|  |  |  |
| --- | --- | --- |
| **SS Timeout Computation in Oracle Clusterware (Doc ID 294430.1)** | [To Bottom](https://support.oracle.com/epmos/faces/DocumentDisplay?_afrLoop=183450221299725&id=294430.1&_adf.ctrl-state=1a25pamq5g_555%20\o%20To%20Bottom) |  |





|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **In this Document**   |  |  | | --- | --- | |  | [Purpose](https://support.oracle.com/epmos/faces/DocumentDisplay?_afrLoop=183450221299725&id=294430.1&_adf.ctrl-state=1a25pamq5g_555%20\l%20PURPOSE) | |  | [Scope](https://support.oracle.com/epmos/faces/DocumentDisplay?_afrLoop=183450221299725&id=294430.1&_adf.ctrl-state=1a25pamq5g_555%20\l%20SCOPE) |  |  |  | | --- | --- | |  | [Details](https://support.oracle.com/epmos/faces/DocumentDisplay?_afrLoop=183450221299725&id=294430.1&_adf.ctrl-state=1a25pamq5g_555%20\l%20BODYTEXT) | |  | [References](https://support.oracle.com/epmos/faces/DocumentDisplay?_afrLoop=183450221299725&id=294430.1&_adf.ctrl-state=1a25pamq5g_555%20\l%20REF) |     **APPLIES TO:**  Oracle Database - Enterprise Edition - Version 10.1.0.2 to 12.2.0.1 [Release 10.1 to 12.2]  Information in this document applies to any platform.  Oracle Clusterware  **PURPOSE**  The purpose of this Note is to document default CSS misscount timeout calculations in 10g  Release 1,  10g Release 2 , 11g and higher versions.  **SCOPE**   * Define misscount parameter * Define the default calculations for the misscount parameter * Describe Cluster Synchronization Service (CSS) heartbeats and their interrelationship * Describe the cases where the default calculation may be too sensitive   **DETAILS**  **MISSCOUNT DEFINITION AND DEFAULT VALUES**  The CSS misscount parameter represents the maximum time, in seconds, that a network heartbeat can be missed before entering into a cluster reconfiguration to evict the node. The following are the default values for the misscount parameter and their respective versions when using Oracle Clusterware\* in seconds:   |  |  |  | | --- | --- | --- | | **OS** | **10g (R1 &R2)** | **11g** | | Linux | 60 | 30 | | Unix | 30 | 30 | | VMS | 30 | 30 | | Windows | 30 | 30 |   \*CSS misscount default value when using vendor (non-Oracle) clusterware is 600 seconds. This is to allow the vendor clusterware ample time to resolve any possible split brain scenarios.  On AIX platforms with HACMP starting with 10.2.0.3 BP#1, the misscount is 30. This is documented in [Note 551658.1](https://support.oracle.com/epmos/faces/DocumentDisplay?parent=DOCUMENT&sourceId=294430.1&id=551658.1)  **CSS HEARTBEAT MECHANISMS AND THEIR INTERRELATIONSHIP**  The synchronization services component (CSS) of the Oracle Clusterware maintains two heartbeat mechanisms 1.) the disk heartbeat to the voting device and 2.) the network heartbeat  across the interconnect which establish and confirm valid node membership in the cluster. Both of these heartbeat mechanisms have an associated timeout value. The disk heartbeat has an internal i/o timeout interval (DTO Disk TimeOut), in seconds, where an i/o to the voting disk must complete. The misscount parameter (MC), as stated above, is the maximum time, in seconds, that a network heartbeat  can be missed. The disk heartbeat i/o timeout interval is directly related to the misscount parameter setting. There has been some variation in this relationship  between versions as described below:   |  |  | | --- | --- | | 9.x.x.x | NOTE, MISSCOUNT WAS A  DIFFERENT ENTITY IN THIS RELEASE | | 10.1.0.2 | No one should be on this version | | 10.1.0.3 | DTO = MC - 15 seconds | | 10.1.0.4 | DTO = MC - 15 seconds | | 10.1.0.4+Unpublished Bug 3306964 | DTO = MC - 3 seconds | | 10.1.0.4 with CRS II Merge patch | DTO =Disktimeout (Defaults to 200 seconds) Normally OR Misscount seconds only during initial Cluster formation or Slightly before reconfiguration | | 10.1.0.5 | IOT = MC - 3 seconds | | 10.2.0.1 +Fix for unpublished Bug 4896338 | IOT=Disktimeout (Defaults to 200 seconds) Normally OR Misscount seconds only during initial Cluster formation or Slightly before reconfiguration | | 10.2.0.2 | Same as above (10.2.0.1 with Patch [Bug:4896338](https://support.oracle.com/epmos/faces/BugDisplay?parent=DOCUMENT&sourceId=294430.1&id=4896338%20\t%20_blank) | | 10.1 - 12.2 | During node join and leave (reconfiguration) in a cluster we need to reconfigure, in that particular case we use Short Disk TimeOut (SDTO) which is in all versions SDTO = MC – reboottime (usually 3 seconds) |   Misscount drives cluster membership reconfigurations and directly effects the availability of the cluster. In most cases, the default settings for MC should be acceptable.  Modifying the default value of misscount not only influences the timeout interval for the i/o to the voting disk, but also influences the tolerance for missed network heartbeats across the interconnect.  **LONG LATENCIES TO THE VOTING DISKS**  If I/O latencies to the voting disk are greater than the default DTO calculations noted above, the cluster may experience CSS node evictions depending on (a)the Oracle Clusterware (CRS) version, (b)whether merge patch has been applied and (c)the state of the Cluster. More details on this are covered in the section "Change in Behavior with CRS Merge PATCH (4896338 on 10.2.0.1)".  These latencies can be attributed to any number of problems in the i/o subsystem or problems with any component in the i/o path. The following is a non exhaustive list of reported problems which resulted in CSS node eviction due to latencies to the voting disk longer than the default Oracle Clusterware i/o timeout value(DTO):   1. QLogic HBA cards with a Link Down Timeout greater than the default misscount. 2. Bad cables to the SAN/storage array that effect i/o latencies 3. SAN switch (like Brocade) failover latency greater than the default misscount 4. EMC Clariion Array when trespassing the SP to the backup SP greater than default misscount 5. EMC PowerPath path error detection and I/O repost and redirect greater than default misscount 6. NetApp Cluster (CFO) failover latency greater than default misscount 7. Sustained high CPU load which effects the CSSD disk ping monitoring thread 8. Poor SAN network configuration that creates latencies in the I/O path.   The most common problems relate to multi-path IO software drivers, and the reconfiguration times resulting from a failure in the IO path. Hardware and (re)configuration issues that introduce these latencies should be corrected. Incompatible failover times with underlying OS, network or storage hardware or software may be addressed given a complete understanding of the considerations listed below.  **Misscount should NOT be modified to workaround the above-mentioned issues. Oracle support recommends that you apply the latest patchset which changes the CSS behaviour.** More details covered in next section.  **Change in Behavior with**[**Bug:4896338**](https://support.oracle.com/epmos/faces/BugDisplay?parent=DOCUMENT&sourceId=294430.1&id=4896338%20\t%20_blank) **applied on top of 10.2.0.1**  Starting with 10.2.0.1+[Bug:4896338](https://support.oracle.com/epmos/faces/BugDisplay?parent=DOCUMENT&sourceId=294430.1&id=4896338%20\t%20_blank), CSS will not evict the node from the cluster due to (DTO) I/O to voting disk taking more than misscount seconds unless it is during the initial cluster formation or slightly before reconfiguration.  So if we have a N number of nodes in a cluster and one of the nodes takes more than misscount seconds to access the voting disk, the node will not be evicted as long as the access to the voting disk is completed within disktimeout seconds. Consequently with this patch, there is no need to increase the misscount at all.  Additionally this merge patch introduces Disktimeout  which is the amount of time that a lack of disk ping to voting disk(s) will be tolerated.  Note:  applying the patch will not change your value for Misscount.  The table below explains in the conditions under which the eviction will occur   |  |  |  | | --- | --- | --- | | **Network Ping** | **Disk Ping** | **Reboot** | | Completes within misscount seconds | Completes within Misscount seconds | N | | Completes within Misscount seconds | Takes more than misscount seconds but less than Disktimeout seconds | N | | Completes within Misscount seconds | Takes more than Disktimeout seconds | Y | | Takes more than Misscount Seconds | Completes within Misscount seconds | Y |                 \* By default Misscount is less than Disktimeout seconds  **CONSIDERATIONS WHEN CHANGING MISSCOUNT FROM THE DEFAULT VALUE**   1. Customers drive SLA and cluster availability. The customer ultimately defines Service Levels and availability for the cluster. Before recommending any change to misscount, the full impact of that change should be described and the impact to cluster availability measured. 2. Customers may have timeout and retry logic in their applications. The impact of delaying reconfiguration may cause 'artificial' timeouts of the application, reconnect failures and subsequent logon storms. 3. Misscount timeout values are version dependent and are subject to change. As we have seen, misscount calculations are variable between releases and between versions within a release. Creating a false dependency on misscount calculation in one version may not be appropriate for later versions. 4. Internal I/O timeout interval (DTO) algorithms may change in later releases as stated above, there exists a direct relationship between the internal I/O timeout interval and misscount. This relationship is subject to change in later releases. 5. An increase in misscount to compensate for i/o latencies directly effects reconfiguration times for network failures. The network heartbeat is the primary indicator of connectivity within the cluster. Misscount is the tolerance level of missed 'check ins' that trigger cluster reconfiguration. Increasing misscount will prolong the time to take corrective action in the event of network failure or other anomalies effecting the availability of a node in the cluster. This directly effects cluster availability. 6. Changing misscount to workaround voting disk latencies will need to be corrected when the underlying disk latency is corrected, misscount needs to be set back to the default The customer needs to document the change and set the parameter back to the default when the underlying storage I/O latency is resolved. 7. Do not change default misscount values if you are  running Vendor Clusterware along with Oracle Clusterware. The default values for misscount should not be changed when using vendor clusterware. Modifying misscount in this environment may cause clusterwide outages and potential corruptions. 8. Changing misscount parameter incurs a clusterwide outage. As note below, the customer will need to schedule a clusterwide outage to make this change. 9. Changing misscount should not be used to compensate for poor configurations or faulty hardware 10. Cluster and RDBMS availability are directly effected by high misscount settings. 11. In case of stretched clusters and stretched storage systems and a site failure where we lose one storage and N number of nodes we go into a reconfiguration state and then we revert to ShortDiskTimeOut value as internal I/O timeout for the votings. Several cases are known with stretched clusters where when a site failure happen the storage failover cannot complete within SDTO. If the I/O to the votings is blocked more than SDTO the result is node evictions on the surviving side.   **To Change MISSCOUNT back to default Please refer to**[**Note:284752.1**](https://support.oracle.com/epmos/faces/DocumentDisplay?parent=DOCUMENT&sourceId=294430.1&id=284752.1)  THIS IS THE ONLY SUPPORTED METHOD. NOT FOLLOWING THIS METHOD RISKS EVICTIONS AND/OR CORRUPTING THE OCR  **10g Release 2 MIRRORED VOTING DISKS AND VENDOR MULTIPATHING SOLUTIONS**  Oracle RAC 10g Release 2 allows for multiple voting disks so that  the customer does not have to rely on a multipathing solution from a storage vendor. You can have n voting disks (up to 31) where n = m\*2+1 where m is the number of disk failures you  want to survive. Oracle recommends each voting disk to be on a separate physical disk.    **Database - RAC/Scalability Community**  To discuss this topic further with Oracle experts and industry peers, we encourage you to review, join or start a discussion in the My Oracle Support [Database - RAC/Scalability Community](https://community.oracle.com/community/support/oracle_database/database_-_rac_scalability%20/o%20Database%20-%20RAC/Scalability%20Community)  **REFERENCES**    [BUG:4896338](https://support.oracle.com/epmos/faces/BugDisplay?parent=DOCUMENT&sourceId=294430.1&id=4896338%20\t%20_blank) - PLACEHOLDER BUG FOR PCW 10.2.0.1 MERGE FOR VERY LOW BROWNOUT - 10.2  [NOTE:284752.1](https://support.oracle.com/epmos/faces/DocumentDisplay?parent=DOCUMENT&sourceId=294430.1&id=284752.1) - Steps To Change CSS Misscount, Reboottime and Disktimeout  [NOTE:551658.1](https://support.oracle.com/epmos/faces/DocumentDisplay?parent=DOCUMENT&sourceId=294430.1&id=551658.1) - CSS errors out with clssnm\_skgxninit: Compatible vendor clusterware not in use  [NOTE:559365.1](https://support.oracle.com/epmos/faces/DocumentDisplay?parent=DOCUMENT&sourceId=294430.1&id=559365.1) - Pre-11.2: Using Diagwait as a diagnostic to get more information for diagnosing Oracle Clusterware Node evictions |