Resource Utilization Prediction: A Proposal for Information Technology Research

Daniel W. Yoas
Nova Southeastern University
Pennsylvania College of Technology
Williamsport, PA 17701
yoas@nova.edu; dyoas@pct.edu

Greg Simco
Graduate School of Computer and Information Sciences
Nova Southeastern University
Ft. Lauderdale, FL 33314
greg@nova.edu

ABSTRACT
Research into predicting long-term resource needs has been faced with a very difficult problem of extending the accuracy period beyond the immediate future. Business forecasting has overcome this limitation by successfully incorporating the concept of human interaction as the basis of prediction patterns at the hourly, daily, weekly, monthly, and yearly time frames. Computer resource utilization is also impacted by human interaction therefore influencing research into predictability of resource usage based on human access patterns. Emulated human web server access data was captured in a feasibility study that used time series analysis to predict future resource usage. For prediction beyond several minutes, results indicate that the majority of projected resource usage was within an 80% confidence level thus supporting the foundation of future resource prediction work in this area.

Categories and Subject Descriptors
D.4.8 [**Performance**]: Measurements, Modeling and Prediction, Monitors;

General Terms
Design, Experimentation, Human Factors, Management, Measurement, Performance, Reliability, Security, Theory.

Keywords
Prediction methods, Demand forecasting.

1. INTRODUCTION
Researchers have been developing methods to provide more accurate computer resource predictability to support improvements in scheduling processes, managing disk IO, and quality of service [1, 5, 6, 8, 9]. This desire to find a more effective method to manage resources has led researchers to explore the predictability of resource usage patterns. Near-term utilization prediction solutions have often provided improvements over previous methodologies [1, 5, 7, 16, 18, 19]. However, the farther into the future a methodology attempts to predict utilization, the less accurate it becomes. Thus research in computer system usage has focused on recent history to predict the near future utilization of the resource. However, business prediction models focus on longer time frames.

The domain of business trend prediction has used resource utilization forecasting to help determine just-in-time manufacturing, employee scheduling, inventory delivery, traffic flow analysis, high-temperature and low-temperature predictions, and mass transit scheduling. These business cases do not experience the lack of extensibility to longer-term predictions as seen in current methods used for predicting computing resource utilization. Because the research methods used by business for long term prediction focus on human interaction patterns they may provide opportunities for Information Technology (IT) researchers to adapt these techniques to help predict computing resource needs for processing, communications, and storage management.

The remaining sections of the paper will address the following: section two will address the current research being conducted in computing and business, section three will discuss the need for the proposed research, section four will provide observational evidence supporting the assertion, section five discusses the setup of a feasibility experiment, section six provides the results of the feasibility experiment, section seven provides theoretical benefits from the research, and section eight discuss future computer resource prediction research.

2. CURRENT RESEARCH

2.1 Computer Resource Management
Operating system scheduling has been a research topic of interest since early operating systems implemented multitasking [18]. Current scheduling algorithms use basic mathematical formulas to determine how to order the resource access. More complicated formulas are reserved for special needs, since the overhead of making the choice often outweighs the benefit gained from the selection [18]. It is very difficult to predict the utilization of a resource by an unknown process or the length of time that a process requires a resource. Research has shown that, as resource loads increase, scheduling algorithms become unfair [19]. Thus the provision of additional resource utilization information may improve upon an algorithms ability to provide an increase in fairness to running tasks. Research remains active in web scheduling [1, 7, 16]. The theme of this research domain is to increase the efficiency of web service responses to provide improved turnaround on tasks. Thus web service load balancing is improved by leveraging operating system scheduling with the addition of utilization prediction.

RIIT'12, October 11–13, 2012, Calgary, Alberta, Canada.

Andreolini and Casolari [1] focused their research on recognizing larger patterns in resource trends. By looking at the recent past they determined that processing needs could be determined using a reasonable prediction pattern such as increasing, decreasing, staircase, or alternating. By better understanding load trends Andreolini and Casolari [1] could provide advanced notice of overloading, offloading, or stabilization leading to better management of load balancing, admission control, scheduling, and graceful degradation.

Google has also recognized the need for its systems to react well in a highly fault-ridden environment [7]. The primary concern is hardware failure, but Google also recognizes that software can be "buggy" or have faults that create availability issues. Failure is deemed a normal state in systems and thus a variety of prevention techniques are used to keep the services active. Thus increased reliability is provided through fault tolerance aided by usage data obtained in the monitoring process.

Schroeder and Harchol-Balter's [16] web server work supports issues with heavily loaded systems noted by Ghemawat et al. at Google [7]. The work recognizes that at times web services may experience transient overloads attributed to events such as national television recognition or a newspaper article. While unpredictable events present random loads on the web servers, other events such as holidays also cause increased loads that are difficult to predict using current research. To help reduce the loads during excessive utilization, the Shortest-Remaining-Process-Time (SRPT) scheduling algorithm is used to estimate the remaining process time for web requests and, when appropriate, raises a process's priority to allow that process to complete more quickly [16]. Schroeder's research monitors system bandwidth usage to determine if SRPT is an improvement over existing web response scheduling methods used in an unmodified system. Both Schroeder's [16] and Google's [7] research have taken a reactive posture over a preventative posture.

2.2 Distributed Resource Management

Load balancing [14], resource requirements [5], and performance [10] research establishes a management framework that is enhanced by prediction. The methods use data collection and prediction techniques to identify resource-scheduling algorithms that improve scheduling effectiveness over current methods.

The Resource Prediction System (RPS) [5] is based on a collection mechanism that considers the cost of collecting the information for prediction could outweigh the benefits of that prediction. Thus, as the RPS toolkit records readings from the system, those numbers are fed into a predictive algorithm that is used to determine the resource's utilization up to approximately 30 seconds into the future.

Most predictive systems use the measurement of mean, median, and standard deviation of recent usage data from the target resources. Istin et al. [10] expanded on Dinda's [6] work, using wave analysis. Like previous work, Istin's focuses on common resources including memory, CPU statistics, disk IO, and communications. In this research, [10] short-term predictions were used over long-term predictions because the authors believe that prediction over the next few seconds for a distributed system is much more important that being able to predict several hours into the future.

Rood and Lewis [14] focused their work on failure prediction. To create the predictor, Rood and Lewis established a state machine that permits the predictor to move between five states:

- system available,
- CPU threshold exceeded,
- job eviction,
- user present,
- and system unavailable.

The probability of the appropriate state is calculated for the target machine using predictors based on the time a process is expected to take. Rood and Lewis [14] did not indicate the amount of history collected from the machine, but they used enough data to identify transitions between states. These transitions resulted in a Markov chain that identifies the probability of a state change. Rood's method was able to provide reasonable predictability of a machines state for about 16 days which improved machine resource use.

2.3 Business' View of Predictability

Business has used resource forecasting for a variety of needs, including traffic flow [9] and short-term water usage [13]. Both Haung [9] and Liu [13] identify a block of data to define a history of resource utilization and these data blocks are used to predict future resource utilization.

In Haung's [9] work, statistics generated from two months of traffic patterns for a toll road were used as a historical base. The traffic data was processed and provided a successful prediction of traffic flow over seven days because human interaction is frequently predictable.

In Liu et al's [13] work, a history of water utilization was collected for approximately six months in one-hour increments. This then provided a single day of predicted use, in which the prediction was compared to the actual use. Liu's research showed that human interaction creates an environment where forecasting can identify future water use over a single day.

In no case is the use of forecasting 100% accurate. Just as Haung's [9] and Liu's [13] research showed predictability over days or weeks is feasible, studies into computing resource use based on similar techniques may show that hourly, monthly, and yearly trends can be reasonably predicted due to human generated access to applications and systems.

3. THE NEED FOR RESEARCH

Computer science algorithms are targeted to control process and resource management, however low-level raw measurement of these resources is deemed random instead of predictable [1]. This is supported by the concept that computers execute many instructions per second, perform multiple context changes between tasks, and service many interrupt responses that predictability becomes difficult. Reinforcement of the randomness of task execution is supported by studies in process selection, resource management, and system communication which show predicting the future state of a computer is very difficult and probably is likely beyond determination.

Attempts have been made [5, 15] to review resource utilization and then apply those observations to predicting the future state of computer system resource utilization. Although there is some success, predictions have not been accurate beyond several minutes. As a result, disk IO, process, and other methods of scheduling are evaluated based on inconclusive results of efficiency thus simplicity guides the ultimate choice of algorithm [12].

The relative execution timeframe for a computer system and a user is very different, thus this difference in time perception should lead to a new direction in the focus of research. Scheduling is done at the system task level and as seen earlier in the discussion has been the time frame used for research on the topic of resource prediction. The review of research in this area has revealed the problem is extremely difficult due to the vast amount of information required to make a prediction. Servers can remain active for months; current research has not considered this large timeframe when reviewing resource utilization.

Research into computers and human behavior has crossed paths [2, 9, 13], but focuses on human patterns instead of the computing patterns. The timeframe of minutes, hours, days, weeks, months, years, and even decades is important to people and their associated tasks. This human timeframe also has no interest in how the computer completes a task, the resources used, or the processes involved in getting the work done. Thus this area can be further explored to adjust the focus of the research in resource predictability to time frames that are better suited to the nature of human and computer task interaction.

4. SUBJECTIVE EVIDENCE

Research supporting longer prediction timeframes has been used to train and predict patterns that contain human interaction [9, 13]. Patterns of usage exist in many other places such as in national phone usage on mother's day, gasoline usage over the summer, heating fuel usage during the winter, electric utilization for air conditioning during the summer, and passenger traffic at the airports between Thanksgiving and New Year's Day. While this research [9, 13] deals with traffic patterns or water usage, there are human patterns that also exist in computing that may also be predicted using similar methods. Businesses currently collect data to find patterns that can be more accurately provide pricing structures, inventory management, and staffing levels. Research hasn't been conducted into how long term computing patterns, generated by the user, drive resource utilization of that equipment.

One published example of patterns in computing use driven by people is a usage analysis of the 1998 World Cup Web Site [2]. Figure 1 clearly shows that shortly after the start of a game, web site usage jumped and then quickly declined. The only exception was on Sundays. This pattern follows through the entire World Cup series, mapped on page 23 of Arlitt and Jin's report [2], the pattern repeats for each game over a 20 day period of the World Cup.

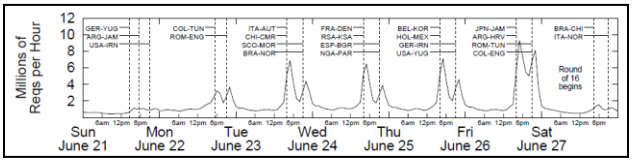

Figure 1: Workload Characteristics of 1998 World Cup Web Site

When a business opens, a pattern develops as the users log onto the system and begin their tasks for the first time each day. Additionally, disk utilization is likely to spike as well as Internet traffic as employees start the day and thus a pattern of use develops throughout the day.

Some colleges provide periodic class periods throughout the day, Monday through Friday of each semester. This class schedule creates very predictable network utilization based on student computer resource use. These patterns are generated by human interaction and are predictable [11], thus supporting the possibility that computer resource utilization is predictable in such cases based on the human timescale.

5. STARTING LONG-TERM PREDICTION

Research into long-term prediction must demonstrate that the basic premise of time-series analysis can be used for server resources before the more complex evaluations used by business can be applied to the process. To address the basics, a web server was set up using Windows 64-bit 2008 R2 Server operating system. This was loaded onto a Gateway E4300 (Intel Pentium 4 dual processor running at 3.4 Ghz). The server was also equipped with a Marvell Yukon Gigabit Ethernet interface and a Western Digital 1600JD ATA hard drive.

Both DNS and DHCP services were provided by the server for the client machines attached to a closed network for this experiment. These units were connected using a Cisco Catalyst 2950 48-port switch so they could access web content provided by IIS 7.0 web services. The raw data was collected on the server using a Windows LogMan, the command-line version of the GUI performance monitor program. LogMan was configured to capture CPU utilization, available memory, network traffic (in bytes per second) to and from the server, and the disk activity resulting from the requests sent to the web server. Each of these metrics was recorded in a log every ten seconds and a new log was started after each 24-hour period.

Twenty-two client machines were used for the initial research. These machines were Dell Precision T3500's with 64 bit openSUSE 11.2 installed on a Toshiba HDDR500E04X USB hard drive. The clients used an Intel Xeon 8-core CPU running at 2.8Ghz, a Broadcom NetXtreme gigabit Ethernet card, and eight gigabytes of RAM. The installation of openSUSE used the text-based interface to maximize the amount of RAM available to the simulator being used for the experiment.

The simulator selected for the experiment was Scalable URL Reference Generator (SURGE) created by Paul Barford [3] to exercise web servers. This simulator was selected after reviewing a number of current traffic generators because it not only exercises web services, but does so in a fashion that mimics human utilization of static web pages. While services have drastically changed since SURGE was created, the load put on the server is still able to provide a patterned level of requests to permit resources to be logged and evaluated for long-term predictability.

Barford [4] created a suite of programs with a variety of functions to emulate human trace data. The first program calculates the number of times a file will be accessed and the number of files the client may access on the server based on the most frequently-accessed document. A second program randomizes a file access list based on the Pareto distribution. Another program determines which files will be accessed as a single request and which files will be requested in a group using the pipelining feature available on web servers today. Another part of the suite will generate random wait times before the next request is made. In the original version this could be up to 15 minutes; however, modern servers set a two-minute timeout for connections that go quiet. The suite was adjusted so that no wait was longer than 90 seconds. Finally, the suite generates each of the files that will be used in the communication with random lengths of a single alphanumeric character. In Barford's [3] original version, every file contained "a"-s. For this experiment, it was desirable to exercise the server hard drive so that each client has its own character and set of files on the server.

The main purpose of SURGE is to generate requests and capture replies from the server. Barford's [3] focus was to create a workload generator that mimicked human behavior based on trace data, while this study will focus on using a controllable simulator to determine if long-term predictability is a viable research area. SURGE accomplishes its load generation by permitting up to eight processes to be started by the program, with up to 250 threads within each process. When each thread is running, it pulls a request from a common queue to generate a web GET statement. This is then transmitted to the web server using a standard TCP connection; the thread then reads the reply from the server. After the read is completed the thread pulls a sleep time from another common queue and waits for up to 90 seconds before initiating the next request. Initial runs indicated that a client could generate, and the server respond, to eight processes with 200 threads, making over 100,000 requests in about five minutes without difficulty.

Each of the SURGE clients was configured to run a request sequence every fifteen minutes using 75 threads. The first iteration starts eight processes and can complete 287,000 requests to the web server during its execution. The second iteration runs six processes, while the third iteration uses four processes and the fourth iteration uses two processes covering approximately 100,000 requests. Each iteration runs for a period of fourteen minutes with the remaining minute used to reset for the next iteration. A script was written to start SURGE in fifteen-minute iterations, restarting the four part sequence every hour. Each iteration level produces an average level of 35%, 28%, 21%, and 14% CPU utilization.

6. FEASABILITY OF FORECASTING

For this short study sample data was collected from the web server every ten seconds over a forty hour period by the LogMan program. The data samples were averaged into one-minute data points before being evaluated to determine if the data was a good fit for long-term forecasting. SAS was used to examine the data distribution for each minute, correlated hour-by-hour, over the forty-eight hours of data.

Each fifteen-minute load level shows a tight CPU utilization load generated by the simulator. These levels often remain within several percent of the mean. The server's clock was two to three minutes faster than the client's, providing the shift (see Figure 2) that accounts for the step-down pattern generated by the simulator. At the end of each fifteen-minute cycle the reset time is also evident within the minute pattern.

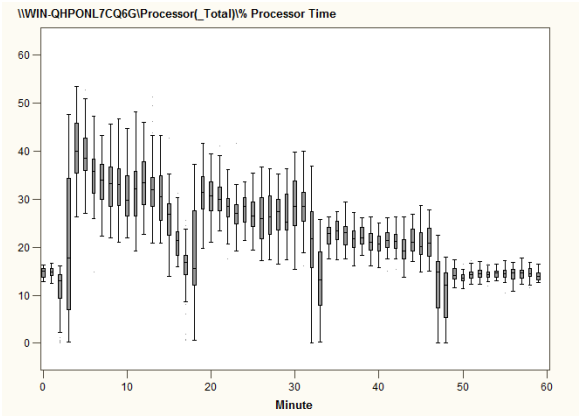

Figure 2: Box with Whiskers for CPU utilization

An additional evaluation was done using linear regression with an 80% confidence banding to provide evidence that the use of an appropriate seasonal evaluation will support the predictability of resources. In this case a single minute – minute 55 – was extracted from each hour over the forty-eight hour period and used to plot out the CPU utilization level of the server (see Figure 3). This minute had a mean of 14.44% and a standard deviation of 1.21%. As the figure indicates, only three samples of CPU utilization were outside the 3.1% range representing an 80% confidence level. The trend line and data distribution provide evidence that forecasting good fit based on the correlation of time and CPU utilization, and that time series analysis may provide an effective forecasting method.

The information provided in sections 5 and 6 is provided as foundation for the need for a much larger body of work. Time-series analysis has been widely used for business forecasting for years. Research toward an understanding of the patterns of computer resource utilization outside of operating system management has not received much attention. Additionally, work lags in predictions that attempt to look beyond several minutes into the future. The system described in section 5 was established to begin the feasibility of using time-series analysis for system resources over long-term periods.

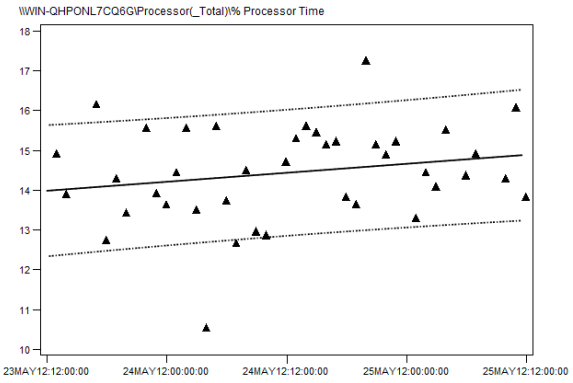

Figure 3: Scatter Plot for one minute over two days.

7. POSSIBLE RESEARCH BENEFITS

Since identifiable patterns already exist for resources, researchers can begin to identify how long-term resource predictions can be accomplished. Future research will focus on predictable use over hours, days, weeks, months, and years instead of only over the next few computing cycles. Once researchers begin to understand where human patterns influence resource utilization, additional IT research will open a variety of benefits to business and systems management.

Given a qualified forecasting system an administrator could identify periods-of-service overload for a web server, so that extra traffic could be diverted to another machine within the system. The administrator would use the history of the web server over the past several months while forecasting could identify times when an overload is likely to occur. The extra services could be deployed shortly before the overload is expected and released once the need has abated. A process of renting services from the cloud for a short time could reduce the cost of having a second system always on standby or by understanding when new systems would need to be added to accommodate growing needs.

At present, resource utilization is considered unpredictable and chances of predicting future utilization appears nearly random.

But with this type of research, some of the resource utilization should be predictable. For example, if the evaluation determines that CPU utilization during the organization's startup period each day, Monday through Friday, normally jumps to 60% and, of that 60%, 30% can be directly attributed to user logins with a ±2% deviation, a "random" CPU utilization of 30% would remain. That last 30% could then be bounded with acceptable deviations so that when the utilization moves outside those bounds, the system would be considered to be behaving abnormally. Then research could focus on another predictable event in the remaining 30% to reduce the randomness further. Eventually research would be able to identify the true randomness within the system and when that system was not acting predictably.

Initially the randomness of any resource utilization will be very large, but as more human generated patterns are discovered, that deviation will shrink and system resource use prediction will rise. By shrinking the remaining randomness of the system resources, a set of probabilities about the state of the machine can be determined (Figure 4). The wellness of a machine can be determined by three states: normal, abnormal, and failure. While resource utilization remains within the acceptable deviation of the prediction, the system remains in the normal state. If the resource utilization falls outside the prediction and deviation, the machine would move into an abnormal state. If the resource returns to utilization between the predicted use and deviation the machine can return to the normal state. Finally, if the machine experiences a hard or soft failure, it would move into the failure state.

Research begins with identifying the common patterns within a system. These could include network traffic patterns, Internet utilization patterns, and web service patterns. As more resources and utilization patterns are identified, the more administrators will be able to balance system health. Research then moves to creating a feedback loop that provides the system with the ability to make future predictions based on the original history plus the resource utilization as it has occurred in real time. In this way, hourly, daily, weekly, monthly, and yearly patterns can be updated to reflect patterns as they change based on human interaction.

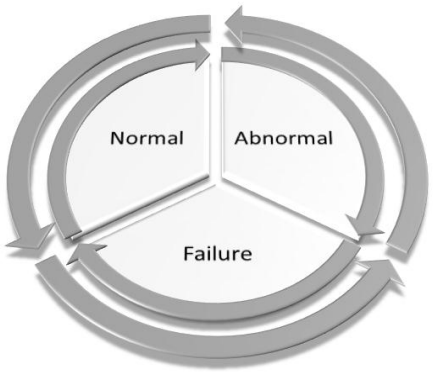

Figure 4: Machine States under Predictable Usage.

Once the forecasting research begins identifying resource utilization, a wide variety of additional research and management opportunities will arise.

- Load balancing could now add the additional dimension of expected resource load, permitting businesses to rent services for short periods after transferring critical data to the rented system.
- Distributed systems could use the forecasting to better balance process loads and, if a machine entered an abnormal state, begin preemptive measures to move critical tasks prior to failure or overload.
- Services could be more properly sized to equipment and provide growth projections for a system retirement or upgrade.
- Virtual machines with opposite resource utilization could be safely matched together to save on hardware costs.
- Survivable systems could begin graceful degradation upon entering the abnormal state and restore full services if the machine returned to the normal state.
- Intrusion Detection/Prevention systems could begin more aggressive scanning upon entry to the abnormal state.
- Resource load prediction could identify the tipping points for the overloading of a resource and take preventative measures as the resource approaches that threshold.

8. CONCLUSION

Five to ten minute prediction methodologies based on computing time frames have been reasonably successful [1, 5, 11, 17], while similar results in longer patterns of resource utilization and prediction remain elusive. System degradation and overloading of services is a constant issue for the service owners that needs to be addressed to guarantee availability [8]. Web services continue to be a key area of interest due to the rate of growth and importance to industry [17]. Additionally, to support Quality of Service issues web services require better resource management. To address these concerns, Internet services should embrace resource prediction by shifting from short-term to long-term time periods. The results of these methods will provide improvements in load balancing, job dispatching, job distribution, and overload prevention [1] by permitting a system to better anticipate resource needs further into the future.

Algorithms currently use real time sampling [5] to solve resource limitation issues and can only address problems as they are detected or based on short-term prediction. The results of research in the area of short-term prediction of resource utilization, has been helpful in supporting scheduling changes [16], load balancing [17], and resource trending [1]. But, researchers have determined that current methods of data collection and analysis deteriorate quickly the further into the future predictions are made [1] and that more research into longer-term prediction is needed to manage system components.

Through the understanding of predictable human computer usage, the state of a machine can be determined. There will always be a small part of randomness in systems as patterns change over time. But business already understands the value of forecasting and knows that research based on human utilization patterns has tremendous value to improving business performance. Computing is no different: it is also driven by predictable human patterns.

Once those patterns are understood, administrators will be able to see which of the three states their equipment currently occupies. Knowing that equipment has entered an abnormal state will provide additional time to react to circumstances instead of reacting only after the system has entered a failing state.

Computer engineer's focus on computing electronics, the computer scientists focus on efficiencies of algorithms and systems, while IT has focused on effective use of equipment and services. The ability to effective use resources, balancing cost

with need, keeping systems alive and protecting data during failures all fall into the concerns of IT. This type of research advances the effectiveness of equipment and services and therefore should be a research concern for those in IT.

9. REFERENCES

[1] Andreolini, M. and S. Casolari, *Load prediction models in web-based systems*, in *Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*. 2006, ACM: Pisa, Italy. p. 27.

[2] Arlitt, M. and H. Jin, *Workload Characterization of the 1998 World Cup Web Site*, in *HPL-1999-35 (R.1)*. 1999, HP Laboratories Palo Alto. p. 90.

[3] Barford, P. and M. Crovella, *Generating representative Web workloads for network and server performance evaluation.* SIGMETRICS Perform. Eval. Rev., 1998. **26**(1): p. 151-160.

[4] Barford, P.R., *Modeling, Measurement and Performance of World Wide Web Transactions*, in *Graduate School of Arts and Sciences*. 2001, Boston University: Boston.

[5] Dinda, P.A., *Design, Implementation, and Performance of an Extensible Toolkit for Resource Prediction in Distributed Systems.* IEEE Transactions on Parallel and Distributed Systems, 2006. **17**: p. 160-173.

[6] Dinda, P.A. and D.R. O'Hallaron, *Host load prediction using linear models.* Cluster Computing, 2000. **3**(4): p. 265-280.

[7] Ghemawat, S., H. Gobioff, and S.-T. Leung, *The Google file system*, in *Proceedings of the nineteenth ACM symposium on Operating systems principles*. 2003, ACM: Bolton Landing, NY, USA. p. 29-43.

[8] Hoffmann, G.A., K.S. Trivedi, and M. Malek. *A Best Practice Guide to Resources Forecasting for the Apache Webserver*. in *12th Pacific Rim International Symposium on Dependable Computing (PRDC'O6)*. 2006.

[9] Huang, J. *Short-Term Traffic Flow Forecasting Based on Wavelet Network Model Combined with PSO*. in *International Conference on Intelligent Computation Technology and Automation*. 2008.

[10] Istin, M., A. Visan, F. Pop, and V. Cristea. *Decomposition Based Algorithm for State Prediction in Large Scale Distributed Systems*. in *Ninth International Symposium on Parallel and Distributed Computing*. 2010. Istanbul, Turkey

[11] Krithikaivasan, B., Y. Zeng, K. Deka, and D. Medhi, *ARCH-based traffic forecasting and dynamic bandwidth provisioning for periodically measured nonstationary traffic*. IEEE/ACM Trans. Netw., 2007. **15**(3): p. 683-696.

[12] Lampson, B.W., *Hints for computer system design*, in *Proceedings of the ninth ACM symposium on Operating systems principles*. 1983, ACM: Bretton Woods, New Hampshire, United States. p. 33-48.

[13] Liu, J., R. Zhang, and L. Wang. *Prediction of Urban Short-Term Water Consumption in Zhengzhou City*. in *International Conference on Intelligent Computation Technology and Automation*. 2010. Changsha, Hunan, China.

[14] Rood, B. and M.J. Lewis. *Resource Availability Prediction for Improved Grid Scheduling*. in *Fourth IEEE International Conference on eScience*. 2008.

[15] Rood, B. and M.J. Lewis. *Availability Prediction Based Replication Strategies for Grid Environments*. in *10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. 2010. Melbourne, VIC, Australia.

[16] Schroeder, B. and M. Harchol-Balter, *Web servers under overload: How scheduling can help.* ACM Trans. Internet Technol., 2006. **6**(1): p. 20-52.

[17] Sharifian, S., S.A. Motamedi, and M.K. Akbari, *An approximation-based load-balancing algorithm with admission control for cluster web servers with dynamic workloads.* J. Supercomput., 2010. **53**(3): p. 440-463.

[18] Silberschatz, A., P.B. Galvin, and G. Gagne, *Operating Systems Concepts, sixth ed.* 2003: John Wiley & Sons, Inc;.

[19] Wierman, A. and M. Harchol-Balter, *Classifying scheduling policies with respect to unfairness in an M/GI/1*, in *Proceedings of the 2003 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*. 2003, ACM: San Diego, CA, USA. p. 238-249.