Designing a Fault-Tolerant Network Using Sun Netra™ CP3x40 Switches

Sun Microsystems, Inc.
www.sun.com

Part No. 820-7346-10
April 2009, Revision 01

Submit comments about this document at: `http://www.sun.com/hwdocs/feedback`

Copyright © 2009 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, U.S.A. All rights reserved.

Sun Microsystems, Inc. has intellectual property rights relating to technology embodied in the product that is described in this document. In particular, and without limitation, these intellectual property rights may include one or more of the U.S. patents listed at http://www.sun.com/patents and one or more additional patents or pending patent applications in the U.S. and in other countries.

This distribution may include materials developed by third parties.

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 in other countries, exclusively licensed through X/Open Company, Ltd.

Sun, Sun Microsystems, the Sun logo, Solaris, Netra and the Netra logo are trademarks or registered trademarks of Sun Microsystems, Inc., or its subsidiaries, in the U.S. and other countries.

Products covered by and information contained in this service manual are controlled by U.S. Export Control laws and may be subject to the export or import laws in other countries. Nuclear, missile, chemical biological weapons or nuclear maritime end uses or end users, whether direct or indirect, are strictly prohibited. Export or reexport to countries subject to U.S. embargo or to entities identified on U.S. export exclusion lists, including, but not limited to, the denied persons and specially designated nationals lists is strictly prohibited.

Use of any spare or replacement CPUs is limited to repair or one-for-one replacement of CPUs in products exported in compliance with U.S. export laws. Use of CPUs as product upgrades unless authorized by the U.S. Government is strictly prohibited.

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 © 2009 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, Etats-Unis. Tous droits réservés.

Sun Microsystems, Inc. détient les droits de propriété intellectuels relatifs à la technologie incorporée dans le produit qui est décrit dans ce document. En particulier, et ce sans limitation, ces droits de propriété intellectuelle peuvent inclure un ou plus des brevets américains listés à l'adresse http://www.sun.com/patents et un ou les brevets supplémentaires ou les applications de brevet en attente aux Etats - Unis et dans les autres pays.

Cette distribution peut comprendre des composants développés par des tierces parties.

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, Solaris, Netra et le logo Netra sont des marques de fabrique ou des marques déposées de Sun Microsystems, Inc., ou ses filiales, aux Etats-Unis et dans d'autres pays.

Ce produit est soumis à la législation américaine en matière de contrôle des exportations et peut être soumis à la règlementation en vigueur dans d'autres pays dans le domaine des exportations et importations. Les utilisations , ou utilisateurs finaux, pour des armes nucléaires, des missiles, des armes biologiques et chimiques ou du nucléaire maritime, directement ou indirectement, sont strictement interdites. Les exportations ou reexportations vers les pays sous embargo américain, ou vers des entités figurant sur les listes d'exclusion d'exportation américaines, y compris, mais de manière non exhaustive, la liste de personnes qui font objet d'un ordre de ne pas participer, d'une façon directe ou indirecte, aux exportations des produits ou des services qui sont régis par la législation américaine en matière de contrôle des exportations et la liste de ressortissants spécifiquement désignés, sont rigoureusement interdites. L'utilisation de pièces détachées ou d'unités centrales de remplacement est limitée aux réparations ou à l'échange standard d'unités centrales pour les produits exportés, conformément à la législation américaine en matière d'exportation. Sauf autorisation par les autorités des Etats-Unis, l'utilisation d'unités centrales pour procéder à des mises à jour de produits est rigoureusement interdite.

LA DOCUMENTATION EST FOURNIE "EN L'ETAT" ET TOUTES AUTRES CONDITIONS, DECLARATIONS ET GARANTIES EXPRESSES OU TACITES SONT FORMELLEMENT EXCLUES, DANS LA MESURE AUTORISEE PAR LA LOI APPLICABLE, Y COMPRIS NOTAMMENT TOUTE GARANTIE IMPLICITE RELATIVE A LA QUALITE MARCHANDE, A L'APTITUDE A UNE UTILISATION PARTICULIERE OU A L'ABSENCE DE CONTREFACON.

Please
Recycle

Adobe PostScript

READ AND DELETE

- Print this file and have it next to you before deleting this text frame.
- Press Control and click once on the red text frame to select it, and delete it. Follow the same procedure to delete the red text frame on the next page.

If you generated a copyright notice yourself in the Legal Notices Tool, follow these instructions:

1. After generating a copyright notice for your document in the Legal Notices Tool, keep it and the browser window open on your desktop.

2. In the browser window, locate the Long English copyright notice.

3. Select and copy all paragraphs in the Long English copyright notice.

4. In the FrameMaker `copyright.`*xxx* template, place the cursor in the empty paragraph at the top of the template. (Select View -> Text Symbols to see the empty paragraph symbol).

5. Paste the copied text into the copyright template.

6. Press Return at end of the last English paragraph (in ALL CAPS) to insert an empty paragraph.

7. In the browser window, locate the Long French copyright notice.

8. Select and copy all paragraphs in the Long French copyright notice.

9. In the copyright template, place the cursor in the empty paragraph under the last English paragraph, and paste the French text into the template.

10. Remove all empty paragraphs from the copyright notices.

    After copying and pasting all copyright notice paragraphs into the copyright template, every other paragraph is empty. You must remove these empty paragraphs.

11. Retag three paragraphs as follows:

    - First paragraph in the English notice containing the "Copyright @ 2008" string:
      Tag this with the "copyright" paragraph tag.

    - Last paragraph in the English notice (in all caps):
      Tag this paragraph with the "Paragraph_Divider" paragraph tag.

    - Last paragraph in the French notice (in all caps).
      Tag this paragraph with the "Paragraph_CreditsCaps" paragraph tag:

    The default paragraph tag for all other text is “Paragraph_Credits.” Tag the other text if necessary.

12. In the first paragraph of the English copyright notice, use the Ctrl-Q) (Shift-0) key sequence to insert a © symbol after the word “Copyright.” Repeat in the first paragraph of the French copyright notice.

13. Save the copyright template as *product-name*`-copyright.`*xxx*.

14. Using standard procedures, place the *product-name*`-copyright.`*xxx* file as the second file in the book

Please Recycle

Adobe PostScript

Contents

ATCA Network Overview 2
Design Principles and Guidelines 3
- Basic 3
- Smart End Points: Channel Bonding 3
- Layer 2 Methods: VLANs and MSTP 4
- Layer 2 VLAN Configuration 6
- Layer 3 Methods: VRRP and VRRP Tracking 8
- Layer 3 VRRP Configuration 10
- Best Practices 12
- All Methods Integrated Configuration 14

Related Documentation 17

Designing a Fault-Tolerant Network

This document provides guidelines for designing a fault-tolerant network using Netra CP3140 switches and Netra CP3240 switches.

This document contains the following topics:

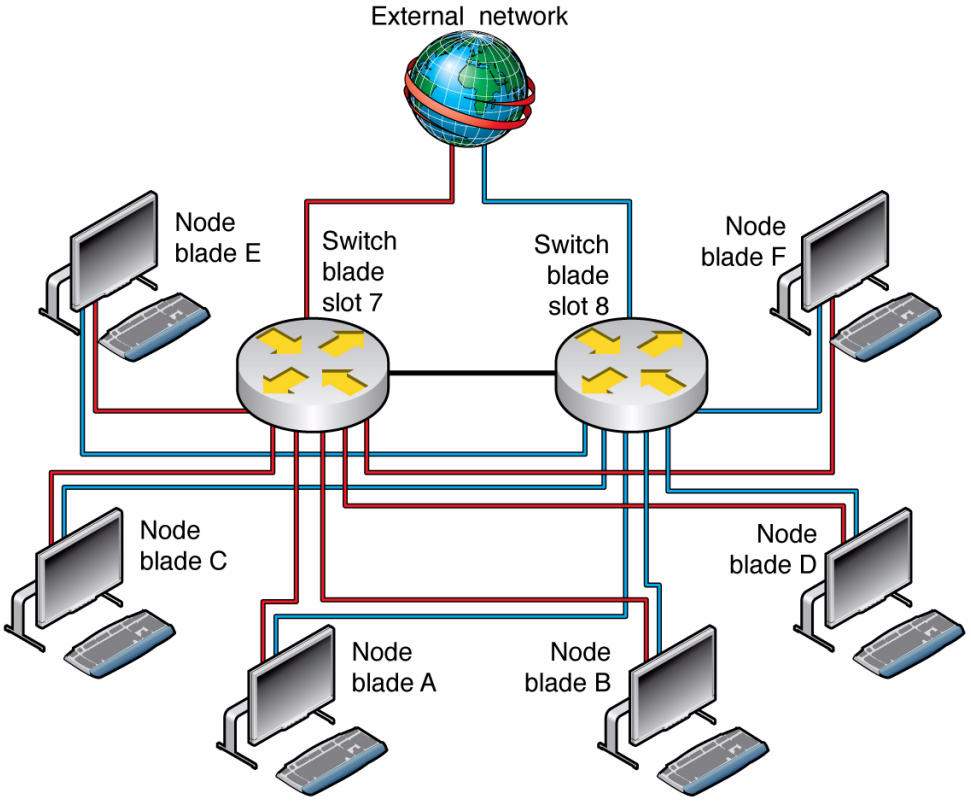
- "ATCA Network Overview" on page 2
- "Design Principles and Guidelines" on page 3
- "Related Documentation" on page 17

ATCA Network Overview

ATCA is designed to be used in highly available (HA) environments. Many HA principles such as redundancy and fault tolerance are designed into ATCA specification. ATCA systems need to be connected to external networks in such a manner that the HA principles applied inside the shelf are also applied to external networks. A system is only as available as its connections to end users. This document outlines some of the techniques to maximize uptime of an ATCA system that can be applied when using Sun's Netra switch.

The following shows a simplified partial-network diagram of an ATCA network.

FIGURE 1 ATCA Network Diagram (Partial)



Design Principles and Guidelines

This section covers fault-tolerant design principles and guidelines.

Basic

Independent of the software used to increase availability, a system should be redundantly cabled, preferably at both the board level and the link level.

ATCA uses a dual-star topology for backplane connections. So, inside the ATCA shelf every node (anything that is a network endpoint) is connected to both switch blades. (For shelf managers, this statement is true only for PICMG 3.0 R2.0 ECN-002 “cross-connect” enabled shelves).

- The redundancy should be replicated outside the ATCA shelf.
- Every external element (any network-enabled equipment such as a node, a switch, or a router) should be connected to both switch blades in the ATCA shelf.
- With these redundant links, the ATCA concept of a dual-star network is extended outside of the shelf.
- While cabling to both switch blades provides fault tolerance for external link failures, running multiple cables to both hubs further increases fault tolerance.
- External cables are one of the more vulnerable parts of a HA system. A full fail-over can be avoided, or at least delayed in some cases, by having multiple links to each switch blade.

Smart End Points: Channel Bonding

The simplest and most powerful way to implement an HA network is by using channel bonding drivers. Because every ATCA node is connected to two hub blades, the node has at least two network interfaces. While these interfaces can be treated as separate interfaces logically, it makes sense to treat them as a single interface. Channel bonding drivers provide that abstraction. Higher level software uses a single virtual-network interface, and the channel bonding driver handles the complex choice of which physical-network interface to use.

Channel bonding drivers choose which port to use at what time, through various decision algorithms. Most of these algorithms put bandwidth at a higher priority than availability. However, the active-standby algorithm is specifically designed for HA, and therefore it is the best algorithm to use in ATCA networks.

The true power of a channel bonding driver is not in its decision algorithm for choosing ports, but rather its decision algorithm that defines whether a port is usable. Only two common algorithms define if a port is usable:

- The first very simply checks the link state of the network interface. This algorithm has limited use because it only monitors the immediate physical link of the port in question.
- The second monitors the availability of an IP. This one can be a very powerful method of monitoring a port, because an entire path (all links and elements between the port and the element being monitored) can be monitored. If any links between the elements fail, the channel bonding driver will failover to the other network interface. In this way, an ATCA node can monitor and act on failures both inside and outside of the ATCA shelf.

One of the great features of channel bonding drivers is that they are generally topology, layer, and protocol independent (decision algorithms are not topology independent, but the overall channel bonding driver is). They work well in complex networks, simple networks, Layer 2 switched networks, and Layer 3 routed networks.

Layer 2 Methods: VLANs and MSTP

Layer 2 switching is used often because it is simple. The simplicity of switching makes it fast and inexpensive, but it puts limits on the design of the network.

A Layer 2 network, without additional protocols, must follow a strict tree structure. No loops are allowed in the network. This limitation is in direct contrast to the key HA principle of redundancy. However, using the VLAN and MSTP protocols on top of a Layer 2 switched network resolves this limitation and allows loops.

- Use VLANs to segment the switched network into smaller networks.
- Configure switches to allow only certain VLAN tagged traffic onto certain ports. The traffic get tags from both switches and nodes as it passes though/to them (switches and nodes can remove tags as needed).

With these rules and tags, traffic can be limited and controlled so that the network loop does not exist on any single VLAN.

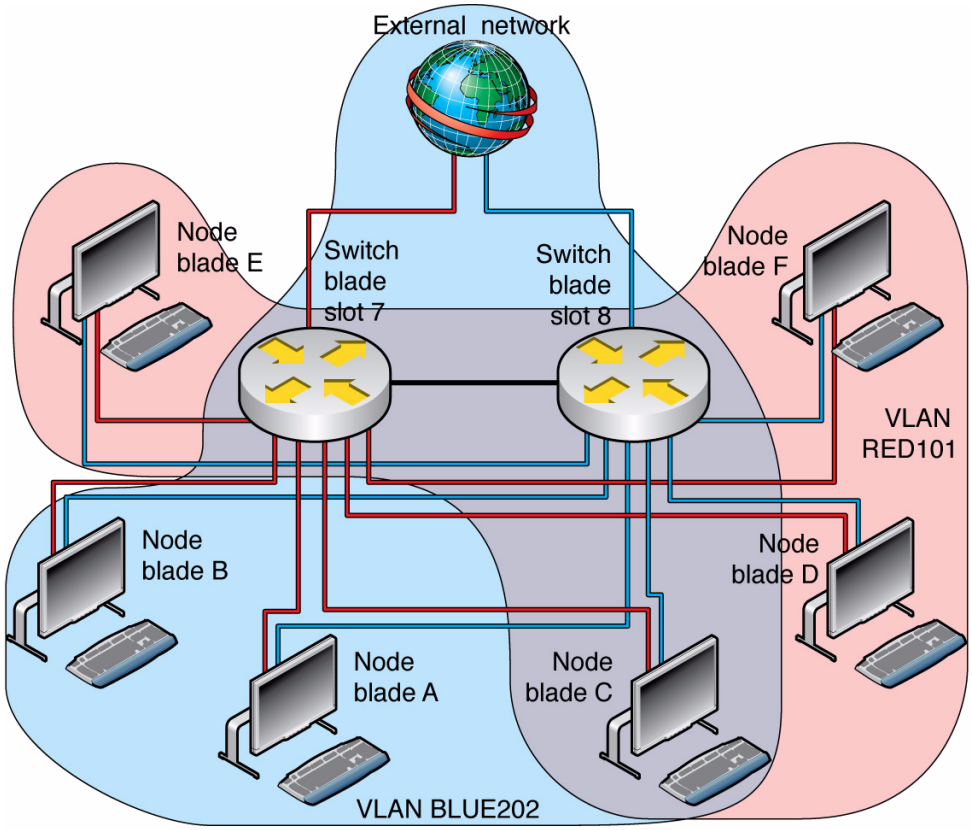
STP (spanning tree protocol) is a protocol that is designed specifically to deal with network loops. STP traverses a network, finds, all loops, and disables the links that created loops. It effectively makes the network graph into a tree graph (no node connect to any other node by more than one link), hence the name.

MSTP (multiple spanning tree protocol) is a protocol designed to improve STP. MSTP is better than STP in two ways. First, MSTP is VLAN aware. Regular STP ignores VLAN settings and thus, even if a network loop has been properly segmented with VLAN, STP will disable the loop link. MSTP can recognize that

while the loop exists, it has been contained by VLAN settings and does not disable the loop link. Secondly, MSTP can converge (reconfigure) more quickly than STP when the network changes. Of the two redundant links in any network loop, one will be active and the other will be inactive. When the active link fails, it is desired to switch to the inactive link as quickly as possible.

The following shows an example Layer 2 network configured with VLAN and MSTP.

FIGURE 2 Layer 2 Network Configured With VLAN and MSTP



In this network, we have three node blades that do not need to communicate with the external network (RED) and three that do (BLUE). One of the BLUE nodes needs to communicate with the RED nodes. VLANs are used to logically separate the RED and BLUE nodes. MSTP is used to prevent a blocking of the inter-switch link due to a loop in VLAN BLUE from affecting VLAN RED.

Layer 2 VLAN Configuration

The following code example shows how to configure the example network shown in FIGURE 2.

CODE EXAMPLE 1 Configuring a Layer 2 VLAN Network With MSTP

```
vlan database
vlan  101
vlan  202
exit

configure

!interswitch link needs to be in both VLANs
interface  0/2
!The port cost of MST 2 is set here are lower then normal
!to signify using this port is preferential over others
spanning-tree mst 2 cost 1800
vlan participation exclude 1
vlan participation include 101
vlan tagging 101
vlan participation include 202
vlan tagging 202
exit

!node blade A is in 202
interface  0/3
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
exit

!node blade B is in 202
interface  0/4
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
exit

!node blade C is in 202 and 101
```

CODE EXAMPLE 1 Configuring a Layer 2 VLAN Network With MSTP *(Continued)*

```
interface  0/5
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
vlan participation include 101
vlan tagging 101
exit

!node blade D is in 101
interface  0/6
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade E is in 101
interface  0/7
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade F is in 101
interface  0/8
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!external network
interface 0/20
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
exit

!MSTP is enabled such that 101 and 202 are in different
!MSTP instances
spanning-tree
spanning-tree configuration name "MSTPexample"
spanning-tree configuration revision 0
spanning-tree mst instance 1
spanning-tree mst vlan 1 1
spanning-tree mst vlan 1 101
spanning-tree mst instance 2
```

CODE EXAMPLE 1 Configuring a Layer 2 VLAN Network With MSTP *(Continued)*

```
spanning-tree mst vlan 2 202

exit
```

Layer 3 Methods: VRRP and VRRP Tracking

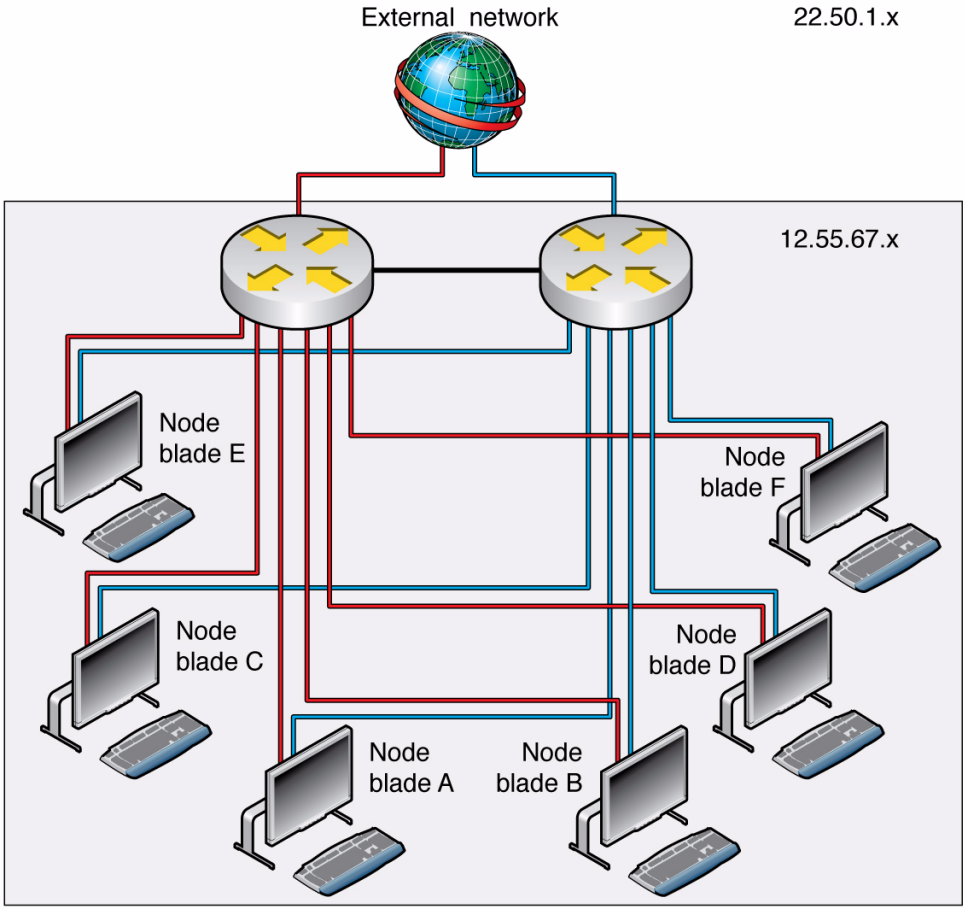
VRRP (virtual router redundancy protocol) lets multiple routers appear as a single virtual router. With VRRP nodes on the Layer 2 side of the hub, blades can be configured with the single virtual IP and network elements on the Layer 3 side, to see two routes to the same subnet. One of the VRRP routers becomes the master of the virtual IP and does the routing. If it fails, the backup router takes over.

Traditional VRRP defines a router failure as a series of missing checkpoint packets sent between VRRP routers. VRRP tracking adds other fail-over conditions. With VRRP tracking, multiple “tracks” can be set up to monitor the status of a link, a local route, or a remote IP. If any track fails, the router forces a fail-over immediately, without waiting for the checkpoint packets.

Layer 3 routing is more complex than Layer 2 switching but is more robust. Loops are expected in Layer 3 networks. In ATCA, hub blades are used as gateway routers. A hub blade connects a Layer 3 external network to a Layer 2 internal network. There are two gateways for the two hub blades in the ATCA shelf. The node blades could be configured to handle two separate gateways; however, the VRRP protocol provides a more elegant solution.

The following shows an example VRRP network configured with VRRP tracking.

FIGURE 3 Layer 3 Network Configured with VRRP



In this example, the two switch blades are redundant gateways between the node blades and the external network. The node blades are in the 12.55.67.x subnet, whereas the external network is in the 22.50.1.x subnet. An instance of VRRP is configured so that if either switch blade fails, the node can still reach the external network. VRRP tracks are added to provide more robust failover conditions.

Layer 3 VRRP Configuration

The following code example shows how to configure the example network shown in FIGURE 3.

CODE EXAMPLE 2 Configuring a Layer 3 Network With VRRP

```
vlan database
vlan  101
vlan routing 101
exit

configure
ip routing
ip vrrp

!interswitch link needs to be in both VLANs
interface  0/2
vlan participation exclude 1
vlan participation include 101
vlan tagging 101
exit

!node blade A is in 101
interface  0/3
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade B is in 101
interface  0/4
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade C is in 101
interface  0/5
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade D is in 101
interface  0/6
vlan participation exclude 1
vlan participation include 101
```

CODE EXAMPLE 2 Configuring a Layer 3 Network With VRRP *(Continued)*

```
vlan pvid 101
exit

!node blade E is in 101
interface  0/7
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade F is in 101
interface  0/8
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!external network
interface 0/20
ip address  22.50.1.5  255.255.255.0
exit

interface  4/1
!the other switch is set to 12.55.67.19
ip address  12.55.67.18  255.255.255.0
ip vrrp 102
ip vrrp 102 ip 12.55.67.20
!the other switch is set to 253
ip vrrp 102 priority 250
ip vrrp 102 mode
exit
!
!track remote server
track 1  ip route 22.50.1.15/24 reachability

!track local link state If any link goes down, failover
track 2 interface 0/3 line-protocol
track 3 interface 0/4 line-protocol
track 4 interface 0/5 line-protocol
track 5 interface 0/6 line-protocol
track 6 interface 0/7 line-protocol
track 7 interface 0/8 line-protocol
track 8 interface 0/20 line-protocol

!track local route
track 9  interface 4/1 ip routing
```

CODE EXAMPLE 2 Configuring a Layer 3 Network With VRRP *(Continued)*

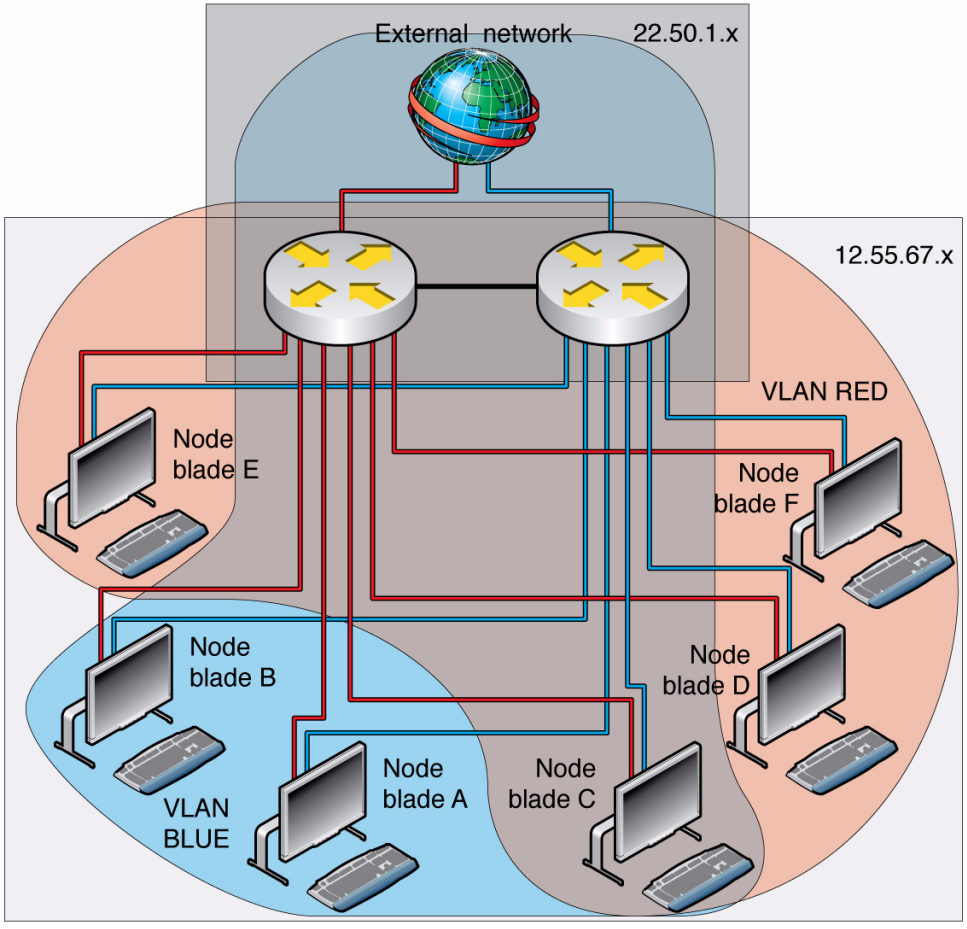
```
!Assign values to each track and assign the tracks to the vrrp
instance
vrrp 102 track 1 decrement 40
vrrp 102 track 2 decrement 40
vrrp 102 track 3 decrement 40
vrrp 102 track 5 decrement 40
vrrp 102 track 6 decrement 40
vrrp 102 track 7 decrement 40
vrrp 102 track 8 decrement 40
vrrp 102 track 9 decrement 40

exit
```

Best Practices

Each of the fault-tolerant network design methods presented (channel bonding drivers, Layer 2 methods, and Layer 3 methods) are best used together to achieve maximum availability.

The following shows an example of all methods combined into a single network configuration.

FIGURE 4 Fault-Tolerant Network Combining All Design Methods



All Methods Integrated Configuration

In the following example, channel bonding, VLANs with MSTP, VRRP, and VRRP tracking are integrated.

CODE EXAMPLE 3 Configuring a Network With All Design Methods Integrated

```
vlan database
vlan  101
vlan routing 101
vlan 202
vlan routing 202
exit

configure
ip routing
ip vrrp

!interswitch link needs to be in both VLANs
interface  0/2
!The port cost of MST 2 is set here are lower then normal
!to signify using this port is preferential over others
spanning-tree mst 2 cost 1800
vlan participation exclude 1
vlan participation include 101
vlan tagging 101
exit

!node blade A is in 202
interface  0/3
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
exit

!node blade B is in 202
interface  0/4
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
exit

!node blade C is in 202 and 101
interface  0/5
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
vlan participation include 101
```

CODE EXAMPLE 3 Configuring a Network With All Design Methods Integrated

```
vlan tagging 101
exit

!node blade D is in 101
interface  0/6
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade E is in 101
interface  0/7
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!node blade F is in 101
interface  0/8
vlan participation exclude 1
vlan participation include 101
vlan pvid 101
exit

!external network
interface 0/20
vlan participation exclude 1
vlan participation include 202
vlan pvid 202
exit

interface  4/1
!the other switch is set to 12.55.67.19
ip address  12.55.67.18  255.255.255.0
ip vrrp 102
ip vrrp 102 ip 12.55.67.20
!the other switch is set to 253
ip vrrp 102 priority 250
ip vrrp 102 mode
exit

interface  4/2
!the other switch is set to 22.50.1.4
ip address  22.50.1.3  255.255.255.0
ip vrrp 202
ip vrrp 202 ip 22.50.1.5
!the other switch is set to 253
```

CODE EXAMPLE 3 Configuring a Network With All Design Methods Integrated

```
ip vrrp 202 priority 250
ip vrrp 202 mode
exit

!
!track remote server
track 1  ip route 22.50.1.15/24 reachability

!track local link state If any link goes down, failover
track 2 interface 0/3 line-protocol
track 3 interface 0/4 line-protocol
track 4 interface 0/5 line-protocol
track 5 interface 0/6 line-protocol
track 6 interface 0/7 line-protocol
track 7 interface 0/8 line-protocol
track 8 interface 0/20 line-protocol

!track local route
track 9  interface 4/1 ip routing
track 10  interface 4/2 ip routing

!Assign values to each track and assign the tracks to the vrrp
instance
vrrp 102 track 1 decrement 40
vrrp 102 track 2 decrement 40
vrrp 102 track 3 decrement 40
vrrp 102 track 5 decrement 40
vrrp 102 track 6 decrement 40
vrrp 102 track 7 decrement 40
vrrp 102 track 8 decrement 40
vrrp 102 track 9 decrement 40
vrrp 102 track 10 decrement 40
vrrp 202 track 1 decrement 40
vrrp 202 track 2 decrement 40
vrrp 202 track 3 decrement 40
vrrp 202 track 5 decrement 40
vrrp 202 track 6 decrement 40
vrrp 202 track 7 decrement 40
vrrp 202 track 8 decrement 40
vrrp 202 track 9 decrement 40
vrrp 202 track 10 decrement 40

!MSTP is enabled such that 101 and 202 are in different
!MSTP instances
spanning-tree
spanning-tree configuration name "MSTPexample"
spanning-tree configuration revision 0
```

CODE EXAMPLE 3 Configuring a Network With All Design Methods Integrated

```
spanning-tree mst instance 1
spanning-tree mst vlan 1 1
spanning-tree mst vlan 1 101
spanning-tree mst instance 2
spanning-tree mst vlan 2 202

exit
```

Related Documentation

The following table lists related documentation. The online documentation is available at:

`http://docs.sun.com/app/docs/prod/cp3240.switch?l=en#hic`

`http://docs.sun.com/app/docs/prod/n900.srvr#hic`

Application	Title	Part Number	Format	Location
Latest information	*Sun Netra CP32x0 Product Notes*	820-3260-*xx*	PDF	Online
Pointer doc	*Sun Netra CP3240 Switch Getting Started Guide*	820-3254-*xx*	Printed	Shipping Kit
Usage	*Sun Netra CP3240 Switch User's Guide*	820-3252-*xx*	PDF	Online
Reference	*Sun Netra CP3240 Switch Software Reference Manual*	820-3253-*xx*	PDF	Online
Safety	*Sun Netra CP3x20 Switch Safety and Compliance Manual*	820-3505	PDF	Online
Latest information	*Netra CT 900 Server Product Notes*	819-1180-*xx*	PDF	Online
Pointer Doc	*Netra CT 900 Server Getting Started Guide*	819-1173-*xx*	Printed	Shipping kit
Overview	*Netra CT 900 Server Overview*	819-1174-*xx*	PDF	Online
Installation	*Netra CT 900 Server Installation Guide*	819-1175-*xx*	PDF	Online
Administration	*Netra CT 900 Server Administration and Reference Manual*	819-1177-*xx*	PDF	Online
Reference	*Netra CP3140 Switch Software Reference Manual*	819-3774-*xx*	PDF	Online

Note – Documentation for the Netra CP3140 switch is within the Netra CT900 server documentation, as listed in the table.