SUN BLUEPRINTS APPROVED

Observability

By Adrian Cockcroft - Enterprise Engineering

Sun BluePrints™ OnLine - December 1999

Sun microsystems
ENTERPRISE
ENGINEERING

`http://www.sun.com/blueprints`

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303 USA
650 960-1300 fax 650 969-9131

Part No.: 806-4051-10
Revision 01, December 1999

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

CETTE PUBLICATION EST FOURNIE "EN L'ETAT" ET AUCUNE GARANTIE, EXPRESSE OU IMPLICITE, N'EST ACCORDEE, Y COMPRIS DES GARANTIES CONCERNANT LA VALEUR MARCHANDE, L'APTITUDE DE LA PUBLICATION A REPONDRE A UNE UTILISATION PARTICULIERE, OU LE FAIT QU'ELLE NE SOIT PAS CONTREFAISANTE DE PRODUIT DE TIERS. CE DENI DE GARANTIE NE S'APPLIQUERAIT PAS, DANS LA MESURE OU IL SERAIT TENU JURIDIQUEMENT NUL ET NON AVENU.

Please
Recycle

Adobe PostScript™

Observability

Observability is the first requirement for performance management and capacity planning. The performance characteristics need to be presented in a way that makes sense to those who are looking at it, so they can observe the behavior of the systems and understand what is happening. There are three different viewpoints that need to be provided for different audiences:

- Operations
- Engineering
- Management

Each needs very different information and can be implemented using different views of the same performance database or using completely different tools. In this article the three views will be explained briefly then an implementation and examples will be provided for the operations viewpoint.

This is the first Sun BluePrint™ Online article to be based upon a forthcoming Sun BluePrint book by Adrian Cockcroft that will discuss Capacity Planning and Performance Management techniques. Publication is expected in the summer of 2000, and several more Sun BluePrint Online articles will preview its content over the coming months.

Operations Viewpoint

The operations room works in the immediate time frame using a real time display, updated every few seconds. Alert based monitoring should be combined with automatic high level problem diagnosis if possible. However this level of sophistication is rare. Simple high level graphs and views of current data will be needed.

Engineering Viewpoint

Engineers need to see large volumes of detailed data at several different time scales. It is used as input to tuning, reconfiguring and future product development, and aids in low level problem diagnosis.

You need to generate and store large volumes of raw data only on systems that are being tuned. Detailed reports with drill down and correlation analysis are needed, and many tools implement this kind of functionality very well.

Management Viewpoint

Management will never have time to read a voluminous report. A single page daily summary of status and problems is more appropriate. This should contain:

- Business oriented matrix
- Future scenario planning
- Concise report with dashboard style status indicators

Marketing and management input is needed to define the scenarios. These form a grid, with things that change the load level down the side, and things that change the capacity across the top. It is then easy to see which tuning and upgrade projects must be implemented to cope with the increased load of the next marketing campaign.

Example Scenarios

In the following sections, techniques and examples for implementing the operations viewpoint is presented. Automated alert monitoring for large numbers of systems is demonstrated using the Sun™ Management Center (SunMC a.k.a. Sun Enterprise™ SyMON 2). An alternative technique for monitoring a small number of systems with customized data views uses the SE Tookit to log processed data and plot it using a spreadsheet.

Operations Viewpoint Implementation

This section describes an implementation of operations management that is generic in the techniques being used, but specific in the tools used to implement in a primarily Sun™ technology-based environment.

The steps covered in this section illustrate how to use a system monitor to automate the process of tracking multiple systems for unusual or problematic conditions.

1. **Enable Rules**

   Many monitoring tools are shipped with their monitoring capability unconfigured or partially enabled so the first step is to enable all relevant rules. If you have BMC patrol, this is performed by loading Knowledge Modules. With SunMC basic rules are enabled but additional health monitoring can be loaded.

2. **Setup Rule Thresholds**

   The default rules may be setup for a small system, in which case they may be generating false alarms on a much larger server. They may also be set too high, so a problem could occur without firing the alarm. For all the rules that you have, make a list of each rule, its threshold rule, and the typical and maximum metric values you see in normal operation. If you can identify the values for these matrix seen during problem situations, then you can set an alert threshold correctly to indicate the transition from normal (green) to high (amber) to problematic (red) levels. If you have never seen a problem with a particular metric, then set its threshold to a little higher than the highest value you have ever seen before. In this way you “surround” the systems being monitored with a behavior envelope. If the behavior goes beyond that envelope in any dimension then an alarm tells you that something abnormal is going on. Document and save the customized rules and thresholds so they can be replicated on similar systems and do not get lost during system upgrades.

3. **Monitor and Reset Alarms**

   The monitoring software usually indicates that a problem state has occurred, so you can acknowledge the alarm and investigate the source of the problem. You can disable alarms that don’t appear to be useful and create new alarms to watch matrix that are especially relevant to your workload. Be sure to archive the alarm log. Collect the number of alarms that occur at each of the problematic levels (amber and red) on a daily basis and report them to management so a trend can be plotted over time.

Implementing with SunMC

SunMC is used in this example as it is becoming increasingly important as a primary platform for Sun to deliver tools that manage and monitor Sun systems. There are many other tools that could be used in this role. Their primary advantage is also their primary disadvantage—they are not Sun-specific and can manage many kinds of systems from several vendors. But because they are not Sun-specific they do not provide anything like the detailed active management and diagnosis capabilities of SunMC.

- SunMC has third party vendor support for managing Windows NT based systems and relational databases as well as for performing generic system administration functions.
- Inter-operation with CA Unicenter, BMC Patrol, and HP IT Operations among others allows SunMC to integrate into existing multi-vendor environments.
- The recent addition of a Software Development Kit (SDK) allows anyone to develop extended modules and user interface components.
- SunMC is secure, using public key authentication by default, and allowing full encryption as a higher overhead option. This allows a remote administrator to modify the system configuration safely. Administrative role definitions and access control lists provide flexible security control.
- Full system configuration information is available via SunMC. This is not at all trivial to implement, and no other tools come close. Both logical and physical views using pictures of components are available. A recent addition is support for display and management of the Sun Enterprise™ 10000 System Service Processor (SSP) and Dynamic System Domains.
- Hardware and software are developed in parallel at Sun, and many new hardware and software systems will use SunMC as their primary management console, or new products will be launched from SunMC.
- Diagnostic tools such as the Sun Configuration Service Tracker (CST) use SunMC as their user interface console.
- Because SunMC is based on the Simple Network Management Protocol (SNMP), it can integrate closely with network management products and tools as a peer, rather than just forwarding SNMP traps. The SunMC SNMP version 2 with usec security implementation is more efficient at bulk transfer of data and adds security to the common SNMP standard.

Alert Monitoring with Sun™ Management Center

SunMC can be used to monitor systems using complex rule based alerts to diagnose problems. The screenshots in this section illustrate how to configure and use SunMC to perform basic monitoring operations that are useful for operations management. The screenshots used in these examples are taken from the Sun Enterprise SyMON 2.0 product as described in the Sun BluePrint™ Resource Management book (ISBN

#0-13-025855-5). They will be updated and extended using the latest SunMC features for inclusion in the Sun BluePrint Capacity Planning book scheduled to publish the year 2000.

The SunMC Health Monitor

SunMC includes a system health monitoring module that can be used by operations to simply see if a system has enough resources to run comfortably. For example if the CPU state is reported as red then either less work or more CPU power may be needed on that system. Similarly if the memory rule reports red then the system may need more memory.

The SunMC health monitor is based on a set of complex rule objects. The health monitor is not enabled by default when SunMC is first installed, as only the basic modules are loaded into the agent.

A simple rule can be placed on any single SunMC metric to monitor its value. Additional rule objects implement more complex rules that refer to several matrix or perform advanced processing on the data.

▼ To Load the Health Monitor Module

1. **Start SunMC with the default administrative domain.**
2. **Select the system and popup a menu.**
3. **Select the Load Module option from the menu.**

   In the example shown in FIGURE 1, a single red alert is present because one of the file systems on this machine is 90 percent full.

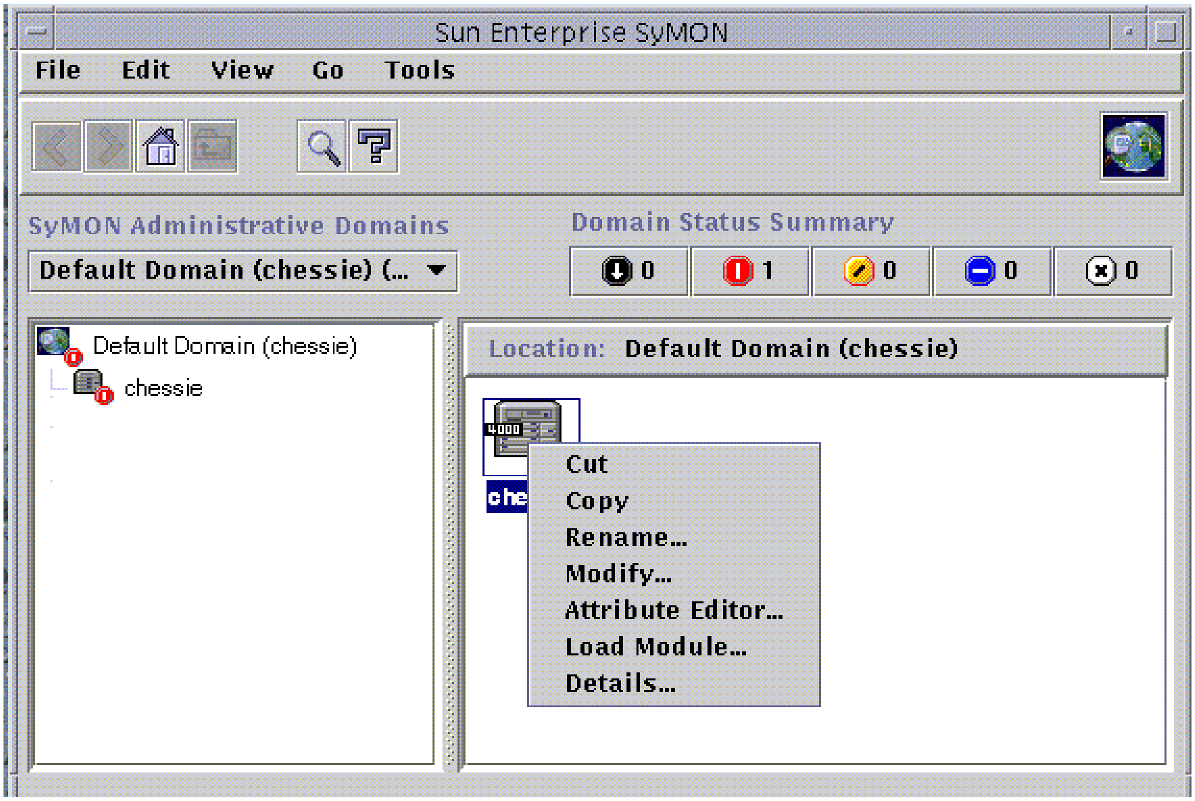
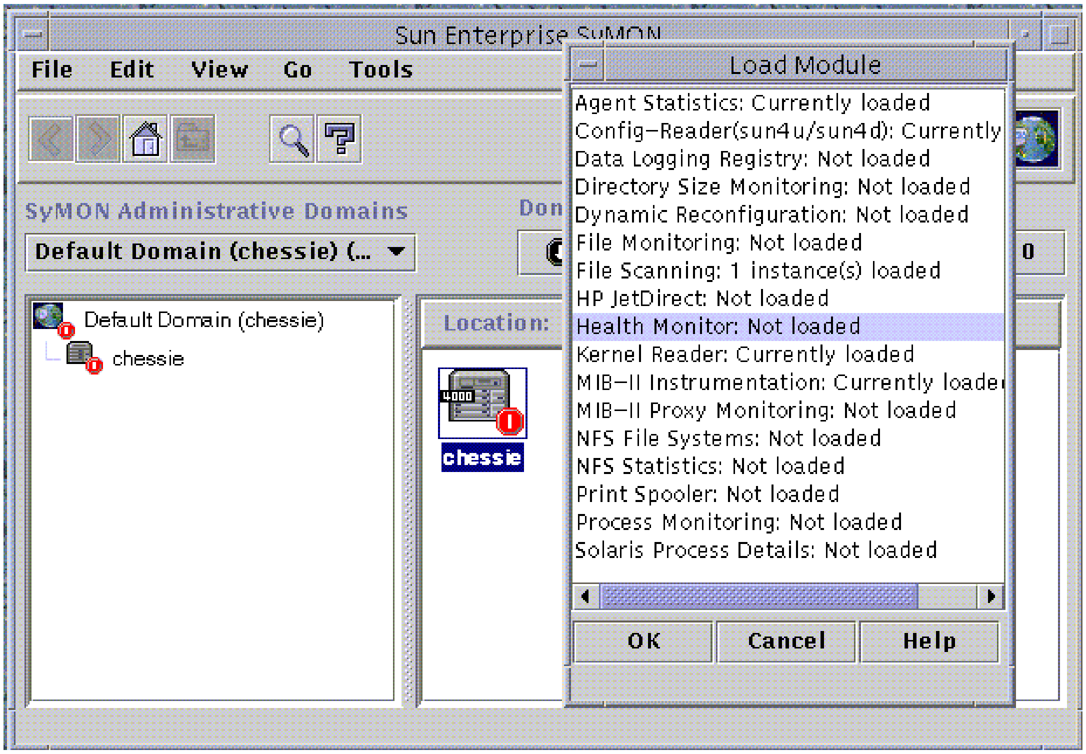


FIGURE 1 SunMC / Sun Enterprise SyMON Console

4. **Scroll down and choose the health monitor module.**

   It may already be loaded if SunMC has been pre-configured (see FIGURE 2).

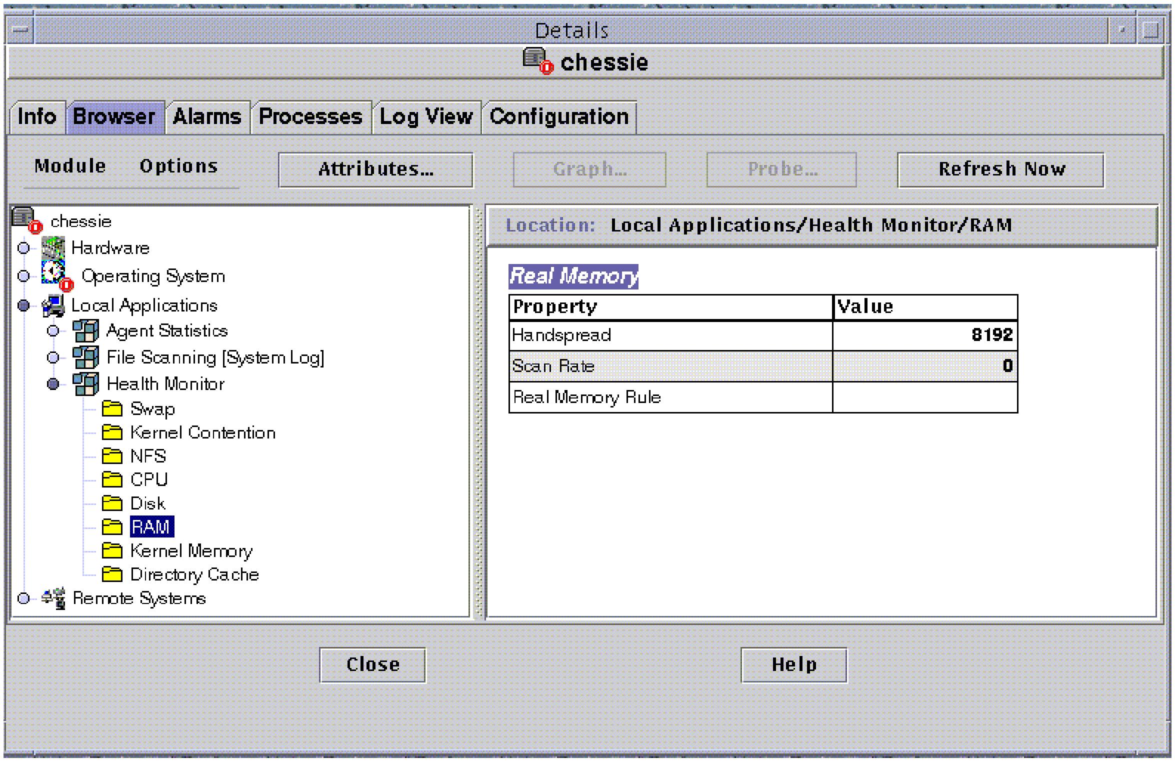
5. **Select it and press the OK button.**



Agent Statistics: Currently loaded
Config–Reader(sun4u/sun4d): Currently
Data Logging Registry: Not loaded
Directory Size Monitoring: Not loaded
Dynamic Reconfiguration: Not loaded
File Monitoring: Not loaded
File Scanning: 1 instance(s) loaded
HP JetDirect: Not loaded
Health Monitor: Not loaded
Kernel Reader: Currently loaded
MIB–II Instrumentation: Currently loade
MIB–II Proxy Monitoring: Not loaded
NFS File Systems: Not loaded
NFS Statistics: Not loaded
Print Spooler: Not loaded
Process Monitoring: Not loaded
Solaris Process Details: Not loaded

FIGURE 2 Load Health Monitor Module

Now any Health Monitor alerts will be logged for this system. You could drill down to the subsystem that caused the alert, but you don't expect any Health Monitor alerts yet, and there is already an unrelated disk space alert on this system. Instead, we will select the system and bring up the detailed view by hand. The Details... menu option was seen previously when we were bringing up the load module window. When it is selected, a second window opens that is specific to the system being monitored.



Location: Local Applications/Health Monitor/RAM

Real Memory

Property	Value
Handspread	8192
Scan Rate	0
Real Memory Rule	

FIGURE 3 Host Details Window

The browser tab of the host details window shows the modules that are loaded. Under local applications (which opens up if you click on the bullet next to it) you will find the health monitor module, and inside that you find the eight rules that are implemented to monitor several system components. Each rule shows a few variables. The RAM rule, displayed in FIGURE 3, shows that the current scan rate is zero, so the rule value is a white box. This box would go red and the red state would propagate up the hierarchy if the ratio of scan rate to handspread went too high and the page residence time dropped below the pre-set threshold. The rule attributes and thresholds can be viewed and edited by popping up a menu over the rule value. The way to use these rules is to increase the thresholds until in normal use on a system that is performing well there are no warnings. As the load increases over time, you will start to get warnings that indicate what subsystem is likely to be the bottleneck. If you have a system that is not performing well to start with, these rules will help you eliminate some problem areas and indicate which subsystems to concentrate on.

This browser mode can be used to explore all the operating system measurements supplied by the kernel reader for this system, including CPU usage, paging rates, and disk utilization.

Handling Alarms in SunMC

SunMC can monitor hundreds of systems from a single console. Systems can be grouped into various hierarchies. For example, they can be grouped by subnet, by application, or by department. Each group is given a cloud icon, or a background image can be loaded to form a map or building plan, and the icons can be positioned on the map. When a problem occurs the status propagates up the hierarchy making the clouds turn the appropriate color, so you can drill down to an individual system. The SunMC console supports multiple administrative domains that can be used by different types of users to see different views of the systems being monitored.

When a simple rule or one of the health monitoring rules generates an alarm it is logged by SunMC. At the domain level console, the worst alarm state for each system being monitored is counted. This means that with only one system being monitored, only one alarm will be indicated. In our example it is in the red state.

If you click on the red indicator shown in FIGURE 4 with a 1 next to it a new window opens that shows all the alarms for all the systems in this domain.

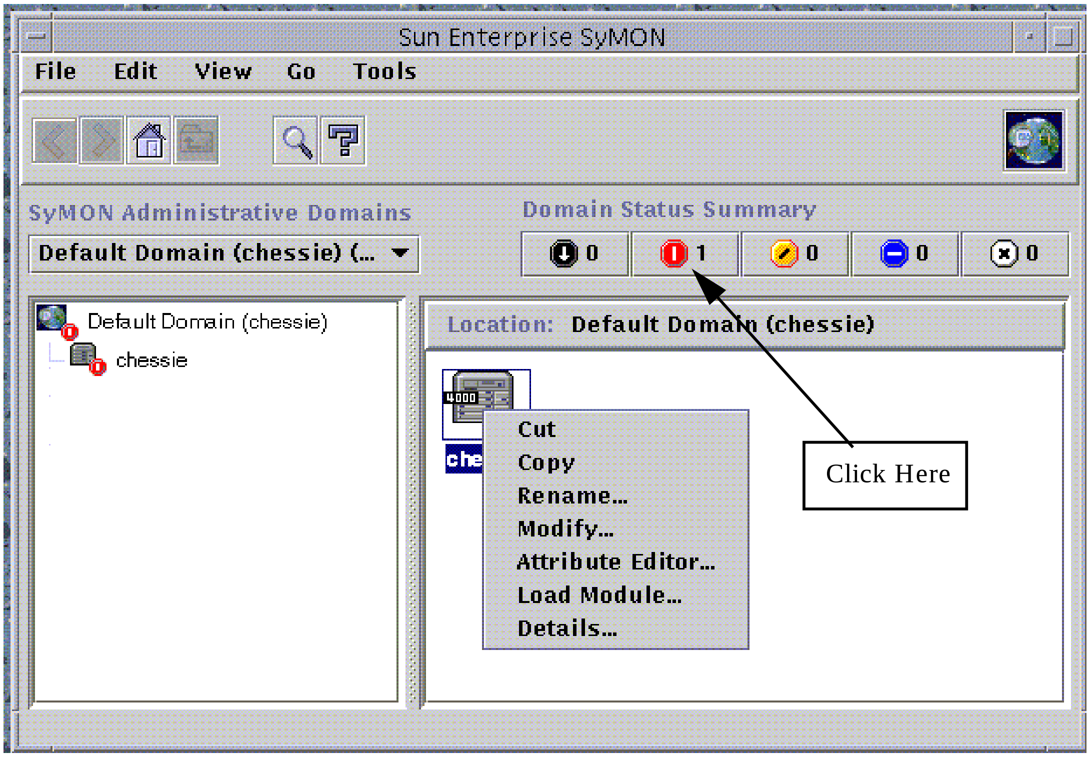
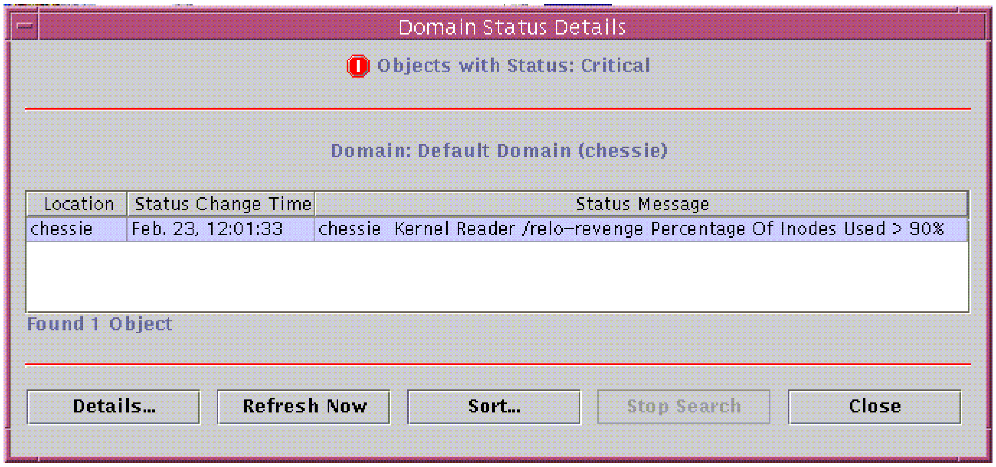


FIGURE 4 SunMC Console (Ignore popup menu shown here)

The screen shot shown in FIGURE 5 displays only systems or other network components that are in the red (critical) state. In this case, one of the file systems has too many small files in it, so it is almost out of capacity. This is a very useful form of filtering, so a flood of alarms from a large number of systems can be handled by selecting a domain and dealing with the critical problems for all systems in that domain first.



Domain Status Details

Objects with Status: Critical

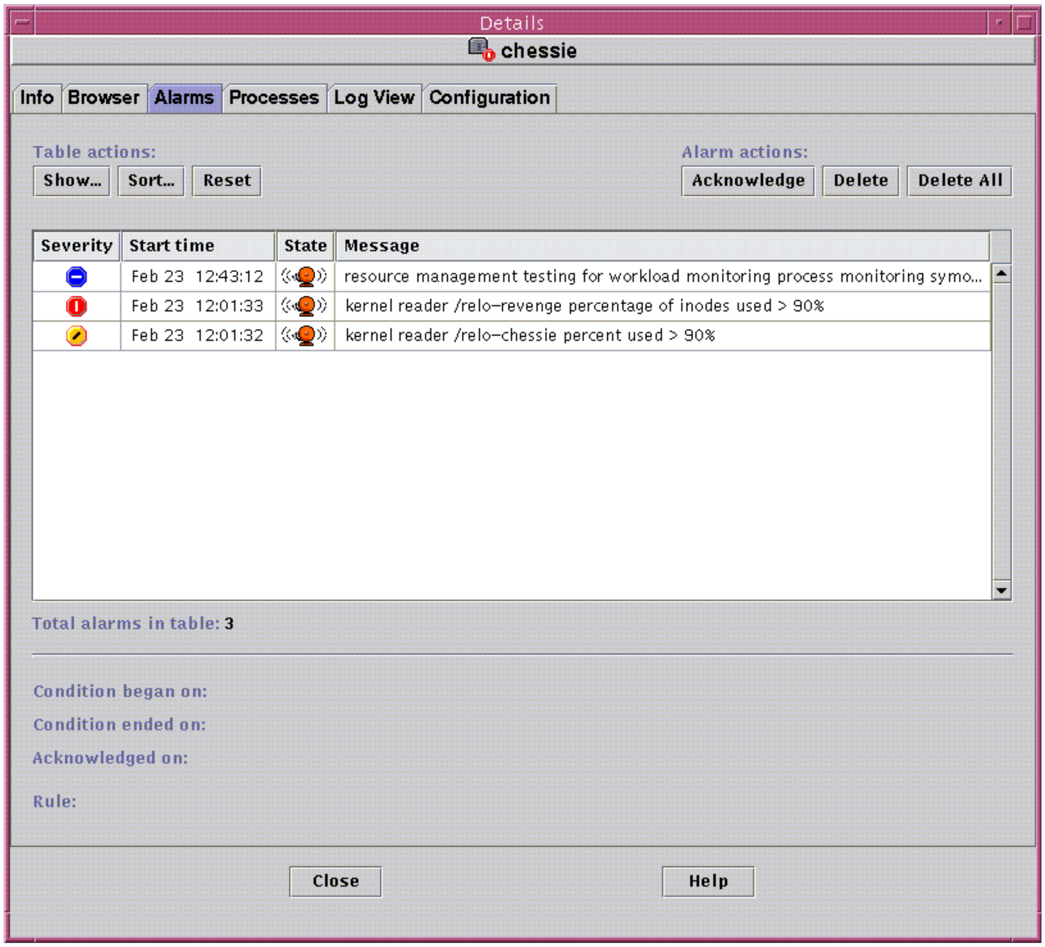
Domain: Default Domain (chessie)

Location	Status Change Time	Status Message
chessie	Feb. 23, 12:01:33	chessie Kernel Reader /relo–revenge Percentage Of Inodes Used > 90%

Found 1 Object

FIGURE 5 Domain Status Details Window

If you either double-click on the alarm or select the alarm and press the Details... button the Details window for that system opens with its alarm display tab selected as shown in FIGURE 6. This shows that in fact there are three alarms on this system, with only the most important one (red) being shown at the domain level.

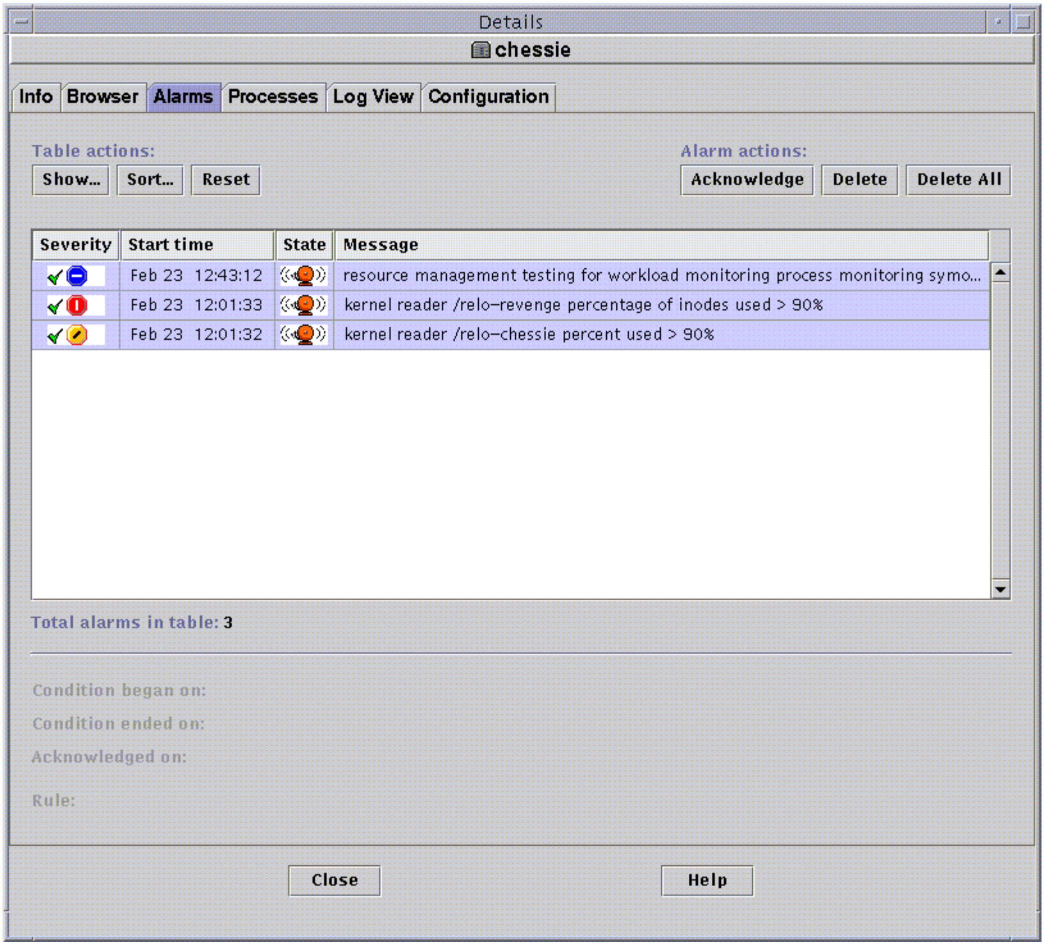


Severity	Start time	State	Message
⊖	Feb 23 12:43:12	🔔	resource management testing for workload monitoring process monitoring symo...
!	Feb 23 12:01:33	🔔	kernel reader /relo–revenge percentage of inodes used > 90%
✎	Feb 23 12:01:32	🔔	kernel reader /relo–chessie percent used > 90%

Total alarms in table: 3

FIGURE 6 Alarm Details Window

The next step is to select one or all of the alarms and acknowledge them by pushing the Acknowledge button. It is best to select them all by clicking on the first one and sliding the mouse down, then acknowledge them all at one time. It takes some time to perform the acknowledgment as it involves communicating all the way back to the agent on the server being monitored. Once an alarm is acknowledged, a tick mark appears by it as shown in FIGURE 7



Severity	Start time	State	Message
✓	Feb 23 12:43:12		resource management testing for workload monitoring process monitoring symo...
✓	Feb 23 12:01:33		kernel reader /relo–revenge percentage of inodes used > 90%
✓	Feb 23 12:01:32		kernel reader /relo–chessie percent used > 90%

Total alarms in table: 3

Condition began on:

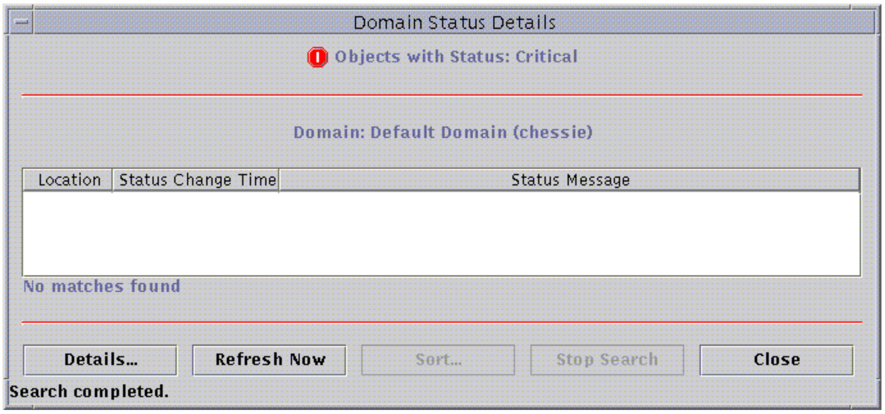
Condition ended on:

Acknowledged on:

Rule:

FIGURE 7 Acknowledged Alarms

Closing this window returns you to the Domain Status window. If the Domain Status window has not changed, press the Refresh Now button, and the alarm entry will go away as shown. Then you can close this window as well. The Domain Console in FIGURE 8 shows that the server no longer has a red marker on it and the Domain Status summary is all zeroes.



Objects with Status: Critical

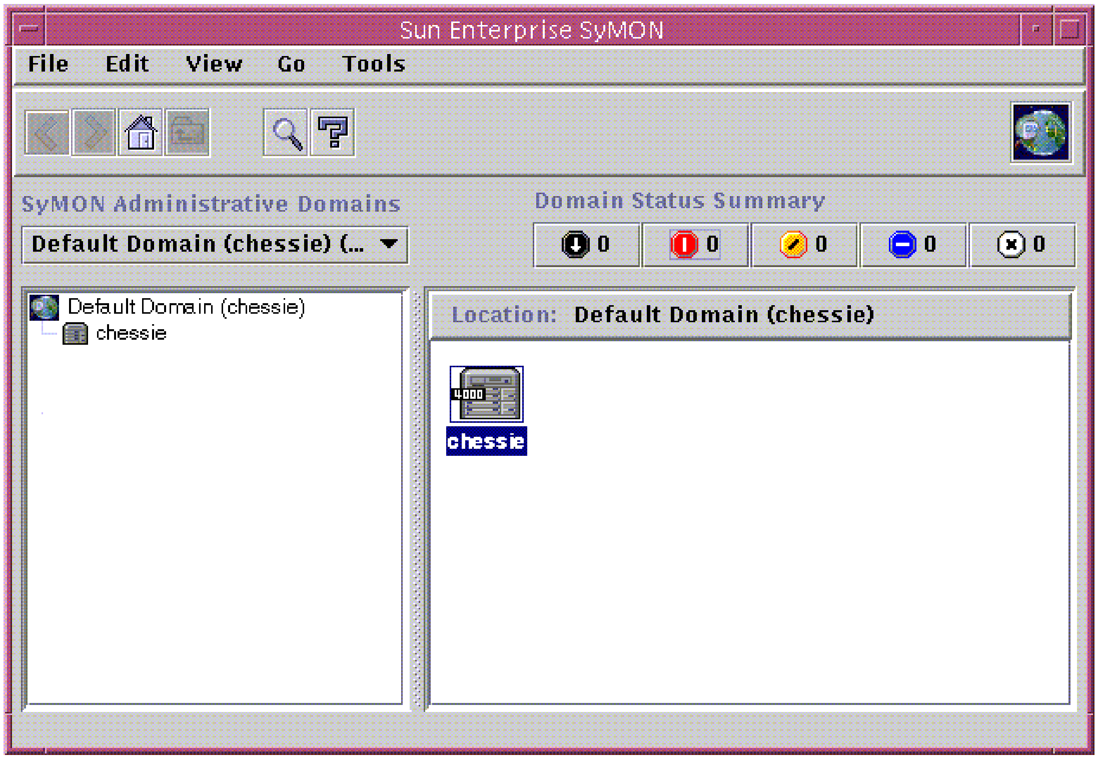
Domain: Default Domain (chessie)

Location	Status Change Time	Status Message

No matches found

Search completed.

FIGURE 8 Refreshed Domain Status Details Window



Default Domain (chessie) (...

Location: Default Domain (chessie)

FIGURE 9 SunMC Domain Console With No Alarms

In the Sun BluePrints Resource Management book (ISBN #0-13-025855-5) we showed how to separate and monitor multiple workloads using SunMC. It is important to use this capability to track application processes and define alarms that will tell you when something goes wrong, for example, if a process goes away or a workload starts to use too much CPU.

Key Performance Indicator Plots

Operations must have access to displays that show daily data for the main components of key systems. These are a great help with problem diagnosis. A simple approach is to log data to a table, import it into a spreadsheet, and plot the key indicators. It is also possible to generate plots on a web page automatically using free tools such as the Multi Router Traffic Grapher (MRTG) and Orca. The disadvantage of these methods is that they do not operate in real time and need to be refreshed manually. Almost all commercial performance tools supply data in real time for display in ways that are suitable for an operations monitoring display.

As an example of the kind of displays that can be useful, a spreadsheet has been developed that reads in data logged to a file using the SE toolkit. Plots should cover the main components of a single system: CPU, disk, network and memory. Plots for these components are shown following:

- CPU utilization, the balance of user CPU time to system CPU time, and the load average (number of runnable jobs) relative to the number of CPUs configured and online.
- Disk utilization for the busiest disk in the whole system, and the average utilization over all the disks in the system. This shows how skewed the load is and highlights the most likely disk bottleneck. Read and write throughput in kbytes shows the data transfer characteristics.
- Network throughput in kbytes can be obtained either by monitoring the TCP stack or by monitoring the network interfaces. Most interfaces report byte level counters nowadays. But many tools just read packet level information, which is far less useful given the wide range of possible packet sizes.
- Memory and swap space can be plotted together, but they need to be interpreted differently. Swap space must always be available so that processes can start and grow, so if it runs out the system will have problems. Main memory is managed by reclaiming once it reaches a low threshold, so it will tend to hover around that threshold. The page residence time is another useful measure, which is explained below.

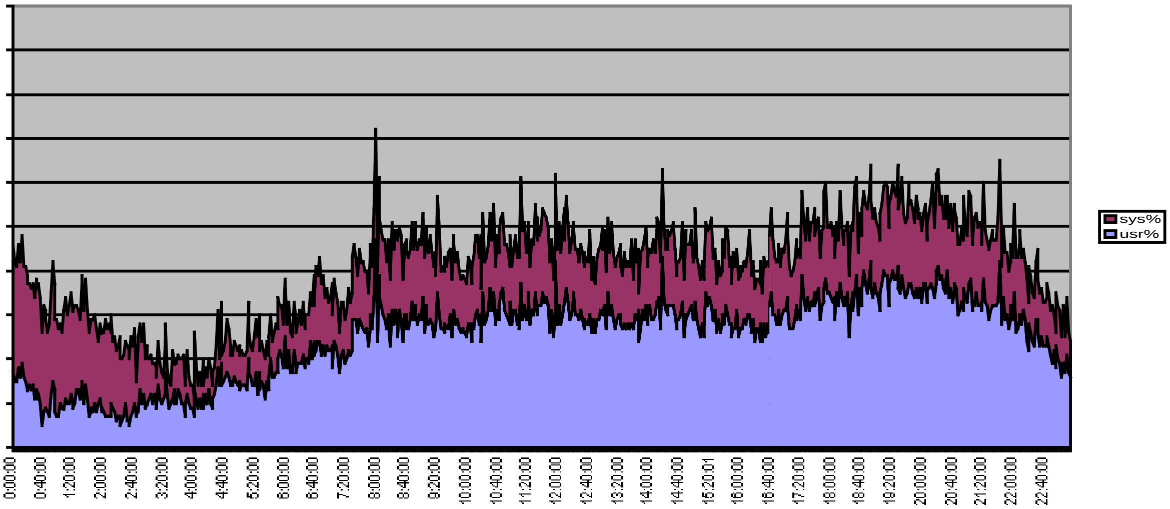


FIGURE 10 CPU User and System Time for a Day

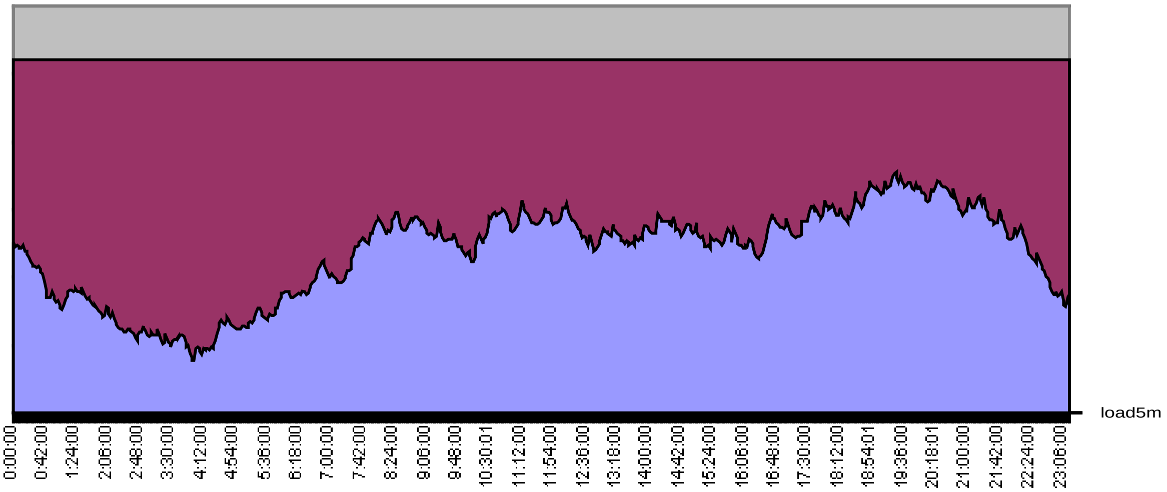


FIGURE 11 Five Minute Load Average and Number of CPU’s Online

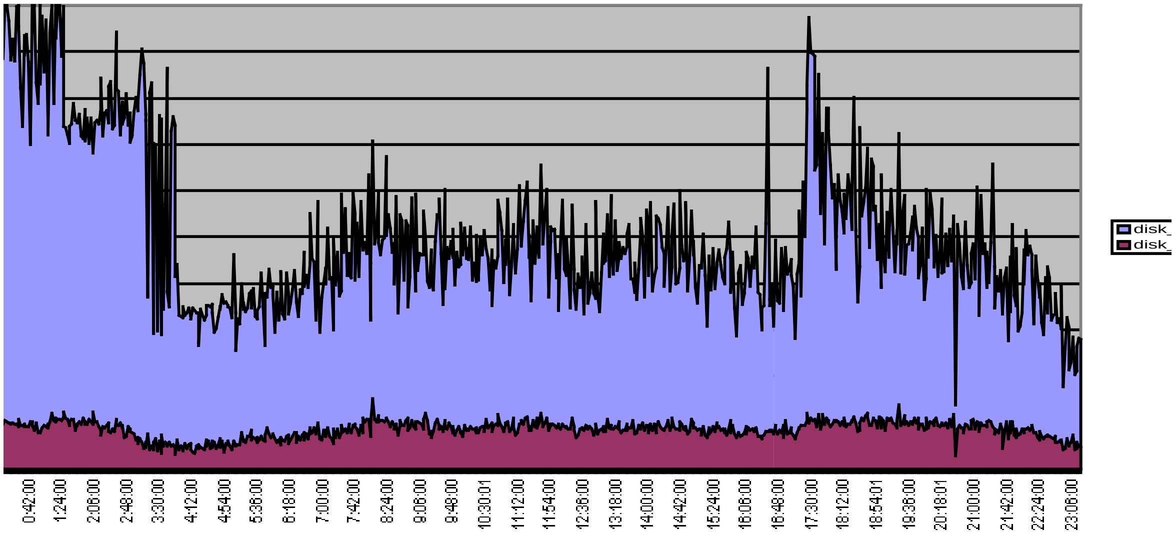


FIGURE 12 Disk Utilization—Busiest Disk and Average Over All Disks

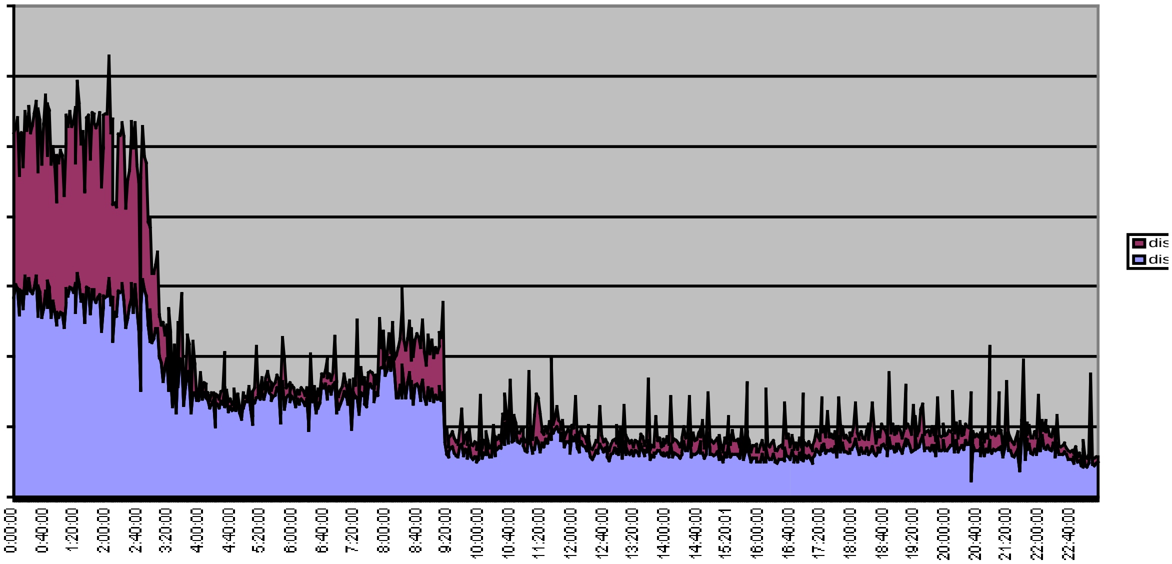


FIGURE 13 Disk Throughput Read and Write Kbytes

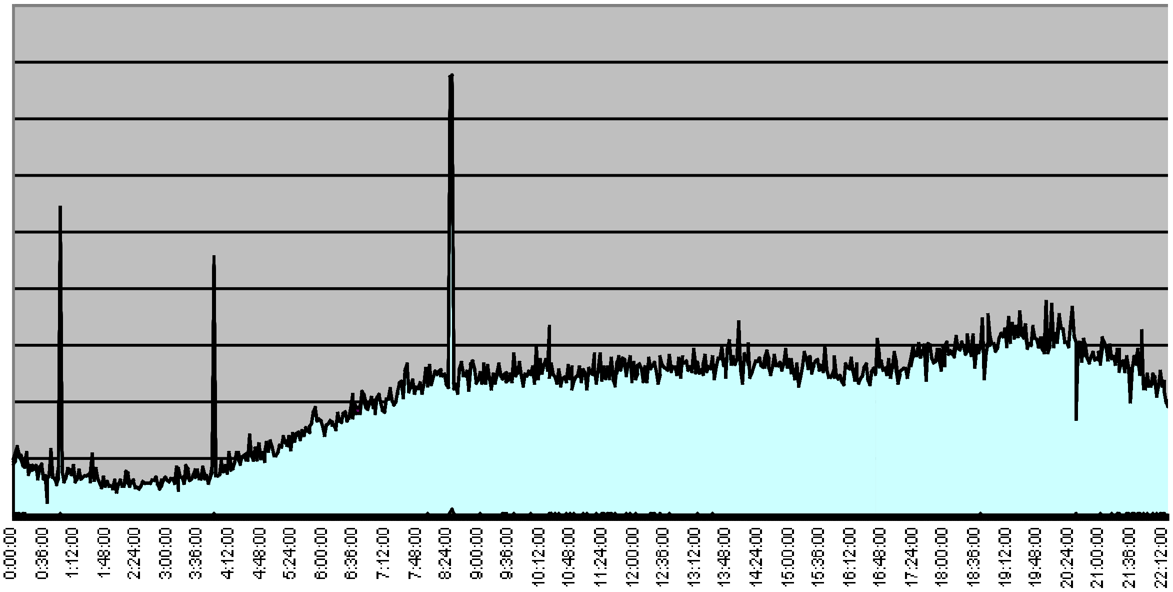


FIGURE 14 Network Throughput Over a Day

The network plot shows a few spikes of high throughput due to file transfers taking place on a local fast network, but its overall shape follows the load average fairly closely in this case. The characteristics of these plots show that this system is active all the time, and it has a load shape that is typical of many internet sites. The system is quiet overnight, gets busy and plateaus during the day, then peaks in the early evening. The CPU and disk plots also show that there is extra system CPU time during the night and a higher disk data rate associated with an online backup taking place. The axis scales have been removed from some of these plots, partly to obscure the source of the example data, but also the absolute values are less important than the shape and the changes from day to day.

The following memory plots show that a regular batch job is creating and deleting a large file, and the memory and swap space is consumed and freed on a regular basis. There are a few times when the page residence time drops to a low level during the backup period. This value is calculated by dividing the page scan rate into a system parameter called handspreadpages. Since the scan rate can be zero, a divide by zero is caught and the residence time is clamped to 600 seconds. Sustained residence times of 30s or less are a problem.

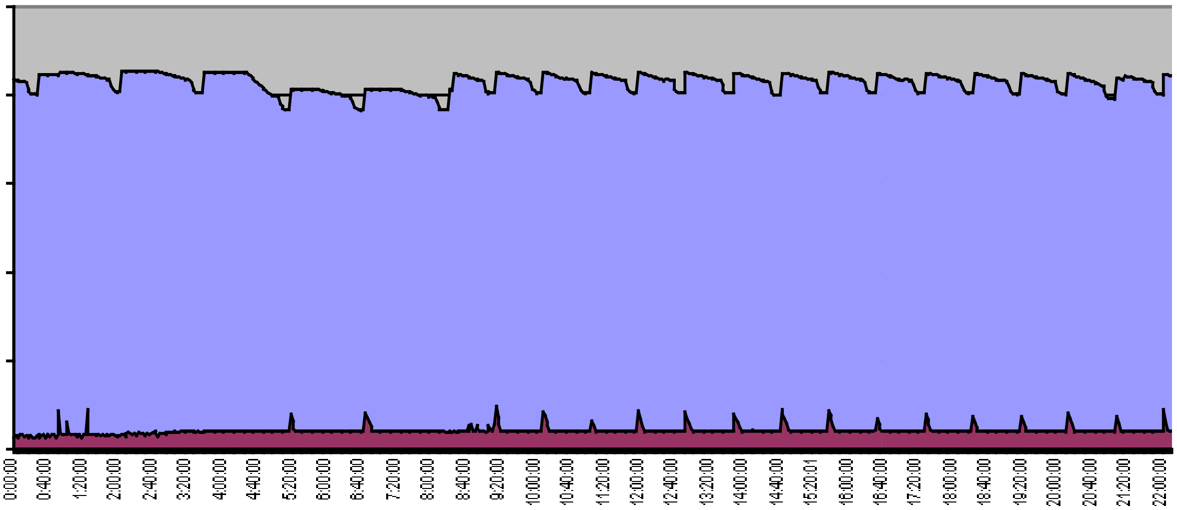


FIGURE 15 Memory and Swap Usage

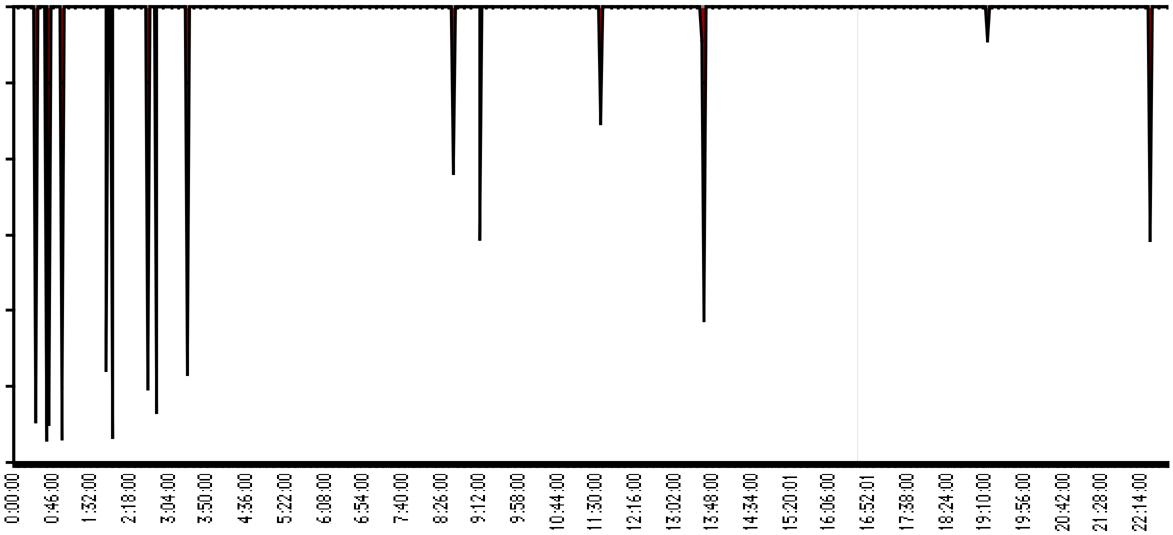


FIGURE 16 Memory Demand Viewed as Page Residence Time

Summary

Operations monitoring concentrates on real-time, alert-based problem diagnosis. It can be implemented using a variety of tools and displays along the lines described here. Its requirements are quite different from those of capacity planners and management reports, although it feeds information to both of them.

Author's Bio: Adrian Cockcroft

The author of Sun Performance And Tuning, Adrian is an accomplished performance specialist for Sun Microsystems and recognized worldwide as an expert on the subject.