Architectures for High Availability and Disaster Tolerant Solutions

Seminar # 3528

Bob Sauers / Chief Architect
Bill Gray / Advanced Technology Consultant
HP High Availability Advanced Technology Center (HA ATC)

Agenda

- Introduction to High Availability and Disaster Tolerance
- Trends in High Availability and Disaster Tolerance
- High Availability Technology Components
- HP High Availability Solutions
- HP Disaster Tolerant Solutions
- High Availability Reference Architectures
- High Availability Solutions in HP-UX 11i v2UP2
- Summary

Introduction to High Availability and Disaster Tolerance

Business drivers for increasingly higher levels of availability

- Society and businesses are moving toward a more dynamic and continuously online world
  - Increased competitive pressures
  - Globalization of services
  - E-commerce
  - Growing customer demand for fast, easy, and continuous information access
  - Business processes are dependent on IT infrastructure. No manual backups: when the infrastructure stops working, the business stops working

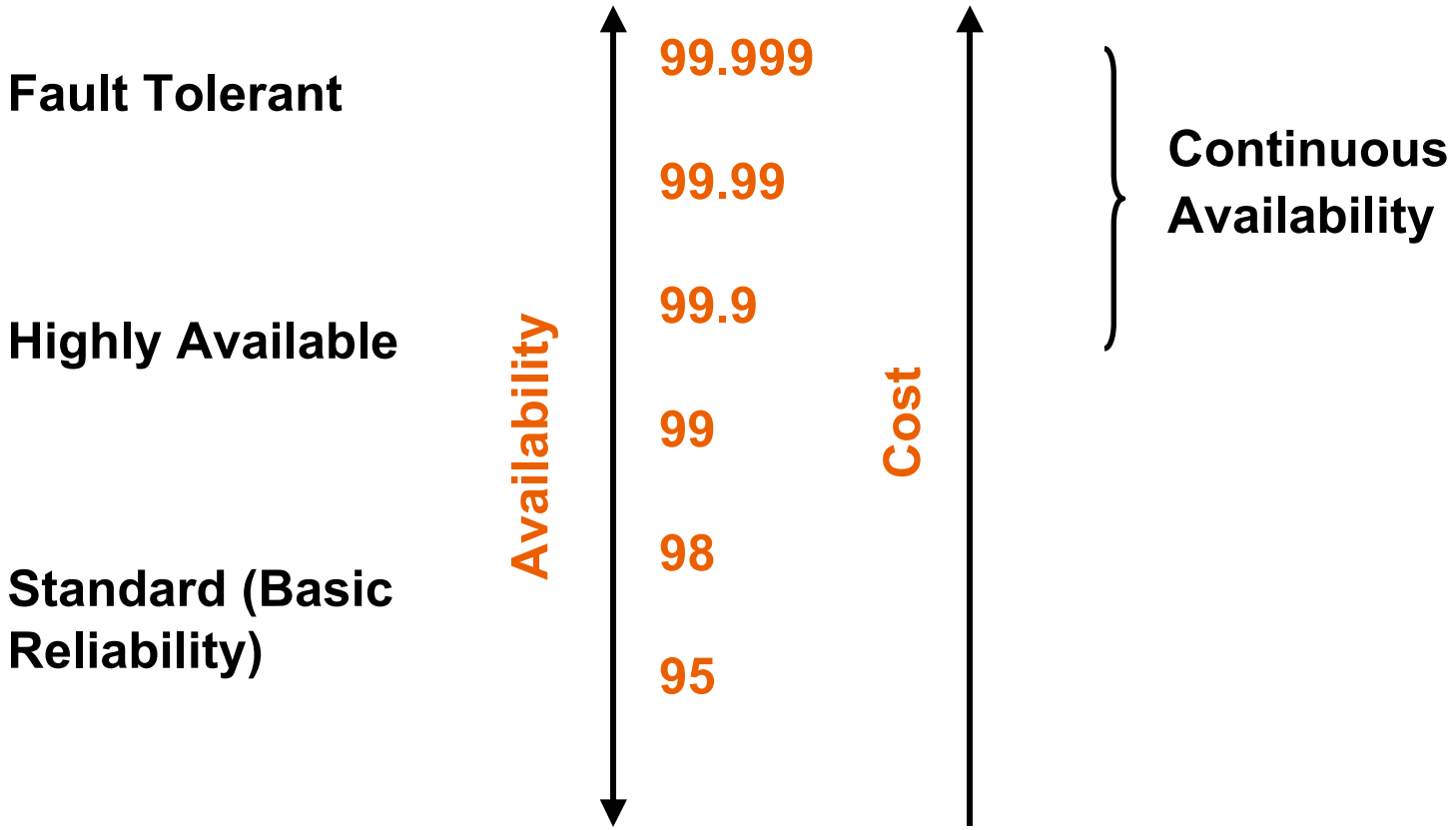
Business Challenge: Develop flexible, scalable, and highly available infrastructure that allows IT to meet the demands of today's and tomorrow's business drivers

Downtime can mean headlines

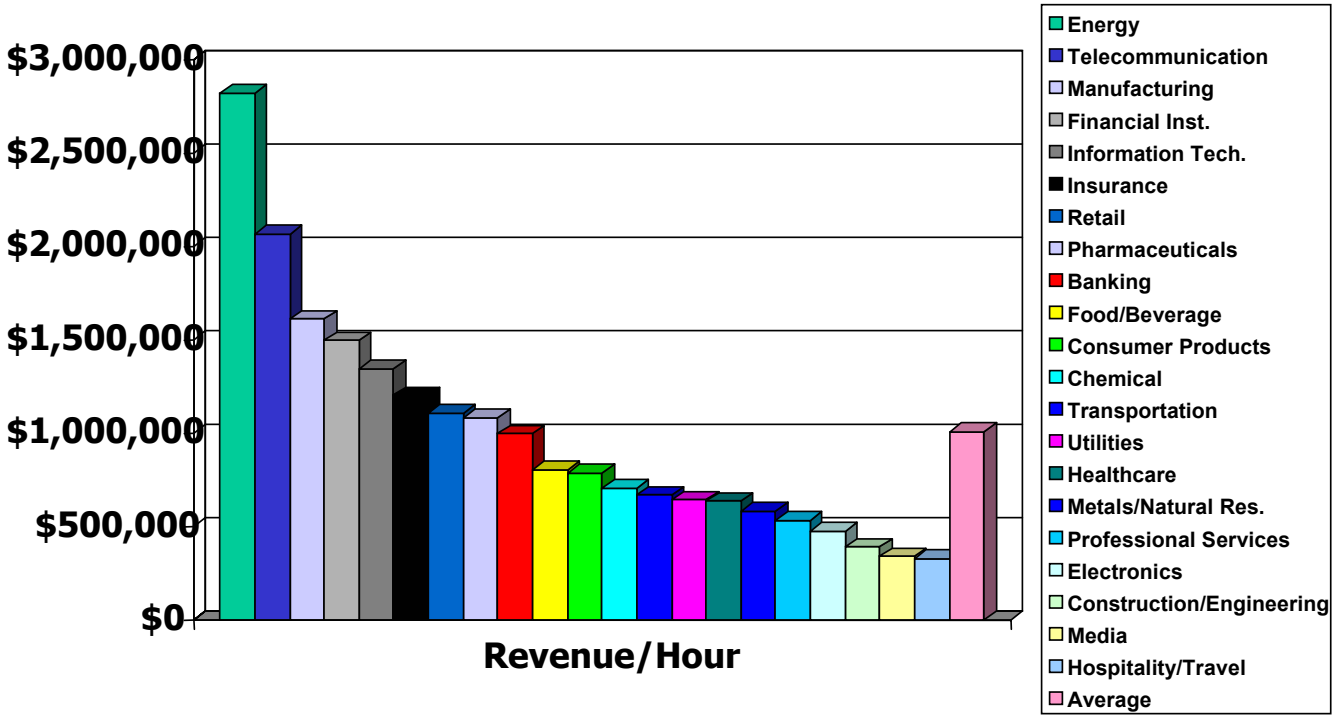
What are the total business consequences of an outage?

- Tarnished company reputation and customer loyalty
- Lost opportunities and revenue
- Idle or unproductive labor
- Cost of restoration
- Penalties
- Litigation
- Loss of stock valuation
- Loss of critical data

Types of HA Systems



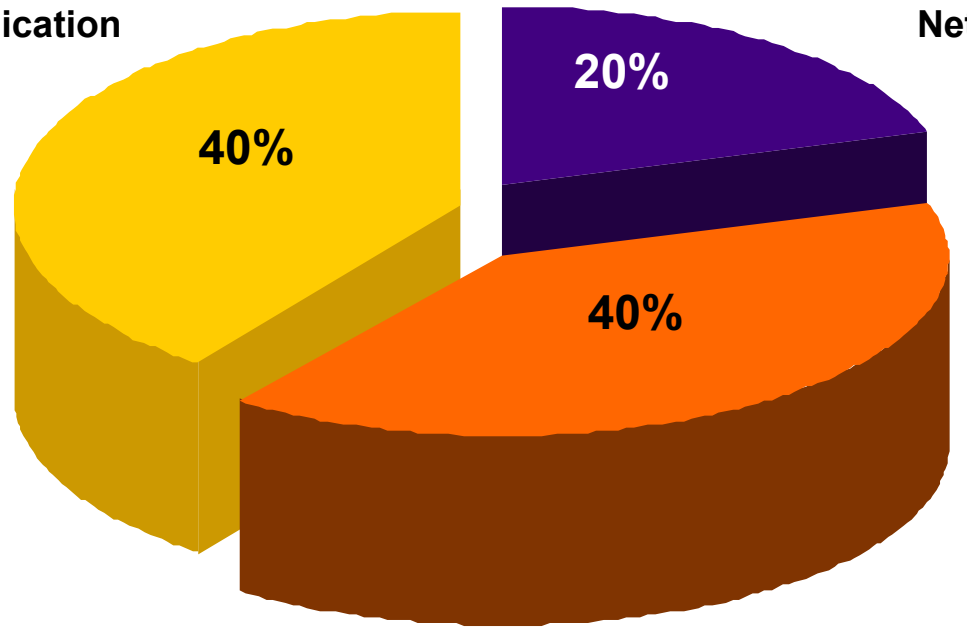
Cost of not implementing HA



Source: Meta Group, 10/2000

Causes of system failure

Software Failures
OS, Middleware, Application



Technology Failures
Network, servers, disk

Human Errors
Operational and Administrative Errors
Errors under duress

Source : Gartner Group

Measuring Availability

- Organizations define availability in many ways
  - Total availability (we'll discuss this in a minute)
  - Hours of operation
    - 5x8, 6x12, 7x23, 7x24, 6x24 + 1x8
  - Incident reporting
  - Automatic
  - Manual
  - Accuracy
  - Tracking the cause
  - What does it include?

Measuring Availability

- Unplanned downtime
  - Network elements
  - Network links
  - Servers
  - Disk storage
  - Operating system
  - Middleware
  - Application software
- Planned downtime
  - Hardware upgrades
  - Software upgrades and patches
  - Database and file system maintenance
  - Backups

Measuring Availability – Point of Measurement

- Point of Measurement
  - The OS prompt
  - Middleware is ready for transactions
  - Application is ready for connections
  - Users see the application on the monitor

- The higher the point the measurement, the greater the cost and complexity of achieving a given availability level

Measuring Availability – How many Nines?

- Number of nines – is this realistic?
  - 99.95% -- four hours of downtime per year
  - 99.99% -- one hour of downtime per year
  - 99.999% -- five minutes of downtime per year

- Reminder: what does it include? The more it includes, the harder and more expensive it is to achieve.
  - Just unplanned downtime?
  - Planned downtime?
  - Disasters?

Service Level Agreements (SLAs)

- Usually defined by the business unit and includes:
  - Application criticality
  - Acceptable planned downtime
  - Acceptable unplanned downtime
  - Hours of availability
  - Duration of outages
  - Average monthly or annual availability
  - Number of users
  - Acceptable response time and other performance metrics
  - Contracted costs (payments)
  - Penalties
  - etc.
- May include disaster-related objectives
  - Recovery Time Objective (RTO)
  - Recovery Point Objective (RPO)

Measuring Availability - Exclusions

- Possible exclusions from SLA and measurement
  - Planned outage windows (weekly, monthly, quarterly, etc.)
  - Emergency outage windows
  - Disaster situations

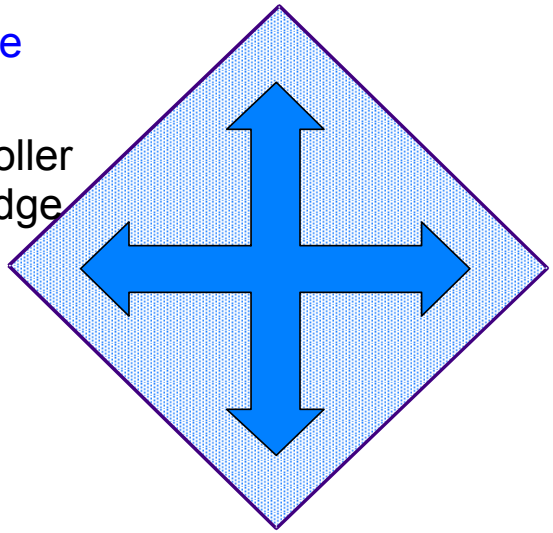
Goal of High Availability solutions

- hardware, system software, applications & middleware, and IT processes designed to minimize both planned and unplanned downtime that minimize time that an application is not available to the end user

Reduce number of outages

Unplanned Service Outage

- Disks – cable, card, controller
- Networks – card, LAN, bridge
- System – OS crash, hdw
- Site Disasters
- Power
- Database Software
- Application Software
- User/Operator Error



Planned Service Outage

- Systems – OS upgrade/patch
- Hardware – disk, LAN, CPU
- Upgrade – HW, OS, DB, App
- Reconfig – HW, OS, DB, App
- Patch – HW, OS, DB, App
- Maintenance – HW, OS, DB, App
- Backup

Reduce duration of outage

Achieving Availability

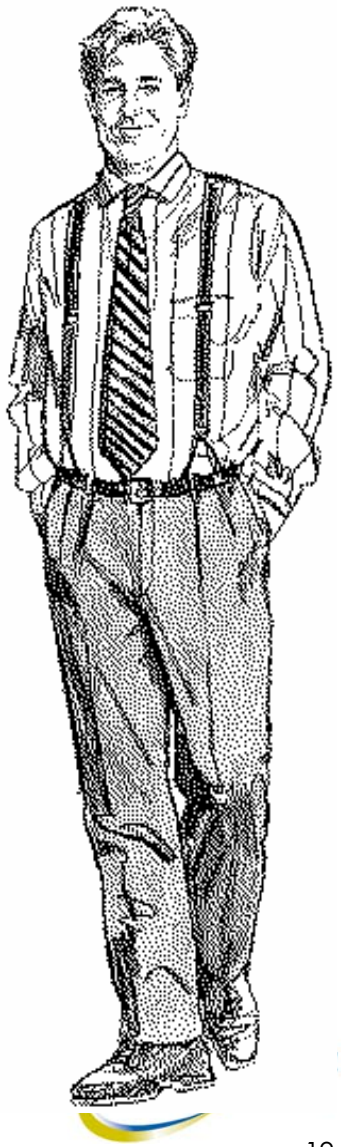
- Combination of 3 major pillars of availability
  - Technology
    - Reliable architectural components (servers, disks, network, etc.)
    - Clustering software
    - Middleware
    - Data replication
  - People and Processes
  - Support Services

High Availability versus Disaster Tolerance

- High Availability
  - Providing redundancy within a data center to maintain the service (with or without a short outage)
    - Hardware failures
    - Software failures
    - Human error
- Disaster Tolerance
  - Providing redundancy between data centers to restore the service quickly (tens of minutes) after certain disasters (dedicated equipment)
    - Power loss
    - Fire, flood, earthquakes
    - Sabotage, terrorism
- Disaster recovery
  - Usually providing a remote site with similar equipment that is shared among multiple organizations (shared equipment model)
  - Personnel fly to the site with tapes and restore the service in days or weeks

High Availability

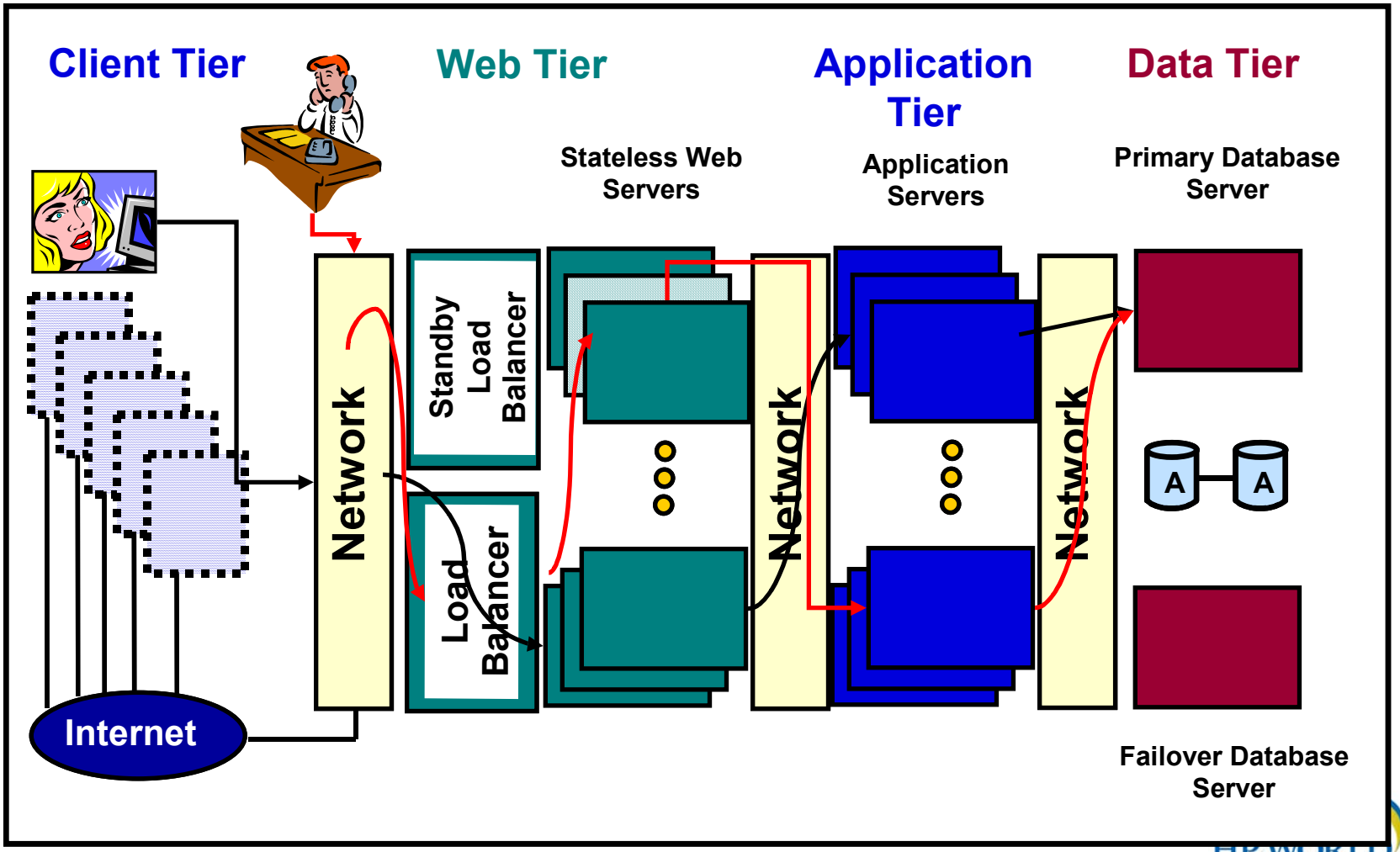
- Redundancy involves multiple pieces of hardware that will take over immediately or within a short time, in case of component failure
  - Increased investment in hardware & software
  - Sometimes results in no disruption of the service
    - networks
    - protected disks (RAID)
  - Other times, a short outage occurs while switching to the redundant hardware
    - Systems
    - Data centers



Types of High Availability solutions

- None
- Duplicated content
- Network-access to content on HA servers
- Load Balancing for stateless servers
- Preservation of transactions in progress
- Failover Cluster
- Disaster Tolerance
- Disaster Recovery

Multi-Tier HA Architecture example



Architecture tiers

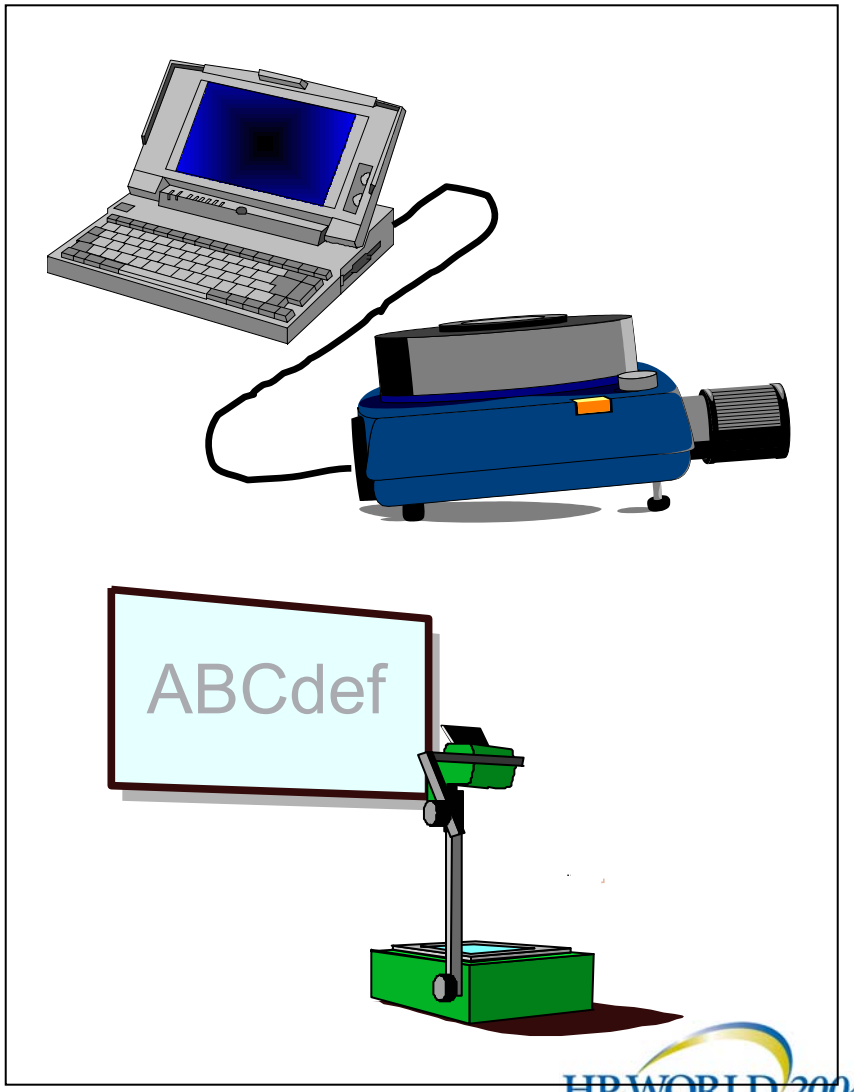
- Access Tier
  - Access
    - Network-based load balancing
    - Network redundancy
- Web / Security Tier
  - Security
    - Security firewall and encryption
    - HA by duplication of servers
  - Web
    - Software-based load balancing and Quality of Service (QoS)
    - HA by duplication of content or NFS access by multiple web

Architecture tiers (continued)

- Application Tier
  - Stateless application servers
    - HA through replicated servers (app tier) and software load balancing (web tier)
  - Stateful application servers
    - HA through failover clustering
      - MC/ServiceGuard
      - MSCS
- Data Tier
  - Database Server
    - HA through failover clustering
      - MC/ServiceGuard
      - MSCS
    - HA through parallel access
      - ServiceGuard OPS Edition
      - HA through data replication

Disaster Tolerance

- Redundancy involves multiple pieces of hardware in another geographic location that will take over within a reasonably short time in case of certain disasters
- An outage occurs while switching to the redundant hardware or location



Choosing the right solution

- Criticality of the application
  - Mission critical, business critical, etc.
- Business needs
  - Hours of operation, importance of the data
- Budget
- Cost of downtime
  - Lost revenue and productivity
  - Loss of customers
- Vulnerability (risk) analysis
- Cost / Benefit analysis
  - Return on investment (ROI)

Industry Trends in High Availability and Disaster Tolerance

Industry Trends in High Availability

- More and more nines
- Measuring higher in the stack
- Manageability improvements
- Automation to reduce human error
- More use of GUIs
- Requests for Cluster File System (CFS)
- Vertical integrated solutions
  - IBM's "Redbooks"
  - HP's:
    - Reference Architectures
    - High Availability Reference Architectures

Industry Trends in High Availability

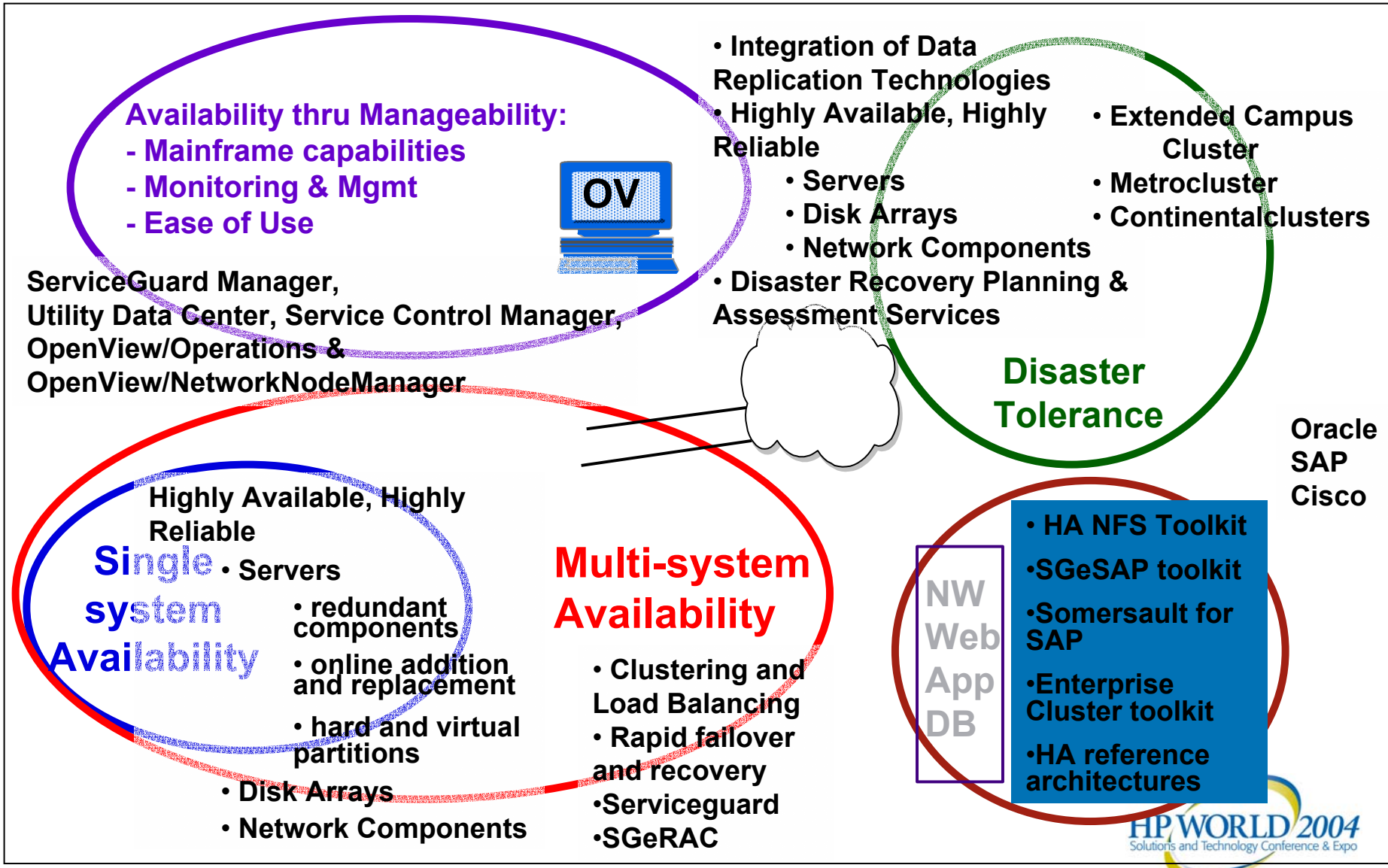
- Vertical scaling continues to increase
- Horizontal scaling is becoming more prevalent
  - Oracle RAC
  - Informix XPS
- Multi-tier architectures
- Use of middleware or applications that may help to preserve transactions in progress
- Multiple availability solutions
  - Clustering
  - Load balancing
  - Duplication of content
- Linux clustering

Industry Trends in Disaster Tolerance

- Movement from Disaster Recovery (days to weeks) to Disaster Tolerance (minutes to hours)
- Specification of Recovery Point Objective (RPO) and Recovery Time Objective (RTO) in SLA
- Dramatic increase in the use of Data Replication
  - Physical replication in hardware or software
  - Logical replication
  - Transaction Processing Managers (TPMs) and Reliable Messaging Middleware
    - BEA Tuxedo for C and C++, BEA WebLogics for Java, IBM WebSphere for Java
    - IBM MQSeries
- Automation versus Manual Intervention
- Network technologies
  - DWDM (dark fiber) – 100 km or more
  - MAN – major cities, looks like an in-house Ethernet network
  - SONET – major cities, switched
  - Long-distance links (OC48, etc.)
- Sarbanes-Oxley act requirements

High Availability Technology Components

HA technology components



High Availability Products

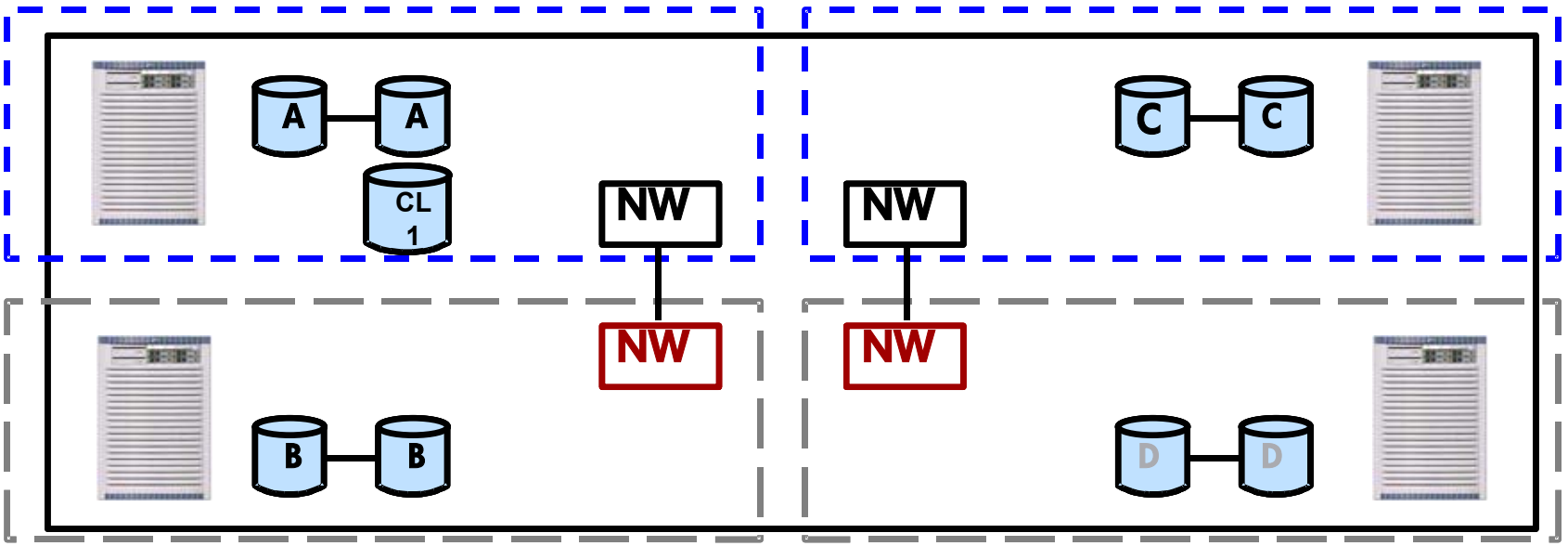
High Availability Products

- Serviceguard
  - Serviceguard Clusters
  - Serviceguard Manager (GUI)
  - HA & Partitions
  - HA & Resource Management with Workload Manager
  - ECM Toolkit
- Serviceguard Extension for Faster Failover (SGeFF)
- Serviceguard Extension for SAP (SGeSAP)
- Serviceguard Extension for RAC (SGeRAC)

Serviceguard

- Single SG Cluster up to 16 nodes
- For use when all nodes are in a single Data Center
- Automatic failover, up to 150 application packages (up to 900 services total); supports up to 200 relocatable package IP addresses per cluster
- Cluster within a single data center
- SCSI or Fibre Channel for disks
- Single IP subnet for heartbeat networks (Ethernet, FDDI, Token Ring) (NOTE: DWDM will provide extended distances for Ethernet)
- IPv6 support (data links only)
  Heartbeat must still use an IPv4 network.
  In SG version 11.15, package IPs may be either IPv4 or IPv6)
- LVM or VxVM, MirrorDisk/UX (for LVM) optional
- Cluster lock disk required for 2 nodes, optional for 3-4 nodes only
- Alternatively can use Quorum Server in place of cluster lock disk with up to 16 nodes

SG cluster (local cluster – shared connectivity)



- All systems are physically connected (cabled) to each disk
- Maximum cluster size is 16 nodes
- Each application runs on only one host at a time
- Failover is possible to any node that is physically connected to the data

Alternative to cluster lock disk: Quorum Server (not member of cluster)

SG cluster (local cluster – shared connectivity)

- Protects against failures of:
  - Hosts
  - Networks
  - Applications (services)
  - User-defined resources
  - OS resource problems (e.g., shared memory)
- Provides:
  - Transparent IP address failover
  - Rolling upgrade for OS and some applications
  - Shared connectivity to single copy of data
  - Integration with Oracle 8i and 9i Server

Failover Models

- Active / Active
  - All nodes are running (different) applications
  - Upon failover, choice of
    - Reduced capacity when multiple applications run on the same node
    - Shutdown less critical applications
    - Optional use of PRM and WLM to guarantee resource entitlements
- Active / Standby
  - One or more nodes are reserved for failover use
  - Upon failover, the applications maintain performance due to spare capacity
- Rotating Standby
  - Upon failover, the standby system becomes the new production system and the repaired system becomes the new standby system

Serviceguard basic unit of work

- Uses a **move the service point** or **multiple service points** paradigm
- The Application Package contains a list of:
  - Servers
    - Prioritized list of one or more nodes where the application is able to run
  - Network Identity
    - Zero or more IP addresses that will failover with the application
  - Data
    - Zero or more disk volume groups (VGs) that the application needs to access (shared connectivity, not shared access)
  - Applications
    - Zero or more application services that are monitored
  - Other Resources
    - Zero or more user-defined EMS resources that the application uses

Serviceguard configuration

The Application Package defines:

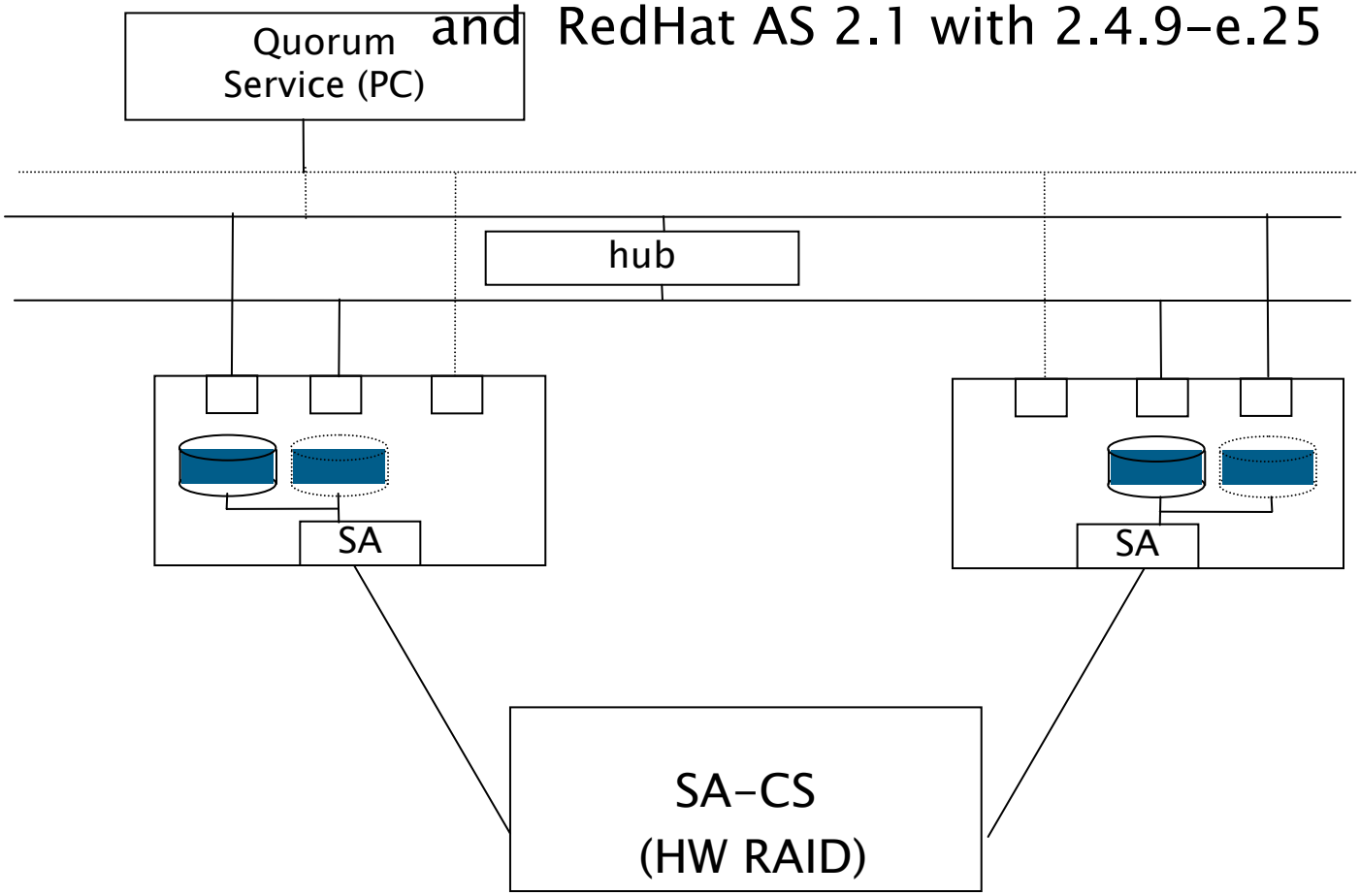
1. How to startup the application
2. How to shutdown the application
3. How to monitor the application
4. What resources are used by the application
   - Nodes
   - Networks
   - Disk Volume Groups
   - Services
   - User-defined resources

Serviceguard for Linux (SG/LX)

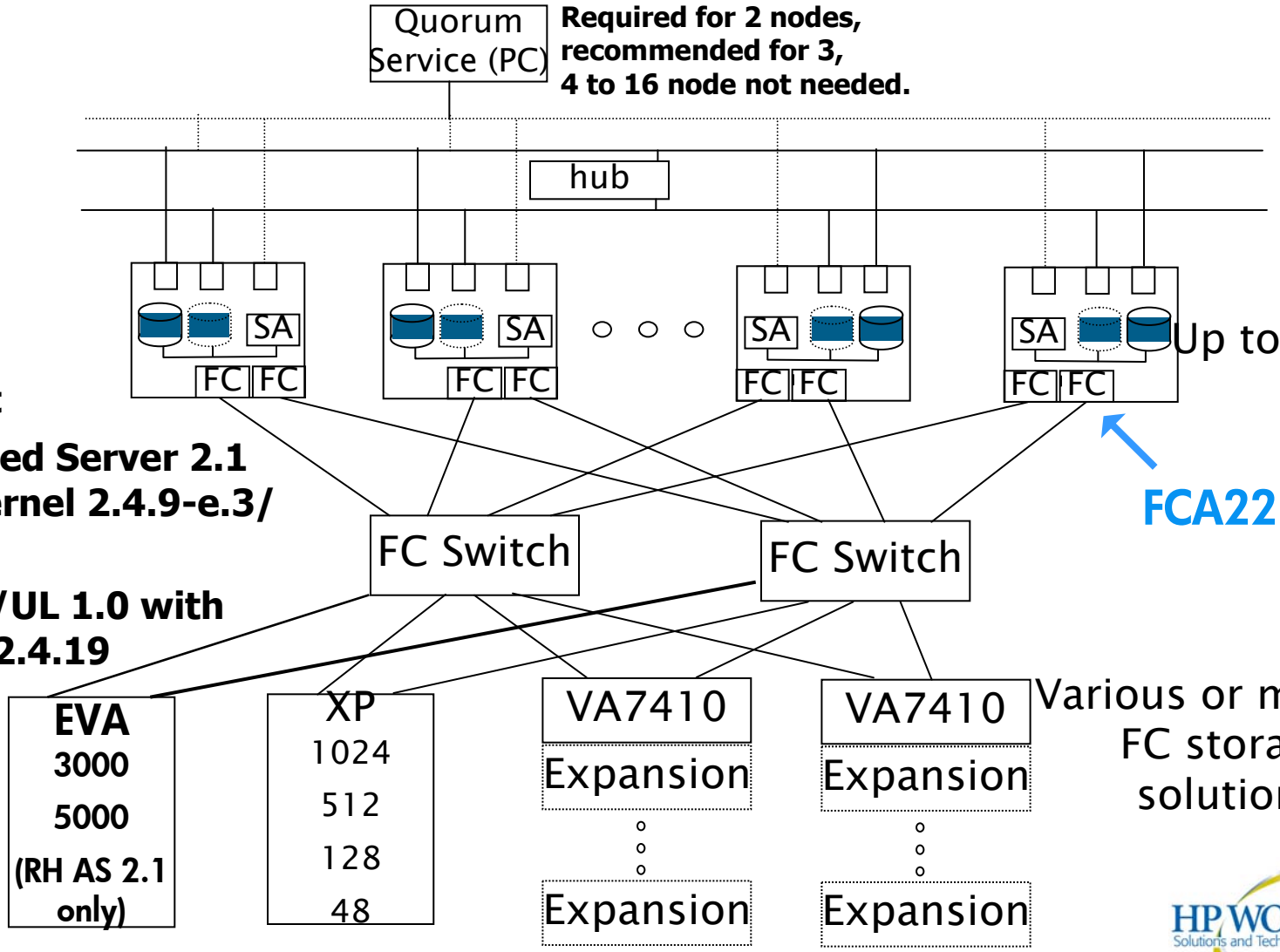
- Package Failover (for node, network, or service)
- Support up to 150 packages and 900 services per cluster (previously up to 30 packages with 30 services per package)
- Heartbeat over Ethernet, supporting up to 7 Heartbeat subnets
- Local Network Switch (Bonding)
- Online Reconfiguration
- Rolling Upgrade Infrastructure
- Quorum Service — supporting maximum of 100 nodes/50 clusters (previously maximum of 2 clusters, 8 nodes)
- Toolkits: NFS, Apache, Samba, SendMail, and Oracle, MySQL
- SG/Mgr (localizable, and capable of monitoring and polling Linux and HP-UX clusters) + administration, multi-subnets and alert summary
- Disk Monitor
- Support reiser file system (a Journaled file system)
- Parallel fsck
- Improved installation process (dynamically loadable modules)

SG Linux SCSI Configuration: Support only 2-node clusters

ProLiant and SA-CS on RedHat 7.3 with 2.4.18-10/-26/-27 kernel and RedHat AS 2.1 with 2.4.9-e.25



SG Linux FC Configuration: Supports up to 16 nodes



RedHat

Advanced Server 2.1 with kernel 2.4.9-e.3/ -e.25,

SLES 8/UL 1.0 with kernel 2.4.19

Protecting against split brain and data corruption (1)

- ServiceGuard uses a "**tie-breaker**" to prevent "**Split-Brain**" of the cluster
- Split-Brain can occur when:
  - a network failure splits the cluster into 2 equal halves

    -OR-
  - exactly half of the servers in the cluster fail all at once
- Causes data corruption if the application runs concurrently on both "halves" of the cluster and modifies the same single copy of the data
- Tie-breaking is implemented either by a lock disk or a quorum server

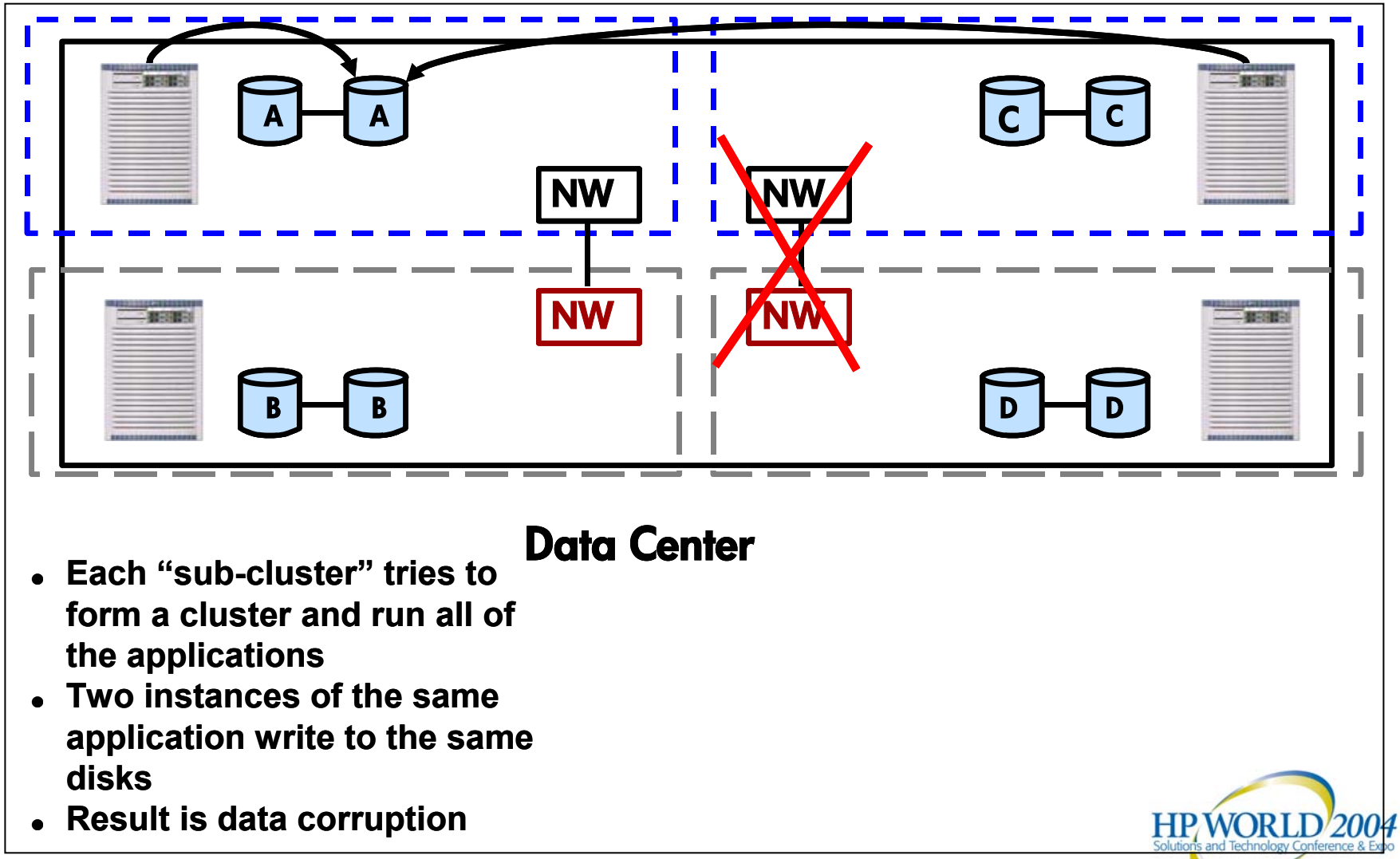
Protecting against split brain and data corruption (2)

- Possible tie-breakers:
  - Cluster lock disk
    - single cluster lock disk (when all servers are in a single data center)
    - dual cluster lock disks (when the servers are distributed across two data centers)
  - Arbitrator systems
    - one or two Arbitrator systems which are fully participating members of the cluster (when the servers are distributed across three data centers)
  - Quorum Server
    - a (small) server that is outside of the cluster

Cluster Lock (HP-UX Only)

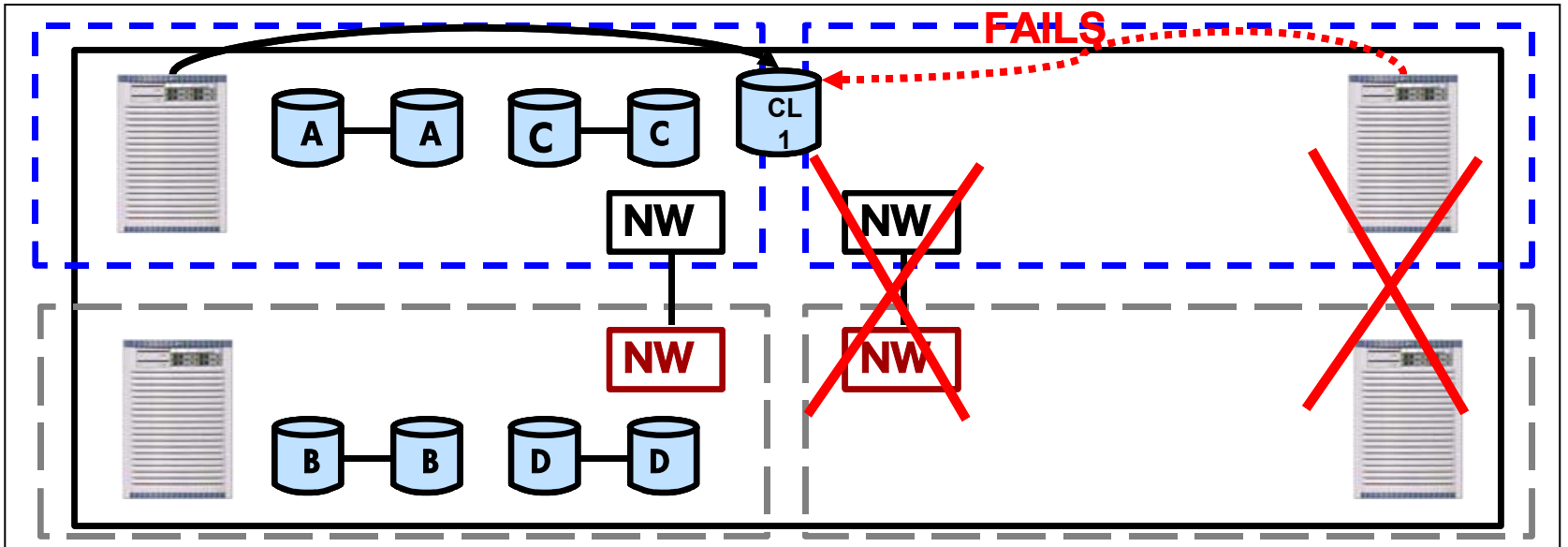
- A special area on an LVM disk located in a volume group that is shareable by all nodes in the cluster
- When a node obtains the cluster lock, this area is marked so that other nodes will recognize the lock as "taken."
- A cluster lock disk for HP-UX can be employed as part of a normal volume group containing user data
- Lock requirements
  - One-node cluster does not require a lock disk
  - Two-node cluster requires a lock disk
  - Three and four-node clusters – lock disk is recommended
  - Greater than four nodes – lock disk is not allowed, however a quorum server may be used

SG cluster – preventing split brain



- Each "sub-cluster" tries to form a cluster and run all of the applications
- Two instances of the same application write to the same disks
- Result is data corruption

SG cluster – preventing split brain (continued)

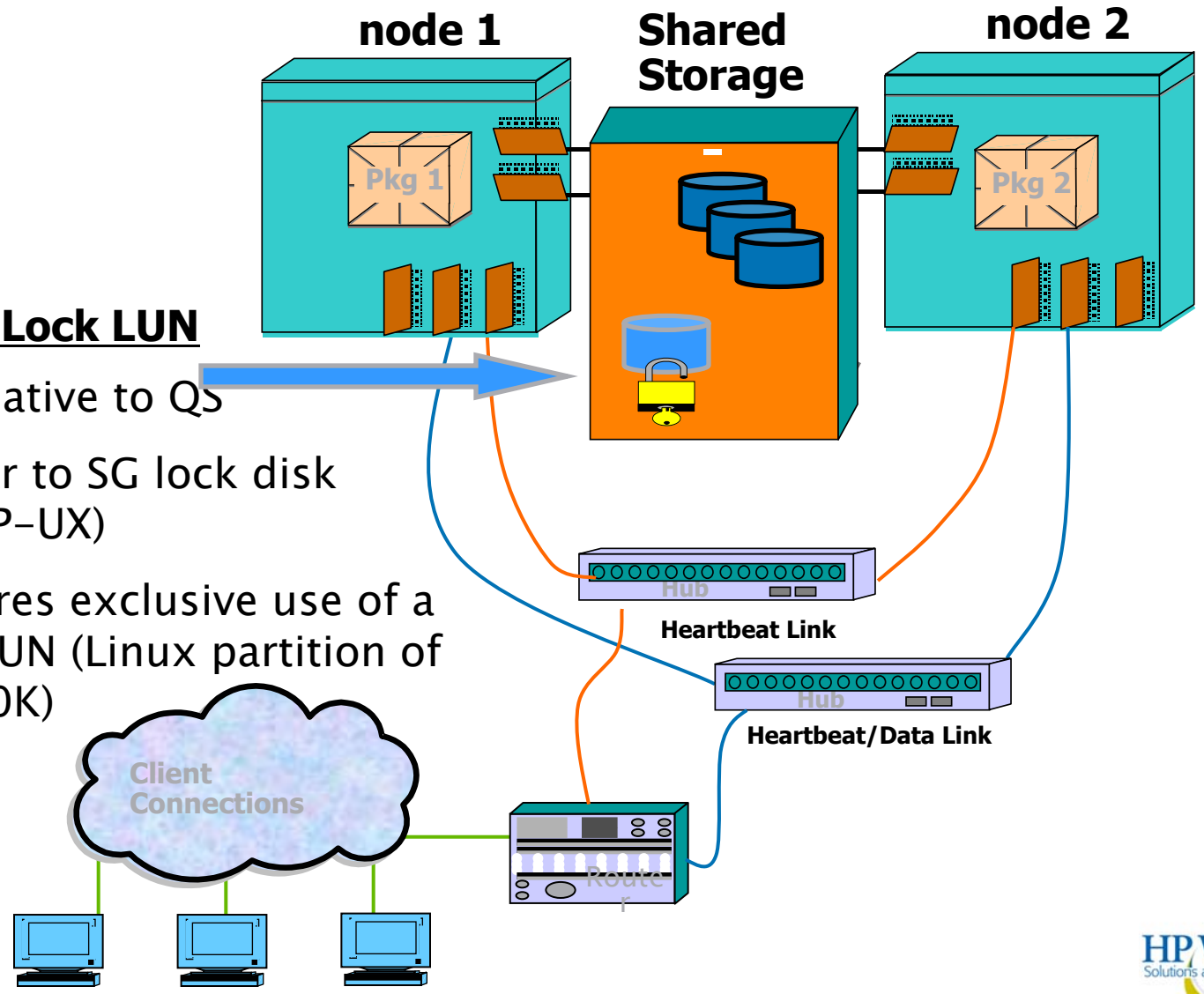


Data Center

- Each "sub-cluster" tries to acquire the cluster lock on the cluster lock disk
- The algorithm guarantees that only one sub-cluster will get it
- One sub-cluster is forced to crash to prevent data corruption !

Alternative to cluster lock disk: Quorum Server (not member of cluster)

Cluster Lock LUN
(for SG/LX clusters)



Cluster Lock LUN

- Alternative to QS
- Similar to SG lock disk (on HP-UX)
- Requires exclusive use of a single LUN (Linux partition of size 100K)

Quorum Service (QS) (HP-UX and Linux)

- Alternative cluster arbitration method
- Supports up to 50 clusters and maximum of 100 nodes
- TCP/IP network connection required
  - (Not required to be in the same subnet, although recommended to minimize network delays)
- Stand-alone HP-UX or Linux-based server outside of the SG cluster
- Runs as a real-time process
- The Quorum Service (QS A.02.00)
  - can be configured in a package in a cluster
  - cannot reside in the same cluster that uses it
  - do not configure two clusters that use the same Quorum Service package

QS Example Configuration

Clusters:

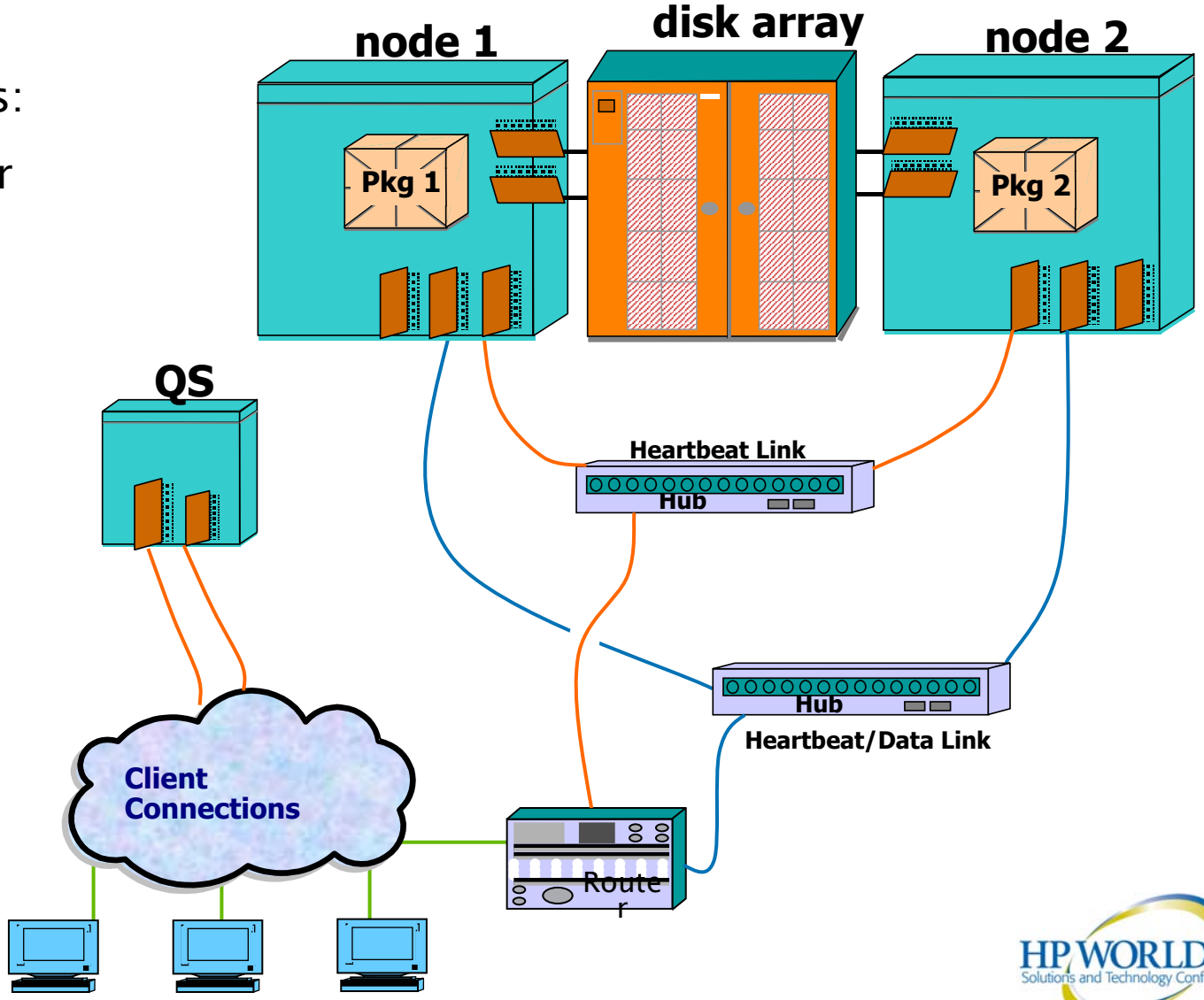
Linux or

HP-UX

QS:

Linux or

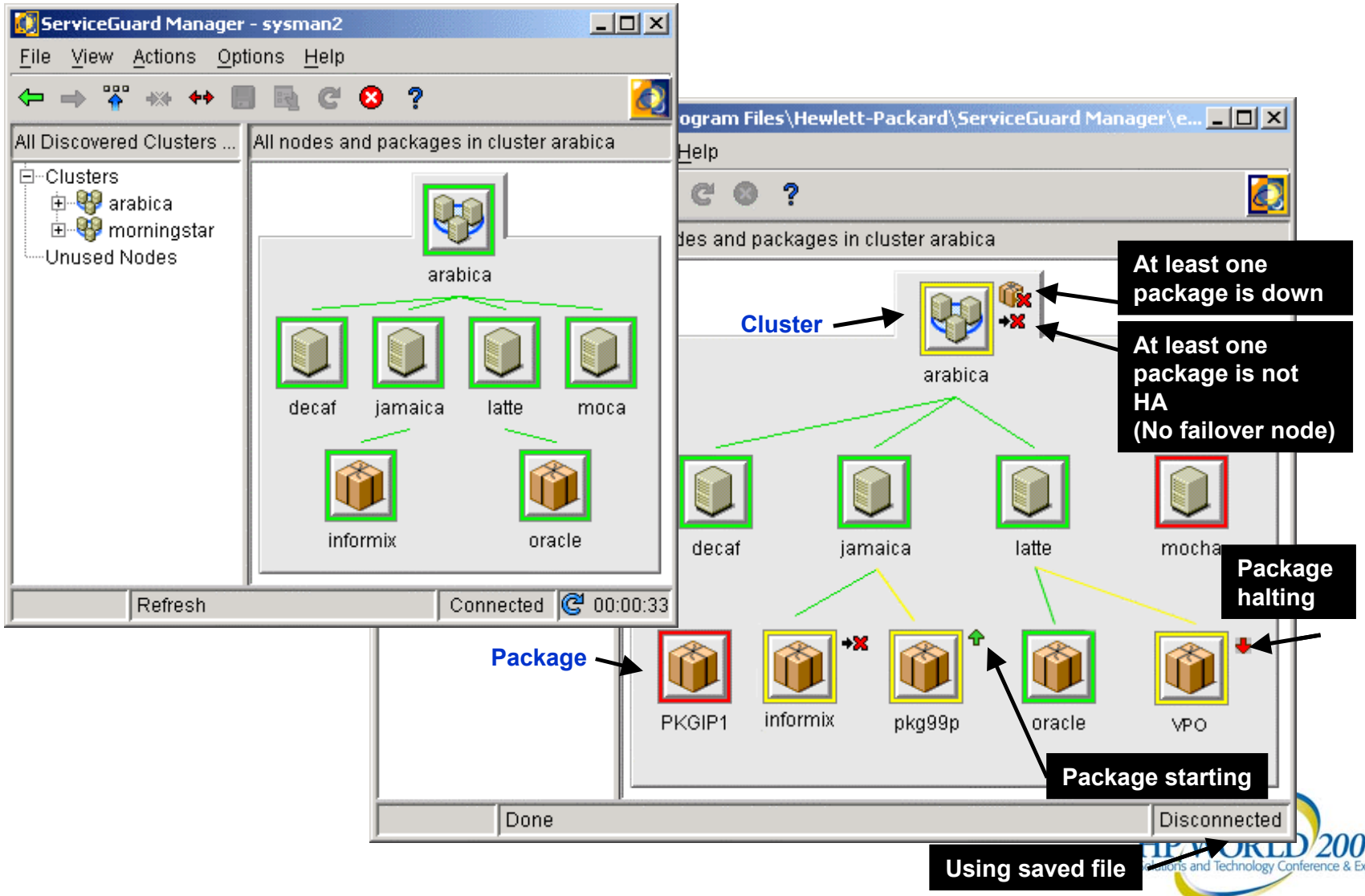
HP-UX



Serviceguard Manager

- An intuitive and easy-to-use Java©-based GUI to display HP Serviceguard clusters on both HP-UX and Linux
  - Integrated with Service Control Manager and OpenView/Operations & NNM
  - Uses color-coded icons to show status information about a cluster, node or package
  - Provides GUI for:
    - Monitoring
    - Administration / management (start/stop/move)
    - Configuration (cluster and package)
  - Role-based access rules for segregating users – Example:
    - Operators can only monitor the clusters
    - Level 2 Support can start/stop/move application packages
    - Administrators have full capabilities
  - It's free !
  - New version planned for integration with System Insight Manager on HP-UX 11i v3

Serviceguard Manager example

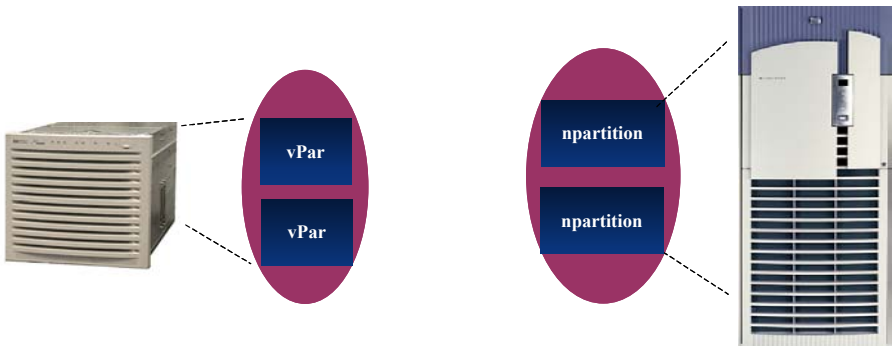


HA and Virtual Server Environment

- Serviceguard integrates with the Virtual Server Environment (VSE):
  - nPARs (hard partitions with electrical isolation)
  - vPARs (soft or virtual partitions with software isolation)
  - Workload Manager (WLM) - automatically adjust resource (CPU, Memory, Disk) allocation in normal and failover situations

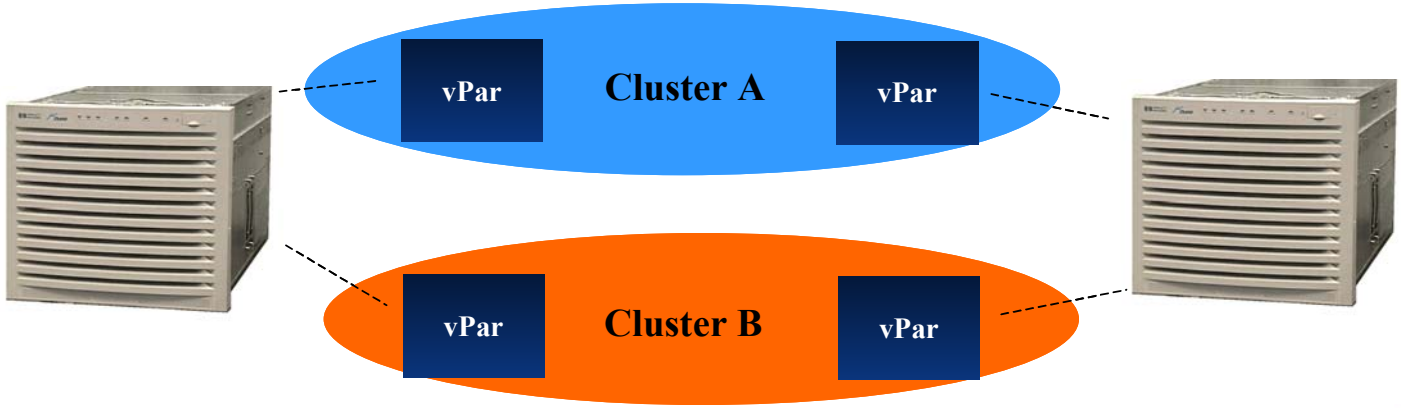
Cluster configuration with partitioned servers

- ServiceGuard "cluster-in-a-box" is supported. Partitions of a cluster are wholly contained within the same hardware cabinet. While this configuration is subject to single points of failure, it may provide adequate availability characteristics for some applications.
- Note: "cluster-in-a-box" consisting exclusively of vPars is more likely to suffer a complete cluster failure than one made up exclusively of nPartitions. nPartitions provides hardware and software isolation where as vPars provides software isolation.
- See "ServiceGuard Cluster Configuration for Partitioned Systems" whitepaper at http://docs.hp.com/hpux/ha22



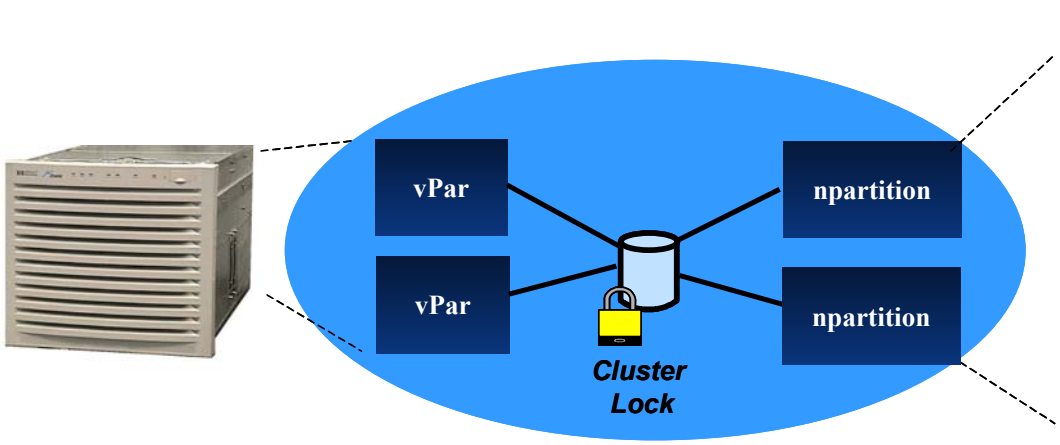
Cluster configuration with partitioned servers (continued)

- To configure a cluster with no SPOF, the membership must extend beyond a single hardware cabinet and comply to the standard Serviceguard configuration rules.
- The cluster must be configured such that a failure of a single hardware cabinet does not result in the failure of a majority of the nodes in the cluster.
- Separate partitions within a physical server can be configured as part of different Serviceguard clusters.



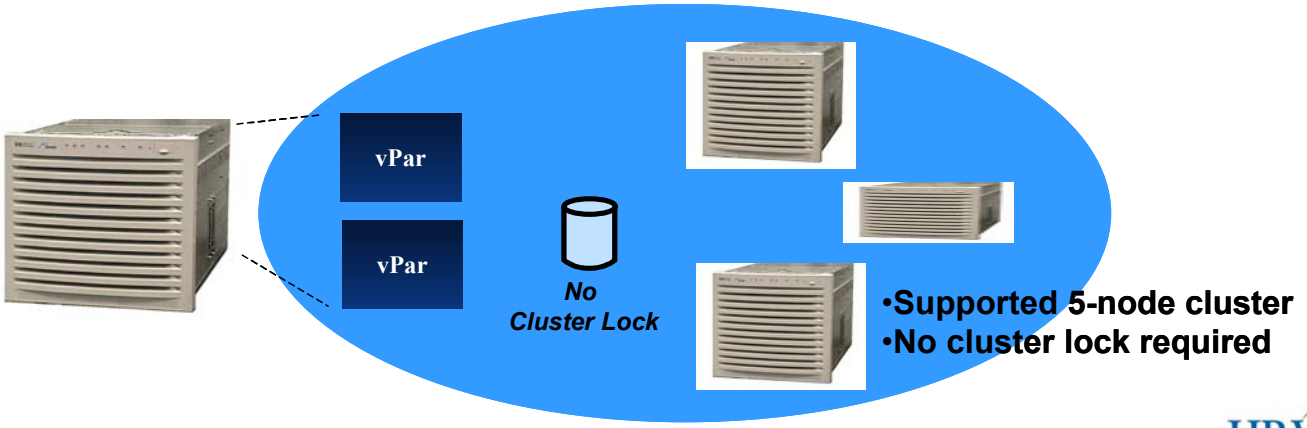
Cluster lock considerations with partitioned servers

- A cluster lock is required if the cluster is wholly contained within two hardware cabinets. This requirement is due to a possible 50% cluster failure.
- Serviceguard only supports cluster lock up to four nodes. Thus, cluster partitions configured within two hardware cabinets is limited to four nodes.
- Cluster lock must be powered independently of either hardware cabinet.
- You must evenly divide nodes between two hardware cabinets configured with partitions (i.e. 3 and 1 is not a legal 4-node configuration).

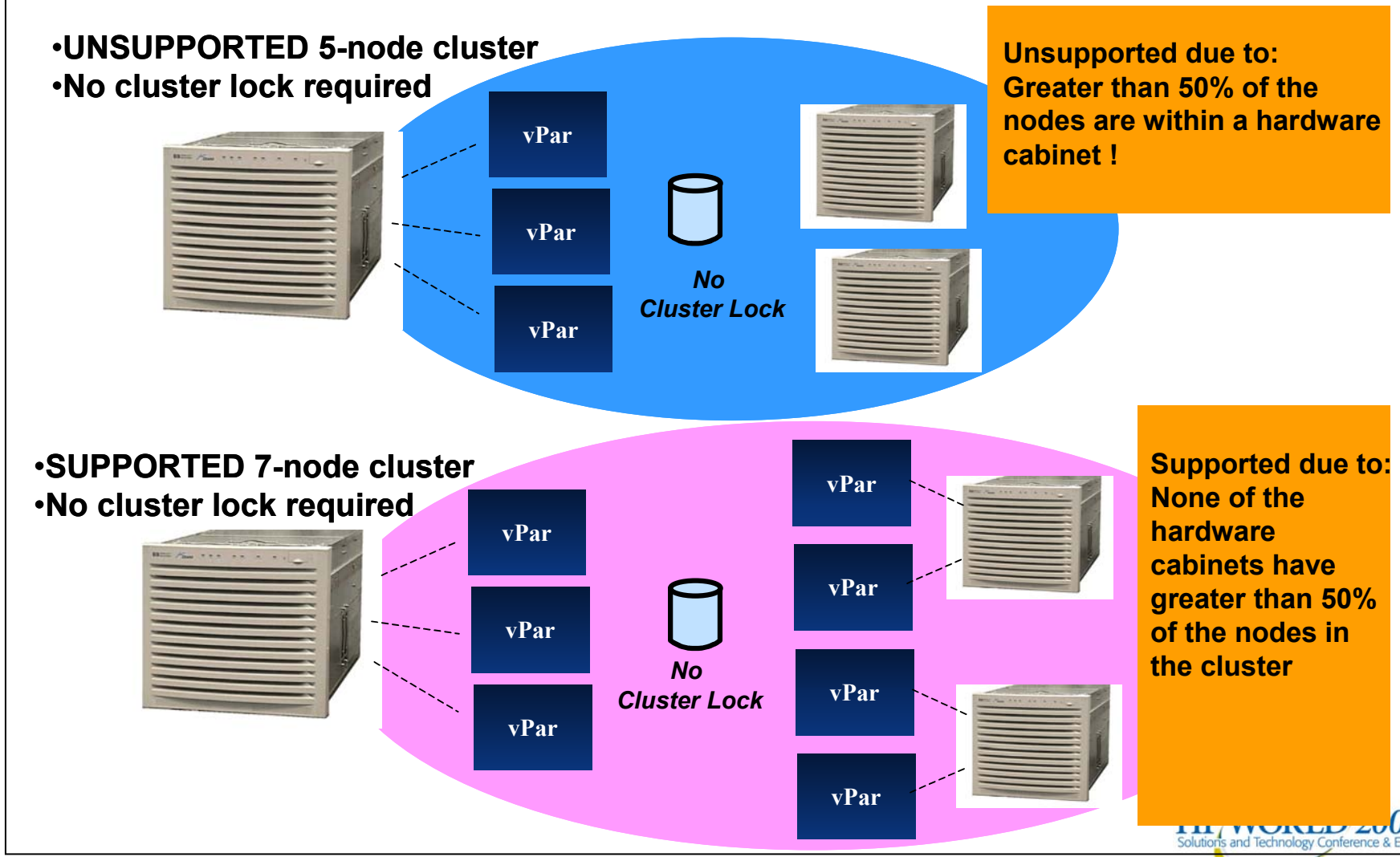


Cluster quorum with partitioned servers

- Care must be taken to configure an even or greater number of nodes outside a hardware cabinet with partitions.
- If half the nodes of a cluster are within a hardware cabinet, a cluster lock is required (4-node maximum cluster size).
- If more than half the nodes of a cluster are outside a hardware cabinet with partitions, no cluster lock is required (16-node maximum cluster size).
- Without a cluster lock, beware of configurations where the failure of a hardware cabinet will cause the remaining nodes to be 50% or less quorum – the cluster will halt.



Cluster quorum with partitioned servers (continued)



Workload Manager (WLM) and Process Resource Manager (PRM) with SG

- Service Level Management solutions that can be easily integrated into an Serviceguard environment to provide management of computing resources (CPU, memory and I/O utilization) according to desired SLOs both before and after application failover among nodes

- Applications can be assigned maximum resources when they run on their "preferred" server.

- After a failover to a different node, the resource utilization for the existing and new applications can be dynamically reset to:
  - favor important applications
  - restrict resource utilization of less important applications

Components of Failover Time

- Node Failover Time
  1. failure detection
  2. node selection
  3. node failover

Serviceguard-Specific

- Application Failover Time
  1. volume group activation
  2. file system recovery (if applicable)
  3. file system mount (if applicable)
  4. application startup
  5. application recovery

Application-Specific

(Not affected by SG)

HP Serviceguard Extension for Faster Failover (SGeFF)

NEW ! – July 2004

- Optimized Serviceguard and Serviceguard Extension for RAC environments may achieve improved node failover times of 5 seconds (current failover 30-45 seconds)
- Expected to be used with
  - An application or a database that has fast (or no) startup time
  - An application that has fast recovery (tens of seconds at most), e.g., Oracle RAC
  - Systems with predictable loads
- Restricted Configuration
  - SG A.11.16 and later
  - 2-node cluster only
  - Multiple heartbeat networks
  - Requires quorum server as tie-breaker
  - LVM, SLVM or VxVM (not CVM)
  - Need to qualify network latency over time

Typical applications for SGeFF

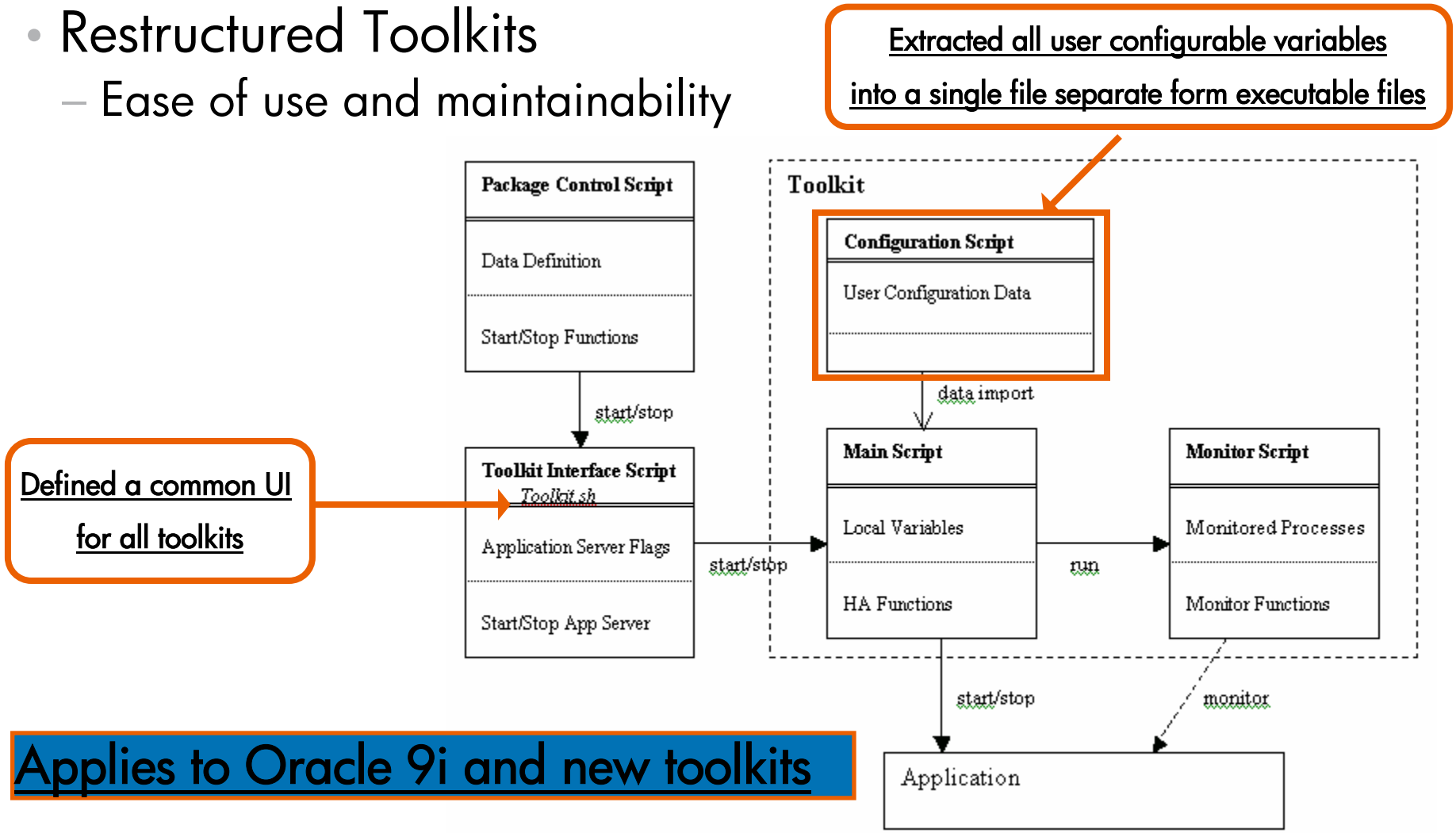
- Oracle RAC:
  - The RAC instances are already running on all of the nodes, so there is no application startup time
  - RAC can be tuned for rapid recovery, often in 10-60 seconds
- Other application factors that are well-suited for SGeFF include:
  - An application where there is no disk-based database
  - An application where you can have the processes already started, but idle and waiting until a failover occurs (i.e., Hot Standby) such as in some Telco applications
  - Systems with fewer volume groups
  - Applications using raw volumes rather than file systems -OR- systems with fewer file systems
  - VxFS file systems with the intent log enabled at least for meta-data (note: this is the default for VxFS, and it should not be changed)

Toolkits

- Enterprise Cluster Master Toolkit (ECMT)
  - Fully-tested and supported collection of integration templates for certain popular third-party applications
- HA NFS Toolkit
  - Pre-tested and supported templates to make NFS servers highly available
- SGeSAP – Serviceguard Extension for SAP

Enterprise Cluster Master Toolkit
version B.02.00

- Restructured Toolkits
  - Ease of use and maintainability



Serviceguard extensions for SAP (SGeSAP)

Integrate SAP R3 with:

- Serviceguard
  - Toolkit template for easily configuring SAP with Serviceguard (HP-UX and Linux)
  - Options to on how to configure the Central Instance (CI) and the Database (DB) servers
- Metrocluster (HP-UX only)
  - Optional template to create a disaster tolerant architecture for SAP

(Configuration example shown in the Metrocluster section of this presentation)

SGeSAP Feature Comparison

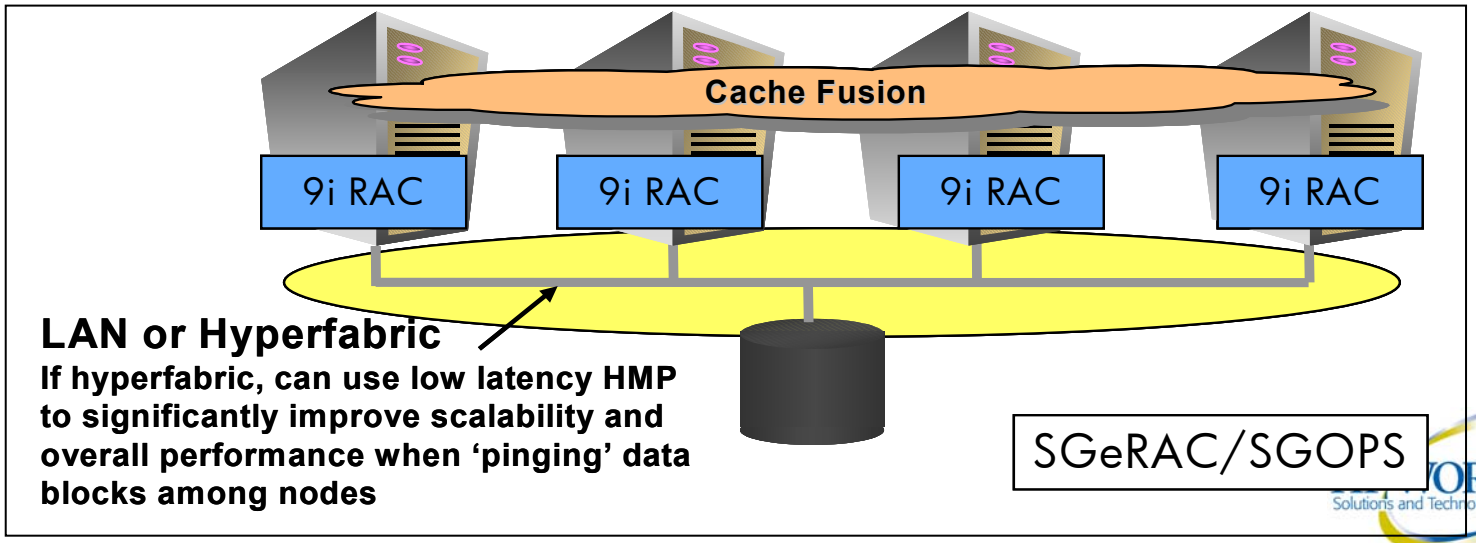
Feature	SGeSAP/UX B.03.09 on PA-RISC B7885BA)	SGeSAP/UX B.03.09on IPF (T2357BA)	SGeSAP/LX A.01.00 on IA-32 (T1227AA)	SGeSAP/LX on IPF (T2392AA)
Supported SAP kernels	3.0x, 3.1x, 4.0x, 4.5x, 4.6x, 6.00,6.10,6.20 6.30,6.40	4.6x, 6.00,6.10,6.20 6.30,6.40	4.6x 6.00, 6.10, 6.20 6.30,6.40	4.6x, 6.00, 6.10, 6.20 6.30,6.40 FY04
Supported database technologies	Oracle SAPDB Informix, UDB (DB/2)	Oracle SAPDB	Oracle SAPDB	Oracle SAPDB
Central System setup support (dbci packages)	X	X	X	Q2/04
DEV/QA/AppServer System shutdown on failover nodes	X	X	X	Q2/04
Mutual failover support (db, ci packages)	X	X	X	Q2/04
Application Server Instance packaging (APP pkgs)	X	X		
SAP System consolidation on cluster nodes	X	X		
Livecache integration (LC packages)	X	FY04		
HP Somersault support	(X)			
SAP Replicated Enqueue Support	support with SAP kernel 6.40	support with SAP kernel 6.40	support with SAP kernel 6.40	support with SAP kernel 6.40

ServiceGuard Extension for RAC (SGeRAC)

- Same protection & functionality for applications as Serviceguard
- Parallel database environment for increased availability and scalability with Oracle Real Application Cluster (RAC)
- Up to 16 nodes with Shared Logical Volume Manager (SLVM) and 4 nodes with Cluster Volume Manager (CVM)

What is SGeRAC?

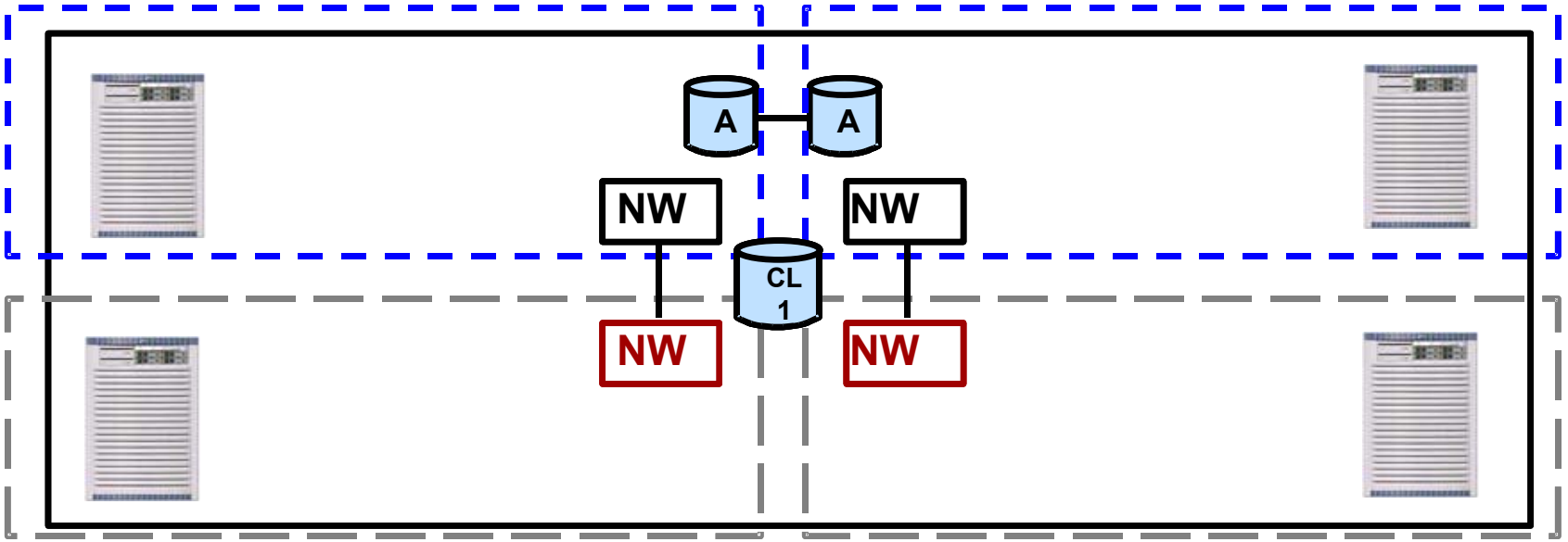
- SG Extension for RAC (SGeRAC) is the repackaged product which currently has the name "SG OPS Edition"
- Allows Oracle RAC/OPS customers to use the Mission Critical Operating Environment (MCOE)
- SGeRAC provides clustering service to Oracle 9i Real Application Cluster (RAC) and 8i Parallel Server (OPS) to meet high availability and scalability requirements of modern enterprise mission critical computing



SGeRAC (Local Cluster – Shared Access)

- Protects against failures of:
  - Hosts
  - Networks
  - Applications (services)
  - User-defined resources
  - OS resource problems (e.g., shared memory)
- Provides:
  - Transparent IP address failover
  - Rolling upgrade for OS
  - Shared access to single copy of data
  - Integration with Oracle Parallel Server 8i / Oracle Parallel Failsafe 8i (OPFS) / Oracle Real Application Cluster (RAC) 9i / RACGuard
  - RACGuard or OPFS provide monitoring of Oracle application hangs and other errors

SGeRAC (local cluster – shared data)



- **All systems are physically connected (cabled) to each disk**
- **Maximum cluster size is 16 nodes with SLVM and 8 nodes with CVM**
- **The same application (OPS, OPFS, RAC 9i) runs concurrently on all nodes**
- **All nodes have concurrent read/write access to the data**
- **Writes are controlled through a distributed lock manager (DLM)**
- **Users are usually distributed among the nodes with a logical partitioning scheme to minimize lock contention**
- **Failover of the user connections is possible among the nodes**

Alternative to cluster lock disk: Quorum Server (not member of cluster)

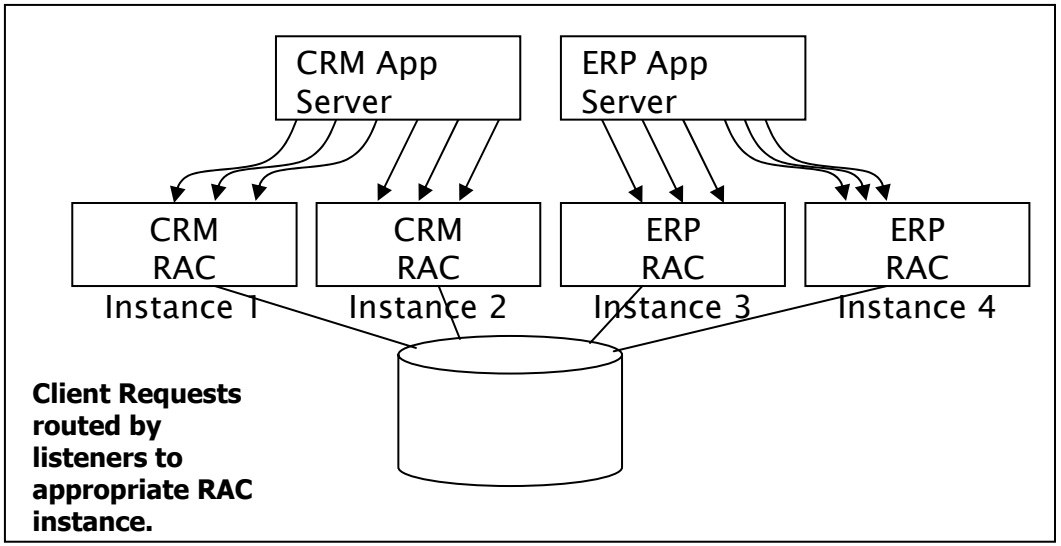
Serviceguard and Oracle 10g RAC

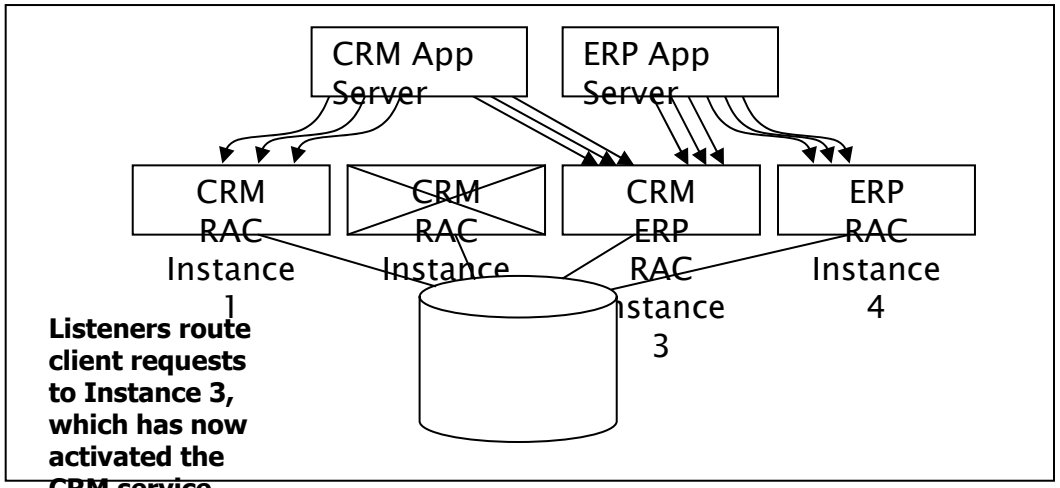
- Oracle Real Application Clusters (RAC) 10g is an option to Oracle 10g Enterprise Edition
- Differences from the previous Oracle9i RAC product include:
  - Integrated clusterware (Cluster Ready Services - CRS)
  - Application Storage Management (ASM) software can be used to manage storage for the RAC database
  - Performance improvements
  - Zero downtime patches for Oracle RAC environments
- NOTE: the use of CRS is OPTIONAL; you can continue to use SGeRAC

Oracle 10g Cluster Ready Services (CRS)

- Introduced with 10g, CRS provides:
  - A standard cluster interface on all platforms for high availability and workload management
  - Management of cluster DB functions, including:
    - Node membership (connectivity)
    - Group services (messaging and locking)
    - Global resource management for Oracle processes
    - High availability (cluster control and recovery)
- Services are created for each application or for major components within complex applications as a method for controlling availability and workloads

Example RAC 10g cluster with two services





Oracle 10g Application Storage Management (ASM)

- Software that can be optionally used to
  - Manage storage for the RAC database
  - Collects LUNs together
  - No file system or volume manager
  - No cluster file system (CFS)

- Use of an HP volume manager and file system
  - Will still be required for non-RAC-database files
  - Can still be used for RAC database files
  - Will be needed if the HP-UX 11i v3 Cluster File System will be used
  - Allows a single tape backup solution

SGeRAC with RAC 10g

- Increases availability for non-RAC processes and applications running on nodes within the RAC cluster
  - since CRS focuses only on Oracle-specific processes and resources
- Is required when using a volume manager
  - Greater flexibility when working with LUNs
  - Features such as mirroring, striping, etc.
  - SLVM or CVM provides concurrent access to the same storage by multiple nodes
- Provides reliable node membership information through
  - Tight kernel integration
  - Real Time priority execution
- Improves network reliability through monitoring and failover management of
  - User LAN
  - Cluster interconnect
  - IPv6 networks

WLM with Oracle RAC 10g

- HP-UX WLM can be used to supplement the workload management capabilities of CRS:
  - CRS manages workloads of only Oracle RAC-related processes
  - WLM can manage resource entitlements of non-Oracle processes and applications
    - Based on defined SLAs and relative priorities within a server's partitioning schemes (e.g., vPARs, nPARs, and psets)
    - Automatically manages Pay Per Use and Temporary iCOD processors for temporary capacity increases
    - For example, WLM can be used to manage resource entitlements between development, test and production environments

Serviceguard and Oracle 10g

- There are several Technical Whitepapers being written that are expected to be available later in 2004
- These whitepapers are intended to help HP field people and customers to choose the best HA solutions for their Oracle 10g environment

Disaster Tolerant Products

Extended/Campus Cluster, Metrocluster and Continentalclusters

Disaster Tolerant Solutions Require Data Replication

- Defined as a scheme by which data is copied from one site to another for disaster tolerance
- Required in disaster tolerant architectures
- Physical Replication
  - Hardware
  - Software
- Logical Replication
  - File system
  - Database
- Issues
  - Data consistency
  - Data currency
  - Data recoverability
  - Data loss

Extended/Campus Clusters

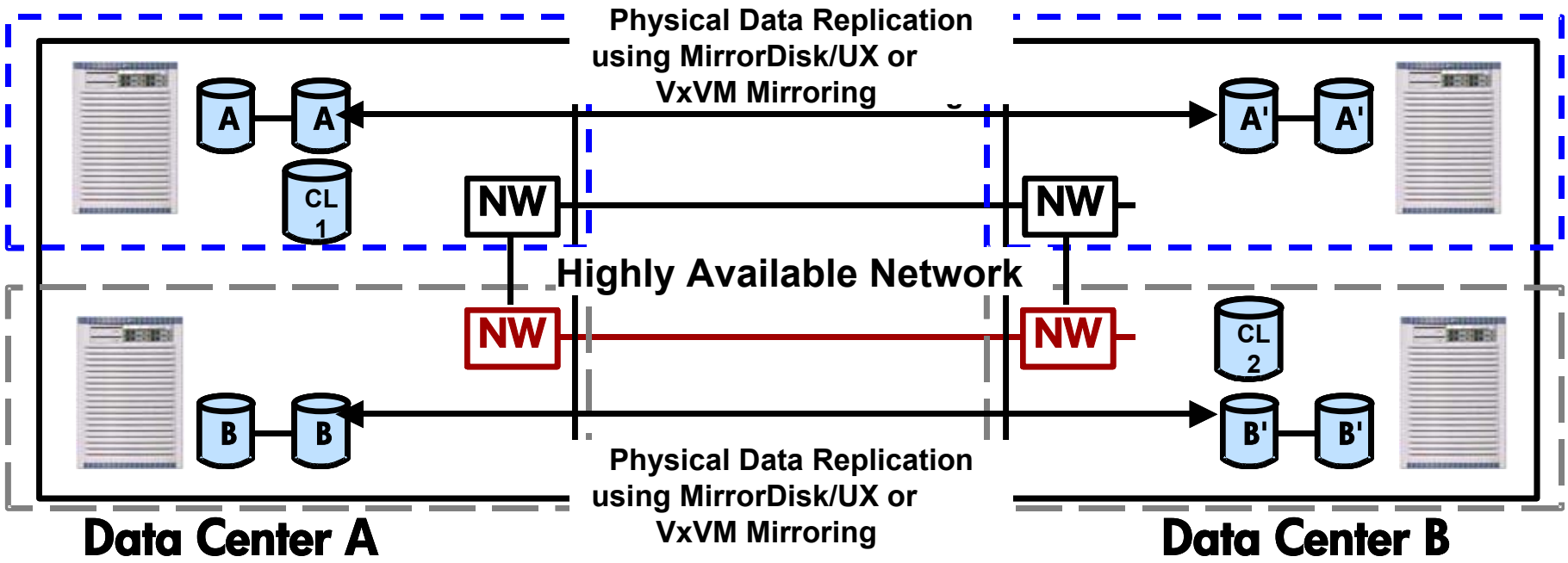
- Single SG Cluster up to 4 nodes with 2 data centers, 16 nodes with 3 data centers (NOTE: 4 nodes with CVM near future)
- Automatic failover, up to 150 application packages
- Campus and city distances
  - Between 2 data centers (up to 10 km)
  - Among 3 data centers
    - Up to 10 km between A & B with FibreChannel Hubs
    - Up to 50 km with FDDI networking & Finisar Long Haul FibreChannel GBICs
    - Up to 100 km using DWDM/CWDM
- Fibre Channel for disks
- Single IP subnet for heartbeat networks (Ethernet or FDDI)
- Dedicated links for network and physical data replication (dark fiber) –
  - no telco switched networks
  - no conversion to other protocols like ATM
- MirrorDisk/UX or Veritas Mirroring REQUIRED
- LVM or VxVM for Serviceguard; SLVM (2-nodes) or CVM (4-nodes) for SGeRAC
- Dual cluster lock disks (LVM) required for up to 4 nodes with 2 data centers
- QS or Arbitrators with 3-data center architecture

Extended/Campus Cluster Architectures

- Two data center architecture
  - Limited to 4 nodes
  - Requires dual cluster lock disks
  - Minute risk of split brain
  - Physical data replication using MirrorDisk/UX

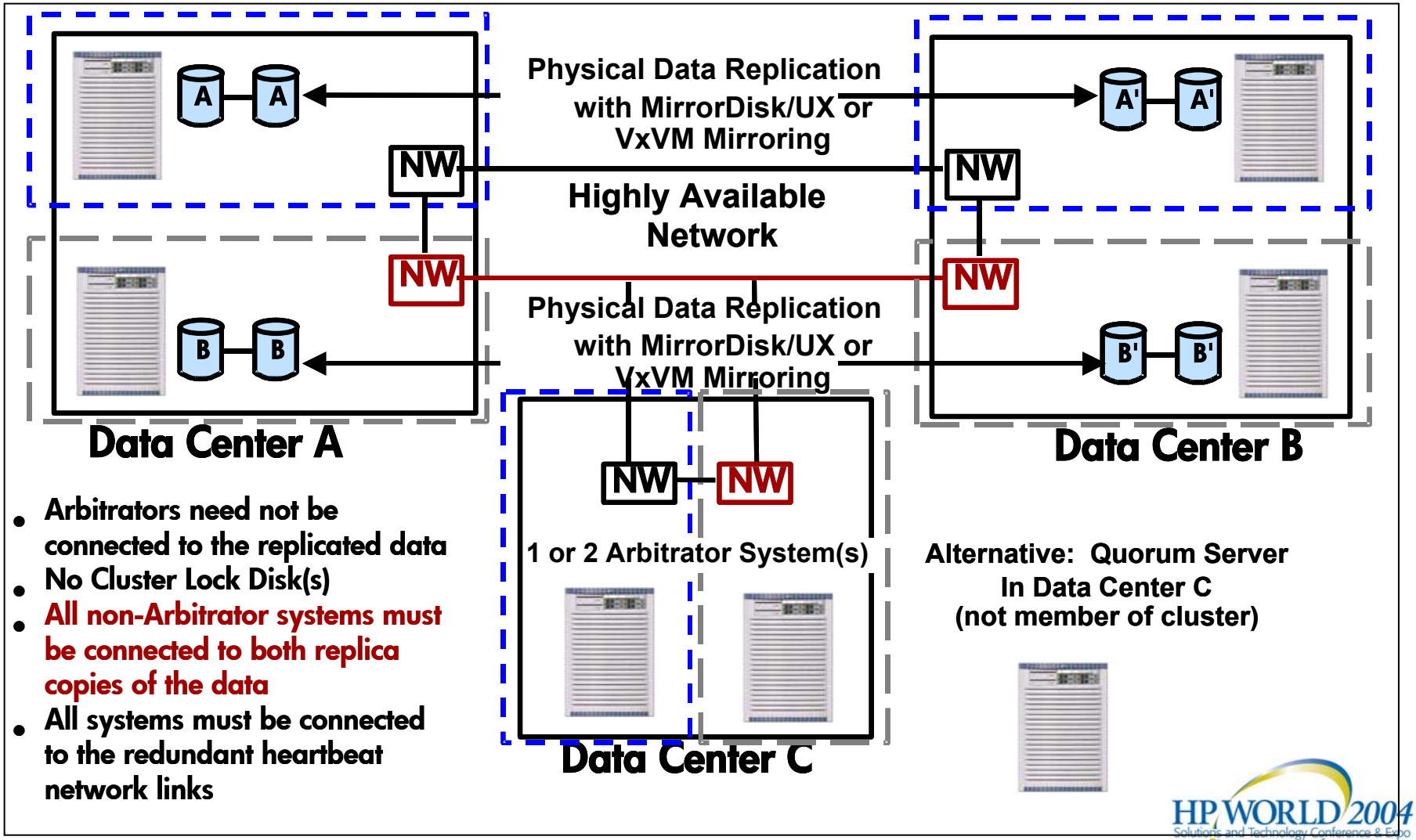
- Three data center architecture
  - Limited to 16 nodes
  - Cluster lock disks not used
  - Arbitrators or Quorum Server act as tie breakers
  - No chance of split brain
  - Physical data replication using MirrorDisk/UX or Veritas Mirroring

Extended Campus Cluster (2 data centers)



- All systems must be connected to both cluster lock disks
- failover systems must be connected to both mirror copies of the data
- all systems must be connected to the redundant heartbeat network links
- MUST have dual cluster lock disks
- MAXIMUM cluster size is currently 4 nodes when using cluster lock disks

Extended Campus Cluster (3 data centers)



- Arbitrators need not be connected to the replicated data
- No Cluster Lock Disk(s)
- All non-Arbitrator systems must be connected to both replica copies of the data
- All systems must be connected to the redundant heartbeat network links

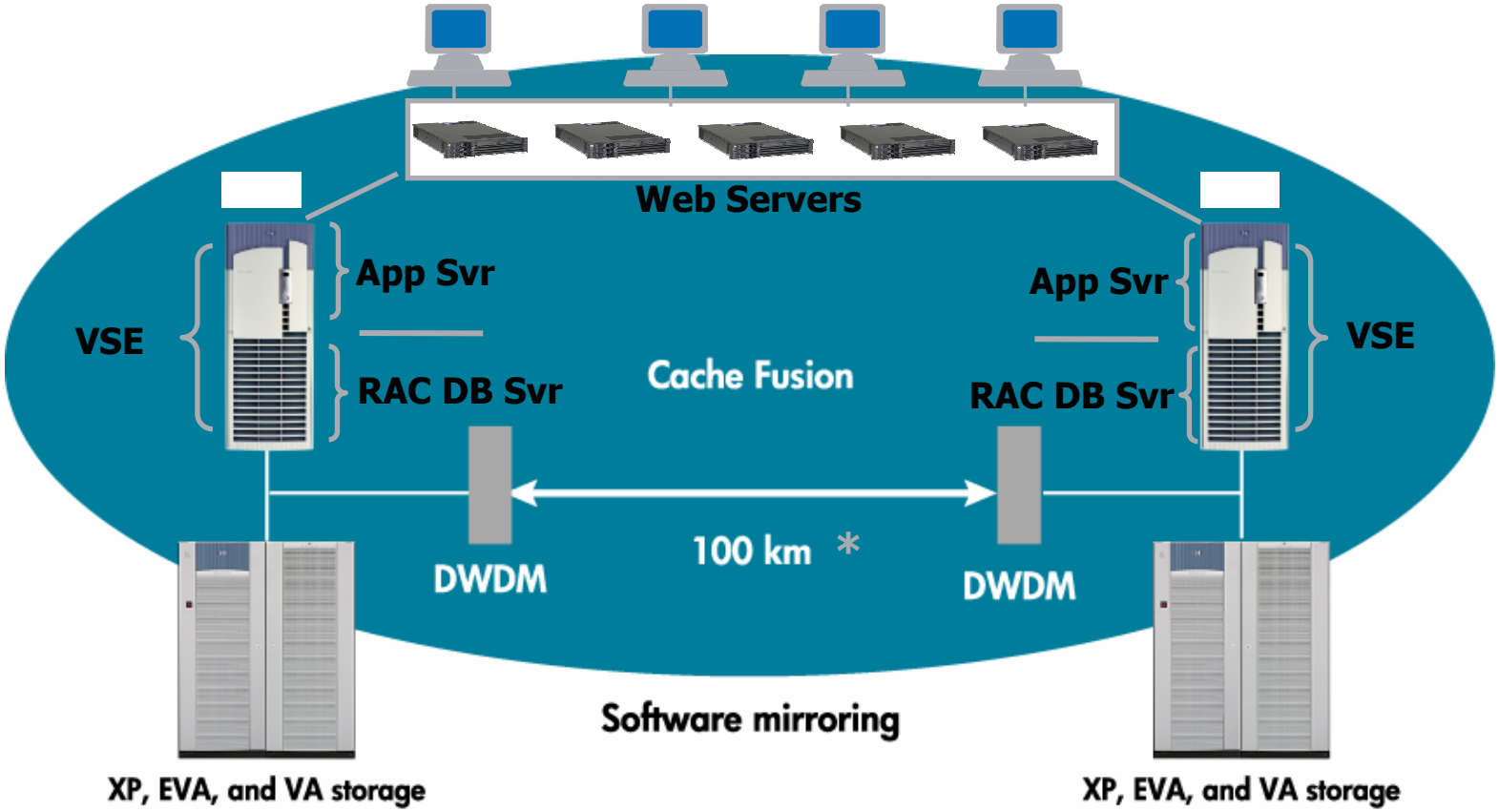
Extended/Campus Cluster Summary of Capabilities (Part 1)

Topology	Extended Campus Cluster		SGeRAC		
Distance	100 km	10 km	Local Cluster – Same Data Center		
Volume Manager	LVM	VxVM	SLVM	CVM	
Data Replication	MirrorDisk/UX	VxVM Mirroring	MirrorDisk/UX	CVM Mirroring	
Storage	VA XP EMC EVA	VA XP EMC	VA XP EMC EVA	VA XP EMC	
Cluster Size in Nodes (HP-UX 11i)	16	16	16 With SLVM patch	4	

Extended/Campus Cluster Summary of Capabilities (Part 2)

Topology			**SGeRAC Stretched**		**SGeRAC Stretched**
Distance			**100 km**		**10 km**
Volume Manager			**SLVM**	**CVM**	**CVM**
Data Replication			**MirrorDisk/UX**	**CVM Mirroring**	**CVM Mirroring**
Storage			**VA XP**	**VA XP**	**VA XP EMC**
Cluster Size in Nodes (HP-UX 11i)			**2**	**4**	**2 or 4**

HP Extended Cluster for RAC
Disaster tolerance with continuous application availability

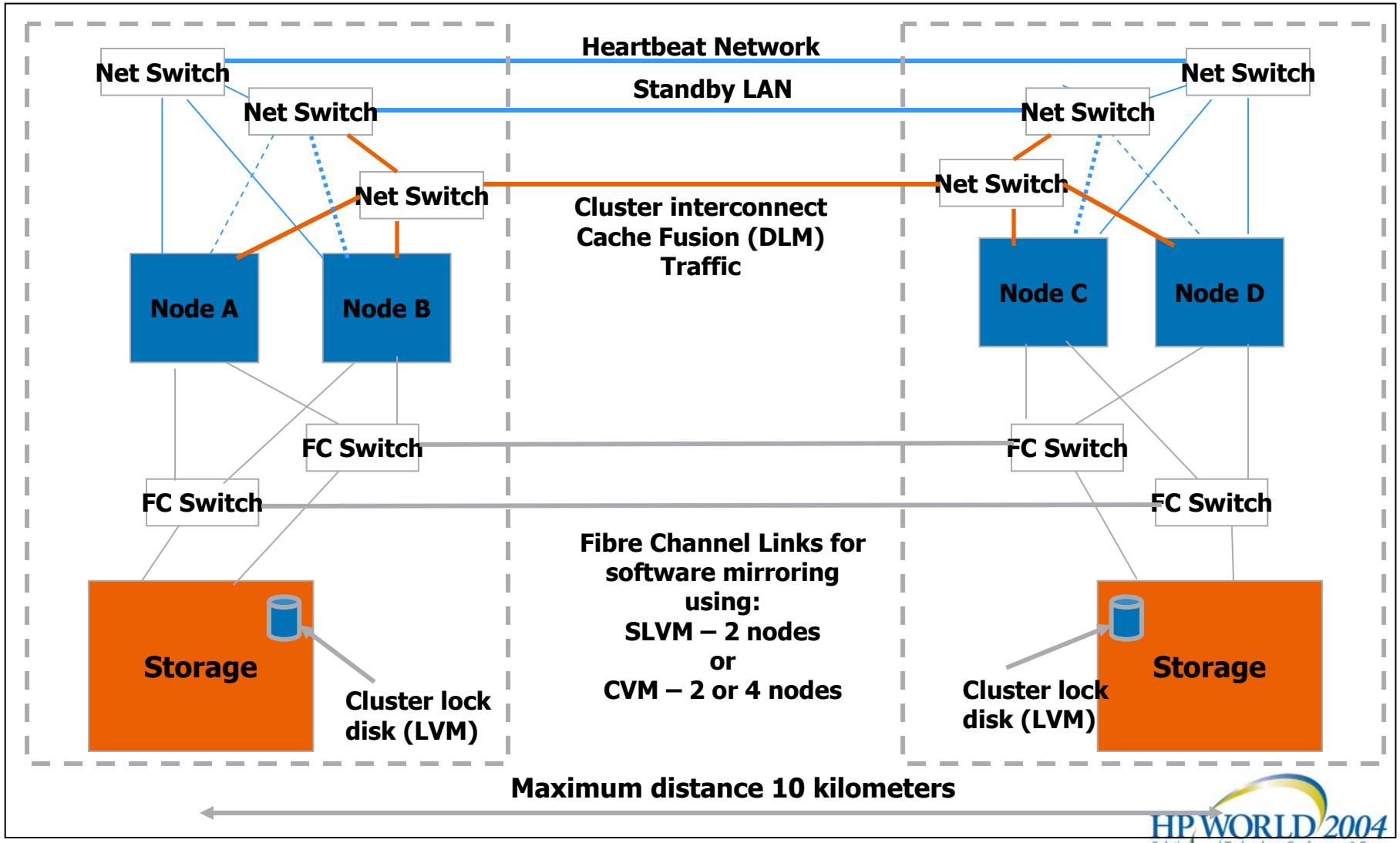


Extended Cluster for RAC

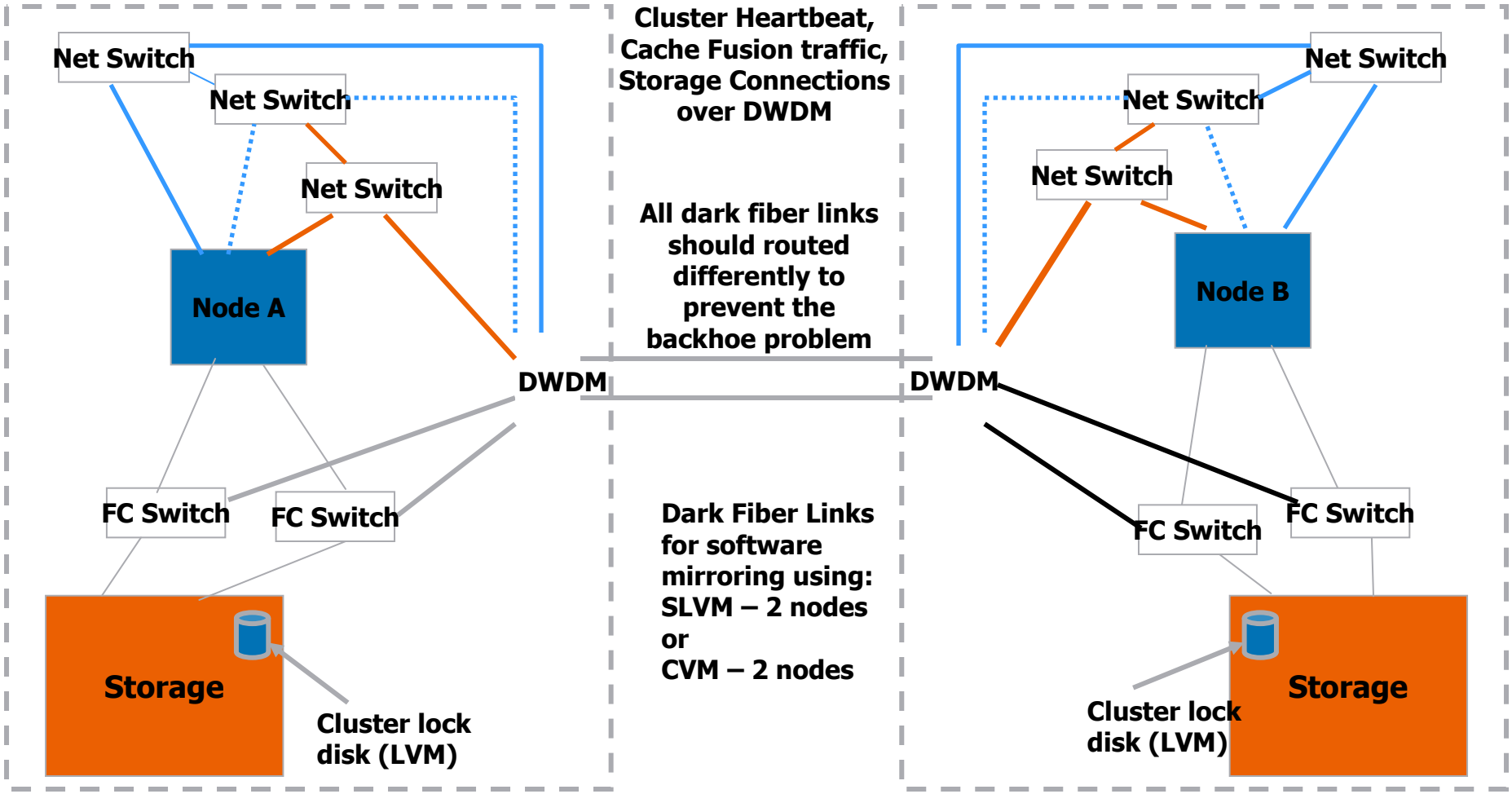
Active/active environment with transparent access to applications and data!

* Plans to extend distance > 100 km

Two Data Center Extended SGeRAC Cluster – up to 10km



Two Data Center Extended SGeRAC Cluster – up to 100km



Arbitrator systems

Used with:

- 3-data center Extended Campus Cluster architecture
- Metrocluster

Arbitrator system(s) for Extended Campus Cluster and Metrocluster

- Arbitrators may be performing important and useful work such as:
  - Another mission-critical application not protected by DT
  - OpenView/Operations or /NetworkNodeManager
  - Network Backup
  - Application Server(s)

- Advantages of using two Arbitrator systems:
  - Provides local failover capability to applications running on the Arbitrator
  - Protects against more multiple points of failure (MPOFs)
  - Provides for planned downtime of a single system anywhere in the cluster

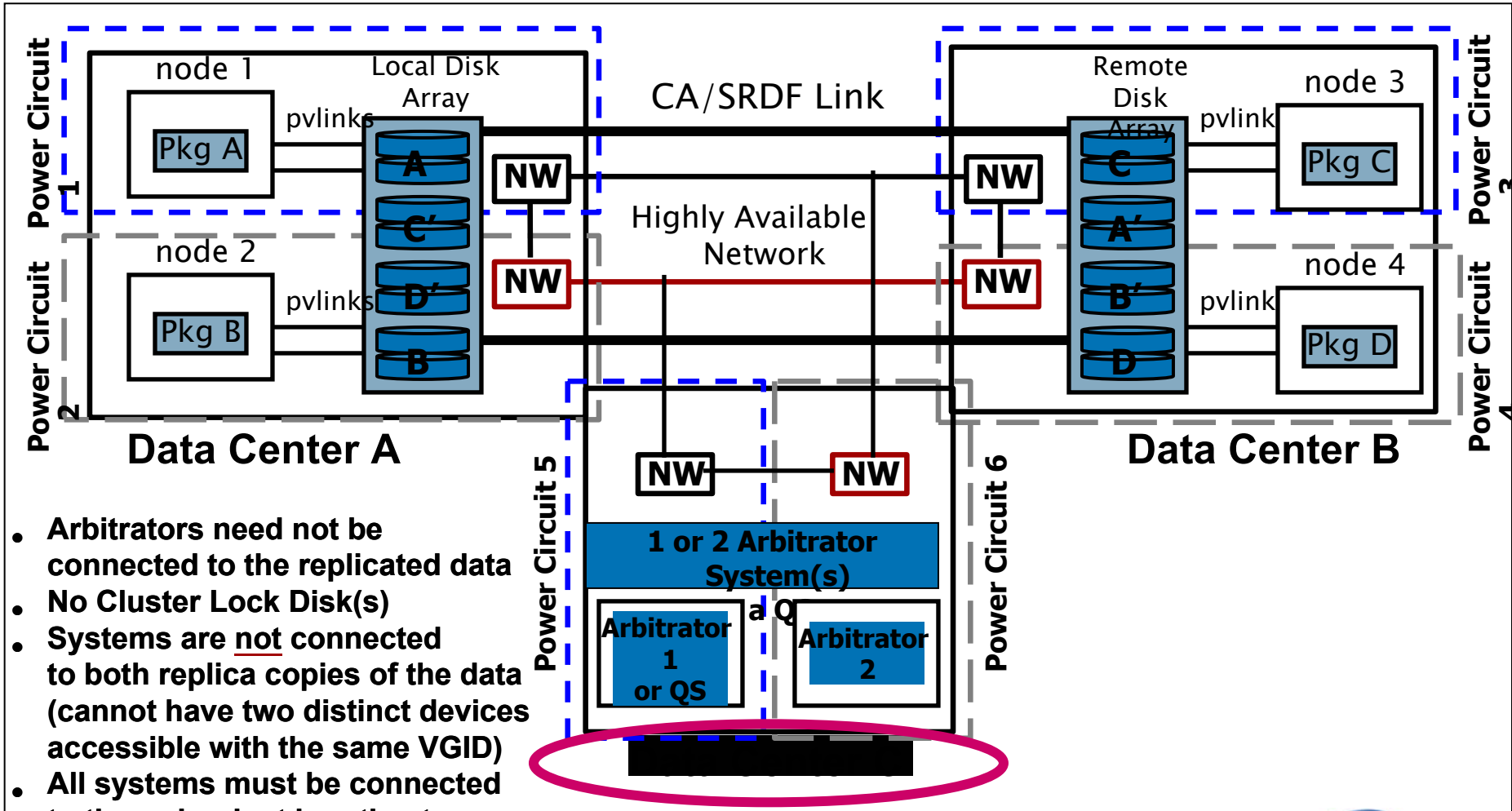
Metrocluster

- Single SG Cluster up to 16 nodes (NO SG OPS)
- Automatic failover, up to 150 application packages among all nodes
- 3 data centers required – Metropolitan (same city) distances
  - up to 50 km between A & B
  - 100 km with DWDM or CWDM
- SCSI or Fibre Channel for local disk connectivity
- Data replication between Data Centers A & B in hardware
  - HP Continuous Access XP (Asynchronous and Synchronous)
  - EMC SRDF (Synchronous ONLY)
  - Toward the end of 2004: EVA CA
- Dedicated links for network and physical data replication (dark fiber) –
  - no telco switched networks
  - no conversion to other protocols like ATM
  - Single IP subnet for heartbeat networks (Ethernet, FDDI)
- MirrorDisk/UX optional for root disks only
- LVM or VxVM ONLY
- No cluster lock disks

Metrocluster (continued)

- Extends protection of Serviceguard cluster to cover certain disasters than affect a data center
- Uses physical data replication in hardware:
- Automated local and remote failover
- MetroCluster works with most Serviceguard applications
- Does not protect against:
  - Human error
  - Database corruption
  - Some application bugs
  - Natural disasters that affect an entire metro area

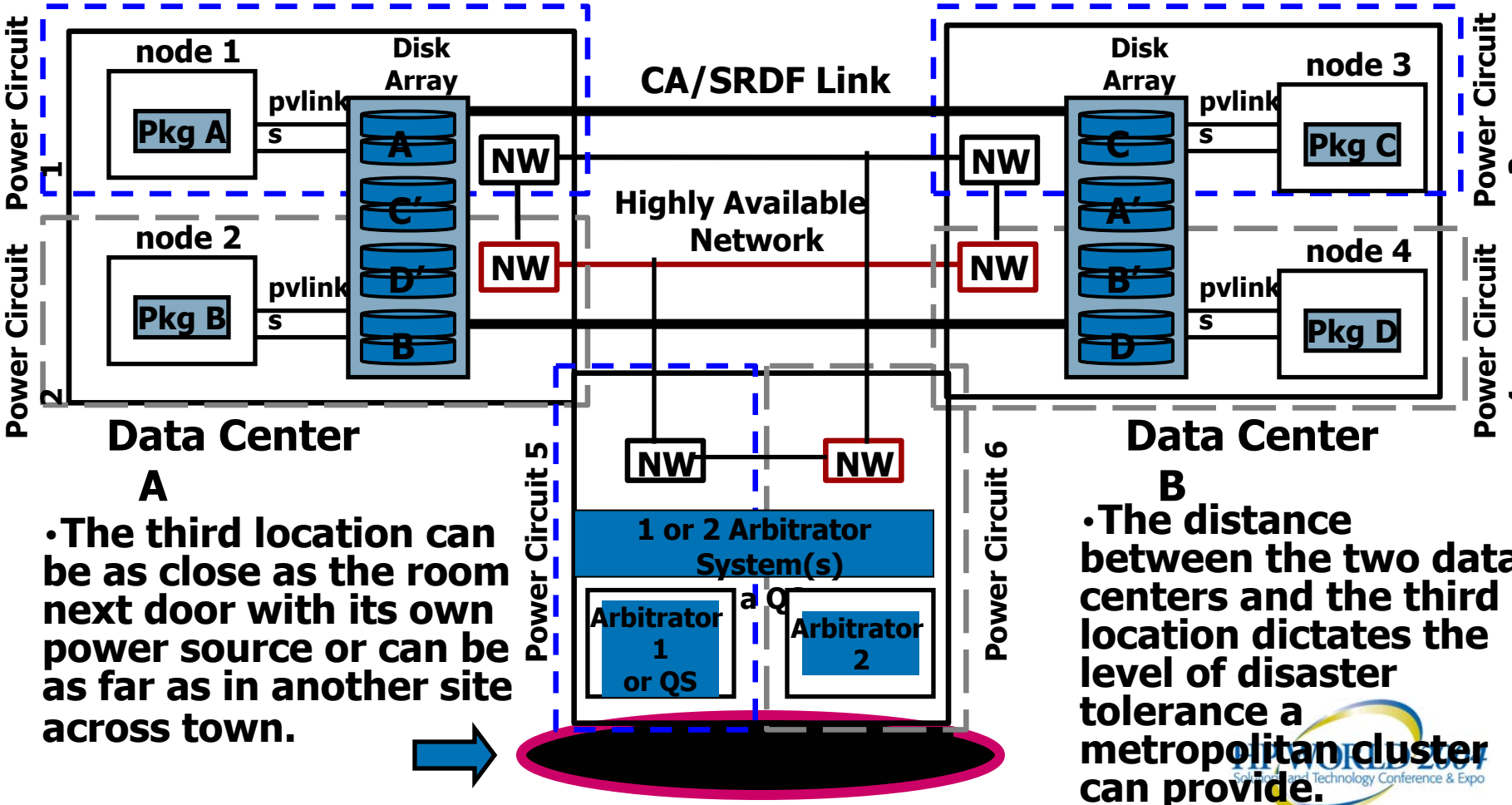
Metrocluster



- **Arbitrators need not be connected to the replicated data**
- **No Cluster Lock Disk(s)**
- **Systems are not connected to both replica copies of the data (cannot have two distinct devices accessible with the same VGID)**
- **All systems must be connected to the redundant heartbeat network links**

Metrocluster example: 2 Main data centers & 3rd location instead of 3 data centers

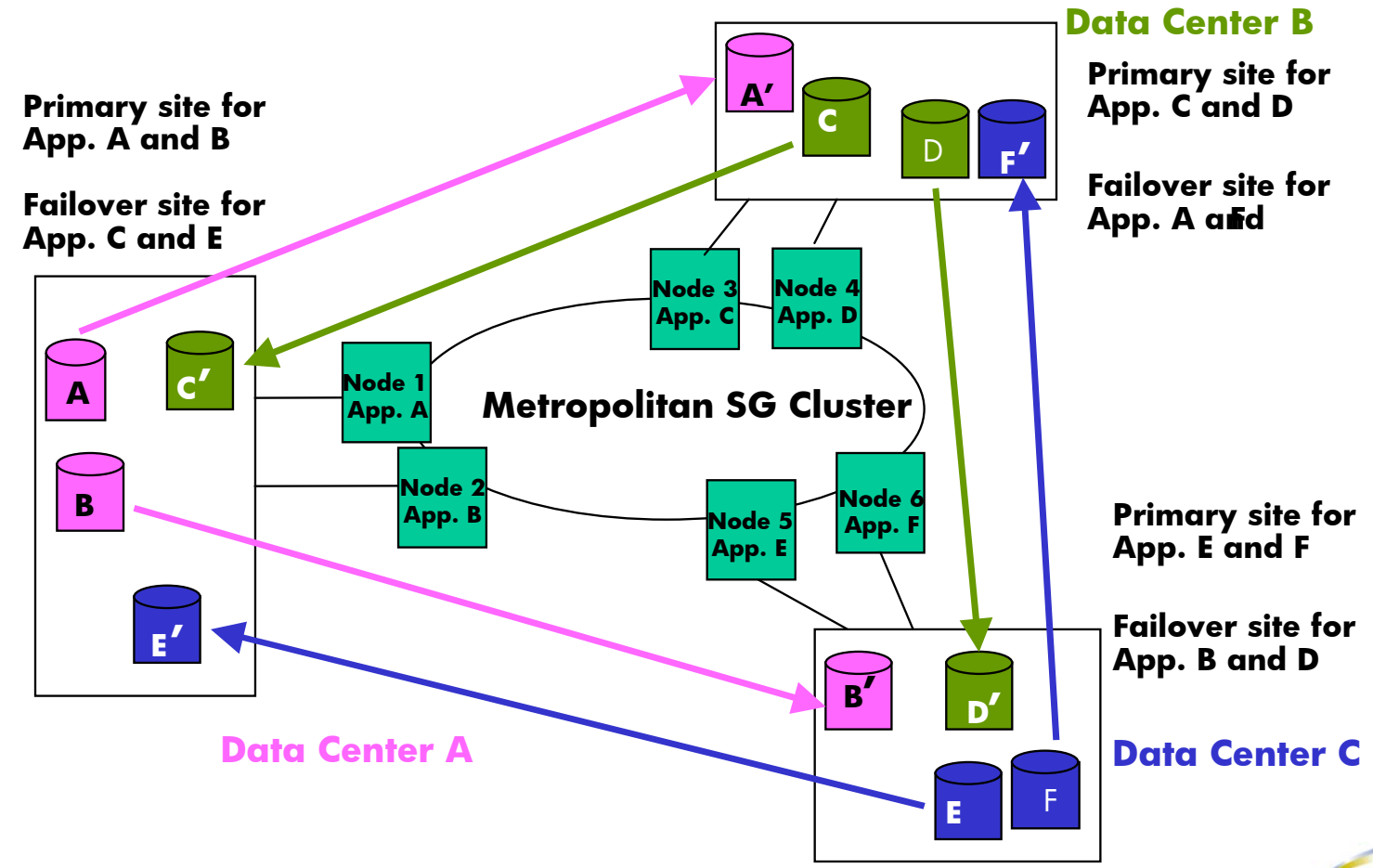
Customer requirements for disaster tolerance dictates the distance to the third location



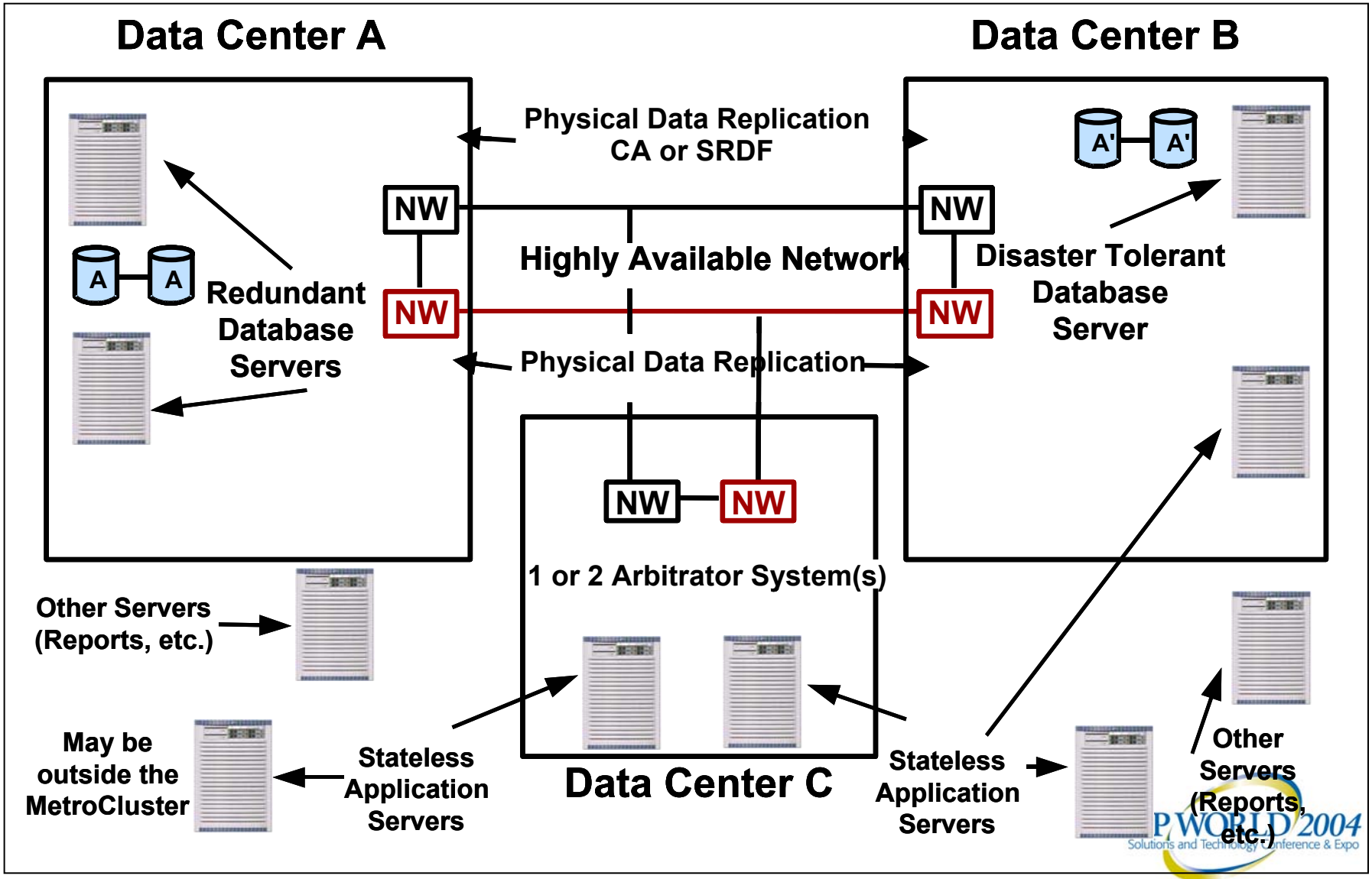
- The third location can be as close as the room next door with its own power source or can be as far as in another site across town.
- The distance between the two data centers and the third location dictates the level of disaster tolerance a metropolitan cluster can provide.

3 active site configuration (with 1-1 data replication) for Metrocluster CA XP

- valid for all supported Metrocluster CA XP versions



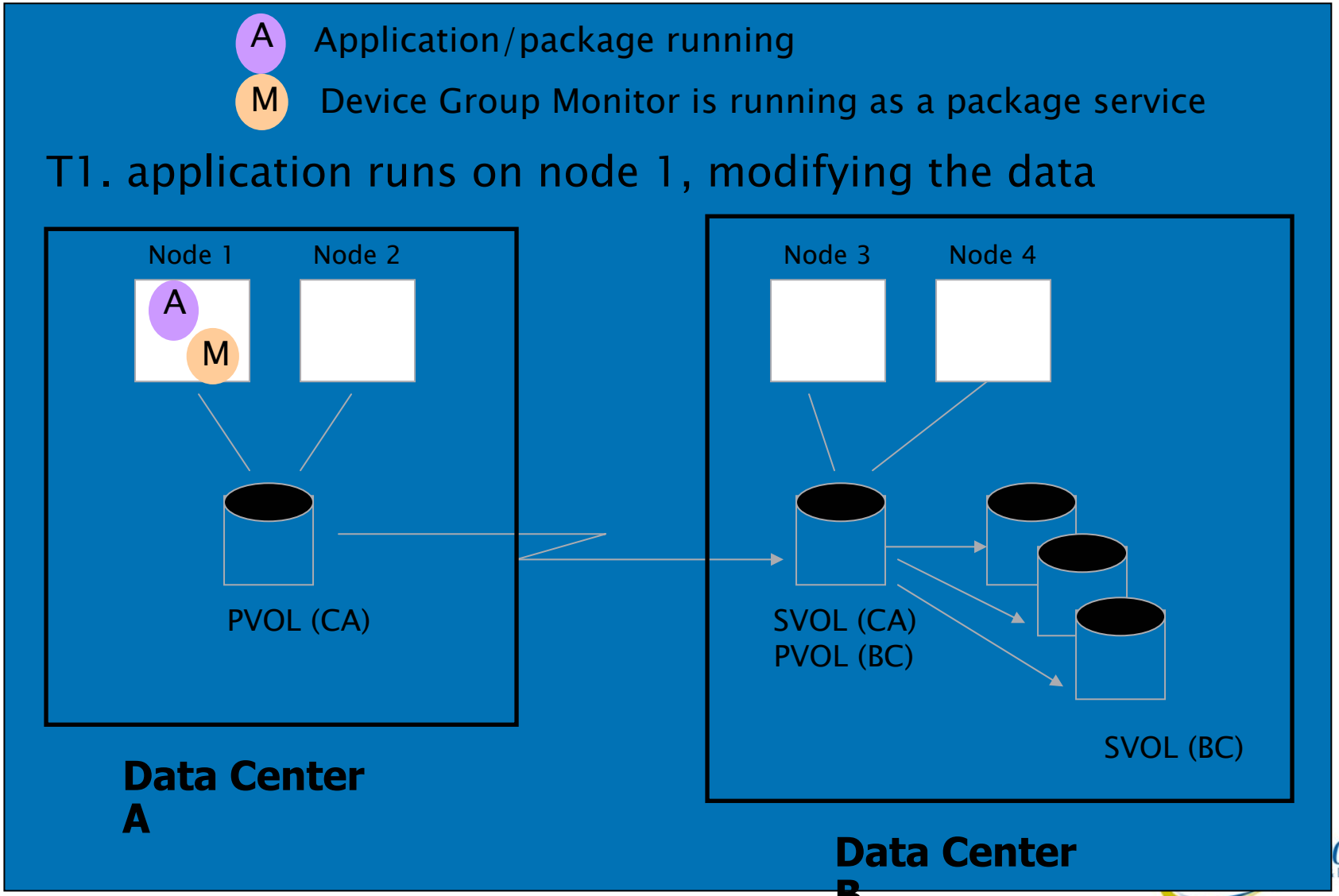
SAP with Metrocluster



Optional Metrocluster/ XP CA architectures

- Support for switched network links (ATM and IP) for XP/CA data replication
  - Besides dark fiber link, MC/CA now also supports switched network links (ATM and IP) for XP/CA data replication.
  - Enables implementation of a Metrocluster/CA solution in an environment where dark fiber is not available.
  - Basic Requirements
    - Maximum distance between the XP disk arrays is 100 kilometers
    - Network latency for cluster network no more than 200 milliseconds
    - Business Copy at both sites for point-in-time copies to recover from a rolling disaster (due to decreased reliability of these links)
    - Cluster network must be pure Ethernet (i.e., no ATM encapsulated Ethernet packets) with all nodes in the same subnet
    - There must be at least two (redundant) supported converter boxes installed at each site
    - For IP configuration, there must be at least two IP routers or LAN switches installed at each site
    - The Device Group Monitor, which detects link failure and splits the BC, must be configured as an application package

XP CA & BC Configuration with Device Group Monitor



XP CA Device Group Monitor Example

T2. All CA links are down, application continues to run on PVOL side

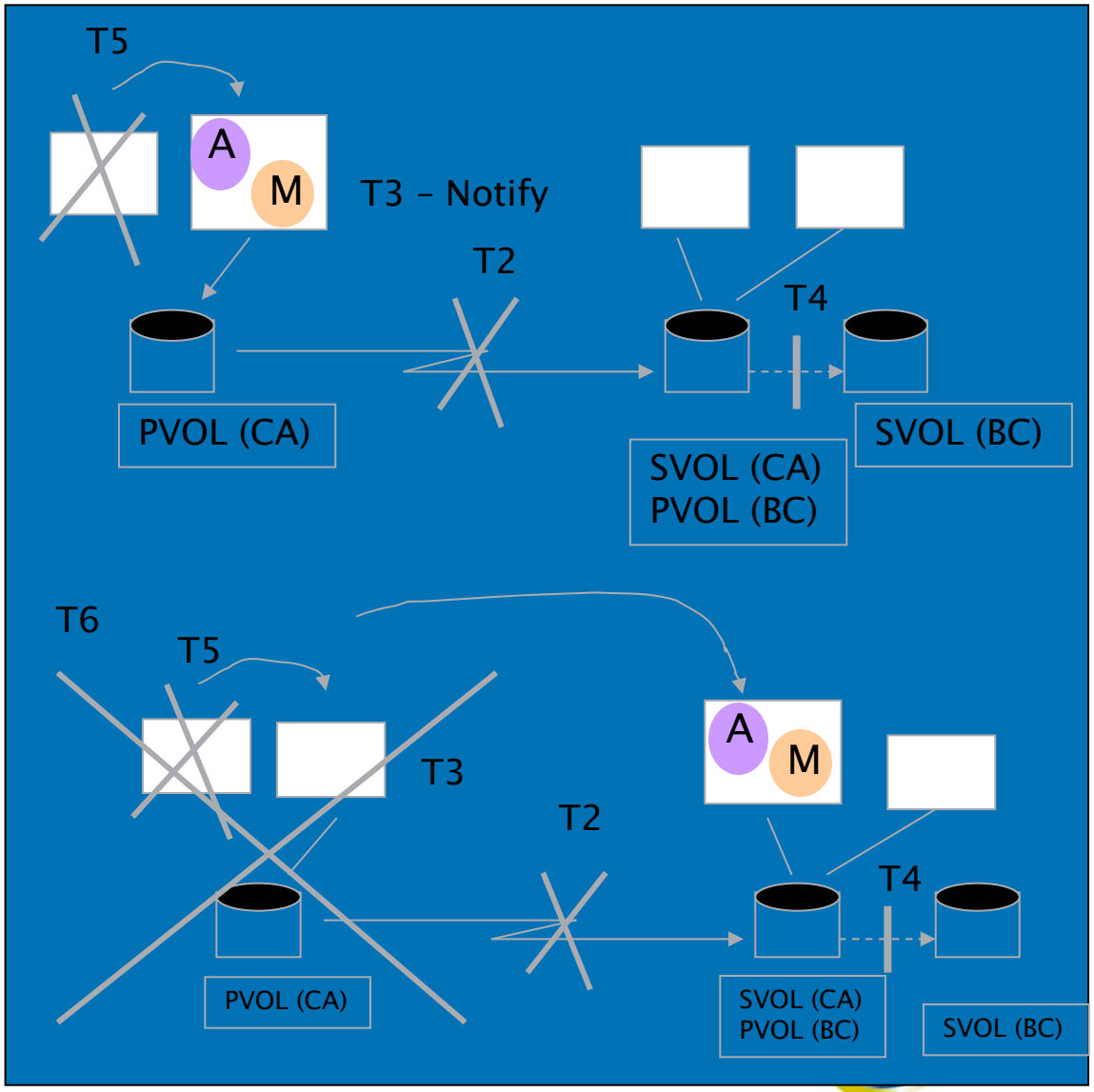
T3. Monitor detects CA links failure, notification generated. AUTO_RESYNC is 0 or not set

T4. IF AUTO_RESYNC is set to 1, monitor automatically splits a configured BC and then periodically tries to resynchronize the data.

T5. Node 1 failure in Data Center A causes Application package and Monitor service to start on local node 2.

T6. Data Center A fails. Application package and Monitor service fail over to Data Center B.

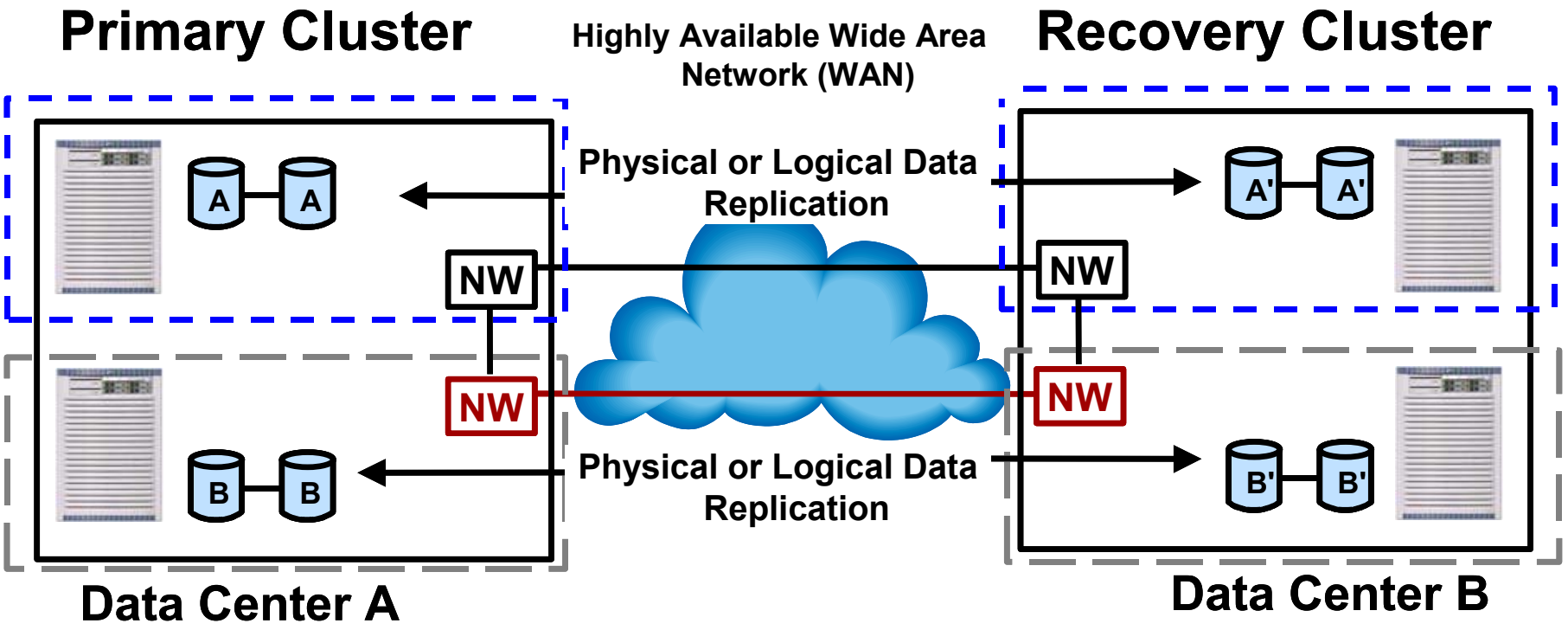
If failure occurs before link recovery or the resynchronization completes, the BC will remain split.



Continentalclusters

- Dual SG Clusters up to 16 nodes each
- Automatic failover, up to 150 application packages within each cluster, push-button failover of pre-configured set of application packages between the clusters (bi-directional)
  - Group based recovery can be used to recover a selected recovery group and gives the option of recovering applications in a desired sequence and time
  - Option to switch primary and recovery package roles in the recovery groups for which the specified cluster is defined as the primary cluster
- 2 data centers required with up to Intercontinental distances
- SCSI or Fibre Channel for local disk connectivity
  - HP Continuous Access XP physical data replication
  - EMC SRDF physical data replication
  - Oracle Standby Database logical replication
- Telco switched networks and protocol conversion to other protocols like ATM OK between clusters ONLY
- Single IP subnet for heartbeat networks (Ethernet, FDDI) within each cluster
- MirrorDisk/UX optional for root disks only when using physical replication
- LVM, VxVM, SLVM or CVM
- Normal cluster lock rules within clusters

Continentalclusters



- **Systems are not connected to both replica copies of the data (hosts in each cluster are connected to only one copy of the data)**
- **Each cluster must separately conform to heartbeat network requirements**
- **Each cluster must separately conform to quorum rules (cluster lock disks, Arbitrators or Quorum Server)**
- **Use of cluster lock disks requires three power circuits in each cluster**

* **HA network is used for both data replication and inter-cluster monitoring**

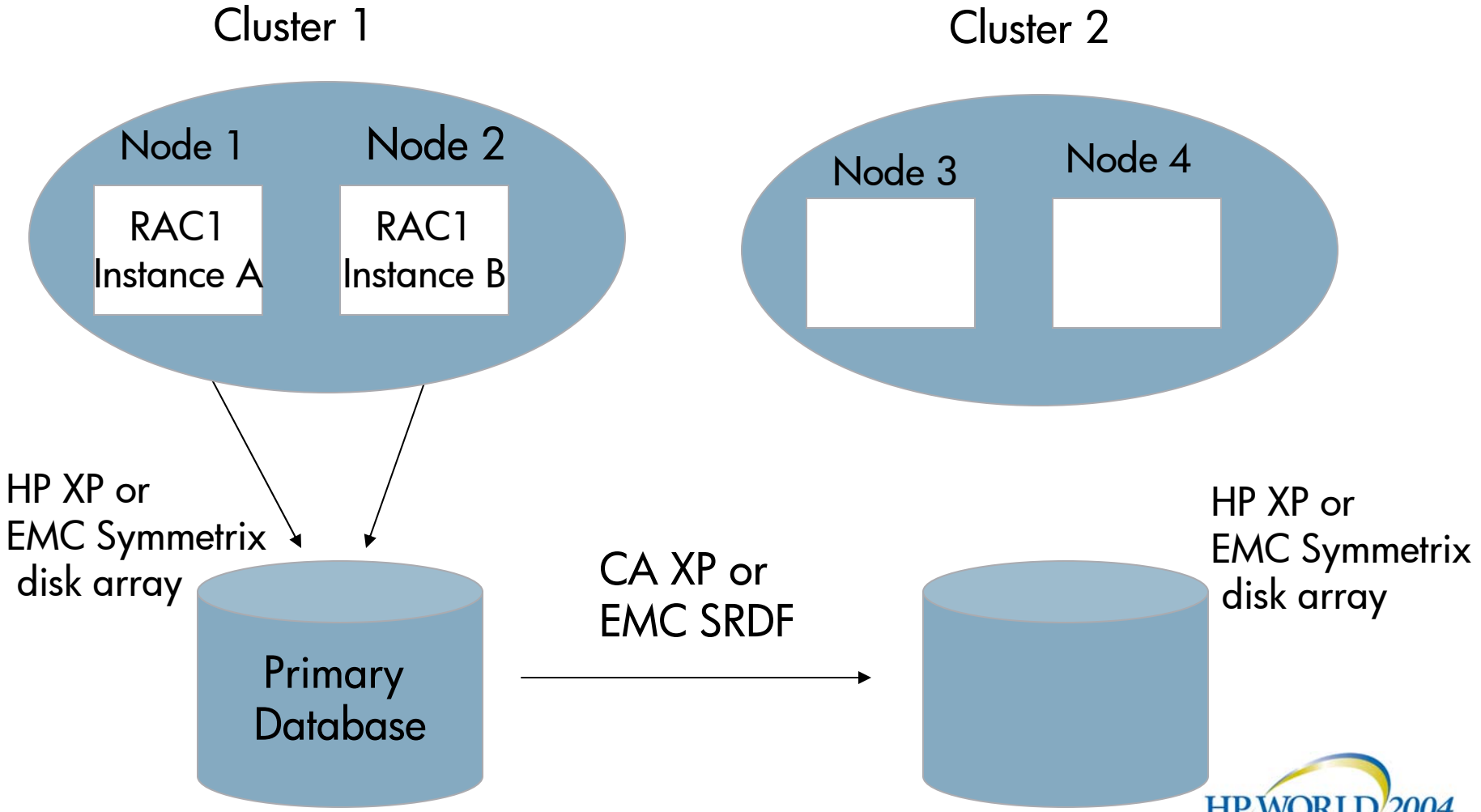
Continentalclusters (continued)

- Each cluster can backup the other cluster (bi-directional)
- Human decision is necessary to initiate the failover
- Cluster problem notification:
  - Text files
  - System console
  - e-mail
  - SNMP trap
  - opcmsg (OpenView/Operations)
- Local failover still occurs within the primary cluster
- Remote failover is used only when the entire primary cluster fails or is unreachable (not if a single node or package fails)
- Choice of data replication methods (physical or logical):
  - HP XP Continuous Access (Async and Sync)
  - EMC SRDF (Sync only)
  - Oracle Standby Database
  - **"Allow model" for other data replication methods to be integrated**

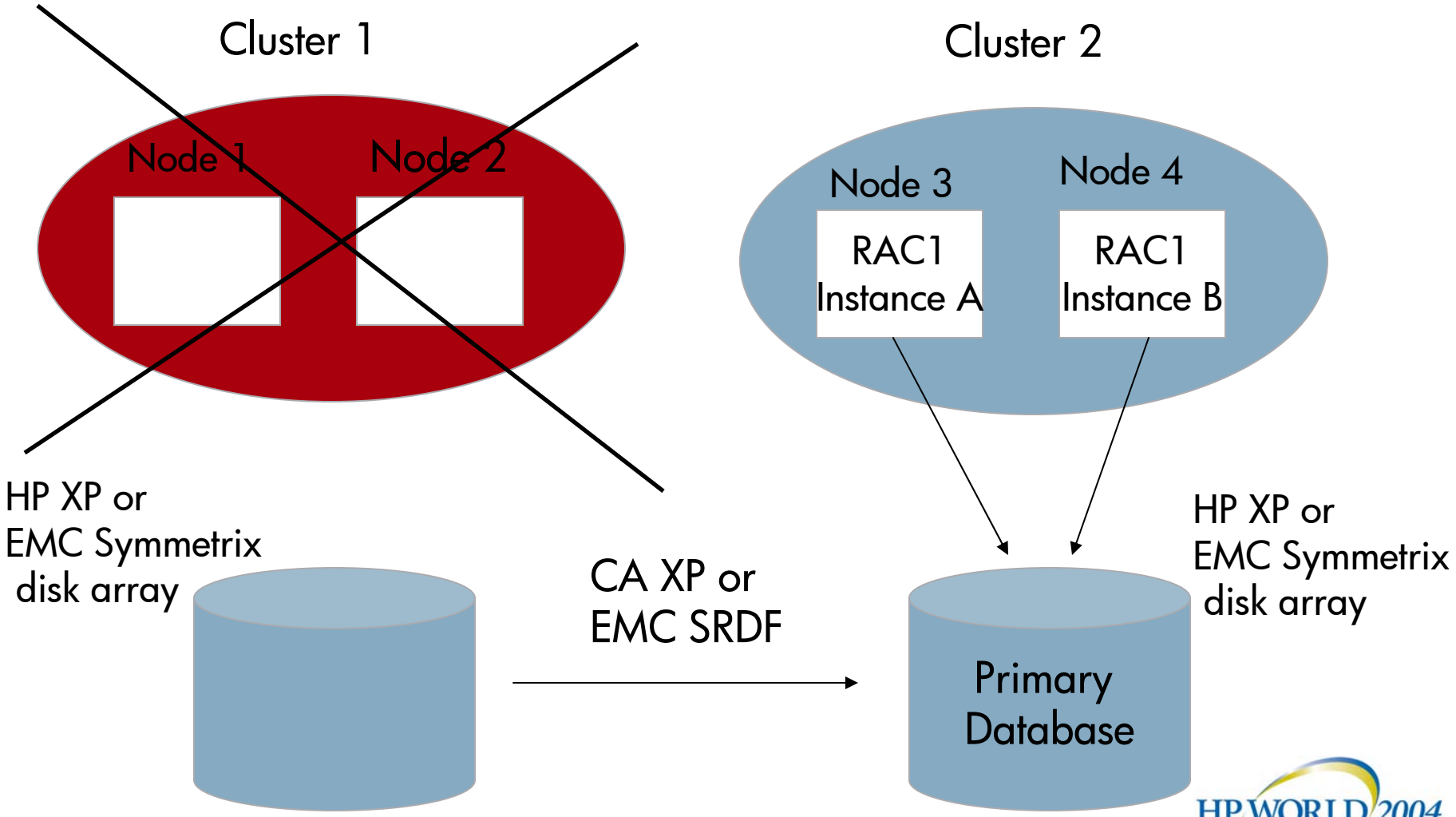
Continentalclusters and Oracle 9i RAC

- Oracle 9i RAC instances running on the primary cluster will be restarted/recovered by CC on the recovery cluster upon primary cluster failure
- This is a bi-directional (mutual) active/standby model where DIFFERENT RAC databases may run in each cluster
- Requirements
  - SLVM for volume management
  - XP CA or EMC SRDF for data replication.
  - VxVM/CVM volume management is not currently supported in this configuration

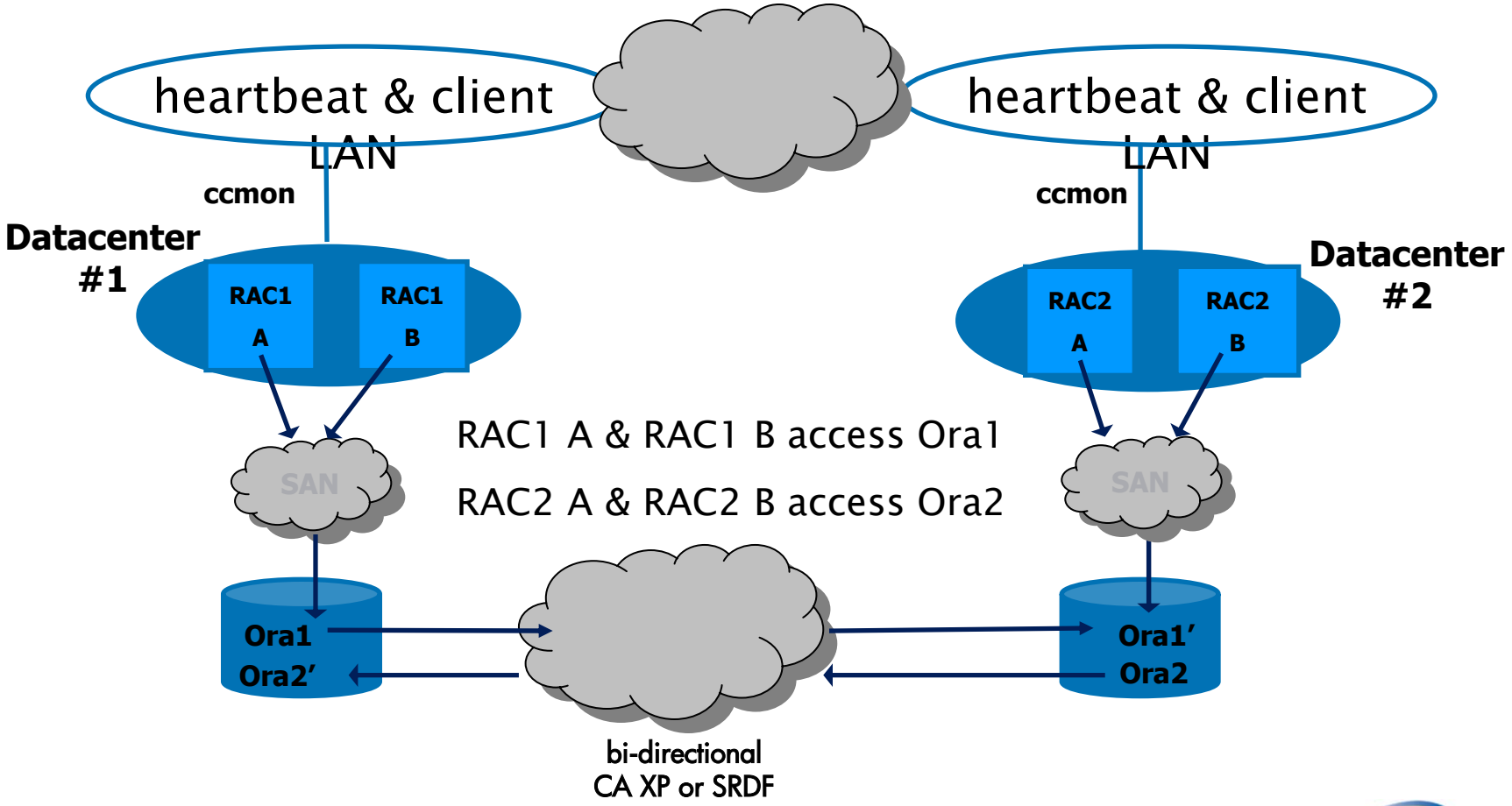
Oracle 9i RAC instances running in CC environment before failover:



Oracle 9i RAC instances running in CC environment after failover:



HP Continentalclusters for RAC Disaster tolerance with failover at unlimited distances

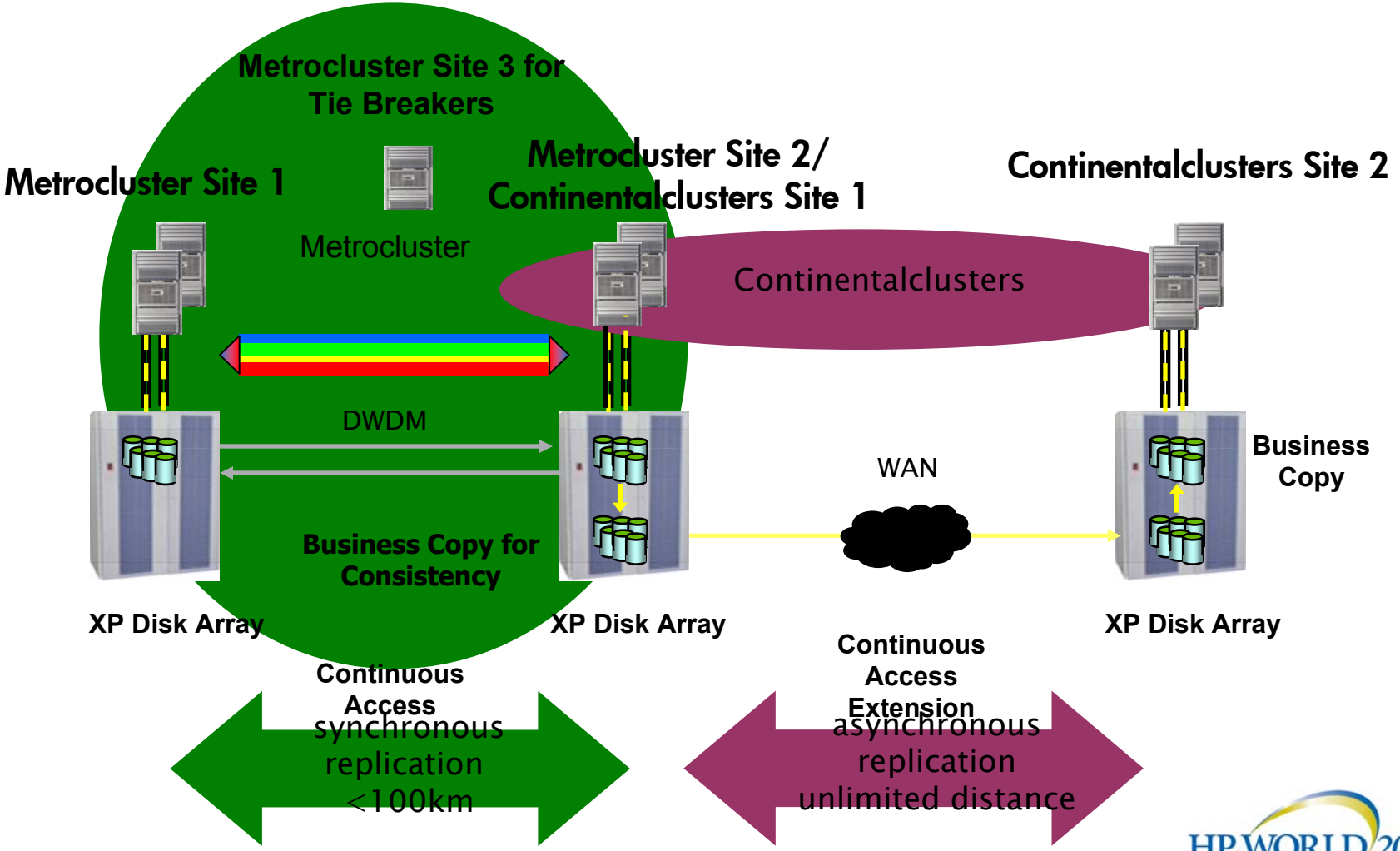


Comparison between various RAC solutions

	Same Data Center RAC	Extended RAC 10 km	Stretched RAC 100 km	CC RAC
# of clusters	1	1	1	2
# of nodes	Up to 16	2, 4	2	2 * local
Distance	Local	10km	100km	unlimited
OS	**SLVM: 11.0, 11.11, 11.23** **CVM: 11.11, 11.23**	**11.11, 11.23**	**SLVM: 11.11, 11.23** **CVM: 11.11**	**11.11**
Volume manager	**SLVM – 16** **CVM – 4***	**SLVM – 2** **CVM – 2,4**	**SLVM, CVM**	**SLVM**
Mirroring	software mirroring or RAID	software mirroring	software mirroring	XP CA sync or async
Storage	No restrictions*	No restrictions*	No restrictions*	XP
Active/active	Yes	Yes	Yes	No

NOTE: EVA storage is not currently compatible with CVM

Combining Metrocluster & Continentalclusters



HA / DT solution comparison chart

	SG	SGeRAC	Ext. SG–2DC	Ext. SG –3DC	Metrocluster	Continentalclusters
# of Clusters	1	1	1	1	1	2
# of DCs	1	1	2	3	3	2
Networks	Ethernet FDDI	Ethernet FDDI	Ethernet FDDI	Ethernet FDDI	Ethernet FDDI	Any between clusters
Disk Mirroring	Optional	Optional	MirrorDisk/UX or Veritas M	MirrorDisk/UX or Veritas M	Root disks only	Within clusters only
Data Replication: Physical/Logical	N/A	N/A	Physical – SW Mirroring	Physical - SW Mirroring	Physical – HP XP CA or EMC SRDF	HP XP CA, EMC SRDF (P) or Oracle S/B (L)
Volume Mgr	LVM, VxVM	SLVM, CVM	LVM, VxVM SLVM, CVM	LVM, VxVM SLVM, CVM	LVM, VxVM	LVM, VxVM
Cluster Lock	Quorum Server or Single Disk	Quorum Server or Single Disk	Quorum Server or Dual Disk	Quorum Server or Arbitrators	Arbitrators or Quorum Server	Quorum Server or Single Disk (each cluster)
Failover Automation	Automatic, Package/Node	Automatic, Package/Node	Automatic, Package/Node	Automatic, Package/Node	Automatic, Package/Node	Operator-initiated, all or nothing
Hyperfabric	No	Yes	Up to 10 km RAC only	Up to 10 km RAC only	No	No
Subnets	Single	Single	Single	Single	Single	Dual
Long Distance Link	N/A	N/A	Dark Fiber, DWDM opt 100 km – SG 10 km - SGeRAC	Dark Fiber, DWDM opt 100 km – SG 10 km - SGeRAC	Dark Fiber, DWDM opt 100 km	Switched Circuit OK Unlimited distance

Solution Matrix for Oracle / 9i RAC

Disaster Tolerance for Oracle 9i & 9i RAC on HP-UX PA RISC & Integrity

	Single Instance				RAC Solutions			Extended RAC		
Topology	Extended Campus Cluster		Metro-cluster	Continental-clusters	Serviceguard Extension for RAC		VERITAS DBeAC	Extended Cluster for RAC	Extended Cluster for RAC	Continental-clusters for RAC
Distance	10 km	100 km	100 km	Unlimited	Local Datacenter		Local DC	10 km	100 km	Unlimited
Volume Manager	VxVM	LVM	LVM VxVM	LVM, VxVM	SLVM	CVM	VERITAS CVM	CVM	LVM, CVM	SLVM
CFS	-	-	-	-	-	-	VCFS	-	-	-
SW Mirroring	VxVM Mirroring	MirrorDisk UX	N/A	Any 3rd party mirroring DataGuard	MirrorDisk UX	CVM Mirroring	CVM Mirroring	CVM Mirroring	MirrorDisk UX, CVM Mirroring	DataGuard Optional
HW Mirroring	N/A	N/A	CA XP, SRDF	CA XP, SRDF	N/A	N/A	N/A	N/A	N/A	CA XP
Storage	VA, XP, EMC	VA, XP, EMC, EVA	XP, EMC	VA,XP, EMC, EVA	VA, XP, EMC, EVA	VA, XP, EMC	XP, EMC	VA, XP, EVA	VA, XP, EVA	XP
# Nodes	16	16	16	16 per cluster / 32 nodes total	16	8	4	2 or 4	2	16 per cluster / 32 nodes total
Bi-directional Failover	Yes	Yes	Yes	Yes	N/A	N/A	N/A	Yes	Yes	Yes
Servers Supported	PA-RISC & Integrity	PA-RISC & Integrity	PA-RISC & Integrity	PA-RISC & Integrity	PA-RISC & Integrity	PA-RISC & Integrity	VERITAS Tested A Class	PA-RISC & Integrity	PA-RISC & Integrity	PA-RISC & Integrity

Criteria for Choosing a Campus Cluster Architecture

- Short distances (campus environment)
  - High throughput links for maximum performance
  - Relatively low cost of implementation
    - Software mirroring is much less expensive than hardware array-based replication
    - Network infrastructure is already in place
    - Separate links for FibreChannel and TCP/IP
  - Require an active/active RAC disaster tolerant solution
  - Meets Disaster Tolerance business requirements
    - Not often subject to natural disasters that would impact the entire campus (perhaps have separate electric service or backup generators)
    - Impact of data and service loss does not warrant the cost of implementing a distant data center

Criteria for Choosing an Extended Cluster Architecture

- Metropolitan distances (up to 100 km)
  - Dark fiber already exists or is readily available
  - High throughput links for maximum performance
  - Medium cost of implementation
    - Software mirroring is much less expensive than hardware array-based replication
    - Network infrastructure costs increase due to
      - Dark fiber
      - Redundant DWDM/CWDM converters
    - Share links for TCP/IP and FibreChannel
  - Require an active/active RAC disaster tolerant solution
  - Meets Disaster Tolerance business requirements
    - Fully automated failover without human intervention
    - Greatest availability
    - Not often subject to natural disasters that would impact the entire metropolitan area (e.g., not in earthquake country)
    - Impact of data and service loss does not warrant the cost of implementing a more distant data center

Criteria for Choosing a Metrocluster Architecture

- Metropolitan distances (up to 100 km)
  - Dark fiber already exists or is readily available
  - High throughput links for maximum performance
  - Medium cost of implementation
    - Must invest in enterprise disk arrays that support array-based replication
    - Network infrastructure costs increase due to
      - Dark fiber
      - Redundant DWDM/CWDM converters
    - Share links for TCP/IP and FibreChannel
  - Meets Disaster Tolerance business requirements
    - Fully automated failover without human intervention
    - Greatest availability
    - Not often subject to natural disasters that would impact the entire metropolitan area (e.g., not in earthquake country)
    - Impact of data and service loss does not warrant the cost of implementing a more distant data center

Criteria for Choosing a Continentalclusters Architecture

- Continental and Intercontinental distances
  - Dark fiber is not readily available or the distance is too great
  - Link cost is the largest component
    - Network infrastructure costs increase due to
      - Bandwidth needed to support both TCP/IP and data replication
      - Long distance
  - Want control over the initiation of the failover
  - Failover only when entire data center fails meets DR model and availability requirements
  - Active/standby RAC configuration is acceptable
  - Meets Disaster Tolerance business requirements
    - Not often subject to natural disasters that would impact the entire metropolitan area (e.g., not in earthquake country)
    - Impact of data and service loss does not warrant the cost of implementing a more distant data center

Data replication

Categories & attributes for data replication

- Categories for data replication include:
  - Off-line: storing data on tape for transport to another data center
  - On-line: copying data from one data center to another over a communications link
- Data replication should provide:
  - Data consistency: data is replicated to maintain the original logical order of the application, making it immediately useable or recoverable
  - Data currency: data is replicated quickly to allow access to most or all of the latest committed disk transactions during a recovery
  - Data recoverability: action can be taken to make the data useable
  - Minimal data loss: replication method addresses consistency, currency and recoverability

Data replication methods

- Physical replication: disk write operations are replicated to a secondary data center, the method of which can be either:
  - Hardware-based (disk storage subsystem performs the replication over a disk-to-disk link)
  - Software-based (system host performs the replication over a system-to-disk link)
- Logical replication: data is replicated by repeating the sequence of transactions that were used to create the original data at a secondary data center

Data replication modes

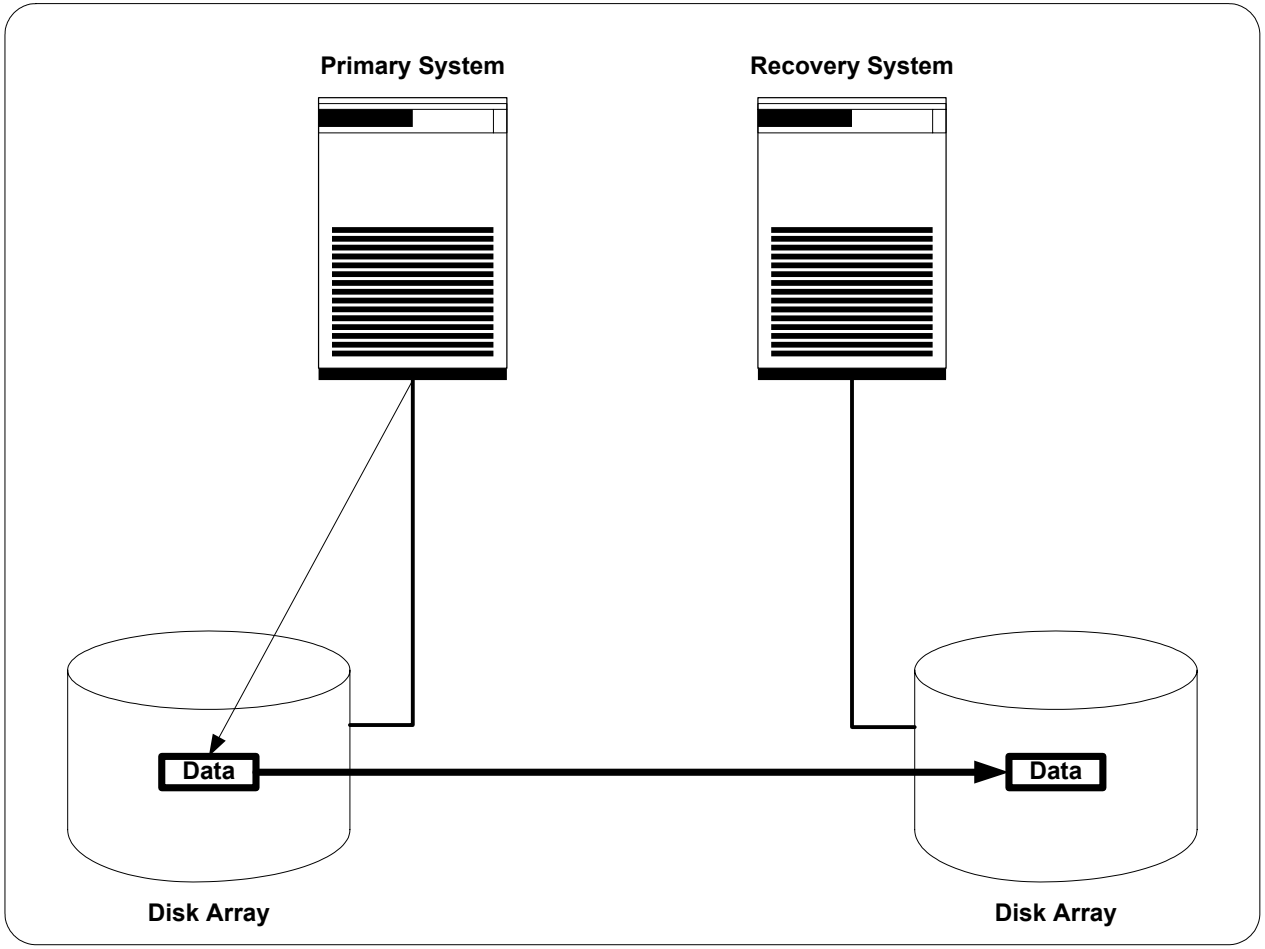
- Both physical and logical data replication can operate in two basic modes, depending on business data integrity requirements:
  - Synchronous Mode: a disk write must be completed locally and replicated remotely before another disk write can be performed
  - Asynchronous Mode: disk writes can occur locally without waiting for the write operations to be replicated remotely
- Some disk array vendors offer other replication modes which are not supported since they do not guarantee write order required for data integrity

Physical Replication

Physical replication

- Each physical disk write operation is replicated to another disk at another site
- Physical replication is not application-dependent, allowing alternate systems to take over and use the replicated data after a disaster (provided the replicated data is current and consistent)
- May be used for file systems and databases

Physical replication: hardware-based



Hardware-based Physical Replication

Advantages of hardware-based physical replication:

- Lower latency writing data to the replica disk array may allow for better data currency (depending on link speed and distance)
- Replication uses no additional CPU resources
- Disk array performs array-to-array resynchronization if the link or disk fails
- Replication can be performed in both directions, allowing restoration of a primary site after the replica disk has taken over during a failure or disaster

Disadvantages of hardware-based physical replication:

- Logical write order of data is not maintained in synchronous mode during a link failure
  - Can increase the risk of data inconsistency in case of a rolling disaster where another failure occurs before resynchronization completes
- Additional costs for specialized data replication hardware
- Distance between sites and data throughput can be limited depending on the disk array interconnect technology used
- Cost of links used for data replication between sites can vary drastically depending on data throughput requirements

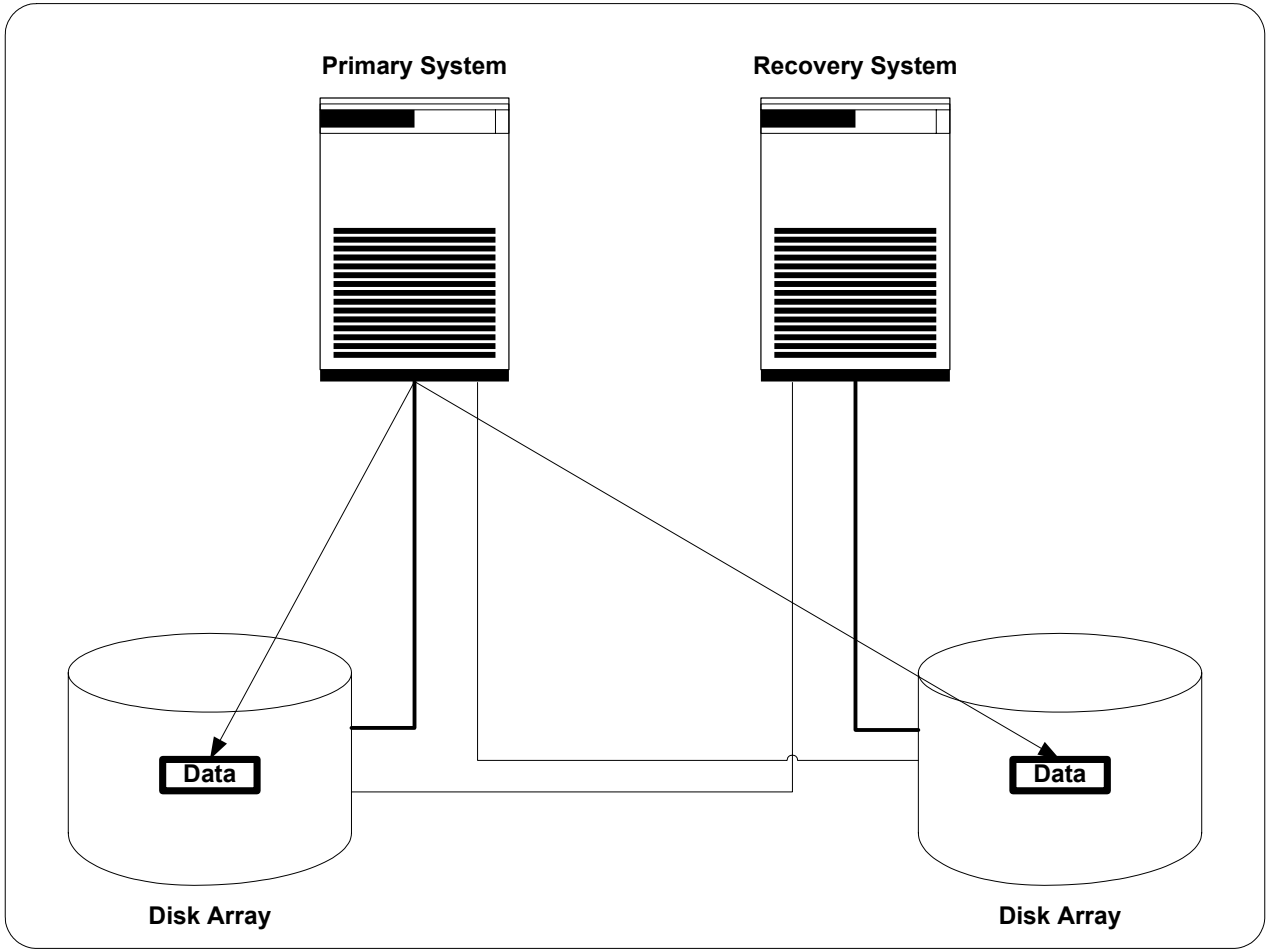
Disk Array Support

- HP SureStore Continuous Access (CA) XP
  - Synchronous and asynchronous copy modes
- EMC Symmetrix Remote Data Facility (SRDF)
  - Only synchronous mode is supported by HP with its Disaster Tolerant solutions
- Remote replication over a wide range of link technologies, such as
  - DWDM and CWDM
  - WAN/LAN converters for ATM (OC-3), DS-3 (T3) and IP
- using
  - ESCON (Enterprise System Connect)
  - FC (Fiber Channel)

Software-based Physical Replication

- Advantages of software-based physical replication:
  - Minimal lag time writing data to the replica disk array allows for better data currency
  - Replication is independent of disk technology used, allowing the use of any type of supported disks
  - Since data copies are peers (all copies are used for both reading and writing), the disk replica does not require reconfiguration to operate as a primary disk after a failure
  - Disk read performance can be improved with system access to both copies of data
- Disadvantages of software-based physical replication:
  - Logical write order of data is not maintained in synchronous mode after a link failure, which can increase the risk of data inconsistency
  - Redundant disk hardware and cabling are required, thus doubling storage costs
  - Distance between sites is limited by physical disk link capabilities
  - System performance can be affected by CPU overhead for mirroring, double disk write I/O operations, and resynchronization that must be performed by the host

Physical replication: software-based



HP MirrorDisk

- HP MirrorDisk/UX is a physical data replication software subsystem available as an add-on product for HP-UX that provides:
  - Striping (RAID 0) and mirroring (RAID 1) for up to three copies (total) of hp Logical Volume Manager (LVM) disk volumes
  - Support for replicating volumes up to a distance of 100km using DWDM (Dense Wave Division Multiplexing) or CWDM

Veritas mirroring in VxVM and CVM

- Veritas Volume Manager (VxVM) and Cluster Volume Manager (CVM) can perform physical data replication in software:
  - Striping (RAID 0) (base product), mirroring (RAID 1) up to 32 mirrors, mirrored stripes (RAID 0+1), striped mirrors (RAID 1+0), RAID 5 (full products)
  - Support for replicating volumes up to a distance of 10km using DWDM (Dense Wave Division Multiplexing)

Veritas Volume Replicator VVR

- Veritas Volume Replicator (VVR) is a physical data replication software subsystem built upon the Veritas Volume Manager and provides:
  - Full support for all commercial DBMS (e.g., Oracle, Informix, …)
  - Full heterogeneous replication on any storage hardware platform across any IP network
  - Synchronous and asynchronous volume group replication modes for up to 32 replication links
  - Replication of only the data blocks that actually change, which makes efficient use of network resources
  - Delta-based resynchronization of data blocks between sites

Veritas FlashSnap

- Add-on software from Veritas that creates point-in-time copies (snapshots) of data for off-host processing, such as backups, reporting and application testing
  - Copies are created independent of disk hardware used
  - Provides two snapshot methods
    - Volume-based snapshot (uses Veritas Volume Manager to create a point-in-time copy of a logical volume)
    - Storage Checkpoint (creates an exact image of a Veritas file system by tracking which data changes in the file system and maintains original data in a block map associated with the checkpoint)
  - Applications must be suspended prior to creating the snapshot

Quest SharePlex FS

- Quest SharePlex FS is a file system replication product that:
  - Replicates changes for any file type, including text, data, log, archive, compressed, multimedia, graphics, HTML, XML, scripts, applications and more via TCP-IP to one or more target systems
  - Distributes byte-by-byte data block changes asynchronously, which reduces CPU usage, network and disk overhead
  - Supports rolling upgrades by allowing the source server to queue production activity and then release the updates to the target systems after upgrades are completed

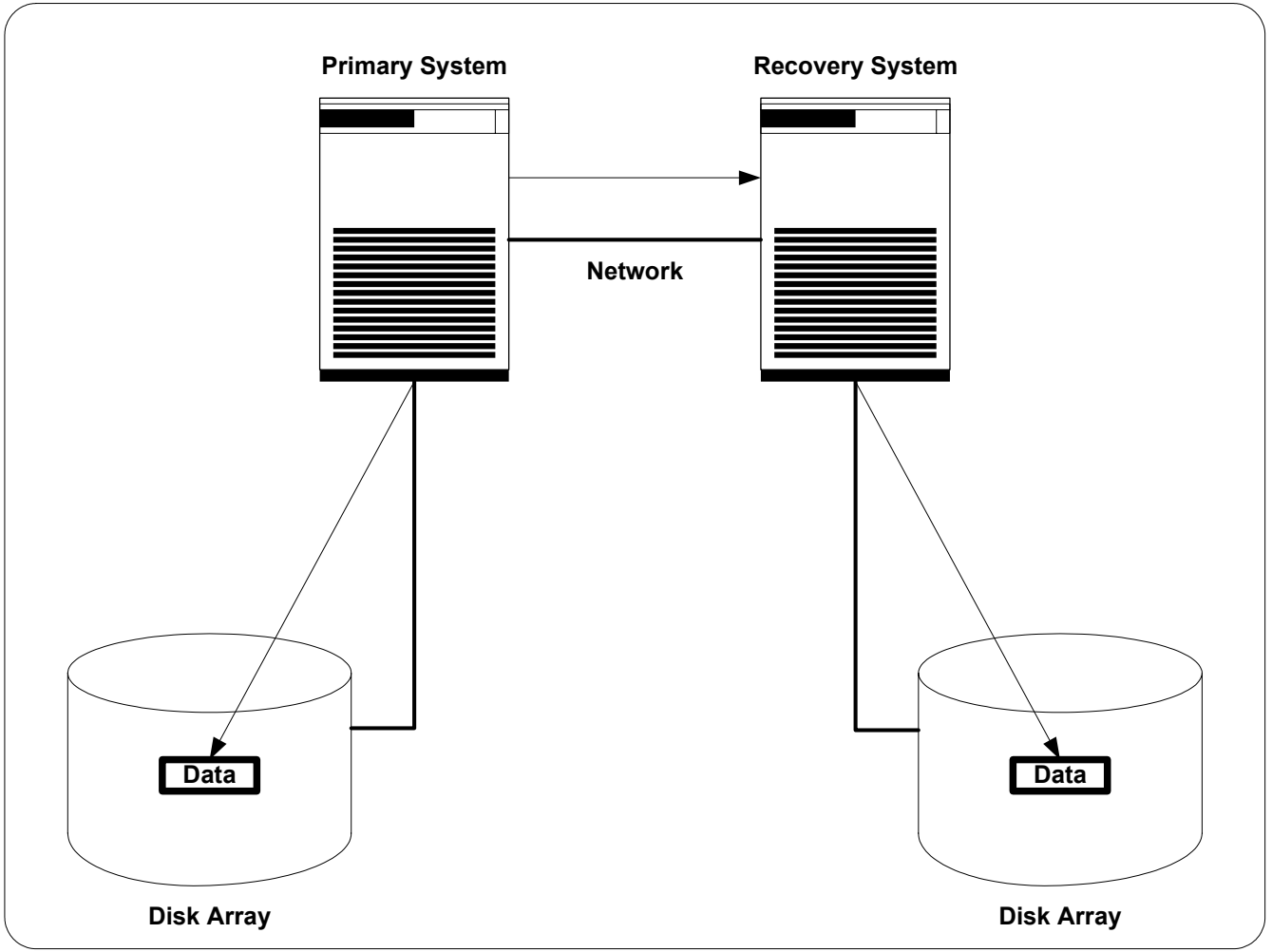
Logical Data Replication

- Data is replicated by repeating at a remote site the sequence of transactions that were used to create the original data
- Logical replication is most often performed at the file system and database level to ensure availability and consistency of application-associated data
- All major database vendors provide software that perform logical data replication
  - Oracle (DataGuard (log-based)
  - Oracle Advanced Replication (transaction-based)
  - IBM Informix HDR
  - IBM UDB (DB/2)
  - Sybase
- Other examples of applications that perform logical replication include:
  - Transaction Processing Monitors (TPMs)
    - BEA Tuxedo (C and C++)
    - BEA WebLogics (Java)
    - IBM WebSphere (Java)
  - Reliable message queuing software (e.g., IBM MQseries)

Logical Replication

- Advantages of logical replication:
  - Distance between sites is limited only by networking technology used and transaction throughput requirements
  - No additional hardware (other than the servers and disks at the remote site) is required to perform replication (unless CPU or network performance is affected by replication overhead)
  - Reduces the chance of human error affecting remote data in some cases by replicating DB transactions and not commands (e.g., adding or removing DB tables)
  - Database transactions can be rolled forward or backward to ensure correct level of data currency
- Disadvantages of logical replication:
  - Replication can cause significant increases in CPU and network utilization, which can affect data currency at the remote site
  - Restoration of a primary database from the remote site can be complex and time consuming
  - "Personality swapping", or "role reversal" (having the remote site become the primary and the former primary site become the secondary) is a very manual process for most logical replication methods

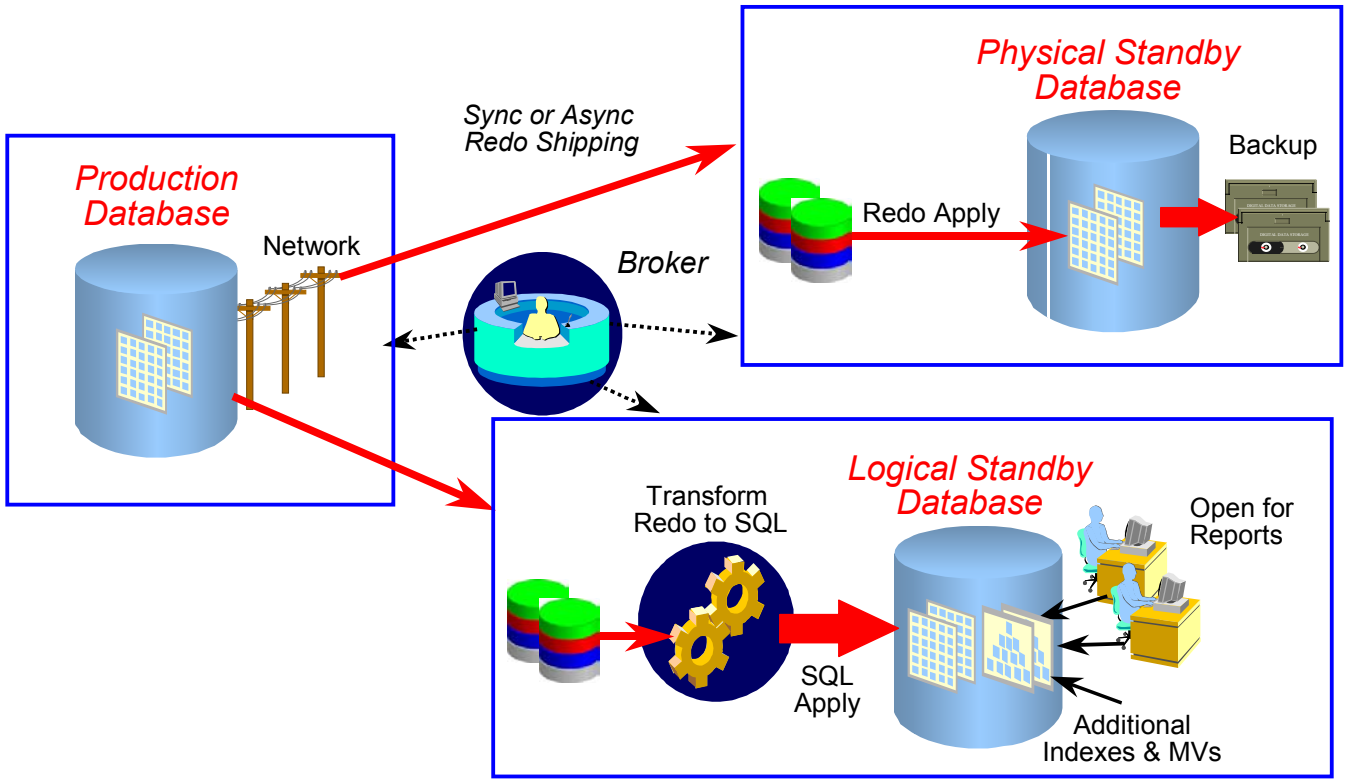
Logical replication



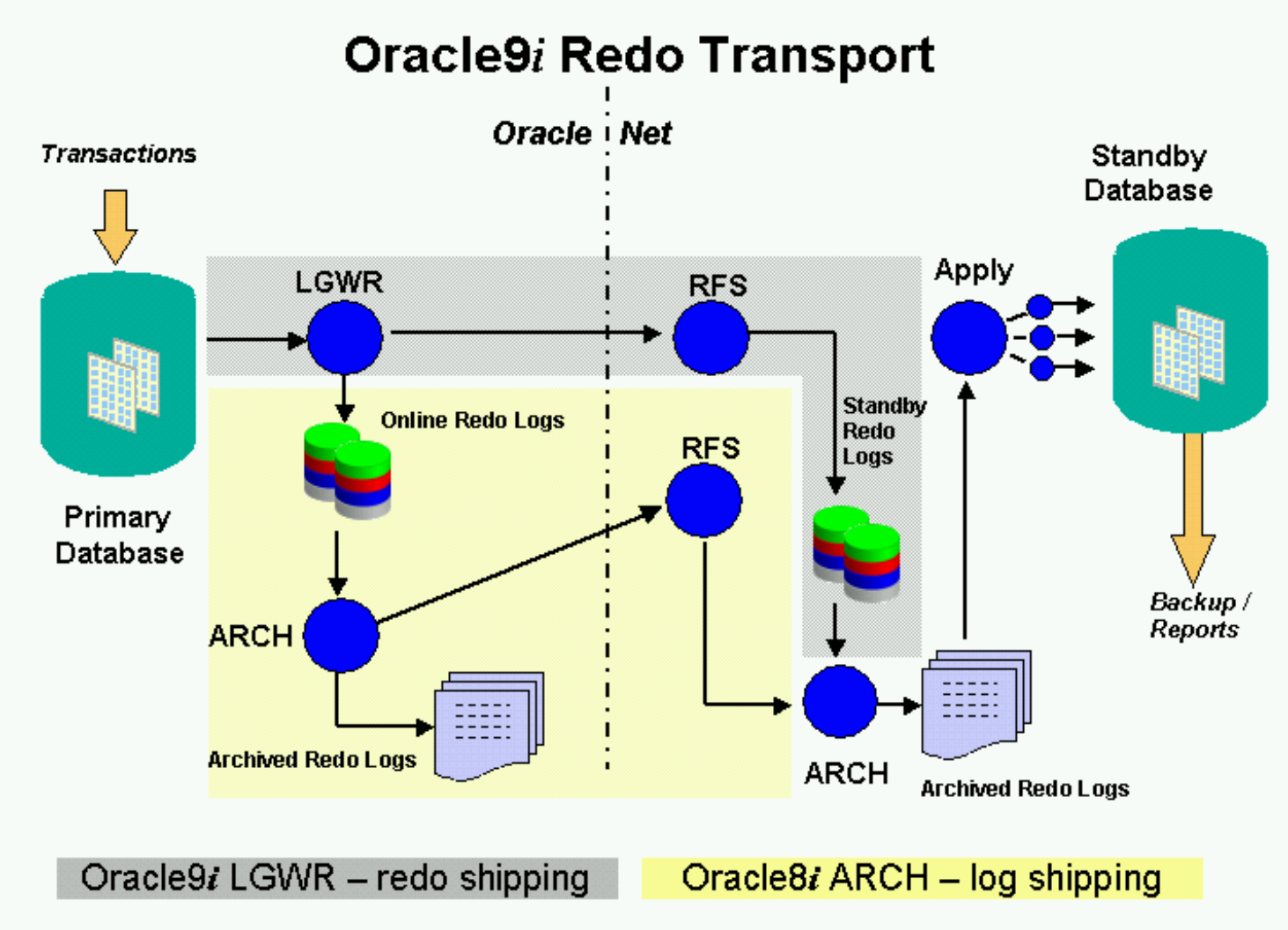
Oracle 9i and 10g DataGuard

- Oracle DataGuard is management, monitoring and automation software that logically replicates data from one Oracle production database to one or more Oracle standby databases
- Standby databases can be
  - physical (block-by-block disk DB structures)
  - logical (independent DB with the same data)
- Transactional consistency between primary and standby DBs is maintained by applying Oracle online redo logs for physical standbys and SQL statements for logical standbys
- Read-only (physical) and read-write (logical) standby DB access is possible (e.g., running reports)
- Integrated with Continentalclusters (contributed)

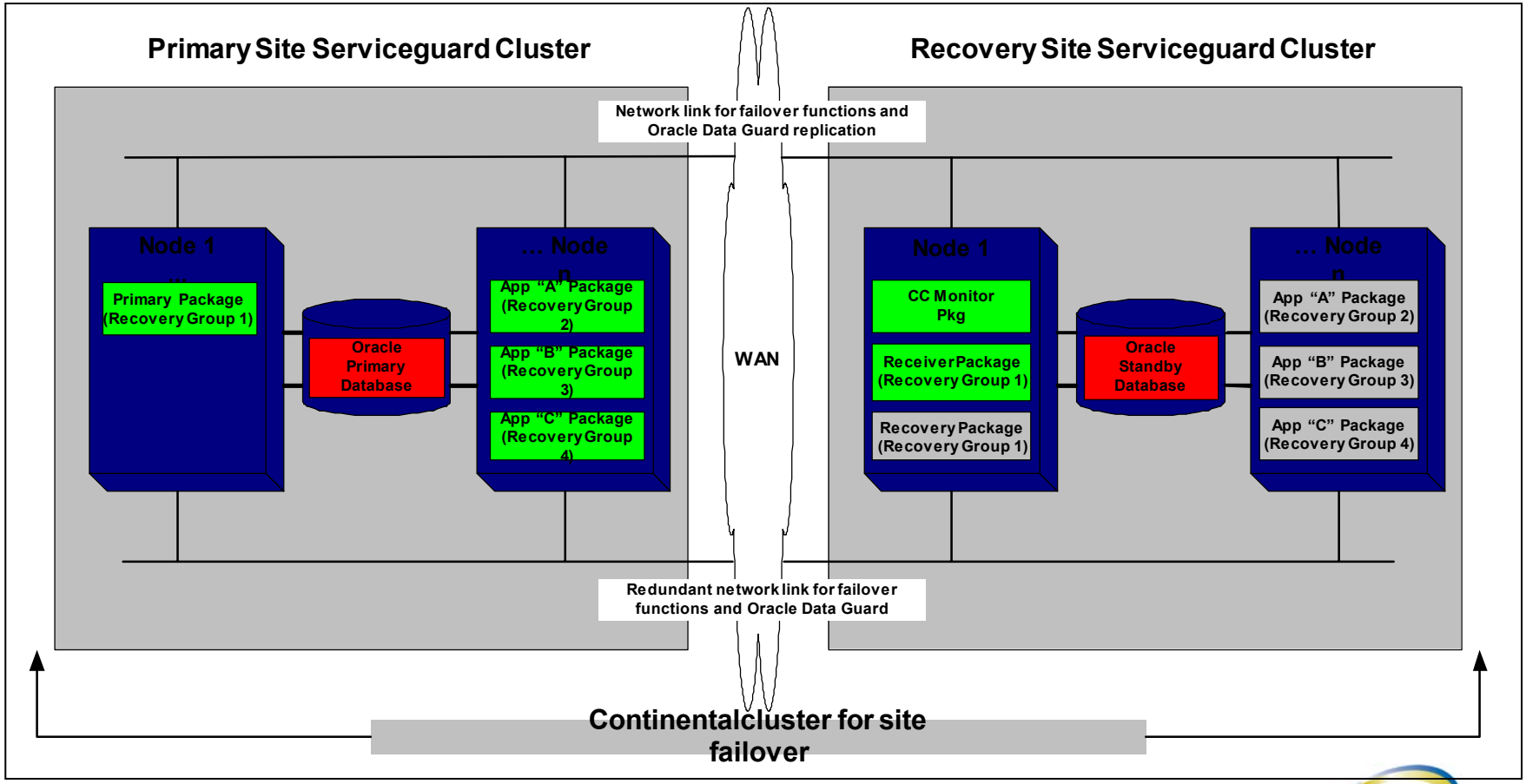
Oracle 9i Data Guard functional components



Oracle 8i versus 9i data replication



Example: Continentalcluster with Oracle 9i Data Guard



Quest SharePlex

- Quest SharePlex provides logical, log-based replication for Oracle databases:
- SharePlex captures modifications to the Oracle log and determines what changes need to be sent to the target system(s)
- Only DB changes are shipped to target systems (via TCP/IP), which reduces network bandwidth
- Target system receives data and uses an import process to construct SQL statements and applies the transactions to the target database instance
- Replication is supported between dissimilar platforms, OS and Oracle versions

Summary of data replication methods for HP Disaster Tolerant Solutions

- Extended distance Serviceguard/SGeRAC clusters
  - Software physical data replication using MirrorDisk/UX
- Metrocluster
  - Hardware physical data replication using XP CA or EMC SRDF
- Continentalcluster
  - Supported Hardware physical data replication using XP CA or EMC SRDF
  - Supported logical replication using Oracle 8i standby archive log file shipping
  - "Allow" model for customers to use any data replication method they choose (customer is responsible for data currency and consistency)

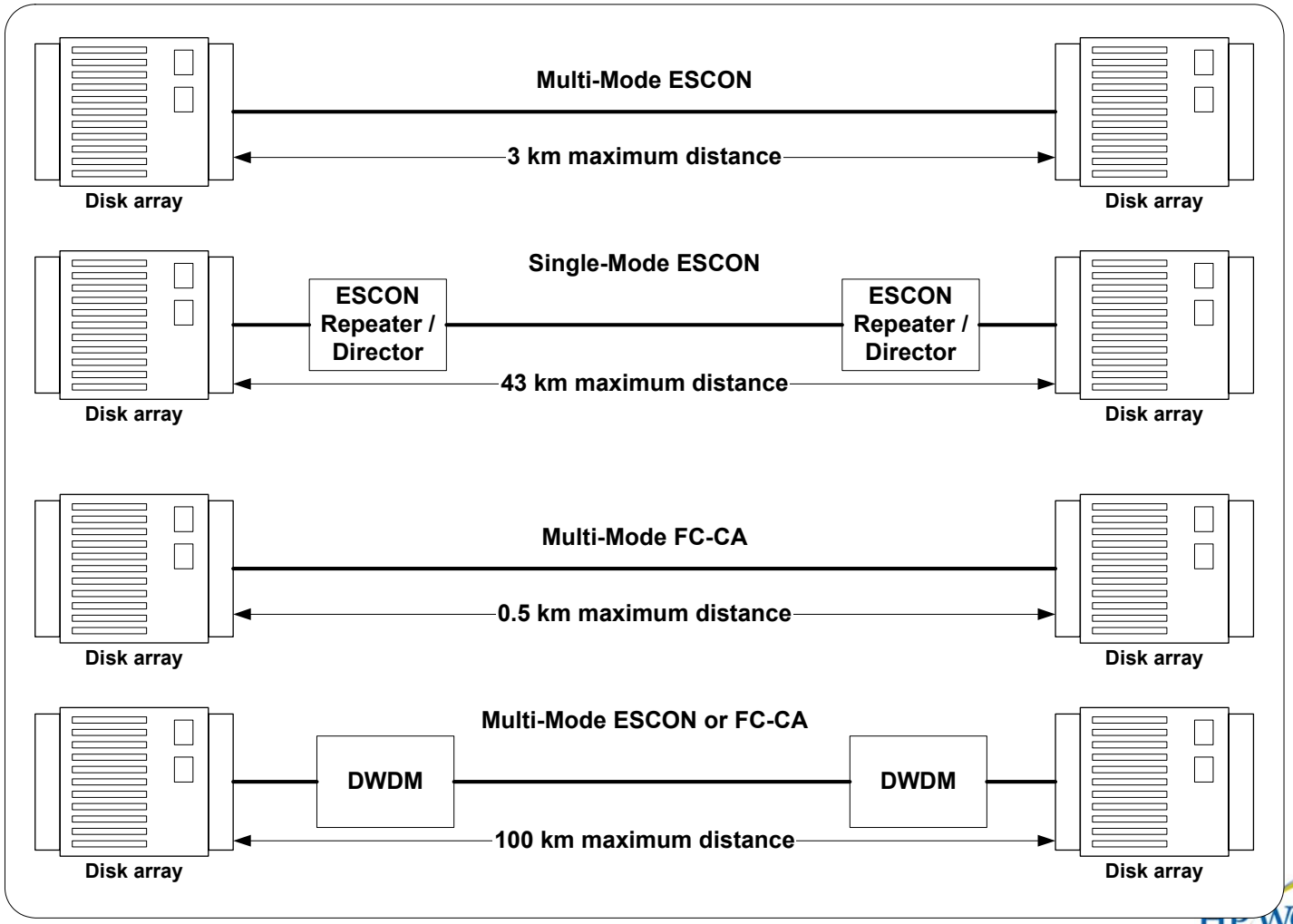
Network technologies used in disaster tolerant data replication solutions

- The distance between data centers will have an impact on the network technologies that can be used in a disaster tolerant design for data replication:
  - Short distances (campus – 100km)
    - Dense Wave Division Multiplexing (DWDM)
  - Long distances (100km – unlimited)
    - DS3 Leased Line
    - OC3 to OC12
    - Ethernet
    - ATM/SONET

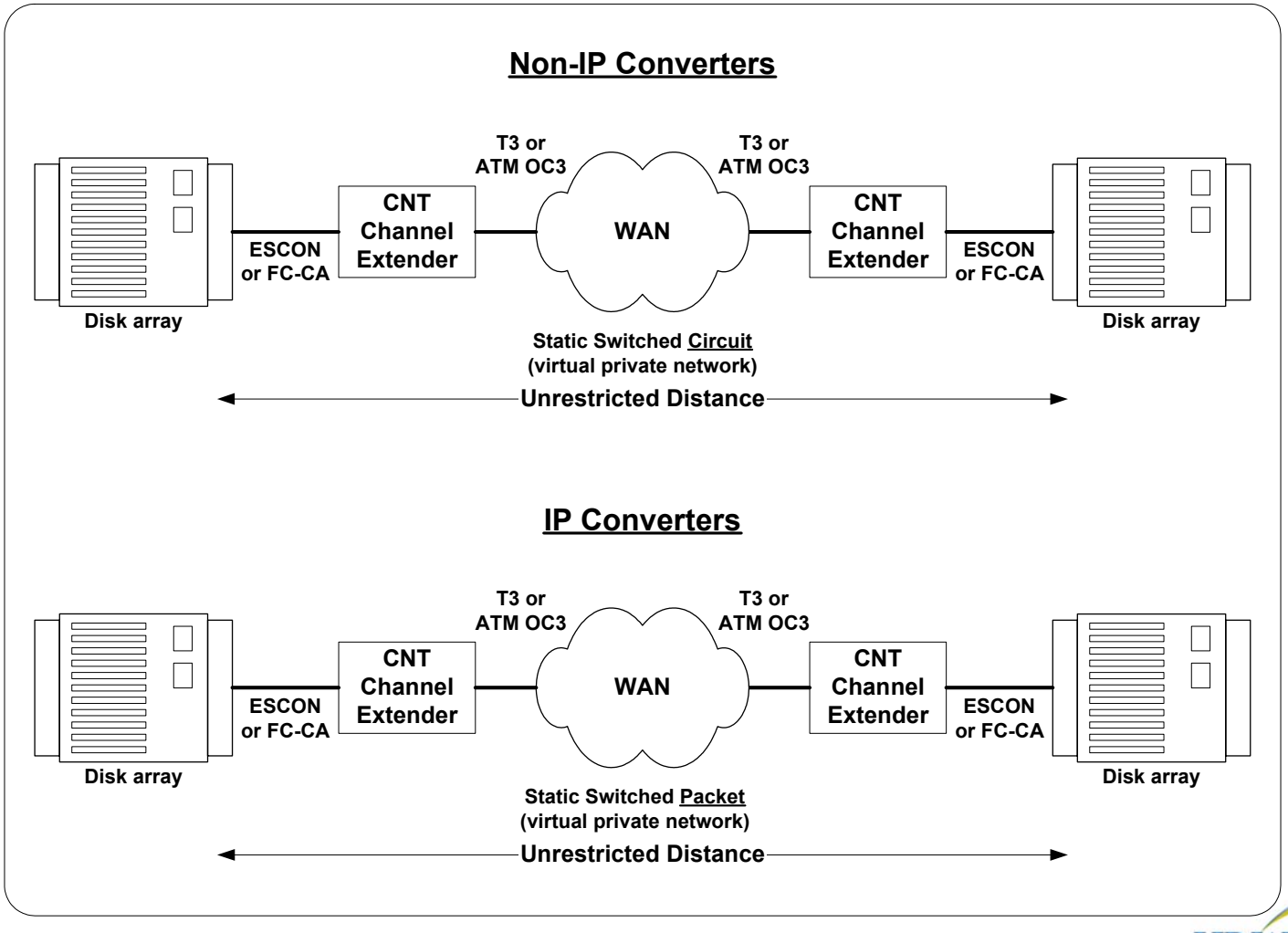
Network comparisons

Network attribute	Leased Lines (circuit switched)	ATM (cell switched)	IP (packet switched)
Typical bandwidth	T1 = 1.54 Mbps E1 = 2.0 Mbps E3 = 34 Mbps T3 = 45 Mbps	(Typically used over SONET networks) OC-1 = 45 Mbps OC-3 = 155 Mbps OC-12 = 622 Mbps OC-24 = 1.24 Gbps OC-48 = 2.48 Gbps	Router network can be WAN-connected using either Leased Lines or ATM
Typical network latency	Approximates the speed of light (1 millisecond per 100 miles)	Approximates the speed of light (1 millisecond per 100 miles)	Depends on the number of router hops through the network, which can add approx. 2 - 10 milliseconds per router

Data replication examples for short distances



Data replication examples for long distances



Factors affecting bandwidth pricing

- Geographic region
- Distance between end points
- Type of provider (IXC, ILEC or CLEC, older telco vs. new upstart)
- Circuit location
  - Within a telco region (LATA)
  - Between telco regions requires ILEC-to-IXC-to-ILEC pricing)
- Capacity and type of service (T1, DS-3, ATM, SONET, etc.)
- Volume purchase agreement
- Length of contract term
- Circuit redundancy
- Degree of managed services available
- Service Level Agreement (restore time, guaranteed performance level, uptime, etc.)

Bandwidth cost considerations

Dedicated bandwidth connection	Approximate U.S. street price (1,000 mile circuit, excluding local loop on each end)
T1 (1.54 Mbps)	$720/month
T3 (45 Mbps)	$15,500/month
OC-3 (155 Mbps)	$40,000/month
OC-12 (622 Mbps)	$120,000/month
OC-48 (2.488 Gbps)	$320,000/month

- Long-haul connection is approx. 1/3 of total connection costs, excluding local loops
  - For Europe, estimate 1.5x U.S. pricing
  - For South America, estimate 2x U.S. pricing (when available)
  - For Pacific Rim, estimate 2.5x U.S. pricing

Source: CNT, 2005

Data replication solution comparison (1)

Replication Solution	Replication Mode	Example	Cost	Recovery Time	Notes
Simple backup with off-site storage	Offline	Tape Backup	Least Expensive Solution $	Hours to Days	Minimum step to protect data
Software based local mirroring	Synchronous	MirrorDisk/UX	Inexpensive $$	Many minutes to hours	Higher data safety; little or no data loss; still requires manual recovery
Hardware-based local mirroring	Synchronous or Asynchronous	HP Business Copy and VA	Fairly Expensive $$$	Many minutes to hours	Higher data safety; little or no data loss; still requires manual recovery

Data replication solution comparison (2)

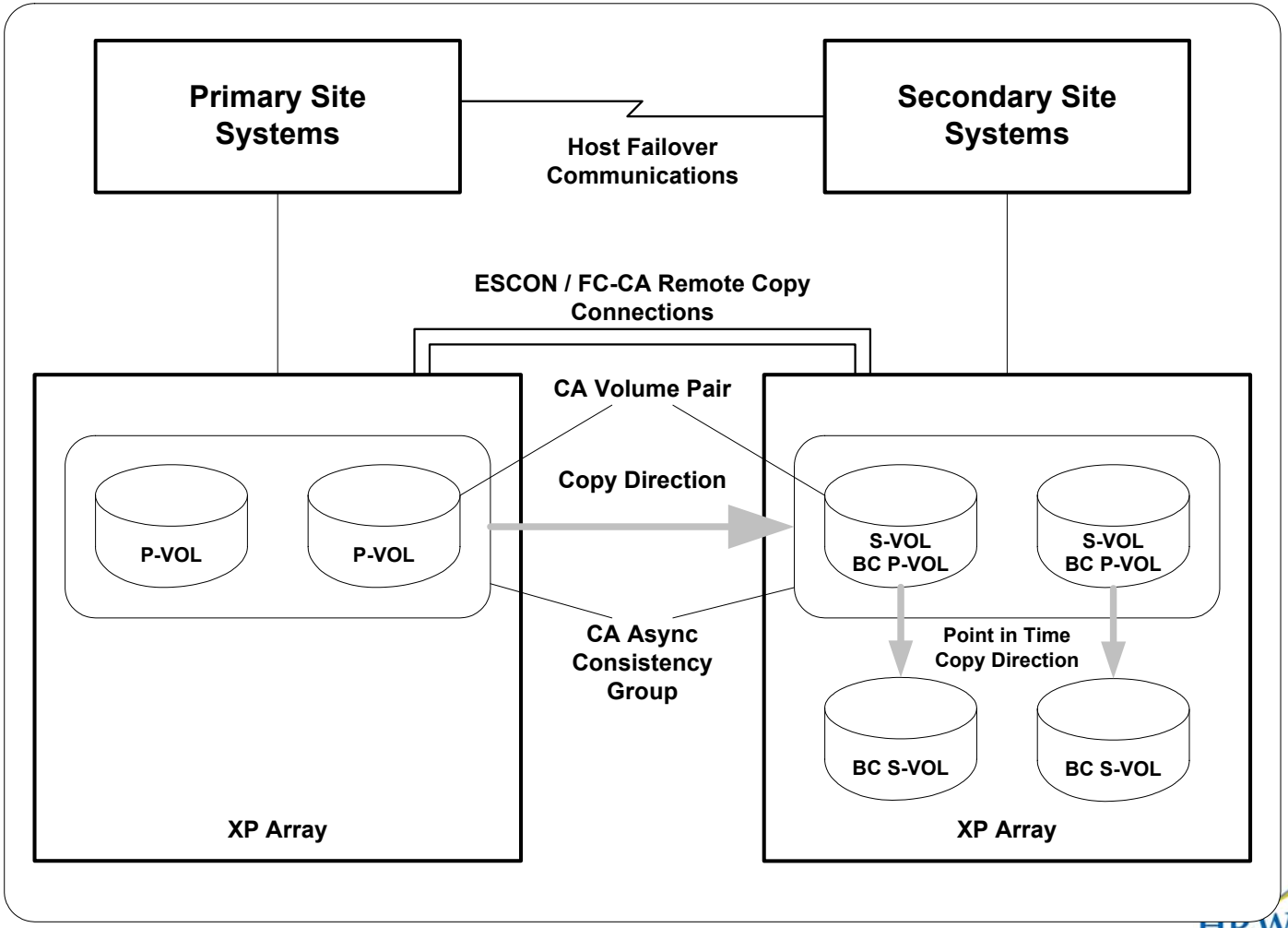
Replication Solution	Replication Mode	Example	Cost	Recovery Time	Notes
Extended distance cluster using software-based mirroring	Synchronous	MC/ServiceGuard Extended Distance Cluster with MirrorDisk/UX	Expensive with redundant hardware $$$$	A few to many minutes depending on solution	Very good data safety with rapid recovery
Geographically -dispersed cluster(s) using hardware-based mirroring	Synchronous or Asynchronous	HP MetroCluster or Continental Clusters	Most expensive with redundant hardware and additional software licenses $$$$$	A few to many minutes depending on solution	Eliminates unrecoverable data errors and minimizes downtime

Ideal solution

An "ideal" data replication solution (budget-permitting)

- For performance and data currency:
  - physical data replication
- For data consistency:
  - use a 2nd physical replication copy as a point-in-time snapshot
  - Use logical replication to protect against some types of database corruption

Example of an "ideal" data replication solution



Business Criteria for Selecting a Data Replication Method

- Recovery Time Objective (RTO) –
  - how quickly must the applications and services be running after a disaster?
- Recovery Point Objective (RPO) –
  - how much recent data must be available?
  - (conversely: how much data can be lost without re-entry?)
- Cost of implementation and operations versus value of the data
    - Inventory the current infrastructure
    - Protecting human resources
    - Hardware and network costs
    - Alternate office space
    - Cost of operation

Design Criteria for Selecting a Data Replication Method

- Backup methodology
  - Minimizing planned downtime
  - Handling open files
- Data consistency and currency
- Bandwidth between sites
- Communication link speed
- Distance between sites
- Implementation assistance
- DR hardware ownership

High Availability Reference Architectures

High Availability Reference Architectures

- Focused on the Complete Solution
  - Multi-tier, multi-OS (HP-UX, Linux & Windows 2000), highly available, and scalable integrated architectures
  - Defined for applications in emerging markets
  - May involve multiple HA solutions (not just ServiceGuard)
  - Initially based on best currently available technology
  - Ultimately, based on improvements in HP and/or ISV software
  - Customer references, where appropriate
  - Examples:
    - Broadvision
    - Oracle CRM
    - Siebel eBusiness Suite
    - Telco Billing (Amdocs Ensemble)

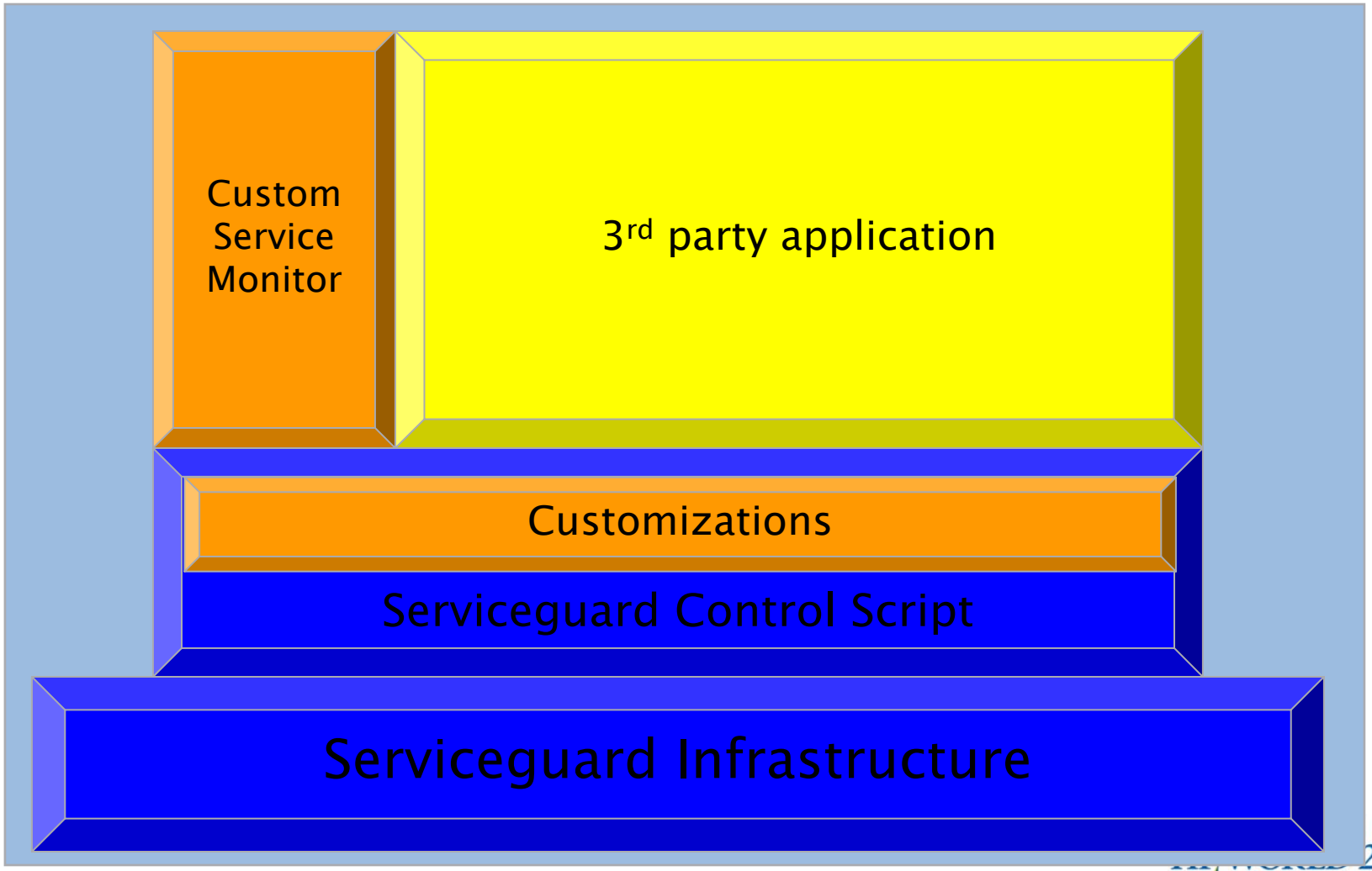
Integrated architectures

- Researched, verified and documented
  - Provides examples of known, proven, pre-tested architectures that can be purchased quickly and with confidence
  - Provides implementers with a documented blueprint for hardware and software architecture and design
- Multiple architectural choices to meet several price points that tradeoff availability, cost, and/or performance
- Ultimately, define improvements in the existing HP and ISV solution by identifying and incorporating new technologies for:
  - Elimination of Single Points of Failure in the hardware and software architecture, if they exist
  - Improving failover times
  - Preserving state during failover
  - Overall goal: improve availability

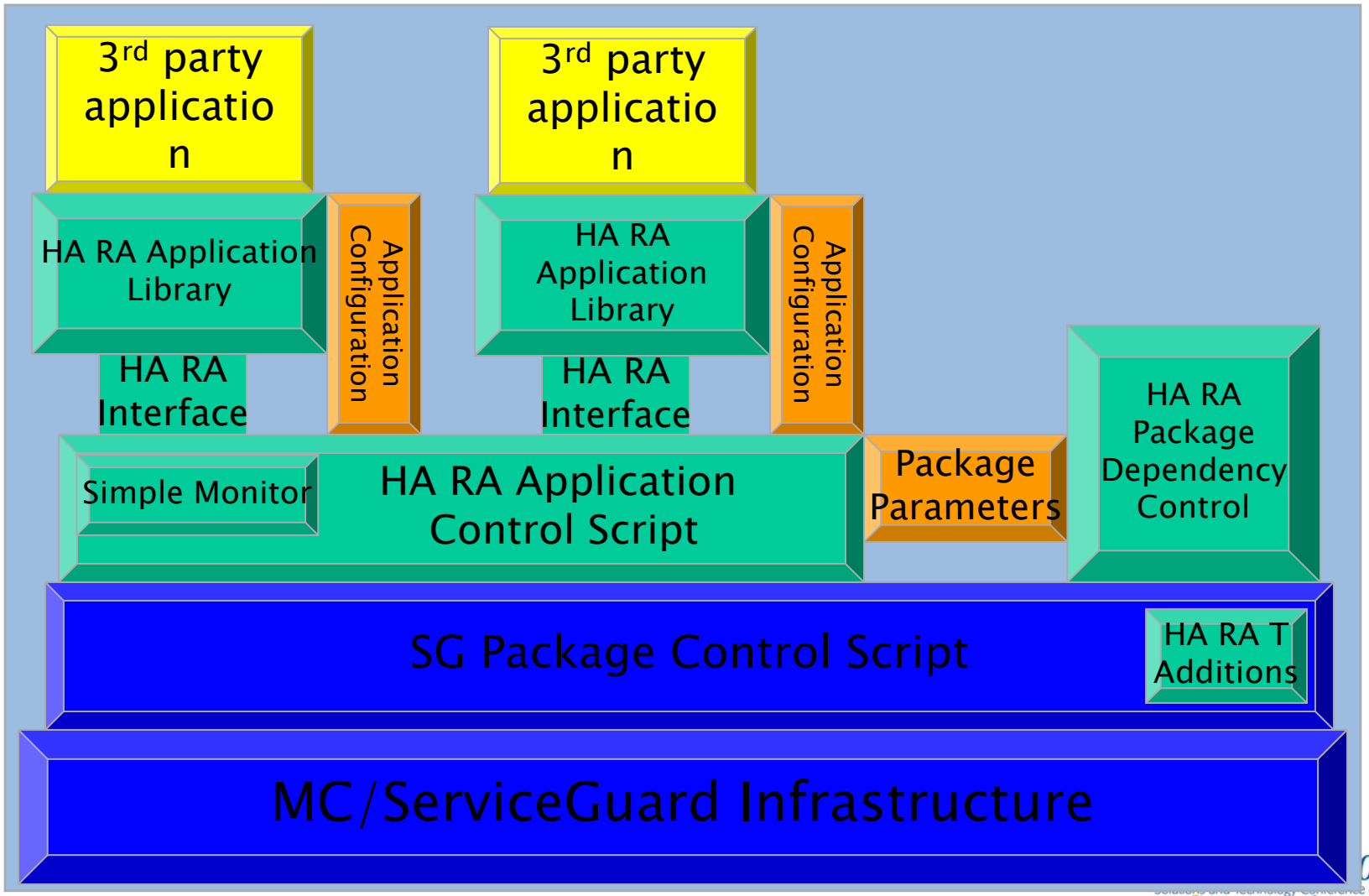
HA RA Toolkit Overview
Features

- Provides package dependencies:
  - Change in state on one package can affect other packages in the cluster
  - Allows applications to be restarted if a service they depend on is restarted
  - Allows packages that compete for the same resources to be prevented from starting on the same system
- Enhanced application monitoring:
  - Improvements to the functionality of the service monitor to allow easier implementation of application monitors
  - Improved restart attempt capability within the monitor
  - Concept of maintenance mode for easily taking an application offline and online to better support installation, upgrade, and maintenance activities

Traditional application integration with Serviceguard



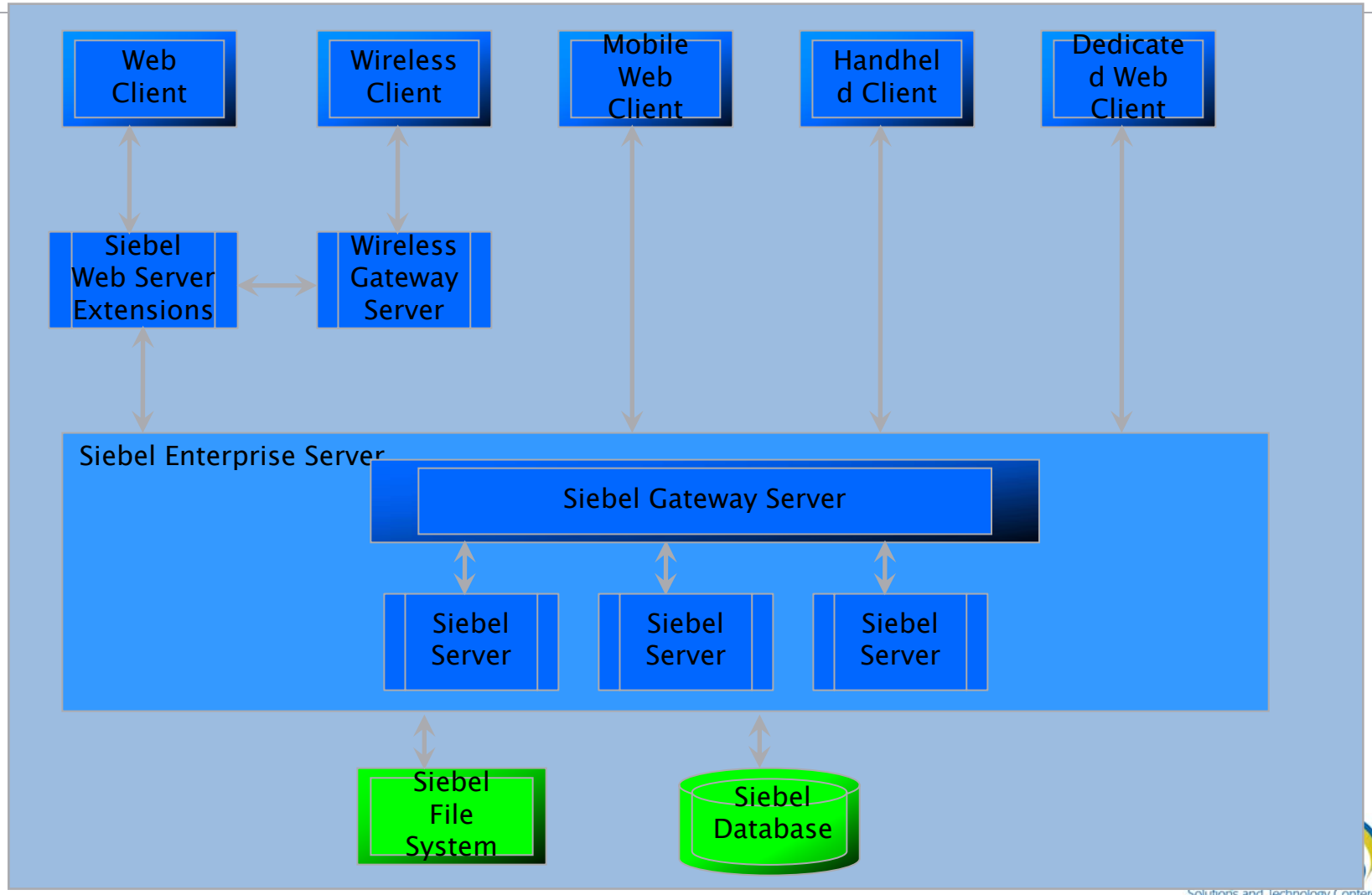
Application integration using HA RA Toolkit



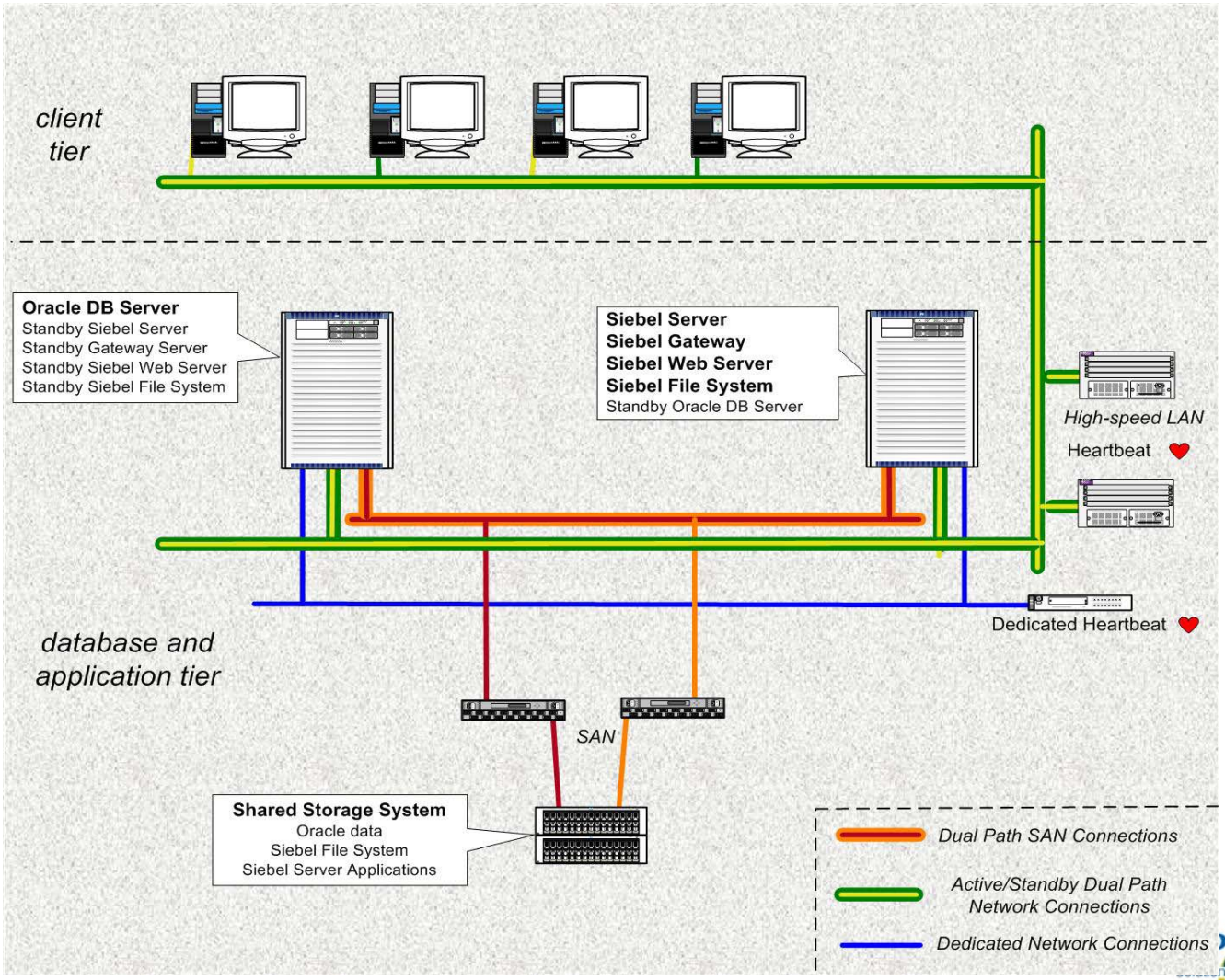
High Availability Reference Architecture: Siebel eBusiness Suite

- Siebel eBusiness applications are a suite of products enabling enterprise class Customer Relationship Management in the areas of multi-channel sales, marketing, service, and partner and employee relationship management.
  - Full Siebel implementations could conceivably include integrating over one hundred application components
  - Components include both Siebel applications as well as partner applications
  - Many different implementation architectures available with differing levels of integration with Serviceguard
  - Minimum configurations include at least five components

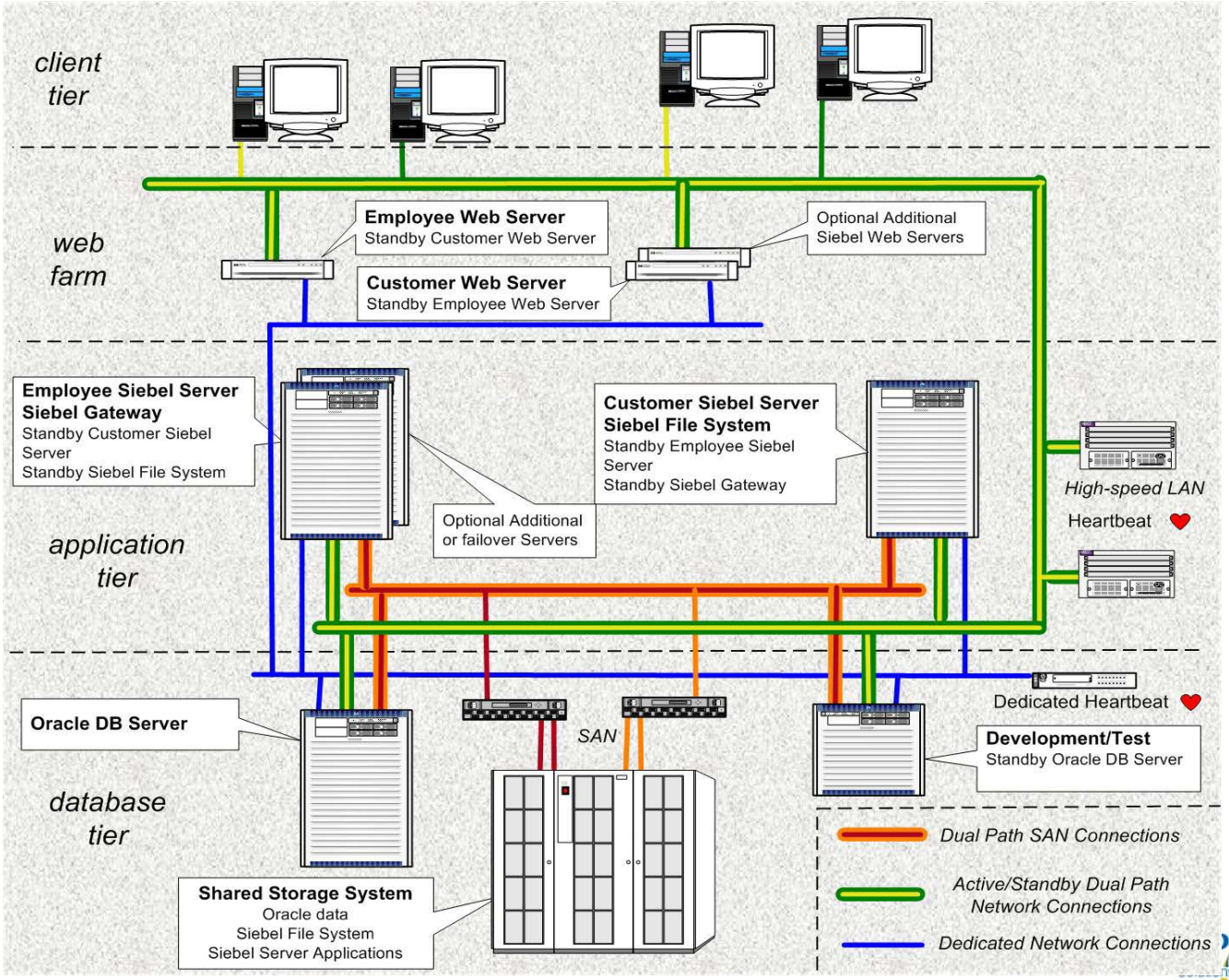
HA Ref Arch: Siebel eBusiness Suite
Siebel Architecture Overview



HA Ref Arch: Siebel eBusiness Suite Sample Configuration - Simple



HA Ref Arch: Siebel eBusiness Suite Sample configuration - Complex



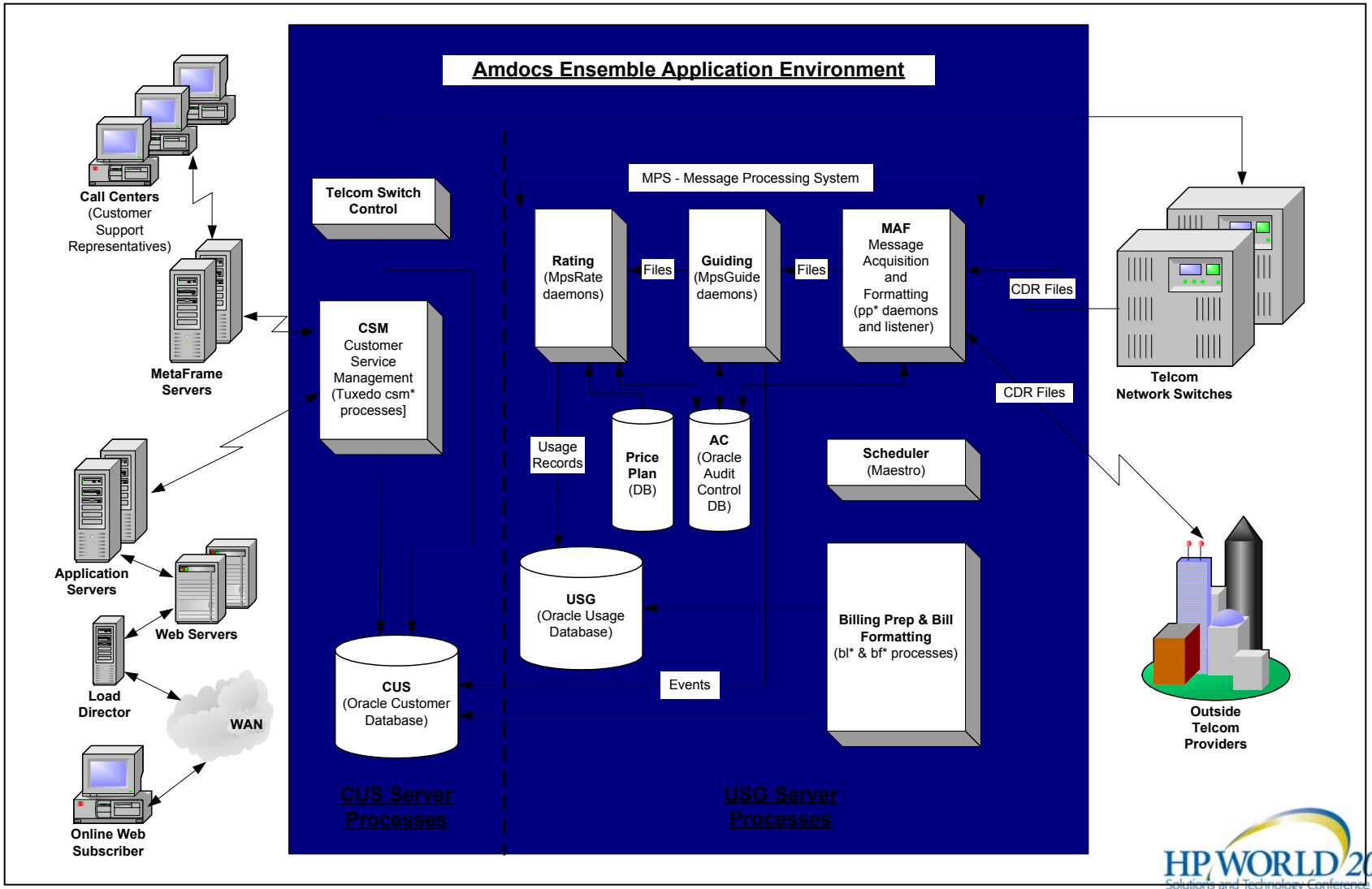
HA Reference Architecture: Amdocs Ensemble telcom billing

Telcom billing outages can impact Telcos in the following areas:

- Business operations
  - Collections / payment processing
  - Usage processing
  - Bill processing
  - Service activation
  - Customer service
  - Fraud detection
- Revenue realization
  - Cash flow
  - Debt management
  - Month-end close of books
  - Revenue leakage
  - Third Party settlement

- To mitigate these business risks, a HA / DT Reference Architecture was developed for the Amdocs Ensemble telcom billing application and is currently sold as the HP Business Continuity for Billing (BCB) Solution using HP HA and DT solution products (SG, Continentalcluster, SG Manager)

Amdocs Ensemble application architecture



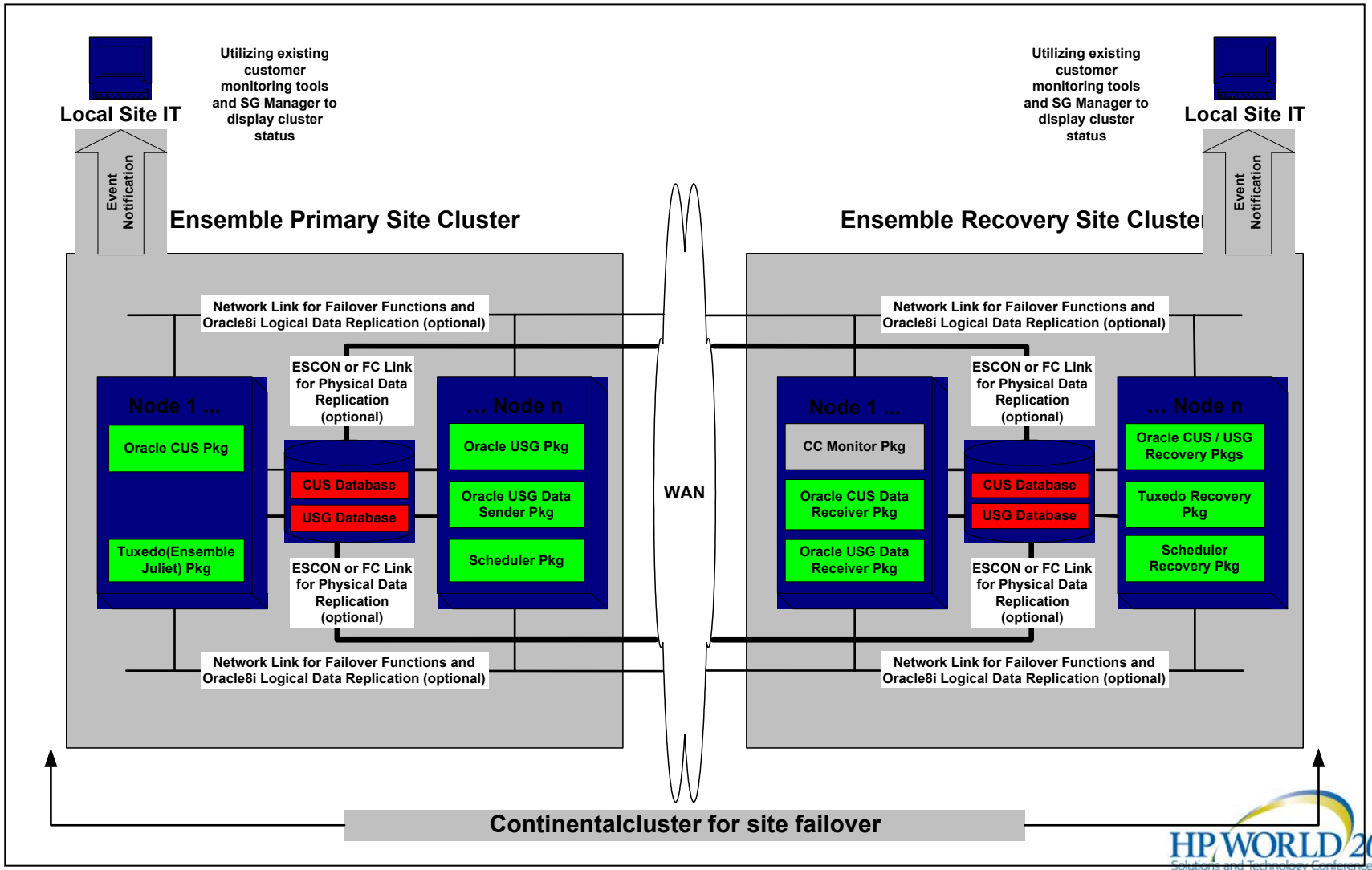
Serviceguard failover for Customer Care processes (Customer database)

- BCB HARA Toolkit scripts provide coordinated failover of Customer Care – related processes:
  - Oracle Customer database instance
  - Tuxedo (Amdocs Juliet) middleware
  - Oracle Switch Control database instance
  - Switch Control programs
  - Any customer-defined processes
- Local failover of the Serviceguard packages for these processes protects the Customer Care portion of Ensemble
- Serviceguard minimizes downtime for users accessing highly visible customer front-end business functions (e.g., activating new services, deactivating services, etc.)

Serviceguard failover for Billing processes (Usage databases)

- BCB HARA Toolkit scripts provide coordinated failover of billing – related subsystems:
  - Oracle Usage database instances
  - Maestro (or other) scheduler
  - Any customer-defined processes
- Local failover of the Serviceguard packages for these processes provide for restarting basic Ensemble subsystems used in bill processing (e.g., usage database instances, schedulers)
- All programs related to the Ensemble Message Processing System (MPS), bill preparation and formatting are not protected by Serviceguard (Amdocs recommendation)
- Full recovery after a failover requires a manual examination of the billing cycle processing at the point of the failure to determine the best point to restart processing (Amdocs recommendation)

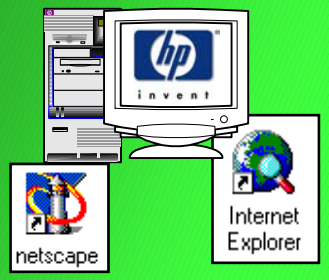
Detailed Ensemble Reference Architecture



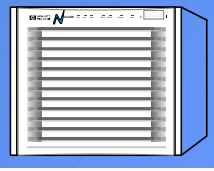
HA Reference Architecture example: Oracle CRM

- All Oracle applications are built on the same Oracle tools
  - Oracle Applications Server
  - Oracle Forms and Reports Server
  - Oracle Discoverer Server
  - Oracle Database Server

Oracle CRM 11i logical architecture



Browser



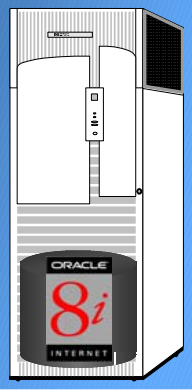
Application Server — Apache 1.3.9

Forms Server — Forms 6i, Graphics 6i

Concurrent Processing Server — Reports 6i

Discoverer Server — Discoverer 3i

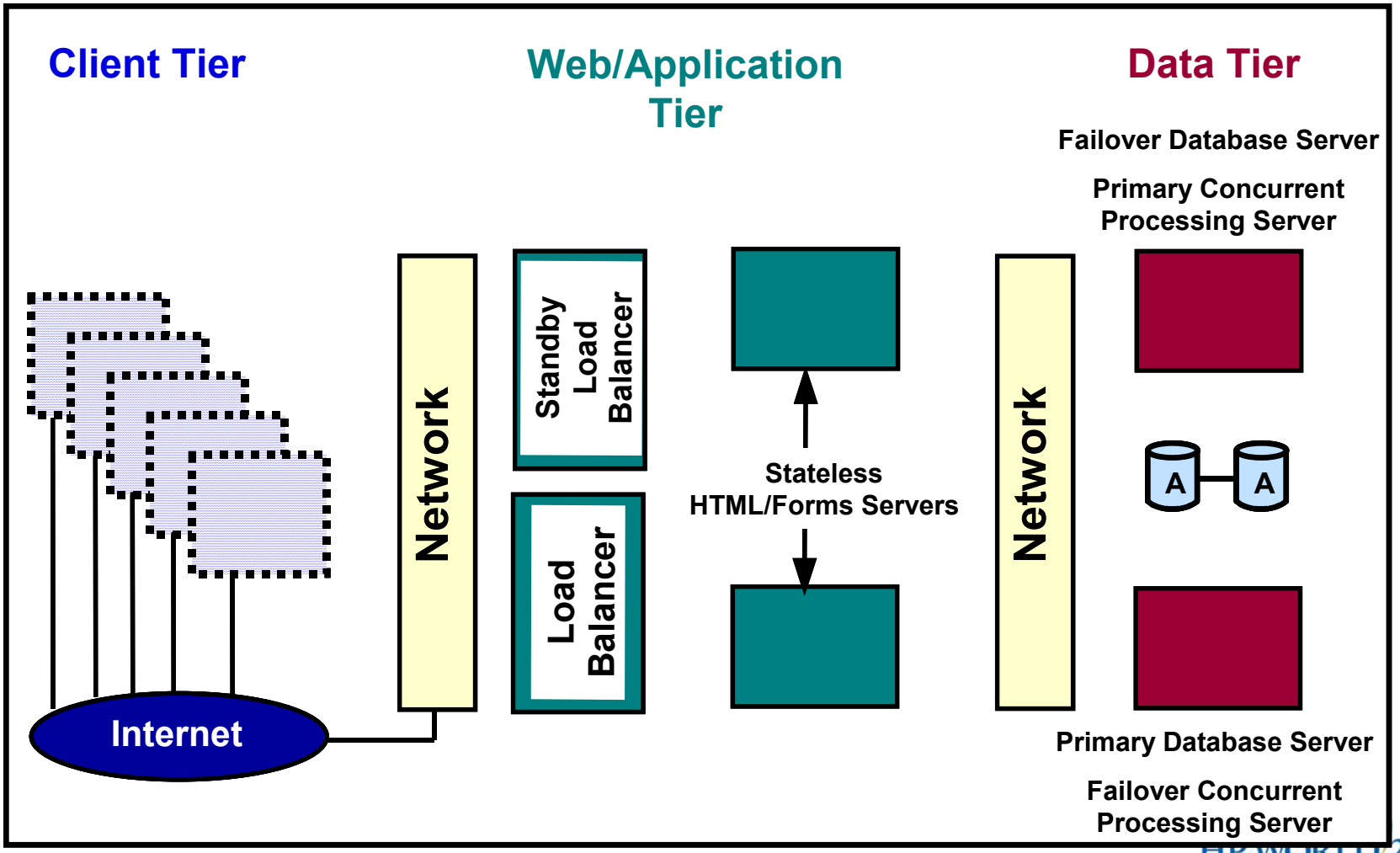
Reports Server — Reports 6i, Graphics 6i



Data Server

Oracle 8i

Oracle CRM example HA Reference Architecture



High Availability Solutions in HP-UX 11i v2 UD2

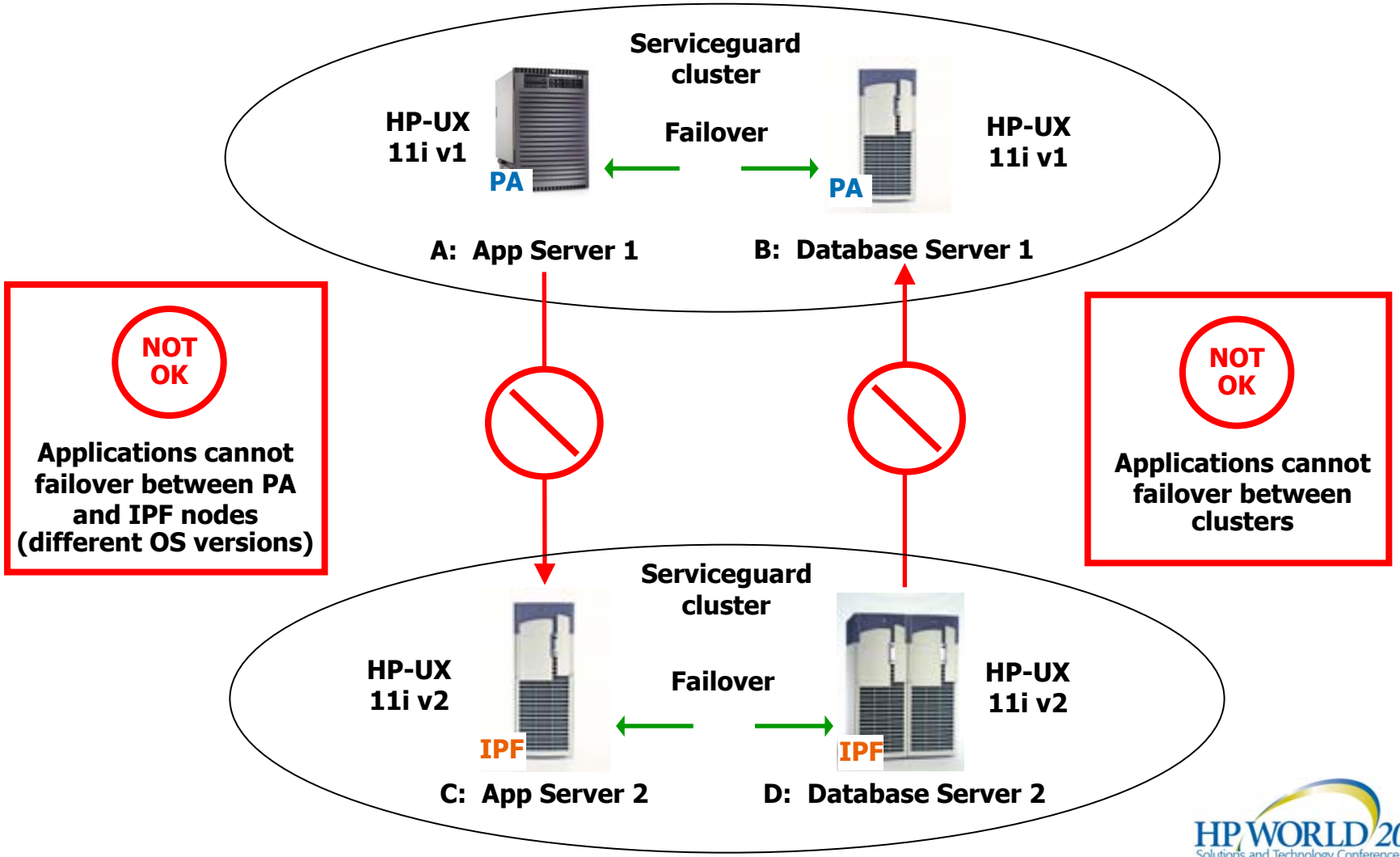
High Availability Solutions in HP-UX 11i Version 2 Update 2

- Planned to ship November 2004
- Support for PA-RISC and Integrity (IPF-based) servers
  - Single OS with the same features
  - Serviceguard 11.16
  - Metrocluster and Continentalclusters
  - Both families of servers in the same cluster

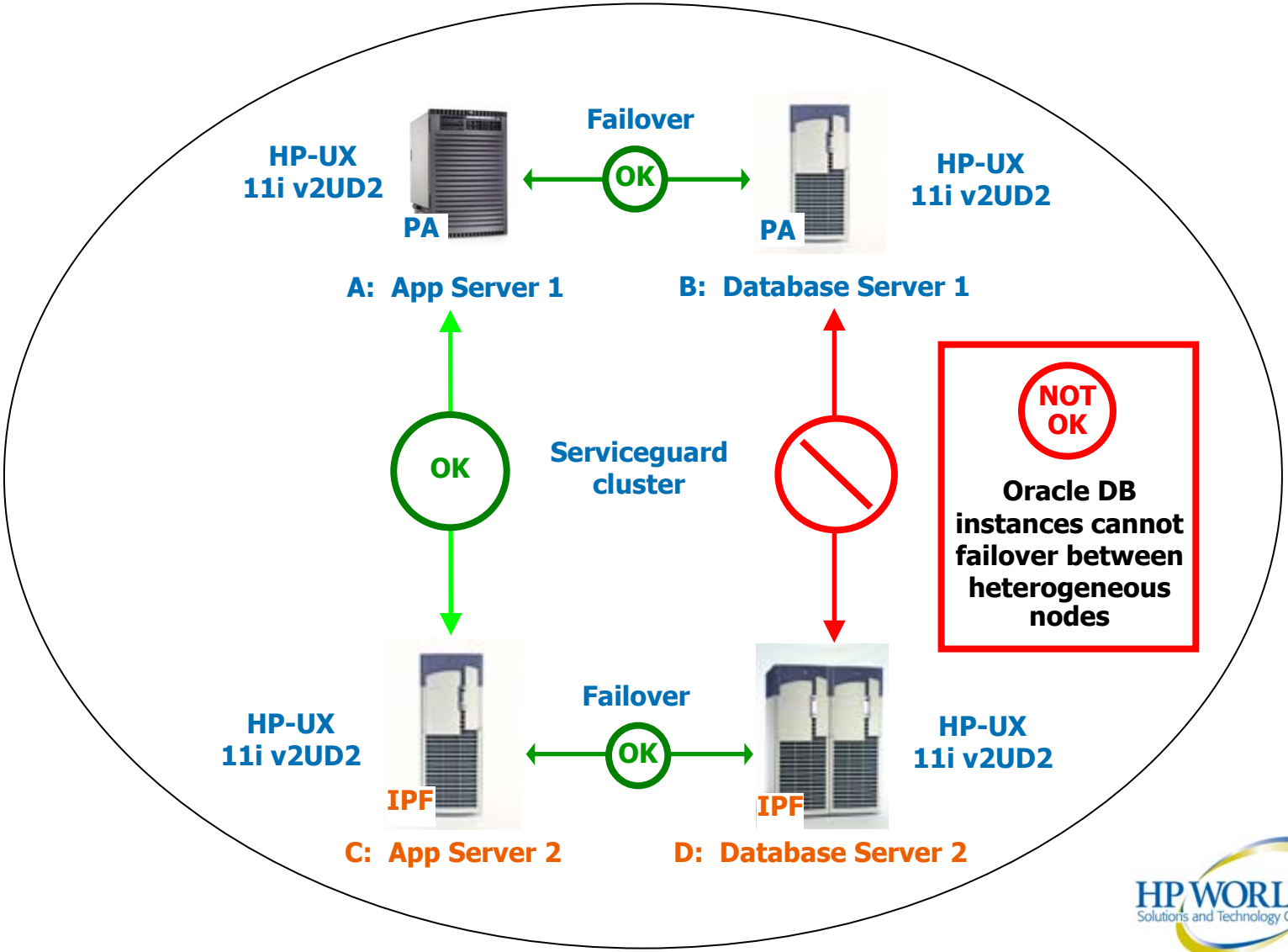
New!

  - Technical Whitepaper will be available to fully describe the supported architectures and restrictions

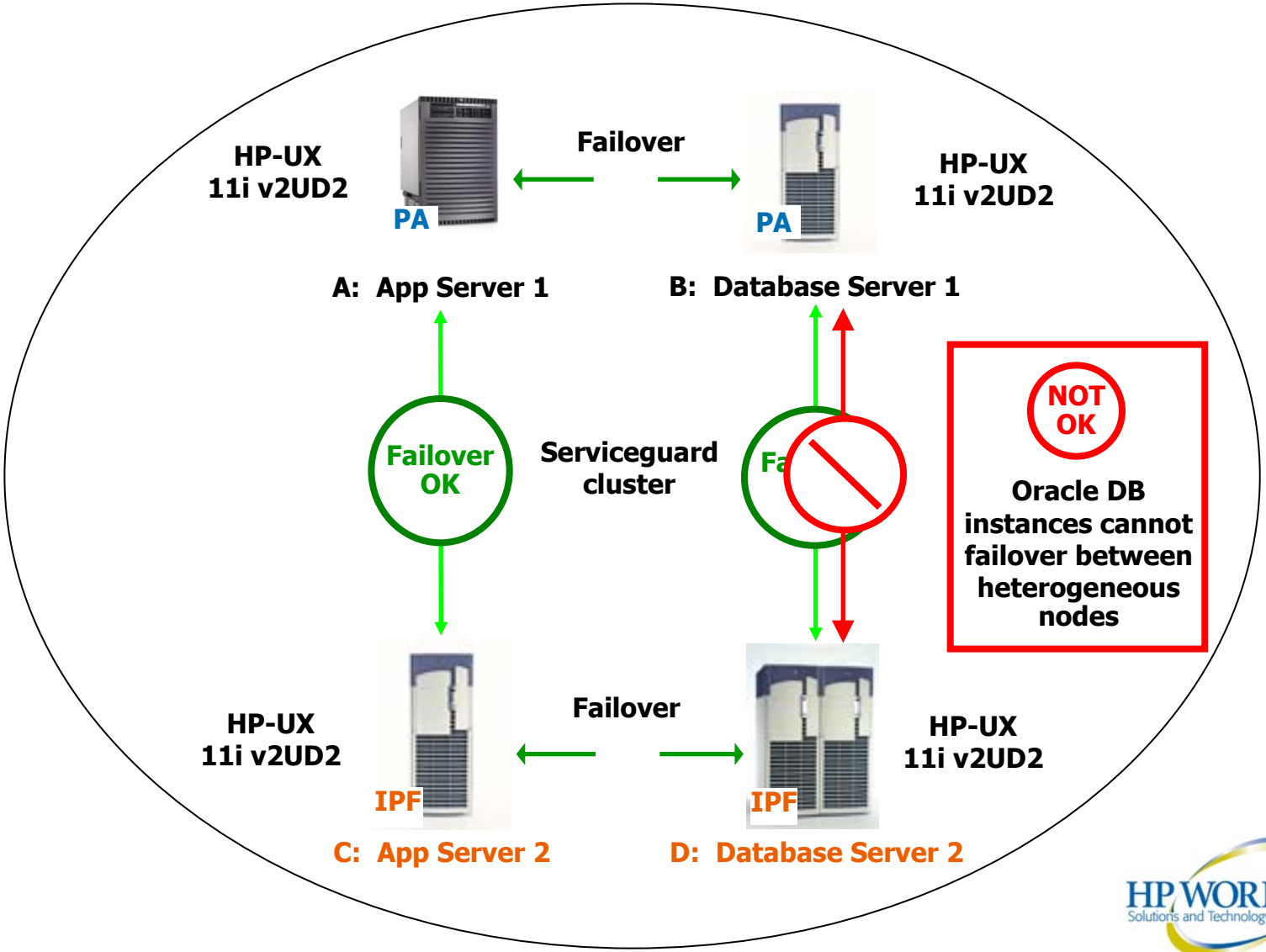
Mixed PA and IPF Clusters are Not Supported Today



Mixed PA and IPF system clusters tomorrow (Q4'2004)



Mixed PA and IPF system clusters tomorrow (Q4'2004)



Mixed cluster advantages

- Simplified administration for both PA and IPF-based systems
  - One common OS (HP-UX 11i v2UD2)
  - one common cluster software (Serviceguard 11.16)
- Easy addition of IPF-based systems to existing PA-based SG clusters
- Ability to gain experience with IPF-based systems in known environments
- Transition from PA- to IPF-based systems can be more easily accomplished

Native Executables and Aries

- Normally, executables are compiled for a particular architecture
  - PA executables for use on a PA-based system
  - IPF executables for use on an IPF-based system
  - Magic numbers in the a.out file are used to code the executable for a particular architecture

- Aries is a binary emulator included with HP-UX 11i v 2
  - transparently emulates 32-bit and 64-bit HP-UX PA-RISC executables on HP-UX Itanium®-based machines
  - HP-UX PA-RISC applications do not have to be recompiled
  - The user does not explicitly invoke Aries
    - The HP-UX kernel running on an IPF system recognizes an HP-UX PA-RISC executable and automatically invokes Aries to emulate the application without any user intervention
    - Note: Application performance under Aries emulation may vary

Aries application emulator

- Consists of 4 shared libraries
  - /usr/lib/hpux32/aries[32/64].so
    - 32/64-bit Aries application emulator
  - /usr/lib/hpux32/pa_boot[32/64].so
    - Loads and executes the appropriate emulator when the HP-UX Itanium-based kernel recognizes a HP-UX PA-RISC executable
- Supports all HP-UX PA-RISC applications
  - Debugging of PA applications on IPF systems using PA gdb
  - Inter-process communication (IPC)
    - Semaphores, shared memory, sockets, etc.
    - Signal/exception behavior
    - IPC between emulated PA applications and native IPF applications
- Includes performance-enhanced Dynamic Translator (beta)
  - Beta translator not invoked by default (enabled via resource file)
  - Measured translation time improvement by 20%, on average
  - Some risk associated with using beta translator on mission-critical apps

Notes on Aries emulation

- Aries PA-RISC emulation performance will vary depending on the application
- Aries included with HP-UX 11i v2 supports PA applications running on HP-UX 11i v1.6 and below
- No support for PA programs that load IPF shared libraries
- Does not support privileged PA-RISC instructions (device drivers, loadable kernel modules)
- Does not guarantee correct emulation of PA applications that make assumptions about the time required to execute certain sections of application code or system calls
- Emulated processor-related system calls return information pertinent to a PA-RISC 2.0 processor
- See Aries(5) man pages and release notes for detailed information on Aries emulation

Requirements for mixed PA/IPF cluster implementations

- All nodes must have HP-UX 11i v2UD2 installed
- All nodes must have Serviceguard 11.16 installed
- The applications are configured to follow the ISV's configuration requirements:
  - Failover between homogeneous nodes in a mixed cluster only, OR
  - Failover between heterogeneous nodes in a mixed cluster using either of the following (not all ISVs will support this):
    - One set of PA binaries running native on PA and the same set using the Aries emulator on IPF, OR
    - Two sets of native binaries (one for PA, one for IPF)
      - either all 32-bit or all 64-bit binaries

Supported servers in a mixed PA/IPF Serviceguard cluster

- PA-RISC-based servers (PCI or PCI-X based backplane only):
  - Entry Level: rp24xx, rp34xx, rp44xx
  - Mid Range: rp54xx, rp74xx, rp84xx
  - High End: Superdome
  - Current A-, L- and N-Class servers
- IPF-based servers:
  - Entry Level: rx16xx, rx26xx, rx46xx, [rx56xx]
  - Mid Range: rx76xx, rx86xx
  - High End: Superdome

Supported storage in a mixed PA/IPF Serviceguard cluster

Disk Arrays:	Interface
XP48	1Gb FC
XP128	2Gb FC
XP128	1Gb iSCSI
XP256 – Fibre Channel ports	1Gb FC
XP512	1Gb FC
XP1024	2Gb FC
XP1024	1Gb iSCSI
XP enhancement	2Gb FC
VA7100	1Gb FC
VA7400	2Gb FC
VA7110	2Gb FC
VA7410	2Gb FC
FC60	1Gb FC
MSA1000	2Gb FC
MSA1500	2Gb FC
SabreJet	1Gb iSCSI
MA8000 (HSG80)	1Gb FC
EMA 12000 & 16000 (HSG80)	1Gb FC
EVA1500 XL	2Gb FC
EVA5000 (HSV110)	2Gb FC
EVA3000 (HSV100)	2Gb FC
JBODs:	
SC10	LVD SCSI U2
FC10	1Gb FC
DS2100	LVD U160
DS2120	LVD U320
DS2300	LVD SCSI U160
DS2405	2Gb **FC-AL only**
Storage works 4314T & 4314R	LVD U160
Storage works 4354R	LVD U160
Storageworks Northstar	LVD U320
Storageworks Southern Cross	LVD U320

SAN Switches and Hubs	
B Series (Brocade)	
Storageworks SAN switch 8-EL	1Gb FC
Storageworks SAN switch 16-EL	1Gb FC
Storageworks SAN switch /8	1Gb FC
Storageworks SAN switch /16	1Gb FC
Storageworks SAN switch Intregrated 32	1Gb FC
Storageworks SAN switch Intregrated 64	1Gb FC
Storageworks SAN switch 2/8-EL	2Gb FC
Storageworks SAN switch 2/8	2Gb FC
Storageworks SAN switch 2/16-EL	2Gb FC
Storageworks SAN switch 2/16	2Gb FC
Storageworks SAN switch 2/32	2Gb FC
Storageworks Core Switch 2/64	2Gb FC
M Series (McDATA)	
McDATA ED5000	1Gb FC
EMC Connectrix DS-16B	1Gb FC
EMC Connectrix ED-1032	1Gb FC
Storageworks SAN Director 64	1Gb FC
Storageworks Edge switch 2/32	2Gb FC
Storageworks Director 2/64	2Gb FC
C Series (CISCO)	
MDS 9216	2Gb FC
MDS 9509	2Gb FC
SR2122-2 iSCSI storage router	iSCSI
Other Switch/Hub	
Surestore L10/S10 hub	1Gb FC

Supported software in a mixed PA/IPF Serviceguard cluster

- HP-UX UX 11i v2UD2
  - Mission Critical Operating Environment (MCOE) installed on all nodes
  - A non-MCOE 11i v2UD2-based system with Serviceguard 11.16
- Same or similar patch set for OS, Serviceguard and ISV's applications
- Same version of volume manager software (if used for shared storage)
  - LVM
  - VxVM
  - CVM
- Same version of file system software (if used for shared storage) Note: The default file system layout versions of VxFS will differ based on the OS release:
  - VxFS created on 11i v1 use file system version 4 and is compatible with older layouts
  - VxFS 3.5 created on all 11i v2 use file system version 5 and is compatible with version 4
  - A migration tool is available to convert version 4 to version 5
- nPARs
  - No special mixed cluster requirements on Serviceguard configurations with nPARs
- vPARs
  - Currently, Serviceguard is supported with vPARs on PA systems only for 11i v2UD2
  - A new version of vPARs should be available in 1H05 on 11i v2 that will add vPAR support for IPF systems
  - Configurations with Serviceguard and virtual partitions within the same server are not recommended
- Campus/Extended Cluster, Metrocluster and Continentalclusters (subject to constraints above)

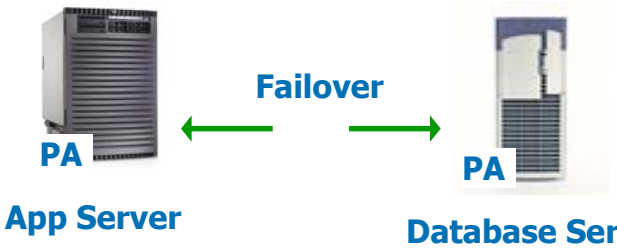
Note: SGeRAC in mixed clusters are not currently supported (under investigation for a future release)

Application code support in mixed clusters

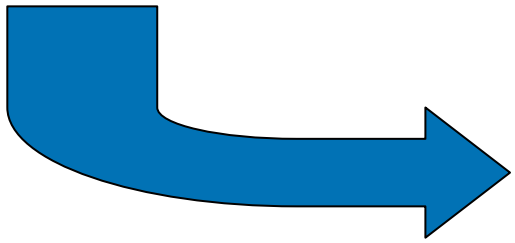
- HP-UX 11i v2UD2 supports applications that have been compiled in 32-bit mode as well as 64-bit mode on both PA and IPF systems

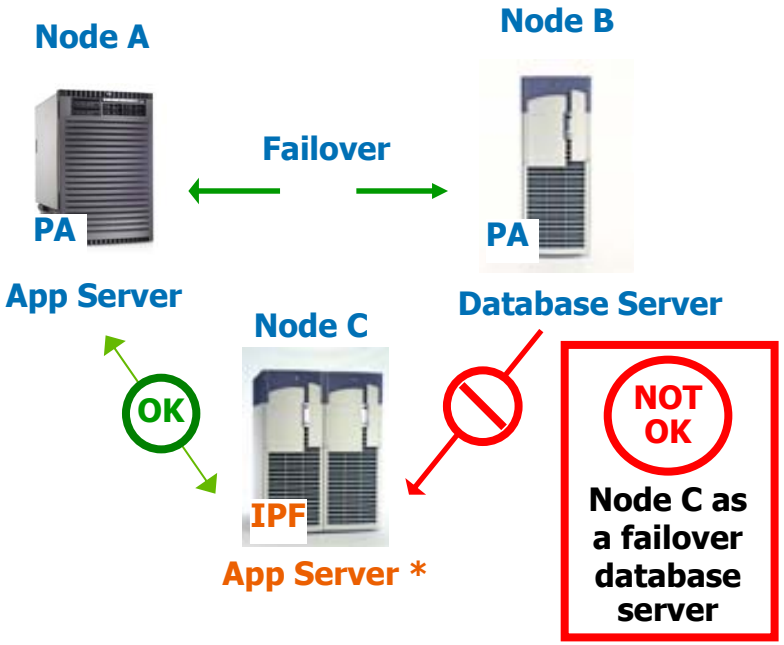
	PA systems with HP-UX 11.i v2UD2	IPF systems with HP-UX 11.i v2UD2
PA 32-bit application	Supported	Supported with Aries
PA 64-bit application	Supported	Supported with Aries
IPF 32-bit application	N/A	Supported
IPF 64-bit application	N/A	Supported

Mixed cluster configuration example:
Adding 1 IPF server to a 2-node PA cluster



Original 2-node PA cluster

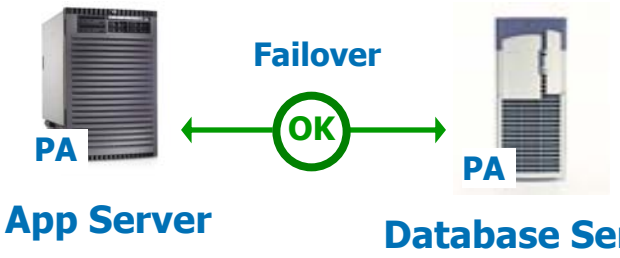




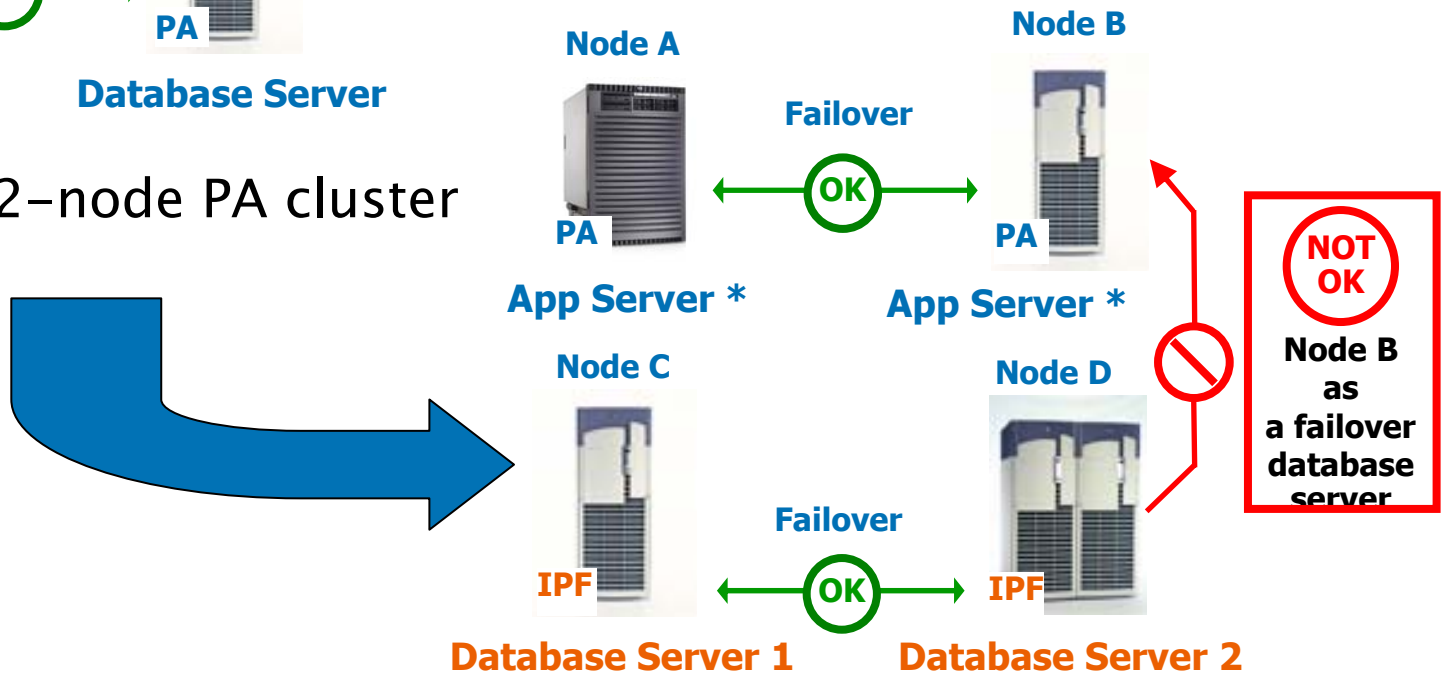
New 3-node mixed PA/IPF cluster

* not all app servers need failover functionality

Mixed cluster configuration example:
Adding 2 IPF servers to a 2-node PA cluster



Original 2-node PA cluster



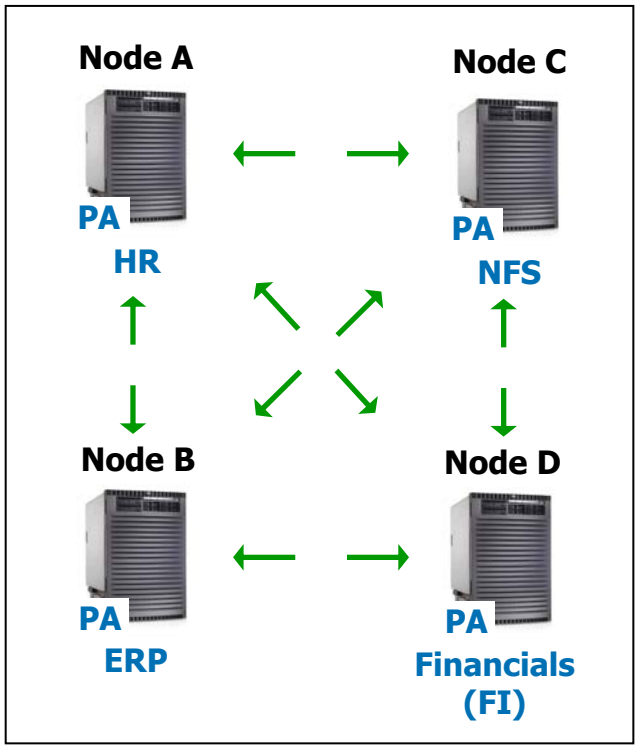
New 4-node mixed PA/IPF cluster

* not all app servers need failover functionality

Mixed cluster configuration example:

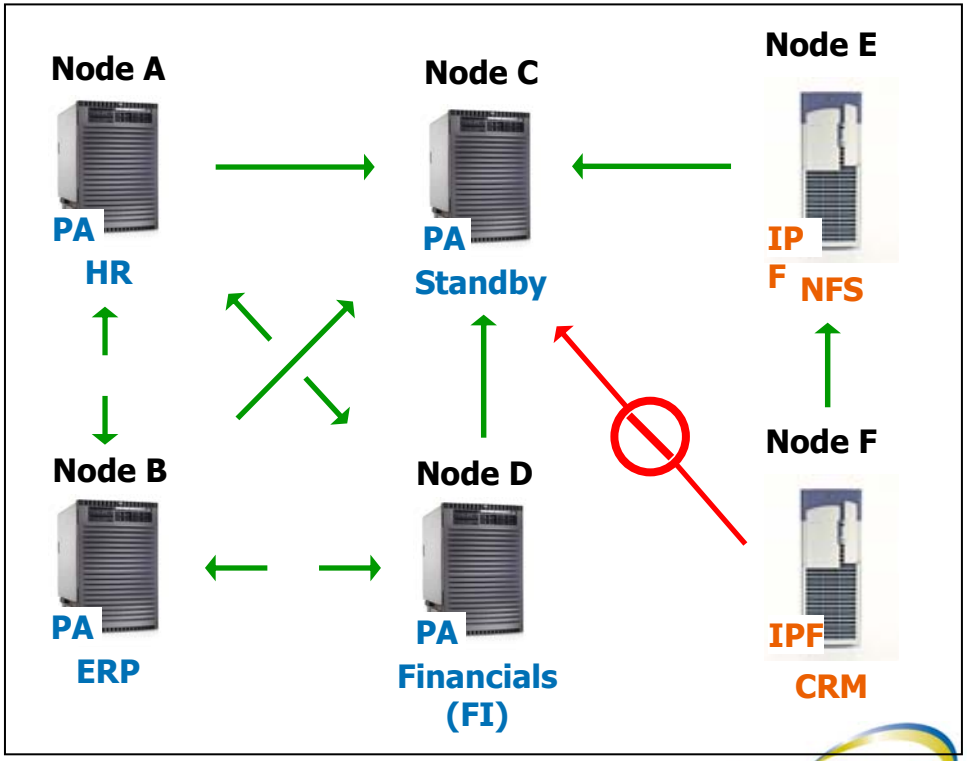
Adding one critical application and two IPF nodes to existing 4 node PA cluster

Existing 4-node PA cluster with 4 critical applications



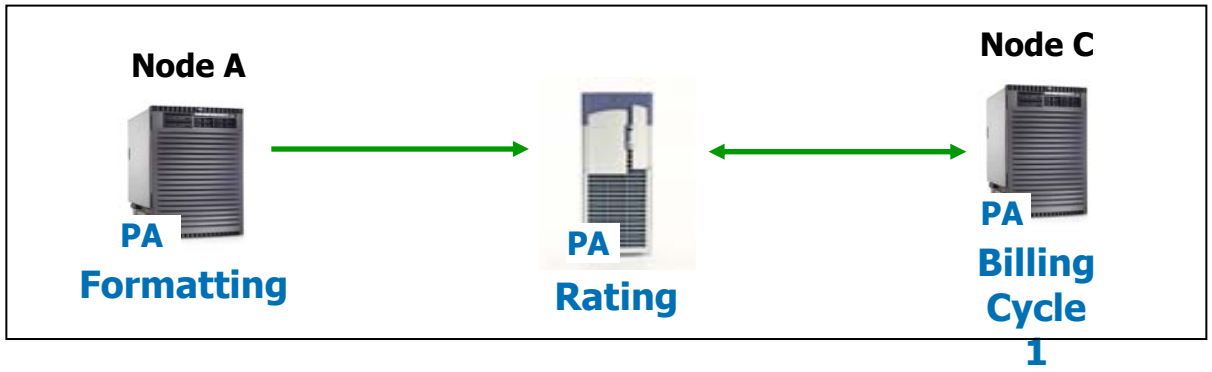
New mixed cluster with:

- **Additional new critical app (CRM)**
- **2 additional IPF nodes**
- **Maintained performance in failover situation**

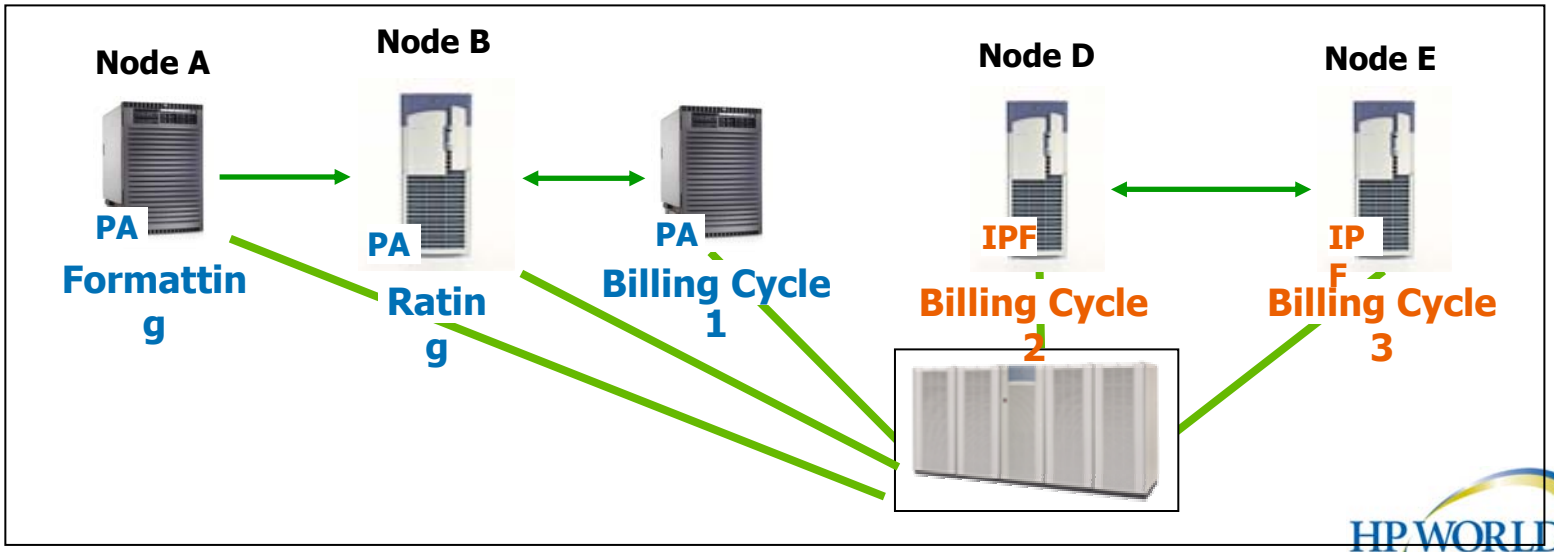


Mixed cluster configuration example:
Add 2 IPF nodes to PA telco cluster for more capacity

Existing 3-node PA telco billing cluster



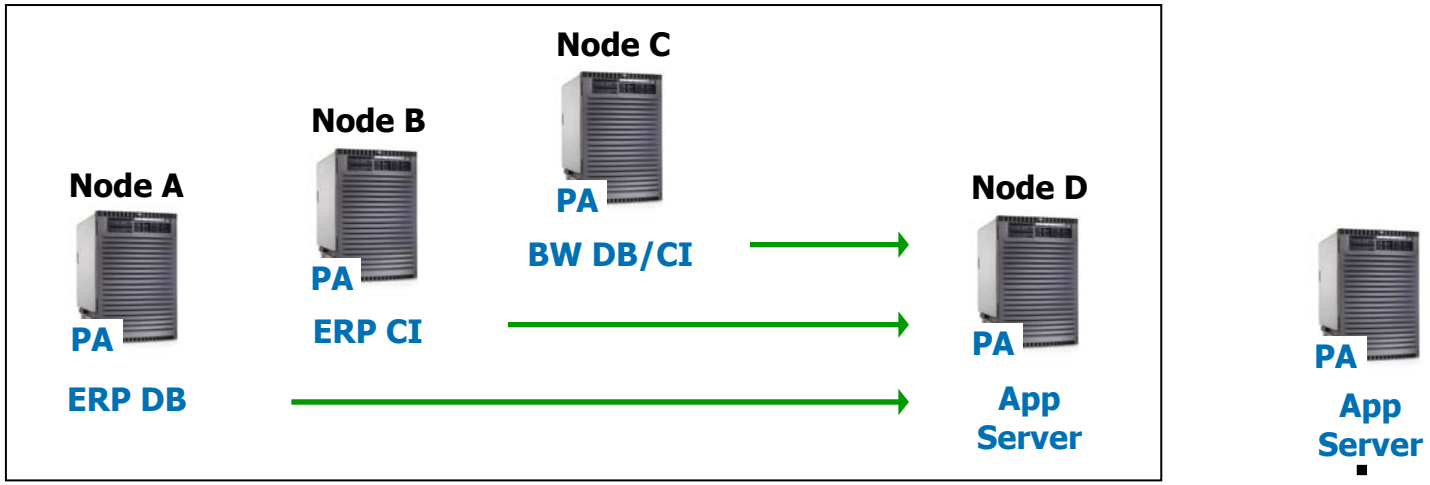
New 5-node PA/IPF cluster with 2 additional billing cycle systems



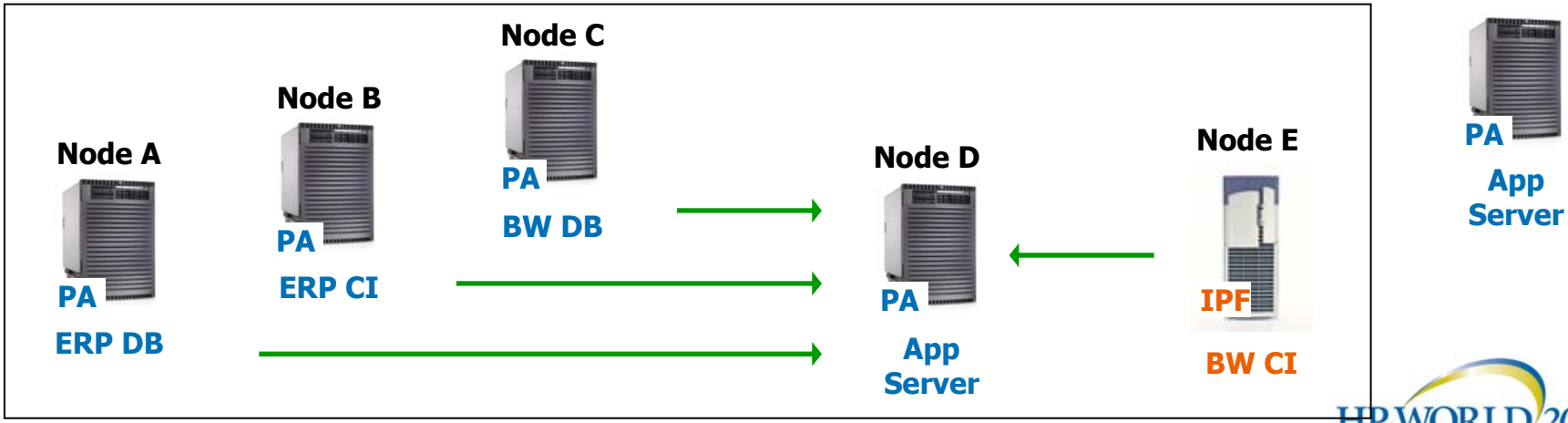
Transition from PA to IPF via mixed clusters

Add 1 IPF node to 4 node SAP PA cluster

Existing 4-node all-PA SAP cluster with ERP and Business Warehouse (BW) systems



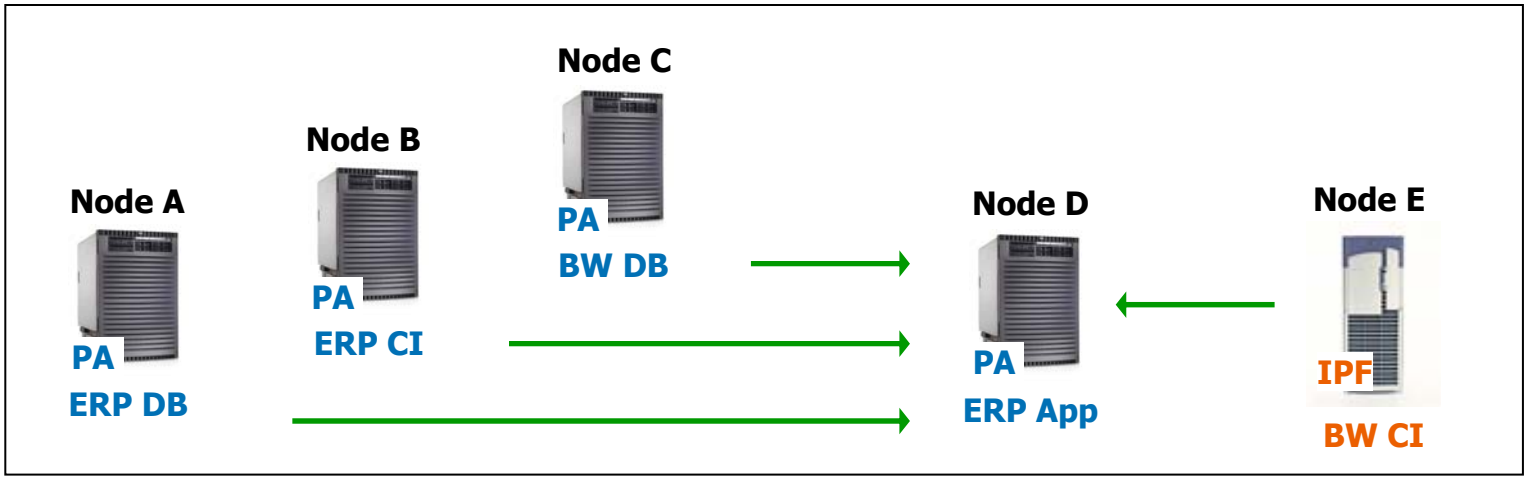
New 5-node PA/IPF SAP cluster with ERP and BW systems



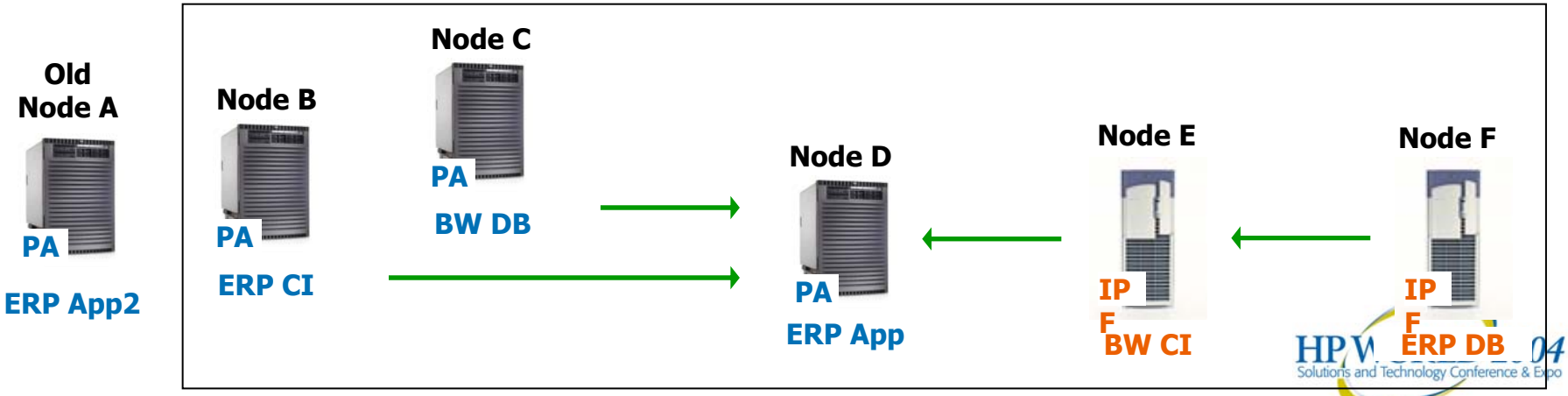
Transition from PA to IPF via mixed clusters

Add 2nd IPF node to 5 node SAP cluster

Existing 5-node PA/IPF SAP cluster with ERP and BW systems

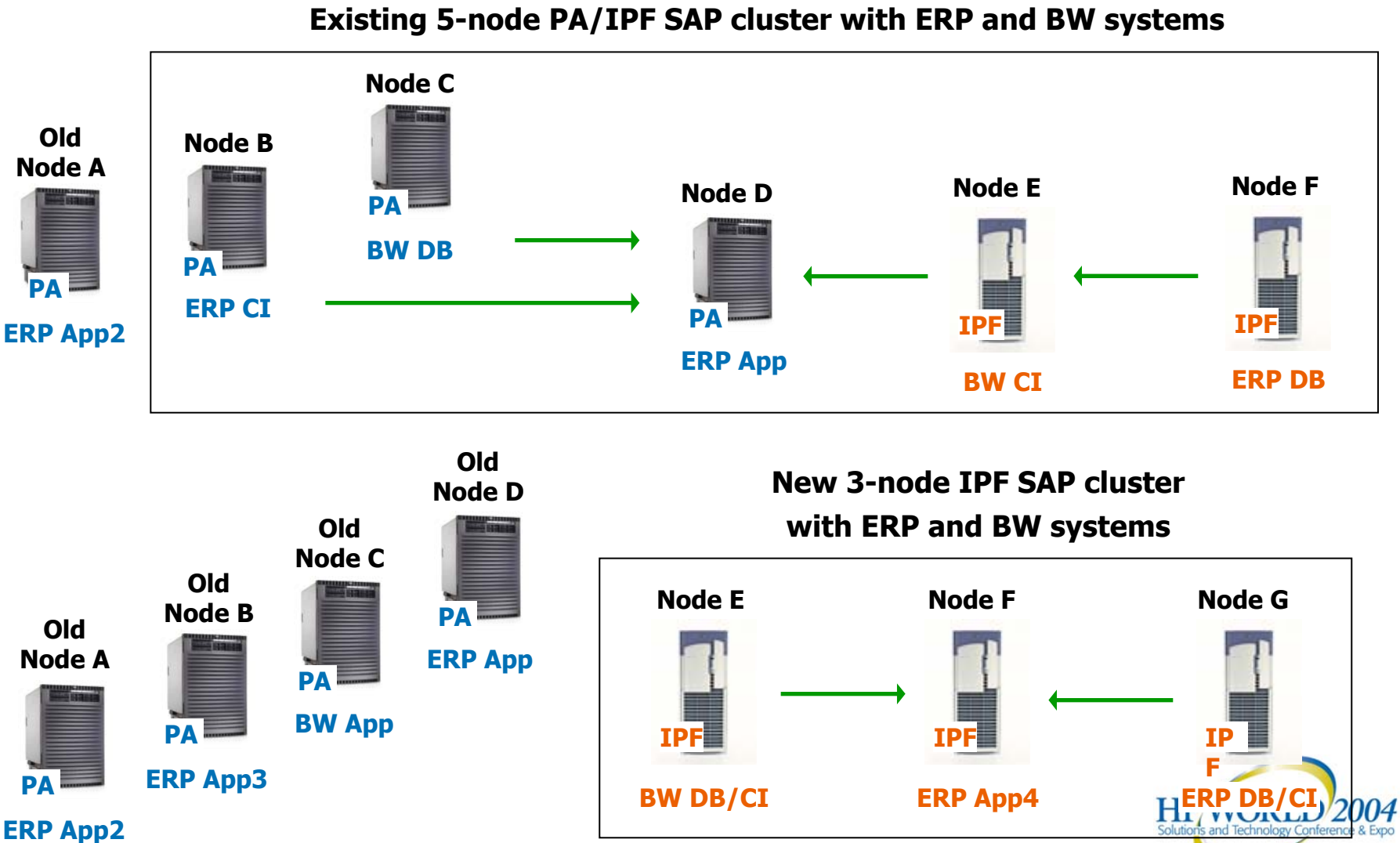


New 5-node PA/IPF SAP cluster with ERP and BW systems



Transition from PA to IPF via mixed clusters

Add 3rd IPF node and recycle all PA nodes as Apps



Summary

HP's strong heritage in clustering - high availability leadership

- Serviceguard (for hp-ux)
  - Widely installed, third-generation solution:
    - #1 Unix sales volume leader (IDC,2001)
    - Over 100,000 licenses sold since 1995
- Oracle Parallel Server
  - First vendor to deploy OPS in1995
  - #1 sales volume leader (Gartner, 4/02)
  - Jointly developed OPFS with Oracle in 1998 for fastest failover
- Oracle 9iRAC
  - Exclusive – value add from HP
  - HMP2 (fast node interconnect)
- Disaster Tolerant Products
  - Metrocluster and Continentalcluster
  - #1 ranking over IBM & Sun, more supported configurations, better integration with other middleware (Gartner, 2/02)
  - Hundreds of real disaster tolerant implementations

HP's strong heritage in clustering - high availability solutions

- SAP solutions w/ SG on HP-UX
  - #1 ranking for HA, (Gartner, 12'01)
  - SGeSAP & Somersault
  - Metrocluster for DR
- Business continuity solutions
  - HP/UX, Linux & Windows solutions
  - Integrated with Utility pricing and Sys Mgmt
- O/S integration
  - "HA in a box", integration with npars and vpars partitioning
- HA Reference Architectures
  - Integrated business availability solutions

Product and solution design criteria

- Data Integrity (#1 Priority)
  - Architected to ensure data integrity and prevent data corruption (prevent split brain)
- Availability
  - Most cluster configuration changes can occur online with the cluster and its applications running
- Performance
  - Constantly improving failover times for rapid failover of the application

For more information on HP's High Availability offerings …

- Visit our web sites at:

http://www.hp.com/go/ha

http://docs.hp.com/hpux/ha

Questions?

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

RECOMMENDED TRAINING VENUE FOR THE

HP Certified Professional


invent