

Intelligent Load Balancing with IBM Multi-site Workload Lifeline

Mike Fitzpatrick – mfitz@us.ibm.com
IBM Raleigh, NC

Thursday, August 11th, 4:30pm
Session: 9257

Trademarks, notices, and disclaimers



The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both:

- | | | | | |
|--|---|---|--|---|
| <ul style="list-style-type: none">Advanced Peer-to-Peer Networking®AlX®alphaWorks®AnyNet®AS/400®BladeCenter®Candle®CICS®DataPower®DB2 ConnectDB2®DRDA®e-business on demand®e-business (logo)e business (logo)®ESCON®FICON® | <ul style="list-style-type: none">GDDM®GDPS®Geographically Dispersed Parallel SysplexHiperSocketsHPR Channel ConnectivityHyperSwapi5/OS (logo)i5/OS®IBM eServerIBM (logo)®IBM®IBM zEnterprise™ SystemIMSInfiniBand ®IP PrintWayIPDSiSeriesLANDP® | <ul style="list-style-type: none">Language Environment®MQSeries®MVSNetView®OMEGAMON®Open PowerOpenPowerOperating System/2®Operating System/400®OS/2®OS/390®OS/400®Parallel Sysplex®POWER®POWER7®PowerVMPR/SMpSeries®RACF® | <ul style="list-style-type: none">Rational Suite®Rational®RedbooksRedbooks (logo)Sysplex Timer®System i5System p5System x®System z®System z9®System z10Tivoli (logo)®Tivoli®VTAM®WebSphere®xSeries®z9®z10 BCz10 EC | <ul style="list-style-type: none">zEnterprisezSeries®z/Architecturez/OS®z/VM®z/VSE |
|--|---|---|--|---|

* All other products may be trademarks or registered trademarks of their respective companies.

The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both:

- Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
- Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
- Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
- Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
- InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
- Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
- ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
- IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

Notes:

- Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
- IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
- All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
- This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
- All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
- Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
- Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Refer to www.ibm.com/legal/us for further legal information.



Agenda

- ❑ Current Disaster Recovery Solutions
- ❑ GDPS Active-Active Sites
- ❑ Multi-site Workload Lifeline
- ❑ Appendix: Configuration Statements



Disclaimer: All statements regarding IBM future direction or intent, including current product plans, are subject to change or withdrawal without notice and represent goals and objectives only. All information is provided for informational purposes only, on an “as is” basis, without warranty of any kind.

Multi-site Workload Lifeline

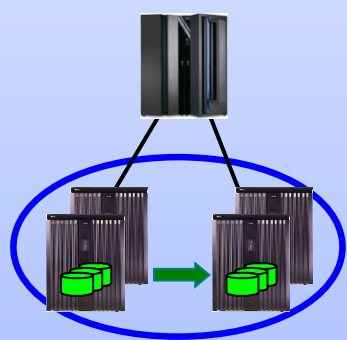
- ⇒ ***Current Disaster Recovery Solutions***
GDPS Active-Active Sites
Multi-site Workload Lifeline
Appendix: Configuration Statements

What are customers doing today ?

Continuous Availability of Data within a Data Center

Single Data Center
Applications remain active

Continuous access to data in the event of a storage subsystem outage

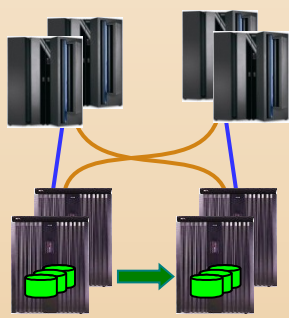


GDPS/HyperSwap Mgr
RPO=0 & RTO=0

Continuous Availability / Disaster Recovery within a Metropolitan Region

Two Data Centers
Systems remain active

Multi-site workloads can withstand site and/or storage failures

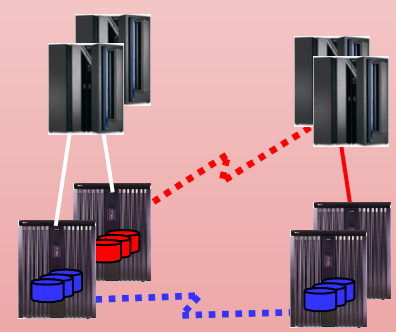


GDPS/PPRC
RPO=0 & RTO<4 hr

Disaster Recovery at Extended Distance

Two Data Centers
Rapid Systems Disaster Recovery with "seconds" of Data Loss

Disaster recovery for out of region interruptions

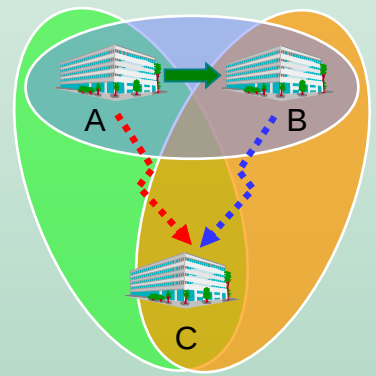


GDPS/GM & GDPS/XRC
RPO secs & RTO <1 hr

Continuous Availability Regionally and Disaster Recovery Extended Distance

Three Data Centers
High availability for site disasters

Disaster recovery for regional disasters



GDPS/MGM & GDPS/MzGM

How Much Interruption can your Business Tolerate?

Ensuring Business Continuity:

- Disaster Recovery
 - Restore business after an unplanned outage
- High-Availability
 - Meet Service Availability objectives e.g., 99.9% availability or 8.8 hours of down-time a year
- Continuous Availability
 - No downtime (planned or not)



Active/Active

Global Enterprises that operate across time-zones no longer have any 'off-hours' window. Continuous Availability is required.

What is the cost of 1 hour of downtime during core business hours?

Cost of Downtime by Industry	
Industry Sector	Loss per Hour
Financial	\$8,213,470
Telecommunications	\$4,611,604
Information Technology	\$3,316,058
Insurance	\$2,582,382
Pharmaceuticals	\$2,058,710
Energy	\$1,468,798
Transportation	\$1,463,128
Banking	\$1,145,129
Chemicals	\$1,071,404
Consumer Products	\$989,795

Source: Robert Frances Group 2006, "Picking up the value of PKI: Leveraging z/OS for Improving Manageability, Reliability, and Total Cost of Ownership of PKI and Digital Certificates."

Customer Requirements

- Want to shift focus from a failover model to a nearly-continuous availability model (RTO near zero)
- Access data from any site (unlimited distance between sites)
- No application changes
- Multi-sysplex, multi-platform solution
 - “Recover my business rather than my platform technology”
- Ensure successful recovery via automated processes (similar to GDPS technology today).
 - Can be handled by less-skilled operators
- Provide workload distribution between sites (route around failed sites, dynamically select sites based on ability of site to handle additional workload).
- Provide application level granularity
 - Some workloads may require immediate access from every site, other workloads may only need to update other sites every 24 hours (less critical data).
 - Current solutions employ an all-or-nothing approach (complete disk mirroring, requiring extra network capacity).

Multi-site Workload Lifeline

Current Disaster Recovery Solutions

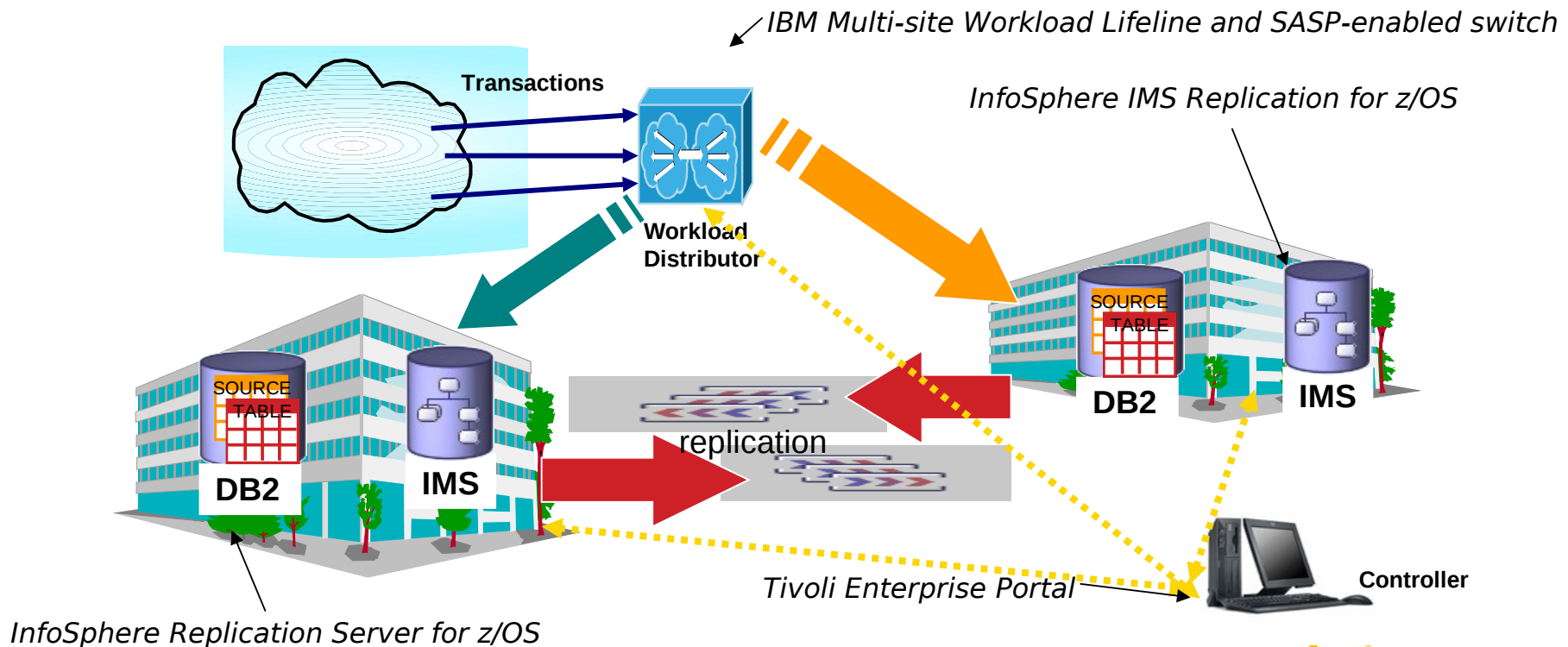
⇒ ***GDPS Active-Active Sites***

Multi-site Workload Lifeline

Appendix: Configuration Statements

GDPS Active-Active Sites – What is it?

- Two or more sites, separated by *unlimited* distances, running the same applications and having the same data to provide cross-site workload balancing and Continuous Availability / Disaster Recovery
- Paradigm shift: failover model => near continuous availability model



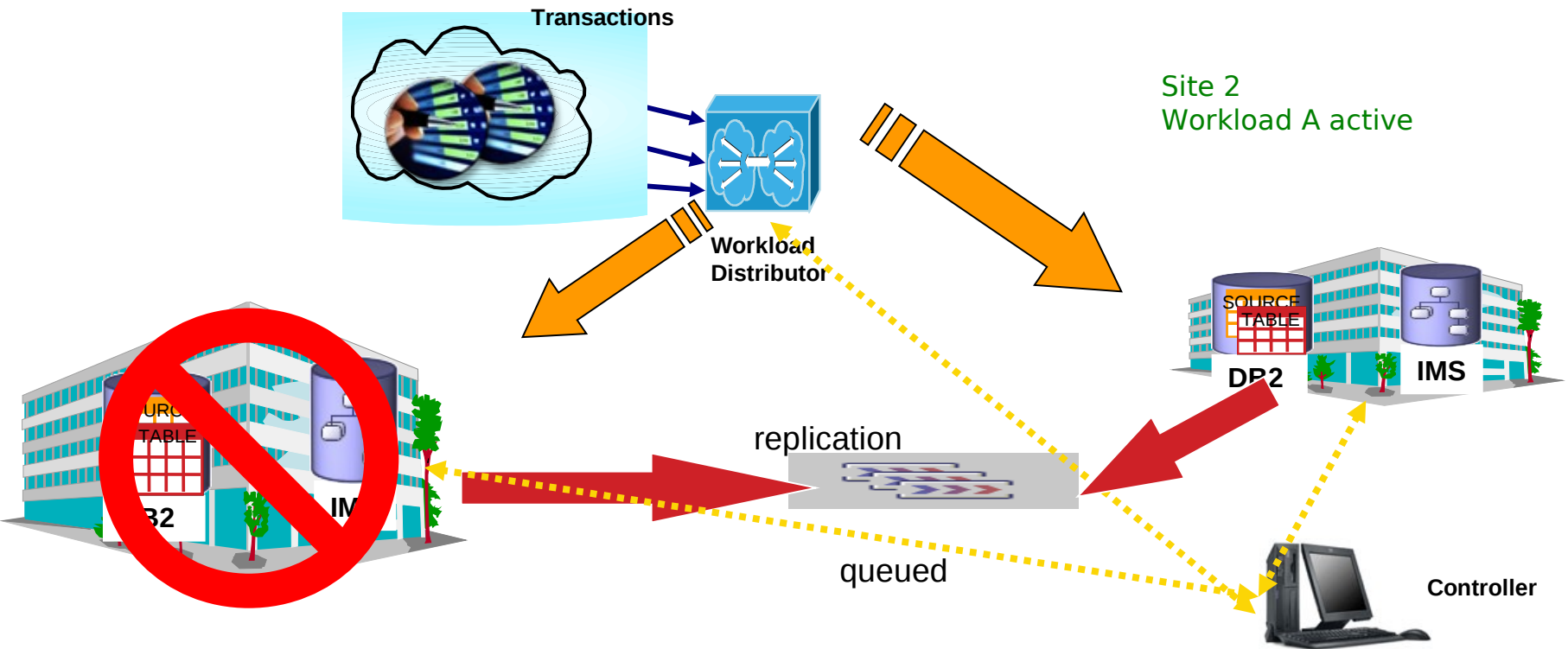
GDPS Active-Active Sites Configurations

- Configurations
 1. Active/Standby
 2. Active/Query (future)
- A configuration is specified on a workload basis
- A workload is the aggregation of these components
 - **Software:** applications (e.g., COBOL program) and the middleware run time environment (e.g., CICS region & DB2 subsystem)
 - **Data:** related set of objects that must preserve transactional consistency (e.g., DB2 Tables)
 - **Network connectivity:** one or more TCP/IP addresses & ports (e.g., 10.10.10.1:80)

Active/Standby Configuration

Site 1
Workload A active

Site 2
Workload A standby



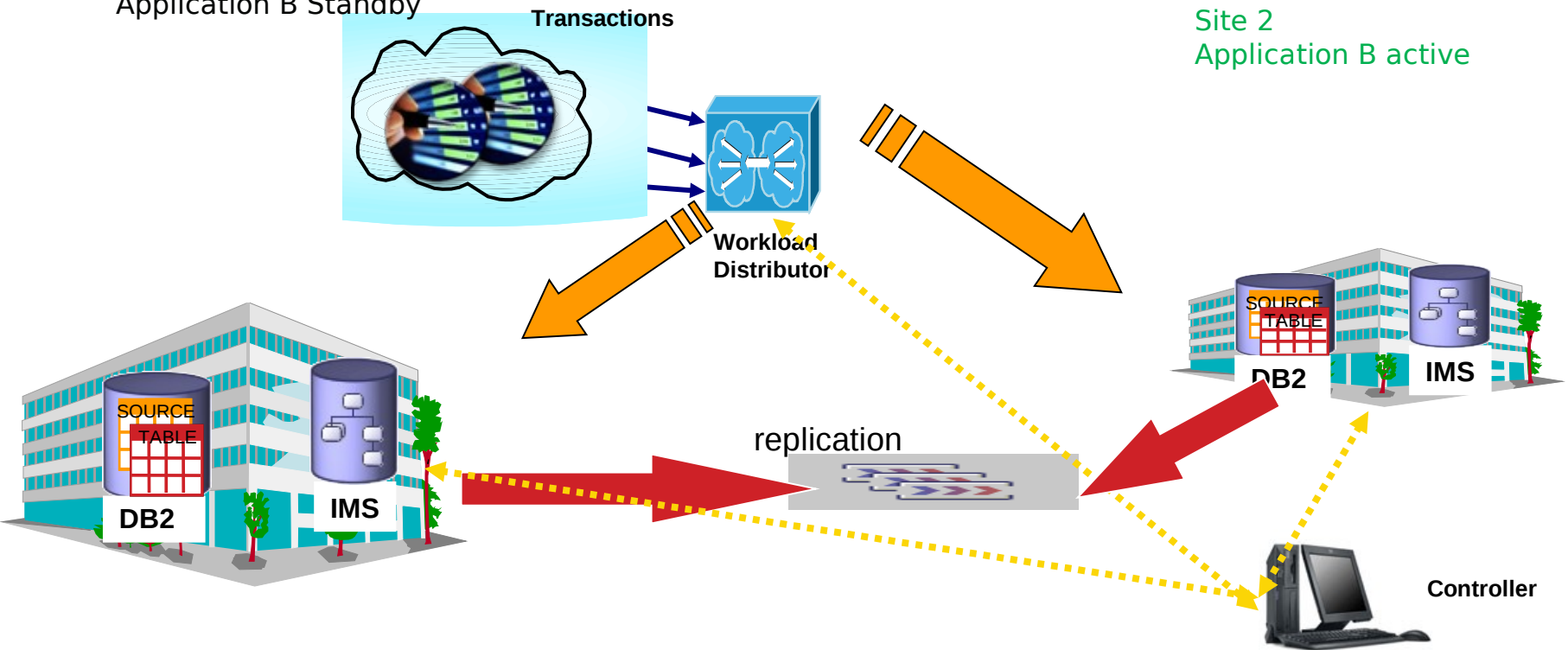
Active/Standby Configuration (multiple workloads)

Site 1
Application A active

Site 1
Application B Standby

Site 2
Application A standby

Site 2
Application B active



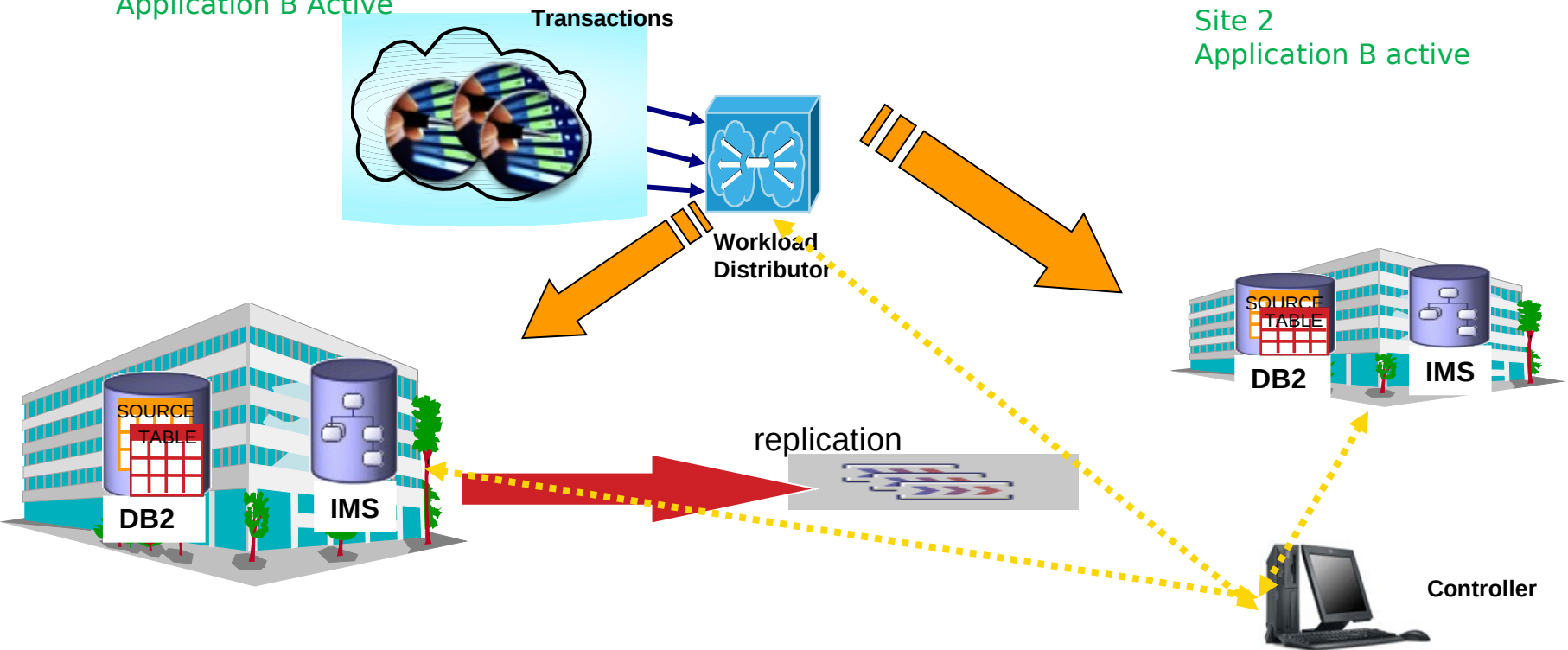
Active/Query Configuration

Site 1
Application A active

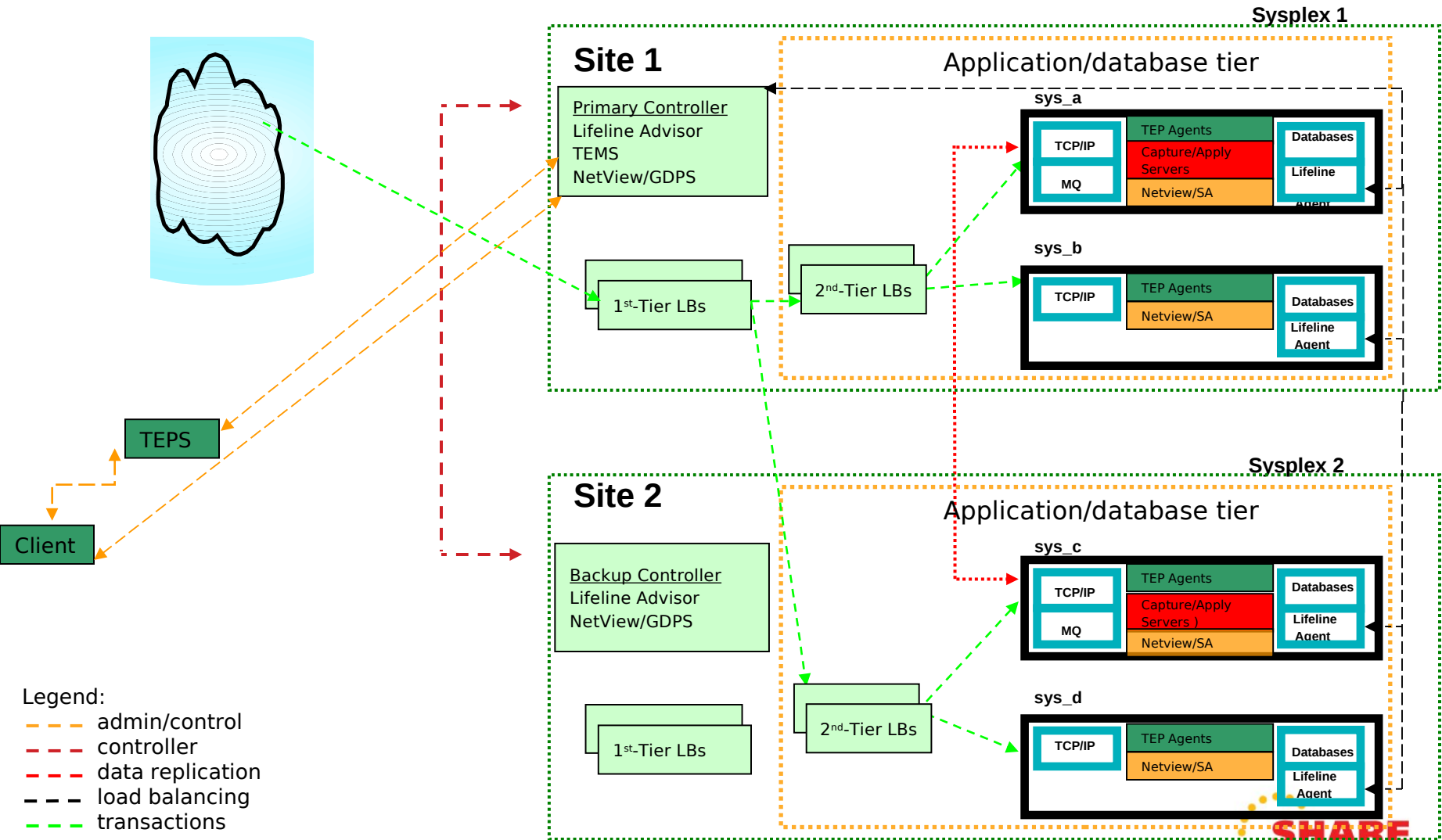
Site 1
Application B Active

Site 2
Application A standby

Site 2
Application B active



Active/Active Sites Structure



Multi-site Workload Lifeline

Current Disaster Recovery Solutions

GDPS Active-Active Sites

⇒ ***Multi-site Workload Lifeline***

Appendix: Configuration Statements

Benefits of intelligent Load Balancing

▪ Performance

- Improves response time (distribute new connections to server applications best able to handle additional work)

▪ Availability

- If one server instance goes down, existing connections to it break, but new connections can be established with remaining server instances

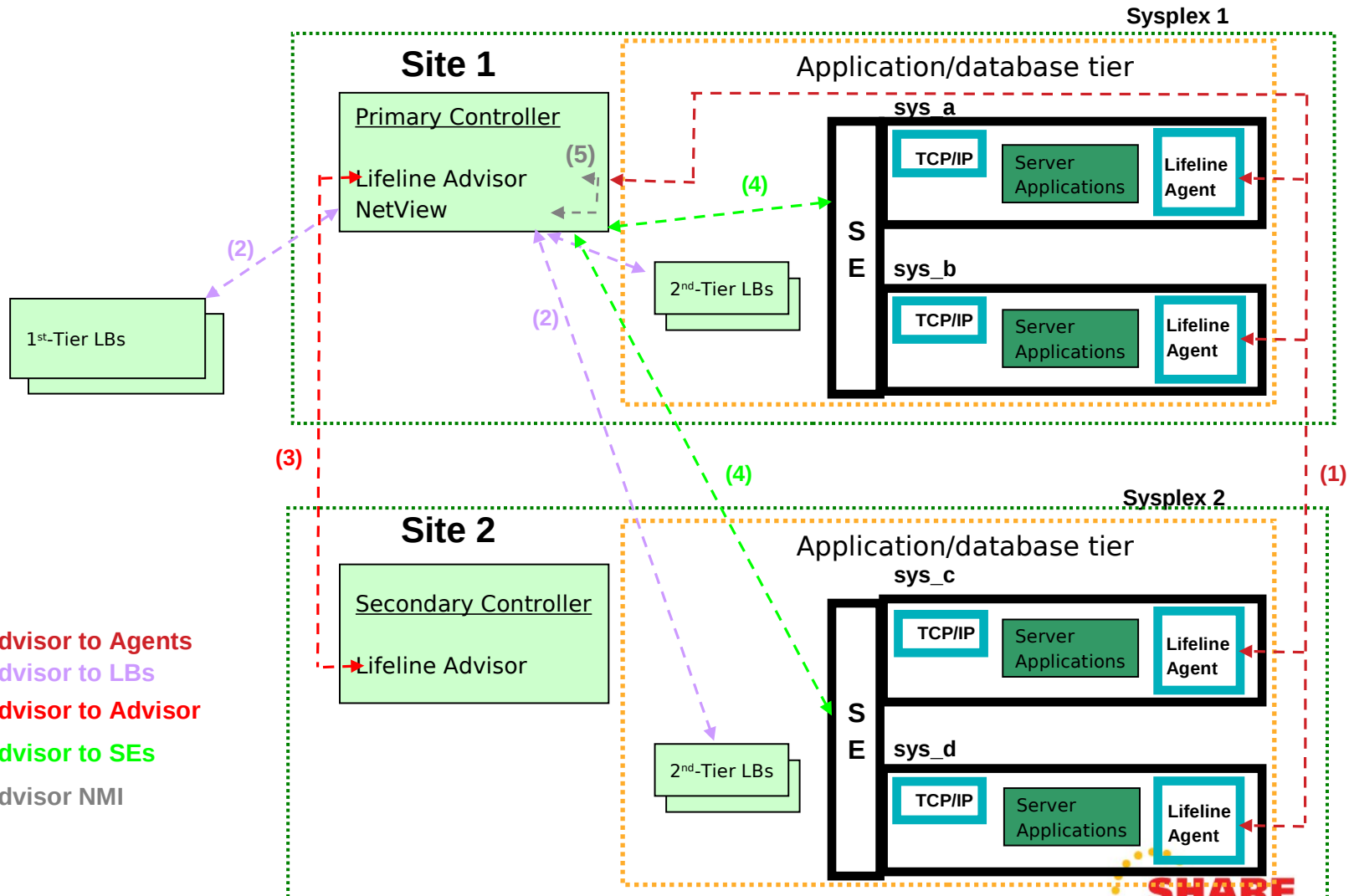
▪ Scalability

- More server instances can be added on demand (horizontal growth)

GDPS Active-Active Sites load balancing requirements

- Ability to distribute workloads between sites (and route around failed sites)
 - Based on capacity/health of sites and server application instances within a site
- Ability to detect workload or site failures
- Ability to switch workloads from one site to another site
 - Perform “graceful” takeover for site maintenance
- Ability to maintain workload configuration states in event of a workload manager failure
 - Keep a peer workload manager in sync with workload states
- Ability to dynamically add/modify workloads
- Ability to surface routing recommendations to network management agents

Workload Lifeline Structure



Workload Lifeline role in Active/Standby Environment

- Advisor provides distribution recommendations to multiple tiers of load balancers
 - Server-specific WLM metrics and Communications Server weights provided by Agents running in all LPARs across both sites are used to build recommendations
- Site recommendations to 1st-tier load balancers
 - Direct 1st-tier load balancers to route new connections for a workload to a 2nd-tier load balancer within a site (using SASP – see RFC 4678)
 - Site selection determined by where the workload is currently active
- Server application recommendations to 2nd-tier load balancers
 - Direct 2nd-tier load balancers to route new connections for a workload to specific server applications within the site (using SASP)
 - Server application selection determined by recommendations provided by the Agents within the site
 - Sysplex Distributor may assume role of 2nd-tier load balancer
 - No server application recommendations provided by Advisor in this case

Workload Lifeline distribution recommendations

- Agents provide relative weights per server application instance
 - WLM weight
 - Server-specific WLM recommendations
 - Reflects how much displaceable capacity is available on the target system at the importance level of the server application
 - Communications Server weight
 - This weight is calculated based on the availability of the actual server instances (are they up and ready to accept workload) and how well TCP/IP and the individual server instances process the workload that is sent to them.
 - Prevent stalled server instance from being sent more work (accepting no new connections and new connections are being dropped due to backlog queue full condition)
 - Proactively react to server instance that is getting overloaded (accepting new connections, but size of backlog queue increases over time approaching the max backlog queue size)
- Advisor uses relative weights of all the server application instances for a workload are available/healthy within a site to determine whether a workload failure has occurred

Workload Lifeline in Active/Standby Environment ...

- Advisor provides ability to group different server applications into a workload
 - Distinguish different workloads and perform different distribution decisions based on the workload (direct each workload to its Active site)
- Advisor responsible for detecting workload failures
 - Monitor the capacity of LPARs within a workload's Active site and availability/health of the server applications that make up the workload
 - Ability to dynamically switch a workload to the alternate site after detecting a failure
- Advisor responsible for detecting site failures
 - Monitor the availability/reachability of the LPARs that make up the site
 - Communication with Agents active on the LPARs verifies IP network connectivity to the site
 - Communication with Support Elements (SE) over HMC network verifies LPAR status
 - Ability to dynamically switch all workloads to the alternate site after detecting a failure

Workload Lifeline in Active/Standby Environment ...

- **Advisor communicates with a peer Advisor**
 - Shares workload state information
 - A workload can be inactive
 - A workload can be active to a specific site
 - Peer Advisor takes over responsibilities in the event the primary Advisor fails
- **Advisor provides graceful movement of a workload to an alternate site (a 'planned' failure)**
 - Prevents new connections for the workload from being distributed to the Active site
 - Terminates any existing connections being distributed to the Active site
 - Reroutes new transactions to the alternate site
- **Advisor has ability to dynamically add or modify existing workloads to an active configuration**
 - Allows changes without recycling the Advisor
- **Advisor provides Network Management Interface (NMI)**
 - Surface workload states, distribution recommendations, and component information to network management agents

Workload Lifeline in Active/Standby Environment ...



- Agents communicate with a Communications Server TCPIP stack
 - Extracts information about available server applications and server application health via documented interfaces

Advisor to Load Balancer communication

- **Server Application State Protocol (SASP)**
 - Open protocol documented in RFC4678
 - Provides a mechanism for workload managers to give distribution recommendations to load balancers
 - Does not handle the transport or actual distribution of work, only provides recommendations
- **1st-tier load balancers register groups it is interested in load balancing**
 - Each group designates a list of 2nd-tier load balancers (members) it will distribute connections to (either another SASP-enabled load balancer or a Sysplex Distributor node)
 - Identified by protocol (TCP/UDP), IP addresses (IPv4/IPv6) of the 2nd-tier load balancers, and the port number used by the server applications that the 2nd-tier load balancer will be load balancing
 - Advisor uses its lb_id_list to verify whether a load balancer is allowed to connect
 - Advisor uses its cross_sysplex_list configuration statement to map groups to a workload

Advisor to Load Balancer communication...

- **2nd-tier load balancers register Groups it is interested in load balancing**
 - Each group designates a list of server applications (members) to be load balanced
 - Identified by protocol (TCP/UDP), IP addresses (IPv4/IPv6) of the target systems the server applications reside on, and the port number used by the server applications
- **Load balancers can request to receive distribution recommendations using two possible methods**
 - Load balancer will periodically “pull” member distribution recommendations from the Advisor
 - Advisor will periodically “push” member distribution recommendations to the load balancer
 - Can be configured to only “push” changed information about members

Advisor to Agent communication

- **Internal protocol for communication**
- **Agents connect to Advisor**
 - Each Agent registers its system name, site name (i.e. sysplex name), and LPAR name
 - Advisor uses its agent_id_list configuration statement to verify Agent is allowed to connect
 - Advisor uses its cross_sysplex_list configuration statement to verify Agent resides in valid site
- **Advisor sends information about all members it wants the Agent to monitor**
 - The IP address, port number, and protocol for all server applications that were registered as group members by 2nd-tier load balancers
 - The IP address and port number for target server applications being distributed by the Sysplex Distributors node that were registered as group members by 1st-tier load balancers

Advisor to Agent communication...

- Agent sends periodic updates about system to Advisor
 - List of active members (server applications) active on its system
 - Server WLM recommendation for each member
 - Communications Server health on its system
- Advisor sends requests to reset connections to Agents
 - In response to a DEACTIVATE command, Advisor sends a list of server applications (that make up a workload) to Agents to direct them to reset any active connections for these server applications

Advisor to Advisor communication

- **Internal protocol**
- **Peer Advisor (secondary) connects to Advisor (primary)**
 - Peer Advisor registers its system name and LPAR name
 - Primary Advisor uses its `advisor_id_list` configuration statement to verify Advisor is allowed to connect
- **Primary Advisor sends information about its active configuration to peer Advisor**
 - Verifies configurations are identical between the two Advisors in case the peer Advisor needs to become primary Advisor
- **Primary Advisor sends workload state changes to peer Advisor**
 - Primary Advisor builds a list of commands that will QUIESCE or ACTIVATE workloads based on their current states
 - Peer Advisor replays this list of commands in event it becomes primary Advisor so that all workloads remain in the same state

Advisor to Support Element (SE) communication

- **Base Control Program Internal Interface (BCPii)**
 - Documented IBM protocol
 - Allows communication between LPAR where Advisor is active and all interconnected Central Processor Complexes (CPCs)
 - Each CPC can be queried to extract list of LPARs and their status
 - Communication occurs over a Hardware Management Console (HMC) network
 - Typically resides on a different physical network than network used for IP communication
- **Advisor uses BCPii address space as a bridge to the SEs**
 - New address space shipped in V1R11
 - Advisor uses LPAR names received from peer Advisor and Agents to build list of LPARs to query status information

Advisor to Network Management App communication

- **Network Management Interface (NMI)**
 - Documented interface
- **Advisor creates AF_UNIX socket and accepts connections from network management applications**
 - Supplies workload state information, site information, load balancer group registrations, connected load balancers and Agents and peer Advisor, and distribution recommendations for server applications

Key Advisor Display commands

- **MODIFY advproc,DISPLAY,ADVISOR,DETAIL**
 - When issued on the primary Advisor, displays the role of the Advisor, the connected load balancers (and whether it is a 1st-tier or 2nd-tier), the connected Agents (including system and site name where the Agents are active), and the connected peer Advisor (including the system name where the peer is active)
 - When issued on the peer Advisor, displays the role of the Advisor and the connected primary Advisor (including the system name where the primary is active)
- **MODIFY advproc,DISPLAY,CONFIG**
 - Displays the current active configuration for the Advisor

Key Advisor Display commands...

- **MODIFY advproc,DISPLAY,LB,DETAIL**
 - Displays the connected load balancers, including the list of groups registered by the load balancer, the members within each group, and the distribution recommendations provided for each member
- **MODIFY advproc,DISPLAY,WORKLOAD,DETAIL**
 - Displays the status of all defined workloads, including the status of all the server applications that make up the workload

Key Advisor State Change commands

- **MODIFY advproc,ACTIVATE,WORKLOAD=...,SITE=...**
 - Signals the Advisor to direct 1st-tier load balancers to distribute new connections for the specified workload to the requested site
- **MODIFY advproc,DEACTIVATE,WORKLOAD=...**
 - Signals the Advisor to direct Agents on the site where the specified workload was last active to reset any existing connections for this workload
- **MODIFY advproc,QUIESCE,WORKLOAD=...**
 - Signals the Advisor to direct 1st-tier load balancers to stop distributing new connections for the specified workload to any site
- **MODIFY advproc,REFRESH**
 - Signals the Advisor to reread its configuration file and apply any updates to its active configuration
- **MODIFY advproc,TAKEOVER**
 - Signal the peer Advisor to take over primary Advisor responsibilities from the current primary Advisor

Key Agent Display commands

- **MODIFY ageproc,DISPLAY,CONFIG**
 - Displays the current active configuration for the Agent
- **MODIFY ageproc,DISPLAY,MEMBERS,DETAIL**
 - Displays information about each of the server applications this Agent was asked to monitor, including whether the server application exists, the jobname of the server application, and current state of the server application

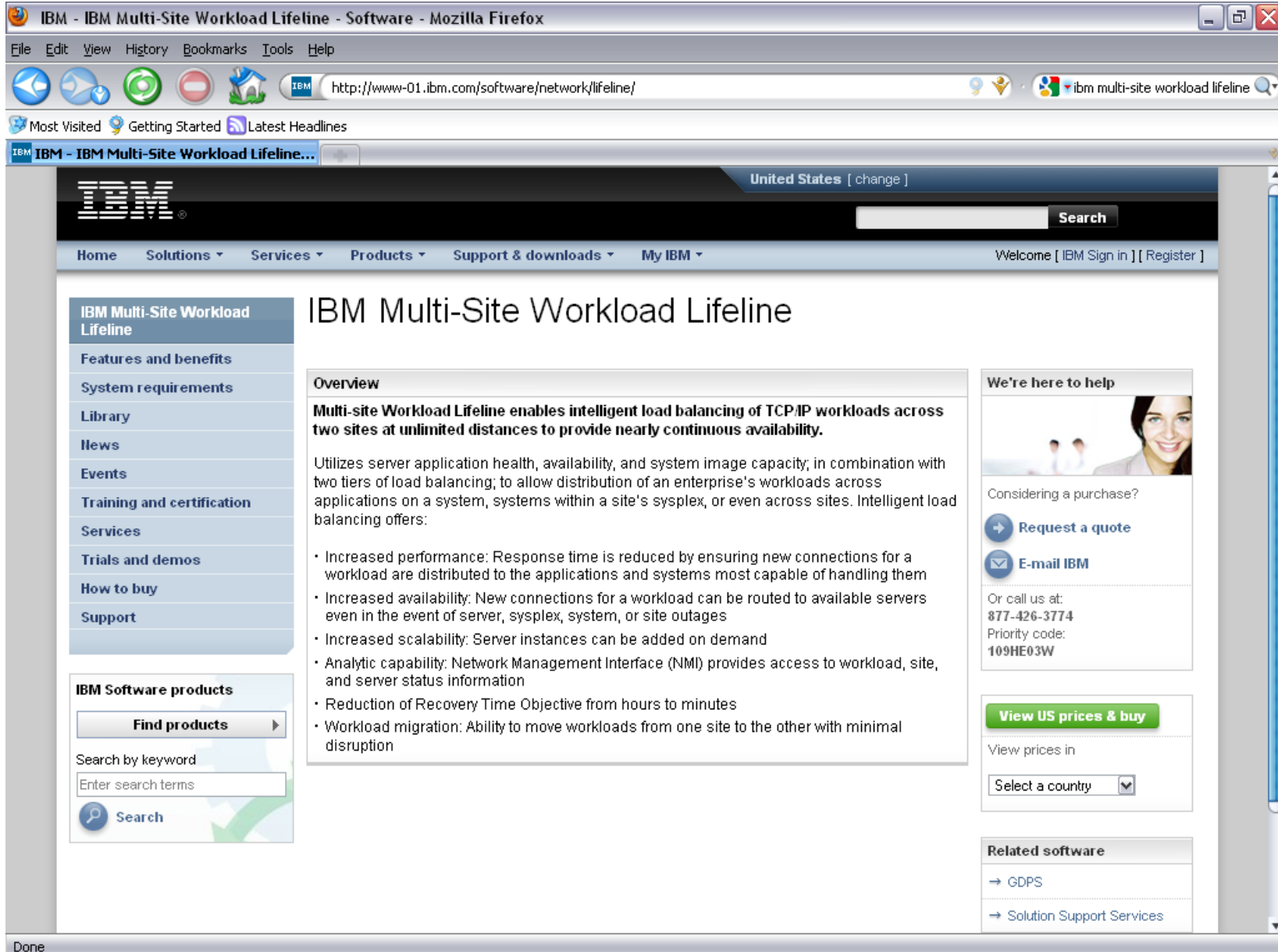
Key Agent State Change commands

- **MODIFY ageproc,ENABLE,...**
 - Signals the Agent to enable server applications (make them available to be load balanced to)
 - Server applications bound to a distributable dynamic VIPA must be enabled using the VARY TCPIP,,SYSPLEX,RESUME command
- **MODIFY ageproc,QUEISCE,...**
 - Signals the Agent to quiesce server applications (make them unavailable to be load balanced to)
 - Server applications bound to a distributable dynamic VIPA must be quiesced using the VARY TCPIP,,SYSPLEX,QUIESCE command

Debugging

- All debugging information recorded in syslogd
 - Requires the syslogd daemon be configured and started
- Enable debugging during startup
 - `debug_level` statement in both Advisor and Agent configuration
- Dynamically enable, disable, change debugging while active
 - **MODIFY advproc,DEBUG,LEVEL=...** for Advisor
 - **MODIFY ageproc,DEBUG,LEVEL=...** for Agent
- Display current debug level
 - **MODIFY advproc,DISPLAY,DEBUG** for Advisor
 - **MODIFY ageproc,DISPLAY,DEBUG** for Agent
- Default level traces errors, warnings, and commands

For more information...



The screenshot shows a Mozilla Firefox browser window with the address bar displaying `http://www-01.ibm.com/software/network/lifeline/`. The page content includes the IBM logo, a navigation menu with items like Home, Solutions, Services, Products, Support & downloads, and My IBM. The main heading is "IBM Multi-Site Workload Lifeline". A left sidebar lists various product details such as Features and benefits, System requirements, Library, News, Events, Training and certification, Services, Trials and demos, How to buy, and Support. The main content area has an "Overview" section with a bolded summary: "Multi-site Workload Lifeline enables intelligent load balancing of TCP/IP workloads across two sites at unlimited distances to provide nearly continuous availability." Below this is a paragraph describing the technology and a bulleted list of benefits including increased performance, availability, scalability, analytic capability, and workload migration. On the right, there is a "We're here to help" section with a "Request a quote" button, an "E-mail IBM" button, and contact information (877-426-3774, Priority code: 109HE03W). A "View US prices & buy" section includes a "View prices in" dropdown menu set to "Select a country". At the bottom right, a "Related software" section lists "GDPS" and "Solution Support Services". The browser's status bar at the bottom left shows "Done".

Multi-site Workload Lifeline

Current Disaster Recovery Solutions

GDPS Active-Active Sites

Multi-site Workload Lifeline

⇒ ***Appendix: Configuration Statements***

Key Advisor configuration statements

- **advisor_id_list**
 - List of IP addresses used by primary Advisor to determine which peer Advisors are permitted to connect to it
 - Used by peer Advisor to select a source IP address when connecting to primary Advisor
- **agent_id_list**
 - List of IP addresses used by the Advisor to determine which Agents are permitted to connect to it
- **cross_sysplex_list**
 - Specifies the IP address of the 2nd-tier load balancer, the site name for that load balancer, the port number of the server application used for the workload, and the workload name
 - Used by the Advisor to map 1st-tier load balancer group registrations with workload names
 - Used by the Advisor to validate the sites where connected Agents reside

Key Advisor configuration statements...

- `failure_detection_interval`
 - Time interval used by Advisor to determine how long to wait before declaring a workload or site failure
- `lb_connection_v4`
 - Specifies the IPv4 address bound by the Advisor to accept connections from load balancers, Agents, and peer Advisor
 - Recommended to be defined as a VIPARANGE dynamic VIPA so that a peer Advisor can take over the dynamic VIPA (when taking over as primary Advisor) without requiring any load balancer or Agent configuration changes
- `lb_connection_v6`
 - Specifies the IPv6 address bound by the Advisor to accept connections from load balancers, Agents, and peer Advisor
 - Recommended to be defined as a VIPARANGE dynamic VIPA so that a peer Advisor can take over the dynamic VIPA (when taking over as primary Advisor) without requiring any load balancer or Agent configuration changes

Key Advisor configuration statements...

- `lb_id_list`
 - List of IP addresses used by the Advisor to determine which load balancers are permitted to connect to it
- `update_interval`
 - Time interval communicated to the Agents to specify how frequently the Advisor should be updated with server application metrics

Key Agent configuration statements

- **advisor_id**
 - The IP address used by the Agent as the destination IP address when connecting to the Advisor
 - Must match the IP address specified in the `lb_connection_v4` or `lb_connection_v6` statement
- **host_connection**
 - The IP address used by the Agent as the source address when connecting to the Advisor
 - Must match one of the IP addresses specified in the `agent_id_list` statement