

# Tru64 UNIX Best Practice

## Tuning the Cluster Transition Time

**June 2003**

**Product Version:**                      **TruCluster Server Software Version  
5.1A or higher**

This Best Practice describes how to tune the cluster transition time for the TruCluster Server software.



---

# Contents

## Tuning the Cluster Transition Time

Is This Best Practice Right for You? .....	1
Before You Begin .....	2
Factors Affecting How Low the cluster_rebuild_delay Attribute Can be Set .....	2
System Size and Load .....	3
Cluster Interconnect Configuration and Usage .....	3
Number and Model Mix of Cluster Members .....	3
Storage Configuration .....	4
System and Component Reliability .....	4
Effects of Lowering the cluster_rebuild_delay Value Too Much .....	4
Test Description and Methodology .....	5
Test Results: Tru64 UNIX Version 5.1B and Patch Kit ....	5
Modifying System Attributes .....	6
Applying the Best Practice .....	6
Verifying Success .....	7
Troubleshooting .....	7
Comments and Questions .....	7
Legal Notice .....	8



---

## Tuning the Cluster Transition Time

This Best Practice describes how to tune a TruCluster Server system to maximize client availability during cluster member node transitions.

Cluster member node transition occurs when members either expectedly or unexpectedly leave or join the cluster. A member may expectedly leave the cluster as a result of a user issuing a shutdown, reboot, or halt command. A member may unexpectedly leave the cluster as the result of a panic, machine check, equipment failure, or loss of power.

During node transition a sequence of events takes place that is required for proper cluster operation. These events may cause client applications, such as NFS and Web browsers, to be either slow to respond or to be unavailable for a period of time.

The length of time required for this sequence is defined by the `cluster_rebuild_delay` attribute. This attribute is a system attribute in the `clubase` subsystem and has a default setting of 240 seconds. This default setting has been optimized for eight-node Memory Channel clusters with complex storage configurations operating under heavy loads. If your cluster is not configured in this fashion, you may be able to lower the `cluster_rebuild_delay` value and reduce the length of time that client services are not available during node transition.

See the Tru64 UNIX Best Practices Web page for more information about Best Practices documentation:

[http://www.tru64unix.compaq.com/docs/best\\_practices/](http://www.tru64unix.compaq.com/docs/best_practices/)

### Is This Best Practice Right for You?

Not all Best Practices apply to all configurations, so you must be sure that this Best Practice is appropriate for your system and circumstances. To use this Best Practice, you must meet the requirements described in the following table:

<b>Requirement</b>	<b>Description</b>
Operating System	Tru64 UNIX Version 5.1A or higher.
Product Version	TruCluster Server Version 5.1A or higher.
Impact on Availability	Requires a clusterwide halt and reboot. Completion of this procedure increases system availability during node transition.
Required Skills	This Best Practice requires manual editing of system configuration files, so you must be an experienced cluster administrator to use this procedure.

If you do not meet these requirements, do not attempt to perform the tasks describe in this Best Practice.

## Before You Begin

Before you apply the Best Practice for tuning the node transition time, you must understand the following background information:

- Factors that affect how low the `cluster_rebuild_delay` attribute can be set.
- The consequences of incorrectly setting the `cluster_rebuild_delay` attribute.
- The results of testing.
- The different methods available for modifying system attributes.

## Factors Affecting How Low the `cluster_rebuild_delay` Attribute Can be Set

The following factors can affect how low you can set the `cluster_rebuild_delay` attribute:

- System size and load
- Cluster interconnect configuration and use
- Number and model mix of cluster members
- Storage configuration
- System and component reliability

## System Size and Load

The system size and load are two of the most important factors to consider when lowering the `cluster_rebuild_delay` value. If all members have been sized or are loaded to run at a maximum load for their configuration, then lowering the `cluster_rebuild_delay` value may result in members being inappropriately removed from the cluster. However, if cluster members are loaded with a significant amount of spare processing capability, the `cluster_rebuild_delay` value may be significantly lowered without cluster members being inappropriately removed from the cluster.

## Cluster Interconnect Configuration and Usage

Differences between Memory Channel and the LAN cluster interconnect technologies result in significant differences in client availability during cluster member node transitions. A Memory Channel interconnect immediately detects a node down transition. A LAN interconnect relies on TCP keep alive timeouts to detect a node down transition. A LAN interconnect may take up to half of the `cluster_rebuild_delay` value to detect a node down condition.

Loading on the cluster interconnect is another factor that may limit how low the `cluster_rebuild_delay` value may be set. Configurations using the interconnect as a failover communication channel, such as NFS severing, will have much greater freedom in lowering the `cluster_rebuild_delay` value. Configurations that transfer large amounts of data across the cluster interconnect may be limited in how low the value can be set before seeing members inappropriately removed.

## Number and Model Mix of Cluster Members

The number of members in a cluster may also affect how low to set the `cluster_rebuild_delay` value. While a lightly loaded two-member cluster may be able to use the minimum value for `cluster_rebuild_delay`, a lightly loaded eight-member cluster may be limited to using half the default value. The reason for this is related to cluster quorum algorithms, which require a series of synchronized transactions between all cluster members during node transitions.

The mix of cluster members also affects how low you can set the `cluster_rebuild_delay` value. To maintain quorum during node transitions, a series of synchronized transactions between all cluster

members is required. This methodology dictates that the low-end systems generally limit the performance during node transitions. Clusters containing homogeneous system configurations may allow greater freedom in lowering the `cluster_rebuild_delay` value. Configurations that have a mix of high- and low-end systems may find that they are limited in how low the `cluster_rebuild_delay` value can be set before seeing inappropriate member removal.

## Storage Configuration

Storage configuration may also affect how low the `cluster_rebuild_delay` value may be set. Having a simple directly connected storage configuration simplifies much of the processing involved with storage and file system failover and therefore will allow for a greater freedom in lowering the `cluster_rebuild_delay` value. Configurations with complex storage configurations, including those with certain RAID and Fibre Channel configurations, may be limited in how low they can go before seeing inappropriate member removal.

## System and Component Reliability

System and component reliability is an additional factor which affecting how low the `cluster_rebuild_delay` value may be set. Although difficult to measure and quantify, having a cluster with questionable or intermittent hardware errors may make lowering the `cluster_rebuild_delay` value to any value but the default value all but impossible.

As hard drives, cables, and connections age, they may fail in a manner that is obvious and immediate or they may fail in a manner that is preceded by intermittent or recoverable failures. Intermittent or recoverable failures may result in retransmission and retries and backoff strategies, which make the series of synchronized transaction between all cluster members during node transitions difficult to complete. Removing and replacing all intermittent hardware will allow greater freedom in lowering the `cluster_rebuild_delay` value

## Effects of Lowering the `cluster_rebuild_delay` Value Too Much

Lowering the `cluster_rebuild_delay` value too much will result in cluster behavior that will limit availability or actually increase the time associated with cluster member node transition. If the `cluster_rebuild_delay` value is lowered too much members may suffer from intermittent loss of communication across the interconnect, which can lead to cluster members being removed from the cluster. Additionally,

members attempting to join a cluster where the `cluster_rebuild_delay` value is lowered too much may find that they are unable to join the cluster. The probability of either of these behaviors is directly related to all of the previous factors.

## Test Description and Methodology

To demonstrate some of the factors limiting the lowering of the `cluster_rebuild_delay` value, a number of clusters were configured using a Gigabit Ethernet interface as the LAN cluster interconnect. Each cluster was configured as an NFS server serving a number of CFS file systems. An NFS client, which is not part of the cluster, was configured and applications were started to create varying amounts of NFS load and monitor NFS availability during member node transitions. Once the client load and monitoring programs are started on the client, the cluster members serving the `cluster_lockd` daemon are halted, forcing a cluster node down transition. Following each test run the `cluster_rebuild_delay` value was lowered and the test repeated.

This process was repeated for each cluster configuration until it was determined that the cluster members were being inappropriately removed from the cluster, or cluster nodes were unable to join the cluster.

## Test Results: Tru64 UNIX Version 5.1B and Patch Kit

The following tables show the failover times (in seconds) associated with different `cluster_rebuild_delay` values and different loads for different cluster configurations:

### Two-Node AlphaServer 4100 Cluster with 100-Mb LAN Interconnect and HSG80 Storage

<code>cluster_rebuild_delay</code> Value	Light Load	Heavy Load
240	257	311
120	166	183
90	144	165
60	76	131
30	60	109
20	52	170

### **Eight-Node Mixed Cluster with 1-Gb LAN Interconnect and HSV110 Storage**

<b>cluster_rebuild_delay Value</b>	<b>Light Load</b>	<b>Heavy Load</b>
240	230	246
120	125	160
90	114	153
60	84	135
30	115	150
20	120	147

## **Modifying System Attributes**

You have two tools that can be used to modify the `cluster_rebuild_delay` attribute:

- `dxkerneltuner` - This is the Kernel Tuner GUI. To access the GUI through the Common Desktop Environment (CDE) Application Manager window, select the `System_Admin` icon and then select the `MonitoringTuning` icon. Choose the subsystem whose attribute you want to modify, and enter the new value in the Current Value field. See `dxkerneltuner(8)` for more information.
- Edit the `/etc/sysconfigtab` file. See `sysconfigtab(4)` for more information.

## **Applying the Best Practice**

Before you apply this best practice, be sure to follow the recommendations in *Before You Begin*.

1. Modify the `cluster_rebuild_delay` attribute for each cluster members `sysconfigtab` file.

### **NOTE**

The `cluster_rebuild_delay` value must be set to the same value on each cluster member.

The minimum value supported for the `cluster_rebuild_delay` is 50.

2. Shut down the entire cluster and then reboot each member.
3. Apply a typical load to the cluster.
4. Halt or power cycle different members.
5. Note client availability.
6. Note members that are not booting or being removed from the cluster.

## Verifying Success

After you apply this Best Practice for tuning the `cluster_rebuild_delay` attribute, you can verify whether it was successful as follows:

1. Verify that the `cluster_rebuild_delay` attribute has been changed:

```
# sysconfig -q clubase cluster_rebuild_delay
```

2. To verify that the value has improved availability during node transitions without the adverse effect previously described, apply a typical load to the system and then reset or halt various members of the cluster. As long as all members are able to boot and no members are unexpectedly removed from the cluster, the lower `cluster_rebuild_delay` value should improve availability during node transitions without other adverse effects.

If the Best Practice was not successful, see *Troubleshooting* for information about identifying and solving problems.

## Troubleshooting

If over time members are unable to boot or if members are unexpectedly removed from the cluster then it is likely that something in the cluster has changed since the `cluster_rebuild_delay` attribute value was lowered. Repeat the Applying the Best Practice procedure, raising the `cluster_rebuild_delay` attribute value as necessary.

## Comments and Questions

We value your comments and questions on the information in this document. Please mail your comments to us at this address:

[best\\_practices@zk3.dec.com](mailto:best_practices@zk3.dec.com)

## Legal Notice

All other product names mentioned herein may be trademarks of their respective owners.

Confidential computer software. Valid license from HP and/or its subsidiaries required for possession, use, or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Neither HP nor any of its subsidiaries shall be liable for technical or editorial errors or omissions contained herein. The information is provided "as is" without warranty of any kind and is subject to change without notice. The warranties for HP products are set forth in the express limited warranty statements accompanying such products. Nothing herein should be construed as constituting an additional warranty.