RAC on Extended Distance Clusters

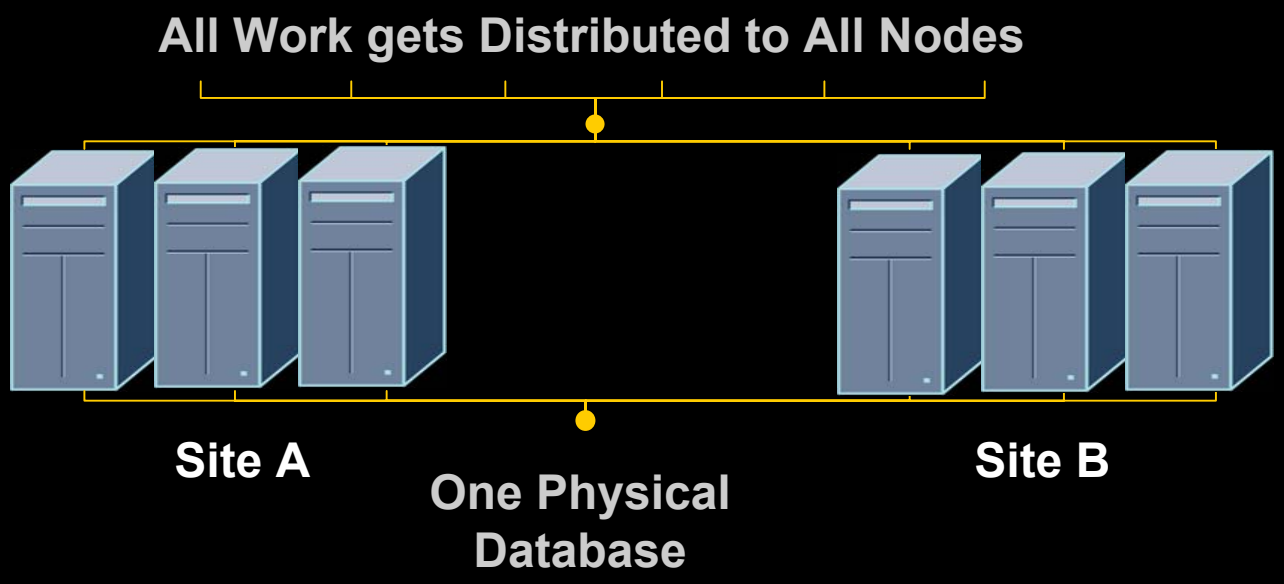
Erik Peterson
RAC Development
Oracle Corporation

Agenda

- Benefits of RAC on extended clusters
- Design considerations
- Empirical performance data
- Live customer examples
- Positioning w.r.t. DataGuard
- Summary

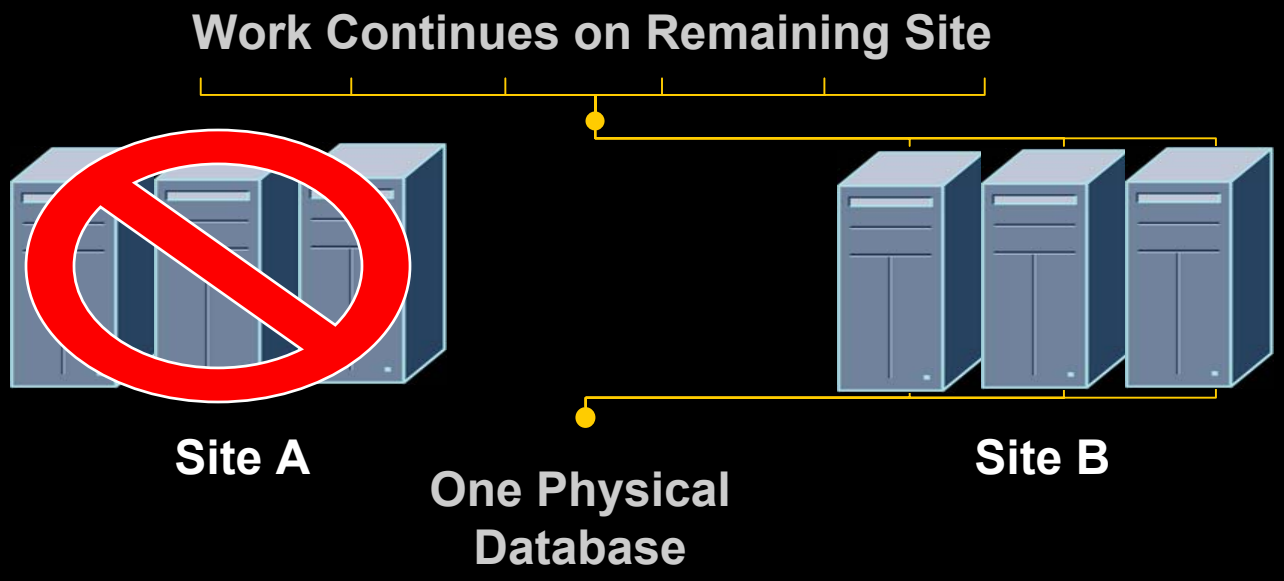
Benefits of RAC on Extended Clusters

- Full utilization of resources no matter where they are located



Benefits of RAC on Extended Clusters

- Faster recovery from site failure than any other technology in the market



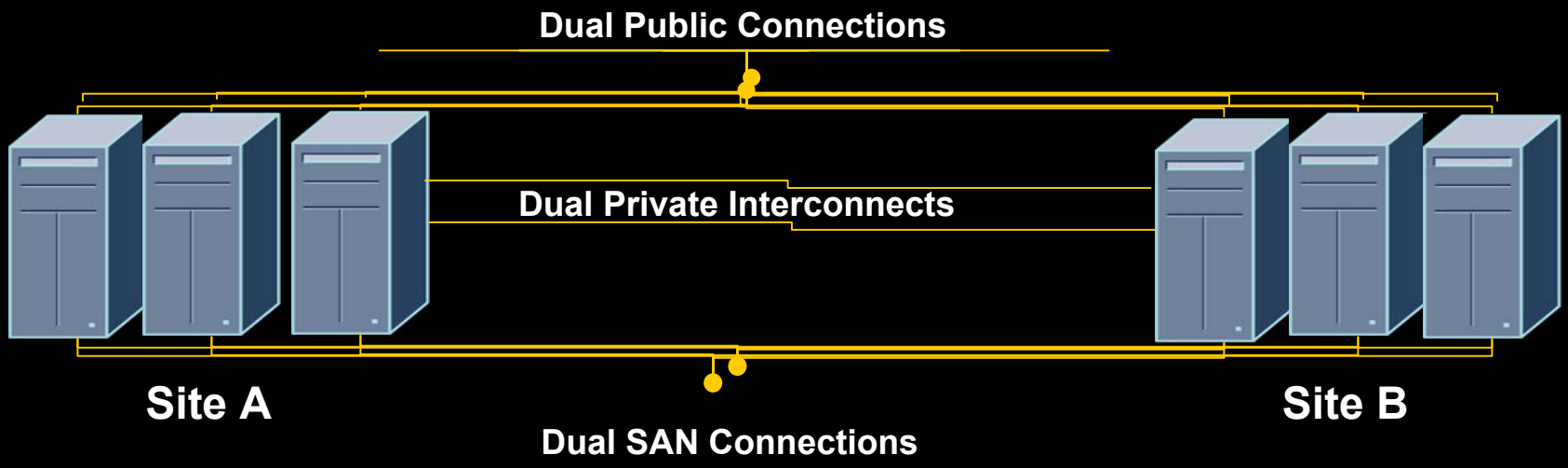
Design Considerations

Design Considerations

- Connectivity
- Disk Mirroring
- Quorum

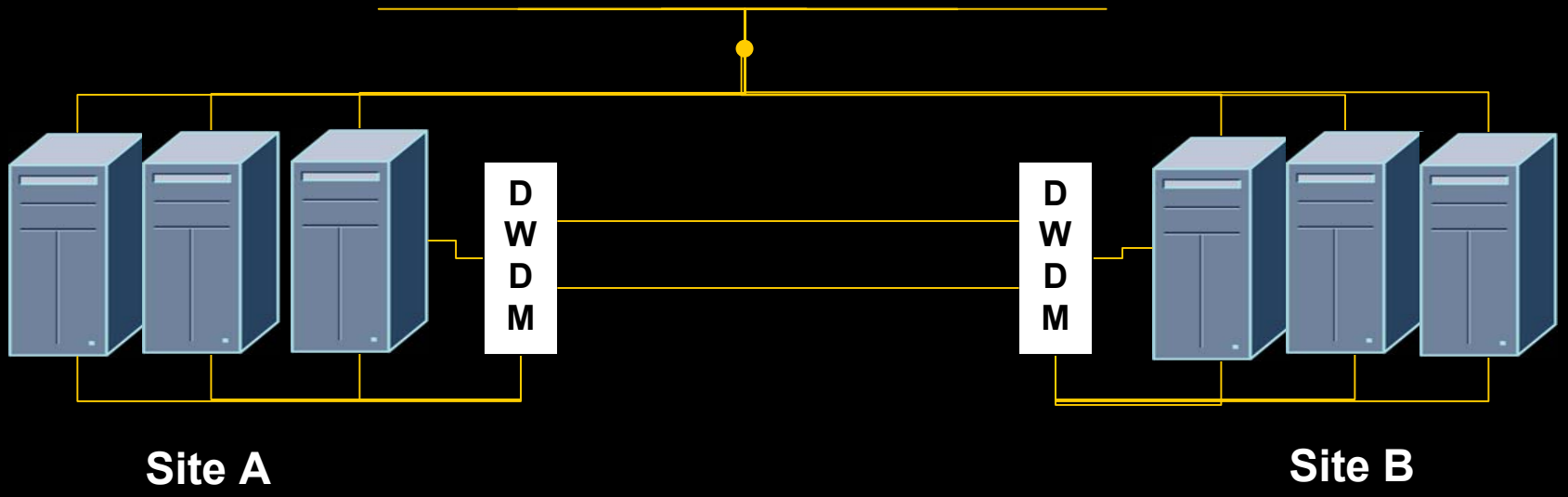
Connectivity

- Redundant connections for public traffic, interconnect and I/O



Connectivity

- Distances > 10km require Dark Fiber (DWDM or CWM).
- Extra benefit of separate dedicated channels on 1 fibre
- Essential to setup buffer credits for large distances



Connectivity Caveats

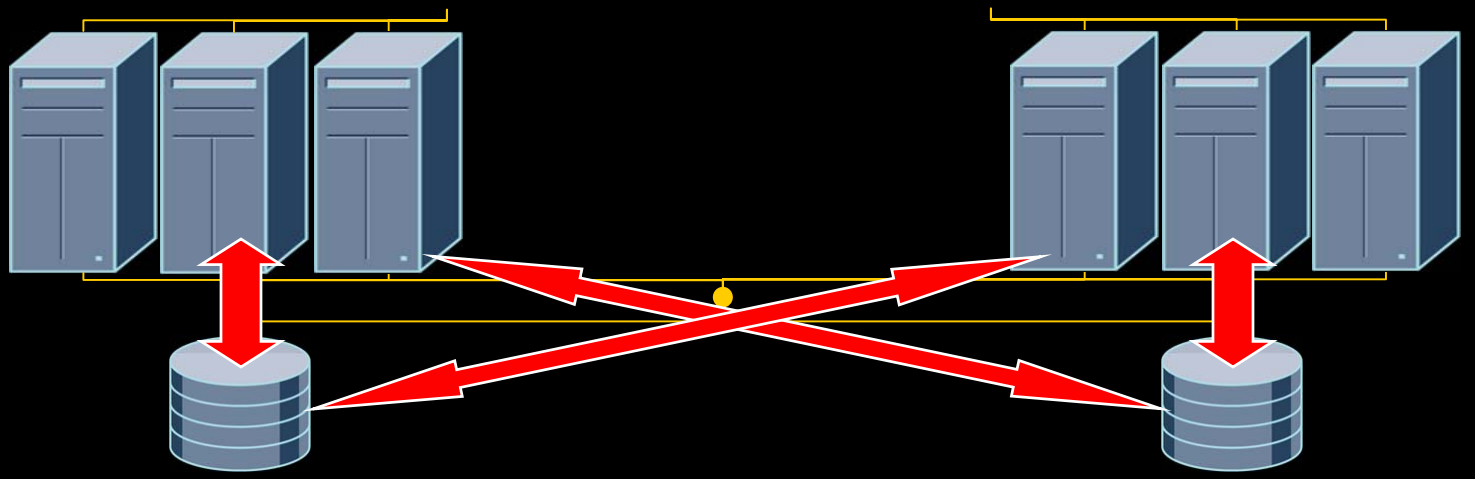
- Distance
  - Single fiber limit (100km?)
- Performance
  - Need to Minimize Latency.
    - Direct effect on synchronous disk mirroring and Cache Fusion operation
    - Direct point to point connection => Additional routers, hubs, or extra switches add latency
- Cost
  - High cost of DWDM if not already present in the infrastructure

Disk Mirroring

- Need copy of data at each location
- 2 options exist:
  - Host Based Mirroring (CLVM)
  - Remote Array Based Mirroring

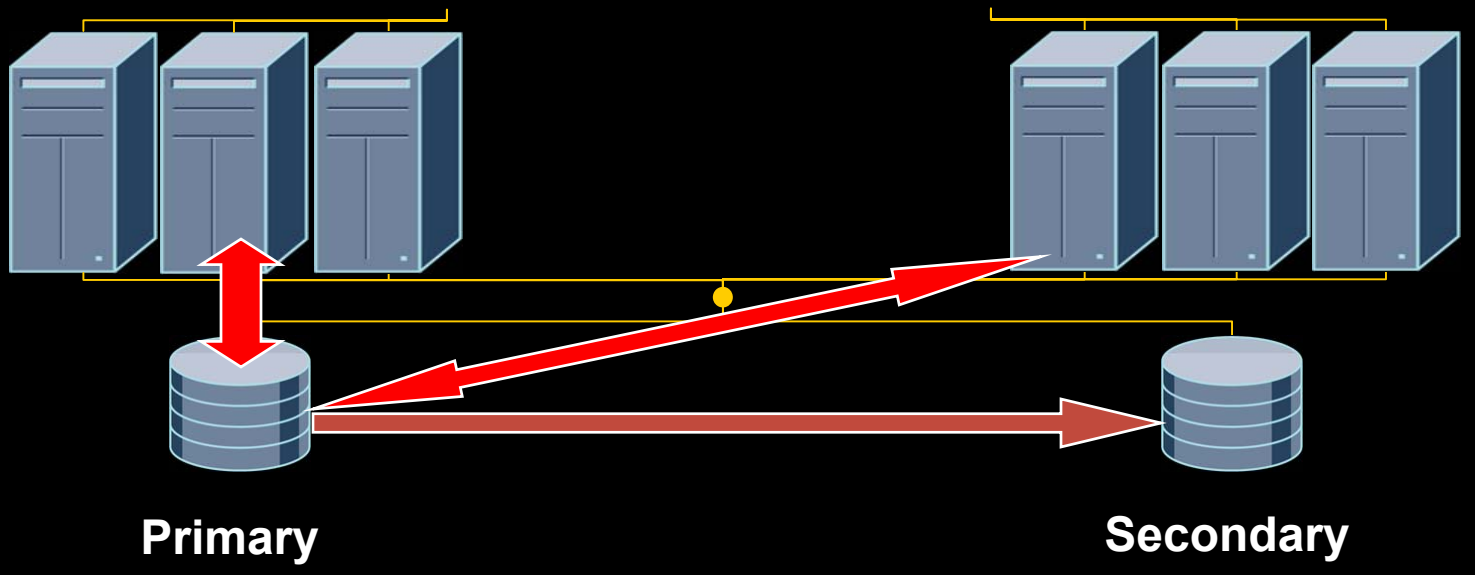
Host Based Mirroring

- Standard *cluster aware* host based LVM solutions (requires a CLVM)
- Disks appear as one set
- All writes get sent to both sets of disks



Array Based Mirroring

- All I/Os get sent to one site, mirrored to other
- Examples: EMC SRDF
- Longer outage in case of failure of primary site



Mirroring Example: Large UK Bank

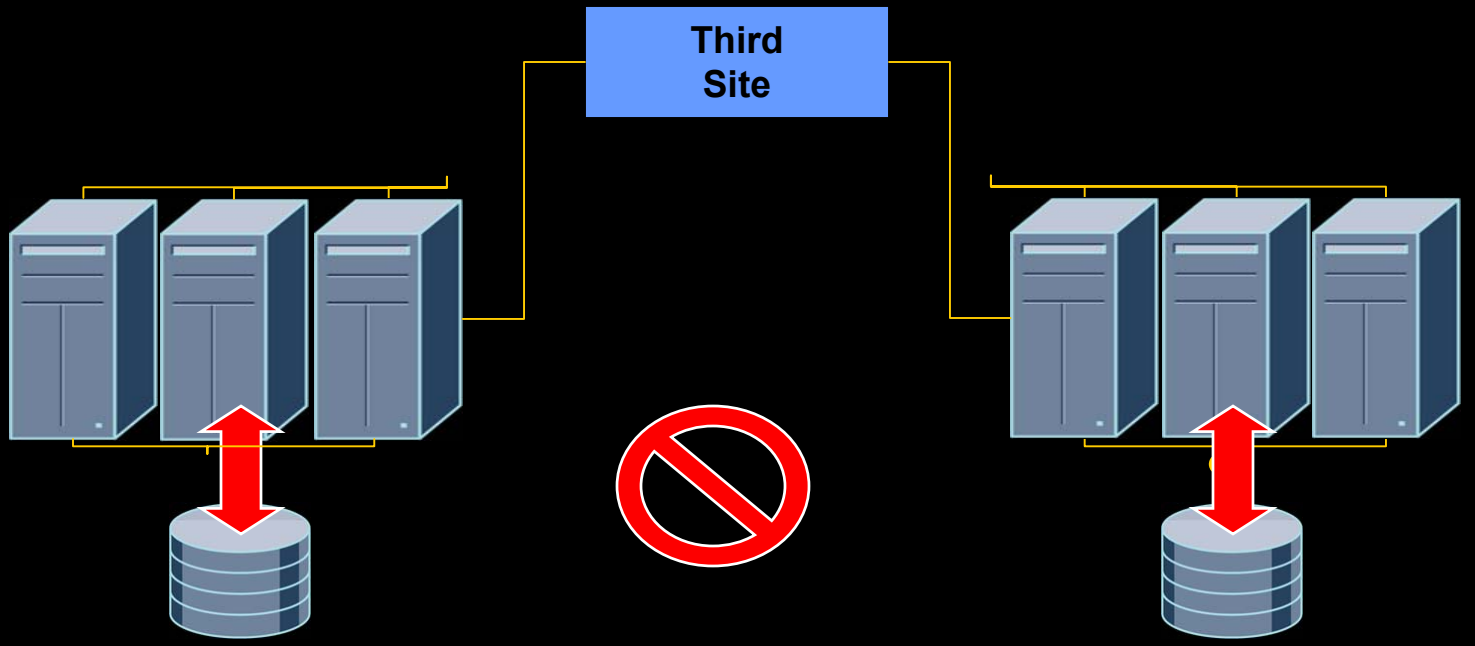
- 2 nodes AIX
- Tested both
- 9 km – Host Based Mirroring – Shark Storage (<1 minute down)
- 20 km – Array Based Mirroring (PPRC) w/ ERCMF (extended remote copy facility) that avoids doing a manual restart by suspending I/Os until PPRC has done the switch. (1-5 minutes down)

Cluster Quorum: Recommendations

- What happens if all communications between sites is lost?

Cluster Quorum: Recommendations

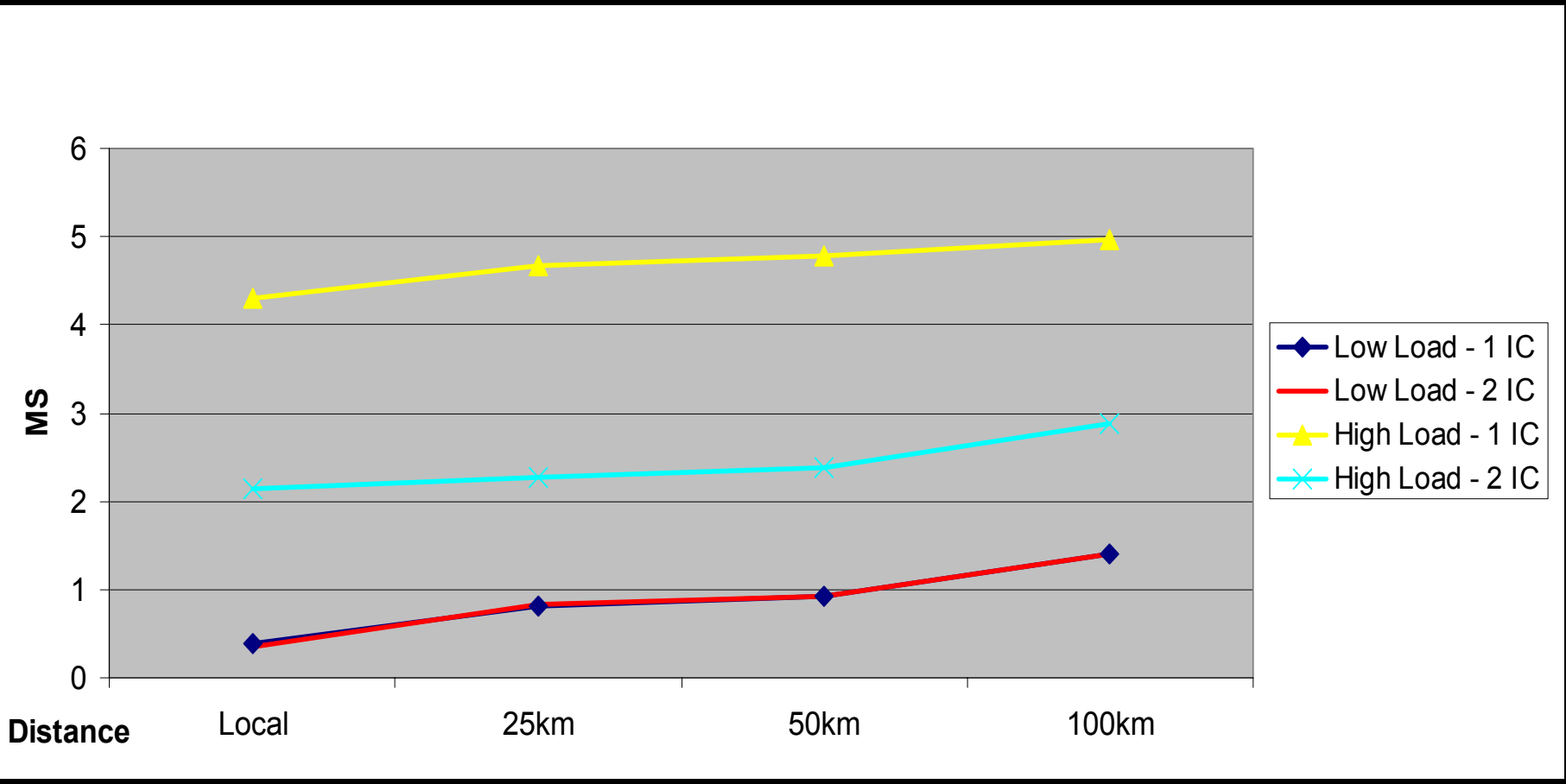
- Use a third site for quorum device for maximum availability



Empirical Performance Data

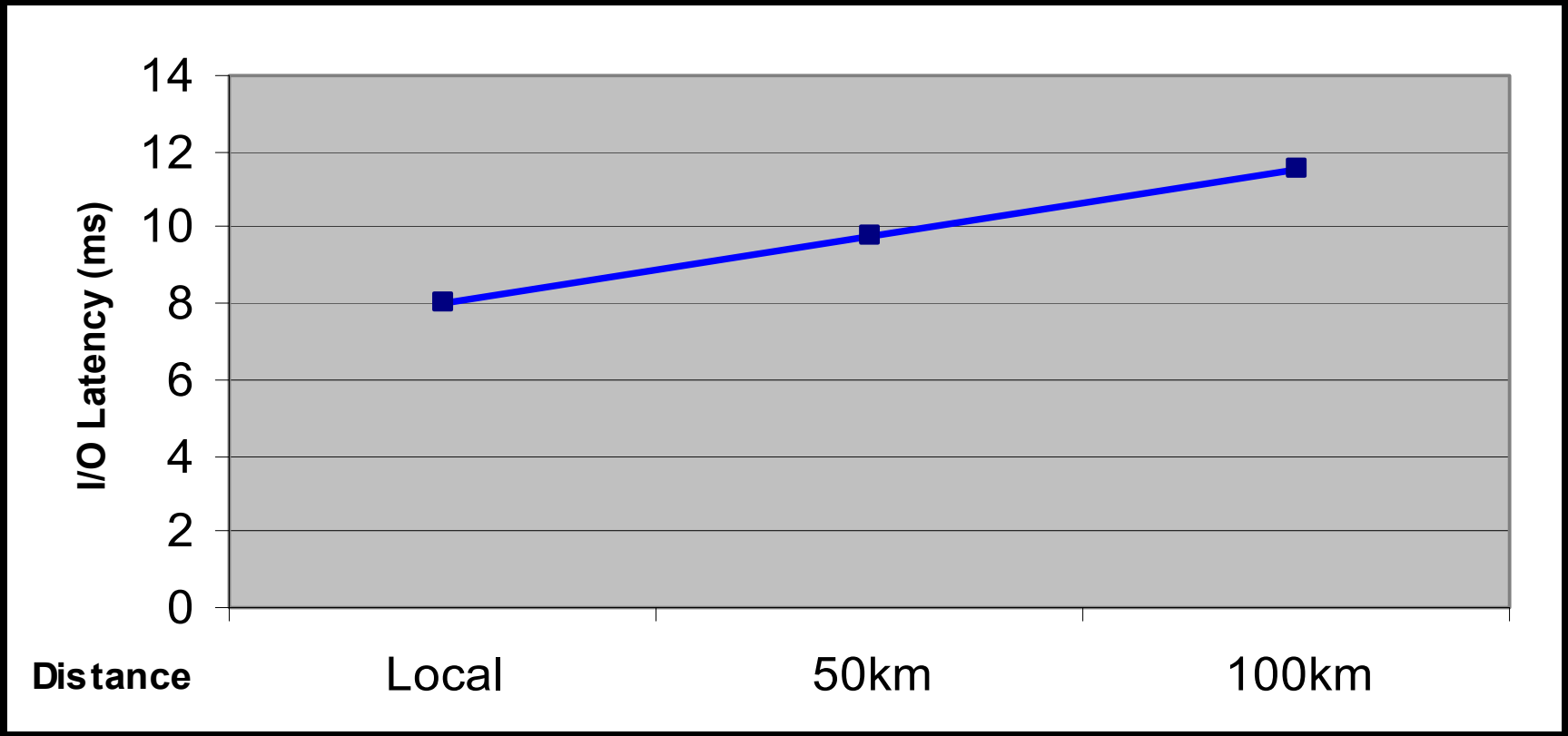
- Unit Tests (Oracle/HP Test results)
  - Cache Fusion
  - I/O
- Overall Application Tests (from 4 different sets of tests)

Empirical Performance Data
Cache Fusion Unit Test



~1ms increased memory-to-memory block transfer latency over 100km for all cases
Results from joint Oracle/HP testing

Empirical Performance Data
I/O Unit Test



I/O latency increased by 43% over 100km.

Note: Without buffer credits this tested at120-270% I/O latency degradation

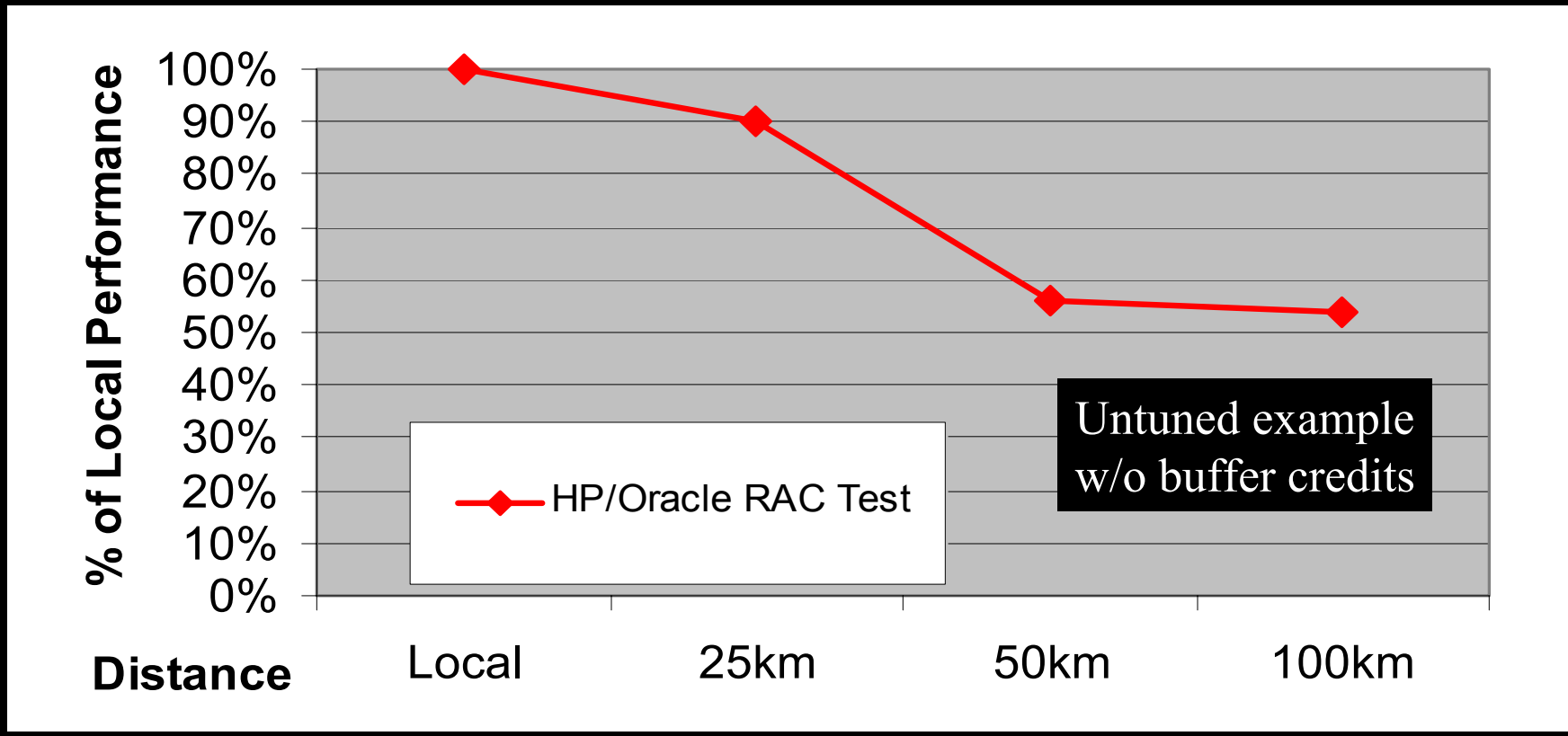
Results from joint Oracle/HP testing

Empirical Performance Data
Overall Results: Joint Oracle/HP Testing

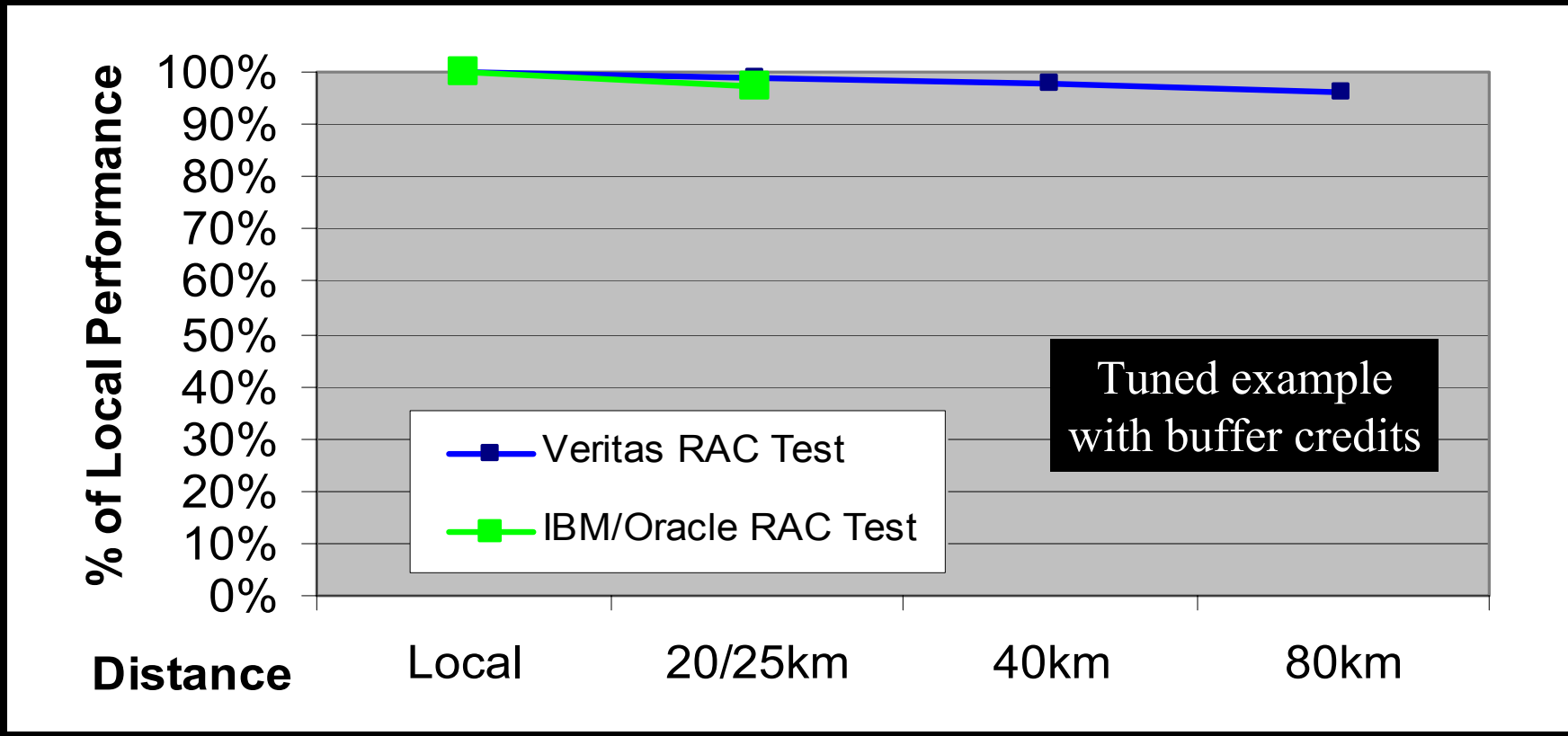
For 100km …

- Memory-to-memory messaging latency increased ~ 1ms
- I/O latency increased in the ballpark of 43% . This is ~ 4-5 ms

Empirical Performance Data
Overall Application Effect



Empirical Performance Data
Overall Application Effect

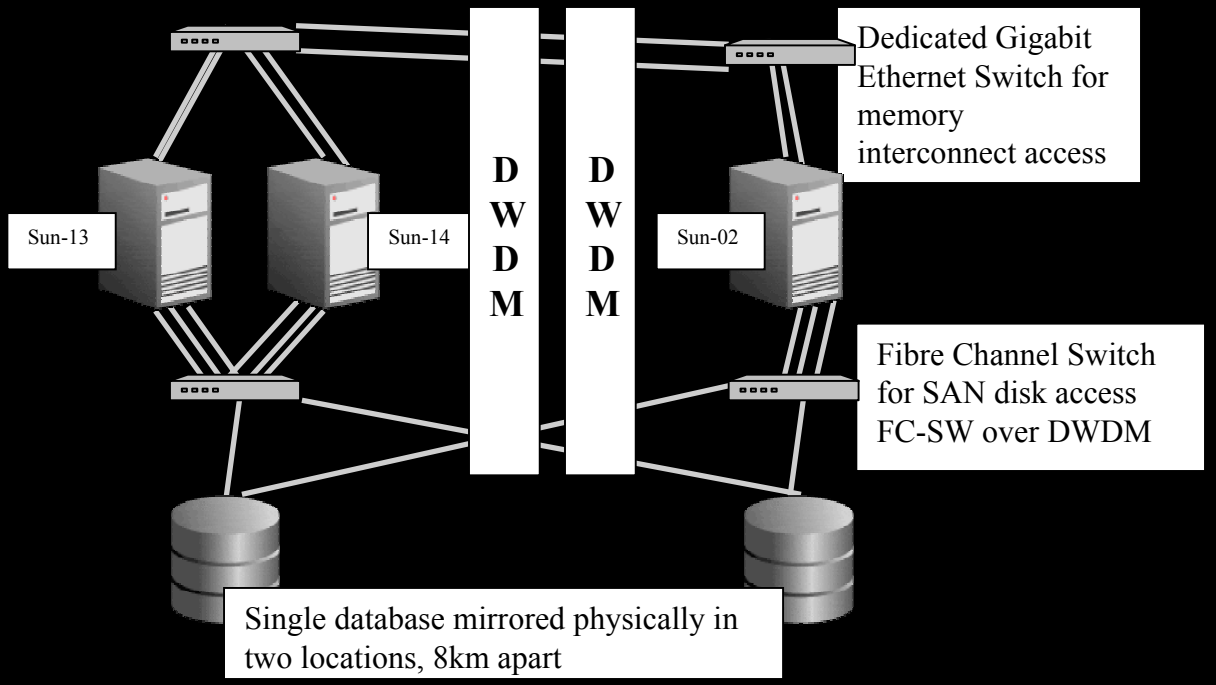


Note: differences in results are due to differences in test cases, not in clusterware used

Comic Relief

- 3 nodes Sun Solaris
- 8km DWDM link
- brownouts of around 11 seconds
- 10% introduced performance hit
- Active/Active: host based mirroring using Veritas Volume Manager

Comic Relief (UK) – Sun 8km



Latency Tests

- ⬆ Oracle instance running with application activity
- ⇧ Oracle instance running but with no application activity

  ⬇ Oracle instance shutdown

sun-13	sun-14	sun-02	Measurements from COMPROD1 running on sun-13	
COMPROD 1	COMPROD 2	COMPROD 3		
⬆	⬆	⬇	395 / s	1.2 ms ①
⬆	⬆	⇧	356 / s	1.4 ms ②
⬆	⬇	⬆	320 / s	1.5 ms ③
⬆	⇧	⬆	310 / s	1.6 ms ④
⬆	⬆	⬆	376 / s	1.6 ms ⑤

Live Customer Examples

First Known Client

- The Rover Group did the first known implementation with a similar architecture in the mid 1990's using Oracle7 Parallel Server.

Austrian Railways

- 6 nodes Tru64
- OPS => RAC migration
- 1.6 km 24 mono-mode fiber optic cable running Memory Channel , 3 nodes on each side
- 2 SAN fabrics
- Host based mirroring
- 13 DB, one RAC, one OPS

ESPN

- American sports broadcasting network
- 9iRAC provides the sports ticker (that shows current scores) that is always on the ESPN channel.
- 2 Node IBM AIX, dual gigabit interconnect
- Distance: Across the Street
- Host Based Mirroring

Strathclyde University

- Running Oracle RAC on 2 node Sun Solaris nodes approximately 1km apart. Previously ran OPS in this environment. Sun Cluster 3, Veritas Volume manager to perform mirroring,

Extended RAC - SAP customers

- **BASF** (Germany, 2 x 2 nodes IBM AIX (8 way)) - 2 TB. Both production and test clusters have nodes 2km apart.

Other Examples

- **Vodaphone Italy** – 2 node Sun Solaris, Sun Cluster, 2.2km, Host Based Mirroring (Veritas)
- **Nordac** - Germany, 4 node HP Tru64, 300m
- **University of Melbourne** - Oracle E*Business Suite 11i on 3 nodes Tru64, 0.8 km
- **China Mobile** (Shanghai) - 3 node IBM AIX using HA GEO for mirroring. 2 corners of Shanghai (15-20km apart) - Host Based Mirroring

Other Examples

- **Western Canada Lotterie Corporation -** 4 node OpenVMS on 10 KM apart
- **Deutsche Bank** (Germany) - 2 node Sun Solaris cluster at 12Km apart.

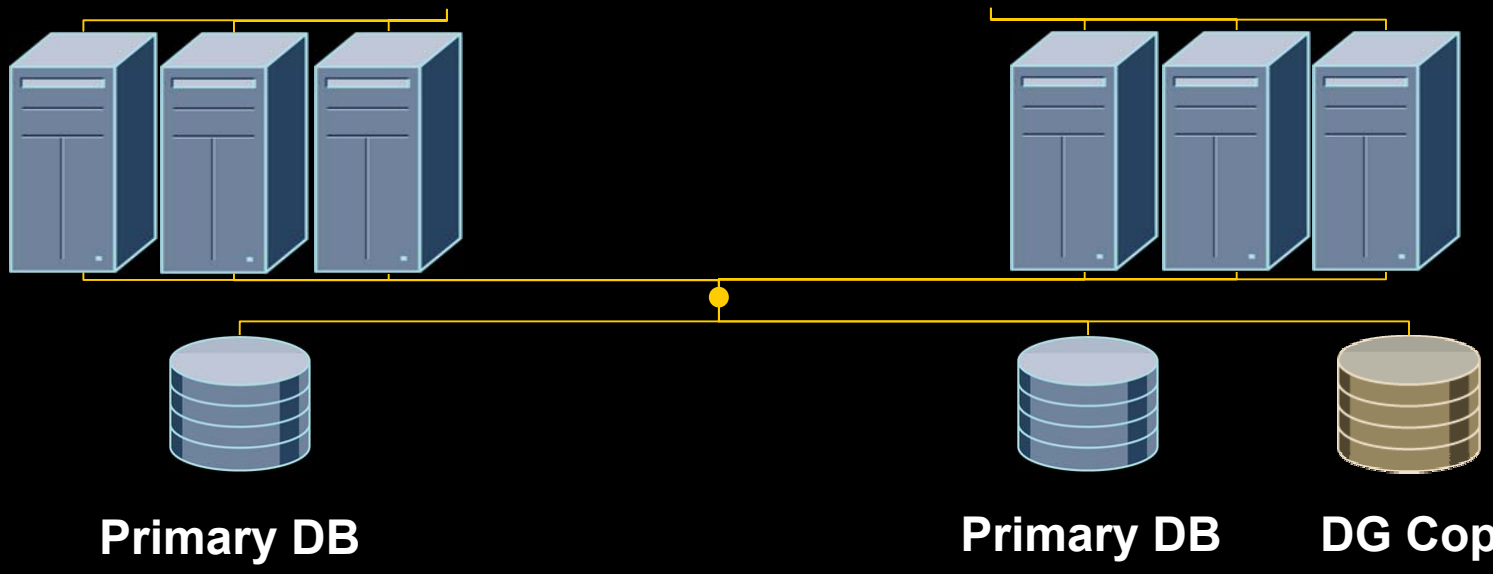
RAC on Extended Clusters Positioning W.R.T. Data Guard

Additional Benefits Data Guard Provides

- Greater Disaster Protection
  - Greater distance
  - Additional protection against corruptions
- Better for Planned Maintenance
  - Full Rolling Upgrades
- More performance neutral at large distances
  - Option to do asynchronous
- If you cannot handle the costs of a DWDM network, Data Guard still works over cheap standard networks.

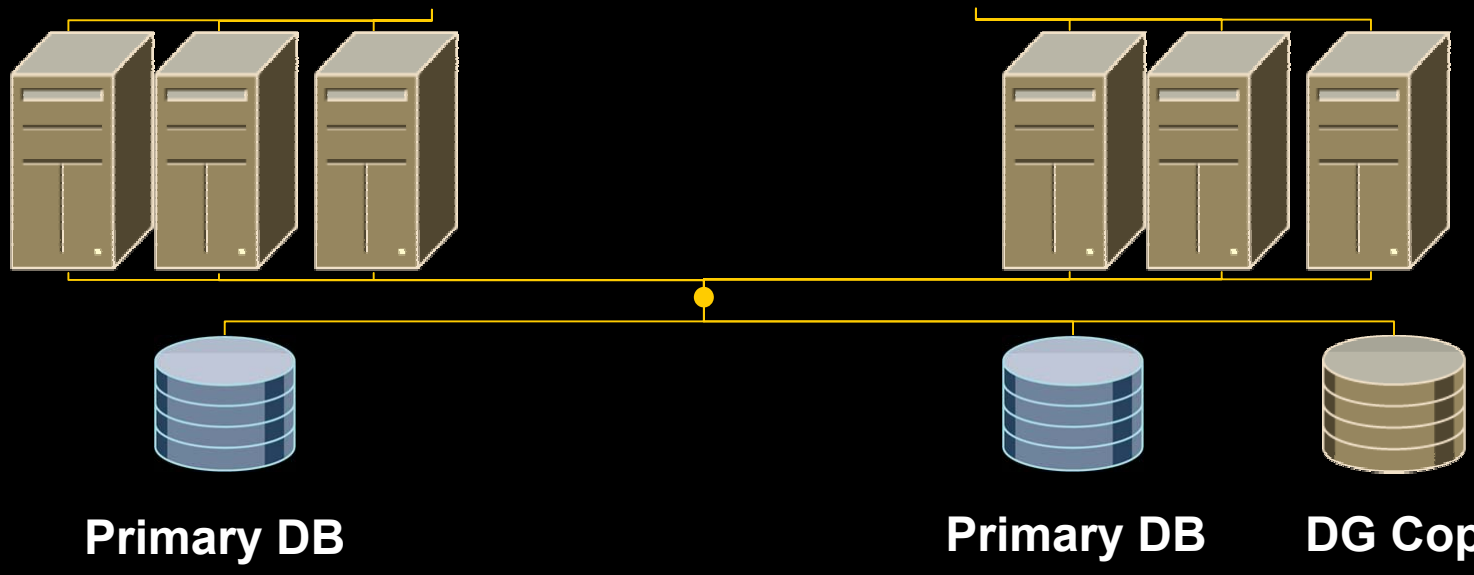
Hybrid: Extended RAC + DataGuard:

- One cluster, one RAC, one primary database
- All nodes are used for the primary RAC
- Separate Data Guard Database, connected to all nodes



Switch to DataGuard

- If need comes to switch to Data Guard copy
- All available nodes can host Data Guard RAC cluster.



Hybrid Advantages

- Protection Against Corruptions
- Better Ability to Support Planned Maintenance (Rolling Upgrades)

- *Distance is still limited*

When does it not work well?

- Distance is too great
  - No fixed cutoff, but as distance increases you are slowing down both cache fusion & I/O activity. The impact of this will vary by application. Prototype first if doing this over ~50km.
- Public Networks
  - To much latency added between the nodes.

Summary

RAC on Extended Cluster

- It works! – proven at customer sites & partner labs.
- Good design is key! Bad design can lead to a badly performing system.
- Data Guard offers additional benefits

References

1. Joseph Algieri & Xavier Dahan, *Extended MC/ServiceGuard cluster configurations (Metro clusters),* Version 1.4, January 2002 <InternalPaper>
2. Sun Microsystems, *Metro clusters Based on Sun Cluster 3.0 Software*, 2002
3. Michael Hallas and Robert Smyth, *Comic Relief Red Nose Day 2003 (RND03), Installing a Three-Node RAC Cluster in a Dual-Site Configuration using an 8 Km DWDM Link,* Issue 1, April 2003
4. Paul Bramy (Oracle), Christine O'Sullivan (IBM), Thierry Plumeau (IBM) at the EMEA Joint Solutions Center Oracle/IBM, *Oracle9i RAC Metropolitan Area Network implementation in an IBM pSeries environment*, July 2003
5. Veritas, *VERITAS Volume Manager for Solaris: Performance Brief – Remote Mirroring Using VxVM*, December 2003
6. CTC TechRep:*How to design a disaster tolerant solution with Oracle9i RAC and HP ContinentalClusters*
7. Mai Cutler (HP), Sandy Gruver (HP), Stefan Pommerenk (Oracle) *"Extended Distance RAC" Eliminating the current physical restriction of Oracle Real Application Cluster*
8. Oracle Maximum Availability Architecture (OTN)

DATABASE 9i CLUSTER

Questions & Answers

Discussion