Executive Summary

The notion of RASIS (Reliability / Availability / Serviceability / Integrity / Security) constitutes an important consideration for cloud service providers and enterprise datacenter operators alike. In a consolidated or virtualized environment, a hardware malfunction can bring down hundreds of virtual machines and dozens of customers of departmental applications.

Dynamic Reconfiguration, a capability available in the mission critical Fujitsu* Server PRIMEQUEST* 2800E2 / 2400E2 running Red Hat* Enterprise Linux* 7 (RHEL 7), significantly reduces the likelihood of such a catastrophic event. This feature allows the hot add and remove of system boards containing CPUs and memory, or I/O units containing PCIe* slots without the need to shut down the operating system. The built in fault tolerance in the PRIMEQUEST servers running RHEL allows the system to continue operating after certain failures such as recoverable hardware faults. Dynamic Reconfiguration allows taking a faulty unit offline and replacing it with a healthy one without unplanned downtime.

Operators in highly consolidated environments run servers with large numbers of virtual machines. These servers need to be highly reliable and fault tolerant because of the high cost of downtime for such a large application footprint. Cloud operators providing infrastructure as a service (IaaS) face a similar situation when they run hundreds of virtual machines on a standard 2-socket server machine. Unfortunately, when a hardware failure occurs, although an unlikely event, it brings down the applications running with it. Operators try to reduce liability by putting disclaimers in their service contracts. While doing so may reduce legal obligations in the short term, the reduced service levels may put the service provider in a competitive disadvantage over another provider able to fulfill these expectations.

Mission critical operators have no such options. They must offer uncompromised quality of service (QoS) using ultra-reliable infrastructure solutions. Dynamic Reconfiguration is a defining feature that makes possible to differentiate PRIMEQUEST servers running the latest version of RHEL from competing offerings through sophisticated fault monitoring and
recovery. The features help mission critical operators reduce unscheduled downtime.

**Use Case Context**

**Transparent Recovery from Hardware Failures**

Hardware failures, although infrequent, do happen. The goal of a highly available system is to enable applications to continue running in the presence of hardware failures. System boards in a PRIMEQUEST 2800E2 / 2400E2 series carry two CPUs and their corresponding memory.

Mirrored memory in a PRIMEQUEST system board allows applications to continue running normally even if one of the mirror sets suffers an unrecoverable memory error and Red Hat Enterprise Linux® 7 delivers the OS support to make the workload migration seamless. At this point a second failure would result in the failure of the system board.

Architecture of PRIMEQUEST can remove and add new system board or I/O unit without system stop, when recoverable faults occur in SB. It is called Dynamic Reconfiguration or DR. As with the memory mirroring feature, RHEL has to detect failed I/O and route the traffic accordingly. The system may experience a short interrupt while the new subsystem comes online. However the operating system and applications continue running with no loss of state.

**On the Fly System Upgrades**

Dynamic Reconfiguration also allows operators to add system boards and I/O units while the system continues operating in anticipation of demand on the machine. The main advantages of this capability are:

- No need to bring a system down outside the maintenance schedule and only because a new subsystem needs to be plugged in.
- No need to reboot the system because the operating system detects the new hardware and adds it to the running configuration.
- Deferred capital spending: the new modules can be purchased when needed, not when the machine is first installed.

**On the Fly Resizing Across Partitions**

The next step in capability is a simultaneous resizing of two partitions. This double operation would grow one partition at the expense of another. This capability allows growing a partition running a critical application at the expense of another partition running a less critical application. A module to be reconfigured may include either a system board or an I/O unit. The system administrator could utilize advanced

---

**Figure 1.** Hot Replacement of System Board

**Figure 2.** Hot Replacement of I/O Units
RHEL capabilities such as control groups (cgroups), to set resource priorities for processes, applications or virtual machines thus ensuring a seamless transition of the workload.

**Actual Use Case**

A large enterprise in Brazil deployed a PRIMEQUEST machine running Red Hat Enterprise Linux to fulfill exacting requirements:

1. **Scaling of hardware resources to meet application demand requirements.** Doing so involves moving computing resources from one partition that can do without “shutdown” to another. These allow running the appropriate ERP subsystems at end of month to close books on time, and scaling back after the demand peak.

2. **Run seasonal applications.** This involves carving out resources from a running machine to make room for the new application and releasing them after the demand has been fulfilled.

3. **Hot Upgrade on most critical applications.** On this approach they migrate non-critical applications to another platforms and use its computing resources to increase the mission-critical applications.

These machines run mission critical applications such as data warehousing, point-of-sale billing systems and web portals. The customer is exploring using PHP (PCI Hot Plug) functionality to add or remove PCI cards (HBAs and LAN cards) from one machine to another. Doing so allows adding extra storage or network bandwidth without shutting down the application or rebooting the partition or the whole machine.

**Figure 3.** Hot Addition of System Boards and I/O Unit

**Figure 4.** Reconfiguration using Dynamic Reconfiguration
Supporting Business Continuity
From the perspective of business continuity planning, there are two key parameters after the interruption of a business process, the recovery time objective (RTO) and the recovery point objective (RPO). RTO is the expected time to restore the business process. Data can be recovered only up to the point of the most recent backup. This defines the return point objective or RPO. Everything else, except for the pending transactions is lost. Operators must decide the acceptable maximum data loss after the business process is restored. Data can be restored up to the last backup after a disaster. This defines the “grain” for data recovery. More frequent backups provide a backstop for the maximum possible data loss.

Short of a disaster where an operator uses backups as a backstop to restore operational data, finer grained mechanisms are available to preserve data integrity. For instance, a transaction may be interrupted due to a machine crash. The recovery process involves rolling back any state in the incomplete transaction and retrying it in another machine or in the same machine after it comes back into operation. Unfortunately, depending on the configuration, (memory, storage, etc.) this process can take from minutes to days.

Restarting a machine and re-synchronizing it with the application can take several minutes. This is defined by the mean time to recover or mean time to repair (MTTR) parameter, a finer grained version of RTO. Fujitsu Server PRIMEQUEST 2800E2 / 2400E2 with RHEL continue operating in the presence of unavoidable hardware failures which would cause an outage in lesser machines by using Intel® Xeon® processor E7 family, which has rich RAS (Reliability, Availability, Serviceability) set of features. For instance, DDDC (Double Device Data Correction) allows a machine to continue operating in the presence of unrecoverable memory errors that would crash other machines, and even before memory errors occur. The high end Intel® Xeon® processor architecture supports Intel® Scalable Memory Interconnect (Intel® SMI) lane failover and packet retries as well as packet retries with Intel® QuickPath Interconnect (Intel® QPI) link transactions. When combined with multi-pathing a capability of RHEL, system availability and, more importantly, data traffic are maintained until the system administrator decides to perform a maintenance event or system performance falls below an established threshold.

The Fujitsu PRIMEQUEST architecture allows operators to replace system boards holding CPU and memory as well as I/O units holding PCIe slots while the system is running, thereby avoiding application interruptions in the process and inconveniencing customers. For these cases MTTR is reduced effectively to zero. The hot replacement of physical components with proper operating system support is possible for cases of scheduled or unscheduled maintenance.

Fujitsu PRIMEQUEST Dynamic Reconfiguration
In order to support the business continuity use cases just described, Fujitsu Server PRIMEQUEST 2800E2 / 2400E2 in combination with Red Hat Enterprise Linux 7 have a capability to carry out on-lining (physically adding) and off-lining (physically removing) system boards and I/O units while the system remains operational.

Solution Architecture
Hardware
Fujitsu Server PRIMEQUEST 2800E2/2400E2 is a mission critical Intel processor-based server, scalable from 2 to 8 sockets, providing high-end capabilities using an advanced Fujitsu architecture evolved and refined over many generations of computer system development. Combined with Red Hat Enterprise Linux 7, Fujitsu Server PRIMEQUEST 2800E2/2400E2 embodies the best characteristics of mainframe and UNIX* server reliability, high-performance computing, in-memory computing, real-time analytics with the cost and flexibility benefits of open systems. Fujitsu Server PRIMEQUEST 2800E2/2400E2 is the world first open server that supports hot addition or removal of system boards and I/O Units on Linux OS. Being able to avoid downtime removes MTTR considerations for certain scenarios.

Red Hat Enterprise Linux
Red Hat Enterprise Linux, the world leading enterprise Linux platform, provides unparalleled stability and flexibility. Red Hat introduced a variety of new software-based functions targeting the requirements of enterprise workloads. Red Hat Enterprise Linux delivers the operating system functionality required to exploit the next generation resilience features in the latest Intel® Xeon® processors.
The combination of hardware and operating system functionality enables servers to withstand failures at the physical memory, I/O and fabric.

Furthermore by deploying the Red Hat’s high-availability cluster solution for Red Hat Enterprise Linux (High Availability Add-On) across multiple systems, customers can address business continuity requirements and implement higher levels of uptime as required for business-critical workloads in their organization.

For Further Reading
Reference Architecture - Go Beyond Five Nines with Dynamic Reconfiguration
Solution Implementation Guide - Go Beyond Five Nines with Dynamic Reconfiguration


Business Considerations
Dynamic Reconfiguration is realized by interaction of hardware and software, therefore both hardware and software support is required to utilize this feature. Dynamic Reconfiguration is fully supported in Red Hat Enterprise Linux 7.1 or a later version.

Summary
Dynamic Reconfiguration feature provides highest class of Reliability and Availability to your system for minimizing business impact. This feature has been implemented in Intel® Xeon® processor-based server, the Fujitsu Server PRIMEQUEST 2800E2/2400E2 running Red Hat Enterprise Linux 7, the first solution of its kind in the world.
To learn more about the Intel® Xeon® Processor E7 Family, visit www.intel.com/xeonE7