Quantifying Cloud Workload Burstiness: New Measures and Models

Abiola Adegboyega
Electrical & Computer Engineering
The University of Calgary
Calgary, Alberta
aadegboy@ucalgary.ca

***Abstract*—diverse cloud applications deployed on-demand make for workload burstiness. Burstiness is quantified statistically through different variance measures. This paper focuses on the statistical measures used to quantify cloud workload burstiness. Using diverse workloads, it identifies different statistical models that uniquely capture workload specific burstiness. Subsequently, it employs recent econometric models described as Auto-regressive Conditional Score (ACS) motivated by their ability to model time-varying parameters that capture burstiness more accurately than existing methods. Furthermore, it has inspired a novel measure of burstiness, the Normalized Score Index (NSI). Compared to existing measures, the NSI captures burstiness specific to statistical features per workload. When standard variance features are observed, the NSI reverts to traditional measures and when nonstandard features are present, it models them accordingly. The NSI has been applied to a diverse workload set and yields both a static metric and a means by which to track burstiness over a workload's lifecycle.**

Keywords—burstiness; variance; workloads; score

I. INTRODUCTION

Workload burstiness arises given that applications with unique characteristics compete for finite cloud resources. Cloud workload variance has many descriptive terms. *Burstiness, spikes, fluctuations, slashdot effects* and *flash-crowds* all quantify variance. While these have been used to model various variance characteristics in cloud workloads, there is no general consensus regarding definition & no *quantitative* consensus on what differentiates a spike in traffic from a flash-crowd as both are often categorized as bursty. In this paper, burstiness is the collective term for variability given its use in the research literature. Existing measures of burstiness largely employ workload models with underlying assumptions of normality. That is, cloud workloads are generally assumed to be composed of variables drawn from the normal distribution. While this is valid in many cloud use-cases [1] recent research has uncovered new observations that are not explained or captured by the normality assumption. These observations are made by a study of the time-series of workloads. Loboz [2] made observations of extreme value distributions in Microsoft Azure cloud computing environments. The study provides statistical evidence of heavy-tailed distributions which are not explained by assumptions of normality. Analysis of cloud storage traffic has been shown to exhibit right-tailed distributions [3].

This paper identifies current measures of cloud workload burstiness by the analysis of their time-series which is elicited from synthetic and real traces. It studies them given the normality assumption and identifies salient statistical features that determine when the assumption is valid. Furthermore, it isolates statistical features that determine when the normality assumption becomes inadequate. Then, it develops a new measure of variance described as the *Normalized Score Index* (NSI). This was motivated by recent econometric models described as conditional score models. Its uniqueness from current methods is discussed as well as its contributions and extensions thereof. Subsequently, its practical realization as solutions employed in cloud computing scenarios is elaborated. This paper's contributions are as follows:

- A novel measure of burstiness, the NSI.
- A methodology to determine when existing burstiness measures & the NSI should be used.
- An extensible algorithm by which to estimate the NSI parameters for models statistically determined to provide the best fit per workload.

This paper is organized as follows. Section II presents the state-of-the-art in existing measures of cloud workload burstiness and details their use in current research. Section III presents the realization of the Normalized Score Index (NSI). Section IV details its evaluation using a diverse set of cloud workloads and different use-case scenarios while conclusions are given in section V.

II. STATE-OF-THE-ART IN MEASURING WORKLOAD BURSTINESS

A. Current Measures of Burstiness

Current measures of burstiness focus on the relationships between the observations of the time-series of workloads. These account for burstiness according to different properties. Table 1 provides the measures of burstiness employed in current research. All the parameters are specified for a time-

Table 1: Current Measures of Cloud Workload Burstiness

Measure	Details
Auto-Correlation Function (ACF)	$ACF(t,b) = \frac{E[(Y_t-\mu_t)(Y_b-\mu_b)]}{\sigma_t \sigma_b}$
Index of Dispersion, I	$I = SCV(1 + 2\sum_{i=1}^{\infty} \rho_i)$
Hurst parameter. H	$VAR[Y^{(m)}] \sim km^{2H-2}$ as $m \rightarrow \infty$
Entropy, $H_S(Y)$	$H_S(Y) = -\sum_{i=1}^{n} P(Y_i) \log P(Y_i)$

series Y_t (t = 1,2,3,..). In table 1, E is the expectation of Y, σ_t and σ_b are standard deviation at lags t and b, SCV is the squared coefficient of variation (σ/μ); ρ is ACF; k is a constant, m the time-series sample size; H is the Hurst parameter with range [0.5, 1]; $P(Y_i)$ is the probability of Y having value at Y_i occur. Their essential features are discussed.

The ACF has been used to model 'burstiness' under two distinct categories: *temporal* and *spatial* burstiness. Temporal burstiness relates to the correlations that exist between the observations of a time-series over lags. This is the standard definition of the ACF. The correlations that occur over lags in a time-series are quantified as burstiness given that the relationship is temporal over *finite* periods of time (lags) [4]. Spatial burstiness [5] characterizes burstiness according to the magnitude or size of the observations of a time-series under study. From a statistical perspective, spatial burstiness arises due to the presence of heavy tails observed in the conditional distribution of a time-series.

The index of dispersion is expressed in terms of the mean value of a random variable to normalize the dispersion captured with the variance. It has been used in current research as a measure of burstiness. The Hurst parameter is a measure of long-range dependence or the long-range memory of a time series. The Hurst parameter relates to the rate at which the temporal correlations vary over time between the observations of a time-series. Table 1 gives the expression in terms of the variance. The Hurst parameter range for burstiness is [0.5, 1]. Increase in value means increase in burstiness. Entropy as a measure of the uncertainty in the content of information models workload bursts given that they represent sudden departures from regular behavior. The Shannon entropy [6] is given in table 1.

B. Current Reseach Methods In Measures of Burstiness

There are important drawbacks regarding the measures of burstiness discussed. The index of dispersion contains an infinite sum given the ACF function as shown in table 1. Practical implementations arbitrarily determine an upper limit given the environment & a practical implementation [7] replaces the infinite sum with request arrivals over a finite window while employing another method for burst detection. The main drawback of the Hurst parameter as a measure of burstiness is the lack of a unified approach in its estimation [8]. Different implementations and realized algorithms yield different values for the same parameter. The Shannon entropy requires a long history of the time-series before calculation of its probability density function subsequent to calculation of the entropy measure. Furthermore, it is unable to distinguish between a slow increase in workload and an sudden increases representing intense workloads [8].

III. THE NORMALIZED SCORE INDEX (NSI)

In order to mitigate some of the drawbacks discussed, 4 traces from production clouds were studied. The aim was the discovery of workload dynamics to enable the determination of the measure of burstiness that captures its salient statistical properties. They are described in brief.

A. Analysis of Diverse Cloud Workloads

Four workloads were selected from production cloud environments identified as series I through IV. Series I is from recent cloud video traffic study in [9]. Series II comes from traces obtained from an Infrastructure-as-a-Service (IaaS) cloud environment [10] containing aggregate data for CPU, memory and network I/O from 1700 VMs over 4 months. Series III is the well-studied trace from Google's compute cluster composed of 12,500 nodes. Series IV comes from the evaluation of personal cloud storage [11]. For the four workloads discussed, the time-series for each was elicited upon which standard statistical methods were employed in their analysis. The signal + residual model is adopted in the analysis of each time-series given its basis for the realization of linear classical models of time-series. This methodology is illustrated in Figure 1a.

B. Nonstandard Statistical Features in Cloud Workloads

With reference to Figure 1a, current measures of burstiness are captured by white noise residuals and squared residuals given the left and right branches of the decision loop for correlation. These are *standard* statistical features representative of traditional linear and nonlinear models such as the well-known Generalized Auto-Regressive Conditional Heteroskedastic model (GARCH). Figure 1b illustrates the residual ACF for all series under study. Series I ACF represents temporal burstiness while the (squared) residuals for series III represent variance that is captured by GARCH models. The residual ACF of series II & IV however represents those statistical features that are not explained by temporal correlations or GARCH variance. These are described as *nonstandard*. The empirical distribution for series II is illustrated in Figure 1b. Both series II and IV (not shown) exhibit a right-tailed distribution as shown in the empirical histogram. These corroborate the extreme value distributions similarly observed in Microsoft cloud environments [2] These observations necessitate a new measure given that existing linear classical & nonlinear models do not explain them.

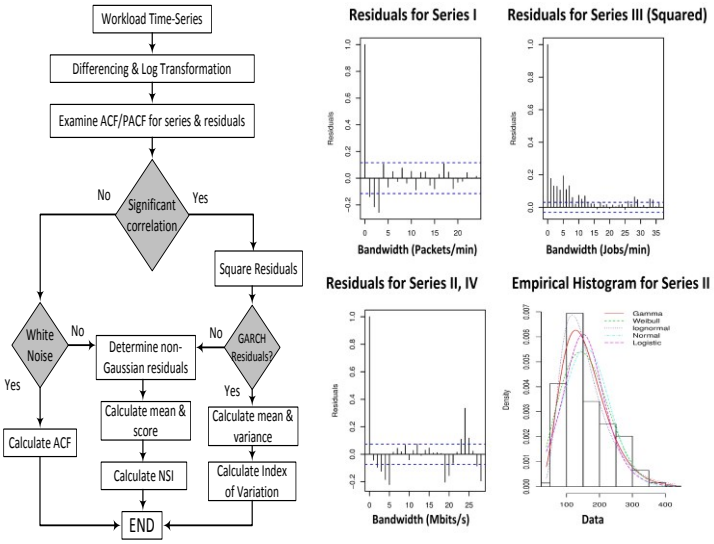

Figure 1a: Methodology 1b: Workload Residual Analysis

C. Conditional Score Models

Research conducted in the field of econometrics has led to new perspectives in modeling non-normal errors as was observed for series II and IV which present characteristics which are more accurately described by measures described in econometrics as the score, the derivative of the time-series done with likelihood estimation. Harvey [12] presents new approaches in modelling time-varying properties, both in the mean and the variance, of a time series described as Autoregressive Conditional Score (ACS) models. The score refers to the derivative of the maximum likelihood estimate of the random variable describing the time-series. To specify the model, the time-series under study Y_t with error term ε_t is assumed to be white noise with zero mean and finite variance. A general distribution is assumed for Y_t, that is not the normal distribution: $Y_t \sim p(Y_t|f_t,\theta_t)$. Here again, f_t is the time-varying parameter; θ_t the corresponding parameter vector. f_t is variously the mean μ for linear models and the variance σ^2 for nonlinear models. Likewise, the parameter vector θ_t in model realization is $\theta(\mu,\sigma^2)$ depending on statistical observations. When analysis determines departures from normality and the GARCH model as illustrated in Figure 1a, then a new measure is realized according to the observations. Writing f_t in the autoregressive form, we get: $f_t = \omega + \sum_{i=1}^{p} \beta_i f_{t-i} + \sum_{j=1}^{q} \alpha_j s_{t-j}$ where $\omega,\{\alpha_i,\beta_j\}$ are coefficients $\{i,j = 1,,p;1,,q\}$ respectively & parameters determined by log-likelihood:

$$L(\theta|y_1,\dots,y_t) = g(y_1,\dots,y_n;\theta) = \prod_{t=1}^{n} g(y_i;\theta) \quad (1)$$

Taking the second derivative of the score function of y_t yields the scaled score:

$$s_t = S_t.\nabla_t = \frac{\partial^2 \ln g(y_t|f_t;\theta)}{\partial^2 f_t} \quad (2)$$

The scaled score function of equation (2) captures burstiness specific to the probability distribution of the time-series.

D. Development of the NSI

The measures of interest, which vary over time, are the mean and the score. The realization given existing work in statistics uses scoring rules which are used to measure the predictive power of a forecasting solution, well-known methods used in the statistics literature [13]. For a time-series Y_t, $t = 1$ to T, a score S_{t+1} is defined for the prediction done at S_t with the scoring rule defined by:

$$S_{t+1} = S(Y_{t+1}, p(Y_{t+1}; \hat{\theta}_{t+1})) \quad (3)$$

where p is the probability density function of Y and θ is the parameter which in the Gaussian case is the variance σ^2. The standard form of scoring rule for (3) from [13] is:

$$\overline{NLS} = -\frac{1}{T}\sum_{t=1}^{t+T-1} \log p(Y_{t+1}; \hat{\theta}_{t+1}) \quad (4)$$

This is the Negative Log Score (NLS), averaged over the observations T of the time-series. With reference to Index of dispersion ($I = \sigma^2/\mu$), if we replace the variance by the score of equation 4, we get the one-step score prediction & moving average of the time-series:

$$NSI = -\frac{1}{T}\sum_{t=1}^{t+T-1} \left(\frac{\log p(Y_{t+1};\hat{\theta}_{t+1})}{\mu_t}\right) \quad (5)$$

This is the normalized Score Index given that it is normalized by the mean value of the time-series over T.

IV. EVALUATION

A. Comparison with Existing Methods

Given the analysis and realization of the NSI, a performance analysis exercise was conducted one in which it was compared with existing measures of burstiness. The four series selected were evaluated for burstiness with the Hurst parameter, Shannon entropy and with the NSI. Table 2 shows

Table 2: Evaluation of Time-Series for Burstiness

Series	NSI	Hurst Parameter	Entropy
I	0.15	0.67	1.9
II	0.16	0.99	1.9
III	0.76	0.99	2.2
IV	0.39	0.72	1.7

the results. While there is some degree of comparison possible for each method, the criticism of the Hurst parameter is the narrow range [0.5, 1] & a lack of real granularity in measuring burstiness. The Shannon entropy does not suffer from the limited range but the returned values conflict with the Hurst parameter. The NSI on the other hand provides a wider range of values which can also be expressed as a percentage of burstiness. For instance, compute cluster workloads (III, IV) exhibit a high degree of variation and a comparison of values in table 2 compare well with the research in [8] given that existing measures do not quantitatively reflect the diversity of cloud applications as well as the usage scenarios. This also provides a way to rank applications based on their measure of variation to plan adequate resource provisioning.

B. Cloud Computing Use-Cases

The static value of the NSI is useful for comparison with existing measures of burstiness but of limited use in practical cloud scenarios. Thus, an algorithm was devised, one that employs predictive score-based method in the computation of the NSI over measurable time-windows. This is better able to

Input: $\boldsymbol{Y} = (Y_i|1{:}T)$
While $t < T$ **DO;** Initialize $w = w_0$; $\alpha = \alpha_K$; $\beta = \beta_K$
LLIK(Y) = $\boldsymbol{L}(\boldsymbol{Y}|\mu, \sigma^2, \mu, \alpha, \beta, w)$
MLE(Y) ← OPTIM ($\boldsymbol{L}(\boldsymbol{Y}|\mu, \sigma^2, \alpha, \beta, w)$); (Iterate MLE to converge)
IF minimum is not achieved **THEN END; ELSE** Return LLIK(Y); estimate score of $Y = (\mu_T, s_T)$ **END IF**
END WHILE

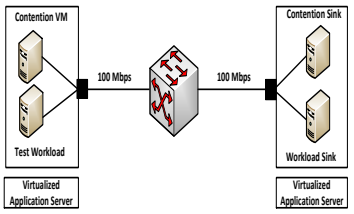

Figure 2a: Algorithm pseudo-code 2b: Test-bed

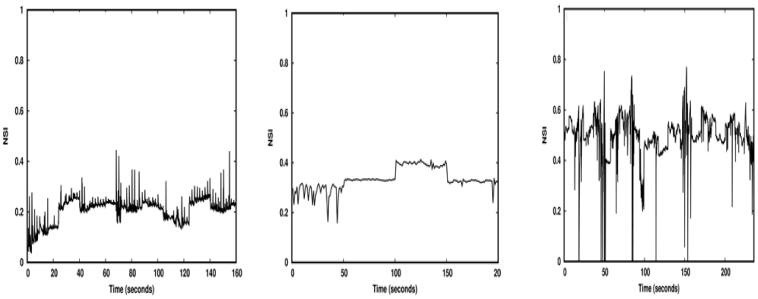

Figure 3(a) IPERF NSI (b) Google NSI (c) Dropbox NSI

adapt in provisioning scenarios. Figure 2a provides the pseudo-code of the predictive score-based algorithm. It takes as input the time-series for the cloud environment under study and computes the score along with the model parameters elaborated in section III. Through log-likelihood, the mean and score of the time-series is calculated. The computation of the score is dependent on the model selected for the time-series. As illustrated in Figure 1b which presents non-Gaussian departures from normality, the model selection process determined the lognormal distribution to provide the best fit for series II and IV. Thus the algorithm of Figure 2a illustrates the expression for the score ($s_t = \log Y_{t-1}^2 - \sigma_{t-1}^2$) and scaled score ($\sigma_t^2 = \omega + B\sigma_{t-1}^2 + \alpha \ln(y_{t-1}^2)$) for the lognormal distribution. The log-likelihood method also returns the mean value for the time-series. Subsequently the one-step-ahead forecast for the mean and the score are calculated. This enables the calculation of the instantaneous NSI for cloud use-cases.

To test the usefulness of the NSI in practical cloud scenarios, the test-bed of Figure 2b was realized. An identified workload type competes with a second workload identified as contention traffic over a 100 Mbps Ethernet link. A baseline experiment is conducted where contention over the 100 Mbps link is incremented from nominal utilization (40%) until it reaches capacity assignment (100%). For each experiment, the mean and score are calculated according to the algorithm of Figure 2a. Four workload types were experimented with: IPERF synthetic traffic, the RUBiS web workload benchmark, Google and Dropbox drive download traffic. Due to space constraints, only the NSI calculated at 80% utilization is reported. The static value of the NSI for each workload is: RUBiS (0.28), Google Drive (0.35), IPERF (0.21), & Dropbox Drive (0.49). Figures 3a to c illustrate the instantaneous NSI for IPERF, Google & Dropbox over the experimentation interval. Comparing the static values with the plots, the instantaneous NSI shown in Figure 3, for the IPERF workload, compares with the instantaneous plot which is within the range of 0.2 to 0.4. For the cloud-based workloads for the Google drive and Dropbox drive downloads, the NSI shows high burstiness for Dropbox for both the static value (0.49) as well as the instantaneous NSI. The NSI has been employed in provisioning for a rate-based algorithm used for QoS maintenance in virtualized cloud environments [14]. There, the methods employed are able to maintain QoS requirements for different applications in the shared cloud environment.

V. Conclusion

This paper introduces the Normalized Score Index (NSI) as a measure that quantifies the burstiness of cloud workloads according to observations of two time-varying properties: its mean and its scaled score, which extends variance beyond its traditional Gaussian definitions. The NSI compared to existing measures presents a more accurate range by which it can be expressed as a percentage [0, 100] or as a score [0, 1] in the definition of variance as applies to cloud workloads. The usefulness of the NSI as a measure of burstiness applies to diverse workloads. The NSI employs statistical features specific to each workload based on a selection methodology.

References

[1] M. Li, J. Bi, and Z. Li, "Virtual-Switching-Aware VM Consolidation in Virtualized Data Centers," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, 2014, pp. 817-822.

[2] C. Z. Loboz, "Cloud resource usage: extreme distributions invalidating traditional capacity planning models," presented at the Proceedings of the 2nd international workshop on Scientific cloud computing, San Jose, California, USA, 2011.

[3] G. Gon, x00E, alves, I. Drago, A. P. C. d. Silva, J. M. Almeida*, et al.*, "Characterizing and Modeling the Dropbox Workload," in *Computer Networks and Distributed Systems (SBRC), 2014 Brazilian Symposium on*, 2014, pp. 266-274.

[4] H. Li, "Realistic Workload Modeling and Its Performance Impacts in Large-Scale eScience Grids," *IEEE Transactions on Parallel and Distributed Systems,* vol. 21, pp. 480-493, 2010.

[5] C. Delimitrou, S. Sankar, A. Kansal, and C. Kozyrakis, "ECHO: Recreating network traffic maps for datacenters with tens of thousands of servers," in *2012 IEEE International Symposium on Workload Characterization (IISWC)*, 2012, pp. 14-24.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*: Wiley, 2006.

[7] A. Caniff, L. Lu, N. Mi, L. Cherkasova, and E. Smirni, "Fastrack for taming burstiness and saving power in multi-tiered systems," in *Teletraffic Congress (ITC), 2010 22nd International*, 2010, pp. 1-8.

[8] A. Ali-Eldin, O. Seleznjev, S. Sj, x00F, L. stedt-de, J. Tordsson*, et al.*, "Measuring Cloud Workload Burstiness," in *Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on*, 2014, pp. 566-572.

[9] R. Yanggratoke, J. Ahmed, J. Ardelius, C. Flinta, A. Johnsson, D. Gillblad*, et al.*, "Predicting service metrics for cluster-based services using real-time analytics," in *Network and Service Management (CNSM), 2015 11th International Conference on*, 2015, pp. 135-143.

[10] S. Siqi, V. van Beek, and A. Iosup, "Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters," in *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on*, 2015, pp. 465-474.

[11] R. Gracia-Tinedo, M. S. Artigas, A. Moreno-Mart, x00Ed, nez, C. Cotes*, et al.*, "Actively Measuring Personal Cloud Storage," in *2013 IEEE Sixth International Conference on Cloud Computing*, 2013, pp. 301-308.

[12] A. C. Harvey, *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*: Cambridge University Press, 2013.

[13] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 69, pp. 243-268, 2007.

[14] A. Adegboyega, "An Adaptive Reservation Scheme for Effective QoS Provisioning in the Cloud SDN Stack," in *Computer Communication and Networks (ICCCN), 2015 24th International Conference on*, 2015, pp. 1-6.