PRIMEPOWER ARMTech Resource Management Provides a Workload Management Solution

September 2002

A D.H. Brown Associates, Inc. White Paper Prepared for

Fujitsu

This document is copyrighted © by D.H. Brown Associates, Inc. (DHBA) and is protected by U.S. and international copyright laws and conventions. This document may not be copied, reproduced, stored in a retrieval system, transmitted in any form, posted on a public or private website or bulletin board, or sublicensed to a third party without the written consent of DHBA. No copyright may be obscured or removed from the paper. D.H. Brown Associates, Inc. and DHBA are trademarks of D.H. Brown Associates, Inc. All trademarks and registered marks of products and companies referred to in this paper are protected.

This document was developed on the basis of information and sources believed to be reliable. This document is to be used "as is." DHBA makes no guarantees or representations regarding, and shall have no liability for the accuracy of, data, subject matter, quality, or timeliness of the content. The data contained in this document are subject to change. DHBA accepts no responsibility to inform the reader of changes in the data. In addition, DHBA may change its view of the products, services, and companies described in this document.

DHBA accepts no responsibility for decisions made on the basis of information contained herein, nor from the reader's attempts to duplicate performance results or other outcomes. Nor can the paper be used to predict future values or performance levels. This document may not be used to create an endorsement for products and services discussed in the paper or for other products and services offered by the vendors discussed.

TABLE OF CONTENTS

PRIMEPOWER ARMTech Resource Management Provides a Workload Management Solution

Fujitsu provides the ARMTech ShareEnterprise resource management software on its PRIMEPOWER servers, which forms a key element of a complete workload management solution. While Fujitsu's XPAR (Extended Partitioning) hardware partitions are necessary if complete isolation is required between multiple critical applications running on a single server, ARMTech allows administrators to put flexible policies into place. This approach assures applications will receive precisely the resources that they are entitled to on a busy system at the time that they need them.

ARMTech allows multiple, dominant applications to coexist in a single instance of the operating system, either on an entire PRIMEPOWER server, or within XPARs of a partitioned PRIMEPOWER server. ARMTech accomplishes this by enabling the running Solaris Operating Environment (OE) to efficiently allocate system CPU resources to different applications with flexible scheduling policies that mirror the requirements of an organization.

This white paper provides an overview of ARMTech's capabilities. It is one in a series of seven PRIMEPOWER white papers that provide a PRIMEPOWER overview, information on PRIMECLUSTER, the Solaris OE, the PRIMEPOWER system architecture, the PRIMEPOWER SPARC64 V microprocessor, and PRIMEPOWER system management.

POSITIONING WORKLOAD MANAGEMENT TECHNIQUES

Mainframes have long represented the ideal for centralized, high-throughput computing. Over time, they have refined the use of kernel-based workload management techniques to optimize large servers for processing dominant jobs. Historically these were handled by batch processes. Mainframes first addressed the need to isolate application environments from each other through physical partitioning.

Mainframe physical partitions could be resized dynamically, but the process involved a certain degree of operator intervention, since applications had to be quiesced before boundaries could be shifted. This explicit management burden limited the use of physical partitions as a tool to respond to fluctuating workload needs.

To provide greater flexibility, mainframe developers replaced physical partitions with logical partitions (LPARs) about 15-20 years ago. Over the years, mainframe developers verified and refined the isolation between LPARs. They also enhanced LPARs by creating an additional layer of resource management across partitions by speciftyin time-slicing parameters.

Time-slicing functions allowed LPARs to provide a much finer degree of granularity than physical partitions, offering time slices of a single processor, or simultaneous time slices across all shared memory processing (SMP) processors. Moreover, if a time slice could not use its allocation (e.g., because it was waiting for I/O), the system could pass these unused resources along to the next-highest priority partition. Thus, LPARs permitted an extremely flexible setting of shares (allocated portions of resources), priorities, preemption, and other parameters.

Simultaneously, mainframe developers produced sophisticated workload management tools. These tools allowed systems to respond dynamically to fluctuating loads, and were implemented as a kernel function within each of the mainframe's SMP partitions. Thanks to their LPAR and workload management capabilities, mainframes now offer a virtually ideal environment for running multiple, heavy-duty applications simultaneously on a single server. For example, they:

- support logical partitions that provide reliable isolation between applications and a great deal of flexibility in specifying boundaries; and
- offer sophisticated workload management functions within each operating environment.

In the mainframe, the LPAR approach merged these two functions successfully. However, in the UNIX space they remain separate. In fact, in UNIX they meet different needs (see *Table 1: XPAR and ARMTech Compared*). In the UNIX space:

- Physical partitions enable isolation between multiple critical applications running on a single server.
- Resource management software allows multiple diverse workloads to efficiently share a common pool of resources.[1]

TABLE 1: XPAR and ARMTech Compared

	PRIMEPOWER New Models Extended Partitioning (XPAR)	**ARMTech ShareEnterprise Resource Management Software**
Isolation	Multiple workloads run in separate physical address space partitions protected from each other by hardware barriers.	Multiple workloads run in the same physical address space.
Operating System Installation	Each partition runs in a separate operating system instance.	All workloads run in the same operating system instance.
Granularity	Each partition requires at least 1 CPU and 1 GB of memory.[1]	Workloads can consume arbitrary shares of CPUs.
Flexibility	Can change size of partition online.	Can dynamically change share of resources consumed by workload.
Manageability	Assign workloads to partition as if it were a stand-alone server.	Assign resources to workload by policy mirroring the requirements of the organization.

[1] XPAR can achieve single processor partitions on the PRIMEPOWER 900 and 1500 models. The minimum PRIMEPOWER 2500 XPAR configuration contains two processors. For a fuller explanation, please refer to the companion white paper *PRIMEPOWER Server Architecture Excels in Scalability and Flexibility*.

 Copyright 2002 © D.H. Brown Associates, Inc.

To satisfy partitioning requirements, Fujitsu's PRIMEPOWER new models XPAR capabilities provide dynamically adjustable physical partitions between multiple instances of the Solaris OE. The XPAR mechanism allows administrators to run multiple instances of the Solaris OE within a single PRIMEPOWER server. Each instance behaves as if it were running on a stand-alone machine.

"Bullet-proof" barriers between the different environments maintain overall system robustness, so that even the most extreme application failure or operating system crash in one partition leaves the others unaffected. The entire environment, i.e., all partitions, can be managed from a single point.

However, XPARs have a partition granularity of one CPU and one GB of memory[2] (a part of a system board), which may not provide precise enough control over resources for some applications. A more fine-grained resource management mechanism may be needed for some types of applications.

Resource management software allows multiple dominant applications to coexist within a single instance of the operating system, running either on the entire server, or within a partition of the server. Instead of partitioning resources at the granularity of entire CPUs or groups of CPUs, resource-management software allows CPU resources to be guaranteed to users or applications, in terms of precise share allocations.

Once a policy has been input, the system will control the scheduling of work so that minimum allocations are met when the system is fully used and when resource use is optimized. Groups or important applications can thus be assured that they will receive the resources on a busy system that they are truly entitled to when they need them.

Resource management tools also provide the ability to control and reduce the impact of occasional runaway processes that would otherwise take over the whole system, either accidentally or as a deliberate denial-of-service attack. Such sudden usage spikes can be controlled online even before the precise origin of the problem is located and fixed on a more permanent basis. This is a useful troubleshooting capability.

Finally, resource management software can become particularly important in high-availability scenarios, in which a node failure results in insufficient capacity to run both critical and non-critical applications. In those cases, resource management software can ensure that critical applications maintain the share of resources they need to meet business-critical performance requirements.

---

2 XPAR can achieve single processor partitions on the PRIMEPOWER 900 and 1500 models. The minimum PRIMEPOWER 2500 XPAR configuration contains two processors. For a fuller explanation, please refer to the companion white paper *PRIMEPOWER Server Architecture Excels in Scalability and Flexibility.*

Copyright 2002 © D.H. Brown Associates, Inc.

RESOURCE MANAGEMENT SOFTWARE BUSINESS BENEFITS

Physical partitions and resource-management software each play useful roles in enterprise IT infrastructures, particularly in the increasingly critical area of server consolidation. The proliferation of servers resulting from the client-server model of distributed computing – coupled with the seductively low acquisition costs of PC-based LAN servers – created a heavy burden of IT administration with a distinct impact on system management costs.

In today's environment, IT administrators increasingly look to server consolidation as a remedy for the administrative overhead and excess computing resources that were side effects of distributed computing. Centralization has once again become fashionable, and the industry is shifting its attention to applying mainframe-like workload management techniques in other environments.

By aggregating multiple applications on a single, large machine or a centrally managed cluster of machines, server consolidation allows management of these servers with greater economies of scale, potentially lowering costs. As systems such as Fujitsu's PRIMEPOWER have increased their SMP range to as many as 128 processors, they have become attractive options for consolidating applications and server functions from multiple smaller (four- and eight-way) servers now predominantly used for managing departmental and branch functions. Such efforts require redeployment of critical applications, which have often been designed to dominate the resources of a dedicated server.

However, administrators implementing server consolidation in large organizations quickly find that it is not sufficient to merely transplant departmental applications onto a single system, even if there is enough hardware resources for the server on which consolidation takes place. Users increasingly view IT administrators as equivalent to internal application service providers (ASPs) who must adhere to specific service level agreements (SLAs).

As a result, users expect IT managers to provide flexible and guaranteed service levels for widely fluctuating workloads. Accountability can become a principal issue, because it provides a basis for negotiation between administrators, who need to explain the performance levels they provide, and users, who need to justify their service-consumption budgets.

When multiple departments with separate budgets share a single physical server, departments may need assurances that they will receive the resources they paid for. In the past, with each department owning its own resources, even if those resources were underused much of the time, each department felt it was "in control of its own destiny." Now that departments are turning over the management of these resources to centralized IT managers, they expect to continue receiving consistent and predictable levels of service.

Copyright 2002 © D.H. Brown Associates, Inc.

The need for flexible application service levels also surfaces in web-based business models, which tend to involve fast-growing, unpredictable workloads. In this space, customers, partners, and suppliers access IT infrastructures, and the quality of their experience directly affects both revenues and online brand image.

Effectively managing workloads on large SMP servers thus requires specialized tools that enable administrators to meet the following objectives:

- guarantee service levels;
- gather detailed information on usage and capacity;
- anticipate changes in workload;
- understand how applications behave under load;
- implement usage policies; and
- maximize their ability to make system changes flexibly.

To meet these requirements, administrators need vastly greater control over application behavior. They must thoroughly understand how applications behave under load, must be able to anticipate what expected loads will be, and must develop policies governing user and department entitlements.

More important, administrators must be able to change resource allocation very rapidly and with maximum precision, employing scripts, traditional system management tools, and other components of their IT infrastructures. Resource management software provides the solution to this challenge.

RESOURCE MANAGEMENT SOFTWARE OPERATION

Resource management tools work within a single operating system instance to effectively manage massive, constantly changing workloads so that multiple dominant applications can coexist in a single operating system instance. The tools work by efficiently allocating system resources to different applications with flexible scheduling policies.

These resource management functions effectively override the operations of the default UNIX scheduler, taking into consideration customized policies instead. Although UNIX has always offered simple resource management tools such as disk quotas and the *nice* command to set application priorities, UNIX system developers have recently invested in far more sophisticated capabilities, modeled in part after long-standing mainframe functions.

Fujitsu provides the ARMTech ShareEnterprise resource management software on its PRIMEPOWER servers, which can provision CPU resources to processes running within an XPAR. ARMTech can be used to offer different service levels to particular classes of users. A low-priority user would receive a smaller than average allotment of resources, and a medium-priority user would receive average resources, while a high-priority user would receive above average resources.

ARMTech accomplishes all this by allowing administrators to define a "policy path" that mirrors the requirements of an organization. Administrators define a policy that specifies the allocation of CPU resources a particular resource consumer should receive. Administrators start by defining classes of resource "consumers" with rule-based memberships. A resource consumer is any entity that directly or indirectly consumes a system resource. These can include users, groups, or applications, which can represent arbitrary groups of UNIX processes. The policy path determines the relationship between resource consumers. Administrators set up policy paths by assigning "shares" of resources to the consumers. Shares entitle a consumer to CPU resources in proportion to the sum of the shares held by all contesting resource consumers. During normal operation, ARMTech continuously monitors the resource usage of all processes running on the system, and schedules them to resources using a "fair share" scheduling mechanism so that the defined policy is achieved.

To accomplish this, ARMTech assigns a process to a particular resource consumer by matching a characteristic of that process with that of the resource consumer. For every resource consumer, ARMTech also records current and historical usage. ARMTech then uses this history and an administrator-defined policy to calculate entitlements for active processes on the server.

Administrators can change the policy dynamically so that if an application is not receiving the necessary service levels, the policy can be continuously readjusted until satisfactory performance is reached. Because administrators specify the entitlement of each resource consumer in terms of shares, rather than percentages, the resources assigned to a consumer will always be maintained relative to the total resources available.

The more resource consumers there are on the system, the smaller the entitlement each consumer's shares will represent. Thus, administrators do not have to concern themselves with the number of resource consumers defined on a given system at any particular time. Regardless of the total number of resource consumers, ARMTech will always enforce a policy that ensures the more deserving resource consumers receive a greater percentage of the resource. This allows new applications to be added to a server any time without having to redefine the existing policies.

ARMTech also provides several additional control mechanisms that can be used to enforce certain boundary conditions for CPU resources, including absolute minimum resource requirements specified with CPU "reservations," and absolute maximum resource allocations, using CPU "caps." The ARMTech resource reservation mechanism allows administrators to specify an absolute fixed percentage of CPU resources that a consumer will receive. Unlike shares, reserved CPU resources are not adjusted dynamically in response to changing demands.

Copyright 2002 © D.H. Brown Associates, Inc.

A consumer who reserves CPU resources will receive those resources at minimum. ARMTech will only provide CPU resources to share-using consumers after it has provided reserved CPU resources to consumers who have requested them.

Conversely, in certain application environments it might be necessary to restrict the consumption of CPU resources to a maximum. ARMTech provides such a limiting facility with its CPU caps feature. When a CPU cap is set for a consumer, ARMTech prevents consumers from ever exceeding the amount of CPU resources set by the cap.

ARMTech allows all of these functions to be configured in several ways, including a graphical user interface (GUI) and a comprehensive command line interface (CLI). The ARMTech GUI allows administrators to manage the resource management functions in a user-friendly fashion, employing a convenient point-and-click interface that visualizes the resources being consumed by various applications with a graph view. The GUI can also be accessed remotely over the web or enterprise networks, allowing administrators to quickly get an overview of load conditions and respond with reconfigurations if necessary, regardless of their location.

The CLI provides a way for more experienced users to configure ARMTech. The CLI is also more amenable to scripting, which allows administrators to put together customized procedures for dynamically adjusting policies in response to certain conditions.

REVIEW OF KEY POINTS

This white paper provides an overview of Fujitsu's PRIMEPOWER new models ARMTech ShareEnterprise resource management software. This paper describes how it meets the needs of the business-critical IT environment for fine-grained resource management. PRIMEPOWER's ARMTech ShareEnterprise software features and attributes are another part of the strong foundation that is the basis for PRIMEPOWER's ability to deliver a high quality of service to the IT environment.

As a result of ARMTech ShareEnterprise resource management software, and additional technologies discussed in the other white papers in this series, it is clear that PRIMEPOWER is a short-list candidate for single SMP server or clustered SMP server operation in the business critical IT infrastructure. (Additional information concerning PRIMEPOWER's ARMTech ShareEnterprise and other PRIMEPOWER attributes can be found on the Fujitsu websites.)

Copyright 2002 © D.H. Brown Associates, Inc.