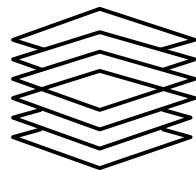


*Fujitsu's PRIMECLUSTER
Facilitates High-Value
IT Infrastructures*

September 2002



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

Fujitsu

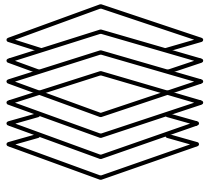
This document is copyrighted © by D.H. Brown Associates, Inc. (DHBA) and is protected by U.S. and international copyright laws and conventions. This document may not be copied, reproduced, stored in a retrieval system, transmitted in any form, posted on a public or private website or bulletin board, 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

INSIDE PRIMECLUSTER.....	3
<i>Figure 1: PRIMECLUSTER Modular Software Architecture.....</i>	<i>3</i>
<i>Table 1: PRIMECLUSTER Module Requirements.....</i>	<i>4</i>
HA SERVER	4
PARALLEL SERVER	5
SCALABILITY SERVER.....	6
ENTERPRISE EDITION.....	7
REVIEW OF KEY POINTS	7



Fujitsu's PRIMECLUSTER Facilitates High-Value IT Infrastructures

Fujitsu's PRIMEPOWER servers equipped with the Solaris Operating Environment (OE) and Fujitsu's SPARC64 V microprocessor and other technical advances achieve the highest levels of availability, flexibility, performance, scalability, and security. PRIMEPOWER nodes are combined in Fujitsu's fourth-generation PRIMECLUSTER cluster designed with state-of-the-art clustering software modules. PRIMECLUSTER allows the highest degree of information technology infrastructure. The CIO, CTO, or other IT executive responsible for delivering the optimum in cluster scalability, manageability, or parallel database operation will find PRIMECLUSTER equally facile in these critical tasks.

Placing PRIMECLUSTER on a short list for purchase does not call for the IT infrastructure manager to understand PRIMECLUSTER in detail. However, an understanding of how the PRIMECLUSTER architecture is set up to ensure cluster availability, scalability, and manageability deepens the IT executive's appreciation of the technology and its value.

This white paper provides this understanding. It is one in a series of seven PRIMEPOWER white papers that provide an overview of the new PRIMEPOWER offerings, PRIMECLUSTER, the Solaris OE, ARMTech resource management software, the PRIMEPOWER system architecture, the PRIMEPOWER SPARC64 V microprocessor, and PRIMEPOWER system management.

PRIMECLUSTER is a single, modular software-based product. Its modules can be mixed and matched to meet cluster node, storage, and network infrastructure needs for high availability, scalability, parallel database operation, and manageability. PRIMECLUSTER is based on Fujitsu's long history of UNIX-based clusters with mainframe-based technology (SynfinityCLUSTER) and mature cluster systems offered by Fujitsu Siemens Computers (Reliant UNIX, which is rooted in technology derived from Pyramid Technologies).

Note that while PRIMECLUSTER is discussed in this paper with reference to Fujitsu's PRIMEPOWER computers and the Solaris OE, PRIMECLUSTER is not limited to this environment. Indeed, PRIMECLUSTER is operating system, hardware platform, and interconnect technology independent. For example, PRIMECLUSTER modules also run on IA Linux.

PRIMECLUSTER modules provide four key services to ensure PRIMECLUSTER clustering benefits. (See *Sidebar 1: PRIMECLUSTER Clustering Benefits*.) These services include,

- Clustering Services

- Global Disk Services
- Global File Services
- Global Link Services

All four services address essential needs. For server node high availability, Clustering Services provide software for rapid node failover and cluster scalability. For storage high availability with SAN capability, Global Disk Services provide a disk volume manager and software mirroring. For further storage high availability with SAN capability, Global File Services provide an all-encompassing file system. For network high availability, Global Link Services provide a redundant network.

All of these capabilities are designed to take advantage of the new architecture and high-availability features in the PRIMEPOWER new series Models 900, 1500, and 2500. These computers operate on Solaris and employ Fujitsu's SPARC64 V microprocessor, which complies with the SPARC International open specification. (These capabilities are discussed in detail in the other white papers in this series.)

Sidebar 1: PRIMECLUSTER Clustering Benefits

PRIMECLUSTER's capabilities provide the IT infrastructure with numerous advantages in today's volatile business environment.

- ⇒ PRIMECLUSTER supports the fast, cost-effective capture of new market opportunities due to its flexible, module-based software architecture that can handle new applications and increased workloads.
- ⇒ PRIMECLUSTER allows the quick change of business strategies because of its fast accommodation of new applications and changing application workloads and growth.
- ⇒ PRIMECLUSTER minimizes user downtime and maximizes user quality of service thanks to its PRIMEPOWER node and high-availability features and functions.
- ⇒ PRIMECLUSTER reduces cluster cost of ownership (especially the "people cost" component) due to its extensive use of GUIs, Wizards, and a single-system image (SSI).
- ⇒ PRIMECLUSTER offers the highest available performance as demonstrated by the PRIMEPOWER node's continued capture of industry-standard benchmarks such as TPC-C and SAP.
- ⇒ PRIMECLUSTER is ideal for server consolidation due to its ability to reduce costs and add redundancy with minimal effort.
- ⇒ PRIMECLUSTER enjoys industry-leading scalability due to its use of state-of-the-art components such as the SPARC64 V microprocessor and a supercomputer-based crossbar switch in its PRIMEPOWER nodes.
- ⇒ PRIMECLUSTER uses industry-standard Solaris and the SPARC International-compliant SPARC64 V microprocessor. As a result, the system offers maximum "openness" and consequent lower costs because of the availability of trained personnel.
- ⇒ PRIMECLUSTER, with its dedicated high-speed internode data transmission facility, is tuned for optimum performance with industry-standard Oracle9i RAC (including Cache Fusion).
- ⇒ PRIMECLUSTER provides tight integration with PRIMEPOWER hardware features such as partitioning, dynamic reconfiguration, error detection, and correction.
- ⇒ PRIMECLUSTER offers multiple industry-leading features in one package to ensure the highest performance, lowest cost of ownership, and ease of use including multiple cluster configurations, auto recognition of resources, single-system image, load balancing, Java-based interface, instant failure detection and lowest Oracle failover time. All of these benefits are available in an offering that is available worldwide. Hence, multinational corporations may enjoy hardware, software, and personnel economies of scale at all of their locations.

INSIDE PRIMECLUSTER

To provide a modular capability to meet the needs discussed above, Fujitsu provides a cost-effective solution, which includes the specification of PRIMECLUSTER software components based on the availability of four PRIMECLUSTER products. These products can be used as required in the web, application, or database tier of an IT infrastructure:

- HA Server
- Parallel Server
- Scalability Server
- Enterprise Edition

The modular software architecture of PRIMECLUSTER is shown in Figure 1. Note that a WebView module provides a Java-based GUI administrative view so that PRIMECLUSTER is managed from a single console. In addition, the modular structure, built on the Cluster Foundation (CF) module, allows the easy addition of future modules to accommodate other cluster needs as they arise.

The particular modules required for the four PRIMECLUSTER products are shown in Table 1. The Enterprise Edition requires all eight available modules. The names and functions of these modules are defined in the sections that follow. All of the servers require the Cluster Foundation (CF) module. Except for the Scalability Server, all of the servers require the CF, Global Disk Services (GDS), Global File Services (GFS), and Global Link Services (GLS) modules.

For the most part, to maximize efficiency, these modules operate at the kernel level and are called through an operating system's system call interface. Kernel-level modules are not subject to the same issues as user space cluster modules. For example, a run-away user process cannot prevent the core cluster services from continuing to operate. Furthermore, since the kernel modules can operate in what amounts to real time, PRIMECLUSTER reacts faster. For example, node failure recovery can be reduced to tens of seconds. This figure includes application database recovery, mirrored disk state, and the application.

FIGURE 1:
*PRIMECLUSTER
Modular Software
Architecture*

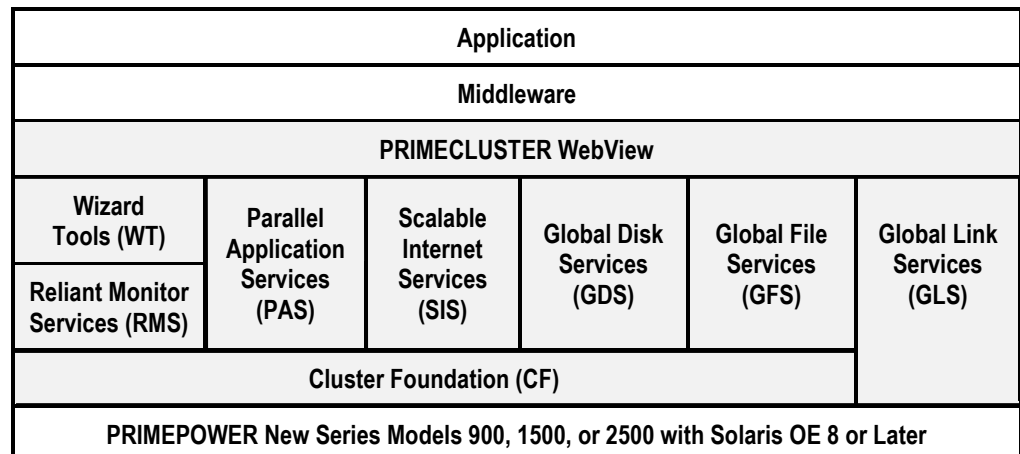


TABLE 1:
PRIMECLUSTER
Module Requirements

PRIMECLUSTER Type	Module							
	CF	RMS	WT	PAS	SIS	GDS	GFS	GLS
HA Server	R	R	R			R	R	R
Parallel Server	R			R		R	R	R
Scalability Server	R				R			
Enterprise Edition	R	R	R	R	R	R	R	R
Key: R = Required								

HA SERVER

The HA Server requires the use of the CF module, the Reliant Monitor Services Module (RMS), and the Wizard Tools (WT) module. As shown in Table 1, the GDS, GFS, and GLS modules are also required. The RMS module delivers application high availability through resource monitoring (by means of “detectors” in the PRIMECLUSTER nodes, network, and applications) and failure detection. This module also provides local recovery and failover services.

The RMS module equips PRIMECLUSTER with the ability to choose the cluster architecture environment’s most suitable configuration including,

- multiple node backup;
- multiple node failover;
- cascading failover;
- selective failover;
- manual or automatic failover; and
- local recovery.

Used in conjunction with the RMS module, the PRIMECLUSTER WT module and application Wizards allow the maintenance of high availability for specific applications. Application Wizards provide predefined detectors, application scripts, and integration with the RMS module for specific commonly used applications such as Oracle, SAP, and EMC SRDF. Fujitsu or third parties may develop custom Wizards. Wizards also simplify the development of the recovery and failover scripts for cluster applications.

The GDS module provides storage management (through a PRIMECLUSTER-based SAN if required). Among its other functions, GDS provides consistent, cluster-wide device name capability, (including intuitive name capability), device access control, and the administration of the cluster SSI. In PRIMECLUSTER, SSI includes a cluster-wide file system, a cluster volume manger, and a single cluster IP address among other functions.

GDS is also the mechanism by which a software cluster volume manager is delivered to PRIMECLUSTER so that it may access more than two nodes with

different RAID capabilities (RAID 0, 1, 0+1) which are either local or shared. In addition, the GDS allows disk software partitioning (up to 256 partitions per volume) and root-disk mirroring. Finally, GDS allows backup and restore by means of detaching and reattaching a mirror disk (or part thereof). This is the equivalent of a file system snapshot. Recovery is speeded after a panic through log-based recovery.

The GFS module directly provides a full file-sharing capability for two or more PRIMEPOWER cluster nodes running multiple user applications in a SAN environment. This file sharing allows the reduction of file-system overhead that is due to the need for the cluster LAN to be used for data exchange in a system without a cluster file system. It also allows eased access to a cluster's dedicated NFS file server(s) (if present) and eases the task of making the cluster as scalable as possible by further software overhead reduction.

The GFS file system is highly available and uses log-based recovery. It supports redundant meta data and provides greater data integrity than the NFS file system. In addition, the file system provides high performance. It allows direct data access from each PRIMECLUSTER node (the interconnect is not required). For high-performance computing, up to 32 TB can reside in the GFS file system.

To operate, the GFS uses a meta data server on one PRIMECLUSTER node (with failover to another node, if required). This server has access to the cluster meta data and the data logs so that it may manage and arbitrate inter-node data access.

The GLS module provides PRIMECLUSTER network redundancy. It accomplishes this task by allowing virtual IP addresses to be allocated to multiple network interface cards (NICs). Thus, applications can continue even in the face of a NIC, cable, or router failure. GLS also allows an application to failover to another node in the unlikely event that all of a node's NICs fail. Through the use of GLS services, a node LAN can failover, LAN loads can be balanced, and LANs can be bundled to improve traffic flow for a particular application.

PARALLEL SERVER

PRIMECLUSTER Parallel Server provides the cluster operating environment for parallel databases such as Oracle9i RAC. For optimum use of Oracle's unique small message and lock caching scheme, Cache Fusion, PRIMECLUSTER employs a private, low-latency, message-passing protocol.

With Parallel Server, the database is available even if a node fails. Moreover, database parallel access is synchronized, and the preferred architectural concept of a single database design with a database instance on each cluster node can be maintained. To support parallel database products such as Oracle9i RAC, the Parallel Server requires the CF module, the Parallel Application Services (PAS) module, the GDS module, the GFS module, and the GLS module.

The PAS module provides the redundant, high-speed interconnect communications needed by Oracle and other parallel applications. It enhances the previously discussed CF features and functions for parallel applications (e.g., cluster membership). The PAS module also contains an application programming interface (API) that allows the accommodation of memory-to-memory transfers over InfiniBand. This virtual interface is not part of the standard issue PAS module but is available.

SCALABILITY SERVER

The PRIMECLUSTER Scalability Server provides service to client requests (via the Internet or direct connections). It is configured with a gateway node (and perhaps a gateway node backup). Various other cluster nodes (service nodes) provide the responses to the user requests.

In the Scalability Server configuration, PRIMECLUSTER provides dynamic load balancing for TCP/IP-based multi-instance applications. This load balancing provides scaling capabilities for these applications, which are often CPU intensive. To provide high availability for this scaling, the load can be redistributed in case a node fails. Persons responsible for the IT infrastructure use Scalability Server to accommodate unpredicted or time-varying loads in their business environment.

To do its job, the Scalability Server requires the basic CF module and the Scalable Internet Services (SIS) module. The CF module provides software to accommodate basic cluster features (for up to 64 nodes) such as cluster membership services. It also provides cluster lock management and event notification as well as redundant node interconnect management. Other CF functions include internode communications across a redundant private PRIMECLUSTER network. All of these functions are easily accommodated. For example, a single command adds a node to a PRIMECLUSTER.

For efficient operation, the CF module is implemented using a dedicated, low-latency protocol rather than the slower and more overhead-laden TCP/IP stack. Nevertheless, it supports Ethernet and Gigabit Ethernet, and is ready for InfiniBand operation. The overall efficiency of the CF module allows the detection of a failed node within ten seconds. This considerably speeds, for example, the recovery of Oracle9i RAC and restoration of user service should a node failure occur.

The CF module also allows a novel node shutdown facility, which ensures that the node being taken down is really down and that there has not been a network problem instead. This eliminates most of the cases of the so-called split-brain wherein cluster nodes do not know for sure if a node is down and more than one node contends for cluster control. Such a situation can result in data corruption, which is to be avoided at all costs.

The SIS module enables dynamic load balancing across cluster nodes. To accomplish this task, the SIS module distributes connections based on a number of load-balancing algorithms (round robin, least system load, weighted system load, and more). SIS converts the cluster IP address to internal IP addresses on the different cluster nodes in PRIMECLUSTER. If a node goes down, connections on surviving nodes are not affected and the load can be redistributed. All this can be GUI controlled.

ENTERPRISE EDITION

The Enterprise Edition uses all of the described modules. This server is the high-end PRIMECLUSTER offering and performs all the required cluster operations for the business-critical IT infrastructure.

It is possible, for example, to put together a two-node PRIMECLUSTER with the recently introduced PRIMEPOWER Model 2500 server. Each of these servers can run a previously unheard of 128 SPARC64 V microprocessors, 512 GB of PC266 ECC RAM, 320 PCI I/O slots, and 15 hard partitions with dynamic degradation and dynamic reconfiguration capability. Such a cluster exceeds the capability of anything else available in the industry.

REVIEW OF KEY POINTS

PRIMECLUSTER is a robust, flexible, module-based solution that addresses IT infrastructure cluster requirements. IT executives responsible for this infrastructure can today purchase what they need to meet their requirements and be confident that they can meet future cluster growth requirements with minimal operational effect. PRIMECLUSTER is also a symmetric software architecture. Through this design, all cluster information is replicated across its nodes and single points of failure (SPOFs) from software are avoided.

There are many other PRIMECLUSTER benefits detailed in this white paper, which has briefly described PRIMECLUSTER's capabilities for software module-based clustering. PRIMECLUSTER provides both state-of-the-art and industry-standard clustering features and functions that directly benefit the IT infrastructure and its ability to deliver a high quality of service in a cost-effective manner.

The combination of PRIMECLUSTER and PRIMEPOWER nodes provides availability, scalability, performance, openness, low cost of ownership, and a state-of-the-art architecture second to none among competitive offerings in the OEM cluster community. As a result, PRIMECLUSTER is a short-list candidate for any infrastructure tier that seeks a cluster architecture. Further information concerning PRIMEPOWER attributes can be found in the other six white papers in this series, and on the Fujitsu websites.