine theorem offers a solution for the case where we are dealing with high numbers of demand sources: for large numbers of equilibrium renewal processes, each with a small intensity, the superposition of such processes will converge to a Poisson distribution. This fact has been used successfully in the CMGimPACt2016 paper by Brady and Gunther [BRGU2016] for generating demand for performance simulations.

However, it being a limit theorem, and the conditions that the underlying processes should be equilibrium and low-intensity, imposes certain serious limitations on its usefulness in dealing with non-stationary demand that is often observed in the infrastructure demand data.

This means that identification of segments in the demand time series where it loses stationarity is of paramount importance in performance monitoring and capacity analytics. Research into methodologies that help detect such segments and identify their root causes is outside the scope of this paper.

An alternative solution is to identify segments of data where a number of signals move together. This is achievable by distribution-based clustering (from something as basic as K-means to e.g., DBSCAN and Gaussian-Mixed Model, or GMM) and factor analysis, e.g., Principal Component Analysis (PCA). Detailed discussion of these methods can be found in any advanced statistics or basic machine learning text and is outside the scope of this paper as well.

References

[BZN2015] Buzen, J. (2015) Rethinking Randomness: A New Foundation for Stochastic Modeling CreateSpace/Amazon, August 2015. ISBN: 978-1-508-43598-3.

[KOEN2001] Koenker, R.; Hallock, K. (2001) Quantile Regression – Journal of Economic Perspectives. Vol. 15, Number 4, Fall 2001. Downloaded from http://master272.com/finance/QR/QRJEP.pdf

[BAHÖ2007] Barosso, L.A., Hölzle, U. The Case for Energy-Proportional Computing. *Computer*. ***40*** *(12): 33–37.doi: 10.1109/mc.2007.443*. Downloaded from http://www.barroso.org/publications/ieee_computer07.pdf

[MWHH1998] Makridakis, S.; Wheelright, S.; Hyndman, R. (1998) Forecasting: Methods and Applications, 3rd Edition. John Wiley & Sons, ISBN: 978-0-471-53233-0

[CRFT2006] Utilization is Virtually Useless as a Metric! Cockcroft, Adrian. International Conference of the Computer Measurement Group (CMG'06). Reno, NV, 2006.

[GIFE2014] Gilgur, A.; Ferrandiz, J. (2014) Predictive SPC – presented at ASA Conference for Statistical Practice (CSP2014. Tampa, FL. February 2014)

[GIEC2014] Gilgur, A.; Eck, B. (2014) Sources of Traffic Demand Variability and Use of Monte Carlo for Network Capacity Planning – Performance and Capacity 39th International Conference by the Computer Measurement Group (CMG'14), Atlanta, GA

[GIGB2015] Gilgur, A.; Gunn, C.S.; Browning, D.; et.al (2015) Percentile-Based Approach to Forecasting Workload Growth – Performance and Capacity 40th International Conference by the Computer Measurement Group (CMG'15). Austin, TX November 2015

[NAMB2013] Nambiar, M. Network Performance Engineering (2013). Performance and Capacity 38th International Conference of the Computer Measurement Group (CMG'13) La Jolla, CA.

[BRGU2016] Brady, J., Gunther, N. (2016). How to Emulate Web Traffic Using Standard Load Testing Tools. 42nd International Conference of the Computer Measurement Group (CMG imPACT'2016). La Jolla, CA.

[DOGI2016] Doni, S., Giglioli, F. (2016). Productivity: a New Metric for Performance Analysis of Multi-Core and Multi-Threaded Processors. 42nd International Conference of the Computer Measurement Group (CMG imPACT'2016). La Jolla, CA.