

Best of Both Worlds: Scale-up with Scale-out

Scaling Very Large Databases with Oracle RAC using Intel EM64T mainstream Servers

Srinagesh Battula, Gayathri Seetharaman, Deepen Chakraborty, Ram Varra

Intel Technology Manufacturing Group

Intel Corporation

November 2007

Best of Both Worlds: Scale-up with Scale-out
Scaling Very Large Databases with Oracle RAC using
Intel EM64T mainstream servers

INTRODUCTION

Intel's Factory Automation Decision Support Systems (DSS) are used for making critical manufacturing decisions. These DSS systems store Intel's operational, planning, engineering analysis and process control data. A huge data explosion in decision support systems is occurring due to advanced manufacturing processes that support the ever-expanding Intel product pipeline. Rapid data retrieval from these systems is critical for timely manufacturing decisions. The current single instance 32-bit database implementation does not have sufficient headroom to accommodate the scalability requirements on hand. Added to that, administration of multiple DSS databases running in silos becomes a manageability nightmare. Intel's automation engineering team has been chartered to devise a solution to address these scalability problems on hand with a goal of reducing total cost of ownership.

CHALLENGES FACED

Increased projected sizes of the database and the number of users contributed to the corresponding increase in system capacity requirements. Existing single database instance implementations lacked sufficient headroom to address the scalability requirements. Some of challenges faced were due to:

- Limitations of 32-bit architecture (process memory foot print constraint; typical server configuration: 8 CPU and 8GB RAM). This limited the scalability of number of users and the query performance Service Level Agreements. Parallelism couldn't have been exploited to the fullest extent due to the resource constraints.
- Administration of multiple DSS databases running in silos became a manageability nightmare.
- Integration of data in multiple databases through cross database joins via DB links posed challenges in query performance.
- Capacity growth forces "re-platformization" i.e. existing system cannot be incrementally grown to accommodate scaling needs.

Rapid data retrieval from these DSS systems is critical for timely manufacturing decisions.

Existing single instance implementations lacked sufficient headroom to address the scalability requirements.

SCALING OPTION ANALYSIS

Intel’s Automation Engineering team had a mission to devise a solution to address scalability problems with lowest TCO, provisioning of linear incremental scalability and compatibility with existing Windows based applications.

The team performed internal benchmarking/scalability study using Oracle RAC on Intel mainstream EM64T servers for a Multi-Terabyte DSS database and observed “near-linear” scalability. Various other options that were considered were: Oracle RAC on Linux, Oracle on 32CPU/Windows Datacenter and Oracle on 32CPU/HPUX.

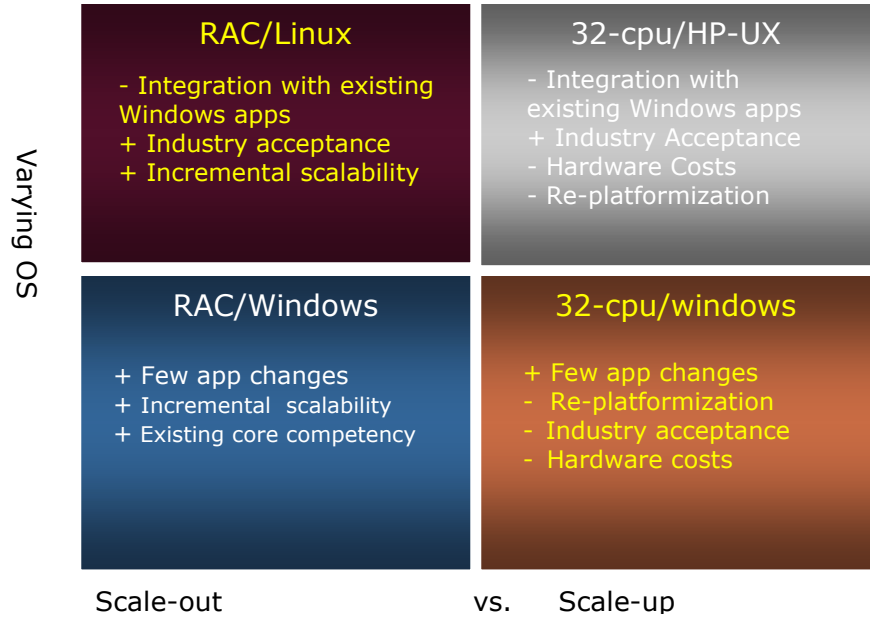


Figure1. Scaling Option Analysis

Pro-Con Analysis of various options resulted in choosing Oracle RAC on Windows 2003/64bit stack using Intel mainstream EM64T servers as a viable option from performance, scalability and TCO perspective. This chosen option allowed Intel Technology Manufacturing Group to obtain incremental scalability to accommodate dynamic capacity growth requirements. This also allowed the team to leverage existing core competency in Windows, and implement the stack with minimal modifications to existing applications.

Solution was to consolidate the multiple DSS databases on a single Oracle RAC cluster. Stack is built on Intel mainstream EM64T servers on 64-bit windows 2003 advanced server and 64-bit Oracle 10g Release 2 RAC/ASM/Clusterware.

Oracle RAC on Windows using Intel mainstream EM64T servers facilitated us to leverage existing core competency in Windows with minimal modifications to existing applications.

The stack allowed us to obtain incremental scalability to accommodate dynamic capacity growth requirements.

SCALABILITY USING INTEL BASED MAINSTREAM SERVERS AND RAC

Architecture has been validated against a projected 20TB database on a 10 node RAC cluster.

Choice was made to “beef-up” every node of the cluster, instead of a pure scale-out with smaller machines. Moving to 64-bit Intel servers enabled each individual node to be “scaled up” providing more processing power and a larger memory footprint essential for resource intensive DSS queries. Each node of the 10-node RAC cluster has 8CPUs and 32GB of memory. This node configuration facilitated us to “scale-up” from the previous configuration which constrained memory footprint due to the 32-bit architectural limitation. *Scaling out* with RAC enabled application workload partitioning. This facilitated resource intensive queries to take advantage of resources across multiple nodes without impacting the applications running on other domain nodes. The combination of scale-up/scale-out provides an effective balance between costs, manageability of the systems and scalability. Overall, this solution allowed us to take advantage of state of the art processors by seamlessly swapping with cost effective powerful Intel processors and realize better price/performance.

CONSOLIDATED DSS RAC CLUSTER

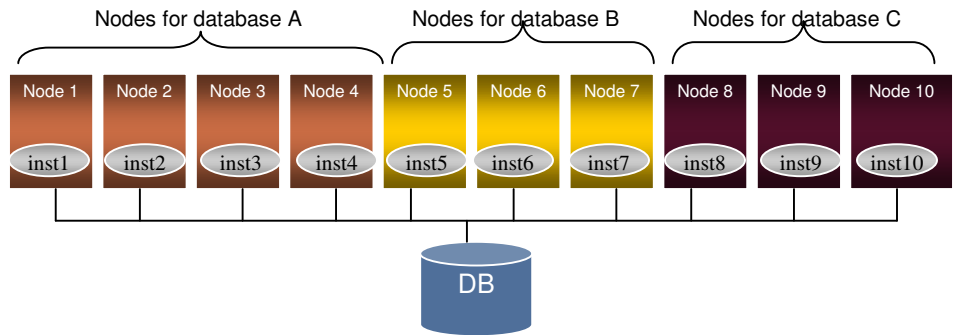


Figure2. Consolidated DSS RAC cluster

The final DSS cluster architecture is composed of ten nodes serving a single consolidated database. The three application domains (A, B, and C) have been consolidated as three schemas (DB_A, DB_B, and DB_C) in the clustered database.

Application Workload Management has been employed to confine workload of individual applications (ETL, Queries and Utilities etc) to separate nodes:

- DB_A : Node 1, 2, 3, 4
- DB_B : Node 5, 6, 7
- DB_C : Node 8, 9, 10

RAC CLUSTER - END TO END ARCHITECTURE

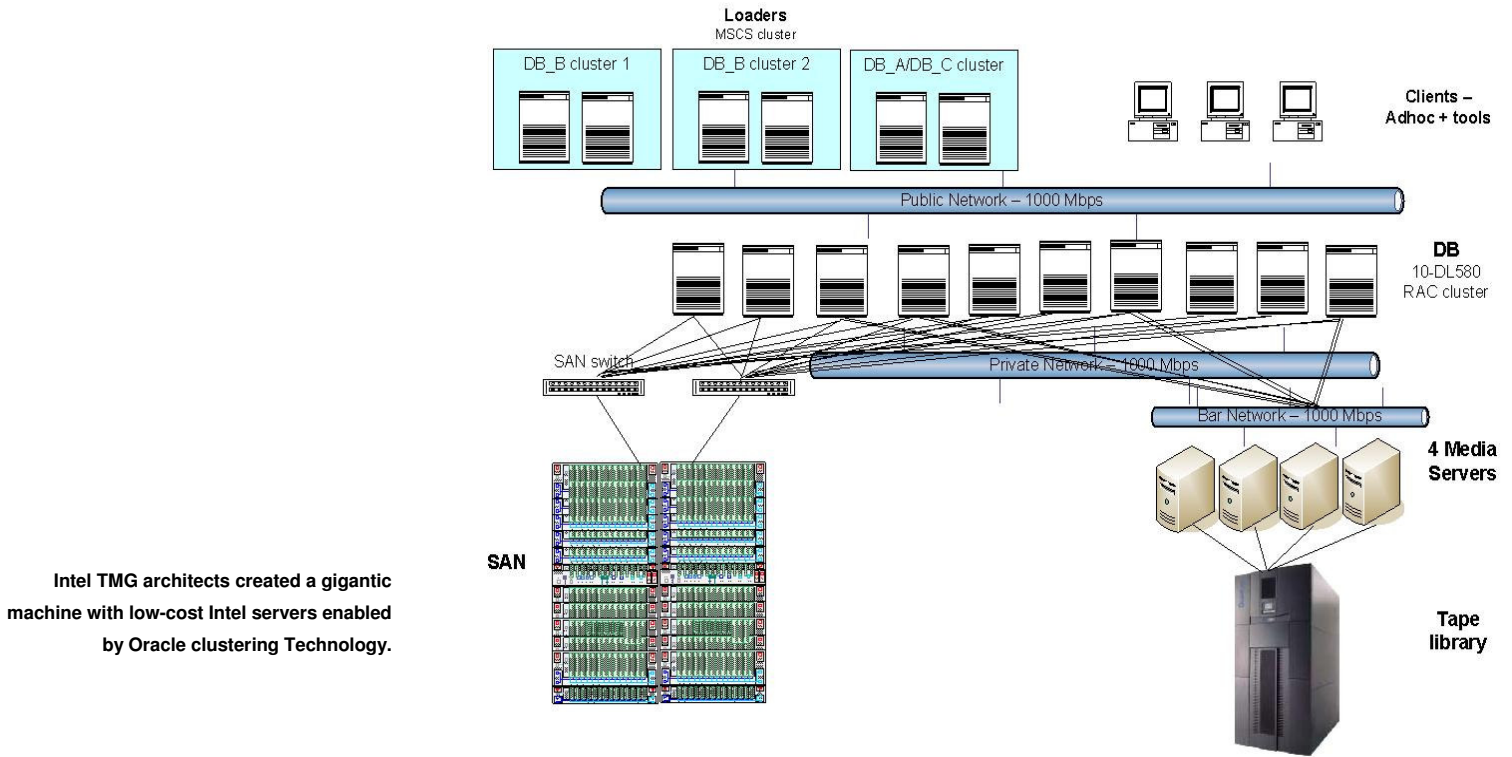


Figure3. End to End Cluster Architecture

The DSS cluster consists of 10 HP DL580/G4 nodes equipped with 4 Dual Core HT 3.4GHz XEONs CPUs running 64 bit Windows 2003 SP1 with “Large Page” memory configuration enabled. The clustering is implemented with 10g Release 2 64 bit Oracle Clusterware. This resulted in a gigantic “machine” with 80 CPU Cores and 320Gig of RAM carved out of Intel EM64T mainstream servers.

Dedicated Gigabit Ethernet VLAN (Virtual LAN) is employed for the private Cluster Interconnect. Another VLAN is created for Backup and Recovery operations to isolate backup network traffic from application workload.

The Consolidated clustered database runs on 10g Release 2. Separate Oracle Homes are used for keeping the Oracle database, ASM and Clusterware binaries. ASM with external redundancy is used for carving out disk groups meant for shared storage. Three raw LUNS are used for keeping three Voting Disks. The Oracle Cluster Registry (OCR) file is kept in two RAW LUNS.

Database Files reside on +DATA ASM disk group. Incrementally updated image copies are kept in the +FRA ASM disk group. Also, one copy of each of the

control files, redo log members and archive logs are stored in each of the +DATA and +FRA disk groups.

End to End backups are driven by RMAN both for Disk to Disk as well as Disk to Tape. Disk to Disk backups are performed from one of the dedicated cluster nodes with 8 RMAN channels streaming the data. Tape backup uses four of the dedicated cluster nodes with 2 RMAN channels per node. The scalability for the tape infrastructure is achieved via multiple dedicated network streams from the database nodes (2 per node), going to multiple media servers (up to 4). These media servers have multiple fiber channel connections to the tape library.

WORKLOAD MANAGEMENT

The usage of the three domains across the ten node cluster is streamlined using RAC's workload management capability. Server Side Connect-Time and Client Side connect-time Load Balancing are enabled to obtain an effective utilization of resources.

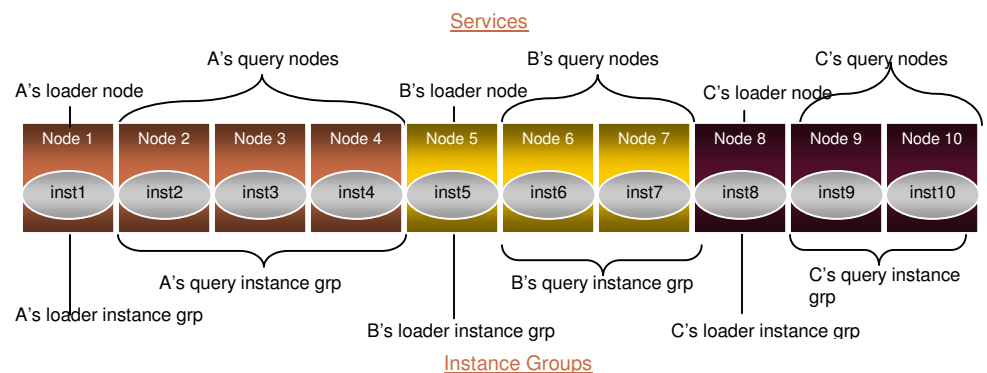


Figure4. Workload Management

Services have been created for each domain's Query and ETL components.

Transparent Application Failover took care of resilience to instance/node failures for the engineering community's queries.

- ▶ Dedicated one instance each for the three database domain's ETL component. (Nodes 1, 5, 8 as "PREFERRED", Nodes 2, 6, 9 as "AVAILABLE")
- ▶ Allocated separate set of nodes for reporting Query Services for each domain (Nodes 2, 3, 4 for App A, Nodes 6, 7 for App B, Nodes 9, 10 for App C).
- ▶ Transparent Application Failover (TAF) enabled for all query services. TAF took care of resilience to instance/node failures for the engineering community's queries.

In addition, we enabled multi-node parallelism via instance groups to match the services setup. Parallel instance groups assisted in exploiting the CPU and memory

resources of the given domain's nodes. Thus, these resources are not limited to a given node but to the given domain specific nodes of the cluster.

By matching the services setup and instance groups, we were also able to ensure that parallel executions of a query of one domain will not impact any other servers of a different domain in the same cluster.

APPLICATION INTEGRATION

In the pre-RAC environment, the ETL client components were configured to run on the database nodes. The high availability for the ETL and the database instance was provisioned by Microsoft Cluster Server (MSCS). As part of migrating to RAC, we decided to avoid the complexity integrating MSCS and Oracle Clusterware by separating the ETL client functions to dedicated MSCS nodes.

Apart from this, the we needed to make very minimal application modifications and mostly to the database connectivity (i.e., TNS entries with TAF and relevant Service Names)

HIGH AVAILABILITY AT ALL LAYERS

The stack has been configured in such a way that each component of the stack is capable of resilient to failures. The following table depicts this vividly.

The stack has been configured in such a way that each component of the stack is capable of resilient to failures.

Level	HA solution Used	Solution resilient to failures
Db Level	Oracle RAC	Db Node failures, db instance failures
Application Level	MSCS cluster	Node failures, application instance failures
App connectivity to Db	Services set to preferred/available and TAF	Db Node failures, db instance failures
Network Public/Private	Teamed NIC, going to two separate switches	NIC failures, Network switch failures
SAN – host level	2 HBA → 2 switches → 2 controllers per SAN	HBA on host failures, SAN switch failures
SAN level	RAID 10, 2 controllers per SAN	SAN , controller, disk failures
Backup level	4 media servers, 4 db server nodes, 2 connections from each server node	Channel failures, tape failures, tape software failures

Queries performed up to 5x faster. Primary contributor is the combination of powerful Intel mainstream servers and RAC's Workload management

RESULTS – SOME EXAMPLES

Data loads improved by 2X primarily due to the performance of 64-bit stack and Workload Management (by dedicating nodes for ETL components and shielding the resource intensive queries dedicated to run on Query service nodes). This architecture also allowed the system to easily scale to 2X concurrent users than original system.

We used a set of reference queries for benchmarking the user experience and it showed 5X improvement in response time. Primary contributor to this gain is the combination of power of each node (8 Intel XEON cores, 32GB RAM), Oracle database 10g 64-bit and RAC's workload management capability. The new architecture also allowed queries to be redesigned to take advantage of multi-node parallelism while maintaining sufficient headroom for the projected throughput.

Disk to Disk Level 0 backups run at 1TB/hr. Disk to Tape Level 0 backups run at 1.4TB/hr. We achieved this level of performance by exploiting multiple nodes in parallel to drive backup and stream the data to tape @ 1.4TB/hour.

We spent significant focus on vendor support for the end to end backup solution by working closely with each component provider (tape library, media management, RAC RDBMS)

KEY LEARNINGS

Goal is to keep the stack simple by minimizing the vendor integration touch points as much as possible.

- ▶ A key component of our success is multi-discipline team work. Integrated multi-domain team (Application owners, Database Administrators, System Administrators) is essential for delivering the system.
 - The creation of the end to end stack that is geared for reliability, availability, manageability, performance and scalability involves collective contributions of the expertise of a team of engineers from various domains. The team ought to be well versed with the understanding of the concepts of ASM Disk group layout, Workload partitioning, Cache Fusion etc., This teamwork also aids in rapid troubleshooting of the stack issues should they arise.
 - Solutions for high availability, reliability/performance evaluation, performance tuning, backup/recovery all were facilitated due to multi-domain team work
- ▶ While engineering the stack, approach in stages if migrating from 8i/9i to RAC 10g.
 - As part of taking the databases from 8i to RAC 10g, there can be performance anomalies that may come about due to various reasons (CBO differences across versions, Histograms due to the methodology of statistics collection, Interconnect bottlenecks etc.). In order to isolate the root cause of the performance gap, it

is important that the applications be migrated (during engineering) in a step-wise fashion.

- ▶ Step 1 – Baseline with 8i (Establish test bench/targets)
- ▶ Step 2 – Move to 10g Stand-alone (Tune to CBO, Re-baseline)
- ▶ Step 3 – Move from 10g Stand-alone to RAC 10g (Tune to RAC)
- This approach aids in isolating issues pertinent to a specific change.
- ▶ Use Oracle RAC out of the box components for smoother integration and supportability
 - RDBMS, Clusterware, ASM, RMAN for e2e backups, EM/Grid Control for Metrics and Monitoring - components of the stack are all from one vendor - Oracle
 - Goal is to keep the stack simple by minimizing the vendor integration touch points as much as possible.
- ▶ Plan for Cross-Vendor Integration surprises
 - End to end stack still includes multiple vendors (Oracle (RDBMS/ASM), Microsoft (OS), Media Management, Tape Library)
 - Making the vendors part of the “stakeholders” allowed quicker resolution of issues uncovered.
 - Spawn parallel channels with all vendors for efficient troubleshooting of the issues in the stack.

CONCLUSION

The solution that Intel Technology Manufacturing Group designed to address the scalability of decision support systems was a single clustered database running on a RAC cluster. The cluster was built on a 10 Node, Intel EM64T servers running 64-bit windows 2003 advanced server and 64-bit Oracle database 10g Release 2. Moving to 64-bit servers enabled each individual node to be *scaled up* from the current configuration providing more processing power and a larger memory footprint essential for resource intensive DSS queries. *Scaling out* with RAC enabled application workload partitioning as well as facilitated resource intensive queries to take advantage of resources across multiple nodes. Apart from solving the near term requirements, this architecture enables us to add incremental capacity on demand and allows us to take advantage of powerful processors in future by seamlessly swapping with cost efficient powerful Intel processors.

This gigantic “machine” is equipped with 80 CPUs and 320Gig of RAM. The machine provided Infrastructural capacity to push more workload.

Intel TMG architects created a gigantic VLDB Database machine with cost effective mainstream Intel EM64T servers on Oracle RAC that together provided Infrastructural capability/capacity to push more workload. The stack enabled:

- Faster Processing power via Intel Mainstream Servers
- Efficient I/O via ASM
- Larger global virtual shared cache with larger SGAs
- Efficient workload management & Multi-instance parallelism
- Incremental scalability



Scaling very Large Databases with Oracle RAC using Intel EM64T mainstream servers

November 2007

Author: Srinagesh Battula (Intel Corporation), Gayathri Seetharaman (Intel Corporation), Deepen Chankraborty (Intel Corporation)

Contributing Authors: Ram Varra (Intel Corporation)

Review : Philip Newlan (Oracle)

Version 1.1

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

Copyright © 2007, Oracle. 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 is a registered trademark of Oracle Corporation and/or its affiliates.

Other names may be trademarks of their respective owners.