

An Oracle White Paper  
June 2010

Oracle Berkeley DB Java Edition  
High Availability -  
Large Configuration and Scalability Testing

Executive Overview .....	1
Introduction .....	1
Test Environment .....	2
Performance Analysis.....	2
Test 1 – Impact of Replication .....	2
Test 2 – Impact of Adding Replicas .....	7
Conclusion .....	10
Appendix .....	11
Typical GC, JVM, and JE Parameters .....	11

## Executive Overview

This whitepaper examines the performance of Berkeley DB Java Edition (JE) in an High Availability (HA) environment on large-system hardware configurations. The tests are broken into two general categories:

- Tests measuring the overhead of JE HA in various configurations
- Tests measuring the overhead of additional JE HA replicas

These results can help estimate JE HA performance for an application. The actual performance of your application may vary based on data size, access patterns, cache size, I/O subsystem, file system, operating system, processor speed, and memory size. This paper only examines record creation (insert) performance.

## Introduction

Oracle Berkeley DB Java Edition is an embeddable database implemented in pure Java. It provides a transactional storage engine that reduces the overhead of object persistence, while improving the flexibility, speed, and scalability of object to relation mapping (ORM) solutions. An High Availability (HA) option provides the ability to scale read performance as well as improve reliability through a failover mechanism. JE has various APIs including positional key/value storage of binary data and the Direct Persistence Layer (DPL) that provides a Plain Old Java Object (POJO) interface through Java annotations.

## Test Environment

All tests are run using an internal testing framework called *ScaleTest*. While *ScaleTest* provides many functions, only the record creation (insert) function is measured.

All tests are conducted using the following configuration:

- Solaris 10 with UFS (except for the SSD tests which used ZFS)
- Sun JDK 1.6.0
- JE 4.0.103
- Key size of 10 bytes and data size of 100 bytes
- Sun/Oracle x86, Intel Xeon (Nehalem) processors for Test 1
- Sun T2000 Sun Niagra hardware for Test 2
- 16 GiB JVM heap and 7GB JE cache (except where specifically stated otherwise)
- Gigabit Ethernet (GbE) network

Typical JVM and GC parameters are specified in the Appendix.

## Performance Analysis

### Test 1 – Impact of Replication

For the first set of tests, the following machines (all Sun/Oracle x4450 and x2270 x86 Nehalem hardware) are used:

TABLE 1. HARDWARE CONFIGURATION FOR TEST 1						
HOST	MAKE	MODEL	PROCESSORS	MEMORY	DISK	
1	SunFire	x4450	4x E7220 dualcore @ 2.93GHz	32GB	ST6140 disk cache disabled	
1c	SunFire	x4450	4x E7220 dualcore @ 2.93GHz	32GB	ST6140 disk cache enabled	
2	SunFire	x4450	4x E7220 dualcore @ 2.93GHz	32GB	4x146GB SAS (10kRPM)	
3	SunFire	x2270	2x E5570 duadcore, HT @ 2.93GHz	48GB	500GB SATA + SSD	
4	Sunfire	x2270	2x E5570 quadcore, HT @ 2.93GHz	32GB	500GB SATA	

Host 1 is configured with an ST6140 disk array in which the disk cache is enabled (noted as *Host 1c* in the tables) and disabled (noted as *Host 1* in the tables). The goal of this set of tests is to start with a baseline configuration representing a fully-durable (Sync transaction durability), non-replicated system and then go on to measure the impact of enabling and disabling the on-board hardware disk cache for both the StorageTek ST6140 storage array as well as a commodity 7200 RPM SATA HDD. In the Sync case HDD, storage array, and SSD performance are also compared. The configuration change from Sync to NoSync is to measure *best case*, albeit non-durable, insert performance. Replication is added to the Sync configuration on the same host to determine the overhead. From there the Sync durability is disabled and replicas are added, first in a single-replica configuration and then a 2-replica configuration. This measures various configurations using the network for durability. The following table summarizes the results of these tests. Each scenario is discussed in more detail below.

TABLE 2. TEST RESULTS OF TEST 1

SCENARIO	RECORDS	AVG OPS/SEC	STDDEV	MIN	MAX	HOST	COMMENTS
1	5M	491	55	76	689	1	Sync, non-replicated, no disk cache
1	5M	5107	173	4385	5327	1c	Sync, non-replicated, disk cache
1-SATA	5M	260				3	Sync, non-replicated, SATA, no wc
1-SATAwc	5M	3076				3	Sync, non-replicated, SATA, wc
1-SATAns	5M	158753				3	NoSync, non-replicated, SATA, wc
1-SATAns	30M	133695				3	NoSync, non-replicated, SATA, wc
1-SSD	5M	8492	349	7396	9040	3	Sync, non-replicated, SSD
2	5M	485	54	54	659	1	Sync, master-only replication
2-SSD	5M	8652	281	7898	9036	3	Sync, master-only replication, SSD
2A	5M	49375	1979	45981	53115	1	NoSync, master-only replication
3	30M	41837	5723	26541	50579	1	NoSync, master-only replication
3	100M	36531	6426	6109	52254	1	NoSync, master-only replication
4	30M	15818	2155	10372	19523	1,2	NoSync, single replica
4-cache	30M	18040	1153	14689	20305	1c,2	NoSync, single replica, disk cache
5	30M	18929	1455	14345	21622	1,2,3	NoSync, two replicas
5	30M	19589	964	16726	21441	1c,2,3	NoSync, two replicas, disk cache

### Scenario 1, 1-SATA, and 1-SATAwc, 1-SATAns (5M), 1-SATAns (30M), 1-SSD

This set of tests measures full sync durability, non-replicated performance and provides a baseline to compare the other tests. A JE application that requires full sync durability using either a disk array or standard direct-attached (for example, SATA) drive can run in this configuration. Several comparisons are made.

TABLE 3. TEST RESULTS SCENARIO 1, 1-SATA, AND 1-SATAWC, 1-SATANS (5M), 1-SATANS (30M), 1-SSD

SCENARIO	RECORDS	AVG OPS/SEC	STDDEV	MIN	MAX	HOST	COMMENTS
1	5M	491	55	76	689	1	Sync, non-replicated, no disk cache
1	5M	5107	173	4385	5327	1c	Sync, non-replicated, disk cache
1-SATA	5M	260				3	Sync, non-replicated, SATA, no wc
1-SATAwc	5M	3076				3	Sync, non-replicated, SATA, wc
1-SATAns	5M	158753				3	NoSync, non-replicated, SATA, wc
1-SATAns	30M	133695				3	NoSync, non-replicated, SATA, wc
1-SSD	5M	8492	349	7396	9040	3	Sync, non-replicated, SSD

First, comparing Scenario 1 to Scenario 1-SATA demonstrates the difference in performance between the StorageTek disk array and a commodity SATA drive (500GB, 7200RPM). (491 versus 260 inserts/sec, respectively)

Comparing 1-SATA to 1-SATAwc shows the effect of enabling the on-board write-cache (260 versus 3076 inserts/sec).

Comparing 1-SATAwc and 1-SATAns (5M) shows the difference (3076 versus 158,753) between sync and no-sync with the write-cache enabled. Comparing 1-SATAns (5M) to 1-SATAns (30M) shows that this naturally degrades for an higher number of records.

Comparing scenario 1 with host 1-nc (the disk cache disabled) and host 1 (disk cache enabled) shows dramatic improvement because of the battery backup cache on the ST6140. With the cache enabled, durability is still maintained.

Finally, comparing 1-SATA to 1-SSD shows commodity HDD performance versus SSD performance (260 versus 8,492 inserts/sec). So expect significant improvements when you use SSDs.

## Scenario 2

This test measures the overhead of adding replication to a single-node sync scenario. Inserts are made on a single machine, which is configured as a replication master (that is, there is no replica). Sync durability is specified.

TABLE 4. TEST RESULTS OF SCENARIO 2

SCENARIO	RECORDS	AVG OPS/SEC	STDDEV	MIN	MAX	HOST	COMMENTS
1	5M	491	55	76	689	1	Sync, non-replicated, no disk cache
1-SSD	5M	8492	349	7396	9040	3	Sync, non-replicated, SSD
2	5M	485	54	54	659	1	Sync, master-only replication
2-SSD	5M	8652	281	7898	9036	3	Sync, master-only replication, SSD

Comparing Scenario 1 to Scenario 2, the throughput drops from 491 to 485, indicating a relatively small overhead for adding replication when you use the ST6140 disk array. When you use an SSD, the change in throughput (actually shown as an increase from 8492 in 1-SSD to 8652 in 2-SSD) is also negligible.

## Scenario 2A-5M

This test measures no-sync performance in a non-replicated mode. This test is run to measure the cost of Sync mode in the single-node master-only replication scenario. 2A is the same as 3 except that it uses 5M records instead of 30M records.

TABLE 5. TEST RESULTS OF SCENARIO 2A-5M

SCENARIO	RECORDS	AVG OPS/SEC	STDDEV	MIN	MAX	HOST	COMMENTS
2	5M	485	54	54	659	1	Sync, master-only replication
2A	5M	49375	1979	45981	53115	1	NoSync, master-only replication

### Scenario 3, 4, 5

These scenarios measure replication in single-node, one-replica, and two-replica configurations. For completeness, Scenarios 1 and 2 are also included in the table excerpt below. Scenario 3-100M shows the effect of more records with a larger heap/cache size.

TABLE 6. TEST RESULTS OF SCENARIO 3, 4, 5

SCENARIO	RECORDS	AVG OPS/SEC	STDDEV	MIN	MAX	HOST	COMMENTS
1	5M	491	55	76	689	1	Sync, non-replicated, no disk cache
2	5M	485	54	54	659	1	Sync, master-only replication
3	30M	41837	5723	26541	50579	1	NoSync, master-only replication
3	100M	36531	6426	6109	52254	1	NoSync, master-only replication
4	30M	15818	2155	10372	19523	1,3,4	NoSync, single replica
4	30M	18040	1153	14689	20305	1c, 3,4	NoSync, single replica, disk cache
5	30M	18929	1455	14345	21622	1,2,3	NoSync, two replicas
5	30M	19589	964	16726	21441	1c,2,3	NoSync, two replicas, disk cache

A result that may be surprising is that the single-replica case (Scenario 4) is slower than the two-replica case (Scenario 5). This is because the master only waits for one acknowledgement from a replica and the faster one is the winner, thereby making the throughput faster. The two Scenarios, 4 and 5, are with and without the disk cache on the master and demonstrate the results when the master disk is not a bottleneck.

## Test 2 – Impact of Adding Replicas

This test runs ScaleTest in a single configuration except that additional replicas are added to the replication group to measure the degradation caused by additional feeder overhead. For this test Sun T2000 hardware is used. All tests are run with a 6GB heap and a 4GiB cache.

TABLE 7. HARDWARE CONFIGURATION FOR TEST 2

MAKE	MODEL	PROCS	MEMORY	DISK
Sun Fire	T2000	8 core x 4 thread (1GHz)	32GB	2x72GB SAS 10k rpm
Sun Fire	T2000	6 core x 4 thread (1GHz)	16GB	3x72GB SAS 10k rpm
Sun Fire	T2000	6 core x 4 thread (1GHz)	8GB	3x72GB SAS 10k rpm
Sun Fire	T2000	6 core x 4 thread (1GHz)	8GB	2x72GB SAS 10k rpm
Sun Fire	T2000	6 core x 4 thread (1GHz)	8GB	3x72GB SAS 10k rpm

While the machines used are not completely homogeneous in terms of processor/memory/disk, the test is sized small enough so that the minor differences in configuration do not make a difference. Further, because the CPU utilization is relatively small, it is possible to run multiple replicas per node when more than 4 replicas are called for.

Two sets of tests are run - one with 3M records to observe the effects of adding replicas in a completely-in-cache scenario, and one with 30M records to observe the same, but with an out-of-cache configuration. In both sets of tests, the disk-write cache is disabled.

TABLE 8. TEST RESULTS FOR 3M RECORDS

N Replicas	Avg inserts/sec	Min	Max
1	3197	2531	3532
2	3189	2628	3517
3	3151	2528	3503
4	3168	2459	3558
5	3122	2589	3419
10	3083	2229	3396
15	2821	1985	3210

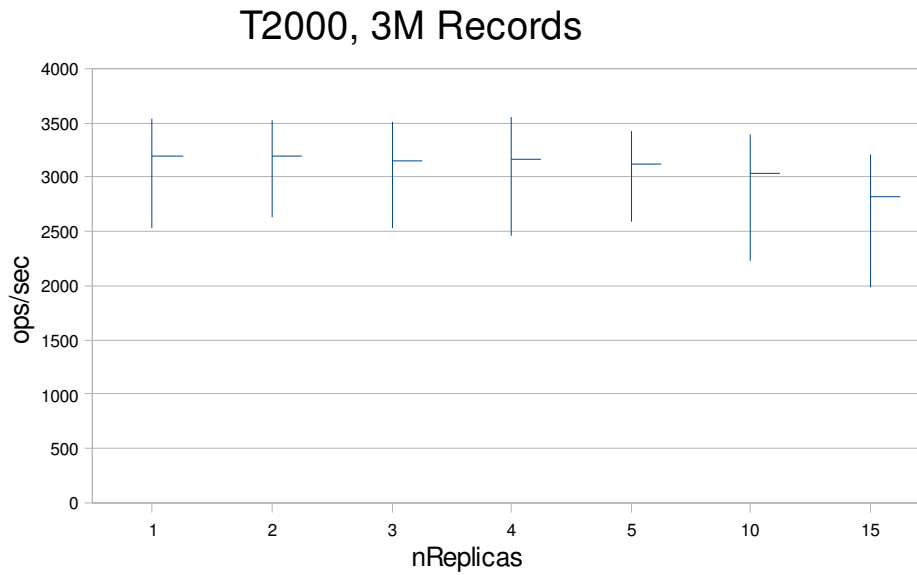
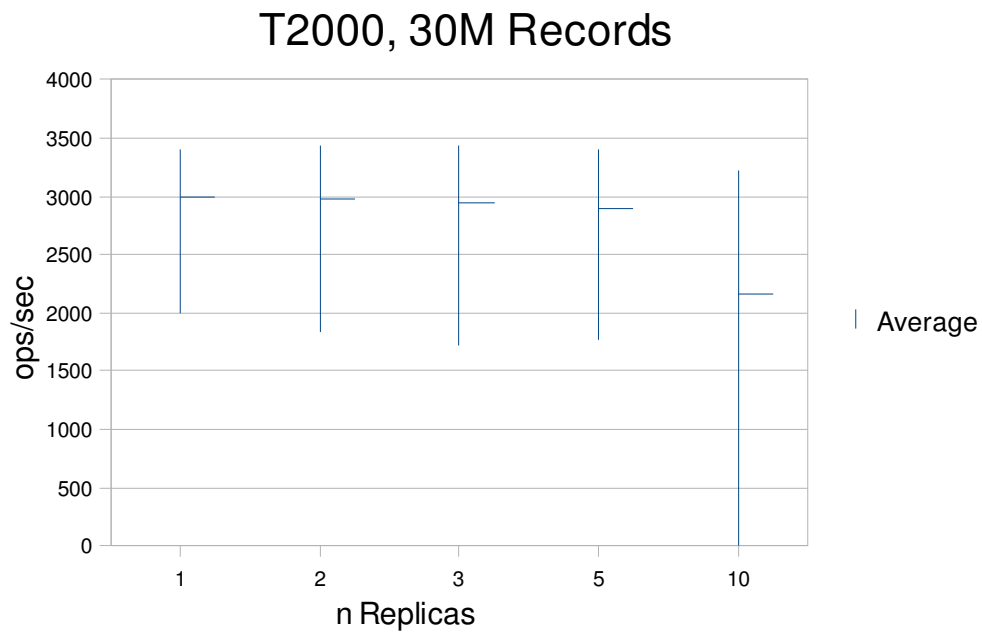


TABLE 9. TEST RESULTS FOR 30M RECORDS

N Replicas	Avg inserts/sec	Min	Max
1	2996	2001	3406
2	2984	1828	3436
3	2943	1725	3432
5	2888	1776	3402
10	2163	1	3216



This demonstrates that an HA environment experiences minimal performance degradation until at least 5 replicas are configured.

## Conclusion

This paper examines the performance of Oracle Berkeley DB Java Edition High Availability during the record creation (insert) operation. The first set of tests demonstrates that the overhead of adding HA to a configuration is minimal and that using an HA replica to provide durability (committing data to the network) can provide a performance improvement over durability with an HDD. The second set of tests demonstrates that adding replicas to a configuration causes minimal degradation.

You can download Oracle Berkeley DB Java Edition at:

<http://www.oracle.com/technology/software/products/berkeley-db/je/index.html>

You can post your comments and questions at the Oracle Technology Network (OTN) forum for Oracle Berkeley DB Java Edition at:

<http://forums.oracle.com/forums/forum.jspa?forumID=273>

For sales or support information email: [berkeleydb-info\\_us@oracle.com](mailto:berkeleydb-info_us@oracle.com)

Find out about new product releases by sending an email to: [bdb-join@oss.oracle.com](mailto:bdb-join@oss.oracle.com)

## Appendix

### Typical GC, JVM, and JE Parameters

```
GC_PARAMS="-XX:+UseConcMarkSweepGC -XX:+UseParNewGC -XX:CMSInitiatingOccupancyFraction=75 -  
XX:+DisableExplicitGC -XX:+PrintTenuringDistribution -XX:+PrintGCDetails -XX:+PrintGCTimeStamps -  
Xloggc:gc.log -XX:+PrintReferenceGC -XX:+UseLargePages"
```

```
HEAP="-d64 -Xms16g -Xmx16g -XX:NewSize=4g -XX:MaxNewSize=4g -XX:+UseCompressedOops"
```

```
JE="-je.maxMemory 7000000000 -je.txn.durability sync, sync, simple_majority -je.log.fileMax 100000000 -  
je.checkpointer.bytesInterval 100000000 -je.cleaner.readSize 1048576 -je.evictor.lruOnly false -  
je.evictor.nodesPerScan 100 -je.nodeMaxEntries 256 -je.checkpointer.highPriority false -je.cleaner.threads 16 -  
je.log.writeQueueSize 2097152 -je.log.fileCacheSize 500"
```



Oracle Berkeley DB Java Edition  
High Availability -  
Large Configuration and Scalability Testing  
June 2010  
Author: Charles Lamb

Oracle Corporation  
World Headquarters  
500 Oracle Parkway  
Redwood Shores, CA 94065  
U.S.A.

Worldwide Inquiries:  
Phone: +1.650.506.7000  
Fax: +1.650.506.7200  
oracle.com



Oracle is committed to developing practices and products that help protect the environment

Copyright © 2010, Oracle and/or its affiliates. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 0110