

RAC **on** **Extended Distance Clusters**

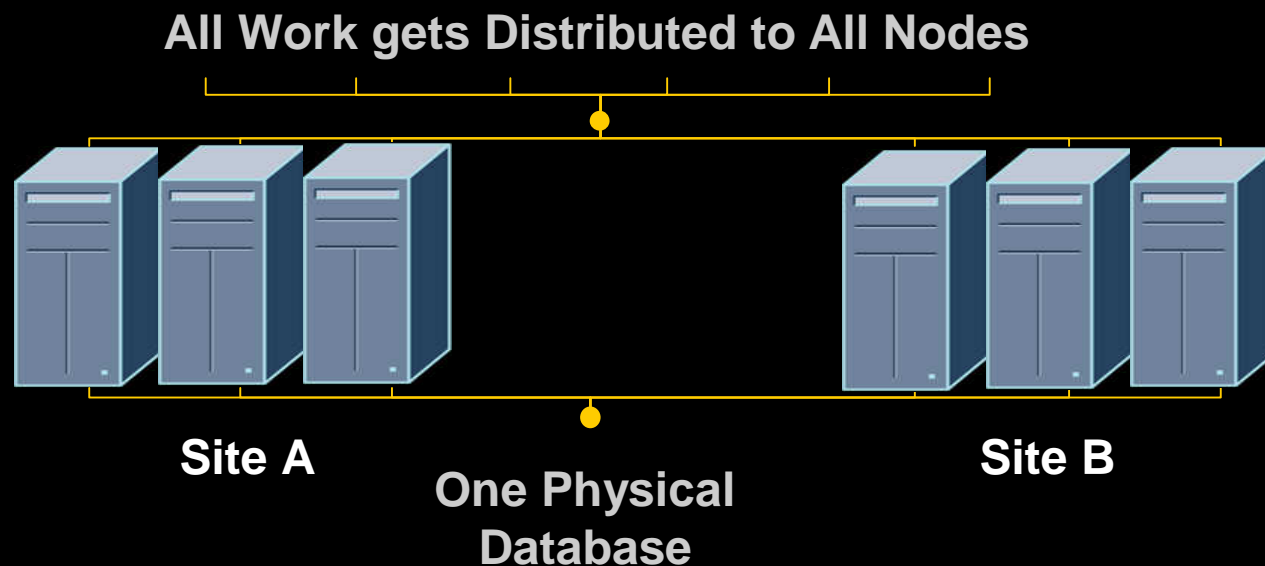
Erik Peterson
RAC Development
Oracle Corporation

Agenda

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

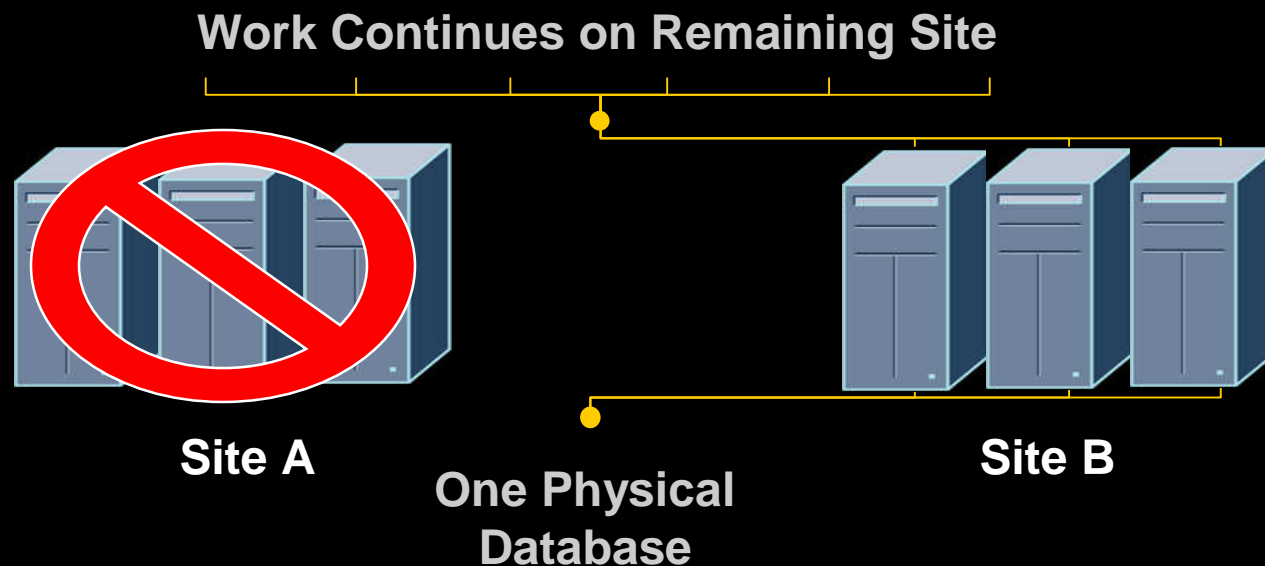
Benefits of RAC on Extended Clusters

- 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



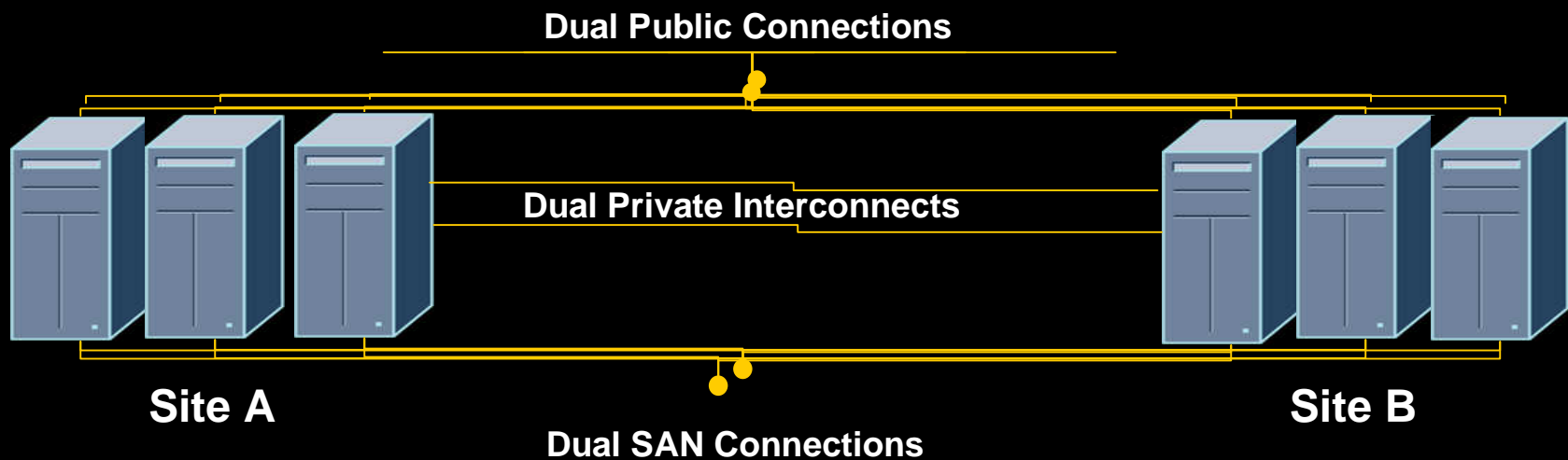
Design Considerations

Design Considerations

- Connectivity
- Disk Mirroring
- Quorum
- Comparing Alternatives
- Other Considerations

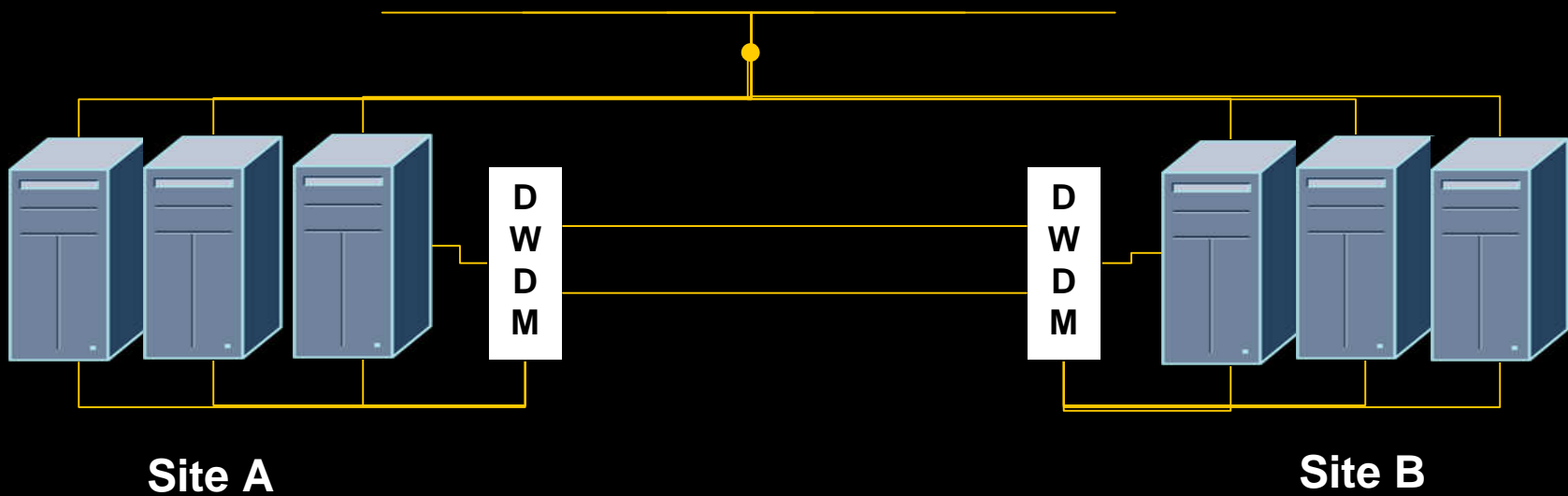
Connectivity

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



Connectivity

- Distances $> 10\text{km}$ 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

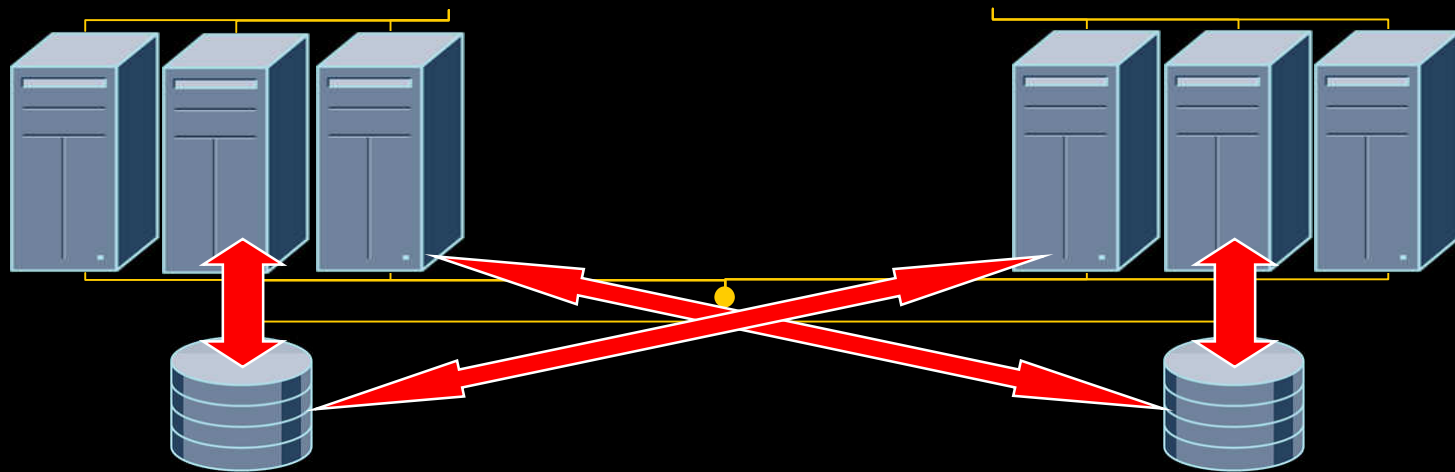
- Distance
 - Single fiber limit (100km?)
- Performance
 - Need to Minimize Latency.
 - Direct effect on synchronous disk mirroring and Cache Fusion operation
 - Direct point to point connection => Additional routers, hubs, or extra switches add latency
- Cost
 - High cost of DWDM if not already present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

Disk Mirroring

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

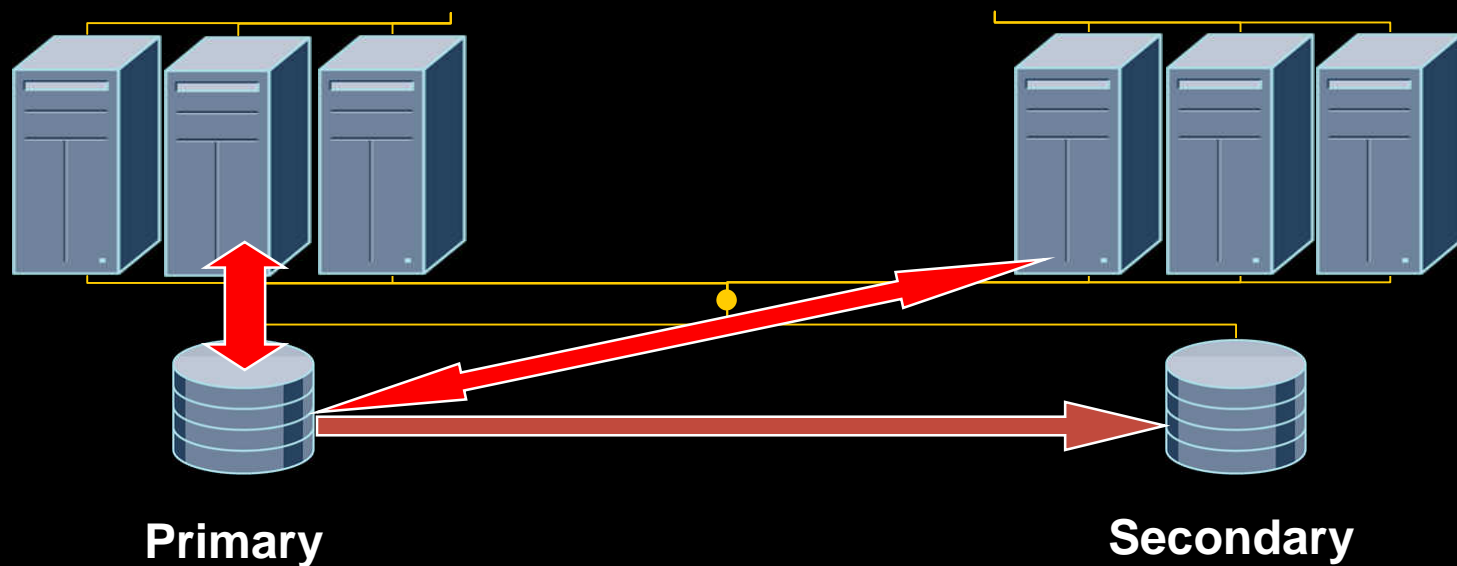
Host Based Mirroring

- 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)

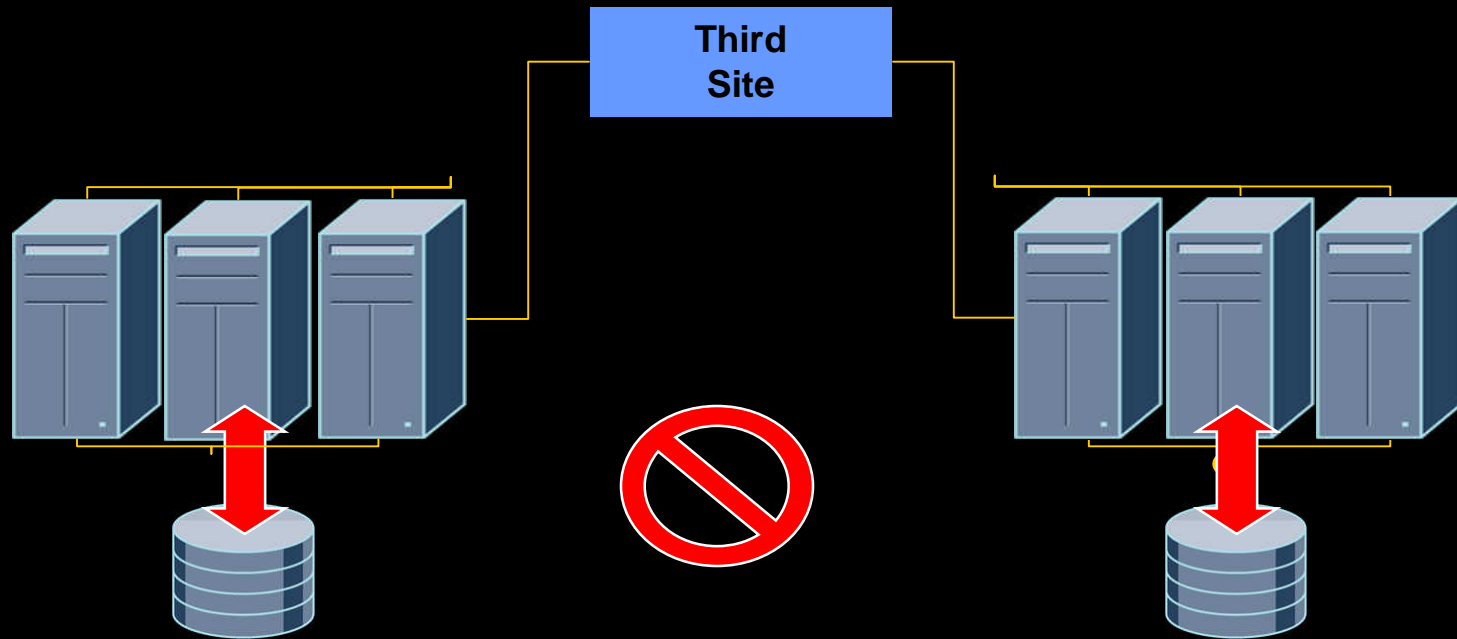
Cluster Quorum: Recommendations

- What happens if all communications between sites is lost?



Cluster Quorum: Recommendations

- Use a third site for quorum device for maximum availability



Primary/Primary or Primary/Secondary?

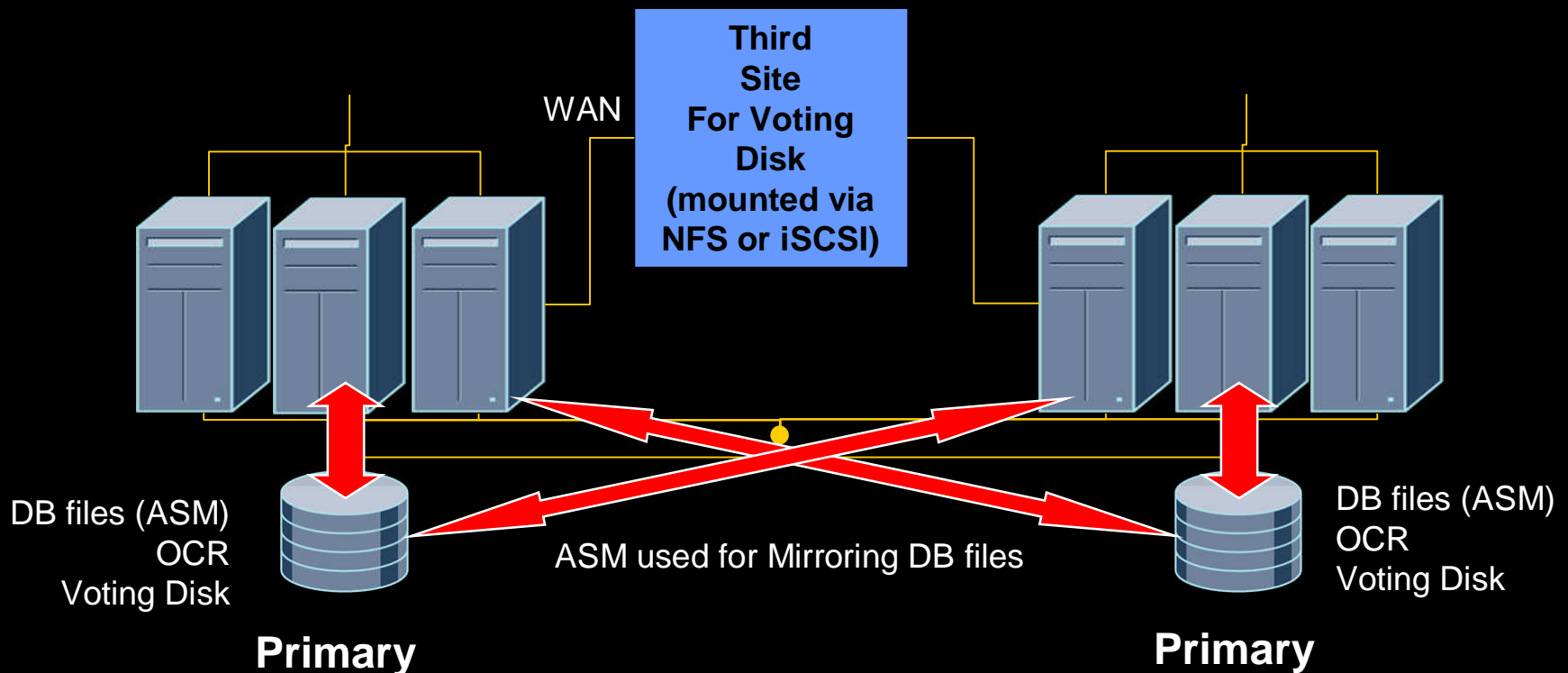
- Active/Active is different from Primary/Primary
- Primary means site continues w/o restart should other site fail
- Primary/Primary – preferred, but requires:
 - 3rd Site for Quorum
 - *and* 10gR2 Oracle Clusterware & ASM for Mirroring
 - *or* 3rd Party Clusterware + Host Based Mirroring
- Primary/Secondary – if any of these conditions exist
 - No 3rd Site
 - Array Based Mirroring
 - 9i or 10gR1 Oracle Clusterware

Clusterware Specifics

Clusterware	Versions	Primary/ Primary	Limitations
Oracle	9i, 10gR1	N	64 Nodes
Oracle	10gR2	Y	100 Nodes
Veritas	All	Y	8 (16) Nodes Not Supported on Linux
HP ServiceGuard	All	Y	16 Nodes < 10km 2 Nodes > 10km Not Supported on Linux
IBM HACMP	All	Y	?
Sun Cluster	All	Y	8 Nodes

10gR2 Extended RAC on pure Oracle Stack

- Any site can fail, and system continues
- Support for Generic NFS for 3rd Voting Disk Currently limited to Linux



Current Limitations of ASM for Extended RAC Mirroring

- Should connectivity between sites be lost, ASM will need to do full resilvering of lost volumes. Partial resilvering is not yet supported.
- ASM currently will read from any available disk group. No optimization is done to do 'local' reads.
- Both items are currently done by some cluster aware LVMs, and are planned for a future release of ASM.

Other Considerations

- Needs to look like a local cluster to Oracle, i.e.:
- Sharing subnets
 - Private Interconnect
 - Public VIPs

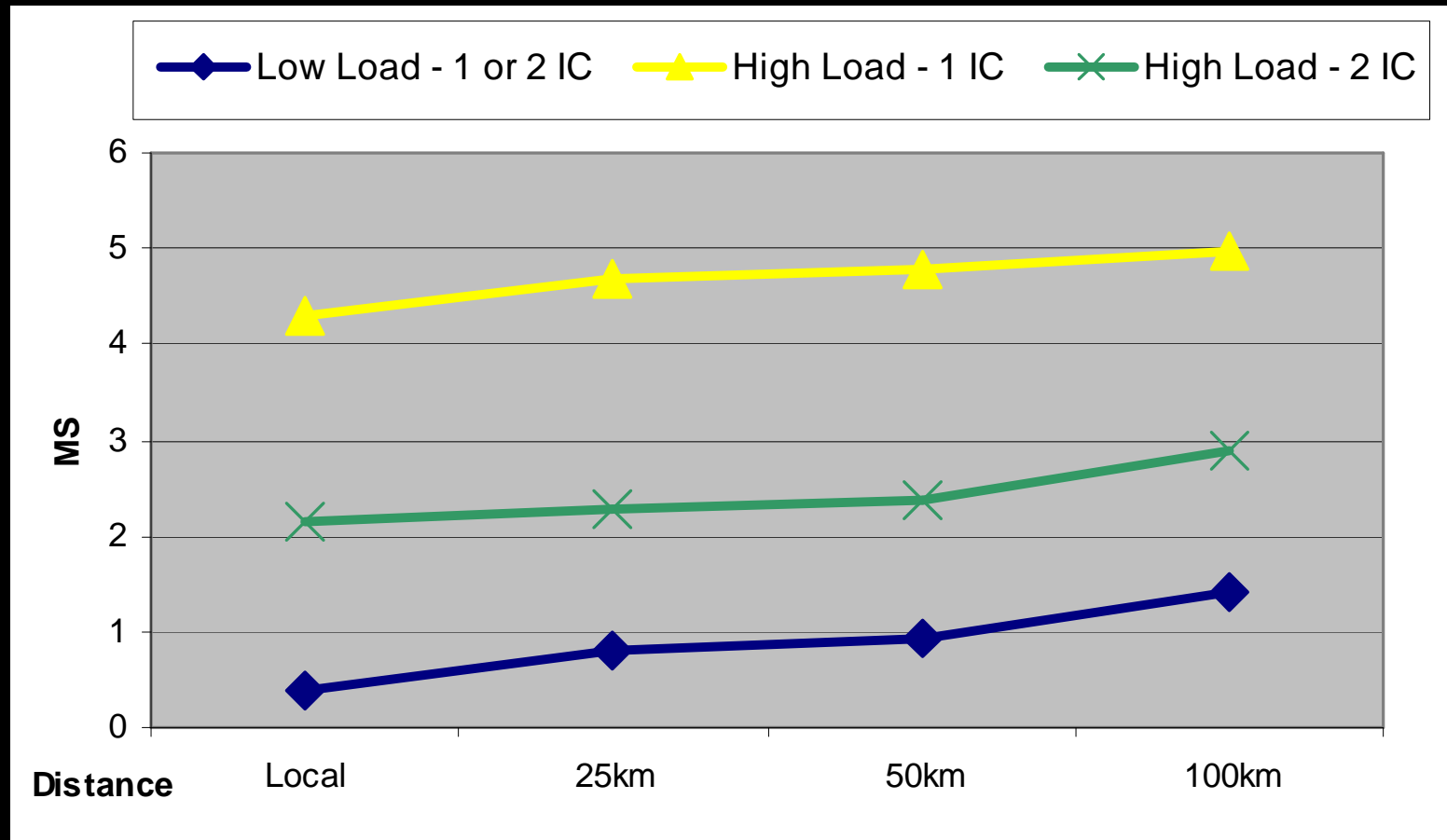
Performance

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

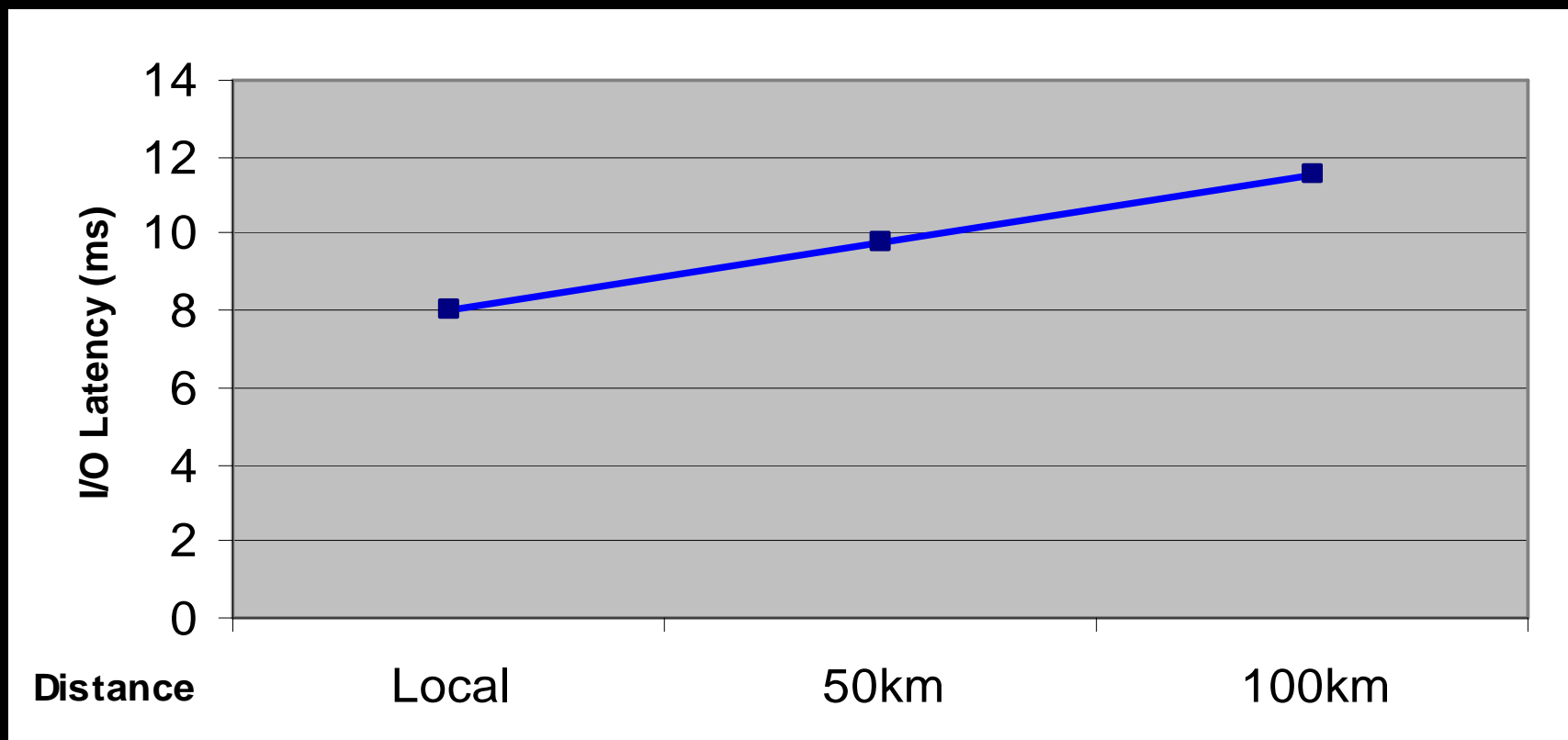
Cache Fusion Unit Test



~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 at 120-270% I/O latency degradation

Results from joint Oracle/HP testing

Empirical Performance Data

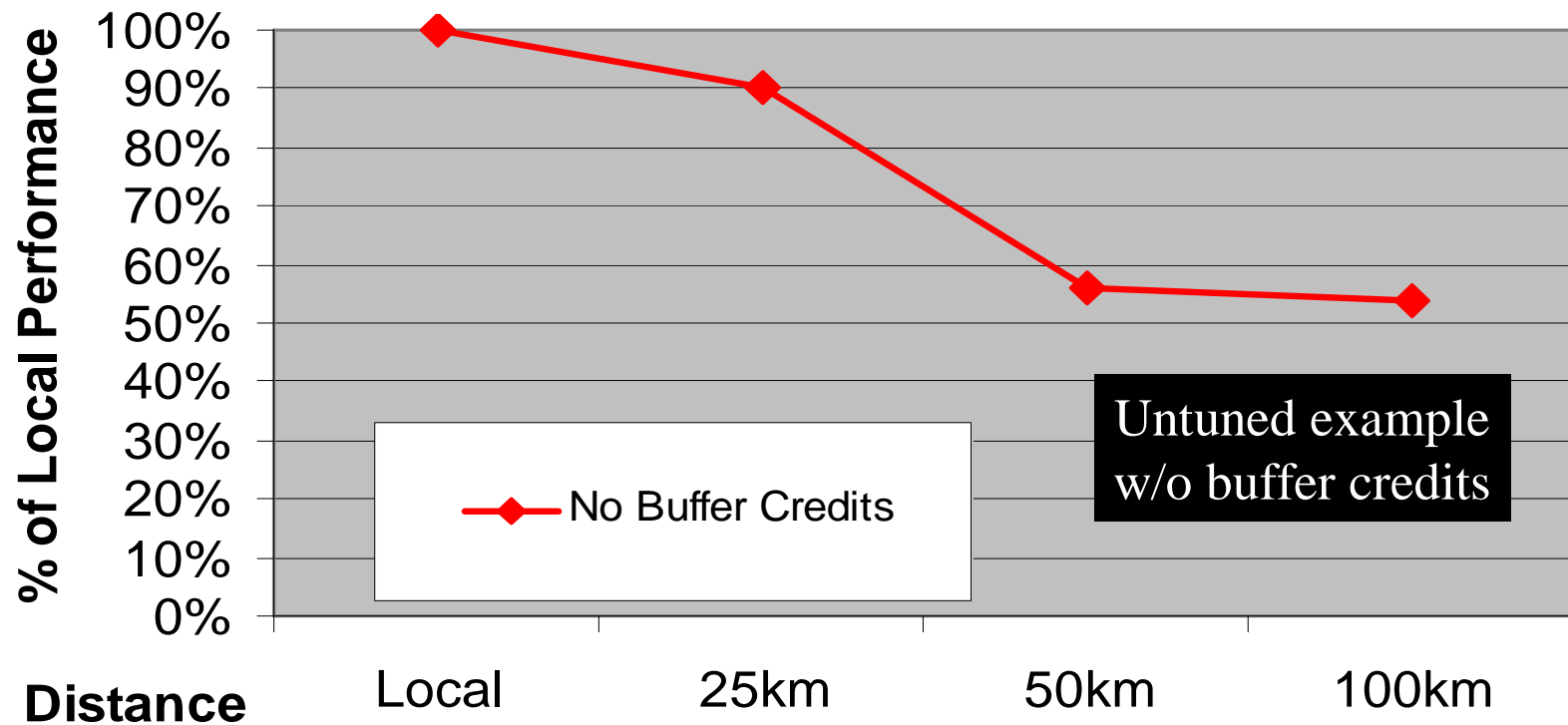
Overall Results: Joint Oracle/HP Testing

For 100km ...

- Memory-to-memory messaging latency increased
~ 1ms
- I/O latency increased in the ballpark of 43% .
This is ~ 3-4 ms

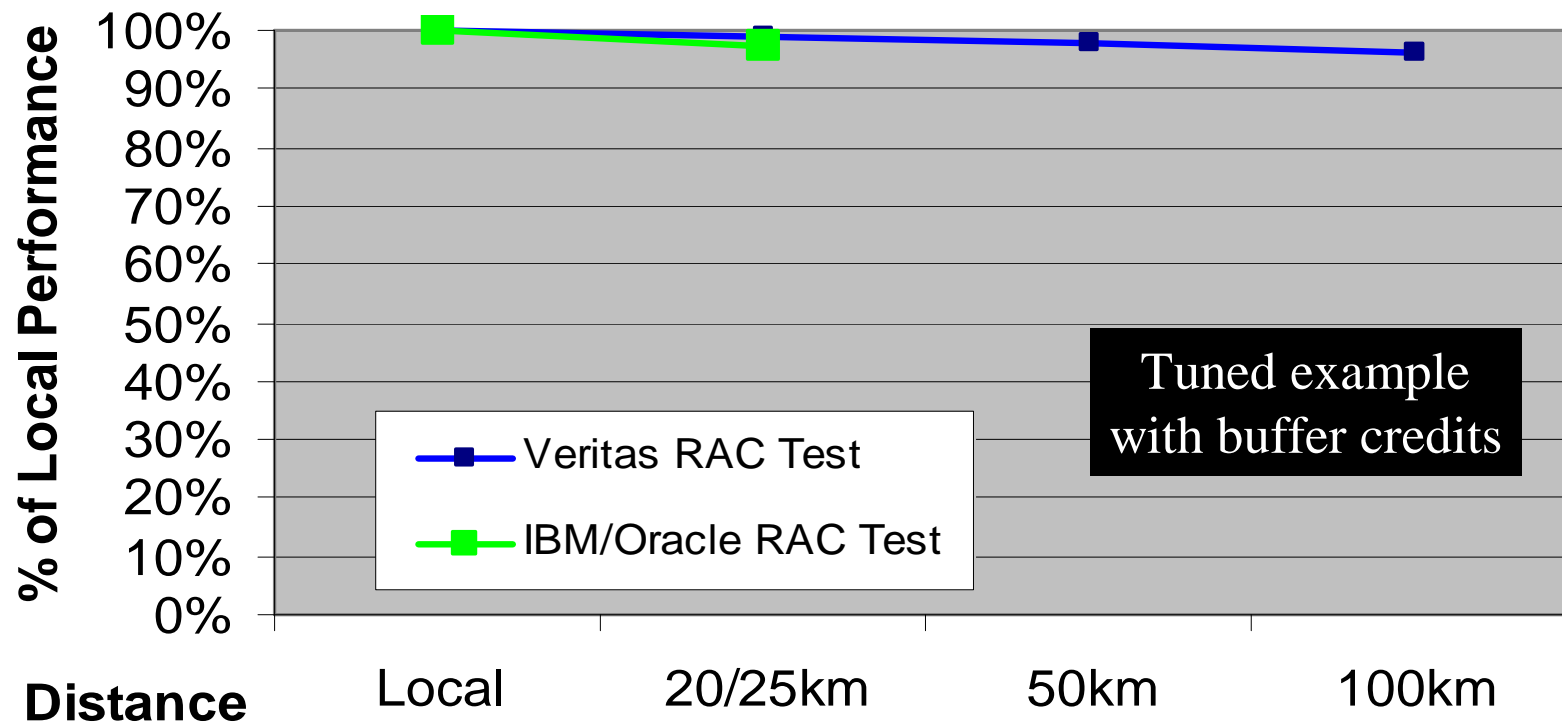
Empirical Performance Data

Overall Application Effect



Empirical Performance Data

Overall Application Effect



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

Oracle 9i - Live Customer Examples

Name	Release	Nodes	Platform	OS	Clusterware	Stretch Distance (KM)
<i>European Electronics firm</i>	9i	2	IBM	AIX	HACMP	8
<i>US Police Department</i>	9i	2	IBM	AIX	HACMP	3
<i>European Government</i>	9i	2	IBM	AIX	HACMP	8
<i>US Broadcaster</i>	9i	2	IBM	AIX	HACMP	0.2
<i>Austrian Hospital</i>	9i	2	IBM	AIX	HACMP	0.6
<i>Brazilian Credit Union Network</i>	9i	3	IBM	AIX	HACMP	10
<i>UzPromStroyBank</i>	9i	2	IBM	AIX	HACMP	1.7
<i>US Fortune 100 firm</i>	9i	2	HP	HP-UX	HP Service Guard	2
<i>Brazilian Hospital</i>	9i	2	HP	HP-UX	HP Service Guard	0.5
<i>North American Lottery</i>	9i	4	HP	OpenVMS		10
<i>European Mobile Operator</i>	9i	3	Sun	Solaris	Veritas Cluster	48
<i>Comic Relief</i>	9i	3	Sun	Solaris	Sun Cluster	8
<i>German Bank</i>	9i	2	Sun	Solaris		12
<i>European Mail</i>	9i	2	Sun	Solaris	Veritas Cluster	12
<i>European Government</i>	9i	2	Sun	Solaris	Sun Cluster	0.4
<i>UK University</i>	9i	2	Sun	Solaris	Sun Cluster	0.8
<i>Italian Telco</i>	9i	2	Sun	Solaris	Sun Cluster	2
<i>Austrian Railways</i>	9i	2	HP	Tru64	TruCluster	1.5
<i>Nordac/ Draeger</i>	9i	4	HP	Tru64	TruCluster	0.3
<i>University of Melbourne</i>	9i	3	HP	Tru64	TruCluster	0.8

Oracle 10g - Live Customer Examples

Name	Release	Nodes	Platform	OS	Clusterware	Stretch Distance (KM)
<i>Italian Financial Services firm</i>	10g	20	IBM	AIX	HACMP	0.2
<i>Groupe Diffusion Plus</i>	10g	2	IBM	AIX	Orade	0.5
<i>Austrian IT Services Provider</i>	10g	2	IBM	AIX	HACMP	1
<i>Daiso Sangyo</i>	10g	2	HP	HP-UX	Orade	10
<i>Italian Manufacturer</i>	10g	4	HP	Linux	Orade	0.8
<i>Swedish Automotive Parts</i>	10g	2	IBM	Linux	Orade	2
<i>Austrian Health Provider</i>	10g	2	IBM	Linux	Orade	0.3
<i>Thomson Legal</i>	10g	8	Sun	Linux	Orade	1
<i>German Telecom</i>	10g	4	Sun	Solaris	Sun Cluster	5
<i>European Bank</i>	10g	2	Sun	Solaris	Orade	5
<i>European Electronics Components firm</i>	10g	2	IBM	Windows	Orade	0.5

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.

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
 - Too 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

- **Roland Knapp, Daniel Dibbets, Amit Das**, Using standard NFS to support a third voting disk on a stretch cluster configuration on Linux, September 2006
- **EMEA Joint Solutions Center Oracle/IBM**, 10gRAC Release2 High Availability Test Over 2 distant sites on xSeries, July 2005
- **Paul Brame (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
- **Veritas**, VERITAS Volume Manager for Solaris: Performance Brief – Remote Mirroring Using VxVM, December 2003
- **HP Oracle CTC**, Extended Serviceguard cluster configurations. Detailed configuration information for extended RAC on HP-UX clusters, November 2003
- **Mai Cutler (HP), Sandy Gruver (HP), Stefan Pommerenk (Oracle)** Eliminate the Current Physical Restrictions of a Single Oracle Cluster, OracleWorld San Francisco 2003
- **Joseph Algieri & Xavier Dahan (HP)**, Extended MC/ServiceGuard cluster configurations (Metro clusters), Version 1.4, January 2002
- **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
- **Ray Dutcher**, Oracle9i Data Guard: Primary Site and Network Configuration Best Practices, October 2003
- **Joseph Meeks, Michael T. Smith, Ashish Ray, Sadhana Kyathappala**, Oracle Data Guard 10g Release 2 Fast-Start Failover Best Practices, November, 2005
- **Tim Read**, Architecting Availability & Disaster Recovery Solutions, Sun BluePrints™ OnLine, April 2006

A large graphic featuring the letters 'Q', '&', and 'A' in a serif font. The 'Q' and 'A' are in a dark grey color, while the '&' is in a bright red color. The word 'Questions' is written in white serif font to the left of the 'Q', 'Answers' is written in white serif font to the right of the 'A', and 'Discussion' is written in white serif font below the '&'.

Questions & Answers
Discussion