SCTE·ISBE CABLE-TEC EXPO 2018 ATLANTA, GA OCTOBER 22-25

2018 Fall Technical Forum SCTE·ISBE • NCTA • CABLELABS

Using Historical Traffic Data to Schedule Service Interruptions for Minimum Customer Impact

A Technical Paper prepared for SCTE•ISBE by

Jason Rupe
Principal Architect
CableLabs
858 Coal Creek Circle, Louisville, CO 80027
720-313-2434
j.rupe@cablelabs.com

Colin Justis
Associate Engineer
CableLabs
858 Coal Creek Circle, Louisville, CO 80027
303-661-3470
c.justis@cablelabs.com

Table of Contents

List of Figures

Introduction

Maintenance is necessary, but service disruption isn't. Some cable system repairs will impact service in ways that customers notice, but can be necessary and urgent. Fixing service while a customer is not using services is far better. But to do that without bothering the customer requires a usage forecast model.

With historical usage data of service classes, we created a simple model that predicts the amount of data being consumed by end devices in a cable plant. Using this model as an indicator of usage by customers, we can determine the best time for a repair interval of a defined duration, allowing a technician to time necessary but disruptive operations to minimize the disruption to customers. We analyzed the data, tested the model, then built a simple application based on the model which can specify the best time(s) to conduct a service disruptive repair for a defined duration, for a given set of end devices to be impacted.

The application is about to be tested in a trial. Customer call-in rate will be used to measure the effectiveness of the projected schedule, as compared to a baseline.

Background

1. Problem Statement

While an outage must be addressed immediately, impaired service is not always immediately addressed for various reasons, and an impaired network providing sufficient service is a Proactive Network Maintenance (PNM) opportunity. Addressing impaired service or PNM work, where service providers will schedule the maintenance work, is the concern of this paper.

Some maintenance required on HFC networks will impact service. But it is not reasonable to coordinate and schedule all maintenance activity with all affected customers directly. Further, customers don't want to be bothered in that way, and would prefer they not be impacted by maintenance at all. Therefore, minimizing the impact of maintenance on customers is a valuable undertaking. But it is also a difficult one; you can't just ask them if they are using the service, then go do the maintenance if they are not. If it is a large number of customers, you can't coordinate the maintenance reasonably either. So, a service usage measurement or forecast method is necessary.

But a forecast is actually better than a real time measurement method. If a truck needs to roll for a measurement to take place, then there is already a cost involved. If you can measure traffic in a center without rolling a truck, then you don't know if the usage will change by the time you decide to send the technician. So, to solve this problem, a forecast is necessary.

An open question is whether it is necessary to predict whether a customer is actively using a service or not, or whether it is better to estimate a level of usage or utilization of services, at a given future time. Certainly, we expected that availability of information would influence our interpretation of the problem, as well as how to address it. Our predictive model, implied by our framing of the question, would need to provide an accurate, actionable result. Therefore, a measure of effectiveness for our solution must address the heart of the issue, which is how do we best avoid reducing our customers' ability to use services over the HFC plant when they want to use them.

If we can use Internet Protocol Detail Record (IPDR) data to identify usage patterns in the data, by edge device in the network, we have an indication of how much disruption would be experienced at a given time if service was disrupted. This in turn could be used to schedule maintenance to avoid impacting

customers' use of the service. But for this to work, customer usage has to be reasonably predictable; we need a useful model and implementable method.

2. Related Work and Models

Forecasting models and methods are a long-standing area of applied mathematics (operations research) work. The classic book "Operations Research in Production Planning, Scheduling, and Inventory Control," by Johnson and Montgomery contains a chapter on basic forecasting methods. The well-known methods explained in this seminal textbook include regression methods, moving average methods, exponential smoothing methods, adaptive control, Bayesian methods, and Box-Jenkins models. While there are many more methods to consider, the above methods are a sufficient set to start with for our consideration.

Craig Marlow and Nick Pinckernell did some initial work to identify the opportunity. Their approach was to build a Bayesian model of whether a customer was using the service at a given time or not. While a yes or no result on usage at a given time is useful toward scheduling maintenance, we thought it more useful to focus on how much a service is being utilized, setting up for future possible enhancements where we prioritize important service classes.

This approach also allows a balance between timely maintenance and service disruption. Waiting to repair might risk a service impacting event. Further, knowing that a customer is using a service would discourage the maintenance event from being scheduled, even when that usage is very minor, perhaps at a minimal level over a long period of time, and perhaps over what is still the best option for maintenance. There could be some customers who never stop using the service in at least some small way, which would prevent any maintenance at all.

Much work has been done for decades with forecasting, and numerous forecasting methods are worthy of consideration for this particular problem. Approaches such as moving average are useful, especially with consideration to time of day, day of week, seasonal, and overall trend effects. Because we expect customer usage to be affected by the day of week and time of day, perhaps even day of month or day of year, we considered methods that would allow for such correlation effects in the forecast model.

3. Data Review and Analysis

We use Internet Protocol Detail Record (IPDR) data to indicate the amount of traffic on the network, by MAC address and by defined service type. We used filtered data which contained 15 minute traffic data by service type and by MAC address. For this early analysis and proof of concept, we decided to aggregate the service types into one estimate of traffic over the 15 minutes, for each MAC address in the data. From these aggregated data, we began examining the aggregated traffic trends among single MAC addresses, and various random groups.

It may be useful in future versions to exclude some types of traffic, or to weight the traffic types according to criticality. We leave those options to future work.

We began by looking at averages of the data by time of day, and day of week, for both single MAC addresses, and groupings such as interface groups. This first step is important for validating and invalidating our assumptions about the data, its quality, and the general behavior of the traffic statistics.

Figure 1 below shows some time of day data for a single MAC address, with bold lines indicating the average (red), 30 minute moving average (orange), one hour moving average (green), and two hour

moving average (blue). Clear time of day trends are observable, but with a high degree of variability during some times of the day more than others. These results told us that time of day would matter clearly.

Figure 2 below shows week-long trends for the same MAC address, for every 15 minute interval in the week. This figure shows similar patterns each day, but clear differences going into and out of the weekends. As in the previous figure, bold lines indicate the average (red), 30 minute moving average (orange), one hour moving average (green), and two hour moving average (blue). Once again, we see the high variability during some times of the day, but clearly there are times of the day that are generally lower than others in usage, and the day of the week matters somewhat too.

After examining several individual MAC addresses over several weeks, we found that specific MAC addresses had very different patterns from others, so clearly not all devices are being used in the same way, under the same usage patterns. We found many examples that demonstrate why it is not sufficient to just predict general usage patterns; the specific MAC address matters. See Figure 3 for some different MAC addresses showing different usage patterns.

Further, predicting individual MAC address traffic would be important for one or small groups of MAC addresses, but larger groups would likely exhibit a general trend. By looking at groupings of MAC addresses, individual differences became less important, and groups of addresses tended to look the same. In other words, general usage patterns were good predictors for when to do maintenance impacting large groups of customers; specific forecasts would be less important. See for example the interface group of MAC addresses shown in Figure 4. At this large of an aggregation, the usage tended to follow very closely this pattern no matter the MAC addresses in the grouping.

While each MAC address did have important differences, there is still much we can say generally which is of use. Ideal service times differ significantly between MACs, but tend to coincide with early morning hours (before 7:00am), generally. When constrained to typical work hours (say, 8:00am to 5:00pm), service times must be analyzed per MAC in order to minimize disruption, as generally the differences between MAC addresses becomes important during those times of the day.

By averaging over daily and weekly usage, an expectancy can be obtained for the time of day and time of week. The most direct way to project on internet usage is to perform a moving average over a rolling period, such as 4 weeks.

Note that we intend to use this information to predict whether services are in use over a particular end device. Generally, we assume a customer location to be aligned to one MAC address, thought that is not a critical assumption to the project.

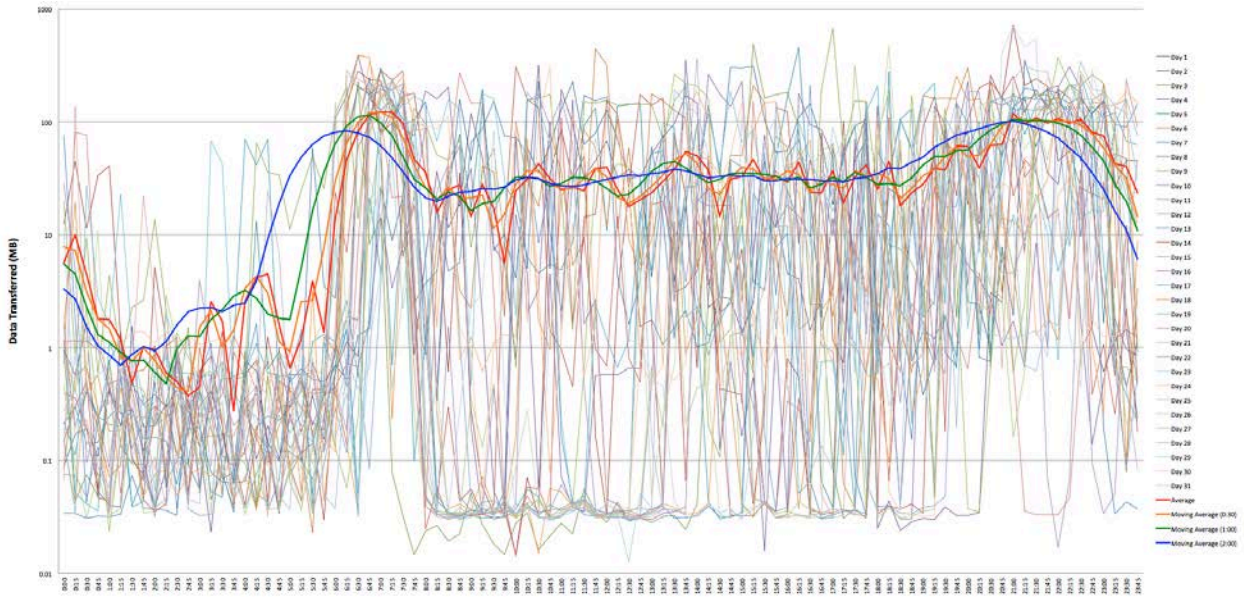


Figure 1 – Usage for a single MAC address, over the time of the day

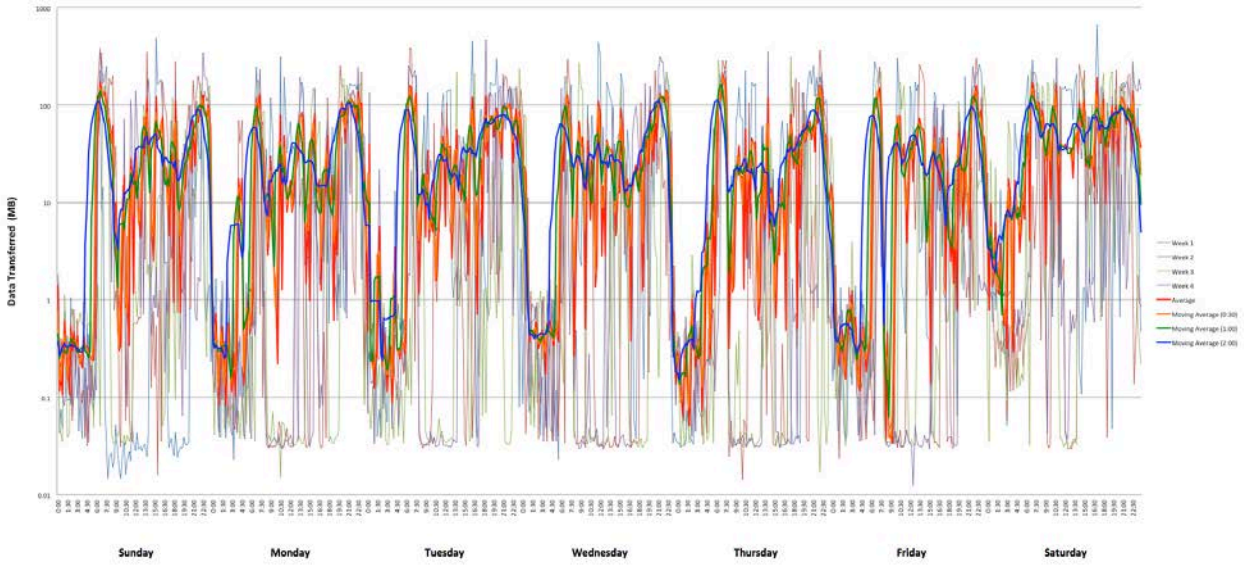
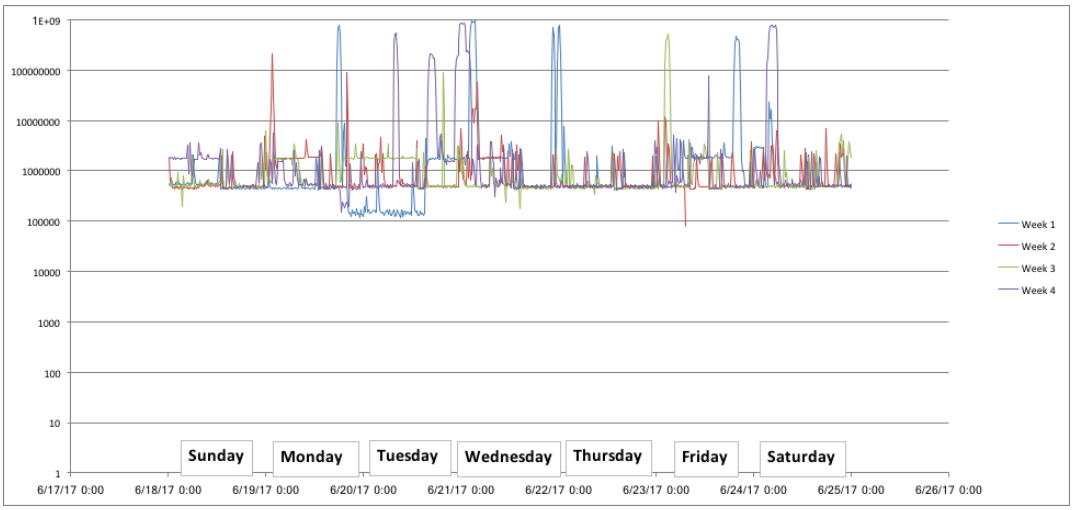
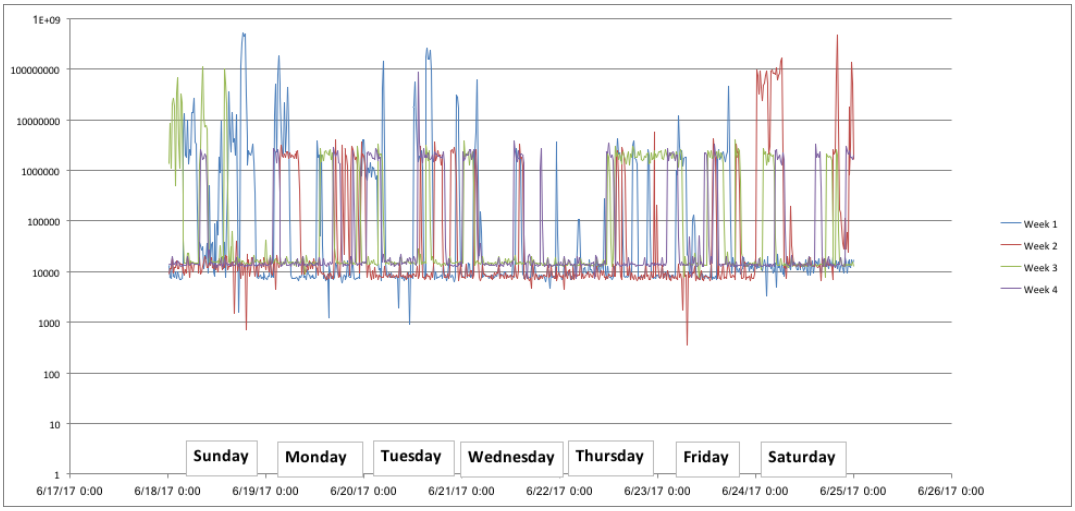
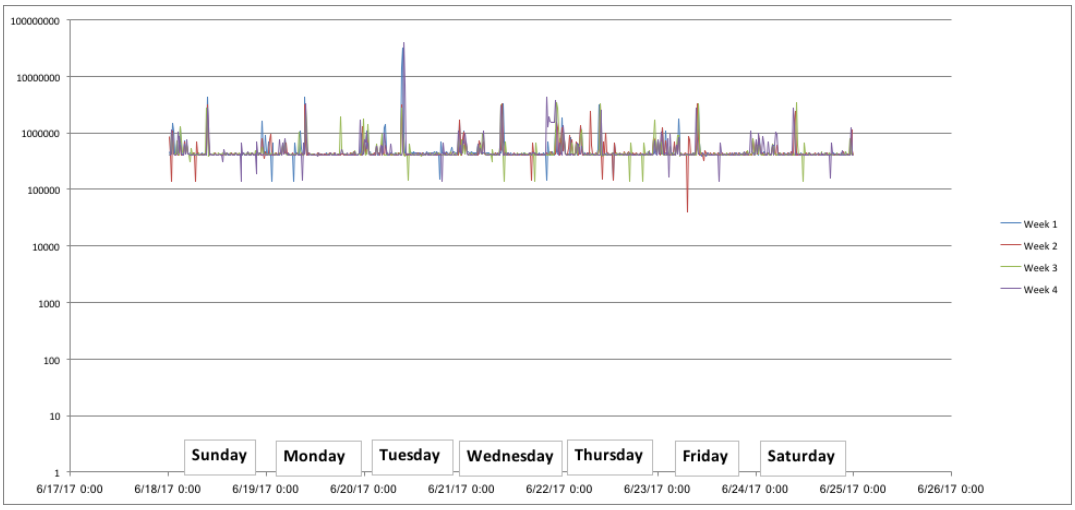


Figure 2 – Usage for a single MAC address, over the time of the week







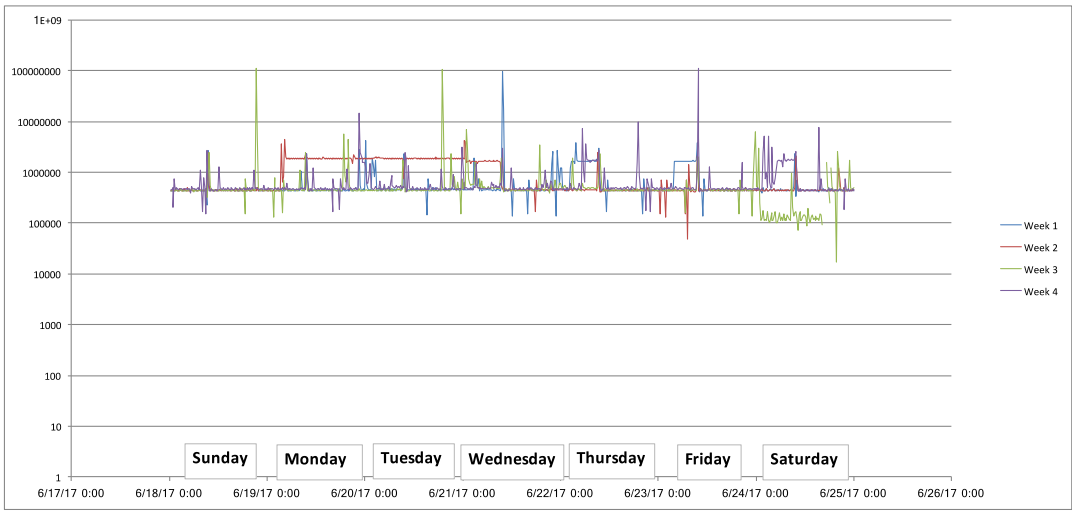


Figure 3 – Four different MAC addresses showing very different usage patterns

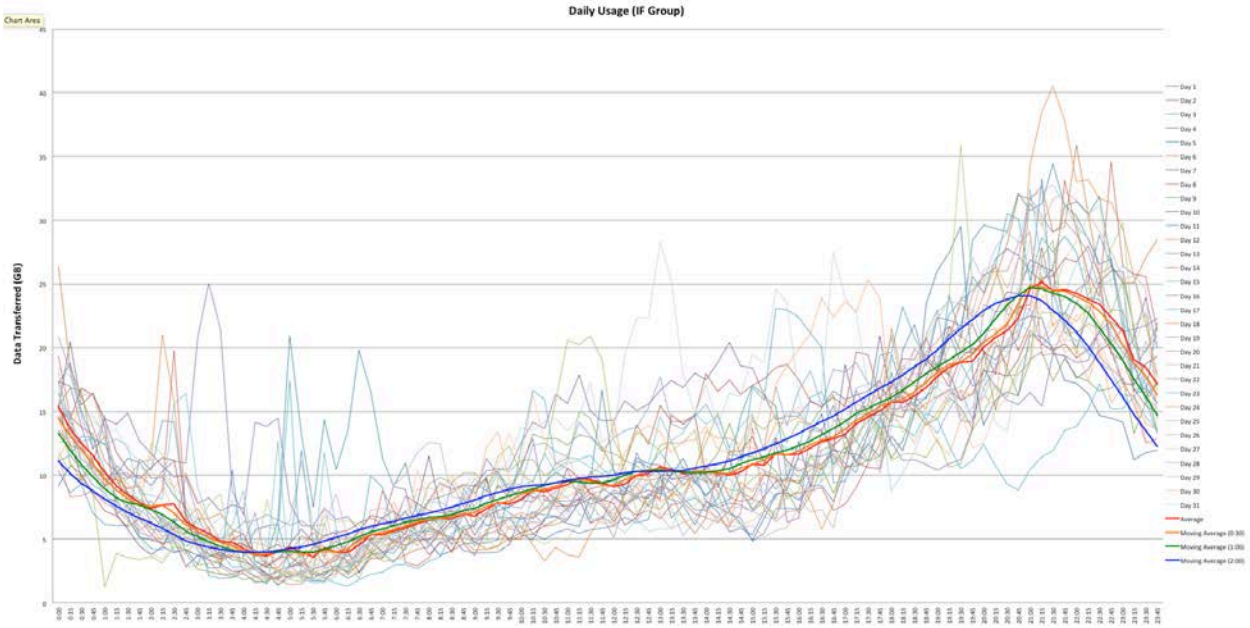


Figure 4 – Daily usage for a single interface group of MAC addresses

4. Formulated Approach to the Problem

Based on our analysis of the data, we learned several ideas which framed our model.

- Usage data exhibit clear patterns which might be exploitable to minimize disruption of service during a repair.
- Clear, significant differences in usage patterns by MAC address leads us to predict individual MAC usage patterns independently.
- Grouping MAC usage models in small groups was important to predict the best times to impact service for the group. But as groups got larger, such as an interface group, a general model was likely sufficient for many times of the day.

- Visually, it was clear to see there were time of day, and day of week effects that were important for almost all MAC addresses studied. We further suspected there was an overall increasing trend of usage too. While we did not have enough data to find an effect for time of the year, we have strong suspicions that there are effects due to holidays, summer vacations, etc. Thus, we recommend:
    - Obtaining a year's worth of data to find annual patterns to add as effects to any chosen model, and
    - Understanding in some way (predicting) the risk of a forecast, especially over days where there are no data to contribute an annual effect to the model.

From these observations, we decided a simple model predicting the amount of usage on each MAC address would be a useful first model. Further, as we could see clear effects, we sought a linear model of these effects as a simple, sufficient way to predict usage for short periods of time into the future, say a day to a few days. While not ideal, it was sufficient for our proof of concept and trial.

Models Description

5. General Model Requirements

We recognized that interrupting lower amounts of usage might still be better than interrupting higher amounts, even if there is usage, so we focused on predicting the amount of usage over whether services were being used or not. Further, we considered classifying usage into discrete levels of usage as a compromise between the on-off and full fidelity of usage level, but later decided to stay with a direct approach as a starting point, not having enough knowledge to set finite usage levels.

6. Competing Methods

We considered a Bayesian approach which would predict whether service was being used or not. This did not meet our criteria for predicting usage level. We considered a Bayesian approach to determine usage level, but again decided to stay simple for our first model, if we found a simple approach that appeared reasonable, which we did.

We also considered several linear prediction models, some using various forms of moving average, as forecast models. Without a large amount of data to do a serious comparison of model methods, we did not have a reason to go with a complicated model for the proof of concept.

7. Chosen Solution

When it comes to predicting internet usage for a single customer, or group of customers, there tends to be consistency in the usage for time of day and time of week. However, the usage can still vary slightly from day to day, or week to week. For short-term projections, it is ample to predict future internet usage using linear projection from a line of best fit for the projected data.

Essentially, the model we used for predicting future usage is as follows. Given a few weeks of data is all we had to work with, we recognized there was no way to model for special days of the year. Therefore, we used the available data to get a weekly effect, day of week effect, and time of day effect. Then, we combined these linearly to form projections into a short future of a week.

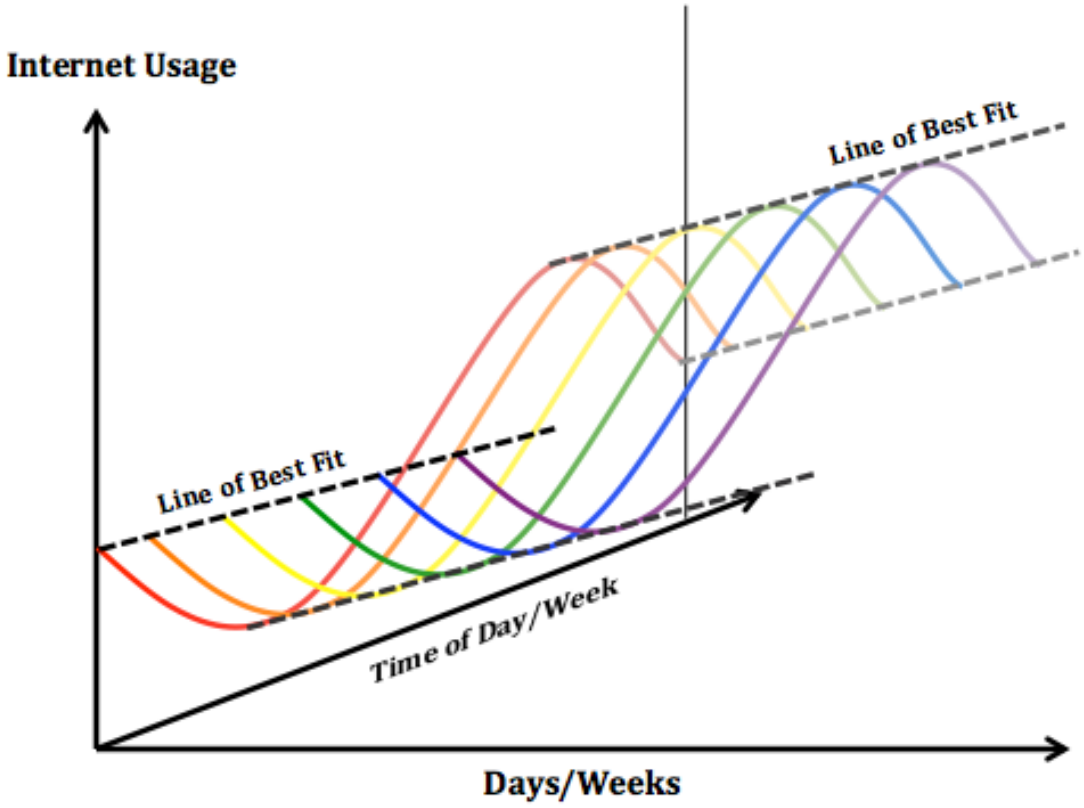


Figure 5 – Rainbow plot of the conceptual model

An example for day-to-day, or week-to-week, projections is shown in Figure 5. This figure is a 3-D rainbow plot where the time of day (or week) is treated as independent of the number of days or weeks passed; color indicates a particular day or week, and paleness represents the depth or time of day or week. The time of day or week can be measured in hour, half-hour, or 15 minute intervals. Internet usage is measured as bytes passed (usually megabits (MB) or gigabits (GB)) for each of those time intervals. Finally, the lines of best fit are calculated independently for each time of day or week; the slope and y-intercept can be different for different time intervals of the day or week.

The simplest way to calculate a line of best fit is through method of least squares. In that case, the line of best fit becomes

$$y^* = corr(x,y)\frac{\sigma_y}{\sigma_x}(x - \langle x \rangle) + \langle y \rangle, \quad (1)$$

where x is the time of day or week, y is the internet usage rate, and y* is the usage for the line of best fit. By convention, standard deviation is denoted by σ, and corr(x,y) is the correlation between x and y. The Root Mean Squared (RMS) error becomes

$$Error_{RMS} = \sqrt{1 - corr(x,y)^2}\sigma_y \quad (2)$$

Because the standard deviation is the RMS error for an average mean, (2) shows that the RMS error for a least squares line of best fit projection is always no greater than a simple flat projection of a moving

average. In other words, there is nothing to lose by performing a linear projection, regardless of whether internet usage varies significantly day-to-day or week-to-week.

Once we have a line of best fit, it can be projected into future days or weeks for some short time period. A hypothetical example for a given time of day or week is shown below in Figure 6.

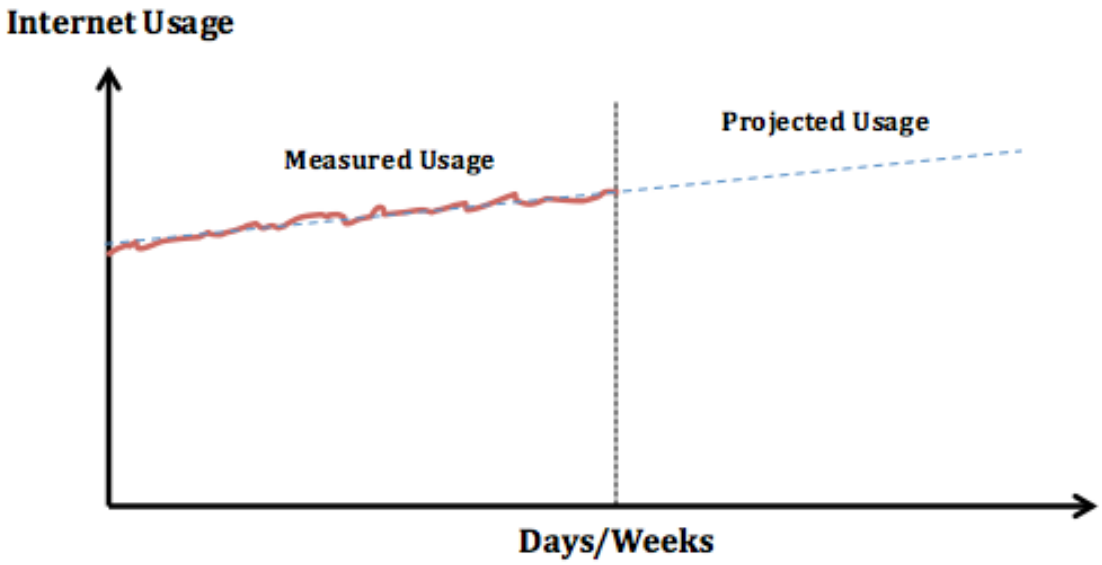


Figure 6 – General concept of forecasting a linear trend, as it applies to the usage model's general exhibited overall trend of usage growth over time

The slope can still vary over long time periods, so it is important not to project too far. At most, the projected usage should not be as long as the measured usage. The measured internet usage would be on a moving time window. For instance, a queue of one month's worth of data could be collected, and then the next day or week would be projected. After the next day or week, a new day's or a week's worth of data would be added to the model. Potentially, as older data becomes less helpful to the prediction, it would be removed from the model. A weighting of older data, such as in an exponentially weighted moving average model, would be best, as in a Box-Jenkings approach.

From looking at simple averages, the best service times can be found, along with a measure of the reliability of a chosen service interval for every grouping of customers possible.

8. Model Validation

We tested the model using different data sets at different CMTSs. Further, we took a trained model and used it to predict results for times which we had data. As the chosen model showed promise for being a sufficiently accurate model as we could measure that, we decided to continue to conduct a trial where we could measure the real impact of the approach against actual customer calls of interrupted service.

Implementation Approach

9. System Description

To support the trial of the model, the model was encoded and implemented with a system and process.

The model described above was implemented in a two-stage process for the sake of the field verification trial; an implementation would be very similar. The two stages involve a processing stage, and then an application stage.

The processing stage receives updated IPDR data periodically, and applies the modeling described above to form a new set of predictions for the forecasted horizon. The output from this stage forms a table of expected usage by MAC address for each of the 15 minute intervals, which is the resolution of the data, over the horizon being predicted. This output table is the input to the next stage.

The application stage is a web-based or locally-cached application which performs table lookups. The application front end created by CableLabs, shown in Figure 7 below, is the user interface which collects the MAC addresses to potentially be disrupted, the window within which the repair needs to be scheduled, and the duration of the service interruption expected. The application performs a table lookup of the MAC addresses, adds the predicted usage for each 15 minute interval over the duration of the scheduling window, then does a moving average of the disruption window size over the duration window, reporting on the lowest usage candidates. The application reports the top few options for minimal impact based on the overall usage statistics calculated.

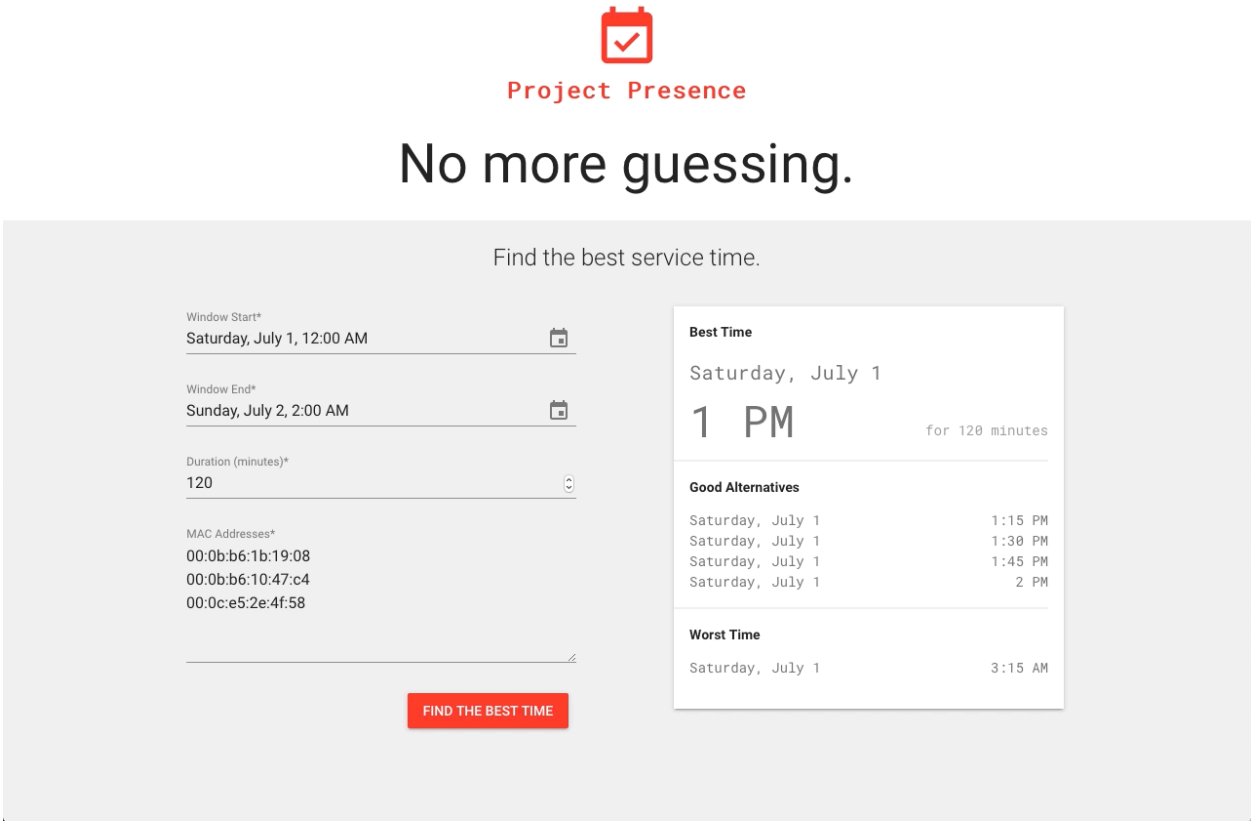


Figure 7 – Application Front End

10. Process Approach

Each evening, new data are gathered for the node, and fed into the model. The model is updated with the new data, creating a new forecast of 15 minute usage predictions for each MAC address on the node. The resulting file is uploaded to the application server. The application is then updated to connect to the new file, and tested. Once confirmed as functional, the application is able to process off of the new model results table. Because the application exists separate from the data file that is updated periodically, there is no duration over which the application is down. Instead, the application simply updates to the new information when made available. The application then processes updated intervals each time it is used.

At the start of the work assignment day, service personnel bring up the application through a browser. They then use the application to assign the work times for the work done in the area covered by the application.

Field Trial Verification

11. Field Trial Plan and Design

For the field trial, we selected two CMTSs in the Logan, Utah area. We pulled data from this location before and after summer break for the local schools, as a comparison to determine whether the model

results change across this known usage change. These data were used to form the reference model for the field trial.

Each evening, we extracted new IPDR data in 15 minute increments for the past 24 hours, and incorporated the new data into the old. This updated data set was used to train a new model for the next work day. The back-end model ingests the updated data, and the output is a table of 15 minute predicted usage for the future week for each MAC address.

As of this writing we have yet to begin the trial, so the rest of the planning and design have yet to be tested. We anticipate the following general activities to follow. All trial participants will need to be briefed as to the changes in the operations steps when assigning work and conducting maintenance. Those assigning the work will need to add the step of using the front-end tool to determine the best times to conduct the maintenance, based on their best information about what needs to be done, how long it will take, and who will be impacted. If changes in the field are necessary, a line of communication needs to be established so that technicians and those assigning the work can agree as to any adjustments. This action will help maintain the integrity of the trial. Further, field technicians who conduct the maintenance will need to record the times when service was disrupted and restored to be sure it did or did not overlap the times indicated by the model. A good model that can't be followed is not very useful, so we must track its usability as well as its accuracy.

12. Performance Measurements

The key measure of performance for this trial is the number of customer call complaints per impacted customer per unit of service interruption time. We track this by collecting from records the number of customers who call in to indicate a service interruption during the maintenance time for the experiment group, and comparing this to the control group. To normalize for each maintenance event, we take the number of customers who call in to complain out of the group of interrupted customers, and divide that by the number of interrupted customers times the interruption duration. Each maintenance event has one measure of performance result; if the maintenance action required more than one outage, we simply add the performance measure for each outage for that maintenance event. We collect this measure of performance for the experimental group, and for a control group as well. Then we calculate basic statistics from the measure of performance including mean, standard deviation, and confidence bounds. Finally, we calculate statistical confidence bounds and conduct statistical tests to determine whether we can say with confidence whether the results indicate effectiveness of the solution.

Given the measure of performance offered, we expect the measure of performance to be distributed Poisson, so simple statistical tests on Poisson parameters should apply.

Findings

13. Importance of the Model

While the trial will reveal some information with which to determine the importance of the model, a single trial will not reveal very much. And because we had access to very limited data, we do not consider our model to be necessarily the best, but sufficient.

Instead of simply adopting the model given in this paper, we suggest, in a full implementation, that multiple models be formed, tested, and used in competition, with a long-term adopted model to result from the experience of field use. The model reported here was the result of a few competing models considered and compared on multiple merits, including accuracy of prediction. Had we more data to work

with in the development of these models, perhaps a different model would result. More experience, and more data, are in order.

A good model can be created from the data only, but there are implementation differences that can make one model better than another, and one approach to implementation better than another as well. Consideration of specific applications is important (network, operations, environment, OSS, etc.).

14. Trial surprises and findings

The trial is planned to begin in September. We hope to have preliminary results to report in our presentation at the Expo.

Conclusion

By using available usage information from the network, a simple model can be created to predict the traffic through an end point on the network. Service class usage in 15 minute increments can be used to form simple predictive models of usage, projecting a few days into the future. This prediction of usage can be used to schedule maintenance so that the impact on traffic is minimized. The expectation is that by impacting the least amount of network traffic we reduce the impact on customers. This lower impact should be measurable through a lower customer call in rate, so that fewer customers will call to complain about service outages when the model is used to plan the maintenance activities. This idea is to be tested in a field trial soon.

The prototype built for the trial demonstrates that a process, using a simple model based on IPDR data, and utilizing a web-based front-end interface, can provide work assignment windows for planned maintenance which would interrupt service.

The analysis we conducted showed definite patterns in usage at network end points (MAC addresses), and in groupings of end points. While there were considerable repeating patterns in the data across MAC addresses, not all MAC address usage patterns were the same. But in most cases, there were large differences in the peak usage compared to the minimum usage, and long periods of time over which usage remained mostly low. The analysis suggested a model to predict usage by MAC address was achievable and could be useful for scheduling maintenance.

By predicting usage on single MAC addresses, then clustering the models for the group of MAC addresses to be impacted, a merged model was created for each maintenance case. By searching the time over which the maintenance was desired to be scheduled, the best times to schedule a maintenance activity for any given duration could be found easily. By creating a front end to the model, and a process by which we could update the model as frequently as daily, we were able to build a prototypical solution which could be trialed to prove the concept further.

Acknowledgements

A special thanks to Larry Wolcott who drove this work and made sure it met success, and for all his input and help with the paper and trial. More thanks go to the software team at CableLabs who supported the front-end for the trial. Also, a big thanks goes to Jay Zhu and John Phillips at CableLabs who stepped in to support the work when a primary author was on leave.

Abbreviations

GB	gigabits
HFC	hybrid fiber coax
IPDR	Internet protocol detail records
MB	megabits
PNM	proactive network maintenance
RMS	root mean squared

Bibliography & References

Lynwood A. Johnson, Douglas C. Montgomery, *Operations Research in Production Planning, Scheduling, and Inventory Control*, John Wiley & Sons, New York, Copyright 1974, ISBN 0-471-44618-1.

Douglas C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, New York, Copyright 1985, 1991, ISBN 0-471-51988-X.

