Understanding Demand Volatility in Large VoD Systems

Di Niu
Department of Electrical and Computer Engineering
University of Toronto
dniu@eecg.toronto.edu

Baochun Li
Department of Electrical and Computer Engineering
University of Toronto
bli@eecg.toronto.edu

Shuqiao Zhao
Multimedia Development Group
UUSee, Inc.
shuqiao.zhao@gmail.com

ABSTRACT

Bandwidth usage in large-scale Video on Demand (VoD) systems varies rapidly over time, due to unpredictable dynamics in user demand and network conditions. Such bandwidth volatility makes it hard to provision the exact amount of server resources that matches the demand in each video channel, posing significant challenges to achieving quality assurance and efficient resource allocation at the same time. In this paper, we seek to statistically model time-varying traffic volatility in VoD servers, leveraging heteroscedastic models first used to interpret economic time series, with the goal of forecasting not only traffic patterns but also traffic volatility. We present the application of volatility forecast to efficient resource allocation that provides probabilistic service level guarantees to user groups. We also discuss volatility reduction from diversification, and its implications to new strategies for cost-effective server management. Our study is based on monitoring the workload of a large-scale commercial VoD system widely deployed on the Internet.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling Techniques; C.2.3 [**Network Operations**]: Network Monitoring; Network Management

General Terms

Measurement, Performance, Reliability

Keywords

Video-on-Demand, Volatility, Traffic Forecast, Demand Prediction, GARCH, Measurement, Resource Allocation, Diversification

1. INTRODUCTION

A large-scale video on demand (VoD) system on the Internet involves millions of users streaming movies, TV episodes, and other on-demand media from a huge library of video channels. However, Internet VoD largely remains a best-effort service, where either a large amount of extra unused server capacity is provisioned with low utilization, or the user experience is at risk. To enjoy smooth playback, a user needs to download at an average rate greater than the video playback rate. It is therefore necessary to ensure the right amount of outgoing bandwidth is available at the servers to meet the instantaneous demand.

To avoid the complication and vast cost in hardware maintenance, more content providers choose to rent public server resources, such as content delivery networks (CDNs), for video streaming. When a VoD application shares the underlying infrastructure with other applications in a "multi-tenant" environment, it is inevitably exposed to random congestion and variance in bandwidth availability. Providing quality assurance to bandwidth-intensive VoD services while not over-provisioning the resources becomes one of the most challenging issues for CDN servers.

As video access patterns exhibit clear trends and periodicity with time-of-day effects [2, 6, 8, 9], the expected bandwidth demand for each video is highly predictable by monitoring the usage history [5, 6]. Demand forecast enables the elastic adjustment of bandwidth allocation to match instantaneous user demand. We envision that a proactive "match strategy" for elastic bandwidth reservation in the presence of time-varying demand is a critical enabling technology to offer service-level assurance to VoD users, while making efficient utilization of resources. Nevertheless, since demand forecast is subject to errors due to unpredictable user dynamics, fast-changing network conditions and inherently noisy traffic, a "risk premium" must be accommodated on top of the expected future demand to tolerate fluctuations, or *volatility*.

In this paper, we argue that in order to elastically book resources for a VoD service, forecasting demand volatility is as important as predicting the expected demand. We seek to statistically model traffic volatility in large-scale VoD systems by analyzing the operational traces of 173 popular video channels collected from UUSee Inc. [1], one of the leading commercial Internet video solutions based in China. Inspection of real-world traces suggests that server bandwidth usage in each video channel exhibits alternating phases of relative tranquility and high variation around its expected value. We thus introduce GARCH [3], a heteroscedastic model originally used to characterize economic time series, to quantify the changing variance in large-scale VoD traffic, with the goal of forecasting volatility. We apply GARCH-based volatility forecasts to bandwidth allocation that economically books resources for each video channel with probabilistic bandwidth guarantees.

As real-world systems typically host very large video libraries, we proceed to study the volatility of the aggregate or mixed traffic of multiple video channels, and observe the volatility reduction phenomenon attributed to diversification. We discuss the implications of such an observation to cost-effective server management and load direction based on financial management tools such as hedging and diversification in real-world VoD systems, which may

NOSSDAV'11, June 1–3, 2011, Vancouver, British Columbia, Canada.

run a large collection of streaming channels over geographically distributed servers.

1.1 Relation to Prior Work

The importance of bandwidth demand estimation to capacity planning in Internet VoD systems has been recognized recently. It is shown that estimating time-varying demands in a large-scale IPTV network can help the system optimally place content on its geographically distributed servers [2]. Toward this goal, the recent demand history is used as an estimate of future demand in each video channel [2]. Apparently, this simple method does not yield accurate forecasts. [6] introduces linear stochastic time series models to capture the periodicity, trends and autocorrelations that exist in the demand history, achieving a high accuracy in demand forecast.

However, traditional forecast methods assume a constant forecast error variance and fail to capture the changing volatility in data. In fact, measurements show that bandwidth demand is subject to rapid changes in some periods, while remaining tranquil and highly predictable in other periods. We therefore introduce GARCH models [3] originated from econometrics to model the volatility persistence phenomenon — the bandwidth demand at a certain time period tends to exhibit similar volatility as in recent time periods.

Volatility reduction in the mixed traffic of multiple channels is similar to the idea of statistical multiplexing and resource overbooking [7] in shared hosting platforms, where the resources are booked to satisfy a certain percentile of demand in each application instead of its worst-case demand, so as to enhance resource utilization. However, the volatility reduction discussed here is novel in three aspects. First, we are concerned with forward-looking resource allocation and volatility forecasts for future demand, while in [7] the resource usage of each application is profiled in an offline and fixed manner, ignoring the change of demand patterns over time. Second, our study focuses on *large-scale* VoD systems, where the concurrent number of users can ramp up by several hundreds or thousands in tens of minutes. In this scenario, any fixed resource usage profiling for small video channels (*e.g.,* those with a user population of 20 in [7]) will be insufficient. Last but not least, we do not assume independence between the demands of channels. Instead, we accurately quantify the conditional demand variance in each channel, which enables the use of financial instruments such as hedging and diversification to achieve cost-effective server management with service level guarantees.

2. TRADITIONAL DEMAND FORECAST: APPLICATIONS AND LIMITATIONS

This research is based on our extensive experiences with UUSee, a real-world on-demand media streaming system widely deployed on the Internet. As one of the leading commercial P2P multimedia solution providers in China, UUSee simultaneously broadcasts tens of thousands of video channels to millions of users distributed across over 40 countries. It implements an optimized peer-assisted delivery structure where users can upload media data to each other, alleviating the server burden. However, servers are still responsible for a large part of the upload and play a critical role in compensating bandwidth shortage and controlling the quality provided [8]. It is worth noting that the observations and mechanisms presented in this paper also apply to any general streaming systems that do not involve peer assistance.

The data for validation in this paper feature the traces collected from 173 popular video channels over 21 days during the 2008 Summer Olympics. The maximum online population in each channel varies from 200 to 8000. The dataset contains server bandwidth consumption in each video channel sampled at a 10-minute frequency, so that there are 144 samples in a day.

From the traces, we note that UUSee users demonstrate diurnal access patterns with time-of-day effects [6, 8], and the popularity of most videos exhibits gradual downward trends after they are released. We can therefore use the so-called Box-Jenkins method [4] to predict the future evolution of server bandwidth demand by learning the trend, periodicity and autocorrelation exhibited in usage history. An accurate prediction of bandwidth requirement can help with server capacity planning and resource provisioning to meet user demands. We now briefly review and generalize a time-series modeling technique specifically tailored for VoD systems first described in [6], and point out its deficiency in handling volatility.

2.1 Forecasting the Expected Demand

Given a time series of interest $\{Y_t\}$, define the backward shift operator $B(\cdot)$ by $BY_t = Y_{t-1}$ and the lag-1 difference operator $\nabla(\cdot)$ by $\nabla Y_t = Y_t - Y_{t-1} = (1 - B)Y_t$. Powers of B and ∇ are defined in the obvious way, *i.e.*,

$$\begin{cases} B^j Y_t = Y_{t-j}, \\ \nabla^j Y_t = \nabla(\nabla^{j-1} Y_t), \quad \text{for } j \geq 1, \quad \text{with } \nabla^0 Y_t = Y_t. \end{cases}$$

We further introduce the lag-d difference operator ∇_d defined by

$$\nabla_d Y_t = Y_t - Y_{t-d} = (1 - B^d)Y_t.$$

The gist of the Box-Jenkins modeling of non-stationary series is to remove periodicity and trends in $\{Y_t\}$, using various differencing transformations, to obtain a stationary series $\{\tilde{Y}_t\}$ that can be modeled by an autoregressive moving-average (ARMA) process [4].

For a particular series $\{Y_t\}$ in VoD systems, *e.g.*, the server bandwidth usage in a video channel, we first apply transformation $\log(\cdot)$ to $\{Y_t\}$ to equalize the fluctuation, and apply ∇_{144} to $\{\log Y_t\}$ to remove daily periodicity. We then difference $\nabla_{144} \log Y_t$ for d times to remove the trend, obtaining a stationary series $\tilde{Y}(t) = \nabla^d \nabla_{144} \log Y_t$, which is well explained by an ARMA(p, q) process. The corresponding seasonal ARIMA model [4] for the original series $\{Y_t\}$ is thus

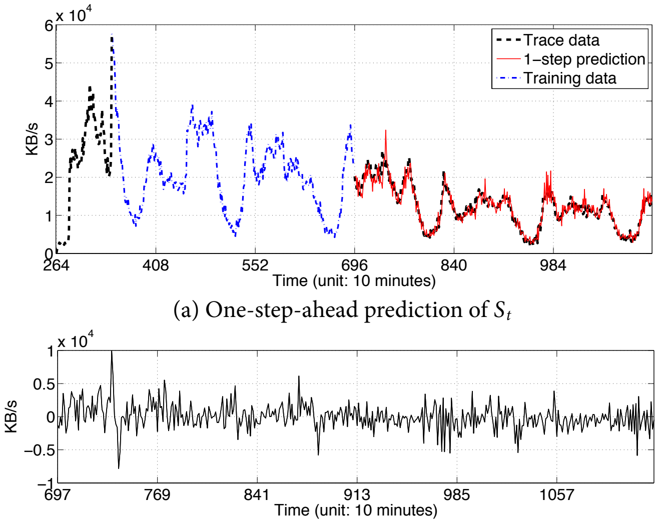
$$\phi(B)\nabla^d \nabla_{144} \log Y_t = \theta(B)Z_t, \quad d \in \{0, 1\}, \tag{1}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ denotes the uncorrelated white noise with zero mean, and $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$ are polynomial operators in B of degrees p and q. The difference order d is chosen from $\{0, 1\}$, depending on whether a trend exists in the daily population variation.

Given $\{Y_1, \ldots, Y_t\}$, let $P_t Y_{t+h}$ $(h > 0)$ denote the h-step-ahead *conditional mean prediction* for Y_{t+h}, *i.e.*, the expected value of Y_{t+h} given observations up to time t. Once the parameters of (1) are learned from the training data, $P_t Y_{t+h}$ is derived as follows. First, we obtain $P_t \tilde{Y}_{t+h}$, the minimum mean square error (MMSE) predictor for $\tilde{Y}_{t+h}$. $P_t Y_{t+h}$ is then calculated by retransforming $P_t \tilde{Y}_{t+h}$ using the inverse of the corresponding operators ∇^d, ∇_{144} and $\log(\cdot)$, *i.e.*,

$$P_t Y_{t+h} = (\nabla^d)^{-1} \nabla_{144}^{-1} \exp(P_t \tilde{Y}_{t+h}). \tag{2}$$

As an example, we make 10-minutes-ahead (one-step) prediction of the server bandwidth $\{S_t\}$ consumed by a popular video channel released at time period $t_0 = 264$ (2008-08-10 10:47:39). The channel has a maximum online population of 2664. The server consumption series of the first 3 days is used as the training data, excluding the initial 80 time periods after the release of the video which may not conform to later evolution patterns. The prediction is tested on the data of 3 days following the training period. We fit model (1) to the training data and obtain parameter estimates through a maximum likelihood estimator [4]. As shown in Fig. 1a, with $d = 0$, $p = 20$,



(a) One-step-ahead prediction of S_t

(b) Prediction errors

Figure 1: 10-minutes-ahead prediction for the server bandwidth consumption S_t of a popular video channel A55FF released at time period 264, compared against the trace data.

$q = 20$, model (1) can yield prediction results that are close to the real server bandwidth required by the channel.

2.2 Applications and Limitations

Demand forecast based on past observations enables the system to allocate a right amount of bandwidth to match the demand. A lightweight online bandwidth monitoring and reservation framework, as shown in Fig. 2, can be unobtrusively implemented in current operational systems. It monitors the server bandwidth consumed by a particular video channel periodically, *e.g., every 10 minutes*, learns models to forecast future demand, and judiciously decides the amount bandwidth to be provisioned in the next time period. As server bandwidth usage of each channel is readily available in server logs, there is no need to collect statistics from users.

However, as the MMSE predictor in Sec. 2.1 only forecasts the conditional mean demand, or the expected demand, the real demand may vary around this predicted conditional mean. The resulted prediction errors have a mean of zero and are plotted in Fig. 1b. Due to the existence of prediction errors, we need to provision an additional "risk premium" to tolerate demand fluctuation, which apparently depends on the variance of prediction errors. A further look into Fig. 1b suggests that forecast errors of server bandwidth consumption do not have a constant variance: there are periods where the prediction is relatively accurate alongside periods with less trustworthy forecasts. In other words, the server usage evolves smoothly and is highly predictable in some periods, but also becomes highly variable and unpredictable in other periods.

The changing volatility in resource consumption poses great challenges to efficient server provisioning. As traditional time-series techniques assume a constant variance for the disturbance Z_t, the risk premium provisioned will be a function of the fixed forecast error variance averaged over the long run. However, such an approach will necessarily result in over-provisioning when the demand is less volatile, but lead to insufficiency when the demand is liable to unpredictable changes. To achieve efficient resource allocation, we need a *conditional variance* forecast mechanism to estimate the time-changing variance of demand around its expected value (conditional mean), and adjust the amount of risk premium provisioned accordingly.

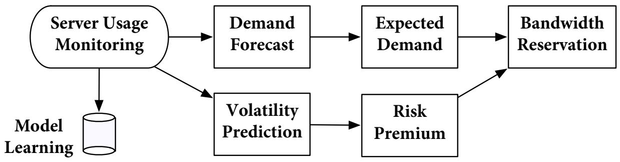
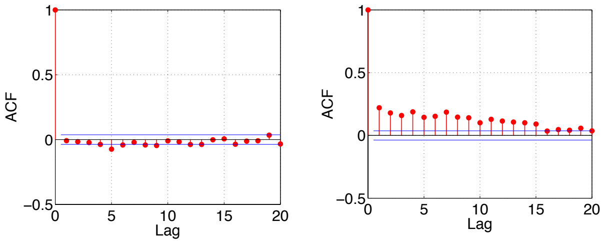


Figure 2: Online server bandwidth monitoring and reservation. The reserved bandwidth should match the sum of the expected demand and a risk premium that tolerates volatility.



(a) ACF of Z_t

(b) ACF of Z_t^2

Figure 3: The ACFs of the disturbance Z_t and Z_t^2 obtained from fitting the seasonal ARIMA model (1) with $d = 0$ and $p = q = 20$ to server bandwidth consumption $\{S_t\}$ in video channel A55FF.

3. MODELING DEMAND VOLATILITY

As a system operator, one may be interested in estimating the conditional variance in server bandwidth demand based on past observations, which is critical to deciding the "risk premium" provisioned to accommodate demand fluctuation. Such a "risk premium" should increase when we are less certain about the accuracy of the conditional mean demand forecast, and decrease otherwise, when the conditional variance of future demand is low. In contrast, the unconditional variance, *i.e.*, the long-run estimate of forecast error variance averaged over time, would not be important if we care about the instantaneous "risk premium" needed.

Although the seasonal ARIMA model (1) is a good conditional mean model that predicts the expectation of server bandwidth consumption S_{t+h} conditioned on $\{S_t, S_{t-1}, \dots\}$, it fails to capture the serial dependency within the disturbance series $\{Z_t\}$ obtained from fitting (1) to $\{S_t\}$. To check such dependency, we plot the autocorrelation functions (ACFs) of Z_t and Z_t^2 in Fig. 3. We can see that although $\{Z_t\}$ is an uncorrelated white noise, it is not IID — the variance term Z_t^2 clearly depends on $Z_{t-1}^2, Z_{t-2}^2, \dots$. In fact, we can observe a persistence of volatility for Z_t from Fig. 1b: Z_t tends to exhibit a similar conditional variance as in recent periods.

To include past variances in the explanation of future variances, we model Z_t using the GARCH (generalized autoregressive conditional heteroscedasticity) process [3], which has been successfully applied to modeling the volatility of stock data for the past decade. Specifically, we model the disturbance Z_t obtained from fitting model (1) to $\{S_t\}$ as a GARCH(P, Q) process:

$$\begin{cases} Z_t = \sqrt{h_t}\, e_t, \quad \{e_t\} \sim \text{IID}\, \mathcal{N}(0,1), \\ h_t = \alpha_0 + \sum_{i=1}^{P} \alpha_i Z_{t-i}^2 + \sum_{j=1}^{Q} \beta_j h_{t-j}, \end{cases} \tag{3}$$

where $\alpha_0 > 0$ and $\alpha_j, \beta_j \geq 0$, $j = 1, 2, \dots$, and h_t is the conditional variance of Z_t given its history $\{Z_s; s < t\}$. The GARCH model reflects the evolution of the variance in data by incorporating correlation in the sequence $\{h_t\}$ of conditional variances.

Taking channel A55FF as an example, we fit a GARCH$(1, 1)$ model to the one-step-ahead prediction errors for server bandwidth usage $\{S_t\}$ shown in Fig. 1b. The model parameters are obtained using

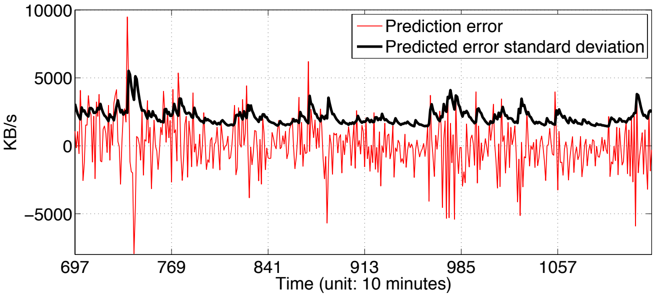


Figure 4: One-step-ahead forecast errors for server bandwidth consumption S_t in the video channel A55FF, and predicted conditional standard deviations for the forecast errors.

maximum likelihood estimation (pp. 417, [4]) based on the prediction errors of model (1) during the training period. We predict the conditional standard deviations $\{\sqrt{h_t}\}$ of the prediction errors for the test period using the trained model, and plot the results in Fig. 4. We can see that the predicted error standard deviation is larger when the demand prediction errors are highly variable.

With the GARCH model, we are able to forecast how much real data will deviate from the predicted conditional mean produced by model (1). It allows us to quantify our certainty about bandwidth consumption forecast so that server provisioning can leverage this fact to enhance resource utilization. When we are certain about the S_t in the next time period, the server bandwidth reserved for the channel should be close to the predicted conditional mean consumption. On the other hand, during periods where S_t is subject to rapid changes and less predictable, we need to provision a higher risk premium to tolerate the demand volatility in the channel.

4. VOLATILITY AND RESOURCE ALLOCATION

In this section, we present the application of volatility forecasts to resource allocation. To achieve service level guarantees to users without over-provisioning, the resource allocated should match the future demand with a conditional mean forecast plus a "risk premium" that tolerates traffic volatility.

In general, the effectiveness of a resource allocation scheme can be evaluated by two performance metrics: 1) *insufficiency ratio* e, which is the ratio of time periods where the booked resource is lower than the actual demand over all the test periods; and 2) *time-averaged utilization* $\overline{U}$, which is the average utilization of the allocated resource over all the test periods.

To provide quality assurance to users, it is expected to maintain the insufficiency ratio e to a low level. On the other hand, for the cost-effectiveness of servers, the cloud service providers expect to keep the average utilization $\overline{U}$ at a high level by booking resources sparingly. Striking a balance between the two conflicting objectives, the key to successful resource booking is to decide the minimum necessary "risk premium" R_{t+1} at time period $t+1$ given observations up to time t that achieves a target insufficiency ratio, *e.g.*, $e \leq 2\%$, under an appropriate volatility model.

4.1 Comparing Five Volatility Models

We propose five proactive server bandwidth reservation schemes for VoD systems, each based on a different volatility model including GARCH and other heuristics. To reserve bandwidth for a video channel at time $t+1$, all these schemes periodically monitor the server bandwidth usage $\{S_1, \ldots, S_t\}$ of this channel by checking server logs (*e.g.*, at a 10-minute frequency), and predict the conditional mean demand $P_t S_{t+1}$ using the same method described in Sec. 2. However, they incorporate different volatility models to determine the "risk premium" provisioned. Specifically, assuming $e = 2\%$, these schemes are described as follows:

Constant variance (baseline method) assumes a constant variance σ^2 in demand forecast errors, which can be learned from training data. As demand forecast errors exhibit Gaussian distribution in the traces, the risk premium is the 98th percentile of the normal distribution $\mathcal{N}(0, \sigma^2)$, *i.e.,* the value below which 98% of samples from the distribution $\mathcal{N}(0, \sigma^2)$ fall.

Probabilistic GARCH predicts the conditional variance h_{t+1} for the demand forecast error using the GARCH model. The risk premium is the 98th percentile of the normal distribution $\mathcal{N}(0, h_{t+1})$.

Deterministic GARCH predicts the conditional variance h_{t+1} for the forecast error using the GARCH model. The risk premium is $\eta\sqrt{h_{t+1}}$, where η is a positive constant determined as the η that achieves an insufficiency ratio $e = 2\%$ in the training data.

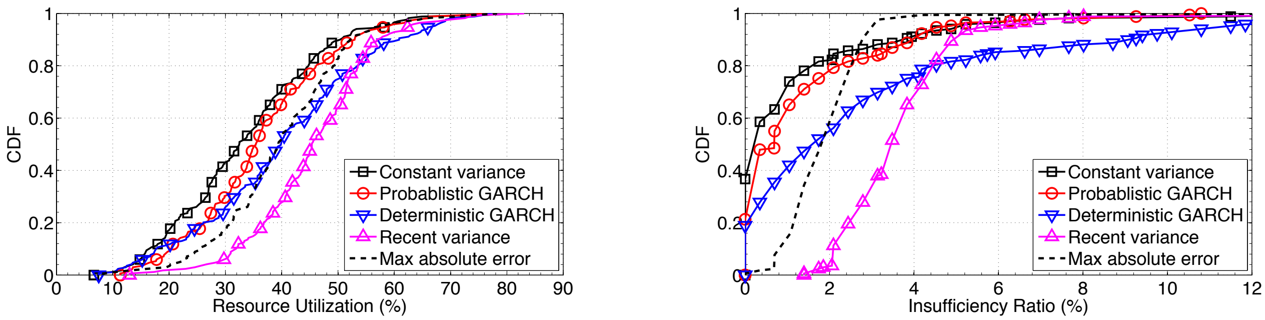
Recent variance is a heuristic method that calculates the sample variance σ_τ^2 of demand forecast errors in the recent τ time periods. The risk premium is the 98th percentile of the normal distribution $\mathcal{N}(0, \sigma_\tau^2)$.

Maximum absolute error is a heuristic method that decides the risk premium as the maximum absolute error of demand forecasts in recent τ time periods.

We test the performance of the above schemes in terms of resource utilization and insufficiency ratio through simulations driven by the real-world traces of 169 popular video channels. For each channel, we train the conditional mean model (1) and conditional variance model (3) based on the data of 1.5 days from the 50th to the 266th time period after the channel is released. The first 500 minutes (50 time periods) are excluded from training as the initial demands may not conform to the later evolution patterns. To test the generalizability of our proposed methods to any VoD demands, we only assume a simple seasonal ARIMA model (1) with $d = 0$, $p = q = 1$ for demand forecast. For the two GARCH-based methods, we train a GARCH$(1,1)$ model for demand forecast errors in the training data. We then use the above 5 methods with the trained parameters to perform one-step-ahead bandwidth reservation in each channel for a test period of 2 days following the training period.

We calculate the average resource utilization $\overline{U}$ and insufficiency ratio e in each of the 169 channels over its corresponding test period, and plot the empirical cumulative distribution functions (CDFs) of these $\overline{U}$'s and e's in Fig. 5a and Fig. 5b, respectively. We can see that "constant variance," as the baseline method, achieves the lowest resource utilization as well as the lowest e, because it aggressively books a high risk premium assuming a large constant variance for demand forecast errors. In contrast, "probabilistic GARCH" can adjust the risk premium dynamically based on the changing forecast error variance. From Fig. 5b, we see that it achieves an insufficiency ratio $e \leq 2\%$ in 80% of the 169 channels, which is close to "constant variance," and an insufficiency ratio $e \leq 4\%$ in more than 90% of the channels, which is even better than "constant variance." In the meantime, "probabilistic GARCH" achieves a markable enhancement in resource utilization, shown in Fig. 5a as it only books the necessary risk premium to tolerate the instantaneous demand volatility instead of the long-run variance.

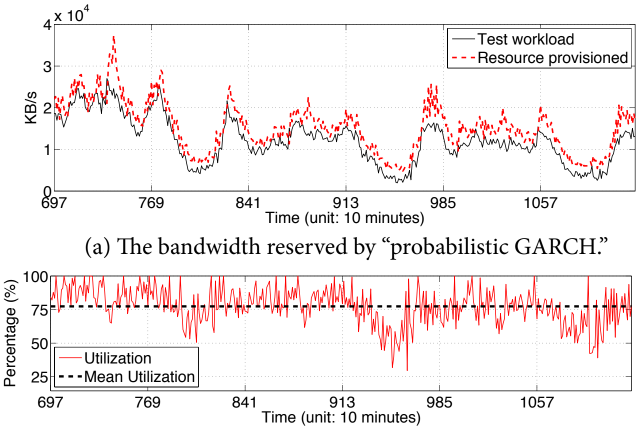
Although "deterministic GARCH" further enhances utilization by booking the risk premium more conservatively, its average e exceeds the target of 2%. Furthermore, the "recent variance" and "maximum absolute error" heuristics, when compared with the first three methods, enjoy an advantageous position, as they are fully adaptive during the test period, while the first three use fixed model parameters estimated from the training data for prediction. Even in such



(a) CDF of resource utilization $\overline{U}$ in 169 channels

(b) CDF of bandwidth insufficiency ratio e in 169 channels

Figure 5: The empirical CDF of the utilization of booked bandwidth and the ratio of time periods where the booked bandwidth is insufficient. Bandwidth reservation is performed in 169 popular video channels independently, each with a test period of 2 days.



(a) The bandwidth reserved by "probabilistic GARCH."

(b) Utilization of the reserved bandwidth.

Figure 6: Bandwidth reservation with "probabilistic GARCH" in channel A55FF, tested on a period of 3 days.

an unfair comparison, both heuristics perform worse than GARCH-based methods, resulting in an e way beyond the target of 2%. As they inherently lack a mechanism to quantitatively tune the trade-off between $\overline{U}$ and e, they are not appealing for the sake of quality assurance. We conjecture that a fully online version of ARIMA and GARCH, which adaptively relearns model parameters as the simulation proceeds in the entire test period, can yield even better utilization while being able to constrain e within the target range.

4.2 Utilization as a Volatility Indicator

From the above comparison, we find that the best bandwidth reservation scheme that strikes a balance in the $\overline{U}$-e tradeoff is "probabilistic GARCH," thanks to its superior ability to adjust the risk premium provisioned as the volatility changes. As an example, we apply "probabilistic GARCH" to a popular channel A55FF over a test period of 3 days, which represents the same test workload as in Fig. 1 in Sec. 2 and Fig. 4 in Sec. 3, with models learned from the preceding 3 days. The bandwidth demand is forecasted as a seasonal ARIMA process with $d = 0$, $p = q = 20$ driven by GARCH$(1,1)$ forecast errors. We plot the bandwidth provisioned by "probabilistic GARCH" in Fig. 6a and the achieved utilization in Fig. 6b under a target insufficiency ratio of $e = 5\%$. The achieved e is 3.71% and $\overline{U} = 75.04\%$ on the test data.

Just as traffic volatility limits the resource utilization efficiency, the best achievable utilization $\overline{U}$ produced by one-step-ahead resource booking, given a target insufficiency ratio e (*e.g.*, $e = 2\%$), is essentially a quantitative indicator of traffic volatility in the long term. A high $\overline{U}$ means that less "risk premium" is needed and the traffic is likely to evolve smoothly with tractable variation. In contrast, a low $\overline{U}$ implies that the traffic is inherently liable to rapid and unpredictable variations, and thus more "risk premium" must be booked to tolerate demand fluctuation. In other words, $\overline{U}$ evaluates volatility by the ratio of the actual bandwidth usage over the sum of expected usage and the volatile part of usage. Therefore, in the following analysis, we use the time-average utilization $\overline{U}$ achieved by resource booking based on "probabilistic GARCH" to evaluate the demand volatility in the long run.

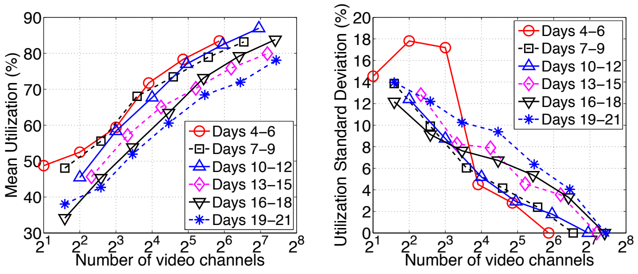
5. VOLATILITY REDUCTION THROUGH DIVERSIFICATION

In this section, we consider the combined or mixed traffic of multiple video channels. We observe that the aggregate traffic volatility decreases as the number of channels to be combined increases. Furthermore, such volatility reduction is not merely a consequence of a higher volume of traffic, but more importantly is due to traffic mixing between channels that may exhibit diverse variations. From the previous section, we see that $\overline{U}$ evaluates volatility by computing the ratio of the actual bandwidth usage over the sum of the expected usage and usage uncertainty. Therefore, in this section, we evaluate the demand volatility of mixed channels by the achievable $\overline{U}$ in one-step-ahead resource booking, given a target of $e = 2\%$.

To study how the number of combined channels affects traffic volatility, we conduct a series of bandwidth reservation simulations driven by the traces of 173 video channels in UUSee over 21 days. We divide the entire 21 days into 7 periods, each of 3 days, and study the traffic volatility of random combinations of video channels in each 3-day test period (except the first 3 days, which are only used for model training).

Now we illustrate the test with the second 3-day period (days 4-6). Suppose that the number of channels to be combined is n. We *randomly* choose n channels that exist in this test period and its previous 3-day training period, and consider their aggregate traffic. We perform bandwidth reservation based on "probabilistic GARCH" in this test period (days 4-6), using the models trained from the previous 3 days (days 1-3). We assume a simple conditional mean model (1) with $d = 0$, $p = q = 1$ and a conditional variance model (3) of GARCH$(1,1)$. We calculate $\overline{U}$ and e in this test period of days 4-6. For this particular value of n, the above experiment is repeated for 60 times to allow for different combinations of channels. The mean and standard deviation of the achieved 60 $\overline{U}$'s and e's are calculated. The above procedure applies to the other 3-day test periods, each with its preceding 3 days as the training period.

From Fig. 7, we see that as bandwidth is reserved for an increasing



(a) Mean of utilization $\overline{U}$

(b) Standard deviation of $\overline{U}$

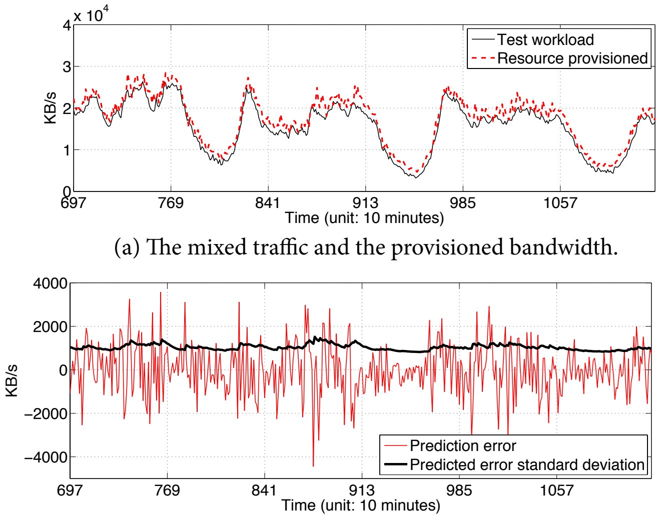
Figure 7: The mean and standard deviation of $\overline{U}$ when different numbers of video channels are randomly combined for bandwidth reservation. The achieved average e is less than 4% in all cases.

number of combined video channels, the mean of the achieved $\overline{U}$'s increases significantly up to close to 90%, and the standard deviation of $\overline{U}$'s decreases in all of the test days 4-21. As the resource utilization in one-step-ahead bandwidth reservation essentially evaluates the degree of volatility in data, the above phenomenon implies that the aggregate traffic of multiple channels demonstrates less volatile. Bandwidth provisioned to multiple channels can thus match the demand more closely. A further check into Fig. 7b shows that when the number of channels is small, the $\overline{U}$'s have a high standard deviation. This means that although all channels follow diurnal evolution in trend, there exist different degrees of correlation between the demand forecast errors of different channels: the volatility of the combined traffic is amplified when they are positively correlated and suppressed when they are negatively correlated.

To verify that volatility reduction is not a consequence of an increased amount of total traffic, but because of mixing different channels, we perform bandwidth reservation for all the 93 channels in a test period from time 697 to 1127, which is the same test period as in Fig. 1, Fig. 4, and Fig. 6a, where bandwidth reservation is performed for a single channel A55FF. However, instead of considering the aggregate traffic all 93 channels (including channel A55FF), we consider a mixture of them by taking 1/10 of user requests from each channel and adding them up. Comparing Fig. 8a with Fig. 6a, we see that the resulted mixed traffic is roughly the same in size as that of the single channel A55FF. But the mixed traffic evolves more smoothly, and its forecast errors shown in Fig. 8b not only have a smaller variance than that of channel A55FF shown in Fig. 4, but also exhibits less heteroscedastic property, *i.e.*, less time variation in terms of variance.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we focus on demand volatility forecasts in large-scale operational VoD systems, with the objective of dynamically and efficiently provisioning bandwidth resources in VoD servers. We introduce GARCH models originated from econometrics to predict demand volatility based on server usage monitoring. We propose and compare five volatility-aware resource provisioning schemes, based on GARCH modeling and other volatility heuristics. It is shown that GARCH models yield the best tradeoff in terms of resource utilization and the service level provided to users. We further study the volatility reduction due to diversification when traffic of multiple video channels is mixed. As a result, allocating resources for a well diversified collection of video channels can improve resource utilization by up to 3 times as opposed to single-channel allocation.



(a) The mixed traffic and the provisioned bandwidth.

(b) Conditional standard deviations for forecast errors.

Figure 8: Bandwidth reservation with "probabilistic GARCH" for the mixed traffic of all the 93 channels in the 6-day period from time 264 to 1127. The first 3 days are the training period, and the other 3 days (time 697-1127) are the test period.

The volatility reduction phenomenon observed in UUSee traces has laid a foundation for using modern portfolio theory such as hedging and diversification to achieve cost-effective server management, as the server cost directly links with both the mean and volatility of its usage. In our on-going work, we consider geographically distributed video servers, such as CDN nodes, and explore the use of hedging to enhance resource efficiency and to reduce usage fluctuation by mixing the video channels with negatively correlated demands.

7. REFERENCES

[1] UUSee Inc. [Online]. Available: http://www.uusee.com.

[2] D. Applegate, A. Archer, V. G. S. Lee, and K. Ramakrishnan. Optimal Content Placement for a Large-Scale VoD System. In *Proc. of ACM CoNEXT*, Philadelphia, USA, November, 2010.

[3] T. Bollerslev. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[4] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control.* WILEY, 2008.

[5] G. Gürsun, M. Crovella, and I. Matta. Describing and Forecasting Video Access Patterns. In *Proc. of IEEE INFOCOM '11 Mini-Conference*, Shanghai, China, April 11-15 2011.

[6] D. Niu, Z. Liu, B. Li, and S. Zhao. Demand Forecast and Performance Prediction in Peer-Assisted On-Demand Streaming Systems. In *Proc. of IEEE INFOCOM '11 Mini-Conference*, Shanghai, China, April 11-15 2011.

[7] B. Urgaonkar, P. Shenoy, and T. Roscoe. Resource Overbooking and Application Profiling in Shared Hosting Platforms. In *Proc. of the 5th Symposium on Operating Systems Design and Implementation (OSDI)*, Boston, Massachusetts, December 9-11, 2002 2002.

[8] C. Wu, B. Li, and S. Zhao. Multi-Channel Live P2P Streaming: Refocusing on Servers. In *Proc. of IEEE INFOCOM '08*, Phoenix, Arizona, 2008.

[9] H. Yin, X. Liu, F. Qiu, N. Xia, C. Lin, H. Zhang, V. Sekar, and G. Min. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics. In *Proc. of Internet Measurement Conference (IMC)*, Chicago, Illinois, November 4–6 2009.