IDS Availability:
A Competitive Analysis

IDS High Availability v/s Competition

Anup Nair
IBM

Session Code: C17
Date and Time: Wednesday April 30th 3:30-4:30 PM

2008 IIUG Informix Conference

www.iiug.org

Overview

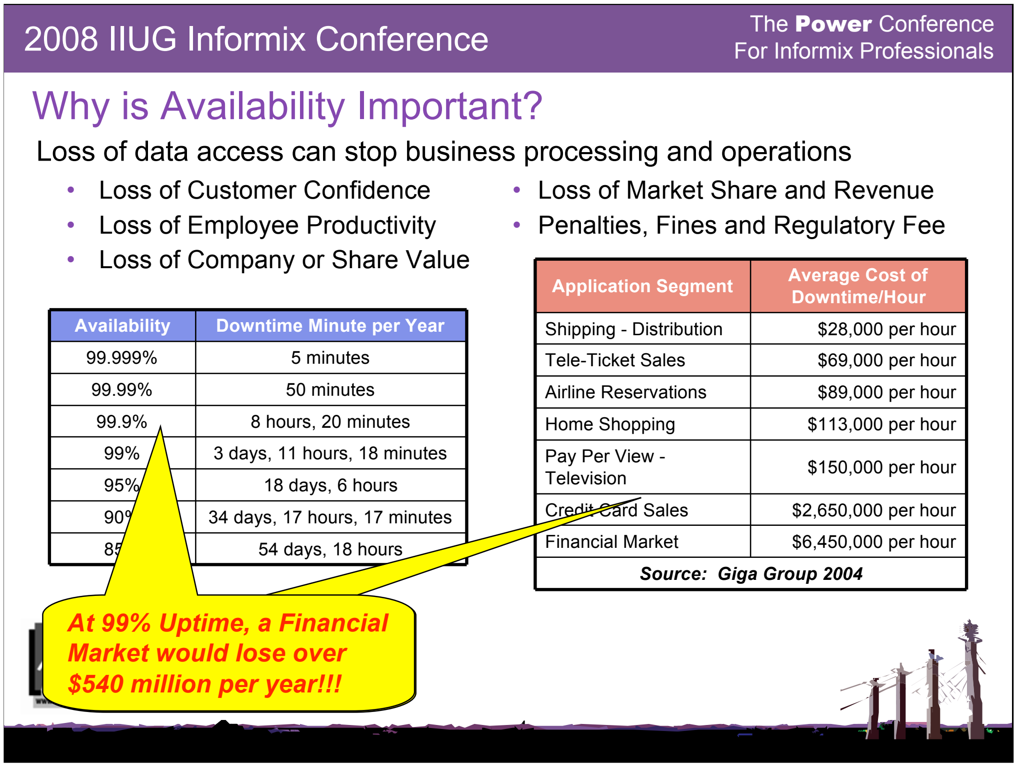
- **What is Important to you?**

  A comprehensive competitive analysis of why IDS is better than Oracle, with emphasis on IDS High-Availability Data Replication (HDR) vs Oracle Real Application Cluster (RAC) and Data Guard (DG)

What is Important to You?

- Data Availability
- Scalability

A comprehensive competitive analysis of why IDS is better than Oracle, with primary emphasis on IDS High-Availability Data Replication (HDR) vs Oracle Real Application Cluster (RAC) and Data Guard (DG)



Why is Availability Important?

Loss of data access can stop business processing and operations

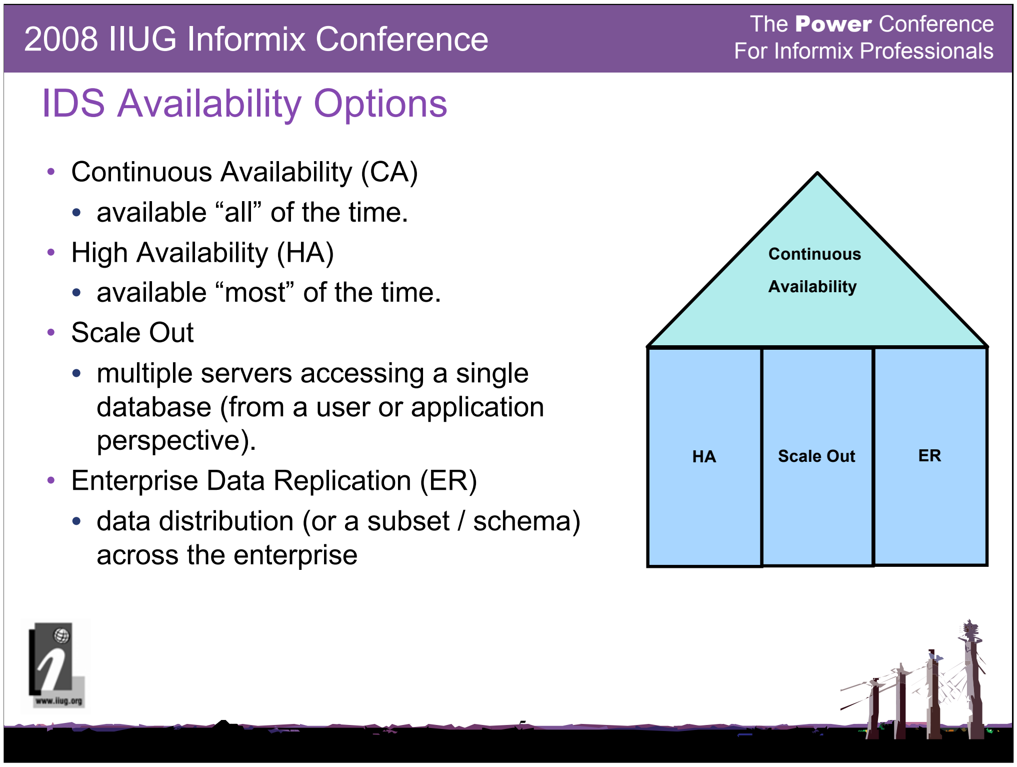
- Loss of Customer Confidence
- Loss of Employee Productivity
- Loss of Company or Share Value
- Loss of Market Share and Revenue
- Penalties, Fines and Regulatory Fee

Availability	Downtime Minute per Year
99.999%	5 minutes
99.99%	50 minutes
99.9%	8 hours, 20 minutes
99%	3 days, 11 hours, 18 minutes
95%	18 days, 6 hours
90	34 days, 17 hours, 17 minutes
85	54 days, 18 hours

Application Segment	Average Cost of Downtime/Hour
Shipping - Distribution	$28,000 per hour
Tele-Ticket Sales	$69,000 per hour
Airline Reservations	$89,000 per hour
Home Shopping	$113,000 per hour
Pay Per View - Television	$150,000 per hour
Credit Card Sales	$2,650,000 per hour
Financial Market	$6,450,000 per hour

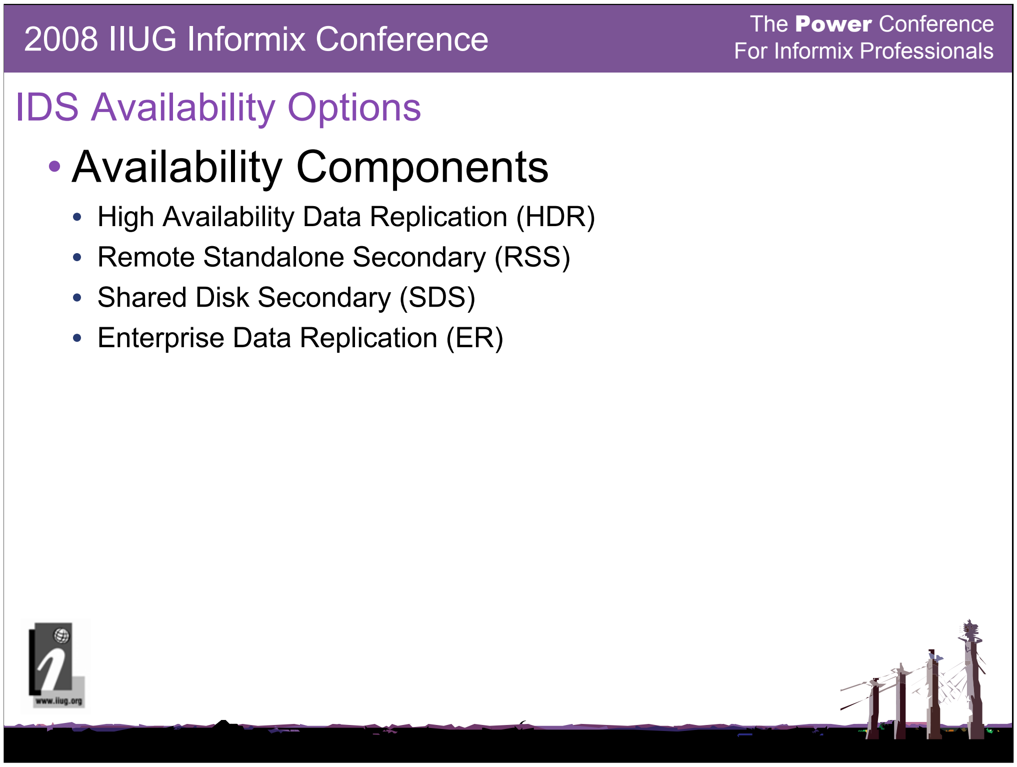
Source: Giga Group 2004

At 99% Uptime, a Financial Market would lose over $540 million per year!!!



IDS Availability Options

- Continuous Availability (CA)
  - available “all” of the time.
- High Availability (HA)
  - available “most” of the time.
- Scale Out
  - multiple servers accessing a single database (from a user or application perspective).
- Enterprise Data Replication (ER)
  - data distribution (or a subset / schema) across the enterprise



2008 IIUG Informix Conference

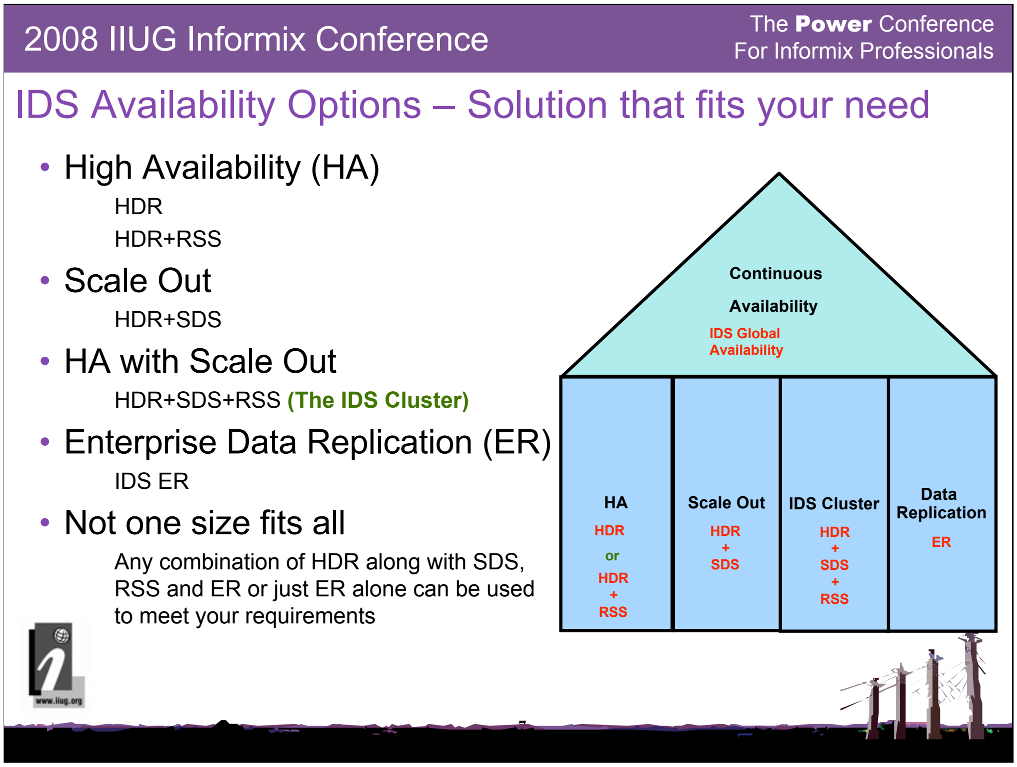
The **Power** Conference
For Informix Professionals

IDS Availability Options

- Availability Components
  - High Availability Data Replication (HDR)
  - Remote Standalone Secondary (RSS)
  - Shared Disk Secondary (SDS)
  - Enterprise Data Replication (ER)

www.iiug.org

Will be discussing about HDR, SDS and RSS in the following slides



2008 IIUG Informix Conference

The **Power** Conference
For Informix Professionals

IDS Availability Options – Solution that fits your need

- High Availability (HA)
  - HDR
  - HDR+RSS
- Scale Out
  - HDR+SDS
- HA with Scale Out
  - HDR+SDS+RSS **(The IDS Cluster)**
- Enterprise Data Replication (ER)
  - IDS ER
- Not one size fits all
  - Any combination of HDR along with SDS, RSS and ER or just ER alone can be used to meet your requirements

Continuous

Availability

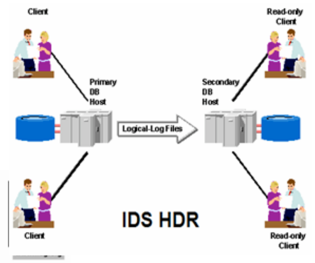
IDS Global Availability

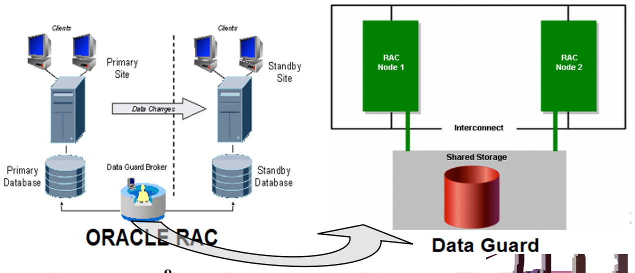
HA	**Scale Out**	**IDS Cluster**	**Data Replication**
HDR **or** **HDR** **+** **RSS**	**HDR** **+** **SDS**	**HDR** **+** **SDS** **+** **RSS**	**ER**

The animation was creating problems hence divided into different slides

Availability Options … Comparative study

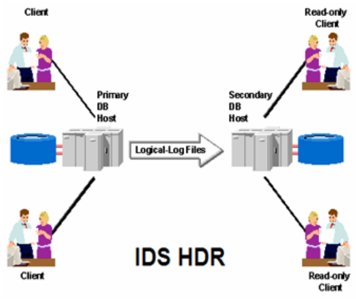
- IDS High Availability Data Replication (HDR)
- Comparison to Oracle Data Guard (DG)





IDS Availability Options … Comparative study

IDS High Availability Data Replication

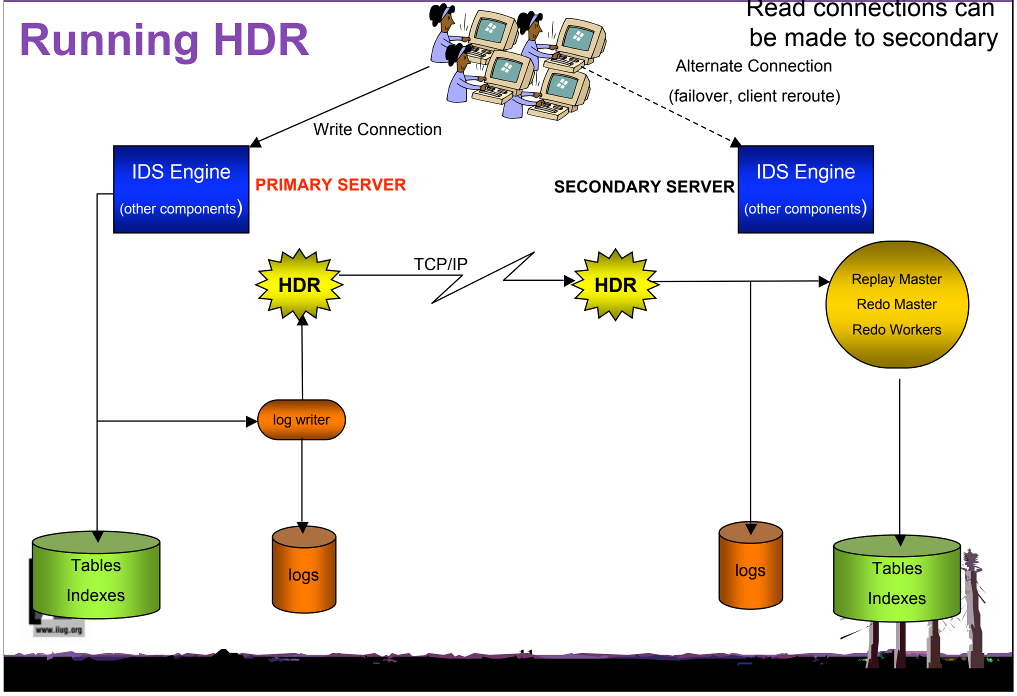


Main Goals of the Design for High Availability (HDR)

- Ultra-fast failover
- Easy administration
- Negligible impact on performance
- Transparent failover and fail back for applications (combined with “client re-route”)

Running HDR

Read connections can be made to secondary



Alternate Connection

(failover, client reroute)

Write Connection

IDS Engine

(other components)

PRIMARY SERVER

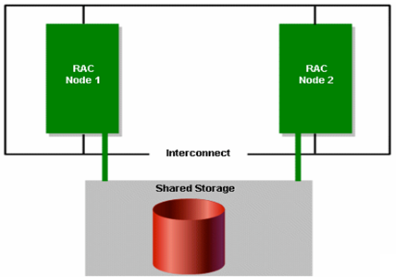
SECONDARY SERVER

IDS Engine

(other components)

IDS Availability Options … Comparative study

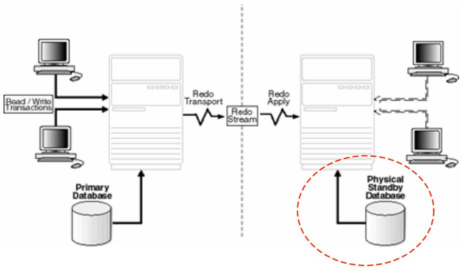
Comparison to Oracle Data Guard



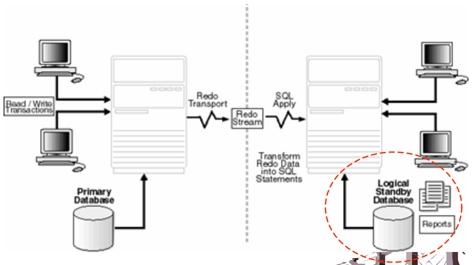
Data Guard

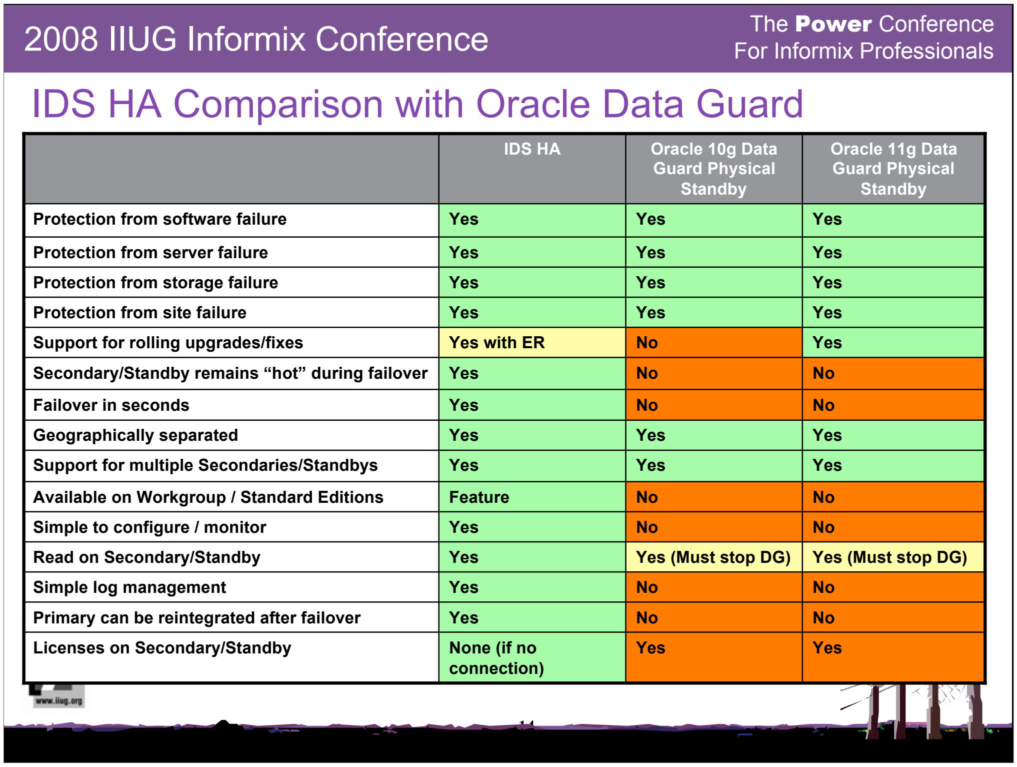
Oracle Data Guard "Flavors"

- Physical Standby Database
  - Two servers are exact copies of each other
  - Log buffers or log files shipped to standby
  - Applied on Standby as redo records
  - "Similar" to HDR



- Logical Standby Database
  - Two servers are logical copies of each other but can differ in many aspects (tables, indexes, disk structure, etc)
  - Log buffers or log files are shipped to standby
  - Applied on standby as SQL statements





2008 IIUG Informix Conference

The **Power** Conference
For Informix Professionals

IDS HA Comparison with Oracle Data Guard

	IDS HA	Oracle 10g Data Guard Physical Standby	Oracle 11g Data Guard Physical Standby
Protection from software failure	Yes	Yes	Yes
Protection from server failure	Yes	Yes	Yes
Protection from storage failure	Yes	Yes	Yes
Protection from site failure	Yes	Yes	Yes
Support for rolling upgrades/fixes	Yes with ER	No	Yes
Secondary/Standby remains “hot” during failover	Yes	No	No
Failover in seconds	Yes	No	No
Geographically separated	Yes	Yes	Yes
Support for multiple Secondaries/Standbys	Yes	Yes	Yes
Available on Workgroup / Standard Editions	Feature	No	No
Simple to configure / monitor	Yes	No	No
Read on Secondary/Standby	Yes	Yes (Must stop DG)	Yes (Must stop DG)
Simple log management	Yes	No	No
Primary can be reintegrated after failover	Yes	No	No
Licenses on Secondary/Standby	None (if no connection)	Yes	Yes

(1) Oracle Real Application Clusters (Oracle RAC), is an option to Oracle Database 11*g* Enterprise Edition and included with Oracle Database 11*g* Standard Edition (on clusters with a maximum of 4 sockets).

In Context of Data Guard

- Are reads on the Oracle Secondary available only the Logical secondary, not the Physical?

Reads are possible on both. Difference is only in its config and the way each one is synced. Phsyical uses redo log, logical uses sql generated out of read log.

- In Oracle 11, Are reads possible on all secondary servers?

Yes.

- Does Oracle Data Guard supports update anywhere and/or multi-master?

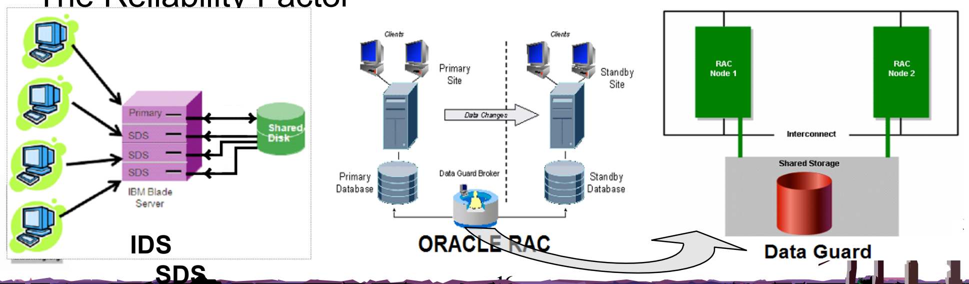
DG per se does not support multi master as it is a simple facility to provide protection and availability. If you need update anywhere streams/replication is the way to go.

What is Important to You?

- Data Availability
- Scalability

IDS Scalability

- IDS Shared Disk Secondaries (SDS) for Scale Out
- Comparison & Issues with Oracle Real Application Cluster (RAC)
- HP/Dell RAC Tests
- The IDS Global Availability: A Complete Picture With Scalability
- The Reliability Factor

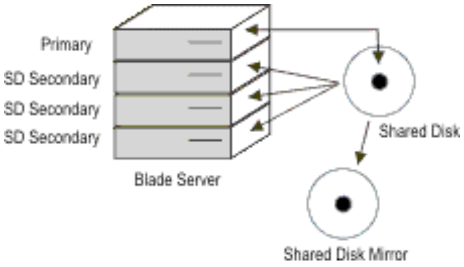


IDS SDS

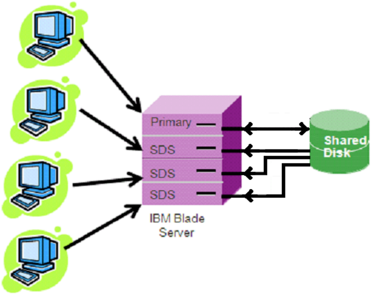
ORACLE RAC

Data Guard

IDS Shared Disk Secondary



IDS Shared Disk Secondaries for Scale Out

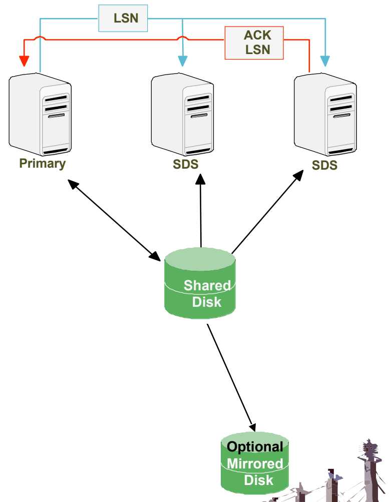


Easily grow capacity as needed

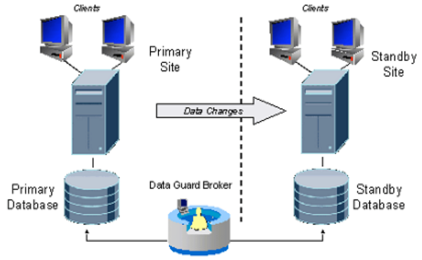
- Shared Disk Secondaries (SDS) provide scale-out using stand alone servers or IBM Blade Center
- A single copy of the database (no replication required)
- Minimal setup time – Only a checkpoint is required to start the SDS
- Primary can failover to any SDS node if needed
- Network exchanges use Logical Sequence Numbers (LSNs), not log pages so no interconnect latency
- Mirroring of the database is optional

Shared Disk Secondary (SDS) in to IDS

- HDR on top of a shared disk subsystem
- Works nicely with a blade center or stand alone servers
- Minimal startup time – only a checkpoint is required to start the SDS
- Primary role can shift to any of the SDS nodes
- Provides additional read capacity without requiring additional disk
- Works by coordination of page flushing to disk
- Network exchanges log position (LSN), not log pages

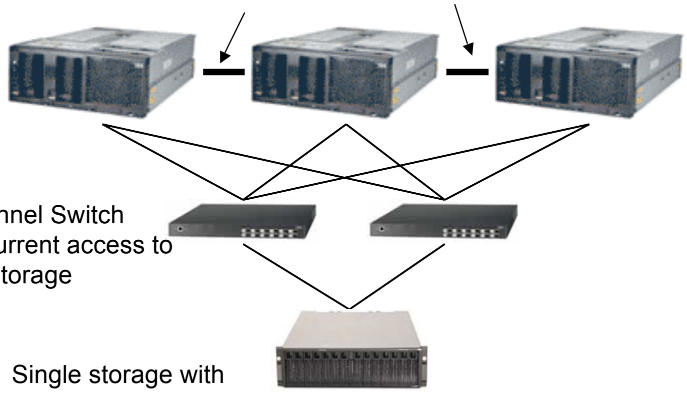


Comparison & Issues with Oracle RAC



ORACLE RAC

Oracle RAC Shared Disk Architecture



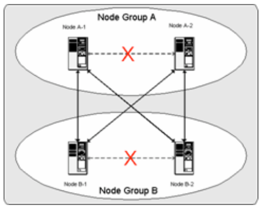
High Speed Interconnect

Separate Servers
- No shared components
- Each running an Oracle Instance

Fiber Channel Switch
- for concurrent access to shared storage

Single storage with
single copy of the database

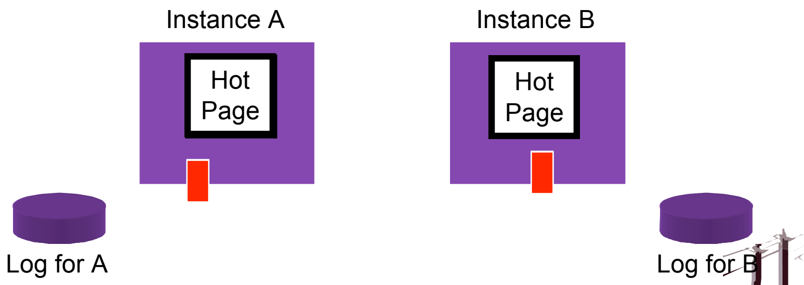
Steps involved in a node failure



- Node failure detection
- Data block re-mastering
- Locking of pages that need recovery
- Redo and undo recovery

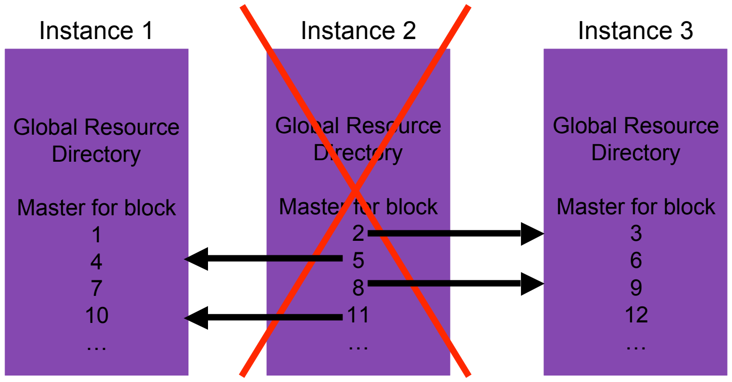
Hot Pages

- Any page that is “hot” becomes a bottleneck between nodes
  - E.g. Index on increasing numbers (unique ids), data on measurements, etc.
  - Data block is pinged back and forth between the caches, each time a log write is required.



Data Block Remastering

- When a node fails, the blocks it mastered must be redistributed to other nodes.
- While this is happening, GCS on all nodes is **frozen**



Instance 1

Global Resource Directory

Master for block
1
4
7
10
…

Instance 2

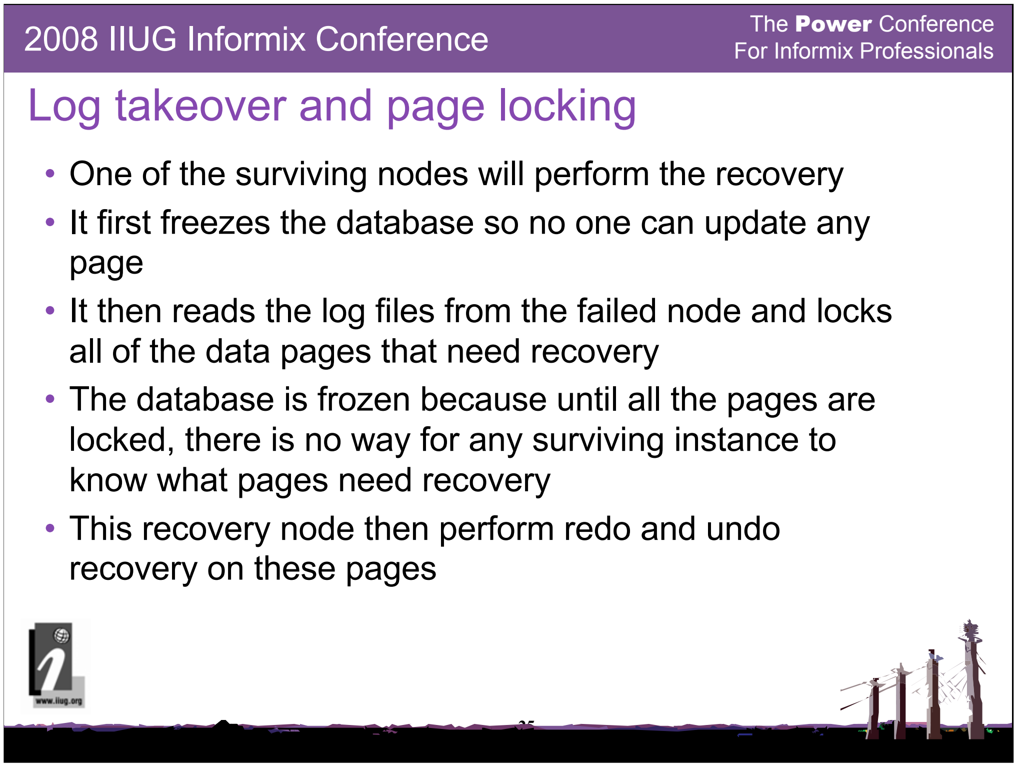
Global Resource Directory

Master for block
2
5
8
11
…

Instance 3

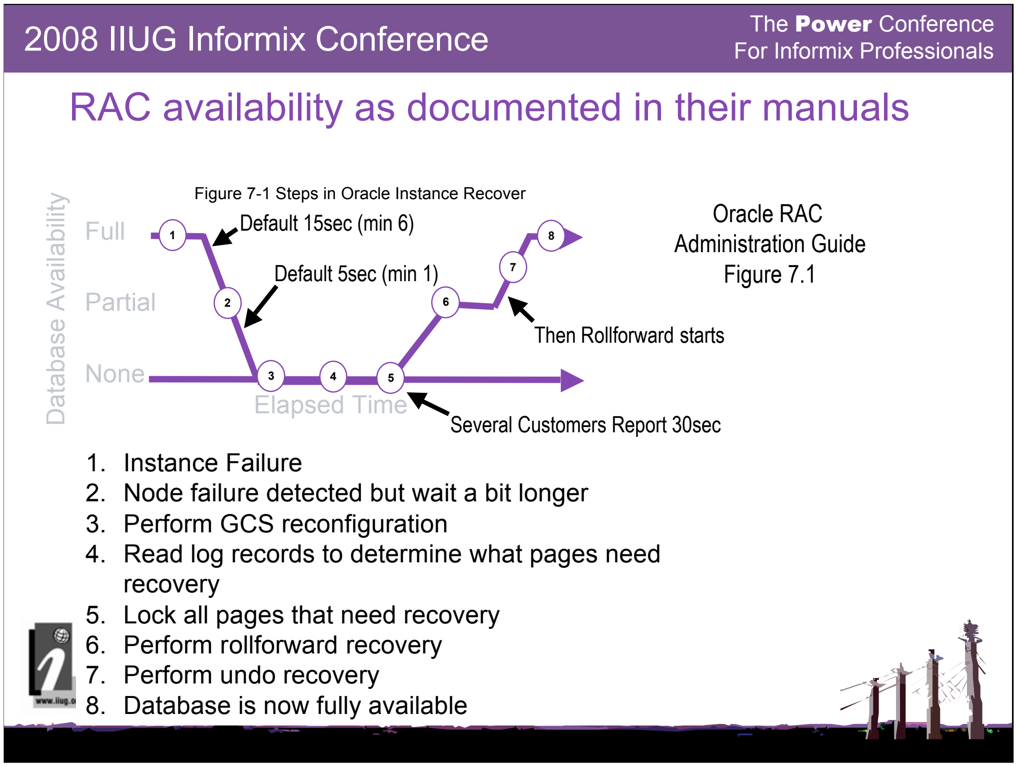
Global Resource Directory

Master for block
3
6
9
12
…



Log takeover and page locking

- One of the surviving nodes will perform the recovery
- It first freezes the database so no one can update any page
- It then reads the log files from the failed node and locks all of the data pages that need recovery
- The database is frozen because until all the pages are locked, there is no way for any surviving instance to know what pages need recovery
- This recovery node then perform redo and undo recovery on these pages



2008 IIUG Informix Conference

The **Power** Conference
For Informix Professionals

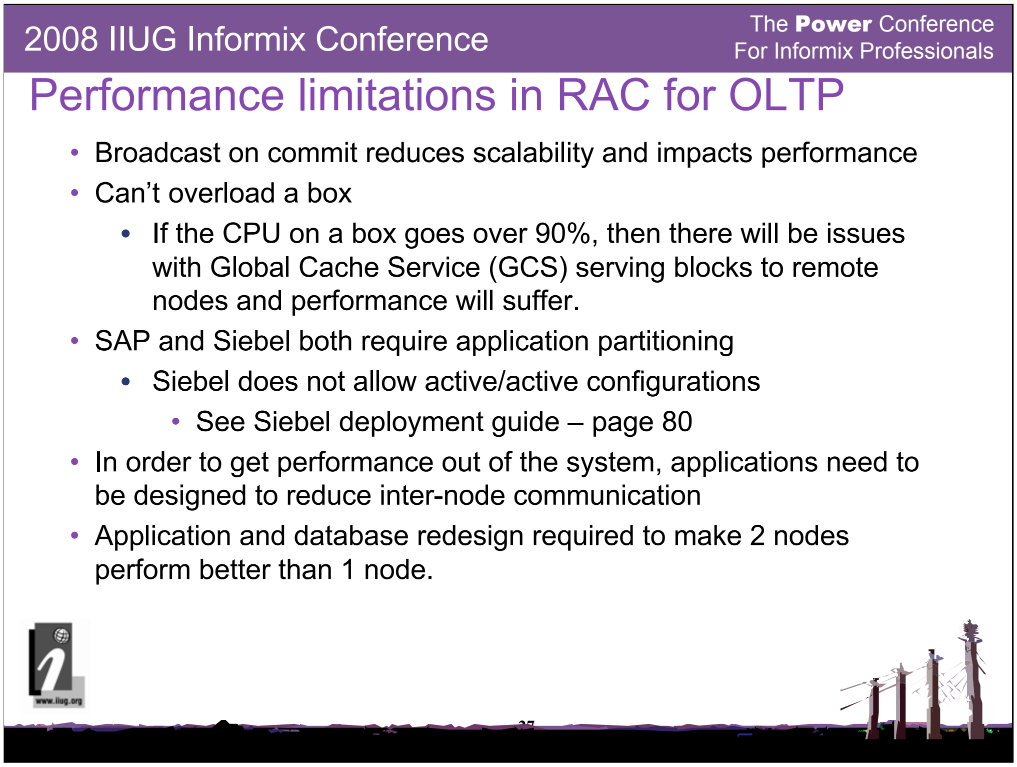
RAC availability as documented in their manuals

Figure 7-1 Steps in Oracle Instance Recover

Oracle RAC
Administration Guide
Figure 7.1

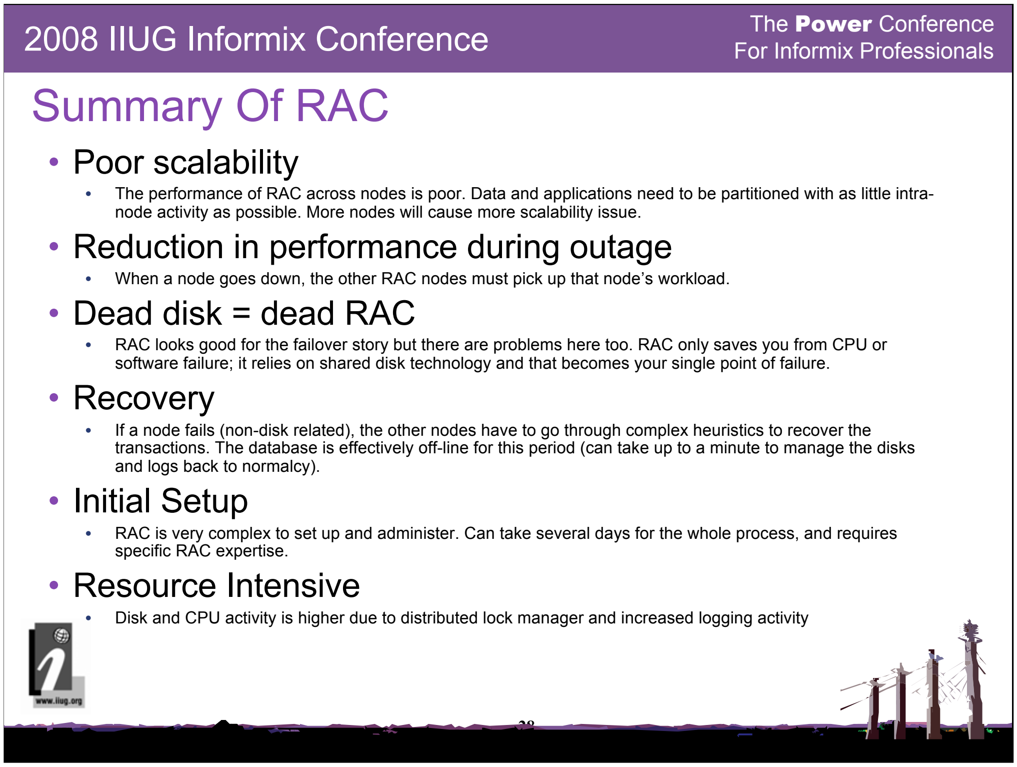
1. Instance Failure
2. Node failure detected but wait a bit longer
3. Perform GCS reconfiguration
4. Read log records to determine what pages need recovery
5. Lock all pages that need recovery
6. Perform rollforward recovery
7. Perform undo recovery
8. Database is now fully available

This comes directly from Oracle's own manuals. Note that in RAC the database has "NO AVAILABILITY" when one node in the cluster fails between points 3 and 5. Also good to note that from the time the failure occurs, Oracle RAC will not detect the failure and begin the remastering for 20 seconds (by default), this can be configured down to a minimum of 7 seconds. During this period, queries that require data in the cache (or managed by) the failed node will hang. Then all queries hang during the data remastering process, then data starts to become available while the data on the crashed node goes through crash recovery processing (point 6-8).



Performance limitations in RAC for OLTP

- Broadcast on commit reduces scalability and impacts performance
- Can’t overload a box
  - If the CPU on a box goes over 90%, then there will be issues with Global Cache Service (GCS) serving blocks to remote nodes and performance will suffer.
- SAP and Siebel both require application partitioning
  - Siebel does not allow active/active configurations
    - See Siebel deployment guide – page 80
- In order to get performance out of the system, applications need to be designed to reduce inter-node communication
- Application and database redesign required to make 2 nodes perform better than 1 node.



Summary Of RAC

- Poor scalability
  - The performance of RAC across nodes is poor. Data and applications need to be partitioned with as little intra-node activity as possible. More nodes will cause more scalability issue.
- Reduction in performance during outage
  - When a node goes down, the other RAC nodes must pick up that node's workload.
- Dead disk = dead RAC
  - RAC looks good for the failover story but there are problems here too. RAC only saves you from CPU or software failure; it relies on shared disk technology and that becomes your single point of failure.
- Recovery
  - If a node fails (non-disk related), the other nodes have to go through complex heuristics to recover the transactions. The database is effectively off-line for this period (can take up to a minute to manage the disks and logs back to normalcy).
- Initial Setup
  - RAC is very complex to set up and administer. Can take several days for the whole process, and requires specific RAC expertise.
- Resource Intensive
  - Disk and CPU activity is higher due to distributed lock manager and increased logging activity

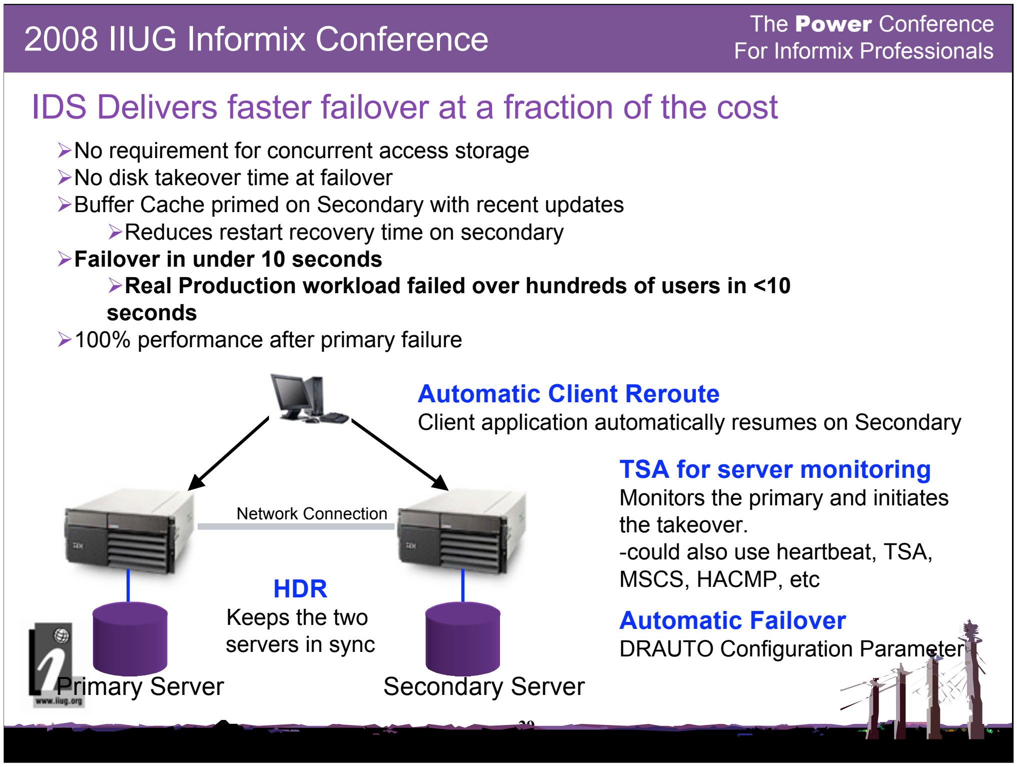
o The performance of RAC across nodes is poor so that you need to partition the data and applications with as little intra-node activity as possible. It doesn't scale - I believe there are very few 4+ node systems.

o RAC only looks good for the failover story but there are problems here too. RAC only really saves you from CPU or software failure, it relies on shared disk technology and that becomes your single point of failure - dead disk = dead RAC.

o If a node does fail (not disk related), the other nodes then have to go through a dance to recover the transactions and the database is effectively off-line for this period - can be a minute to sort out the disks and logs.

o RAC is very complex to set up - 4 day install time - and administer.

o It is very resource intensive - the log spaces have to sustain very high I/O rates. O‹#›ne customer was told to use solid state RAM disk technology (very expensive) to keep up with the transaction rate.



IDS Delivers faster failover at a fraction of the cost

- No requirement for concurrent access storage
- No disk takeover time at failover
- Buffer Cache primed on Secondary with recent updates
  - Reduces restart recovery time on secondary
- **Failover in under 10 seconds**
  - **Real Production workload failed over hundreds of users in <10 seconds**
- 100% performance after primary failure

Automatic Client Reroute
Client application automatically resumes on Secondary

TSA for server monitoring
Monitors the primary and initiates the takeover.
-could also use heartbeat, TSA, MSCS, HACMP, etc

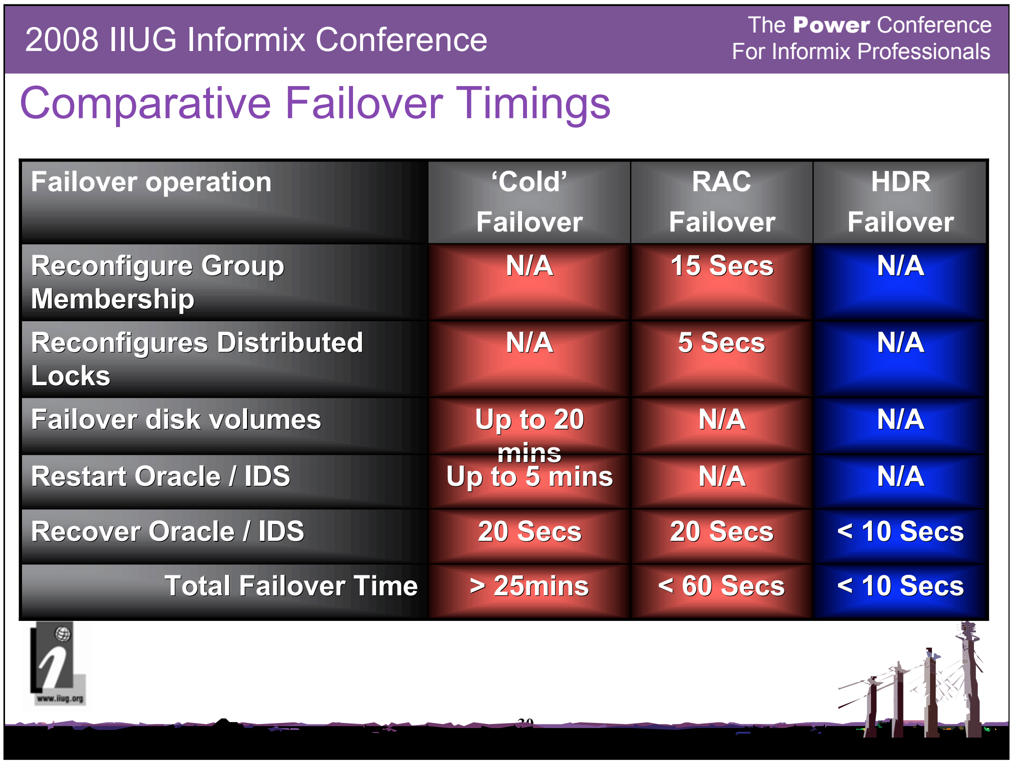
Automatic Failover
DRAUTO Configuration Parameter

Network Connection

HDR
Keeps the two servers in sync

Primary Server

Secondary Server

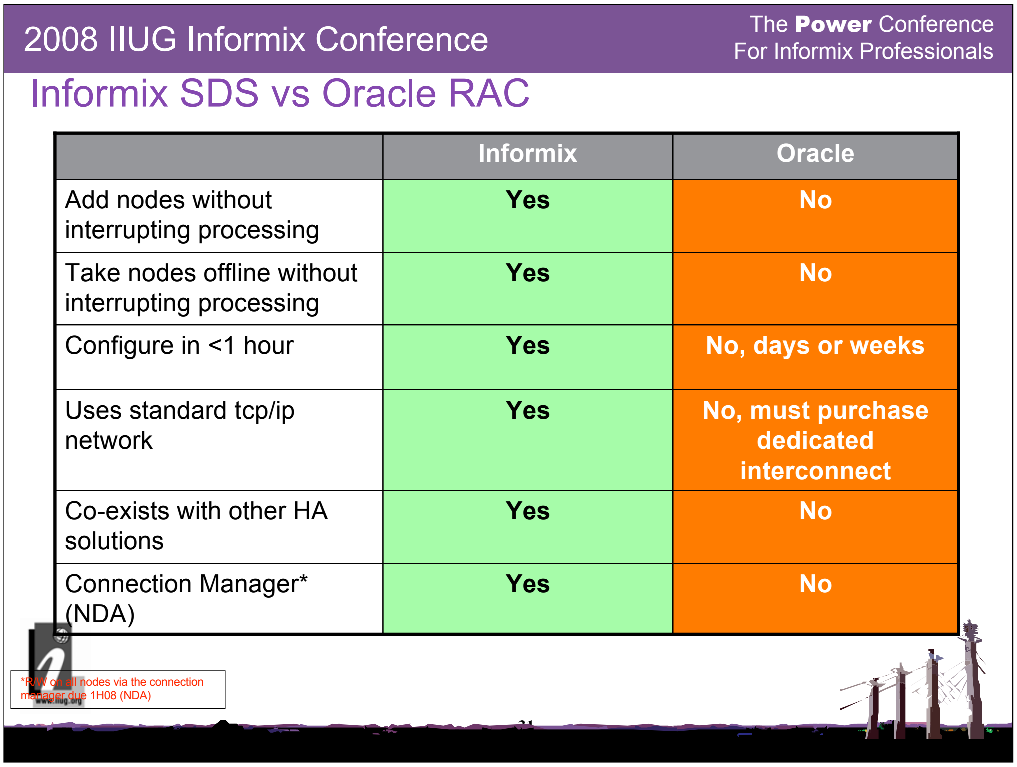


Comparative Failover Timings

Failover operation	'Cold' Failover	RAC Failover	HDR Failover
Reconfigure Group Membership	N/A	15 Secs	N/A
Reconfigures Distributed Locks	N/A	5 Secs	N/A
Failover disk volumes	Up to 20 mins	N/A	N/A
Restart Oracle / IDS	Up to 5 mins	N/A	N/A
Recover Oracle / IDS	20 Secs	20 Secs	< 10 Secs
Total Failover Time	> 25mins	< 60 Secs	< 10 Secs

Presentation given by Marshall Presser Principal Technologist Oracle Corporation to the Beowolf users group 11 May, 2004

Source http://www.bwbug.org/docs/1



Informix SDS vs Oracle RAC

	Informix	Oracle
Add nodes without interrupting processing	**Yes**	**No**
Take nodes offline without interrupting processing	**Yes**	**No**
Configure in <1 hour	**Yes**	**No, days or weeks**
Uses standard tcp/ip network	**Yes**	**No, must purchase dedicated interconnect**
Co-exists with other HA solutions	**Yes**	**No**
Connection Manager* (NDA)	**Yes**	**No**

*R/W on all nodes via the connection manager due 1H08 (NDA)

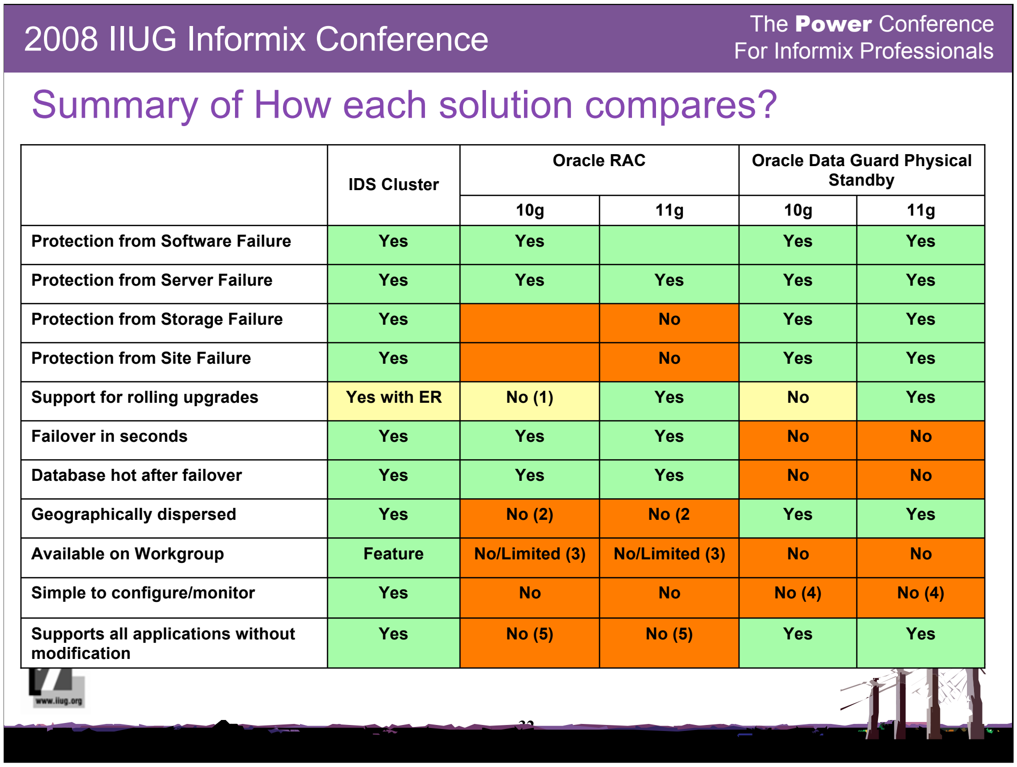
o The performance of RAC across nodes is poor so that you need to partition the data and applications with as little intra-node activity as possible. It doesn't scale - I believe there are very few 4+ node systems.

o RAC only looks good for the failover story but there are problems here too. RAC only really saves you from CPU or software failure, it relies on shared disk technology and that becomes your single point of failure - dead disk = dead RAC.

o If a node does fail (not disk related), the other nodes then have to go through a dance to recover the transactions and the database is effectively off-line for this period - can be a minute to sort out the disks and logs.

o RAC is very complex to set up - 4 day install time - and administer.

o It is very resource intensive - the log spaces have to sustain very high I/O rates. O‹#›ne customer was told to use solid state RAM disk technology (very expensive) to keep up with the transaction rate.



2008 IIUG Informix Conference

The **Power** Conference
For Informix Professionals

Summary of How each solution compares?

	IDS Cluster	Oracle RAC		Oracle Data Guard Physical Standby	
		10g	11g	10g	11g
Protection from Software Failure	Yes	Yes		Yes	Yes
Protection from Server Failure	Yes	Yes	Yes	Yes	Yes
Protection from Storage Failure	Yes		No	Yes	Yes
Protection from Site Failure	Yes		No	Yes	Yes
Support for rolling upgrades	Yes with ER	No (1)	Yes	No	Yes
Failover in seconds	Yes	Yes	Yes	No	No
Database hot after failover	Yes	Yes	Yes	No	No
Geographically dispersed	Yes	No (2)	No (2	Yes	Yes
Available on Workgroup	Feature	No/Limited (3)	No/Limited (3)	No	No
Simple to configure/monitor	Yes	No	No	No (4)	No (4)
Supports all applications without modification	Yes	No (5)	No (5)	Yes	Yes

1 - Oracle does allow “limited” fixes to be applied in a rolling manner. This does not apply to patchsets (aka fixpacks) but rather just to individual fixes that don’t affect the database engine.

2 – A RAC cluster must have all nodes connected to the same shared storage. It is possible to have nodes separated by up to 10km (6 miles) but this would seriously impact performance AND would require very costly storage.

3 – Oracle RAC is only available on up to 4 cpus with Standard edition (so 2 2ways or 4 1ways). You cannot use RAC with Standard Edition one. Note also that standard edition (even with RAC) has many other limitations (no Data Guard, no parallelism, no gui tools, etc).

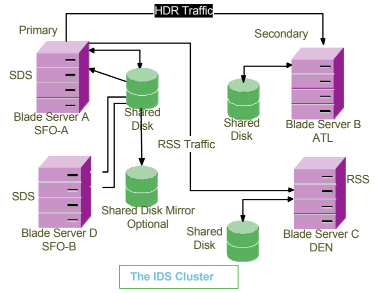
4 – in order to configure data guard using a gui tool, you must first install Oracle Grid Control. This requires that you also install Oracle application server, a separate database for the management server and on some systems (like AIX) you must download Java Cryptographic Services from the sun website. After all that you may be able to use grid control to setup and manage your Data Guard environment. But your management repository now also needs to be configured for HA. Note that with IDS all you need is any IDS administration client to use the GUI tools to setup and manage HDR.

5 – Oracle claims that you don’t have to change your application. However there is significant evidence that this is not true if you want to scale the application across multiple nodes (put in references).

6 – You're stretching the truth if you say the HDR supports rolling upgrades. ER does support rolling upgrades, however. Also - with Oracle 11, the physical secondary can be temporary switched to a logical dataguare secondary - and thus supports rolling upgrades.

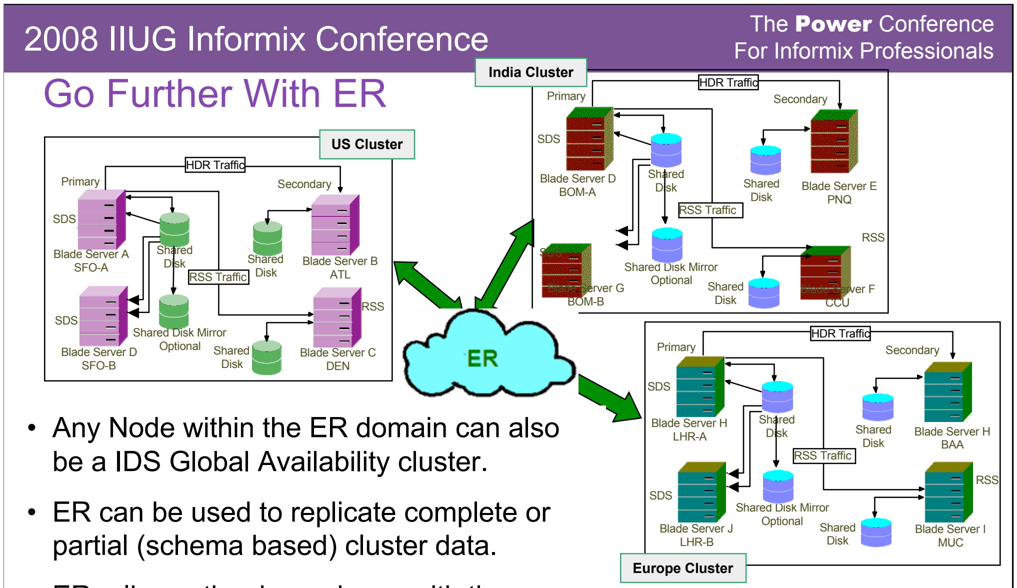
The IDS Global Availability

A Complete Picture With Scalability



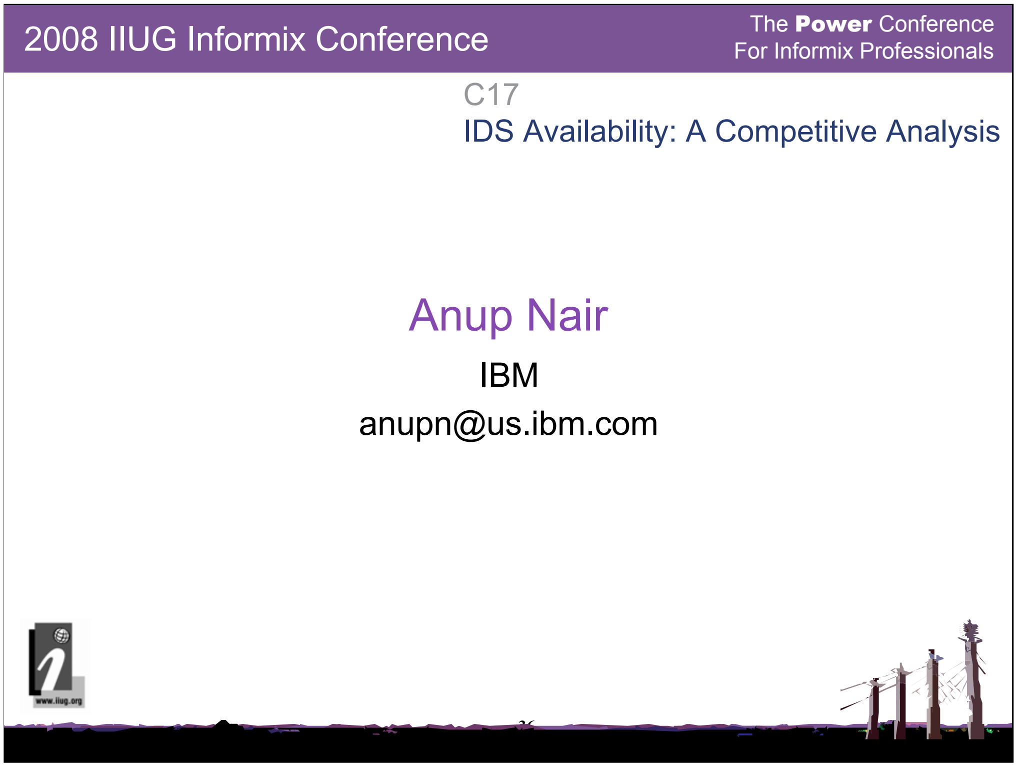
www.iiug.org

Go Further With ER



- Any Node within the ER domain can also be a IDS Global Availability cluster.
- ER can be used to replicate complete or partial (schema based) cluster data.
- ER relieves the dependency with the Primary in situation like network outages.

www.iiug.org



2008 IIUG Informix Conference

The **Power** Conference
For Informix Professionals

C17

IDS Availability: A Competitive Analysis

Anup Nair

IBM

anupn@us.ibm.com