Paper 206-28

Spectral Decomposition of Performance Variables for Dynamic System Characterization of Web Servers

K. C. Gross, W. Lu, and K. Mishra
RAS Computer Analysis Laboratory
Sun Microsystems, Inc., San Diego, CA

ABSTRACT

Using SAS® software for data analysis and presentation, a study was undertaken in which subtle, multifrequency sinusoidal perturbations were superimposed on the normal chaotic workloads for a large eCommerce web server. The system developed in this study, which uses SAS as its foundation software, is a novel sinusoidal-impulsion technique that enables accurate dynamical system characterization of huge datacenter web servers, producing a wealth of diagnostic performance information such as transfer functions, coupling coefficients, and lead-lag phase relationships between system resource and performance metrics**.** SAS macros that implement PROC SPECTRA univariate and bivariate spectral decomposition techniques were employed to analyze time series performance metrics captured by a customized system telemetry harness. Digitized performance variables with disparate sampling rates and formats were aligned, preprocessed, and analytically resampled using SAS macros. Cross correlation analyses were carried out using PROC CORR to eliminate non-correlated and duplicate system performance variables, resulting in a set of highly correlated variables for continuous health-monitoring. The system presented here has been demonstrated in empirical studies to provide valuable dynamical system characterization information that can be indispensable to datacenter architects for the functions of performance management, capacity planning, quality-of-service (QoS) assurance, resource provisioning, and root cause analyses.

INTRODUCTION

An e-Commerce web server is a complex system with hundreds of resource, performance, and throughput parameters, making the study of relationships among variables quite difficult using traditional "static analysis" approaches. Currently, system performance for servers is characterized by testing its operation under maximum load, random load, and using performance benchmarks that mimic typical user loads. These conventional approaches are not able to fully characterize transfer function relationships among performance variables to answer vital questions for data center architects and capacity planners, such as: What effects do my 300 email users have on 50 users FTPing files; What effect do my 200 users with web browser banners have on 400 users doing online transaction processing? Or, in general, what effect does system demand vector X have on system response vector Y? To determine such vital cause/effect relationships using conventional approaches, one could introduce a large step-function perturbation in demand vector X, and measure the QoS responses in a variety of response vectors of interest. However, if such experiments were to be conducted during times of high user activities, the step-function would have to be quite large to infer accurate coupling coefficients due to poor signal-to-noise relationships. Such maneuvers would likely cause system overload events, and would certainly interfere with the normal day-to-day operation of the system one is seeking to characterize (a macroscopic variant to Heisenberg's Uncertainty Principle).

Dynamic characterization of complex systems such as web servers can be achieved by introducing perturbations in one or more "input" variables, and measuring the time-dependent responses in one or more "response" variables. We quantify this relationship between input variables and response variables with a "dynamic coupling coefficient," which may be a function of load, or, more generally, a multivariate function of multiple input variables. In a dynamically changing system such as a web server, distributed synthetic transaction generators can be employed for real-time continuous monitoring of system transaction latencies. These "canary tests" provide QoS performance metrics on a 24/7 basis as a dynamical function of system load. Specifically, in order to measure the impact of some performance parameter X on another performance parameter Y, the synthetic transactions introduce an (ideally small) perturbation in X, from which the resulting effect on parameter Y, if any, can be measured.

A new technique for dynamical system characterization of large, multi-processor web servers has been adapted from a technique developed in the 1980's for dynamic system characterization of chaotic, nonlinearly interacting physics variables in nuclear power plants [1,2]. The *M*ultifrequency *S*inusoid Tool for *S*ervice-*L*evel *A*vailability *M*onitoring, MS-SLAM, superimposes an envelope of multifrequency periodic perturbations with very small amplitudes on X, and the spectral decomposition of resulting responses in a wide range of performance and QoS metrics Y(i), i = 1,2,3,...N, determines the coupling coefficients among all elements of the system, thereby characterizing the dynamical system behavior [3]. Consequently, the technique of superimposing small sinusoidal perturbations on normal load is an elegant and powerful method with which to characterize systems, particularly the relationships among various system parameters.

ANALYTICAL APPROACH

Performance metrics related to load on the system CPUs, memory, cache, queue lengths, transaction latencies, and I/O parameters are collected over a sequence of experiments conducted while the servers are fully loaded with dynamic user transactions. In addition to this workload, very subtle multifrequency periodic variations are introduced into the system that are sinusoidal with respect to time. For the present investigation, three sinusoidal perturbations with periods of 20, 15, and 12 minutes are simultaneously generated in multiple input variables. All three of the injected sinusoids have very small amplitudes in comparison with normal variations in user load patterns (typically < 1% of nominal variations). One or several synthetic clients (called the 'canary' variables) are launched to commit typical user transactions. The response times for these canary tests are recorded to produce continuous time series that reflect QoS from the end-user perspective. In parallel with the canary tests, a large suite of system performance, throughput, and transaction latency variables are recorded on a 24/7 basis using the comprehensive system-monitoring tool, SEToolkit [4].

EXPERIMENTAL DATA & PRE-PROCESSING

The foregoing system telemetry harness produces a plethora of performance variables to monitor. Their number is restricted to about 150 by eliminating those that are either very poorly correlated, or those that are redundant (e.g. variables that may report the same metric, but with different units). Detailed cross-correlation and coherence analyses are made use of in this step using **PROC CORR** (a one-time sensitivity analysis conducted before applying MS-SLAM to a new system configuration). Using the **Base SAS** Software, the various date and time formats of data from all the sources are converted to a common format: the number of

seconds from Jan. 1, 1960. The **SAS** Macros for manipulating time series data are then utilized for data management. Since sampling of the various performance metrics is done at disparate (and sometimes time-varying) rates, an analytical resampling algorithm using **SAS** Software is invoked to align all the parameters and bring them to a uniform, synchronized sampling rate. Specifically, an empty time series with the desired interval is generated and merged with the actual data using the statement **MERGE**, thereby creating missing values. The time series thus generated is sorted by calling **PROC SORT.** The sorted data with missing values are interpolated to fill in the missing values by using **PROC EXPAND**, which is a part of the **SAS/ETS** software [5].

Once the data are made uniform, we can then plot the time series of the three sinusoids and the canary variable. Figure 1 illustrates data collected during a typical experiment. The figure was produced using **PROC GPLOT** in the **SAS/GRAPH** system with the **ANNOTATE** option. The three-plot format was created by calling the procedure **PROC GREPLAY,** which allowed the three sinusoids and the canary variable to be displayed in a single graph [6].

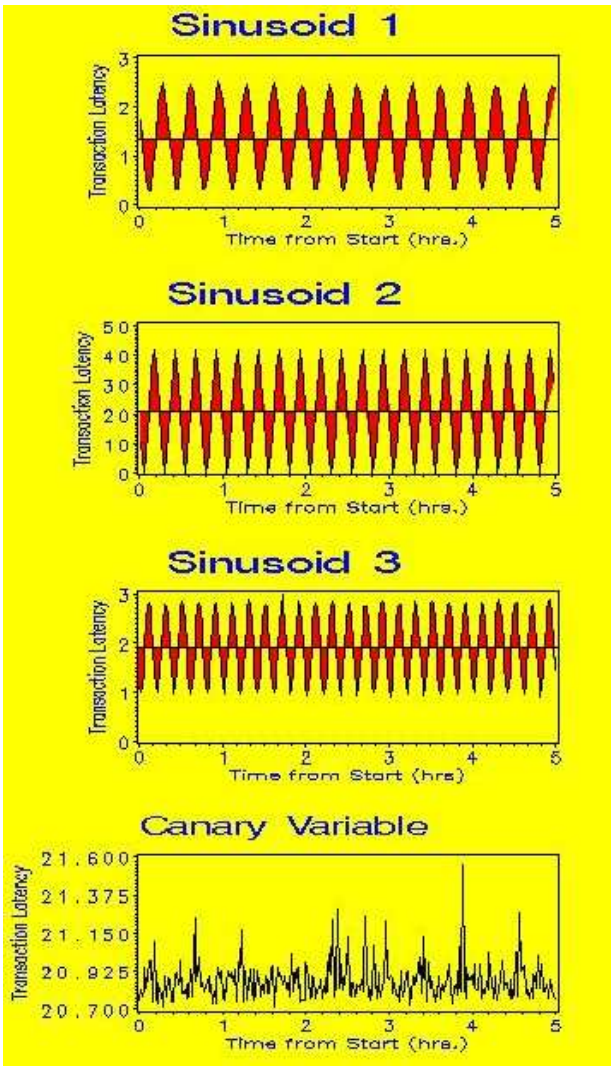

Figure 1. Time Series of the Three Sinusoids and the Canary Variable

UNIVARIATE AND BIVARIATE SPECTRAL DECOMPOSITION ANALYSIS

To determine dynamic coupling coefficients among system variables, univariate spectral decomposition analysis of the preprocessed data was performed using **PROC SPECTRA**, part of the SAS/ETS software. Output from **PROC SPECTRA** for univariate analyses is provided in the form of periodograms and spectrograms, which plot the power spectral density (PSD) for the time series as a function of period, P, or frequency, w, respectively. A physical interpretation of the PSD function is that f(w)dw represents the contribution to variance of components within the frequency range (w,w+dw). When the spectrum is plotted, the total area under the curve is equal to the variance of the time series. A peak in the spectrum indicates an important contribution to variance at frequencies in the appropriate region [7,8].

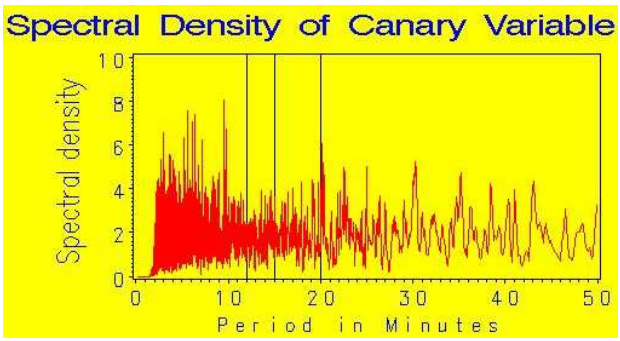

Figure 2. Spectral Density of the Canary Variable

Figure 2 plots the PSD of the response time of the synthetic client (canary variable) as a function of the period, in minutes. The multitude of peaks in the plot, resulting from chaotic user load, makes the peaks at the period of the sinusoids less prominent. The stochastic noise (chaotic user load) associated with the canary signal is so high that the test periods of 20, 15, and 12 minutes are not discernable in the univariate PSD. This illustrates that characterization of typical web-server performance metrics is not amenable to univariate spectral decomposition calculations via conventional Fourier analysis because the signal-to-noise ratio is too small to discern the sinusoidal perturbation in the response variables.

To overcome this limitation of conventional Fourier analysis methods, we employ in MS-SLAM the Normalized Cross Power Spectral Density (NCPSD) technique, a bivariate diagnostic technique that is highly sensitive, even to weakly coupled parameters with very poor signal-to-noise ratios. The NCPSD technique dramatically and selectively amplifies the input sinusoid harmonics in response variables such that the period of the sinusoidal perturbation in the control variable is readily apparent with excellent peak resolution, with low noise and no "side lobe" contamination [8,9]. The presence of a peak in the NCPSD is evidence of a common periodicity and hence a cause-and-effect relationship between the sinusoidal perturbations in the load and the system variables as well as the canary variable.

Bivariate spectral decomposition analyses were carried out by employing **PROC SPECTRA** with the **CROSS** option. In the bivariate analysis, the cross power spectral density function between two time series is computed, which shows whether the frequency components in one series are associated with large or small amplitudes at the same frequency in the other series. The presence of a peak at any frequency indicates this association, and the amplitude of the peak is a measure of the coupling coefficient between the two series.

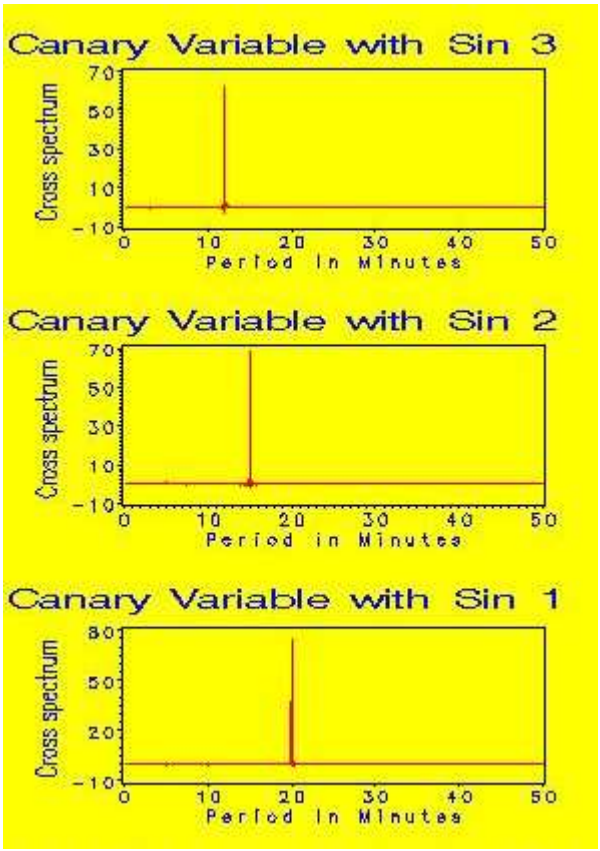

Figure 3. Cross-Spectrum Estimates of the Canary Variable with Each of the Sinusoids

Figure 3 plots the cross spectrum estimate of the canary variable with each of the three sinusoids. In each subplot, a well-defined peak corresponding to the period of the sinusoid is readily observable, implying a common periodicity and hence a cause-and-effect relationship between the sinusoidal perturbations in the load and the

synthetic client's response time. In other words, the periodicity in the perturbations introduced a corresponding periodicity in the canary variable. The univariate spectral decomposition analysis could not conclusively identify the presence of these perturbations (though there were peaks at those periods), due to the presence of a number of other phenomena and large random noise components in that frequency range.

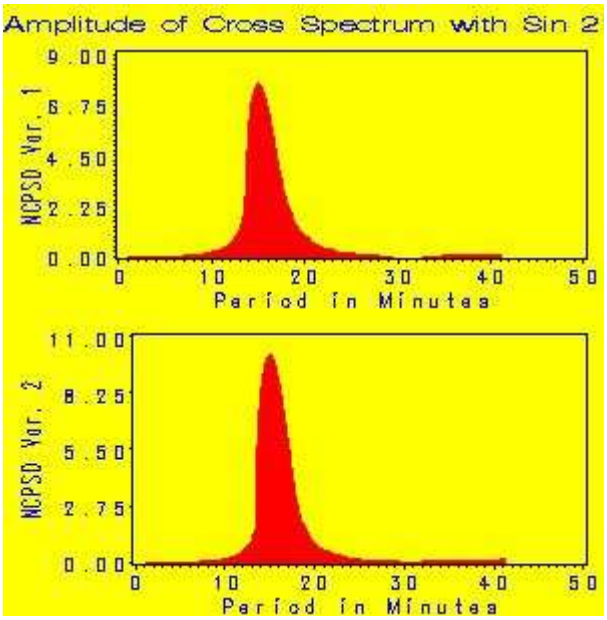

Figure 4. Amplitude of Cross Spectrum Estimates of Two System Variables with Sinusoid of Period = 12 Minutes

Figure 4 shows the amplitude of the smoothed cross spectrum estimate of two of the system variables with one of the sinusoids (period = 12 minutes). A peak at 12 minutes illustrates an induced periodicity in the variables. The amplitude of the cross spectrum estimate is a measure of how closely the variable is coupled with the input perturbations. Separate output metrics from the NCPSD computation give the phase lags between variables. To have a common datum for comparing the amplitudes of the cross spectrum estimates, all the parameters in consideration are first normalized to have both mean and variance as unity, using **PROC STANDARD**.

CONCLUSION

To achieve accurate characterization of the inter-dependencies between and among service level QOS metrics in a dynamically changing system, distributed synthetic transactions that simulate user response times can be generated for real-time continuous monitoring of system parameters. These "canary tests" of external variables, such as user transaction latency times, reflect user-facing performance metrics and system stress as a dynamical function of system load. Specifically, in order to measure the impact of some performance parameter X on another performance parameter Y, the synthetic transactions introduce a perturbation in X, from which the resulting effect on parameter Y, if any, can be measured. The multi-frequency sinusoidal excitation approach superimposes a very subtle periodic impulsional perturbation in X, and the bivariate spectral decomposition of resulting responses in other performance variables Y(i), i = 1,2,3,...N, determines the coupling coefficients among all elements of the system, thereby characterizing the dynamical system behavior.

Most current methods for qualification testing of enterprise computing systems involve putting a maximum expected load on one or more input variables, and seeing if the system crashes. While this type of qualification testing is necessary, we have found that dynamical response testing can provide a far greater wealth of information that is useful for designing robust systems that deliver optimal QoS metrics over a large range of system performance. The MS-SLAM tool described in this paper can determine dynamic coupling coefficients, transfer functions, and phase relationships among a wide range of throughput and performance variables. Experiments documented herein with large, multi-processor web servers have demonstrated that complex systems comprising chaotic performance dynamics and high noise levels are not amenable to univariate spectral decomposition calculations via conventional Fourier analy-

sis because the signal to noise ratio is too small. The NCPSD technique introduced in this paper, which has been adapted from dynamical system characterization of reactor physics variables, can accurately assess cross correlation and coherence relationships among multiple, dynamic system parameters, even those characterized with extremely poor signal to noise ratios. In addition to cross correlation among response variables, the technique introduced here also provides dynamical coupling coefficients between input variables and response variables, and phase shifts (which can be expressed differently to give lag times) between "cause" and "effect" variables throughout the dynamic system.

REFERENCES

[1] K. C. Gross and L. K. Polley. "Investigation of Nonrecoil Fission-Product Release Phenomena Using Multifrequency Source-Perturbation Experiments in EBR-II," *Annals of Nuclear Energy*, Vol. 18 (3), pp. 419-441, 1986.

[2] K. C. Gross and L. K. Drenth. "Response of EBR-II's Delayed Neutron Monitoring Systems During Sinusoidal Reactivity-Oscillation Experiments," *Trans. of the Amer. Nuclear Soc.*, Vol. 52, 1986

[3] K.C. Gross and R. V. Strain. "Sinusoidal Source Perturbation Experiments With a Breached Fuel Subassembly in the Experimental Breeder Reactor II," *Nuclear Technology Journal*, Vol. 98, April 1992.

[4] A. Cockroft and R. Pettit, http://www.setoolkit.com/

[5] SAS/ETS User's Guide Version 8, SAS Institute, Inc., Cary, NC, 1999.

[6] SAS/GRAPH User's Guide Version 8, SAS Institute, Inc., Cary, NC, 1999.

[7] K. C. Gross, H. P. Planchon and J. Poloncsik. "Time Series Investigation of Anomalous Thermocouple Responses in a Liquid-Metal-Cooled Reactor," *Proc. SAS Users Group Int'l Conf.* pp. 732-735, Orlando, April 1988.

[8] K. C. Gross, W. Lu, and D. Huang. "Time-Series Investigation of Anomalous CRC Error Patterns in Fibre Channel Arbitrated Loops," *Proc. 2002 IEEE Int'l Conf. on Machine Learning and Applications (ICMLA),* Las Vegas, NV, June 2002.

[9] G. M. Jenkins and D. G. Watts. *Spectral Analysis and its Applications*, Emerson-Adams Press, 2000.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kenny C. Gross, Ph.D.
RAS Computer Analysis Laboratory
Sun Microsystems, Inc.
9525 Towne Centre Drive, USAN10-103
San Diego, CA 92121
Kenny.Gross@sun.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.