# A Ensuring SLAs and SLOs - How to commit Computer Resources

Udo Schmitt, Hewlett-Packard Corporation

| | |
|---|---|
| Street: | Herrenbergerstraße 130 |
| Zip City: | 71277 Boeblingen |
| Country: | Germany |
| Work Phone: | +49 7031 14 8229 |
| Fax: | +49 7031 14 1496 |
| Email: | udo_Schmitt@hp.com |

# A Ensuring SLAs and SLOs - How to commit computer resources
## Udo Schmitt, Hewlett-Packard Corporation

## ABSTRACT

HP has developed a system management software product called HP-UX Workload Manager (WLM). HP-UX WLM is a dynamic front-end tool to HP's Process Resource Manager (PRM). HP-UX WLM automatically adjusts PRM parameters so that system resources are reallocated in response to changing conditions, time of day, and varying workloads.

With HP-UX WLM, applications are partitioned into various workloads (groups). In addition, the system administrator specifies goals for the various groups, priorities of the goals (i.e. all priority 1 goals are met before priority 2 goals are addressed), conditions under which goals are met, and means to measure to what extent goals are being met. As the system runs, the workloads are monitored according to user-specific metrics in order to determine whether they are meeting defined objectives. Additionally, HP-UX WLM can adapt the CPU allocation in response to virtually any other measurable event on your system. CPU utilization, metrics measured by GlancePlus, a GUI-based resource-meter, or other user-provided measurements can be used to tailor resource allocation policies. These policies determine when to give a workload more CPU, and when a workload has excess CPU. WLM's dynamic policy management allows more efficient use of computer resources and gives the user greater control.

This document shall explain the capabilities and benefits of HP-UX WLM in a SAS environment.

## INTRODUCTION - THE PROBLEM

The traditional open systems usage model has been one SAS application or data warehouse resident per server. This has led to a proliferation of servers, which cannot be managed effectively. Generally each server is sized to provide "headroom" for peak capacity and future growth. New applications lead to new servers and more surplus capacity--with no possibility to share the "excess" capacity among applications. The solution to this wasted excess is to run multiple applications on a shared server. Doing so creates a new issue**: resource management on the shared server**.



*Figure 1:* *Handling peak demand for critical applications - a Snapshot from GlancePlus*

Traditional UNIX resource management treats all users and applications as equals, i.e. resources are shared equally among all applications. This leads to the critical situation that--during times of peak resource utilization--important applications may not get the resources they need to run. Users who have bought the same amount of server capacity may find that they get unequal resources. With the proper management of resource allocation, multiple applications can be run simultaneously and effectively on a single server, thereby making it possible to consolidate data centers.

A typical scenario could be:

- Department A runs a SAS data mart with 100 GB.

- Department B runs a SAS data mining application such as CRM with 5 power users doing heavy analysis and 100 standard users retrieving reports and results.

- Department C has all its production data in a data base and runs a SAS application on top of it.

The users and departments are complaining about unpredictable response-times. WLM is not going to make the entire machine go faster - but WLM can make sure some operations go fast at the expense of others. At month's end everyone is getting nervous about being able to present the departments results to upper management on time.

## THE SOLUTION

When consolidating servers, one of the challenges is resource allocation. How do you ensure that one application will not steal needed resources from another? The existing HP PRM tool allows you to specify CPU, real memory and disk I/O bandwidth allocations to make sure workloads get predefined amounts of resources. The server is therefore partitioned into application domains to guarantee the committed computer resources.

Additionally, with HP-UX WLM, the user is allowed to define performance goals for each workload as well as conditions, constraints, and limitations so that critical applications may perform at the expense of less important ones in the event of peak utilization (bottleneck). These parameters (performance goal, conditions, constraints and priority) make up the service level objectives (SLOs) of a group.

An example of an SLO might be

| | |
|---|---|
| **Performance goal:** | response time - report retrieval to be carried out in less than 2 seconds |
| **Performance conditions:** | goal applies Monday-Friday, 9:00 a.m. - 5:00 p.m. |
| **Resource constraints:** | no less than 10% CPU and no more than 50% CPU to be used |
| **Priority:** | 1 |

With such a goal-based SLO, WLM can provide adequate resources dynamically to workloads (applications or users/groups). For example with WLM you might be able to define cascaded goals so that:

- A SAS job which lasts longer than 10 seconds would be moved into another group with lower allocation policy.

- A SAS job which lasts longer than 1 minute might then be moved into still another group with an even lower allocation policy.

In this way you can ensure response time for short reports and short analyses and also continue the longer lasting jobs without

adversely affecting the overall system thoughput.

Or you may have named SAS users, which connect via SAS/CONNECT and therefore are known UNIX users on the system. You can then define goals to be achieved for the specific users or user groups.

The ability to assign an SLO to a workload is a significant enhancement to PRM for several reasons:

- **Higher degree of system utilization**

  By giving each workload only the CPU resource it needs, when it is needed, excess capacity is efficiently shared across all workloads. This means users can minimize server capacity as large amounts of headroom become unnecessary.

- **Prioritization of goals with each workload**

  Whereas PRM treats all workloads with single, equal priority, thereby allocating only the configured entitlement to jobs which may have their peak demands simultaneously, WLM groups allow multiple prioritized goals. This means that when there is a shortage of resources, top priority goals are satisfied before lower priority goals addressed.

- **Service level objectives are met even during peak utilization**

  Along with each goal, the HP-UX system administrator provides a performance monitor which measures the extent to which the goal is being met, exceeded or underachieved. For the newly developed SAS toolkit, HP begins to provide some performance monitors for major applications. Thus, system administrators may only need to configure the performance monitors provided by the SAS or other toolkits to deliver the proper metrics making implementation of WLM in a SAS environment a simple task. With the appropriate metric provided, WLM will then increase CPU entitlements of under performing workloads and decreases the CPU entitlement of over performing ones. CPU entitlements are constrained by the minimum and maximum values specified for the respective goal. Because a system can have many workloads and each workload may have multiple goals with different priorities, HP-UX WLM arbitrates among the different CPU entitlement requests. If resources are limited, the entitlements needed to achieve the higher priority goals will be satisfied first. In the event of a surplus, the excess resources are then allocated to groups with goals having lower priorities. In this way the HP-UX WLM facility regularly calculates new CPU entitlements and then reconfigures PRM. In this way, resources are optimally allocated when needed, applications never overachieve, and yet even during times of peak demand, minimum SLO goals are met.

### TYPES OF SUPPORTED SERVICE LEVEL OBJECTIVES

HP-UX WLM provides automatic PRM configuration maintenance based on the system administrator's predefined performance goals. The software supports the following types of SLOs:

- **Achievement-based**

  These SLOs are suitable for applications like reporting or OLTP. Possible goals might have objectives that specify the response time for a particular transaction or the desired duration of batch jobs. In a SAS environment, it means that the system administrator can create response-time goals for particular reports, or even for particular procedures or data-steps in addition also be differentiated by individual users.

- **Dynamic Policies**

  Set Policies that adapts the CPU allocation to a group of processes. For example, you might want to say that group CRM00 should be given a minimum of 10% of the CPU, and also be given an additional 2% for every process running in

that group, up to a maximum of 30%. You can create policies such as this using any of the metrics reported by GlancePlus.
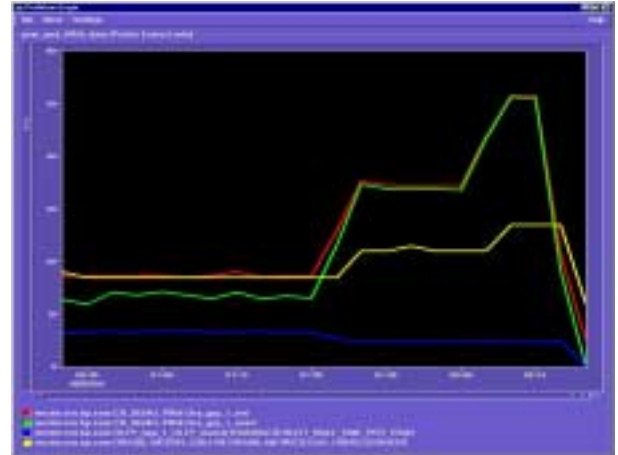


*Figure 2:  Allocate CPU based on the number of users - a snapshot form GlancePlus*

- **Scheduled Resources**

  You can schedule a resource allocation for regularly scheduled jobs. For example, on the 28th day of every month, you can allocate 20% of the CPU for group End_of_Month_Financials.

- **Entitlement-based**

  Entitlement-based SLOs are suitable for workloads that have no measurable performance objective. Such applications are given a fixed CPU entitlement determined by the specified CPU constraints. This SLO type could be used to allocate resources to optional or discretionary work. These entitlements can be enabled or disabled automatically. The metrics provided by GlancePlus, as well as any user-supplied metric can be used in a simple expression to enable or disable an entitlement.
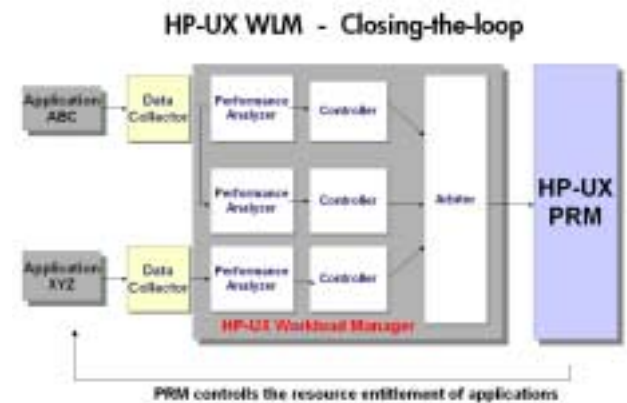


*Figure 3:  HP-UX WLM Architecture*

### EXAMPLES

- **On-Time-Job-Completion**

  There are several ways to accomplish an On-Time-Job-Completion goal. Having the metric for this implemented WLM is able to dynamically change the resource assignments to the workload to ensure the goal.

    o **Using ARM Information**

Any value that is passed to WLM via ARM can be incorporated into performance metrics and then used to define the goals. The On-Time-Job-Completion goal is just one example of how ARM values could be used. Another example could be transaction response times (TxCpu) as reported by the SAS ARM implementation.

o **Progress Estimation**

The programmer might have knowledge about the completion ratio of his job so he can give this specific information to WLM and WLM can act accordingly e.g. long where-clauses could be calculated using the ratio of scanned rows divided by total rows.

o **Profiling**

WLM profiles the queries and will estimate the expected runtime and report therefore the percent complete.

- **Automatic Workload Assignment**
  Based on the credentials of an applications or the user, the job is moved at the beginning of its run into its proper workload group. Identified by:

  o **UID/GID**
    - Carly Fiorina/CEO of HP
    - Marketing Department
    - Regional Sales Manager
    - R&D Division

  o **Application or Function**
    - Based on the application or function name
    - Based on arguments supplied to the application

  o **Class of Processes**
    - C/S Applications
    - Web Applications

- **Express Lane**

  An express lane gives the system administrator an ability to manually increase resource assignments to specific applications by moving them into a different workload group that we call an Express-Lane.

- **Aging Processes**

  As described in the text above, the ability to implement an easy job aging mechanism and have WLM act automatically to ensure system resources for the quick and hot queries. This capability is vital to some businesses. Distinguishing between long running SAS Enterprise Miner algorithms and short web reports helps to utilize the system properly.
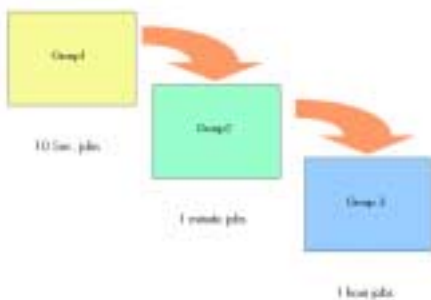


*Figure 4: Cascading Groups*

**THE IDEAL ENVIRONMENT FOR THE HP-UX WORKLOAD MANAGER**

Those who will benefit most from what HP-UX WLM offers are those who:

- run more than one workload concurrently on a server. These workloads could be multiple SAS applications or data servers such as SPDS, a data warehouse and an SAS application, or any other combination of workloads, provided that they are using HP server-class systems running HP-UX 11.0

- have workloads that can be managed by PRM

- have measurable performance goals for at least one workload. Use WLM whenever (a) you want to adapt CPU allocations based on what's going on, or (b) you want to control the way multiple workloads share excess CPU capacity.

- would like to automate changing PRM entitlements in response to measurable data/events.

## CONCLUSION

HP's WLM superior supports and enhances SAS based solutions and gives real benefit to SAS users and the business. HP and SAS are jointly working to enhance this integration. This project is driven by topics issued from the clients and addressed to both companies. If there is anything that is needed for your business do not hesitate to contact us. We appreciate every comment and every idea you are sharing with us.

## REFERENCES

Customer references:

- Staples - Reseller in the USA
- First Union - 6th largest Bank in USA
- Nesbitt Burns - leading Canadian investment firm
- POSCO - worldwide largest company in the steel industry
- Novistar - Accounting and IT Industry

White paper references:

- Paper 298 from SAS Institute at SUGI 24
  *Obtaining Response Time and Service Level Information About Your SAS System Applications*
- HP WLM Home Page
  http://www.unix.hp.com/management/workloadmanagement/wlm/index.html
- HP WLM White Paper
  *Tuning WLM*
  http://www.unix.hp.com/management/workloadmanagement/wlm/infolibrary/wlm_tuning.html
- HP WLM White Paper
  *Workload Performance Monitor Considerations*
  http://www.unix.hp.com/management/workloadmanagement/wlm/infolibrary/wlm_considerations.html

## ACKNOWLEDGEMENTS

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:

| | |
|---|---|
| Author: | Udo Schmitt |
| Company: | Hewlett-Packard |
| Street: | Herrenbergerstraße 130 |
| Zip City: | 71277 Boeblingen |
| Country: | Germany |
| Work Phone: | +49 7031 14 8229 |
| Fax: | +49 7031 14 1496 |
| Email: | udo_Schmitt@hp.com |
| Web: | www.hp.com |

Contact the HP WLM-SAS Integration Specialist:

| | |
|---|---|
| Author: | Joe Presley |
| Company: | Hewlett-Packard |
| Street: | 3000 Waterview Parkway |
| Zip City: | Plano, TX  75080 |
| Country: | USA |
| Work Phone: | +1 972 497 4609 |
| Fax: | +1 972 497 3123 |
| Email: | joe_presley@hp.com |
| Web: | www.hp.com |

Or from the *HP Workload and Resource Management* division:

Email: wlmfeedback@rsn.hp.comWeb: www.hp.com/go/wlm