Paper 286

Autoregressive Integrated Moving Average Models for Comparing Forecasted to Actual Values of CPU Workloads for Open Systems

Anthony C. Waclawski, Ph.D.
MCI WorldCom

ABSTRACT

The marriage of MCI and WorldCom has created an enormous worldwide computing network with over one billion dollars in computing assets. The MCI WorldCom computing environment includes several very large central electronic complexes containing arrays of central and distributed computers. The entire network is remotely managed from Colorado Springs, Colorado. It serves thousands of distributed interactive users at hundreds of locations throughout the world. To effectively manage these corporate assets, decision-makers need accurate forecasts of midrange workload performance in order to justify requests for acquisition of new central processing units.

This paper describes our use of the Autoregressive Integrated Moving Average (ARIMA) technique to accurately forecast workload consumption of midrange resources with 95% statistical confidence. In order to demonstrate the robustness of the ARIMA construct, and to facilitate calibration of the models, we graphically illustrate how they are used to compare forecasted to actual values of CPU consumption.

EXECUTIVE SUMMARY

For the past twelve months the MCI WorldCom Midrange Capacity Planning organization in Colorado Springs, Colorado has conducted an ongoing scientific evaluation of the ability of the Autoregressive Integrated Moving Average technique to accurately forecast midrange workload consumption. Our research philosophy was predicated on the intrinsic value of time series data as an aid in policy formulation as well as our confidence in the epistemological validity of ARIMA models as viable forecasting tools. The purpose of this study was to develop, test, and apply an ARIMA regression equation to forecast CPU consumption over time.

One of the salient issues relative to the development of useful time series models are irregular fluctuations in the underlying data caused by random effects. At MCI WorldCom, data irregularities can be a major statistical concern due to the rapidly evolving nature of the computing environment. For example, during the past twelve months there have been hundreds of configuration changes, processor moves, replacements, and/or upgrades. This evolutionary churn in the configuration of the computing environment has facilitated the implementation of a powerful computing infrastructure whose data centers are currently rated as the industry leader by both the Meta and

Gartner groups. Although these numerous network changes are driven by the extremely competitive nature of the telecommunications industry and by MCI WorldComs corporate commitment to maintain a “state of the art” computing environment, they greatly complicate the process of constructing useful probability algorithms.

In order to determine the universe of performance data for the application platform analyzed by this document, and to improve the explanatory power of the experimental design, univariate statistical analysis was performed on relevant performance data from 35 UNIX production nodes for the past twelve months. These data show that the mean weekly workload demand on these systems causes UNIX central processing units (CPUs) to operate within a range of 16% to 43% of capacity. This information was used as input to several iterations of autoregressive models.

The research results indicate that a stochastic order 1 mixed autoregressive model with a moving average parameter accurately forecast’s midrange CPU consumption with 95% accuracy. Moreover, the results suggest that time series modeling in general, and ARIMA forecasting in particular, can be useful planning tools for projecting CPU utilization for mainframe (legacy) platforms.

PROCESS and METHODOLOGY

The BMC Patrol UNIX system monitor collects data on central processing unit performance at one-minute intervals. These records are similar to the performance metering data obtained by “real time” hardware monitors. They contain measurements of central processing unit activity and active process metrics. These data, as well as thousands of other statistical descriptors of midrange performance and resource utilization, are stored in a SAS IT Service Vision ® performance database. Presently, IT Service Vision processes, statistically collapses, and archives data from hundreds of client (i.e. customer) machines. In order to reduce storage requirements, IT Service Vision statistically collapses all performance data by automatically calculating the mean for all numeric-metering records at fifteen-minute intervals before archiving the information. This information was used as source data to evaluate the ability of the Autoregressive Integrated Moving Average technique to accurately forecast workload consumption.

The purpose of this study was to develop, test, and apply an ARIMA regression equation, using UNIX central processing unit metering records. The specific research objective was to determine if the percentage a given CPU is busy processing work (the criterion variable) is an accurate predictor of future CPU consumption based its values obtained earlier in time.

INSTRUMENTATION

In order to construct univariate probability models of CPU consumption, accurate historical data on resource consumption needed to be obtained. The

programming and mechanical machinations required to obtain historically accurate UNIX descriptors of CPU consumption are straightforward if one uses our patented META Collector technology. To illustrate, raw CPU workload performance data is obtained from BMC Patrol UNIX system monitor's located on hundreds of IBM, Hewlett-Packard, DEC, and SUN client (i.e. sending) machines. These monitors continuously collect "raw" performance metering data (i.e. ASCII files) and temporarily stores them in a "local directory" until they are transmitted to a host machine at regular intervals using the native UNIX scheduler.

From a philosophical, financial, and operational perspective MCI WorldCom, like many companies with very large heterogeneous computing environments, cannot always expend the resources required to support the plethora of proprietary collectors coterminous with the various flavors of UNIX in order to perform scientific capacity planning. Our solution to this political and engineering conundrum was to write and patent our own "META Collector." This collector is machine and vendor independent and completely absolves users from the IT Service Vision requirement that they configure a particular manufacturers proprietary collector to process performance data before insertion into Service Vision. The META Collector is software that takes standard CSV (i.e. comma separated files) data feeds of metering records from any machine running any flavor of UNIX and prepares it for insertion into IT Service Vision or any other commercial data base product (i.e. MICS, ORACLE, SYBASE, INFORMIX et al). The META Collector's batch jobs perform all necessary variable mapping, table construction, and statistical calibration so that the data can be "seamlessly" inserted into a data warehouse. Moreover, as part of the preparatory or staging process the META Collector code transposes (i.e. converts variables to observations and observations to variables) and statistically summarizes the raw data according to the date time stamp of each metering record received. By transposing or "rectangularizing" the data in this manner, it becomes more useful to users who want to perform statistical analysis as well as empowering analysts with the ability to focus on time granularity down to one-one thousand of a second by node, application, instance, and/or parameter. From a performance standpoint, the META Collector reduces the volume of raw/detail data stored in the performance database by approximately 66%.

In order to forecast CPU consumption for the customer SAS code was written to extract the relevant performance records from IT Service Vision. These records include the node or individual machine identifier, the processor busy variable, and the date time stamp attached to the metric. Finally, in order to prepare the data for input into the modeling algorithms, it was statistically collapsed into one record for each week and written to an external SAS data set.

The theoretical constructs tested in this study, were a series of ARIMA time series modeling algorithms that were modified to use information technology data. The usefulness of this approach to the study and development of time series models was originally advanced by G.E.P. Box and G.C. Tiao

in their 1975 *Journal of the American Statistical Association* article, "Intervention Analysis with Applications to Economic and Environmental Problems." Box and Tiao correctly suggest the following general strategy for model development:

1. Frame a model for change that describes what is expected to occur given knowledge of a known intervention;
2. Work out the appropriate data analysis based on that model;
3. If diagnostic checks show no inadequacy in the model, make appropriate inferences; if serious deficiencies are uncovered make appropriate model modifications and repeat the process.

ANALYSIS

Actual historical performance data from one of MCI WorldCom's Midrange Planning customers were used for the models presented in this paper. The number of raw data points used to determine CPU consumption was 18,396,000 (i.e. 60 data points per hour 24 hours per day x 365 days x 35 machines). Next, univariate statistical procedures were used to mathematically collapse these data into one data point per week. Extracting, post-processing and presenting data in this manner has the advantage of producing statistically valid, reliable and easily understood measures of central tendency.

Average CPU consumption for the 35 machines used in this study ranges from 16% to 43% for the weeks studied. In order to probe the nature of the relationship between the CPU consumption time series a number of inferential statistical techniques were used. The first step in developing useful time series models is to determine if the time series being analyzed is statistically stationary. This determination is typically made by analyzing the probability values for the *Chi-Square* tests as well as by reviewing the raw autocorrelation and covariance statistics. At MCI WorldCom, these tests indicate a statistically significant correlation ($p <= .0005$) between CPU consumption and the value for CPU consumption for the previous period. That is, since the expected values for the time series and its autocovariance function (*white noise*) are not independent of time, we cannot reject the hypothesis that the residuals are not correlated. Because these tests confirm that the CPU consumption data, like most other time series, are nonstationary, they were statistically transformed into a stationary series by differencing. That is, instead of modeling the CPU consumption series itself, we model the statistically *differenced values* of CPU consumption from one period to the next.

After converting the time series from nonstationary to stationary, we identified a number of potential ARIMA constructs. As part of the model identification phase autocorrelation, inverse autocorrelation, and partial autocorrelation functions were calculated and compared with theoretical correlation functions expected from different kinds of ARIMA models.

These tests indicated that a series of order 1 mixed autoregressive models with moving average parameters accurately forecast CPU consumption. This "fitting of the data," or matching the theoretical autocorrelation functions of different autoregressive modeling constructs to the autocorrelation functions computed from the response series is the heart of autoregressive modeling. The ARIMA (1, 1, 1) models suggested by these analyses predicts CPU consumption as a function of mean CPU consumption over time, plus a moving average parameter estimate for CPU consumption divided by the coefficient of the lagged autoregressive value of CPU consumption and its estimated value, plus a random error. An ARIMA (1, 1, 1) model for the *level* of CPU consumption is the same as an ARMA (1, 1) model for the *change* in CPU consumption. An analysis of the *Conditional Least Squares* regression estimates in Table 2 shows that both the moving average parameters (labeled MA1, 1) and the autoregressive parameters (labeled AR1, 1) have significant *t-ratios*. Moreover, the *Chi-Square* tests for the white noise residuals (see Table 3) show that we cannot reject the hypothesis that the residuals are not correlated. Therefore, we conclude that an ARIMA (1, 1, 1) model is an adequate predictor for the CPU consumption time series, and there is no useful purpose served by trying models that are more complex.

CONCLUSIONS

The research results indicate that stochastic autoregressive models can be useful capacity planning tools within the context of the MCI WorldCom midrange system-computing environment. A series of order 1 mixed autoregressive models with moving average parameters accurately forecast midrange CPU consumption with 95% accuracy. In contradistinction to causal models, which require accurate and often difficult to obtain confidential information relative to future strategic business initiatives, autoregressive constructs rely on electronically generated empirical data (CPU metering records) for model input. This suggests that autoregressive models should be employed in tandem with causal models in order to validate forecasts. Moreover, the results also suggest that time series modeling in general, and ARIMA forecasting in particular, can be useful planning tools for projecting CPU utilization for mainframe platforms.

Table 1

Comparison of Actual to Forecasted Values of Maximum Percent CPU Busy with 95% Confidence Limits for the XXXXXXX Application

Date	Median % CPU Busy	Forecast	Lower 95% Confidence Bound	Upper 95% Confidence Bound
12APR98	28.09	42.21	31.49	52.58
19APR98	23.77	35.87	25.80	45.94
26APR98	27.54	29.03	18.96	39.10
03MAY98	26.66	28.12	18.06	38.19
10MAY98	28.91	28.16	18.09	38.22
17MAY98	31.07	32.10	22.03	42.17
24MAY98	29.53	31.80	21.73	41.87
31MAY98	28.13	30.05	19.98	40.12
07JUN98	23.34	26.47	16.40	36.54
14JUN98	25.28	27.22	17.15	37.29
21JUN98	22.42	26.81	16.74	36.88
28JUN98	17.43	21.67	11.60	31.74
05JUL98	17.59	22.18	12.11	32.25
12JUL98	26.16	24.61	14.54	34.68
19JUL98	42.17	40.78	30.71	50.84
26JUL98	39.04	40.85	30.78	50.92
02AUG98	35.09	35.62	25.55	45.69
09AUG98	38.37	39.17	29.11	49.24
16AUG98	32.73	33.11	23.04	43.18
23AUG98	39.95	39.95	29.88	50.02
30AUG98	37.99	39.73	29.66	49.79
06SEP98	38.58	40.23	30.16	49.30
13SEP98	36.09	39.71	29.64	49.77
20SEP98	32.75	35.03	24.96	45.10
27SEP98	27.01	28.82	18.75	38.89
04OCT98	26.14	28.83	18.76	38.90
11OCT98	24.92	29.14	19.07	39.21
18OCT98	26.89	29.66	19.59	39.72
25OCT98	30.13	32.09	22.02	42.16
01NOV98	29.94	33.24	23.17	43.31
08NOV98	31.62	33.66	23.59	43.73
15NOV98	30.74	33.95	23.88	44.02
22NOV98	28.99	34.03	23.96	44.10
29NOV98	30.58	32.58	22.51	42.65
06DEC98	34.62	37.22	27.15	47.29
13DEC98	33.17	36.74	26.67	46.81

20DEC98	30.39	34.85	24.78	44.92
27DEC98	31.49	35.85	25.78	45.92
03JAN99	31.48	34.06	23.99	44.13
10JAN99	.	33.71	19.06	48.36
17JAN99	.	33.42	15.26	51.57
24JAN99	.	33.03	12.14	53.92
31JAN99	.	32.64	9.34	55.94
07FEB99	.	32.25	6.76	57.73
14FEB99	.	31.86	4.36	59.35
21FEB99	.	31.46	2.10	60.83
28FEB99	.	31.07	0.23	62.20
07MAR99	.	30.68	0.00	63.48
14MAR99	.	30.29	0.00	64.67
21MAR99	.	29.90	0.00	65.79
28MAR99	.	29.51	0.00	66.86
04APR99	.	29.12	0.00	67.87
11APR99	.	28.73	0.00	68.83
18APR99	.	28.34	0.00	69.74
25APR99	.	27.95	0.00	70.62
02MAY99	.	27.56	0.00	71.46
09MAY99	.	27.17	0.00	72.26
16MAY99	.	26.81	0.00	72.98

Table 2

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	T Ratio	Lag
MU	-0.07192	0.12181	-0.59	0
MA1,1	0.80654	0.07024	11.48	1
AR1,1	0.50225	0.10271	4.89	1

Constant Estimate = -0.035797

Table 3

Chi-Square Values for the Autocorrelation Check of Residuals

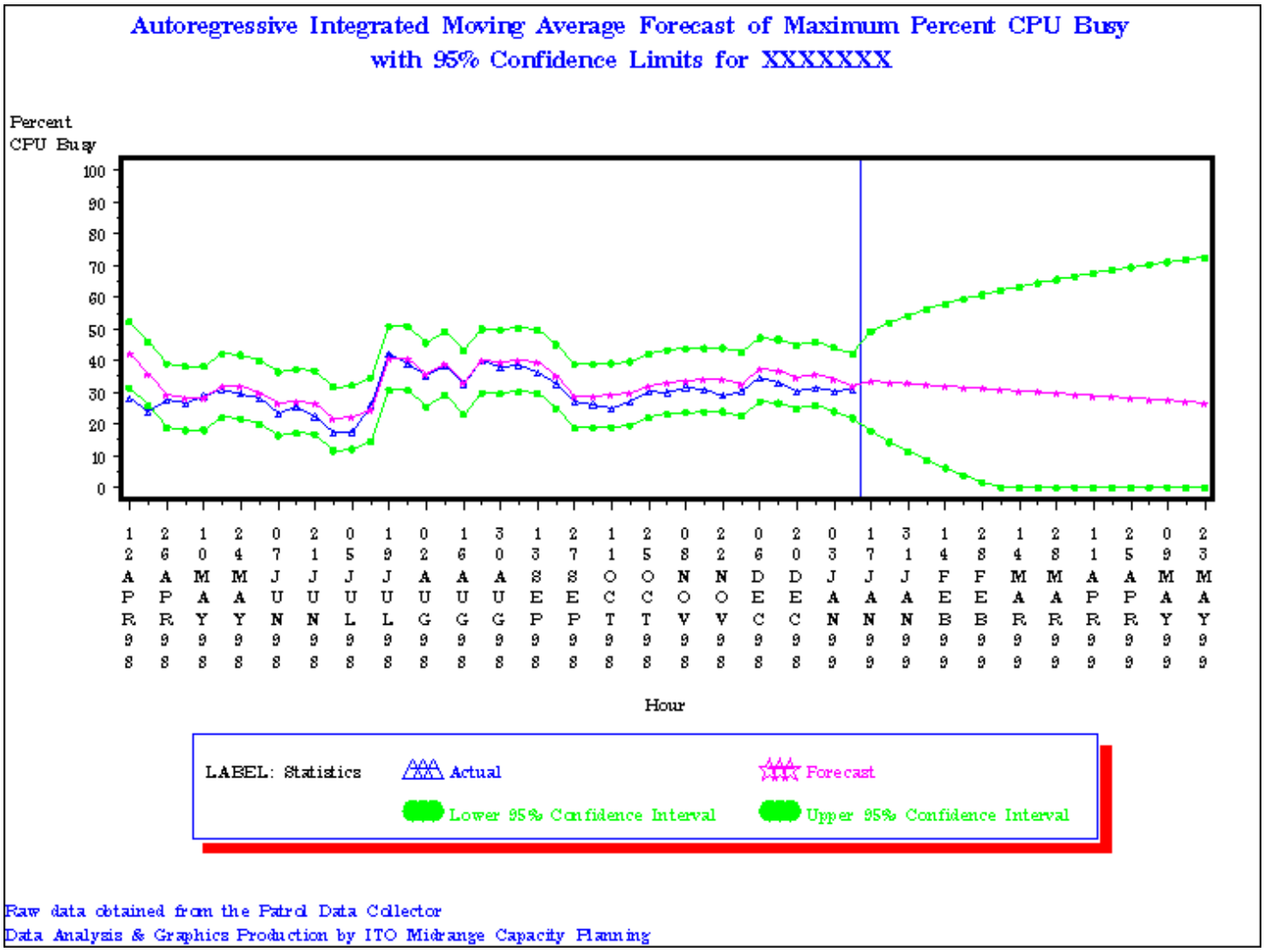
To Lag	Chi Square	DF	Probability	Autocorrelations					
6	12.00	4	0.017	.048	- .080	- .028	.022	.160	.084
12	31.05	10	0.001	.203	.103	- .096	- .008	.041	- .063
18	39.90	16	0.001	-.008	.134	.027	- .069	- .081	- .001
24	49.16	22	0.001	-.052	.011	.072	.050	-.142	- .008
30	52.71	28	0.003	.047	- .054	.005	.012	- .051	- .060
36	58.22	34	0.006	-.093	.042	- .021	.028	.069	.035
42	60.85	40	0.018	-.050	.047	- .042	- .036	- .015	- .015
48	63.71	46	0.043	.054	.026	.027	.048	- .006	- .045
54	65.87	52	0.094	-.019	- .021	- .011	.064	.035	- .008
60	67.98	58	0.174	-.004	- .012	- .020	- .020	.070	.010

Table 4

Autoregressive Moving Average Model Notation for the XXXXXXX Application

$$(1 - B)\,\text{cpubusy}_t = -0.07192 + \frac{(1 + 0.80654\,B)}{(1 - 0.50225\,B)}\,a_t$$

Figure 1


Autoregressive Integrated Moving Average Forecast of Maximum Percent CPU Busy
with 95% Confidence Limits for XXXXXXX
Percent
CPU Busy
100
90
80
70
60
50
40
30
20
10
0
Hour
LABEL: Statistics
Actual
Forecast
Lower 95% Confidence Interval
Upper 95% Confidence Interval
Raw data obtained from the Patrol Data Collector
Data Analysis & Graphics Production by ITO Midrange Capacity Planning

REFERENCES

Beyer, William. H. ed., (1976) *Handbook of Tables for Probability and Statistics*, Cleveland: CRC Press.

Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*, San Franciso: Holden Day.

Box, G.E.P. and Tiao, G.C. (1975), "Intervention Analysis with Applications to Economic and Environmental Problems," JASA, 70, 70-79.

Rrocklebank, J.C. and Dickey, D.A. (1986) *SAS System for Forecasting Time Series, 1986 Edition*, Cary, North Carolina: SAS Institute Inc.

Harvey, A.C. (1981), *Time Series Models*, New York: John Wiley & Sons, Inc.

Jones, Richard H. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations," *Technometrics, 22, 389-396.*

Kachigan, S.H. (1983), *Statistical Analysis*, New York: Radius Press.

Merrill, H.W. (1984) *Merrill's Expanded Guide to Computer Performance Evaluation Using the SAS System*, Cary, NC, USA: SAS Institute Inc.

The author may be contacted at:

Anthony C. Waclawski, Ph.D.
MCI WorldCom
2424 Garden of the Gods Rd.
Colorado Springs, Co. 80919 USA
(719) 535-1721
E-mail: anthony.c.waclawski@mci.com