The Effect of Packet Traffic Prediction Limitations on DAMA Schemes for Future Satellite Packet Networks

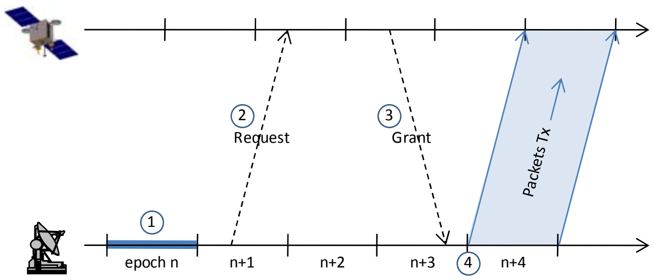
Max Williams
Kathleen Nichols
Pollere Inc.
Menlo Park, CA

ABSTRACT

Future satellite-based packet networks (FSPN) will utilize a request/grant MAC protocol referred to as Demand Assigned Multiple Access (DAMA) for uplink allocations. DAMA is conceptually similar to Bandwidth-on-Demand (BoD) protocols used in terrestrial networks, such as 802.16 and DOCSIS, which flexibly share total bandwidth according to actual loads and usage policies. However, unique characteristics of FSPN result in challenges not faced by these protocols. DAMA request/grant cycles are on the order of seconds instead of tens of milliseconds. This makes the process one of prediction of future needs based on past traffic loading measurements rather than grants for specific packets. Over a decade of packet traffic study has shown the difficulty of fine-grained load prediction. If these requests are heavily relied upon for grant allocations, load prediction errors can result in unfair and non-deterministic terminal allocations that may violate terminal SLAs. This study assesses the predictability of packet traffic at FSPN request/grant timescales and compares the performance of DAMA approaches of varying request granularity, starting from coarse measures of inactivity and activity and focusing on meeting SLA commitments. The complexity of fine-grained approaches with respect to any potential gain is also considered. Rejecting fine-grained approaches to load prediction is shown to result in more deterministic DAMA performance and lower overhead while achieving efficient use of shared RF resources.

INTRODUCTION

The FSPN DAMA scheme is structured as a request-and-grant MAC layer, like 802.16 or DOCSIS. As shown in Figure 1, terminals send rate requests to the satellite DAMA controller based on demand and the satellite controller responds with radio frequency (RF) resource allocations that provide a specific average data rate for the uplink. An important difference between FSPN DAMA and other terrestrial BoD protocols is the request/grant process is much longer due to constraints imposed by GEO propagation, Media Access Control (MAC) layer delays, processing for protected waveforms and support for disadvantaged communications-on-the-move (COTM) terminals. With epochs on the order of 500 milliseconds, total response time of around 2 seconds is optimistic for FSPN. This lag between the time traffic measurements are taken at the ground terminal and when the corresponding allocation goes into effect renders the simple, efficient BoD model where a packet of n bytes arrives and is buffered while a request for n bytes of resources is sent unworkable. At these lag times, buffering delays would be unacceptable, thus terminal requests become predictions of terminal needs in the future.



1 – Arrivals measured

2 – Request derived from past loading measurements sent to satellite

3 – Grant sent back to terminal

4 – Grant implemented in epoch n+4

Figure 1 - FSPN DAMA Request / Grant Timeline

Ideally, precise and accurate predictions of what a terminal will need each epoch would allow DAMA to make allocations that meet SLA commitments while maximizing overall resource efficiency. However, precise prediction of packet traffic arrivals over these timescales is not possible in most cases. Many previous studies have shown that the bursty nature of Internet traffic results in high or near infinite variability of arrivals. This makes it improbable that any FSPN DAMA can make allocations that are both accurate and precise in matching terminal demand. Prediction errors can lead to either over-allocations, reducing resource efficiency or more importantly under-allocations which can result in performance degradation and SLA violations. They also interfere with providing predictable and stable service to the terminal user.

Despite this barrier to precise statistical prediction, there are two simple traffic loading states that tend to persist at the FSPN DAMA timescales, inactivity and activity. That is, active (terminals with traffic loading over some minimal threshold) or conversely inactive conditions are likely to persist over the 2 second request/grant cycle and thus are relatively good, though coarse, predictors. Prediction errors are confined to the transition states which vary by terminal aggregation. DAMA resource efficiency is realized using this metric of active vs. inactive by making resources not used by inactive terminals available for other terminals that are active. As this study shows, a DAMA approach that uses this simple metric provides for accurate and stable allocations that support terminal SLAs with low relative complexity. To improve overall resource efficiency further, the possible use of additional granularity in the load prediction used by DAMA is systematically examined. Gains in resource efficiency however, must be weighed against any negative effects on SLA compliance caused by increased prediction errors and increased complexity of these approaches.

This study first analyzes predictability of packet traffic at the target timescales, focusing on less aggregated traffic expected to be typical of many smaller FSPN terminals. A correlation analysis is presented that provides a quantitative assessment of the ability to predict traffic loading for several types of common applications. Next, a range of possible DAMA request strategies are presented from those that minimize reliance on traffic prediction to those that attempt fine grained prediction. Finally, simulation-based data is presented assessing the performance of several types of DAMA request strategies with varying traffic prediction granularities. Each scheme is assessed based on a variety of evaluation criteria defined in this study.

TRAFFIC PREDICTABILITY

The ability to predict loading several 500 ms epochs into the future would offer obvious advantages to any DAMA strategy by allowing terminals to precisely request what will be needed to satisfy demand. Internet traffic has been shown in many studies, including [1], to exhibit long-range dependence and self similarity, meaning it has memory or dependence over long timescales and is bursty with infinite variability. Long-range dependence does not imply predictability over short timeframes as is discussed in [3]. Traffic unpredictability is indicated by the very large or infinite variance found in samples.

Previous studies, e.g. [2], have shown that in order for linear prediction models (commonly proposed for DAMA) to work effectively, there should be relatively strong correlation between values of the time series at the lag time of interest (in our case about 2 seconds). Correlation analysis also gives insight into how well more complex and computationally intensive non-linear prediction models, such as neural networks, will perform. To quantify the correlation between traffic arrivals 4 epochs apart, the number of bytes arriving in each epoch was compared to the number arriving 4 epochs later for different types of packet traffic. Studying time series correlation can provide quantifiable evidence of the ability to do accurate fine-grained predictions, independent of the specific linear prediction model chosen.

The predictability of packet traffic was studied for terrestrial network traces of aggregated WAN links as well as Ethernet LAN segments in [2, 4]. In [2], packet captures from production networks are binned over certain non-overlapping time intervals and an autocorrelation function is computed to determine the correlation coefficient as a function of the lag time between sample bins. This indicates the degree to which a relationship exists between traffic levels at certain time separations. [2] found mixed results with some traces from aggregated links showing relatively strong correlation and others including the LAN traces showing poor correlation. Congested links will, of course, show strong correlation.

Following the approach in [2], the Pearson product-moment correlation coefficient was derived between 2 data sets for a specific terminal and application type as:

$$r_{X,Y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \qquad (1)$$

where X is the data set representing the number of bytes that arrived at a queue in each epoch e and Y is the data set representing the number of bytes that arrived at the queue in epoch e+4. So X and Y represent the arrivals each epoch for a particular terminal and particular type of traffic, where Y is simply the X data set time shifted by 4 epochs. n is the number of epochs (data pairs) included in the derivation.

The correlation coefficient is a measure of the linear relationship between the two variables, ranging from -1 to +1. A 1 indicates perfect positive correlation while a -1 indicates perfect negative correlation. A 0 indicates no linear correlation of the 2 data sets. Interpretation of the correlation coefficient value is somewhat dependent on the process being evaluated but in general values below 0.5 represent very low correlation, 0.51 – 0.79 represents low correlation, 0.8 – 0.89 moderate correlation and > 0.89 high correlation. As seen in the coefficient described next, even a 0.9 or higher coefficient of correlation can allow a fair amount of independence between the data sets.

A complementary coefficient was calculated in addition to the correlation coefficient called the Coefficient of Determination. This is the square of the Pearson product-moment correlation coefficient and represents the percent of common variance between the 2 data sets. This is used to gauge the percentage of variability of Y that can be explained by variance in the X variable. The remaining percentage is unexplained and is a measure of the prediction error when predicting Y from X.

Only sample pairs where either x or y were non-zero were evaluated. Sample pairs in which both were zero were excluded in order to concentrate on traffic arrival prediction during active periods since terminal inactivity tends to persist. Full stack application models were built in Opnet and configured based on behavioral parameters derived from trace studies of application traffic in real operational networks including large and small FTPs, web traffic, texting and VBR streaming video. For each application type, the model was run multiple times, varying random seeds, until between 1500 and 5000 epochs worth of data was collected. Results are shown in Table 1.

Table 1 - Correlation Analysis of Pkt Traffic Arrivals

Traffic Type	Correlation Coefficient *r*	Coefficient of Determination r^2*x100*
Large FTPs (1MB – 12MB)	0.872830	76%
Small FTPs (10KB – 500KB)	0.477526	22.8%
Web Browsing (HTTP)	0.101825	1%
Texting	-0.335131	11.2%
Streaming Video - H.263 codec Low Motion Scenes (talking head)	-0.061915	0.4%
Streaming Video - H.263 codec Dynamic Scenes (action movie)	0.011479	0.01%

The results show very low correlation for all but the Large FTP application as expected. The Large FTP correlation coefficient was close to the highly correlated range, indicating the arrivals four epochs apart exhibit a near linear relationship. The coefficient of determination indicates that over 75% of the variation of epoch n+4 is accounted for by the variation in epoch n. The correlation seen with this application configuration is due in large part to the fact that it allows the TCP session to progress through slow start and fill the pipe. Once there it adjusts its sending rate to the available bandwidth and keeps a fairly steady arrival rate over the remainder of the file transfer.

This contrasts with smaller FTPs as well as the web browsing and texting applications in which the underlying TCP connection never leaves slow start and doesn’t fill the pipe before the message transfer completes. These short transactions are much burstier and thus more difficult to predict 4 epochs out as evidenced by their very low correlation coefficients. Less than a quarter (with web browsing substantially less) of the variation in the arrivals in epoch n+4 can be accounted for by the variation in arrivals in epoch n for traffic generated by these applications. That leaves a substantial amount of uncertainty in the possible arrivals that can occur 4 epochs into the future which will result in higher prediction errors.

Streaming video traffic showed the least amount of correlation in arrivals 4 epochs apart. The codec analyzed, H.263, is a variable bit rate encoder that generates variable size frames with variable inter-frame times based on the content being encoded. This results in a very dynamic arrival rate with the difference in arrivals 4 epochs apart quite variable with large magnitudes.

Another way to see the relationship between arrivals is to use a scatterplot with arrivals in epoch n (the independent variable) along the x-axis and arrivals in epoch n+4 (the dependent variable) along the y-axis. This shows the degree of correlation (or lack thereof) that exists between 2 data sets. Linear correlation would show in the graph as data points scattered closely around (or directly on) a line, while non-linear correlation might show up as points clustered around a curve representing the non-linear relationship between the data sets. The closer the data points are clustered around a line (or curve for non-linear dependence) the stronger the linear correlation. The exception is a vertical or horizontal line which shows independence since one of the variables would be constant regardless of the value of the other.

A scatter plot of streaming video (Figure 2) clearly shows no linear relationship between these data sets and thus the extremely poor correlation coefficient seen in Table 1 above.

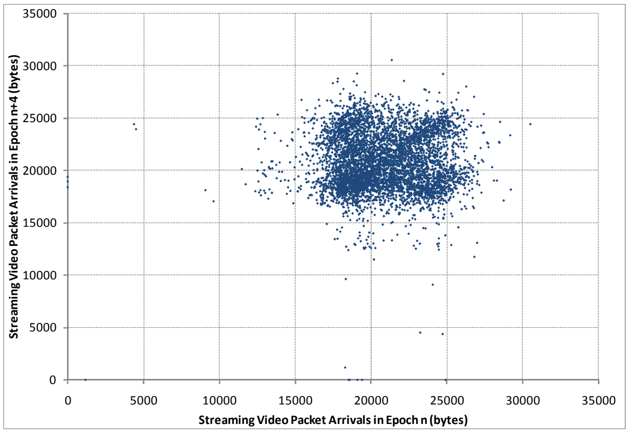


Figure 2 - Streaming Video Packet Arrivals Scatterplot

This analysis was conducted with terminal workloads showing very little aggregation, representative of those with a relatively small number of simultaneous users behind them. Terminals with more traffic aggregation may have different characteristic uplink loads. Such aggregation is likely for larger terminals servicing a large number of users, possibly from multiple tactical networks. Network characteristics outside the FSPN domain could also lead to different results, e.g. a congested bottleneck along the path in a user network could cause resulting traffic arrivals into the FSPN appear smooth. As a result, the correlation might be higher. However, these cases likely represent transients that would not be generally useful for more fine grained predictions.

The correlation analysis detailed in this section showed that except for large and long-lived FTPs, traffic arrivals show very low correlation at 4 epoch lags. In most cases, less than 25% of the variability of arrivals in epoch n+4 can be explained by variability in epoch n. In many cases it is substantially less. Much of arrival dynamics 4 epochs into the future are unlikely to be precisely predictable with linear prediction models and at best difficult for the more computationally intensive non-linear models

DAMA REQUEST STRATEGIES

DAMA request strategies can be categorized based on the request quanta utilized (Figure 3). There are advantages and disadvantages associated with request granularity. A fine-grained request quantum implies use of more precise traffic prediction, while a coarse-grained request quantum implies use of lower precision prediction. As seen in the last section, precision does not imply prediction accuracy. Coarse-grained requests, i.e. binary activity/inactivity indications, are more likely to be accurate which leads to more deterministic performance. Coarse-grained requests also need fewer request updates but may result in more unused resources and thus reduce effective capacity. More granular requests, when accurate, can allow more harvesting of unused resources but may suffer frequent prediction errors and will require a higher request frequency.

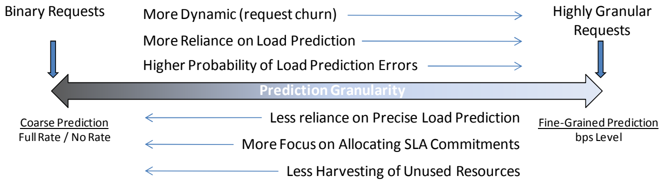

Binary Requests
More Dynamic (request churn)
More Reliance on Load Prediction
Higher Probability of Load Prediction Errors
Highly Granular Requests
Prediction Granularity
Coarse Prediction
Full Rate / No Rate
Less reliance on Precise Load Prediction
More Focus on Allocating SLA Commitments
Less Harvesting of Unused Resources
Fine-Grained Prediction
bps Level

Figure 3 - DAMA Request Approaches

The DAMA request strategies evaluated in this study range from utilizing a very coarse metric of activity/inactivity (whether a terminal will have any packet arrivals or not) to the most granular (how many bytes will arrive). The former approach is termed Activity Based (AB) and the request (prediction) is simply based on the presence or absence of arrivals during the last epoch. The latter is termed Load Based (LB) and the request is based on an exponential weighted moving average of the arrivals from past epochs. These represent opposite ends of the spectrum and serve to frame the problem space. Additional variants, that fall in between, were evaluated by incorporating varying levels of traffic prediction granularity. These used multiple discrete request levels or quanta that are selected based on past arrivals, where one level is equivalent to the AB approach described above. These variants are referred to as Multi-level Activity Based (MLAB). Figure 4 shows how request levels are selected based on past arrivals for a 2-level MLAB approach. When arrivals fall within an activity level range for a terminal in an epoch (1 or 2 in this case), the corresponding request level is used. Simple dampening mechanisms can be used to prevent rapid oscillations between request levels.

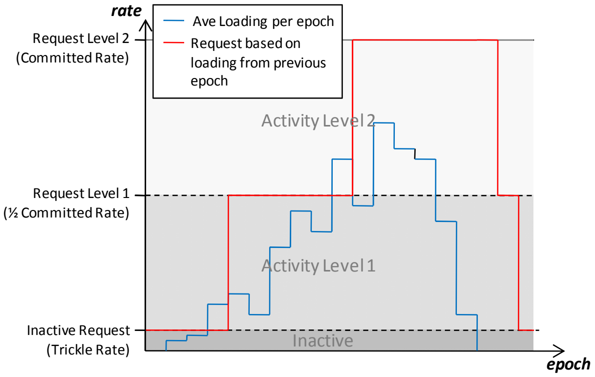


Figure 4 - MLAB Request Levels (2-level example)

EVALUATION CRITERIA

To evaluate the performance of DAMA request strategies, a mix of high level aggregate metrics that represent performance directly observable by FSPN users and low level metrics for the MAC layer DAMA behavior were used. Together, these give an evaluation of the ability of each approach to maintain commitments in terminal SLAs and maximize effective system capacity with least complexity.

QoS Metrics (Delay and Packet Loss)

Edge-to-edge QoS metrics are used as one assessment of ability to support terminal SLSs. Since delay and loss metrics will be directly observable by the FSPN user and will be part of a service commitment, this metric will give insight into how the allocations affect the QoS commitments achievable across FSPN.

Under-Allocations

One DAMA objective is to allocate rates so it appears to users that the SLA rate is always there. That is, the rate allocated should at least be equivalent to the arrival rate of the traffic up to the SLA commitment (i.e., number of bytes arriving in an epoch ≤ number of bytes allocated for transmission in the epoch). Allocations less than the arrival rate (i.e., more bytes arriving than leaving over an epoch) result in a queue sustained across epochs, with longer delays and/or packet loss that can affect the performance commitments. This condition is termed an under-allocation and it is measured in number of epochs of occurrence across all terminals in each experiment.

Misallocated Bytes

The Misallocated Bytes metric is a measure of how efficiently resources were allocated to terminals in order to maximize the effective capacity. It is a measure of the number of unused bytes allocated to terminals (allocated bytes > arriving bytes) in an epoch that theoretically could have instead been allocated to terminals that were under-allocated (allocated bytes < arriving bytes) in that same epoch. The misallocated bytes are tallied in each epoch of a simulation across all terminals and are summed over all epochs in each simulation.

Overhead/Complexity

This is a measure of the DAMA requests and grants that were needed for each request strategy used. The uplink request frequency provides a quantifiable measure of the DAMA overhead necessary and also gives some insight into the required complexity and processing in the terminals and satellite to generate and utilize these DAMA request updates to allocate resources each epoch. When hundreds of terminals share uplink DAMA resources, the complexity of receiving and processing requests, completing the uplink channelization process to derive individual allocations and sending the assignments back to all terminals is not trivial especially for size weight and power (SWAP) constrained payload processors.

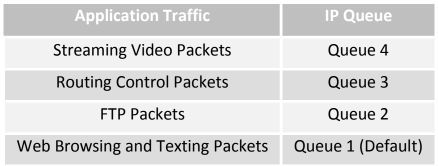
SIMULATION SETUP

The simulation scenario utilized contains 100 FSPN terminals under a single satellite sharing a common pool of uplink RF resources through DAMA. Each terminal has a router and a DAMA agent that sends a request to the satellite DAMA controller based on the packet arrivals at the terminal router queues each epoch. The DAMA controller receives all requests from terminals and issues grants to each based on available RF resources using a multi-frequency TDMA channelization process. Most terminals in the simulation have 1 Mbps uplink committed rates while a smaller subset has 4 Mbps uplink commitments. The resources available on the uplink are constrained, meaning that not all terminal commitments can be met in most epochs. Space – ground links include GEO propagation delays and various lower layer processing delays. All packet paths include both an uplink and downlink. The uplink is the focus of this study.

Application models are used to generate the packet traffic. The models include FTP, web browsing, routing control traffic, texting and video streaming. Inelastic real-time traffic such as interactive voice and video conferencing was not included in this study, since the strict QoS metrics and other operational policies typically required with these traffic types necessitate the use of resource allocation strategies different from the epoch-by-epoch DAMA approaches assessed in this study.

Since application models with stochastic behavior were used, a Monte Carlo simulation technique was employed to allow for convergence and thus higher confidence in the resulting data. Each test case was run up to 100 times using different random number seeds. Each simulation run is 3600 seconds in duration.

Table 2 - Traffic to Queue Mapping



Application Traffic	IP Queue
Streaming Video Packets	Queue 4
Routing Control Packets	Queue 3
FTP Packets	Queue 2
Web Browsing and Texting Packets	Queue 1 (Default)

Terminal routers are configured with 4 queues with a class based weighted fair queue scheduler. Packets generated from the application models were marked with Diffserv Code Points (DSCPs) and placed in the four queues in each router as indicated in **Error! Reference source not found.**.

SIMULATION RESULTS

Results are organized with respect to the four metrics of Evaluation Criteria. Multi-level AB (MLAB) was evaluated using 1, 2, 4, 6, 8 and 10 request levels and the performance was compared to the most granular LB request approach.

The objective of these tests was to establish baseline DAMA performance using a simple, low risk approach that toggles between the full committed rate and a low trickle rate when the terminal moves between active data transmission and data inactivity. Additional levels of request granularity are then introduced to quantify the effects of attempting more precise load prediction and evaluate whether gains in performance can be achieved. Gains were expected going from a single level (binary request) to multiple levels, with diminishment at some point as predictions become increasingly granular and prediction errors rise. It was expected that a few coarse levels would offer the majority of gains in harvesting unused allocated bytes and thereby increase effective system capacity without the negative side-effects of the very granular LB DAMA request approach.

Test results largely support this hypothesis. As can be seen in the results presented here, all metrics showed improvements going from 1 level (AB) to multiple levels, but improvements leveled off and even reversed in some cases as the request granularity increased towards the LB request approach. Eventually the negative effects of rising prediction error probabilities, specifically under-request errors, and increases in request frequency overshadow the additional minor resource efficiency gains that might result from increasing the prediction granularity further.

Figure 5 shows the packet delay metric for FTP traffic where improvements can be seen going from 1 level to multiple levels. This is due to the increased ability to harvest unused allocations and thus increase effective capacity (i.e., resources were utilized more efficiently to allow more terminal demand to be met). This led to lower packet delays. However, the LB request approach did not see this improvement. Its requests were the most granular, but they suffered more prediction errors due to the inherent uncertainty with traffic arrivals as described previously. Under-request errors were particularly harmful since these caused DAMA to allocate less than the terminals needed and thus increased packet delays as packets backed up in the router queues.

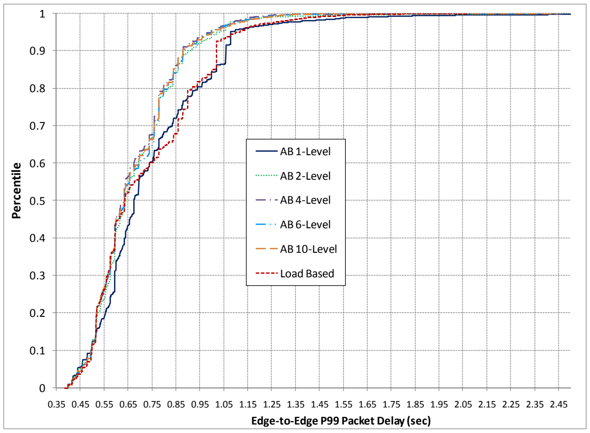


Figure 5 - Packet Delay Metric (FTP)

This same effect can be seen in the next two metrics – under-allocations Figure 6 and misallocated bytes Figure 7. Starting from a single request level, the under-allocations and misallocated bytes are relatively high due to the inability to accurately harvest unused allocations from terminals using less than their full committed rates. This prevents these unused allocations from being allocated to other terminals to satisfy actual demand and thus increase effective capacity. Increasing the request granularity to 2 levels provides a fairly dramatic improvement in both metrics as DAMA is able to identify terminals that only need half or less of their commitments, thus reducing unused allocations.

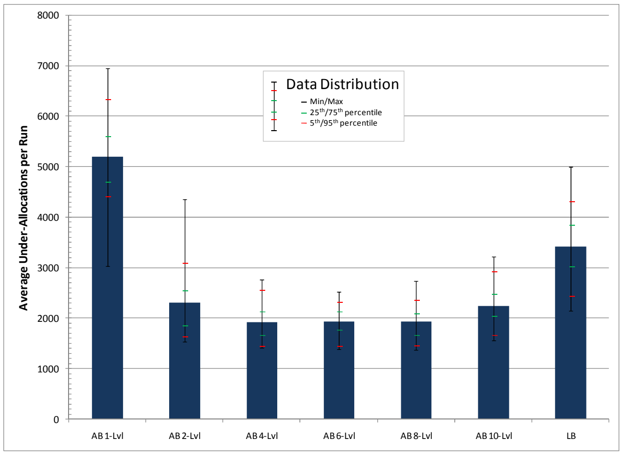


Figure 6 - Under-Allocations Metric

Further levels provide minor improvements up to a point as the average number of under-allocations and misallocated bytes drops slightly. As the number of levels rises approaching 10 and LB, the performance actually worsens. This is the result of increasing under-request errors due to an attempted DAMA request precision that is not supported by the uncertainties involved in traffic arrivals.

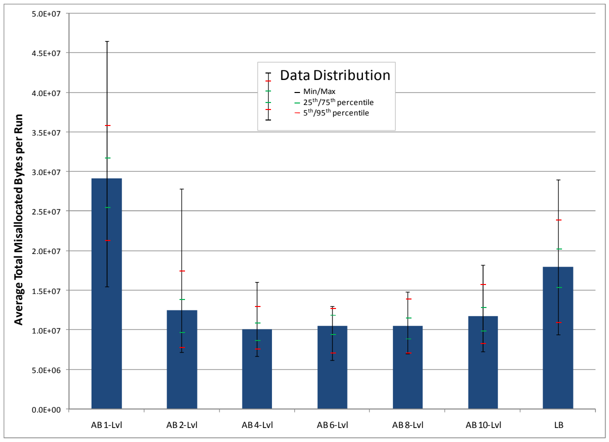


Figure 7 - Misallocated Bytes Metric

The last metric evaluated is the frequency of DAMA request updates necessary with each approach. Since the LB request approach attempts to predict the arrivals precisely, it updates the request every epoch for those terminals with traffic. Other approaches issue request updates only if the arrival rate predicted falls outside of the range predicted previously. The range is based on the number of request levels used. Generally the more request levels used, the more frequently the requests must be updated to reflect changing traffic levels. Figure 8 shows the result.

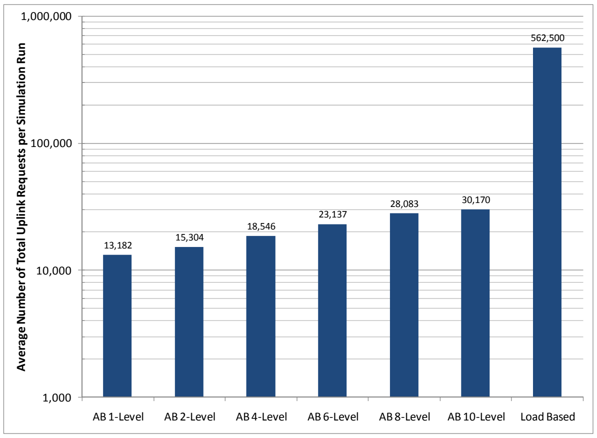


Figure 8 - DAMA Request Frequency Metric

As expected, the AB 1-Level approach required the least request frequency as updates were only necessary if a terminal had traffic before and now does not or vice-versa. Adding additional request levels increases the request frequency with over twice as many requests needed on average using 10 request levels as compared to a single level. The LB approach requires a far higher frequency, with over an order of magnitude more DAMA requests than any of the MLAB approaches. This can offer a significant decrease in both overhead to transmit the requests and the updated allocations as well as the processing and complexity required in the terminal and satellite DAMA.

SUMMARY

Due to long request/grant cycles, FSPN DAMA faces challenges not faced by BoD terrestrial networks. Terminal queue measurements are predictions of a terminal's needs seconds into the future rather than request to service particular packets in queue. This raises concerns for DAMA schemes that attempt precise prediction. As shown here, packet traffic over these timescales is not very predictable in most cases.

DAMA request approaches were evaluated that incorporate load prediction of varying granularity from a simple binary request structure (all or nothing) to a fine grained request scheme that attempts precise prediction. A comprehensive set of evaluation metrics was developed to characterize the ability of each approach to meet terminal SLAs and utilize RF resources efficiently with the least complexity/overhead. A parametric analysis showed that by increasing the prediction granularity only slightly through the use of a few coarse request levels, most of the resource efficiency realizable from FSPN DAMA is achieved without incurring the majority of the negative side-effects of fine-grained load prediction observed for LB request approach such as frequent under-request errors and high request frequency. In general FSPN DAMA requests should recognize the nature of packet traffic by not attempting traffic prediction unsupported by the uncertainties involved and instead focus on meeting terminal commitments as efficiently as possible. As seen, this permits the use of simple terminal reporting and bandwidth allocation models that are effective in delivering SLA commitments while sharing unused resources.

REFERENCES

[1] Vern Paxson, Sally Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", IEEE/ACM Transactions on Networking, vol 3, issue 3, pp 226-244, June 1995.

[2] Yi Qiao, Jason Skicewicz, Peter Dinda, "An Empirical Study of the Multiscale Predictability of Network Traffic", In IEEE Proceedings of HPDC, 2003.

[3] S. A. M. Ostring and H. Sirisena, "The influence of long-range dependence on traffic prediction," in Proc. IEEE ICC'01, vol. 4, Helsinki, Finland, June 2001, pp. 1000–1005

[4] SANG, A., and LI, S., "Predictability Analysis of Network Traffic", In the IEEE Proceedings of INFOCOM 2000 (2000), pp. 342–351.