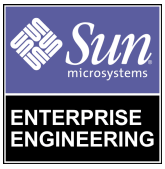
Starfire™ Server DR-Detach and DR-Attach Requirements

By Enrique Vargas - Enterprise Engineering

Sun BluePrints™ OnLine - August 1999



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.: 806-4635-10
Revision 01, August 1999

Copyright 1999 Sun Microsystems, Inc. 901 San Antonio Road, Palo Alto, California 94303 U.S.A. All rights reserved.

This product or document is protected by copyright and distributed under licenses restricting its use, copying, distribution, and decompilation. No part of this product or document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any. Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the U.S. and other countries, exclusively licensed through X / Open Company, Ltd.

Sun, Sun Microsystems, the Sun logo, The Network Is The Computer, Starfire, Sun Enterprise, OpenBoot, Sun BluePrints and Solaris are trademarks, registered trademarks, or service marks of Sun Microsystems, Inc. in the U.S. and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

The OPEN LOOK and Sun™ Graphical User Interface was developed by Sun Microsystems, Inc. for its users and licensees. Sun acknowledges the pioneering efforts of Xerox in researching and developing the concept of visual or graphical user interfaces for the computer industry. Sun holds a non-exclusive license from Xerox to the Xerox Graphical User Interface, which license also covers Sun's licensees who implement OPEN LOOK GUIs and otherwise comply with Sun's written license agreements.

RESTRICTED RIGHTS: Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g)(2)(6 / 87) and FAR 52.227-19(6 / 87), or DFAR 252.227-7015(b)(6 / 95) and DFAR 227.7202-3(a).

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.

Copyright 1999 Sun Microsystems, Inc., 901 San Antonio Road, Palo Alto, Californie 94303 Etats-Unis. Tous droits réservés.

Ce produit ou document est protégé par un copyright et distribué avec des licences qui en restreignent l'utilisation, la copie, la distribution, et la décompilation. Aucune partie de ce produit ou document ne peut être reproduite sous aucune forme, par quelque moyen que ce soit, sans l'autorisation préalable et écrite de Sun et de ses bailleurs de licence, s'il y en a. Le logiciel détenu par des tiers, et qui comprend la technologie relative aux polices de caractères, est protégé par un copyright et licencié par des fournisseurs de Sun.

Des parties de ce produit pourront être dérivées des systèmes Berkeley BSD licenciés par l'Université de Californie. UNIX est une marque déposée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X / Open Company, Ltd.

Sun, Sun Microsystems, le logo Sun, The Network Is The Computer, Starfire, Sun Enterprise, OpenBoot, Sun BluePrints, et Solaris sont des marques de fabrique ou des marques déposées, ou marques de service, de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays. Toutes les marques SPARC sont utilisées sous licence et sont des marques de fabrique ou des marques déposées de SPARC International, Inc. aux Etats-Unis et dans d'autres pays. Les produits portant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

L'interface d'utilisation graphique OPEN LOOK et Sun™ a été développée par Sun Microsystems, Inc. pour ses utilisateurs et licenciés. Sun reconnaît les efforts de pionniers de Xerox pour la recherche et le développement du concept des interfaces d'utilisation visuelle ou graphique pour l'industrie de l'informatique. Sun détient une licence non exclusive de Xerox sur l'interface d'utilisation graphique Xerox, cette licence couvrant également les licenciés de Sun qui mettent en place l'interface d'utilisation graphique OPEN LOOK et qui en outre se conforment aux licences écrites de Sun.

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

Starfire™ Server DR-Detach and DR-Attach Requirements

This is the third article in the *Dynamic Reconfiguration* (DR) series, and it provides a complete list of Starfire™ server prerequisites for enabling attach and detach operations on *any* system board. Keep in mind that we are providing a set of general prerequisites to address *all* I/O interfaces and only a subset of those requirements will apply to your situation. This article also provides details of the sequential actions that take place on the Dynamic System Domain (DSD) and System Service Processor (SSP) to provide a better understanding of the DR framework functionality.

It is extremely important that you upgrade both the DSD and the SSP to the latest patch level before attempting any DR operations (contact your local Sun Service representative for current Starfire server patch information). Use the two figures included in this article as a *requirement checklist* before executing any attach or detach operations on system boards.

If you have already implemented DR in your datacenter and are interested in sharing your experiences, or if you would like us to address a topic that will help you take better advantage of DR, please give us your feedback through the *Questions/Comments* button located on the navbar to the left at the Sun BluePrints™ OnLine web site (http://www.sun.com/blueprints)

Starfire Server DR-Detach Details

The software mechanism by which Dynamic Reconfiguration reduces resources available to the operating system is called detaching (see FIGURE 1).

The DR-detach process requires you to do strategic planning because it reduces the overall CPU and memory resources available to the DSD, and it may involve removing I/O interfaces that are critical to the operation of the machine. Before removing a system board, it is critical that all on-board I/O devices be closed since I/O drivers will fail to detach when an application has open device instances.

Keep in mind that removing CPU and memory resources from an already busy system (a system running short of memory and using excessive amounts of CPU) will significantly delay the DR-detach process since the removed memory will have to be transferred to the swap area (disk or memory).

DR-Detach Prerequisites and SSP/DSD Actions

To guarantee a successful DR-detach process the following rerequisites must be fulfilled:

Pre-Boot Requirements

The following prerequisites require a system reboot to take effect:

- All system memory must not be interleaved between system boards.
- In the Solaris™ 2.5.1 and Solaris 2.6 5/98 Operating Environments, the `dr_max_mem` OBP environment variable has to be set to a value equal to or greater than 512 Mbytes. The `dr_max_mem` OBP environment variable is not required after Solaris 7 5/99 Operating Environment because the required VM structures are dynamically allocated.
- The `soc` driver requires the following entry in the `/etc/system` file to enable the suspend-resume/detach-attach features:

  set soc:`soc_enable_detach_suspend=1`
- In the Solaris 2.6 5/98 Operating Environment, the `socal` driver requires the following entry in the `/etc/system` file to enable the suspend-resume features:

  set socal:`socal_enable_suspend=1`
- The `pln` driver requires the following entry in the `/etc/system` file to enable the suspend-resume/detach-attach features:

  set pln:`pln_enable_detach_suspend=1`

- In the Solaris 2.5.1 Operating Environment, the `ssd` driver requires the following entry in the `/etc/system` file to enable the suspend-resume / detach-attach features:

  set ssd:`ssd_enable_detach_suspend=1`

- Activation of the *kernel cage* requires the following entry in the `/etc/system` file:

```
set kernel_cage_enable=1
```

System Requirements

The following prerequisites must be fulfilled on the running system immediately before the DR-Detach operation takes place:

- All on-board non-DR-compliant drivers must be unloaded using the `modunload(1M)` utility.
- Real-time processes *are not* allowed whenever the *kernel cage* is being relocated because their behavior becomes unpredictable with the operating system quiescence. All real-time processes must be reassigned an alternate scheduling class using the `priocntl(1)` command or killed using the `kill(1)` command. The `hostview(1M)` interface provides the `force` option to override this requirement

---

Note – The *kernel cage* is a special data structure (normally contained by a single system board) which controls the dynamic growth of all non-relocatable memory, including the OpenBoot™ PROM (OBP) and kernel memory. When detaching a system board containing the kernel cage, it is necessary to quiesce the operating system to ensure that no I/O or kernel activity exists while the kernel cage is being relocated. The operating system quiescence involves the suspension of *all* device driver activity and user / kernel threads to avoid corruption of such critical memory.

- Processor sets or processes bound to CPUs hosted by the system board to be detached are *not* allowed. Processor sets must be torn down using the `psrset(1M)` command and processes bound to on-board processors have to be unbounded using the `pbind(1M)`command.
- All swap devices contained by the system board to be detached must be deleted using the `swap(1M)` command and removed from the `/etc/vfstab` file.
- All non-network I/O interfaces contained by the system board to be detached must be brought down.
- All file systems hosted by the system board to be detached must be unmounted.
- If AP is not used, disks hosted by the system board to be detached must be removed from Veritas VxVM control using the `vxdiskadm(1M)` command with the `replace` option.

- Tape drives hosted by the system board to be detached must be empty or have a loaded tape sitting at `BOT`.

DR-Detach Actions

The following steps are automatically executed on both the SSP and DSD sides during the DR-Detach operation:

1. All free pages associated with the system board to be detached are locked to avoid further use by applications. All dirty pages are flushed to swap (swap could be either memory or disk).
2. ISM *must* be relocated to memory available on the remaining system. If there isn't available memory, dirty pages from the remaining system are flushed to disk to make room for the ISM footprint.

Note – Many databases make use of the intimate Shared Memory (ISM) capability in the Solaris Operating Environment where shared memory can be exclusively reserved and locked into RAM to prevent paging. ISM makes use of the shared page table, to make more efficient use of the page translation lookaside buffers (TLBs).

---

3. If the kernel cage is located on the system board to be detached, the `dr` driver invokes the `DDI_SUSPEND` driver function on all I/O devices in the system and the operating system is quiesced (suspended) to allow the safe relocation of the entire address range associated with the system board to be detached. The relocation process involves releasing the memory content from the *target system board* as defined in steps 1-2 above. Once the memory content is evacuated out of the target system board, the entire memory range for the board to be detached is transferred to the target system board through the copy-rename routine (running out of a single CPU cache). When the copy-rename routine completes, the memory controller on the target system board is updated to reflect the physical address range of the copied memory and the whole system is then *resumed*
4. If the board to be detached *does not* contain the kernel cage, its associated address range is removed from the system. If the kernel cage is located on the system board to be detached, the address range associated with the *target system board* is removed from the system.
5. The memory controller on the detached board is disabled and the removed memory is taken out of the kernel memory pool.
6. All processors resident on the system board to be detached are offlined and removed from the kernel configuration. The Boot processor is reassigned if needed.

7. The `dr_daemon` automatically downs and unplumbs all network devices resident on the system board.
8. The `dr` driver invokes the `DDI_DETACH` function on all onboard I/O devices. The kernel updates the operating system device tree (`devinfo`) to reflect the removed I/O devices.
9. The SSP reprograms the detached system board and centerplane hardware registers to remove any DSD association. The `domain_config(4)`is executed to remove the detached board from the DSD configuration.

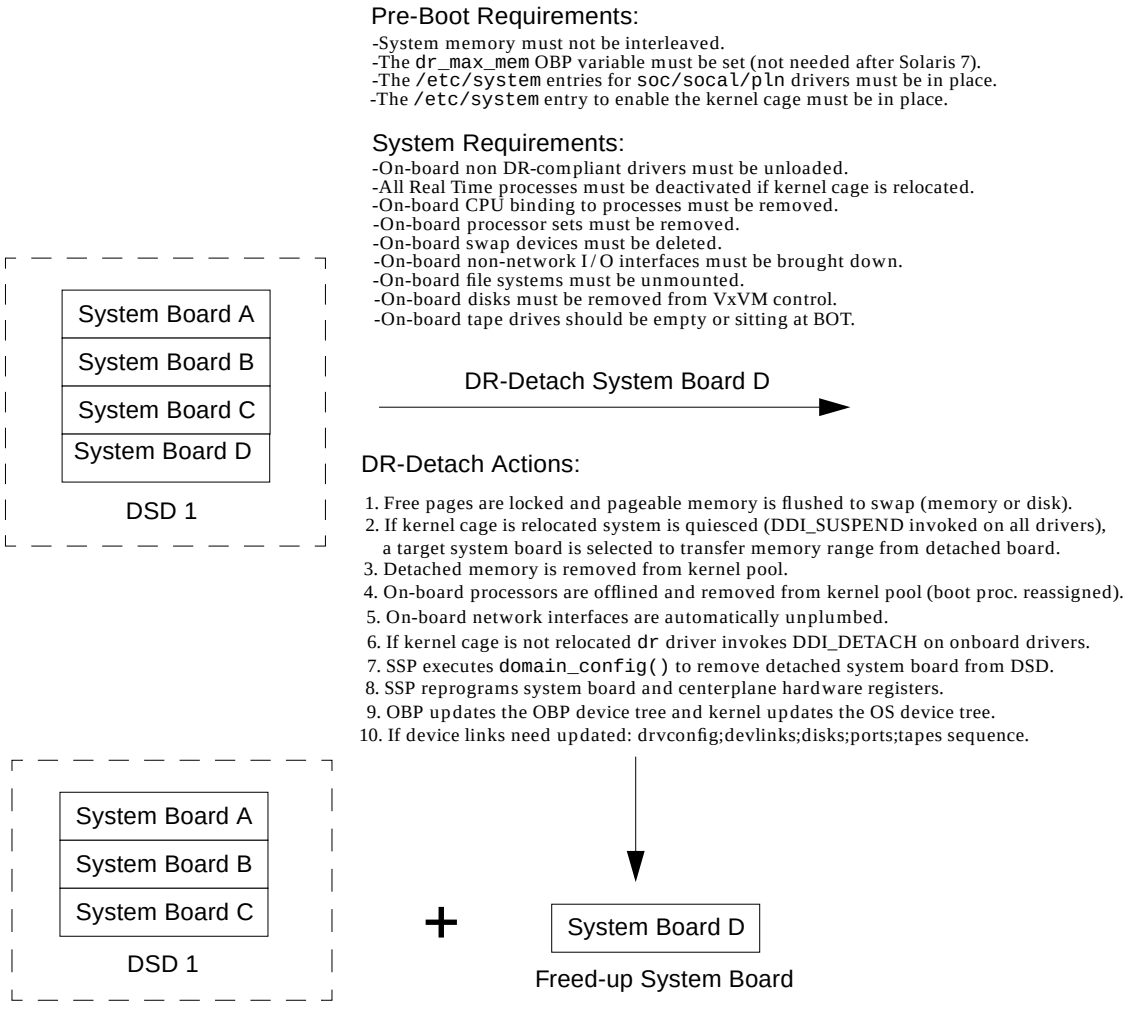


FIGURE 1 DR-Detaching a Single System Board from a DSD

Starfire Server DR-Attach Details

The software mechanism by which Dynamic Reconfiguration increases resources available to the operating system is called attaching (see FIGURE 2).

The DR-Attach mechanism provides a way for the operating system to logically integrate new CPU, memory and I/O resources into its resource pool and make it immediately available to applications requiring the increased capacity. The DR-Attach process creates the appropriate device file links and loads the appropriate drivers to support the newly introduced I/O interfaces contained by the system board.

If there's a non-DR-compliant driver associated with the system board to be attached, the DR operation will not fail, but it will not load the appropriate drivers either. A workaround for detaching system boards with non-DR-compliant drivers is to load the driver manually using the `modload(1M)` command.

Note – The `modinfo(1M)` command must be used to verify an existing driver instance for the non-DR-compliant driver (`modinfo|grep sg`). The `modunload(1M)`command must be executed before the `modload(1M)`command if there is an existing instance of a non-DR-compliant driver associated with the system board to be attached.

DR-Attach Prerequisites and Actions

To guarantee a successful DR-attach process the following prerequisites must be fulfilled:

Pre-Boot Requirements

The following prerequisites require a system reboot to take effect:

- All system memory should not be interleaved between system boards.
- In the Solaris 2.5.1 and Solaris 2.6 5/98 Operating Environments, the `dr_max_mem` OBP environment variable must be set to a value equal to or greater than 512 Mbytes. The `dr_max_mem` OBP environment variable is not required after Solaris 7 5/99 Operating Environment because the required VM structures are dynamically allocated.
- The `soc` driver requires the following entry in the `/etc/system` file to enable the suspend-resume/detach-attach features:

  set soc:`soc_enable_detach_suspend=1`

- In the Solaris 2.6 5/98 Operating Environment, the `socal` driver requires the following entry in the `/etc/system` file to enable the suspend-resume features:

  set socal:`socal_enable_suspend=1`

- The `pln` driver requires the following entry in the `/etc/system` file to enable the suspend-resume/detach-attach features:

  set pln:`pln_enable_detach_suspend=1`

- In the Solaris 2.5.1 Operating Environment, the `ssd` driver requires the following entry in the `/etc/system` file to enable the suspend-resume/detach-attach features:

  set ssd:`ssd_enable_detach_suspend=1`

- Activation of the kernel cage requires the following entry in the `/etc/system` file:

```
set kernel_cage_enable=1
```

System Requirements

The only prerequisites to be fulfilled before the DR-attach process are:

- The system board to be attached must be powered up and not be a part of any domain.
- The system board to be attached must host at least one CPU to enable POST execution.

DR-Attach Actions

The following steps are automatically executed on both the SSP and DSD sides during the DR-Attach operation:

- The `domain_config(4)`is executed to add the new system board to the DSD.
- `hpost` (POST through the JTAG interface) is executed against the newly attached system board to guarantee full functionality before configuring components.
- CPU code is downloaded to the `BBSRAM` (Boot Bus Static Random Access Memory) `NVRAM` (Non-Volatile Random Access Memory) `mailbox` to handle the newly attached processors. CPUs are taken out of reset and start executed downloaded code.
- The SSP reprograms the attached system board and centerplane hardware registers to create the DSD association.
- OBP probes newly attached board devices and updates the OBP (`dnode`) device tree. The Solaris Operating Environment kernel updates the corresponding operating system (`devinfo`) device tree.

- The Solaris Operating Environment kernel starts the newly attached processors and adds the newly attached memory to the memory pool.
- The `dr` driver invokes the DDI_ATTACH function on all device drivers associated with the newly attached system board.

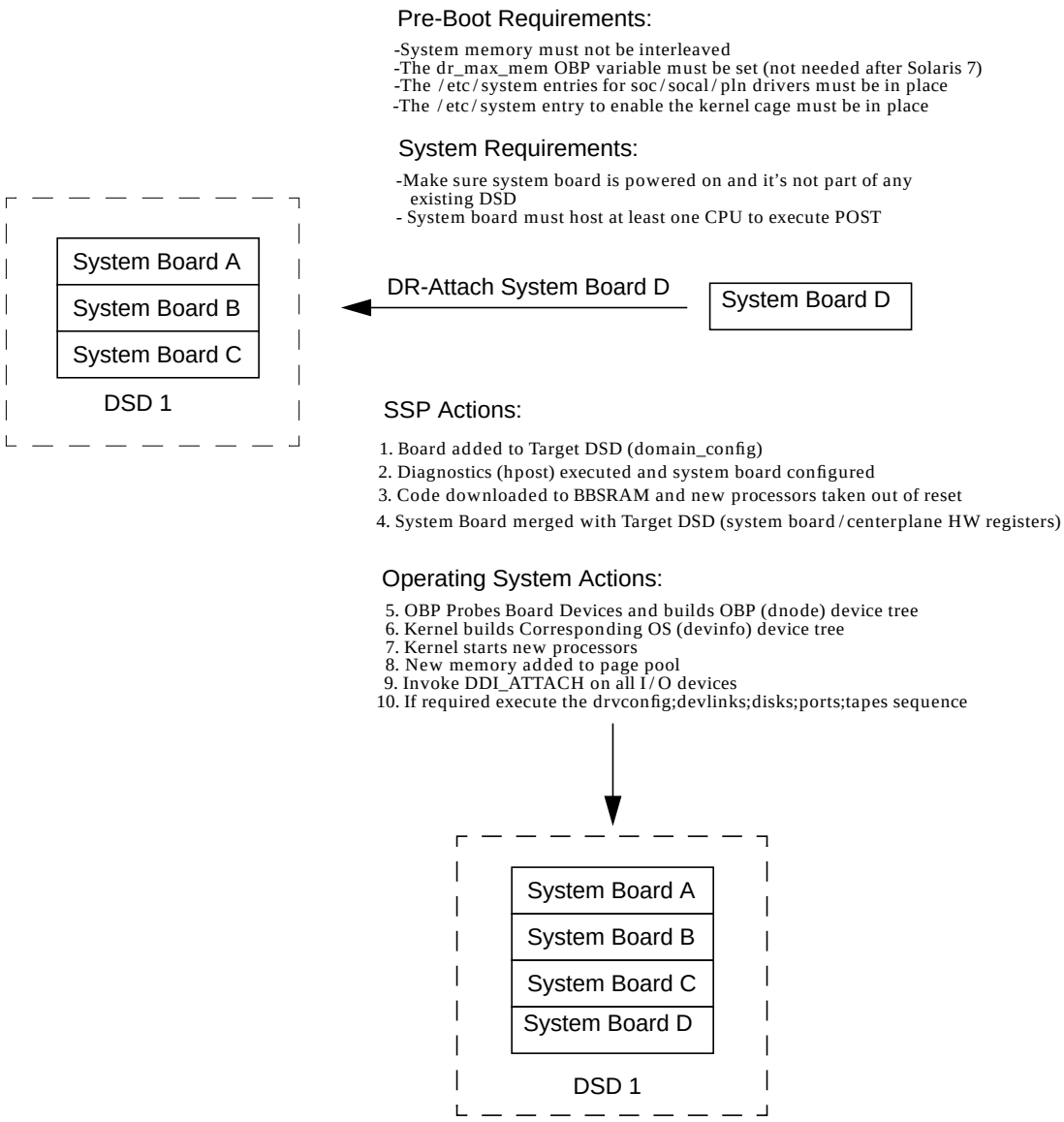


FIGURE 2 DR-Attaching a System Board to a DSD

Conclusion

The DR-detach process requires strategic planning because it reduces the overall CPU and memory resources available to the DSD, and it involves removing I/O interfaces which may be critical to the operation of the machine. Both the DR-attach and DR-detach operations require that established prerequisites be fulfilled to avoid operation failures.

Author's Bio: Enrique Vargas

Enrique brings a wealth of large systems experience to Sun and specializes in high end UNIX offerings including the Sun Enterprise™ 10000 Server. Enrique came to Sun from Amdahl where he also focused on the high end Solaris systems